HUMA: Superordinate Genetic Code

Cordula Sachse-Seeboth*

Department of General Medicine, University of Greifswald, Bad Lauterberg, Germany

INTRODUCTION

Subject

Are mathematical functions hidden within the genome that are superordinate to the genetic code? Does mathomics form the basis of all downstream applications?

Scientific disciplines

Bioinformatics, genetics, coding theory, number theory (interdisciplinary).

Hypotheses

Main thesis: There are codes superordinate to the genetic code, which can be described as mathematical functions.

Argumentation for the hypotheses

In nature fractal patterns occur ubiquitously. This observation raises the question of how and whether fractal patterns are stored within the genetic code to determine the morphology of cells, organs and the general appearance, or whether an explanation can be deduced from previous knowledge alone [1].

- Palindromic sequences within DNA fulfil important biological functions. Palindromes and prime palindromes represent a special entity in mathematics.
- The main thesis should not be confused with the concept of DNA computing. In this technology, the genetic material is used as a storage and processing medium. The entire internet would fit into just one shoe box.
- Boolean calculations can be carried out on the basis of biomolecules such as DNA and RNA.
- Solving complicated calculations is possible. It is a parallel computer. DNA computing is thus the biological competition to quantum computers.
- Due to the fact that DNA can be used for complex calculations, the question arises whether this mechanism is not also used by the cell itself.

General conclusions and definitions

Under the assumption that the hypothesis turns out to be true, the following can be derived:

- Living cells continuously make calculations using their genetic material to control vital processes.
- Natura calculat.

In the following, the ability to calculate as well as the totality of possible calculations will be referred to as Mathome. The tools still to be established to analyze these processes are called mathomics.

Within the mathome, theoretically all mathematical structures known to us could appear and be associated with specific biological functions. Mathematical constants like number Pi, the Euler number, mathematical functions, prime numbers (series), mathematical series and natural constants would be of particular interest [2-5].

DERIVABLE HYPOTHESES

Checksum and checksum calculation

Hypothesis: The genetic material uses the principle of checksum calculation to detect defective DNA sections or foreign DNA.

Background

DNA polymerases check during replication whether a DNA copy has been copied correctly. Before any mitotic cell division, the cell's genome must be copied completely. Other enzyme systems repair DNA sections after damage (UV light, for example). Supercoiled DNA cannot be read (dormant genes), uncoiled DNA can be read by RNA polymerase (DNA can be copied or gene products can be produced) [6-8].

Consequences

Just as each computer file has its own checksum, this could apply by analogy to genes, gene families and the whole genome (for species with small genome size).

If the hypothesis is correct, it is necessary to check how the checksum calculation can be influenced. A faulty checksum calculation leads to an increased incidence of errors within the genetic code, which could manifest themselves in the form of diseases (e.g. tumors, autoimmune diseases) or a changed evolutionary speed [9,10].

Derived hypothesis: A faulty checksum calculation can result in mutations in the genetic material, a change in the speed of evolution and susceptibility to pathogens.

It is known that modifications of nucleotides control gene expression (how many copies of 80 the gene are transcribed and translated by the cell). Modifications (de)activate genes by changing the steric DNA structure, e.g. enzymatic DNA methylation.

Any modification could give different results in the checksum calculation, which could cause the enzyme and other transcription initiators not to start their regular activity.

It would therefore be necessary to find out whether the gene expression is under the influence of any checksum calculation. Important structural motifs of regulatory proteins that play a role in the binding to DNA include the helix-turn helix motif and zinc fingers [11].

Derived hypothesis: The result of the checksum calculation influences gene expression.

Defective or dysfunctional DNA repair systems contribute to the development of cancer (mutagenesis, carcinogenesis) and accelerate the ageing process. Over time, due to the limited accuracy of all DNA repair systems, errors accumulate, leading to physiological ageing.

Derived hypothesis: DNA/RNA polymerases use, amongst other things, the principle of checksum calculation.

The functionality of all processes involved in checksum calculation would be impaired if the checksums were to malfunction. This would lead to new diagnostic and therapeutic approaches. Furthermore, the question arises whether and to what extent an organism tolerates errors within checksum calculation.

The variable occurrence of 20 repetitions to 2000 repetitions of GCT triplets in the 3'UTR region is correlated with the severity and onset of the genetically determined myotonic dystrophy.

Derived hypothesis: The greater the deviation of the calculated checksum from the expected checksum (normal findings), the more severe the effect of the present change.

Programs and programming languages

Hypothesis: The genetic material contains modular programs (genes and gene groups) in the sense of software, which are controlled by superordinate structures or programs. Superordinate structures are able to form new gene segments and integrate these into existing programs.

DNA/RNA molecules etc. are the equivalent of hardware. DNA would thus assume a double role: It would be software and hardware at the same time. If a mathome exists, different

modules or mathematical functions could be understood as software units. Examples of programs that are controlled by superordinate modules would be apoptosis (controlled cell death), gluconeogenesis (new sugar production in the liver), signal cascades such as NF kappa B pathway (immune response, cell proliferation) and so on [12].

Hypothesis: The genetic material uses several programming languages. Replication (making copies of 125 DNA), transcription (reading DNA and translating it into RNA) and translation (translating mRNA into proteins) are only three of the already known programming languages.

The genetic material encodes three-dimensional information

Hypothesis: Morphological (architectural structure of a folded protein, cell organelles, a defined cell type or organ) and topographical (geographical position of cell organelles, cells, tissues and organs relative to each other) information is encoded within the DNA.

How location and image information is stored in the genetic material is not clarified in detail.

It might be promising to analyse whether the macroscopically existing fractal patterns can be found on the microscopic (DNA) level.

Worthwhile approaches are: turtle geometry, substitution rules or Lindenmayer languages (L-systems) and holography. DNA has been used to simulate holograms. L-systems are used in computer graphics to create fractals and for realistic modeling of plants.

Hypothesis: The interference patterns generated by DNA and other biomolecules contain spatial information.

Such interference patterns could play a role in the folding of proteins or compression of DNA on a small scale, but could also be useful for global functions such as spatial coding and morphology.

Checking the hypotheses

Software: The software to be developed serves as a basis for the quick verification of the above mentioned hypotheses. The software components are tools of mathomics.

- Programs that automatically transform numbers (rows) into different number systems (e.g. from decimal to binary or number systems on other bases and vice versa).
- Programs for automating the transformation of DNA sequences into numbers (series) in different number systems and vice versa.

In the context of DNA computing, the genetic code is digitized (representation of the genetic code as 0 column and 1 column). However, the genetic code has 4 different letters whose information content could be lost if extrapolated to 2 different pieces of information (0 or 1) [13-15].

Hypothesis: The genetic code uses different number systems.

If a mathome can be detected, it should be checked which number systems are used by the genetic material (e.g. binary number system, number system on base 4 or 8, decimal number system, hexadecimal system). Theoretically, several codes that are (un)dependent on each other are conceivable.

Software development for the representation of nucleotide sequences in multidimensional image information and image analysis (similarity).

- For the transformation of nucleotide sequences into fractal patterns and vice versa decoding of fractals into a nucleotide sequence.
- For transformation of a nucleotide sequence into turtle geometry patterns and vice versa for the transformation of a nucleotide sequence in Lindenmayer languages and vice versa.

Automatic alignment of nucleotide sequences in the corresponding databases (BLAST: Basic Local Alignment Search Tool, FASTA, ALIGN, SSEARCH, etc.) according to pre-defined search criteria.

Output of only relevant results with an E-value to be determined in advance.

Automatic comparison of calculated three-dimensional structures with those of real biomolecules [16-18].

Examples

Digitization of the genetic code: There are two nucleotide bases that are complementary to each other: Adenine (A) and Thymine (T), Cytosine (C) and Guanine (G). There could be several (un)dependent digital codes.

The base pair (A/T) is interpreted as 0, while the base pair (C/G) is interpreted as 1 and vice versa.

Other number systems: Any number Z can be represented to the base g as the sum of a sequence of powers of the base g:

$Z = x_0 * g^n + x1 * g^{n-1} +... + x_{n-1} * g1 + x_n * g^0$.

Number system on base 4: The number system on the base g=4 (4 bases=4 characters) consists of four different numbers: x=0, 1, 2 or 3.

The assignment of the digits to the four naturally occurring nucleotide bases (A, T, C, G) is arbitrary. In this case there are 4! =24 possible combinations, which initially have to be considered as being of equal rank. This means that from a previously defined number (sequence) 24 different number sequences can be generated in the number system on base each of which must be analyzed individually:

1 number sequence in the number system on the base of g=10 → Different number series in number system on the base g=4 →

24 different nucleotide sequences → 24x Alignment → 24 different results.

If highly significant alignment results are obtained several times under a defined assignment code (digit nucleotide code), it must be checked whether a universal code can be detected.

Hypothesis: An universal numerical nucleotide code or digit-nucleotide code exists.

Example of transformation of numbers in the decimal number system into the number system on base four and transformation into a nucleotide sequence with a predefined code and vice versa

To represent the prime number 883 in the number system on base 4:

$(Z)_4 = x * 4^n + x * 4^{n-1} +... + x * 4^1 + x * 4^0$

$(883)_{10} = 3 * 4^4 + 1 * 4^3 + 3 * 4^2 + 0 * 4^1 + 3 * 4^0 = (31303)_4$

- Assuming a previously agreed code applies (one out of 24): A =0, T=1, C=2, G=3
- The sequence of numbers 31303 is transformed into GTGAG.
- Since the reading direction must be considered, the sequence is entered in reverse order to the alignment tool. This results in the nucleotide sequence GAGTG.

Therefore, the sequence of the »calculated« nucleotide sequence must always be considered, since the reading direction is from right to left (smallest number is noted right-aligned).

In highly repetitive sequences, short sequences are often repeated n times:

Example: $(T_xG_y)_n$ with x and y between 1 and 4 at the 5' yeast telomer. Such sequences should be analyzed especially for mathematical functions.

- The nucleotide sequence TTGGGTTGGGTTGGGTGGTT GGGTTGGG given is $(T_2G_3)_5$.

- This results in the following sequence of numbers recording to the code.

A=0,T=1,C=2,G=3:

→ 11333113331133311333311333

- Reversal of the order due to the reading direction to be considered: $(33311333113331133311133311)_4 = (1114894042650613)_{10}$
- If only the short sequence $(33311)_4$ is considered as a corresponding number, the result is $(1013)_{10}$.
- If the above nucleotide sequence is considered a function in which the number of shortsequences 33311 is read from one to five times in succession, the following sequence of numbers results:
- 1013-1038325-1063245813-1088763713525-1114894042650613
- Such sequences of numbers could be relevant. A graphical representation of the results facilitates interpretation.

HUMA

Equivalent to the human genome project, the initiation of a HUMA project should be considered. HUMA stands for human

mathome project. This is undoubtedly only useful if the above hypotheses can be verified.

If the genetic material can perform complex calculations that find a physiological equivalent, this goes far beyond the applications of DNA computing.

The first phase would be to first capture all logical functions within the human genome. To facilitate this, the analysis of smaller genomes such as those of the bacterium *E. coli* will initially will simplify the process. In later phases, the genomes would be compared within the human species, followed by comparisons between different species.

Example of practical applications

Each gene or even individual gene segments would form a mathematical unit with defined functions. In order to ensure through redundancy that a meaningful gene sequence is not lost, it can be expected that individual gene segments would become independent at regular intervals. This could explain the co-evolution of prokaryotes/eukaryotes and viruses.

Hypothesis: The gene-virus link would lead to the fact that with the evolution of each new gene, a virus would automatically be created that could switch that gene on or off and thus cause damage.

For each gene there is a suitable virus. For example, the corona virus paralyses the gene for NF-kB, which plays a key role in the body's defense reaction and immune system. A virus could therefore be interpreted as malware.

Conversely, a prediction could be made as to which codes a virus must have if it corresponds to a defined gene segment. It would be possible to create specific drugs based on these codes in advance [19,20].

All hypotheses at a glance

- Are mathematical functions hidden within the genome that are superordinate to the genetic code?
- Living cells continuously make calculations with the help of their genetic material to control vital processes.
- The genetic material uses the principle of checksum calculation to detect defective DNA segments or foreign DNA.
- A faulty checksum calculation can result in mutations in the genetic material, a change in the speed of evolution and susceptibility to pathogens.
- The result of the checksum calculation influences the gene expression.
- DNA/RNA polymerases use the principle of checksum calculation.
- The greater the deviation of the calculated checksum from the expected checksum (normal result), the more serious the effect of the change in question.
- The genetic material contains modular programs (genes and gene groups) in the sense of software, which are controlled by superordinate structures or programs. Superordinate structures are able to form new gene segments and integrate these into existing programs.

- The genetic material makes use of various programming languages. Replication (making copies of DNA), transcription (reading DNA and translating it into RNA) and translation (translating mRNA into proteins) are only three of the programming languages already known.
- Morphological (architectural structure of a folded protein, cell organelles, a defined cell type or an organ) and topographical (geographical position of cell organelles, cells, tissues and organs in relation to each other) information is encoded within the DNA.
- The interference patterns generated by the DNA and other biomolecules contain spatial information.
- The genetic code uses different number systems.
- It exists an universal number nucleotide code or digit-nucleotide code.
- The gene-virus link would mean that with the evolution of each new gene, a virus would be automatically created that could switch that gene on or off and thus cause damage.

REFERENCES

1. Arber W. Promotion and limitation of genetic exchange. Sci. 1979;205(4404):361-365.

2. Weinstein JN. Searching for pharmacogenomic markers: the synergy between omic and hypothesis-driven research. Dis. Markers. 2001;17(2):77-88.

3. Hutchinson GB. The prediction of vertebrate promotor regions using differential hexamer frequency analysis. Comput Appl Biosci. 1996;12(5):391-8.

4. Yang CN, Yang CB. A DNA solution of SAT problem by a modified sticker model. BioSyst. 2005;81(1):1-9.

5. Lee JY, Shin SY, Park TH, Zhang BT. Solving traveling salesman problems with DNA molecules encoding numerical values. BioSystems. 2004;78(1-3):39-47.

6. Schmidt KA, Henkel CV, Rozenberg HG, Spaink HP. DNA computing using single-molecule hybridization detection. Nucleic Acids Res. 2004;32(17):4962-4968.

7. Su X, Smith LM. Demonstration of a universal surface DNA computer. Nucleic Acids Res. 2004;32(10):3115-3123.

8. ZhiXiang Y, Fengyue Z, Jin X. The general form of 0–1 programming problem based on DNA computing. Biosyst. 2003;70(1):73-78.

9. Braich RS, Chelyapov N, Johnson C, Rothemund PW, Adleman L. Solution of a 20-variable 3-SAT problem on a DNA computer. Sci. 2002;296(5567):499-502.

10. Lee JY, Bae S, Myoung J. Middle east respiratory syndrome Coronavirus encoded accessory proteins impair MDA5-and TBK1-mediated activation of NF-κB. J Microbiol Biotechnol. 2019;29(8):1316-1323.

11. Adleman LM. Molecular computation of solutions to combinatorial problems. Sci. 1994;266(5187):1021-1024.

12. Day GR, Blake RD. Statistical significance of symmetrical and repetitive segments in DNA. Nucleic Acids Res.1982;10(24): 8323-8339.

13. Lisnic B, Svetec IK, Saric H, NikolicI, Zgaga Z. Palindrome content of the yeast *Saccharomyces cerevisiae* genome. Curr Genet. 2005;47:289-297.

14. Gibas C, Jambeck P. Einfuhrung in die praktische Bioinformatik. O'Reilly; 2002.

15. Gunther K, Grelck K, Zilm T. Linux User Referenz. mitp-Verlag, Bonn, 2002.

16. West GB, Brown JH, Enquist BJ. The fourth dimension of life: fractal geometry and allometric scaling of organisms. sci. 1999;284(5420):1677-1679.

17. West GB, Brown JH, Enquist BJ. A general model for the origin of allometric scaling laws in biology. Sci. 1997;276(5309):122-126.

18. Pansera F. Fractals and cancer. Med Hypotheses 1994;42.

19. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Sci. 2001;291(5507): 1304-1351.

20. West GB, Brown JH, Enquist BJ. The fourth dimension of life: fractal geometry and allometric scaling of organisms. Sci. 1999;284(5420):1677-1679.