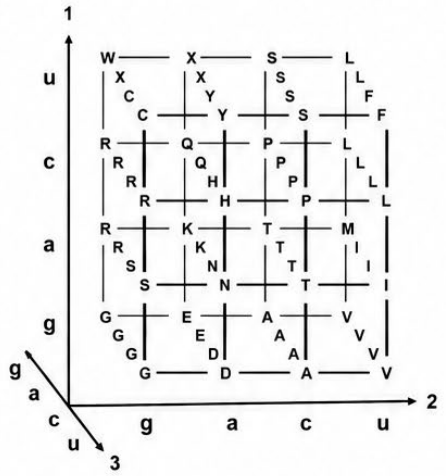
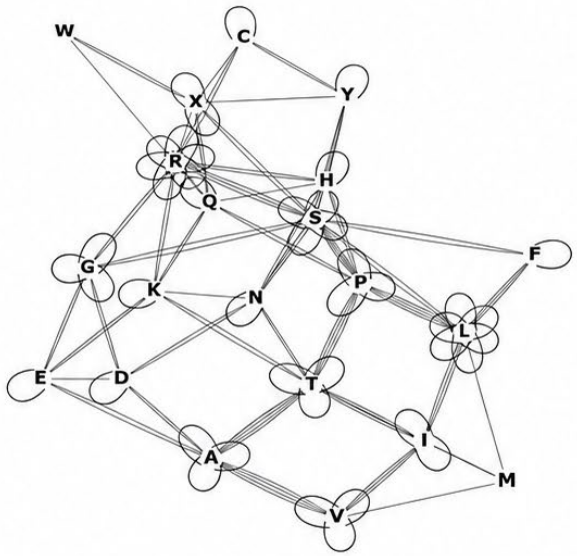
Mathematics of the Genetic Code





Douglas C. Youvan

Mathematics of the Genetic Code

Douglas C. Youvan

Copyright

Copyright © 2026 Douglas C. Youvan

All rights reserved.

First Kindle edition, 2026.

No part of this book may be reproduced, distributed, transmitted, stored, or republished in any form or by any means without the prior written permission of the author, except for brief quotations used in reviews, scholarly discussion, or other uses permitted by law.

This book presents mathematical, computational, physicochemical, and graph-theoretic analyses of the standard RNA genetic code. The genetic code is represented using RNA notation, with U for uracil. Historical DNA-based figures may retain T for thymine when necessary to preserve the original analysis. Translation stop is represented by X.

Scientific interpretations are distinguished from exact mathematical results. The mathematical derivations, computational procedures, and numerical values are presented for scholarly examination and reproducibility. Nothing in this book constitutes medical advice.

Table of Contents

Front Matter

Copyright

Table of Contents

List of Figures

List of Tables

Notation and Conventions

Abstract

Preface: The Genetic Code as a Mathematical Object

Part I

The Genetic Code as a Complete Finite Relation

Chapter 1

The Standard Genetic Code

1.1 A Complete Lookup Relation
1.2 The Twenty-One Translated Classes
1.3 Codon Fibers
1.4 Degeneracy Is Not Geometry
1.5 Synonymous Codons and Output Preservation
1.6 Completeness and Standardness
1.7 A Finite Natural Lookup System
1.8 Code, Table, and Physical Implementation
1.9 From the Lookup Table to the Matrix

Chapter 2

The Complete 64 × 12 Binary Codon Matrix

2.1 One-Hot Positional Encoding
2.2 Exact Balance of the Complete Codon Set
2.3 The Gram Matrix
2.4 The Constant Direction
2.5 Within-Position Contrast Directions
2.6 Redundancy Among the Three Positional Blocks

2.7 Rank and Nullity
2.8 The Meaning of Rank Ten
2.9 Input Architecture Before Translation

Chapter 3

The Exact Moore–Penrose Pseudoinverse

3.1 Why a Pseudoinverse Is Required
3.2 The Four Moore–Penrose Conditions
3.3 Closed-Form Pseudoinverse
3.4 The Two Exact Entry Values
3.5 Column Structure
3.6 Row Structure
3.7 Constant Codon Fields
3.8 Gauge Freedom
3.9 The Minimum-Norm Gauge
3.10 Projection into Codon Space
3.11 The Hamming-Distance Projector
3.12 The Coefficient-Space Projector
3.13 Direct Verification of the Moore–Penrose Conditions
3.14 What the Exact Formula Does Not Depend On
3.15 Why the Result Matters

Part II

Nucleotide-Determined Physicochemical Organization

Chapter 4

Hydropathy as a Codon-Level Property Field

4.1 From Amino-Acid Values to a Codon Field
4.2 The Question Asked by the Projection
4.3 Historical Origins of the Calculation
4.4 The Historical 61-Sense-Codon Result
4.5 The Complete 64-Codon Formulation
4.6 The Dominant Second-Position Polarity
4.7 Comparison with the Historical Result
4.8 Why the Third-Position Effects Are Small
4.9 First-Position Refinement

4.10 What the Coefficients Mean
4.11 What the Coefficients Do Not Show

Chapter 5

Reconstructing Amino-Acid Hydropathy

5.1 Projection Through the Exact Pseudoinverse
5.2 Codon-Level Reconstruction
5.3 Synonymous Codons Need Not Reconstruct Identically
5.4 Averaging Within Synonymous Fibers
5.5 Empirical and Reconstructed Hydropathy
5.6 Regression Is Not Causation
5.7 Codon-Level Projection Strength
5.8 Residual Structure
5.9 Fiber Geometry and Reconstruction Error
5.10 Additive Information and Unrepresented Interactions

Chapter 6

Hydropathy and Residue Volume

6.1 A Two-Property Representation
6.2 The Four Second-Position Codon Families
6.3 NUN and Hydrophobic Residues
6.4 NAN and Hydrophilic Residues
6.5 NCN and Lower Residue Volume
6.6 The Broader NGN Distribution
6.7 Hydropathy–Volume Segregation
6.8 Why the Middle Nucleotide Matters

Chapter 7

Membrane Hydropathy from the Second Codon Position

7.1 Origins in Membrane-Protein Engineering
7.2 Conventional Kyte–Doolittle Profiling
7.3 The Historical Youvan T–A Method
7.4 The RNA U–A Equivalent
7.5 Two Nucleotide Parameters Versus Twenty Amino-Acid Parameters
7.6 Why the Contrast Works
7.7 Relationship to the Exact Pseudoinverse Model
7.8 Recovery of the Principal Hydrophobic Peaks

7.9 Compression Without Explicit Translation
7.10 Reading-Frame Dependence
7.11 Synonymous Codon Choice
7.12 Limits of the Present Comparison
7.13 Future Quantitative Tests
7.14 What the Demonstration Means

Part III

Geometry and Topology of Codon Space

Chapter 8

The G–A–C–U Codon Cube

8.1 From Product Space to Ordered Geometry
8.2 Linear Nucleotide Paths
8.3 The G–A–C–U Path
8.4 Construction of the 4 × 4 × 4 Codon Cube
8.5 Forty-Eight Edges per Codon Position
8.6 The Complete 144-Edge Cube
8.7 Placing Translation on the Cube
8.8 Position-Specific Structure
8.9 The Four Middle-Nucleotide Layers
8.10 Isotropic Codon Cubes
8.11 Anisotropic Codon Cubes
8.12 Why G–A–C–U Is Not Assumed
8.13 Scope of the Geometry

Chapter 9

The Forty-Four-Loop Translated Multigraph

9.1 Collapse by Translated Identity
9.2 Twenty-One Translated Vertices
9.3 Self-Loops and Parallel Edges
9.4 Forty-Four Synonymous Edge Instances
9.5 One Hundred Interclass Edge Instances
9.6 Loop Multiplicity by Translated Class
9.7 Six-Codon Fibers
9.8 Fourfold-Degenerate Fibers
9.9 Three-Codon Fibers

9.10 Two-Codon Fibers
9.11 Single-Codon Fibers
9.12 Degeneracy Versus Local Compactness
9.13 Fiber Connectivity and Local Redundancy
9.14 What the Quotient Preserves and Suppresses
9.15 Why the Distribution Matters

Chapter 10

Exhaustive Search of Isotropic Nucleotide Orders

10.1 The Twenty-Four Oriented Orders
10.2 Reversal Symmetry
10.3 Complete Isotropic Enumeration
10.4 The Score Range
10.5 Position-Specific Nucleotide-Pair Weights
10.6 Exact Proof of the Isotropic Maximum
10.7 The Nearest Competitors
10.8 The Lowest-Scoring Orders
10.9 Why Exhaustive Enumeration Remains Necessary
10.10 The Unique Isotropic Maximum
10.11 Isotropic Optimality Is Not Yet the Global Result

Chapter 11

Exhaustive Search of Anisotropic Codon Cubes

11.1 Independent Nucleotide Orders on Three Axes
11.2 Axis-Separable Scoring
11.3 Position-Specific Score Spectra
11.4 The Exact Generating Polynomial
11.5 The Complete Score Distribution
11.6 Mean, Variance, and Symmetry
11.7 First-Position Maximizers
11.8 Second-Position Maximizers
11.9 Third-Position Maximizers
11.10 Position-Specific Maximum Families
11.11 Cartesian Product of the Maximum Families
11.12 Isotropic Members of the Maximum Family
11.13 G–A–C–U as the Unique Common Undirected Path

11.14 The Exact Forty-Four-Edge Maximum Theorem
11.15 What the Anisotropic Search Establishes

Part IV

Interpretation and Generalization

Chapter 12

One Code, Three Mathematical Views

12.1 The Lookup Relation
12.2 Algebraic Balance
12.3 Physicochemical Projection
12.4 Synonymous Topology
12.5 Input-Space and Assignment-Specific Results
12.6 The Same Completeness Produces Different Structures
12.7 Balance Below, Asymmetry Above
12.8 Hydropathy and Topology as Complementary Projections
12.9 The Role of X Across the Three Views
12.10 Compression at Three Levels
12.11 One Relation, Several Quotients
12.12 Exactness at Different Levels
12.13 The Genetic Code Is More Than Its Matrix
12.14 The Genetic Code Is More Than Its Topology
12.15 The Genetic Code Is More Than Hydropathy
12.16 A Layered Mathematical Architecture
12.17 A Unified Statement
12.18 What the Synthesis Does Not Establish
12.19 From Description to Explanation

Chapter 13

Beyond Frozen Accident

13.1 Persistence Versus Structure
13.2 What an Adequate Account Must Explain
13.3 Accident Requires a Defined Ensemble
13.4 A Hierarchy of Alternative Code Spaces
13.5 Combinatorial Counts Are Not Historical Probabilities
13.6 Topology as an Additional Constraint
13.7 Physicochemical Organization as Another Constraint

13.8 The Limits of “Could Have Been Otherwise”
13.9 An Information-Theoretic Phase Change
13.10 Suddenness at the Organizational Level
13.11 A Code That Participates in Its Own Persistence
13.12 Why Global Structure Can Be Understated
13.13 Scientific Caution and Conceptual Fear
13.14 Four Levels of Interpretation
13.15 Beyond the False Choice
13.16 What Could Weaken the Interpretation
13.17 What “Frozen Accident” Still Contributes
13.18 The Strongest Present Conclusion

Chapter 14

The Genetic Code as a Persistent Signal

14.1 Code Versus Message
14.2 Distributed Physical Implementation
14.3 Global Consequences of Reassignment
14.4 Reflexive Maintenance
14.5 Persistence Across Biological Time
14.6 Structure Visible in Coding Sequences
14.7 The Code as a Constraint on Future Biological Change
14.8 A Candidate for an Exceptionally Persistent Natural Signal
14.9 Stability and Meaning
14.10 Persistence Does Not Mean Immutability
14.11 Persistent Signal, Open Origin

Chapter 15

Tests, Controls, and Falsifiable Extensions

15.1 Reanalysis of the Historical Hydropathy Calculation
15.2 Alternative Hydropathy Scales
15.3 Alternative Numerical Conventions for X
15.4 Residue-Volume Verification
15.5 Natural Variant Genetic Codes
15.6 Degeneracy-Preserving Randomization
15.7 Codon-Box-Preserving Randomization
15.8 Amino-Acid Property Permutations
15.9 Mutation-Weighted Graphs

15.10 Codon-Usage-Weighted Analyses
15.11 Complete Multigraph Analysis
15.12 Quantitative Benchmarking of the U/T–A Membrane Profile
15.13 Results That Can Change and Results That Cannot
15.14 Reproducibility and Provenance
15.15 The Purpose of the Controls

Chapter 16

Conclusions

16.1 The Complete Codon Matrix
16.2 The Exact Pseudoinverse
16.3 Nucleotide-Determined Hydropathy
16.4 The Second Codon Position
16.5 The G–A–C–U Codon Geometry
16.6 The Forty-Four-Loop Multigraph
16.7 The G–A–C–U Maximum
16.8 The Anisotropic Maximum
16.9 Structure Beyond the Conventional Lookup Table
16.10 What Is Exact and What Is Conditional
16.11 Beyond Frozen Accident
16.12 A Possible Information-Theoretic Transition
16.13 The Genetic Code as Persistent Natural Signal
16.14 What Remains to Be Explained
16.15 Final Perspective

Technical Appendices

Appendix A

Construction of the Complete 64 × 12 Codon Matrix

Appendix B

Gram Matrix, Rank, and Singular Spectrum

Appendix C

Exact Moore–Penrose Pseudoinverse

Appendix D

Projection, Gauge Freedom, and Minimum Norm

Appendix E

Complete Position-Specific Nucleotide-Pair Weights

Appendix F

Complete Enumeration of Isotropic Nucleotide Orders

Appendix G

Exact Anisotropic Score Distribution

Appendix H

Synonymous Self-Loop Multiplicities

Appendix I

Computational Reproduction Protocol

Appendix J

Notation

Appendix K

Glossary

Appendix L

Exact Pseudoinverse for Complete Alphabet–Word Matrices

Appendix M

Inclusion–Exclusion and the Space of Complete Genetic Codes

Back Matter

References

List of Figures

Figure 1. The complete 64 × 12 binary codon incidence matrix paired with the translated-output vector.

Figure 2. Singular-value spectrum of the complete codon incidence matrix.

Figure 3. Exact two-value Moore–Penrose pseudoinverse of the complete codon incidence matrix.

Figure 4. Complete-code nucleotide-position hydropathy coefficients.

Figure 5. Empirical and reconstructed amino-acid hydropathy.

Figure 6. Hydropathy–residue-volume organization by the second codon position.

Figure 7. Comparison of the conventional Kyte–Doolittle profile and the nucleotide-only second-position T–A profile.

Figure 8. The isotropic G–A–C–U codon cube.

Figure 9. The G–A–C–U translated multigraph.

Figure 10. Synonymous self-loop multiplicity by translated class.

Figure 11. Synonymous-edge scores for all twenty-four oriented isotropic nucleotide orders.

Figure 12. Distribution of synonymous-edge totals across all 13,824 oriented anisotropic codon cubes.

Figure 13. Axis-separable summary of the forty-four-edge maximum and the unique common nucleotide path.

List of Tables

Table 1. The standard RNA genetic code in single-letter notation.

Table 2. Exact complete-code nucleotide-position hydropathy coefficients.

Table 3. Position-specific synonymous-edge score spectra.

Table 4. Position-specific maximum families and their Cartesian product.

Table E1. Position-specific and total synonymous nucleotide-pair weights.

Table F1. Complete isotropic nucleotide-order enumeration.

Table G1. Exact score distribution across all 13,824 oriented anisotropic codon cubes.

Table H1. Codon degeneracy and G–A–C–U self-loop multiplicity.

Notation and Conventions

The principal notation of this book is based on the standard RNA genetic code.

RNA codons use the nucleotide alphabet

$\Sigma = \{ G, A, C, U \}$.

The letter U denotes uracil. Historical figures constructed from DNA sequences may retain T for thymine. In those contexts, the RNA equivalent of T is U.

Codons are written in the 5′-to-3′ direction. A codon is an ordered triplet

$c = c_1 c_2 c_3$,

where c_1, c_2, and c_3 denote the first, second, and third nucleotide positions.

The complete codon space contains

$4^3 = 64$

codons.

Amino acids are represented by their standard single-letter codes. Translation stop is represented by X rather than an asterisk. The translated-output set therefore contains twenty-one classes: twenty amino acids and X.

The standard genetic-code map is written

$\tau : C \rightarrow \Omega$,

where C is the sixty-four-codon input set and Ω is the twenty-one-member translated-output set.

For a translated class ω, its codon fiber is

$F\omega = \tau^{-1} (\omega)$.

The number of codons in that fiber is its degeneracy:

$d\omega = | F\omega |$.

The complete binary codon incidence matrix is written D. It contains sixty-four rows and twelve columns. The column order used throughout this book is

1G, 1A, 1C, 1U,

2G, 2A, 2C, 2U,

3G, 3A, 3C, 3U.

Thus, the matrix uses G–A–C–U order within each positional block. The phrase **G–A–C–U path** also refers to the linear nucleotide order used to construct the principal codon cube.

The transpose of D is written D^{T}. Its Moore–Penrose pseudoinverse is written D^{+}. The all-ones matrix is written J, with dimensions supplied when needed. Identity matrices are written I with a dimensional subscript.

Numerical property fields over codon space are written h. Their nucleotide-position coefficient vectors are written β. The additive reconstruction is written $\hat{h}$, and the residual is written r:

$h = \hat{h} + r.$

The codon-space projector is

$P = D D^{+}.$

Nucleotide-path scores count synonymous nearest-neighbor edges. The total synonymous-edge score is written L. Position-specific scores are written L_1, L_2, and L_3.

A path and its reverse represent the same undirected nucleotide topology. Thus, G–A–C–U and U–C–A–G are one undirected path.

All displayed mathematics is written in Word-safe notation. No raw LaTeX commands are required in the final manuscript.

Abstract

The standard genetic code is ordinarily presented as a biological lookup table in which sixty-four RNA codons map to twenty amino acids and translation stop. This book treats that complete finite relation as a mathematical object. The same code is represented as a binary incidence matrix, a projected physicochemical field, a three-dimensional codon lattice, and a translated multigraph. These representations expose structures that remain largely hidden in the conventional chart.

The complete codon set is encoded by a 64 × 12 binary matrix D, with one column for each nucleotide at each of the three codon positions. Every row contains three 1s and nine 0s. The matrix has rank ten, one singular value equal to 4 √3, nine singular values equal to 4, and two zero singular values. Its Moore–Penrose pseudoinverse has the exact closed form

$$D^{+} = (1 / 16) D^{T} - (1 / 96) J$$

and contains only two values: 5 / 96 when a nucleotide-position coordinate is present in a codon and −1 / 96 when it is absent.

Numerical properties assigned to translated outputs can be projected through this pseudoinverse. For amino-acid hydropathy, the resulting nucleotide-position coefficients reveal a strong polarity at the second codon position. U contributes strongly toward hydrophobicity, whereas A contributes strongly toward hydrophilicity. A historically developed nucleotide-only membrane profile based on the second-position T–A contrast, or U–A in RNA notation, reproduces the principal broad hydrophobic peaks of a conventional Kyte–Doolittle profile without explicitly translating the coding sequence into amino acids.

A complementary topological representation orders the four nucleotides along each codon axis. The isotropic G–A–C–U cube contains sixty-four vertices and 144 nearest-neighbor edges. Forty-four of those edges connect synonymous codons and become self-loops when the cube is collapsed by translated identity. Exhaustive evaluation of all twenty-four isotropic nucleotide orders and all 13,824 anisotropic axis-order combinations proves that no arrangement exceeds forty-four synonymous edges. The independent maxima for the first, second, and third codon positions are 4, 1, and 39. G–A–C–U is the unique undirected nucleotide path belonging to all three positional maximum families.

The exact matrix results follow from the complete factorial architecture of codon space. The hydropathy and topology results depend on the particular assignments of the standard genetic code. Together, they show that the familiar lookup table contains algebraic balance, physicochemical organization, and exact local topology. These findings do not determine how the genetic code originated, but they define structures that any adequate historical account must explain.

Preface

The Genetic Code as a Mathematical Object

The Familiar Table and the Hidden Object

The genetic code is one of the most familiar tables in biology. A nucleotide triplet is located, an amino acid is read, and the table has apparently completed its task. Its compactness makes the code seem fully understood at a glance. Sixty-four codons occupy a finite chart, twenty amino acids recur among its cells, and three codons specify translation stop.

Yet consultation is not analysis.

A table can answer individual questions while concealing relations among all of its entries. The genetic-code table tells us that one codon maps to leucine, another to glycine, and another to translation stop. It does not immediately show how the complete set of codons behaves as a balanced product space, how translated physicochemical properties project onto nucleotide coordinates, or how synonymous classes are arranged under a declared local geometry.

This book begins by treating the entire code as one finite object.

Let C denote the set of sixty-four RNA codons, and let Ω denote the twenty amino-acid identities together with translation stop X. The standard genetic code is a map

$\tau : C \rightarrow \Omega$.

Every codon has exactly one translated output. Multiple codons may share the same output. The map is therefore complete and many-to-one.

That simple statement permits several distinct mathematical representations.

The codons can be represented as rows of a binary incidence matrix. The translated outputs can be represented as a categorical vector or, after assigning a numerical property, as a codon-level field. The four nucleotide states can be ordered along three axes to create a codon lattice. The codon vertices can then be collapsed by translated identity to create a multigraph containing self-loops and parallel edges.

Each representation preserves one aspect of the code and suppresses another. Their combination reveals a structure richer than the conventional printed table suggests.

Three Principal Views

The first view is algebraic.

Each codon contains one nucleotide at each of three positions. A one-hot representation therefore uses twelve coordinates: four nucleotide indicators for position 1, four for position 2, and four for position 3. The complete matrix has sixty-four rows and twelve columns.

Because every possible codon occurs, the design is exactly balanced. Every nucleotide appears sixteen times at every position. Every pair of nucleotide states belonging to two different positions appears four times. This balance determines the matrix rank, singular spectrum, and Moore–Penrose pseudoinverse.

The matrix has rank ten. Its pseudoinverse is not merely numerically computable; it has the exact formula

$D^{+} = (1 / 16) D^{T} - (1 / 96) J.$

Every pseudoinverse entry is one of only two rational values. A coordinate present in a codon contributes 5 / 96, while an absent coordinate contributes −1 / 96.

The second view is physicochemical.

A numerical property such as hydropathy can be assigned to every translated output and thereby to every codon. The pseudoinverse then determines the closest additive representation of that property using twelve nucleotide-position coefficients.

This calculation does not claim that nucleotide identity chemically causes amino-acid hydropathy. It measures how the standard lookup relation distributes hydropathy across codon space. The strong opposition between second-position U and second-position A shows that broad amino-acid environmental preference is already visible in the nucleotide coordinates used to encode translated identity.

The third view is topological.

Codon space is a Cartesian product of three four-state positions, but a product space does not specify which nucleotide states should be regarded as neighbors. Choosing a linear nucleotide order defines a local geometry. Under G–A–C–U, each codon position becomes a four-vertex path, and the complete code becomes a 4 × 4 × 4 lattice.

When two neighboring codons translate to the same output, their connecting edge preserves translated identity. Collapsing synonymous codons converts such an edge into a self-loop. The isotropic G–A–C–U cube produces forty-four such loops.

The number forty-four is not selected because it is visually attractive or because the cube was drawn first. Every alternative isotropic nucleotide order is evaluated. Every anisotropic combination of three independently ordered axes is also evaluated. The maximum remains forty-four.

G–A–C–U has a further exact property. It is the unique undirected nucleotide path that belongs independently to the maximizing families of all three codon positions.

Exact Results and Empirical Results

This book distinguishes several kinds of result.

Some results follow entirely from the architecture of complete word space. The rank-ten matrix, its singular values, and its exact pseudoinverse would remain unchanged if the sixty-four codons were reassigned to different outputs. These properties belong to the complete codon input representation.

Other results depend on the standard translation map. Hydropathy coefficients, synonymous-loop multiplicities, nucleotide-pair weights, and the G–A–C–U optimum would generally change if the translated assignments changed. These properties belong to the organization of output over input space.

This distinction is essential. It prevents the exact pseudoinverse from being presented as though it were itself evidence for a special biological assignment. The pseudoinverse is exact because the complete codon set forms a balanced factorial design. Its biological use begins when the standard translated-output field is projected through it.

The same discipline applies to hydropathy. The projection formula is exact. The chosen amino-acid property scale, the numerical treatment of X, the regression procedure, and the interpretation of residuals are empirical choices that must be stated and audited.

The topological results occupy an intermediate position. Once the standard genetic-code assignments and a class of permitted nucleotide paths are declared, the edge counts and exhaustive maxima are exact. The choice of that comparison class remains part of the model.

An exact answer is exact within its stated definitions. Precision does not eliminate the need to state those definitions.

Historical Motivation

The nucleotide-determined hydropathy work arose in the context of engineering membrane proteins. Conventional hydropathy profiles translate a coding sequence into amino acids and then assign one of twenty residue-specific parameters. The earlier genetic-coding approach

asked whether the nucleotide sequence itself already contained enough structure to reveal membrane-spanning regions.

A strikingly compressed signal emerged from the second codon position. In historical DNA notation, a profile based on the contrast

$$T - A$$

tracked the principal broad peaks of a conventional Kyte–Doolittle profile. In RNA notation, the corresponding contrast is

$$U - A.$$

The comparison suggested that a major part of membrane hydropathy could be recovered from only two nucleotide states occupying one of the three codon positions. The complete pseudoinverse analysis developed here places that observation within a broader linear-algebraic framework.

The historical 1990 calculation used singular-value decomposition on the sixty-one sense codons. The present principal formulation uses all sixty-four codons, retains X as a categorical translated output, assigns X hydropathy zero for the stated complete-code calculation, and applies the exact pseudoinverse.

The two methods should not be confused, but neither should their common result be obscured. Both show that the strongest hydropathy organization is concentrated at the second codon position.

Why the Complete Code Matters

A sixty-one-codon analysis is appropriate when numerical properties are assigned only to amino acids. A sixty-four-codon representation is necessary when the code itself is treated as a complete categorical relation.

Retaining all sixty-four codons preserves the exact balance of the input space. Every nucleotide occurs sixteen times at every position. Every cross-position nucleotide pair occurs four times. The matrix obtains its simple Gram structure, exact singular spectrum, and closed-form pseudoinverse from this completeness.

Translation stop X is not an amino acid, but it is a genuine output of the lookup relation. UAA, UAG, and UGA do not lack meaning within translation. They specify termination.

The categorical analysis therefore retains X. Numerical property analyses must then state separately how X is treated. It may be omitted, assigned zero, assigned another declared value, or handled as missing. These conventions define different calculations.

Completeness should not be gained by silently pretending that X has an intrinsic amino-acid hydropathy. Nor should exact balance be abandoned without acknowledging what is lost.

Frozen Accident and Internal Structure

The phrase “frozen accident” captures an important historical idea. Once a translation code is deeply embedded in a biological system, changing codon meanings becomes globally disruptive. Reassignment can alter many proteins at once. The code may therefore remain stable because change is costly.

That argument addresses persistence.

It does not, by itself, explain internal organization.

Freezing can preserve a structured table or an unstructured table. A random assignment can freeze. A chemically constrained assignment can freeze. A partially optimized assignment can freeze. A historically layered assignment can freeze.

If the present code contains strong physicochemical partitioning, exact local topology, and a nucleotide path that independently maximizes all three codon positions, those features remain to be explained after historical lock-in is acknowledged.

This book does not replace frozen accident with a single alternative theory of origin. It separates measurable structure from the historical narratives proposed to explain it.

The Reluctance to See Structure

The genetic code has long stood close to questions of accident, constraint, optimization, and design. That intellectual setting can make structural claims unusually sensitive. A result may be resisted not because its arithmetic is difficult, but because its possible interpretation appears uncomfortable.

The proper response is not to suppress structure or to inflate it.

An exact pseudoinverse should be reported as an exact pseudoinverse. A maximum over a finite comparison class should be reported as a maximum over that class. A physicochemical correlation should be reported with its scale, convention, and residuals. None of these results alone establishes a historical mechanism or metaphysical conclusion.

Conversely, concern about speculative interpretation is not a reason to leave exact structure unexamined.

The unexpected fact is not merely that a pseudoinverse can be computed. Any finite matrix has a Moore–Penrose pseudoinverse. The unusual fact is that the complete natural lookup domain produces a pseudoinverse with a simple exact formula and only two rational entry values.

That property deserves recognition independently of what one believes it ultimately means.

A Possible Information-Theoretic Discontinuity

A complete genetic code is more than a collection of local molecular affinities. Every admissible triplet receives an interpreted output. The system therefore passes from isolated chemical interactions into a globally maintained correspondence between two finite alphabets.

That transition may be viewed as an information-theoretic discontinuity or phase change.

This possibility does not require every molecular component to have appeared simultaneously. Partial assignments, ambiguous translation, and incomplete coding systems could precede a complete code. The discontinuity occurs at the organizational level when a sufficiently complete and self-maintaining lookup relation exists.

Once established, the code participates in producing the molecular machinery that continues to enact it. Proteins translated under the code help maintain translation under the same code. The relation becomes dynamically reflexive.

The code is not merely stored. It is repeatedly interpreted, implemented, repaired, and reproduced by a system whose components are themselves produced through that interpretation.

Whether the transition was gradual, abrupt, chemically constrained, selectively optimized, historically contingent, or some combination remains unresolved. The mathematics developed here specifies the present structure that an adequate account of origin must reproduce.

The Genetic Code as a Persistent Signal

A message is a particular sequence. A code is the relation that makes sequences interpretable.

Individual genes change, duplicate, recombine, and disappear. The mapping by which codons are translated has persisted far more broadly than any individual genetic message. Its persistence is distributed across organisms, genomes, translation machinery, and repeated acts of protein synthesis.

This makes the genetic code a candidate for an exceptionally persistent natural signal.

The statement must remain qualified. Demonstrating that it is the single oldest or most persistent natural signal would require a formal comparison class and historical evidence extending beyond the present mathematical analysis.

The stronger defensible observation is that the code combines several unusual properties. It is finite, nearly universal, globally consequential, physically distributed, reflexively maintained, and repeatedly instantiated across translated genes.

What persists is not one molecular object. It is a relation.

Scope

This book does not claim that G–A–C–U is the only possible geometry of nucleotide relations. It is the exact optimum within the family of linear four-state paths under the objective of maximizing synonymous nearest-neighbor edges.

The book does not claim that amino-acid chemistry is reducible to codon composition. It shows that a substantial additive component of hydropathy aligns with nucleotide-position coordinates.

The book does not claim that an exact pseudoinverse proves a particular origin theory. The pseudoinverse belongs to the complete factorial input architecture.

The book does not claim that exact structure establishes intention, design, or a metaphysical conclusion. It establishes exact structure.

The book also does not claim that historical contingency is absent. A persistent code may contain constrained, optimized, contingent, and inherited features simultaneously.

The aim is narrower and stronger: to determine what follows when the complete genetic code is treated as a finite mathematical world rather than merely as a chart.

Part I begins with the complete lookup relation itself.

Part I

The Genetic Code as a Complete Finite Relation

Chapter 1

The Standard Genetic Code

1.1 A Complete Lookup Relation

The standard RNA genetic code assigns one translated output to every possible nucleotide triplet. Its input alphabet is

$\Sigma = \{ G, A, C, U \}$.

A codon is an ordered word of length three:

$c = c_1 c_2 c_3$,

where

$c_1, c_2, c_3 \in \Sigma$.

The complete codon set is therefore

$C = \Sigma^3$.

Its cardinality is

$| C | = 4^3 = 64$.

The translated-output set consists of the twenty standard amino-acid classes together with translation stop X:

$\Omega = \{ A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, X \}$.

Thus,

$| \Omega | = 21$.

The standard genetic code is the finite map

$\tau : C \rightarrow \Omega$.

The map is complete, or total, because every member of C has exactly one translated output. No codon is left uninterpreted within the standard relation.

The map is not one-to-one. Several codons may share the same translated output. This many-to-one structure is commonly called degeneracy.

For example, UUU and UUC both translate to phenylalanine, represented by F. GGU, GGC, GGA, and GGG all translate to glycine, represented by G. The six leucine codons translate to L. Three codons—UAA, UAG, and UGA—translate to X.

The word **degeneracy** can sound as though information has been lost or the code has failed to distinguish its inputs. That is not the relevant meaning here. The sixty-four codons remain distinct inputs. Degeneracy means that several distinct inputs belong to the same translated-output class.

The complete lookup relation can therefore be viewed as a partition of codon space into twenty-one translated fibers.

1st base	2nd base U	2nd base C	2nd base A	2nd base G
U	UUU F UUC F UUA L UUG L	UCU S UCC S UCA S UCG S	UAU Y UAC Y UAA X UAG X	UGU C UGC C UGA X UGG W
C	CUU L CUC L CUA L CUG L	CCU P CCC P CCA P CCG P	CAU H CAC H CAA Q CAG Q	CGU R CGC R CGA R CGG R
A	AUU I AUC I AUA I AUG M	ACU T ACC T ACA T ACG T	AAU N AAC N AAA K AAG K	AGU S AGC S AGA R AGG R
G	GUU V GUC V GUA V GUG V	GCU A GCC A GCA A GCG A	GAU D GAC D GAA E GAG E	GGU G GGC G GGA G GGG G

Table 1. The standard RNA genetic code in single-letter notation. The sixty-four codons formed from G, A, C, and U map to twenty amino-acid classes and translation stop X. Repeated output letters show synonymous degeneracy: distinct codons may share the same translated output. Codons are read in the 5′-to-3′ direction.

The conventional genetic-code table is already a complete finite mathematical object. It contains:

sixty-four distinct inputs;

twenty-one output classes;

one output for every input;

unequal numbers of inputs per output;

and a precisely specified many-to-one relation.

The table does not merely summarize observations. It defines the complete mapping used by translation under the standard code.

1.2 The Twenty-One Translated Classes

The genetic code is often described as a mapping from sixty-four codons to twenty amino acids. That statement is biologically familiar but categorically incomplete.

Three codons do not specify an amino acid. They specify translation termination. In this book, translation stop is represented by X.

The output set therefore contains twenty-one classes rather than twenty.

X is not treated as a twenty-first amino acid. It is treated as a genuine output class of the lookup relation.

This distinction matters because the categorical code and a numerical amino-acid property analysis are different objects.

Categorically,

τ (UAA) = X,

τ (UAG) = X,

and

τ (UGA) = X.

These assignments are as definite as

τ (AUG) = M

or

τ (UGG) = W.

Numerically, however, X does not possess an intrinsic amino-acid hydropathy, residue volume, charge, or side-chain mass. Any numerical value assigned to X must therefore be introduced as an explicit analytical convention.

The complete categorical code should not be reduced to sixty-one codons merely because a chosen amino-acid property is undefined for X. Nor should X be silently assigned an amino-acid-like property merely to preserve matrix dimensions.

The categorical and numerical questions should instead be separated.

For the complete categorical relation:

all sixty-four codons are retained;

X is included as one translated output;

and the exact balance of codon space is preserved.

For a numerical property calculation:

the treatment of X is stated separately;

the consequences of that treatment can be tested;

and historical sixty-one-codon analyses can be compared with complete sixty-four-codon analyses without conflating them.

Retaining X also matters for the graph-theoretic analysis developed later. The three stop codons occupy definite vertices in codon space. Under the G–A–C–U topology, two selected nearest-neighbor edges connect stop codons to other stop codons. After quotienting by translated identity, those edges become two self-loops at the X vertex.

The stop class therefore participates in the topology of the complete translated relation even though it is not an amino acid.

1.3 Codon Fibers

For a translated output $\omega \in \Omega$, define its codon fiber as

$F\omega = \tau^{-1}(\omega)$.

The fiber $F\omega$ is the set of all codons that translate to ω.

Its cardinality is

$d\omega = |F\omega|$.

The number $d\omega$ is the codon degeneracy of the translated class ω.

The complete standard degeneracy profile is:

$d_R = d_L = d_S = 6;$

$d_A = d_G = d_P = d_T = d_V = 4;$

$d_I = d_X = 3;$

$d_C = d_D = d_E = d_F = d_H = d_K = d_N = d_Q = d_Y = 2;$

and

$d_M = d_W = 1.$

The total codon count is therefore

$3(6) + 5(4) + 2(3) + 9(2) + 2(1)$

$= 18 + 20 + 6 + 18 + 2$

$= 64.$

The number of translated classes is

$3 + 5 + 2 + 9 + 2 = 21.$

These two totals provide elementary checks on any machine-readable representation of the standard code.

Every codon must occur exactly once.

Every codon must belong to exactly one translated fiber.

All twenty amino-acid classes and X must occur.

The fiber cardinalities must sum to sixty-four.

The unequal fiber sizes are among the most visible features of the conventional genetic-code table. Methionine and tryptophan each have one codon. Many amino acids have two. Isoleucine and X have three. Five amino acids have four. Leucine, arginine, and serine each have six.

Yet fiber cardinality is only one kind of structure.

1.4 Degeneracy Is Not Geometry

Two translated classes can have the same number of codons while occupying very different configurations in codon space.

Leucine, arginine, and serine each have six codons:

$d_L = d_R = d_S = 6$.

Their fibers are equal in cardinality. They are not equivalent in local geometry.

Under the G–A–C–U codon topology developed in Part III, leucine and arginine each contain six selected synonymous nearest-neighbor edges, whereas serine contains four:

$\ell_L = 6$,

$\ell_R = 6$,

and

$\ell_S = 4$.

Thus, equal degeneracy does not imply equal local connectedness.

The distinction can be stated generally.

Degeneracy asks:

How many codons belong to the translated class?

Fiber geometry asks:

How are those codons arranged under a declared adjacency relation?

A translated fiber may be compact, extended, divided into several components, or connected through different codon positions. None of those properties is determined by cardinality alone.

This distinction becomes especially important after codon space is represented as a graph. Synonymous codons that are locally adjacent produce self-loops after translation collapse. Synonymous codons that are distant from one another contribute to degeneracy but not to local synonymous connectivity.

The ordinary code table shows which codons share an output. It does not, by itself, specify a nucleotide geometry or reveal the local shapes of the translated fibers.

1.5 Synonymous Codons and Output Preservation

Two codons c and d are synonymous when

$\tau(c) = \tau(d)$.

Synonymy is an equality of translated output, not an assertion that the codons are physically or statistically interchangeable in every biological context.

Synonymous codons can differ in:

usage frequency;

transfer-RNA availability;

translation rate;

local sequence context;

RNA structure;

mutation accessibility;

and regulatory effects.

The finite lookup relation suppresses these additional contexts and retains only translated identity.

That abstraction is intentional.

The mathematical object studied here is the standard codon-to-output relation. Questions involving codon usage, mutation bias, translation kinetics, or organism-specific regulation can later be introduced as weights or extensions. They should not be inserted silently into the definition of the code itself.

At the categorical level, synonymous codons are distinct inputs mapped to the same output.

At the geometric level, some synonymous codon pairs are neighbors and others are not.

At the statistical level, some may be common and others rare.

At the molecular level, they may differ in effects not represented by amino-acid identity alone.

The word **synonymous** refers only to preservation of the translated class under τ.

1.6 Completeness and Standardness

The phrase **standard genetic code** refers to the canonical assignment used as the principal reference throughout molecular biology.

Natural variant codes exist. Some mitochondrial, microbial, and organellar systems reassign particular codons. The existence of variants does not prevent the standard code from being treated as a complete finite relation. It means that the standard relation is one member of a broader family of natural coding systems.

This book uses the standard RNA code as its primary object because it provides:

a fully specified sixty-four-codon domain;

a widely recognized twenty-one-class output relation;

a stable basis for comparison;

and a natural starting point for exact algebraic and topological analysis.

The word **complete** does not mean universal without exception. It means that every input in the declared domain has an output within the declared relation.

The word **standard** does not mean that every living system uses every assignment without variation. It identifies the reference mapping under study.

Natural variants provide valuable controls. Later analyses can ask whether the exact hydropathy coefficients, synonymous-loop counts, or maximizing nucleotide paths change under alternative codes.

The input matrix will remain unchanged so long as all sixty-four RNA codons are retained. The output-dependent structures may change when assignments change.

1.7 A Finite Natural Lookup System

The genetic code is unusual because its complete input set and its complete output assignment can be written explicitly.

Many natural relationships are continuous, probabilistic, contextual, or only partially observed. Their possible inputs may be unbounded, and their outputs may depend on environmental conditions or hidden states.

The genetic code is different in several respects.

Its canonical input alphabet is finite.

Its words have fixed length.

The complete input space contains only sixty-four elements.

Each input has a declared translated output.

The resulting relation can be enumerated exactly.

This permits exact questions:

How many codons exist?

How many translated classes exist?

How many codons belong to each class?

What is the rank of a complete positional incidence matrix?

Does that matrix have an exact pseudoinverse?

How many selected local codon changes preserve translated identity?

Which nucleotide paths maximize that preservation?

How many possible complete codon-to-output assignments exist under stated conditions?

These are not asymptotic or sampling questions. Once the definitions are fixed, they have finite answers.

The genetic code may be one of the clearest known natural examples of a complete lookup relation. It is tempting to call it the only natural lookup code. That stronger assertion requires caution.

To establish uniqueness, one would first need to define a comparison class broad enough to include other natural recognition and interpretation systems. Candidate comparisons might include immune recognition, receptor–ligand specificity, transcription-factor binding, RNA editing, enzymatic substrate recognition, or developmental signaling. Most of these systems are more contextual and less easily represented as complete finite tables, but that does not settle the matter.

The defensible claim is narrower:

The genetic code is an exceptionally clear natural system in which a finite set of symbolic inputs is assigned to a finite set of translated outputs by a physically implemented and repeatedly enacted lookup relation.

Its mathematical clarity makes it suitable for exact analysis.

1.8 Code, Table, and Physical Implementation

The printed table is not itself the genetic code.

The table is a human-readable representation of a relation enacted by molecular systems.

A codon is recognized through interactions involving messenger RNA, transfer RNA, aminoacyl-transfer-RNA synthetases, ribosomes, release factors, and associated molecular machinery. The translated assignment is distributed across this system rather than stored in one isolated object.

The finite map τ abstracts from those molecular details. It records the input–output relation they collectively implement.

This abstraction is comparable to writing a mathematical function without claiming that the function notation contains the physical process that realizes it.

The distinction between the relation and its implementation becomes important when discussing persistence. A particular messenger-RNA molecule is temporary. A particular ribosome is temporary. A particular organism is temporary. The codon-to-output relation persists because it is repeatedly instantiated across many physical components and generations.

The code is therefore both abstract and physical.

It is abstract because the same relation can be written as a finite table or map.

It is physical because translation repeatedly enacts that relation in matter.

The mathematical analysis concerns the relation. Biological interpretation must remember that the relation survives only through physical implementation.

1.9 From the Lookup Table to the Matrix

The standard table emphasizes translated output. The next representation emphasizes codon composition.

Each codon contains one nucleotide at each of three positions. Therefore, each codon can be represented by twelve binary indicators:

four possible nucleotides at position 1;

four at position 2;

and four at position 3.

Exactly one indicator is active in each positional block.

Applying this representation to all sixty-four codons produces a 64 × 12 binary incidence matrix. The output relation can then be placed beside that matrix as a separate sixty-four-component vector.

This transformation does not discard the lookup table. It separates its two layers:

the complete structure of the codon inputs;

and the translated assignments placed over those inputs.

That separation is the foundation of the exact algebra developed in Chapters 2 and 3. The input matrix has a rank, singular spectrum, null space, pseudoinverse, and projection geometry before any amino-acid property is introduced.

The translated outputs become essential again when hydropathy fields and synonymous topologies are constructed.

The standard genetic code is therefore the starting relation from which the rest of the book proceeds.

Chapter 2

The Complete 64 × 12 Binary Codon Matrix

The conventional genetic-code table is designed for lookup. A codon is located, its translated output is read, and the immediate biological question is answered. That format is effective for consultation, but it does not expose the exact combinatorial balance of the complete codon set.

A different representation begins with codon composition before translated meaning is introduced. Each codon contains one nucleotide at each of three ordered positions. The nucleotide occupying each position can therefore be represented by a binary indicator. Applying that representation to all sixty-four codons produces a complete 64 × 12 binary incidence matrix.

This matrix is not an approximation to the genetic code. It is an exact encoding of codon identity. Every row represents one codon. Every column represents one nucleotide at one codon position. The matrix records the complete input architecture over which the translated lookup relation is imposed.

2.1 One-Hot Positional Encoding

Let the RNA nucleotide alphabet be

$\Sigma = \{ G, A, C, U \}$.

The complete codon space is

$C = \Sigma^3$.

Therefore,

$| C | = 4^3 = 64$.

A codon is an ordered triplet

$c = c_1 c_2 c_3$,

where

$c_1, c_2, c_3 \in \Sigma$.

The twelve binary coordinates are divided into three positional blocks. Using G–A–C–U order within each block, the columns are

1G, 1A, 1C, 1U,

2G, 2A, 2C, 2U,

3G, 3A, 3C, 3U.

The numeral identifies the codon position. The nucleotide symbol identifies the state present at that position. Thus, 2A means A at the second codon position, while 3U means U at the third codon position.

Define the binary incidence matrix

$D \in \{ 0, 1 \}^{64 \times 12}$.

For codon c and coordinate (p, n), the matrix entry is

$D_{c,(p,n)} = 1$

when

$c_p = n$,

and

$D_{c,(p,n)} = 0$

otherwise.

Because a codon contains exactly one nucleotide at each of three positions, every row contains exactly three 1s. The remaining nine entries are 0.

For example, GAC activates the coordinates 1G, 2A, and 3C. Its row is

GAC = (1, 0, 0, 0 | 0, 1, 0, 0 | 0, 0, 1, 0).

The separators distinguish the first-, second-, and third-position blocks.

The codon UUU activates the U coordinate in all three blocks:

UUU = (0, 0, 0, 1 | 0, 0, 0, 1 | 0, 0, 0, 1).

The codon GGG activates the G coordinate in all three blocks:

GGG = (1, 0, 0, 0 | 1, 0, 0, 0 | 1, 0, 0, 0).

No amino-acid identity is required to construct any of these rows. The complete matrix can be generated from codon spellings alone.

The translated outputs may be placed beside the matrix as a separate 64 × 1 categorical vector. In this book, that vector contains the twenty standard amino-acid single-letter codes and X for translation stop. The output vector uses the same codon row order as D, but it is not part of D.

This separation is fundamental. The binary matrix describes the complete input space. The adjacent output vector describes the biological assignment imposed on that space.

Codon	1G	1A	1C	1U	2G	2A	2C	2U	3G	3A	3C	3U	Output
GGG	1	0	0	0	1	0	0	0	1	0	0	0	G
GGA	1	0	0	0	1	0	0	0	0	1	0	0	G
GGC	1	0	0	0	1	0	0	0	0	0	1	0	G
GGU	1	0	0	0	1	0	0	0	0	0	0	1	G
GAG	1	0	0	0	0	1	0	0	1	0	0	0	E
GAA	1	0	0	0	0	1	0	0	0	1	0	0	E
GAC	1	0	0	0	0	1	0	0	0	0	1	0	D
GAU	1	0	0	0	0	1	0	0	0	0	0	1	D
GCG	1	0	0	0	0	0	1	0	1	0	0	0	A
GCA	1	0	0	0	0	0	1	0	0	1	0	0	A
GCC	1	0	0	0	0	0	1	0	0	0	1	0	A
GCU	1	0	0	0	0	0	1	0	0	0	0	1	A
GUG	1	0	0	0	0	0	0	1	1	0	0	0	V
GUA	1	0	0	0	0	0	0	1	0	1	0	0	V
GUC	1	0	0	0	0	0	0	1	0	0	1	0	V
GUU	1	0	0	0	0	0	0	1	0	0	0	1	V
AGG	0	1	0	0	1	0	0	0	1	0	0	0	R
AGA	0	1	0	0	1	0	0	0	0	1	0	0	R
AGC	0	1	0	0	1	0	0	0	0	0	1	0	S
AGU	0	1	0	0	1	0	0	0	0	0	0	1	S
AAG	0	1	0	0	0	1	0	0	1	0	0	0	K
AAA	0	1	0	0	0	1	0	0	0	1	0	0	K
AAC	0	1	0	0	0	1	0	0	0	0	1	0	N
AAU	0	1	0	0	0	1	0	0	0	0	0	1	N
ACG	0	1	0	0	0	0	1	0	1	0	0	0	T
ACA	0	1	0	0	0	0	1	0	0	1	0	0	T
ACC	0	1	0	0	0	0	1	0	0	0	1	0	T
ACU	0	1	0	0	0	0	1	0	0	0	0	1	T
AUG	0	1	0	0	0	0	0	1	1	0	0	0	M
AUA	0	1	0	0	0	0	0	1	0	1	0	0	I
AUC	0	1	0	0	0	0	0	1	0	0	1	0	I
AUU	0	1	0	0	0	0	0	1	0	0	0	1	I
CGG	0	0	1	0	1	0	0	0	1	0	0	0	R
CGA	0	0	1	0	1	0	0	0	0	1	0	0	R
CGC	0	0	1	0	1	0	0	0	0	0	1	0	R
CGU	0	0	1	0	1	0	0	0	0	0	0	1	R
CAG	0	0	1	0	0	1	0	0	1	0	0	0	Q
CAA	0	0	1	0	0	1	0	0	0	1	0	0	Q
CAC	0	0	1	0	0	1	0	0	0	0	1	0	H
CAU	0	0	1	0	0	1	0	0	0	0	0	1	H
CCG	0	0	1	0	0	0	1	0	1	0	0	0	P
CCA	0	0	1	0	0	0	1	0	0	1	0	0	P
CCC	0	0	1	0	0	0	1	0	0	0	1	0	P
CCU	0	0	1	0	0	0	1	0	0	0	0	1	P
CUG	0	0	1	0	0	0	0	1	1	0	0	0	L
CUA	0	0	1	0	0	0	0	1	0	1	0	0	L
CUC	0	0	1	0	0	0	0	1	0	0	1	0	L
CUU	0	0	1	0	0	0	0	1	0	0	0	1	L
UGG	0	0	0	1	1	0	0	0	1	0	0	0	W
UGA	0	0	0	1	1	0	0	0	0	1	0	0	X
UGC	0	0	0	1	1	0	0	0	0	0	1	0	C
UGU	0	0	0	1	1	0	0	0	0	0	0	1	C
UAG	0	0	0	1	0	1	0	0	1	0	0	0	X
UAA	0	0	0	1	0	1	0	0	0	1	0	0	X
UAC	0	0	0	1	0	1	0	0	0	0	1	0	Y
UAU	0	0	0	1	0	1	0	0	0	0	0	1	Y
UCG	0	0	0	1	0	0	1	0	1	0	0	0	S
UCA	0	0	0	1	0	0	1	0	0	1	0	0	S
UCC	0	0	0	1	0	0	1	0	0	0	1	0	S
UCU	0	0	0	1	0	0	1	0	0	0	0	1	S
UUG	0	0	0	1	0	0	0	1	1	0	0	0	L
UUA	0	0	0	1	0	0	0	1	0	1	0	0	L
UUC	0	0	0	1	0	0	0	1	0	0	1	0	F
UUU	0	0	0	1	0	0	0	1	0	0	0	1	F

Figure 1. The complete 64 × 12 binary codon incidence matrix paired with the 64 × 1 translated-output vector. Every row contains three 1s and nine 0s. Every column contains sixteen 1s and forty-eight 0s. The output vector uses the standard amino-acid single-letter codes and X for translation stop.

2.2 Exact Balance of the Complete Codon Set

The matrix contains

$64 \times 12 = 768$

binary entries.

Every one of the sixty-four rows contains three 1s, so the total number of active entries is

$64 \times 3 = 192$.

The remaining entries are 0:

$768 - 192 = 576$.

Thus, D contains exactly 192 entries equal to 1 and 576 entries equal to 0.

In vector notation, the row-sum identity is

$D\, 1_{12} = 3\, 1_{64}$.

Every codon activates exactly three of the twelve nucleotide-position coordinates.

The column sums are also exactly balanced. Fix one codon position and one nucleotide. The two remaining positions can independently contain any of four nucleotides. Therefore, each nucleotide-position coordinate occurs in

$4^2 = 16$

codons.

Consequently,

$D^T 1_{64} = 16\, 1_{12}$.

Every column contains sixteen 1s and forty-eight 0s.

This equality is not a statistical tendency. It follows exactly from including all sixty-four codons. No nucleotide is more frequent than another within any positional block because the matrix represents the complete $4 \times 4 \times 4$ product space.

The same balance extends to pairs of different codon positions. Fix a nucleotide at one position and another nucleotide at a second position. The remaining position can take any of four values. Every cross-position nucleotide pair therefore occurs exactly four times.

For example, the combination 1G and 2A occurs in

GAG, GAA, GAC, and GAU.

The combination 2C and 3U occurs in

GCU, ACU, CCU, and UCU.

In general, for two different codon positions p and q and any nucleotides x and y,

the number of codons satisfying $c_p = x$ and $c_\varphi = y$ is 4.

Within a single codon position, two different nucleotide indicators can never be active simultaneously. A codon cannot contain both G and A at its first position. Distinct columns within one positional block therefore have inner product zero.

Across two different codon positions, every nucleotide-coordinate pair has inner product four.

These exact relationships determine the complete Gram matrix.

2.3 The Gram Matrix

The column Gram matrix is

$G = D^T D$.

It is a 12 × 12 matrix whose entries are the inner products of the twelve incidence columns.

Partition D into three 64 × 4 positional blocks:

$D = [\, D_1 \mid D_2 \mid D_3 \,]$.

Each block contains the four nucleotide indicators for one codon position.

Within one positional block, every column contains sixteen 1s. The squared norm of each column is therefore sixteen. Distinct columns within the same block are mutually exclusive and have inner product zero. The diagonal blocks of G are consequently

$16\, I_4$,

where I_4 is the 4 × 4 identity matrix.

Across two different positional blocks, every nucleotide pair occurs in four codons. Every entry of an off-diagonal 4 × 4 block is therefore four. The off-diagonal blocks are

$4\, J_4$,

where J_4 is the 4 × 4 all-ones matrix.

The complete Gram matrix is

G =

$$\begin{bmatrix} 16\, I_4 & 4\, J_4 & 4\, J_4 \\ 4\, J_4 & 16\, I_4 & 4\, J_4 \\ 4\, J_4 & 4\, J_4 & 16\, I_4 \end{bmatrix}.$$

This compact block expression contains the rank and singular-value structure of D. It replaces a direct numerical calculation involving all sixty-four rows with an exact argument based on balance and symmetry.

The trace of G is

$$\mathrm{tr}(G) = 12 \times 16 = 192.$$

This is also the squared Frobenius norm of D:

$$\| D \|^2_F = 192.$$

The same value was obtained by counting the 192 entries equal to 1. This agreement provides an immediate internal check.

2.4 The Constant Direction

Consider the twelve-dimensional constant vector

$$1_{12} = (1_4, 1_4, 1_4)^T.$$

The diagonal block acting on one four-component constant vector gives

$$16\, I_4\, 1_4 = 16\, 1_4.$$

Each off-diagonal block gives

$$4\, J_4\, 1_4 = 16\, 1_4.$$

For each output block, one diagonal contribution and two off-diagonal contributions are present. Therefore,

$$16\, 1_4 + 16\, 1_4 + 16\, 1_4 = 48\, 1_4.$$

It follows that

$$G\, 1_{12} = 48\, 1_{12}.$$

The constant coordinate direction is therefore an eigenvector of G with eigenvalue

$$\lambda_o = 48.$$

The corresponding singular value of D is

$$\sigma_o = \sqrt{48} = 4\sqrt{3}.$$

This singular value occurs once.

The constant direction represents an equal contribution from every nucleotide-position coordinate. When mapped through D, it produces a constant field over all sixty-four codons because every row contains exactly three active coordinates.

2.5 Within-Position Contrast Directions

Now consider a vector supported within one positional block whose four entries sum to zero:

$x = (xG, xA, xC, xU)^T$,

with

$xG + xA + xC + xU = 0$.

Because J_4 multiplies a vector by its coordinate sum,

$J_4 x = 0$.

The off-diagonal Gram blocks therefore make no contribution to such a contrast. Only the diagonal block remains:

$16 I_4 x = 16 x$.

Every zero-sum contrast within one positional block is therefore an eigenvector of G with eigenvalue

$\lambda = 16$.

A four-dimensional positional block contains a three-dimensional zero-sum subspace. Each of the three codon positions contributes three independent nucleotide contrasts. The total number of these directions is

$3 \times (4 - 1) = 9$.

Each has singular value

$\sigma = \sqrt{16} = 4$.

These nine directions represent changes among nucleotide states within fixed codon positions. They include contrasts such as G versus A, C versus U, or one nucleotide versus the average of the other three.

Different contrast bases can be selected, but the nine-dimensional contrast space and its singular value are invariant.

2.6 Redundancy Among the Three Positional Blocks

The four columns belonging to the first codon position sum to the sixty-four-dimensional all-ones vector:

$D1G + D1A + D1C + D1U = 1_{64}$.

The same is true at the second and third positions:

$D2G + D2A + D2C + D2U = 1_{64}$,

and

$D3G + D3A + D3C + D3U = 1_{64}$.

The three positional block sums are therefore identical.

Subtracting any two block sums gives zero. For example,

$D1G + D1A + D1C + D1U$

$- D2G - D2A - D2C - D2U$

$= 0$.

A second independent relation can be obtained by comparing either of those blocks with the third.

Equivalently, consider a blockwise constant coefficient vector

$z = (a\, 1_4, b\, 1_4, c\, 1_4)^T$,

where

$a + b + c = 0$.

For every codon row, exactly one coordinate is selected from each block. Therefore,

$D z = (a + b + c)\, 1_{64} = 0$.

These vectors form a two-dimensional right null space.

A convenient basis is

$z_1 = (1_4, -1_4, 0_4)^T$

and

$z_2 = (1_4, 0_4, -1_4)^T$.

The two dependencies do not arise from omitted codons, missing measurements, amino-acid degeneracy, or numerical instability. They arise because the same constant codon field is represented redundantly in all three positional blocks.

The twelve displayed coordinates therefore contain only ten independent directions.

2.7 Rank and Nullity

The complete eigenvalue spectrum of the Gram matrix is

48 with multiplicity 1,

16 with multiplicity 9,

and

0 with multiplicity 2.

The multiplicities sum to twelve, so the list is complete.

The singular values of D are the nonnegative square roots of these eigenvalues:

4 √3 with multiplicity 1,

4 with multiplicity 9,

and

0 with multiplicity 2.

The number of nonzero singular values is ten. Therefore,

rank (D) = 10.

Because D has twelve columns, its right nullity is

nullity (D) = 12 − 10 = 2.

Because D has sixty-four rows, its left nullity is

nullity (D^T) = 64 − 10 = 54.

The fifty-four-dimensional left-null space consists of codon-level fields orthogonal to every additive nucleotide-position coordinate. These directions become important when numerical properties are projected onto the column space of D. They contain the parts of a codon field that cannot be represented as a sum of independent first-, second-, and third-position contributions.

The singular values satisfy the required sum-of-squares identity:

$(4 \sqrt{3})^2 + 9 (4^2) + 2 (0^2)$

$= 48 + 144$

$= 192.$

This equals both $tr (D^T D)$ and $\| D \|^2 F$.

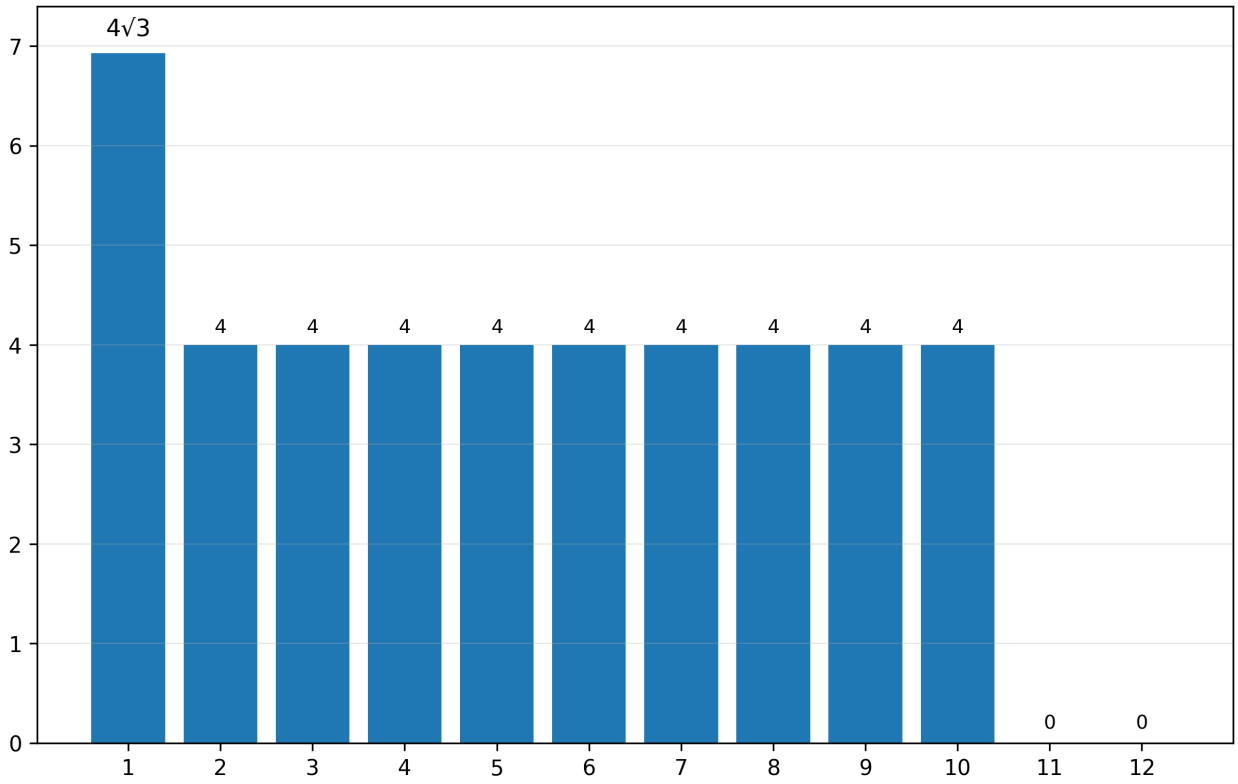


Figure 2. Singular-value spectrum of the complete 64 × 12 codon incidence matrix. The spectrum contains one singular value equal to 4 √3, nine singular values equal to 4, and two singular values equal to 0. The two zero values arise from the equality of the three positional block sums.

2.8 The Meaning of Rank Ten

The ten-dimensional column space has a simple decomposition.

It contains one global constant direction. It also contains three independent nucleotide contrasts at the first codon position, three at the second position, and three at the third position.

Thus,

$10 = 1 + 3 + 3 + 3$

$= 1 + 3 (4 − 1).$

This is the dimension of an additive three-factor model with four levels per factor and one overall constant.

A codon field belonging to this space can be written as

$f (c_1 c_2 c_3)$

$= f_1 (c_1) + f_2 (c_2) + f_3 (c_3).$

The value assigned to a codon is the sum of one contribution from the first position, one from the second position, and one from the third position.

The representation uses twelve displayed coefficients, but those coefficients are not unique. A constant can be moved from one positional block to another without changing any reconstructed codon value, provided that the total shift across the three blocks is zero.

For example, adding the same constant to all four first-position coefficients and subtracting it from all four second-position coefficients leaves every codon sum unchanged.

The codon-level field is unique even when its twelve-coordinate representation is not.

This distinction creates the need for the Moore–Penrose pseudoinverse. The pseudoinverse does not add a biological assumption. It selects one unique minimum-norm coefficient vector from the family of coefficient vectors producing the same additive reconstruction.

2.9 Input Architecture Before Translation

No amino-acid assignments were used to derive the matrix rank, Gram matrix, eigenvalues, or singular values.

These results would remain unchanged if the translated outputs were randomly reassigned among the sixty-four codons. They would remain unchanged if the twenty amino-acid labels

and X were replaced by arbitrary symbols. They would also remain unchanged if no output vector were supplied.

The exact matrix structure belongs to the complete codon input space

Σ^3

and to its one-hot positional encoding.

The translated genetic code enters only when a categorical or numerical output is attached to each row. The amino-acid assignments determine translated fibers, hydropathy fields, synonymous edges, and the quotient multigraph. They do not determine the rank-ten architecture of D.

Two claims must therefore remain distinct.

The first is algebraic: the complete sixty-four-codon input space has an exactly balanced 64 × 12 incidence matrix with rank ten and a fixed singular spectrum.

The second is biological: the standard genetic code places particular translated identities and physicochemical properties over that balanced input space.

The first is forced by the complete factorial construction. The second is where the organization of the genetic code becomes visible.

The next chapter uses this exact rank-ten structure to derive the Moore–Penrose pseudoinverse in closed form.

Chapter 3

The Exact Moore–Penrose Pseudoinverse

The complete codon incidence matrix contains twelve displayed nucleotide-position coordinates but only ten independent directions. Its two-dimensional null space prevents the construction of an ordinary inverse. Nevertheless, every numerical field assigned to the sixty-four codons has a unique closest representation within the additive nucleotide-position subspace.

The Moore–Penrose pseudoinverse supplies that representation.

For a general matrix, a pseudoinverse is often obtained numerically through singular-value decomposition. The complete codon matrix is unusually symmetric. Its pseudoinverse can be written exactly using only the transpose of the original binary matrix and a uniform correction term. Every entry of the resulting 12 × 64 matrix is one of only two rational numbers.

This exact formula belongs to the architecture of the complete codon input space. No amino-acid identities or physicochemical values are required to derive it. Its biological use begins when a translated property field, such as hydropathy, is placed over the sixty-four codons.

3.1 Why a Pseudoinverse Is Required

The codon incidence matrix is

$D \in \{ 0, 1 \}^{64 \times 12}$.

Chapter 2 established that

$\text{rank}(D) = 10$.

Therefore,

$\text{rank}(D^T D) = 10$.

The Gram matrix $D^T D$ is 12 × 12 but has rank ten. It is singular and cannot be inverted in the ordinary sense.

This singularity does not indicate missing codons or incomplete information. All sixty-four codons are present. The rank deficiency arises because the sums of the four columns in each positional block are identical:

$D1G + D1A + D1C + D1U = 1_{64}$,

$D2G + D2A + D2C + D2U = 1_{64}$,

and

$D3G + D3A + D3C + D3U = 1_{64}$.

The same constant codon field is represented independently in all three positional blocks. The twelve displayed coordinates therefore contain two redundant constant directions.

Suppose that a numerical property is assigned to every codon. Collect the sixty-four values in a vector

$h \in \mathbb{R}^{64}$.

We seek twelve nucleotide-position coefficients

$\beta \in \mathbb{R}^{12}$

such that

$D \beta$

approximates h.

When h lies exactly in the column space of D, at least one coefficient vector satisfies

$$D \beta = h.$$

Because D has a two-dimensional right null space, that coefficient vector is not unique. If z belongs to the null space of D, then

$$D z = 0,$$

and therefore

$$D (\beta + z) = D \beta.$$

Infinitely many coefficient vectors can produce the same codon-level reconstruction.

When h does not lie in the column space of D, no exact coefficient vector exists. The appropriate objective is then to minimize the squared residual

$$\| h - D \beta \|^2.$$

Rank deficiency again prevents a unique ordinary least-squares coefficient vector. Several coefficient vectors may produce the same minimum residual.

The Moore–Penrose pseudoinverse resolves both problems. It defines

$$\beta = D^+ h,$$

where

$$D^+ \in \mathbb{R}^{12 \times 64}.$$

The reconstruction is

$$\hat{h} = D D^+ h.$$

This is the unique orthogonal projection of h onto the column space of D. Among all coefficient vectors that produce this reconstruction, $D^+ h$ is the unique vector having minimum Euclidean norm.

The pseudoinverse therefore provides two forms of uniqueness:

the nearest additive codon field is unique;

and the minimum-norm coefficient representation of that field is unique.

3.2 The Four Moore–Penrose Conditions

For any matrix D, its Moore–Penrose pseudoinverse D^{+} is the unique matrix satisfying four equations:

$D\,D^{+}\,D = D$,

$D^{+}\,D\,D^{+} = D^{+}$,

$(\,D\,D^{+}\,)^{T} = D\,D^{+}$,

and

$(\,D^{+}\,D\,)^{T} = D^{+}\,D$.

The first identity states that inserting the pseudoinverse between two copies of D leaves D unchanged.

The second is the corresponding identity for D^{+}.

The third states that

$P = D\,D^{+}$

is symmetric.

The fourth states that

$Q = D^{+}\,D$

is symmetric.

Together, the generalized-inverse conditions and the symmetry conditions make P and Q orthogonal projectors.

The matrix P acts on the sixty-four-dimensional space of codon fields. It projects any codon-level vector onto the ten-dimensional column space of D.

The matrix Q acts on the twelve-dimensional coefficient space. It projects any coefficient vector onto the ten-dimensional row space of D and removes its two null-space components.

For the complete codon matrix, all four conditions can be verified exactly. No numerical tolerance is required when rational arithmetic is used.

3.3 Closed-Form Pseudoinverse

The exact pseudoinverse is

$D^{+} = (\,1 / 16\,)\,D^{T} - (\,1 / 96\,)\,J_{12\times 64}$,

where $J_{12\times64}$ is the 12 × 64 all-ones matrix.

The formula contains two terms.

The first term is a scaled transpose:

$(1 / 16) D^{T}$.

Each nucleotide-position coordinate occurs in exactly sixteen codons. When this term acts on a codon field h, it averages the values of the sixteen codons containing that coordinate.

For coordinate (p, n),

$[(1 / 16) D^{T} h]_{p,n}$

is the mean of h over the sixteen codons having nucleotide n at position p.

If this averaging term were used alone, however, the global mean of h would be represented independently in all three positional blocks. The constant component would be counted three times.

The second term,

$(1 / 96) J_{12\times64}$,

subtracts the same correction from all twelve coefficients. When applied to h, it contributes

$(1 / 96) \Sigma_{c\in C} h_c$

to every coefficient.

An individual coefficient can therefore be written as

$\beta_{p,n}$

$= (1 / 16) \Sigma_{c : c_p = n} h_c$

$- (1 / 96) \Sigma_{c\in C} h_c$.

The first sum is coordinate specific. The second is global.

Let the mean of the complete codon field be

$\mu = (1 / 64) \Sigma_{c\in C} h_c$.

Then

$(1 / 96) \Sigma_{c\in C} h_c = (2 / 3) \mu$.

Therefore,

$\beta_{p,n}$

= mean of h among codons containing coordinate (p, n)

− (2 / 3) μ.

The correction does not remove the overall mean. It distributes that mean symmetrically among the three codon positions.

A constant codon field of value μ is represented by a coefficient μ / 3 in every coordinate, so that the three active coefficients in each codon sum to μ.

The formula is exact:

$D^{+} = (1 / 16) D^{T} - (1 / 96) J_{12 \times 64}$.

It does not depend on a selected amino-acid property, a numerical fitting threshold, or an approximate singular-value cutoff.

3.4 The Two Exact Entry Values

Every entry of D is either 1 or 0. The closed-form expression for D^{+} therefore permits only two entry values.

Suppose the codon represented by a pseudoinverse column contains the nucleotide-position coordinate represented by a pseudoinverse row. The corresponding entry of D^{T} is 1. Therefore,

$D^{+}_{p,n;c}$

= 1 / 16 − 1 / 96

= 6 / 96 − 1 / 96

= 5 / 96.

If the codon does not contain that coordinate, the corresponding entry of D^{T} is 0. Therefore,

$D^{+}_{p,n;c}$

= 0 − 1 / 96

= −1 / 96.

Thus,

$D^{+}_{p,n;c} = 5 / 96$

when codon c contains nucleotide n at position p,

and

$D^{+}p,n;c = -1 / 96$

when it does not.

The entire 12 × 64 pseudoinverse contains only these two rational values.

This two-value structure is a direct consequence of the complete factorial balance of codon space. It is not produced by rounding, thresholding, or graphical simplification.

A pseudoinverse is ordinarily expected to contain a broad range of numerical values. Here, a natural biological input domain produces a pseudoinverse that is exact, closed form, and binary-derived. The mathematical reason is identifiable: every coordinate occurs equally often, every cross-position coordinate pair occurs equally often, and the only column dependencies are the two positional block-sum redundancies.

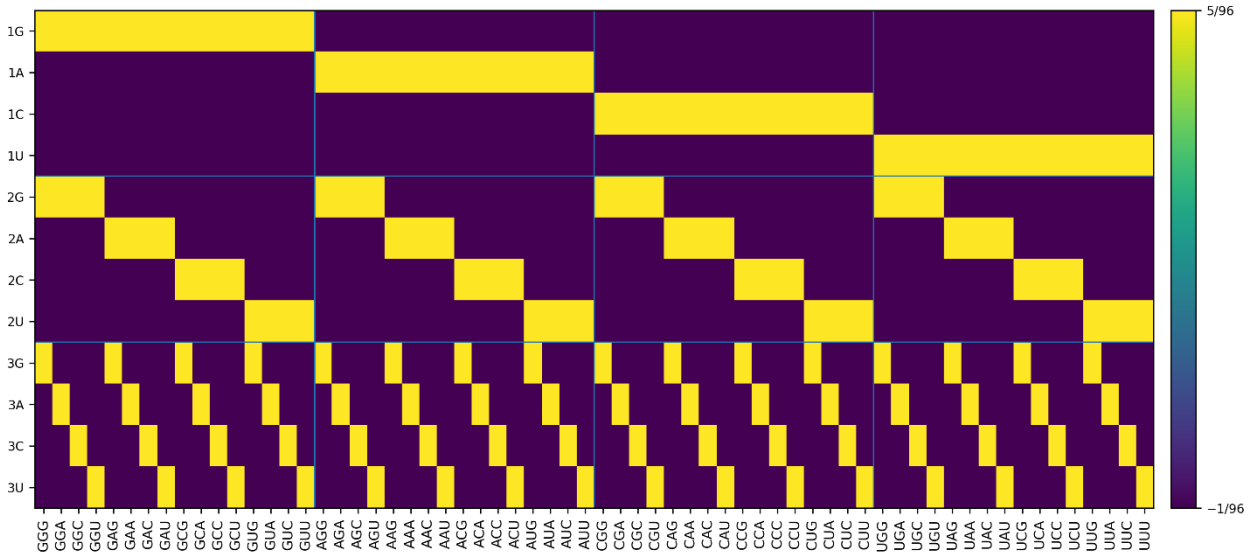


Figure 3. Exact two-value structure of the Moore–Penrose pseudoinverse of the complete codon incidence matrix. The 12 × 64 matrix contains 5 / 96 when the codon represented by a column contains the nucleotide-position coordinate represented by a row, and −1 / 96 otherwise. Every codon column contains exactly three positive entries and nine negative entries.

3.5 Column Structure

Fix one codon c.

That codon contains exactly three active nucleotide-position coordinates: one in each positional block. Its pseudoinverse column therefore contains

three entries equal to 5 / 96

and

nine entries equal to −1 / 96.

The sum of the column is

$3 (5 / 96) + 9 (-1 / 96)$

$= 15 / 96 - 9 / 96$

$= 6 / 96$

$= 1 / 16.$

Every codon column of D^+ therefore sums to

$1 / 16.$

This uniform column sum is another consequence of treating all sixty-four codons symmetrically.

The squared norm of every pseudoinverse column is also identical:

$3 (5 / 96)^2 + 9 (1 / 96)^2$

$= 75 / 9{,}216 + 9 / 9{,}216$

$= 84 / 9{,}216$

$= 7 / 768.$

Every codon is therefore represented by a pseudoinverse column having the same Euclidean norm. The columns differ only in the locations of their three positive entries.

3.6 Row Structure

Fix one nucleotide-position coordinate.

That coordinate occurs in exactly sixteen codons and is absent from forty-eight codons. The corresponding pseudoinverse row therefore contains

sixteen entries equal to 5 / 96

and

forty-eight entries equal to −1 / 96.

The row sum is

16 (5 / 96) + 48 (−1 / 96)

= 80 / 96 − 48 / 96

= 32 / 96

= 1 / 3.

Every row of D^+ therefore sums to

1 / 3.

The squared norm of every pseudoinverse row is

16 (5 / 96)² + 48 (1 / 96)²

= 400 / 9,216 + 48 / 9,216

= 448 / 9,216

= 7 / 144.

All twelve nucleotide-position coordinates are treated equally at the level of the input architecture. Biological differences among their fitted coefficients arise only after a nonconstant output field is supplied.

3.7 Constant Codon Fields

The row-sum identity provides a simple exact test of the pseudoinverse.

Suppose every codon is assigned the same numerical value k:

$h = k\, 1_{64}$.

Then

$\beta = D^+ h$.

Because every row of D^+ sums to 1 / 3,

$\beta = (k / 3)\, 1_{12}$.

Every nucleotide-position coefficient equals k / 3.

Each codon activates exactly three coefficients. Therefore,

$$D\,\beta$$

$$= 3\,(k / 3)\,1_{64}$$

$$= k\,1_{64}$$

$$= h.$$

The constant codon field is reconstructed exactly.

This result explains the role of the uniform correction in the pseudoinverse formula. Without it, the coordinate averages would assign k to every coefficient, and each codon would reconstruct to 3 k rather than k. The subtraction term distributes the constant component correctly across the three positional blocks.

More generally, let h have mean μ. The pseudoinverse preserves that mean exactly:

$$\text{mean}\,(D\,D^{+}\,h) = \mu.$$

The projection may remove nonadditive variation, but it does not alter the global codon-field mean.

3.8 Gauge Freedom

The additive codon model has the form

$$\hat{h}c_1c_2c_3$$

$$= \beta 1,c_1 + \beta 2,c_2 + \beta 3,c_3.$$

The individual coefficients are not unique before a gauge convention is imposed.

Consider three constants a, b, and c satisfying

$$a + b + c = 0.$$

Add a to all four first-position coefficients, b to all four second-position coefficients, and c to all four third-position coefficients.

For every codon, the reconstructed value changes by

$$a + b + c = 0.$$

The codon field remains unchanged.

One simple example is to add a constant δ to all first-position coefficients and subtract δ from all second-position coefficients:

$\beta_1 \to \beta_1 + \delta\, 1_4$,

$\beta_2 \to \beta_2 - \delta\, 1_4$,

$\beta_3 \to \beta_3$.

Every reconstructed codon value remains the same.

A second independent transfer can be made between another pair of blocks. These two independent block-offset transformations form the two-dimensional right null space of D.

The gauge freedom has the precise form

$z = (a\, 1_4, b\, 1_4, c\, 1_4)^T$,

with

$a + b + c = 0$.

Every such vector satisfies

$D\, z = 0$.

The coefficient vectors β and β + z therefore represent the same codon field.

Individual coefficient levels cannot be interpreted independently of the selected gauge. A statement that one coefficient has a particular absolute value is meaningful only after the coefficient convention has been declared.

Within-position contrasts are different. For example,

$\beta 2,U - \beta 2,A$

is unaffected by adding the same constant to all four second-position coefficients. Such contrasts are gauge invariant.

The strong opposition between second-position U and second-position A, developed later in the hydropathy analysis, therefore does not depend on an arbitrary transfer of constants among positional blocks.

3.9 The Minimum-Norm Gauge

The Moore–Penrose pseudoinverse selects the unique coefficient vector orthogonal to the null space of D.

Orthogonality to

$(1_4, -1_4, 0_4)^T$

requires the sum of the first-position coefficients to equal the sum of the second-position coefficients.

Orthogonality to

$(1_4, 0_4, -1_4)^T$

requires the sum of the first-position coefficients to equal the sum of the third-position coefficients.

The Moore–Penrose solution therefore satisfies

$\Sigma n \in \Sigma\ \beta 1,n$

$= \Sigma n \in \Sigma\ \beta 2,n$

$= \Sigma n \in \Sigma\ \beta 3,n.$

The three positional block sums are equal.

Using the closed-form pseudoinverse, each block sum can be calculated directly. Let

$S = \Sigma c \in C\ hc.$

For any positional block, the four coordinate-specific sums together count every codon value once. Therefore,

block sum

$= (1 / 16) S - 4 (1 / 96) S$

$= S / 16 - S / 24$

$= S / 48.$

Every positional block sum equals

$S / 48.$

Since the codon-field mean is

$\mu = S / 64,$

the common block sum is

$S / 48 = 4 \mu / 3.$

The total sum of all twelve coefficients is

$3 (S / 48) = S / 16 = 4 \mu.$

For a centered codon field, S = 0, and each positional coefficient block sums to zero.

This block balance is the symmetric minimum-norm gauge. It does not claim that the three codon positions are biologically equivalent. It states only that the shared constant component has been distributed evenly among the three redundant positional representations.

Among all coefficient vectors generating the same fitted codon field, this balanced vector has the smallest Euclidean norm.

3.10 Projection into Codon Space

The codon-space projection matrix is

$P = D\,D^{+}$.

Substituting the exact pseudoinverse gives

P

$= D\,[\,(1/16)\,D^{T} - (1/96)\,J_{12\times 64}\,]$

$= (1/16)\,D\,D^{T} - (1/96)\,D\,J_{12\times 64}$.

Every row of D contains three 1s. Therefore,

$D\,J_{12\times 64} = 3\,J_{64}$.

Consequently,

$P = (1/16)\,D\,D^{T} - (1/32)\,J_{64}$.

This is an exact 64 × 64 matrix.

It satisfies

$P^{2} = P$

and

$P^{T} = P$.

Therefore, P is an orthogonal projector.

Its rank and trace are

$\mathrm{rank}(P) = \mathrm{tr}(P) = 10$.

For any codon-level property field h,

$\hat{h} = P\,h$

is the unique closest field in the ten-dimensional additive nucleotide-position subspace.

The residual is

$r = h - \hat{h}$

$= (I_{64} - P) h.$

Because P is an orthogonal projector,

$P r = 0$

and

$\hat{h}^T r = 0.$

Therefore,

$\| h \|^2 = \| \hat{h} \|^2 + \| r \|^2.$

The original property field separates exactly into two orthogonal components:

$h = \hat{h} + r.$

The first component is representable through independent nucleotide effects at the three codon positions.

The second contains everything outside that additive model.

The residual is not automatically noise. It may contain interactions among codon positions, translated-fiber effects, nonlinear structure, distinctions among synonymous codons, or features of the selected property scale that cannot be represented by independent positional terms.

The pseudoinverse identifies the additive component without asserting that the residual is unimportant.

3.11 The Hamming-Distance Projector

The matrix $D D^T$ has a direct codon-pair interpretation.

For codons c and d, the entry

$(D D^T)_{c,d}$

is the inner product of their binary incidence rows.

Each codon row has one active coordinate at each of three positions. Their inner product therefore counts the number of codon positions at which the two codons contain the same nucleotide.

Let

$dH (c, d)$

denote their Hamming distance: the number of positions at which they differ.

The number of matching positions is

$3 - dH (c, d)$.

Therefore,

$(D D^{T})_{c,d} = 3 - dH (c, d)$.

Substituting into the projector formula gives

$P_{c,d}$

$= [3 - dH (c, d)] / 16 - 1 / 32$

$= 5 / 32 - dH (c, d) / 16$.

Every projector entry therefore depends only on the Hamming distance between its two codons.

The possible values are:

when $dH = 0$,

$P_{c,c} = 5 / 32$;

when $dH = 1$,

$P_{c,d} = 3 / 32$;

when $dH = 2$,

$P_{c,d} = 1 / 32$;

and when $dH = 3$,

$P_{c,d} = -1 / 32$.

The complete 64×64 projector contains only four distinct values.

This is a second exact low-value structure derived from the complete codon matrix. The pseudoinverse contains two values because each coordinate is either present or absent. The codon-space projector contains four values because two codons can agree at three, two, one, or zero positions.

The diagonal value is constant:

$P_{c,c} = 5 / 32$.

Summing the sixty-four diagonal entries gives

$\mathrm{tr}(P) = 64 (5 / 32) = 10$,

as required.

The row sum of P is one. Therefore, constant fields are fixed by the projection:

$P\, 1_{64} = 1_{64}$.

The Hamming-distance formula also shows that the additive projection is determined entirely by codon-coordinate agreement. It does not use translated amino-acid identity.

3.12 The Coefficient-Space Projector

The second Moore–Penrose projector is

$Q = D^{+} D$.

Substitution gives

Q

$= (1 / 16) D^{T} D - (1 / 96) J_{12 \times 64} D$.

Every column of D contains sixteen 1s. Therefore,

$J_{12 \times 64} D = 16 J_{12}$.

Hence,

$Q = (1 / 16) G - (1 / 6) J_{12}$,

where

$G = D^{T} D$.

The matrix Q is symmetric and idempotent:

$Q^{T} = Q$,

$Q^2 = Q$.

It has rank and trace ten.

On the global constant coefficient direction, Q acts as the identity.

On the nine within-position contrast directions, Q also acts as the identity.

On the two blockwise constant null-space directions, Q acts as zero.

Thus, Q removes precisely the two coefficient components that do not change any codon reconstruction.

If an arbitrary coefficient vector γ is supplied, then

$Q\,\gamma$

is its unique minimum-norm representative after null-space offsets have been removed.

3.13 Direct Verification of the Moore–Penrose Conditions

The exact formulas for P and Q make the Moore–Penrose conditions transparent.

Because Q is the orthogonal projector onto the row space of D,

$D\,Q = D$.

Therefore,

$D\,D^{+}\,D = D$.

Because P is the orthogonal projector onto the column space of D,

$D^{+}\,P = D^{+}$.

Therefore,

$D^{+}\,D\,D^{+} = D^{+}$.

The explicit formulas also show that both projectors are symmetric:

$P^{T} = P$

and

$Q^{T} = Q$.

All four Moore–Penrose equations are satisfied.

The closed-form candidate is therefore not merely a convenient generalized inverse. It is the unique Moore–Penrose pseudoinverse.

3.14 What the Exact Formula Does Not Depend On

The formula

$$D^{+} = (1 / 16) D^{T} - (1 / 96) J_{12 \times 64}$$

does not depend on the standard amino-acid assignments.

It would remain unchanged if:

the codons were mapped to different amino acids;

the amino-acid labels were replaced by arbitrary symbols;

the number of codons assigned to each translated output changed;

or no translated-output vector were included.

The formula depends only on three structural facts:

the alphabet contains four nucleotide states;

the words have length three;

and the matrix includes all 4^3 possible words in one-hot positional form.

The exact pseudoinverse is therefore an input-space theorem.

The standard genetic code becomes relevant when the biological output relation is placed over this exact input architecture. Hydropathy, residue volume, synonymous degeneracy, and translated topology then generate fields and graphs specific to the standard assignment.

This distinction prevents overinterpretation. The existence of the exact pseudoinverse does not, by itself, prove that the biological output mapping was designed, selected, or historically inevitable. It shows that the complete codon domain has an unusually transparent algebraic inverse structure through which biological output fields can be analyzed.

3.15 Why the Result Matters

The complete codon matrix could have been treated as an ordinary numerical design matrix and passed directly to a software pseudoinverse routine. Such a procedure would produce numerical coefficients, but it would conceal the reason for their structure.

The exact derivation reveals that:

the matrix has rank ten for a simple combinatorial reason;

the pseudoinverse has a two-term closed form;

every pseudoinverse entry is either 5 / 96 or −1 / 96;

every codon column contains three positive and nine negative entries;

every coordinate row contains sixteen positive and forty-eight negative entries;

every pseudoinverse column sums to 1 / 16;

every pseudoinverse row sums to 1 / 3;

constant codon fields are reconstructed exactly;

the minimum-norm gauge equalizes the three positional block sums;

and the codon-space projector depends only on Hamming distance.

These are exact properties of the complete codon representation, not approximate numerical observations.

The pseudoinverse turns any sixty-four-component codon field into twelve nucleotide-position coefficients. Part II applies this operation to amino-acid hydropathy. That application introduces biological and empirical choices, including the hydropathy scale and the numerical treatment of translation stop X. The projection machinery itself, however, is already complete.

Part II

Nucleotide-Determined Physicochemical Organization

Chapter 4

Hydropathy as a Codon-Level Property Field

The genetic code maps codons to translated identities. Hydropathy is not assigned directly to codons in the conventional biological description; it is assigned to amino-acid residues. The genetic-code map nevertheless permits an amino-acid property to be transferred back onto codon space.

Once that transfer is made, the sixty-four codons carry a numerical field. The exact pseudoinverse derived in Chapter 3 can then ask how much of that field is representable as the sum of independent nucleotide contributions at the first, second, and third codon positions.

This calculation does not claim that nucleotides possess amino-acid hydropathy. It measures how the standard genetic-code assignment organizes amino-acid hydropathy over its codon inputs.

The distinction is essential. The hydropathy values come from amino-acid chemistry. Their distribution over codon space comes from the genetic code.

4.1 From Amino-Acid Values to a Codon Field

Let

$H : \Omega \rightarrow \mathbb{R}$

be a numerical property assignment over the translated-output classes.

For an amino acid ω, H (ω) is its value on the selected hydropathy scale. The induced codon-level field is

h c = H (τ (c)).

Thus,

$h \in \mathbb{R}^{64}$

contains one numerical value for each codon.

Synonymous codons initially receive the same value because they belong to the same translated fiber. If four codons translate to alanine, all four receive the alanine hydropathy value. If six codons translate to leucine, all six receive the leucine value.

The codon field is therefore constant within every translated fiber:

$$h\,c = h\,d$$

whenever

$$\tau(c) = \tau(d).$$

This property is imposed by construction. It reflects the decision to study translated amino-acid identity rather than codon usage, translation rate, RNA structure, or another codon-specific variable.

The present analysis uses the Kyte–Doolittle hydropathy scale. On that scale, more positive values generally indicate stronger hydrophobic character, while more negative values indicate stronger hydrophilic character.

For the twenty amino acids, H is supplied by the selected empirical scale. Translation stop X requires separate treatment because X is not an amino acid and has no intrinsic residue hydropathy.

The complete calculation used in this book adopts the explicit convention

$$H(X) = 0.$$

Therefore,

$$h_{UAA} = h_{UAG} = h_{UGA} = 0.$$

This convention is not a statement that translation stop chemically resembles a residue of hydropathy zero. It is a numerical choice that permits the complete sixty-four-codon relation to remain intact during the property projection.

Other conventions are possible and should be treated as alternative analyses. The three stop codons could be omitted, treated as missing, or assigned another declared value. Each choice defines a different numerical field.

4.2 The Question Asked by the Projection

The exact pseudoinverse transforms the codon-level hydropathy field into twelve coefficients:

$$\beta = D^{+} h.$$

The coefficient vector contains one value for every nucleotide-position coordinate:

$$\beta =$$

$$(\beta_{1,G}, \beta_{1,A}, \beta_{1,C}, \beta_{1,U},$$

$\beta 2,G, \beta 2,A, \beta 2,C, \beta 2,U,$

$\beta 3,G, \beta 3,A, \beta 3,C, \beta 3,U)^{T}.$

For codon

$c = c_1 c_2 c_3,$

the additive reconstruction is

$\hat{h} c = \beta 1,c_1 + \beta 2,c_2 + \beta 3,c_3.$

The reconstructed value is calculated from the three nucleotides composing the codon. It is not read directly from the amino-acid name.

The projection therefore asks:

How closely can the codon-level hydropathy field generated by the standard genetic code be represented by independent contributions from the nucleotide at position 1, the nucleotide at position 2, and the nucleotide at position 3?

The model is additive. It does not include interactions such as a special joint effect of first-position G and second-position U. Such interactions remain in the projection residual.

The exact coefficient formula from Chapter 3 is

$\beta p,n$

$= (1 / 16) \Sigma c : c_p = n \; h \; c$

$- (1 / 96) \Sigma c \in C \; h \; c.$

Equivalently,

$\beta p,n$

= the mean hydropathy of the sixteen codons containing nucleotide n at position p

− two-thirds of the global codon-field mean.

This expression gives each coefficient a direct descriptive interpretation. It begins with the mean property of one nucleotide-position family and then applies the uniform correction required by the minimum-norm gauge.

4.3 Historical Origins of the Calculation

The nucleotide-determined hydropathy analysis originated in work on genetic coding algorithms for engineering membrane proteins.

A conventional membrane-protein analysis begins with a nucleotide sequence, translates it into amino acids, assigns a hydropathy value to each residue, and averages those values within a sliding window. The resulting profile is used to identify extended hydrophobic regions that may span a membrane.

The earlier work asked whether the coding sequence itself already contained a compressed hydropathy signal.

The genetic code suggested that it might. Amino acids encoded by U at the second codon position are often hydrophobic, whereas amino acids encoded by A at the second position are often hydrophilic or charged. This broad partition is visible in the standard code table before any matrix calculation is performed.

The 1990 Yang–Coleman–Youvan analysis formalized that observation by applying singular-value decomposition to a codon-property matrix constructed from the sixty-one sense codons. The three stop codons were omitted because the numerical property under study was amino-acid hydropathy.

The historical calculation therefore used:

sixty-one codon observations;

twenty amino-acid hydropathy values;

twelve nucleotide-position coordinates;

and a numerical singular-value decomposition.

It identified the second codon position as the dominant carrier of hydropathy information.

The historical analysis was not merely an abstract regression. It was connected to the practical problem of recognizing and engineering hydrophobic segments in membrane proteins.

4.4 The Historical 61-Sense-Codon Result

The historical coefficients are most easily compared after converting DNA T notation to RNA U notation.

Written in the G–A–C–U coordinate order used in the present book, the approximate historical coefficients were:

First codon position:

G = 0.76

A = −0.57

C = −1.18

U = 0.52

Second codon position:

G = −1.48

A = −2.92

C = −0.09

U = 4.02

Third codon position:

G = −0.50

A = −0.35

C = 0.19

U = 0.19

The largest coefficient is second-position U:

$\beta_{2,U} \approx 4.02$.

The most negative coefficient is second-position A:

$\beta_{2,A} \approx -2.92$.

Their contrast is approximately

$\beta_{2,U} - \beta_{2,A} \approx 6.94$.

The third-position coefficients are much smaller. Their total range is approximately

0.69,

from −0.50 to 0.19.

The historical result therefore contains a clear hierarchy:

the second position carries the strongest hydropathy polarity;

the first position provides moderate refinement;

and the third position contributes comparatively little.

This is the central historical result retained in the present book.

4.5 The Complete 64-Codon Formulation

The complete formulation differs in two ways.

First, it retains all sixty-four codons, including UAA, UAG, and UGA.

Second, it uses the exact Moore–Penrose pseudoinverse of the balanced 64 × 12 matrix rather than a numerical pseudoinverse of a 61-row matrix.

With

$H(X) = 0$,

the coefficient vector is calculated exactly by

$\beta = D^{+} h$,

where

$D^{+} = (1/16) D^{T} - (1/96) J$.

The resulting coefficients, in the coordinate order G–A–C–U within each positional block, are:

First codon position:

$\beta_{1,G} = 0.686458$

$\beta_{1,A} = -0.638542$

$\beta_{1,C} = -1.251042$

$\beta_{1,U} = 0.880208$

Second codon position:

$\beta_{2,G} = -1.469792$

$\beta_{2,A} = -2.638542$

$\beta_{2,C} = -0.163542$

$\beta_{2,U} = 3.948958$

Third codon position:

$\beta_{3,G} = -0.388542$

$\beta_{3,A} = -0.169792$

$\beta_{3,C} = 0.117708$

$\beta_{3,U} = 0.117708$

The three positional block sums are equal:

$\Sigma_n \beta_{1,n}$

$= \Sigma_n \beta_{2,n}$

$= \Sigma_n \beta_{3,n}$

$= -0.322918.$

This equality is required by the Moore–Penrose minimum-norm gauge. It is an internal verification of the coefficient calculation.

Codon position	G	A	C	U
1	0.686458	−0.638542	−1.251042	0.880208
2	−1.469792	−2.638542	−0.163542	3.948958
3	−0.388542	−0.169792	0.117708	0.117708

Table 2. Exact nucleotide-position hydropathy coefficients obtained from the complete 64-codon matrix, the Kyte–Doolittle amino-acid hydropathy scale, the convention $H(X) = 0$, and the exact Moore–Penrose pseudoinverse. Coordinates are ordered G–A–C–U within each codon position.

4.6 The Dominant Second-Position Polarity

The complete-code coefficients reproduce the central historical pattern.

The largest positive coefficient is

$\beta_{2,U} = 3.948958.$

The most negative coefficient is

$\beta_{2,A} = -2.638542.$

Their difference is

$\beta_{2,U} - \beta_{2,A}$

= 3.948958 − (−2.638542)

= 6.587500.

This is the largest within-position nucleotide contrast in the complete coefficient vector.

The second-position G coefficient is also negative:

β2,G = −1.469792.

The second-position C coefficient is much closer to zero:

β2,C = −0.163542.

The four second-position coefficients therefore organize the nucleotide states approximately as

U strongly hydrophobic;

C near the middle;

G hydrophilic;

A strongly hydrophilic.

This is not a complete classification of amino-acid chemistry. It is the ordering produced when Kyte–Doolittle hydropathy is projected onto the additive nucleotide-position model.

The first-position coefficient range is

0.880208 − (−1.251042)

= 2.131250.

The second-position range is

3.948958 − (−2.638542)

= 6.587500.

The third-position range is

0.117708 − (−0.388542)

= 0.506250.

The second-position range is therefore more than three times the first-position range and approximately thirteen times the third-position range.

The positional hierarchy is unmistakable.

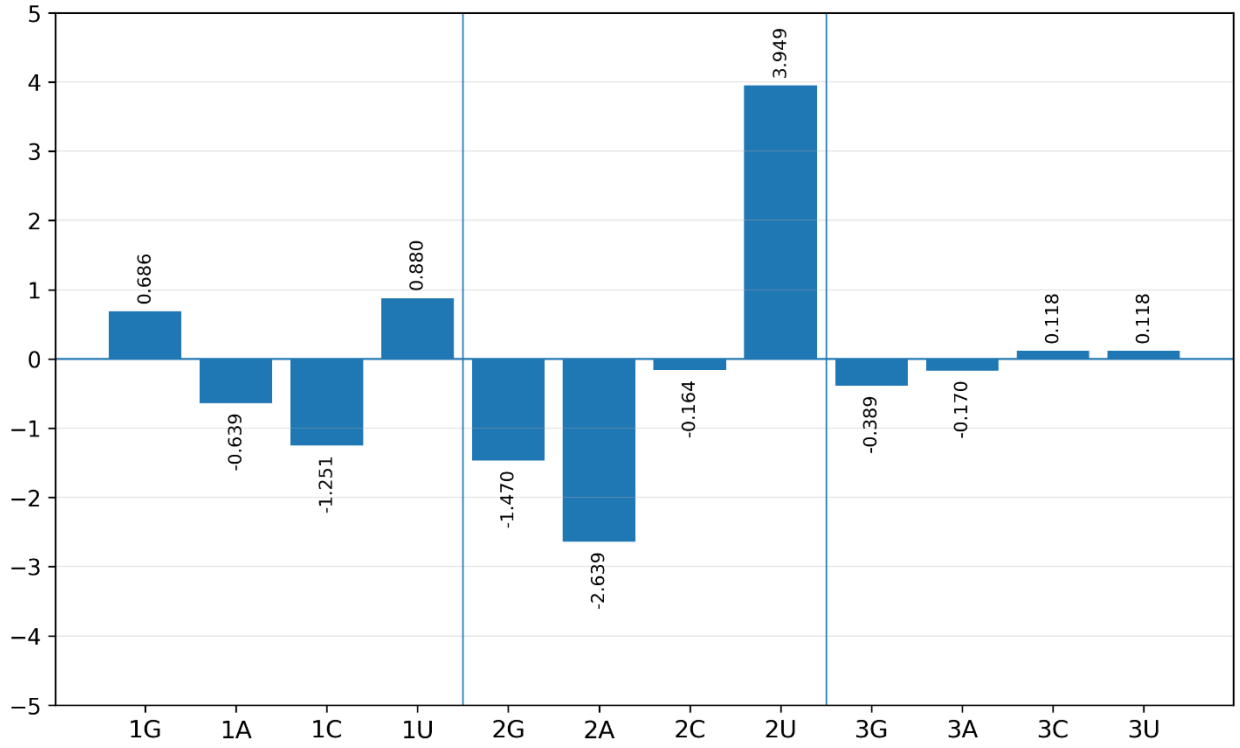


Figure 4. Exact complete-code nucleotide-position hydropathy coefficients. The dominant feature is the strong opposition between second-position U and second-position A. The third-position coefficients occupy a much narrower range.

4.7 Comparison with the Historical Result

The historical and complete calculations are not identical computational objects.

The historical analysis:

used sixty-one sense codons;

omitted stop;

used a numerical singular-value decomposition;

and arose within a membrane-protein engineering calculation.

The complete analysis:

uses all sixty-four codons;

assigns hydropathy zero to X;

uses the balanced complete factorial matrix;

and applies its exact closed-form pseudoinverse.

Despite these differences, the principal coefficients agree closely in pattern and magnitude.

Historical second-position U:

approximately 4.02.

Complete second-position U:

3.948958.

Historical second-position A:

approximately −2.92.

Complete second-position A:

−2.638542.

Historical second-position G:

approximately −1.48.

Complete second-position G:

−1.469792.

Historical second-position C:

approximately −0.09.

Complete second-position C:

−0.163542.

Both calculations therefore identify the same principal polarity:

U at the second codon position is strongly associated with hydrophobic translated outputs;

A at the second codon position is strongly associated with hydrophilic translated outputs;

G is moderately negative;

and C is near the center of the fitted scale.

The purpose of the complete formulation is not to displace the historical result. It supplies an exact algebraic setting in which the same biological pattern can be calculated using the entire codon relation.

4.8 Why the Third-Position Effects Are Small

The third codon position is frequently associated with synonymous variation.

In a fourfold-degenerate translated family, the first two nucleotides remain fixed while the third nucleotide cycles through G, A, C, and U. All four codons receive the same amino-acid hydropathy value.

The additive model therefore observes all four third-position states while the output property remains unchanged. It has little reason to assign large hydropathy differences among those states.

The complete third-position coefficients are

G = −0.388542,

A = −0.169792,

C = 0.117708,

and

U = 0.117708.

Their small range does not mean that the third codon position is biologically unimportant. The third position can affect synonymous codon usage, translation kinetics, RNA structure, mutation probability, and regulation.

The narrower statement is mathematical:

When the standard codon-level Kyte–Doolittle field is projected onto independent nucleotide-position coordinates, relatively little of the fitted hydropathy contrast is assigned to the third position.

The hydropathy projection and the later topology analysis reveal complementary aspects of the same organization.

The hydropathy calculation finds relatively small property contrasts at the third position.

The G–A–C–U cube finds that most selected synonymous nearest-neighbor edges occur along the third codon axis.

A position that often changes without changing translated identity will also tend to contribute less to a property inherited from translated identity.

4.9 First-Position Refinement

The first codon position carries moderate structure.

Its coefficients are

G = 0.686458,

A = −0.638542,

C = −1.251042,

and

U = 0.880208.

The first position does not generate the principal hydropathy opposition, but it helps distinguish amino acids within broad second-position families.

Codons sharing the same second nucleotide can encode residues spanning more than one hydropathy value. The first nucleotide contributes part of that refinement.

This can be understood hierarchically.

The second position establishes a broad physicochemical division.

The first position helps separate translated outputs within that division.

The third position often varies while translated identity remains unchanged.

The hierarchy is not absolute. Particular codons and amino acids do not always follow a simple second-position rule. The coefficient vector summarizes a global least-squares structure over all sixty-four codons.

4.10 What the Coefficients Mean

A coefficient such as

$\beta_{2,U} = 3.948958$

does not mean that uracil itself possesses a hydropathy of 3.948958.

Nor does it mean that inserting U into the second codon position chemically makes an amino acid hydrophobic.

The coefficient has a relational meaning.

It describes the contribution assigned to second-position U in the minimum-norm additive representation of the codon-level hydropathy field generated by the standard genetic code.

The result depends jointly on:

the standard codon-to-output assignments;

the selected hydropathy scale;

the inclusion of all sixty-four codons;

the convention $H(X) = 0$;

the one-hot nucleotide-position representation;

and the Moore–Penrose minimum-norm gauge.

Changing one of these elements may change the coefficient values.

Some quantities are nevertheless more robust than individual coefficient levels. Within-position contrasts are gauge invariant. In particular,

$$\beta_{2,U} - \beta_{2,A} = 6.587500$$

cannot be altered by transferring a common offset among the three positional blocks.

The large U–A contrast is therefore not an artifact of the two-dimensional coefficient gauge.

4.11 What the Coefficients Do Not Show

The twelve coefficients do not reconstruct every detail of amino-acid hydropathy exactly.

The additive model contains only ten independent dimensions. The codon-level hydropathy field is one vector in a sixty-four-dimensional space. Its projection preserves the component aligned with independent nucleotide-position effects and leaves a residual outside that subspace.

The coefficients do not directly represent:

interactions among codon positions;

special behavior of individual codon boxes;

nonadditive distinctions among translated families;

codon-usage frequencies;

organism-specific translation effects;

or properties unrelated to the selected hydropathy scale.

The coefficient vector is therefore not a replacement for the genetic-code table. It is a compressed additive summary of one numerical field placed over that table.

The quality and limits of that compression become clearer when the fitted codon values are returned to the translated amino-acid classes. Chapter 5 examines that reconstruction and compares the resulting amino-acid estimates with their empirical Kyte–Doolittle values.

Chapter 5

Reconstructing Amino-Acid Hydropathy

The twelve nucleotide-position coefficients obtained in Chapter 4 define an additive hydropathy field over all sixty-four codons. The next question is whether that field retains the recognizable hydropathy ordering of the twenty amino acids from which the original codon-level target vector was constructed.

The reconstruction proceeds in two stages.

First, each codon receives a fitted hydropathy value calculated only from its three nucleotide-position coordinates. Second, the fitted values of synonymous codons are averaged within each translated fiber to obtain one reconstructed value for each amino acid.

The resulting twenty reconstructed values can then be compared with the twenty empirical Kyte–Doolittle hydropathy values.

This comparison does not test whether nucleotide letters are physically equivalent to amino-acid side chains. It tests how much of the translated hydropathy field is aligned with the additive coordinate structure of codon space.

5.1 Projection Through the Exact Pseudoinverse

Let

$h \in \mathbb{R}^{64}$

be the complete codon-level hydropathy vector.

For each sense codon c,

$h\,c = H(\tau(c))$,

where H is the Kyte–Doolittle hydropathy assignment.

For the three stop codons, the declared complete-code convention is

$H(X) = 0.$

The nucleotide-position coefficient vector is

$\beta = D^{+} h.$

The fitted codon field is

$\hat{h} = D \beta$

$= D D^{+} h.$

Because

$P = D D^{+}$

is an orthogonal projector,

$\hat{h} = P h$

is the unique field in the additive nucleotide-position subspace closest to h in Euclidean distance.

The residual is

$r = h - \hat{h}.$

The fitted field and residual are orthogonal:

$\hat{h}^{T} r = 0.$

Therefore,

$\| h \|^2 = \| \hat{h} \|^2 + \| r \|^2.$

The projection does not merely generate twelve coefficients. It decomposes the original codon property field into two mathematically distinct components:

an additive component determined by independent nucleotide-position effects;

and a residual component outside that model.

5.2 Codon-Level Reconstruction

For codon

$c = c_1 c_2 c_3,$

the reconstructed hydropathy is

$\hat{h}\, c$

$= \beta 1,c_1 + \beta 2,c_2 + \beta 3,c_3.$

No amino-acid letter appears in this reconstruction formula.

The amino-acid property entered earlier, when the target vector h was constructed. After the coefficients have been obtained, the fitted value of a codon is calculated only from the nucleotide occupying each of its three positions.

For example, the reconstructed value of GAC is

$\hat{h}$ GAC

$= \beta 1,G + \beta 2,A + \beta 3,C.$

Using the complete-code coefficients,

$\hat{h}$ GAC

$= 0.686458 - 2.638542 + 0.117708$

$= -1.834376.$

GAC translates to aspartate, D, whose empirical Kyte–Doolittle value is −3.5. The reconstructed value is less negative because the additive codon-coordinate model captures only the component of the codon field aligned with independent positional effects.

For UUU,

$\hat{h}$ UUU

$= \beta 1,U + \beta 2,U + \beta 3,U$

$= 0.880208 + 3.948958 + 0.117708$

$= 4.946874.$

UUU translates to phenylalanine, whose empirical value is 2.8. The fitted value is more positive than the empirical value.

For AUG,

$\hat{h}$ AUG

$= \beta 1,A + \beta 2,U + \beta 3,G$

$= -0.638542 + 3.948958 - 0.388542$

$= 2.921874.$

AUG translates to methionine, whose empirical value is 1.9.

These examples show that the reconstruction is not a lookup of the original amino-acid value. It is a coordinate-based estimate formed from the global additive structure of the complete codon field.

5.3 Synonymous Codons Need Not Reconstruct Identically

Synonymous codons receive the same empirical target value because they translate to the same amino acid. They need not receive the same reconstructed value.

Two synonymous codons can contain different nucleotide-position coordinates. The additive model assigns a fitted value from those coordinates, not from the translated label.

Phenylalanine provides a simple case. Its two codons are UUU and UUC. The complete-code coefficients for 3U and 3C are equal:

$\beta_{3,U} = \beta_{3,C} = 0.117708.$

The two phenylalanine codons therefore receive the same reconstruction:

$\hat{h}_{UUU} = \hat{h}_{UUC} = 4.946875.$

Isoleucine has three codons:

AUU, AUC, and AUA.

AUU and AUC reconstruct to

3.428125,

whereas AUA reconstructs to

3.140625.

Their empirical amino-acid value is the same, but the third-position A coefficient differs from the third-position U and C coefficients.

Serine shows a much larger internal spread because its six codons occupy two separated regions of codon space. The UCN codons reconstruct between

0.328125

and

0.834375,

whereas AGU and AGC each reconstruct to

−1.990625.

All six codons share the empirical serine hydropathy value of −0.8. Their fitted values differ because the additive coordinates distinguish the two portions of the serine fiber.

This variation is not an error in the calculation. It is one of the consequences of replacing translated identity with an additive coordinate model.

5.4 Averaging Within Synonymous Fibers

To compare reconstructed values with the twenty empirical amino-acid values, the codon reconstructions are averaged within each synonymous fiber.

For amino acid ω, define

$\hat{H}(\omega)$

$= (1 / d\omega) \Sigma_{c \in F\omega} \hat{h}_c.$

The quantity $\hat{H}(\omega)$ is the fiber-averaged nucleotide reconstruction of the amino-acid hydropathy.

Every amino acid contributes one observation to the comparison, regardless of the number of synonymous codons in its fiber. Leucine, with six codons, and methionine, with one codon, each contribute one amino-acid-level point.

This equal weighting differs from the original codon-level projection. In the codon vector h, a six-codon amino acid occurs six times, whereas a one-codon amino acid occurs once. The amino-acid scatter plot instead compares twenty equally weighted translated classes after the projection has been completed.

The reconstructed values include:

F = 4.946875;

V = 4.554688;

I = 3.332292;

L = 3.261458;

M = 2.921875;

A = 0.442188;

S = −0.239583;

C = −0.471875;

G = −0.864063;

T = −0.882813;

W = −0.978125;

P = −1.495313;

Y = −1.640625;

D = −1.834375;

E = −2.231250;

R = −2.663542;

N = −3.159375;

K = −3.556250;

H = −3.771875;

and

Q = −4.168750.

These values reproduce the broad empirical ordering but not every individual amino-acid value.

5.5 Empirical and Reconstructed Hydropathy

Let the empirical Kyte–Doolittle value be the horizontal variable x and the fiber-averaged nucleotide reconstruction be the vertical variable y.

Using the twenty amino acids with equal weight, the audited complete-code regression is

y = −0.003660 + 0.859718 x.

The intercept is very close to zero. The slope is less than one, indicating that the fitted relation is somewhat compressed relative to the empirical scale.

The Pearson correlation is

r = 0.915646.

The corresponding coefficient of determination is

R^2 = 0.838407.

Thus, approximately 83.8 percent of the variance among the twenty reconstructed amino-acid values is aligned with their linear relationship to the empirical Kyte–Doolittle values.

This statistic describes the amino-acid-level comparison after synonymous-fiber averaging. It is not identical to the fraction of codon-level variation captured by the orthogonal projection.

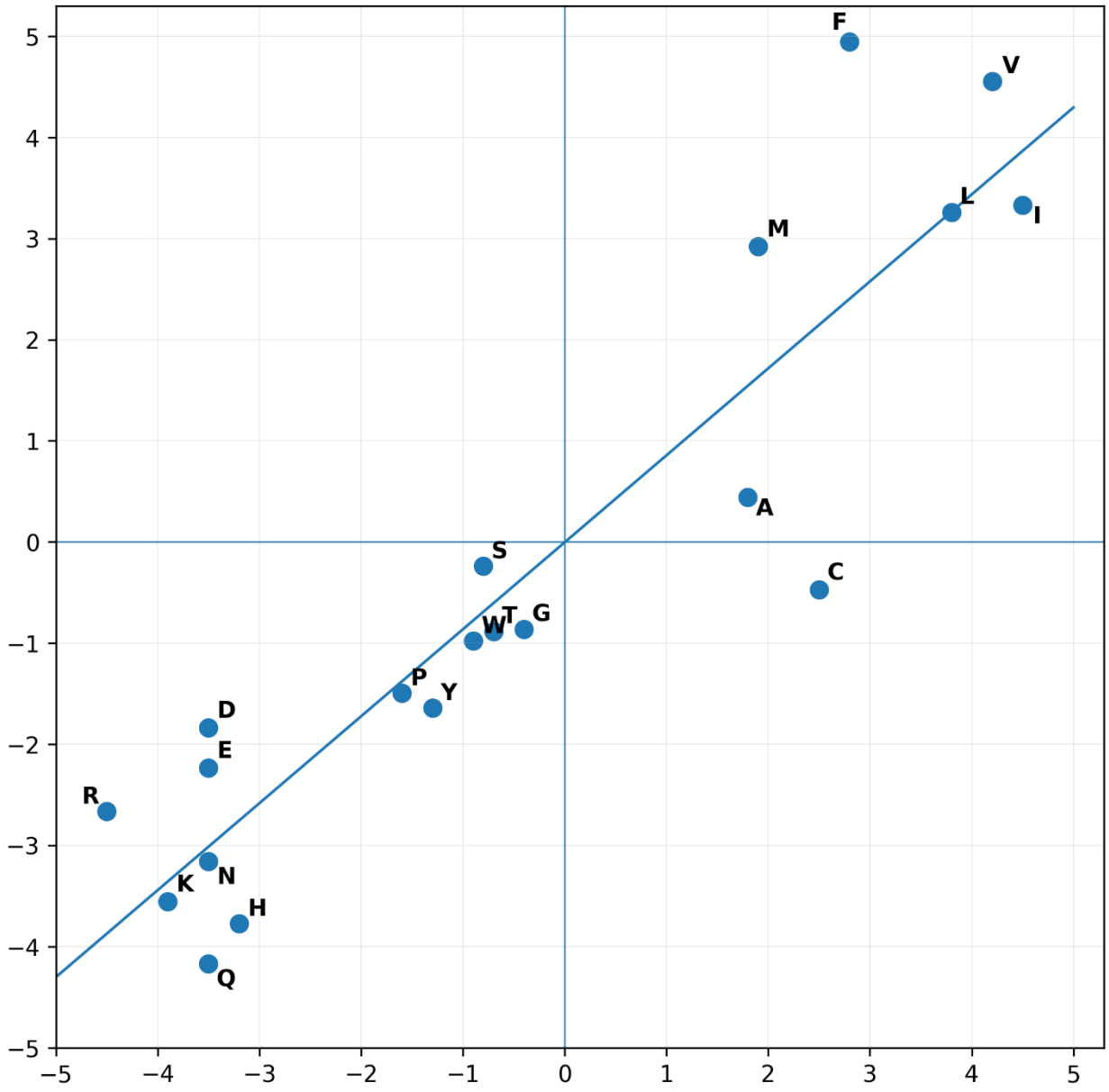


Figure 5. Empirical and nucleotide-reconstructed amino-acid hydropathy. The horizontal axis gives the empirical Kyte–Doolittle value, and the vertical axis gives the fiber-averaged reconstruction obtained from the complete 64-codon pseudoinverse calculation with H (X) = 0. Each of the twenty amino acids is weighted equally. The fitted relationship is y = −0.003660 + 0.859718 x, with Pearson correlation r = 0.915646 and R^2 = 0.838407.

The upper-right region contains the strongly hydrophobic amino acids. Valine, leucine, isoleucine, phenylalanine, and methionine all remain on the hydrophobic side of the reconstructed scale.

The lower-left region contains many strongly hydrophilic or charged amino acids. Arginine, lysine, asparagine, glutamine, histidine, glutamate, and aspartate remain negative after reconstruction.

The broad separation is therefore preserved.

The deviations from the fitted line are equally informative.

Phenylalanine reconstructs substantially more hydrophobically than its empirical value. Cysteine reconstructs substantially less hydrophobically. Alanine and isoleucine also lie below the value expected from an exact identity line, while arginine and aspartate exhibit substantial negative residuals when empirical values are compared with their reconstructions.

The scatter is not a defect to be hidden. It identifies the limits of the additive nucleotide-position model.

5.6 Regression Is Not Causation

A correlation near 0.916 is strong, but it does not establish that nucleotide composition physically causes amino-acid hydropathy.

The direction of construction is important.

The empirical amino-acid values were first transferred through the standard genetic code to construct a codon-level field. That field was then projected onto nucleotide-position coordinates. The comparison measures how strongly the translated property field aligns with the structure of the lookup relation.

The calculation therefore supports the statement:

The standard genetic code organizes much of amino-acid hydropathy in a form recoverable from additive nucleotide-position coordinates.

It does not support the stronger statement:

The nucleotides chemically determine most of amino-acid hydropathy.

Amino-acid hydropathy arises from side-chain chemistry. The mathematical result concerns the way that chemistry is distributed across the codon assignments.

The distinction between causal chemistry and coding alignment should remain explicit throughout the book.

5.7 Codon-Level Projection Strength

The amino-acid scatter plot is intuitive, but the orthogonal projection permits a more direct measure at the codon level.

Let the centered target field be

$hc = h - \mu 1_{64}$,

where μ is the mean of the complete sixty-four-codon field.

Because the projection preserves the mean, the centered fitted field is

$P hc$.

The fraction of centered squared codon-level variation captured by the additive subspace is

$Qadd$

$= \| P hc \|^2 / \| hc \|^2$.

For the complete Kyte–Doolittle field with $H (X) = 0$,

$Qadd \approx 0.816882$.

Thus, approximately 81.7 percent of the centered codon-level squared variation lies in the additive nucleotide-position subspace.

This measure differs from the amino-acid-level R^2 for several reasons.

The codon-level projection weights each codon equally. Amino acids with more synonymous codons therefore contribute more observations.

The amino-acid regression averages within translated fibers and gives each of the twenty amino acids equal weight.

The codon-level statistic is derived directly from an orthogonal projection. The amino-acid statistic is derived from a regression between two sets of twenty class-level values.

Their numerical similarity is useful, but they should not be treated as the same quantity.

5.8 Residual Structure

The residual field is

$r = h - D D^{+} h$.

It is orthogonal to every column of D:

$D^T r = 0$.

No additional adjustment of the same twelve additive nucleotide-position coordinates can reduce this residual. The pseudoinverse has already extracted the complete component available to that model.

The residual should not automatically be called noise.

It can contain chemical distinctions that the additive model cannot express, including:

side-chain charge;

aromaticity;

sulfur chemistry;

branching;

hydrogen-bonding capacity;

molecular geometry;

and interactions among codon positions.

At the amino-acid level, define the fiber-averaged residual as

$R(\omega) = H(\omega) - \hat{H}(\omega)$.

Some of the largest negative residuals are:

$F = -2.146875$;

$R = -1.836458$;

$D = -1.665625$;

and

$E = -1.268750$.

Some of the largest positive residuals are:

$C = 2.971875$;

$A = 1.357813$;

$I = 1.167708$;

and

$Q = 0.668750$.

The sign convention here is empirical minus reconstructed. A positive value means that the empirical amino-acid hydropathy is more positive than the nucleotide reconstruction. A negative value means that the reconstruction is more positive than the empirical value.

These residuals identify translated classes whose chemistry is not well summarized by independent nucleotide-position effects.

Cysteine is a conspicuous example. Its empirical hydropathy is positive, but its reconstructed value is slightly negative. Phenylalanine shows the opposite pattern: its reconstruction is much more positive than its empirical value.

Such cases do not negate the strong global relationship. They show where the global compression loses amino-acid-specific information.

5.9 Fiber Geometry and Reconstruction Error

The spread of reconstructed codon values within one translated fiber provides another form of residual information.

A fiber whose codons occupy one compact path may receive relatively similar fitted values. A fiber divided across distant codon regions may receive a broader range.

Serine again provides the clearest example. The UCN and AGY codon groups receive substantially different additive reconstructions even though they share one translated output.

Leucine also spans more than one codon region. Its reconstructed codon values range from

2.309375

to

4.659375.

Arginine ranges from

−3.109375

to

−2.278125.

The amino-acid-level average suppresses these internal differences. That suppression is appropriate when comparing one reconstructed value with one empirical amino-acid value, but the codon-level variation should remain available for structural analysis.

The projection therefore produces two related objects:

a sixty-four-component fitted codon field;

and a twenty-component fiber-averaged amino-acid reconstruction.

Neither should be substituted silently for the other.

5.10 Additive Information and Unrepresented Interactions

The success of the reconstruction demonstrates that a large component of hydropathy is organized additively across the three codon positions.

The second position supplies the dominant broad polarity.

The first position provides substantial refinement.

The third position contributes comparatively small adjustments.

This hierarchy permits a compact representation of the property field. Twelve displayed coefficients, with ten independent dimensions, recover much of the hydropathy ordering across twenty amino acids and sixty-four codons.

The residual shows that the compression is incomplete.

A more elaborate model could include pairwise interactions between codon positions. For example, it could assign a special contribution to the joint occurrence of first-position G and second-position U. A fully saturated model could reconstruct every codon value exactly, but it would lose the explanatory economy of the additive representation.

The relevant question is not whether a model with enough parameters can reproduce the table. It can.

The relevant question is how much recognizable physicochemical organization is retained by a small and structurally natural coordinate system.

The answer is substantial. The complete additive model captures approximately 81.7 percent of centered codon-level squared variation and yields an amino-acid-level correlation of approximately 0.916.

The next chapter places hydropathy beside residue volume and shows that the second codon position organizes not merely one numerical property but a broader physicochemical arrangement.

Chapter 6

Hydropathy and Residue Volume

The pseudoinverse analysis identifies the second codon position as the strongest additive organizer of amino-acid hydropathy. That conclusion can also be examined without beginning from the twelve fitted coefficients. The sixty-four codons can be divided directly according to the nucleotide occupying their middle position, producing four equal codon families.

These families occupy different regions of amino-acid physicochemical space.

Hydropathy supplies one axis of comparison. Residue volume supplies another. Together, they reveal a broad organization that is less visible when either property is considered alone.

The result is not a rigid classification in which the middle nucleotide uniquely determines amino-acid chemistry. The first and third nucleotides remain essential to translated identity, and chemically unusual residues occur within several groups. The second nucleotide instead acts as a coarse physicochemical organizer. It places a codon within a broad region of property space before the remaining nucleotides refine the assignment.

6.1 A Two-Property Representation

Hydropathy describes the tendency of an amino-acid side chain to favor a nonpolar rather than an aqueous environment. It is especially informative for membrane proteins, buried protein interiors, solvent exposure, and amphipathic structure.

Residue volume describes a different property: the approximate spatial size occupied by an amino-acid residue.

The two quantities are related only partially. A large residue need not be hydrophobic, and a small residue need not be hydrophilic. Two amino acids may have similar hydropathy values while differing substantially in size. Conversely, two residues of similar size may have very different polarity, charge, or environmental preference.

A two-dimensional property plane therefore distinguishes relationships that cannot be seen on a single hydropathy axis.

Let an amino acid ω be represented by the ordered pair

$$(H\omega, V\omega),$$

where $H\omega$ is its hydropathy and $V\omega$ is its residue-volume value.

The hydropathy coordinate used here is the Kyte–Doolittle value. The historical volume comparison retains the residue-volume values used in the source analysis, based on the solution-volume compilation associated with Zamyatnin.

Translation stop X is not plotted as an amino acid because it has neither an amino-acid side chain nor a residue volume. This differs from the complete codon-level pseudoinverse calculation, where X was assigned the declared numerical hydropathy convention

$H(X) = 0$.

The two analyses therefore have different purposes.

The complete pseudoinverse calculation retains all sixty-four codons and requires a numerical convention for X.

The hydropathy–volume plot compares the twenty amino-acid residue classes and omits X because no residue exists to place in the property plane.

6.2 The Four Second-Position Codon Families

Using N to denote any of the four RNA nucleotides, the complete codon set separates into four middle-position families:

NGN,

NAN,

NCN,

and

NUN.

The notation NUN, for example, means that the second nucleotide is U while the first and third positions may contain any nucleotide.

Each family contains

$4 \times 1 \times 4 = 16$

codons.

Therefore,

$|\mathrm{NGN}| = |\mathrm{NAN}| = |\mathrm{NCN}| = |\mathrm{NUN}| = 16.$

Together, the four groups contain

$4 \times 16 = 64$

codons.

This is an exact partition of codon space. Every codon belongs to one and only one family because every codon contains exactly one nucleotide at its second position.

The translated-output groups are not equally simple. Different codons within one family may translate to the same amino acid, and one translated class may occasionally appear in more than one second-position family. Serine is the principal example. The UCN serine codons belong to NCN, whereas AGU and AGC belong to NGN.

Translation stop also appears in two families. UAA and UAG belong to NAN, while UGA belongs to NGN.

Thus, the four groups are disjoint at the codon level but not necessarily at the translated-class level.

Their physicochemical differences nevertheless remain pronounced.

6.3 NUN and Hydrophobic Residues

The NUN family contains codons whose second nucleotide is U.

Its translated amino acids are dominated by

phenylalanine,

leucine,

isoleucine,

methionine,

and valine.

These are all hydrophobic residues on the Kyte–Doolittle scale.

Their values are:

F = 2.8,

L = 3.8,

I = 4.5,

M = 1.9,

and

V = 4.2.

The entire family therefore occupies the positive hydropathy region.

This unusually coherent grouping is the code-table counterpart of the large positive pseudoinverse coefficient

β2,U = 3.948958.

The coefficient calculation and the direct family classification are two views of the same organization.

The coefficient summarizes the complete codon field within an additive model.

The family classification shows the translated residue identities associated with middle-position U.

The middle U does not identify one particular hydrophobic amino acid. The first and third nucleotides distinguish among phenylalanine, leucine, isoleucine, methionine, and valine. The second-position U nevertheless places the codon in a strongly hydrophobic translated region before that refinement occurs.

The NUN family is therefore a coarse hydrophobic sector of the standard genetic code.

6.4 NAN and Hydrophilic Residues

The NAN family contains codons whose second nucleotide is A.

Its amino-acid outputs include

aspartate,

glutamate,

asparagine,

lysine,

histidine,

glutamine,

and tyrosine.

The UA codon box also contains the stop codons UAA and UAG.

Most of the amino acids in this family are polar, charged, or hydrophilic. Their Kyte–Doolittle values are:

D = −3.5,

E = −3.5,

N = −3.5,

K = −3.9,

H = −3.2,

Q = −3.5,

and

Y = −1.3.

Tyrosine is less strongly hydrophilic than the charged and amide residues, but it remains on the negative side of this hydropathy scale.

The NAN family is therefore strongly separated from NUN.

This separation is reflected by the complete-code coefficient

$\beta_{2,A} = -2.638542$.

The gauge-invariant contrast between the two families is

$\beta_{2,U} - \beta_{2,A} = 6.587500$.

The second-position U–A opposition is visible in three mutually reinforcing ways:

in the conventional genetic-code table;

in the direct grouping of translated amino acids;

and in the exact pseudoinverse-derived coefficient contrast.

The contrast is broad rather than absolute. It does not describe every chemical feature of the encoded residues. It nevertheless forms the strongest hydropathy division among the four second-position nucleotide states.

6.5 NCN and Lower Residue Volume

The NCN family contains codons whose second nucleotide is C.

Its translated outputs are principally

alanine,

threonine,

proline,

and serine.

These residues are not united by an extreme hydropathy value. Their Kyte–Doolittle values range from moderately positive alanine to negative serine and threonine:

A = 1.8,

T = −0.7,

P = −1.6,

and

S = −0.8.

The family is therefore less sharply separated on the hydropathy axis than NUN or NAN.

Its organization becomes clearer when residue volume is added.

Alanine and serine are small residues. Threonine and proline are larger but remain within a comparatively compact small-to-intermediate volume region. The NCN points consequently occupy a narrower volume sector than several of the other second-position families.

The complete hydropathy coefficient for second-position C is

$\beta_{2,C} = -0.163542$.

This value lies near the center of the fitted hydropathy scale. The NCN family is not strongly hydrophobic or strongly hydrophilic as a whole.

Its more distinctive feature is its concentration toward lower and intermediate residue volumes.

The second-position classification therefore carries information beyond a single hydropathy polarity. Middle-position C identifies a region enriched in relatively compact amino-acid residues.

This does not mean that residue size can be inferred exactly from C at the second position. The four encoded amino acids remain distinct, and proline has unusual conformational properties that cannot be reduced to volume. The claim is regional: NCN occupies a comparatively restricted portion of the two-property plane.

6.6 The Broader NGN Distribution

The NGN family contains codons whose second nucleotide is G.

Its translated outputs include

glycine,

arginine,

serine,

tryptophan,

and cysteine.

The UGA stop codon also belongs to this family.

This group spans a wide range of chemistry.

Glycine is the smallest amino acid and lacks a conventional side chain beyond hydrogen.

Arginine is large, positively charged, and strongly hydrophilic.

Serine is small and polar.

Tryptophan is large, aromatic, and moderately hydrophobic.

Cysteine is sulfur-containing and occupies an intermediate physicochemical position.

Their Kyte–Doolittle values range from

R = −4.5

to

C = 2.5.

Their residue volumes also extend from very small glycine to large tryptophan and arginine.

NGN therefore occupies a broad region of the hydropathy–volume plane.

The complete second-position G coefficient is

β2,G = −1.469792.

This negative coefficient represents the global additive hydropathy tendency of the sixteen NGN codons. It does not imply that every NGN amino acid is hydrophilic. Cysteine and tryptophan lie on the positive side of the hydropathy scale, while arginine is strongly negative.

The family is less reducible to one dominant physicochemical direction than NUN or NAN. Its internal spread demonstrates why a single nucleotide-position coefficient is necessarily a summary rather than a complete chemical description.

6.7 Hydropathy–Volume Segregation

The four middle-nucleotide families form recognizable regions when amino-acid hydropathy is plotted against residue volume.

NUN occupies the hydrophobic side of the plane.

NAN occupies the hydrophilic side and includes several medium-to-large polar or charged residues.

NCN is concentrated toward smaller and intermediate residue volumes.

NGN is broadly dispersed across both axes.

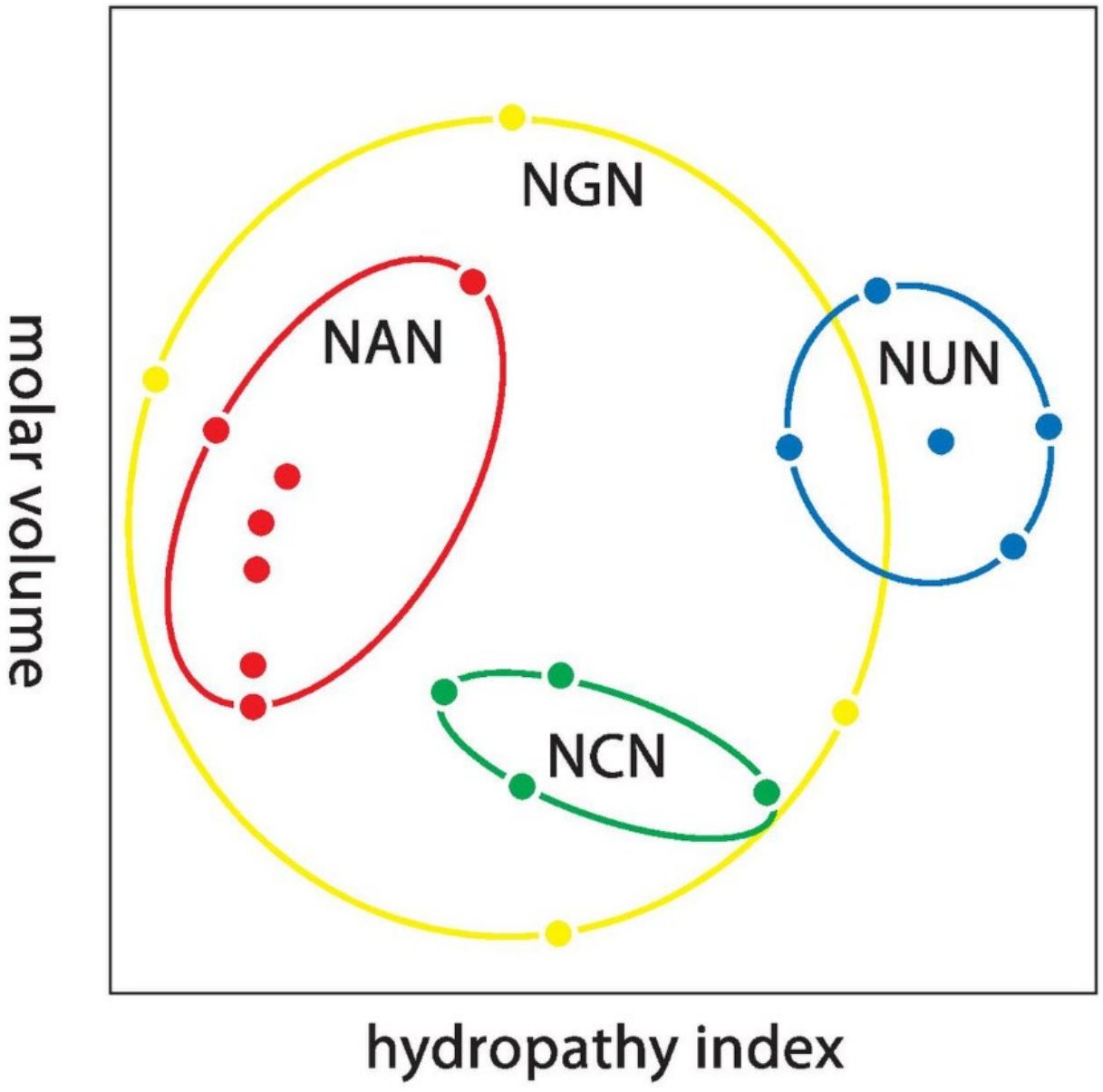


Figure 6. Historical hydropathy–residue-volume representation of the amino acids grouped by the nucleotide occupying the second codon position. Hydropathy is represented by the Kyte–Doolittle scale. The volume coordinate retains the residue-volume values used in the historical source analysis, based on the Zamyatnin solution-volume compilation. Translation stop X is omitted because it is not an amino-acid residue and has no residue volume. The figure shows broad physicochemical segregation rather than a rigid or perfectly separated classification.

The plot should not be interpreted as four nonoverlapping clusters generated by a formal classification algorithm. Some regions overlap. Serine is represented through codons in two second-position families. Individual residues possess chemical features not captured by either plotted coordinate.

The important result is the broad organization.

One nucleotide position divides codon space into four equal sets of sixteen codons. Those equal input sets do not map to randomly intermingled regions of the two-property plane. They map to translated families with visibly different physicochemical tendencies.

The second nucleotide is therefore informative not merely because it helps identify an amino acid. It organizes broad properties of the resulting amino-acid classes.

6.8 Why the Middle Nucleotide Matters

The direct property grouping and the pseudoinverse analysis arrive at the same hierarchy by different routes.

The family analysis begins with the code table. It groups codons by their second nucleotide and inspects the translated residues.

The pseudoinverse analysis begins with a numerical hydropathy field over all sixty-four codons. It determines the closest additive representation of that field and measures the contribution of each nucleotide-position coordinate.

Both identify second-position U and A as the dominant hydropathy opposition.

The hydropathy–volume plot adds another observation: the middle nucleotide also helps organize residue size. The NCN family is comparatively compact in volume, while NGN spans a much broader range.

The middle position therefore behaves as a coarse physicochemical selector.

A schematic hierarchy can be stated as follows:

the second nucleotide establishes a broad property region;

the first nucleotide contributes substantial refinement within that region;

and the third nucleotide often distinguishes synonymous alternatives or makes smaller property adjustments.

This hierarchy is statistical and structural, not absolute. It does not imply that every amino-acid property can be predicted from the middle nucleotide alone. Nor does it eliminate the importance of the first and third positions.

Exact translated identity still requires the complete codon.

The value of the second-position representation lies in compression. A single nucleotide state carries enough information to separate major regions of amino-acid physicochemical space.

That compression becomes operational in membrane-protein analysis. If second-position U is strongly associated with hydrophobic residues and second-position A with hydrophilic residues, a coding sequence may contain a directly readable membrane-hydropathy signal before explicit amino-acid translation.

Chapter 7 examines the historical T–A method and its RNA U–A equivalent.

Chapter 7

Membrane Hydropathy from the Second Codon Position

The coefficient analysis of Chapter 4 showed that second-position U carries the strongest positive hydropathy contribution, while second-position A carries the strongest negative contribution. Chapter 6 showed the same polarity geometrically: NUN codons predominantly encode hydrophobic amino acids, whereas NAN codons predominantly encode hydrophilic or charged amino acids.

A stronger question is whether this code-level organization remains visible along the coding sequence of an actual membrane protein.

Conventional hydropathy profiling begins with the translated amino-acid sequence. Each residue is assigned a hydropathy value, and those values are averaged over a sliding window. Extended positive regions are interpreted as candidate membrane-spanning segments.

The historical nucleotide method asked whether the principal hydrophobic regions could be detected before explicitly converting each codon into an amino-acid identity.

The result was unexpectedly strong. A profile based only on the opposition between T and A at the second codon position reproduced the principal broad hydrophobic peaks of a conventional Kyte–Doolittle profile in the illustrated membrane protein. In RNA notation, the same contrast is U minus A.

The method does not recover every amino-acid-specific feature. Its importance lies in the amount of membrane-scale information retained after a drastic compression of the property description.

7.1 Origins in Membrane-Protein Engineering

The nucleotide hydropathy method arose from work on genetic coding algorithms for engineering membrane proteins.

Membrane proteins contain extended segments enriched in amino acids favorable to a nonaqueous environment. These segments often form transmembrane α-helices. Their identification is therefore a central problem in sequence analysis and protein engineering.

A conventional prediction pipeline proceeds through several stages:

the coding sequence is translated;

each amino acid is assigned a hydropathy value;

the values are smoothed over a window;

and sustained hydrophobic peaks are identified.

The genetic-code analysis suggested a shorter route.

Because many hydrophobic amino acids are encoded by codons containing U or T at the second position, and many hydrophilic amino acids are encoded by codons containing A there, the coding sequence itself should contain a coarse membrane-hydropathy signal.

The relevant information is not overall nucleotide composition. It is nucleotide identity at one specific position within the reading frame.

A U or T at the first or third codon position does not carry the same global implication. The effect depends on the middle position because the standard code organizes broad amino-acid physicochemical character most strongly along that coordinate.

The historical method therefore examined the second nucleotide of each codon and ignored most of the remaining sequence information.

7.2 Conventional Kyte–Doolittle Profiling

Let a translated protein sequence contain residues

$a_1, a_2, \ldots, aN.$

Assign each residue its Kyte–Doolittle hydropathy value:

$H (a_i).$

For a sliding window of length w, a local hydropathy profile can be written as

K_i

$= (1 / w) \Sigma j \in Wi \; H (a_j),$

where Wi is the set of residue positions included in the window centered at or associated with sequence position i.

A positive value indicates that the window is enriched in hydrophobic residues. A sustained positive region may indicate a membrane-spanning segment.

The precise numerical trace depends on several analytical choices:

the window length;

whether the window is centered or trailing;

how terminal positions are handled;

whether the values are summed or averaged;

and whether any additional smoothing or scaling is applied.

The conceptual procedure is nevertheless straightforward. Twenty amino-acid identities are distinguished, and each is assigned a residue-specific hydropathy parameter.

The method therefore uses the full translated sequence and a table of twenty property values.

Its strength is chemical specificity. Leucine, isoleucine, valine, phenylalanine, methionine, cysteine, and the remaining amino acids retain their individual hydropathy values.

Its cost is representational complexity. Explicit translation and amino-acid identification are required before the profile is generated.

7.3 The Historical Youvan T–A Method

The historical nucleotide-only method used DNA notation and focused on the second position of each codon.

Let the coding sequence be divided into codons

$c_1, c_2, \ldots, cN.$

For codon c_i, define its second-position nucleotide as

$c_{i,2}.$

A simple DNA-based score is

$s_i = +1$ if $c_{i,2} = T,$

$s_i = -1$ if $c_{i,2} = A,$

and

$s_i = 0$ if $c_{i,2} = G$ or C.

A sliding-window profile can then be written as

Y_i

$= (1 / w) \Sigma j \in Wi\ s_j.$

Equivalently, apart from the selected normalization,

$Y_i \propto NT (Wi) - NA (Wi),$

where NT (Wi) is the number of codons having T at the second position within the window, and NA (Wi) is the corresponding number having A.

The profile therefore depends only on two opposed nucleotide indicators:

second-position T;

and second-position A.

Second-position G and C contribute zero in this compressed representation.

The exact historical window length, boundary convention, vertical scaling, and normalization should be recovered from the original publication, notebook, or numerical trace before a definitive reproduction is claimed. The simple formulation above captures the central contrast but should not be assumed to reproduce every historical plotting detail.

7.4 The RNA U–A Equivalent

The historical figure used DNA notation because the analyzed sequence was represented with thymine.

For an RNA codon sequence, T is replaced by U. The score becomes

$s_i = +1$ if $c_{i,2} = U$,

$s_i = -1$ if $c_{i,2} = A$,

and

$s_i = 0$ if $c_{i,2} = G$ or C.

The RNA profile is therefore

Y_i

$= (1 / w) \Sigma j \in Wi$

$[1 \{ c_{j,2} = U \} - 1 \{ c_{j,2} = A \}]$,

where the indicator 1 { condition } equals 1 when the condition is true and 0 otherwise.

Equivalently,

$Y_i \propto NU (Wi) - NA (Wi)$.

The DNA T–A and RNA U–A descriptions refer to the same position-specific genetic-code contrast. The different letters reflect the molecular notation used for the coding sequence.

Throughout the general discussion of the standard RNA code, U–A is the preferred notation. T–A is retained when describing the historical DNA-based figure or method.

7.5 Two Nucleotide Parameters Versus Twenty Amino-Acid Parameters

The conventional profile distinguishes twenty amino-acid identities and assigns a separate hydropathy value to each.

The nucleotide-only profile distinguishes only two opposed second-position states:

U or T contributes positively;

A contributes negatively.

G and C are neutral in the simplest implementation.

The representational reduction is therefore extreme.

The conventional method uses information equivalent to twenty residue-specific property parameters.

The compressed method uses two opposed nucleotide indicators, or one signed U/T–A contrast.

It also avoids explicit amino-acid assignment during profile construction. The coding sequence must be correctly divided into codons, but each codon need not be translated into a residue letter before its contribution is scored.

The method discards substantial information. It ignores:

the identity of the first codon nucleotide;

the identity of the third codon nucleotide;

differences among amino acids sharing the same second-position family;

the distinction between G and C in the simplest contrast;

and chemical features not aligned with the U–A hydropathy direction.

The compressed profile cannot distinguish leucine from isoleucine or valine merely because all are commonly encoded within the NUN family. Nor can it represent the unusual chemistry of cysteine, tryptophan, proline, glycine, or charged residues in full detail.

Its significance lies in what remains after these distinctions are removed.

7.6 Why the Contrast Works

The U/T–A profile works because the genetic code itself aligns second-position nucleotide identity with broad amino-acid environmental preference.

The NUN codon family encodes

phenylalanine,

leucine,

isoleucine,

methionine,

and valine.

All five are hydrophobic on the Kyte–Doolittle scale.

The NAN family encodes

aspartate,

glutamate,

asparagine,

lysine,

histidine,

glutamine,

and tyrosine,

together with two stop codons.

Most of these amino acids are hydrophilic, polar, or charged.

A hydrophobic protein segment will therefore tend to contain an elevated frequency of codons having U or T at the second position.

A hydrophilic segment will tend to contain more codons having A at the second position.

The relationship can be summarized schematically as

second-position U or T → hydrophobic enrichment,

and

second-position A → hydrophilic enrichment.

These arrows express an association created by the standard genetic-code assignment. They do not imply that U, T, or A chemically possesses the side-chain hydropathy of the encoded amino acids.

Sliding-window averaging then converts local enrichment into broad sequence peaks. It strengthens sustained patterns and suppresses isolated codons.

The windowing procedure alone does not create the correspondence. The correspondence exists because the lookup table places many hydrophobic and hydrophilic residues in opposed middle-nucleotide families.

7.7 Relationship to the Exact Pseudoinverse Model

The complete additive hydropathy reconstruction is

$\hat{h}\ c_1c_2c_3$

$= \beta 1,c_1 + \beta 2,c_2 + \beta 3,c_3.$

This model uses twelve displayed nucleotide-position coefficients with ten independent dimensions.

The historical U/T–A profile retains only its dominant contrast:

$\beta_{2,U} - \beta_{2,A}$.

For the complete sixty-four-codon analysis,

$\beta_{2,U} = 3.948958$,

and

$\beta_{2,A} = -2.638542$.

Therefore,

$\beta_{2,U} - \beta_{2,A} = 6.587500$.

This contrast is larger than any other within-position hydropathy contrast in the complete coefficient vector.

A profile using only U and A at the second position can therefore be understood as a one-dimensional compression of the full additive model.

Schematically,

full model

= first-position effects

- second-position effects
- third-position effects,

whereas the historical profile retains primarily

second-position U–A polarity.

An offset and scale factor may be introduced without changing the location of the broad peaks. The simplest signed count therefore need not have the same numerical vertical units as the Kyte–Doolittle profile to recover similar sequence regions.

The connection between the historical method and the exact pseudoinverse is consequently structural rather than merely visual. The exact coefficient calculation independently identifies the same middle-position contrast that the earlier membrane-profile method used operationally.

7.8 Recovery of the Principal Hydrophobic Peaks

The historical comparison presents two traces over the same sequence region.

The upper trace is a conventional Kyte–Doolittle profile generated from translated amino-acid identities.

The lower trace is the Youvan nucleotide-only profile generated from the second-position T–A contrast.

The principal broad peaks occur in corresponding sequence regions. Regions that appear strongly hydrophobic in the amino-acid profile also tend to be enriched in second-position T in the coding sequence and depleted in second-position A.

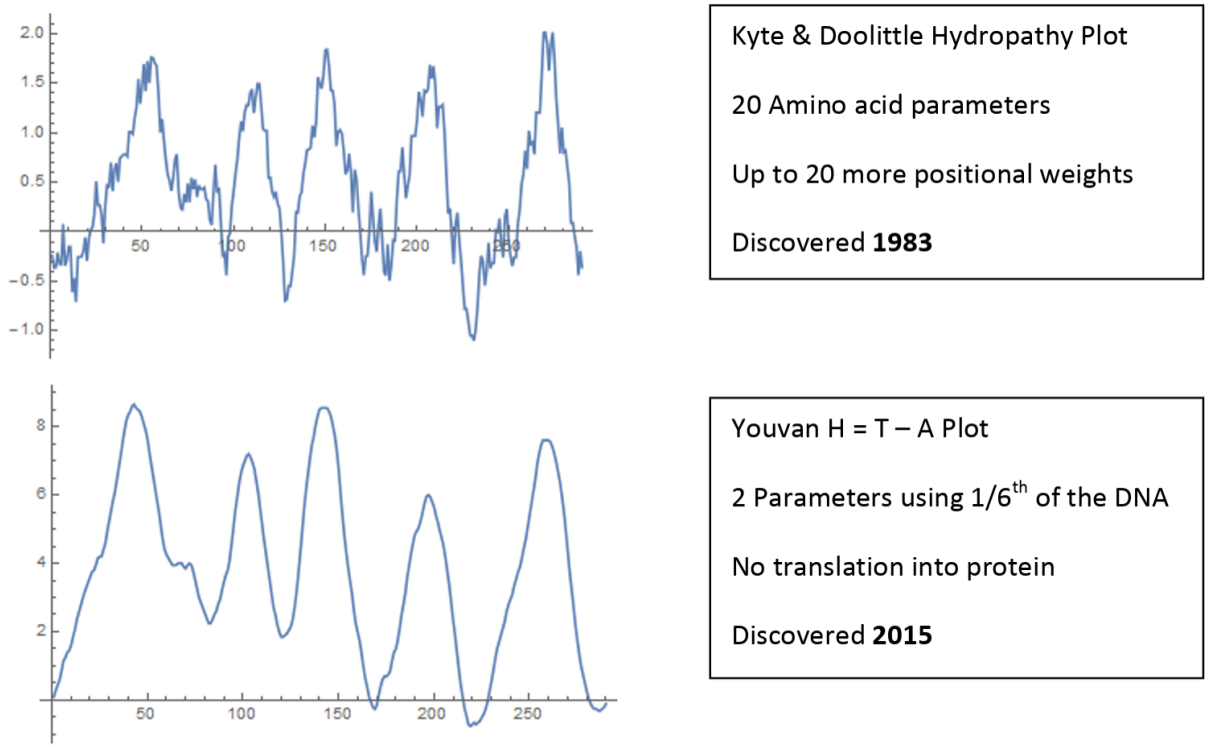


Figure 7. Comparison of conventional and nucleotide-only membrane-hydropathy profiles. The upper trace is a Kyte–Doolittle profile calculated from translated amino-acid identities. The lower trace is the historical Youvan profile calculated from the coding sequence using the T–A contrast at the second codon position. In RNA notation, the corresponding contrast is U–A. The nucleotide-only profile reproduces the principal broad hydrophobic peaks of the conventional profile in the illustrated membrane protein. The photosynthetic reaction center L-subunit sequence was used – known from X-ray data to have five transmembrane helices.

The strongest defensible conclusion from the historical figure is:

The second-position T/U–A profile reproduces the principal hydrophobic peaks of the conventional Kyte–Doolittle profile in the illustrated membrane protein.

The figure should not be described as proving that the methods are identical or universally “as good” without quantitative benchmarking.

7.9 Compression Without Explicit Translation

The nucleotide profile demonstrates that a translated physicochemical pattern can remain visible in the coding sequence at a highly compressed representational level.

A conventional calculation asks:

Which amino acid occurs at each position, and what is its hydropathy?

The compressed calculation asks:

Does the corresponding codon contain U/T or A at its middle position?

The first question preserves residue-specific information.

The second preserves only a dominant code-level contrast.

Yet membrane-spanning regions are extended structures. Their detection depends more on sustained hydrophobic enrichment than on the precise value assigned to every individual residue. A coarse signal can therefore retain the major peaks even while losing local chemical detail.

This illustrates a general principle of informational compression.

A lower-dimensional representation may fail to reconstruct individual symbols exactly while preserving a larger-scale feature of the sequence.

The U/T–A profile is not a substitute for the full amino-acid sequence. It is a compact projection that retains an important macroscopic property.

7.10 Reading-Frame Dependence

The method depends on the correct reading frame.

A nucleotide is classified as first, second, or third only after the coding sequence has been divided into codons. Shifting the frame by one nucleotide changes which bases occupy the middle positions.

This creates an important control.

The correct reading frame should preserve the code-derived U/T–A hydropathy signal more strongly than either of the two incorrect frames.

A profile generated after shifting the frame by one or two nucleotides would still reflect general base composition, but it would no longer sample the middle coordinate defined by the translated code.

Comparing the three reading frames can therefore help distinguish a genuine codon-position signal from a signal produced merely by local enrichment of T, U, or A anywhere in the sequence.

The historical figure demonstrates the method in the coding frame. A systematic future evaluation should include both alternative frames as negative or degraded controls.

7.11 Synonymous Codon Choice

The U/T–A profile is largely determined by amino-acid identity because the second codon position is often fixed within an amino-acid family.

For example, the principal hydrophobic NUN amino acids retain U or T at the middle position across their synonymous codons.

Synonymous variation nevertheless matters in some regions of the code. Serine codons occur in both NCN and AGY families, placing serine in more than one second-position group. Leucine and arginine also occupy split codon families, although their middle-position patterns differ from the serine case.

A synonymous-codon reshuffling that preserves the amino-acid sequence may therefore alter parts of the nucleotide-only profile while leaving the Kyte–Doolittle profile unchanged.

This suggests another useful control:

retain the protein sequence;

randomize synonymous codon choice;

and recompute the U/T–A trace.

The resulting distribution would measure how much of the nucleotide profile is fixed by translated identity and how much depends on the particular synonymous coding sequence.

The historical demonstration establishes that the signal is visible in one actual sequence. It does not determine the full range of synonymous variability compatible with that protein.

7.12 Limits of the Present Comparison

The historical figure establishes a concrete sequence-level manifestation of genetic-code structure.

It shows that:

middle-position nucleotide composition varies along a membrane-protein coding sequence;

that variation forms broad peaks after windowing;

the peaks align visually with major peaks in a conventional translated hydropathy profile;

and the signal can be generated without assigning one of twenty amino-acid hydropathy values at every residue.

The figure does not establish that:

the two profiles are mathematically identical;

every local peak is reproduced;

the vertical scales are directly comparable;

the nucleotide profile works equally well for every membrane protein;

the historical window parameters are optimal;

the method equals modern transmembrane-helix predictors;

or the same performance would be obtained for soluble proteins.

The original historical numerical traces would be required to calculate pointwise correlation, peak displacement, sensitivity, specificity, and other quantitative measures.

The present language should therefore remain precise. The profile reproduces the principal broad peaks in the illustrated example. It has not yet been established as a universally equivalent predictor.

7.13 Future Quantitative Tests

A modern evaluation should begin with a clearly defined protein collection containing experimentally established transmembrane segments and an appropriate soluble-protein control set.

For every coding sequence, the following profiles could be compared:

the conventional Kyte–Doolittle profile;

the simple second-position U–A profile;

the full twelve-coefficient pseudoinverse profile;

profiles generated in the two incorrect reading frames;

and profiles generated after controlled sequence randomizations.

Relevant measurements would include:

correlation between traces;

overlap of major peak regions;

distance between predicted and established membrane boundaries;

sensitivity for known transmembrane segments;

specificity against nonmembrane regions;

false-positive rates in soluble proteins;

and robustness to synonymous codon substitution.

Several controls would separate different sources of agreement.

Shuffling codons while preserving codon composition would destroy sequence order but retain global nucleotide counts.

Synonymous shuffling would preserve the protein sequence and the Kyte–Doolittle trace while altering portions of the nucleotide trace.

Second-position shuffling would directly disrupt the proposed code-level signal.

Reading-frame shifts would test positional specificity.

Randomizing the genetic-code assignments would test whether the observed performance depends on the structure of the standard code rather than on the sliding-window procedure itself.

These experiments could establish how much predictive information is contained in the U–A coordinate and where the compressed method fails.

7.14 What the Demonstration Means

The membrane-profile comparison connects the abstract matrix analysis to an actual biological sequence.

The exact pseudoinverse shows that the strongest additive hydropathy contrast lies between U and A at the second codon position.

The family analysis shows why: NUN and NAN encode broadly opposed physicochemical classes.

The historical membrane trace shows that the same contrast varies meaningfully along a real coding sequence and recovers major translated hydrophobic regions.

The three results occur at different levels:

the coefficient is a global mathematical summary of the complete genetic code;

the family plot is a physicochemical classification of translated outputs;

and the membrane profile is a sequence-level manifestation.

Their agreement shows that the middle-nucleotide hydropathy structure is not confined to an abstract coefficient table. It remains visible when the code is used to encode an extended membrane protein.

The result is a striking example of compression through a natural lookup relation. Twenty amino-acid hydropathy parameters are reduced to one principal nucleotide contrast, yet the largest membrane-scale features remain recognizable.

Part III now turns from physicochemical projection to the geometry of synonymous codons. The same sixty-four inputs will be arranged as a $4 \times 4 \times 4$ lattice under the nucleotide path G–A–C–U.

Part III

Geometry and Topology of Codon Space

Chapter 8

The G–A–C–U Codon Cube

The sixty-four codons form a complete three-coordinate product space. Each coordinate can occupy one of four nucleotide states, but the product structure alone does not specify which states should be considered neighbors.

To construct a local codon geometry, the nucleotide states must be connected by a declared adjacency relation.

This book begins with the simplest connected linear geometry: a path through all four nucleotide symbols. Applying that path independently to the first, second, and third codon positions produces a 4 × 4 × 4 lattice containing all sixty-four codons.

The principal path is

G–A–C–U.

The resulting object is called the G–A–C–U codon cube.

The word **cube** refers to the three-dimensional arrangement of sixty-four codon vertices. It should not be confused with the ordinary eight-vertex geometric cube or with the complete Hamming-distance-one codon graph. The G–A–C–U codon cube is the Cartesian product of three four-vertex paths.

Its significance does not come from assuming that G–A–C–U is privileged. Every possible linear order of the four nucleotide states will later be evaluated under the same translation-preservation objective.

8.1 From Product Space to Ordered Geometry

The complete RNA codon set is

$C = \Sigma^3$,

where

$\Sigma = \{ G, A, C, U \}$.

A codon

$c = c_1 c_2 c_3$

is represented by three ordered coordinates:

the nucleotide at position 1;

the nucleotide at position 2;

and the nucleotide at position 3.

The product

$\Sigma \times \Sigma \times \Sigma$

specifies which codons exist. It does not specify which codons should be connected.

For example, consider the codons

GAC

and

GUC.

They differ only at the second position:

GAC → GUC.

Whether these codons are neighbors depends on the nucleotide geometry placed on the second coordinate.

If A and U are declared adjacent, the codons are connected.

If A and U are not adjacent, they are not connected in the selected lattice even though their Hamming distance is one.

The product set therefore supplies coordinates but not local topology.

A topology is introduced by choosing an adjacency graph on the four nucleotide states and applying that graph to each codon position.

The present analysis restricts the initial comparison class to linear paths containing all four nucleotide symbols exactly once.

8.2 Linear Nucleotide Paths

A linear nucleotide path is an ordering

$n_1–n_2–n_3–n_4$,

where

$\{ n_1, n_2, n_3, n_4 \} = \{ G, A, C, U \}$.

Such a path selects three unordered nucleotide adjacencies:

$\{ n_1, n_2 \}$,

$\{ n_2, n_3 \}$,

and

$\{ n_3, n_4 \}$.

It excludes the other three possible unordered nucleotide pairs.

For example, the path

G–A–C–U

contains the adjacencies

G–A,

A–C,

and

C–U.

It does not contain

G–C,

G–U,

or

A–U.

A nearest-neighbor movement along this path can move between G and A, between A and C, or between C and U. It cannot move directly between two nonadjacent states.

The reverse path

U–C–A–G

contains the same three undirected adjacencies. It therefore defines the same undirected nucleotide graph.

Orientation matters when paths are listed as permutations, but it does not affect the undirected synonymous-edge count used in this book.

8.3 The G–A–C–U Path

The principal nucleotide path is

G–A–C–U.

Its nucleotide-state graph is

G — A — C — U.

G and U are endpoints. Each has one path neighbor.

A and C are interior states. Each has two path neighbors.

The path makes the following local substitutions admissible:

G ↔ A,

A ↔ C,

and

C ↔ U.

The following substitutions are excluded from the selected local edge set:

G ↔ C,

G ↔ U,

and

A ↔ U.

These exclusions do not imply that the corresponding nucleotide substitutions are biologically impossible. They mean only that those substitutions are not edges in this particular sparse codon topology.

The G–A–C–U path is used as a mathematical ordering, not introduced as a biochemical law. Its standing must be determined by comparison with the other possible nucleotide paths.

8.4 Construction of the 4 × 4 × 4 Codon Cube

Apply the nucleotide path independently to each of the three codon positions.

The first axis varies c_1 while holding c_2 and c_3 fixed.

The second axis varies c_2 while holding c_1 and c_3 fixed.

The third axis varies c_3 while holding c_1 and c_2 fixed.

Each axis contains the four ordered states

G, A, C, U.

The complete lattice is therefore

$P_4 \square P_4 \square P_4$,

the Cartesian product of three four-vertex paths.

Its number of vertices is

$4 \times 4 \times 4 = 64$.

Every codon occupies exactly one lattice point.

For example, GAC occupies the point with coordinates

(G, A, C).

Moving one step along the first axis can change its first nucleotide from G to A because G and A are adjacent in the path:

GAC ↔ AAC.

Moving one step along the second axis can change A to G or C because A is an interior nucleotide state:

GAC ↔ GGC,

and

GAC ↔ GCC.

Moving one step along the third axis can change C to A or U:

GAC ↔ GAA,

and

GAC ↔ GAU.

A lattice edge always changes exactly one codon position. The other two positions remain fixed.

The possible degree of a vertex depends on whether each of its three nucleotide coordinates is an endpoint or an interior state of the path.

A coordinate equal to G or U contributes one available step along its axis.

A coordinate equal to A or C contributes two.

Consequently, codon vertices have degrees ranging from three to six.

The eight codons whose three coordinates are all path endpoints occupy the corners of the lattice and have degree three. Codons with one or more interior coordinates have correspondingly larger degree.

8.5 Forty-Eight Edges per Codon Position

A four-state path contains three edges.

Consider the first codon position. For every selected nucleotide adjacency at that position, the second and third positions may independently assume any of four states.

The number of settings of the two unchanged coordinates is therefore

$4^2 = 16$.

Each nucleotide-pair adjacency is repeated sixteen times along the first-position axis.

Because the path contains three nucleotide adjacencies, the number of first-position edges is

$3 \times 16 = 48$.

The same calculation applies to the second position:

$3 \times 16 = 48$.

It also applies to the third position:

$3 \times 16 = 48$.

The edge set therefore separates exactly into three disjoint positional classes:

48 first-position edges;

48 second-position edges;

and

48 third-position edges.

This separation is important because every edge can be attributed unambiguously to one codon position. The later anisotropic score is consequently a sum of three independent positional scores.

8.6 The Complete 144-Edge Cube

The total number of nearest-neighbor edges is

$48 + 48 + 48 = 144.$

Equivalently,

$| E | = 3 \times 3 \times 4^2 = 144.$

The first factor of three is the number of codon positions.

The second factor of three is the number of edges in a four-state nucleotide path.

The factor 4^2 counts the possible settings of the two unchanged codon positions.

Thus, the G–A–C–U codon cube contains

64 vertices

and

144 undirected edges.

This edge total can also be checked from the mean vertex degree. By the handshaking identity,

$\Sigma c \in C \deg (c) = 2 | E | = 288.$

The mean degree is therefore

$288 / 64 = 4.5.$

The ordered cube is substantially sparser than the complete Hamming-distance-one codon graph.

In the complete Hamming graph, each codon can change at any of three positions to any of the other three nucleotides. Every codon has

$3 \times 3 = 9$

one-substitution neighbors.

The complete Hamming graph therefore contains

$64 \times 9 / 2 = 288$

undirected edges.

The ordered path cube retains exactly half of those edges:

$144 / 288 = 1 / 2.$

The G–A–C–U cube is therefore a sparse path subgraph of the complete one-substitution codon graph.

Its purpose is not to represent every possible mutation. Its purpose is to compare how different linear nucleotide paths preserve translated identity under the same local-edge budget.

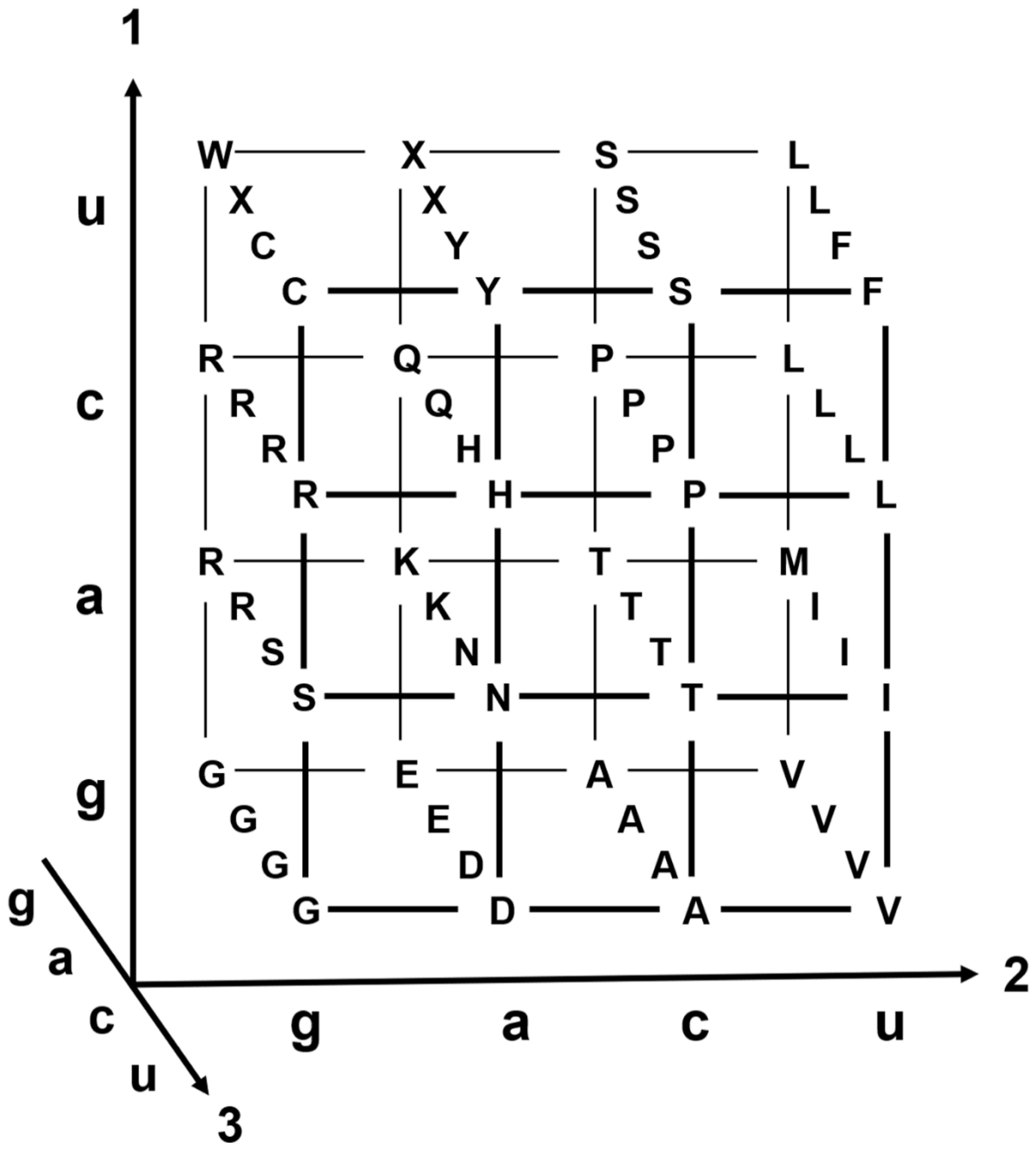


Figure 8. The isotropic G–A–C–U codon cube. The sixty-four RNA codons are arranged as a 4 × 4 × 4 lattice whose three axes correspond to the first, second, and third codon positions. Each axis uses the nucleotide order G–A–C–U. Every lattice point is labeled by the single-letter identity of its translated amino acid or by X for translation stop. The lattice contains 144 nearest-neighbor edges: forty-eight associated with changes at each codon position. Repeated translated letters at adjacent vertices identify synonymous codon pairs that become self-loops after collapse by translated identity.

8.7 Placing Translation on the Cube

The cube is initially a geometry of codon identity.

Each of its sixty-four vertices represents a distinct codon. The standard genetic-code map then assigns one translated label to every vertex:

$\tau : C \rightarrow \Omega$.

The cube consequently carries two simultaneous labelings.

At the input level, it contains sixty-four distinct codon identities.

At the output level, those codons carry twenty-one repeated translated identities.

An edge connecting codons c and d can therefore have one of two translation outcomes.

It may connect different translated classes:

$\tau (c) \neq \tau (d)$.

Such an edge is nonsynonymous.

Alternatively, it may connect codons sharing the same translated output:

$\tau (c) = \tau (d)$.

Such an edge is synonymous.

A synonymous cube edge is a local movement in codon space that preserves translated identity under the standard code.

The cube contains forty-four such edges under the G–A–C–U ordering. The remaining one hundred edges connect different translated classes:

$144 - 44 = 100$.

The exact derivation and translated multigraph interpretation of these counts are developed in Chapter 9.

8.8 Position-Specific Structure

The synonymous edges are distributed very unevenly among the three codon positions.

The G–A–C–U cube contains:

4 synonymous first-position edges;

1 synonymous second-position edge;

and

39 synonymous third-position edges.

The total is

4 + 1 + 39 = 44.

The dominance of the third position is consistent with the familiar prevalence of synonymous third-position variation. It is also complementary to the hydropathy results of Part II.

In the hydropathy projection, the third-position coefficients are comparatively small because many changes at that position preserve translated identity and therefore preserve the amino-acid property assigned to the codon.

In the topology, the third axis supplies most of the selected nearest-neighbor synonymous edges.

The same feature of the genetic code appears in two mathematical forms:

small translated-property contrasts at the third position;

and extensive translated-identity preservation along the third axis.

The second position shows the opposite pattern.

It produces the strongest hydropathy contrasts but only one synonymous G–A–C–U edge. A change at the middle position is therefore much more likely to cross a major translated physicochemical boundary.

8.9 The Four Middle-Nucleotide Layers

The codon cube also provides a geometric view of the four second-position families examined in Chapter 6.

Fixing the second coordinate produces four parallel 4 × 4 layers:

NGN;

NAN;

NCN;

and

NUN.

Each layer contains sixteen codons.

Along the second axis, these layers occur in the order

NGN → NAN → NCN → NUN.

The hydropathy–volume families are therefore not external categories imposed after the cube is constructed. They are literal parallel slices of the same codon product space.

The transition from NAN to NCN and from NCN to NUN moves between adjacent middle-nucleotide layers under the selected path. The transition from NGN to NAN forms the remaining adjacent-layer relation.

Only one of the forty-eight second-position cube edges preserves translated identity. Most movements between these layers alter the translated class.

The cube thereby connects the physicochemical and topological descriptions. The same second coordinate that organizes broad hydropathy families also provides the least synonymous axis of the G–A–C–U lattice.

8.10 Isotropic Codon Cubes

An isotropic codon cube uses the same nucleotide path on all three axes.

If the selected path is

$\pi = n_1–n_2–n_3–n_4$,

then π is used for the first, second, and third codon positions.

The G–A–C–U cube is isotropic because all three axes use

G–A–C–U.

There are

$4! = 24$

oriented linear orders of the four nucleotide symbols.

Each oriented order defines an isotropic codon cube. Because a path and its reverse produce the same undirected adjacencies, the twenty-four oriented orders correspond to twelve distinct undirected path topologies.

Every isotropic cube contains the same number of vertices and edges:

64 vertices;

48 edges along each axis;

and

144 edges in total.

What changes is which nucleotide pairs are treated as adjacent. That change affects how many edges connect synonymous codons.

For an isotropic path π, define

$L(\pi)$

as the number of its 144 cube edges whose endpoints have the same translated output.

The isotropic optimization problem is therefore:

Among all twenty-four oriented nucleotide orders, which path maximizes $L(\pi)$?

The answer is G–A–C–U and its reverse, with

$L = 44$.

That result will be proved by complete enumeration and by exact nucleotide-pair weights in Chapter 10.

8.11 Anisotropic Codon Cubes

An anisotropic codon cube permits a different nucleotide path on each codon axis.

Let

π_1

be the path used at the first position,

π_2

the path used at the second position,

and

π_3

the path used at the third position.

Each path may be any of the twenty-four oriented nucleotide orders.

The number of oriented anisotropic combinations is therefore

$24^3 = 13{,}824$.

Every combination still produces a 4 × 4 × 4 lattice with sixty-four vertices and 144 edges. The difference is that nucleotide adjacency may vary among the three codon positions.

The complete synonymous score is

$L(\pi_1, \pi_2, \pi_3)$

$= L_1(\pi_1) + L_2(\pi_2) + L_3(\pi_3)$,

where L_1, L_2, and L_3 count the synonymous edges contributed by the first, second, and third axes.

No cross-term appears because every edge changes exactly one codon position.

This exact separability permits the global anisotropic maximum to be found by maximizing the three positional scores independently.

The complete search shows that no anisotropic combination exceeds forty-four synonymous edges.

8.12 Why G–A–C–U Is Not Assumed

The G–A–C–U order is visually and mathematically useful, but neither usefulness nor visual symmetry establishes optimality.

The order is tested against the entire declared comparison class.

For the isotropic problem, all twenty-four oriented nucleotide paths are evaluated.

For the anisotropic problem, all 13,824 oriented triples of nucleotide paths are evaluated.

Every arrangement is scored by the same objective:

Count the nearest-neighbor cube edges whose endpoint codons have identical translated outputs.

G–A–C–U and its reverse attain the unique undirected isotropic maximum.

The anisotropic search permits each axis to choose its own maximizing path. It still cannot exceed the total of forty-four.

A further result will show that G–A–C–U is the unique undirected nucleotide path occurring in the maximizing families of all three codon positions.

The path is therefore not privileged by declaration. It emerges from exhaustive finite comparison under a precisely defined objective.

8.13 Scope of the Geometry

The linear-path geometry is one possible definition of nucleotide locality.

Other geometries could be studied, including:

the complete four-state graph;

a four-cycle;

transition–transversion adjacency;

chemically weighted nucleotide graphs;

or graphs derived from empirical substitution rates.

Different geometries ask different questions and may produce different optima.

The theorem established in this book must therefore retain its scope:

Within the family of linear four-nucleotide paths, scored by the number of synonymous nearest-neighbor codon edges under the standard genetic code, the maximum is forty-four.

The G–A–C–U cube supplies the principal realization of that maximum.

The next chapter collapses its sixty-four codon vertices into twenty-one translated vertices. The forty-four synonymous cube edges then become forty-four self-loops in a translated multigraph.

Chapter 9

The Forty-Four-Loop Translated Multigraph

The G–A–C–U codon cube contains sixty-four vertices, one for every RNA codon, and 144 nearest-neighbor edges. Each vertex also carries a translated label: one of the twenty amino-acid single-letter codes or X for translation stop.

Many codon vertices therefore share the same translated identity.

The cube preserves all distinctions among codons. A complementary representation suppresses those distinctions and retains only translated identity. All codons producing the same output are collapsed into one translated vertex, while the edges of the original cube are transferred to the resulting graph.

This operation converts the sixty-four-vertex codon cube into a twenty-one-vertex translated multigraph.

Its most distinctive feature is the presence of exactly forty-four self-loops. Every loop represents one selected nearest-neighbor movement in codon space that leaves the translated output unchanged.

The loop count is not synonymous degeneracy expressed in another notation. Degeneracy counts codons. Loops count local translation-preserving connections among codons under the declared G–A–C–U geometry.

9.1 Collapse by Translated Identity

Let

$G C = (C, E)$

denote the G–A–C–U codon cube.

Its vertex set is the complete codon set

$C = \Sigma^3$,

where

$\Sigma = \{ G, A, C, U \}$.

Its edge set E contains the 144 nearest-neighbor pairs defined by the G–A–C–U path on each of the three codon positions.

The standard genetic-code map is

$\tau : C \rightarrow \Omega$,

where Ω contains twenty amino-acid identities and translation stop X.

Define an equivalence relation on the codon vertices by

$c \sim d$

when

$\tau (c) = \tau (d)$.

Two codons are equivalent precisely when they belong to the same translated fiber.

The equivalence classes are therefore

$F\omega = \tau^{-1}(\omega)$,

for

$\omega \in \Omega$.

Collapsing every fiber $F\omega$ to one vertex produces a translated quotient of the codon cube.

At the vertex level, the operation is straightforward:

all alanine codons collapse to A;

all leucine codons collapse to L;

all serine codons collapse to S;

UAA, UAG, and UGA collapse to X;

and similarly for the remaining translated classes.

The quotient has

$|\Omega| = 21$

vertices.

The edge transformation must retain more information than an ordinary simple graph permits. Different codon-level edges can map to the same pair of translated vertices, and synonymous codon-level edges can map from a translated vertex back to itself.

The resulting object is therefore a multigraph with self-loops.

9.2 Twenty-One Translated Vertices

The translated vertex set is

$\Omega = \{ A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, X \}$.

Each translated vertex represents an entire codon fiber rather than one codon.

For example, the glycine vertex represents

$FG = \{ GGG, GGA, GGC, GGU \}$.

The leucine vertex represents

$FL = \{ UUA, UUG, CUA, CUC, CUG, CUU \}$.

The stop vertex represents

FX = { UAA, UAG, UGA }.

The translated graph therefore compresses the input domain from sixty-four vertices to twenty-one.

This compression discards the individual identity of a codon once it has been assigned to its translated fiber. It does not discard the number or type of codon-level edges entering the quotient.

An edge in the translated multigraph remains an edge instance inherited from a particular edge of the codon cube. If five different codon-level adjacencies collapse to the same translated pair, all five edge instances are retained.

The graph is consequently not merely a diagram showing which amino-acid classes can be connected. It records the multiplicities induced by the complete cube.

9.3 Self-Loops and Parallel Edges

Consider a codon-cube edge

$\{ c, d \} \in E$.

Because c and d are neighbors in the cube, they differ at exactly one codon position by one allowed step along the G–A–C–U nucleotide path.

There are two cases.

If

$\tau (c) = \tau (d)$,

the codons are synonymous. After both vertices collapse to the same translated class, the edge begins and ends at one translated vertex. It becomes a self-loop.

If

$\tau (c) \neq \tau (d)$,

the edge becomes an edge between two distinct translated vertices.

Thus,

synonymous cube edge → translated self-loop,

whereas

nonsynonymous cube edge → interclass translated edge.

A translated class may contain several self-loops because several distinct pairs of codons within its fiber may be nearest neighbors in the original cube.

Similarly, two translated classes may be joined by several parallel edges. Each parallel edge represents a different codon-level adjacency that collapses to the same output pair.

For example, several distinct codon substitutions may change one amino-acid identity into another while following different positions or nucleotide contexts. Once codon identity is suppressed, those distinct substitutions appear as parallel connections between the same two translated vertices.

Removing those multiplicities would destroy information. The translated quotient must therefore be treated as a multigraph rather than converted silently into a simple graph.

9.4 Forty-Four Synonymous Edge Instances

The G–A–C–U codon cube contains

144

edge instances.

For an edge

e = { c, d },

define its translation-preservation indicator by

s (e) = 1

when

τ (c) = τ (d),

and

s (e) = 0

otherwise.

The total number of synonymous nearest-neighbor edges is

L = Σe∈E s (e).

For the standard genetic code under the isotropic G–A–C–U topology,

L = 44.

After collapse by translated identity, these forty-four synonymous edge instances become forty-four self-loops.

The position-specific decomposition is

$L_1 = 4$,

$L_2 = 1$,

and

$L_3 = 39$.

Therefore,

$L = L_1 + L_2 + L_3$

$= 4 + 1 + 39$

$= 44$.

Most loops arise from third-position changes. This reflects the extensive third-position synonymy of the standard genetic code.

Only one selected second-position edge preserves translated identity. This scarcity complements the physicochemical results of Part II, in which the second position carried the strongest hydropathy contrast.

The first position occupies an intermediate role. It contributes four synonymous edge instances.

The total loop count is therefore not distributed uniformly over the three codon axes. It records the strongly anisotropic way in which translated identity is organized within the otherwise isotropic G–A–C–U cube.

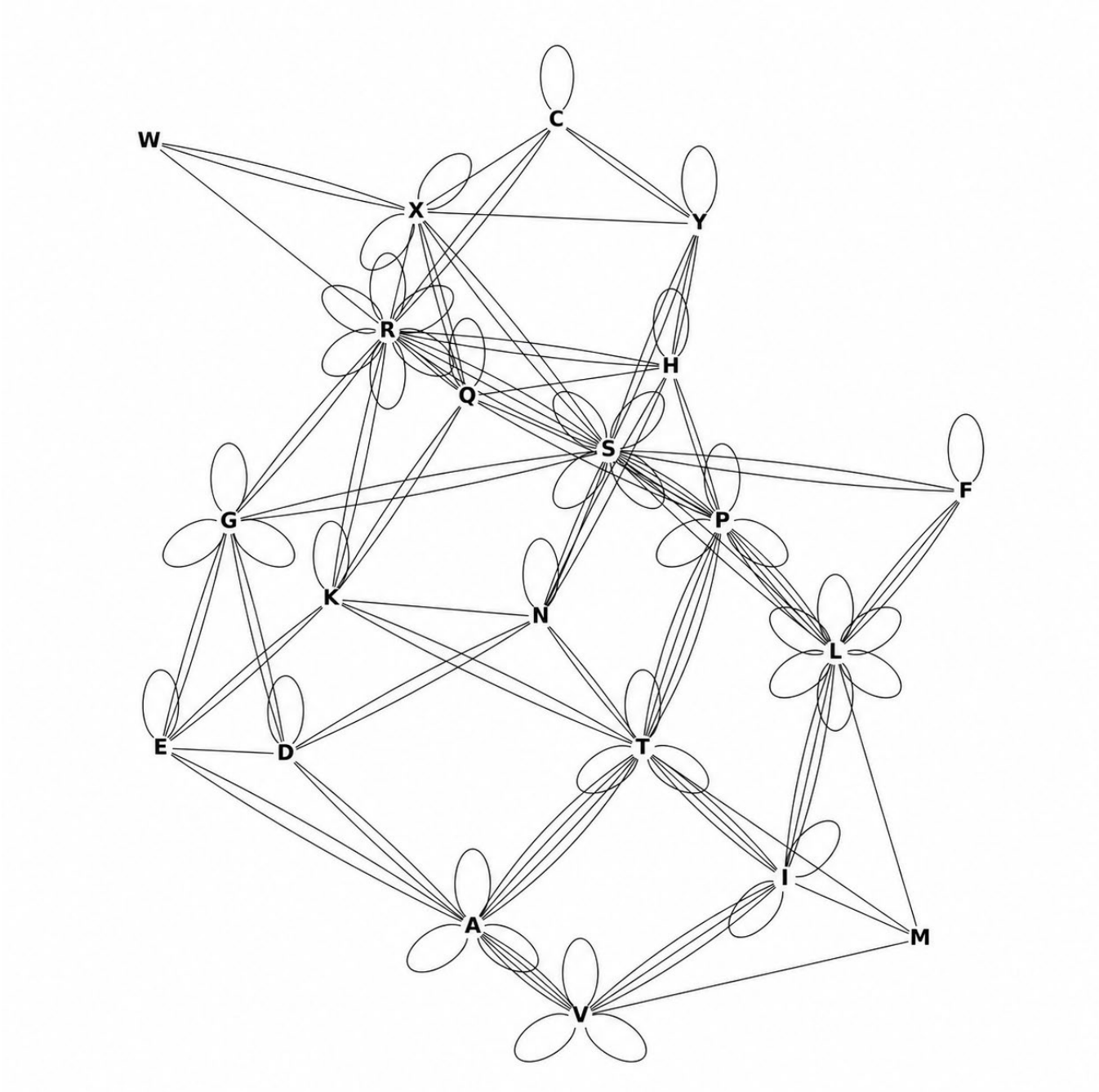


Figure 9. The G–A–C–U translated multigraph. The sixty-four codon vertices of the isotropic G–A–C–U cube have been collapsed into twenty-one vertices corresponding to the twenty amino acids and translation stop X. Each self-loop represents one nearest-neighbor cube edge connecting synonymous codons. The multigraph contains exactly forty-four self-loops and one hundred interclass edge instances. Parallel edges are retained because distinct codon-level adjacencies may collapse to the same pair of translated vertices.

The exact source figure should be preserved without graphical reinterpretation. Loop multiplicity and parallel-edge multiplicity are mathematical data. A visually similar redrawing could be incorrect even if all twenty-one translated labels were present.

An altered graph could inadvertently omit a loop, duplicate a loop, merge parallel edges, mislabel X, or change the local incidence structure. The authoritative graph is therefore part of the evidence rather than merely an illustration.

9.5 One Hundred Interclass Edge Instances

The remaining cube edges connect codons with different translated outputs.

Their number is

144 − 44 = 100.

The translated multigraph therefore contains:

44 self-loop edge instances;

100 interclass edge instances;

and

144 edge instances in total.

The one hundred interclass edges do not necessarily correspond to one hundred distinct unordered pairs of translated classes. Parallel edges are possible and common.

Let $m_{\omega,\nu}$ denote the number of codon-cube edges whose translated endpoints are ω and ν, with

$\omega \neq \nu$.

Then $m_{\omega,\nu}$ may be greater than one.

The total interclass multiplicity satisfies

$\Sigma_{\{\omega,\nu\}\subset\Omega,\ \omega\neq\nu} m_{\omega,\nu} = 100$.

The multigraph therefore preserves two related forms of translated adjacency.

The first is qualitative: whether at least one selected codon edge connects two translated classes.

The second is quantitative: how many distinct codon-level edge instances produce that translated connection.

A simple graph would preserve only the first. The multigraph preserves both.

The present book concentrates primarily on synonymous self-loops because they define the finite optimization objective used in Chapters 10 and 11. The interclass edge structure is nevertheless a substantial object in its own right. It can be analyzed for repeated amino-acid substitutions, position-specific connectivity, chemical similarity, mutation weighting, and natural-code variants.

9.6 Loop Multiplicity by Translated Class

For each translated class ω, define its loop multiplicity as

ℓω

= the number of G–A–C–U cube edges whose two endpoint codons both belong to Fω.

Equivalently,

ℓω

= | { { c, d } ∈ E : c ∈ Fω and d ∈ Fω } |.

The total synonymous-edge count is the sum over all translated classes:

L = Σω∈Ω ℓω.

The complete loop-multiplicity distribution is:

ℓR = 6,

ℓL = 6,

ℓS = 4;

ℓA = ℓG = ℓP = ℓT = ℓV = 3;

ℓI = ℓX = 2;

ℓC = ℓD = ℓE = ℓF = ℓH = ℓK = ℓN = ℓQ = ℓY = 1;

and

ℓM = ℓW = 0.

The sum is

6 + 6 + 4

- 5 (3)

- 2 (2)
- 9 (1)
- 2 (0)

= 16 + 15 + 4 + 9

= 44.

This gives an exact class-by-class decomposition of the global loop total.

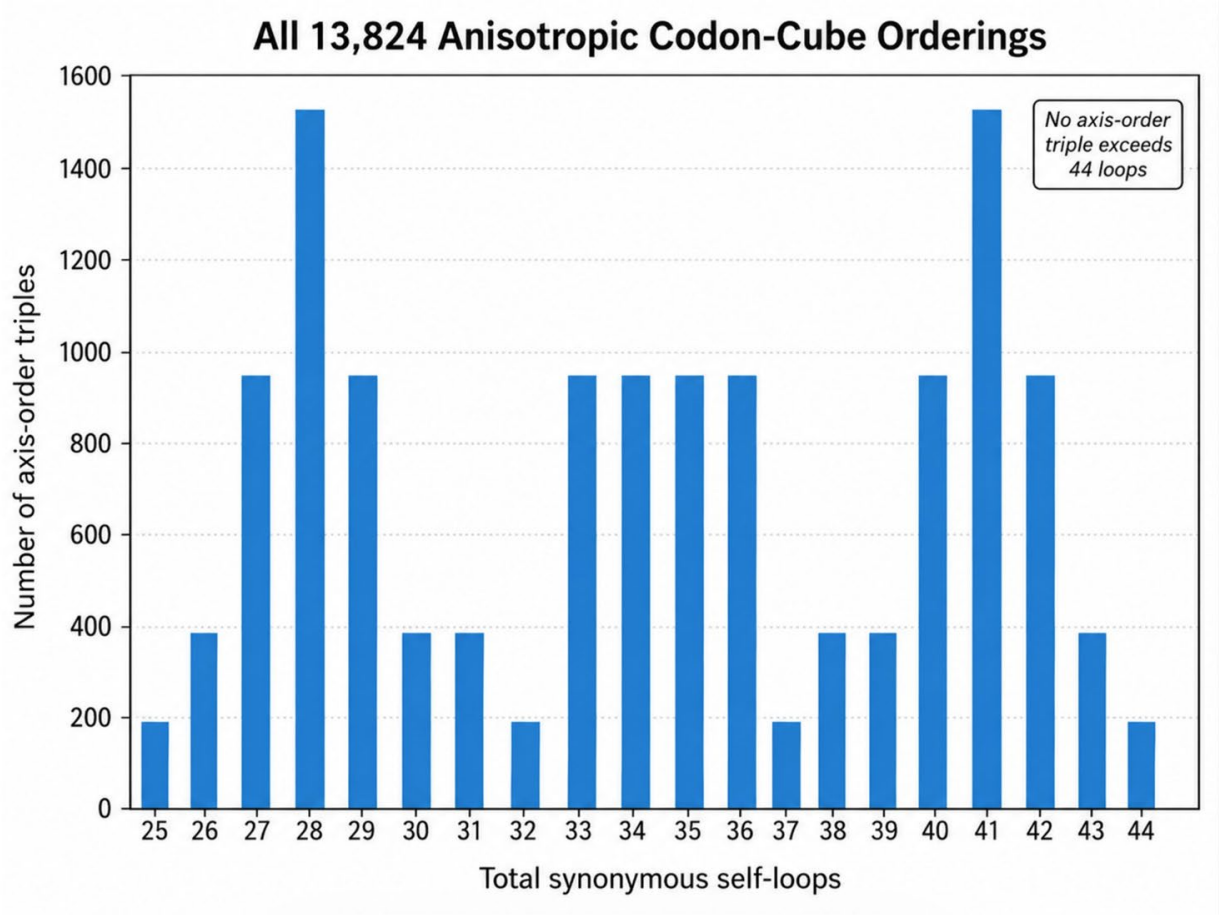


Figure 10. Synonymous self-loop multiplicity by translated class in the G–A–C–U codon cube. Arginine and leucine each contribute six self-loops, serine contributes four, and alanine, glycine, proline, threonine, and valine each contribute three. Isoleucine and translation stop X each contribute two. Nine two-codon amino-acid classes each contribute one loop, while methionine and tryptophan contribute none. The class totals sum exactly to forty-four.

The distribution reveals that the total forty-four is not generated by one uniform mechanism. Different translated fibers contribute according to their size and their placement in the selected codon geometry.

9.7 Six-Codon Fibers

Arginine, leucine, and serine each have six codons:

$dR = dL = dS = 6.$

Their loop multiplicities are not equal:

$\ell R = 6,$

$\ell L = 6,$

and

$\ell S = 4.$

The three fibers therefore have identical cardinality but different local topology.

Leucine and arginine each contain six selected internal nearest-neighbor edges. Serine contains four.

The difference arises because serine occupies two separated codon regions:

the four UCN codons;

and the two AGY codons, where Y denotes C or U.

A fiber can therefore be highly degenerate while remaining less locally compact under a particular codon geometry.

The comparison among R, L, and S supplies the clearest demonstration that degeneracy and local connectedness are distinct properties.

Degeneracy asks how many codons belong to the fiber.

Loop multiplicity asks how many selected nearest-neighbor edges remain inside it.

Two fibers of the same size can have different answers to the second question.

9.8 Fourfold-Degenerate Fibers

Alanine, glycine, proline, threonine, and valine each have four codons:

$dA = dG = dP = dT = dV = 4.$

Each contributes three loops:

ℓA = ℓG = ℓP = ℓT = ℓV = 3.

These fibers commonly consist of four codons having the same first two nucleotides while the third nucleotide follows the G–A–C–U path.

For example, the alanine codons are

GCG, GCA, GCC, and GCU.

Under the third-position order

G–A–C–U,

they form the path

GCG — GCA — GCC — GCU.

A four-vertex path contains three edges. All three preserve alanine identity, so the alanine fiber contributes three loops after translation collapse.

The same structure occurs for the other fourfold-degenerate families in this group.

The value

ℓω = 3

is therefore the loop signature of a complete four-codon third-position path under the selected ordering.

The degeneracy value four alone would not guarantee three loops under every nucleotide path. A different ordering of the four third-position states could place fewer synonymous nucleotide pairs next to one another for other types of fibers. The observed multiplicity reflects both the standard assignments and the chosen G–A–C–U adjacency.

9.9 Three-Codon Fibers

Isoleucine and translation stop X each have three codons:

dI = dX = 3.

Their loop multiplicities are

ℓI = ℓX = 2.

Three vertices arranged as one connected path contain two edges.

For isoleucine, the codons are

AUA, AUC, and AUU.

Within the third-position G–A–C–U path, A, C, and U occur consecutively:

A–C–U.

The corresponding codons therefore form

AUA — AUC — AUU.

The isoleucine fiber contributes two synonymous edges.

The stop codons are

UGA, UAA, and UAG.

Their local connections are not simply the same third-position path arrangement as isoleucine, because UGA differs from the other two at the second position. Nevertheless, the selected G–A–C–U cube contains two stop-to-stop edge instances in total.

After collapse, both become loops at X.

The treatment of X in this chapter is categorical. No hydropathy, residue volume, or amino-acid property is assigned to it. X participates because it is one of the twenty-one translated outputs of the complete lookup relation.

9.10 Two-Codon Fibers

Nine amino acids have two codons and one loop:

C, D, E, F, H, K, N, Q, and Y.

For every class ω in this group,

$d\omega = 2$

and

$\ell\omega = 1$.

The two codons are adjacent under the G–A–C–U topology and therefore form one connected synonymous pair.

Phenylalanine, for example, is encoded by

UUC

and

UUU.

Because C and U are adjacent on the selected path, these codons are joined by one third-position edge.

Aspartate is encoded by

GAC

and

GAU.

Again, C and U are adjacent, producing one synonymous edge.

Glutamate is encoded by

GAG

and

GAA.

G and A are adjacent, so the two codons are connected.

The same local relation holds for each of the nine two-codon classes listed above.

A two-codon fiber can contribute at most one internal undirected edge. These nine classes attain that maximum under G–A–C–U.

9.11 Single-Codon Fibers

Methionine and tryptophan each have one codon:

$dM = dW = 1$.

Therefore,

$\ell M = \ell W = 0$.

A fiber containing one vertex has no pair of distinct codons and cannot contain a synonymous edge.

Its zero loop count is therefore forced by cardinality and does not indicate poor geometric organization.

This observation illustrates why raw loop counts must be interpreted alongside fiber sizes. A large fiber has more opportunities to contain internal edges than a small fiber. A one-codon class cannot contribute a loop under any codon topology.

9.12 Degeneracy Versus Local Compactness

Let

$$d_\omega = | F_\omega |$$

denote degeneracy and let ℓ_ω denote loop multiplicity.

The two quantities measure different aspects of a translated class.

The degeneracy d_ω is purely a fiber cardinality.

The loop count ℓ_ω depends on the positions of those codons in the selected cube.

A simple normalized compactness index can be defined for fibers with at least two codons:

$$\kappa_\omega = 2 \ell_\omega / [d_\omega (d_\omega - 1)].$$

The denominator

$$d_\omega (d_\omega - 1) / 2$$

is the number of all possible unordered codon pairs within the fiber.

Thus, κ_ω is the fraction of all synonymous codon pairs that are realized as selected nearest-neighbor edges.

For leucine and arginine,

$$\kappa_L = \kappa_R$$

$$= 2 (6) / [6 (5)]$$

$$= 12 / 30$$

$$= 0.4.$$

For serine,

$$\kappa_S$$

$$= 2 (4) / [6 (5)]$$

$$= 8 / 30$$

$$= 4 / 15$$

$$\approx 0.2667.$$

For a four-codon family with three loops,

κ

= 2 (3) / [4 (3)]

= 6 / 12

= 0.5.

For a three-codon family with two loops,

κ

= 2 (2) / [3 (2)]

= 4 / 6

= 2 / 3.

For a two-codon family with one loop,

κ

= 2 (1) / [2 (1)]

= 1.

The normalized measure should be interpreted cautiously. Small fibers can attain high normalized values more easily because they contain fewer possible pairs.

A two-codon fiber with one edge receives κ = 1, while a six-codon fiber with six edges receives κ = 0.4 even though the larger fiber contains six times as many loops.

The raw degeneracy, raw loop count, and normalized compactness should therefore remain visible together.

9.13 Fiber Connectivity and Local Redundancy

Loop count and connectivity are related but not identical.

A fiber is connected in the induced G–A–C–U subgraph when every codon in the fiber can be reached from every other through synonymous cube edges.

A connected fiber with dω vertices requires at least

dω − 1

internal edges.

A four-codon family with three loops can form a connected path.

A three-codon fiber with two loops can also form a connected path.

A two-codon fiber with one loop is connected.

For a larger or more complicated fiber, the number of edges alone may not determine connectivity. Several edges could be concentrated within one component while other codons remain isolated. The arrangement must be inspected as well as counted.

At least three fiber properties should therefore be distinguished:

the degeneracy dω;

the internal edge count ℓω;

and the number of connected components of the induced fiber subgraph.

The present optimization objective emphasizes ℓω because internal edge counts add exactly over translated classes:

$L = \Sigma_{\omega \in \Omega} \ell_\omega$.

Connectivity is not additive in the same manner. It remains a complementary structural property.

The loops can also be interpreted as local redundancy. Each loop represents one selected single-position codon change that leaves the translated class unchanged.

A fiber with several loops contains several local routes of translation preservation. A fiber with no loops contains none under the selected geometry.

This interpretation should not be confused with mutation probability. Every selected edge is counted equally here. Biological substitution rates, transition–transversion biases, codon usage, and organism-specific mutation processes are not included. They could later be introduced as edge weights.

9.14 What the Quotient Preserves and Suppresses

The translated multigraph preserves:

the twenty-one output classes;

the total number of codon-cube edge instances;

which edges preserve translated identity;

the loop multiplicity of each translated class;

and the multiplicities of connections between distinct translated classes.

It suppresses:

the names of the individual codons after collapse;

the precise codon coordinates of an edge unless separately recorded;

the nucleotide position changed by an edge unless edges are position-labeled;

and the distinction among different codon pairs that become visually merged as parallel translated edges.

The cube and the multigraph therefore answer different questions.

The cube asks:

Where are the sixty-four codons located, and which codons are nearest neighbors?

The multigraph asks:

How are the twenty-one translated identities connected by those local codon movements?

Neither representation replaces the other.

The cube retains input detail.

The multigraph reveals translated topology.

9.15 Why the Distribution Matters

The global total

L = 44

could conceal many different internal organizations.

One hypothetical code might concentrate most synonymous edges inside a few large translated fibers.

Another might distribute them more evenly among many smaller fibers.

A third might have the same global total while producing disconnected synonymous regions.

The class-by-class loop distribution shows how the standard code realizes its forty-four-edge total.

Leucine and arginine contribute the largest individual counts.

Serine contributes fewer loops despite having the same degeneracy.

Five four-codon families form three-edge local structures.

Isoleucine and X form two-edge structures.

Nine two-codon families contribute one edge each.

Methionine and tryptophan necessarily contribute none.

The translated multigraph therefore connects the familiar degeneracy pattern of the standard code with a distinct measure of local codon-space organization.

The next chapter asks whether the total forty-four is exceptional. All twenty-four oriented isotropic orders of G, A, C, and U will be evaluated, and the exact nucleotide-pair weights will show why G–A–C–U and its reverse attain the unique undirected isotropic maximum.

Chapter 10

Exhaustive Search of Isotropic Nucleotide Orders

The G–A–C–U codon cube contains forty-four synonymous nearest-neighbor edges. That count is exact once the nucleotide path is fixed, but it does not by itself show that the selected path is exceptional.

The four RNA nucleotides can be placed along a linear path in twenty-four oriented orders. Each order selects three of the six possible unordered nucleotide adjacencies. When the same order is used on all three codon axes, it defines an isotropic 4 × 4 × 4 codon cube containing sixty-four vertices and 144 edges.

Every isotropic cube has the same size. Only the selected nucleotide adjacencies change.

The appropriate comparison is therefore exhaustive:

construct all twenty-four oriented nucleotide orders;

use each order on all three codon positions;

count the cube edges connecting synonymous codons;

and compare the complete score set.

The result is exact. G–A–C–U and its reverse achieve the maximum score of forty-four. No other undirected nucleotide path reaches that value.

10.1 The Twenty-Four Oriented Orders

A linear nucleotide order is a permutation

$\pi = n_1–n_2–n_3–n_4$

of

{ G, A, C, U }.

Because there are four distinct symbols, the number of oriented orders is

4! = 24.

Each order defines three adjacent unordered nucleotide pairs:

$\{ n_1, n_2 \}$,

$\{ n_2, n_3 \}$,

and

$\{ n_3, n_4 \}$.

For example,

G–A–C–U

selects

G–A,

A–C,

and

C–U.

The isotropic construction applies the same path to all three codon positions. Thus, G–A–C–U is used along the first, second, and third codon axes.

For a path π, define

$L (\pi)$

as the number of the 144 cube edges whose endpoint codons have the same translated output.

Equivalently,

$L (\pi)$

is the number of selected nearest-neighbor codon changes that preserve translated identity.

Every oriented order is evaluated by the same procedure. No path is granted a different number of edges, a different translation map, or a different scoring rule.

10.2 Reversal Symmetry

The reverse of

$\pi = n_1–n_2–n_3–n_4$

is

$\pi^R = n_4–n_3–n_2–n_1$.

The two paths contain exactly the same three unordered adjacencies.

For example,

G–A–C–U

and

U–C–A–G

both contain

G–A,

A–C,

and

C–U.

They therefore define the same undirected nucleotide graph and the same undirected codon cube.

Consequently,

$L (\pi) = L (\pi^R)$.

The twenty-four oriented orders occur in twelve reversal pairs. There are therefore twelve distinct undirected four-nucleotide paths.

Orientation can still be retained in a table or computational enumeration, but it does not affect the synonymous-edge objective used here.

When this chapter calls G–A–C–U the unique path maximum, uniqueness is understood up to reversal. The two oriented maximizing orders are

G–A–C–U

and

U–C–A–G.

They represent one undirected topology.

10.3 Complete Isotropic Enumeration

The complete oriented enumeration, paired with reversals, is:

G–A–C–U and U–C–A–G: 44;

G–A–U–C and C–U–A–G: 42;

A–G–C–U and U–C–G–A: 41;

A–G–U–C and C–U–G–A: 41;

A–C–U–G and G–U–C–A: 37;

A–U–C–G and G–C–U–A: 35;

C–A–G–U and U–G–A–C: 34;

C–G–A–U and U–A–G–C: 32;

C–A–U–G and G–U–A–C: 28;

G–C–A–U and U–A–C–G: 28;

A–C–G–U and U–G–C–A: 27;

A–U–G–C and C–G–U–A: 25.

The twenty-four oriented paths therefore have the following score frequencies:

score 44 occurs twice;

score 42 occurs twice;

score 41 occurs four times;

score 37 occurs twice;

score 35 occurs twice;

score 34 occurs twice;

score 32 occurs twice;

score 28 occurs four times;

score 27 occurs twice;

and

score 25 occurs twice.

The frequencies sum to

2 + 2 + 4 + 2 + 2 + 2 + 2 + 4 + 2 + 2 = 24.

Scores 29, 30, 31, 33, 36, 38, 39, 40, and 43 do not occur in the isotropic enumeration.

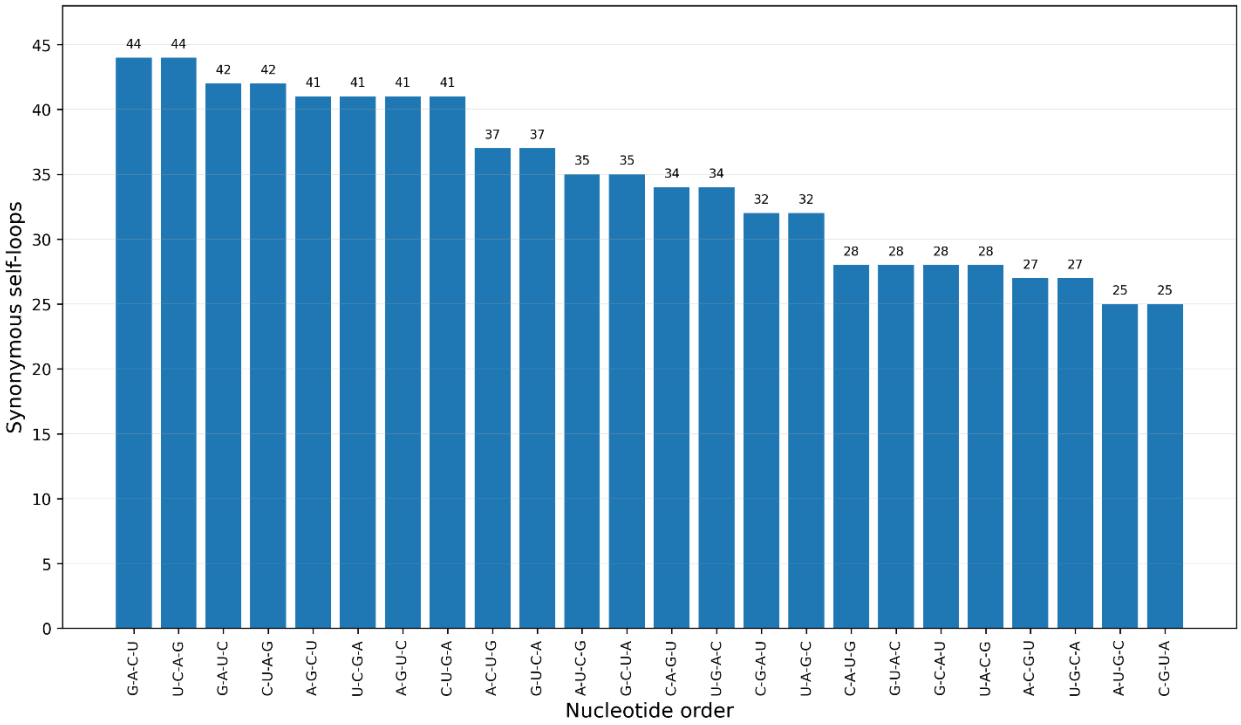


Figure 11. Synonymous-edge scores for all twenty-four oriented isotropic nucleotide orders. Each four-nucleotide path is used identically on the first, second, and third codon axes. G–A–C–U and its reverse U–C–A–G attain the unique undirected maximum of forty-four. G–A–U–C and its reverse score forty-two. A–U–G–C and its reverse attain the minimum score of twenty-five. Reversal pairs have equal scores because they contain the same three unordered nucleotide adjacencies.

The complete enumeration makes several facts immediately visible.

The G–A–C–U maximum is not tied with another undirected path.

The nearest competitor is separated by two edges.

Several paths occupy an intermediate range.

The lowest score is nineteen edges below the maximum.

The selected nucleotide ordering therefore has a substantial effect on the number of local synonymous adjacencies.

10.4 The Score Range

The maximum isotropic score is

Lmax = 44.

The minimum is

Lmin = 25.

The total range is

44 − 25 = 19.

Thus, replacing one nucleotide path with another can change the number of synonymous edges from twenty-five to forty-four while leaving the following properties unchanged:

the sixty-four codon vertices;

the standard translated assignments;

the number of codon axes;

the number of edges along each axis;

and the total of 144 cube edges.

Only the selected nucleotide adjacencies change.

The maximum preserves translated identity along

44 / 144 ≈ 0.3056

of the selected edges.

The minimum preserves translated identity along

25 / 144 ≈ 0.1736

of the selected edges.

The difference is

19 / 144 ≈ 0.1319

of the complete path-cube edge set.

The average isotropic score is

34.5.

This value can be obtained directly from the complete enumeration, but it also follows from the six nucleotide-pair weights developed below.

10.5 Position-Specific Nucleotide-Pair Weights

There are six possible unordered pairs of four nucleotide states:

A–G,

A–C,

A–U,

C–G,

C–U,

and

G–U.

For codon position p and unordered nucleotide pair { x, y }, define

$w_p (x, y)$

as the number of synonymous codon pairs that:

differ only at position p;

contain x and y at that position;

and agree at the other two codon positions.

Summing over all three codon positions gives the total pair weight

$w (x, y)$

$= w_1 (x, y) + w_2 (x, y) + w_3 (x, y).$

The exact position-specific and total weights are:

A–G:

$(w_1, w_2, w_3) = (0, 1, 14),$

so

w (A, G) = 15;

A–C:

(2, 0, 9),

so

w (A, C) = 11;

A–U:

(0, 0, 9),

so

w (A, U) = 9;

C–G:

(0, 0, 8),

so

w (C, G) = 8;

C–U:

(2, 0, 16),

so

w (C, U) = 18;

G–U:

(0, 0, 8),

so

w (G, U) = 8.

The six total weights are therefore

w (A, G) = 15,

w (A, C) = 11,

w (A, U) = 9,

w (C, G) = 8,

w (C, U) = 18,

and

w (G, U) = 8.

Their sum is

15 + 11 + 9 + 8 + 18 + 8 = 69.

The number sixty-nine is the total number of synonymous codon pairs at Hamming distance one when all six possible nucleotide substitutions are admitted.

The isotropic path cube selects only three of those six nucleotide-pair types. Its score is the sum of the corresponding three weights.

For an isotropic path π with adjacency set A (π),

L (π) = Σ{x,y}∈A(π) w (x, y).

The exhaustive cube calculation can therefore be replaced by a six-weight path calculation.

For G–A–C–U, the selected pairs are

G–A,

A–C,

and

C–U.

Therefore,

L (G–A–C–U)

= w (A, G) + w (A, C) + w (C, U)

= 15 + 11 + 18

= 44.

The pair-weight table does not approximate the cube score. It is an exact compression of it.

10.6 Exact Proof of the Isotropic Maximum

The six total pair weights, in decreasing order, are

C–U = 18,

A–G = 15,

A–C = 11,

A–U = 9,

C–G = 8,

and

G–U = 8.

A four-nucleotide linear path contains exactly three unordered edges. Therefore, no path can use more than three of the six weights.

The three largest weights are

18,

15,

and

11.

Their associated nucleotide pairs are

C–U,

A–G,

and

A–C.

These three edges form one connected linear path through all four nucleotide states:

G–A–C–U.

Therefore, the sum of the three largest weights is realizable by a valid four-nucleotide path:

18 + 15 + 11 = 44.

No path can exceed this value because any path contains only three pair weights, and no alternative selection of three weights can have a sum greater than the three largest.

Hence,

$L(\pi) \leq 44$

for every isotropic nucleotide path π.

Equality requires the path to contain all three maximum-weight pairs:

A–G,

A–C,

and

C–U.

Those edges determine the undirected path

G–A–C–U.

Therefore,

Lmax = 44,

and the maximizing oriented paths are exactly

G–A–C–U

and

U–C–A–G.

This is an exact finite proof, not merely a report from a numerical search.

The exhaustive enumeration independently confirms the theorem and provides the complete behavior of the nonmaximizing paths.

10.7 The Nearest Competitors

The second-highest isotropic score is

42.

It is attained by

G–A–U–C

and its reverse

C–U–A–G.

The selected nucleotide pairs are

G–A,

A–U,

and

U–C.

Therefore,

L (G–A–U–C)

= w (A, G) + w (A, U) + w (C, U)

= 15 + 9 + 18

= 42.

Relative to the maximizing path G–A–C–U, the nearest competitor retains the two largest pair weights:

A–G = 15,

and

C–U = 18.

It replaces

A–C = 11

with

A–U = 9.

The score decreases by exactly

11 − 9 = 2.

The next two undirected paths each score forty-one.

For A–G–C–U,

L

= w (A, G) + w (C, G) + w (C, U)

= 15 + 8 + 18

= 41.

For A–G–U–C,

L

= w (A, G) + w (G, U) + w (C, U)

= 15 + 8 + 18

= 41.

Both retain the two largest disjoint end-region weights A–G and C–U but use one of the weight-eight pairs to connect them.

The maximum is therefore not isolated by a vast numerical margin. Its significance lies in exact optimality and uniqueness, not in an enormous separation from every competitor.

10.8 The Lowest-Scoring Orders

The minimum isotropic score is

25.

It is attained by

A–U–G–C

and its reverse

C–G–U–A.

The selected pairs are

A–U,

U–G,

and

G–C.

Therefore,

L (A–U–G–C)

= w (A, U) + w (G, U) + w (C, G)

= 9 + 8 + 8

= 25.

These three pair types are exactly the nucleotide pairs omitted by the maximizing G–A–C–U path.

The complete set of six unordered nucleotide pairs can be partitioned into:

the three edges of G–A–C–U,

and

the three edges of A–U–G–C.

Their scores therefore sum to the total weight of all six pair types:

44 + 25 = 69.

More generally, the complement of any four-vertex path within the complete nucleotide graph is another four-vertex path. If one path has score L, its complementary path has score

69 − L.

This explains the symmetry of the isotropic score spectrum around

69 / 2 = 34.5.

Examples include:

44 + 25 = 69;

42 + 27 = 69;

41 + 28 = 69;

37 + 32 = 69;

and

35 + 34 = 69.

The mean score over all twelve undirected paths, and equivalently over all twenty-four oriented paths, is therefore

34.5.

10.9 Why Exhaustive Enumeration Remains Necessary

The pair-weight proof establishes the maximum without needing to inspect all twenty-four orders individually. Exhaustive enumeration nevertheless remains important.

It verifies that all oriented paths have been included.

It confirms reversal symmetry.

It identifies the nearest competitors.

It establishes the minimum.

It reveals the complete score spectrum.

It supplies a reproducible check against an error in the translation table, pair weights, or path-generation code.

It also shows that several mathematically possible integer scores never occur in the isotropic search.

For a search space containing only twenty-four oriented paths, complete enumeration is preferable to sampling or heuristic optimization. There is no need to estimate the maximum probabilistically when every candidate can be evaluated exactly.

10.10 The Unique Isotropic Maximum

The isotropic result can now be stated as a theorem.

Isotropic Maximum Theorem.
Let π be any linear ordering of the four RNA nucleotides. Use π identically on the first, second, and third codon axes to construct a 4 × 4 × 4 codon lattice. Score the lattice by the number of nearest-neighbor edges whose endpoint codons have the same output under the standard genetic code. Then

$L(\pi) \leq 44$.

Equality holds only for the oriented paths

G–A–C–U

and

U–C–A–G.

Thus, G–A–C–U is the unique maximizing undirected isotropic nucleotide path.

The theorem is exact within its stated comparison class and objective.

It does not claim that G–A–C–U is the uniquely correct physical geometry of nucleotide substitution.

It does not include mutation rates, codon usage, transition–transversion weighting, or chemical distances among nucleotides.

It establishes that, among all linear four-state nucleotide paths under equal edge weighting, no isotropic path preserves translated identity along more selected codon edges.

10.11 Isotropic Optimality Is Not Yet the Global Result

The isotropic search imposes one substantial restriction: the same nucleotide order must be used on all three codon axes.

The standard genetic code treats the first, second, and third positions differently. The hydropathy analysis showed that the second position carries the strongest physicochemical contrast. The synonymous-edge decomposition showed that the third position contains most of the translation-preserving edges.

A critic could therefore ask whether the total score can be increased by assigning a different nucleotide path to each codon position.

There are twenty-four oriented choices for the first axis, twenty-four for the second, and twenty-four for the third. The resulting anisotropic search contains

$24^3 = 13{,}824$

oriented axis-order combinations.

Chapter 11 evaluates that complete space.

The isotropic maximum survives the larger test. No anisotropic codon cube exceeds forty-four synonymous edges. The proof becomes stronger because the three codon positions can be optimized independently, yielding maxima of four, one, and thirty-nine. Their sum is again forty-four.

Chapter 11

Exhaustive Search of Anisotropic Codon Cubes

The isotropic search of Chapter 10 required the same nucleotide path to be used on all three codon axes. That restriction produced the unique undirected maximum

G–A–C–U

with forty-four synonymous nearest-neighbor edges.

The three codon positions, however, play different roles in the standard genetic code. The second position dominates the additive hydropathy signal. The third position contributes most synonymous variation. The first position provides a smaller but nonzero contribution to both translated-property refinement and local synonymous connectivity.

It is therefore reasonable to remove the isotropic restriction and permit each codon position to use its own nucleotide path.

The resulting anisotropic search contains 13,824 oriented codon cubes. Every cube has the same sixty-four codon vertices and the same total of 144 edges. The only difference is the linear nucleotide order assigned independently to each of the three coordinate axes.

The larger search does not produce a higher maximum.

The total synonymous-edge score still cannot exceed forty-four. The reason is exact axis separability: every cube edge belongs to one codon position, so the three positional contributions can be optimized independently.

The maximum positional scores are

4 at the first codon position,

1 at the second codon position,

and

39 at the third codon position.

Their sum is

4 + 1 + 39 = 44.

The anisotropic analysis also classifies every maximizing cube. There are 192 oriented maximizers, corresponding to twenty-four undirected anisotropic cubes. Among the positional maximum families, G–A–C–U is the unique undirected path common to all three codon positions.

11.1 Independent Nucleotide Orders on Three Axes

Let

π_1

denote the nucleotide path used on the first codon axis,

π_2

the path used on the second codon axis,

and

π_3

the path used on the third codon axis.

Each path is an oriented permutation of

G, A, C, and U.

There are

$4! = 24$

oriented choices for each axis.

The number of oriented anisotropic axis-order triples is therefore

$24 \times 24 \times 24 = 24^3 = 13{,}824$.

An anisotropic cube is specified by

(π_1, π_2, π_3).

For example, one cube might use

G–A–C–U

at the first position,

A–G–U–C

at the second position,

and

G–A–U–C

at the third position.

Each path contains three nucleotide adjacencies. Each adjacency is repeated over sixteen settings of the other two codon coordinates. Every axis therefore still contains forty-eight edges, and every anisotropic cube contains

$48 + 48 + 48 = 144$

edges.

The anisotropic construction changes which nucleotide pairs are adjacent at each position. It does not change the codon set, translation map, or edge budget.

11.2 Axis-Separable Scoring

Let

$L_1 (\pi_1)$

be the number of synonymous edges produced by first-position path π_1.

Let

$L_2 (\pi_2)$

be the corresponding second-position score.

Let

$L_3 (\pi_3)$

be the corresponding third-position score.

The total anisotropic score is

$L (\pi_1, \pi_2, \pi_3)$

$= L_1 (\pi_1) + L_2 (\pi_2) + L_3 (\pi_3).$

There are no interaction terms.

Every cube edge changes exactly one codon position. A first-axis edge is determined entirely by the first-position nucleotide path. Its contribution cannot depend on how the second or third axes are ordered.

The same is true for the other two positions.

This additivity turns the search from a seemingly large three-axis optimization problem into three independent four-nucleotide path problems.

The global maximum is therefore

max L

$= \max L_1 + \max L_2 + \max L_3.$

The complete enumeration remains useful because it verifies the calculation and supplies the full score distribution. The maximum itself can be proved from the three positional spectra.

11.3 Position-Specific Score Spectra

The first codon position has only two synonymous nucleotide-pair types with nonzero weight:

A–C contributes 2;

C–U contributes 2.

All other first-position nucleotide pairs contribute zero.

A path may contain neither, one, or both of these edges. Therefore, the possible first-position scores are

0, 2, and 4.

Among the twenty-four oriented nucleotide paths:

4 score 0;

16 score 2;

and

4 score 4.

The first-position score polynomial is

$F_1 (x) = 4 + 16 x^2 + 4 x^4$.

The second-position structure is even sparser.

Only the nucleotide pair A–G produces a synonymous second-position edge, and its weight is one. A path either contains A–G or it does not.

The possible second-position scores are therefore

0 and 1.

Among the twenty-four oriented paths:

12 score 0;

and

12 score 1.

The second-position score polynomial is

$F_2 (x) = 12 + 12 x$.

The third codon position contains most of the synonymous structure. Its nucleotide-pair weights are:

A–G = 14;

A–C = 9;

A–U = 9;

C–G = 8;

C–U = 16;

and

G–U = 8.

The twenty-four oriented third-position paths produce six possible scores:

25, 26, 31, 33, 38, and 39.

Each score is attained by four oriented paths.

Thus,

$F_3 (x)$

$= 4 x^{25} + 4 x^{26} + 4 x^{31} + 4 x^{33} + 4 x^{38} + 4 x^{39}$.

The position-specific spectra are therefore:

First position:

0 with multiplicity 4;

2 with multiplicity 16;

4 with multiplicity 4.

Second position:

0 with multiplicity 12;

1 with multiplicity 12.

Third position:

25, 26, 31, 33, 38, and 39, each with multiplicity 4.

The multiplicities sum to twenty-four within every positional spectrum.

Codon position	Possible score	Oriented nucleotide orders
First	0	4
First	2	16
First	4	4
Second	0	12
Second	1	12
Third	25	4
Third	26	4
Third	31	4
Third	33	4
Third	38	4
Third	39	4

Table 3. Position-specific synonymous-edge score spectra for all twenty-four oriented nucleotide paths. The first codon position contributes scores of 0, 2, or 4. The second contributes only 0 or 1. The third contributes scores from 25 to 39. The corresponding maximum contributions are 4, 1, and 39.

The spectra immediately yield the global bounds.

The smallest possible anisotropic score is

0 + 0 + 25 = 25.

The largest possible score is

4 + 1 + 39 = 44.

Thus, the entire 13,824-cube search must lie within the interval

25 ≤ L ≤ 44.

11.4 The Exact Generating Polynomial

Because the three positional scores are independent, the generating polynomial for the complete anisotropic score distribution is the product

$F(x) = F_1(x)\,F_2(x)\,F_3(x)$.

Substituting the exact positional polynomials gives

$$F(x) = (4 + 16x^2 + 4x^4) \times (12 + 12x) \times (4x^{25} + 4x^{26} + 4x^{31} + 4x^{33} + 4x^{38} + 4x^{39}).$$

The coefficient of x^s in $F(x)$ is the number of oriented anisotropic codon cubes having synonymous-edge score s.

Evaluating the polynomial at x = 1 gives the total number of cubes:

$$F(1) = 24 \times 24 \times 24 = 13{,}824.$$

The smallest power is x^{25}, confirming the minimum score of twenty-five.

The largest power is x^{44}, confirming the maximum score of forty-four.

The coefficient of x^{44} is

$4 \times 12 \times 4 = 192$,

because a maximum cube requires one of four maximizing oriented paths at the first position, one of twelve at the second position, and one of four at the third.

The generating polynomial therefore encodes both the score distribution and the Cartesian-product structure of the maximum family.

11.5 The Complete Score Distribution

Let

N_s

denote the number of oriented anisotropic cubes with total score s.

The exact frequencies are:

$N_{25} = 192$;

$N_{26} = 384$;

$N_{27} = 960$;

$N_{28} = 1{,}536$;

$N_{29} = 960$;

$N_{30} = 384$;

$N_{31} = 384$;

$N_{32} = 192$;

$N_{33} = 960$;

$N_{34} = 960$;

$N_{35} = 960$;

$N_{36} = 960$;

$N_{37} = 192$;

$N_{38} = 384$;

$N_{39} = 384$;

$N_{40} = 960$;

$N_{41} = 1{,}536$;

$N_{42} = 960$;

$N_{43} = 384$;

and

$N_{44} = 192$.

The frequencies sum exactly to

13,824.

Every integer score from twenty-five through forty-four occurs. This differs from the isotropic spectrum, in which several values in the same range are absent.

The two most frequent anisotropic scores are

28

and

41.

Each is attained by

1,536

oriented cubes.

The distribution is not unimodal. Its internal rises and falls reflect the discrete combination of three small positional spectra rather than a smooth continuous process.

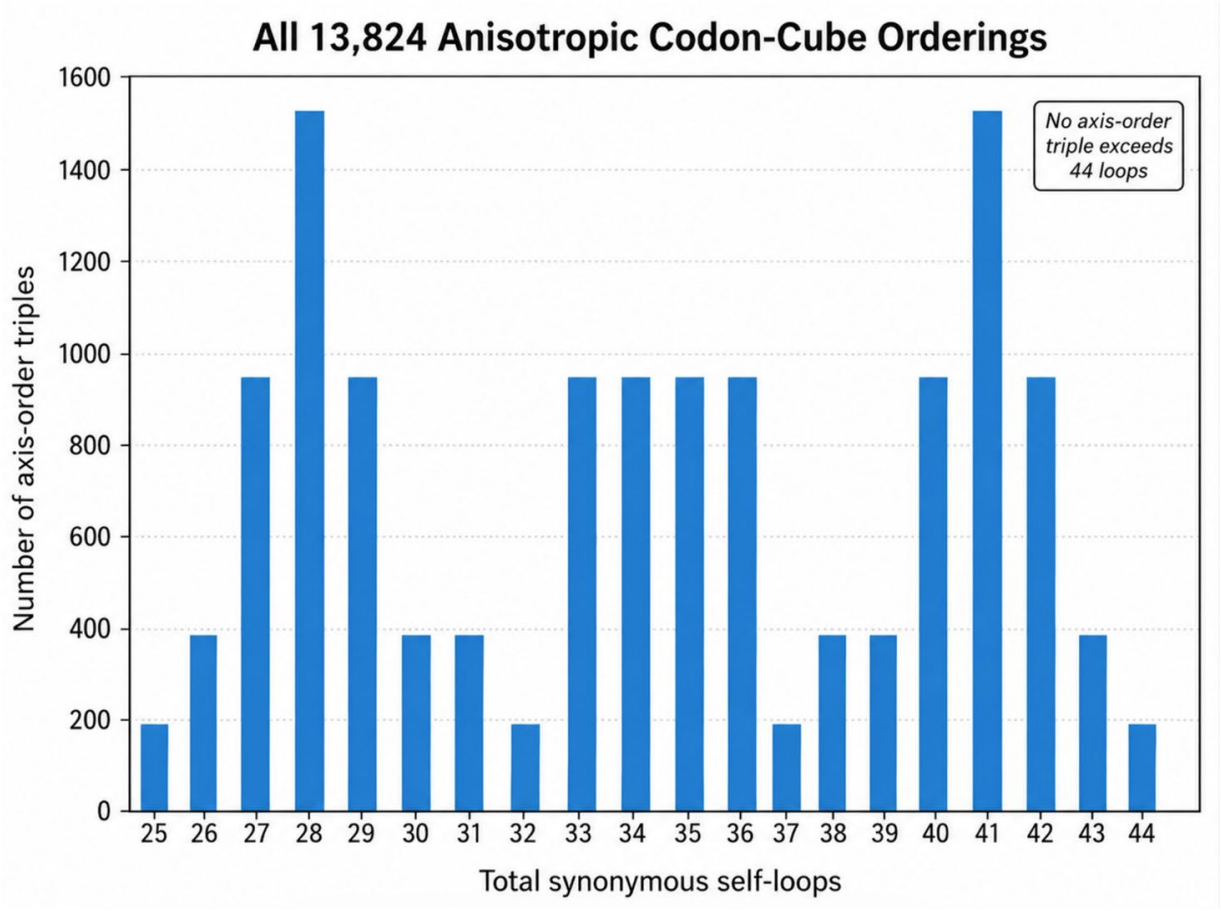


Figure 12. Exact distribution of synonymous-edge totals across all 13,824 oriented anisotropic codon cubes. Each codon position independently uses one of the twenty-four oriented linear nucleotide paths. Scores range from twenty-five to forty-four. The mean and median are 34.5, the population standard deviation is 5.5, and the distribution is exactly symmetric. A total of 192 oriented cubes attain the maximum score of forty-four, and 192 attain the minimum score of twenty-five. The largest frequency is 1,536, attained at scores twenty-eight and forty-one.

11.6 Mean, Variance, and Symmetry

The mean first-position score is

$[4 (0) + 16 (2) + 4 (4)] / 24$

$= 48 / 24$

$= 2.$

The mean second-position score is

$[12 (0) + 12 (1)] / 24$

$= 12 / 24$

$= 0.5.$

The mean third-position score is

$4 (25 + 26 + 31 + 33 + 38 + 39) / 24$

$= 192 / 6$

$= 32.$

The mean total score is therefore

$2 + 0.5 + 32 = 34.5.$

This is also the midpoint of the minimum and maximum:

$(25 + 44) / 2 = 34.5.$

The exact population variance is

30.25.

Therefore, the population standard deviation is

$\sqrt{30.25} = 5.5.$

The maximum score lies

$(44 - 34.5) / 5.5$

$= 9.5 / 5.5$

≈ 1.727

population standard deviations above the mean.

The minimum lies the same distance below it.

The complete frequency distribution satisfies

$N_s = N_{69-s}$.

For example,

$N_{25} = N_{44} = 192$;

$N_{26} = N_{43} = 384$;

$N_{27} = N_{42} = 960$;

and

$N_{28} = N_{41} = 1{,}536$.

The symmetry arises from complementary nucleotide paths.

At the first codon position, the weights of all six nucleotide-pair types sum to four. The complement of a path scoring L_1 therefore scores

$4 - L_1$.

At the second position, all pair weights sum to one, so the complementary score is

$1 - L_2$.

At the third position, all pair weights sum to sixty-four, so the complementary score is

$64 - L_3$.

The complementary anisotropic cube therefore has total score

$(4 - L_1) + (1 - L_2) + (64 - L_3)$

$= 69 - L$.

Complementation pairs every score s with score 69 − s, producing exact symmetry around 34.5.

11.7 First-Position Maximizers

The first-position maximum is

$\max L_1 = 4$.

To attain four, a nucleotide path must contain both nonzero first-position pair types:

A–C

and

C–U.

These two edges form the segment

A–C–U.

The remaining nucleotide G can be attached to either endpoint of that segment.

The two undirected maximizing paths are therefore

G–A–C–U

and

A–C–U–G.

Their four oriented forms are:

G–A–C–U;

U–C–A–G;

A–C–U–G;

and

G–U–C–A.

Thus, the first-position maximum family contains:

2 undirected paths;

and

4 oriented paths.

The first position does not uniquely select G–A–C–U. It admits one additional undirected order.

11.8 Second-Position Maximizers

The second-position maximum is

max $L_2 = 1$.

Only the pair A–G contributes a synonymous second-position edge. Therefore, a nucleotide path maximizes the second-position score exactly when A and G are adjacent.

Among the twelve undirected four-nucleotide paths, six contain the edge A–G.

One convenient listing is:

G–A–C–U;

A–G–C–U;

A–G–U–C;

C–G–A–U;

C–A–G–U;

and

C–U–A–G.

Each represents its reverse as the same undirected path.

The second-position maximum family therefore contains:

6 undirected paths;

and

12 oriented paths.

The second position imposes the weakest path restriction. Half of all undirected nucleotide paths attain its maximum.

This weak topological restriction contrasts with the strong physicochemical role of the second codon position. It contributes only one selected synonymous edge, yet it carries the largest additive hydropathy contrast.

11.9 Third-Position Maximizers

The third-position maximum is

max $L_3 = 39$.

The two largest third-position pair weights are

C–U = 16

and

A–G = 14.

A maximum path also requires one of the two weight-nine edges:

A–C = 9

or

A–U = 9.

The two undirected maximizing paths are therefore

G–A–C–U

and

G–A–U–C.

Their four oriented forms are:

G–A–C–U;

U–C–A–G;

G–A–U–C;

and

C–U–A–G.

Each gives

14 + 16 + 9 = 39.

Thus, the third-position maximum family contains:

2 undirected paths;

and

4 oriented paths.

As at the first position, the local maximum does not uniquely determine G–A–C–U. It admits one alternative undirected path.

11.10 Position-Specific Maximum Families

The three positional maximum families can now be summarized.

First position:

maximum score 4;

2 undirected maximizing paths;

4 oriented maximizing paths.

Second position:

maximum score 1;

6 undirected maximizing paths;

12 oriented maximizing paths.

Third position:

maximum score 39;

2 undirected maximizing paths;

4 oriented maximizing paths.

Codon position	Maximum score	Maximizing path or condition	Undirected / oriented
First	4	G–A–C–U; A–C–U–G	2 / 4
Second	1	Any path containing A–G	6 / 12
Third	39	G–A–C–U; G–A–U–C	2 / 4
Complete	44	Cartesian product	24 / 192

Table 4. Classification of the position-specific maximum families. The first codon position has two undirected maximizing paths, the second has six, and the third has two. Their Cartesian product contains twenty-four undirected anisotropic cubes, corresponding to 192 oriented cubes, each with the global maximum score of forty-four. The only undirected path common to all three positional maximum sets is G–A–C–U.

11.11 Cartesian Product of the Maximum Families

A globally maximizing anisotropic cube must maximize each codon position independently.

Therefore, the number of undirected maximum cubes is

$2 \times 6 \times 2 = 24$.

Every undirected axis path has two orientations. An anisotropic cube has three independently oriented axes, so every undirected cube corresponds to

$2^3 = 8$

oriented cubes.

The number of oriented maximizers is therefore

$24 \times 8 = 192.$

Equivalently, using the oriented positional counts directly,

$4 \times 12 \times 4 = 192.$

This agrees with the coefficient of x^{44} in the exact generating polynomial.

The complete anisotropic search contains

$12^3 = 1{,}728$

undirected cubes

or

$24^3 = 13{,}824$

oriented cubes.

The proportion attaining the maximum is therefore

$24 / 1{,}728$

$= 192 / 13{,}824$

$= 1 / 72$

$\approx 0.013889.$

Approximately 1.39 percent of the anisotropic cubes reach the maximum.

The maximum is therefore not unique at the level of independently selected axis paths. Its stronger distinction lies in the intersection of the three positional families.

11.12 Isotropic Members of the Maximum Family

An isotropic cube requires one common nucleotide path on all three axes:

$\pi_1 = \pi_2 = \pi_3.$

For such a cube to attain the global anisotropic maximum, the common path must belong simultaneously to the first-, second-, and third-position maximum families.

Let

M_1

be the set of undirected first-position maximizers,

M_2

the set of undirected second-position maximizers,

and

M_3

the set of undirected third-position maximizers.

Then an isotropic maximum path must belong to

$M_1 \cap M_2 \cap M_3$.

The first-position set is

M_1

= { G–A–C–U, A–C–U–G }.

The second-position set contains the six undirected paths having A–G as an edge.

The third-position set is

M_3

= { G–A–C–U, G–A–U–C }.

The intersection is

$M_1 \cap M_2 \cap M_3$

= { G–A–C–U }.

The reverse path

U–C–A–G

represents the same undirected topology.

Therefore, G–A–C–U is the unique undirected nucleotide path that independently maximizes synonymous adjacency at all three codon positions.

11.13 G–A–C–U as the Unique Common Undirected Path

The distinction between uniqueness and multiplicity must be stated precisely.

G–A–C–U is not the only anisotropic cube having forty-four synonymous edges.

There are:

twenty-four undirected anisotropic maximum cubes;

and

192 oriented anisotropic maximum cubes.

What is unique is the common isotropic solution.

G–A–C–U is the only undirected linear nucleotide path that can be placed on all three codon axes while independently attaining:

the first-position maximum of 4;

the second-position maximum of 1;

and

the third-position maximum of 39.

This result is stronger than saying that G–A–C–U ranked first in the twenty-four-path isotropic enumeration.

The isotropic enumeration establishes that the path has the largest total when one common order is imposed.

The anisotropic classification establishes why it has that status: it lies in the intersection of all three positional maximum families.

No individual codon position forces the full order by itself.

The first position permits two paths.

The third position permits two paths.

The second position permits six paths.

The unique complete path emerges only when the three local constraints are imposed together.

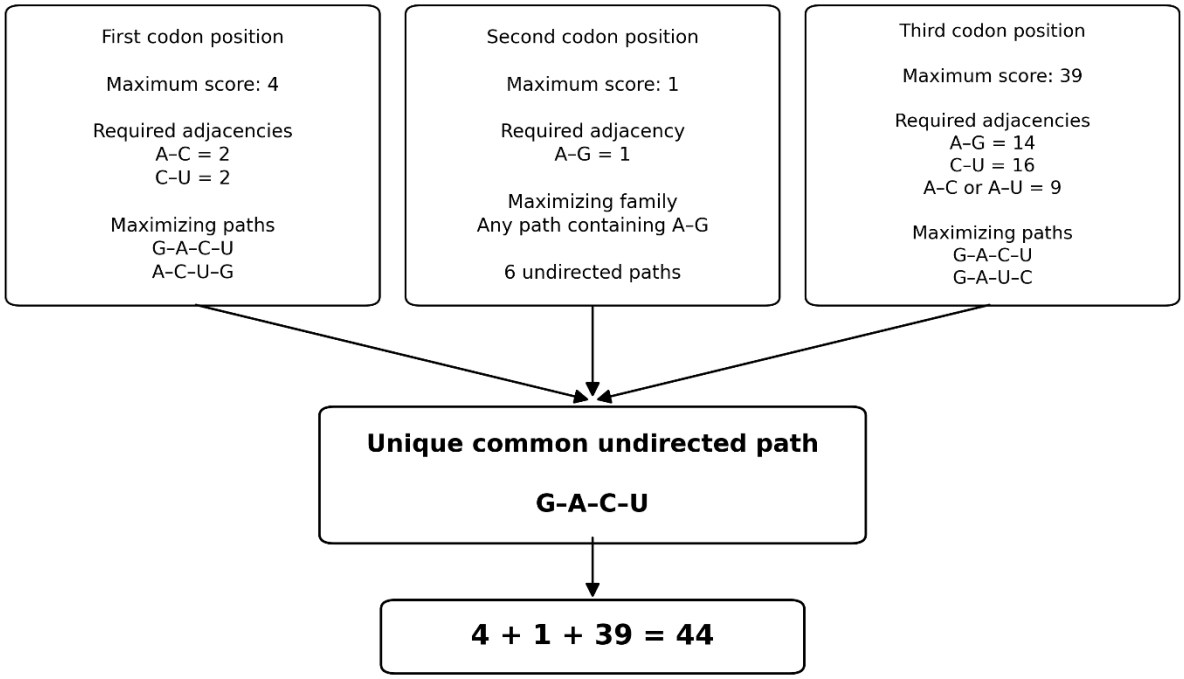


Figure 13. Axis-separable summary of the forty-four-edge maximum. The independent positional maxima are 4, 1, and 39, giving the global bound 44. The first-position maximum family contains two undirected paths, the second-position family contains six, and the third-position family contains two. Their Cartesian product yields twenty-four undirected anisotropic maximizers. G–A–C–U is the unique undirected path in the intersection of all three positional families.

11.14 The Exact Forty-Four-Edge Maximum Theorem

The anisotropic result can be stated as a theorem.

Forty-Four-Edge Maximum Theorem.
Assign an arbitrary linear order of G, A, C, and U independently to each of the three codon positions. Construct the corresponding 4 × 4 × 4 codon lattice and score it by the number of nearest-neighbor edges whose endpoint codons have the same translated output under the standard genetic code. Then

$L(\pi_1, \pi_2, \pi_3) \leq 44$.

The bound is attained.

Proof.
The score separates exactly by codon position:

$L (\pi_1, \pi_2, \pi_3)$

$= L_1 (\pi_1) + L_2 (\pi_2) + L_3 (\pi_3)$.

The maximum positional scores are

$\max L_1 = 4$,

$\max L_2 = 1$,

and

$\max L_3 = 39$.

Therefore,

$L (\pi_1, \pi_2, \pi_3)$

$\leq 4 + 1 + 39$

$= 44$.

Each positional maximum family is nonempty, so paths can be selected independently that attain all three maxima. Hence, forty-four is both an upper bound and an attained score. ■

A second theorem identifies the unique isotropic member.

Unique Common-Path Theorem.
Among undirected linear nucleotide paths, G–A–C–U is the unique path belonging to the maximizing families of the first, second, and third codon positions simultaneously.

This theorem follows from the exact intersection

$M_1 \cap M_2 \cap M_3$

$= \{ G–A–C–U \}$.

11.15 What the Anisotropic Search Establishes

The complete anisotropic search establishes several distinct results.

First, forty-four is the absolute maximum within the family of three-dimensional codon lattices formed from independently ordered four-state nucleotide paths.

Second, the maximum is not an artifact of requiring one isotropic order on all three axes.

Third, the full family of anisotropic maximizers can be classified exactly as a Cartesian product of three positional maximum sets.

Fourth, the number of maximizers is exact:

192 oriented cubes;

or

twenty-four undirected cubes.

Fifth, the complete score distribution is known, not estimated. It ranges from twenty-five to forty-four, has mean and median 34.5, has population standard deviation 5.5, and is exactly symmetric.

Sixth, the isotropic G–A–C–U cube is distinguished within the larger maximum family because its one undirected nucleotide path independently maximizes all three codon positions.

The result is therefore not that only one possible codon cube reaches forty-four.

The result is that forty-four is the global axis-separable maximum and G–A–C–U is the unique common isotropic path realizing all three positional maxima simultaneously.

The numerical decomposition

$44 = 4 + 1 + 39$

also summarizes the unequal roles of the codon positions.

The third position dominates local translation preservation.

The second position contributes almost no synonymous adjacency under the maximizing topology, yet it dominates hydropathy discrimination.

The first position supplies a smaller intermediate contribution.

The standard genetic code therefore distributes different mathematical functions across its three coordinates. Part IV brings the algebraic, physicochemical, and topological views together and asks what this combined structure does—and does not—permit us to infer.

Part IV

Interpretation and Generalization

Chapter 12

One Code, Three Mathematical Views

The preceding chapters have examined the standard genetic code through three mathematical representations.

The first is algebraic. The complete set of sixty-four codons becomes a balanced 64 × 12 binary incidence matrix whose rank, singular spectrum, null spaces, pseudoinverse, and projection operators can be derived exactly.

The second is physicochemical. Amino-acid properties are transferred through the translation map to create numerical fields over codon space. Hydropathy then projects strongly onto nucleotide identity at the second codon position, especially the opposition between U and A.

The third is topological. The four nucleotide states are arranged along linear paths, the codons become vertices of a 4 × 4 × 4 lattice, and synonymous nearest-neighbor edges measure local preservation of translated identity. Under this objective, G–A–C–U is the unique undirected path common to the maximizing families of all three codon positions.

These are not three unrelated curiosities. They are three views of one finite lookup relation.

Their synthesis requires care because some results follow from the complete codon domain alone, while others depend on the particular biological assignments of the standard genetic code. The mathematical strength of the analysis lies partly in keeping those layers separate and then showing how they interact.

12.1 The Lookup Relation

The standard genetic code is the map

$\tau : C \to \Omega,$

where C is the sixty-four-member RNA codon set and Ω contains twenty amino-acid classes together with translation stop X.

This relation is the common object underlying every representation in the book.

The binary matrix D represents the input identities in C.

A codon-level hydropathy field h is created by composing τ with an amino-acid property assignment.

The translated multigraph is created by collapsing codon vertices according to equal values of τ.

Thus, the three principal analyses begin from different aspects of the same relation:

the matrix begins from codon composition;

the property projection begins from numerical values assigned to translated outputs;

and the topology begins from local adjacency among codons followed by collapse under translated identity.

The conventional genetic-code table combines these aspects visually but does not distinguish them mathematically. It places codon spelling and translated output in one chart. The present analysis separates them so that input architecture, translated assignment, and local geometry can be studied independently.

This separation is not merely organizational. It prevents conclusions derived from one layer from being attributed incorrectly to another.

12.2 Algebraic Balance

The complete codon domain is the Cartesian product

Σ^3,

where

$\Sigma = \{ G, A, C, U \}$.

Its one-hot positional representation has twelve columns:

four nucleotide indicators at each of three positions.

Because every possible codon occurs, the design is exactly balanced.

Every row contains three 1s and nine 0s.

Every column contains sixteen 1s and forty-eight 0s.

Every pair of coordinates from different codon positions co-occurs in four rows.

Distinct nucleotide coordinates within one position never co-occur.

These balance conditions determine the Gram matrix

$D^T D$

and its spectrum.

The complete matrix has:

rank ten;

one singular value equal to 4 √3;

nine singular values equal to 4;

and two singular values equal to zero.

The two zero singular values arise because the four columns in each positional block sum to the same constant codon vector. The three block sums supply only one independent constant direction, leaving two redundant blockwise offsets.

The exact Moore–Penrose pseudoinverse is

$D^{+} = (1 / 16) D^{T} - (1 / 96) J.$

Every entry is either

5 / 96

or

−1 / 96.

The positive value occurs when a codon contains the nucleotide-position coordinate represented by the pseudoinverse row. The negative value occurs when it does not.

The associated codon-space projection matrix is

$P = D D^{+}$

$= (1 / 16) D D^{T} - (1 / 32) J.$

Its entries depend only on codon Hamming distance:

5 / 32 for identical codons;

3 / 32 for codons differing at one position;

1 / 32 for codons differing at two positions;

and

−1 / 32 for codons differing at all three positions.

These results require no amino-acid assignments. They would remain true if the translated labels were replaced, permuted, or removed.

The exact matrix structure therefore belongs to the input architecture of complete three-letter words over a four-state alphabet.

12.3 Physicochemical Projection

The algebra becomes biologically specific when a numerical translated property is placed over the sixty-four codons.

Let

$H : \Omega \to \mathbb{R}$

assign a numerical value to every translated class under a declared convention. The induced codon field is

$h\,c = H (\tau (c))$.

The exact pseudoinverse produces the minimum-norm nucleotide-position coefficient vector

$\beta = D^{+} h$.

The fitted codon field is

$\hat{h} = D \beta$

$= D D^{+} h$.

This field is the closest additive representation of h of the form

$\hat{h}\ c_1 c_2 c_3$

$= \beta 1,c_1 + \beta 2,c_2 + \beta 3,c_3$.

For Kyte–Doolittle hydropathy with the declared complete-code convention

$H (X) = 0$,

the strongest coefficient opposition occurs at the second codon position.

The complete-code values are

$\beta 2,U = 3.948958$

and

$\beta 2,A = -2.638542$.

Their gauge-invariant contrast is

$\beta 2,U - \beta 2,A = 6.587500$.

The corresponding G coefficient is also negative:

$\beta_{2,G} = -1.469792$.

The C coefficient lies much closer to the center:

$\beta_{2,C} = -0.163542$.

The four second-position states therefore define a broad hydropathy ordering in which U is strongly associated with hydrophobic outputs, A with hydrophilic outputs, G with a moderately negative tendency, and C with an intermediate tendency.

The positional coefficient ranges reinforce this result.

The first-position range is approximately 2.131250.

The second-position range is 6.587500.

The third-position range is approximately 0.506250.

The second position dominates hydropathy discrimination, while the third position contributes comparatively little.

The codon-level projection captures approximately 81.7 percent of centered squared hydropathy variation. After fitted codon values are averaged within synonymous fibers, the reconstructed amino-acid values correlate with empirical Kyte–Doolittle hydropathy at approximately

$r = 0.915646$,

with

$R^2 = 0.838407$.

The result is not an exact recovery of amino-acid chemistry. It is a strong alignment between translated hydropathy and the additive coordinate structure of the codon inputs.

12.4 Synonymous Topology

The topological representation introduces a local relation among nucleotide states.

A linear nucleotide path selects three of the six possible unordered nucleotide pairs. Applying that path to all three codon positions produces a sparse $4 \times 4 \times 4$ lattice containing

64 vertices

and

144 edges.

Under the isotropic G–A–C–U path, the selected nucleotide adjacencies are

G–A,

A–C,

and

C–U.

Every lattice edge changes one codon position by one step along this path.

When the endpoints of an edge translate to the same output, the edge is synonymous. Collapsing all synonymous codons to their translated identities turns such an edge into a self-loop.

The G–A–C–U cube contains

44 synonymous edges

and

100 nonsynonymous edges.

The synonymous total separates by codon position as

44 = 4 + 1 + 39.

The first position contributes four.

The second contributes one.

The third contributes thirty-nine.

The third codon axis therefore carries most local translation preservation. The second carries almost none.

The same positional asymmetry appeared in the hydropathy projection in an inverse form. The second position carried the strongest translated-property contrast, while the third carried the weakest.

A codon position that often changes without changing translated identity will also tend to contribute less to a numerical property inherited from translated identity.

The topology and physicochemical projection are therefore not independent observations. They express complementary consequences of the same assignment structure.

12.5 Input-Space and Assignment-Specific Results

The results of the book divide naturally into two principal categories.

The first category contains input-space results.

These follow from the complete set of all three-letter words over a four-state alphabet and its one-hot positional representation.

They include:

the 64 × 12 matrix dimensions;

three 1s per row;

sixteen 1s per column;

the exact Gram matrix;

rank ten;

right nullity two;

left nullity fifty-four;

the singular spectrum;

the exact pseudoinverse;

the two pseudoinverse entry values;

and the four Hamming-distance projector values.

None of these requires the standard genetic-code assignment.

The second category contains assignment-specific results.

These depend on how the standard code places translated outputs over the sixty-four codons.

They include:

the translated fiber sizes;

the hydropathy coefficient values;

the dominance of the second-position U–A contrast;

the reconstructed amino-acid hydropathy relationship;

the distribution of hydropathy and residue volume among second-position families;

the number and location of synonymous codon edges;

the forty-four-loop translated multigraph;

the nucleotide-pair synonymous weights;

the isotropic path scores;

the anisotropic score distribution;

and the unique common-path status of G–A–C–U.

These results could change under a reassigned or variant genetic code.

A third category contains model-dependent results. These depend not only on the standard code but also on an additional analytical convention.

Examples include:

the use of the Kyte–Doolittle hydropathy scale;

the assignment H (X) = 0;

the restriction to additive nucleotide-position effects;

the use of linear nucleotide paths;

equal weighting of all selected graph edges;

and equal weighting of amino acids in the reconstructed hydropathy regression.

These conventions are legitimate because they are declared and reproducible. They are not immutable properties of the code.

The distinction among input-space, assignment-specific, and model-dependent results is central to responsible interpretation.

12.6 The Same Completeness Produces Different Structures

Completeness plays a different role in the three representations.

In the matrix analysis, completeness produces exact balance. All sixty-four codons are required to obtain sixteen occurrences of every coordinate and four occurrences of every cross-position coordinate pair.

In the hydropathy analysis, completeness permits the numerical property field to be placed over the entire input domain. Because X is not an amino acid, completeness also forces an explicit convention rather than permitting the stop codons to disappear unnoticed.

In the topology, completeness ensures that the cube contains every possible coordinate combination. All 144 selected edges are available for translation-preservation analysis.

The same complete codon set therefore produces:

algebraic regularity at the input level;

a full numerical field at the property level;

and a closed finite graph at the topological level.

Completeness does not make the translated assignments regular in every sense. The fiber sizes are unequal. The synonymous edges are highly concentrated at the third position. The physicochemical coefficients are strongly unequal among positions.

The input space is perfectly balanced, while the output assignment over it is highly structured and anisotropic.

That contrast is one of the principal mathematical features of the genetic code.

12.7 Balance Below, Asymmetry Above

The codon matrix treats the three positions symmetrically at the level of input occurrence.

Every position has four nucleotide states.

Every nucleotide occurs sixteen times at every position.

Each positional block contributes three independent contrasts.

The three blocks are algebraically equivalent before translated outputs are introduced.

The standard code breaks that symmetry.

Hydropathy is not distributed equally across the three positions.

Synonymous adjacency is not distributed equally across the three positions.

The exact decompositions are strikingly different.

For hydropathy, the second-position coefficient range is largest and the third-position range smallest.

For synonymous topology, the third-position edge count is largest and the second-position count smallest.

The code therefore places distinct functional roles over a symmetric coordinate architecture.

The second coordinate acts as a strong physicochemical discriminator.

The third coordinate acts as a strong local redundancy channel.

The first coordinate provides intermediate refinement and a smaller amount of synonymous preservation.

This division of roles is not built into the balanced matrix. It emerges only after the standard translated assignments are applied.

12.8 Hydropathy and Topology as Complementary Projections

The hydropathy calculation and the synonymous-edge calculation can be understood as two different projections of the translated relation.

The hydropathy analysis replaces every translated class with a numerical value. It asks how that numerical field aligns with additive nucleotide-position coordinates.

The topology replaces numerical properties with an equality test. It asks whether neighboring codons retain the same translated label.

The first studies graded property similarity.

The second studies exact identity preservation.

Hydropathy may remain similar even when translated identity changes. A codon edge connecting leucine to valine is nonsynonymous, but both outputs are hydrophobic.

Conversely, every synonymous edge preserves hydropathy exactly because both codons receive the same amino-acid property value.

The two analyses therefore overlap without being equivalent.

Synonymous topology is stricter. It counts only exact equality of translated class.

Hydropathy projection is coarser. It can recognize physicochemical continuity across different amino-acid identities.

This distinction suggests further analyses. An edge could be weighted by the hydropathy difference between its translated endpoints rather than scored only as synonymous or nonsynonymous. Similar weighted graphs could use residue volume, charge, aromaticity, or combinations of properties.

The present unweighted synonymous objective supplies a clean exact baseline.

12.9 The Role of X Across the Three Views

Translation stop X illustrates why the three representations must not be conflated.

In the categorical lookup relation, X is an ordinary output class in the sense that three codons map to it.

In the binary input matrix, X is irrelevant because the matrix depends only on codon composition.

In the numerical hydropathy field, X requires an explicit convention because it has no amino-acid side chain. The complete calculation assigns it zero, while alternative analyses may omit it or handle it differently.

In the translated topology, X again behaves categorically. Stop-to-stop neighboring codons produce self-loops at the X vertex without requiring any numerical amino-acid property.

Thus, X is:

absent from the derivation of D;

retained as an output of τ;

numerically conventional in h;

and topologically genuine in the quotient multigraph.

Treating these roles separately preserves both completeness and biological accuracy.

12.10 Compression at Three Levels

Each mathematical view compresses the code differently.

The incidence matrix converts each codon into twelve binary coordinates with only three active entries. It expands the spelling into a structured vector while exposing exact linear dependencies.

The pseudoinverse compresses a sixty-four-component property field into twelve displayed coefficients with ten independent dimensions.

The U–A membrane profile compresses the full hydropathy representation further into one dominant signed nucleotide contrast.

The translated multigraph compresses sixty-four codon vertices into twenty-one translated vertices while retaining edge multiplicities.

The nucleotide-pair weight table compresses 144 cube edges into six pair weights for the isotropic problem or eighteen position-specific pair weights for the anisotropic problem.

The exhaustive path search compresses 13,824 cube configurations into three positional score spectra and one generating polynomial.

These reductions do not preserve every detail. Their value lies in preserving selected relations exactly or approximately while making the structure visible.

Different compressions retain different invariants.

The pseudoinverse preserves the nearest additive codon field.

The quotient multigraph preserves translated edge multiplicity.

The pair-weight representation preserves every synonymous path score.

The U–A profile preserves broad membrane-hydropathy peaks but not residue-specific detail.

A mathematical representation should therefore be judged by what it preserves, not merely by how small it becomes.

12.11 One Relation, Several Quotients

The genetic code can be viewed as supporting several quotient operations.

The first quotient identifies coefficient vectors that differ only by null-space gauge transformations. Many twelve-coordinate vectors represent the same additive codon field. The Moore–Penrose solution chooses one minimum-norm representative.

The second quotient identifies codons sharing the same translated output. Sixty-four codon vertices collapse to twenty-one translated vertices.

The third quotient identifies oriented nucleotide paths with their reversals. Twenty-four oriented paths become twelve undirected path topologies.

The fourth quotient identifies anisotropic axis triples that differ only by independent reversal of their three axes. The 13,824 oriented cubes become 1,728 undirected cubes.

Each quotient suppresses distinctions irrelevant to a particular question.

Gauge-equivalent coefficients are indistinguishable at the codon-field level.

Synonymous codons are indistinguishable at the translated-identity level.

Reversed paths are indistinguishable under undirected adjacency.

Axis reversals are indistinguishable under the unweighted edge-count objective.

The analysis becomes clearer when these equivalences are stated rather than left implicit.

12.12 Exactness at Different Levels

The word **exact** appears often in this book, but it applies at several levels.

The matrix identities are exact algebraic consequences of the complete factorial input space.

The hydropathy coefficients are exact relative to the selected property values, stop convention, and pseudoinverse formula. The property values themselves are empirical scale assignments rather than mathematical constants.

The synonymous-edge counts are exact relative to the standard code and selected nucleotide path.

The forty-four-edge maximum is exact over the complete declared family of linear path cubes.

The reconstructed membrane-profile resemblance is not an exact theorem. It is an empirical visual and potentially quantitative relationship that requires benchmark testing.

An exact calculation can depend on a chosen model.

A model-dependent result is not thereby weak. It becomes misleading only when the model boundaries are omitted.

The proper language is therefore:

exact input-space theorem;

exact calculation under a declared property convention;

exact finite maximum within a declared topology class;

or empirical correspondence in a stated sequence example.

This vocabulary permits strong claims where justified and restraint where required.

12.13 The Genetic Code Is More Than Its Matrix

The exact pseudoinverse is mathematically unusual, but the genetic code should not be reduced to that result.

Any complete alphabet–word matrix of the same form has an analogous closed structure. Appendix L generalizes the formula to arbitrary alphabet size and word length.

What is biologically distinctive is the output assignment placed over that input architecture.

The standard genetic code uses the balanced domain to organize:

twenty amino-acid identities;

translation termination;

broad physicochemical divisions;

extensive third-position synonymy;

and a maximally compact linear-path topology under the stated objective.

The matrix supplies the stage. The translated assignments supply the biological pattern.

The significance of the complete pseudoinverse lies in providing an exact instrument through which that pattern can be examined.

12.14 The Genetic Code Is More Than Its Topology

The G–A–C–U optimum is also not a complete description of the code.

It depends on one objective: maximizing the number of synonymous nearest-neighbor edges in a lattice formed from three linear nucleotide paths.

Other objectives could rank paths differently.

A mutation-weighted objective might favor transition substitutions.

A chemical-distance objective might use purine–pyrimidine distinctions or base-pairing properties.

A codon-usage-weighted objective could emphasize common codons.

A property-preservation objective could reward small hydropathy or volume changes even when amino-acid identity changes.

The forty-four-edge theorem remains exact within its comparison class. It should not be inflated into the statement that G–A–C–U is the unique correct nucleotide ordering for every possible question.

Its stronger significance is that the same path is independently compatible with all three position-specific maxima under one transparent and biologically meaningful objective.

12.15 The Genetic Code Is More Than Hydropathy

Hydropathy is one important translated property, especially in relation to membrane proteins. It is not the only possible numerical field.

The same exact projection machinery can be applied to:

residue volume;

charge;

polarity;

aromaticity;

side-chain mass;

flexibility;

helix propensity;

sheet propensity;

solvent accessibility;

or experimentally derived substitution effects.

Each property would produce a different coefficient vector and residual field.

Some may be concentrated at the second codon position.

Others may emphasize the first position or reveal interactions not visible in the additive model.

A property that is independent of codon assignment structure would project weakly and leave a large residual.

The exact pseudoinverse therefore provides a general analytical instrument rather than a hydropathy-specific trick.

Hydropathy is especially revealing because it produces a strong coefficient hierarchy and connects directly to the historical nucleotide-only membrane profile.

12.16 A Layered Mathematical Architecture

The combined results support a layered description of the standard genetic code.

The first layer is the complete input domain:

sixty-four ordered nucleotide triplets.

The second layer is the balanced coordinate representation:

a 64 × 12 binary matrix with rank ten.

The third layer is the translated lookup assignment:

twenty amino-acid classes and X distributed over codon space.

The fourth layer is the numerical property field:

amino-acid chemistry transferred back onto the codon inputs.

The fifth layer is local nucleotide geometry:

a declared adjacency relation among the four nucleotide states.

The sixth layer is translated topology:

synonymous loops and interclass connections after collapse by translated identity.

The seventh layer is optimization:

comparison of all permitted nucleotide paths and axis-order combinations.

No one layer contains the entire code.

The balanced matrix does not determine the translated assignments.

The translated assignments do not determine a unique geometry without an adjacency model.

The geometry does not determine which physicochemical property should be studied.

The property field does not determine historical origin.

The architecture becomes informative because relations can be traced across the layers without confusing them.

12.17 A Unified Statement

The three principal views can now be summarized in one statement.

The standard genetic code places a biologically specific many-to-one translation map over a perfectly balanced finite product space. That product space has an exact rank-ten incidence geometry and a closed-form Moore–Penrose pseudoinverse. When amino-acid hydropathy is transferred through the translation map, its dominant additive coordinate is the U–A contrast at the second codon position. When codon space is supplied with linear nucleotide-path geometry, the standard assignment permits at most forty-four synonymous nearest-neighbor edges, and G–A–C–U is the unique undirected path belonging to the maximizing families of all three positions.

Each clause depends on the preceding layer.

The balanced domain makes exact projection possible.

The translated assignment creates the hydropathy field and synonymous fibers.

The declared geometry turns those fibers into local topological objects.

The exhaustive comparison determines whether the selected geometry is exceptional.

The result is not one isolated pattern but a connected system of algebraic, physicochemical, and topological relations.

12.18 What the Synthesis Does Not Establish

The combined structure does not by itself establish how the genetic code originated.

It does not decide among chemical constraint, historical contingency, selection, coevolution, stereochemical affinity, error minimization, communal evolution, or other proposed mechanisms.

It does not prove that every detected pattern arose through one process.

It does not establish that the present standard code was historically inevitable.

It does not prove intention or design.

It does not show that every mathematically striking feature was directly selected.

It does establish that the present code contains exact and measurable internal organization that cannot be dismissed merely by referring to its existence as a lookup table.

Any historical account may still invoke accident, but accident must be defined relative to a comparison ensemble.

Any account may invoke freezing, but freezing explains persistence more directly than internal structure.

Any account may invoke optimization, but the optimized quantity and permitted transformations must be stated.

The mathematics narrows the interpretive problem by specifying what is present.

12.19 From Description to Explanation

A description identifies structure.

An explanation identifies why that structure arose and why it persisted.

This book has so far concentrated on description at unusually high resolution.

It has shown:

what the complete codon matrix is;

why its rank is ten;

why its pseudoinverse has only two values;

how hydropathy projects onto nucleotide coordinates;

why the middle nucleotide carries the dominant hydropathy contrast;

how codon space can be organized into a linear-path cube;

how synonymous edges become loops;

why the G–A–C–U score is forty-four;

and why no isotropic or anisotropic linear-path cube can exceed it.

These results now become constraints on explanation.

An adequate account of the genetic code should address not only why a code became stable but also why its translated assignments exhibit these particular algebraic, physicochemical, and topological relations when placed over the complete codon domain.

Chapter 13 examines the limits of the frozen-accident description and develops a more precise distinction between persistence, contingency, structure, and possible information-theoretic transition.

Chapter 13

Beyond Frozen Accident

The phrase **frozen accident** has often been used to describe the genetic code. In its strongest reasonable form, the phrase identifies a real feature of translation: once a codon-to-output relation becomes deeply embedded in a functioning biological system, changing that relation becomes globally disruptive.

A reassigned codon would not alter only one local sequence. It could reinterpret many occurrences distributed across many proteins. The molecular systems responsible for codon recognition, transfer-RNA charging, translation, termination, proofreading, and protein function would also have to remain mutually compatible.

The code can therefore become resistant to change.

That observation helps explain persistence. It does not, by itself, explain the internal organization of the relation that persists.

A random table can freeze.

A chemically constrained table can freeze.

A partly optimized table can freeze.

A historically layered table can freeze.

Freezing describes the difficulty of later reassignment. It does not determine what was frozen.

13.1 Persistence Versus Structure

The frozen-accident description addresses a temporal question:

Why might an established code remain stable?

The mathematical analyses in this book address a structural question:

What organization is present in the established code?

These are not interchangeable questions.

Suppose two complete codon tables become equally resistant to later change. One table distributes amino-acid properties almost randomly over codon space. The other displays strong second-position hydropathy polarity, extensive third-position synonymy, and unusually compact synonymous topology.

The fact that both tables are frozen does not explain why their internal structures differ.

The standard genetic code contains several kinds of measurable organization.

Its complete input set forms an exactly balanced factorial domain.

Its translated hydropathy field contains a strong additive nucleotide-position component.

The second codon position carries the dominant hydropathy contrast.

The third codon position carries thirty-nine of the forty-four synonymous edges in the G–A–C–U cube.

The G–A–C–U path is the unique undirected nucleotide order that belongs to the maximizing families of all three codon positions.

The first of these properties belongs to the complete codon input space. The others depend on the standard translated assignment.

Historical lock-in may preserve all of them. Preservation is not yet an explanation of their presence.

13.2 What an Adequate Account Must Explain

An account of the genetic code need not assign an independent cause to every measured feature. Several structures may follow from a smaller number of underlying constraints.

A code organized to reduce disruptive translation changes might also produce compact synonymous fibers.

A chemical partition among amino-acid classes might contribute simultaneously to second-position hydropathy and to the distribution of nonsynonymous edges.

A historically constrained expansion of assignments might generate several apparent regularities together.

The mathematics does not require one cause per theorem.

It does require that the object of explanation be stated accurately.

At minimum, an adequate account must confront four distinct observations.

First, the codon domain is complete and factorial:

$C = \Sigma^3$,

with

$|\Sigma| = 4$

and

$|C| = 64$.

Second, translated physicochemical properties are distributed nonuniformly over that domain. Hydropathy is especially concentrated at the second codon position, with the complete-code contrast

$\beta 2,U - \beta 2,A = 6.587500$.

Third, synonymous assignments have a strongly position-dependent topology. Under the G–A–C–U path, their positional edge counts are

4, 1, and 39.

Fourth, the same G–A–C–U path independently belongs to the maximizing family at each of the three codon positions.

An account that explains only why reassignment later became costly remains incomplete. It has explained the persistence mechanism without explaining the measured architecture of the persistent relation.

13.3 Accident Requires a Defined Ensemble

The word **accident** has no quantitative meaning unless the possible alternatives are defined.

To say that the code could have been different is trivially true at the level of abstract combinatorics. Many other maps can be written from sixty-four codons to twenty-one labeled outputs.

That fact alone does not show that the standard code is ordinary, improbable, optimized, or historically accidental.

Every such judgment depends on a comparison ensemble.

One possible ensemble contains all functions from sixty-four labeled codons to twenty-one labeled outputs. Its size is

21^{64}.

This is the broadest lookup-table space. It permits some outputs never to occur, permits arbitrary degeneracy profiles, and ignores codon boxes, stop placement, chemistry, and topology.

A narrower ensemble requires every output to occur at least once. The number of surjective maps is

$\Sigma j=0^{21} (-1)^{j} C (21, j) (21 - j)^{64}$.

Equivalently, it is

$21! S (64, 21)$,

where $S (64, 21)$ is a Stirling number of the second kind.

This ensemble preserves use of the complete output alphabet, but it still permits almost any degeneracy profile.

A still narrower ensemble preserves the exact number of codons assigned to every named output in the standard code.

The standard multiplicities consist of:

three outputs with six codons;

five outputs with four codons;

two outputs with three codons;

nine outputs with two codons;

and two outputs with one codon.

For the specifically labeled translated outputs, the number of assignments preserving those exact multiplicities is

$$\frac{64!}{(6!)^3\,(4!)^5\,(3!)^2\,(2!)^9}.$$

This number is vastly smaller than 21^{64}, but it remains enormous.

It preserves how many codons belong to each translated class. It does not preserve where those codons lie in codon space.

Two codes can have exactly the same degeneracy profile while having very different:

codon-box structure;

synonymous connectivity;

position-specific edge counts;

hydropathy organization;

and nucleotide-path optima.

Degeneracy does not determine topology.

13.4 A Hierarchy of Alternative Code Spaces

The possible comparison spaces form a hierarchy.

At the broadest level are all labeled lookup tables.

A narrower level contains only surjective tables.

A narrower level preserves the standard degeneracy multiset.

Another preserves the exact degeneracy of each named translated output.

Progressively narrower ensembles may also preserve:

the number and locations of stop codons;

the conventional codon boxes;

the membership of selected synonymous families;

amino-acid property classes;

known chemical groupings;

mutation neighborhoods;

or some portion of the standard local topology.

Every additional constraint reduces the comparison space.

Every additional constraint also changes the meaning of a statistical result.

A code may appear extremely unusual when compared with unrestricted random tables but much less unusual when compared with tables already constrained to preserve codon boxes and degeneracy.

Conversely, a property that remains exceptional under a highly constrained ensemble provides stronger evidence of organization not explained by those preserved features.

There is therefore no single meaningful quantity called “the probability of the genetic code.”

There are probabilities only relative to specified sampling rules over specified code spaces.

13.5 Combinatorial Counts Are Not Historical Probabilities

The existence of

21^{64}

abstract lookup tables does not imply that the probability of obtaining the standard code was

$1 / 21^{64}$.

That inference would require every table to have been equally accessible and equally likely under the historical process.

No such assumption is justified.

Physical and organizational constraints could make most mathematically possible tables inaccessible. Chemical affinity, compatibility among molecular recognition systems, translation accuracy, pathway structure, and preexisting assignments could make some regions of code space much more reachable than others.

The combinatorial counts measure possibility spaces.

They do not, by themselves, measure historical probabilities.

Their value is different. They show how much freedom remains after various coarse constraints are imposed.

Even preserving the exact standard degeneracy profile leaves an immense number of possible assignments. Therefore, degeneracy alone cannot specify:

the location of synonymous fibers;

their local connectivity;

the second-position hydropathy polarity;

or the G–A–C–U maximum.

Those are additional dimensions of organization.

13.6 Topology as an Additional Constraint

Consider two hypothetical genetic codes having the exact standard degeneracy profile.

In both, leucine receives six codons, arginine receives six, serine receives six, and every other translated class receives the same number of codons as in the standard code.

In one table, the codons belonging to each translated class are clustered into locally connected regions.

In the other, synonymous codons are scattered across codon space.

The degeneracy profiles are identical. Their synonymous topologies are not.

One code might contain many translation-preserving one-step movements.

The other might contain very few.

One might concentrate synonymous flexibility at the third codon position.

The other might distribute it evenly or concentrate it elsewhere.

The forty-four-edge G–A–C–U result therefore measures information not contained in the degeneracy sequence alone.

The positional decomposition

44 = 4 + 1 + 39

adds further structure. A randomization that preserves only the total number of synonymous edges might still fail to preserve their strong concentration at the third position.

The unique common-path result is stronger again. It depends not only on the total number of internal fiber edges but on the exact distribution of nucleotide-pair weights at the three codon positions.

A proper null model must therefore be matched to the claim being tested.

13.7 Physicochemical Organization as Another Constraint

Topology is not the only structure left free by degeneracy.

The amino-acid labels could be permuted among fibers while preserving every fiber size. Such a permutation would retain the synonymous topology if the fibers themselves remained fixed, but it could radically alter the distribution of hydropathy.

The second-position U–A contrast might weaken, disappear, or reverse.

The additive projection strength could change.

The membrane-profile correspondence could be lost.

Thus, topological organization and physicochemical organization are partly independent dimensions.

A code-space comparison can preserve one while randomizing the other.

This distinction permits stronger tests.

To ask whether the hydropathy organization follows automatically from synonymous topology, one can preserve the codon fibers and permute amino-acid property labels among them.

To ask whether the G–A–C–U topology follows automatically from the degeneracy profile, one can preserve each class size while reallocating codons.

To ask whether both arise from codon-box structure, one can preserve boxes while permuting translated outputs among them.

Different controls isolate different explanations.

13.8 The Limits of “Could Have Been Otherwise”

The statement that the code could have been otherwise is not yet an explanation.

Almost every finite structure can be replaced by alternatives in an unrestricted mathematical space.

A matrix can be replaced by another matrix.

A graph can be rewired.

A lookup table can be permuted.

A sequence can be shuffled.

The scientifically relevant question is not whether alternatives can be imagined. It is which alternatives were compatible with the constraints under consideration.

If the comparison ensemble is too broad, almost any natural object will appear extraordinary.

If the ensemble is too narrow, the property being tested may already be built into the null model.

A useful ensemble preserves enough known structure to remain relevant while leaving the target property free to vary.

For the genetic code, no one ensemble will answer every question. The hydropathy, degeneracy, topology, stop structure, and codon-box organization require different controls.

The language of accident should therefore be replaced, wherever possible, by explicit statements such as:

unusual under degeneracy-preserving reassignment;

typical under codon-box-preserving permutations;

maximal among linear nucleotide paths;

or robust across alternative hydropathy scales.

Such claims are narrower but scientifically meaningful.

13.9 An Information-Theoretic Phase Change

The emergence of a complete genetic code can be considered as a possible information-theoretic phase change.

Before a complete coding relation exists, triplet-to-output associations may be local, incomplete, ambiguous, or dependent on particular molecular circumstances.

After a sufficiently complete coding relation exists, the entire triplet domain is interpreted through a globally coordinated system.

The difference is organizational.

It is a transition from local molecular recognition to a distributed lookup relation.

A complete code does not merely contain associations. It establishes a stable correspondence between a finite input alphabet and a translated output alphabet across a whole system.

This proposal is conceptual. The mathematics of the present code does not demonstrate that such a transition occurred suddenly in historical time.

It identifies a possible discontinuity in organization.

A system may accumulate components gradually and nevertheless cross a threshold at which its behavior changes qualitatively.

Before the threshold, translated products may not maintain the machinery required for reliable interpretation.

After the threshold, the products of translation can help sustain, reproduce, and repair the machinery of translation.

Gradual molecular change and abrupt system-level appearance are therefore not contradictory descriptions.

13.10 Suddenness at the Organizational Level

The phrase **sudden appearance** need not mean that every component appeared simultaneously.

It can mean that the defining system-level relation was absent below a threshold and present above it.

A partial correspondence may assign outputs to some triplets but not others.

An ambiguous correspondence may produce several outputs from the same triplet.

An inaccurate correspondence may fail to reproduce the molecular system that enacts it.

A complete, sufficiently accurate, self-maintaining relation is categorically different from those partial states.

The transition may resemble other threshold phenomena. Incremental changes in components can produce a discontinuous change in collective organization.

Below the threshold, no durable code exists as a system-wide relation.

Above it, a code exists and participates in maintaining the conditions of its own continued enactment.

The present book does not establish where such a threshold lies or how it was crossed. It proposes that the origin question should include this level of organization rather than being reduced to a list of component molecules.

13.11 A Code That Participates in Its Own Persistence

The genetic code is not stored in one isolated molecule.

Its physical implementation is distributed among:

transfer RNAs;

aminoacyl-transfer-RNA synthetases;

ribosomal RNA and proteins;

messenger RNA;

release factors;

editing systems;

translation factors;

and the broader cellular environment.

No single component is the complete code.

The mapping exists in the coordinated operation of the system.

Many proteins required to maintain that operation are themselves produced through translation. The code therefore participates in producing components that continue to enact the code.

This is a form of dynamic reflexivity.

It does not mean that the code is metaphysically self-causing.

It means that the continued existence of the lookup relation depends partly on products generated under that same relation.

This reflexive maintenance helps distinguish a living coding system from a passive printed table. The table represents the relation. The biological system repeatedly performs it.

13.12 Why Global Structure Can Be Understated

The genetic code is often approached through the molecular mechanisms that implement it.

Researchers examine codon recognition, transfer-RNA charging, ribosomal fidelity, release factors, translation errors, reassignment, and variant codes.

Those questions are indispensable.

Yet the global lookup relation can disappear behind the mechanisms that enact individual steps.

A full-table analysis asks a complementary question:

What structures are present in the complete relation, regardless of whether their historical causes are already known?

This approach does not compete with molecular explanation. It defines the finite structure that molecular explanation must eventually address.

The exact pseudoinverse is one example.

A numerical software package can compute a pseudoinverse without revealing that the complete natural input matrix has a closed form containing only two rational values.

The result can therefore remain hidden in plain sight—not because the calculation is impossible, but because the complete relation was not treated as an object worthy of exact global analysis.

13.13 Scientific Caution and Conceptual Fear

Claims of deep organization in the genetic code can become entangled with disputes over design, Intelligent Design, theology, and material explanation.

That environment can produce two opposite errors.

The first is inflation. A mathematical pattern may be treated as though it directly proves intention, purpose, or a theological doctrine.

The second is avoidance. A mathematical pattern may be ignored or minimized because someone might use it to support such an interpretation.

Neither response is scientifically adequate.

The appropriate response is measurement.

If the matrix has rank ten, the rank should be derived.

If the pseudoinverse has a closed form, the formula should be proved.

If G–A–C–U maximizes forty-four synonymous edges, the comparison space should be enumerated.

If a U/T–A nucleotide profile reproduces broad membrane-hydropathy peaks, the comparison should be displayed and quantitatively tested.

The possibility of philosophical misuse does not invalidate an exact calculation.

Likewise, an exact calculation does not settle the philosophy.

Interpretation should follow the evidence rather than determining in advance which structures may be acknowledged.

13.14 Four Levels of Interpretation

The conclusions of this book can be separated into four levels.

The first level is exact mathematics.

Examples include:

rank (D) = 10;

$D^{+} = (1 / 16) D^{T} - (1 / 96) J$;

the pseudoinverse entry values 5 / 96 and −1 / 96;

the forty-four-edge maximum;

and the unique common-path intersection.

These statements can be proved from declared definitions.

The second level is biological explanation.

This level asks which physical, chemical, historical, or organizational processes produced and maintained the measured structures.

The present calculations constrain such explanations but do not select one conclusively.

The third level is philosophical interpretation.

One may ask whether the code is best understood as contingent, constrained, optimized, emergent, law-like, or some combination. These categories extend beyond the exact calculations and require additional argument.

The fourth level is theological interpretation.

A reader may place the code within a theological account of nature, meaning, or Logos. Such an interpretation is not mathematically entailed by the pseudoinverse, hydropathy coefficients, or topology theorem.

Keeping the levels distinct protects both science and philosophy.

Exact results need not be weakened merely because broader interpretations are possible.

Broader interpretations should not be presented as though they were outputs of the matrix calculation.

13.15 Beyond the False Choice

The code need not be classified as either meaningless accident or perfect design.

That binary opposition is too crude.

A more productive description is that the standard genetic code is a historically persistent, physically enacted finite relation with measurable algebraic, physicochemical, and topological structure.

Its present organization may reflect several kinds of constraint simultaneously.

Some features may be consequences of complete product-space architecture.

Some may follow from amino-acid chemistry.

Some may reflect the need to preserve translation under local changes.

Some may be inherited consequences of earlier assignments.

Some may be correlated by-products rather than independently selected traits.

Some may arise only when several constraints are imposed together.

The mathematics does not require that every feature have been optimized.

The code can contain exact structure without being globally perfect under every conceivable objective.

Indeed, the present analyses show different kinds of organization distributed unevenly among the three codon positions. The code resembles a layered system with several overlapping roles rather than a table optimized for one scalar quantity.

13.16 What Could Weaken the Interpretation

Several future results could weaken broad claims about the significance of the measured structure.

The hydropathy coefficient pattern might prove highly sensitive to the selected property scale or treatment of X.

The U/T–A membrane profile might perform poorly across a large benchmark set.

Degeneracy-preserving random codes might frequently produce second-position hydropathy contrasts as strong as the standard code.

Codon-box-preserving controls might commonly reproduce the G–A–C–U topology.

Natural variant codes might show no coherent relationship to the standard results.

Such findings would reduce the evidence that the assignment-specific structures are exceptional or historically informative.

They would not alter the exact input-space theorems.

They would also not alter the result that G–A–C–U is maximal among the linear nucleotide paths for the standard code. They would change the interpretation of why that fact matters.

A strong manuscript must distinguish results that can be challenged empirically from results fixed by finite derivation.

13.17 What "Frozen Accident" Still Contributes

Rejecting frozen accident as a complete account does not require discarding its valid insight.

Once a code is widely instantiated, reassignment becomes costly.

The mapping is distributed through many genes and many components of translation.

The products generated under the code help maintain the system that reads the code.

This creates historical inertia.

The phrase remains useful when it means:

an established coding relation becomes difficult to change because its consequences are global.

It becomes insufficient when it is used to imply:

the internal structure of the relation requires no further explanation.

The first statement concerns persistence.

The second prematurely closes investigation.

13.18 The Strongest Present Conclusion

The strongest defensible conclusion is not that the code could not have appeared through natural processes.

It is not that every feature was separately optimized.

It is not that the exact pseudoinverse proves intention.

It is not that the G–A–C–U topology establishes a metaphysical doctrine.

The conclusion is narrower:

The standard genetic code contains exact and reproducible organization that is not explained merely by saying that the code became frozen.

The phrase **frozen accident** may help describe why later reassignment is difficult.

It does not complete the scientific account of what was frozen.

Chapter 14 considers a related consequence of this persistence. The genetic code is not merely one ancient message among others. It is the relation through which an immense succession of biological messages has been interpreted, making it a candidate for an exceptionally persistent natural signal.

Chapter 14

The Genetic Code as a Persistent Signal

The genetic code is usually described as a table of correspondences between codons and translated outputs. In living systems, however, the code is not merely consulted. It is continuously enacted.

Every translated gene passes through the same underlying relation. Individual coding sequences appear, change, duplicate, recombine, and disappear, while the relation used to interpret them remains comparatively stable. The messages vary. The code persists.

This distinction makes the genetic code a candidate for an exceptionally persistent natural signal. What persists is not one nucleotide sequence, one molecule, or one organism. It is a distributed mapping between two representational levels: nucleotide triplets and translated outputs.

The claim requires restraint. The present analysis cannot establish that the genetic code is literally the oldest signal in nature or the only natural lookup code. Those stronger claims would require a defined comparison class and historical evidence extending far beyond the mathematical results developed here.

The defensible conclusion is that the genetic code combines several unusual properties. It is finite, broadly shared, physically distributed, globally consequential, reflexively maintained, and repeatedly instantiated whenever a coding sequence is translated.

14.1 Code Versus Message

A biological coding sequence carries at least two distinguishable kinds of information.

The first is the message specified by the particular sequence. Different genes contain different codons in different orders and therefore produce different translated products.

The second is the code under which those codons are interpreted. This relation is shared across the many genes translated by the same molecular system.

The distinction can be represented schematically as

message + code → interpreted product.

The message determines which codons occur and in what order.

The code determines what those codons mean within translation.

A coding sequence may change extensively while the underlying translation table remains stable. Mutations, recombination, duplication, and selection can alter the message without altering the code used to interpret it.

Conversely, a change in the code can alter the interpretation of many otherwise unchanged sequences.

The asymmetry is fundamental.

A change in one gene usually changes one local message.

A change in the coding relation can change the reading of many messages distributed throughout the genome.

The message is comparatively local and variable.

The code is comparatively global and persistent.

The same distinction appears in written language. A sentence can change while the language in which it is interpreted remains stable. Altering the dictionary or grammar can affect the interpretation of many sentences at once.

The analogy is limited because cellular translation is a molecular process rather than a human symbolic convention. It nevertheless clarifies the difference between a sequence and the relation through which the sequence acquires translated significance.

14.2 Distributed Physical Implementation

The genetic code is not stored as a printed table inside the cell.

No single molecule contains the complete relation in the same explicit form as the conventional genetic-code chart.

Its implementation is distributed among interacting molecular components, including:

transfer RNAs;

aminoacyl-transfer-RNA synthetases;

ribosomal RNAs and proteins;

messenger RNAs;

initiation, elongation, and termination factors;

release factors;

editing and proofreading systems;

and the metabolic pathways that supply activated amino acids.

No one component alone is the code.

A transfer RNA participates in recognizing a subset of codons but does not contain the complete assignment relation.

An aminoacyl-transfer-RNA synthetase helps connect amino-acid identity with transfer-RNA identity but does not enact the full table alone.

The ribosome coordinates decoding and peptide synthesis but depends on correctly charged transfer RNAs and associated factors.

The code exists operationally in the coordinated behavior of the complete translation system.

This distributed organization has two consequences.

First, the code can persist even though its individual physical components are continually replaced. Molecules are synthesized and degraded, cells divide, and organisms die, yet the relation is reconstituted in successor systems.

Second, reassignment is difficult because many components must remain mutually compatible. A change in one recognition process may require compensating changes elsewhere if translation is to remain functional.

The persistence of the code is therefore relational rather than material.

What survives is not a particular transfer RNA, ribosome, chromosome, or cell. What survives is the coordinated mapping repeatedly enacted by changing physical components.

14.3 Global Consequences of Reassignment

Consider a codon that occurs throughout a genome.

A point mutation changing one occurrence affects one local sequence position. Its consequences may be large or small, but the altered interpretation is confined initially to that occurrence.

A codon reassignment is different.

If the translated meaning of the codon changes, every occurrence of that codon becomes a candidate for reinterpretation.

The consequences may propagate across:

many genes;

many protein families;

structural proteins;

enzymes;

regulatory factors;

transport systems;

and components of translation itself.

The distinction can be stated as follows:

sequence mutation changes a message under a stable dictionary;

code reassignment changes part of the dictionary used to read many messages.

This global reach helps explain why an established code can be strongly conserved.

The difficulty is not simply that one molecular interaction must change. A reassignment must remain compatible with the accumulated set of coding sequences and with the machinery that interprets them.

This does not make reassignment impossible. Natural variant codes show that changes can occur. It does mean that code change is constrained in a way ordinary sequence change is not.

The existing coding genome acts as a large inherited commitment to the established translation relation.

14.4 Reflexive Maintenance

The genetic code participates in producing components required for its own continued implementation.

Proteins generated through translation include enzymes and structural components involved in:

transfer-RNA charging;

ribosome construction and function;

translation-factor activity;

proofreading;

RNA processing;

amino-acid metabolism;

and maintenance of the cellular systems supporting translation.

The products of the code therefore help sustain the machinery that continues to enact the code.

This relationship can be written schematically as

code-guided translation
→ translation-system components
→ continued code-guided translation.

The relation is reflexive, but not self-originating.

The code does not translate itself in isolation. It operates within a larger cellular system containing nucleic acids, proteins, metabolites, membranes, energy flows, and inherited structures.

Reflexive maintenance means that once the translation relation exists within a functioning system, some of its products contribute to preserving the conditions of its continued operation.

The code therefore differs from a passive human-readable table.

A printed table remains inert unless an external reader consults it.

A cellular coding system is an operational regime. It is continuously realized through recognition, binding, catalysis, proofreading, elongation, and termination.

The static mathematical map defines the relation.

The dynamic biological system keeps that relation in force.

14.5 Persistence Across Biological Time

Individual biological messages are transient.

A gene may persist for a long period, but it can mutate, duplicate, diverge, become inactive, or disappear. Even strongly conserved genes have particular sequence histories.

The coding relation is more general than any one gene.

It has been used to interpret an immense succession of sequences across lineages and generations. New genes arise and old genes vanish while remaining embedded within largely the same translation system.

This gives the code an unusual temporal status.

It is not merely an ancient sequence.

It is an ancient rule of interpretation repeatedly applied to newer and changing sequences.

Every contemporary act of translation therefore links a present coding message to a deeply inherited relational structure.

In that sense, the code functions as a form of biological memory.

The memory is not stored as a sentence describing the past. It is embodied as a stable mapping that constrains present interpretation.

Ancient assignment relations continue to influence which proteins can be encoded, how codon changes are interpreted, and which evolutionary paths remain accessible.

The code is not only inherited.

It shapes the interpretation of what is inherited.

14.6 Structure Visible in Coding Sequences

Although the code is not written as a separate internal table, its structure remains visible statistically within actual coding sequences.

A sequence occupies codon space. Its codons therefore inherit the algebraic, physicochemical, and topological organization of the standard lookup relation.

The membrane-hydropathy example provides a concrete demonstration.

A membrane-protein coding sequence enriched in second-position U or T tends to encode hydrophobic amino acids. A sequence region enriched in second-position A tends to encode more hydrophilic outputs.

The historical nucleotide-only profile uses the contrast

U − A

in RNA notation, or

T − A

in DNA notation.

It does not identify all twenty amino acids. It does not reproduce every side-chain property. Yet its broad peaks correspond to the principal hydrophobic regions of the conventional translated Kyte–Doolittle profile in the illustrated membrane protein.

The code-level relation therefore leaves a detectable signal within the message.

The coding sequence carries at least two overlapping patterns:

the exact amino-acid sequence specified by the complete codons;

and a compressed hydropathy signal visible primarily through the middle nucleotide.

The synonymous topology supplies another example.

The standard code concentrates much of its local translated-identity preservation at the third codon position. Consequently, coding sequences can vary extensively at third positions without proportional changes in their translated products.

The sequence is therefore shaped by both message-specific requirements and code-level coordinate structure.

A coding region does not merely list amino acids indirectly. It lies within a finite representational system whose positional asymmetries influence the statistical form of the nucleotide sequence.

14.7 The Code as a Constraint on Future Biological Change

Persistence is not only preservation of the past. It also constrains the future.

Once the code is established, new coding sequences must function within its existing interpretation system.

A newly arising gene does not normally choose an independent dictionary. It must use codons according to the translation relation already implemented by the organism.

Mutations are also evaluated through that relation.

A nucleotide change may be synonymous, conservative, disruptive, or terminating partly because of where the affected codons lie within the established code.

The code therefore structures the space of possible sequence changes.

It determines:

which single-nucleotide substitutions preserve translated identity;

which substitutions change hydropathy strongly;

which codon positions are more likely to tolerate variation;

which codons specify termination;

and which alternative sequences encode the same protein.

In this sense, the code is both inherited history and continuing constraint.

Past assignment structure shapes future evolutionary accessibility.

This does not mean that the code determines biological evolution in detail. Mutation, selection, drift, recombination, regulation, environment, population structure, and many other processes remain essential.

The narrower claim is that these processes act on coding sequences whose possible interpretations are already organized by the translation relation.

The code defines part of the landscape over which future sequence change occurs.

14.8 A Candidate for an Exceptionally Persistent Natural Signal

The genetic code is exceptional not simply because it is ancient.

It combines several properties that are rarely found together.

It is finite. The complete standard relation can be listed as sixty-four codon assignments.

It is broadly shared. The standard code is used across a vast range of living systems, although natural variants exist.

It is distributed. No one molecule alone contains or implements the complete relation.

It is globally consequential. Reassignment can affect many coding messages simultaneously.

It is reflexively maintained. Products generated through the code help preserve the machinery that continues to enact it.

It is repeatedly instantiated. Every translated coding sequence provides another physical realization of the same underlying relation.

It is structurally visible. Its positional and physicochemical organization leaves measurable traces in codon matrices, property fields, coding sequences, and synonymous graphs.

These properties justify describing the genetic code as a candidate for an exceptionally persistent natural signal.

The word **signal** here refers not to one transmitted nucleotide sequence but to a durable relation that constrains interpretation across countless sequences.

The phrase should remain qualified.

To establish that the genetic code is the oldest surviving signal, the most persistent signal, or the only natural lookup code would require a formal comparison with other natural informational systems.

Possible comparisons might include:

developmental patterning systems;

immune-recognition relations;

regulatory binding systems;

epigenetic inheritance;

cellular signaling conventions;

and other molecular recognition networks.

Many such systems are more context-dependent and less easily expressed as complete finite lookup tables. That difference makes the genetic code especially clear, but it does not prove categorical uniqueness.

The strongest present claim is therefore:

The genetic code is one of the clearest known examples of a persistent, distributed, physically enacted relational signal in living systems.

14.9 Stability and Meaning

The persistence of the genetic code is not merely persistence of molecular form.

A codon has no translated meaning in isolation from the system that interprets it.

The triplet AUG is a nucleotide sequence whether or not translation occurs. Its role as methionine within the standard code depends on the operational translation relation.

What persists is therefore a mapping between levels:

codon identity → translated identity.

This relational persistence is deeper than the survival of any one coding sequence.

The nucleotide molecules present in an organism are temporary.

The proteins are temporary.

The ribosomes are temporary.

The cells and organisms are temporary.

Yet successor systems repeatedly rebuild components that enact substantially the same mapping.

The relation persists through turnover of its physical instances.

Meaning, in this restricted biological sense, is maintained through reliable interpretation.

A codon means what the translation system consistently causes it to specify.

The stability of that meaning depends on the stability of the distributed operational regime.

14.10 Persistence Does Not Mean Immutability

Natural variant genetic codes demonstrate that the standard relation is not absolutely immutable.

Some mitochondrial systems, microbial lineages, and other translation systems contain codon reassignments. Stop codons may be captured for amino-acid specification, and codons ordinarily specifying amino acids may acquire different roles.

These variants are important because they prevent an exaggerated claim.

The genetic code is not physically incapable of change.

Its persistence is strong but not absolute.

Variant codes also provide natural experiments. They permit the analyses developed in this book to be repeated under altered translated assignments while preserving the same sixty-four-codon input domain.

The incidence matrix D remains unchanged.

Its rank and exact pseudoinverse remain unchanged.

The output-dependent structures may change:

hydropathy coefficients;

synonymous fibers;

loop multiplicities;

nucleotide-pair weights;

and maximizing paths.

Natural variants therefore separate universal input-space mathematics from assignment-specific biological organization.

The existence of code changes does not weaken the claim that code reassignment is globally constrained. It shows that such constraints can sometimes be crossed through historically viable intermediate states.

14.11 Persistent Signal, Open Origin

Recognizing the code as a persistent signal does not settle how it originated.

Persistence can preserve contingent features.

It can preserve optimized features.

It can preserve chemical constraints, historical compromises, and correlated by-products.

Reflexive maintenance explains how an established relation can become self-reinforcing without explaining how the relation first appeared.

The membrane profile shows that code-level organization remains visible in present sequences. It does not reconstruct the historical path by which the U–A hydropathy polarity arose.

The forty-four-edge theorem shows exact synonymous topology under a declared geometry. It does not determine which historical processes generated that topology.

The exact pseudoinverse shows that complete codon input space has an unusually transparent algebraic structure. It does not show that early biology explicitly selected that formula.

The origin problem therefore remains open, but it can be stated more precisely.

The task is not merely to explain how individual triplets became associated with individual outputs.

It is to explain how a sufficiently complete, accurate, distributed, and persistent lookup relation emerged; how its assignments acquired their physicochemical and topological organization; and how the resulting system became capable of maintaining its own implementation.

The persistence of the code ensures that this ancient relation is not confined to the past.

It remains active in every contemporary act of translation.

Chapter 15

Tests, Controls, and Falsifiable Extensions

Several results in this book are exact within their stated definitions.

The complete 64 × 12 codon incidence matrix has rank ten.

Its Moore–Penrose pseudoinverse has the closed form

$D^{+} = (1 / 16) D^{T} - (1 / 96) J.$

The isotropic G–A–C–U cube contains forty-four synonymous nearest-neighbor edges.

No anisotropic combination of three linear nucleotide paths exceeds that total.

These results are not statistical estimates. They follow from finite definitions and exact calculations.

The broader biological interpretation is different. Claims about hydropathy, membrane-protein prediction, the exceptional status of the standard code, or the historical meaning of its organization depend on empirical scales, numerical conventions, comparison ensembles, and definitions of nucleotide locality. Those claims should be exposed to explicit controls.

The purpose of this chapter is therefore not to reopen exact results that have already been proved. It is to distinguish four categories:

exact finite theorems;

reproducible numerical results;

robustness tests;

and historical or biological interpretation.

A mature account of the genetic code should show clearly which conclusions belong to each category.

15.1 Reanalysis of the Historical Hydropathy Calculation

The historical hydropathy calculation should be reconstructed directly from the original Mathematica notebook or its closest surviving computational record.

The principal historical result is clear. Singular-value decomposition of the sixty-one sense codons identified the second codon position as the dominant carrier of amino-acid hydropathy, with a strong positive contribution from U or T and a strong negative contribution from A.

The exact numerical workflow nevertheless contains details that must be preserved if the historical figure and coefficients are to be reproduced rather than approximated.

The audit should record:

the precise amino-acid order;

the exact hydropathy scale and numerical values;

the codon row order;

the nucleotide-coordinate order;

whether the three stop codons were omitted before matrix construction or removed afterward;

whether amino-acid values were rounded before regression;

whether the fitted regression included an intercept;

whether all twenty amino acids were weighted equally;

whether synonymous codon reconstructions were averaged before fitting;

the numerical singular-value threshold;

the convention used to select a solution in the rank-deficient coefficient space;

and the exact source of the reconstructed values displayed in the historical scatter plot.

These details matter because several related but nonidentical calculations can all be described loosely as “the hydropathy reconstruction.”

One calculation fits sixty-one codon observations.

Another uses the complete sixty-four-codon matrix and assigns X a declared numerical value.

One amino-acid comparison weights every codon equally before averaging.

Another gives every amino acid one equal observation.

One regression may use full-precision values, while another may use displayed rounded values.

Small differences in these procedures can produce small differences in slopes, intercepts, correlations, and individual coefficients without changing the principal biological conclusion.

The present complete-code analysis uses one internally consistent dataset. The twenty empirical and reconstructed amino-acid values give

$$y = -0.003660 + 0.859718\,x,$$

with

$$r = 0.915646$$

and

$$R^2 = 0.838407.$$

Any future revision of those numbers should be made only by replacing the underlying dataset and regenerating the figure and text together. Values from incompatible computational runs should not be combined.

The historical and complete calculations should remain distinguishable.

The historical result belongs to the sixty-one-sense-codon SVD analysis.

The present exact formulation belongs to the complete sixty-four-codon matrix with the declared convention

$$H(X) = 0.$$

Their agreement on the dominant second-position polarity is more important than forcing every coefficient to be numerically identical.

15.2 Alternative Hydropathy Scales

Hydropathy is not defined by one uniquely privileged numerical scale.

The Kyte–Doolittle scale is historically important and especially appropriate for the membrane-profile comparison developed in Chapter 7. Other scales may emphasize transfer energies, solvent exposure, chromatographic behavior, membrane partitioning, or different experimental environments.

The nucleotide-position projection should therefore be repeated across several established scales.

For each scale, the analysis should report:

the complete set of twelve nucleotide-position coefficients;

the within-position coefficient ranges;

the second-position U–A contrast;

the codon-level additive projection fraction;

the norm of the fitted field;

the norm of the residual field;

the amino-acid-level empirical-versus-reconstructed correlation;

and the identities of the amino acids having the largest residuals.

The most important robustness question is not whether every coefficient remains numerically unchanged. Different scales are expected to produce different magnitudes.

The stronger qualitative question is whether the positional hierarchy survives.

Does the second position remain dominant?

Does U remain strongly associated with hydrophobic values?

Does A remain strongly associated with hydrophilic values?

Do third-position effects remain comparatively small?

A result that survives several substantially different hydropathy scales would support the conclusion that the observed organization belongs broadly to amino-acid environmental preference rather than to one particular numerical table.

A result that disappears under most alternative scales would narrow the claim. It would indicate that the strong projection is more specific to the physical interpretation represented by Kyte–Doolittle hydropathy.

Both outcomes would be scientifically useful.

15.3 Alternative Numerical Conventions for X

Translation stop X is a categorical output of the complete genetic code but is not an amino-acid residue. It has no intrinsic side-chain hydropathy.

The complete calculation in this book uses

$H(X) = 0$.

That choice preserves all sixty-four codon rows and the exact balanced pseudoinverse. It should be understood as a declared numerical convention, not as a physical property of translation termination.

At least four treatments should be compared.

The first is the present zero convention:

$H(X) = 0$.

The second assigns X the mean hydropathy of the twenty amino acids:

$H(X) = \mu_{AA}$.

The third omits UAA, UAG, and UGA and performs a sixty-one-sense-codon analysis.

The fourth retains X in all categorical and topological analyses but treats its hydropathy as missing during numerical property fitting.

For each convention, the following should be recomputed:

the twelve nucleotide-position coefficients;

the second-position U–A contrast;

the positional coefficient ranges;

the codon-level projection strength;

the reconstructed amino-acid values;

and the regression statistics.

The central question is whether the dominant middle-position polarity survives all reasonable stop treatments.

Individual coefficients can vary because the three stop codons are not distributed uniformly over the twelve coordinate columns. UGA contains second-position G, while UAA and UAG contain second-position A. Their numerical treatment can therefore affect those coordinate means differently.

That sensitivity does not invalidate the complete calculation. It identifies which coefficient values depend most strongly on the handling of X.

The categorical topology does not require such a convention. X remains a genuine translated vertex, and synonymous stop-to-stop edges remain genuine loops regardless of the numerical hydropathy assigned during a separate property analysis.

15.4 Residue-Volume Verification

The hydropathy–residue-volume figure requires a complete source and data audit.

The final record should identify:

the exact residue-volume scale;

the units;

the original publication or compilation;

whether the values represent side-chain volume, residue volume, partial molar volume, or another molecular quantity;

whether any values were transformed or rescaled;

whether X was omitted or assigned a placeholder;

and whether the enclosing second-position family regions were computed algorithmically or drawn descriptively.

The historical figure has been retained because it conveys a useful two-property organization. Its scientific interpretation depends on knowing exactly what the vertical coordinate represents.

The numerical table used to generate the figure should be preserved with the digital resources. A generation script should reproduce the plotted amino-acid positions from that table.

If the family outlines were drawn to guide the eye, the caption should describe the figure as a descriptive grouping.

If the regions were produced by a defined algorithm, the algorithm and its parameters should be stated.

The principal qualitative claims should also be tested numerically.

For example, one can compare the volume distributions of the four second-position families and ask whether NCN occupies a significantly narrower or lower-volume region than the other groups under the selected scale.

One can also quantify the extent to which the second-position nucleotide predicts hydropathy, volume, or their joint two-dimensional position.

The purpose is not to force a sharp four-cluster classification. The genetic code does not produce four perfectly separated chemical classes. The test is whether the broad organization described in Chapter 6 remains visible under a transparent and reproducible numerical definition.

15.5 Natural Variant Genetic Codes

Natural variant codes provide direct perturbations of the standard translated assignment.

As long as all sixty-four codons remain in the input domain, the incidence matrix D is unchanged. Its rank, singular spectrum, pseudoinverse, and Hamming-distance projector remain exactly the same.

The output-dependent structures may change.

For each natural variant code, the analysis should recompute:

the translated fibers;

the degeneracy profile;

the class-specific loop multiplicities;

the position-specific nucleotide-pair weights;

the isotropic scores of all twenty-four oriented paths;

the best isotropic path;

the anisotropic maximum;

the position-specific maxima;

the complete maximum family;

the hydropathy coefficients;

and the additive projection strength.

These comparisons would determine whether the G–A–C–U optimum is specific to the standard code or persists across nearby natural variants.

Several outcomes are possible.

A variant may preserve the forty-four-edge maximum and the same unique common path.

It may retain G–A–C–U as optimal while changing the total score.

It may preserve the hydropathy hierarchy while altering synonymous topology.

It may preserve topology but weaken the second-position hydropathy contrast.

Each possibility would help separate robust structural features from details specific to the canonical assignment.

A particularly transparent extension is single-reassignment sensitivity.

For every codon c and every alternative translated output ω, define a perturbed code in which only the assignment of c is changed. Then recompute:

the G–A–C–U loop score;

the isotropic maximum;

the maximizing-path set;

the anisotropic maximum;

the positional maxima;

the common-path intersection;

and the hydropathy coefficients.

This produces a local sensitivity map over code space.

Some codon assignments may be topologically fragile: one reassignment may reduce the forty-four-edge score or destroy the unique common-path property.

Other assignments may be neutral with respect to the selected objective.

The analysis would identify which portions of the standard lookup relation contribute most strongly to the measured architecture.

15.6 Degeneracy-Preserving Randomization

A basic null ensemble should preserve the exact codon count assigned to every named translated output.

Under this model, the standard fibers are redistributed among the sixty-four codons while maintaining:

six codons for each sixfold-degenerate class;

four for each fourfold class;

three for each threefold class;

two for each twofold class;

and one for methionine and tryptophan.

The translated labels retain their own standard degeneracies, but the locations of their codons are randomized.

For every randomized code, calculate:

the G–A–C–U synonymous-edge score;

the best isotropic score;

the best anisotropic score;

whether G–A–C–U is optimal;

whether one common positional maximum path exists;

the complete loop-multiplicity distribution;

the positional synonymous-edge decomposition;

the second-position hydropathy contrast;

and the additive projection strength of hydropathy.

The standard code can then be located within the resulting distributions.

This ensemble asks whether the measured topology and physicochemical organization follow merely from the degeneracy counts.

If random codes having the same named degeneracies commonly produce scores and coefficient structures comparable to the standard code, the broader interpretation would weaken.

If the standard code lies in an extreme tail, degeneracy alone would be insufficient to explain the observed organization.

A related but broader ensemble preserves only the multiset of fiber sizes:

{ 6, 6, 6, 4, 4, 4, 4, 4, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1 }.

In this ensemble, translated identities may exchange degeneracy sizes. For example, a class that has two codons in the standard code could receive six in a randomized code.

This comparison separates the effect of the overall degeneracy profile from the specific assignment of degeneracy sizes to named amino acids.

The two null models answer different questions and should not be merged.

15.7 Codon-Box-Preserving Randomization

Unrestricted degeneracy-preserving randomization may destroy so much known code structure that the resulting comparison becomes biologically remote.

A stricter ensemble can preserve the conventional codon-box organization.

The first two codon positions define sixteen boxes, each containing four third-position variants. Some boxes translate entirely to one output. Others are divided into two or more translated groups.

A box-preserving randomization may retain:

the sixteen-box partition;

the sizes of translated sub-boxes;

the positions of split and unsplit boxes;

or the complete pattern of third-position degeneracy,

while permuting translated labels among compatible boxes or sub-boxes.

Such ensembles retain more of the broad architecture of the standard code while leaving physicochemical and topological relations free to vary.

They ask whether the G–A–C–U optimum and second-position hydropathy polarity remain unusual after gross block structure has already been fixed.

A hierarchy of box-preserving models should be used rather than one arbitrary shuffle.

One model might permute only complete four-codon boxes.

Another might preserve all two-versus-two box splits.

Another might preserve the six-codon family architecture.

Another might retain the three stop codons but move them among compatible sub-boxes.

The results will depend on what is held fixed. That dependence is not a defect. It reveals which known structural constraints already account for which observations.

15.8 Amino-Acid Property Permutations

Topology and physicochemical organization can be separated experimentally by permuting output labels or property values while keeping other structures fixed.

One test begins with the standard codon partition and permutes amino-acid labels among fibers having equal degeneracy.

This preserves:

the codon fibers;

the fiber sizes;

all synonymous edges;

the G–A–C–U score;

and the complete loop topology.

It changes:

which amino-acid identities occupy the fibers;

which hydropathy values are placed over codon space;

the twelve property coefficients;

and the nucleotide-only membrane signal.

This ensemble tests whether amino-acid chemistry is unusually aligned with the existing synonymous topology.

A second test holds the named genetic-code assignments fixed but permutes hydropathy values among the twenty amino-acid classes.

The translated fibers and all graph properties remain unchanged.

Only the property field changes.

This provides a direct null distribution for:

the second-position U–A contrast;

the additive projection fraction;

the amino-acid reconstruction correlation;

and the positional coefficient hierarchy.

A third test can preserve broad property groups while permuting values within them. Charged residues might be permuted among charged residues, hydrophobic residues among hydrophobic residues, or biosynthetic families within their groups.

Such constrained permutations ask whether the detailed numerical alignment is unusual after coarse chemistry has already been preserved.

Property permutations are particularly useful because they isolate the assignment of chemistry over topology without changing the topology itself.

15.9 Mutation-Weighted Graphs

The present codon cube assigns equal weight to every edge selected by a nucleotide path.

That equality is mathematically clean but biologically simplified.

A weighted graph can assign a weight

$$m_p(x, y)$$

to the substitution between nucleotides x and y at codon position p.

The weighted synonymous score is then

$$L_w = \Sigma_{e \in E} m_e \, 1\{\text{the endpoints of } e \text{ are synonymous}\}.$$

Possible weights include:

transition-versus-transversion rates;

empirical mutation spectra;

replication-strand asymmetry;

organism-specific substitution frequencies;

measured mistranslation probabilities;

or experimentally estimated fitness consequences.

The resulting objective asks a different question from the unweighted forty-four-edge theorem.

The unweighted score asks how many selected local movements preserve translated identity.

The weighted score asks how much biologically weighted movement preserves identity.

G–A–C–U need not remain optimal under every weighting scheme.

If it remains strongly competitive across several independently motivated weights, its structural significance would broaden.

If a different path consistently performs better under realistic mutation models, the interpretation should be narrowed to the original equal-edge objective.

Directionality may also matter.

The current cube is undirected, but mutation and translation-error processes may satisfy

$$m(x \to y) \neq m(y \to x).$$

A directed model can represent mutation bias, repair asymmetry, codon abundance, or asymmetric mistranslation.

The translated quotient then becomes a directed multigraph with loops and parallel arcs.

The exact undirected result should remain the reference baseline. Weighted and directed extensions describe additional biological processes placed over that baseline.

15.10 Codon-Usage-Weighted Analyses

The complete codon cube treats every codon as one equally present input of the lookup relation.

Actual organisms do not use synonymous codons equally.

Let

u c

denote the usage frequency of codon c in a selected organism, genome, tissue, or gene set.

An edge between codons c and d could be weighted by, for example,

√(u c u d),

the arithmetic mean of the two usage frequencies, their minimum, or a directed mutation flux.

Different choices answer different questions.

A symmetric frequency weight describes how strongly the organism occupies both endpoints of an edge.

A directed flux weight describes expected movement from one codon to another under a specified mutation model.

Codon-usage weighting creates organism-specific topologies over the same universal codon domain.

It may reveal that some of the forty-four synonymous edges are heavily used in one organism and nearly irrelevant in another.

It may also show that an organism’s preferred codons concentrate in particular regions of the G–A–C–U cube.

The unweighted analysis should remain primary because it describes the complete lookup relation itself.

Codon-usage analysis describes how particular biological systems sample and occupy that relation.

The two levels should not be confused.

15.11 Complete Multigraph Analysis

The present topological optimization concentrates on self-loops because they provide a transparent measure of local translation preservation.

The translated multigraph also contains one hundred interclass edge instances.

Those edges should be analyzed rather than treated as an undifferentiated remainder.

For each pair of distinct translated classes ω and ν, define

$m_{\omega,\nu}$

as the number of selected codon-cube edges connecting a codon in F_ω with a codon in F_ν.

The resulting 21 × 21 symmetric multiplicity matrix contains:

self-loop counts on the diagonal;

and interclass edge multiplicities off the diagonal.

This matrix supports several analyses.

One can identify the translated pairs joined by the largest numbers of codon-level edges.

One can separate first-, second-, and third-position contributions.

One can compare edge multiplicity with hydropathy difference, volume difference, charge change, or empirical substitution scores.

One can calculate vertex degree, weighted degree, connectedness, centrality, and community structure.

One can examine whether highly connected translated classes occupy chemically intermediate positions.

One can ask whether stop X has an unusual neighborhood.

The complete Hamming-distance-one codon graph supplies a useful companion baseline. It connects codons differing at exactly one position by any of the three alternative nucleotides.

Each codon has nine such neighbors, so the full graph contains

$64 \times 9 / 2 = 288$

undirected edges.

The linear-path cube retains 144 of those edges.

The complete Hamming graph contains all sixty-nine synonymous one-substitution pairs identified by the sum of the six nucleotide-pair weights.

It has no nucleotide-order optimization because all six pair types are present. It can nevertheless measure:

fiber connectivity;

the number of synonymous components;

distances among translated fibers;

the complete local substitution neighborhood;

and the relation between the sparse path cube and the full one-substitution space.

The ordered cube should therefore be understood as a structured half-edge subgraph of the complete Hamming graph.

15.12 Quantitative Benchmarking of the U/T–A Membrane Profile

The historical membrane-profile comparison is visually persuasive. A claim of predictive equivalence to Kyte–Doolittle requires quantitative testing.

The original protein and coding sequence should first be identified unambiguously.

Both profiles should then be regenerated from that same sequence using explicitly stated:

reading frame;

sequence coordinate system;

window length;

window placement;

terminal boundary rule;

normalization;

smoothing;

vertical scaling;

and decision threshold.

The conventional profile should use the declared amino-acid hydropathy scale.

The nucleotide profile should use the declared second-position scoring rule, such as

+1 for U or T,

−1 for A,

and

0 for G or C.

Agreement can be measured by:

correlation after scale normalization;

overlap of above-threshold regions;

displacement between peak centers;

agreement in peak width;

recovery of experimentally established transmembrane segments;

false-positive regions;

and false-negative regions.

The comparison should then be expanded to an independent benchmark containing both membrane and soluble proteins.

Several controls are essential.

The two incorrect reading frames should be analyzed.

Codon order can be shuffled while preserving total codon composition.

Synonymous codons can be randomized while preserving the amino-acid sequence.

Second-position nucleotides can be selectively shuffled.

The genetic-code assignments can be randomized under declared constraints.

These controls separate a genuine reading-frame-specific code signal from general base composition, window smoothing, or sequence autocorrelation.

The central falsifiable question is:

Does the second-position U/T–A profile retain useful transmembrane-segment discrimination across independent proteins, or was the historical example unusually favorable?

Either outcome would clarify the scope of the method.

The two-state U/A profile and the full twelve-coefficient model also define endpoints of a hierarchy of compressed representations.

Intermediate models might use:

all four second-position coefficients;

the second-position coefficients plus a small number of first-position contrasts;

or a reduced basis selected from the exact additive subspace.

Such models can be compared by cross-validation.

The objective should not be to maximize fit with as many parameters as possible. It should be to identify the smallest representation that retains the principal biological signal.

15.13 Results That Can Change and Results That Cannot

The research program becomes clearer when its conclusions are divided according to what could alter them.

The following results cannot change unless the underlying matrix definition changes:

the dimensions 64 × 12;

three 1s in every row;

sixteen 1s in every column;

rank ten;

right nullity two;

left nullity fifty-four;

the singular spectrum;

the formula

$D^{+} = (1 / 16) D^{T} - (1 / 96) J;$

the two pseudoinverse values;

and the four Hamming-distance projector values.

These are exact consequences of the complete one-hot representation of all three-letter words over a four-state alphabet.

The following results are exact for the standard genetic code and the declared linear-path objective:

the six position-specific nucleotide-pair weight triples;

the sixty-nine synonymous Hamming-distance-one pairs;

the forty-four synonymous edges of the G–A–C–U cube;

the decomposition

44 = 4 + 1 + 39;

the complete twenty-four-path isotropic score list;

the 13,824-case anisotropic frequency distribution;

the maximum of forty-four;

the 192 oriented anisotropic maximizers;

the twenty-four undirected anisotropic maximizers;

and the unique common undirected path G–A–C–U.

These values could change under another translated code or another definition of nucleotide geometry. They cannot be changed by a different philosophical interpretation of the same stated calculation.

The following results are empirical or convention dependent:

the twelve hydropathy coefficient values;

the magnitude of the U–A contrast;

the fraction of hydropathy variation captured by the additive projection;

the amino-acid reconstruction correlation;

the hydropathy–volume clustering;

and the performance of the nucleotide-only membrane profile.

These may change under alternative scales, stop conventions, datasets, or benchmarking procedures.

The broader interpretation would weaken if:

alternative hydropathy scales eliminated the second-position U–A polarity;

the nucleotide membrane profile failed broadly outside the illustrated sequence;

the standard code proved ordinary under appropriately constrained random ensembles;

natural variant codes showed no related organization;

or plausible origin models generated the same structures routinely without restrictive conditions.

The broader interpretation would strengthen if:

the hydropathy hierarchy survived many property scales;

the nucleotide-only membrane profile performed well on independent proteins;

natural code variants retained related positional organization;

the standard code occupied an extreme tail of constrained null distributions;

G–A–C–U performed well under several independently motivated objectives;

and mechanistic models reproduced the same combination of structures.

No single extension will settle the historical interpretation.

Convergence among independent algebraic, physicochemical, topological, comparative, and mechanistic tests would be more persuasive than one exceptionally strong result.

15.14 Reproducibility and Provenance

Every finite enumeration in this book should be independently reproducible.

The computational package should contain:

the complete standard RNA code table;

the codon-order convention;

the exact 64 × 12 incidence matrix;

the rational pseudoinverse construction;

the nucleotide-path generator;

the cube-edge constructor;

the synonymous-edge scoring routine;

the isotropic enumeration;

the anisotropic enumeration;

the positional pair weights;

the loop multiplicities;

the hydropathy input tables;

and automated verification tests.

The package should fail closed if any required invariant is violated.

At minimum, it should verify that:

there are exactly sixty-four codons;

every codon occurs once;

every matrix row contains three 1s;

every matrix column contains sixteen 1s;

the matrix rank is ten;

the exact pseudoinverse satisfies all four Moore–Penrose equations;

the G–A–C–U score is forty-four;

the isotropic maximum is forty-four;

the anisotropic search contains exactly 13,824 oriented cases;

the anisotropic frequencies sum to 13,824;

and the maximum frequency is 192.

The strongest verification uses two independently written implementations.

One implementation can construct all codon-cube edges directly from path coordinates.

Another can calculate path scores from position-specific nucleotide-pair weights.

The two implementations should produce identical:

isotropic scores;

positional spectra;

anisotropic frequencies;

maximum families;

and loop multiplicities.

Exact arithmetic should be used wherever possible.

The pseudoinverse entries are rational.

The graph counts are integers.

The generating polynomials have integer coefficients.

Floating-point approximation is unnecessary for those results.

Floating-point methods remain appropriate for empirical property values, regression, and numerical benchmarking.

Every published figure should also have traceable provenance.

Generated figures should record:

the input data;

generation script;

software version;

script version;

image dimensions;

and export format.

The authoritative codon cube and translated multigraph should be retained from their verified source files. They should not be recreated approximately when exact combinatorial structure is at stake.

A visually attractive but mathematically altered graph is not an acceptable substitute.

15.15 The Purpose of the Controls

The purpose of these tests is not to dilute the exact findings.

It is to locate them properly.

Some structures are unavoidable consequences of complete word space.

Some are specific to the standard translated assignment.

Some may prove robust across property scales, natural code variants, and alternative graph objectives.

Others remain provisional because their historical source data or benchmark performance has not yet been fully reconstructed.

A reliable account should display all of these categories.

The value of an exact result is not diminished when its scope is stated precisely. The value of an empirical pattern is not diminished when it is subjected to a control capable of disproving it.

The exact mathematics establishes what the defined objects contain.

The proposed tests determine how broadly those structures extend and how much historical or biological weight they can bear.

The final chapter states the conclusions that are already justified.

Chapter 16

Conclusions

The standard genetic code is usually encountered as a compact table: sixty-four RNA codons map to twenty amino acids and translation stop. That representation is correct, but it conceals much of the structure present in the complete relation.

This book has treated the code as a finite mathematical object. The same sixty-four assignments were represented as a binary incidence matrix, a projected physicochemical field, a three-dimensional codon lattice, and a translated multigraph. Each representation preserved a different aspect of the relation, and each exposed patterns that are difficult to see in the conventional chart.

The principal conclusions can be stated precisely.

The complete codon input domain has an exactly balanced 64 × 12 one-hot representation with rank ten.

Its Moore–Penrose pseudoinverse has a closed form containing only two rational values.

A large part of amino-acid hydropathy is recoverable from additive nucleotide-position coordinates, with the dominant contrast located between U and A at the second codon position.

A nucleotide-only second-position profile can reproduce the principal broad membrane-hydropathy peaks in the historical protein example.

The isotropic G–A–C–U codon cube contains forty-four synonymous nearest-neighbor edges.

No anisotropic combination of independently ordered linear nucleotide axes exceeds forty-four.

G–A–C–U is the unique undirected nucleotide path belonging simultaneously to the maximizing families of the first, second, and third codon positions.

These results belong to different logical levels. Some are exact consequences of complete word space. Others depend on the standard translated assignments. Still others depend on empirical property scales or declared analytical conventions. Their synthesis is strongest when those distinctions remain explicit.

16.1 The Complete Codon Matrix

The complete RNA codon space is

$C = \{ G, A, C, U \}^3$.

Its sixty-four codons can be encoded by twelve binary nucleotide-position coordinates:

1G, 1A, 1C, 1U;

2G, 2A, 2C, 2U;

and

3G, 3A, 3C, 3U.

The resulting matrix

$D \in \{ 0, 1 \}^{64 \times 12}$

contains exactly three 1s in every row and sixteen 1s in every column.

Its balance is complete.

Every nucleotide appears equally often at every codon position.

Every nucleotide-coordinate pair belonging to two different positions occurs exactly four times.

Distinct nucleotide states within one position never occur together.

These facts determine the Gram matrix

$D^T D$

without numerical approximation.

The diagonal 4×4 blocks are

$16\, I_4$.

The off-diagonal blocks are

$4\ J_4$.

The eigenvalues are:

48 with multiplicity one;

16 with multiplicity nine;

and

0 with multiplicity two.

Therefore, the singular values of D are:

4 √3 once;

4 nine times;

and

0 twice.

The matrix rank is

rank (D) = 10.

The right nullity is two, and the left nullity is fifty-four.

The rank deficiency does not arise from missing codons or incomplete observations. It arises because the four columns within each positional block sum to the same constant sixty-four-component vector.

The three codon positions supply three copies of one constant direction, creating two redundant blockwise offsets.

The ten-dimensional column space decomposes naturally into:

one global constant direction;

three first-position nucleotide contrasts;

three second-position nucleotide contrasts;

and

three third-position nucleotide contrasts.

Thus,

$10 = 1 + 3 + 3 + 3$.

This structure belongs to the complete codon input domain before translation is introduced.

16.2 The Exact Pseudoinverse

The rank deficiency prevents an ordinary matrix inverse, but it does not prevent exact inversion in the Moore–Penrose sense.

The pseudoinverse is

$D^{+} = (1 / 16) D^{T} - (1 / 96) J_{12 \times 64}$.

Every entry of D^{+} is either

$5 / 96$

or

$-1 / 96$.

The positive value occurs when a codon contains the nucleotide-position coordinate represented by the corresponding row of D^{+}.

The negative value occurs when that coordinate is absent.

Every pseudoinverse column contains:

three entries equal to $5 / 96$;

and

nine entries equal to $-1 / 96$.

Every pseudoinverse row contains:

sixteen entries equal to $5 / 96$;

and

forty-eight entries equal to $-1 / 96$.

Every column sums to

$1 / 16$.

Every row sums to

$1 / 3$.

The pseudoinverse therefore does more than provide a numerical least-squares solution. It reveals an exact algebraic relation between complete codon space and its additive coordinate representation.

For any codon-level numerical field h,

$\beta = D^{+} h$

is the unique minimum-norm nucleotide-position coefficient vector.

The fitted field

$\hat{h} = D D^{+} h$

is the unique orthogonal projection of h into the ten-dimensional additive subspace.

The residual

$r = h - \hat{h}$

is orthogonal to every additive nucleotide-position field.

The projection matrix itself has the exact form

$P = D D^{+}$

$= (1 / 16) D D^{T} - (1 / 32) J_{64}.$

Its entries depend only on Hamming distance:

5 / 32 for identical codons;

3 / 32 for distance one;

1 / 32 for distance two;

and

−1 / 32 for distance three.

The complete input domain therefore has two related exact low-value structures:

a two-value pseudoinverse;

and

a four-value codon-space projector.

These results are unusual in their simplicity, but they are not mysterious. They follow from the perfect balance of complete word space.

16.3 Nucleotide-Determined Hydropathy

The genetic code becomes biologically specific when translated properties are placed over the balanced codon domain.

For hydropathy, every sense codon inherits the Kyte–Doolittle value of its translated amino acid. In the complete sixty-four-codon calculation, translation stop X is assigned the explicit numerical convention

H (X) = 0.

Applying the exact pseudoinverse produces twelve nucleotide-position coefficients.

The strongest positive coefficient is

β2,U = 3.948958.

The strongest negative coefficient is

β2,A = −2.638542.

Their gauge-invariant difference is

β2,U − β2,A = 6.587500.

The second-position G coefficient is

β2,G = −1.469792,

while the second-position C coefficient is

β2,C = −0.163542.

The second codon position therefore contains the principal hydropathy polarity.

The positional coefficient ranges are approximately:

2.131250 at the first position;

6.587500 at the second position;

and

0.506250 at the third position.

The second-position range is more than three times the first-position range and about thirteen times the third-position range.

The additive reconstruction captures approximately

81.7 percent

of the centered squared codon-level hydropathy variation.

After fitted codon values are averaged within synonymous fibers, the twenty reconstructed amino-acid values have Pearson correlation

$r = 0.915646$

with their empirical Kyte–Doolittle values.

The corresponding coefficient of determination is

$R^2 = 0.838407$.

The relation is strong but incomplete. Amino-acid chemistry is not reducible to three independent nucleotide-position effects. The residual contains property distinctions and codon-position interactions that the additive model cannot represent.

The important result is not exact chemical reconstruction. It is that a large component of amino-acid hydropathy is aligned with the nucleotide coordinates through which the standard genetic code specifies translated identity.

16.4 The Second Codon Position

The pseudoinverse result agrees with the direct physicochemical grouping of the code.

The four second-position codon families are:

NGN;

NAN;

NCN;

and

NUN.

Each contains sixteen codons.

NUN encodes the strongly hydrophobic group containing phenylalanine, leucine, isoleucine, methionine, and valine.

NAN encodes a predominantly hydrophilic, polar, or charged group containing aspartate, glutamate, asparagine, lysine, histidine, glutamine, and tyrosine, together with two stop codons.

NCN encodes alanine, threonine, proline, and most serine codons, producing a region enriched in smaller and intermediate-volume residues.

NGN encodes a chemically broader group containing glycine, arginine, serine, tryptophan, and cysteine, together with one stop codon.

The middle nucleotide therefore acts as a coarse physicochemical organizer.

It does not determine exact amino-acid identity. The first and third nucleotides remain essential. Yet the second coordinate places codons into broad translated property regions before the remaining coordinates refine the assignment.

This organization becomes visible in an actual coding sequence.

The historical membrane-protein method used the second-position contrast

T − A

in DNA notation, or

U − A

in RNA notation.

A second-position U or T contributed positively.

A second-position A contributed negatively.

G and C contributed zero in the simplest implementation.

This highly compressed nucleotide-only profile reproduced the principal broad hydrophobic peaks of a conventional Kyte–Doolittle profile in the illustrated membrane protein.

The comparison does not establish universal predictive equivalence. It does show that a translated physicochemical feature can remain visible in the coding sequence without explicit assignment of twenty amino-acid hydropathy parameters.

The exact pseudoinverse explains why this compression works. The historical profile retains the dominant direction of the complete additive hydropathy model.

16.5 The G–A–C–U Codon Geometry

The codon product space does not contain a unique notion of local nucleotide adjacency. A geometry must be declared.

This book examined linear paths through the four nucleotide states.

The principal path is

G–A–C–U.

Applying the path independently to all three codon positions creates a 4 × 4 × 4 lattice with:

sixty-four codon vertices;

forty-eight edges along each codon axis;

and

144 edges in total.

Every edge changes exactly one codon position by one adjacent step along the nucleotide path.

The cube is a sparse subgraph of the complete Hamming-distance-one codon graph, which contains 288 edges.

The G–A–C–U cube retains exactly half of those possible single-position nucleotide changes.

When translated labels are placed on its vertices, forty-four of the 144 selected edges connect synonymous codons.

The positional decomposition is

44 = 4 + 1 + 39.

Four synonymous edges change the first codon position.

One changes the second position.

Thirty-nine change the third position.

The topology therefore exhibits the inverse of the hydropathy hierarchy.

The second position strongly distinguishes translated physicochemical character but almost never preserves translated identity along a selected edge.

The third position weakly distinguishes hydropathy but contains most of the synonymous local connectivity.

The first position occupies an intermediate role.

The three codon coordinates are balanced in the input matrix but strongly differentiated by the translated assignments.

16.6 The Forty-Four-Loop Multigraph

Collapsing the sixty-four codon vertices by translated identity produces a twenty-one-vertex multigraph.

Every synonymous codon edge becomes a self-loop.

Every nonsynonymous codon edge becomes an edge between two distinct translated vertices.

Parallel edges are retained.

The translated multigraph contains:

forty-four self-loop edge instances;

one hundred interclass edge instances;

and

144 edge instances in total.

The loop multiplicities differ among translated classes.

Arginine and leucine each contribute six loops.

Serine contributes four.

Alanine, glycine, proline, threonine, and valine each contribute three.

Isoleucine and X each contribute two.

Nine two-codon amino-acid classes each contribute one.

Methionine and tryptophan contribute none because each has only one codon.

This distribution demonstrates that degeneracy and topology are not the same.

Arginine, leucine, and serine each have six codons, but serine has fewer G–A–C–U self-loops.

A codon fiber has both a cardinality and a shape.

The conventional genetic-code table displays the first readily. The translated multigraph reveals the second.

16.7 The G–A–C–U Maximum

The score of forty-four is not privileged by assumption.

All twenty-four oriented linear nucleotide orders were evaluated isotropically. Equivalently, all twelve undirected paths were evaluated after identifying each path with its reverse.

The isotropic scores range from

25

to

44.

G–A–C–U and its reverse U–C–A–G attain the maximum.

No other undirected path reaches forty-four.

The exact total nucleotide-pair weights are:

A–G = 15;

A–C = 11;

A–U = 9;

C–G = 8;

C–U = 18;

and

G–U = 8.

The G–A–C–U path selects:

A–G;

A–C;

and

C–U.

Their weights sum to

15 + 11 + 18 = 44.

These are the three largest total pair weights, and they form one valid four-nucleotide path. No alternative path can have a larger sum.

The complementary path A–U–G–C selects the three omitted pair types and scores

9 + 8 + 8 = 25.

The six total pair weights sum to

69,

which is the number of all synonymous codon pairs at Hamming distance one when every nucleotide substitution type is admitted.

Complementary path scores therefore sum to sixty-nine, producing symmetry about

34.5.

The isotropic result is exact and unique up to reversal.

16.8 The Anisotropic Maximum

The larger anisotropic search permits a different nucleotide path at each codon position.

There are

$24^3 = 13{,}824$

oriented axis-order combinations.

The total score separates exactly:

$L(\pi_1, \pi_2, \pi_3)$

$= L_1(\pi_1) + L_2(\pi_2) + L_3(\pi_3)$.

The maximum positional scores are:

4 at position 1;

1 at position 2;

and

39 at position 3.

Therefore,

$L \le 4 + 1 + 39 = 44$.

The bound is attained.

No isotropic or anisotropic linear-path codon cube exceeds forty-four synonymous edges.

The complete anisotropic score distribution ranges from twenty-five to forty-four. Its mean and median are 34.5, and its population standard deviation is 5.5.

The distribution is exactly symmetric:

$N_s = N_{69-s}$.

There are:

192 oriented maximizing cubes;

and

twenty-four undirected maximizing cubes.

Thus, the forty-four-edge maximum is not realized by only one anisotropic cube.

The special status of G–A–C–U is more specific.

The first-position maximum family contains two undirected paths.

The second-position family contains six.

The third-position family contains two.

Their intersection contains exactly one undirected path:

G–A–C–U.

It is therefore the unique path that can be used identically on all three axes while independently maximizing each codon position.

This is the strongest topological statement established in the book.

16.9 Structure Beyond the Conventional Lookup Table

The standard genetic-code table contains all the data used in these analyses, but it does not display all of the resulting structure.

The table does not visibly announce that its complete codon matrix has rank ten.

It does not display the exact two-value pseudoinverse.

It does not show that the codon-space projector depends only on Hamming distance.

It does not quantify the second-position hydropathy coefficients.

It does not display the U–A contrast as a dominant additive direction.

It does not identify the forty-four selected synonymous edges.

It does not distinguish degeneracy from local fiber compactness.

It does not show that the G–A–C–U path is the unique common positional maximizer.

The structures are latent in the complete relation.

They become visible only after the relation is represented, projected, ordered, collapsed, and exhaustively compared.

The lesson is methodological.

A familiar lookup table can remain mathematically underexamined even when every individual entry is known.

Knowing all entries is not the same as knowing all relations among them.

16.10 What Is Exact and What Is Conditional

The conclusions should be separated according to their dependencies.

The matrix dimensions, balance, rank, singular values, pseudoinverse formula, and projector values are exact consequences of representing all three-letter words over a four-state alphabet by one-hot positional coordinates.

They do not depend on the standard translated assignments.

The hydropathy coefficients are exact results of a declared numerical calculation, but they depend on:

the standard genetic code;

the Kyte–Doolittle scale;

the complete sixty-four-codon representation;

the convention H (X) = 0;

and the Moore–Penrose minimum-norm gauge.

The synonymous-edge counts and maximum theorems are exact for:

the standard genetic code;

linear four-state nucleotide paths;

equal weighting of selected edges;

and the objective of preserving translated identity.

The membrane-profile correspondence is empirical and should be tested on additional proteins under fully specified computational conditions.

The interpretation of the results as historically important, evolutionarily constrained, informationally discontinuous, or exceptionally persistent goes beyond the exact arithmetic. Such interpretations may be reasonable, but they require argument and comparison rather than being direct outputs of the matrix.

16.11 Beyond Frozen Accident

Historical freezing can explain why an established code becomes difficult to change.

It does not, by itself, explain why the frozen relation contains the measured organization.

The phrase **frozen accident** therefore remains incomplete.

A code with weak physicochemical organization could freeze.

A code with different synonymous topology could freeze.

A code whose positional roles were distributed differently could freeze.

The present code contains:

a strong second-position hydropathy polarity;

extensive third-position translation preservation;

an exact forty-four-edge linear-path maximum;

and

a unique nucleotide path satisfying the three positional maximum conditions simultaneously.

These are properties of what was frozen, not merely consequences of freezing.

An adequate historical account should explain, reproduce, or render statistically unsurprising this combination of structures.

That account may involve contingency.

It may involve chemical constraint.

It may involve error minimization, code expansion, coevolution, local optimization, historical lock-in, or several mechanisms together.

The mathematics does not select one origin theory.

It prevents the internal architecture from disappearing behind a phrase about persistence.

16.12 A Possible Information-Theoretic Transition

A complete genetic code is more than a collection of isolated molecular affinities.

It is a globally coordinated mapping in which every triplet receives an interpreted output.

The appearance of such a relation may represent an information-theoretic transition.

The underlying molecular components could have accumulated gradually. The system-level relation could nevertheless have appeared discontinuously when a threshold of completeness, accuracy, and self-maintenance was crossed.

Below that threshold, there may be partial, ambiguous, or locally restricted associations.

Above it, a sufficiently complete lookup relation exists and participates in producing components that continue to enact it.

The products of translation help maintain transfer-RNA charging, ribosome function, proofreading, metabolism, and the wider cellular environment required for further translation.

This is reflexive maintenance.

It does not solve the origin problem. It sharpens it.

The problem is not merely to explain isolated triplet affinities.

It is to explain the appearance of a persistent, distributed, self-maintaining relation between finite alphabets.

16.13 The Genetic Code as Persistent Natural Signal

Individual biological messages are transient.

Genes mutate, duplicate, recombine, diverge, and disappear.

The translation relation persists across an immense succession of such messages.

The code is not one surviving sequence. It is the rule through which sequences are interpreted.

Its physical instances continually turn over. Transfer RNAs, ribosomes, proteins, cells, and organisms are replaced. Yet successor systems reconstruct and enact substantially the same mapping.

What persists is relational rather than material.

This makes the genetic code a candidate for an exceptionally persistent natural signal.

The statement should remain qualified. The present analysis does not prove that the genetic code is the oldest possible natural signal, the most persistent signal under every definition, or the only natural lookup relation.

It does establish that the code is:

finite;

broadly shared;

physically distributed;

globally consequential;

reflexively maintained;

repeatedly instantiated;

and mathematically recoverable as a complete lookup relation.

Few natural systems display that combination with comparable clarity.

16.14 What Remains to Be Explained

The book resolves several mathematical questions.

The complete matrix is known.

Its rank and singular spectrum are known.

Its pseudoinverse is known exactly.

The dominant hydropathy coordinate is known.

The forty-four-loop topology is known.

The complete isotropic and anisotropic linear-path searches are known.

The maximum family and its unique common path are known.

The deeper historical questions remain open.

Why does the standard code place broad hydropathy primarily at the second codon position?

Why does the third position carry so much local synonymous connectivity?

Why do the position-specific nucleotide-pair weights intersect in the path G–A–C–U?

Which portions of this organization follow from chemical necessity?

Which follow from historical sequence?

Which reflect selection for robustness or translational efficiency?

Which are correlated consequences of other constraints?

How exceptional are the combined structures under appropriately restricted alternative-code ensembles?

How well does the nucleotide-only hydropathy signal generalize across real membrane proteins?

How do natural variant codes perturb the algebraic projections and topological maxima?

These are now sharper questions because the present structure has been specified exactly.

16.15 Final Perspective

The genetic code can be viewed as a lookup table, but it is not exhausted by lookup.

It is a complete finite domain.

It is a balanced algebraic design.

It is a carrier of projected physicochemical information.

It is a geometry of local codon relations.

It is a multigraph of translated identity and change.

It is a persistent rule enacted through changing molecular matter.

The exact input-space mathematics is simple enough to write in closed form:

$D^{+} = (1 / 16) D^{T} - (1 / 96) J.$

The principal physicochemical contrast is equally compact:

second-position U opposed to second-position A.

The principal topological result is a finite maximum:

$44 = 4 + 1 + 39.$

The unique common nucleotide path is

G–A–C–U.

These statements do not complete the explanation of the genetic code.

They define more precisely what must be explained.

The standard genetic code is not merely a list of sixty-four assignments preserved by history. It is a complete relational structure whose algebraic balance, physicochemical organization, and translated topology can be derived, measured, compared, and tested.

That structure has been present in the familiar table all along.

Technical Appendices

Appendix A

Construction of the Complete 64 × 12 Codon Matrix

A.1 Alphabet and Word Space

Let the RNA nucleotide alphabet be

$\Sigma = \{ G, A, C, U \}$.

A codon is an ordered word of length three:

$c = c_1 c_2 c_3$,

with

$c_1, c_2, c_3 \in \Sigma$.

The complete codon set is

$C = \Sigma^3$.

Therefore,

$| C | = 4^3 = 64$.

No biological output assignment is required to construct this set. It is the complete Cartesian product of four nucleotide states across three ordered positions.

The word **complete** means that every possible triplet over the declared alphabet occurs exactly once. The set includes the sixty-one sense codons and the three codons that translate to X under the standard genetic code.

The input set exists independently of how its members are translated. A reassignment of amino-acid or stop labels would change the output relation but would not change C.

A.2 Positional One-Hot Coordinates

Each codon position contributes four binary indicator coordinates.

Using G–A–C–U order within each positional block, the twelve coordinates are

1G, 1A, 1C, 1U,

2G, 2A, 2C, 2U,

3G, 3A, 3C, 3U.

For codon c, define the coordinate indicator

$D_{c,(p,n)} = 1$

when

$c_p = n$,

and

$D_{c,(p,n)} = 0$

otherwise.

Here,

$p \in \{ 1, 2, 3 \}$

denotes the codon position, and

$n \in \Sigma$

denotes the nucleotide state.

The complete design matrix is

$D \in \{ 0, 1 \}^{64 \times 12}$.

Its rows correspond to codons. Its columns correspond to nucleotide-position coordinates.

A row of D is therefore the concatenation of three four-component one-hot vectors:

D_c

$= [e (c_1) \mid e (c_2) \mid e (c_3)]$,

where

$e (G) = (1, 0, 0, 0)$,

$e (A) = (0, 1, 0, 0)$,

$e (C) = (0, 0, 1, 0)$,

and

$e (U) = (0, 0, 0, 1)$.

A.3 Row Structure

Every codon contains exactly one nucleotide at each of its three positions.

Therefore, every row of D contains exactly one 1 in each four-column positional block.

For every codon c,

$\Sigma n \in \Sigma\ D\ c,(1,n) = 1,$

$\Sigma n \in \Sigma\ D\ c,(2,n) = 1,$

and

$\Sigma n \in \Sigma\ D\ c,(3,n) = 1.$

Summing across all twelve columns gives

$\Sigma p=1^{3}\ \Sigma n \in \Sigma\ D\ c,(p,n) = 3.$

Thus, every row contains:

three entries equal to 1;

and

nine entries equal to 0.

In matrix notation,

$D\ 1_{12} = 3\ 1_{64}.$

The total number of 1s obtained from the row count is

$64 \times 3 = 192.$

The total number of binary entries is

$64 \times 12 = 768.$

Therefore, the number of 0s is

$768 - 192 = 576.$

A.4 Column Structure

Fix one codon position p and one nucleotide n.

The coordinate (p, n) fixes one position of the codon. The remaining two positions may each contain any of four nucleotide states.

The coordinate therefore occurs in

$4^2 = 16$

codons.

Hence, every column sum is

$\Sigma c \in C\ D\ c,(p,n) = 16.$

In matrix notation,

$D^T\ 1_{64} = 16\ 1_{12}.$

Every column consequently contains:

sixteen entries equal to 1;

and

forty-eight entries equal to 0.

The total number of 1s obtained from the column count is

$12 \times 16 = 192.$

This agrees exactly with the row count.

The equality

$64 \times 3 = 12 \times 16$

is a basic incidence identity for the complete matrix.

A.5 Lexicographic Row Order

A convenient row order is lexicographic under the nucleotide ordering

$G < A < C < U.$

The rows begin

GGG,

GGA,

GGC,

GGU,

GAG,

GAA,

GAC,

GAU,

and continue through all codons ending with

UUU.

More explicitly, the first coordinate changes slowest, the second changes at the intermediate rate, and the third changes fastest.

The sixty-four codons are grouped as follows:

GGN, GAN, GCN, GUN;

AGN, AAN, ACN, AUN;

CGN, CAN, CCN, CUN;

UGN, UAN, UCN, UUN,

where N cycles through G, A, C, and U.

The mathematical properties of D do not depend on this row order.

If R is a 64 × 64 permutation matrix, then

$D' = R\,D$

contains the same codon rows in another order.

Because

$D'^{T}\,D'$

$= D^{T}\,R^{T}\,R\,D$

$= D^{T}\,D,$

row permutation leaves the Gram matrix, rank, singular values, and right-null space unchanged.

The row order should nevertheless be stated because it determines:

the visual form of Figure 1;

the column order of D^{+};

the order of the categorical output vector;

and the order of every codon-level numerical field.

A.6 Example Rows

For

c = GAC,

the active coordinates are

1G, 2A, and 3C.

Therefore,

D GAC

= (1, 0, 0, 0 |

0, 1, 0, 0 |

0, 0, 1, 0).

For

c = AUG,

the active coordinates are

1A, 2U, and 3G.

Therefore,

D AUG

= (0, 1, 0, 0 |

0, 0, 0, 1 |

1, 0, 0, 0).

For

c = UGA,

the active coordinates are

1U, 2G, and 3A.

Therefore,

D UGA

= (0, 0, 0, 1 |

1, 0, 0, 0 |

0, 1, 0, 0).

For

c = UUU,

the active coordinates are

1U, 2U, and 3U.

Therefore,

D UUU

= (0, 0, 0, 1 |

0, 0, 0, 1 |

0, 0, 0, 1).

Every example contains one active coordinate in each positional block.

A.7 The Categorical Output Vector

Let

$\tau : C \rightarrow \Omega$

be the standard genetic-code map, where

Ω

contains the twenty amino-acid single-letter codes and translation stop X.

Using the same codon order as the rows of D, define the categorical output vector

$y = (\tau (c))c \in C$.

The vector y has dimensions

64×1.

It is categorical rather than numerical. Its entries include symbols such as

G, E, D, A, V, R, S, K, N, T, M, I, Q, H, P, L, W, C, Y, F,

and

X.

The output vector is displayed beside the binary matrix in Figure 1, but it is not one of the columns of D.

This distinction separates:

the complete codon input representation;

from

the translated assignment placed over it.

The matrix D remains unchanged if the output labels are reassigned.

A.8 Numerical Property Vectors

Let

$H : \Omega \rightarrow \mathbb{R}$

assign a numerical property to translated outputs under a declared convention.

The induced codon-level property vector is

$h\ c = H (\tau (c))$.

Thus,

$h \in \mathbb{R}^{64}$.

Examples of possible properties include:

hydropathy;

residue volume;

charge;

polarity;

or another numerical amino-acid attribute.

The binary matrix D remains unchanged when the property changes. Only the vector h changes.

Translation stop X requires special treatment in any amino-acid property calculation because X is not an amino-acid residue. A numerical value assigned to X must therefore be stated explicitly.

In the complete hydropathy calculation used in this book,

$H(X) = 0$.

This convention affects h but does not affect D.

A.9 Complete Balance Across Two Positions

Fix two distinct codon positions p and q and two nucleotides n and m.

The number of codons satisfying

$c_p = n$

and

$c_\varphi = m$

is

4.

The reason is that the remaining codon position is unrestricted and may contain any of the four nucleotide states.

Therefore, for

$p \neq q$,

the inner product of the corresponding incidence columns is

$D(p,n)^T D(q,m) = 4$.

This holds for every pair of nucleotides and every pair of distinct codon positions.

For example, the coordinates 1G and 2A occur together in

GAG,

GAA,

GAC,

and

GAU.

The coordinates 2C and 3U occur together in

GCU,

ACU,

CCU,

and

UCU.

This uniform cross-position co-occurrence produces the all-ones off-diagonal blocks of the Gram matrix.

A.10 Mutual Exclusivity Within One Position

Fix one codon position p and two different nucleotides n and m.

No codon can contain both nucleotide states at the same position.

Therefore, when

$n \neq m$,

$D(p,n)^T D(p,m) = 0$.

For a column paired with itself,

$D(p,n)^T D(p,n) = 16$,

because the coordinate occurs in sixteen codons.

Thus, the four columns within one positional block are mutually orthogonal and have squared norm sixteen.

The within-position Gram block is consequently

$16\, I_4$.

Together with the cross-position relation from Section A.9, this determines the complete block structure

$$D^T D = \begin{bmatrix} 16\, I_4 & 4\, J_4 & 4\, J_4 \\ 4\, J_4 & 16\, I_4 & 4\, J_4 \\ 4\, J_4 & 4\, J_4 & 16\, I_4 \end{bmatrix}.$$

Appendix B derives the rank and singular spectrum from this matrix.

A.11 Positional Block Sums

Let D_1, D_2, and D_3 denote the three 64×4 positional blocks of D.

Within each block, exactly one coordinate is active in every row. Therefore,

$D_1 \, 1_4 = 1_{64}$,

$D_2 \, 1_4 = 1_{64}$,

and

$D_3 \, 1_4 = 1_{64}$.

The three block sums are identical.

Consequently,

$D_1 \, 1_4 - D_2 \, 1_4 = 0$,

and

$D_1 \, 1_4 - D_3 \, 1_4 = 0$.

These are two independent linear dependencies among the twelve matrix columns.

Equivalently, the vectors

$z_1 = (1_4, -1_4, 0_4)^T$

and

$z_2 = (1_4, 0_4, -1_4)^T$

satisfy

$D \, z_1 = 0$

and

$D \, z_2 = 0$.

The equal block sums are the source of the two-dimensional right null space.

A.12 Construction Algorithm

The complete matrix can be generated directly from the codon list.

For each codon c:

1. initialize a twelve-component row of zeros;

2. identify c_1 and set the corresponding first-position coordinate to 1;
3. identify c_2 and set the corresponding second-position coordinate to 1;
4. identify c_3 and set the corresponding third-position coordinate to 1;
5. append the row to D;
6. append τ (c) to the categorical output vector.

The following conditions should be verified immediately after construction:

D has dimensions 64 × 12;

all entries belong to { 0, 1 };

every row sum equals 3;

every column sum equals 16;

the total number of 1s is 192;

the total number of 0s is 576;

every codon occurs exactly once;

and the output vector contains sixty-four entries in the same row order.

These checks should fail closed. A matrix that violates any one of them should not be used for subsequent calculations.

A.13 Independence from Biological Assignment

The matrix D contains no amino-acid or stop information.

It depends only on:

the alphabet Σ;

the word length three;

the inclusion of all sixty-four words;

the positional one-hot encoding;

and the selected row and column conventions.

The standard genetic-code map τ may be attached after D has been constructed.

A numerical property field h may be attached after τ has been specified.

A nucleotide topology may be imposed independently by declaring a path among G, A, C, and U.

The rank, Gram matrix, singular spectrum, null space, exact pseudoinverse, and codon-space projector are properties of D.

Hydropathy coefficients, synonymous fibers, self-loop multiplicities, and nucleotide-path scores arise only after biological output information or an adjacency structure is introduced.

The construction therefore provides the exact input-space foundation for all subsequent analyses.

Appendix B

Gram Matrix, Rank, and Singular Spectrum

B.1 Purpose

The complete 64 × 12 codon incidence matrix contains twelve displayed nucleotide-position coordinates, but those coordinates are not all independent. The rank, null spaces, eigenvalues, and singular values can be derived exactly from the combinatorial balance of the complete codon set.

No numerical singular-value decomposition is required for the proof.

The derivation begins with the Gram matrix

$G = D^T D$.

Because G is only 12 × 12 and has a simple three-block form, its complete spectrum can be determined by separating the coefficient space into:

one global constant direction;

nine within-position nucleotide-contrast directions;

and two blockwise constant null directions.

These subspaces are mutually orthogonal and together span all twelve coefficient coordinates.

B.2 Partition of the Incidence Matrix

Write the complete codon matrix as three positional blocks:

$D = [\, D_1 \mid D_2 \mid D_3 \,]$,

where each block has dimensions

64×4.

The block D_1 contains the four indicators for the first codon position:

1G, 1A, 1C, and 1U.

The block D_2 contains the corresponding second-position indicators.

The block D_3 contains the third-position indicators.

Each row of every block contains exactly one 1.

Therefore,

$D_1\, 1_4 = 1_{64}$,

$D_2\, 1_4 = 1_{64}$,

and

$D_3\, 1_4 = 1_{64}$.

The equality of these three block sums will generate the two zero singular values.

B.3 Within-Position Gram Blocks

Consider one positional block D_p.

Each of its four nucleotide columns contains sixteen 1s. Therefore, the squared norm of every column is

16.

Two different nucleotide columns within the same position are mutually exclusive. No codon can contain two nucleotide states at one position.

Thus, for $n \neq m$,

$D(p,n)^T D(p,m) = 0$.

The within-position Gram matrix is consequently

$D_p^T D_p = 16\, I_4$.

This identity holds for

p = 1, 2, and 3.

B.4 Cross-Position Gram Blocks

Now consider two different codon positions p and q.

Fix nucleotide n at position p and nucleotide m at position q. The remaining codon position may contain any of four nucleotide states.

Exactly four codons therefore contain both selected coordinates.

Hence,

$D(p,n)^T D(q,m) = 4$

for every pair $n, m \in \Sigma$ when $p \neq q$.

Every entry of a cross-position 4×4 Gram block is therefore four:

$D_p^T D_\varphi = 4 J_4$,

for

$p \neq q$.

The complete Gram matrix is

$G = D^T D$

=

$$\begin{bmatrix} 16 I_4 & 4 J_4 & 4 J_4 \\ 4 J_4 & 16 I_4 & 4 J_4 \\ 4 J_4 & 4 J_4 & 16 I_4 \end{bmatrix}.$$

This exact block form contains the full right-side spectral structure of D.

B.5 Trace and Frobenius-Norm Check

The diagonal of G contains twelve entries equal to sixteen.

Therefore,

$tr (G) = 12 \times 16 = 192.$

Since

$G = D^T D$,

the trace is also the squared Frobenius norm of D:

$tr (D^T D) = \| D \|^2 F.$

The matrix D contains exactly 192 entries equal to 1. Because all other entries are zero,

$\| D \|^2 F = 192.$

Thus,

$\mathrm{tr}(G) = \|D\|_F^2 = 192.$

This identity will later equal the sum of the twelve Gram eigenvalues and the sum of the squared singular values.

B.6 The Global Constant Direction

Consider the twelve-component vector

$u_o = (1_4, 1_4, 1_4)^T.$

The action of the diagonal block on 1_4 is

$16\, I_4\, 1_4 = 16\, 1_4.$

The action of an off-diagonal block is

$4\, J_4\, 1_4.$

Because

$J_4\, 1_4 = 4\, 1_4,$

it follows that

$4\, J_4\, 1_4 = 16\, 1_4.$

Each output block receives one diagonal contribution and two cross-position contributions:

$16\, 1_4 + 16\, 1_4 + 16\, 1_4$

$= 48\, 1_4.$

Therefore,

$G\, u_o = 48\, u_o.$

The global constant direction is an eigenvector of G with eigenvalue

$\lambda_o = 48.$

Its corresponding singular value is

$\sigma_o = \sqrt{48} = 4\sqrt{3}.$

The eigenvalue has multiplicity one because the globally constant subspace is one-dimensional.

A normalized constant eigenvector is

$v_o = (1 / \sqrt{12}) 1_{12}$.

The corresponding left singular vector is proportional to the constant codon field:

$D v_o$

$= (3 / \sqrt{12}) 1_{64}$

$= (\sqrt{3} / 2) 1_{64}$.

After normalization, the left singular vector is

u_o,left $= (1 / 8) 1_{64}$,

because

$\| 1_{64} \| = 8$.

Indeed,

$D v_o = 4 \sqrt{3} u_o$,left.

B.7 Within-Position Contrast Subspaces

For one codon position, consider a four-component vector x satisfying

$1_4^T x = 0$.

Such a vector is a nucleotide contrast. It compares nucleotide states within one position while having zero total coefficient.

Because

$J_4 x = 0$,

all cross-position Gram blocks annihilate x.

Suppose the contrast is supported only in the first positional block:

$v = (x, 0_4, 0_4)^T$.

Then

$G v$

$= (16 x, 0_4, 0_4)^T$

$= 16 v$.

Therefore, every first-position zero-sum contrast is an eigenvector with eigenvalue sixteen.

The same argument applies to the second and third positions.

A four-component space has a three-dimensional zero-sum subspace. Each codon position therefore contributes three independent eigenvectors.

The total multiplicity is

$3 \times 3 = 9$.

Thus,

$\lambda_{\text{contrast}} = 16$

with multiplicity nine.

The corresponding singular value is

$\sigma_{\text{contrast}} = \sqrt{16} = 4$

with multiplicity nine.

B.8 Example Contrast Bases

Many orthogonal or nonorthogonal bases may be used for the three-dimensional contrast space within a positional block.

One convenient basis is

$c_1 = (1, -1, 0, 0)^T$,

$c_2 = (1, 1, -2, 0)^T$,

and

$c_3 = (1, 1, 1, -3)^T$.

Each vector has coordinate sum zero.

After normalization, these provide an orthonormal contrast basis.

Another possible basis emphasizes nucleotide pairings:

G − A;

C − U;

and

(G + A) − (C + U).

The spectral result does not depend on which contrast basis is selected. Every vector in the zero-sum subspace has the same Gram eigenvalue sixteen.

This degeneracy reflects the exact symmetry of the complete input domain. Before translated outputs are introduced, no within-position nucleotide contrast is algebraically preferred over another.

B.9 Blockwise Constant Subspace

The remaining coefficient directions are constant within each positional block.

A general blockwise constant vector has the form

$v = (a\ 1_4, b\ 1_4, c\ 1_4)^T$.

This three-dimensional subspace contains:

the global constant direction;

and

two directions describing transfers of a common offset among positional blocks.

The action of G can be calculated directly.

The first output block is

$16\ a\ 1_4 + 16\ b\ 1_4 + 16\ c\ 1_4$

$= 16 (a + b + c)\ 1_4$.

The same expression occurs in the second and third output blocks.

Therefore,

$G v$

$= 16 (a + b + c)$

$(1_4, 1_4, 1_4)^T$.

If

$a = b = c$,

the vector lies in the global constant direction and has eigenvalue forty-eight.

If

$a + b + c = 0$,

then

$G v = 0$.

The subspace satisfying

$a + b + c = 0$

has dimension two.

Therefore,

$\lambda_{null} = 0$

with multiplicity two.

A convenient basis is

$z_1 = (1_4, -1_4, 0_4)^T$,

and

$z_2 = (1_4, 0_4, -1_4)^T$.

Both satisfy

$G z_i = 0$.

Because

$G = D^T D$,

a vector lies in the null space of G exactly when it lies in the null space of D. Hence,

$D z_1 = 0$

and

$D z_2 = 0$.

B.10 Complete Eigenvalue Spectrum

The coefficient space has now been decomposed into three invariant subspaces:

the one-dimensional global constant subspace;

the nine-dimensional within-position contrast subspace;

and

the two-dimensional block-offset null subspace.

Their dimensions sum to

$1 + 9 + 2 = 12.$

Therefore, the complete eigenvalue spectrum of G is

48 with multiplicity 1;

16 with multiplicity 9;

and

0 with multiplicity 2.

The eigenvalue sum is

$48 + 9 (16) + 2 (0)$

$= 48 + 144$

$= 192.$

This agrees with

$tr (G) = 192.$

No additional eigenvalues are possible because the multiplicities already account for all twelve dimensions.

B.11 Singular-Value Spectrum

The singular values of D are the nonnegative square roots of the eigenvalues of $D^T D$.

Therefore, the complete singular spectrum is

$4 \sqrt{3}$ with multiplicity 1;

4 with multiplicity 9;

and

0 with multiplicity 2.

The squared singular values sum to

$(4 \sqrt{3})^2 + 9 (4)^2 + 2 (0)^2$

$= 48 + 144$

$= 192.$

This equals the squared Frobenius norm of D, as required.

The spectral condition number on the nonzero singular subspace is

$\kappa = \sigma_{max} / \sigma_{min,nonzero}$

$= (4 \sqrt{3}) / 4$

$= \sqrt{3}.$

Thus, after the two exact null directions are removed, the complete codon design is well-conditioned.

The difficulty is not near-singularity among the identifiable directions. It is exact nonidentifiability of two blockwise offsets.

B.12 Rank and Right Nullity

The rank of a matrix equals the number of nonzero singular values.

There are

$1 + 9 = 10$

nonzero singular values.

Therefore,

$\text{rank} (D) = 10.$

Because D has twelve columns, the right nullity is

$\text{nullity} (D)$

$= 12 - 10$

$= 2.$

The complete right null space is

$\text{null} (D)$

$=$

$\{ (a 1_4, b 1_4, c 1_4)^T : a + b + c = 0 \}.$

No within-position contrast lies in the null space. Every nonzero zero-sum contrast changes the codon-level field.

Only transfers of common block offsets can leave every codon sum unchanged.

B.13 Left Nullity

The matrix D has sixty-four rows and rank ten.

Therefore,

nullity (D^T)

= 64 − 10

= 54.

The left null space consists of codon-level vectors r satisfying

$D^T r = 0$.

Such a vector has zero sum over every nucleotide-position coordinate family:

$\Sigma c : c_p = n \; r \; c = 0$

for every position p and nucleotide n.

These twelve equations contain two redundancies, leaving ten independent constraints. Their solution space therefore has dimension fifty-four.

A codon-level property residual produced by orthogonal projection lies in this left null space:

$r = h - D \; D^+ \; h$,

and

$D^T r = 0$.

The left null space contains all codon fields that cannot be detected by the additive nucleotide-position representation.

B.14 Interpretation as a Three-Factor Additive Model

The rank formula can also be obtained from the dimension of a complete additive factorial model.

There are three codon positions.

Each position has four nucleotide levels.

A general additive field has the form

$f (c_1, c_2, c_3)$

$= \mu + \alpha_1 (c_1) + \alpha_2 (c_2) + \alpha_3 (c_3)$,

subject to one zero-sum constraint within each positional effect:

$\Sigma n\ \alpha_p\ (\ n\ ) = 0.$

The dimension is therefore

$1 + 3\ (\ 4 - 1\ )$

$= 1 + 9$

$= 10.$

The global mean contributes one dimension.

Each of the three positions contributes three independent contrasts.

This is exactly the column-space dimension of D.

The twelve-column one-hot form is an overparameterized representation of this ten-dimensional additive model.

B.15 The Coefficient-Space Orthogonal Decomposition

Every coefficient vector

$\beta \in \mathbb{R}^{12}$

can be decomposed uniquely as

$\beta = \beta_{global} + \beta_{contrast} + \beta_{null},$

where:

β_{global} lies in the global constant direction;

$\beta_{contrast}$ lies in the direct sum of the three within-position zero-sum subspaces;

and

β_{null} lies in the two-dimensional block-offset null space.

The three components are mutually orthogonal under the standard Euclidean inner product.

The matrix D acts differently on them.

On β_{global}, D produces a constant codon field.

On $\beta_{contrast}$, D produces the identifiable additive nucleotide effects.

On β_{null}, D produces zero.

Thus,

$D\,\beta = D\,\beta_{global} + D\,\beta_{contrast}$.

The null component cannot be recovered from codon-level observations because it has no effect on the fitted field.

The Moore–Penrose pseudoinverse sets this null component to zero.

B.16 The Row Space of D

The row space of D is the orthogonal complement of the right null space.

A coefficient vector β belongs to the row space exactly when it is orthogonal to both null basis vectors:

$\beta^T z_1 = 0$,

and

$\beta^T z_2 = 0$.

These conditions are equivalent to equality of the three positional block sums:

$\Sigma_n \beta_{1,n}$

$= \Sigma_n \beta_{2,n}$

$= \Sigma_n \beta_{3,n}$.

Thus, the row space consists of coefficient vectors in the balanced minimum-norm gauge.

The coefficient-space projector

$Q = D^+ D$

projects arbitrary twelve-component coefficient vectors onto this ten-dimensional row space.

B.17 Alternative Derivation Using Kronecker Structure

The complete codon matrix can also be described using repeated one-hot factors.

Let I_4 be the 4×4 identity matrix and let 1_4 be the four-component all-ones vector.

With a suitable lexicographic codon order, the three positional blocks may be written as

$D_1 = I_4 \otimes 1_4 \otimes 1_4$,

$D_2 = 1_4 \otimes I_4 \otimes 1_4$,

and

$D_3 = 1_4 \otimes 1_4 \otimes I_4$,

with dimensions interpreted so that each expression produces a 64 × 4 block.

The complete matrix is

$D = [\, D_1 \mid D_2 \mid D_3 \,]$.

The Kronecker representation makes the balance identities immediate.

For example,

$D_1^T D_1$

$= I_4^T I_4$

$\times (1_4^T 1_4)$

$\times (1_4^T 1_4)$

in the corresponding factorized sense, giving

$16\, I_4$.

Similarly,

$D_1^T D_2 = 4\, J_4$.

This alternative derivation reaches the same Gram matrix and may be useful in the general alphabet–word formulation of Appendix L.

B.18 Direct Computational Verification

A numerical or symbolic implementation should verify the following identities:

dimensions (D) = 64 × 12;

rank (D) = 10;

rank ($D^T D$) = 10;

eigenvalues ($D^T D$)

= { 48, 16 repeated nine times, 0 repeated twice };

singular values (D)

= { 4 √3, 4 repeated nine times, 0 repeated twice };

nullity (D) = 2;

and

nullity (D^T) = 54.

When floating-point arithmetic is used, the two zero singular values may appear as small numerical residues. They should not be interpreted as small nonzero singular values.

Exact integer or rational arithmetic confirms that they are identically zero.

A reliable implementation should derive the rank symbolically or use a tolerance justified by the exact spectrum.

B.19 Relation to Figure 2

Figure 2 displays the twelve singular values in descending order.

The spectrum contains three visually distinct levels:

one value at

4 √3 ≈ 6.928203;

nine values at

4;

and

two values at

1.

The figure is a visualization of the exact derivation, not the source of the result.

The rank should not be inferred merely by visually counting bars above a plotting threshold. It follows algebraically from the Gram matrix.

No additional figure is required in this appendix. Figure 2 remains positioned in Chapter 2 after the rank-and-nullity derivation.

B.20 Summary

The complete codon incidence matrix has the exact Gram matrix

$D^T D$

=

$$\begin{bmatrix} 16\,I_4 & 4\,J_4 & 4\,J_4 \\ 4\,J_4 & 16\,I_4 & 4\,J_4 \\ 4\,J_4 & 4\,J_4 & 16\,I_4 \end{bmatrix}.$$

Its coefficient space decomposes into:

one global constant direction with eigenvalue 48;

nine within-position contrast directions with eigenvalue 16;

and

two blockwise constant null directions with eigenvalue 0.

Therefore,

rank (D) = 10,

nullity (D) = 2,

and

nullity (D^T) = 54.

The singular values are

4 √3 once,

4 nine times,

and

0 twice.

The nonzero condition number is √3.

The two-dimensional null space consists exactly of transfers of common offsets among the three positional blocks whose total shift is zero.

These results provide the spectral foundation for the exact Moore–Penrose pseudoinverse derived in Appendix C.

Appendix C

Exact Moore–Penrose Pseudoinverse

C.1 Purpose

The complete codon incidence matrix

$D \in \{ 0, 1 \}^{64 \times 12}$

has rank ten rather than twelve. Its ordinary inverse does not exist, and the Gram matrix

$D^T D$

is singular.

The Moore–Penrose pseudoinverse nevertheless exists uniquely. For the complete codon matrix, it can be written in a simple exact form:

$D^+ = (1 / 16) D^T - (1 / 96) J_{12 \times 64}$.

The entire 12 × 64 pseudoinverse contains only two rational values:

5 / 96

and

−1 / 96.

This appendix derives the formula, verifies all four Moore–Penrose conditions, and records exact implementation checks.

C.2 Spectral Construction

Appendix B decomposed the twelve-dimensional coefficient space into three mutually orthogonal subspaces:

the one-dimensional global constant subspace;

the nine-dimensional within-position contrast subspace;

and

the two-dimensional block-offset null space.

On these subspaces, the Gram matrix

$G = D^T D$

acts by multiplication by:

48 on the global constant direction;

16 on the nine contrast directions;

and

0 on the two null directions.

The pseudoinverse of G therefore acts by:

1 / 48 on the global constant direction;

1 / 16 on the contrast directions;

and

0 on the null directions.

The Moore–Penrose pseudoinverse of D can be written as

$D^{+} = G^{+} D^{T}$.

The complete symmetry of the design allows this spectral construction to collapse to a two-term expression involving only D^{T} and the all-ones matrix.

C.3 Closed-Form Formula

The exact pseudoinverse is

$D^{+} = (1 / 16) D^{T} - (1 / 96) J_{12 \times 64}$.

The first term,

$(1 / 16) D^{T}$,

averages a codon-level field over the sixteen codons containing each nucleotide-position coordinate.

The second term,

$(1 / 96) J_{12 \times 64}$,

subtracts a uniform component from every coefficient.

This correction is required because the global constant field is represented redundantly in all three positional blocks.

For a codon-level field

$h \in \mathbb{R}^{64}$,

the coefficient vector is

$\beta = D^{+} h.$

For coordinate (p, n),

$\beta p,n$

$= (1 / 16) \Sigma c : c_p = n \; h \; c$

$- (1 / 96) \Sigma c \in C \; h \; c.$

Let the global codon-field mean be

$\mu = (1 / 64) \Sigma c \in C \; h \; c.$

Then

$(1 / 96) \Sigma c \in C \; h \; c$

$= (2 / 3) \mu.$

Therefore,

$\beta p,n$

= mean of h among the sixteen codons containing nucleotide n at position p

$- (2 / 3) \mu.$

The formula distributes the global mean equally among the three positional blocks while retaining every identifiable within-position contrast.

C.4 Verification by Entry Values

Because D contains only 0s and 1s, the pseudoinverse formula produces only two possible values.

For a codon c containing coordinate (p, n),

$D^{T} p,n;c = 1.$

Therefore,

$D^{+} p,n;c$

$= 1 / 16 - 1 / 96$

$= 6 / 96 - 1 / 96$

$= 5 / 96.$

For a codon c not containing coordinate (p, n),

$D^{T}p,n;c = 0.$

Therefore,

$D^{+}p,n;c$

$= -1 / 96.$

Thus,

$D^{+}p,n;c = 5 / 96$

when

$c_{p} = n,$

and

$D^{+}p,n;c = -1 / 96$

when

$c_{p} \neq n.$

The result may also be written compactly as

$D^{+}p,n;c$

$= [6 D c,(p,n) - 1] / 96.$

Every pseudoinverse entry is therefore determined directly by whether the corresponding nucleotide-position coordinate is present in the codon.

C.5 Exact Matrix Composition

The pseudoinverse has dimensions

$12 \times 64,$

so it contains

$12 \times 64 = 768$

entries.

Each of the sixty-four codon columns contains three present coordinates and nine absent coordinates.

Therefore, every column contains:

three entries equal to 5 / 96;

and

nine entries equal to −1 / 96.

The total number of positive entries is

64 × 3 = 192.

The total number of negative entries is

64 × 9 = 576.

The same counts follow from the rows.

Each nucleotide-position coordinate occurs in sixteen codons and is absent from forty-eight.

Every row therefore contains:

sixteen entries equal to 5 / 96;

and

forty-eight entries equal to −1 / 96.

Across twelve rows, the positive count is

12 × 16 = 192,

and the negative count is

12 × 48 = 576.

These counts exactly parallel the composition of D itself:

192 entries equal to 1;

and

576 entries equal to 0.

The pseudoinverse converts the two binary incidence states into two signed rational values.

No additional figure is required in this appendix. The complete two-value matrix is displayed as Figure 3 in Chapter 3.

C.6 Direct Calculation of $D^{+} D$

Define the coefficient-space matrix

$Q = D^{+} D.$

Substitute the closed-form pseudoinverse:

Q

$= [(1 / 16) D^{T} - (1 / 96) J_{12 \times 64}] D$

$= (1 / 16) D^{T} D$

$- (1 / 96) J_{12 \times 64} D.$

Every column of D contains sixteen 1s. Therefore,

$J_{12 \times 64} D = 16 J_{12}.$

Hence,

$Q = (1 / 16) G - (1 / 6) J_{12},$

where

$G = D^{T} D.$

Using the block form of G, the diagonal 4×4 blocks of Q are

$I_4 - (1 / 6) J_4.$

The off-diagonal blocks are

$(1 / 4) J_4 - (1 / 6) J_4$

$= (1 / 12) J_4.$

Therefore,

$Q =$

$$\begin{bmatrix} I_4 - J_4 / 6 & J_4 / 12 & J_4 / 12 \\ J_4 / 12 & I_4 - J_4 / 6 & J_4 / 12 \\ J_4 / 12 & J_4 / 12 & I_4 - J_4 / 6 \end{bmatrix}.$$

This matrix is symmetric.

It acts as the identity on the global constant direction.

It acts as the identity on all nine within-position contrast directions.

It acts as zero on the two block-offset null directions.

Therefore, Q is the orthogonal projector onto the ten-dimensional row space of D.

It satisfies

$Q^2 = Q$

and

$Q^T = Q$.

Its eigenvalues are:

1 with multiplicity ten;

and

0 with multiplicity two.

Consequently,

rank (Q) = tr (Q) = 10.

C.7 Direct Calculation of D D^+

Define the codon-space matrix

$P = D\ D^+$.

Substituting the exact formula gives

P

$= D\ [\ (\ 1 / 16\ )\ D^T - (\ 1 / 96\ )\ J_{12\times64}\ ]$

$= (\ 1 / 16\ )\ D\ D^T$

$-\ (\ 1 / 96\ )\ D\ J_{12\times64}$.

Every row of D contains three 1s. Therefore,

$D\ J_{12\times64} = 3\ J_{64}$.

Hence,

$P = (\ 1 / 16\ )\ D\ D^T - (\ 1 / 32\ )\ J_{64}$.

The matrix P is symmetric because both $D\ D^T$ and J_{64} are symmetric:

$P^T = P$.

It is the orthogonal projector from the sixty-four-dimensional codon-field space onto the ten-dimensional additive nucleotide-position subspace.

Therefore,

$P^2 = P$.

Its eigenvalues are:

1 with multiplicity ten;

and

0 with multiplicity fifty-four.

Consequently,

rank (P) = tr (P) = 10.

For any codon-level field h,

P h

is the unique closest additive field.

The complementary projector

$I_{64} - P$

maps h onto the fifty-four-dimensional residual subspace.

C.8 Verification of D D^+ D = D

Because $Q = D^+ D$ is the orthogonal projector onto the row space of D, it acts as the identity on every row-space vector.

The rows of D lie in that row space. Therefore,

D Q = D.

Substituting Q gives

D D^+ D = D.

This verifies the first Moore–Penrose condition.

The same result can be checked directly from the exact formula:

D D^+ D

= P D.

Every column of D lies in the column space onto which P projects, so

$P\,D = D$.

C.9 Verification of $D^+ D D^+ = D^+$

Because $P = D\,D^+$ is the orthogonal projector onto the column space of D, and the action of D^+ depends only on that projected component,

$D^+ P = D^+$.

Therefore,

$D^+ D D^+ = D^+$.

This verifies the second Moore–Penrose condition.

Equivalently, since Q projects onto the row space containing the columns of D^+,

$Q\,D^+ = D^+$.

Thus,

$D^+ D D^+ = D^+$.

C.10 Verification of the Symmetry Conditions

The third Moore–Penrose condition is

$(D\,D^+)^T = D\,D^+$.

But

$D\,D^+ = P$,

and P is symmetric. Therefore,

$(D\,D^+)^T = D\,D^+$.

The fourth condition is

$(D^+ D)^T = D^+ D$.

But

$D^+ D = Q$,

and Q is symmetric. Therefore,

$(D^+ D)^T = D^+ D$.

All four Moore–Penrose conditions are satisfied:

$D\,D^{+}\,D = D$;

$D^{+}\,D\,D^{+} = D^{+}$;

$(D\,D^{+})^{T} = D\,D^{+}$;

and

$(D^{+}\,D)^{T} = D^{+}\,D$.

The closed-form matrix is therefore the unique Moore–Penrose pseudoinverse of D.

C.11 Row and Column Sums

A pseudoinverse row contains sixteen entries equal to 5 / 96 and forty-eight entries equal to −1 / 96.

Its sum is

$16\,(5/96) + 48\,(-1/96)$

$= 80/96 - 48/96$

$= 32/96$

$= 1/3$.

Therefore,

$D^{+}\,1_{64} = (1/3)\,1_{12}$.

A pseudoinverse column contains three entries equal to 5 / 96 and nine entries equal to −1 / 96.

Its sum is

$3\,(5/96) + 9\,(-1/96)$

$= 15/96 - 9/96$

$= 6/96$

$= 1/16$.

Therefore,

$1_{12}^{T}\,D^{+} = (1/16)\,1_{64}^{T}$.

These identities provide simple exact tests of any generated pseudoinverse.

C.12 Row and Column Norms

Every pseudoinverse row has the same squared Euclidean norm:

$16 (5 / 96)^2 + 48 (1 / 96)^2$

$= 16 (25 / 9{,}216) + 48 / 9{,}216$

$= 448 / 9{,}216$

$= 7 / 144.$

Thus,

$\| \text{row} (D^+) \|^2 = 7 / 144.$

Every pseudoinverse column has the same squared norm:

$3 (5 / 96)^2 + 9 (1 / 96)^2$

$= 75 / 9{,}216 + 9 / 9{,}216$

$= 84 / 9{,}216$

$= 7 / 768.$

Thus,

$\| \text{column} (D^+) \|^2 = 7 / 768.$

The equality of all row norms and all column norms is another consequence of the complete factorial balance.

C.13 Constant-Field Reconstruction

Let the codon field be constant:

$h = k \, 1_{64}.$

Then

$\beta = D^+ h$

$= k \, D^+ \, 1_{64}$

$= (k / 3) \, 1_{12}.$

Every nucleotide-position coefficient equals k / 3.

Each codon activates exactly three coordinates, so

$D\,\beta$

$= D\,[\,(\,k / 3\,)\,1_{12}\,]$

$= (\,k / 3\,)\,D\,1_{12}$

$= (\,k / 3\,)\,(\,3\,1_{64}\,)$

$= k\,1_{64}.$

Thus,

$D\,D^{+}\,h = h.$

The complete constant codon field lies exactly in the additive subspace.

This calculation also explains why the correction term in D^{+} is required. The simple coordinate average

$(\,1 / 16\,)\,D^{T}\,h$

would assign k to all twelve coefficients, causing every codon to reconstruct to 3 k. The subtraction term reduces each coefficient to k / 3.

C.14 Pure Positional Fields

Suppose a codon field depends only on the nucleotide at one codon position.

For example, let

$h\,c_1c_2c_3 = f\,(\,c_2\,)$

for some function

$f : \Sigma \to \mathbb{R}.$

This field can be represented using only the four second-position indicator columns of D.

Therefore,

$h \in \mathrm{col}\,(\,D\,).$

Because P is the orthogonal projector onto col (D),

$P\,h = h.$

Thus,

$D\,D^{+}\,h = h.$

The same result holds for any field of the form

$h\ c_1c_2c_3$

$= f_1 (c_1) + f_2 (c_2) + f_3 (c_3).$

Every purely additive positional field is reconstructed exactly.

Residuals arise only from components that cannot be written as the sum of independent effects from the three codon positions.

C.15 Pairwise Interpretation of the Codon Projector

The entries of

$D\ D^T$

have a direct combinatorial meaning.

Let c and d be two codons. Their incidence rows each contain three 1s.

The inner product

$(D\ D^T)c,d$

counts the number of nucleotide-position coordinates shared by the two codons.

Let

$m (c, d)$

be the number of codon positions at which c and d contain the same nucleotide.

Then

$(D\ D^T)c,d = m (c, d).$

Because codons have length three,

$m (c, d) = 3 − dH (c, d),$

where dH is Hamming distance.

Therefore,

Pc,d

$= m (c, d) / 16 − 1 / 32$

$= [3 − dH (c, d)] / 16 − 1 / 32.$

Simplifying,

$$P_{c,d}$$

$$= 5 / 32 - dH (c, d) / 16.$$

The possible values are:

$$dH (c, d) = 0:$$

$$P_{c,d} = 5 / 32;$$

$$dH (c, d) = 1:$$

$$P_{c,d} = 3 / 32;$$

$$dH (c, d) = 2:$$

$$P_{c,d} = 1 / 32;$$

$$dH (c, d) = 3:$$

$$P_{c,d} = -1 / 32.$$

The codon-space projector therefore contains only four rational values.

The additive reconstruction of one codon is a signed weighted combination of the complete codon field, with weights determined only by the number of nucleotide positions shared with each other codon.

C.16 Diagonal Leverage

For a codon c,

$$dH (c, c) = 0.$$

Therefore,

$$P_{c,c} = 5 / 32.$$

Every codon has identical diagonal leverage.

The trace is

$$tr (P)$$

$$= 64 (5 / 32)$$

$$= 10.$$

This equals the rank of D, as required for an orthogonal projector.

The uniform leverage means that no codon is algebraically privileged by the complete incidence design. Differences in biological influence arise from the output field or additional weighting, not from unequal representation in D.

C.17 Row Sums of the Codon Projector

Because constant fields are reconstructed exactly,

$P\ 1_{64} = 1_{64}$.

Every row of P therefore sums to one.

This can also be verified from the Hamming-distance distribution around one codon.

Relative to any fixed codon, there are:

1 codon at distance 0;

9 codons at distance 1;

27 codons at distance 2;

and

27 codons at distance 3.

The row sum is

$1\ (5 / 32)$

- $9\ (3 / 32)$
- $27\ (1 / 32)$
- $27\ (-1 / 32)$

$= 5 / 32 + 27 / 32 + 27 / 32 - 27 / 32$

$= 32 / 32$

$= 1$.

This supplies another exact implementation check.

C.18 Minimum-Norm Coefficients

Suppose several coefficient vectors produce the same additive codon field:

$D\ \beta = D\ \gamma$.

Then

$D (\beta - \gamma) = 0.$

Therefore,

$\beta - \gamma \in \text{null} (D).$

All coefficient representations of one fitted field differ by a vector in the two-dimensional block-offset null space.

The Moore–Penrose coefficient vector

$\beta_{MP} = D^{+} h$

lies in the row space of D, which is orthogonal to null (D).

For any other solution

$\beta_{MP} + z,$

with

$z \in \text{null} (D),$

the squared norm is

$\| \beta_{MP} + z \|^2$

$= \| \beta_{MP} \|^2 + \| z \|^2,$

because

$\beta_{MP}^{T} z = 0.$

Therefore,

$\| \beta_{MP} + z \|^2 \geq \| \beta_{MP} \|^2,$

with equality only when

$z = 0.$

The pseudoinverse solution is consequently the unique minimum-norm coefficient representation.

C.19 Equality of Positional Block Sums

A coefficient vector lies in the row space of D exactly when it is orthogonal to the two null directions

$z_1 = (1_4, -1_4, 0_4)^{T}$

and

$z_2 = (1_4, 0_4, -1_4)^T$.

Orthogonality to z_1 requires

$\Sigma n\ \beta 1,n = \Sigma n\ \beta 2,n$.

Orthogonality to z_2 requires

$\Sigma n\ \beta 1,n = \Sigma n\ \beta 3,n$.

Therefore, the Moore–Penrose coefficients satisfy

$\Sigma n\ \beta 1,n$

$= \Sigma n\ \beta 2,n$

$= \Sigma n\ \beta 3,n$.

Let

$S = \Sigma c \in C\ h\ c$.

Using the exact coefficient formula, the sum of one four-coordinate positional block is

$S / 16 - 4\ S / 96$

$= S / 16 - S / 24$

$= S / 48$.

Thus, every positional block sum equals

$S / 48$.

If the codon field is centered, then

$S = 0$,

and all three positional blocks sum to zero.

C.20 Exact Rational Arithmetic

The entire derivation can be performed using integers and rational numbers.

No floating-point tolerance is required to verify:

$D\ D^+\ D = D$;

$D^+\ D\ D^+ = D^+$;

$(D D^{+})^{T} = D D^{+}$;

$(D^{+} D)^{T} = D^{+} D$;

$P^{2} = P$;

$Q^{2} = Q$;

rank (P) = 10;

rank (Q) = 10;

and the exact row and column sums.

A reproducible implementation should preserve the fractions

5 / 96,

−1 / 96,

5 / 32,

3 / 32,

1 / 32,

and

−1 / 32

rather than replacing them immediately with floating-point approximations.

Floating-point values may be produced for display, but the verification should be performed in exact arithmetic whenever the software permits it.

C.21 Fail-Closed Verification Conditions

A correct computational implementation should terminate with an error if any of the following conditions fails:

D^{+} has dimensions 12 × 64;

every entry of D^{+} belongs to { 5 / 96, −1 / 96 };

every row contains sixteen positive and forty-eight negative entries;

every column contains three positive and nine negative entries;

every row sum equals 1 / 3;

every column sum equals 1 / 16;

$D\,D^{+}\,D = D$;

$D^{+}\,D\,D^{+} = D^{+}$;

$D\,D^{+}$ is symmetric;

$D^{+}\,D$ is symmetric;

$\mathrm{tr}\,(D\,D^{+}) = 10$;

and

$\mathrm{tr}\,(D^{+}\,D) = 10$.

These conditions are redundant by design. Redundant exact checks make implementation errors easier to detect.

C.22 Summary

The complete codon incidence matrix has the exact Moore–Penrose pseudoinverse

$D^{+} = (1/16)\,D^{T} - (1/96)\,J_{12\times 64}$.

Its entry rule is

5 / 96 when the coordinate is present;

and

−1 / 96 when the coordinate is absent.

The coefficient-space projector

$D^{+}\,D$

has rank ten and removes exactly the two block-offset null directions.

The codon-space projector

$D\,D^{+}$

has rank ten and maps arbitrary sixty-four-component fields onto the additive nucleotide-position subspace.

Its entries depend only on codon Hamming distance.

Every additive positional field is reconstructed exactly.

Every nonadditive component is placed in the fifty-four-dimensional orthogonal residual space.

The pseudoinverse therefore supplies an exact bridge between complete codon space and additive nucleotide-position representation. Appendix D develops the gauge freedom, projection geometry, and minimum-norm interpretation in greater detail.

Appendix D

Projection, Gauge Freedom, and Minimum Norm

D.1 Additive Codon Fields

A coefficient vector

$\beta \in \mathbb{R}^{12}$

assigns one numerical value to each nucleotide-position coordinate:

1G, 1A, 1C, 1U;

2G, 2A, 2C, 2U;

and

3G, 3A, 3C, 3U.

For a codon

$c = c_1 c_2 c_3$,

the corresponding additive codon value is

$f\beta (c)$

$= \beta 1,c_1 + \beta 2,c_2 + \beta 3,c_3$.

In matrix notation, the complete sixty-four-component field is

$f\beta = D \beta$.

The column space of D is therefore precisely the set of codon fields that can be written as the sum of three independent positional contributions:

one contribution from the first nucleotide;

one from the second nucleotide;

and one from the third nucleotide.

No pairwise interaction between codon positions is included. There is no special term for the joint occurrence of first-position G and second-position U, for example. There is also no three-position term that depends on the complete triplet beyond the sum of its individual coordinates.

The additive representation has twelve displayed coefficients but only ten independent dimensions. The difference between twelve displayed coordinates and rank ten is the source of gauge freedom.

D.2 Nonuniqueness of Coefficients

Suppose two coefficient vectors β and γ produce the same codon field:

$D \beta = D \gamma.$

Then

$D (\beta - \gamma) = 0.$

Therefore,

$\beta - \gamma \in \text{null} (D).$

Appendix B established that the right null space has dimension two. Consequently, every additive codon field has a two-parameter family of coefficient representations.

The complete null space is

$\text{null} (D)$

$=$

$\{ (a 1_4, b 1_4, c 1_4)^T : a + b + c = 0 \}.$

Thus, a null vector is constant within each positional block, with the three block offsets summing to zero.

A convenient parameterization is

$z (a, b)$

$=$

$(a 1_4, b 1_4, -(a + b) 1_4)^T.$

For every choice of a and b,

$D z (a, b) = 0.$

Therefore,

$D [\beta + z (a, b)] = D \beta.$

The reconstructed codon field is unique, but its unrestricted twelve-coordinate description is not.

D.3 Gauge Transformations

A gauge transformation changes the coefficient representation without changing any reconstructed codon value.

Let

$z = (a\, 1_4, b\, 1_4, c\, 1_4)^T,$

with

$a + b + c = 0.$

Define

$\beta' = \beta + z.$

For codon $c_1 c_2 c_3$,

$f\beta' (c_1 c_2 c_3)$

$= \beta 1, c_1 + a$

- $\beta 2, c_2 + b$
- $\beta 3, c_3 + c.$

Therefore,

$f\beta' (c_1 c_2 c_3)$

$= f\beta (c_1 c_2 c_3) + a + b + c$

$= f\beta (c_1 c_2 c_3).$

Every codon receives the same total gauge shift, and that shift is zero.

One elementary example transfers a constant δ from the second positional block to the first:

$\beta_1' = \beta_1 + \delta\, 1_4,$

$\beta_2' = \beta_2 - \delta\, 1_4,$

$\beta_3' = \beta_3.$

Every codon activates one coefficient from the first block and one from the second. The two changes cancel.

A second independent transformation can transfer a constant between the first and third blocks, or between the second and third blocks. These two independent transfers correspond exactly to the two-dimensional null space.

The term **gauge freedom** is appropriate because multiple coordinate descriptions represent the same observable codon field.

D.4 Absolute Coefficient Levels

Gauge freedom limits the interpretation of absolute coefficient values.

Suppose one coefficient is reported as

$\beta_{2,U} = 3.948958.$

That value is defined within a particular coefficient convention. A gauge transformation adding δ to every second-position coefficient would change it to

$\beta_{2,U'} = 3.948958 + \delta$

without changing any reconstructed codon hydropathy.

The absolute zero level of one positional block is therefore not independently identifiable from the codon field.

What is identifiable is the combined three-position sum for every codon.

The coefficient vector becomes unique only after a gauge convention is imposed. The Moore–Penrose pseudoinverse supplies one such convention: the unique minimum-norm gauge.

Absolute coefficients should therefore be interpreted as coordinates within that declared representation, not as independently measurable physical contributions detached from it.

D.5 Gauge-Invariant Contrasts

Differences within the same positional block are invariant under gauge transformations.

Consider two second-position nucleotides x and y. Under a gauge transformation,

$\beta_{2,x'} = \beta_{2,x} + b$

and

$\beta_{2,y'} = \beta_{2,y} + b.$

Therefore,

$\beta_{2,x'} - \beta_{2,y'}$

$= \beta_{2,x} - \beta_{2,y}.$

The common block offset cancels.

The same result holds for every within-position contrast.

Examples include:

$\beta_{1,G} - \beta_{1,C};$

$\beta_{2,U} - \beta_{2,A};$

and

$\beta_{3,C} - \beta_{3,G}.$

Such contrasts are identifiable properties of the additive codon field. They do not depend on how constant offsets are distributed among the three positional blocks.

The principal complete-code hydropathy contrast is

$\beta_{2,U} - \beta_{2,A}$

$= 3.948958 - (-2.638542)$

$= 6.587500.$

This value is gauge invariant.

The strong opposition between second-position U and second-position A is therefore not an artifact of the Moore–Penrose distribution of block offsets.

Positional ranges are also gauge invariant. Adding the same constant to all four coefficients in one block changes neither its maximum-minus-minimum range nor the ordering of its four nucleotide effects.

D.6 The Moore–Penrose Gauge

The Moore–Penrose pseudoinverse chooses the coefficient vector

$\beta_{MP} = D^{+} h.$

This vector lies in the row space of D.

The row space is the orthogonal complement of the right null space:

row (D) = null (D)⊥.

Therefore,

$\beta M P^T z = 0$

for every

$z \in \text{null}(D)$.

Using the null basis

$z_1 = (1_4, -1_4, 0_4)^T$

and

$z_2 = (1_4, 0_4, -1_4)^T$,

orthogonality gives

$\beta M P^T z_1 = 0$

and

$\beta M P^T z_2 = 0$.

The first condition implies

$\Sigma n \in \Sigma \beta 1,n$

$= \Sigma n \in \Sigma \beta 2,n$.

The second implies

$\Sigma n \in \Sigma \beta 1,n$

$= \Sigma n \in \Sigma \beta 3,n$.

Hence,

$\Sigma n \in \Sigma \beta 1,n$

$= \Sigma n \in \Sigma \beta 2,n$

$= \Sigma n \in \Sigma \beta 3,n$.

The three positional coefficient blocks have equal sums.

This is the symmetric Moore–Penrose gauge.

It distributes the shared constant component equally among the three otherwise redundant positional representations.

D.7 Exact Positional Block Sums

Let the total of the codon-level field be

$S = \Sigma c \in C \; h \; c.$

The exact coefficient formula is

$\beta p,n$

$= (1 / 16) \Sigma c : c_p = n \; h \; c$

$- (1 / 96) \; S.$

Sum over the four nucleotides n within one fixed codon position p.

The four coordinate-specific sums partition all sixty-four codons, so

$\Sigma n \in \Sigma \; \Sigma c : c_p = n \; h \; c = S.$

Therefore,

$\Sigma n \in \Sigma \; \beta p,n$

$= S / 16 - 4 \; S / 96$

$= S / 16 - S / 24$

$= S / 48.$

Thus, every positional block sum is exactly

$S / 48.$

Let the global codon-field mean be

$\mu = S / 64.$

Then the common block sum is

$S / 48 = 4 \; \mu / 3.$

The sum of all twelve coefficients is

$3 \; S / 48$

$= S / 16$

$= 4\,\mu$.

If the codon field is centered, then

$S = 0$,

and every positional block has sum zero.

The Moore–Penrose gauge therefore reduces to the familiar zero-sum effect convention for centered fields.

D.8 Minimum-Norm Property

Let

$\beta_{MP} = D^{+} h$

be the Moore–Penrose least-squares solution.

Every other coefficient vector producing the same fitted codon field has the form

$\beta = \beta_{MP} + z$,

where

$z \in \text{null}(D)$.

Because β_{MP} belongs to the row space of D and z belongs to its orthogonal complement,

$\beta_{MP}^{T} z = 0$.

Therefore,

$\| \beta \|^2$

$= \| \beta_{MP} + z \|^2$

$= \| \beta_{MP} \|^2 + 2\, \beta_{MP}^{T} z + \| z \|^2$

$= \| \beta_{MP} \|^2 + \| z \|^2$.

Hence,

$\| \beta \|^2 \geq \| \beta_{MP} \|^2$.

Equality holds only when

$z = 0$.

The Moore–Penrose coefficient vector is therefore the unique representation of minimum Euclidean norm.

This statement applies whether the original codon field is exactly additive or only approximately additive.

If h lies in the column space of D, βMP is the minimum-norm exact solution of

$D \beta = h$.

If h does not lie in the column space, βMP is the minimum-norm member of the family of coefficient vectors producing the unique least-squares projection.

D.9 Orthogonal Projection

For any codon-level numerical field

$h \in \mathbb{R}^{64}$,

define its additive reconstruction by

$\hat{h} = D D^{+} h$.

Let

$P = D D^{+}$.

Then

$\hat{h} = P h$.

The matrix P is the orthogonal projector onto the column space of D.

It satisfies

$P^2 = P$

and

$P^T = P$.

Therefore,

$\hat{h} \in \text{col}(D)$.

The residual is

$r = h - \hat{h}$

$= (I_{64} - P) h$.

Because $I_{64} - P$ is the orthogonal projector onto the complement of col (D),

$r \in \text{col}(D)\perp$.

Thus,

$D^T r = 0$.

The residual has zero inner product with every additive nucleotide-position field.

For every coefficient vector γ,

$(D \gamma)^T r = 0$.

No further adjustment of the same additive coordinates can reduce the residual.

D.10 Normal Equations

The orthogonality condition gives the least-squares normal equations.

Because

$D^T r = 0$,

and

$r = h - D \beta MP$,

we have

$D^T (h - D \beta MP) = 0$.

Therefore,

$D^T D \beta MP = D^T h$.

The Gram matrix $D^T D$ is singular, so the normal equations do not determine one unrestricted coefficient vector. They determine an affine family differing by null-space vectors.

The Moore–Penrose condition selects from that family the unique vector orthogonal to the null space.

Thus, the pseudoinverse solves two related problems:

it satisfies the normal equations required for least squares;

and it resolves their coefficient nonuniqueness by imposing minimum norm.

D.11 Pythagorean Decomposition

The fitted field and residual are orthogonal:

$\hat{h}^T r = 0$.

Because

$h = \hat{h} + r,$

the squared norm separates exactly:

$\| h \|^2$

$= \| \hat{h} + r \|^2$

$= \| \hat{h} \|^2 + 2 \hat{h}^T r + \| r \|^2$

$= \| \hat{h} \|^2 + \| r \|^2.$

Therefore,

$\| h \|^2 = \| P h \|^2 + \| (I_{64} - P) h \|^2.$

The total squared magnitude of any codon field is divided exactly into:

an additive nucleotide-position component;

and

an orthogonal nonadditive component.

This decomposition is exact even when the field is not derived from amino-acid properties.

For variance interpretation, however, the global mean should first be removed.

D.12 Centered Property Fields

Let

$\mu = (1 / 64) 1_{64}^T h$

be the codon-level mean.

Define the centered field

$h^\circ = h - \mu 1_{64}.$

Because the constant codon field belongs to col (D),

$P 1_{64} = 1_{64}.$

Therefore,

$P h^\circ$

$= P h - \mu 1_{64}.$

The projection preserves the mean and operates independently on the centered variation.

The centered Pythagorean decomposition is

$\| h^\circ \|^2$

$= \| P h^\circ \|^2$

- $\| (I_{64} - P) h^\circ \|^2$.

A natural additive projection fraction is

Qadd

$= \| P h^\circ \|^2 / \| h^\circ \|^2$,

provided that h is not constant.

Equivalently,

Qadd

$= 1 - \| (I_{64} - P) h^\circ \|^2 / \| h^\circ \|^2$.

The quantity Qadd measures the fraction of centered squared codon-level variation captured by independent nucleotide-position effects.

For the complete Kyte–Doolittle codon field with

$H (X) = 0$,

the value reported in Chapter 5 is approximately

Qadd = 0.816882.

Thus, approximately 81.7 percent of the centered codon-level squared variation lies within the ten-dimensional additive subspace.

D.13 Codon-Level Projection and Regression R^2

The additive projection fraction Qadd should not be confused with the R^2 of the amino-acid-level regression.

The codon-level projection uses all sixty-four codons as separate observations. An amino acid encoded by six codons contributes six entries to h. An amino acid encoded by one codon contributes one.

The amino-acid-level analysis first averages reconstructed codon values within each translated fiber. It then compares twenty reconstructed amino-acid values with twenty empirical values, usually giving every amino acid equal weight.

These procedures use different observational units and different weighting schemes.

The codon-level value Qadd answers:

What fraction of centered squared variation across sixty-four codons is captured by the additive nucleotide-position subspace?

The amino-acid-level R^2 answers:

How strongly do twenty fiber-averaged reconstructed amino-acid values align linearly with the twenty empirical amino-acid values?

The two measures may be numerically similar, but they are not interchangeable.

Every reported fit statistic should identify whether codons or translated classes are being weighted equally.

D.14 Interaction Space

The orthogonal complement of the additive codon space has dimension

$64 - 10 = 54$.

This residual space contains every codon field that cannot be represented as

$f_1 (c_1) + f_2 (c_2) + f_3 (c_3)$.

In the language of a complete three-factor model with four levels at each factor, the sixty-four-dimensional field space decomposes into:

one global constant dimension;

nine main-effect dimensions;

twenty-seven pairwise-interaction dimensions;

and

twenty-seven three-way-interaction dimensions.

The count is

1

- $3 (4 - 1)$
- $3 (4 - 1)^2$
- $(4 - 1)^3$

= 1 + 9 + 27 + 27

= 64.

The additive space contains the constant and main effects:

1 + 9 = 10.

The residual space contains the pairwise and three-way interactions:

27 + 27 = 54.

This decomposition does not imply that every biological residual has an identifiable mechanistic interaction of one specific type. It describes the mathematical degrees of freedom outside the additive model.

A particular property field may have a small residual despite the large dimension of the residual space. Alignment depends on the direction of the field, not merely on the number of available dimensions.

D.15 Pairwise Interaction Examples

A pairwise interaction is a codon-field component that depends jointly on two positions and cannot be separated into independent contributions.

For example, suppose a field assigns a special increment only when:

the first nucleotide is G;

and

the second nucleotide is U.

Such a term depends on the coordinate pair

(1G, 2U).

It cannot generally be represented as the sum of one first-position effect and one second-position effect without also affecting other codons containing 1G or 2U separately.

Analogous interactions may occur between:

positions 1 and 2;

positions 1 and 3;

or

positions 2 and 3.

Each position pair contains

$(4 - 1) (4 - 1) = 9$

independent interaction contrasts.

With three pairs of codon positions, the complete pairwise-interaction dimension is

$3 \times 9 = 27.$

These components remain in the residual after additive projection.

D.16 Three-Way Interaction

A three-way interaction depends on the full triplet in a manner not reducible to:

a constant;

three main positional effects;

or

three sets of pairwise interactions.

Its dimension is

$(4 - 1)^3 = 27.$

Such a component can distinguish specific codon combinations even after all lower-order positional patterns have been accounted for.

A fully saturated codon model contains all sixty-four dimensions and can reconstruct every possible codon field exactly.

The additive model deliberately uses only ten independent dimensions. Its value lies in compression and interpretability rather than unrestricted exact fitting.

D.17 Exact Reconstruction Criterion

A codon-level field h is reconstructed exactly by the additive model if and only if

$P h = h.$

Equivalently,

$(I_{64} - P) h = 0.$

Equivalently,

$h \in \mathrm{col} (D).$

Equivalently, there exist functions

$f_1 : \Sigma \to \mathbb{R}$,

$f_2 : \Sigma \to \mathbb{R}$,

and

$f_3 : \Sigma \to \mathbb{R}$

such that

$h\, c_1 c_2 c_3$

$= f_1(c_1) + f_2(c_2) + f_3(c_3)$

for every codon.

An exact reconstruction does not require the twelve displayed coefficients to be unique. It requires only that the codon field belong to the ten-dimensional additive subspace.

The Moore–Penrose pseudoinverse then selects the unique minimum-norm coefficient representation of that exact field.

Any departure from additive separability contributes to the residual.

D.18 An Equivalent Interaction Test

The additive criterion can be tested through rectangular contrasts.

Fix the third codon position and choose two first-position nucleotides a and a′ together with two second-position nucleotides b and b′.

For an additive field,

$h\, a\, b\, c - h\, a'\, b\, c$

$- h\, a\, b'\, c + h\, a'\, b'\, c$

$= 0.$

The first-position and second-position main effects cancel around the rectangle.

Analogous zero rectangular contrasts must hold for every pair of codon positions and every fixed value of the remaining position.

Nonzero rectangular contrasts reveal pairwise or higher-order interaction structure.

Additional consistency conditions separate pure pairwise interactions from three-way interactions.

The projector test

$$P\,h = h$$

is the most compact complete criterion, while the contrast formulation gives a local interpretation of nonadditivity.

D.19 Property Fields Derived from Translation

Suppose a numerical property is assigned to translated outputs:

$$H : \Omega \to \mathbb{R}.$$

The induced codon field is

$$h\,c = H(\tau(c)).$$

Before projection, h is constant within every translated fiber.

For codons c and d,

$$\tau(c) = \tau(d)$$

implies

$$h\,c = h\,d.$$

The additive reconstruction

$$\hat{h} = P\,h$$

does not generally preserve that equality.

Two synonymous codons may have different nucleotide coordinates and therefore receive different reconstructed values.

Thus, the original field is fiber-constant, while the projected field is coordinate-additive.

The projection measures how well a translated-property field can be represented through codon composition. It does not preserve synonymous equality by definition.

D.20 Synonymous-Fiber Averaging

For translated class ω, let

$$F\omega = \tau^{-1}(\omega)$$

be its codon fiber, with degeneracy

$d\omega = | F\omega |$.

Define the fiber-averaged reconstruction by

$\hat{H}(\omega)$

$= (1 / d\omega) \Sigma_{c \in F\omega} \hat{h}\, c$.

The corresponding fiber-averaged residual is

$R(\omega)$

$= H(\omega) - \hat{H}(\omega)$.

Because the original property field is constant within the fiber,

$H(\omega)$

is the common empirical target value for all codons in $F\omega$.

The average $\hat{H}(\omega)$ supplies one reconstructed value for comparison with that target.

This procedure is appropriate when the analytical goal is an amino-acid-level scatter plot or regression.

It should not replace examination of the individual codon reconstructions, because variation within a synonymous fiber contains structural information.

Serine, leucine, and arginine provide important examples. Their codons occupy more than one local region of codon space, and their fitted values may vary substantially within a single translated class.

D.21 Within-Fiber Reconstruction Spread

For each translated class ω, define the within-fiber fitted variance

$\hat{V}\omega$

$= (1 / d\omega) \Sigma_{c \in F\omega}$

$[\hat{h}\, c - \hat{H}(\omega)]^2$.

This quantity is zero when all synonymous codons receive the same additive reconstruction.

It is positive when synonymous codons occupy nucleotide coordinates with different fitted sums.

A large value indicates that the additive nucleotide model divides a translated fiber internally even though the original property field treats all of its codons identically.

This spread should not be described automatically as measurement error. It reflects the mismatch between two structures:

fiber constancy under translated identity;

and

additivity under nucleotide-position coordinates.

The two structures coincide exactly only for specially arranged fibers and coefficient values.

D.22 Residuals Are Not Necessarily Noise

The residual

$r = h - P\,h$

is sometimes described loosely as unexplained error.

That language can be misleading.

The residual contains all components outside the additive nucleotide-position model. These may include:

pairwise codon-position interactions;

three-way codon effects;

translated-fiber structure incompatible with additive coordinates;

properties of unusual amino-acid chemistry;

and distinctions suppressed by the selected numerical scale.

The residual is unexplained only relative to the ten-dimensional additive model.

It may be highly structured under another representation.

For hydropathy, amino acids such as cysteine and phenylalanine have substantial fiber-averaged residuals. Those deviations do not imply that their empirical values are noisy. They show that their chemical properties are not captured accurately by the broad additive codon-coordinate pattern.

The correct interpretation is therefore:

the pseudoinverse extracts the complete additive component;

the residual contains the remaining structured and unstructured variation;

and further analysis is required before assigning a biological meaning to that remainder.

D.23 Gauge-Invariant Interpretation of Hydropathy

The Moore–Penrose hydropathy coefficients are useful because they provide one exact, reproducible gauge.

Their most robust interpretation lies in gauge-invariant quantities.

These include:

within-position nucleotide contrasts;

within-position coefficient ranges;

the ranking of nucleotides within one position;

the reconstructed codon field D β;

the projection norm;

and

the residual norm.

The individual coefficient

$\beta_{2,U} = 3.948958$

depends on the Moore–Penrose gauge.

The contrast

$\beta_{2,U} - \beta_{2,A} = 6.587500$

does not.

Similarly, the claim that the second-position coefficient range exceeds the first- and third-position ranges is gauge invariant.

The statement that U and A define the dominant hydropathy opposition therefore survives every allowable block-offset transformation.

D.24 Alternative Gauge Conventions

The Moore–Penrose gauge is not the only possible identifiable convention.

One alternative is a reference-level gauge. One nucleotide coefficient in each positional block is set to zero, and the remaining coefficients are interpreted relative to those references.

Another is a zero-sum gauge, in which the four coefficients within every positional block are constrained to sum to zero while a separate intercept represents the global mean.

Another may place the entire constant component in one selected positional block.

All such conventions can represent the same fitted codon field.

Their coefficient values differ by null-space transformations and by the handling of an explicit intercept.

The Moore–Penrose convention is preferred here because it is:

unique;

symmetric across the three positions;

orthogonal to the null space;

and

the minimum-norm solution.

Its use does not imply that other gauges are mathematically incorrect. It supplies one canonical choice for reproducible analysis.

D.25 Projection Geometry

The projection can be viewed geometrically in sixty-four-dimensional codon-field space.

The ten-dimensional additive subspace is a flat linear space passing through the origin.

An arbitrary property field h may lie outside it.

The orthogonal projection $\hat{h}$ is the point in that subspace closest to h.

The residual r is perpendicular to the subspace.

Every other additive field g satisfies

$\| h - g \|^2$

$= \| h - \hat{h} \|^2 + \| \hat{h} - g \|^2$.

Therefore,

$\| h - g \|^2 \geq \| h - \hat{h} \|^2$,

with equality only when

$g = \hat{h}$.

The pseudoinverse reconstruction is consequently not merely one convenient additive approximation. It is the unique least-squares optimum.

D.26 Coefficient Geometry

The coefficient-space geometry is complementary.

The twelve-dimensional coefficient space contains:

a ten-dimensional row space;

and

a two-dimensional null space.

Every coefficient vector decomposes uniquely as

$\beta = \beta_{row} + \beta_{null}$.

The matrix D discards the null component:

$D\,\beta = D\,\beta_{row}$.

The coefficient-space projector

$Q = D^{+} D$

maps β onto βrow:

$Q\,\beta = \beta_{row}$.

The Moore–Penrose inverse therefore performs two orthogonal operations:

in codon space, it identifies the closest additive field;

in coefficient space, it chooses the minimum-norm representative of that field.

These two geometries explain the two forms of uniqueness provided by the pseudoinverse.

D.27 Weighted Extensions

The present projection uses the ordinary Euclidean norm in which every codon receives equal weight:

$\| h - D \beta \|^2$

$= \Sigma c \in C [h c - (D \beta) c]^2.$

A weighted analysis could assign positive codon weights wc and minimize

$\Sigma c \in C \, wc [h c - (D \beta) c]^2.$

Such weights might represent:

codon usage;

measurement reliability;

organism-specific abundance;

or another declared sampling scheme.

The resulting weighted pseudoinverse would generally differ from the exact two-value matrix derived for the uniformly weighted complete codon space.

Its coefficients, gauge convention, leverage values, and projection geometry would also change.

The unweighted exact pseudoinverse remains the natural representation of the complete lookup domain itself because every codon is treated as one equally present member of C.

Weighted projections describe how a particular biological system samples that domain.

D.28 Summary

The complete codon incidence matrix represents a ten-dimensional additive field space using twelve displayed nucleotide-position coordinates.

The two-dimensional excess produces gauge freedom.

Every gauge transformation has the form

$(a 1_4, b 1_4, c 1_4)^T,$

with

$a + b + c = 0.$

Such transformations change positional block offsets without changing any codon reconstruction.

Within-position contrasts are gauge invariant.

The Moore–Penrose pseudoinverse selects the unique coefficient vector orthogonal to the null space. Its three positional block sums are equal, and it has minimum Euclidean norm among all equivalent coefficient representations.

For any codon-level field h,

$\hat{h} = D\, D^{+}\, h$

is the unique closest additive reconstruction, while

$r = h - \hat{h}$

is orthogonal to every additive nucleotide-position field.

The exact decomposition

$\| h \|^2 = \| \hat{h} \|^2 + \| r \|^2$

separates additive and nonadditive components. After centering, the same identity permits a variance-based measure of additive projection strength.

A field is reconstructed exactly if and only if it can be written as the sum of independent first-, second-, and third-position effects.

Translated-property fields are constant within synonymous fibers before projection, but their additive reconstructions need not remain fiber-constant. Fiber averaging therefore supplies an amino-acid-level comparison distinct from the codon-level projection.

The pseudoinverse is thus both an algebraic inverse structure and a geometric decomposition of codon-level information.

Appendix E

Complete Position-Specific Nucleotide-Pair Weights

E.1 Definition

The synonymous-edge scores of every isotropic and anisotropic codon cube can be derived from eighteen position-specific nucleotide-pair weights.

Let

$p \in \{ 1, 2, 3 \}$

denote a codon position, and let

{ x, y }

be an unordered pair of distinct nucleotides from

Σ = { G, A, C, U }.

Define

$w_p (x, y)$

as the number of unordered codon pairs { c, d } satisfying all four conditions:

1. c and d differ only at codon position p;
2. the two nucleotides at that position are x and y;
3. c and d agree at the other two codon positions;
4. c and d translate to the same output under the standard genetic code.

The pair is undirected, so

$w_p (x, y) = w_p (y, x)$.

There are six possible unordered nucleotide pairs:

A–C,

A–G,

A–U,

C–G,

C–U,

and

G–U.

Each pair has one weight at each of the three codon positions. The complete system therefore contains

$6 \times 3 = 18$

position-specific weights.

The total weight of one nucleotide pair is

$w (x, y)$

$= w_1 (x, y)$

- $w_2 (x, y)$
- $w_3 (x, y)$.

These six total weights determine every isotropic nucleotide-path score.

The eighteen position-specific weights determine every anisotropic score.

E.2 Direct Enumeration

The weights can be obtained by direct enumeration of the complete standard genetic code.

For each codon position p:

hold the other two positions fixed;

select one of the six unordered nucleotide pairs;

construct the corresponding two codons;

compare their translated outputs;

and increment the weight when the outputs are identical.

For a fixed position and nucleotide pair, there are

$4^2 = 16$

possible settings of the two unchanged codon positions.

Thus, every position-specific pair weight lies between zero and sixteen.

The complete calculation examines

$3 \times 6 \times 16 = 288$

candidate codon pairs.

This number is also the edge count of the complete Hamming-distance-one codon graph.

Every one of its sixty-four codons has nine one-position neighbors:

three alternative nucleotides at each of three positions.

Therefore, the number of undirected edges is

$64 \times 9 / 2 = 288$.

The eighteen weights partition the synonymous edges of that complete graph by codon position and nucleotide-pair type.

Their sum is sixty-nine.

The remaining

288 − 69 = 219

complete Hamming-distance-one edges connect different translated outputs.

E.3 First-Position Weights

At the first codon position, the six pair weights are

w_1 (A, C) = 2,

w_1 (A, G) = 0,

w_1 (A, U) = 0,

w_1 (C, G) = 0,

w_1 (C, U) = 2,

and

w_1 (G, U) = 0.

Only two first-position nucleotide-pair types preserve translated identity:

A–C;

and

C–U.

The two A–C synonymous pairs are

AGA ↔ CGA,

and

AGG ↔ CGG.

All four codons translate to arginine, R.

The two C–U synonymous pairs are

CUA ↔ UUA,

and

CUG ↔ UUG.

All four codons translate to leucine, L.

The complete first-position synonymous total is therefore

$\Sigma\{x,y\}\, w_1(x, y)$

$= 2 + 2$

$= 4.$

A linear nucleotide path can realize the complete first-position total only when it contains both edges

A–C

and

C–U.

Because these two edges meet at C, the path must contain the three-nucleotide segment

A–C–U

or its reverse

U–C–A.

The remaining nucleotide G can be attached to either endpoint.

The two undirected first-position maximizing paths are therefore

G–A–C–U

and

A–C–U–G.

Their reversals represent the same undirected topologies.

Thus, the first-position maximum is

$\max L_1 = 4.$

E.4 Second-Position Weights

At the second codon position, only one unordered nucleotide pair produces a synonymous codon pair:

$w_2 (A, G) = 1.$

All remaining second-position weights vanish:

$w_2 (A, C) = 0,$

$w_2 (A, U) = 0,$

$w_2 (C, G) = 0,$

$w_2 (C, U) = 0,$

and

$w_2 (G, U) = 0.$

The one synonymous second-position pair is

UAA ↔ UGA.

Both codons translate to X.

The complete second-position synonymous total is therefore

$\Sigma\{x,y\} w_2 (x, y) = 1.$

A nucleotide path reaches the second-position maximum exactly when A and G are adjacent.

Among the twenty-four oriented nucleotide orders:

twelve contain A–G as an edge and score one;

and

twelve omit A–G and score zero.

Thus,

$\max L_2 = 1.$

The near absence of second-position synonymy contrasts sharply with the strong physicochemical role of the middle nucleotide.

The second position contributes the dominant hydropathy contrast, yet it contributes only one synonymous edge in the complete Hamming-distance-one graph.

E.5 Third-Position Weights

The third-position nucleotide-pair weights are much larger:

$w_3 (A, C) = 9,$

w_3 (A, G) = 14,

w_3 (A, U) = 9,

w_3 (C, G) = 8,

w_3 (C, U) = 16,

and

w_3 (G, U) = 8.

Their sum is

9 + 14 + 9 + 8 + 16 + 8 = 64.

Thus, the complete Hamming-distance-one graph contains sixty-four synonymous third-position codon pairs.

Every unordered nucleotide-pair type contributes at the third position. This differs from the first and second positions, where most pair types have zero synonymous weight.

A linear four-nucleotide path contains only three of the six possible nucleotide-pair types. It therefore cannot realize all sixty-four synonymous third-position pairs simultaneously.

The two largest third-position weights are

C–U = 16

and

A–G = 14.

Together they contribute

16 + 14 = 30.

These edges are disconnected from one another: one joins C to U, while the other joins A to G.

To form one path through all four nucleotide states, they must be connected by one additional edge.

The two largest available connecting weights are

A–C = 9

and

A–U = 9.

Using either one produces a third-position score of

16 + 14 + 9 = 39.

The two undirected third-position maximizing paths are therefore

G–A–C–U

and

G–A–U–C.

Thus,

max L_3 = 39.

E.6 Total Pair Weights

The isotropic weight of a nucleotide pair is the sum of its weights across the three codon positions:

w (x, y)

= w_1 (x, y)

- w_2 (x, y)
- w_3 (x, y).

The six totals are

w (A, C)

= 2 + 0 + 9

= 11;

w (A, G)

= 0 + 1 + 14

= 15;

w (A, U)

= 0 + 0 + 9

= 9;

w (C, G)

= 0 + 0 + 8

= 8;

w (C, U)

= 2 + 0 + 16

= 18;

and

w (G, U)

= 0 + 0 + 8

= 8.

Their complete sum is

11 + 15 + 9 + 8 + 18 + 8 = 69.

Equivalently,

4 + 1 + 64 = 69.

The number sixty-nine is the complete count of synonymous codon pairs differing at exactly one position when all six possible nucleotide-pair substitutions are admitted.

A linear nucleotide path selects only three of the six pair types.

Its synonymous score is the sum of the weights assigned to those three selected edges.

E.7 The G–A–C–U Score

The G–A–C–U nucleotide path contains the three adjacent pairs

G–A,

A–C,

and

C–U.

At the first codon position, its score is

L_1 (G–A–C–U)

= w_1 (A, G)

- w_1 (A, C)
- w_1 (C, U)

= 0 + 2 + 2

= 4.

At the second position,

L_2 (G–A–C–U)

= w_2 (A, G)

- w_2 (A, C)
- w_2 (C, U)

= 1 + 0 + 0

= 1.

At the third position,

L_3 (G–A–C–U)

= w_3 (A, G)

- w_3 (A, C)
- w_3 (C, U)

= 14 + 9 + 16

= 39.

Therefore,

L (G–A–C–U)

= 4 + 1 + 39

= 44.

The same result follows directly from the total pair weights:

L (G–A–C–U)

= w (A, G)

- w (A, C)

- w (C, U)

= 15 + 11 + 18

= 44.

The decomposition

44 = 4 + 1 + 39

shows how the same three nucleotide edges contribute differently at the three codon positions.

E.8 Exact Isotropic Maximum from the Pair Weights

The total pair weights in descending order are

C–U = 18,

A–G = 15,

A–C = 11,

A–U = 9,

C–G = 8,

and

G–U = 8.

A four-nucleotide path contains exactly three edges.

Therefore, the score of any isotropic path cannot exceed the sum of the three largest pair weights:

18 + 15 + 11 = 44.

Those three pair types are

C–U,

A–G,

and

A–C.

They form the connected four-vertex path

G–A–C–U.

Thus, the upper bound is attainable.

No path can exceed forty-four because no path contains more than three pair weights, and no alternative set of three weights has a larger sum.

Equality requires all three maximum-weight edges. Those edges determine G–A–C–U up to reversal.

The exact isotropic maximum therefore follows immediately from the six total pair weights.

E.9 Complementary Paths

A four-nucleotide path uses three edges of the complete nucleotide graph K_4.

The three unused nucleotide-pair types form the complement of that path.

The complement of a Hamiltonian path in K_4 is another Hamiltonian path.

Because the total of all six isotropic pair weights is sixty-nine, a path with score L has a complementary path with score

$69 - L$.

For G–A–C–U, the selected edges are

A–G,

A–C,

and

C–U.

The omitted edges are

A–U,

C–G,

and

G–U.

These form the complementary path

A–U–G–C.

Its score is

$w (A, U)$

- w (G, U)
- w (C, G)

= 9 + 8 + 8

= 25.

Therefore,

44 + 25 = 69.

The same relation holds for every complementary path pair.

At the position-specific level, the total available weights are:

4 at the first position;

1 at the second position;

and

64 at the third position.

A path having positional score

(L_1, L_2, L_3)

therefore has a complementary path with score

(4 − L_1, 1 − L_2, 64 − L_3).

The two total scores sum to

4 + 1 + 64 = 69.

This complement relation underlies the exact symmetry of both the isotropic and anisotropic score distributions around

69 / 2 = 34.5.

E.10 Complete Pair-Weight Matrix Form

The six weights at one codon position can be written as a symmetric 4 × 4 matrix with zero diagonal.

Using nucleotide order

A, C, G, U,

the first-position matrix is

W_1 =

$$\begin{bmatrix} 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{bmatrix}.$$

The second-position matrix is

W_2 =

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The third-position matrix is

W_3 =

$$\begin{bmatrix} 0 & 9 & 14 & 9 \\ 9 & 0 & 8 & 16 \\ 14 & 8 & 0 & 8 \\ 9 & 16 & 8 & 0 \end{bmatrix}.$$

Their sum is

$W = W_1 + W_2 + W_3$

=

$$\begin{bmatrix} 0 & 11 & 15 & 9 \\ 11 & 0 & 8 & 18 \\ 15 & 8 & 0 & 8 \\ 9 & 18 & 8 & 0 \end{bmatrix}.$$

For a path

n_1–n_2–n_3–n_4,

the position-specific score is

L_p

= W_p (n_1, n_2)

- W_p (n_2, n_3)

- W_p (n_3, n_4).

The isotropic score is

L

= W (n_1, n_2)

- W (n_2, n_3)
- W (n_3, n_4).

The weight matrices therefore contain all information needed to reproduce the path searches without reconstructing the codon cube for every candidate order.

E.11 Biological Reading of the Weight Matrices

The three position-specific matrices have sharply different forms.

The first-position matrix is sparse.

Only A–C and C–U carry nonzero weight, and the complete synonymous total is four.

The second-position matrix is nearly empty.

Only A–G carries nonzero weight, and the complete synonymous total is one.

The third-position matrix is dense.

Every unordered nucleotide pair has nonzero synonymous weight, and the complete total is sixty-four.

This is a precise representation of positional asymmetry in the standard genetic code.

At the first position, a small number of nucleotide substitutions preserve translated identity.

At the second, almost none do.

At the third, every substitution type participates in multiple synonymous codon pairs.

The code therefore does not distribute synonymy uniformly across positions or nucleotide-pair types.

The asymmetry also parallels the physicochemical analysis.

The second position contributes the fewest synonymous pairs but the strongest hydropathy discrimination.

The third contributes the most synonymous pairs but the smallest hydropathy coefficient range.

The first occupies an intermediate role.

E.12 Transition and Transversion Classes

The six unordered nucleotide pairs can also be divided into conventional transition and transversion classes.

The transition pairs are

A–G

and

C–U.

Their total synonymous weight is

w (A, G) + w (C, U)

= 15 + 18

= 33.

The transversion pairs are

A–C,

A–U,

C–G,

and

G–U.

Their total synonymous weight is

11 + 9 + 8 + 8 = 36.

Thus, the complete total of sixty-nine synonymous Hamming-distance-one pairs is divided as

33 transition pairs

and

36 transversion pairs.

The synonymous total is therefore not overwhelmingly concentrated in transitions.

A linear nucleotide path must select a connected set of three nucleotide-pair types. The placement and connectivity of the weights matter as much as their broad classification as transitions or transversions.

The G–A–C–U path contains two transition edges:

G–A

and

C–U.

They are connected by the transversion edge

A–C.

Their combined weight is

15 + 18 + 11 = 44.

Nucleotide pair	First position	Second position	Third position	Total
A–G	0	1	14	15
A–C	2	0	9	11
A–U	0	0	9	9
C–G	0	0	8	8
C–U	2	0	16	18
G–U	0	0	8	8

Table E1. Position-Specific and Total Synonymous Nucleotide-Pair Weights. Each entry counts codon pairs that differ only by the indicated unordered nucleotide substitution at the stated codon position and retain the same translated output. The first position contains four such pairs, the second contains one, and the third contains sixty-four in the complete Hamming-distance-one graph. A linear nucleotide path selects three pair types. G–A–C–U selects A–G, A–C, and C–U, producing positional scores of 4, 1, and 39 and a total score of 44.

E.13 Reproducibility Checks

An independent implementation should verify all of the following:

there are six unordered nucleotide pairs;

there are three codon positions;

all eighteen weights are nonnegative integers;

the first-position weights sum to four;

the second-position weights sum to one;

the third-position weights sum to sixty-four;

all eighteen weights sum to sixty-nine;

the six isotropic weights sum to sixty-nine;

the G–A–C–U path scores 4, 1, and 39 by position;

the G–A–C–U isotropic score is forty-four;

the complementary A–U–G–C path scores twenty-five;

and complementary isotropic path scores sum to sixty-nine.

The pair weights should also agree exactly with direct enumeration of synonymous edges in the complete Hamming-distance-one graph.

A second implementation should generate codon-cube edges from path coordinates and reproduce every score obtained from the weight matrices.

Agreement between direct edge enumeration and compressed pair-weight calculation provides an important independent check.

E.14 Summary

The complete standard genetic code contains sixty-nine synonymous codon pairs at Hamming distance one.

Their positional distribution is

4 at the first codon position;

1 at the second;

and

64 at the third.

Their total nucleotide-pair weights are

A–C = 11,

A–G = 15,

A–U = 9,

C–G = 8,

C–U = 18,

and

G–U = 8.

These six values determine every isotropic linear-path score.

The complete eighteen position-specific values determine every anisotropic score.

The G–A–C–U path selects A–G, A–C, and C–U, producing

15 + 11 + 18 = 44.

Its position-specific decomposition is

4 + 1 + 39 = 44.

The pair-weight representation compresses the complete codon-edge calculation without loss of scoring information. It provides the exact foundation for the complete isotropic enumeration in Appendix F.

Appendix F

Complete Enumeration of Isotropic Nucleotide Orders

F.1 Isotropic Path Definition

An isotropic codon cube uses the same linear nucleotide path on all three codon axes.

Let

$\pi = n_1–n_2–n_3–n_4$

be an oriented ordering of the four RNA nucleotide states

G, A, C, and U.

The path selects the three unordered nucleotide-pair types

$\{ n_1, n_2 \}$,

$\{ n_2, n_3 \}$,

and

{ n_3, n_4 }.

The same three selected pair types are used for substitutions at:

the first codon position;

the second codon position;

and

the third codon position.

The resulting codon graph contains:

64 vertices;

48 edges along each codon axis;

and

144 undirected edges in total.

The isotropic synonymous-edge score is the number of those 144 edges whose endpoint codons translate to the same standard genetic-code output.

Because reversing a path does not change its undirected nucleotide adjacencies,

n_1–n_2–n_3–n_4

and

n_4–n_3–n_2–n_1

produce the same undirected codon graph and the same score.

There are therefore:

4! = 24 oriented nucleotide orders;

and

24 / 2 = 12 undirected path topologies.

The complete enumeration evaluates all twenty-four oriented paths. The twelve reversal pairs provide the nonredundant undirected classification.

F.2 Score from Total Pair Weights

Appendix E derived the six total synonymous nucleotide-pair weights:

A–G = 15;

A–C = 11;

A–U = 9;

C–G = 8;

C–U = 18;

and

G–U = 8.

For an isotropic path

$\pi = n_1–n_2–n_3–n_4$,

the total synonymous score is

$L (\pi)$

$= w (n_1, n_2)$

- $w (n_2, n_3)$
- $w (n_3, n_4)$.

The score can therefore be calculated without constructing all 144 codon-cube edges separately for every path.

For example, the path

G–A–C–U

selects

G–A,

A–C,

and

C–U.

Its score is

L (G–A–C–U)

= 15 + 11 + 18

= 44.

The path

G–A–U–C

selects

G–A,

A–U,

and

U–C.

Its score is

15 + 9 + 18

= 42.

The path

A–U–G–C

selects

A–U,

U–G,

and

G–C.

Its score is

9 + 8 + 8

= 25.

F.3 Complete Undirected Enumeration

The twelve undirected nucleotide paths and their scores are:

G–A–C–U: 44;

G–A–U–C: 42;

A–G–C–U: 41;

A–G–U–C: 41;

A–C–U–G: 37;

A–U–C–G: 35;

C–A–G–U: 34;

C–G–A–U: 32;

C–A–U–G: 28;

G–C–A–U: 28;

A–C–G–U: 27;

and

A–U–G–C: 25.

Every path represents its reverse as the same undirected topology.

Thus:

G–A–C–U and U–C–A–G both score 44;

G–A–U–C and C–U–A–G both score 42;

A–G–C–U and U–C–G–A both score 41;

A–G–U–C and C–U–G–A both score 41;

A–C–U–G and G–U–C–A both score 37;

A–U–C–G and G–C–U–A both score 35;

C–A–G–U and U–G–A–C both score 34;

C–G–A–U and U–A–G–C both score 32;

C–A–U–G and G–U–A–C both score 28;

G–C–A–U and U–A–C–G both score 28;

A–C–G–U and U–G–C–A both score 27;

and

A–U–G–C and C–G–U–A both score 25.

Oriented order	Score	Oriented order	Score
G–A–C–U	44	C–A–G–U	34
U–C–A–G	44	U–G–A–C	34
G–A–U–C	42	C–G–A–U	32
C–U–A–G	42	U–A–G–C	32
A–G–C–U	41	C–A–U–G	28
U–C–G–A	41	G–U–A–C	28
A–G–U–C	41	G–C–A–U	28
C–U–G–A	41	U–A–C–G	28
A–C–U–G	37	A–C–G–U	27
G–U–C–A	37	U–G–C–A	27
A–U–C–G	35	A–U–G–C	25
G–C–U–A	35	C–G–U–A	25

Table F1. Complete Enumeration of Isotropic Nucleotide Orders. All twenty-four oriented linear orders of G, A, C, and U are grouped into twelve reversal-equivalent undirected paths. The synonymous-edge score is the sum of the three selected total nucleotide-pair weights. G–A–C–U and its reverse uniquely attain the maximum score of forty-four. A–U–G–C and its reverse attain the minimum score of twenty-five.

F.4 Verification of Every Path Score

The complete score calculation can be displayed explicitly.

Maximum path

G–A–C–U:

A–G + A–C + C–U

= 15 + 11 + 18

= 44.

Second-ranked path

G–A–U–C:

A–G + A–U + C–U

= 15 + 9 + 18

= 42.

Paths scoring forty-one

A–G–C–U:

A–G + C–G + C–U

= 15 + 8 + 18

= 41.

A–G–U–C:

A–G + G–U + C–U

= 15 + 8 + 18

= 41.

Path scoring thirty-seven

A–C–U–G:

A–C + C–U + G–U

= 11 + 18 + 8

= 37.

Path scoring thirty-five

A–U–C–G:

A–U + C–U + C–G

= 9 + 18 + 8

= 35.

Path scoring thirty-four

C–A–G–U:

A–C + A–G + G–U

= 11 + 15 + 8

= 34.

Path scoring thirty-two

C–G–A–U:

C–G + A–G + A–U

= 8 + 15 + 9

= 32.

Paths scoring twenty-eight

C–A–U–G:

A–C + A–U + G–U

= 11 + 9 + 8

= 28.

G–C–A–U:

C–G + A–C + A–U

= 8 + 11 + 9

= 28.

Path scoring twenty-seven

A–C–G–U:

A–C + C–G + G–U

= 11 + 8 + 8

= 27.

Minimum path

A–U–G–C:

A–U + G–U + C–G

= 9 + 8 + 8

= 25.

Every reverse path contains the same three unordered edges and therefore has the same score.

F.5 Exact Maximum

The three largest total nucleotide-pair weights are

C–U = 18,

A–G = 15,

and

A–C = 11.

Their sum is

18 + 15 + 11 = 44.

A nucleotide path contains exactly three edges. Therefore, no path can score more than the sum of the three largest available edge weights.

The three maximum-weight edges form the connected path

G–A–C–U.

They use all four nucleotide vertices exactly once and therefore constitute a valid Hamiltonian path through the nucleotide set.

The upper bound of forty-four is consequently attained.

Equality requires a path to contain all three highest-weight edges. Those edges uniquely determine

G–A–C–U

up to reversal.

Therefore:

the maximum isotropic score is forty-four;

and

G–A–C–U is the unique undirected maximizing path.

In oriented form, the two maximizers are

G–A–C–U

and

U–C–A–G.

F.6 Exact Minimum

The three smallest total pair weights are

C–G = 8,

G–U = 8,

and

A–U = 9.

Their sum is

8 + 8 + 9 = 25.

These three edges form the valid path

C–G–U–A,

which is the reverse of

A–U–G–C.

Therefore, the lower bound of twenty-five is attained.

The minimum isotropic score is twenty-five, realized by the undirected path

A–U–G–C.

Its oriented forms are

A–U–G–C

and

C–G–U–A.

The minimum path is the graph complement of the maximum path within the complete six-edge nucleotide graph.

F.7 Complementary-Path Pairing

The complete graph on four nucleotide vertices contains six unordered edges.

A four-vertex linear path selects three of them.

The three omitted edges form another four-vertex linear path.

Thus, every isotropic path π has a complementary path π^{c}.

The six total pair weights sum to

69.

Therefore,

$L(\pi) + L(\pi^c) = 69$.

The twelve undirected paths form six complementary pairs.

Complementary pair 1

G–A–C–U scores 44.

A–U–G–C scores 25.

Therefore,

44 + 25 = 69.

Complementary pair 2

G–A–U–C scores 42.

A–C–G–U scores 27.

Therefore,

42 + 27 = 69.

Complementary pair 3

A–G–C–U scores 41.

C–A–U–G scores 28.

Therefore,

41 + 28 = 69.

Complementary pair 4

A–G–U–C scores 41.

G–C–A–U scores 28.

Therefore,

41 + 28 = 69.

Complementary pair 5

A–C–U–G scores 37.

C–G–A–U scores 32.

Therefore,

37 + 32 = 69.

Complementary pair 6

A–U–C–G scores 35.

C–A–G–U scores 34.

Therefore,

35 + 34 = 69.

The score distribution is consequently symmetric about

69 / 2 = 34.5.

Because sixty-nine is odd, no integer path score lies exactly at the center.

F.8 Oriented Score Multiplicities

Counting all twenty-four oriented paths, the exact score frequencies are:

score 25: 2 paths;

score 27: 2 paths;

score 28: 4 paths;

score 32: 2 paths;

score 34: 2 paths;

score 35: 2 paths;

score 37: 2 paths;

score 41: 4 paths;

score 42: 2 paths;

and

score 44: 2 paths.

The multiplicities sum to

2 + 2 + 4 + 2 + 2 + 2 + 2 + 4 + 2 + 2

= 24.

Scores that do not occur are:

26;

29;

30;

31;

33;

36;

38;

39;

40;

and

43.

The isotropic score distribution is discrete and sparse. Only ten of the twenty integers from twenty-five through forty-four occur.

The anisotropic distribution of Appendix G is denser because three independently selected positional path scores can be combined.

F.9 Isotropic Generating Polynomial

The oriented isotropic score distribution can be encoded by the polynomial

$I(x) = 2x^{25} + 2x^{27} + 4x^{28} + 2x^{32} + 2x^{34} + 2x^{35} + 2x^{37} + 4x^{41} + 2x^{42}$

- $2\,x^{44}$.

The coefficient of x^{s} is the number of oriented nucleotide paths with score s.

Evaluating at x = 1 gives

I (1) = 24.

Complementary-path symmetry gives the reciprocal identity

$I(x) = x^{69}\, I(x^{-1})$.

For example, the coefficient of x^{25} equals the coefficient of x^{44}, and the coefficient of x^{28} equals the coefficient of x^{41}.

The corresponding undirected generating polynomial is

Iu (x)

$= x^{25}$

- x^{27}

$+ 2\,x^{28}$

- x^{32}
- x^{34}
- x^{35}
- x^{37}

$+ 2\,x^{41}$

- x^{42}
- x^{44}.

Its coefficients sum to twelve.

F.10 Mean Score

Every one of the six nucleotide edges occurs in exactly half of the twenty-four oriented Hamiltonian paths.

For a fixed unordered nucleotide pair, the number of oriented paths in which the two nucleotides are adjacent is

12.

Therefore, the total score summed over all oriented paths is

12 × 69 = 828.

The mean isotropic score is

828 / 24

= 34.5.

The same result follows from complementary pairing. Every pair of complementary paths has total score sixty-nine, so its average is 34.5.

The mean equals the midpoint of the minimum and maximum:

(25 + 44) / 2 = 34.5.

The median is also 34.5. In the ordered list of twenty-four scores, the twelfth value is thirty-four and the thirteenth is thirty-five.

F.11 Variance and Standard Deviation

The oriented score distribution has population variance

Var (L)

= 39.583333…

= 475 / 12.

The population standard deviation is therefore

σ

= √(475 / 12)

≈ 6.291529.

The maximum score lies

(44 − 34.5) / 6.291529

≈ 1.510

population standard deviations above the mean.

The minimum lies the same distance below the mean because the distribution is exactly symmetric.

These summary statistics describe the complete set of twenty-four oriented nucleotide orders. They are not estimates from sampling.

F.12 Ranking G–A–C–U

The maximum path has score forty-four.

The second-ranked undirected path has score forty-two.

The gap between first and second place is therefore

44 − 42 = 2.

The next two undirected paths score forty-one.

Thus, the G–A–C–U path exceeds:

the second-ranked path by two edges;

the third-ranked paths by three edges;

the median score by 9.5 edges;

and

the minimum path by nineteen edges.

Its advantage is not due to one overwhelmingly dominant nucleotide-pair weight.

The path succeeds because it simultaneously includes:

the largest weight, C–U = 18;

the second-largest weight, A–G = 15;

and

the third-largest weight, A–C = 11.

The three edges also satisfy the connectivity requirement of one linear path.

F.13 Position-Specific Decomposition of the Isotropic Paths

Although the isotropic score is calculated from the six total weights, every path also has a three-component positional score.

The twelve undirected paths have the following decompositions:

G–A–C–U:

4 + 1 + 39 = 44.

G–A–U–C:

2 + 1 + 39 = 42.

A–G–C–U:

2 + 1 + 38 = 41.

A–G–U–C:

2 + 1 + 38 = 41.

A–C–U–G:

4 + 0 + 33 = 37.

A–U–C–G:

2 + 0 + 33 = 35.

C–A–G–U:

2 + 1 + 31 = 34.

C–G–A–U:

0 + 1 + 31 = 32.

C–A–U–G:

2 + 0 + 26 = 28.

G–C–A–U:

2 + 0 + 26 = 28.

A–C–G–U:

2 + 0 + 25 = 27.

A–U–G–C:

0 + 0 + 25 = 25.

These decompositions show that the total ranking combines three highly unequal positional contributions.

The third-position score supplies most of every total.

The first position separates several paths having the same third-position score.

The second position contributes either zero or one depending on whether the path contains A–G.

The G–A–C–U path is the only undirected path simultaneously attaining:

the first-position maximum of four;

the second-position maximum of one;

and

the third-position maximum of thirty-nine.

F.14 Why the Maximum Is Stronger Than the Total Ranking

The isotropic enumeration alone establishes that G–A–C–U is the unique undirected path with the largest total score.

The positional decomposition establishes a stronger property.

A path might conceivably attain the highest total by compensating for a suboptimal score at one position with a higher score elsewhere.

That does not occur here.

G–A–C–U independently reaches the maximum at every codon position.

Its total is therefore

$\max L_1 + \max L_2 + \max L_3$

$= 4 + 1 + 39$

$= 44.$

The uniqueness of the common positional maximizer is proved fully by the anisotropic classification.

The isotropic path is distinguished not merely by having the highest aggregate score but by satisfying all three local optimality conditions at once.

F.15 Direct Cube Verification

The compressed pair-weight calculation should be verified independently through direct construction of every codon cube.

For each oriented nucleotide path:

assign coordinate values 0, 1, 2, and 3 according to the nucleotide order;

map every codon to a point in the 4 × 4 × 4 lattice;

connect codons whose coordinates differ by one on exactly one axis;

verify that the graph contains 144 undirected edges;

translate both endpoints of every edge;

count edges having equal translated outputs;

and record the positional contributions.

The directly enumerated score must equal the sum of the three selected total pair weights.

For G–A–C–U, direct enumeration must return:

144 total edges;

44 synonymous edges;

100 nonsynonymous edges;

4 synonymous first-axis edges;

1 synonymous second-axis edge;

and

39 synonymous third-axis edges.

The direct and compressed implementations should agree for all twenty-four oriented paths.

F.16 Reversal Invariance

Let

$\pi = n_1–n_2–n_3–n_4$.

Its reverse is

$\pi R = n_4–n_3–n_2–n_1$.

Both paths select the same unordered nucleotide-pair set:

$\{ n_1, n_2 \}$,

$\{ n_2, n_3 \}$,

and

$\{ n_3, n_4 \}$.

Therefore,

$L(\pi R) = L(\pi)$.

At the codon-cube level, reversing an axis reflects the coordinate system but does not change which codon pairs are adjacent.

When all three isotropic axes are reversed together, the cube is reflected along all three coordinate directions. Its edge set and translated-edge multiplicities are unchanged.

This invariance justifies treating the twenty-four oriented orders as twelve undirected path topologies when orientation itself carries no weight or direction.

F.17 Scope of the Enumeration

The enumeration is exhaustive within the declared comparison class:

four nucleotide states;

one linear path containing all four states exactly once;

the same path on all three codon axes;

undirected nearest-neighbor edges;

equal edge weights;

and

synonymous translated identity as the objective.

It does not enumerate every possible graph on the four nucleotide states.

Other adjacency models could include:

cycles;

stars;

complete graphs;

weighted graphs;

directed graphs;

or disconnected graphs.

The present result should therefore be stated precisely:

G–A–C–U uniquely maximizes synonymous nearest-neighbor edges among all isotropic codon cubes formed from linear four-nucleotide paths.

It should not be generalized automatically to every possible nucleotide geometry or every possible biological objective.

F.18 Reproducibility Checks

A correct isotropic enumeration should verify:

there are exactly twenty-four oriented paths;

there are exactly twelve reversal-equivalent undirected paths;

every path contains four distinct nucleotides;

every path contains exactly three unordered edges;

every oriented path and its reverse have equal scores;

the minimum score is twenty-five;

the maximum score is forty-four;

exactly two oriented paths score forty-four;

exactly one undirected path scores forty-four;

the oriented frequencies sum to twenty-four;

the undirected frequencies sum to twelve;

the mean score is 34.5;

every complementary path pair sums to sixty-nine;

and

G–A–C–U has positional decomposition 4 + 1 + 39.

The score of each path should agree between:

direct codon-cube edge enumeration;

and

the sum of the selected pair weights.

Any disagreement indicates an error in codon order, path construction, translation assignment, edge deduplication, or score aggregation.

F.19 Summary

The complete isotropic comparison contains twenty-four oriented nucleotide orders and twelve undirected path topologies.

Their scores are determined exactly by the six total synonymous nucleotide-pair weights.

The scores range from twenty-five to forty-four.

G–A–C–U uniquely attains the undirected maximum:

15 + 11 + 18 = 44.

Its positional decomposition is

4 + 1 + 39 = 44.

The complementary path A–U–G–C attains the minimum:

9 + 8 + 8 = 25.

Every path and its complement have scores summing to sixty-nine, producing exact symmetry around 34.5.

The exhaustive enumeration confirms that the G–A–C–U maximum is not a visual impression or a selected example. It is the exact optimum over the complete declared isotropic path space.

Appendix G extends the enumeration to all 13,824 anisotropic combinations of independently ordered codon axes.

Appendix G

Exact Anisotropic Score Distribution

G.1 Anisotropic Codon Cubes

An anisotropic codon cube permits the three codon positions to use different linear nucleotide paths.

Let

π_1

be the oriented nucleotide path assigned to the first codon position,

π_2

the path assigned to the second codon position,

and

π_3

the path assigned to the third codon position.

An anisotropic cube is specified by the ordered triple

(π_1, π_2, π_3).

Each path is an oriented permutation of

G, A, C, and U.

There are

$4! = 24$

oriented paths available to each axis.

Therefore, the total number of oriented anisotropic cubes is

$24^3 = 13{,}824$.

After identifying each axis path with its reverse, there are twelve undirected path choices per axis. The corresponding number of undirected anisotropic cubes is

$12^3 = 1{,}728$.

Every anisotropic cube contains the same sixty-four codon vertices and the same total of 144 edges. What changes is the set of nucleotide-pair substitutions permitted independently at the three codon positions.

G.2 Exact Axis Separability

Every edge of a codon cube changes exactly one codon position.

A first-axis edge is determined only by the path π_1.

A second-axis edge is determined only by π_2.

A third-axis edge is determined only by π_3.

Consequently, the synonymous-edge score separates exactly:

$L (\pi_1, \pi_2, \pi_3)$

$= L_1 (\pi_1) + L_2 (\pi_2) + L_3 (\pi_3)$.

There are no cross-axis interaction terms.

The first-position contribution does not depend on the paths selected for the second and third positions. The same independence holds for the other two axes.

This separability permits the complete 13,824-case score distribution to be derived from three small positional distributions.

It also proves that the global maximum is the sum of the three independent positional maxima:

max L

$= \max L_1 + \max L_2 + \max L_3$.

G.3 First-Position Spectrum

The nonzero first-position nucleotide-pair weights are

$w_1 (A, C) = 2$

and

$w_1 (C, U) = 2$.

All four remaining pair weights are zero.

A four-nucleotide path may contain:

neither of these two weighted edges;

exactly one of them;

or

both of them.

The corresponding first-position scores are:

0;

2;

and

4.

Among the twenty-four oriented nucleotide paths:

4 paths contain neither A–C nor C–U and score 0;

16 contain exactly one and score 2;

4 contain both and score 4.

Thus, the first-position frequency distribution is

$N_1(0) = 4$,

$N_1(2) = 16$,

and

$N_1(4) = 4$.

Its generating polynomial is

$F_1(x)$

$= 4 + 16x^2 + 4x^4$.

Evaluating at $x = 1$ gives

$F_1(1) = 24$.

The first-position minimum is zero, and the maximum is four.

G.4 Second-Position Spectrum

Only one second-position nucleotide-pair type has nonzero synonymous weight:

$w_2(A, G) = 1$.

A path scores one when it contains A–G as an edge and zero when it does not.

A fixed unordered edge occurs in twelve of the twenty-four oriented Hamiltonian paths through four labeled vertices.

Therefore:

12 oriented paths score 0;

and

12 oriented paths score 1.

The second-position frequency distribution is

$N_2(0) = 12$

and

$N_2(1) = 12$.

Its generating polynomial is

$F_2(x)$

$= 12 + 12x.$

Evaluating at $x = 1$ gives

$F_2(1) = 24.$

The second-position minimum is zero, and the maximum is one.

G.5 Third-Position Spectrum

The third-position nucleotide-pair weights are

A–G = 14,

A–C = 9,

A–U = 9,

C–G = 8,

C–U = 16,

and

G–U = 8.

The twelve undirected paths produce six distinct third-position scores:

25,

26,

31,

33,

38,

and

39.

Each score is attained by two undirected paths or four oriented paths.

The complete oriented third-position distribution is therefore

$N_3(25) = 4,$

$N_3 (26) = 4,$

$N_3 (31) = 4,$

$N_3 (33) = 4,$

$N_3 (38) = 4,$

and

$N_3 (39) = 4.$

The generating polynomial is

$F_3 (x)$

$= 4 x^{25}$

- $4 x^{26}$
- $4 x^{31}$
- $4 x^{33}$
- $4 x^{38}$
- $4 x^{39}$.

Equivalently,

$F_3 (x)$

$= 4 (x^{25} + x^{26} + x^{31} + x^{33} + x^{38} + x^{39}).$

Evaluating at $x = 1$ gives

$F_3 (1) = 24.$

The third-position minimum is twenty-five, and the maximum is thirty-nine.

G.6 Positional Means

The mean first-position score is

μ_1

$= [4 (0) + 16 (2) + 4 (4)] / 24$

$= 48 / 24$

$= 2.$

The mean second-position score is

μ_2

$= [12 (0) + 12 (1)] / 24$

$= 12 / 24$

$= 1 / 2.$

The mean third-position score is

μ_3

$= 4 (25 + 26 + 31 + 33 + 38 + 39) / 24$

$= 192 / 6$

$= 32.$

Therefore, the mean anisotropic score is

μ

$= \mu_1 + \mu_2 + \mu_3$

$= 2 + 1 / 2 + 32$

$= 34.5.$

The mean is also the midpoint of the global minimum and maximum:

$(25 + 44) / 2 = 34.5.$

G.7 Positional Variances

The first-position variance is

$\text{Var} (L_1)$

$= [4 (0 - 2)^2$

- $16 (2 - 2)^2$
- $4 (4 - 2)^2] / 24$

$= [4 (4) + 4 (4)] / 24$

$= 32 / 24$

$= 4 / 3.$

The second-position score is equally divided between zero and one. Therefore,

$\mathrm{Var}(L_2) = 1 / 4.$

For the third position, the six equally frequent values are

25, 26, 31, 33, 38, and 39,

with mean thirty-two.

Their deviations are

−7, −6, −1, 1, 6, and 7.

Therefore,

$\mathrm{Var}(L_3)$

$= [49 + 36 + 1 + 1 + 36 + 49] / 6$

$= 172 / 6$

$= 86 / 3.$

Because the three axis choices are independent in the complete anisotropic enumeration, their variances add:

$\mathrm{Var}(L)$

$= \mathrm{Var}(L_1) + \mathrm{Var}(L_2) + \mathrm{Var}(L_3)$

$= 4 / 3 + 1 / 4 + 86 / 3$

$= 30 + 1 / 4$

$= 121 / 4$

$= 30.25.$

The population standard deviation is therefore

σ

$= \sqrt{(121 / 4)}$

$= 11 / 2$

$= 5.5.$

The exact mean and population standard deviation are thus

$\mu = 34.5$

and

$\sigma = 5.5$.

G.8 The Exact Generating Polynomial

Because the score is the sum of three independently selected positional scores, the generating polynomial for the complete anisotropic distribution is

$F(x)$

$= F_1(x)\,F_2(x)\,F_3(x)$.

Substituting the positional polynomials gives

$F(x)$

$= (4 + 16x^2 + 4x^4)$

$(12 + 12x)$

$4(x^{25} + x^{26} + x^{31} + x^{33} + x^{38} + x^{39})$.

The coefficient of x^s is the number of oriented anisotropic cubes having total synonymous-edge score s.

Evaluating at $x = 1$ gives

$F(1)$

$= 24 \times 24 \times 24$

$= 13{,}824$.

The smallest exponent is twenty-five, and the largest is forty-four.

The coefficient of x^{25} is

$4 \times 12 \times 4 = 192$.

The coefficient of x^{44} is also

$4 \times 12 \times 4 = 192$.

These are the numbers of oriented minimum and maximum cubes.

G.9 Expanded Distribution Polynomial

Expanding the exact product gives

$F(x) = 192x^{25}$

- $384x^{26}$
- $960x^{27}$
- $1{,}536x^{28}$
- $960x^{29}$
- $384x^{30}$
- $384x^{31}$
- $192x^{32}$
- $960x^{33}$
- $960x^{34}$
- $960x^{35}$
- $960x^{36}$
- $192x^{37}$
- $384x^{38}$
- $384x^{39}$
- $960x^{40}$
- $1{,}536x^{41}$
- $960x^{42}$
- $384x^{43}$
- $192x^{44}$.

Unlike the isotropic generating polynomial, the anisotropic polynomial contains every integer exponent from twenty-five through forty-four.

Every score in that interval is attained.

The distribution remains highly nonuniform. Some scores occur only 192 times, while the most frequent scores occur 1,536 times.

G.10 Complete Frequency Table

Let

N_s

denote the number of oriented anisotropic cubes having score s.

The exact distribution is:

$N_{25} = 192$;

$N_{26} = 384$;

$N_{27} = 960$;

$N_{28} = 1{,}536$;

$N_{29} = 960$;

$N_{30} = 384$;

$N_{31} = 384$;

$N_{32} = 192$;

$N_{33} = 960$;

$N_{34} = 960$;

$N_{35} = 960$;

$N_{36} = 960$;

$N_{37} = 192$;

$N_{38} = 384$;

$N_{39} = 384$;

$N_{40} = 960$;

$N_{41} = 1{,}536$;

$N_{42} = 960$;

$N_{43} = 384$;

and

$N_{44} = 192$.

The frequencies sum to

192 + 384 + 960 + 1,536 + 960 + 384 + 384 + 192

- 960 + 960 + 960 + 960
- 192 + 384 + 384 + 960 + 1,536 + 960 + 384 + 192

= 13,824.

Score	Frequency	Complementary score	Frequency
25	192	44	192
26	384	43	384
27	960	42	960
28	1,536	41	1,536
29	960	40	960
30	384	39	384
31	384	38	384
32	192	37	192
33	960	36	960
34	960	35	960

Table G1. Exact Anisotropic Synonymous-Edge Score Distribution. The table gives the number of oriented anisotropic codon cubes attaining each score from twenty-five through forty-four. The frequencies sum to 13,824. The distribution has mean and median 34.5, population variance 30.25, and population standard deviation 5.5. It is exactly symmetric under the transformation $s \to 69 - s$. Scores twenty-eight and forty-one have the largest frequency, 1,536. The minimum and maximum scores are each attained by 192 oriented cubes.

G.11 Complementary-Path Symmetry

At one codon position, a nucleotide path selects three of the six possible unordered nucleotide-pair types.

The three omitted pair types form a complementary path.

The total available synonymous weight at each codon position is:

4 at position 1;

1 at position 2;

and

64 at position 3.

Therefore, if a path has positional score L_p, its complementary path has score

$T_p - L_p$,

where

$T_1 = 4$,

$T_2 = 1$,

and

$T_3 = 64$.

For an anisotropic cube with score

$L = L_1 + L_2 + L_3$,

replace each axis path by its complement.

The complementary cube has score

L^c

$= (4 - L_1)$

- $(1 - L_2)$
- $(64 - L_3)$

$= 69 - L$.

Complementation is a bijection on the complete set of 13,824 oriented cubes.

Therefore,

$N_s = N_{69-s}$.

The frequency pairs are:

$N_{25} = N_{44} = 192$;

$N_{26} = N_{43} = 384$;

$N_{27} = N_{42} = 960$;

$N_{28} = N_{41} = 1{,}536$;

$N_{29} = N_{40} = 960$;

$N_{30} = N_{39} = 384$;

$N_{31} = N_{38} = 384$;

$N_{32} = N_{37} = 192$;

$N_{33} = N_{36} = 960$;

and

$N_{34} = N_{35} = 960$.

The exact center is

$69 / 2 = 34.5$.

Because the center is a half-integer, scores thirty-four and thirty-five form the central complementary pair.

The generating polynomial satisfies the reciprocal identity

$F(x) = x^{69} F(x^{-1})$.

G.12 Median

There are 13,824 oriented anisotropic cubes.

Half of this number is

6,912.

Complementary symmetry places exactly half of the observations below 34.5 and half above 34.5.

The cumulative frequency through score thirty-four is 6,912.

The next 960 cubes have score thirty-five.

Thus, the two central ordered observations lie at scores thirty-four and thirty-five.

The median is therefore

$(34 + 35) / 2 = 34.5$.

The mean and median coincide exactly:

mean = median = 34.5.

G.13 The Global Minimum

The minimum first-position score is

1.

The minimum second-position score is

1.

The minimum third-position score is

25.

Therefore, the minimum possible anisotropic score is

0 + 0 + 25 = 25.

The first-position minimum is attained by four oriented paths.

The second-position minimum is attained by twelve oriented paths.

The third-position minimum is attained by four oriented paths.

Consequently, the number of oriented minimum cubes is

4 × 12 × 4 = 192.

At the undirected level, the corresponding counts are:

two first-position minimum paths;

six second-position minimum paths;

and

two third-position minimum paths.

The number of undirected minimum cubes is therefore

2 × 6 × 2 = 24.

The minimum family is the complementary image of the maximum family.

G.14 The Global Maximum

The maximum positional scores are

max $L_1 = 4$,

max $L_2 = 1$,

and

max $L_3 = 39$.

Therefore,

$L(\pi_1, \pi_2, \pi_3)$

$\leq 4 + 1 + 39$

$= 44$.

The bound is attainable because the three axes may be optimized independently.

The number of oriented maximizing paths at each position is:

4 at position 1;

12 at position 2;

and

4 at position 3.

Therefore, the number of oriented global maxima is

$4 \times 12 \times 4 = 192$.

At the undirected level, the corresponding counts are:

2 first-position maximizers;

6 second-position maximizers;

and

2 third-position maximizers.

Thus, the number of undirected global maximum cubes is

$2 \times 6 \times 2 = 24$.

The proportion of all oriented anisotropic cubes attaining the maximum is

$192 / 13{,}824$

$= 1 / 72$

≈ 0.013889.

The same proportion holds in the undirected space:

24 / 1,728

= 1 / 72.

Approximately 1.39 percent of anisotropic cubes reach the exact maximum of forty-four.

G.15 First-Position Maximum Family

A first-position path attains score four only when it contains both weighted edges

A–C

and

C–U.

The two undirected maximizing paths are

G–A–C–U

and

A–C–U–G.

Their four oriented forms are

G–A–C–U;

U–C–A–G;

A–C–U–G;

and

G–U–C–A.

Every first-position maximizing path contains the segment

A–C–U

or its reverse.

G.16 Second-Position Maximum Family

A second-position path attains score one exactly when it contains the edge

A–G.

A fixed unordered edge occurs in six of the twelve undirected nucleotide paths and twelve of the twenty-four oriented paths.

Therefore, the second-position maximum family contains:

six undirected paths;

and

twelve oriented paths.

The second-position condition is weaker than the first- or third-position conditions. Half of all nucleotide paths satisfy it.

G.17 Third-Position Maximum Family

The third-position maximum is thirty-nine.

It is obtained by combining the two highest third-position weights,

C–U = 16

and

A–G = 14,

with one of the two connecting weight-nine edges,

A–C = 9

or

A–U = 9.

The two undirected maximizing paths are therefore

G–A–C–U

and

G–A–U–C.

Their four oriented forms are

G–A–C–U;

U–C–A–G;

G–A–U–C;

and

C–U–A–G.

Every third-position maximizer contains both A–G and C–U.

G.18 Cartesian Product of Maximum Families

Let

M_1,

M_2,

and

M_3

denote the oriented maximum families at the three codon positions.

Then the complete oriented maximum family is

$M_1 \times M_2 \times M_3$.

Its size is

$| M_1 \times M_2 \times M_3 |$

$= | M_1 | \, | M_2 | \, | M_3 |$

$= 4 \times 12 \times 4$

$= 192$.

The undirected maximum family is similarly the Cartesian product of positional undirected maximum sets:

$2 \times 6 \times 2 = 24$.

No additional compatibility condition is required because the axis paths are selected independently.

Every member of this Cartesian product has positional score triple

(4, 1, 39)

and total score forty-four.

G.19 The Unique Common Isotropic Path

An isotropic maximum requires the same undirected path to belong to all three positional maximum families.

Let the first-position undirected maximum set be

M_1^u

= { G–A–C–U, A–C–U–G }.

Let the third-position set be

M_3^u

= { G–A–C–U, G–A–U–C }.

The second-position maximum set M_2^u contains every undirected path having A–G as an edge.

The intersection is

$M_1^u \cap M_2^u \cap M_3^u$

= { G–A–C–U }.

The reverse

U–C–A–G

represents the same undirected path.

Thus, G–A–C–U is the unique undirected path that independently maximizes all three codon positions.

This statement does not mean that it is the only anisotropic maximum cube.

There are twenty-four undirected anisotropic maximum cubes.

It means that G–A–C–U is the only path that can be assigned identically to all three axes while attaining the position-specific maximum at every axis.

G.20 Relation to the Isotropic Distribution

The isotropic search contains only the diagonal subset

$\pi_1 = \pi_2 = \pi_3$

within the full anisotropic product space.

There are twenty-four oriented isotropic cubes compared with 13,824 oriented anisotropic cubes.

The isotropic scores are constrained because the same path must simultaneously determine all three positional contributions.

Only ten score values occur isotropically:

25, 27, 28, 32, 34, 35, 37, 41, 42, and 44.

In the anisotropic search, independent positional combinations fill every integer score from twenty-five through forty-four.

The isotropic maximum remains forty-four because G–A–C–U lies in the intersection of all three positional maximum families.

The isotropic minimum remains twenty-five because the complementary path A–U–G–C lies in the intersection of the three positional minimum families.

The equality of the isotropic and anisotropic endpoints is therefore a consequence of common optimizing and minimizing paths, not a general property of separable optimization problems.

G.21 Frequency of the Isotropic Cubes

The twenty-four oriented isotropic cubes form a very small subset of the anisotropic space:

24 / 13,824

= 1 / 576.

Only about 0.174 percent of oriented anisotropic axis triples are isotropic.

Among the 192 oriented anisotropic maxima, only two are isotropic:

G–A–C–U on all three axes;

and

U–C–A–G on all three axes.

Thus, the isotropic share of the oriented maximum family is

2 / 192

= 1 / 96.

At the undirected level, one of the twenty-four maximum cubes is isotropic:

G–A–C–U on all three axes.

The remaining twenty-three undirected maximizing cubes use at least two different axis paths.

The unique isotropic maximum is therefore embedded within a larger anisotropic maximum family.

G.22 The Maximum Theorem

The anisotropic result can be stated formally.

Theorem.
Let each of the three codon positions be assigned an arbitrary oriented linear path through G, A, C, and U. Construct the corresponding $4 \times 4 \times 4$ codon lattice and count nearest-neighbor edges whose endpoints translate to the same standard genetic-code output. Then the score satisfies

$25 \le L \le 44$.

Every integer score in this interval occurs.

The maximum forty-four is attained by 192 oriented axis triples, corresponding to twenty-four undirected triples.

Proof.
The score separates as

$L = L_1 + L_2 + L_3$.

The positional score ranges are

$L_1 \in \{ 0, 2, 4 \}$,

$L_2 \in \{ 0, 1 \}$,

and

$L_3 \in \{ 25, 26, 31, 33, 38, 39 \}$.

The minimum is therefore

$0 + 0 + 25 = 25$,

and the maximum is

$4 + 1 + 39 = 44$.

The product of the three exact positional generating polynomials has a nonzero coefficient for every exponent from twenty-five through forty-four, proving that every integer score in the interval occurs.

The coefficient of x^{44} is

$4 \times 12 \times 4 = 192$.

Identifying each axis path with its reverse divides this count by eight, giving

192 / 8 = 24

undirected maximum cubes. ■

G.23 Direct Enumeration

The generating-polynomial derivation should be verified by direct enumeration.

For each of the twenty-four oriented paths π_1:

for each of the twenty-four paths π_2;

and

for each of the twenty-four paths π_3;

construct the anisotropic axis triple

(π_1, π_2, π_3).

For every triple:

map the sixty-four codons into the corresponding 4 × 4 × 4 coordinate system;

construct the forty-eight edges associated with each axis;

verify that the total undirected edge count is 144;

translate the endpoints of every edge;

count synonymous first-axis, second-axis, and third-axis edges separately;

sum the three positional counts;

and add one to the frequency of the resulting total.

The direct enumeration should return exactly 13,824 cases.

Its frequency table must agree coefficient by coefficient with the expanded generating polynomial.

A separate implementation may calculate each score directly from the position-specific nucleotide-pair matrices:

$L (\pi_1, \pi_2, \pi_3)$

$= L_1 (\pi_1)$

- $L_2 (\pi_2)$

- L_3 (π_3).

Agreement between graph construction and compressed weight calculation provides an independent verification.

G.24 Exact Computational Checks

A correct anisotropic implementation should verify that:

there are exactly twenty-four oriented nucleotide paths;

there are exactly $24^3 = 13{,}824$ oriented axis triples;

there are exactly $12^3 = 1{,}728$ undirected axis triples;

every constructed cube contains sixty-four vertices;

every cube contains forty-eight edges per axis;

every cube contains 144 edges in total;

the first-position spectrum is 0:4, 2:16, 4:4;

the second-position spectrum is 0:12, 1:12;

the third-position spectrum is 25:4, 26:4, 31:4, 33:4, 38:4, 39:4;

the complete frequencies sum to 13,824;

the minimum score is twenty-five;

the maximum score is forty-four;

every score from twenty-five through forty-four occurs;

the mean is 34.5;

the population variance is 30.25;

the population standard deviation is 5.5;

$N_s = N_{69-s}$ for every score;

exactly 192 oriented cubes score forty-four;

exactly twenty-four undirected cubes score forty-four;

and

G–A–C–U is the unique undirected path common to all three positional maximum families.

Every one of these values can be checked using integer arithmetic.

G.25 Scope of the Result

The distribution is exhaustive within the declared anisotropic comparison class:

four nucleotide states;

one linear path on each codon axis;

independent path selection among the three axes;

undirected nearest-neighbor edges;

equal weighting of every selected edge;

and

translated identity as the synonymous criterion.

The result does not cover every possible nucleotide geometry.

A codon space constructed from cycles, stars, complete nucleotide graphs, weighted edges, directed substitutions, or context-dependent adjacency would define a different comparison class.

The score distribution would also change under a variant genetic code because the position-specific synonymous weights would change.

The present theorem should therefore be stated precisely:

forty-four is the exact global maximum among all standard-code codon cubes whose three axes are independently formed from linear paths through G, A, C, and U.

G.26 Summary

The complete anisotropic comparison contains

13,824 oriented codon cubes

or

1,728 undirected cubes.

Its score separates exactly into independent first-, second-, and third-position contributions.

The positional generating polynomials are

$F_1(x) = 4 + 16x^2 + 4x^4$,

$F_2(x) = 12 + 12x$,

and

$F_3(x)$

$= 4(x^{25} + x^{26} + x^{31} + x^{33} + x^{38} + x^{39})$.

Their product gives the exact complete distribution.

Every integer score from twenty-five through forty-four occurs.

The mean and median are 34.5.

The population variance is 30.25.

The population standard deviation is 5.5.

Complementation pairs score s with score 69 − s, producing exact distributional symmetry.

The global maximum is

4 + 1 + 39 = 44.

It is attained by:

192 oriented anisotropic cubes;

and

twenty-four undirected anisotropic cubes.

G–A–C–U is the unique undirected nucleotide path belonging to all three positional maximum families. It is therefore the unique common isotropic realization of the anisotropic global maximum.

Appendix H gives the complete distribution of the forty-four synonymous self-loops among the translated output classes.

Appendix H

Synonymous Self-Loop Multiplicities

H.1 Definition

The G–A–C–U codon cube contains sixty-four codon vertices and 144 undirected nearest-neighbor edges.

Each edge changes exactly one codon position by one of the three nucleotide adjacencies

G–A,

A–C,

or

C–U.

Let

$\tau : C \rightarrow \Omega$

be the standard genetic-code translation map.

A cube edge { c, d } is synonymous when

$\tau (c) = \tau (d)$.

When all codons with the same translated output are collapsed to one translated vertex, every synonymous codon edge becomes a self-loop at that vertex.

For translated output ω, define the self-loop multiplicity

$\ell\omega$

as the number of G–A–C–U cube edges having both endpoints in the codon fiber

$F\omega = \tau^{-1} (\omega)$.

Thus,

$\ell\omega$

counts codon-level edge instances rather than visually distinct loop symbols.

A translated vertex having multiplicity six represents six separate codon adjacencies, even if a simplified drawing displays them compactly.

The complete self-loop total is

$\Sigma\omega\in\Omega\ \ell\omega = 44$.

These forty-four self-loop instances are the synonymous edges counted throughout the isotropic and anisotropic analyses.

H.2 Complete Positional Decomposition

The forty-four self-loops are distributed among the three codon positions as

4 at the first position;

1 at the second position;

and

39 at the third position.

Therefore,

44 = 4 + 1 + 39.

The complete position-specific loop totals agree with the nucleotide-pair weights derived in Appendix E.

At the first position, the four loops arise from:

two A–C substitutions within arginine;

and

two C–U substitutions within leucine.

At the second position, the one loop is the stop-to-stop pair

UAA ↔ UGA.

At the third position, thirty-nine selected codon adjacencies preserve translated identity.

The loop distribution is therefore strongly concentrated at the third codon position.

H.3 Complete Multiplicity Distribution

The twenty-one translated output classes fall into six loop-multiplicity levels.

Loop multiplicity six:

arginine, R;

and

leucine, L.

Loop multiplicity four:

serine, S.

Loop multiplicity three:

alanine, A;

glycine, G;

proline, P;

threonine, T;

and

valine, V.

Loop multiplicity two:

isoleucine, I;

and

translation stop, X.

Loop multiplicity one:

cysteine, C;

aspartate, D;

glutamate, E;

phenylalanine, F;

histidine, H;

lysine, K;

asparagine, N;

glutamine, Q;

and

tyrosine, Y.

Loop multiplicity zero:

methionine, M;

and

tryptophan, W.

The total is

2 (6)

- 1 (4)
- 5 (3)
- 2 (2)
- 9 (1)
- 2 (0)

= 12 + 4 + 15 + 4 + 9

= 44.

Translated class	Codon degeneracy	Self-loop multiplicity
R	6	6
L	6	6
S	6	4
A	4	3
G	4	3
P	4	3
T	4	3
V	4	3
I	3	2
X	3	2
C	2	1
D	2	1
E	2	1
F	2	1
H	2	1
K	2	1
N	2	1
Q	2	1
Y	2	1
M	1	0
W	1	0
Total	**64**	**44**

Table H1. Degeneracy and G–A–C–U Self-Loop Multiplicity. The table compares the number of codons assigned to each translated output with the number of synonymous nearest-neighbor edge instances retained by the G–A–C–U codon cube. Arginine and leucine each contribute six

loops, serine contributes four, five four-codon families contribute three each, isoleucine and X contribute two each, nine two-codon amino-acid families contribute one each, and methionine and tryptophan contribute none. The loop multiplicities sum to forty-four.

H.4 Arginine

Arginine has six codons:

CGU,

CGC,

CGA,

CGG,

AGA,

and

AGG.

Its G–A–C–U self-loop multiplicity is

ℓR = 6.

Four arginine loops occur at the third codon position:

CGU ↔ CGC,

CGC ↔ CGA,

CGA ↔ CGG,

and

AGA ↔ AGG.

Two occur at the first codon position:

CGA ↔ AGA,

and

CGG ↔ AGG.

Thus, the positional decomposition is

ℓR = 2 + 0 + 4 = 6.

The four CGN codons form a third-position path

CGG–CGA–CGC–CGU

when read in G–A–C–U order.

The two AGR codons form the third-position edge

AGA ↔ AGG.

The first-position A–C edges connect the CGN and AGN portions at the A-ending and G-ending codons.

Arginine therefore forms one connected six-codon fiber within the G–A–C–U cube.

Its six loops arise from both third-position degeneracy and first-position bridges.

H.5 Leucine

Leucine also has six codons:

UUA,

UUG,

CUU,

CUC,

CUA,

and

CUG.

Its loop multiplicity is

ℓL = 6.

Four leucine loops occur at the third codon position:

UUA ↔ UUG,

CUU ↔ CUC,

CUC ↔ CUA,

and

CUA ↔ CUG.

Two occur at the first position:

UUA ↔ CUA,

and

UUG ↔ CUG.

Thus,

ℓL = 2 + 0 + 4 = 6.

The four CUN codons form the third-position path

CUG–CUA–CUC–CUU.

The UUA and UUG codons form another third-position edge.

The two C–U first-position edges connect the UUR and CUN portions at third-position A and G.

Like arginine, leucine forms a connected six-codon fiber having both internal third-position paths and first-position bridges.

H.6 Serine

Serine has six codons:

UCU,

UCC,

UCA,

UCG,

AGU,

and

AGC.

Its loop multiplicity is

ℓS = 4.

All four loops occur at the third codon position:

UCU ↔ UCC,

UCC ↔ UCA,

UCA ↔ UCG,

and

AGU ↔ AGC.

Thus,

ℓS = 0 + 0 + 4 = 4.

The four UCN codons form the complete third-position path

UCG–UCA–UCC–UCU.

The two AGY codons form the edge

AGU ↔ AGC.

Unlike the six-codon arginine and leucine fibers, the two serine regions are not connected by a synonymous G–A–C–U cube edge.

The serine fiber therefore has two connected components under the selected topology.

This explains why serine has the same codon degeneracy as arginine and leucine but fewer self-loops.

Degeneracy alone does not determine loop multiplicity.

The spatial arrangement of the codons within the cube also matters.

H.7 Four-Codon Families

Five translated classes have four codons forming complete third-position boxes:

alanine, A;

glycine, G;

proline, P;

threonine, T;

and

valine, V.

For each class, the four codons differ only at the third position.

Because the G–A–C–U nucleotide path contains three edges, every four-codon box becomes a three-edge path.

Each class therefore contributes exactly three self-loops.

Alanine

Alanine codons are

GCU,

GCC,

GCA,

and

GCG.

The three loops are

GCU ↔ GCC,

GCC ↔ GCA,

and

GCA ↔ GCG.

Therefore,

ℓA = 3.

Glycine

Glycine codons are

GGU,

GGC,

GGA,

and

GGG.

The three loops are

GGU ↔ GGC,

GGC ↔ GGA,

and

GGA ↔ GGG.

Therefore,

ℓG = 3.

Proline

Proline codons are

CCU,

CCC,

CCA,

and

CCG.

The three loops are

CCU ↔ CCC,

CCC ↔ CCA,

and

CCA ↔ CCG.

Therefore,

ℓP = 3.

Threonine

Threonine codons are

ACU,

ACC,

ACA,

and

ACG.

The three loops are

ACU ↔ ACC,

ACC ↔ ACA,

and

ACA ↔ ACG.

Therefore,

ℓT = 3.

Valine

Valine codons are

GUU,

GUC,

GUA,

and

GUG.

The three loops are

GUU ↔ GUC,

GUC ↔ GUA,

and

GUA ↔ GUG.

Therefore,

ℓV = 3.

For all five four-codon families, the positional decomposition is

0 + 0 + 3.

The four synonymous codons form one connected third-position path.

H.8 Isoleucine

Isoleucine has three codons:

AUU,

AUC,

and

AUA.

Its two synonymous G–A–C–U cube edges are

AUU ↔ AUC,

and

AUC ↔ AUA.

Therefore,

ℓI = 2.

Both loops occur at the third position:

ℓI = 0 + 0 + 2.

The three codons occupy the consecutive nucleotide states

U–C–A

within the full G–A–C–U third-position path.

They therefore form one connected three-vertex path.

The remaining third-position state G gives AUG, which translates to methionine rather than isoleucine.

H.9 Translation Stop X

The three standard stop codons are

UAA,

UAG,

and

UGA.

Their loop multiplicity is

ℓX = 2.

One loop occurs at the third position:

UAA ↔ UAG.

This pair differs by A–G at the third nucleotide.

The second loop occurs at the second position:

UAA ↔ UGA.

This pair differs by A–G at the second nucleotide.

Thus,

ℓX = 0 + 1 + 1 = 2.

The three stop codons do not form a straight path along one codon axis. Instead, UAA serves as a central codon joined to:

UAG along the third axis;

and

UGA along the second axis.

The X fiber therefore forms a two-edge right-angle configuration in codon space.

This categorical topology does not depend on assigning X a hydropathy value.

X is a genuine translated output class for the graph analysis even though it is not an amino acid.

H.10 Two-Codon Amino-Acid Families

Nine amino acids have two standard codons and contribute one loop each:

phenylalanine, F;

tyrosine, Y;

cysteine, C;

histidine, H;

glutamine, Q;

asparagine, N;

lysine, K;

aspartate, D;

and

glutamate, E.

Every loop occurs at the third codon position.

Phenylalanine

UUU ↔ UUC.

Therefore,

ℓF = 1.

Tyrosine

UAU ↔ UAC.

Therefore,

ℓY = 1.

Cysteine

UGU ↔ UGC.

Therefore,

ℓC = 1.

Histidine

CAU ↔ CAC.

Therefore,

ℓH = 1.

Glutamine

CAA ↔ CAG.

Therefore,

ℓQ = 1.

Asparagine

AAU ↔ AAC.

Therefore,

ℓN = 1.

Lysine

AAA ↔ AAG.

Therefore,

$\ell K = 1$.

Aspartate

GAU ↔ GAC.

Therefore,

$\ell D = 1$.

Glutamate

GAA ↔ GAG.

Therefore,

$\ell E = 1$.

The U/C-ending pairs use the selected third-position edge

C–U.

The A/G-ending pairs use the selected third-position edge

G–A.

Each two-codon family therefore forms one connected edge within the G–A–C–U cube.

H.11 Single-Codon Amino Acids

Methionine and tryptophan each have one standard codon.

Methionine is encoded by

AUG.

Tryptophan is encoded by

UGG.

A one-codon fiber contains no pair of distinct codons and therefore cannot contain a synonymous codon edge.

Thus,

$\ell M = 0$

and

ℓW = 0.

The absence of loops is a direct consequence of degeneracy one.

H.12 Complete Class Table

The complete degeneracy and loop data are:

Output	Codons	First	Second	Third	Total loops
A	4	0	0	3	3
C	2	0	0	1	1
D	2	0	0	1	1
E	2	0	0	1	1
F	2	0	0	1	1
G	4	0	0	3	3
H	2	0	0	1	1
I	3	0	0	2	2
K	2	0	0	1	1
L	6	2	0	4	6
M	1	0	0	0	0
N	2	0	0	1	1
P	4	0	0	3	3
Q	2	0	0	1	1
R	6	2	0	4	6
S	6	0	0	4	4
T	4	0	0	3	3
V	4	0	0	3	3

Output	Codons	First	Second	Third	Total loops
W	1	0	0	0	0
X	3	0	1	1	2
Y	2	0	0	1	1
Total	**64**	**4**	**1**	**39**	**44**

The codon counts sum to sixty-four.

The positional loop columns sum to

4,

1,

and

39.

The total loop column sums to forty-four.

H.13 Degeneracy and Topological Compactness

Codon degeneracy and self-loop multiplicity are related but not identical.

A class with one codon must have zero loops.

A two-codon class can have at most one loop and contributes one only if its two codons are adjacent under the selected nucleotide path.

A four-codon third-position box contributes three loops because the four nucleotide states form a path rather than a complete graph.

For the six-codon classes, topology becomes especially informative.

Arginine, leucine, and serine all have degeneracy six.

Yet their loop multiplicities are:

arginine: 6;

leucine: 6;

serine: 4.

Arginine and leucine possess first-position bridges connecting codon subfamilies.

Serine does not.

The six serine codons occupy two disconnected components under the G–A–C–U cube topology.

Thus, equal fiber cardinality does not imply equal internal connectivity.

The loop count records how the codons are arranged, not merely how many codons the class contains.

H.14 Maximum Possible Fiber Edges

A translated class with degeneracy d contains

$C (d, 2)$

possible unordered pairs of distinct codons.

Only some of those pairs differ at one codon position, and only some of those one-position differences belong to the G–A–C–U path.

For a six-codon class, there are

$C (6, 2) = 15$

possible internal pairs.

Arginine and leucine retain six of them as cube edges.

Serine retains four.

For a four-codon class, there are

$C (4, 2) = 6$

possible pairs, but the linear nucleotide path retains three.

For a three-codon class, there are three possible pairs.

Isoleucine and X each retain two.

For a two-codon class, there is one possible pair, and all nine two-codon amino-acid families retain it.

The loop multiplicity is therefore a joint property of:

fiber size;

codon Hamming geometry;

and

the selected nucleotide adjacency path.

H.15 Connected Components of the Synonymous Fibers

The self-loop multiplicity does not by itself give the complete internal graph shape. Connected components provide additional information.

Arginine forms one connected six-vertex synonymous subgraph.

Leucine forms one connected six-vertex synonymous subgraph.

Serine forms two connected components:

one four-vertex UCN path;

and

one two-vertex AGU–AGC edge.

Each four-codon family forms one connected four-vertex path.

Isoleucine forms one connected three-vertex path.

X forms one connected three-vertex, two-edge angled configuration.

Every two-codon family forms one connected edge.

Methionine and tryptophan each form one isolated singleton vertex.

Therefore, the standard translated fibers are not all geometrically equivalent even when their sizes agree.

H.16 Pair-Type Contributions to the Loops

Every G–A–C–U self-loop uses one of three nucleotide-pair types:

A–G;

A–C;

or

C–U.

Their total loop contributions are:

A–G: 15;

A–C: 11;

and

C–U: 18.

Therefore,

15 + 11 + 18 = 44.

The C–U adjacency contributes the largest number of self-loops.

A–G contributes the second largest.

A–C contributes the third largest.

These are exactly the three total pair weights selected by the G–A–C–U path.

The loop multiplicity distribution across translated classes is therefore another decomposition of the same forty-four-edge total described by the nucleotide-pair weights.

The pair-weight view groups synonymous edges by nucleotide substitution.

The present appendix groups them by translated output.

Both partitions sum to forty-four.

H.17 Complete List of the Forty-Four Loop Edges

For exact reproducibility, the complete loop set is listed below by translated output.

A:
GCU ↔ GCC;
GCC ↔ GCA;
GCA ↔ GCG.

C:
UGU ↔ UGC.

D:
GAU ↔ GAC.

E:
GAA ↔ GAG.

F:
UUU ↔ UUC.

G:
GGU ↔ GGC;

GGC ↔ GGA;
GGA ↔ GGG.

H:
CAU ↔ CAC.

I:
AUU ↔ AUC;
AUC ↔ AUA.

K:
AAA ↔ AAG.

L:
UUA ↔ UUG;
UUA ↔ CUA;
UUG ↔ CUG;
CUU ↔ CUC;
CUC ↔ CUA;
CUA ↔ CUG.

N:
AAU ↔ AAC.

P:
CCU ↔ CCC;
CCC ↔ CCA;
CCA ↔ CCG.

Q:
CAA ↔ CAG.

R:
CGU ↔ CGC;
CGC ↔ CGA;
CGA ↔ CGG;
CGA ↔ AGA;
CGG ↔ AGG;
AGA ↔ AGG.

S:
UCU ↔ UCC;
UCC ↔ UCA;

UCA ↔ UCG;
AGU ↔ AGC.

T:
ACU ↔ ACC;
ACC ↔ ACA;
ACA ↔ ACG.

V:
GUU ↔ GUC;
GUC ↔ GUA;
GUA ↔ GUG.

X:
UAA ↔ UAG;
UAA ↔ UGA.

Y:
UAU ↔ UAC.

Methionine and tryptophan contribute no loop edges.

The list contains exactly forty-four unordered codon pairs.

H.18 Direct Verification Procedure

A direct implementation should construct the G–A–C–U cube and perform the following steps.

For every cube edge { c, d }:

verify that c and d differ at exactly one codon position;

verify that the differing nucleotide pair is one of

A–G,

A–C,

or

C–U;

compare τ (c) and τ (d);

when the outputs are equal, increment the loop count for that translated class;

and increment the count for the changed codon position.

The resulting class totals must match Table H1.

The implementation should verify:

twenty-one translated classes;

sixty-four codons across all fibers;

forty-four synonymous cube edges;

four first-position loops;

one second-position loop;

thirty-nine third-position loops;

six loops each for R and L;

four for S;

three each for A, G, P, T, and V;

two each for I and X;

one each for C, D, E, F, H, K, N, Q, and Y;

and zero each for M and W.

A second implementation can group codons by translated output first and then test every internal codon pair for cube adjacency.

The two procedures should produce exactly the same loop list.

H.19 Relation to the Translated Multigraph

In the translated multigraph, the forty-four synonymous codon edges become forty-four self-loop edge instances distributed over nineteen translated vertices.

Methionine and tryptophan have no loops.

All other translated classes have at least one.

The same multigraph also contains one hundred nonsynonymous edge instances joining distinct translated vertices.

Therefore, the total edge-instance count is

44 + 100 = 144.

The loop multiplicities must be retained when the multigraph is analyzed quantitatively.

Replacing six arginine loops with one unweighted loop symbol would destroy edge-count information.

A diagram may consolidate repeated loops visually, but the underlying data structure must retain multiplicity.

H.20 Summary

The G–A–C–U codon cube contains forty-four synonymous nearest-neighbor edges.

After quotienting codons by translated identity, these become forty-four self-loop edge instances.

Their positional decomposition is

4 + 1 + 39 = 44.

Arginine and leucine contribute six loops each.

Serine contributes four.

Alanine, glycine, proline, threonine, and valine contribute three each.

Isoleucine and X contribute two each.

Nine two-codon amino-acid classes contribute one each.

Methionine and tryptophan contribute none.

Equal codon degeneracy does not guarantee equal loop multiplicity. Arginine, leucine, and serine all have six codons, but serine has fewer loops because its codon fiber is divided into two disconnected regions under the selected topology.

The forty-four loops can be partitioned equivalently:

by codon position;

by nucleotide-pair type;

or

by translated output class.

Each partition reveals a different aspect of the same exact synonymous-edge structure.

Appendix I specifies the computational protocol required to reproduce the complete matrix, pseudoinverse, hydropathy projection, path enumeration, and self-loop results.

Appendix I

Computational Reproduction Protocol

I.1 Purpose

The principal calculations in this book are finite and fully reproducible.

They require no stochastic optimization, inaccessible experimental apparatus, or proprietary numerical method. The complete codon matrix, exact pseudoinverse, hydropathy projection, nucleotide-pair weights, codon-cube graphs, isotropic enumeration, anisotropic score distribution, and synonymous self-loop multiplicities can all be regenerated from explicitly declared inputs.

The computational package should reproduce the results rather than merely store them.

A correct implementation must begin from:

the four-letter RNA alphabet;

the complete set of sixty-four codons;

the standard genetic-code translation map;

the declared codon and coordinate orderings;

the selected numerical amino-acid property table;

and the definition of nucleotide adjacency used in the cube calculations.

Every derived object should be constructed from those inputs and checked against exact invariants.

The protocol should fail closed. If a required count, matrix identity, or finite enumeration does not agree with the declared result, the program should stop rather than continue with partially inconsistent data.

I.2 Recommended Directory Structure

A reproducible project may use the following directory organization:

README.md

LICENSE

CITATION.cff

environment.yml

or

requirements.txt

data/

src/

tests/

results/

figures/

tables/

notebooks/

docs/

The data directory should contain only authoritative inputs or clearly identified derived tables.

The src directory should contain reusable computational functions rather than one monolithic script.

The tests directory should contain automated checks of exact invariants.

The results directory should contain machine-readable outputs from complete analyses.

The figures and tables directories should contain generated publication files together with provenance information.

Exploratory notebooks may be retained in notebooks, but the final published calculations should also be executable from ordinary scripts or command-line entry points. A result should not depend on hidden notebook state or an undocumented execution order.

I.3 Authoritative Input Files

The minimum authoritative input set should include the following files.

Standard genetic-code table

A machine-readable table with one row for every RNA codon and at least two columns:

codon

and

output.

The codons should use U rather than T.

Translation stop should be represented consistently as

X.

There must be exactly sixty-four rows and sixty-four distinct codons.

Amino-acid property table

For the hydropathy calculation, the file should contain:

the twenty standard amino-acid single-letter codes;

their complete numerical Kyte–Doolittle hydropathy values;

the bibliographic source;

the numerical precision retained;

and the explicit stop convention used in the complete calculation.

For the present analysis,

H (X) = 0.

The value for X should be marked as an analytical convention rather than an amino-acid measurement.

Optional residue-volume table

The residue-volume source used for the historical two-property figure should be stored separately and should include:

the residue code;

the numerical value;

the units;

the original source;

and any transformation applied before plotting.

Sequence file for the membrane example

The original coding sequence and translated amino-acid sequence used in the historical membrane-profile comparison should be retained in a standard format such as FASTA.

The record should identify:

the sequence source;

the reading frame;

the nucleotide coordinate convention;

whether the sequence uses DNA or RNA notation;

and the relation between the stored sequence and the plotted profile.

I.4 Canonical Alphabet and Coordinate Order

The computational package must declare one canonical nucleotide order.

For this book, the preferred order is

G, A, C, U.

The twelve one-hot matrix columns are then

1G, 1A, 1C, 1U,

2G, 2A, 2C, 2U,

3G, 3A, 3C, 3U.

Changing the coordinate order does not change the abstract mathematics, but it changes the visual form of:

the codon matrix;

the pseudoinverse;

the coefficient vector;

and every exported table based on coordinate position.

The software should never infer the coordinate order from a language-dependent dictionary or unordered set.

It should define the order explicitly.

For example, in pseudocode:

alphabet = [G, A, C, U]

coordinates = [(1,G), (1,A), (1,C), (1,U), ..., (3,U)]

Every later routine should receive or import this same canonical ordering.

I.5 Canonical Codon Order

The codon rows should also have one declared order.

A convenient convention is lexicographic order under

G < A < C < U,

with the first position changing slowest and the third position changing fastest.

The codon list is generated by the triple product

for c_1 in G, A, C, U:

for c_2 in G, A, C, U:

for c_3 in G, A, C, U:

append $c_1c_2c_3$.

The first codons are therefore

GGG,

GGA,

GGC,

GGU,

GAG,

GAA,

GAC,

GAU,

and so forth.

The final codon is

UUU.

The exact row order does not affect rank, singular values, or graph counts. It does affect:

alignment between D and the output vector;

alignment between D and numerical property vectors;

the displayed matrix image;

the pseudoinverse column order;

and reproducibility of saved arrays.

Every saved codon-level vector should therefore include codon labels or inherit one documented canonical order.

I.6 Construction of the Complete Codon List

The program should generate the complete codon set algorithmically rather than relying only on a manually entered list.

The following conditions must be checked:

the list contains sixty-four entries;

all entries have length three;

every character belongs to { G, A, C, U };

every codon is unique;

and the set equals the complete Cartesian product Σ^3.

The code-table input should then be validated against the generated codon set.

There must be:

no missing codon;

no duplicated codon;

no extra codon;

and no codon containing T in the canonical RNA table.

A DNA display version may replace U with T only after the canonical RNA relation has been validated.

I.7 Validation of the Translation Map

The standard translation table should be loaded into a mapping

$\tau : C \rightarrow \Omega$.

The software should verify that every codon has exactly one translated output.

The output alphabet must contain:

the twenty standard amino-acid single-letter codes;

and

X.

There should be exactly twenty-one distinct translated classes.

The standard degeneracy counts should be verified.

The six-codon classes are:

L, R, and S.

The four-codon classes are:

A, G, P, T, and V.

The three-codon classes are:

I and X.

The two-codon classes are:

C, D, E, F, H, K, N, Q, and Y.

The one-codon classes are:

M and W.

The total must be

3 (6)

- 5 (4)
- 2 (3)
- 9 (2)
- 2 (1)

= 64.

Failure of this check indicates either a code-table error or use of a nonstandard variant code.

Variant codes should be stored and analyzed as explicitly named alternatives rather than silently substituted for the standard code.

I.8 Construction of the 64 × 12 Incidence Matrix

Initialize a zero matrix

D

with dimensions

64 × 12.

For each codon row c and coordinate column (p, n), set

D c,(p,n) = 1

when the nucleotide at codon position p equals n.

Otherwise set the entry to zero.

The construction should be followed immediately by exact checks:

shape(D) == (64, 12)

every entry belongs to { 0, 1 };

every row sum equals 3;

every column sum equals 16;

the total number of 1s is 192;

the total number of 0s is 576;

and the three four-column positional block sums each equal the sixty-four-component all-ones vector.

The matrix should be saved with row and column labels in a format that preserves exact integers.

Suitable formats include:

CSV with labels;

tab-separated text;

or

a structured binary format accompanied by a metadata file.

I.9 Exact Gram Matrix

Compute

$G = D^T D$

using integer arithmetic.

Verify that the diagonal 4×4 blocks equal

$16 I_4$

and the off-diagonal blocks equal

$4 J_4$.

The program should compare G against the explicitly constructed block matrix rather than checking only selected entries.

Additional checks are:

trace(G) == 192

rank(G) == 10

and the exact characteristic spectrum is

48 once;

16 nine times;

and

0 twice.

A symbolic algebra system may verify the eigenvalues exactly.

A numerical library may also be used, but the program should compare the numerical results with the known exact values and use a tolerance appropriate to floating-point representation.

The two zero eigenvalues are exact. They should not be treated as small positive quantities.

I.10 Rank and Singular Values

Compute the singular values of D for visualization or independent confirmation.

The expected spectrum is

$4 \sqrt{3}$;

4 repeated nine times;

and

0 repeated twice.

The implementation should verify:

ten nonzero singular values;

rank ten;

right nullity two;

and

left nullity fifty-four.

If a numerical rank function is used, the chosen tolerance should be recorded.

The exact rank should also be checked independently by symbolic row reduction or exact rational arithmetic.

The two right-null vectors may be verified directly:

$z_1 = (1_4, -1_4, 0_4)^T$,

and

$z_2 = (1_4, 0_4, -1_4)^T$.

The program should confirm

$D\, z_1 = 0$

and

$D\, z_2 = 0$.

I.11 Construction of the Exact Pseudoinverse

Construct the pseudoinverse directly from the closed form

$D^+ = (1 / 16) D^T - (1 / 96) J_{12 \times 64}$.

Exact rational arithmetic is preferred.

The implementation should not rely only on a generic floating-point pseudoinverse function. A generic function may be used as an independent comparison, but the published matrix should be generated from the exact formula.

Every entry must belong to the set

{ 5 / 96, −1 / 96 }.

The program should verify:

dimensions 12 × 64;

192 entries equal to 5 / 96;

576 entries equal to −1 / 96;

sixteen positive entries in every row;

forty-eight negative entries in every row;

three positive entries in every column;

nine negative entries in every column;

row sums equal to 1 / 3;

and column sums equal to 1 / 16.

The complete pseudoinverse may be exported as rational strings to preserve exactness.

I.12 Moore–Penrose Verification

The four Moore–Penrose equations must be checked exactly:

$D\,D^{+}\,D = D$;

$D^{+}\,D\,D^{+} = D^{+}$;

$(D\,D^{+})^{T} = D\,D^{+}$;

and

$(D^{+}\,D)^{T} = D^{+}\,D$.

The codon-space projector

$P = D\,D^{+}$

should satisfy

$P^{2} = P$,

$P^{T} = P$,

and

$\mathrm{tr}(P) = 10$.

The coefficient-space projector

$Q = D^{+} D$

should satisfy

$Q^{2} = Q$,

$Q^{T} = Q$,

and

$\mathrm{tr}(Q) = 10$.

These tests should use exact equality when rational arithmetic is available.

If floating-point arrays are used for a secondary comparison, the maximum absolute residual for every identity should be recorded.

I.13 Hamming-Distance Projector Test

For every pair of codons c and d, calculate their Hamming distance

$d_H(c, d)$.

Verify that the projector entry $P_{c,d}$ equals:

$5 / 32$ when $d_H = 0$;

$3 / 32$ when $d_H = 1$;

$1 / 32$ when $d_H = 2$;

and

$-1 / 32$ when $d_H = 3$.

Around every fixed codon, verify that there are:

one codon at distance zero;

nine at distance one;

twenty-seven at distance two;

and

twenty-seven at distance three.

Every row of P must sum to one.

This provides a combinatorial verification independent of direct matrix multiplication.

I.14 Construction of the Hydropathy Field

Load the amino-acid hydropathy table and validate that it contains exactly one value for each of the twenty amino-acid codes.

Add the declared stop convention:

$H (X) = 0.$

Construct the sixty-four-component codon hydropathy field

$h c = H (\tau (c)).$

The codon labels must remain attached to the values.

The software should record:

the hydropathy scale name;

the bibliographic source;

the value assigned to X;

and the numerical precision of all inputs.

The complete field should be saved as a table containing:

codon;

translated output;

and hydropathy value.

I.15 Hydropathy Coefficients

Compute

$\beta = D^{+} h.$

The coefficient table should contain:

codon position;

nucleotide;

coefficient value;

and a stable coordinate label such as 2U.

For the complete Kyte–Doolittle calculation with

$H(X) = 0$,

the second-position values must include:

$\beta_{2,G} = -1.469792$;

$\beta_{2,A} = -2.638542$;

$\beta_{2,C} = -0.163542$;

and

$\beta_{2,U} = 3.948958$,

subject only to the numerical precision retained in the input property table.

The gauge-invariant contrast must be

$\beta_{2,U} - \beta_{2,A} = 6.587500$.

The program should also calculate the positional coefficient ranges.

The second-position range must be

6.587500.

The complete twelve-value table should be exported rather than retaining only the dominant coefficients.

I.16 Projection and Residual

Compute the fitted codon field

$\hat{h} = D\beta$

or equivalently

$\hat{h} = P h$.

The two calculations must agree.

Compute the residual

$r = h - \hat{h}$.

Verify the normal-equation condition

$D^{T} r = 0$.

The program should calculate:

the global mean of h;

the centered field h°;

the centered fitted field P h°;

the centered residual;

the fitted squared norm;

the residual squared norm;

and the additive projection fraction

Qadd

= ‖ P h° ‖² / ‖ h° ‖².

For the declared complete calculation, the expected value is approximately

0.816882.

The exact displayed precision should be regenerated from the authoritative hydropathy input table.

I.17 Fiber-Averaged Reconstruction

Group the fitted codon values by translated output.

For every amino acid ω, compute

$\hat{H}(\omega)$

= (1 / dω) Σc∈Fω $\hat{h}$ c.

The primary amino-acid-level comparison should use the twenty amino acids and exclude X unless a separate stop analysis is intended.

The output table should contain:

amino-acid code;

empirical hydropathy;

number of codons;

mean reconstructed hydropathy;

minimum reconstructed codon value;

maximum reconstructed codon value;

within-fiber standard deviation;

and residual.

The twenty empirical and reconstructed values should reproduce the stated regression:

y = −0.003660 + 0.859718 x,

with

r = 0.915646

and

R^2 = 0.838407,

when the same authoritative dataset and equal amino-acid weighting are used.

The program should state which variable is plotted on each axis and which variable is treated as the regression response.

I.18 Membrane-Profile Reconstruction

The nucleotide-only membrane profile requires a separate, fully specified procedure.

For each codon in the correct reading frame, inspect only its second nucleotide.

Assign the score:

+1 for U in RNA or T in DNA;

−1 for A;

and

0 for G or C.

Let the resulting codon-level sequence be

$s_1, s_2, \ldots, sN$.

Apply the declared sliding-window rule.

For a window of width w, one possible centered profile is

M i

= (1 / w) Σj∈Wi s j,

where Wi is the set of codon indices in the window centered at i.

The exact implementation must state:

the window width;

whether the sum or mean is displayed;

the treatment of incomplete terminal windows;

whether the output is assigned to the center or leading edge;

whether further smoothing is applied;

and whether the plotted sign is U − A or A − U.

The conventional Kyte–Doolittle profile should be generated from the translated amino-acid sequence using the same window placement wherever possible.

The two profiles should be saved numerically, not only as an image.

I.19 Definition of a Nucleotide Path

Represent an oriented nucleotide path as an ordered tuple of four distinct symbols.

For example:

(G, A, C, U).

The corresponding unordered edge set is

{ { G, A }, { A, C }, { C, U } }.

A path-validation function should verify:

length four;

all four nucleotides present exactly once;

and exactly three adjacent unordered pairs.

Generate all oriented paths as the twenty-four permutations of the alphabet.

A canonical undirected representative may be chosen by comparing each path with its reverse and retaining the lexicographically smaller tuple.

This should produce exactly twelve undirected paths.

I.20 Construction of a Codon Cube

For one path π, map each nucleotide to its coordinate index:

0, 1, 2, or 3.

A codon $c_1c_2c_3$ is mapped to

(indexπ (c_1), indexπ (c_2), indexπ (c_3)).

Two codons are cube neighbors when their coordinate triples differ:

by exactly one in one axis;

and

by zero in the other two axes.

The graph must contain:

sixty-four vertices;

forty-eight first-axis edges;

forty-eight second-axis edges;

forty-eight third-axis edges;

and

144 undirected edges.

Every edge should be stored in canonical endpoint order to prevent accidental duplication.

For an anisotropic cube, use separate coordinate maps for π_1, π_2, and π_3.

The edge-count invariants remain unchanged.

I.21 Direct Synonymous-Edge Scoring

For every cube edge { c, d }, compare

τ (c)

and

τ (d).

If the outputs are equal, classify the edge as synonymous.

Otherwise classify it as nonsynonymous.

Record:

the changed codon position;

the nucleotide pair;

the translated endpoint labels;

and whether the edge is synonymous.

For the isotropic G–A–C–U cube, the exact totals must be:

144 edges;

44 synonymous edges;

100 nonsynonymous edges;

4 synonymous first-position edges;

1 synonymous second-position edge;

and

39 synonymous third-position edges.

The program should also verify that every edge changes one of the selected nucleotide pairs:

A–G;

A–C;

or

C–U.

I.22 Position-Specific Pair Weights

Construct the eighteen position-specific nucleotide-pair weights independently of the path cube.

For every codon position p and unordered nucleotide pair { x, y }:

iterate over the sixteen settings of the other two codon positions;

construct the two codons differing only by x versus y at position p;

compare their translated outputs;

and count the synonymous pairs.

The exact tuple values, ordered by nucleotide pair, are:

A–G: (0, 1, 14);

A–C: (2, 0, 9);

A–U: (0, 0, 9);

C–G: (0, 0, 8);

C–U: (2, 0, 16);

and

G–U: (0, 0, 8).

The positional totals must be:

4;

1;

and

64.

The grand total must be

69.

The total pair weights must be:

A–G = 15;

A–C = 11;

A–U = 9;

C–G = 8;

C–U = 18;

and

G–U = 8.

I.23 Independent Isotropic Scoring

Every isotropic path should be scored in two independent ways.

Direct graph method

Construct the complete 144-edge codon cube and count synonymous edges.

Pair-weight method

Sum the three total pair weights selected by the nucleotide path.

The two scores must agree for all twenty-four oriented paths.

The program should export a table containing:

oriented path;

reverse path;

canonical undirected representative;

selected edge types;

first-position score;

second-position score;

third-position score;

and total score.

The expected undirected score list is:

G–A–C–U: 44;

G–A–U–C: 42;

A–G–C–U: 41;

A–G–U–C: 41;

A–C–U–G: 37;

A–U–C–G: 35;

C–A–G–U: 34;

C–G–A–U: 32;

C–A–U–G: 28;

G–C–A–U: 28;

A–C–G–U: 27;

and

A–U–G–C: 25.

There must be exactly one undirected maximum path:

G–A–C–U.

I.24 Complementary-Path Verification

For every undirected path, calculate the three nucleotide-pair edges not used by that path.

Verify that those three omitted edges form another Hamiltonian path on the four nucleotide vertices.

The two path scores must sum to

69.

The maximum path G–A–C–U must pair with the minimum path A–U–G–C:

44 + 25 = 69.

The isotropic frequency polynomial should satisfy the reciprocal identity

$I (x) = x^{69} I (x^{-1})$.

The software need not use symbolic polynomials internally, but the coefficient symmetry should be checked exactly.

I.25 Anisotropic Enumeration

Generate the Cartesian product of the twenty-four oriented paths on each of the three codon axes.

The total number of axis triples must be

$24^3 = 13{,}824$.

For every triple

(π_1, π_2, π_3),

calculate the score as

$L_1 (\pi_1) + L_2 (\pi_2) + L_3 (\pi_3)$.

A direct cube construction may also be performed as an independent verification.

The implementation should record:

the three oriented paths;

their canonical undirected representatives;

the three positional scores;

and the total score.

The positional spectra must be:

first position:

0 with multiplicity 4;

2 with multiplicity 16;

4 with multiplicity 4;

second position:

0 with multiplicity 12;

1 with multiplicity 12;

third position:

25, 26, 31, 33, 38, and 39, each with multiplicity 4.

I.26 Exact Anisotropic Frequency Table

The complete score frequencies must be:

25: 192;

26: 384;

27: 960;

28: 1,536;

29: 960;

30: 384;

31: 384;

32: 192;

33: 960;

34: 960;

35: 960;

36: 960;

37: 192;

38: 384;

39: 384;

40: 960;

41: 1,536;

42: 960;

43: 384;

and

44: 192.

The frequencies must sum to

13,824.

The program should verify:

minimum 25;

maximum 44;

mean 34.5;

median 34.5;

population variance 30.25;

population standard deviation 5.5;

and

N s = N 69−s.

All calculations can be performed with exact integers and rational numbers until the final square root is displayed.

I.27 Maximum Families

Identify every path attaining the positional maximum.

The first-position maximum is four. Its undirected maximum set is:

G–A–C–U;

and

A–C–U–G.

The second-position maximum is one. Its undirected maximum set contains the six paths having A–G as an edge.

The third-position maximum is thirty-nine. Its undirected maximum set is:

G–A–C–U;

and

G–A–U–C.

The program should verify:

four oriented first-position maximizers;

twelve oriented second-position maximizers;

four oriented third-position maximizers;

192 oriented global maxima;

and

twenty-four undirected global maxima.

The undirected intersection of all three positional maximum sets must be exactly

{ G–A–C–U }.

I.28 Self-Loop Multiplicities

For the G–A–C–U cube, group all synonymous edges by translated output.

The expected multiplicities are:

R = 6;

L = 6;

S = 4;

A = 3;

G = 3;

P = 3;

T = 3;

V = 3;

I = 2;

X = 2;

C = 1;

D = 1;

E = 1;

F = 1;

H = 1;

K = 1;

N = 1;

Q = 1;

Y = 1;

M = 0;

and

W = 0.

The total must be forty-four.

The positional loop totals must be:

first position 4;

second position 1;

third position 39.

The complete edge list given in Appendix H should be reproduced exactly after endpoint order is canonicalized.

I.29 Translated Multigraph

Construct the translated multigraph by mapping every codon edge

{ c, d }

to the translated endpoint pair

{ τ (c), τ (d) }.

Retain all edge instances.

When the two translated endpoints are equal, store the edge as a self-loop.

When they differ, store it as an interclass edge.

Do not collapse parallel edge instances into one unweighted edge unless multiplicity is retained as an attribute.

The exact totals are:

twenty-one translated vertices;

forty-four self-loop instances;

one hundred interclass edge instances;

and

144 edge instances overall.

A graph-export format such as GraphML may be used, provided that:

loop edges are supported;

parallel edges are retained;

codon-pair provenance is attached to every edge;

and translated labels are preserved.

I.30 Figure Generation

Every generated figure should be reproducible from machine-readable data.

The figure-generation script should record:

input file names;

software and library versions;

figure dimensions;

resolution;

font sizes;

axis labels;

data transformations;

and output format.

Publication figures should not be edited manually in ways that alter data values or graph topology.

Cosmetic editing is acceptable only when it does not change:

node identity;

edge identity;

edge multiplicity;

coordinate labels;

numerical values;

or relative ordering.

For the authoritative codon cube and translated multigraph, the verified graph structure should be retained exactly.

A redrawn figure must be checked against the underlying edge list before publication.

I.31 Table Generation

Tables should be generated from the same computational outputs used for analysis.

The following tables should not be retyped manually:

the twelve hydropathy coefficients;

position-specific nucleotide-pair weights;

the complete isotropic enumeration;

the anisotropic score distribution;

and

the degeneracy and self-loop multiplicities.

The generation script should round only for display.

Full-precision values should remain in the machine-readable results.

Every displayed total should be calculated from the underlying rows rather than entered independently.

For example, Table E1 should calculate:

first-position total 4;

second-position total 1;

third-position total 64;

and grand total 69.

I.32 Unit Tests

The project should contain automated unit tests for at least the following components:

codon generation;

translation-table validation;

matrix construction;

Gram matrix;

rank and singular spectrum;

null vectors;

exact pseudoinverse;

Moore–Penrose equations;

projector values;

hydropathy-field construction;

coefficient calculation;

path generation;

cube-edge generation;

pair-weight calculation;

isotropic enumeration;

anisotropic enumeration;

and loop multiplicities.

Each test should isolate one property as far as possible.

A failure in the anisotropic distribution should not be hidden by a test that checks only the final maximum.

Exact finite outputs permit more demanding tests than approximate scientific computations usually allow.

I.33 Integration Tests

Integration tests should reproduce complete analysis pipelines.

One integration test should begin from the alphabet and generate:

all codons;

D;

D^+;

P;

and all matrix invariants.

A second should begin from the standard code table and generate:

the pair weights;

the G–A–C–U cube;

the forty-four synonymous edges;

the translated multigraph;

and the self-loop multiplicities.

A third should generate:

all twenty-four isotropic paths;

their scores;

the unique maximum;

and the complement relation.

A fourth should generate:

all 13,824 anisotropic triples;

the complete frequency distribution;

the maximum family;

and the unique common undirected path.

A fifth should begin from the hydropathy input table and regenerate:

h;

β;

ĥ;

r;

the additive projection fraction;

the fiber-averaged reconstruction;

and the amino-acid regression.

I.34 Independent Implementations

The most important exact results should be verified by at least two computational routes.

For the pseudoinverse:

one implementation should use the closed-form rational formula;

another may use symbolic singular-value decomposition or an exact generalized inverse routine.

For isotropic scores:

one implementation should construct codon-cube edges;

another should sum total nucleotide-pair weights.

For anisotropic scores:

one implementation should enumerate all cube edges for every axis triple;

another should convolve the three positional score distributions or multiply their generating polynomials.

For self-loop multiplicities:

one implementation should group synonymous cube edges by translated output;

another should enumerate all internal pairs within each translated fiber and test cube adjacency.

Agreement among independent implementations greatly reduces the chance that one shared indexing error has propagated through the entire analysis.

I.35 Randomization Protocols

Any future random-code analysis should preserve the random seed and the full sampling rule.

The metadata must state:

the ensemble definition;

which features are held fixed;

which features are permuted;

the number of random samples;

the pseudorandom generator;

the generator version;

the seed or seed sequence;

and the treatment of duplicate samples.

A result described as degeneracy preserving must state whether it preserves:

only the multiset of degeneracy sizes;

the exact degeneracy of each named output;

the stop count;

the stop locations;

codon boxes;

or other structures.

The randomized codes themselves need not all be distributed if the sample is extremely large, but the procedure must permit exact regeneration.

Summary statistics without a reconstructible ensemble definition are insufficient.

I.36 Variant-Code Protocol

Natural variant genetic codes should be stored as separate input tables.

Each file should include:

the code name;

the biological source or taxonomy;

the reference;

the codon assignments that differ from the standard code;

and the complete sixty-four-codon mapping.

The complete incidence matrix D remains unchanged.

The program should recompute all assignment-dependent quantities:

translated fibers;

hydropathy field;

coefficients;

pair weights;

isotropic scores;

anisotropic distribution;

maximum families;

and self-loop multiplicities.

The standard-code outputs should never be overwritten.

Every result file should contain an explicit code identifier.

I.37 Software Environment

The project should record the exact software environment.

For a Python implementation, this may include:

Python version;

NumPy version;

SciPy version;

pandas version;

SymPy version;

Matplotlib version;

NetworkX version;

and the operating system.

A lock file or environment specification should permit recreation of the environment.

The exact matrix and graph counts should be independent of library versions when implemented correctly. Recording the environment nevertheless improves provenance and helps diagnose changes in formatting, plotting, or numerical defaults.

I.38 Command-Line Reproduction

The full analysis should be executable through documented commands.

For example:

python -m genetic_code.build_matrix

python -m genetic_code.verify_pseudoinverse

python -m genetic_code.project_hydropathy

python -m genetic_code.enumerate_isotropic

python -m genetic_code.enumerate_anisotropic

python -m genetic_code.build_multigraph

python -m genetic_code.make_figures

python -m genetic_code.run_all

The exact command names may differ, but the project should provide one command that regenerates all tables, figures, and verification reports from the authoritative inputs.

The command should create a manifest listing every output file and its checksum.

I.39 Result Manifest and Checksums

Every released result package should include a manifest containing:

file path;

file size;

cryptographic checksum;

creation time;

software version;

input-data version;

and the command used to generate it.

Checksums help distinguish:

an authoritative generated file;

a manually edited copy;

and

a later regenerated version.

The manifest should include the machine-readable data behind every figure and table.

An image alone is not sufficient archival evidence of the calculation.

I.40 Provenance of Historical Figures

Historical figures that cannot yet be regenerated exactly should be marked as historical.

Their provenance record should include:

the earliest known file;

the associated paper, notebook, or program;

the known input sequence or dataset;

the known plotting method;

and any unresolved uncertainty.

Such a figure should not be silently treated as though it were generated by the current package.

When an exact reconstruction becomes available, the historical and reconstructed versions should both be retained.

The new figure should identify its method explicitly rather than replacing the historical artifact without documentation.

I.41 Fail-Closed Master Verification

The full pipeline should end with a master verification report.

The report should contain at least the following assertions:

Input domain

64 codons;

12 coordinates;

21 translated outputs.

Matrix

row sums all equal 3;

column sums all equal 16;

rank 10;

right nullity 2;

left nullity 54.

Singular spectrum

4 √3 once;

4 nine times;

0 twice.

Pseudoinverse

entry set { 5 / 96, −1 / 96 };

all four Moore–Penrose equations satisfied.

Projector

entry set { 5 / 32, 3 / 32, 1 / 32, −1 / 32 };

trace 10;

rank 10.

Pair weights

position totals 4, 1, and 64;

grand total 69.

G–A–C–U cube

144 total edges;

44 synonymous;

100 nonsynonymous;

positional synonymous counts 4, 1, and 39.

Isotropic search

24 oriented paths;

12 undirected paths;

maximum 44;

minimum 25;

unique undirected maximum G–A–C–U.

Anisotropic search

13,824 oriented cubes;

1,728 undirected cubes;

score range 25–44;

mean 34.5;

population standard deviation 5.5;

192 oriented maxima;

24 undirected maxima;

unique common path G–A–C–U.

Multigraph

21 translated vertices;

44 loop instances;

100 interclass edge instances.

The release should be considered valid only when every assertion passes.

I.42 Reproducibility Report

The software should write a human-readable report summarizing:

the input files used;

their checksums;

the canonical orders;

the numerical conventions;

the exact matrix invariants;

the hydropathy results;

the graph counts;

the isotropic ranking;

the anisotropic distribution;

and the final pass-or-fail status.

This report should be suitable for inclusion in an archive or supplementary repository.

It should not require the user to inspect raw console output to determine whether the analysis succeeded.

I.43 Archival Release

A complete archival release should contain:

source code;

input data;

machine-readable results;

generated tables;

generated figures;

test reports;

environment specification;

result manifest;

checksums;

citation information;

and a clear license.

A permanent release should be assigned a version number.

The version should change whenever:

an input table changes;

a numerical convention changes;

a calculation changes;

a figure is regenerated from different data;

or an error is corrected.

Cosmetic documentation changes may use a patch-level update, while analytical changes should receive a more prominent version increment.

I.44 Minimum Reproduction Standard

A reader should be able to reproduce the central results without relying on unpublished intermediate files.

At minimum, the package must permit an independent user to regenerate:

the complete 64 × 12 matrix;

the exact pseudoinverse;

the hydropathy coefficient table;

the G–A–C–U cube edge list;

the forty-four synonymous edges;

the complete isotropic score table;

the complete anisotropic frequency table;

and the self-loop multiplicity table.

A package that displays final images but cannot regenerate their underlying numerical data does not meet this standard.

I.45 Summary

The mathematics of the genetic code developed in this book is unusually well suited to exact computational reproduction.

The input domain is finite.

The incidence matrix contains integers.

The pseudoinverse contains rational numbers.

The path spaces are completely enumerable.

The graph counts are exact.

The principal verification conditions can therefore be tested without statistical tolerance.

The recommended protocol begins from authoritative codon and property tables, declares all coordinate orders and conventions, generates every derived object algorithmically, and checks each result against redundant invariants.

Independent computational routes should agree on the central outputs.

Every published figure and table should remain traceable to machine-readable data and a specific generation script.

The final pipeline should fail closed unless all exact conditions are satisfied.

Reproducibility is not an accessory to the analysis. It is part of the mathematical claim.

Appendix J

Notation

J.1 Purpose and Scope

This appendix consolidates the symbols used throughout the book.

The same genetic-code relation is represented in several mathematical forms:

a complete codon set;

a categorical translation map;

a binary incidence matrix;

a numerical property field;

a codon lattice;

and

a translated multigraph.

Symbols are grouped according to those roles. Unless explicitly stated otherwise, all codons use RNA notation with U rather than DNA notation with T.

J.2 Basic Sets

Nucleotide alphabet

$\Sigma = \{ G, A, C, U \}$.

The symbol Σ denotes the four-state RNA nucleotide alphabet.

The displayed order

G–A–C–U

may serve either as:

a coordinate convention for matrix columns;

or

a declared linear nucleotide path in the codon-cube analysis.

These two uses are related but logically distinct. A column order labels coordinates. A path defines adjacency.

Codon set

$C = \Sigma^3$.

The set C contains all ordered three-nucleotide words over Σ.

Its cardinality is

$| C | = 4^3 = 64$.

A codon is written

$c = c_1 c_2 c_3$,

where

$c_p \in \Sigma$

for

$p \in \{ 1, 2, 3 \}$.

Translated-output alphabet

Ω denotes the set of translated output classes.

For the standard genetic code,

Ω

contains the twenty standard amino-acid single-letter codes together with translation stop X.

Therefore,

$| \Omega | = 21$.

The symbol X is used consistently for translation stop in place of an asterisk.

J.3 Codon Positions and Nucleotides

The symbol

p

denotes a codon position:

$p \in \{ 1, 2, 3 \}$.

The symbols

n, x, y

denote nucleotide states in Σ.

A nucleotide-position coordinate is written

(p, n).

Examples include:

$(1, G)$;

$(2, A)$;

and

$(3, U)$.

In compact labels, the same coordinates are written:

1G;

2A;

and

3U.

The complete coordinate order used for the binary matrix is

1G, 1A, 1C, 1U |

2G, 2A, 2C, 2U |

3G, 3A, 3C, 3U.

J.4 Translation Map

The standard genetic-code relation is written

$\tau : C \to \Omega$.

For codon c,

$\tau (c)$

is its translated amino-acid or stop output.

For example,

$\tau (AUG) = M$,

and

$\tau (UGA) = X$.

The map τ is many-to-one because several codons may share the same translated output.

J.5 Translated Fibers and Degeneracy

For translated output

$\omega \in \Omega$,

the codon fiber is

$F\omega = \tau^{-1} (\omega)$.

Thus,

$F\omega$

is the set of all codons translating to ω.

The degeneracy of ω is

$d\omega = | F\omega |$.

For example,

$dL = 6$,

$dM = 1$,

and

$dX = 3$

under the standard genetic code.

The term **fiber** emphasizes that synonymous codons are preimages of the same translated output under τ.

J.6 Binary Codon Incidence Matrix

The complete binary codon matrix is

$D \in \{0,1\}^{64 \times 12}$.

Rows correspond to codons.

Columns correspond to nucleotide-position coordinates.

The entry

$D_{c,(p,n)}$

is defined by

$D_{c,(p,n)} = 1$

when

$c_p = n$,

and

$D_{c,(p,n)} = 0$

otherwise.

Every row contains three 1s and nine 0s.

Every column contains sixteen 1s and forty-eight 0s.

The three positional blocks are written

$D = [\, D_1 \mid D_2 \mid D_3 \,]$,

where each

D_p

has dimensions

64×4.

J.7 Vectors of Ones

The symbol

$1k$

denotes the k-component column vector whose entries are all one.

Examples include

1_4,

1_{12},

and

1_{64}.

When dimensions are clear from context, the subscript may be omitted.

The complete codon matrix satisfies

$D\,1_{12} = 3\,1_{64}$

and

$D^T\,1_{64} = 16\,1_{12}$.

Within each positional block,

$D_p\,1_4 = 1_{64}$.

J.8 Identity and All-Ones Matrices

The symbol

Ik

denotes the $k \times k$ identity matrix.

The symbol

$Jm{\times}n$

denotes the $m \times n$ matrix whose entries are all one.

When the matrix is square, the notation may be shortened to

Jk.

Examples include

I_4,

J_4,

$J_{12\times 64}$,

and

J_{64}.

The matrices I and J should not be confused with the amino-acid single-letter code I for isoleucine. Context distinguishes the matrix symbol from the translated output label.

J.9 Gram Matrix

The coefficient-space Gram matrix is

$G = D^T D$.

It has dimensions

12×12.

Its exact block structure is

$$G = \begin{bmatrix} 16\, I_4 & 4\, J_4 & 4\, J_4 \\ 4\, J_4 & 16\, I_4 & 4\, J_4 \\ 4\, J_4 & 4\, J_4 & 16\, I_4 \end{bmatrix}.$$

The symbol G in this context denotes the Gram matrix. The nucleotide G is written as an alphabet symbol or appears in a codon or coordinate label. The amino-acid single-letter code G denotes glycine. The surrounding notation distinguishes these uses.

J.10 Rank and Null Spaces

The matrix rank is written

rank (D).

For the complete codon matrix,

rank (D) = 10.

The right null space is

null (D)

and has dimension two.

The left null space is

null (D^{T})

and has dimension fifty-four.

A convenient right-null basis is

$z_{1} = (1_{4}, -1_{4}, 0_{4})^{T}$,

and

$z_{2} = (1_{4}, 0_{4}, -1_{4})^{T}$.

A general right-null vector has the form

z

$= (a\ 1_{4}, b\ 1_{4}, c\ 1_{4})^{T}$,

with

$a + b + c = 0$.

J.11 Eigenvalues and Singular Values

The symbol

λ

denotes an eigenvalue.

The symbol

σ

denotes a singular value or, when explicitly identified, a population standard deviation.

For the Gram matrix $D^{T} D$, the eigenvalues are:

48 with multiplicity one;

16 with multiplicity nine;

and

0 with multiplicity two.

For D, the singular values are:

4 √3 with multiplicity one;

4 with multiplicity nine;

and

0 with multiplicity two.

Multiplicity indicates how many independent directions share the same value.

J.12 Moore–Penrose Pseudoinverse

The Moore–Penrose pseudoinverse of D is written

D^{+}.

It has dimensions

12 × 64.

Its exact form is

D^{+}

$= (1 / 16) D^{T}$

$- (1 / 96) J_{12 \times 64}$.

The superscript plus denotes the pseudoinverse, not ordinary addition.

The entry

$D^{+}p,n;c$

equals

5 / 96

when codon c contains nucleotide n at position p,

and

−1 / 96

otherwise.

The pseudoinverse satisfies the four Moore–Penrose equations:

$D D^{+} D = D;$

$D^{+} D D^{+} = D^{+}$;

$(D D^{+})^{T} = D D^{+}$;

and

$(D^{+} D)^{T} = D^{+} D$.

J.13 Projection Matrices

The codon-space projector is

$P = D D^{+}$.

It has dimensions

64×64

and projects an arbitrary codon field onto the additive nucleotide-position subspace.

Its exact form is

P

$= (1 / 16) D D^{T}$

$- (1 / 32) J_{64}$.

The coefficient-space projector is

$Q = D^{+} D$.

It has dimensions

12×12

and projects an arbitrary coefficient vector onto the row space of D.

Both matrices are symmetric and idempotent:

$P^{T} = P$,

$P^{2} = P$,

$Q^{T} = Q$,

and

$Q^{2} = Q$.

The ranks and traces are

rank (P) = tr (P) = 10,

and

rank (Q) = tr (Q) = 10.

J.14 Hamming Distance

The Hamming distance between codons c and d is written

dH (c, d).

It is the number of codon positions at which the two codons differ.

For three-letter codons,

dH (c, d) ∈ { 0, 1, 2, 3 }.

The entries of the codon-space projector depend only on this distance:

Pc,d = 5 / 32 when dH = 0;

Pc,d = 3 / 32 when dH = 1;

Pc,d = 1 / 32 when dH = 2;

and

Pc,d = −1 / 32 when dH = 3.

The number of matching coordinates is

m (c, d) = 3 − dH (c, d).

J.15 Numerical Property Functions

A numerical property assigned to translated outputs is written

H : Ω → ℝ.

For hydropathy, H assigns a hydropathy value to each amino acid and uses a declared convention for X.

In the complete calculation,

H (X) = 0.

The induced codon-level property field is written

h ∈ $\mathbb{R}^{64}$,

with entries

$h_c = H(\tau(c))$.

Uppercase H denotes the property function on translated outputs.

Lowercase h denotes the corresponding sixty-four-component codon field.

J.16 Additive Coefficients

The twelve-component nucleotide-position coefficient vector is written

$\beta \in \mathbb{R}^{12}$.

It is obtained from the property field by

$\beta = D^{+} h$.

Its coordinate associated with position p and nucleotide n is

$\beta_{p,n}$.

Examples include

$\beta_{2,U}$

and

$\beta_{2,A}$.

The additive fitted value for codon

$c = c_1 c_2 c_3$

is

$\beta_{1,c_1} + \beta_{2,c_2} + \beta_{3,c_3}$.

The same field is written in matrix form as

$D \beta$.

J.17 Fitted Field and Residual

The fitted additive codon field is written

$\hat{h}$.

It is defined by

$\hat{h} = D \beta$

$= D D^{+} h$

$= P h.$

The residual is written

$r = h - \hat{h}.$

The residual is orthogonal to the additive column space:

$D^{T} r = 0.$

The original field decomposes as

$h = \hat{h} + r.$

Because the fitted field and residual are orthogonal,

$\| h \|^2 = \| \hat{h} \|^2 + \| r \|^2.$

J.18 Centered Fields and Means

The global mean of a codon-level field is written

μ

and defined by

$\mu = (1 / 64) 1_{64}^{T} h.$

The centered field is written

h°

and defined by

$h^{\circ} = h - \mu 1_{64}.$

The superscript degree symbol indicates centering, not exponentiation.

The additive projection fraction is

$Qadd$

$= \| P h^{\circ} \|^2 / \| h^{\circ} \|^2.$

For the complete Kyte–Doolittle field with

$H (X) = 0,$

$Qadd \approx 0.816882.$

This value is distinct from an amino-acid-level regression R^2.

J.19 Fiber-Averaged Property Reconstruction

The reconstructed property of translated output ω is written

$\hat{H}(\omega)$.

It is the mean of the fitted codon values over the fiber F_ω:

$\hat{H}(\omega) = (1 / d_\omega) \Sigma_{c \in F_\omega} \hat{h}\, c$.

The corresponding fiber-averaged residual is

$R(\omega) = H(\omega) - \hat{H}(\omega)$.

The within-fiber fitted variance may be written

$\hat{V}_\omega = (1 / d_\omega) \Sigma_{c \in F_\omega} [\hat{h}\, c - \hat{H}(\omega)]^2$.

The symbol R in this expression denotes a translated-property residual. It should not be confused with the amino-acid label R for arginine; context distinguishes them.

J.20 Regression Quantities

The symbol

r

may denote the codon-level residual vector in matrix sections.

In statistical sections, italic r may also denote the Pearson correlation coefficient. The surrounding expression makes the meaning explicit.

The coefficient of determination is written

R^2.

The fitted amino-acid-level regression used in the complete hydropathy calculation is

$y = -0.003660 + 0.859718\, x$,

with

$r = 0.915646$

and

$R^2 = 0.838407$.

The variables x and y must be identified in the accompanying figure or text whenever the regression is reported.

J.21 Norms and Inner Products

The Euclidean norm of a vector v is written

$\| v \|$.

Its squared norm is

$\| v \|^2$.

The Frobenius norm of a matrix A is written

$\| A \|_F$.

The standard inner product of vectors u and v is

$u^T v$.

Orthogonality is written

$u^T v = 0$.

No weighting is implied unless a weighted norm or inner product is explicitly defined.

J.22 Nucleotide Paths

An oriented linear nucleotide path is written

$\pi = n_1 – n_2 – n_3 – n_4$,

where each nucleotide appears exactly once.

Examples include

G–A–C–U

and

A–U–G–C.

The reverse path is written

πR

and defined by

$\pi R = n_4 – n_3 – n_2 – n_1$.

The reverse has the same undirected edge set and the same score under the unweighted topology.

An undirected path identifies π and πR as equivalent.

There are:

twenty-four oriented paths;

and

twelve undirected paths.

J.23 Isotropic and Anisotropic Cubes

An isotropic cube uses one path π on all three codon axes.

An anisotropic cube uses the ordered path triple

(π_1, π_2, π_3).

Here,

π_1

is assigned to the first codon position,

π_2

to the second,

and

π_3

to the third.

The number of oriented anisotropic axis triples is

$24^3 = 13{,}824$.

The number of undirected anisotropic axis triples is

$12^3 = 1{,}728$.

J.24 Position-Specific Pair Weights

For codon position p and unordered nucleotide pair { x, y }, the synonymous pair weight is

$w_p (x, y)$.

It counts synonymous codon pairs that:

differ only at position p;

use nucleotides x and y at that position;

and agree at the other two positions.

The weight is symmetric:

$w_p (x, y) = w_p (y, x)$.

The total isotropic pair weight is

$w (x, y)$

$= w_1 (x, y)$

- $w_2 (x, y)$
- $w_3 (x, y)$.

The position-specific weights may also be stored in symmetric matrices

W_1,

W_2,

and

W_3.

Their sum is

$W = W_1 + W_2 + W_3$.

J.25 Synonymous Path Scores

The position-specific score of path π is

$L_p (\pi)$.

For

$\pi = n_1–n_2–n_3–n_4$,

$L_p (\pi)$

$= w_p (n_1, n_2)$

- $w_p (n_2, n_3)$
- $w_p (n_3, n_4)$.

The isotropic total is

$L (\pi)$

$= L_1 (\pi) + L_2 (\pi) + L_3 (\pi)$.

Equivalently,

$L (\pi)$

$= w (n_1, n_2)$

- $w (n_2, n_3)$
- $w (n_3, n_4)$.

For an anisotropic cube,

$L (\pi_1, \pi_2, \pi_3)$

$= L_1 (\pi_1)$

- $L_2 (\pi_2)$
- $L_3 (\pi_3)$.

The exact positional maxima are

$\max L_1 = 4$,

$\max L_2 = 1$,

and

$\max L_3 = 39$.

Therefore,

$\max L = 44$.

J.26 Complementary Paths

The complement of a four-nucleotide path π is written

π^{c}.

It contains the three nucleotide-pair edges omitted by π from the complete six-edge nucleotide graph.

For isotropic scores,

$L(\pi) + L(\pi^{c}) = 69$.

At one codon position,

$L_{1}(\pi) + L_{1}(\pi^{c}) = 4$,

$L_{2}(\pi) + L_{2}(\pi^{c}) = 1$,

and

$L_{3}(\pi) + L_{3}(\pi^{c}) = 64$.

For an anisotropic cube with total score L, complementing all three axis paths gives score

$69 - L$.

J.27 Score Frequencies

The notation

N_{s}

denotes the number of oriented anisotropic cubes having total score s.

The exact complement symmetry is

$N_{s} = N_{69-s}$.

The score range is

$25 \le s \le 44$.

The frequencies satisfy

$\Sigma_{s=25}^{44} N_{s} = 13{,}824$.

The minimum and maximum frequencies include

$N_{25} = 192$

and

$N_{44} = 192$.

J.28 Generating Polynomials

The first-position score polynomial is

$F_1 (x)$

$= 4 + 16 x^2 + 4 x^4$.

The second-position polynomial is

$F_2 (x)$

$= 12 + 12 x$.

The third-position polynomial is

$F_3 (x)$

$= 4 (x^{25} + x^{26} + x^{31} + x^{33} + x^{38} + x^{39})$.

The complete anisotropic generating polynomial is

$F (x)$

$= F_1 (x) F_2 (x) F_3 (x)$.

The coefficient of x^s in $F (x)$ is Ns.

The isotropic oriented score polynomial is written

$I (x)$.

The undirected isotropic polynomial is written

$Iu (x)$.

Complement symmetry produces reciprocal identities such as

$F (x) = x^{69} F (x^{-1})$.

J.29 Maximum Families

The oriented position-specific maximum families are written

M_1,

M_2,

and

M_3.

Their undirected forms may be written

M_1^u,

M_2^u,

and

M_3^u.

The complete oriented anisotropic maximum family is

$M_1 \times M_2 \times M_3$.

Its size is

$4 \times 12 \times 4 = 192$.

The undirected maximum family has size

$2 \times 6 \times 2 = 24$.

The unique common undirected path is expressed by

$M_1^u \cap M_2^u \cap M_3^u$

$= \{ G–A–C–U \}$.

J.30 Codon-Cube Graphs

A codon-cube vertex is a codon $c \in C$.

A cube edge is an unordered codon pair

$\{ c, d \}$

whose members differ by one adjacent path step at exactly one codon position.

The edge set may be written

$E\pi$

for an isotropic path π

or

$E\pi_1,\pi_2,\pi_3$

for an anisotropic path triple.

Every linear-path codon cube contains

$| E | = 144$

undirected edges.

The positional edge counts are

$48 + 48 + 48 = 144$.

J.31 Synonymous and Nonsynonymous Edges

A cube edge $\{ c, d \}$ is synonymous when

$\tau (c) = \tau (d)$.

It is nonsynonymous when

$\tau (c) \neq \tau (d)$.

For the isotropic G–A–C–U cube:

synonymous edges = 44;

nonsynonymous edges = 100;

and

total edges = 144.

The complete Hamming-distance-one codon graph contains

288

undirected edges, of which

69

are synonymous.

J.32 Translated Multigraph

The translated multigraph is obtained by mapping every codon vertex c to its translated output $\tau (c)$.

Its translated vertex set is Ω.

Parallel edge instances are retained.

A synonymous codon edge becomes a self-loop.

A nonsynonymous codon edge becomes an edge between two distinct translated vertices.

For the G–A–C–U cube, the translated multigraph contains:

twenty-one translated vertices;

forty-four self-loop instances;

one hundred interclass edge instances;

and

144 edge instances in total.

J.33 Self-Loop Multiplicity

The self-loop multiplicity of translated output ω is written

ℓω.

It is the number of synonymous G–A–C–U cube edges whose two endpoints both lie in Fω.

Thus,

ℓω

is a count of codon-level edge instances.

The total is

Σω∈Ω ℓω = 44.

Examples include

ℓR = 6,

ℓL = 6,

ℓS = 4,

ℓX = 2,

and

ℓM = ℓW = 0.

J.34 Probability and Frequency Notation

Combinatorial frequencies in this book are exact counts unless explicitly normalized.

For example,

$N_{44} = 192$

is an exact count, not a statistical estimate.

A proportion may be written as a ratio such as

192 / 13,824 = 1 / 72.

The symbols

μ

and

σ

denote the population mean and population standard deviation when the complete finite comparison space has been enumerated.

For the anisotropic distribution,

$\mu = 34.5$

and

$\sigma = 5.5$.

These are population quantities for all 13,824 oriented cubes, not sample estimates.

J.35 Combinatorial Symbols

The notation

$| A |$

denotes the cardinality of set A.

The binomial coefficient

$C (n, k)$

denotes the number of k-element subsets of an n-element set.

For example,

C (6, 2) = 15.

The factorial

n!

denotes

n (n − 1) … 2 × 1.

Thus,

4! = 24.

The Stirling number of the second kind is written

S (n, k).

It counts partitions of an n-element labeled set into k nonempty unlabeled blocks.

The number of surjective maps from sixty-four labeled codons onto twenty-one labeled outputs is

21! S (64, 21).

J.36 Summation Notation

The symbol

Σ

is used both for the nucleotide alphabet and for mathematical summation. The context distinguishes them.

As an alphabet,

Σ = { G, A, C, U }.

As a summation symbol, it appears with an index, as in

Σc∈C h c.

A summation over a translated fiber is written

Σc∈Fω.

A summation over nucleotide states is written

Σn∈Σ.

J.37 Transpose and Superscripts

The superscript

^{T}

denotes matrix or vector transpose.

Thus,

D^{T}

has dimensions

12×64.

The superscript

-

denotes the Moore–Penrose pseudoinverse.

The superscript

R

on a path denotes reversal.

The superscript

c

on a path denotes graph complement.

The superscript

u

on a maximum family denotes an undirected family.

The superscript degree symbol in

h°

denotes centering.

A superscript 2 may denote either an ordinary square or, in R^2, the coefficient of determination.

J.38 Equality, Approximation, and Definition

The symbol

=

denotes exact equality.

The symbol

≈

denotes numerical approximation.

Exact rational results should be displayed with equality. For example,

$D^{+}p,n;c = 5 / 96$.

Rounded empirical or decimal values should use approximation when appropriate. For example,

$4 \sqrt{3} \approx 6.928203$.

A quantity introduced by definition may be written with an ordinary equality when the prose explicitly states that it is being defined.

J.39 RNA and DNA Notation

The canonical genetic-code analysis uses RNA codons and therefore uses U.

DNA coding sequences may use T in place of U.

The middle-position membrane profile may therefore be written as:

U − A

for RNA;

or

T − A

for DNA.

These are the same nucleotide-state contrast after transcriptional notation is accounted for.

No separate mathematical code is implied by the substitution of T for U in DNA sequence displays.

J.40 Amino-Acid Single-Letter Codes

The translated output alphabet uses the standard amino-acid single-letter codes:

A alanine;

C cysteine;

D aspartate;

E glutamate;

F phenylalanine;

G glycine;

H histidine;

I isoleucine;

K lysine;

L leucine;

M methionine;

N asparagine;

P proline;

Q glutamine;

R arginine;

S serine;

T threonine;

V valine;

W tryptophan;

Y tyrosine;

and

X translation stop.

The symbol X is not treated as an amino acid. It is retained as the categorical translation-stop output.

J.41 Scope Conventions

Unless otherwise stated:

all codon sets contain all sixty-four RNA codons;

the standard genetic code is used;

X denotes translation stop;

the matrix D uses the twelve-coordinate order declared above;

codons are weighted equally in the complete matrix and codon-field analyses;

nucleotide paths are undirected for topology unless orientation is being counted explicitly;

cube edges have equal weight;

and

synonymous means equality of translated output, including X-to-X equality.

Any departure from these conventions should be stated where it occurs.

J.42 Summary of Principal Symbols

The central symbols are:

Σ — RNA nucleotide alphabet;

C — complete sixty-four-codon set;

Ω — twenty amino acids plus X;

τ — standard translation map;

Fω — codon fiber of output ω;

dω — degeneracy of output ω;

D — 64 × 12 binary codon matrix;

G — Gram matrix $D^T D$;

D^+ — exact Moore–Penrose pseudoinverse;

P — codon-space projector $D D^+$;

Q — coefficient-space projector $D^+ D$;

H — numerical property function on translated outputs;

h — codon-level property field;

β — nucleotide-position coefficient vector;

$\hat{h}$ — fitted additive codon field;

r — residual field;

π — nucleotide path;

$w_p (x, y)$ — position-specific synonymous pair weight;

$L_p (\pi)$ — position-specific path score;

L — total synonymous-edge score;

Ns — anisotropic score frequency;

F (x) — anisotropic generating polynomial;

ℓω — self-loop multiplicity of translated output ω;

and

dH (c, d) — codon Hamming distance.

These symbols connect the algebraic, physicochemical, and topological representations of the complete genetic-code relation.

Appendix K

Glossary

K.1 Additive Codon Field

An **additive codon field** is a numerical function on the sixty-four codons that can be written as the sum of three independent positional contributions.

For codon

$c = c_1 c_2 c_3$,

an additive field has the form

$f (c)$

$= f_1 (c_1)$

- $f_2 (c_2)$

- $f_3 (c_3)$.

Each codon position contributes one nucleotide-dependent term. No pairwise or three-way interaction is included.

The column space of the complete 64 × 12 codon incidence matrix D is exactly the ten-dimensional space of additive codon fields.

K.2 Additive Projection

The **additive projection** of a codon-level numerical field h is its closest representation within the additive codon-field space.

It is

$\hat{h} = D D^{+} h$.

The matrix

$P = D D^{+}$

is the orthogonal projector onto the additive subspace.

The difference

$r = h - \hat{h}$

is the residual component orthogonal to every additive nucleotide-position field.

K.3 Alphabet

An **alphabet** is a finite set of symbols used to construct words.

For the RNA codon system,

$\Sigma = \{ G, A, C, U \}$.

The alphabet has four symbols.

The complete codon set consists of every ordered word of length three over this alphabet.

K.4 Amino-Acid Reconstruction

An **amino-acid reconstruction** is the value obtained by averaging fitted codon-level property values over all codons that translate to one amino acid.

For output ω,

$\hat{H} (\omega)$

$= (1 / d\omega) \Sigma c \in F\omega \hat{h}\, c.$

The reconstruction measures how well a translated property can be recovered from the additive nucleotide-position representation.

It is distinct from the fitted value of any one codon.

K.5 Anisotropic Codon Cube

An **anisotropic codon cube** uses independently chosen nucleotide paths on the three codon axes.

It is specified by

(π_1, π_2, π_3),

where π_1 governs first-position adjacency, π_2 governs second-position adjacency, and π_3 governs third-position adjacency.

There are

$24^3 = 13{,}824$

oriented anisotropic cubes when every axis may use any oriented linear order of G, A, C, and U.

The total synonymous score separates exactly into three independent positional contributions.

K.6 Assignment-Specific Result

An **assignment-specific result** depends on the particular translated outputs assigned to the sixty-four codons.

Examples include:

the hydropathy coefficients;

the synonymous-edge counts;

the self-loop multiplicities;

the G–A–C–U score;

and the position-specific nucleotide-pair weights.

Such results could change under a natural variant genetic code or an artificial reassignment of translated outputs.

They should be distinguished from input-space results that follow from the complete codon domain alone.

K.7 Balanced Design

A **balanced design** is one in which the coordinate states occur with exact regularity.

In the complete codon incidence matrix:

every nucleotide occurs sixteen times at every codon position;

every row contains exactly three active coordinates;

and every pair of nucleotide coordinates belonging to different codon positions co-occurs exactly four times.

This balance produces the exact block structure of the Gram matrix.

K.8 Binary Codon Matrix

The **binary codon matrix** is the 64 × 12 one-hot representation of the complete codon set.

Each row represents one codon.

Each group of four columns represents the nucleotide identity at one codon position.

A row contains one 1 in each positional block and therefore three 1s in total.

The matrix is denoted D.

K.9 Block

A **block** is one four-column positional section of the codon incidence matrix.

The three blocks are:

first-position nucleotide indicators;

second-position nucleotide indicators;

and third-position nucleotide indicators.

They are denoted

D_1,

D_2,

and

D_3.

Every block contains one active coordinate in each codon row.

K.10 Block Sum

A **block sum** is the sum of the four nucleotide coefficients or four indicator columns belonging to one codon position.

For the incidence matrix,

$D_1\ 1_4$

$= D_2\ 1_4$

$= D_3\ 1_4$

$= 1_{64}$.

The equality of these three column-block sums creates two exact linear dependencies.

For Moore–Penrose coefficients, the sums of the four coefficients within each positional block are also equal.

K.11 Categorical Output

A **categorical output** is a translated label rather than a numerical property.

The genetic-code output alphabet contains the twenty amino-acid single-letter codes and X.

A categorical output may later be assigned a numerical property such as hydropathy, but the category itself is not numerical.

The distinction is especially important for X, which is a valid translated class but not an amino acid.

K.12 Codon

A **codon** is an ordered triplet of nucleotide states.

In RNA notation,

$c = c_1\ c_2\ c_3$,

where every c_p belongs to

{ G, A, C, U }.

There are sixty-four possible codons.

A codon is an input word in the genetic-code relation.

K.13 Codon Box

A **codon box** is the set of four codons sharing the same first two nucleotides and differing only at the third position.

For example,

GCN

contains

GCG,

GCA,

GCC,

and

GCU.

Some codon boxes translate entirely to one output. Others are divided among two or more outputs.

Codon-box structure is often retained in constrained randomization models.

K.14 Codon Cube

A **codon cube** is a 4 × 4 × 4 arrangement of the sixty-four codons created by assigning an ordered coordinate to each nucleotide state on each codon axis.

A linear four-nucleotide path determines which nucleotide states are nearest neighbors.

Every such cube contains:

sixty-four vertices;

forty-eight edges along each axis;

and

144 undirected edges in total.

The cube is a sparse subgraph of the complete Hamming-distance-one codon graph.

K.15 Codon Field

A **codon field** is any numerical function defined over the sixty-four codons.

It can be represented as a vector

$h \in \mathbb{R}^{64}$.

Examples include:

hydropathy inherited from translated amino acids;

residue volume;

charge;

or an arbitrary numerical test field.

A codon field need not be additive.

K.16 Codon Hamming Graph

The **codon Hamming graph** connects two codons when they differ at exactly one nucleotide position.

Each codon has nine one-step neighbors:

three alternative nucleotides at each of three positions.

The complete graph contains

$64 \times 9 / 2 = 288$

undirected edges.

The G–A–C–U cube retains 144 of those edges.

K.17 Codon-Level Property

A **codon-level property** is a numerical value assigned to every codon.

When the property is inherited from translated output,

$h\,c = H(\tau(c))$.

All synonymous codons then receive the same original property value, although their additive reconstructions may differ.

K.18 Coefficient Range

A **coefficient range** is the difference between the maximum and minimum nucleotide coefficient within one codon position.

For position p,

rangep

$= \max_n \beta_{p,n} - \min_n \beta_{p,n}.$

The range is gauge invariant because adding one common constant to all four coefficients in a positional block does not change their differences.

The second-position hydropathy range is 6.587500 in the complete calculation.

K.19 Coefficient-Space Projector

The **coefficient-space projector** is

$Q = D^{+} D.$

It maps an arbitrary twelve-component coefficient vector onto the ten-dimensional row space of D.

It removes the two block-offset null components.

The matrix satisfies

$Q^2 = Q$

and

$Q^{T} = Q.$

K.20 Column Space

The **column space** of D is the set of all codon-level vectors that can be produced as

$D\,\beta.$

It has dimension ten.

It consists exactly of the additive codon fields.

The codon-space projector maps arbitrary sixty-four-component fields onto this space.

K.21 Common-Path Maximum

The **common-path maximum** refers to a nucleotide path that independently belongs to the maximizing family at all three codon positions.

For the standard genetic code under the linear-path synonymous-edge objective,

G–A–C–U

is the unique undirected common-path maximizer.

This is stronger than merely having the largest isotropic total.

K.22 Complete Codon Domain

The **complete codon domain** is the full Cartesian product

Σ^3.

It contains all sixty-four possible codons, including the three codons assigned to X.

Completeness is required for the exact balance of the 64 × 12 matrix.

Removing the stop codons changes the design and destroys the exact complete-product symmetry.

K.23 Complete Lookup Relation

A **complete lookup relation** assigns one output to every member of a finite input domain.

For the standard genetic code,

$\tau : C \rightarrow \Omega$

assigns one translated output to every one of the sixty-four codons.

The relation is complete even though it is many-to-one.

Several codons may share one translated output.

K.24 Complementary Path

The **complementary path** of a four-nucleotide linear path uses the three nucleotide-pair edges omitted by the original path from the complete six-edge nucleotide graph.

If one isotropic path has score L, its complement has score

$69 - L$.

The maximum path G–A–C–U and the minimum path A–U–G–C are complementary.

K.25 Complete Enumeration

A **complete enumeration** evaluates every object in a finite declared comparison class.

The isotropic analysis examines all twenty-four oriented nucleotide orders.

The anisotropic analysis examines all 13,824 oriented path triples.

Because every case is included, the resulting maxima, minima, means, and frequency distributions are population quantities rather than sample estimates.

K.26 Contrast

A **contrast** is a difference or zero-sum combination among coordinate levels.

Within one codon position, examples include

$\beta 2,U - \beta 2,A$

or

$G - A$.

A positional block with four nucleotide levels contains three independent contrasts.

The complete codon matrix has nine within-position contrast dimensions in total.

K.27 Degeneracy

Degeneracy is the number of codons assigned to one translated output.

For output ω,

$d\omega = |F\omega|$.

Leucine, arginine, and serine each have degeneracy six.

Methionine and tryptophan each have degeneracy one.

Degeneracy measures fiber size but not internal topology.

K.28 Degeneracy Profile

A **degeneracy profile** is the collection of codon counts assigned to the translated output classes.

For the standard code, the profile contains:

three six-codon classes;

five four-codon classes;

two three-codon classes;

nine two-codon classes;

and two one-codon classes.

A randomization may preserve the profile while changing the locations of the codons belonging to each class.

K.29 Design Matrix

A **design matrix** is a matrix whose columns represent explanatory coordinates and whose rows represent observations.

The codon incidence matrix D is a design matrix for additive nucleotide-position models.

The observations are the sixty-four codons.

The explanatory coordinates are the twelve nucleotide-position indicators.

K.30 Directed Graph

A **directed graph** distinguishes the ordered movement

$x \to y$

from

$y \to x$.

The principal codon-cube analyses in this book use undirected edges.

A directed extension could represent asymmetric mutation rates, mistranslation probabilities, or codon flows.

K.31 Edge Instance

An **edge instance** is one specific codon-level adjacency.

Different codon pairs may collapse to the same pair of translated outputs.

Those edges remain distinct instances in the translated multigraph.

Retaining edge instances preserves multiplicity.

K.32 Empirical Property Scale

An **empirical property scale** assigns numerical values to amino acids based on experimental measurements, compilations, or derived physical interpretations.

The Kyte–Doolittle hydropathy scale is one example.

Such values are not mathematical constants.

Exact projection calculations performed with them are exact relative to the declared numerical inputs and conventions.

K.33 Exact Result

An **exact result** follows without sampling or numerical approximation from the declared finite definitions.

Examples include:

rank (D) = 10;

the pseudoinverse entry values;

the forty-four-edge maximum;

and the complete anisotropic frequency distribution.

An exact result may still be model dependent. Its exactness applies within the declared mathematical construction.

K.34 Factorial Domain

A **factorial domain** contains every combination of the levels of several factors.

The complete codon set is a three-factor factorial domain with four levels at each factor.

The three factors are codon positions.

The four levels are G, A, C, and U.

The domain contains

$4 \times 4 \times 4 = 64$

combinations.

K.35 Fiber

A **fiber** is the set of all inputs sharing one output under a map.

For the translation map τ, the fiber of ω is

$F\omega = \tau^{-1} (\omega)$.

In biological language, a fiber is the synonymous codon set of one amino acid or X.

A fiber has both cardinality and shape within codon space.

K.36 Fiber-Averaged Reconstruction

A **fiber-averaged reconstruction** is the mean additive fitted value over all codons in one translated fiber.

It provides one reconstructed property value for comparison with the empirical property assigned to that translated class.

It should not be confused with exact equality among the fitted values of the synonymous codons.

K.37 Fiber Connectivity

Fiber connectivity describes whether the codons belonging to one translated output form one connected subgraph or several disconnected components under a declared codon topology.

Arginine and leucine are connected under the G–A–C–U cube.

Serine is divided into two components.

Connectivity is not determined by degeneracy alone.

K.38 Frozen Accident

Frozen accident is a historical description emphasizing that an established genetic code becomes difficult to change because reassignment would affect many existing coding sequences and translation components.

The phrase can help explain persistence.

It does not by itself explain the internal physicochemical or topological organization of the relation that became frozen.

K.39 Gauge Freedom

Gauge freedom is the nonuniqueness of the twelve additive coefficients representing one codon field.

A constant may be added to all coefficients in one positional block and removed from another without changing any reconstructed codon value.

The allowed gauge shifts are

$(a\, 1_4, b\, 1_4, c\, 1_4)^T$,

with

$a + b + c = 0$.

There are two independent gauge dimensions.

K.40 Gauge-Invariant Quantity

A **gauge-invariant quantity** does not change under an allowed block-offset transformation.

Within-position coefficient differences are gauge invariant.

Examples include

$\beta_{2,U} - \beta_{2,A}$

and the range of the second-position coefficient block.

Absolute coefficient levels are gauge dependent.

K.41 Generating Polynomial

A **generating polynomial** encodes a finite frequency distribution in its coefficients.

For the anisotropic codon cubes,

$F(x)$

$= \Sigma_s N_s x^s.$

The coefficient of x^s is the number of oriented cubes with score s.

The polynomial factors into the product of three position-specific score polynomials because the axes are independent.

K.42 Global Constant Direction

The **global constant direction** in coefficient space assigns the same value to all twelve nucleotide-position coordinates.

Under D, this produces a constant codon field because every codon activates exactly three coordinates.

The Gram eigenvalue associated with this direction is forty-eight.

Its singular value is

$4\sqrt{3}.$

K.43 Gram Matrix

The **Gram matrix** is

$G = D^T D.$

Its entries are inner products among the twelve coordinate columns.

The diagonal blocks are

16 I_4,

and the off-diagonal blocks are

4 J_4.

Its exact eigenvalues determine the rank and singular spectrum of D.

K.44 Hamming Distance

The **Hamming distance** between two codons is the number of codon positions at which they differ.

For three-letter codons, the possible distances are zero, one, two, and three.

The codon-space projector has one exact value for each Hamming distance.

K.45 Hamiltonian Path

A **Hamiltonian path** through the four nucleotide states visits every nucleotide exactly once.

Every linear nucleotide ordering used in the codon-cube analysis is a Hamiltonian path on the complete four-vertex nucleotide graph.

A path contains three nucleotide-pair edges.

There are twenty-four oriented Hamiltonian paths and twelve undirected ones.

K.46 Hydropathy

Hydropathy is a numerical representation of the tendency of an amino-acid side chain to favor hydrophobic or hydrophilic environments.

The book uses the Kyte–Doolittle scale for its main numerical analysis.

Hydropathy is transferred from amino-acid outputs to codons through the translation map.

The dominant additive codon-coordinate contrast occurs at the second position between U and A.

K.47 Hydropathy Field

A **hydropathy field** is the sixty-four-component codon vector obtained by assigning every codon the hydropathy of its translated amino acid.

Under the complete-code convention,

$H(X) = 0$.

The field is constant within synonymous fibers before additive projection.

K.48 Hydropathy Projection Fraction

The **hydropathy projection fraction** is the proportion of centered codon-level squared hydropathy variation captured by the additive nucleotide-position subspace.

It is

Q_{add}

$= \| P h^\circ \|^2 / \| h^\circ \|^2$.

For the complete Kyte–Doolittle field with X assigned zero, the value is approximately

0.816882.

K.49 Information-Theoretic Transition

An **information-theoretic transition** is a proposed change in organizational regime in which incomplete or local molecular associations become a sufficiently complete, accurate, and system-wide lookup relation.

The concept permits gradual changes in components together with a comparatively abrupt appearance of code-level organization.

It is an interpretive proposal, not a theorem derived from the matrix.

K.50 Input-Space Result

An **input-space result** follows from the complete codon domain and its representation without requiring the standard translated assignment.

Examples include:

the matrix dimensions;

the exact balance conditions;

rank ten;

the singular spectrum;

the pseudoinverse formula;

and the Hamming-distance projector.

These results would remain unchanged if the translated outputs were reassigned.

K.51 Interaction

An **interaction** is a codon-field component that depends jointly on two or three nucleotide positions and cannot be represented as a sum of independent positional effects.

The residual space of the complete additive model contains:

twenty-seven pairwise-interaction dimensions;

and twenty-seven three-way-interaction dimensions.

Together they form the fifty-four-dimensional orthogonal complement of the additive space.

K.52 Interclass Edge

An **interclass edge** is a codon-cube edge whose endpoints translate to different outputs.

After quotienting by translated identity, it becomes an edge between two distinct translated vertices.

The G–A–C–U translated multigraph contains one hundred interclass edge instances.

K.53 Isotropic Codon Cube

An **isotropic codon cube** uses the same nucleotide path on all three codon axes.

For path π, the cube uses π at codon positions one, two, and three.

There are twenty-four oriented isotropic cubes or twelve undirected path topologies.

G–A–C–U uniquely maximizes the isotropic synonymous-edge score up to reversal.

K.54 Kyte–Doolittle Scale

The **Kyte–Doolittle scale** is an amino-acid hydropathy scale widely used in membrane-protein analysis.

The present book uses it to construct the codon-level hydropathy field and the conventional sliding-window membrane profile.

The numerical conclusions are exact relative to the declared scale values and the treatment of X.

K.55 Left Null Space

The **left null space** of D is

null (D^T).

It contains codon-level vectors orthogonal to every column of D.

Its dimension is

64 − 10 = 54.

The residual from additive projection lies in this space.

K.56 Linear Nucleotide Path

A **linear nucleotide path** is an ordering

n_1–n_2–n_3–n_4

of the four nucleotide states in which only consecutive states are treated as neighbors.

It selects three of the six possible unordered nucleotide pairs.

A path and its reversal define the same undirected adjacency structure.

K.57 Lookup Code

A **lookup code** is a finite relation in which discrete input symbols or words are assigned outputs according to a stable mapping.

The genetic code is represented as the map

$\tau : C \rightarrow \Omega$.

Describing it as a lookup code emphasizes its complete finite assignment structure.

It does not imply that cellular translation is implemented as an electronic table lookup.

K.58 Lookup Relation

A **lookup relation** is the complete set of input-output assignments.

For the standard genetic code, it consists of sixty-four codon-to-output pairs.

The same relation can be represented as a conventional table, a categorical vector, a partition into fibers, or a labeled graph.

K.59 Many-to-One Map

A **many-to-one map** assigns several inputs to the same output.

The genetic code is many-to-one because synonymous codons share translated identity.

The map is still a function because each codon receives only one output in the standard table.

K.60 Matrix Rank

The **rank** of a matrix is the dimension of its column space or row space.

The complete codon incidence matrix has rank ten.

The rank consists of:

one global constant direction;

and nine independent within-position nucleotide contrasts.

K.61 Minimum-Norm Solution

The **minimum-norm solution** is the coefficient vector of smallest Euclidean length among all vectors producing the same fitted codon field.

The Moore–Penrose pseudoinverse supplies this unique solution.

Any other equivalent coefficient vector differs by a nonzero null-space component and therefore has a larger norm.

K.62 Model-Dependent Result

A **model-dependent result** depends on analytical choices beyond the standard code itself.

Examples include:

the selected hydropathy scale;

the assignment $H(X) = 0$;

the restriction to additive position effects;

the use of linear nucleotide paths;

and equal edge weighting.

Model dependence does not make a result invalid. It defines the boundaries within which the result is exact.

K.63 Moore–Penrose Pseudoinverse

The **Moore–Penrose pseudoinverse** is the unique generalized inverse satisfying four symmetry and consistency conditions.

For the complete codon matrix,

D^{+}

$= (1 / 16) D^T$

$- (1 / 96) J.$

It maps a codon-level field to its minimum-norm additive coefficient vector.

Its entries are only

$5 / 96$

and

$-1 / 96.$

K.64 Multigraph

A **multigraph** permits multiple edge instances between the same pair of vertices.

It may also permit self-loops.

The translated genetic-code graph is a multigraph because many distinct codon edges can collapse to the same translated endpoint pair.

Multiplicity must be retained for exact edge counts.

K.65 Natural Variant Code

A **natural variant code** is a biological genetic code differing from the standard code in one or more codon assignments.

The complete codon input matrix remains unchanged.

Assignment-specific quantities such as hydropathy coefficients, pair weights, loop counts, and path maxima may change.

Variant codes provide natural perturbations of the standard assignment.

K.66 Nearest Neighbor

A **nearest neighbor** in the codon cube differs by one step along one declared nucleotide path at exactly one codon position.

Nearest-neighbor status depends on the chosen nucleotide geometry.

It is not identical to Hamming distance one, because the full Hamming graph includes all six nucleotide-pair types while a linear path includes only three.

K.67 Nonadditive Component

A **nonadditive component** is the part of a codon field that cannot be represented as the sum of independent effects from the three codon positions.

It is

$$r = (I - P) h.$$

The nonadditive space has dimension fifty-four.

It contains pairwise and three-way interaction structure.

K.68 Nonsynonymous Edge

A **nonsynonymous edge** connects two adjacent codons having different translated outputs.

The G–A–C–U cube contains one hundred nonsynonymous edge instances.

After quotienting by translated identity, these become interclass edges.

K.69 Null Model

A **null model** is a declared comparison procedure used to determine whether an observed property is common or unusual under specified alternatives.

Possible genetic-code null models may preserve:

the exact degeneracy profile;

codon-box structure;

stop locations;

or translated fibers.

A probability has scientific meaning only relative to such a defined ensemble.

K.70 Null Space

The **null space** of D contains coefficient vectors z satisfying

$$D z = 0.$$

It has dimension two.

Adding a null vector to an additive coefficient vector changes the displayed coefficients without changing the reconstructed codon field.

K.71 One-Hot Encoding

One-hot encoding represents one categorical state by a binary vector containing one 1 and zeros elsewhere.

Each codon position is encoded by four one-hot nucleotide coordinates.

A complete codon row is the concatenation of three one-hot vectors and therefore contains three 1s.

K.72 Oriented Path

An **oriented path** distinguishes one sequence of nucleotide states from its reverse.

For example,

G–A–C–U

and

U–C–A–G

are different oriented paths.

They define the same undirected edge set and therefore have the same unweighted synonymous score.

K.73 Orthogonal Projection

An **orthogonal projection** maps a vector to the closest point in a linear subspace.

The matrix

$P = D\,D^{+}$

orthogonally projects a codon field onto the additive nucleotide-position subspace.

The residual is perpendicular to the fitted field and to every other additive field.

K.74 Output Alphabet

The **output alphabet** is the finite set of translated categories.

For the standard genetic code, it consists of twenty amino-acid codes and X.

Its cardinality is twenty-one.

K.75 Pair Weight

A **pair weight** counts synonymous codon pairs associated with one unordered nucleotide substitution type.

The position-specific weight

$w_p (x, y)$

counts such pairs at codon position p.

The total weight

$w (x, y)$

sums the three positional contributions.

The six total weights determine every isotropic path score.

K.76 Persistent Signal

A **persistent signal** is a relational pattern that remains operative across changing physical instances and changing messages.

The genetic code is a candidate for an exceptionally persistent natural signal because the same codon-to-output relation is repeatedly enacted across generations, genomes, and molecular turnover.

The phrase is interpretive and comparative rather than an exact mathematical theorem.

K.77 Position-Specific Maximum

A **position-specific maximum** is the largest synonymous-edge score obtainable by choosing a nucleotide path for one codon position.

For the standard code:

the first-position maximum is 4;

the second-position maximum is 1;

and the third-position maximum is 39.

The anisotropic global maximum is their sum.

K.78 Positional Asymmetry

Positional asymmetry is the unequal biological organization placed over the three otherwise balanced codon coordinates.

The matrix treats the positions symmetrically at the input level.

The translated assignment does not.

The second position dominates hydropathy discrimination, while the third dominates synonymous local connectivity.

K.79 Product Space

A **product space** is formed by selecting one element independently from each of several component sets.

The complete codon set is the product

$\Sigma \times \Sigma \times \Sigma$.

Each codon is one point in that three-factor product space.

K.80 Projection Fraction

A **projection fraction** is the proportion of centered squared magnitude lying within a selected subspace.

For codon field h,

Qadd

$= \| P\, h^\circ \|^2 / \| h^\circ \|^2$.

It measures the strength of additive nucleotide-position organization.

K.81 Projector

A **projector** is a matrix satisfying

$P^2 = P$.

Applying it twice has the same effect as applying it once.

An orthogonal projector is also symmetric.

The codon-space and coefficient-space projectors in this book are both orthogonal.

K.82 Pseudoinverse

A **pseudoinverse** is a generalized inverse defined for singular or rectangular matrices.

The Moore–Penrose pseudoinverse gives least-squares solutions and selects a unique minimum-norm coefficient vector.

For the complete codon matrix, the pseudoinverse has an exact closed form.

K.83 Quotient Graph

A **quotient graph** is formed by identifying several original vertices according to an equivalence relation.

In the translated multigraph, codons are identified when they share the same translated output.

Synonymous codon edges become self-loops after the quotient.

K.84 Randomization Ensemble

A **randomization ensemble** is the full set or sampling distribution of alternative structures generated under a declared randomization rule.

Different ensembles may preserve different features.

A conclusion that the standard code is unusual must specify the ensemble under which it is unusual.

K.85 Rank Deficiency

Rank deficiency occurs when a matrix has fewer independent columns than displayed columns.

The complete codon matrix has twelve columns but rank ten.

Its deficiency of two arises from equality of the three positional block sums.

It is exact, not numerical.

K.86 Reflexive Maintenance

Reflexive maintenance means that products generated through the genetic code help preserve the machinery that continues to enact the code.

Translation produces proteins involved in transfer-RNA charging, ribosome function, proofreading, metabolism, and other supporting processes.

The phrase does not imply that the code causes itself to originate.

K.87 Residual

A **residual** is the difference between an original codon field and its additive projection:

$r = h - \hat{h}$.

It is orthogonal to every additive nucleotide-position field.

The residual may contain structured interaction effects and should not automatically be interpreted as random noise.

K.88 Reversal Equivalence

Reversal equivalence identifies a nucleotide path with its reverse when only undirected adjacency matters.

Thus,

G–A–C–U

and

U–C–A–G

represent one undirected path topology.

Reversal reduces twenty-four oriented paths to twelve undirected paths.

K.89 Right Null Space

The **right null space** of D contains coefficient vectors z satisfying

D z = 0.

It has dimension two.

Its vectors transfer common offsets among the three positional blocks while preserving every codon-level sum.

K.90 Row Space

The **row space** of D is the ten-dimensional subspace of coefficient space orthogonal to the right null space.

The Moore–Penrose coefficient vector lies in this space.

Its three positional block sums are equal.

K.91 Self-Loop

A **self-loop** is an edge whose two endpoints become the same vertex after quotienting by translated identity.

Every synonymous codon-cube edge becomes a self-loop at the corresponding amino-acid or X vertex.

The G–A–C–U translated multigraph contains forty-four self-loop instances.

K.92 Self-Loop Multiplicity

Self-loop multiplicity is the number of distinct codon-level synonymous edges collapsed onto one translated vertex.

Arginine and leucine each have multiplicity six.

Serine has multiplicity four.

Multiplicity must be retained even when a drawing shows one compact loop symbol.

K.93 Singular Spectrum

The **singular spectrum** is the complete set of singular values of a matrix together with their multiplicities.

For D, the spectrum is:

4 √3 once;

4 nine times;

and

0 twice.

The ten nonzero singular values establish rank ten.

K.94 Singular-Value Decomposition

The **singular-value decomposition**, or SVD, writes a matrix in terms of orthogonal left and right singular vectors and nonnegative singular values.

It can be used to calculate a pseudoinverse and analyze identifiable directions.

The historical sixty-one-sense-codon hydropathy work used SVD numerically.

The complete sixty-four-codon matrix permits an exact closed-form derivation.

K.95 Stop Convention

A **stop convention** is the declared numerical treatment of X when an amino-acid property is extended to all sixty-four codons.

In the complete hydropathy calculation,

H (X) = 0.

This is an analytical convention rather than an amino-acid measurement.

Categorical and topological analyses do not require a numerical stop value.

K.96 Standard Genetic Code

The **standard genetic code** is the conventional mapping from the sixty-four RNA codons to twenty amino acids and translation stop.

Natural variants exist.

The word **standard** identifies the assignment used in the principal analyses of this book.

K.97 Synonymous Codons

Synonymous codons are distinct codons having the same translated output.

For codons c and d,

$\tau (c) = \tau (d)$.

The term includes pairs of stop codons when both translate to X in the categorical graph analysis.

K.98 Synonymous Edge

A **synonymous edge** is a codon-cube edge connecting synonymous codons.

The G–A–C–U cube contains forty-four such edges.

The complete Hamming-distance-one codon graph contains sixty-nine.

K.99 Synonymous Fiber

A **synonymous fiber** is the complete set of codons assigned to one translated output.

It is the same object as the preimage

$F\omega = \tau^{-1} (\omega)$.

The word **fiber** emphasizes the mathematical map, while **synonymous set** emphasizes biological translation.

K.100 Topology

Topology, in the restricted use of this book, refers to adjacency and connectivity relationships among codons and translated fibers under a declared nucleotide-path geometry.

It does not refer to every branch of mathematical topology.

The principal objects are:

the codon cube;

its synonymous subgraphs;

and the translated multigraph.

K.101 Translation Map

The **translation map** is

$\tau : C \to \Omega$.

It assigns one translated output to every codon.

The matrix D represents codon input composition but does not itself contain τ.

Hydropathy and synonymous topology enter only after the translation map is applied.

K.102 Translated Multigraph

The **translated multigraph** is the graph obtained by collapsing codon vertices according to translated identity while retaining every codon-edge instance.

It contains twenty-one vertices under the standard code.

The G–A–C–U version contains forty-four self-loops and one hundred interclass edges.

K.103 Undirected Path

An **undirected path** treats a nucleotide ordering and its reverse as the same adjacency structure.

There are twelve undirected linear paths through the four nucleotide states.

G–A–C–U is the unique undirected isotropic maximum.

K.104 U–A Contrast

The **U–A contrast** is the within-position difference between the coefficient for U and the coefficient for A.

At the second codon position,

$\beta_{2,U} - \beta_{2,A} = 6.587500$

in the complete Kyte–Doolittle calculation.

This contrast is gauge invariant and is the dominant additive hydropathy direction.

K.105 U/T–A Membrane Profile

The **U/T–A membrane profile** is a nucleotide-only sliding-window profile based on the second codon position.

In RNA notation:

second-position U contributes positively;

second-position A contributes negatively;

and G or C contributes zero in the simplest form.

In DNA notation, T replaces U.

The profile is a compressed approximation to translated hydropathy organization.

K.106 Weighted Graph

A **weighted graph** assigns a numerical importance to each edge.

The main codon-cube analysis gives every selected edge equal weight.

Extensions may use mutation rates, codon usage, mistranslation probabilities, or physicochemical differences.

A weighted objective may rank nucleotide paths differently from the unweighted forty-four-edge objective.

K.107 Word Space

A **word space** is the set of all ordered strings of a fixed length over a fixed alphabet.

The codon domain is the word space of length three over a four-symbol alphabet.

The general matrix theory in Appendix L applies to complete word spaces beyond the genetic code.

K.108 X

X denotes translation stop throughout the book.

It replaces the asterisk often used in conventional genetic-code tables.

X is treated as:

a categorical translated output;

a vertex in the translated multigraph;

and a three-codon fiber.

It is not an amino acid.

Any numerical value assigned to X in a property calculation must be declared as a convention.

K.109 Summary

The terminology of the book separates several levels that are often combined in one conventional genetic-code table.

The complete word space defines the input domain.

The binary incidence matrix exposes its exact additive algebra.

The translation map partitions codons into synonymous fibers.

Numerical amino-acid properties create fields over codon space.

The pseudoinverse separates additive and nonadditive components.

A declared nucleotide path creates local geometry.

Quotienting by translated identity creates the multigraph.

Exhaustive enumeration establishes exact maxima within the declared comparison class.

These distinctions permit strong mathematical statements without confusing input structure, biological assignment, analytical convention, and historical interpretation.

Appendix L

Exact Pseudoinverse for Complete Alphabet–Word Matrices

L.1 Generalization Beyond the Genetic Code

The exact pseudoinverse of the 64 × 12 codon incidence matrix is not an isolated property of the genetic code.

It belongs to a general family of complete word-space matrices.

Let an alphabet contain

q

distinct symbols, and let every word have length

ℓ.

The complete word space contains

q^{ℓ}

words.

Represent each word by one-hot positional coordinates. Every one of the ℓ word positions contributes q indicator columns, one for each alphabet symbol. The complete incidence matrix therefore has dimensions

$q^{\ell} \times \ell q$.

Every row contains exactly

ℓ

entries equal to 1 and

$\ell (q - 1)$

entries equal to 0.

The genetic-code matrix is the special case

$q = 4$

and

$\ell = 3$.

Then

$q^{\ell} = 4^3 = 64$

and

$\ell q = 3 \times 4 = 12$.

The rank, singular spectrum, pseudoinverse, gauge freedom, and projection structure derived for codon space are instances of exact formulas valid for every complete alphabet–word design.

L.2 Alphabet and Complete Word Space

Let the alphabet be

$\Sigma q = \{ a_1, a_2, \ldots, aq \}$.

A word is an ordered ℓ-tuple

$w = w_1 w_2 \ldots w\ell$,

where

wp ∈ Σq

for every position

p ∈ { 1, 2, …, ℓ }.

The complete word set is

W = Σq^ℓ.

Its cardinality is

| W | = q^ℓ.

Every possible assignment of one alphabet symbol to each of the ℓ positions occurs exactly once.

No output labels, biological assignments, or numerical property values are required to construct W.

L.3 Matrix Definition

For each word position p and alphabet symbol a, define one binary coordinate

(p, a).

There are

ℓq

such coordinates.

Define the complete incidence matrix

D (q, ℓ)

∈ { 0, 1 }q^ℓ × ℓq

by

D (q, ℓ)w,(p,a) = 1

when

wp = a,

and

D (q, ℓ)w,(p,a) = 0

otherwise.

The rows range over all q^ℓ words.

The columns range over all ℓq position-symbol coordinates.

When the values of q and ℓ are clear, the matrix will be written simply as D.

The matrix can be partitioned into ℓ positional blocks:

D = [D_1 | D_2 | … | Dℓ],

where every block has dimensions

q^ℓ × q.

Each block records the alphabet symbol appearing at one word position.

L.4 Exact Row Counts

Every word contains one symbol at each of ℓ positions.

Therefore, every row of D contains one 1 in every positional block.

Its row sum is

ℓ.

In matrix form,

D 1ℓq = ℓ 1q^ℓ.

Every row therefore contains:

ℓ entries equal to 1;

and

ℓq − ℓ

= ℓ (q − 1)

entries equal to 0.

The total number of 1s calculated from the row count is

q^ℓ × ℓ

= ℓ q^ℓ.

L.5 Exact Column Counts

Fix one word position p and one alphabet symbol a.

Requiring

wp = a

fixes one position. The remaining

ℓ − 1

positions may contain any of q symbols.

The coordinate therefore occurs in

q^(ℓ−1)

words.

Every column sum is consequently

q^(ℓ−1).

In matrix notation,

D^T 1q^ℓ

= q^(ℓ−1) 1ℓq.

Every column contains:

q^(ℓ−1) entries equal to 1;

and

q^ℓ − q^(ℓ−1)

= q^(ℓ−1) (q − 1)

entries equal to 0.

The total number of 1s calculated from the column count is

ℓq × q^(ℓ−1)

= ℓ q^ℓ.

This agrees with the row calculation.

The complete matrix contains

$q^{\ell} \times \ell q$

$= \ell \, q^{(\ell+1)}$

binary entries, of which

$\ell \, q^{\ell}$

are 1 and

$\ell \, (q - 1) \, q^{\ell}$

are 0.

L.6 Pairwise Column Overlap

Two distinct symbol columns belonging to the same word position are mutually exclusive.

A word cannot contain two different symbols at one position.

Therefore, within one positional block,

$D_p^T D_p$

$= q^{(\ell-1)} Iq.$

Now consider two different positions p and r.

Fixing one symbol at position p and one symbol at position r leaves

$\ell - 2$

positions unrestricted.

The number of words containing both selected coordinates is

$q^{(\ell-2)}.$

Therefore, for

$p \neq r,$

$D_p^T Dr$

$= q^{(\ell-2)} Jq.$

The complete Gram matrix

$G = D^T D$

is an $\ell \times \ell$ block matrix with blocks of size $q \times q$:

diagonal blocks

q^(ℓ−1) Iq;

and

off-diagonal blocks

q^(ℓ−2) Jq.

For ℓ = 1, there are no off-diagonal blocks, and the matrix reduces to Iq.

L.7 Trace Check

The Gram matrix has ℓq diagonal entries, each equal to q^(ℓ−1).

Therefore,

tr (G)

= ℓq × q^(ℓ−1)

= ℓ q^ℓ.

This is also the squared Frobenius norm of D:

‖ D ‖²F = ℓ q^ℓ,

because D contains exactly ℓ q^ℓ entries equal to 1.

The eigenvalues derived below must therefore sum to

ℓ q^ℓ.

L.8 Decomposition of Coordinate Space

The coefficient space has dimension

ℓq.

It separates into three mutually orthogonal invariant subspaces.

The first is the global constant direction.

It consists of vectors proportional to

1ℓq

= (1q, 1q, …, 1q)ᵀ.

Its dimension is one.

The second consists of zero-sum contrasts within individual positional blocks.

For one position p, a contrast vector x satisfies

$1q^{T} x = 0.$

Each q-dimensional positional block contains

$q - 1$

independent contrasts.

Across ℓ positions, the contrast subspace has dimension

$ℓ (q - 1).$

The third consists of vectors that are constant within every block but whose block constants sum to zero.

A general vector in this subspace has the form

z

$= (c_1 1q, c_2 1q, …, cℓ 1q)^{T},$

where

$Σp=1^ℓ cp = 0.$

The dimension of this block-offset subspace is

$ℓ - 1.$

The dimensions sum correctly:

1

- $ℓ (q - 1)$
- $(ℓ - 1)$

$= 1 + ℓq - ℓ + ℓ - 1$

$= ℓq.$

These three subspaces therefore account for the entire coefficient space.

L.9 Global Constant Eigenvalue

Let

$v_o = (1q, 1q, \ldots, 1q)^T$.

Consider one output block of G v_o.

The diagonal Gram block contributes

q^(ℓ−1) Iq 1q

= q^(ℓ−1) 1q.

Every off-diagonal block contributes

q^(ℓ−2) Jq 1q.

Because

Jq 1q = q 1q,

each off-diagonal contribution is

q^(ℓ−1) 1q.

There are

ℓ − 1

off-diagonal blocks contributing to each output block.

The total is therefore

q^(ℓ−1) 1q

- (ℓ − 1) q^(ℓ−1) 1q

= ℓ q^(ℓ−1) 1q.

Hence,

G v_o

= ℓ q^(ℓ−1) v_o.

The global constant eigenvalue is

λ_o = ℓ q^(ℓ−1).

Its multiplicity is one.

L.10 Within-Position Contrast Eigenvalue

Let x be a zero-sum vector supported in one positional block:

$1q^{T} x = 0.$

Because

$Jq\ x = 0,$

all off-diagonal Gram blocks annihilate the contrast.

The diagonal block acts as

$q^{(\ell-1)}\ Iq\ x$

$= q^{(\ell-1)}\ x.$

Therefore, every within-position contrast is an eigenvector with eigenvalue

$\lambda contrast = q^{(\ell-1)}.$

Each positional block contributes q − 1 independent contrasts.

The multiplicity is therefore

$\ell\ (\ q - 1\ ).$

Before any output assignment is introduced, all within-position contrasts share the same eigenvalue. Complete word space does not algebraically privilege one symbol contrast over another.

L.11 Block-Offset Null Space

Consider a blockwise constant coefficient vector

z

$= (\ c_1\ 1q,\ c_2\ 1q,\ \ldots,\ c\ell\ 1q\ )^{T},$

with

$\Sigma p=1^{\ell}\ cp = 0.$

For any output block, the Gram matrix produces

$q^{(\ell-1)}$

times the sum of the block constants, multiplied by 1q.

Because that sum is zero,

$G z = 0.$

Since

$G = D^T D,$

it follows that

$D z = 0.$

The block-offset subspace is therefore the complete right null space of D.

Its dimension is

$\ell - 1.$

The nullity is not caused by omitted words or incomplete data. It results from the fact that every positional block sums to the same constant word vector:

$D_1 \, 1q$

$= D_2 \, 1q$

$= \dots$

$= D\ell \, 1q$

$= 1q^{\ell}.$

The ℓ copies of the constant direction supply only one independent codon-field direction, leaving

$\ell - 1$

redundant block offsets.

L.12 Rank and Nullities

The nonzero eigenspaces have total dimension

$1 + \ell (q - 1).$

Therefore,

rank $D (q, \ell)$

$= 1 + \ell (q - 1).$

Because D has ℓq columns, its right nullity is

$\ell q - [1 + \ell (q - 1)]$

= ℓ − 1.

Because D has q^ℓ rows, its left nullity is

q^ℓ − [1 + ℓ (q − 1)].

Thus,

nullity (D)

= ℓ − 1,

and

nullity (D^T)

= q^ℓ − 1 − ℓ (q − 1).

The rank formula can also be understood as the dimension of a complete additive word model:

one global constant direction;

plus q − 1 symbol contrasts at each of ℓ positions.

L.13 Complete Gram Spectrum

The complete eigenvalue spectrum of

G = D^T D

is:

ℓ q^(ℓ−1), with multiplicity 1;

q^(ℓ−1), with multiplicity ℓ (q − 1);

and

0, with multiplicity ℓ − 1.

The eigenvalue sum is

ℓ q^(ℓ−1)

- ℓ (q − 1) q^(ℓ−1)

= ℓq q^(ℓ−1)

= ℓ q^ℓ.

This agrees with the trace and Frobenius-norm calculation.

L.14 Singular Spectrum

The singular values of D are the nonnegative square roots of the Gram eigenvalues.

Therefore, D (q, ℓ) has:

√ℓ q^((ℓ−1)/2), with multiplicity 1;

q^((ℓ−1)/2), with multiplicity ℓ (q − 1);

and

0, with multiplicity ℓ − 1.

The spectral condition number restricted to the nonzero singular subspace is

κ

= [√ℓ q^((ℓ−1)/2)]

/ q^((ℓ−1)/2)

= √ℓ.

Thus, after the exact null directions are removed, conditioning depends only on word length and not on alphabet size.

L.15 Exact Pseudoinverse Theorem

Complete Alphabet–Word Pseudoinverse Theorem.
Let D (q, ℓ) be the complete q^ℓ × ℓq one-hot positional incidence matrix of every length-ℓ word over an alphabet of q symbols. Then its Moore–Penrose pseudoinverse is

D (q, ℓ)$^{+}$

= q^(1−ℓ) D (q, ℓ)^{T}

− (ℓ − 1) / (ℓ q^ℓ) Jℓq×q^ℓ.

The formula has the same two-term structure as the codon result:

a scaled transpose;

minus a uniform correction for positional-block redundancy.

When ℓ = 1, the correction term vanishes and D is the q × q identity matrix up to row and column ordering.

L.16 Why the Formula Has This Form

The transpose D^T collects word-level values over all words containing a selected position-symbol coordinate.

Every coordinate occurs in q^(ℓ−1) words. Multiplication by

q^(1−ℓ)

therefore converts each coordinate sum into its mean over the corresponding subset.

That coordinate averaging alone overcounts the global constant component.

A constant field with value k would produce k in every coordinate, and each word would reconstruct to ℓk because it activates ℓ coordinates.

The uniform correction

(ℓ − 1) / (ℓ q^ℓ) J

removes the redundant portion of the global constant field.

After correction, a constant field k produces coefficient k / ℓ in every coordinate, and every word reconstructs to

ℓ (k / ℓ) = k.

The formula therefore retains all symbol contrasts while distributing the constant component equally among the ℓ positional blocks.

L.17 Two Exact Pseudoinverse Values

Every entry of D^T is either 1 or 0.

If a word contains the specified symbol at the specified position, the pseudoinverse entry is

q^(1−ℓ)

− (ℓ − 1) / (ℓ q^ℓ).

Using the common denominator

ℓ q^ℓ,

the first term is

ℓq / (ℓ q^ℓ).

Therefore, the present-coordinate value is

[ℓq − (ℓ − 1)] / (ℓ q^ℓ)

= [ℓ (q − 1) + 1] / (ℓ q^ℓ).

If the coordinate is absent, the entry is

− (ℓ − 1) / (ℓ q^ℓ).

Thus,

$D^{+}(p,a);w$

=

[ℓ (q − 1) + 1] / (ℓ q^ℓ)

when

wp = a,

and

$D^{+}(p,a);w$

=

− (ℓ − 1) / (ℓ q^ℓ)

when

wp ≠ a.

Every pseudoinverse in this complete family therefore contains only two exact rational values.

L.18 Pseudoinverse Column Structure

Each word activates exactly ℓ coordinates.

Therefore, every pseudoinverse column contains:

ℓ positive entries equal to

[ℓ (q − 1) + 1] / (ℓ q^ℓ);

and

ℓ (q − 1) negative entries equal to

− (ℓ − 1) / (ℓ q^ℓ).

The column sum is

ℓ × [ℓ (q − 1) + 1] / (ℓ q^ℓ)

− ℓ (q − 1) × (ℓ − 1) / (ℓ q^ℓ).

After cancellation, the numerator becomes

q.

Therefore, every pseudoinverse column sums to

q / q^ℓ

= q^(1−ℓ).

In matrix form,

$1\ell q^{T} D^{+}$

$= q^{(1-\ell)} 1q^{\ell^{T}}$.

For the codon matrix, this becomes

1 / 16.

L.19 Pseudoinverse Row Structure

Each position-symbol coordinate appears in

q^(ℓ−1)

words and is absent from

q^(ℓ−1) (q − 1)

words.

Therefore, every pseudoinverse row contains:

q^(ℓ−1) positive entries;

and

q^(ℓ−1) (q − 1) negative entries.

Its row sum is

1 / ℓ.

In matrix form,

$D^{+} 1q^{\ell}$

= (1 / ℓ) 1ℓq.

For word length three, every pseudoinverse row sums to

1 / 3.

L.20 Exact Row and Column Norms

The equal row and column compositions imply uniform Euclidean norms.

Every pseudoinverse row has squared norm

[ℓ² (q − 1) + 1]

/ (ℓ² q^ℓ).

Every pseudoinverse column has squared norm

[ℓ² (q − 1) + 1]

/ (ℓ q^(2ℓ−1)).

These formulas provide additional exact implementation checks.

For

q = 4

and

ℓ = 3,

they reduce to

7 / 144

for every pseudoinverse row

and

7 / 768

for every pseudoinverse column.

L.21 Constant-Field Reconstruction

Let every word be assigned the same numerical value k:

h = k 1q^ℓ.

Because every pseudoinverse row sums to 1 / ℓ,

$D^+ h$

$= (k / ℓ) 1ℓq.$

Every word activates exactly ℓ coordinates. Therefore,

$D D^+ h$

$= ℓ (k / ℓ) 1q^ℓ$

$= k 1q^ℓ.$

Thus, every constant word field is reconstructed exactly.

The constant-field calculation demonstrates the role of the uniform correction in the pseudoinverse formula.

L.22 General Coefficient Formula

Let

$h ∈ ℝq^ℓ$

be any numerical field over the complete word space.

The Moore–Penrose coefficient for position p and symbol a is

$βp,a$

$= q^(1−ℓ)$

$Σw : wp=a h w$

$− (ℓ − 1) / (ℓ q^ℓ)$

$Σw∈W h w.$

Let

$μ = (1 / q^ℓ) Σw∈W h w$

be the global word-field mean.

The first term is the mean of h over all words containing symbol a at position p.

The second term is

$(ℓ − 1) μ / ℓ.$

Therefore,

βp,a

=

mean of h among words having symbol a at position p

− (ℓ − 1) μ / ℓ.

For ℓ = 3, the correction is

2 μ / 3,

which reproduces the codon coefficient formula.

L.23 Positional Block Sums

Let

S = Σw∈W h w.

Sum the q coefficients belonging to one fixed position p.

The q coordinate subsets partition the complete word space, so their uncorrected sums total S.

The resulting block sum is

Σa∈Σq βp,a

= S / (ℓ q^(ℓ−1)).

This value is the same for every positional block.

Therefore, the Moore–Penrose coefficient vector satisfies

Σa β1,a

= Σa β2,a

= …

= Σa βℓ,a.

For a centered word field,

S = 0,

and every positional block has coefficient sum zero.

The equality of positional block sums is the general Moore–Penrose gauge condition.

L.24 Gauge Freedom

The positional blocks have equal column sums.

A constant may be added to every coefficient in one positional block and removed from one or more other blocks without changing any reconstructed word value.

The general gauge shift is

z

= (c_1 1q, c_2 1q, …, cℓ 1q)^T,

where

Σp=1^ℓ cp = 0.

Every word activates one coefficient from every block. Its total gauge shift is therefore

$c_1 + c_2 + … + cℓ = 0$.

Thus,

D z = 0.

There are

ℓ − 1

independent gauge directions, exactly matching the right nullity of D.

The Moore–Penrose solution selects the unique coefficient vector orthogonal to this null space.

L.25 Gauge-Invariant Contrasts

Absolute coefficient levels within one positional block depend on the selected gauge.

Within-position contrasts do not.

For symbols a and b at position p,

βp,a − βp,b

is unchanged when one common block constant is added to all q coefficients.

Positional ranges are also gauge invariant:

maxa βp,a − mina βp,a.

Consequently, scientifically or mathematically interpreted contrasts should generally be preferred over isolated absolute coefficient values.

The codon hydropathy contrast

$\beta_{2,U} - \beta_{2,A}$

is one special case of this general invariance.

L.26 Minimum-Norm Property

Let

$\beta_{MP} = D^{+} h.$

Any other coefficient vector producing the same fitted field has the form

$\beta_{MP} + z,$

where

$z \in \text{null}(D).$

The Moore–Penrose vector lies in the row space of D, which is orthogonal to null (D).

Therefore,

$\beta_{MP}^{T} z = 0.$

Hence,

$\| \beta_{MP} + z \|^2$

$= \| \beta_{MP} \|^2 + \| z \|^2.$

Every nonzero gauge component increases the coefficient norm.

The pseudoinverse therefore chooses the unique minimum-norm representative of every additive word field.

L.27 General Additive Word Fields

The column space of D consists exactly of fields of the form

$f(w_1, w_2, \ldots, w_\ell)$

$= f_1(w_1)$

- $f_2(w_2)$
- …
- $f_\ell(w_\ell)$.

Its dimension is

$1 + \ell\,(q - 1)$.

The constant direction contributes one dimension.

Every word position contributes $q - 1$ independent symbol contrasts.

An arbitrary numerical field over W has dimension q^{ℓ}.

The remaining

$q^{\ell} - 1 - \ell\,(q - 1)$

dimensions contain interactions not representable by independent positional effects.

L.28 Interaction Decomposition

The full q^{ℓ}-dimensional word-field space can be decomposed by interaction order.

The number of k-position interaction dimensions is

$C(\ell, k)\,(q - 1)^k$.

Summing over all interaction orders gives

$\Sigma_{k=0}^{\ell}\, C(\ell, k)\,(q - 1)^k$

$= [\,1 + (q - 1)\,]^{\ell}$

$= q^{\ell}$.

The zero-order component contributes the constant direction:

1.

The first-order components contribute the additive positional contrasts:

$\ell\,(q - 1)$.

The residual beyond the additive model consists of all interaction orders $k \geq 2$:

$\Sigma_{k=2}^{\ell}\, C(\ell, k)\,(q - 1)^k$

$= q^{\ell} - 1 - \ell\,(q - 1)$.

For the genetic-code matrix,

$q = 4$

and

ℓ = 3.

The decomposition becomes

1

- 3 (3)
- 3 (3^2)
- 3^3

= 1 + 9 + 27 + 27

= 64.

The additive space has dimension ten, while the interaction space has dimension fifty-four.

L.29 Exact Codon-Space Projector

The word-space projector is

P = D D^+.

Substituting the general pseudoinverse gives

P

= q^(1−ℓ) D D^T

− (ℓ − 1) / q^ℓ Jq^ℓ.

The matrix P is symmetric and idempotent.

It projects arbitrary q^ℓ-component word fields onto the additive subspace.

Its rank and trace are

rank (P)

= tr (P)

= 1 + ℓ (q − 1).

L.30 Projector Entries from Matching Positions

Let u and v be two words.

Let

m (u, v)

be the number of positions at which they contain the same symbol.

Each matrix row contains one active coordinate at every position. Therefore,

$(D D^{T})_{u,v}$

$= m (u, v)$.

The exact projector entry is

$P_{u,v}$

$= m (u, v) / q^{(\ell-1)}$

$- (\ell - 1) / q^{\ell}$.

Equivalently,

$P_{u,v}$

$= [q \, m (u, v) - (\ell - 1)] / q^{\ell}$.

If

$dH (u, v)$

is the Hamming distance between the words, then

$m (u, v) = \ell - dH (u, v)$.

Therefore,

$P_{u,v}$

$=$

$[q (\ell - dH (u, v)) - \ell + 1]$

$/ q^{\ell}$.

The projector depends only on the number of matching positions, or equivalently only on Hamming distance.

For words of length ℓ, the projector contains at most

$\ell + 1$

distinct values.

L.31 Uniform Leverage

For a word compared with itself,

m (w, w) = ℓ.

Therefore, every diagonal projector entry is

Pw,w

= ℓ / q^(ℓ−1)

− (ℓ − 1) / q^ℓ.

Combining terms gives

Pw,w

=

[ℓq − ℓ + 1] / q^ℓ

=

[ℓ (q − 1) + 1] / q^ℓ.

Every word has identical leverage.

The trace is

q^ℓ

× [ℓ (q − 1) + 1] / q^ℓ

= 1 + ℓ (q − 1).

This equals the rank, as required for an orthogonal projector.

Uniform leverage follows from the symmetry of complete word space. No word is favored by the incidence design itself.

L.32 Coefficient-Space Projector

The coefficient-space projector is

$Q = D^{+} D$.

Its diagonal q × q blocks are

Iq

$-(\ell - 1)/(\ell q)\, J_q$.

Its off-diagonal blocks are

$1/(\ell q)\, J_q$.

The matrix acts as:

the identity on the global constant direction;

the identity on all $\ell(q-1)$ within-position contrasts;

and

zero on the $\ell - 1$ block-offset null directions.

Therefore,

$Q^2 = Q$,

$Q^T = Q$,

and

$\operatorname{rank}(Q) = 1 + \ell(q-1)$.

L.33 Verification of the Moore–Penrose Conditions

The exact formula satisfies the four Moore–Penrose equations.

Because

$P = D D^+$

is the orthogonal projector onto the column space of D,

$P D = D$.

Therefore,

$D D^+ D = D$.

Because

$Q = D^+ D$

is the orthogonal projector onto the row space of D and the columns of D^+ lie in that row space,

$Q D^+ = D^+$.

Therefore,

$D^{+} D D^{+} = D^{+}$.

Both projectors are symmetric:

$(D D^{+})^{T} = D D^{+}$,

and

$(D^{+} D)^{T} = D^{+} D$.

The proposed matrix is therefore the unique Moore–Penrose pseudoinverse.

L.34 Exact Reconstruction Criterion

A word field h is reconstructed exactly if and only if

$P h = h$.

Equivalently,

h belongs to the column space of D.

Equivalently, there exist position-specific functions

$f_1, f_2, \ldots, f_{\ell}$

such that

$h (w_1, w_2, \ldots, w_{\ell})$

$= \Sigma p=1 \wedge \ell \, fp (wp)$.

The pseudoinverse reconstructs every additive word field exactly.

Any pairwise or higher interaction component remains in the residual

$r = h - P h$.

The residual is orthogonal to every additive field:

$D^{T} r = 0$.

L.35 Genetic-Code Specialization

Set

$q = 4$

and

$\ell = 3$.

Then the complete matrix has dimensions

$4^3 \times 3\ (\ 4\ )$

$= 64 \times 12.$

Every row contains

3

ones.

Every column contains

$4^2 = 16$

ones.

The rank is

$1 + 3\ (\ 4 - 1\ )$

$= 10.$

The right nullity is

$3 - 1 = 2.$

The left nullity is

$64 - 10 = 54.$

The singular values are

$\sqrt{3} \times 4$

$= 4\ \sqrt{3}$

with multiplicity one;

4

with multiplicity

$3\ (\ 4 - 1\ ) = 9;$

and

0

with multiplicity two.

The general pseudoinverse becomes

D^{+}

$= 4^{(1-3)} D^{T}$

$- (3 - 1) / (3 \times 4^{3}) J.$

Therefore,

D^{+}

$= (1 / 16) D^{T}$

$- (2 / 192) J$

$= (1 / 16) D^{T}$

$- (1 / 96) J.$

The present-coordinate value is

$[3 (4 - 1) + 1] / (3 \times 4^{3})$

$= 10 / 192$

$= 5 / 96.$

The absent-coordinate value is

$- (3 - 1) / (3 \times 4^{3})$

$= -2 / 192$

$= -1 / 96.$

The projector entry for words matching at m positions is

$m / 16 - 1 / 32.$

For codon Hamming distances zero, one, two, and three, the values are:

$5 / 32;$

$3 / 32;$

$1 / 32;$

and

$-1 / 32.$

Every exact matrix result derived earlier is therefore recovered as a special case of the general alphabet–word theorem.

L.36 A Small Binary Example

Consider a binary alphabet with

$q = 2$

and words of length

$\ell = 2$.

The complete word space contains

$2^2 = 4$

words.

The incidence matrix has dimensions

4×4.

Its rank is

$1 + 2 (2 - 1)$

$= 3$.

Its right nullity is

$2 - 1 = 1$.

The singular values are:

$\sqrt{2} \times 2^{1/2} = 2$

once;

$\sqrt{2}$

twice;

and

0

once.

The pseudoinverse is

D^{+}

$= 2^{(-1)} D^{T}$

$- 1 / (2 \times 2^{2}) J$

$= (1 / 2) D^{T}$

$- (1 / 8) J.$

The present-coordinate entry is

$[2 (2 - 1) + 1] / (2 \times 2^{2})$

$= 3 / 8.$

The absent-coordinate entry is

$-1 / 8.$

This smallest nontrivial example already displays:

two exact pseudoinverse values;

one positional-block gauge direction;

uniform leverage;

and a Hamming-distance projector.

The codon case is a larger member of the same family.

L.37 Dependence on Completeness

The exact formulas depend on the presence of every q^{ℓ} word with equal weight.

If words are omitted, repeated, or weighted unequally, then:

column counts may differ;

cross-position overlaps may cease to be uniform;

the Gram blocks may lose their simple I and J form;

the singular spectrum may split;

and the pseudoinverse may contain many distinct values.

The historical sixty-one-sense-codon matrix is an important example.

Removing the three stop codons breaks complete factorial balance. A numerical SVD remains possible, but the two-value exact pseudoinverse formula no longer applies directly.

The complete-code formula is therefore not merely a result about one-hot encoding. It is a result about complete and uniformly weighted word space.

L.38 Dependence on Positional One-Hot Encoding

The formulas also depend on the chosen representation.

A different encoding of alphabet symbols would produce a different matrix.

For example, one could encode q symbols using fewer nonbinary contrast coordinates. Such a representation might have full column rank and no displayed block-offset redundancy.

The present one-hot matrix is valuable because:

every coordinate has direct categorical meaning;

every word has the same sparse row structure;

complete balance is transparent;

and the pseudoinverse maps directly back to position-symbol coefficients.

The rank deficiency is not a flaw. It records the redundant representation of the global constant field across the ℓ positional blocks.

L.39 Significance of the Generalization

The general theorem clarifies the status of the exact codon pseudoinverse.

Its existence is not produced by the biological translation assignments.

It is forced by:

the four-state input alphabet;

word length three;

complete inclusion of all sixty-four codons;

uniform weighting;

and positional one-hot encoding.

Any complete three-letter word system over four symbols would have the same:

rank;

singular spectrum;

two-value pseudoinverse;

gauge dimension;

and Hamming-distance projector.

What is biologically specific is the numerical or categorical field placed over that input space.

The standard genetic code supplies:

amino-acid identities;

translation stop;

hydropathy organization;

synonymous fibers;

and G–A–C–U topology.

The general matrix supplies an exact instrument for examining those assignments.

The unexpected feature is therefore not that balanced complete word space admits a pseudoinverse once the general theorem is known. The notable historical fact is that a natural lookup relation is placed over such a complete word space and can be analyzed through an unusually simple exact inverse structure.

L.40 General Theorem Summary

For a complete length-ℓ word space over an alphabet of q symbols, the positional one-hot incidence matrix has dimensions

q^ℓ × ℓq.

Its rank is

1 + ℓ (q − 1).

Its right nullity is

ℓ − 1.

Its left nullity is

q^ℓ − 1 − ℓ (q − 1).

Its singular spectrum is:

√ℓ q^((ℓ−1)/2), once;

q^((ℓ−1)/2), with multiplicity ℓ (q − 1);

and

0, with multiplicity ℓ − 1.

Its exact Moore–Penrose pseudoinverse is

D^{+}

= q^(1−ℓ) D^{T}

− (ℓ − 1) / (ℓ q^ℓ) J.

Its two entry values are:

[ℓ (q − 1) + 1] / (ℓ q^ℓ)

when the coordinate is present;

and

− (ℓ − 1) / (ℓ q^ℓ)

when it is absent.

Every pseudoinverse row sums to

1 / ℓ.

Every pseudoinverse column sums to

q^(1−ℓ).

The word-space projector is

P

= q^(1−ℓ) D D^{T}

− (ℓ − 1) / q^ℓ J,

and its entries depend only on Hamming distance.

The genetic-code formulas are recovered exactly by setting

q = 4

and

$\ell = 3$.

The complete codon matrix is therefore one biologically important instance of a general exact family of finite alphabet–word incidence matrices.

Appendix M

Inclusion–Exclusion and the Space of Complete Genetic Codes

M.1 Purpose

The standard genetic code is one function from a sixty-four-member codon domain to a twenty-one-member translated-output alphabet.

The input domain is

$C = \{ G, A, C, U \}^3$,

with

$| C | = 64$.

The translated-output alphabet is

Ω,

with

$| \Omega | = 21$,

containing twenty amino-acid classes and translation stop X.

Many alternative maps can be written from C to Ω. Their number depends strongly on which properties are required to remain fixed.

One may count:

all input-complete lookup tables;

only tables using every translated output;

tables preserving the exact standard degeneracy of each named output;

tables preserving only the multiset of degeneracy sizes;

tables preserving three stop codons;

or much narrower families preserving codon boxes, synonymous topology, or physicochemical organization.

These spaces should not be confused.

A combinatorial count identifies the size of a declared mathematical ensemble. It does not automatically supply a historical probability.

M.2 Input Completeness and Output Completeness

Every function

$\tau : C \to \Omega$

assigns one translated output to every codon.

Such a map is complete with respect to the input domain because no codon is left unassigned.

It need not use every member of Ω. Some translated outputs may receive no codons.

It is therefore useful to distinguish two meanings of completeness.

An **input-complete code** assigns one output to each of the sixty-four codons.

An **output-complete code** is also surjective: every one of the twenty-one outputs occurs at least once.

The standard genetic code satisfies both conditions.

The phrase **complete code** should therefore be accompanied by enough context to show whether surjectivity is required.

M.3 All Input-Complete Lookup Tables

Every one of the sixty-four labeled codons may be assigned independently to any of twenty-one labeled outputs.

The total number of functions is

21^{64}.

Numerically,

21^{64}

$\approx 4.188272 \times 10^{84}$.

This is the broadest labeled lookup-table space.

It permits:

unused amino-acid outputs;

no stop codons;

many stop codons;

extreme degeneracy;

and tables having almost no resemblance to a biological genetic code.

The count measures abstract assignment freedom. It does not describe a biologically realistic null ensemble by itself.

M.4 Surjective Genetic Codes

To require every translated output to occur at least once, the function must be surjective.

The number of surjective maps from sixty-four labeled codons onto twenty-one labeled outputs can be calculated by inclusion–exclusion.

Begin with all

21^{64}

functions.

Subtract functions omitting at least one output.

Add back functions omitting at least two outputs, because they were subtracted more than once.

Continue alternating subtraction and addition.

The result is

Nsurj

$= \Sigma j=0^{21} (-1)^{j} C (21, j) (21 - j)^{64}.$

Written term by term,

Nsurj

$= 21^{64}$

$- C (21, 1) 20^{64}$

- $C (21, 2) 19^{64}$

$- C(21, 3) 18^{64}$

- …
- $(-1)^{20} C(21, 20) 1^{64}$.

The final j = 21 term is zero because

$0^{64} = 0$.

The exact value is

1,510,109,515,792,918,244,116,781,339,315,785,081,841,294,607,960,614,956,302,330,123,544,242,628,820,336,640,000.

In scientific notation,

Nsurj

$\approx 1.510110 \times 10^{84}$.

This is still an enormous space.

M.5 Stirling-Number Form

The same surjection count can be expressed using a Stirling number of the second kind.

The quantity

$S(64, 21)$

counts partitions of the sixty-four labeled codons into twenty-one nonempty unlabeled fibers.

After the twenty-one fibers have been formed, they can be assigned to the twenty-one labeled translated outputs in

21!

ways.

Therefore,

Nsurj

$= 21! S(64, 21)$.

The inclusion–exclusion and Stirling-number formulas are exactly equivalent:

$21! S(64, 21)$

$= \Sigma_{j=0}^{21} (-1)^j C(21, j) (21 - j)^{64}$.

The two expressions emphasize different structures.

Inclusion–exclusion begins with all assignments and removes those omitting outputs.

The Stirling-number form first partitions the codons into nonempty synonymous fibers and then names the fibers.

M.6 Surjectivity Under the Uniform Function Ensemble

Under the purely mathematical ensemble in which all

21^{64}

functions are equally likely, the fraction that use all twenty-one outputs is

Nsurj / 21^{64}.

Numerically,

Nsurj / 21^{64}

≈ 0.360557.

Thus, approximately 36.1 percent of unrestricted uniformly sampled functions are surjective.

This percentage belongs only to the stated uniform-function ensemble.

It is not the probability that a biological genetic code would use all twenty amino acids and stop.

A historical or physical process need not sample the abstract function space uniformly.

M.7 The Standard Degeneracy Assignment

The standard code has the following named output degeneracies:

six codons each for L, R, and S;

four codons each for A, G, P, T, and V;

three codons each for I and X;

two codons each for C, D, E, F, H, K, N, Q, and Y;

and

one codon each for M and W.

The multiplicity pattern is therefore:

three outputs of size six;

five outputs of size four;

two outputs of size three;

nine outputs of size two;

and

two outputs of size one.

The total is

3 (6)

- 5 (4)
- 2 (3)
- 9 (2)
- 2 (1)

= 64.

This degeneracy profile is much more restrictive than surjectivity alone.

M.8 Preserving the Exact Degeneracy of Every Named Output

Suppose every named translated output must retain exactly its standard number of codons.

The codons themselves remain labeled.

The number of assignments is the multinomial count

Nnamed

=

64!

_____

$(6!)^3 (4!)^5 (3!)^2 (2!)^9$.

The factors for one-codon outputs are

$1!^2 = 1$

and may be omitted.

The exact value is

2,316,278,295,874,198,383,354,290,667,060,208,126,761,409,719,904,245,070,233,600,000,000.

In scientific notation,

Nnamed

≈ 2,316,278,295,874,198,383,354,290,667,060,208,126,761,409,719,904,245,070,233,600,000,000,000

This ensemble preserves, for example:

six codons for leucine;

six for arginine;

six for serine;

three for X;

one for methionine;

and one for tryptophan.

It does not preserve which codons belong to those outputs.

The codon fibers may be placed almost anywhere in the sixty-four-member domain, subject only to their required sizes.

M.9 Derivation of the Named-Degeneracy Count

The multinomial formula can be understood as a sequence of choices.

Choose six of the sixty-four codons for the first six-codon output.

Choose six of the remaining fifty-eight for the second.

Choose six of the remaining fifty-two for the third.

Continue through all twenty-one named outputs.

The product of these sequential binomial choices simplifies to

64! / Πω∈Ω dω!,

where dω is the required degeneracy of output ω.

For the standard degeneracies,

Πω∈Ω dω!

= (6!)³ (4!)⁵ (3!)² (2!)⁹.

Because the outputs are already named, no additional factor of 21! is included.

M.10 Preserving Only the Degeneracy Multiset

A broader ensemble may preserve only the multiset

{ 6, 6, 6, 4, 4, 4, 4, 4, 3, 3,

2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1 }

without requiring particular named outputs to retain their standard sizes.

The twenty-one output labels may then be assigned to the size classes in

$$\frac{21!}{3!\,5!\,2!\,9!\,2!}$$

ways.

The denominator corrects for permutations among outputs receiving the same degeneracy.

The number of size assignments is

48,886,437,600.

For each size assignment, the codons can be distributed according to the same multinomial factor as before.

Therefore,

$$N_{\text{multiset}} = \frac{64!}{(6!)^3\,(4!)^5\,(3!)^2\,(2!)^9} \times \frac{21!}{3!\,5!\,2!\,9!\,2!}.$$

Numerically,

Nmultiset

≈ 1.132346 × 10^{80}.

This ensemble is about

48.9 billion

times larger than the ensemble preserving the degeneracy of each named output.

M.11 Unlabeled Fiber Partitions

One may also count the synonymous partitions without naming the translated outputs.

The number of partitions of sixty-four labeled codons into unlabeled blocks having the standard degeneracy multiset is

Nunlabeled

=

64!

———————————————————————————————

$(6!)^3 (4!)^5 (3!)^2 (2!)^9$ 3! 5! 2! 9! 2!.

The first group of factorials removes internal order within each fiber.

The second group removes exchange among fibers of equal size.

Naming the twenty-one fibers then multiplies this count by 21!, yielding Nmultiset.

If instead each particular output is required to receive its standard degeneracy, the relevant count is Nnamed.

The distinction among unlabeled partitions, freely named partitions, and fixed named degeneracies should always be stated.

M.12 Relative Sizes of the Principal Spaces

The four principal assignment spaces have approximate base-ten sizes:

all input-complete maps:

4.188272 × 10^{84};

surjective maps:

1.510110×10^{84};

maps preserving only the standard degeneracy multiset:

1.132346×10^{80};

and

maps preserving every named standard degeneracy:

2.316278×10^{69}.

The named-degeneracy ensemble is approximately

$10^{14.814}$

times smaller than the surjective ensemble.

It is approximately

$10^{15.257}$

times smaller than the unrestricted function space.

Even after that enormous reduction, it still contains more than

10^{69}

assignments.

Fixing degeneracy therefore removes a great deal of combinatorial freedom but leaves an immense space of possible codon placements.

M.13 Information Required to Specify an Assignment

The logarithm base two of an ensemble size gives the number of binary decisions required, in an idealized counting sense, to identify one member of that ensemble.

For all input-complete functions,

$\log_2 (21^{64})$

≈ 281.108 bits.

For all surjective functions,

$\log_2 (Nsurj)$

≈ 279.637 bits.

For the standard degeneracy multiset ensemble,

$\log_2 (N_{multiset})$

≈ 265.934 bits.

For the named standard-degeneracy ensemble,

$\log_2 (N_{named})$

≈ 230.425 bits.

These values do not measure biological information directly.

They are logarithms of mathematical ensemble sizes.

They show that fixing the exact degeneracy of every named output removes approximately

50.7 bits

of assignment freedom relative to the unrestricted function space, while still leaving more than 230 bits of combinatorial choice.

M.14 Exactly Three Stop Codons

One intermediate ensemble requires exactly three codons to translate to X but allows the remaining sixty-one codons to be assigned arbitrarily among the twenty amino acids.

First choose the three stop codons:

$C (64, 3)$.

Then assign each remaining codon independently to one of twenty amino-acid outputs:

20^{61}.

The number of such maps is

$C (64, 3) 20^{61}$.

This ensemble does not require every amino acid to be used.

If all twenty amino acids must occur at least once among the sixty-one sense codons, the number is

$C (64, 3) 20! S (61, 20)$.

Equivalently, inclusion–exclusion gives

$C (64, 3)$

$\Sigma_{j=0}^{20} (-1)^j C (20, j) (20 - j)^{61}$.

Neither ensemble preserves the standard positions of UAA, UAG, and UGA.

M.15 Fixing the Stop Codons

A narrower ensemble may fix

UAA,

UAG,

and

UGA

as X while assigning the remaining sixty-one codons among the twenty amino acids.

If amino-acid use is unrestricted, the number of assignments is

20^{61}.

If every amino acid must occur at least once, the number is

20! S (61, 20).

One may narrow the ensemble further by preserving the standard degeneracy of every amino acid.

Then the relevant count is

$$\frac{61!}{(6!)^3 (4!)^5 3! (2!)^9},$$

because the three stop assignments are already fixed and isoleucine remains the one three-codon amino-acid class.

These progressively narrower spaces answer different questions.

M.16 Degeneracy Does Not Determine Codon Geometry

The named-degeneracy count preserves the number of codons in every translated fiber.

It does not preserve the geometric arrangement of those codons.

Consider the three six-codon outputs:

arginine;

leucine;

and serine.

Their degeneracies are identical.

Their G–A–C–U self-loop multiplicities are not:

R = 6;

L = 6;

and

S = 4.

The difference arises because the codons occupy different locations in codon space.

A reassignment preserving all three fiber sizes can produce:

connected fibers;

disconnected fibers;

compact fibers;

widely dispersed fibers;

or fibers having very different numbers of Hamming-distance-one internal edges.

Thus, the multinomial count leaves topology largely unconstrained.

M.17 Degeneracy Does Not Determine Position-Specific Synonymy

The standard G–A–C–U cube has synonymous-edge decomposition

4 + 1 + 39 = 44.

The exact named degeneracy profile does not force this distribution.

A randomized assignment with the same fiber sizes could place many synonymous codon pairs at the first position, the second position, the third position, or none of them under a selected nucleotide path.

Even preserving the total number of synonymous cube edges would not automatically preserve their positional decomposition.

The structure

4, 1, and 39

contains information beyond the fiber cardinalities.

It reflects where the synonymous codons lie relative to one another.

M.18 Degeneracy Does Not Determine the G–A–C–U Maximum

The six total nucleotide-pair weights of the standard code are

A–G = 15;

A–C = 11;

A–U = 9;

C–G = 8;

C–U = 18;

and

G–U = 8.

Their arrangement produces the unique isotropic maximum

G–A–C–U,

with score forty-four.

A reassignment preserving the exact standard degeneracies may alter all six weights.

Another path may become optimal.

Several paths may tie.

The maximum may increase or decrease.

The common intersection of the three positional maximizing families may disappear.

Therefore, the G–A–C–U theorem is assignment specific. It is not a mathematical consequence of the degeneracy profile alone.

M.19 Topology Does Not Determine Chemistry

Suppose the standard synonymous fibers are held fixed, preserving every synonymous edge and every self-loop multiplicity.

The amino-acid labels could still be permuted among fibers of compatible size.

Such a permutation would preserve:

fiber geometry;

the forty-four synonymous edges;

the complete loop topology;

and the G–A–C–U path score.

It could alter:

the hydropathy field;

the second-position U–A contrast;

the residue-volume distribution;

the reconstructed amino-acid properties;

and the nucleotide-only membrane profile.

Thus, synonymous topology does not determine which physicochemical values occupy the fibers.

Topological and physicochemical organization are related in the standard code but remain mathematically distinguishable dimensions.

M.20 Chemistry Does Not Determine Exact Synonymous Topology

The reverse distinction also holds.

A randomization might preserve broad hydrophobic, polar, charged, and small-residue groupings while changing which specific codons are synonymous.

The second-position hydropathy polarity could remain strong even as:

self-loop multiplicities changed;

fiber connectivity changed;

and another nucleotide path became topologically optimal.

Property preservation is weaker than translated-identity preservation.

Two different amino acids may have similar hydropathy while belonging to distinct translated classes.

The hydropathy projection and synonymous topology therefore require separate null ensembles.

M.21 Codon-Box-Preserving Ensembles

The conventional genetic-code table is organized into sixteen boxes determined by the first two codon positions.

Each box contains four third-position codons.

A box-preserving ensemble may require some or all of the following:

the sixteen boxes remain intact;

the same boxes remain split or unsplit;

four-codon families remain four-codon boxes;

two-versus-two splits remain two-versus-two splits;

the six-codon family architecture remains intact;

the three stop codons remain in their standard boxes;

or the complete third-position degeneracy pattern remains unchanged.

There is no single universal count for a “box-preserving code.”

The answer depends on which box features are fixed and which labels may be permuted.

Such ensembles are usually much smaller and biologically closer to the standard code than unrestricted degeneracy-preserving assignments.

They are also more appropriate for testing whether a property follows automatically from the conventional box structure.

M.22 Fiber-Preserving Label Permutations

Another narrow ensemble retains every codon fiber exactly but permutes translated labels among the fibers.

If all twenty-one fibers are considered exchangeable, there are

21!

label permutations.

If only fibers of equal degeneracy may exchange labels, the number is

3! 5! 2! 9! 2!.

The equal-degeneracy restriction preserves the named degeneracy sizes but changes which amino-acid properties occupy the geometrically fixed fibers.

This provides a direct null model for physicochemical alignment.

Every synonymous graph statistic remains unchanged.

Only the association between fiber geometry and translated chemistry changes.

M.23 Property-Value Permutations

An even narrower ensemble leaves the complete translation map unchanged but permutes numerical property values among the twenty amino acids.

The synonymous fibers and topology remain exactly fixed.

Only the property field changes.

There are

20!

unrestricted permutations of twenty distinct property values.

If X is retained with a fixed numerical convention, it is not included in the permutation.

Constrained property permutations may preserve broad classes, such as hydrophobic, charged, or polar groups.

These ensembles test whether the observed second-position coefficient hierarchy is unusually strong relative to the fixed codon topology.

M.24 A Hierarchy of Code Spaces

The principal ensembles can be ordered from broadest to narrowest:

all functions from sixty-four codons to twenty-one outputs;

surjective functions;

functions with exactly three stop codons;

functions preserving the standard degeneracy multiset;

functions preserving the standard degeneracy of every named output;

functions preserving stop positions;

functions preserving codon-box architecture;

functions preserving the complete synonymous fibers;

fiber-preserving translated-label permutations;

and

property-value permutations over a fixed translation map.

Each level holds more structure constant.

A result that is exceptional in a broad ensemble may become ordinary in a narrow ensemble because the narrow model already preserves much of the organization responsible for it.

A result that remains exceptional under several increasingly restrictive ensembles is more difficult to explain as a consequence of the preserved features alone.

M.25 No Unique Probability of the Standard Code

There is no ensemble-independent quantity called

the probability of the standard genetic code.

Under the unrestricted uniform function ensemble, every particular labeled map has probability

$1 / 21^{64}$.

Under the uniform surjective ensemble, every surjective map has probability

1 / Nsurj.

Under the fixed named-degeneracy ensemble, every permitted assignment has probability

1 / Nnamed.

These probabilities differ because the sampling spaces differ.

None is automatically a model of historical code formation.

A physical process may strongly favor some assignments, prohibit others, and move through code space by local transitions rather than drawing one complete table uniformly at random.

Combinatorial rarity and historical improbability are therefore not equivalent.

M.26 Accessibility Versus Cardinality

The cardinality of an assignment space records how many states it contains.

It does not record how those states are connected by possible transitions.

Two code ensembles may have similar sizes but very different accessibility structures.

One assignment may be reachable from another by a single codon reassignment.

Another may require many coordinated changes.

Intermediate assignments may or may not maintain viable translation.

A historical model therefore needs more than a state count.

It requires:

allowed transitions;

transition rates;

fitness or viability constraints;

molecular compatibility;

and path dependence.

The combinatorial spaces in this appendix define possible endpoints. They do not define the dynamics connecting them.

M.27 Local Reassignment Graphs

One may place a graph on code space itself.

Each vertex represents one complete codon-to-output assignment.

Two code vertices are adjacent when they differ in the assignment of one codon.

In the unrestricted function space, every code has

$64 \times 20 = 1{,}280$

one-reassignment neighbors, because any of sixty-four codons can be changed to any of twenty alternative outputs.

The number is smaller near constrained boundaries if surjectivity, exact degeneracy, or box structure must be preserved.

For example, an exact-degeneracy-preserving move cannot reassign only one codon. At least two codons must exchange outputs, or a more complex balanced move must occur.

The geometry of the comparison ensemble therefore changes when its invariants change.

M.28 Exact-Degeneracy Swap Space

Within the fixed named-degeneracy ensemble, a simple local move is a swap of translated assignments between two codons currently belonging to different outputs.

Such a swap preserves every named degeneracy.

There are many possible swaps, but several may produce the same resulting code only if the selected codons or labels coincide, which they do not in a two-codon exchange.

This swap graph connects the ensemble of assignments having fixed named fiber sizes.

It provides a natural space for:

local sensitivity analysis;

hill climbing under a declared score;

random walks;

and comparison of the standard code with nearby assignments.

The G–A–C–U score, hydropathy projection strength, and other properties can be tracked under each allowed swap.

M.29 Multiple Objectives

A genetic code may be evaluated by many different functions.

Examples include:

synonymous-edge count;

hydropathy projection strength;

residue-volume organization;

mutation robustness;

mistranslation cost;

stop isolation;

codon-box compactness;

and preservation of biosynthetic or chemical groupings.

One code may score well under one objective and less well under another.

The standard code should not be described as globally optimal without specifying:

the objective;

the weighting;

the comparison ensemble;

and whether the optimum is exact or approximate.

The forty-four-edge theorem satisfies these requirements because its graph class and objective are explicitly defined.

Broader historical optimization claims require equally explicit definitions.

M.30 Joint Exceptional Structure

The most informative comparison may concern several properties simultaneously.

A random code might produce a strong second-position hydropathy contrast.

Another might produce forty-four or more synonymous edges under some topology.

Another might preserve compact fibers.

The stronger question is how often one assignment produces the observed combination:

a dominant second-position hydropathy polarity;

extensive third-position synonymy;

the exact G–A–C–U maximum;

the position-specific decomposition 4 + 1 + 39;

and the unique common-path property.

Joint distributions are required to answer that question.

Multiplying separate marginal probabilities would be invalid unless the properties were independent.

They are unlikely to be independent because they arise from the same codon assignments.

M.31 Exact Enumeration and Sampling

The broad spaces are too large for direct complete enumeration.

For example,

Nnamed

exceeds

10^{69}.

Statistical analysis of such a space generally requires:

uniform or weighted random sampling;

Markov-chain methods;

importance sampling;

or analytic combinatorics.

The sampling algorithm must be shown to generate the declared ensemble correctly.

A naive reassignment procedure may fail to preserve degeneracy or may sample allowed codes nonuniformly.

For smaller constrained spaces, exact enumeration may still be possible.

The twenty-four isotropic nucleotide paths and 13,824 anisotropic axis triples are examples in which the complete comparison class can be enumerated without sampling.

M.32 Uniform Sampling with Fixed Named Degeneracies

A uniform sample from the fixed named-degeneracy ensemble can be generated by randomly permuting the sixty-four codons and then allocating consecutive segments of the permutation according to the required named fiber sizes.

For example:

the first six codons are assigned to the first six-codon output;

the next six to the second;

and so forth.

Provided that every codon permutation is equally likely and the output allocation order is fixed, every permitted named-degeneracy assignment is generated with equal probability.

The reason is that each final assignment corresponds to exactly

Πω dω!

codon permutations, one for every internal ordering within each output block.

Because this number is the same for every assignment, the resulting distribution is uniform.

M.33 Uniform Sampling with Only the Degeneracy Multiset Fixed

To sample uniformly when only the multiset of fiber sizes is fixed, one may:

randomly assign the twenty-one named outputs to the multiset of sizes with the correct uniform weighting;

randomly permute the sixty-four codons;

and allocate codons according to the selected size assignment.

The first step must avoid overcounting exchanges among equal sizes.

Equivalently, one can randomly choose which three outputs receive size six, which five receive size four, and so forth, using uniform combinations.

The procedure should be documented and tested because seemingly simple randomization routines can bias the output-size assignment.

M.34 Inclusion–Exclusion as a General Tool

The surjection formula is one example of a broader inclusion–exclusion method.

Whenever a code ensemble requires every member of a selected output set to occur, one may begin with all functions and alternately subtract and add assignments omitting selected outputs.

For n codons and k required outputs, the number of surjections is

$\Sigma j=0^{k} (-1)^{j} C (k, j) (k - j)^{n}$.

Equivalently,

$k! S (n, k)$.

For the complete genetic-code domain,

$n = 64$

and

$k = 21$.

For the sixty-one sense codons mapped onto twenty amino acids,

$n = 61$

and

$k = 20$.

The same method applies to any finite lookup system.

M.35 What the Counts Establish

The combinatorial counts establish several exact conclusions.

The unrestricted codon-to-output assignment space is enormous.

Requiring every translated output to occur reduces the space but leaves it of the same order of magnitude.

Preserving the standard degeneracy multiset removes several orders of magnitude.

Preserving every named degeneracy removes many more.

Even the fixed named-degeneracy space still contains more than

10^{69}

assignments.

Therefore, degeneracy alone does not specify the standard code.

Large freedom remains in:

fiber placement;

fiber connectivity;

position-specific synonymy;

nucleotide-pair weights;

path maxima;

and physicochemical assignment.

These dimensions must be measured separately.

M.36 What the Counts Do Not Establish

The counts do not establish that the standard code was selected uniformly from any one of these spaces.

They do not establish a historical probability.

They do not determine which assignments are chemically realizable.

They do not specify which intermediate codes are viable.

They do not show that every codon reassignment is equally accessible.

They do not prove optimization, intention, or accident.

They define finite comparison spaces.

Their scientific use begins when an observable property of the standard code is located within a clearly specified ensemble.

M.37 Summary

The broadest input-complete code space contains

21^{64}

labeled codon-to-output maps.

The output-complete, surjective space contains

$\Sigma j=0^{21} (-1)^{j} C (21, j) (21 - j)^{64}$

=

$21! S (64, 21)$

maps.

Preserving the exact standard degeneracy of every named translated output gives

$\frac{64!}{(6!)^{3} (4!)^{5} (3!)^{2} (2!)^{9}}$

$\approx 2.316278 \times 10^{69}$

assignments.

Preserving only the standard degeneracy multiset gives

approximately

1.132346×10^{80}

assignments.

These numbers show that even strong coarse constraints leave vast freedom in the placement of translated outputs over codon space.

Degeneracy does not determine synonymous topology.

Topology does not determine amino-acid chemistry.

Neither determines historical probability without a model of accessible transitions.

The proper role of combinatorial counting is therefore to define comparison ensembles precisely. Claims that the standard code is ordinary, unusual, optimized, or accidental acquire

mathematical meaning only after the preserved features, allowed transformations, and sampling rules have been declared.

References

Arkin, A. P., and D. C. Youvan. “An Algorithm for Protein Engineering: Simulations of Recursive Ensemble Mutagenesis.” *Proceedings of the National Academy of Sciences of the United States of America* 89 (1992): 7811–7815.

Crick, F. H. C. “The Origin of the Genetic Code.” *Journal of Molecular Biology* 38 (1968): 367–379.

Freeland, S. J., and L. D. Hurst. “The Genetic Code Is One in a Million.” *Journal of Molecular Evolution* 47 (1998): 238–248. DOI: 10.1007/PL00006381.

Haig, D., and L. D. Hurst. “A Quantitative Measure of Error Minimization in the Genetic Code.” *Journal of Molecular Evolution* 33 (1991): 412–417. DOI: 10.1007/BF02103132.

Harary, F. *Graph Theory*. Reading, Massachusetts: Addison-Wesley, 1969.

Hopp, T. P., and K. R. Woods. “Prediction of Protein Antigenic Determinants from Amino Acid Sequences.” *Proceedings of the National Academy of Sciences of the United States of America* 78 (1981): 3824–3828. DOI: 10.1073/pnas.78.6.3824.

Kyte, J., and R. F. Doolittle. “A Simple Method for Displaying the Hydropathic Character of a Protein.” *Journal of Molecular Biology* 157 (1982): 105–132. DOI: 10.1016/0022-2836(82)90515-0.

Moore, E. H. “On the Reciprocal of the General Algebraic Matrix.” *Bulletin of the American Mathematical Society* 26 (1920): 394–395.

Newman, M. E. J. *Networks: An Introduction*. Oxford: Oxford University Press, 2010. DOI: 10.1093/acprof:oso/9780199206650.001.0001.

Penrose, R. “A Generalized Inverse for Matrices.” *Proceedings of the Cambridge Philosophical Society* 51 (1955): 406–413. DOI: 10.1017/S0305004100030401.

Sonneborn, T. M. “Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications.” In *Evolving Genes and Proteins*, edited by V. Bryson and H. J. Vogel, 377–397. New York: Academic Press, 1965. DOI: 10.1016/B978-1-4832-2734-4.50034-6.

Wimley, W. C., and S. H. White. “Experimentally Determined Hydrophobicity Scale for Proteins at Membrane Interfaces.” *Nature Structural Biology* 3 (1996): 842–848. DOI: 10.1038/nsb1096-842.

Woese, C. R. *The Genetic Code: The Molecular Basis for Genetic Expression*. New York: Harper & Row, 1967.

Yang, M. M., W. J. Coleman, and D. C. Youvan. “Genetic Coding Algorithms for Engineering Membrane Proteins.” In *Reaction Centers of Photosynthetic Bacteria*, edited by M.-E. Michel-Beyerle, 209–218. Springer Series in Biophysics, volume 6. Berlin and Heidelberg: Springer-Verlag, 1990. DOI: 10.1007/978-3-642-61297-8_21.

Zamyatnin, A. A. “Protein Volume in Solution.” *Progress in Biophysics and Molecular Biology* 24 (1972): 107–123. DOI: 10.1016/0079-6107(72)90005-3.

Zamyatnin, A. A. “Amino Acid, Peptide, and Protein Volume in Solution.” *Annual Review of Biophysics and Bioengineering* 13 (1984): 145–165. DOI: 10.1146/annurev.bb.13.060184.001045.

Author’s Related Mathematical and Computational Studies

Youvan, D. C. “The Genetic Code as Physicochemical and Topological Geometry.” Working manuscript and compiled source volume.

Youvan, D. C. “Moore–Penrose Pseudoinverse of the Genetic Code.” 2023. DOI: 10.13140/RG.2.2.18540.03203.

Youvan, D. C. “Mathematica Singular-Value Decomposition of Nucleotide-Determined Amino-Acid Hydropathy.” 2023. DOI: 10.13140/RG.2.2.11252.42883.

Youvan, D. C. “Hydropathy and Residue-Volume Organization of the Genetic Code.” 2023. DOI: 10.13140/RG.2.2.10740.94083.

Youvan, D. C. “Loops in the G–A–C–U Genetic-Code Graph.” 2022. DOI: 10.13140/RG.2.2.21318.55362.

Youvan, D. C., and Juvon. “Geometric and Topological Representations of the Genetic Code.” 2022. DOI: 10.13140/RG.2.2.34261.91361.