

Quantum-like Modelling of the Genetic Code

Elena Fimmel (joint work with Sergey V. Petoukhov)

Belgrade, May, 29th, 2025

Contents

- 1 Introduction
- 2 Definitions and Results
- 3 Conclusions and Prospects

A paradigmatic account of the uses of mathematics in the natural sciences comes, in deliberately oversimplified fashion, from the classic sequence of Brahe, Kepler, Newton: observed facts, patterns that give coherence to the observations, fundamental



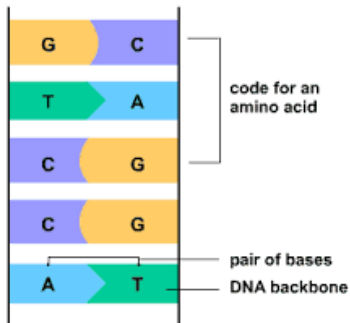
laws that explain the patterns.[...] The virtue of mathematics in such a context is that it forces clarity and precision upon the conjecture, thus enabling meaningful comparison between the consequences of basic assumptions and the empirical facts. Here mathematics is seen in its quintessence: no more, but no less, than a way of thinking clearly.

Robert M. May , “Uses and Abuses of Mathematics in Biology”, 2004 VOL 303 SCIENCE, pp 790-793

- *The term “quantum biology” was introduced in 1932 by one of creators of quantum mechanics P. Jordan. Jordan was convinced he could extend quantum indeterminism from the subatomic world to macroscopic biology.*
- *In quantum mechanics and quantum informatics one of the most important mathematical operations is the Kronecker tensor multiplication of two matrices.*
- *This talk describes special applications of the Kronecker tensor multiplication, which allow developing mathematical models and analytical approaches in bioinformatics and algebraic biology, first of all for analyzing structured alphabets of genetic molecules DNA.*

Structure of DNA

DNA is usually a double-helix and has two strands running in opposite directions. (There are some examples of viral DNA which are single-stranded). The rungs of the ladder are made up of pairs of base molecules connected to each other. There are only 4 different types of bases. Each is usually known by the first letter of its name: Adenine (A), Cytosine (C), Guanine (G), Thymine (T)(replaced by Uracil (U) in RNA).



How big is the human genome?

Remark

The human genome is about 3 billion base pairs long and contains around 30,000 genes. Since every base pair can be coded by 2 bits, this is about 750 megabytes of data. If one stretched the DNA in one cell all the way out, it would be about 2m long and all the DNA in all cells of one human put together would be about twice the diameter of the Solar System or nearly 70 trips from the earth to the sun and back.

Genetic code(s)

- The genetic code is the set of rules by which information encoded in genetic material (DNA or RNA sequences) is translated into proteins (amino acid sequences) by living cells.
- Each group of 3 bases on one side of the DNA, called codons, carries the genetic code for one of the 20 different amino acid molecules that build proteins. Once the whole code for one gene is read, the cell can make a specific protein.
- Because the vast majority of genes are encoded with exactly the same code, this particular code is often referred to as the canonical or standard genetic code, or simply the genetic code, though in fact there are many variant codes. While slight variations on the standard genetic code had been predicted earlier, none was discovered until 1979, when researchers studying human mitochondrial genes determined that they used an alternative code - the vertebrate mitochondrial code.

Standard genetic code

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
U	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
U	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
U	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
C	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
C	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
C	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
A	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
A	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
A	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
G	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
G	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
G	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Table: Standard genetic code

Vertebrate mitochondrial genetic code

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
U	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
U	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Trp	A
U	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
C	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
C	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
C	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
A	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
A	AUA	Met	ACA	Thr	AAA	Lys	AGA	Stop	A
A	AUG	Met	ACG	Thr	AAG	Lys	AGG	Stop	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
G	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
G	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
G	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

Table: Vertebrate mitochondrial code

Binary Oppositional Characteristics Of Nucleotide Bases

Let us denote the nucleotide bases alphabet as

$$\mathcal{B} = \{U(T), C, A, G\}.$$

- $\mathcal{B} = \{C, G\} \cup \{A, U\}$ 'strong' and 'weak' bases, **S/W**-partition
- $\mathcal{B} = \{C, A\} \cup \{U, G\}$ amino and keto bases, **Am/K** -partition
- $\mathcal{B} = \{C, U\} \cup \{A, G\}$ pyrimidine and purine bases, **Y/R**-partition.

Matrix representation of the genetic alphabet

We can arrange the DNA-Alphabet of 4 nucleotides as a 2×2 -matrix e.g. as follows:

$$\begin{pmatrix} C & A \\ T & G \end{pmatrix}$$

Definition of the Kronecker tensor product

Definition

Let A be a $m \times n$ - and B a $p \times q$ -matrix, $m, n, p, q \in \mathbb{N}$. The Kronecker product, denoted by \otimes , is an operation on two matrices of arbitrary size resulting in a block $m \cdot p \times n \cdot q$ - matrix:

$$A \otimes B = \begin{pmatrix} a_{11}B & \dots & a_{1n}B \\ & \dots & \\ a_{m1}B & \dots & a_{mn}B \end{pmatrix}$$

Remark

The Kronecker product should not be confused with the usual matrix multiplication, which is an entirely different operation.

2nd tensor power of the genetic alphabet

If the concatenation of the symbols is taken as the multiplication of the symbols, the 2-e tensor power of the matrix representing the genetic alphabet looks as follows:

$$\begin{pmatrix} C & A \\ T & G \end{pmatrix} \otimes \begin{pmatrix} C & A \\ T & G \end{pmatrix} = \begin{pmatrix} CC & CA & AC & AA \\ CT & CG & AT & AG \\ TC & TA & GC & GA \\ TT & TG & GT & GG \end{pmatrix}$$

3rd tensor power of the genetic alphabet

$$\begin{pmatrix} C & A \\ T & G \end{pmatrix} \otimes \begin{pmatrix} C & A \\ T & G \end{pmatrix} \otimes \begin{pmatrix} C & A \\ T & G \end{pmatrix} =$$

$$\begin{pmatrix} CCC & CCA & CAC & CAA & ACC & ACA & AAC & AAA \\ CCT & CCG & CAT & CAG & ACT & ACG & AAT & AAG \\ CTC & CTA & CGC & CGA & ATC & ATA & AGC & AGA \\ CTT & CTG & CGT & CGG & ATT & ATG & AGT & AGG \\ TCC & TCA & TAC & TAA & GCC & GCA & GAC & GAA \\ TCT & TCG & TAT & TAG & GCT & GCG & GAT & GAG \\ TTC & TTA & TGC & TGA & GTC & GTA & GGC & GGA \\ TTT & TTG & TGT & TGG & GTT & GTG & GGT & GGG \end{pmatrix}$$

A remarkable symmetry in the Vertebrate Mitochondrial Code

The entire set of 8 rows shows itself as a complect of 4 pairs of adjacent rows with identical lists of amino acids and stop-codons in each pair

CCC Pro	CCA Pro	CAC His	CAA Gln	ACC Thr	ACA Thr	AAC Asn	AAA Lys
CCT Pro	CCG Pro	CAT His	CAG Gln	ACT Thr	ACG Thr	AAT Asn	AAG Lys
CTC Leu	CTA Leu	CGC Arg	CGA Arg	ATC Ile	ATA Met	AGC Ser	AGA Stop
CTT Leu	CTG Leu	CGT Arg	CGG Arg	ATT Ile	ATG Met	AGT Ser	AGG Stop
TCC Ser	TCA Ser	TAC Tyr	TAA Stop	GCC Ala	GCA Ala	GAC Asp	GAA Glu
TCT Ser	TCG Ser	TAT Tyr	TAG Stop	GCT Ala	GCG Ala	GAT Asp	GAG Glu
TTC Phe	TTA Leu	TGC Cys	TGA Trp	GTC Val	GTA Val	GGC Gly	GGA Gly
TTT Phe	TTG Leu	TGT Cys	TGG Trp	GTT Val	GTG Val	GGT Gly	GGG Gly

How does the Standard Genetic Codes look in this representation?

CCC Pro	CCA Pro	CAC His	CAA Gln	ACC Thr	ACA Thr	AAC Asn	AAA Lys
CCT Pro	CCG Pro	CAT His	CAG Gln	ACT Thr	ACG Thr	AAT Asn	AAG Lys
CTC Leu	CTA Leu	CGC Arg	CGA Arg	ATC Ile	ATA Iso	AGC Ser	AGA Arg
CTT Leu	CTG Leu	CGT Arg	CGG Arg	ATT Ile	ATG Met	AGT Ser	AGG Arg
TCC Ser	TCA Ser	TAC Tyr	TAA Stop	GCC Ala	GCA Ala	GAC Asp	GAA Glu
TCT Ser	TCG Ser	TAT Tyr	TAG Stop	GCT Ala	GCG Ala	GAT Asp	GAG Glu
TTC Phe	TTA Leu	TGC Cys	TGA Stop	GTC Val	GTA Val	GGC Gly	GGA Gly
TTT Phe	TTG Leu	TGT Cys	TGG Trp	GTT Val	GTG Val	GGT Gly	GGG Gly

The tensor product and long DNA sequences

In the DNA double helix, complementary nucleobases A and T are connected by 2 hydrogen bonds and are called weak bases (W), while the other two, C and G (strong bases (S)), are connected by 3 hydrogen bonds. In long DNA, numbers of weak and strong bases are met with certain percentages denoted by %W and %S, which satisfy the condition %W + %S = 100%:

$$\begin{pmatrix} \%S & \%W \\ \%W & \%S \end{pmatrix} \otimes \begin{pmatrix} \%S & \%W \\ \%W & \%S \end{pmatrix} = \begin{pmatrix} \%S\%S & \%S\%W & \%W\%S & \%W\%W \\ \%S\%W & \%S\%S & \%W\%W & \%W\%S \\ \%W\%S & \%W\%W & \%S\%S & \%S\%W \\ \%W\%W & \%W\%S & \%S\%W & \%S\%S \end{pmatrix}$$

Hypothesis

Remark

- *Matrices, received above by means of the Kronecker tensor product, provokes a hypothesis that percentages of hydrogen doublets SS, SW, WS, WW in long DNA sequences can be modelled as the product of percentages of hydrogen monoplets %S and %W*
- *Since the multiplication of numbers is commutative, the following must apply*

$$\%SW \approx \%WS.$$

Testing the Hypothesis

Remark

The testing of this hypothesis on the materials of the eukaryotic and prokaryotic genomes confirmed its validity: the model values of percentages almost coincided with the phenomenological values of percentages in all tested genomes:

Testing the Hypothesis

Example

In the human chromosome Nr. 1, phenomenological percentages of hydrogen doublets are the following:

$$\%SS = 0.1626, \%SW = 0.2546, \%WS = 0.2546 \text{ and } \%WW = 0.3282.$$

Using phenomenological percentages of hydrogen monolets

$$\%S = 0.4172 \text{ and } \%W = 0.5828$$

in this chromosome, we calculate







$$\%S \cdot \%S = 0.1741 \approx \%SS, \quad \%S \cdot \%W = 0.2431 \approx \%SW,$$

$$\%W \cdot \%S = 0.2431 \approx \%WS, \quad \%W \cdot \%W = 0.3397 \approx \%WW$$

Conclusions

- *In the talk it was shown that with the help of the Kronecker product some genetic connections can be demonstrated very well*
- *It remains to be seen whether a coherent theory can be established on this basis.*

References

-  Robert M. May: Uses and Abuses of Mathematics in Biology, 2004 VOL 303 SCIENCE, pp 790-793.
-  Fimmel E., Petoukhov S.: Genetic Code Modeling from the Perspective of Quantum Informatics. – In: Advances in Artificial Systems for Medicine and Education II. / Hu Z.B., He M., Petoukhov S.V. (Eds), pp. 117-126, Springer (2020),
-  Hu Z.B., Petoukhov S.V., Petukhova E.S. On symmetries, resonances and photonic crystals in morphogenesis. - Biosystems, vol. 173, pp.165-173 (2018).
-  Fimmel, E., Gumbel, M., Karpuzoglu, A., Petoukhov S.V.: On comparing composition principles of long DNA sequences with those of random ones. BioSystems, Volume 180, June 2019, Pages 101-108
-  Petoukhov S.V., He M. Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications. IGI Global, Hershey, USA (2009)
-  Petoukhov S.V. Genetic coding and united-hypercomplex systems in the models of algebraic biology. Biosystems, v. 158, p. 31–46 (2017)