CROATICA CHEMICA ACTA CCACAA **72** (4) 999–1008 (1999)

ISSN-0011-1643
*CCA-2632* *Original Scientific Paper*

# On Circular Coding Properties of Gene and Protein Sequences

*Nikola Štambuk**

*Rudjer Bošković Institute, Bijenička 54, HR-10001 Zagreb, Croatia*

Received June 11, 1999; revised July 16, 1999; accepted July 27, 1999

The algorithms and equations that define and link circular coding properties of the genetic code and amino acid structural properties are derived according to the model of Cantor dynamics based automaton. It is shown that the model defines a unifying concept of the genetic code, which incorporates Crick's code without comma and the evolutionary code concept. Arithmetic for codes is defined via the number theory results in coding theory, Smale's horseshoe map and the dynamics on fractal lattices. The method has been denoted SCA (Symbolic Cantor Algorithm) and defined with respect to the principles of Molecular Recognition Theory, Grafstein's hypothesis of the stereochemical origin of the genetic code and Siemion's mutation ring. Underlying Fibonacci dynamics is extracted and mathematically defined considering the Cantor set and Farey tree codon and amino acid projections. Two digit specification of the codon positions, by means of the binary group subdivision, is particularly analysed with respect to octal coding and defined according to the purine-pyrimidine, amino-keto and strong-week H bonding discrimination principles.

*Key words:* Circular codes, gene, protein, DNA, RNA, Cantor set, necklace, horseshoe map, model

## INTRODUCTION

Recent investigations of gene and protein coding confirmed the existance of circular patterns in different sequences of prokaryotes and eukaryotes.[1,2] In this investigation, we extend the previously described method of

* E-mail: stambuk@rudjer.irb.hr

gene and protein analysis, based on binary coding.[3] The method is based on the solution of the four-colour necklace problem. The results of the codon recombination are linked to the binary tree locations on the Cantor set and octal coding.

## METHODS AND MODEL

### *1. Circular Code Arithmetic and Necklace Coding*

The genetic and protein circular code is defined by means of a combinatorial necklace model.[4] This structure consists of 64 beads of 4 different colours representing 4 nucleotide bases (U or T, C, A, G). The coulored beads form decorations consisting of vertically hanging chains of $x = 3$ beads, which represent each of the codons. Consequently, there are $y = 4^3$ distinct vertical chains that can be formed (*i.e.* number of words of length $x = 3$ with alphabet of size $y = 4$). The total number of possible vertical decorations containing at least two colours each is $y^x - y = 60$, and $y = 4$ decorations contain beads of the same colour.

One of the characteristics of this system is that we may define »beheading« as the process where the top bead is taken and replaced on the bottom.[4] After some repetitions, we observe the initial pattern. Let $b$ be the smallest positive number of successive beheadings (including the reverse ones) needed to get back the original; we have:

$$1 < b \leq x,\ x = ab + c\ (0 \leq c < b). \tag{1}$$

The initial pattern is restored by $x$ behadings followed by $a$ lots of $b$ reverse beheadings. For $c = 0$, $x = ab$, if $x$ is prime and $b > 1$, we have $b = x$, $a = 1$. By observing the chains and their first $x$–1 beheadings, different collections are formed (that cannot be transformed into each other). Thus, when $y^x$–$y$ chains have been accounted, we get a total of $n$ collections $y^x$–$y = nx$ and $y^x = y \ mod\ x$, from which we obtain Fermat's theorem.[4]

Table I (a-f) presents the circular coding patterns for all possible codon collections. It contains two and three colouring collections consisting of 3 transformed/beheaded codons (12 and 8 collections of 3 triplets, *i.e.* 60, of 2 same and 3 different colours, respectively) and four triplets of the same colour. It is shown that there is a codon arrangement for each of the 3 horizontal necklace frames (mod 1, 2, 3) that is 100% identical to the empirically detected one by Arques and Michel[1,2] (Table I and Table II). Four triplets of the same colour link the endpoints of the frames enabling the construction of the three frame automaton (Table II).

TABLE I

Circular coding patterns of the 4 colour necklace algorithm. The necklace consists of three horizontal frames with vertically hanging triplets (made up of 1, 2 or 3 colours). It is shown that permutations of triplets in horizontal and in vertical direction from 8 collections of 3 colours, 12 collections of 2 colours, and 4 triplets of 1 colour. The exact location of each triplet in frames given in Table II is indicated. For the sake of simplicity, in all tables U denotes U or T situation, *i.e.* the model is valid for both RNA and DNA coding.

a)

| *aa* | *Codon* | *Frame* | *aa* | *Codon* | *Frame* |
|---|---|---|---|---|---|
| G | GGU | 1 | P | CCA | 2 |
| V | GUG | 2 | H | CAC | 3 |
| W | UGG | 3 | T | ACC | 1 |
| | | complements ⇔ | | | |
| N | AAC | 1 | L | UUG | 2 |
| T | ACA | 2 | C | UGU | 3 |
| Q | CAA | 3 | V | GUU | 1 |

b)

| *aa* | *Codon* | *Frame* | *aa* | *Codon* | *Frame* |
|---|---|---|---|---|---|
| G | GGC* | 1 | P | CCG* | 2 |
| A | GCG | 2 | R | CGC | 3 |
| R | CGG* | 3 | A | GCC* | 1 |
| | | complements ⇔ | | | |
| N | AAU* | 1 | L | UUA* | 2 |
| I | AUA | 2 | Y | UAU | 3 |
| ochre | UAA* | 3 | I | AUU* | 1 |

c)

| *aa* | *Codon* | *Frame* | *aa* | *Codon* | *Frame* |
|---|---|---|---|---|---|
| F | UUC* | 1 | K | AAG* | 2 |
| S | UCU | 2 | R | AGA | 3 |
| L | CUU*[§] | 3 | E | GAA*[§] | 1 |
| | | complements ⇔ | | | |
| P | CCU* | 3 | G | GGA* | 3 |
| L | CUC[§] | 1 | E | GAG[§] | 1 |
| S | UCC* | 2 | R | AGG* | 2 |

TABLE I (continued)

d)

| *aa* | *Codon* | *Frame* | *aa* | *Codon* | *Frame* |
|---|---|---|---|---|---|
| Y | UAC | 1 | M | AUG | 2 |
| T | ACU* | 2 | opal | UGA* | 3 |
| L | CUA* | 3 | D | GAU* | 1 |
| | | complements ⇔ | | | |
| R | CGU* | 3 | A | GCA* | 3 |
| V | GUC* | 1 | Q | CAG* | 1 |
| S | UCG | 2 | S | AGC | 2 |

e)

| *aa* | *Codon* | *Frame* | *aa* | *Codon* | *Frame* |
|---|---|---|---|---|---|
| I | AUC[§] | 1 | amber | UAG[§] | 2 |
| S | UCA* | 2 | S | AGU* | 3 |
| H | CAU* | 3 | V | GUA* | 1 |
| | | complements ⇔ | | | |
| A | GCU | 1 | R | CGA | 3 |
| L | CUG | 2 | D | GAC | 1 |
| C | UGC* | 3 | T | ACG* | 2 |

f)

| *aa* | *Codon* | *Frame* | *aa* | *Codon* | *Frame* |
|---|---|---|---|---|---|
| F | UUU | 1 | K | AAA | 1 |
| P | CCC | 2 | G | GGG | 3 |
| | | complements ⇔ | | | |

*amino acid – amino acid dimer assigned codon inversion (*exo-endo*, *exo-endo*)[3,18]
[§]pairs with stop codons and *exo-exo*, *endo-endo* inversion[3,18]

Table II shows that the arrangement of the codons in the frames, according to their projection on the Cantor set,[3] transforms each frame in such a way that when one letter shift is performed the next frame is automatically retrieved (a-d). Few letter changes that occur during the transformations are permissible and predicted according to a Molecular Recognition Theory[3] (R⇔S, Q⇔H, D⇔E) or N-end rule (K⇔N), *i.e.* the coding pattern is consistent with theoretical and empirical observations.[3] Four letter-colour notation of the necklace codons (U or T, C, A, G) may be also expressed by two

TABLE II

Necklace model of the genetic code. In horizontal direction, we observe circular coding patterns of 3 necklace frames that make up Crick's comma-less code while vertical direction defines 64 hanging codons of the evolutional code (a-d). The model is also valid for three letter permutations (Y = pyrimidine, R = purine, N = Y or R; row e). *Wst* = opal or W(e), *Yst* = amber, ochre or Y(e).

a)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $m_1 \rightarrow$ | *A* | *F* | *Y* | *L* | *L* | *Q* | *V* | *V* | *V* | *A* | *G* | *G* | *D* | *D* | *E* | *E* | *I* | *I* | *T* | *N* | *N* | $\rightarrow m_2$ |
| | GCU | UUC | UAC | CUC | CUG | CAG | GUU | GUC | GUA | GCC | GGU | GGC | GAU | GAC | GAG | GAA | AUU | AUC | ACC | AAU | AAC | A |
| $m_2 \rightarrow$ | *T* | *S* | *T* | *S* | *C* | *S* | *L* | *S* | *Yst* | *P* | *V* | *A* | *M* | *T* | *R* | *K* | *L* | *S* | *P* | *I* | *T* | $\rightarrow m_3$ |
| | CU | UCU | ACU | UCC | UGC | AGC | UUG | UCG | UAG | CCG | GUG | GCG | AUG | ACG | AGG | AAG | UUA | UCA | CCA | AUA | ACA | AG |
| $m_3 \rightarrow$ | *S* | *L* | *L* | *P* | *A* | *A* | *C* | *R* | *S* | *R* | *W* | *R* | *Wst* | *R* | *G* | *R* | *Y* | *H* | *H* | *Yst* | *Q* | $S \rightarrow m_1$ |
| | U | CUU | CUA | CCU | GCU | GCA | UGU | CGU | AGU | CGC | UGG | CGG | UGA | CGA | GGA | AGA | UAU | CAU | CAC | UAA | CAA | AGC |

b)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $m_1 \rightarrow$ | *L* | *S* | *T** | *S* | *C* | *R** | *L* | *S* | *Yst* | *P* | *V* | *A* | *M* | *T* | *R* | *K** | *L* | *S* | *P* | *I* | *T* | $\rightarrow m_2$ |
| $\Leftarrow 1$ | CUU | UCU | ACC | UCC | UGC | AGG | UUG | UCG | UAG | CCG | GUG | GCG | AUG | ACG | AGG | AAA | UUA | UCA | CCA | AUA | ACA | G |
| $m_2 \rightarrow$ | $A^{\dagger}$,*T* | *S* | *T* | *S* | *C* | *S* | *L* | *S* | *Yst* | *P* | *V* | *A* | *M* | *T* | *R* | *K* | *L* | *S* | *P* | *I* | *T* | $\rightarrow m_3$ |
| | CU | UCU | ACU | UCC | UGC | AGC | UUG | UCG | UAG | CCG | GUG | GCG | AUG | ACG | AGG | AAG | UUA | UCA | CCA | AUA | ACA | AG |

c)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $m_2 \rightarrow$ | *V*,$I^{\dagger}$ | *L** | *L** | *P* | *A** | *A** | *C* | *R* | *S** | *R* | *W* | *R** | *Wst* | *R* | *G* | *S** | *Y* | *H** | *Q** | *Yst* | *H** | $\rightarrow m_3$ |
| $\Leftarrow 1$ | UU | CUA | CUU | CCU | GCA | GCU | UGU | CGU | AGC | CGG | UGG | CGA | UGA | CGA | GGA | AGU | UAU | CAC | CAA | UAA | CAA | GC |
| $m_3 \rightarrow$ | *A*,$S^{\dagger}$ | *L* | *L* | *P* | *A* | *A* | *C* | *R* | *S* | *R* | *W* | *R* | *Wst* | *R* | *G* | *R* | *Y* | *H* | *H* | *Yst* | *Q* | $S \rightarrow m_1$ |
| | U | CUU | CUA | CCU | GCU | GCA | UGU | CGU | AGU | CGC | UGG | CGG | UGA | CGA | GGA | AGA | UAU | CAU | CAC | UAA | CAA | AGC |

d)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $m_3 \rightarrow$ | *A*,$S^{\dagger}$ | *F* | *Y* | *L** | *L* | *H** | *V** | *V** | *V** | *A** | *G** | *G** | *D** | *E** | *E** | *D** | *I** | *I* | *T** | *N** | *K** | $A \rightarrow m_1$ |
| $\Leftarrow 1$ | C | UUC | UAC | CUG | CUG | CAU | GUC | GUA | GUC | GCU | GGC | GGU | GAC | GAG | GAA | GAU | AUC | AUC | ACU | AAC | AAA | GCU |
| $m_1 \rightarrow$ | *A* | *F* | *Y* | *L* | *L* | *Q* | *V* | *V* | *V* | *A* | *G* | *G* | *D* | *D* | *E* | *E* | *I* | *I* | *T* | *N* | *N* | $\rightarrow m_2$ |
| | GCU | UUC | UAC | CUC | CUG | CAG | GUU | GUC | GUA | GCC | GGU | GGC | GAU | GAC | GAG | GAA | AUU | AUC | ACC | AAU | AAC | A |

TABLE II (continued)

e)

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $m_1 \rightarrow$ | *A* | *F* | *Y* | *L* | *L* | *Q* | *V* | *V* | *V* | *A* | *G* | *G* | *D* | *D* | *E* | *E* | *I* | *I* | *T* | *N* | *N* | $\rightarrow m_2$ |
| | NYN | YYY | NRY | NYY | NNR | YNR | RNY | RNY | RNN | NYN | RRN | RRN | NNY | NNY | NRR | NRR | RYN | RYN | RYY | RRY | RRY | R |
| $m_2 \rightarrow$ | | *P* | *I* | *P* | *Y* | *R* | *M* | *M* | *S* | *Q* | *E* | *E* | *A* | *A* | *R* | *K* | *Q* | *Q* | *L* | *T* | *T* | $\rightarrow m_3$ |
| $\Leftarrow 1$ | NY | YYN | RYN | YYN | NRY | NRR | NYR | NYR | NNN | YNR | RNR | RNN | NYN | NYN | RRN | RRR | YNR | YNR | YYR | RYR | RYR | NY |
| $m_3 \rightarrow$ | | *Q* | *S* | *H* | *I* | *G* | *Yst* | *Yst* | *D* | *R* | *R* | *R* | *H* | *Q* | *G* | *N* | *Y* | *Y* | *Wst* | *Wst* | *Yst* | $S \rightarrow m_1$ |
| $\Leftarrow 1$ | Y | YNR | YNY | YNN | RYN | RRN | YRN | YRN | NNY | NRR | NRR | NNN | YNN | YNR | RNR | RRY | NRY | NRY | YRR | YRR | YRN | YNY |

* letter changes that occur during the transformations

TABLE III

Binary coding of the genetic code and protein structure[3] may be transformed into octal coding system based on 3-dimensional cube permutations. The octal coding system has several advantages, *e.g.* database compression or simple two-dimensional map projection.[8]

| Codon | *aa* | code | Codon | *aa* | code | Codon | *aa* | code | Codon | *aa* | code | Codon | *aa* | code | Codon | *aa* | code | Codon | *aa* | code | Codon | *aa* | code |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *UUU* | *F* | *00* | *UGC* | *C* | *11* | *CUG* | *L* | *22* | *CGA* | *R* | *33* | *GCU* | *A* | *44* | *GAC* | *D* | *55* | *ACG* | *T* | *66* | *AAA* | *K* | *77* |
| *UUU* | *F* | *00* | *UGC* | *C* | *11* | *CUG* | *L* | *22* | *CGA* | *R* | *33* | *GCU* | *A* | *44* | *GAC* | *D* | *55* | *ACG* | *T* | *66* | *AAA* | *K* | *77* |
| UUC | F | 01 | UGG | W | 12 | CUA | L | 23 | CAU | H | 34 | GCC | A | 45 | GAG | E* | 56 | ACA | T* | 67 | | | |
| UGU | C* | 10 | CUC | L* | 21 | CGG | R | 32 | GUA | V | 43 | GAU | D | 54 | ACC | T | 65 | AAG | K | 76 | | | |
| UUG | L | 02 | UGA | opal | 13 | CCU | P | 24 | CAC | H* | 35 | GCG | A* | 46 | GAA | E | 57 | | | | | | |
| CUU | L | 20 | CGC | R* | 31 | GUG | V* | 42 | GGA | G | 53 | ACU | T | 64 | AAC | N | 75 | | | | | | |
| UUA | L | 03 | UAU | Y* | 14 | CCC | P* | 25 | CAG | Q | 36 | GCA | A | 47 | | | | | | | | | |
| CGU | R | 30 | GUC | V | 41 | GGG | G* | 52 | AUA | I* | 63 | AAU | N | 74 | | | | | | | | | |
| UCU | S* | 04 | UAC | Y | 15 | CCG | P | 26 | CAA | Q | 37 | | | | | | | | | | | | |
| GUU | V | 40 | GGC | G | 51 | AUG | M | 62 | AGA | R* | 73 | | | | | | | | | | | | |
| UCC | S | 05 | UAG | amber | 16 | CCA | P | 27 | | | | | | | | | | | | | | | |
| GGU | G | 50 | AUC | I | 61 | AGG | R | 72 | | | | | | | | | | | | | | | |
| UCG | S | 06 | UAA | ochre | 17 | | | | | | | | | | | | | | | | | | |
| AUU | I | 60 | AGC | S | 71 | | | | | | | | | | | | | | | | | | |
| UCA | S | 07 | | | | | | | | | | | | | | | | | | | | | |
| AGU | S | 70 | | | | | | | | | | | | | | | | | | | | | |

*Terminating Cantor set decimals[3]; aa = amino acids

symbols (*e.g.* circle and triangle) of different weights (*e.g.* black and white). In this context, the symbols may define structural codon discrimination to pyrimidine and purine groups, while the colours define the base weak and strong H bonding or amino-keto groups, respectively. The latter links the four letter nucleotide notation to the two letter one.

### *2. Three Letter Alphabet and Octal Coding*

The same coding situation valid for the 4 letter alphabet is also valid for the 3 letter alphabet (R = purine, Y = pyrimidine, N = R or Y). The transcription of particular codons in the 3 letter alphabet is performed according to the procedure of Štambuk,[3] which satisfies the principles of the Molecular Recognition Theory, Stereochemical Theory of the Genetic Code and Siemion's mutation ring. Results in Table II (e) prove that the applied 3 letter notation satisfies the code arithmetic since frames 2 and 3 can generate 20 amino acids from frame 1 (of Table II (a)) transcribed into 3 letter notation.

The metric of the Sierpinski gasket[5] defines the sequences of the set [0, 1] by choosing the $c = 3$ letters out of 4. The metric $d_\alpha = 2^{-3} = 1/8$ and measure $M_\alpha = 3^{-3} = 1/27$ with $M_\alpha = d_\alpha{}^s$, $s = 1.58$ ($8^s = 27$, $2^s = 3$) define the number of contained sets as:[5] $m \leq 36\,\pi / \sqrt{3}$, *i.e.* $65.3 \cong 64$, which by analogy explains the fact that intramolecular protein forces follow Sierpinski's gasket model.[6] On the other hand, the two dimensional $x$-$y$ coordinate interaction[7] of two Cantor sets by means of square coordinate geometry (00), (01), (10) and (11) defines also the $t = 3$ triplet (sequence) selection out of $N = 4$ possible letters as $N = t^s$, $4 = 3^s$, $64 = 27^s$, $s = 1.26$.

Further extention of the coding system in Table III defines 28 pairs of all possible 8 (node) 3-dimensional cube permutations (*i.e.* as $8 \times 8$ codon octades).[3] Three binary digits define the octal number coding, and $28 = 8!/2!\,(8–2)!$ pairing combinations are obtained from the permutations of the 2 corresponding octal numbers[8] (or 2 binary triplets). Eight identical doublets, in addition to 56 different dublets define all 64 codons. This pattern is consistent, with the three letter alphabet permutation consisting of two binary, *i.e.* 0 or 1, choices in the truth table ($2^3 = 8$).[9]

If the coding process is described via graph theory by partitioning sequence $N$ into $c = 3$ colour classes and assagning colour to each integer, Schur's lemma[10] gives the solution for the 3 colour problem *via* Schur's number $S(3) = 14$, which is consistent with 14 amino acid pairs (codon groups) of the genetic code.[3] Not surprisingly, Schur's result was obtained in an attempt to prove Fermat's last theorem.[10] Van der Waerden's theorem[10] extends Schur's lemma and the number $W(c,t) = 27$ for $c = t = 3$ ($t$ being the length of the sequence and $c$ representing 3 letter classes). This number corresponds to 13 amino acid pairs of the codon groups ($13 \times 2 = 26$) and one self-similar pair for serine and its complements (1).[3] The pairs define 3-letter permutations with respect to triplets ($3^3 = 27$).[3]

## DISCUSSION

Presented results indicate that the concepts of code without comma or of evolutionary code, based on different premises, strongly depend on the level of the observation/analysis. In the necklace model, Crick's comma-less code represents three horizontal frames that define necklace chains, while the evolutionary code makes vertically hanging beads (codon triplets). Therefore the circular coding necklace algorithm represents a unifying concept of the genetic code. This method, denoted SCA (Symbolic Cantor Algorithm), enables the genetic code and protein analysis *via* number theory arithmetic for codes. Interestingly, the symbolic dynamic binary notation on the Cantor set defined by Štambuk[3] represents the Gray code solution to the $n = 6$ ($2^n = 64$) ring (digit) puzzle as described by Gardner.[11]

Two dimensional Cantor set projection of the binary (squere) notation *via* Smale's horseshoe map[12] reconstructs the classic table of the genetic code from the symbolic dynamic addresses[3] and octal coding elements of Table III, which proves our result and opens the possibility of the gene and protein analyses as chaotic dynamical systems. Additionally, the closest intersections of the Cantor set (binary & symbolic) codon projections and Farey tree codon projections define »golden amino acids«, related to the Fibonacci dynamics.[13] The Fibonacci dynamics noticed in the binary tree algorithms of the genetic code[3,13,14] and in the long range DNA correlation exponents[15] arises from the two frequencies of the Cantor and Farey tree dynamics. The frequences of the Cantor set projection/recombination of the codons of 2/3 are mixed with the Farey tree frequency of 1/2 that splits the amino acid/codon groups upon each Cantor set projection level ($2^6 = 64$). The resulting frequency is the golden ratio 0.618 = 2+1/3+2 = 3/5, which explains the previously mentioned phenomena. Some mathematical and dynamical aspects of these interactions have been discussed by Schroeder and Štambuk.[16,17]

The fact that the 0⇔1 and 3⇔5 changes of the complementary strand represent symmetric digit replacements in the binary notation with respect to the initial (stationary) strand[3] indicates that the complementary strand may represent the coded D-amino acid isomer of the initial strand, which is not transcribed due to the 3⇔5 complementary signal inversion. In that sense, the explanation of the evolutional dissappearance of D-amino acid isomers may be due to the fact that D-isomers remained coded (but usually not transcribed/expressed) in the complementary DNA strand. Rare examples of D-amino acid systems that confirm the theory of stereochemical origin of the genetic code have been reviewed by Grafstein.[18–20] Complementary strand beside the initial D-amino acid sequence defines also the L-amino acid complement,[3,18] which is usually synthesised if this strand is tran-

scribed (probably due to the much more available *t*RNA for complementary L-amino acids then for initial D-amino acids). Considering this, the complementary DNA strand represents a duplicated D-isomer of its initial (stationary) strand or its RNA analogue. It remains an open question if the origin of the double DNA helix may be searched in the evolutional fusion of single and identical L-RNA and D-RNA helices, according to Grafstein's remarks.[18–21]

The evolutional disapearance of D-amino acid RNA (*e.g.* of D-amino acid isomers linked to the corresponding *t*RNAs) might have led to two facts: 1. possible transcription of L-complementary instead of D-identical (stationary) proteins, 2. non-transcribed D-identical (stationary) strand became an evolutionally more preserving isomer while the transcribing L-strand became the one more susceptable to mutation. It remains an open question if the discussed phenomena related to the stereochemical hypothesis may be responsible for different evolutional mutation impacts of coding/leading and non-coding/lagging strands as discussed by Radman,[22,23] especially since recent investigations linked the *t*RNA behaviour to the presented algorithms of the genetic code.[3]

The permutations of the binary coding quadratic algorithm defined by Štambuk[3] for the nucleotide notation are in agreement with the universal rule of TG/CT excess and TA/CG deficiency in coding and noncoding DNA regions.[24,25] Another important aspect of this study is related to the discovery that non-coding DNA sequences possess properties characteristic of natural languages, while the coded DNA sequences correspond to the coded language structures.[26,27] In this context, the concept presented in this study may contribute to the extraction/decoding of the programming language of DNA and RNA strings.

## REFERENCES

1. D. G. Arques and C. J. Michel, *J. Theor. Biol*. **182** (1996) 45–58.
2. D. G. Arques, J. P. Fallot, and C. J. Michel, *J. Theor. Biol*. **185** (1997) 241–253.
3. N. Štambuk, *Croat. Chem. Acta* **71** (1998) 573–589.
4. J. Baylis, *Error Correcting Codes*, Chapman & Hall, London, 1998, pp. 38–42.
5. G. A. Edgar, *Measure, Topology and Fractal Geometry*, Springer-Verlag, New York, 1990, pp. 157–160.
6. M. Kurzynski, K. Palacz, and P. Chelminiak, *Proc. Natl. Acad. Sci. USA* **95** (1998) 11685–11690.
7. K. J. Falconer, *The Geometry of Fractal Sets*, Cambridge University Press, Cambridge, 1990, pp. 14–17.
8. A. R. Plantz and M. Berman, *IEEE Trans. Computers*, May (1971) 593–597.
9. M. Gardner, *Mathematical Magic Show*, Penguin books, Harmondsworth, pp. 97–99.

10. M. J. Erickson, *Introduction to Combinatorics*, J. Wiley & Sons, New York, 1996, pp. 63–68.
11. M. Gardner, *Sci. American,* August (1972) 106–109.
12. N. B. Tufillaro, T. Abbott, and J. Reilly, *An Experimental Approach to Nonlinear Dynamics and Chaos*, Addison-Wesley, Redwood City, 1992, pp. 218–230.
13. M. M. Rakočević, *BioSystems* **46** (1998) 283–291.
14. I. Siemion, *Amino Acids* **8** (1995) 1–13.
15. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Nature* **356** (1994) 168–170.
16. M. Schroeder, *Fractals Chaos, Power Laws*, W. H. Freeman, New York, 1991, pp. 334–340.
17. N. Štambuk, *Mathl. Comput. Modelling* **14** (1990) 565–570.
18. D. Grafstein, *J. Theor. Biol.* **105** (1983) 157–174.
19. D. Grafstein, *J. Theor. Biol.* **114** (1985) 11–20.
20. D. Grafstein, *J. Theor. Biol.* **115** (1985) 415–427.
21. D. Grafstein, *Origins of Life* **14** (1984) 589–596.
22. M. Radman, *Proc. Natl. Acad. Sci. USA* **95** (1998) 9718–9719.
23. I. J. Fijalkowska, P. Jonczyk, M. Maliszewska Tkaczyk, M. Bialoskorska, and R. M. Schaaper, *Proc. Natl. Acad. Sci. USA* **95** (1998) 10020–10025.
24. S. Ohno, *Proc. Natl. Acad. Sci. USA* **85** (1988) 9630–9634.
25. T. Yomo and S. Ohno, *Proc. Natl. Acad. Sci. USA* **86** (1989) 8452–8456.
26. R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, *Physical Review E* **52** (1995) 2939–2950.
27. A. Czirok, R. N. Mantegna, S. Havlin, and H. E. Stanley, *Physical Review E* **52** (1995) 446–452.

## SAŽETAK

### O cirkularnim kodovima gena i proteina

*Nikola Štambuk*

Istražen je model genetičkog koda zasnovan na dinamici Cantor-ova skupa. Definirane su jednadžbe i algoritmi koji opisuju spomenuti cirkularni kod i povezuju ga sa strukturnim osobinama aminokiselina i proteina. Pokazano je da model definira zajednički koncept genetičkog koda koji uključuje Crickov kod bez zareza i evolucijski model genetičkog koda. Istaknuto je da se genske strukture mogu analizirati matematičkim aparatom teorije kodiranja, Smale-ovom potkovastom mapom te fraktalnom analizom. Metoda je nazvana SCA (Simbolički Cantorov Algoritam), te je definirana s obzirom na teoriju molekulskog prepoznavanja, Grafsteinovu hipotezu o stereokemijskom podrijetlu genetičkog koda i Siemionov mutacijski prsten. Objašnjena je i Fibonacci-jeva dinamika rekombinacije kodona na Farey-evu drvu. Dani su uvjeti pod kojima se u binarnoj i oktalnoj notaciji odvija razlučivanje parova na osnovi purinskih i pirimidinskih baza te njihovih jakih i slabih vodikovih veza.