CROATICA CHEMICA ACTA CCACAA **71** (3) 573–589 (1998)

ISSN-0011-1643
CCA-2513 *Original Scientific Paper*

On the Genetic Origin of Complementary Protein Coding

Nikola Štambuk

Rugjer Bošković Institute, Bijenička 54, HR-10001 Zagreb, Croatia
(e-mail: stambuk@rudjer.irb.hr)

Received February 20, 1998; revised May 29, 1998; accepted June 30, 1998

The relations of protein coding and hydropathy are investigated considering the principles of the molecular recognition theory and Grafstein's hypothesis of the stereochemical origin of the genetic code. It is shown that the coding of RNA and DNA requires 14 distinct groups of codon-anticodon pairs, which define all possible complementary amino acids. The molecular recognition theory is redefined considering the codon-anticodon relations of *m*RNAs, DNAs, *t*RNAs and Siemion's mutation ring of the genetic code. A model of DNA, RNA and protein coding (and decoding) based on two fundamental properties of DNA/RNA, denoted as *complementary* and *stationary principles,* is presented. *Stationary* DNA/RNA coding defines the nucleotide relationship of the same (self) DNA/RNA strand and *complementary* coding defines nucleotide distribution related to other (non-self) strand. Combinations of 2 digits, denoting *primary* and *secondary characteristics* of each nucleotide, specify codon positions according to the group subdivision (discrimination) principle. The process of coding is related to the hypercube node codon representations and dynamics of their binary tree locations. The relations between binary tree locations and Cantor set representations of different codon points are discussed in the context of quadratic mappings, Feigenbaum dynamics and signal analysis. Combinations of hypercube nodes and different binary tree positions define the words, sentences and syntax of DNA, RNA and protein language. Possible applications of this method may be related to network analysis and the design, gene, protein and drug modelling.

INTRODUCTION

Recent experimental results indicate that peptides specified by the complementary RNAs and DNAs bind to each other with high efficacy and specificity.[1–4] The theoretical background explaining these experimental findings has been named Molecular Recognition Theory (MRT).[1–4] Based on the analysis of more than 38 complementary peptide-receptor systems, MRT states that the codons for hydrophilic and hydrophobic amino acids are complemented by the codons for the reversely charged ones.[1–4] Neutral amino acids are complemented by the similarly charged ones.[1–4] The application of this concept has provided valuable tools for the analysis of new biologically active peptides, defining of protein folding and modelling of gene segments.[1–12] However, due to the degeneracy of the genetic code, each amino acid is usually coded by several codons, which leads to a large number of possible complementary peptides, even for relatively short peptide motifs.[1–6]

In this paper, an alternative approach to the protein and gene modelling is defined. The method is based on the extraction of the most probable amino acid and nucleotide pairs, instead of focusing on a large number of possible ones interacting with different affinity.[5,6] A model of DNA, RNA and protein coding related to the molecular recognition theory will be also discussed. This model is based on two fundamental properties of DNA/RNA, denoted as *complementary* and *stationary principles*. Stationary DNA/RNA coding defines nucleotide relationship of the same (self) DNA/RNA strand, while complementary coding defines nucleotide distribution related to other (non-self) strand.

METHODS AND MODEL

1. Molecular Recognition Theory and the Related Amino Acid Coding

Changes in the hydropathic scores of different complementary amino acid pairs were analyzed by comparing 34 different motifs that had been derived and experimentally verified by MRT (Figure 1). Seventeen out of 34 peptide-receptor complexes were receptor sequences and 17 represented complementary messages (ligands) to the receptors. Analyzed motifs belonged to 7 peptide-receptor systems: epidermal growth factor, interleukin 2, transferrin, von Willebrand factor, angiotensin II/III, vitronectin and prolactin.[8–13] Matching transcripts for different codon pairs and the related amino acids were obtained by the method of Blalock *et al.*[1–6]

The probability of appearance for each amino acid complementary pair (P) within peptide motifs was defined as:

$$P = n/N \tag{1}$$

with n being a number of detected pairs of the same type and N being the total number of all matching pairs.

The groups of the most probably matching amino acid pairs and their complementary transcripts were compared to the complementary transcripts of the consensus *t*RNA genes reported by Rodin *et al.*[14] Transcripts were obtained according to Blalock *et al.*[1–4] (Figure 1). The method was tested by evaluating the lymphocyte proliferative response to met-enkephalin (peptid-M, LUPEX®, Biofactor, Germany) and its complementary peptide derived by means of Figure 1 and Table I (peptide-D; IPPKY). Peptide-D was synthesized using the standard solid-phase method and analyzed by HPLC and amino acid analysis (Biofactor, Germany; Lot No. B-0158). Synthesized peptide-D had a molecular mass of 616.8 D and > 97% purity. Blocking of the peptid-M induced lymphocyte proliferation by means of different concentrations of its complementary peptide-D is presented in Figure 2.

Lymphocyte proliferation was performed by means of 5-day cell cultures (2×10^5 cells/250 µL well, 5% FCS, sodium citrate anticoagulant) and ^{3}H-Thymidine incorporation, as described by deSmet *et al.*[15]

Hydropathic indexes after Kyte and Doolittle were used for the calculation and comparison of hydrophilicity and hydrophobicity of the amino acid pairs.[1,2,16] Values of the Chou-Fasman parameters $P_\alpha + P_\beta$ and $P_\alpha - P_\beta$ were used for evaluation of the amino acid structural importance ($P_\alpha + P_\beta$) and helix forming potency ($P_\alpha - P_\beta$), with respect to the 3rd base impact on amino acid coding (Figure 3b).[17–19]

2. Model of Complementary and Stationary DNA and RNA Coding

The model is based on the fact that each nucleotide letter may be defined and positioned according to the 2-digit binary principle. Three letter codons out of the set of all 4 letter combinations ($64 = 4^3 = 2^6$) define particular codons by means of a 6 digit binary representation. The combination of 2 digits (0 or 1), denoting *primary* and *secondary characteristics* of the nucleotide, describes each of the letters according to the group subdivision/discrimination principles (purine-pyrimidine and strong-weak H bond discrimination). For the first digit discrimination, pyrimidines (Y) are denoted by 0 and purines (R) by 1. Pyrimidine pairs (C, U or T) and purine (A, G) preserve the first digit notation (Figure 4). In the second digit subdivision to the weak H bonding bases (A, U or T) and the strong H bonding ones (C, G), the second digit notation is complementary (Figure 4).

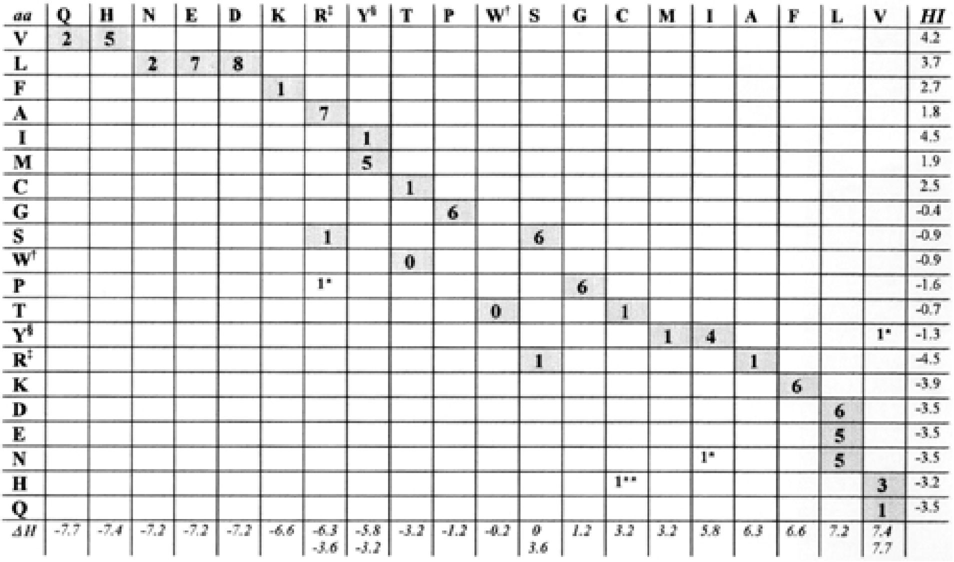
Applied alphabet is based on the binary or Boolean vectors (tuples), which assume values from the set $\{0, 1\}$. *Tuples* are defined by the length of the vector and the coding set represents an n-dimensional cube.[20] *Hamming distance* between the vertices of the cube defines their links[20] and the coding process is described by means of a $n = 6$ dimensional hypercube (C^n) consisting of 2^6 nodes, where each node i, $0 \le i \le 2^n - 1$, is represented by $n=6$ bit binary representation of i for each of the 64 codons.

To preserve the symmetry within *complementary* (matching non-self) and *stationary* (self) DNA/RNA coding strands, the following binary notation was applied: U or T = 00, C = 01, G = 10, A = 11 (Figure 4). This notation ensures that 0–1 digit replacements define complementary signal changes with respect to the stationary one by means of purine-pyrimidine (A, G – C, U or T) and strong-weak H bonding distinction (C, G – A, U or T). The codons are transcribed into binary notation according to the letter appearance (1st, 2nd, 3rd; Table IV).

RESULTS AND DISCUSSION

1. Evaluation of the Molecular Recognition Theory

The analysis of 34 complementary amino acid motifs belonging to 7 different peptide-receptor complexes is shown in Figure 1 and Table I.



aa	Q	H	N	E	D	K	R‡	Y§	T	P	W†	S	G	C	M	I	A	F	L	V	*HI*
V	2	5																			4.2
L			2	7	8																3.7
F						1															2.7
A							7														1.8
I								1													4.5
M								5													1.9
C									1												2.5
G										6											-0.4
S							1					6									-0.9
W†									0												-0.9
P							1*						6								-1.6
T											0			1							-0.7
Y§															1	4				1*	-1.3
R‡												1					1				-4.5
K																		6			-3.9
D																			6		-3.5
E																			5		-3.5
N																1*			5		-3.5
H														1**						3	-3.2
Q																				1	-3.5
ΔH	-7.7	-7.4	-7.2	-7.2	-7.2	-6.6	-6.3 -3.6	-5.8 -3.2	-3.2	-1.2	-0.2	0 3.6	1.2	3.2	3.2	5.8	6.3	6.6	7.2	7.4 7.7	

§ also amber or ochre (stop codons)
† also opal (stop codon)
‡ stop, codon used in human mitochondria (see footnotes to Table I)
ΔH = differences in hydropathic indexes (column aa/ligand – row aa/receptor scores)
HI = hydropathic indexes of the amino acids (aa) according to Kyte and Doolittle[16]
* possible but not probable (as for shaded pairs)
** failures, not possible by the molecular recognition theory (MRT)

Figure 1. Correlation of frequent complementary amino acid pairs in peptide ligand-receptor systems according to MRT ($r = 0.76$, $p < 0.05$). Shaded squares represent *m*RNA transcripts of 32 complementary consensus *t*RNA genes according to Rodin *et al.*[14] (Table II).

TABLE I

Fourteen complementary amino acid (aa) pairs define all codons and Grafstein's[27] stereochemical pairs.

Codon	L-aa (*exo*)	P[‡]	R-aa dimer (*exo*)	Codon	R-aa (*endo*)	L-aa dimer (*endo*)	P[‡]
GUU *RNY* GUC*	V	0.0217	Q	CAA *YNR* CAG	V	Q	0.0109
GUG *RNN* GUA*	V	0.0543	H	CAC *YNN* CAU*	V	H	0.0326
UUG *YYR* UUA*	L	0.0217	N	AAC *RRY* AAU*	L	N	0.0543
GAG *NRR* GAA*	E[§]	0.0761	L	CUC *NYY* CUU*	E[§]	L	0.0543
CUG *NNR* CUA*	L	0.0870	D	GAC *NNY* GAU*	L	D	0.0652
UUU *YYY* UUC*	F	0.0109	K	AAA *RRR* AAG*	F	K	0.0109
GCG *NYN* GCU GCC* GCA*	A	0.0761	R	CGC *NRN* CGA CGG* CGU*	A	R	0.0761
AUG***NYR*	M	0.0543	Y	UAC *NRY*	M	Y	0.0109
UAU *YRN* UAG UAA*	Y[§] amber ochre	0.0430	I	AUA *RYN* AUC AUU*	Y[§] ambe ochre	I	0.0109
UGU *YRY* UGC*	C	0.0109	T	ACA *RYR* ACG*	C	T	0.0109
GGG *RRN* GGU GGA* GGC*	G	0.0652	P	CCC *YYN* CCA CCU* CCG*	G	P	0.0652
UGG *YRR* UGA*	W[§] opal	0	T	ACC *RYY* ACU*	W[§] opal	T	0
UCG *NNN* UCA*	S	0.0652	S	AGC *NNN* AGU*	S	S	0.0652
UCU *YNY* UCC*	S	0.0109	R	AGA *RNR* AGG*	S[†]	R[†]	0.0109

* amino acid – amino acid dimer assigned codon inversion (*exo-endo*, *exo-endo*)[27]

** start pairs

§ pairs with stop codons and *exo-exo, endo-endo* inversion[27]

† stop codon in human mitochondria

‡ probabilities of appearance, for complementary amino acid pairs and codons (P)
Y = U (T) or C, R = A or G, N = Y or R; Michel's[28] nucleotide specification alphabet

Probabilities of appearance for different amino acid pairs ranged from 0.011 to 0.087 (Table I), which is consistent with literature data concerning different protein motifs.[21] Hydropatic scores of the complementary ligand and receptor amino acid pairs (and transcripts) of the analyzed motifs were significantly correlated ($r = 0.76$, $p < 0.01$; Figure 1).

Complementary ligand-receptor interaction presented in Figure 1 was evaluated by designing a peptide complementary to met-enkephalin (YGGFM). This peptide was named peptide-D (IPPKY) and it represents a complementary transcript of met-enkephalin, *i.e.* its possible receptor (Figure 1). By means of cellular proliferative bioassay based on ^{3}H-Thymidine DNA incorporation, it was proved that peptide-D blocks the met-enkephalin induced lymphocyte proliferation in a dose-dependent manner (Figure 2). This confirmed the theoretically predicted ligand-receptor interaction of met-enkephalin (peptid-M) and peptide-D. In the context of MRT, similar results have been obtained by several authors for different short peptide motifs and DNA/RNA transcripts.[1,8–13]

The results presented in Figure 2 also indicate that specific met-enkephalin receptors correspond to human calpastatin location (ICAL_HUMAN, residues 201–205) and share molecular homology to short sequences of rapamycin-selective 25 kD immunophilin FKBP 25, *i.e.* Rapamycin and FK506 binding protein (FKB3_HUMAN, sequence PPKY, residues 108–111).

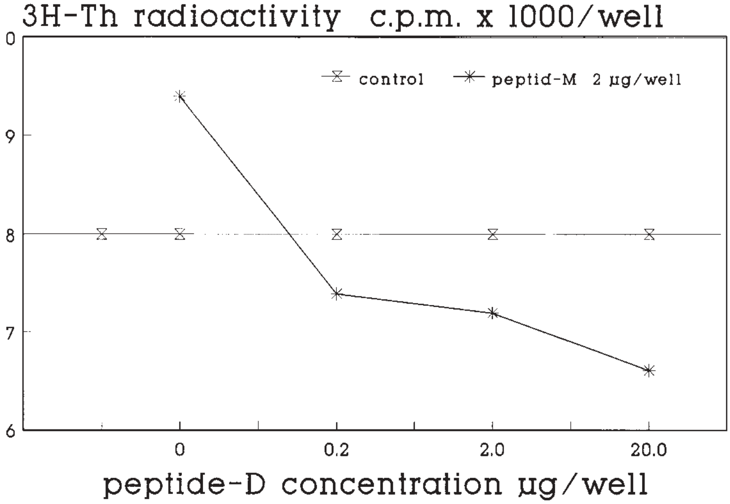


Figure 2. Lymphocyte proliferative response to met-enkephalin (peptid-M; YGGFM) and its complementary peptide defined by means of Figure 1 and Table I (peptide-D; IPPKY). Blockade of the peptid-M induced human lymphocyte proliferation, by means of different concentrations of its complementary peptide-D, is dose-dependent (c.p.m. = counts per minute, control = cell cultures without peptides).

TABLE II

Thirty two complementary codon pairs of the *t*RNA genes (Rodin *et al.*[14,26]) define complementary amino acid arrangements from Table I and confirm that *m*RNA/DNA coding may be evolutionally related to *t*RNA and circular algorithms of the genetic code.[5–7,18,19,29–35]

Class I × Class I	
Inner (left)	Outer (right)
st (*C*, UAA) – **L** (*G*, UUA) **I** (AUU) – **N** (AAU)	**st** (*C*, UAG) – **L** (*G*, CUA) **I** (AUC) – **D** (GAU)
Y (*G*, UAC) – **V** (*S*, GUA) **I/M**[a] (AUG) – **H** (CAU)	**Y** (*T*, UAC) – **V** (*S*, GUA) **M/start** (AUG) – **H** (CAU)
Q (*Y*, CAG) – **L** (*G*, CUG) **V** (GUC) – **D** (GAC)	**Q** (*C*, CAA) – **L** (*G*, UUG) **V** (GUU) – **N** (AAC)
E (*G*, GAG) – **L** (*C*, CUC) **L** (CUC) – **E** (GAG)	

Class II × Class II	
Inner (left)	Outer (right)
S (*C*, AGU) – **T** (*G*, ACU) **S** (UCA) – **st/W**[a] (UGA)	**S** (*G*, AGC) – **A** (*G*, GCU) **S** (UCG) – **R** (CGA)
G (*C*, GGU) – **T** (*C*, ACC) **P** (CCA) – **W** (UGG)	**G** (*G*, GGA) – **S** (*C*, UCC) **P** (CCU) – **R/st**[a] (AGG)
G (*G*, GGC) – **A** (*C*, GCC) **P**(CCG) – **R** (CGG)	
K (*C*, AAA) – **F** (*G*, UUU) **F** (UUU) – **K** (AAA)	
G (*G*, GGG) – **P** (*C*, CCC) **P** (CCC) – **G** (GGG)	

Class I × Class II	
Inner (left)	Outer (right)
R (*G*, CGC) – **A** (*T*, GCG) **A** (GCG) – **R** (CGC)	**R/st**[a] (*G*, AGA) – **S** (*C*, UCU) **S** (UCU) – **R/st**[a] (AGA)
R (*C*, CGU) – **T** (*G*, ACG) **A** (GCA) – **C** (UGC)	**C** (*G*, UGC) – **A** (*S*, GCA) **T** (ACG) – **R** (CGU)
R (*G*, CGC) – **P** (*C*, CCG) **A** (GCG) – **G** (GGC)	**R/st**[a] (*G*, AGG) – **P**(*C*, CCU) **S** (UCC) – **G** (GGA)
W (*G*, UGG) – **P** (*S*, CCA) **T** (ACC) – **G** (GGU)	**C** (*C*, UGU) – **T** (*G*, ACA) **T** (ACA) – **C** (UGU)
R (*G*, CGA) – **S** (*C*, UCG) **A** (GCU) – **S** (AGC)	
st/W[a] (*G*, UGA) – **S** (*S*, UCA) **T** (ACU) – **S** (AGU)	
E (*C*, GAA) – **F** (*C*, UUC) **L** (CUU) – **K** (AAG)	**K** (*G*, AAG) – **L** (*C*, CUU) **F** (UUC) – **E** (GAA)
D (*G*, GAU) – **I** (*C*, AUC) **L** (CUA) – **st** (UAG)	**D** (*G*, GAC) – **V** (*C*, GUC) **L** (CUG) – **Q** (CAG)
N (*G*, AAC) – **V** (*C*, GUU) **L** (UUG) – **Q** (CAA)	**N** (*G*, AAU) – **I** (*C*, AUU) **L** (UUA) – **st** (UAA)
H (*G*, CAC) – **V/M**[b] (*Y*, GUG) **V/M**[b] (GUG) – **H** (CAC)	**H** (*G*, CAU) – **V/M**[b] (*Y*, AUG) **V** (GUA) – **Y** (UAC)

[a]codons in mitochondria,[b]codes for V but can code for M to initiate translation from *m*RNA, st = stop codon; *S* = G or C, correlated complementarity of the 2nd base in anticodon and acceptor[14] is denoted by italics

The latter might explain the fact that immunosuppressive effects of met-enkephalin (peptid-M) on mitotic division and cell cycle[22] may be reversed by short peptides containing the PPK sequence.[23] The evaluation of Ca^{2+} mediated signalling, and links between met-enkephalin (peptid-M) and calpain-calpastatin system will be of importance for the drug design related to several immune-mediated, degenerative and genetic diseases.[24–25]

The appearance of amino acid pairs in 34 complementary motifs was additionally compared with amino acid pairs defined by *m*RNA transcripts of all 32 complementary consensus *t*RNA genes (shaded squares in Figure 1; gene transcripts are given in Table II). This analysis revealed the existence of 27 frequent amino acid codon groups, *i.e.* 2 × 13 amino acid groups of different letters and one single letter group (SS) (Table I, Figure 1). Only three detected pairs out of the 96 analyzed (RP, IN, VY; Figure 1) did not match the results of Rodin *et al.*,[14] *i. e.* related *t*RNA gene transcripts presented in Table II. According to MRT, they were possible, but not highly probable ($P < 0.05$). One pair (CH) was not possible after MRT. Total error of the procedure was < 5%, *i.e.* $P = 4/96$, and the correlation of the hydropathic scores for those extracted and most probable matching pairs ($r = 0.76$, $p < 0.01$) remained almost the same as for all pairs by Blalock *et al.*[1,2] This indicates that 14 distinct groups of complementary RNA/DNA coding pairs define all 64 codons, transcripts of *t*RNA genes and correlated complementarity of the second base in anticodon and acceptor (Table II). The latter confirms the assumption of Rodin *et al.*[14,26] that *m*RNA coding might be evolutionally related to *t*RNA, *i.e.* that *t*RNA genes and aminoacyl-*t*RNA syntethases could have originated from the complementary strands of primordial RNAs.[14,26] Table II and Figure 1 additionally confirm the present subdivision of aminoacyl-*t*RNA syntethases into Class I, Class II and Class I × II enzymes according to their relations to amino acids of different hydropathy and size.[14,26]

2. *Characteristics of Amino Acid and Codon Recombination*

Fourteen pairs of complementary amino acid codon groups define:

1. all 64 possible codons (Table I),
2. Grafstein's hypothesis[27] of the stereochemical origin of the genetic code for 20 left (L) and 20 right (R) amino acid isomers (Table I and Table III),
3. minimal number of 27 codon combinations ($3^3 = 27$) within a three-letter codon alphabet, *i.e.* the nucleotide triplet language, according to Michel[28] (R = purine, Y = pyrimidine, N = R or Y; Table I).

Pairs in Table I and Table II define 64 codons by forming 4 amino acid families (U, C, A, G). Amino acids are classified into 4 families by means of the 1st and 2nd bases, while the 3rd base discriminates when the first two are identical for at least two amino acids (Table III). This also holds for

TABLE III

Complementary amino acid and codon pairs are defined by means of the first and second nucleotide base while the third one discriminates if they are identical.

	L aa – dimer L aa (*exo-endo*)				R aa dimer – R aa (*exo-endo*)			
Base 1st/2nd	**U**	**C**	**A**	**G**	**U**	**C**	**A**	**G**
U	F**, L	S, S*	Y*, Y, st§	C*,W, st†	N, K**	S, R*	I*, M	T, T*
C	L, L*	P	H*, Q	R	D, E*	G	V*, V	A
A	I*, M	T, T*	N, K**	S, R*	Y*, Y, st§	C*, W, st†	F**, L	S, S*
G	V*, V	A	D, E*	G	H*, Q	R	L, L*	P

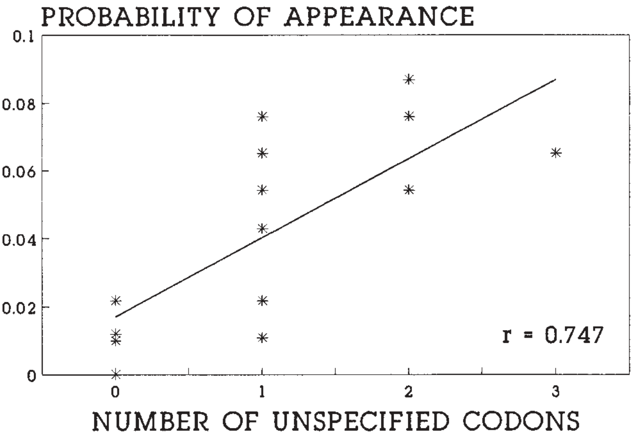
* 1st & 3rd bases identical, ** triplet; § amber-all three different, ochre-2nd & the 3rd same
† opal-all three different; st = stop codon

Grafstein's stereochemical pairs in Table III. As shown in Figure 3b, this 3rd base discrimination between different amino acid letters, of close hypercube nodes possessing identical first two bases, is linked to the purine-pyrimidine subdivision principle. The correlation of conformational and structurally important Chou-Fasman parameter $P_\alpha + P_\beta$ was significant ($r = 0.893$, $p < 0.01$) for all, *i.e.* 8 discriminating purine-pyrimidine pairs of the 3rd base in Table III (LF, MI, WC, stop(Q)Y, QH, ED, RS, KN; Figure 3b). Correlations of these 8 pairs for other Chou-Fasman parameters $P_\alpha - P_\beta$, P_α and P_β were not significant ($p > 0.05$). This is due to the fact that amino acids with purine (R) in the 3rd base position prefer P_α and with pyrimidine (Y) favour P_β parameter.[19]

Lack of correlation for parameter $P_\alpha - P_\beta$ may be due to it characterizing only the helix-forming potency of the amino acids and not the structural importance as, *e.g.*, $P_\alpha + P_\beta$.[19] Consequently, our data support the results of Siemion[18,19] regarding the influence of the third base in the coding process and amino acid discrimination.

Table IV shows that permutations of the 4 amino acid families in Table III (CUGA, UGAC, GACU and ACUG) identify Siemion's one-step mutation ring of the genetic code by means of UC/CU, AG/GA replacements (U row/column), C/G–G/C mutation (C row/column), A/U mutation (A row/column) and C/G mutation (G row/column). Permutations of 4 amino acid families are defined by the simple algorithm presented in Figure 4. Our results in Tables I–IV and Siemion's mutation ring[18] share a marked similarity to several circular algorithms of the genetic code analyzed by Rakočević,[7] Arques and Michel,[30,31] Siemion,[18] Štambuk,[5,6,29] Jiménez-Montano *et al.*,[32] Swanson[33] and Zhang.[34] It is worth mentioning that amino acid codon replacements in the U row/column and their changes (mutations) in C, A, G rows/columns of Table IV result in golden ratio symmetry within the codon

a) CODONS *vs.* PAIR APPEARANCE



b) CHOU-FASMAN PARAMETERS
THIRD BASE DISCRIMINATION

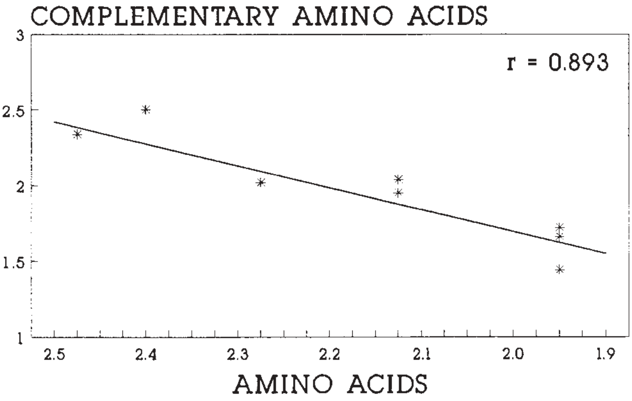


Figure 3. **a)** Correlation of the probability of appearance and the number of unspecified codons for amino acid complementary pairs presented in Table I ($r = 0.747$, $p < 0.01$). **b)** Correlation of the conformational Chou-Fasman parameter $P_{\alpha} + P_{\beta}$ for 8 complementary purine-pyrimidine pairs of the 3rd base (LF, MI, WC, stop(Q)Y, QH, ED, RS, KN). This structurally important parameter discriminates in cases when the first two codon bases are identical ($r = 0.893$, $p < 0.01$).

family groups of Siemion's ring (column+row/diagonal = 3/2, 5/3, 8/5; Table IV). This supports Siemion' s assumption that amino acid codon distribution within the mutation ring is related to golden ratio symmetry.[18]

TABLE IV

Permutations of Table III by the 4 nucleotide families of the first two bases define positions of the closest amino acids (**a**), and places of all amino acid and stop codons (**b**) in Siemion's one-step mutation ring.[18] Figure 4 presents the algorithm defining the permutations.

a.

	Base	**00**	**01**	**11**	**10**
	↓ 1st/2nd →	**U**	**C**	**A**	**G**
00	**U**	C	U	G	A
01	**C**	U	G	A	C
11	**A**	G	A	C	U
10	**G**	A	C	U	G

b.

2nd →	**U**	2 →	**C**	2 →	**A**	2 →	**G**
1st	3rd	1↓	3 →	1↓	3 →	1↓	3 →
↓ **C**	→ G, A, U, C	**U**	C, U, A, G	**G**	G, A, U, C	**A**	C, U, A, G
↓ **U**	→ C, U, A, G	**G**	G, A, U, C	**A**	C, U, A, G	**C**	G, A, U, C
↓ **G**	→ G, A, U, C	**A**	C, U, A, G	**C**	G, A, U, C	**U**	C, U, A, G
↓ **A**	→ C, U, A, G	**C**	G, A, U, C	**U**	C, U, A, G	**G**	G, A, U, C

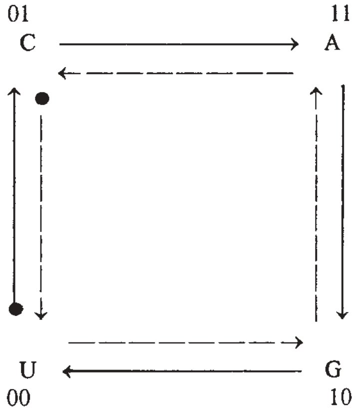


Figure 4. Binary notation of the 4 nucleotides based on purine-pyrimidine and strong-weak H bonding principles. Complementary codon pairs in Tables I–II and Siemion's one step amino acid mutation ring[18] are defined by means of the hypercube node distances (Table V) and related permutations of Table III with 4 nucleotide families, as shown in Table IV. Solid line = 2nd base permutation, dotted line = 1st base permutation, • = start, 3rd base permutation involves CU and AG pairs.

The codon specification analysis defines also the appearance of coded amino acid pairs. Three-letter alphabet combinations according to Michel, with 4 pairs of specified codons, 6 pairs of codons with 1 unspecified base, 3 pairs with 2 unspecified bases (total of 13 pairs) and 1 self-similar/identical pair of 3 unspecified bases give 14 pairs, *i.e.* the previously mentioned 27 letters (Table I, Figure 3a). Probabilities of appearance for these 14 amino acid pairs, presented in Table I, are significantly correlated ($r = 0.747$, $p < 0.01$; Figure 3a) to the number of unspecified codons, which is in agreement with the statistical results of Michel[28] stating that older and less specified codons have a stronger prevalence in present-day genes.

3. *Defining of Ligand-Receptor Recognition by MRT*

Considering the results presented in sections 1 and 2, we may assume that coded peptides/proteins compose X possible classes of complementary amino acids (*e.g.* charged +, –, ≈0, *etc.*). Consequently, binding of each receptor amino acid could be done by $20/X$ peptide amino acids, or with the probability $p = X/20$. For the r amino acid long receptor motif the probability of recognition is $p_r = (X/20)^r$. If the receptor motif is of r length, then there might exist 20^r possible epitopes per N receptors, *i.e.* the probability that the receptor recognizes the epitope is $p_r = N/20^r$.

Thus,
$$p_r = (X/20)^r = N/20^r,\ X = \sqrt[r]{N}. \tag{2}$$

This implies that in the context of the MRT number of receptors N and the length of the coded motif r define the probability of recognition for interacting ligand-receptor systems coded by complementary DNA/RNA strands. Subdivision of the total set of 20 possible complementary amino acids into X observed classes of particular types depends on the length r of the coded peptide motif and on the number of available receptors.

4. *Model of DNA and RNA Coding*

Signal analysis by means of different hypercube nodes constitutes a binary coding system based on 64 codons/codewords. It may be described by a 6-dimensional cube or by a series of 8 (node) 3-dimensional cube permutations of the initial 3-dimensional cube (*i.e.* as 8 × 8 codon octades). Binary tree location of different nodes/codons is presented in Table V. Nucleotides were coded according to their natural sequence appearance (1st, then 2nd, then 3rd base) and by means of the binary notation described in the Methods and Model section. Both procedures are consistent with the results of Jiménez-Montano *et al.*[32] and Halitsky[35] regarding the Gray code structure of the genetic code and *m*RNA – *t*RNA discrimination with respect to the base partition.

TABLE V

Binary and symbolic notation of the DNA, RNA and protein coding language.

aa ↓	codon ↓	Cantor points	binary notation	*symbolic notation*			aa ↑	codon ↑	Cantor points	binary notation	*symbolic notation*		
F	UUU	0	00 00 00	*F*	*UUU*	00 00 00	K	AAA	1	11 11 11	*V*	*GUU*	10 00 00
F	UUC	1/243	00 00 01	*F*	*UUC*	00 00 01	K	AAG	242/243	11 11 10	*V*	*GUC*	10 00 01
L	UUG	2/243	00 00 10	*L*	*UUA*	00 00 11	N	AAC	241/243	11 11 10	*V*	*GUA*	10 00 11
L	UUA	3/243	00 00 11	*L*	*UUG*	00 00 10	N	AAU	240/243	11 11 00	*V*	*GUG*	10 00 10
S	UCU	6/243*	00 01 00	*S*	*UCG*	00 01 10	R	AGA	237/243*	11 10 11	*A*	*GCG*	10 01 10
S	UCC	7/243	00 01 01	*S*	*UCA*	00 01 11	R	AGG	236/243	11 10 10	*A*	*GCA*	10 01 11
S	UCG	8/243	00 01 10	*S*	*UCC*	00 01 01	S	AGC	235/243	11 10 01	*A*	*GCC*	10 01 01
S	UCA	9/243	00 01 11	*S*	*UCU*	00 01 00	S	AGU	234/243	11 10 00	*A*	*GCU*	10 01 00
C	UGU	18/243*	00 10 00	*Y*	*UAU*	00 11 00	T	ACA	225/243*	11 01 11	*D*	*GAU*	10 11 00
C	UGC	19/243	00 10 01	*Y*	*UAC*	00 11 01	T	ACG	224/243	11 01 10	*D*	*GAC*	10 11 01
W	UGG	20/243	00 10 10	*ochre*	*UAA*	00 11 11	T	ACC	223/243	11 01 01	*E*	*GAA*	10 11 11
opal	UGA	21/243	00 10 11	*amber*	*UAG*	00 11 10	T	ACU	222/243	11 01 00	*E*	*GAG*	10 11 10
Y	UAU	24/243*	00 11 00	*W*	*UGG*	00 10 10	I	AUA	219/243*	11 00 11	*G*	*GGG*	10 10 10
Y	UAC	25/243	00 11 01	*opal*	*UGA*	00 10 11	M	AUG	218/243	11 00 10	*G*	*GGA*	10 10 11
amber	UAG	26/243	00 11 10	*C*	*UGC*	00 10 01	I	AUC	217/243	11 00 01	*G*	*GGC*	10 10 01
ochre	UAA	27/243	00 11 11	*C*	*UGU*	00 10 00	I	AUU	216/243	11 00 00	*G*	*GGU*	10 10 00
L	CUU	54/243	01 00 00	*R*	*CGU*	01 10 00	E	GAA	189/243	10 11 11	*S*	*AGU*	11 10 00
L	CUC	55/243*	01 00 01	*R*	*CGC*	01 10 01	E	GAG	188/243*	10 11 10	*S*	*AGC*	11 10 01
L	CUG	56/243	01 00 10	*R*	*CGA*	01 10 11	D	GAC	187/243	10 11 01	*R*	*AGA*	11 10 11
L	CUA	57/243	01 00 11	*R*	*CGG*	01 10 10	D	GAU	186/243	10 11 00	*R*	*AGG*	11 10 10
P	CCU	60/243	01 01 00	*Q*	*CAG*	01 11 10	G	GGA	183/243	10 10 11	*K*	*AAG*	11 11 10
P	CCC	61/243*	01 01 01	*Q*	*CAA*	01 11 11	G	GGG	182/243*	10 10 10	*K*	*AAA*	11 11 11
P	CCG	62/243	01 01 10	*H*	*CAC*	01 11 01	G	GGC	181/243	10 10 01	*N*	*AAC*	11 11 01
P	CCA	63/243	01 01 11	*H*	*CAU*	01 11 00	G	GGU	180/243	10 10 00	*N*	*AAU*	11 11 00
R	CGU	72/243	01 10 00	*P*	*CCU*	01 01 00	A	GCA	171/243	10 01 11	*T*	*ACU*	11 01 00
R	CGC	73/243*	01 10 01	*P*	*CCC*	01 01 01	A	GCG	170/243*	10 01 10	*T*	*ACC*	11 01 01
R	CGG	74/243	01 10 10	*P*	*CCA*	01 01 11	A	GCC	169/243	10 01 01	*T*	*ACA*	11 01 11
R	CGA	75/243	01 10 11	*P*	*CCG*	01 01 10	A	GCU	168/243	10 01 00	*T*	*ACG*	11 01 10
H	CAU	78/243	01 11 00	*L*	*CUG*	01 00 10	V	GUA	165/243	10 00 11	*I*	*AUG*	11 00 10
H	CAC	79/243*	01 11 01	*L*	*CUA*	01 00 11	V	GUG	164/243*	10 00 10	*M*	*AUA*	11 00 11
Q	CAG	80/243	01 11 10	*L*	*CUC*	01 00 01	V	GUC	163/243	10 00 01	*I*	*AUC*	11 00 01
Q	CAA	81/243	01 11 11	*L*	*CUU*	01 00 00	V	GUU	162/243	10 00 00	*I*	*AUU*	11 00 00

*Corresponding positions of Wall's terminating decimals in the Cantor set.[37] Values denote positions 1/40, 3/40, 1/10, 9/40, 1/4, 3/10, 13/40, 27/40, 7/10, 3/4, 31/40, 9/10, 37/40 and 39/40, respectively. aa = amino acids; U = T

The aim of the presented coding/decoding procedure was to provide a simple tool for defining the words, sentences and syntax of DNA, RNA and protein language. Table V shows that the *complementary principle* based on the symmetry of purine-pyrimidine pairs and weak-strong H bonding holds for this notation. The same is valid for the presence of codon octades within the genetic code and the related stereochemical pairings of particular amino acids, wich was first noticed and analyzed by Grafstein.[5–7,27] The data from Tables I–III and Table V confirm Grafstein's observations and link them to the Molecular Recognition Theory. Additionally, closely related positions of the codons for G, A and V in Table V confirm the assumption of Rodin and Ohno[26] that *t*RNA for these amino acids originated in pair.

5. Binary Trees and Related Dynamics

To define more precisely the positions of particular codon intervals, notation based on the *binary tree* with respect to the Cantor set (Table V) was applied. This was performed since the Cantor set possesses two properties related to the binary nucleotide coding:[36]

1. binary decomposition of the initial segment into 2^n segments projected on $(n–1)^{\mathrm{th}}$ binary tree level,

2. partitioning of the observed set by excluding 1/3 of its original length per each of the tree levels.

The relative location of different coding intervals and their orientation is additionally specified in Table V by the nodes of the *alternating binary tree* and their symbolic co-ordinates (names).[36] Briefly, the left half of the unit interval was labelled 0 and the right one 1. For $x < 1/2$ and its derivative $f_\lambda'(x) > 0$, with $f(x) = \lambda x(1–x)$, $\lambda \geq 4$, the pairs of the initial binary tree preserved orientation and for $x > 1/2$, $f_\lambda'(x) < 0$ they reversed orientation in the alternating binary tree.

By means of this notation, it was shown (Table V) that the projection of particular groups of codons from Tables I and II corresponds to the result of Wall[37] defining 14 numbers, which have a terminating decimal expansion in the Cantor set. The latter links complementary pairs of the genetic code presented in Table I and Table II to the mathematics of the Cantor set and chaotic dynamical systems. Pairs were defined by transformation from the codon *binary representation* to the codon *symbolic representation* with respect to Wall's terminating decimals of the Cantor set and its 0 and 1 bounds. This result is valid for all pairs except LN, probably due to the fact that, for the LN combination, two alternative pairs (LD, LE) and a similar one (KF) already exist. These pairs represent connected and/or neighbouring nodes in several circular algorithms defining the genetic code[5–7,18,29–35] and seem to be linked through the N-end rule of selective protein degradation (since in this system N is converted into D and E by deamidase).[38,39]

CONCLUSION

The model of DNA/RNA coding and MRT may provide a better understanding of the evolutional and biomedical aspects of the gene and protein structure or function, and contribute to a better understanding of physiological and pathological processes related to the treatment of different diseases. Networks based on hypercube arhitecture[40] and computer aided-drug design are some of the possible tools for optimizing gene and protein modelling.

Acknowledgements. – The author wishes to thank Dr. Neven Žarković and Mrs. Nevenka Hiršl for the contribution to their experimental work presented in Figure 2.

REFERENCES

1. J. E. Blalock and K. L. Bost, *Biochem. J.* **234** (1986) 679–683.
2. J. E. Blalock and E. M. Smith, *Biochem. Biophys. Res. Commun.* **121** (1984) 203–207.
3. J. E. Blalock, *Nature Medicine* **1** (1995) 876–878.
4. L. Baranyi, W. Campbell, K. Ohishima, S. Fujimoto, M. Boros, and H. Okada, *Nature Medicine* **1** (1995) 894–901.
5. N. Štambuk, *Int. J. Thymology* **5** (1997) 487–491.
6. N. Štambuk, *On the Optimization of Complementary Protein Coding*, in: S. Ohno, K. Aoki, M. Usui, and E. Uchio (Eds.), *Uveitis Today*, Elsevier, Amsterdam, 1998, pp. 315–318.
7. M. M. Rakočević, *Geni Molekuli Jezik*, Naučna Knjiga, Beograd, 1984, pp. 3–225.
8. K. J. Bost, E. M. Smith, and J. E. Blalock, *Biochem. Biophys. Res. Commun.* **128** (1985) 1373–1380.
9. T. K. Gartner, R. Laudon, and D. B. Taylor, *Biochem. Biophys. Res. Commun.* **180** (1991) 1446–1452.
10. G. J. More, R. C. Ganter, and K. J. Franklin, *Biochem. Biophys. Res. Commun.* **160** (1989) 1387–1391.
11. D. A. Weigent, P. D. Hoeprich, K. L. Bost, T. K. Brunck, W. E. Reiher III, and J. E. Blalock, *Biochem. Biophys. Res. Commun.* **139** (1986) 367–374.
12. A. Bajpai, K. P. Hooper, and K. E. Ebner, *Biochem. Biophys. Res. Commun.* **180** (1991) 1312–1317.
13. R. L. Soffer, B. Bandyopadhyay, E. Rosenberg, P. Hoeprich, A. Teitelbaum, T. Brunck, C. B. Colby, and C. G. Gloff*, Proc. Natl. Acad. Sci. USA* **84** (1987) 9219–9222.
14. S. Rodin, A. Rodin, and S. Ohno, *Proc. Natl. Acad. Sci. USA* **93** (1996) 4537–4542.
15. M. D. deSmet, J. H. Yamamoto, M. Mochizuki, V. K. Singh, T. Shinohara, B. Wiggert, G. J. Chader, and R. B. Nussenblatt, *Am. J. Ophthalmol.* **110** (1990) 135–142.
16. J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157** (1982) 105–132.
17. G. D. Fasman, *Development of the Prediction of Protein Structure*, in: G. D. Fasman (Ed.), *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum, New York, 1989, p. 232.

18. I. Siemion, *Amino Acids* **8** (1995) 1–13.
19. I. Z. Siemion and P. J. Siemion, *BioSystems* **33** (1994) 139–148.
20. S. V. Yablonsky, *Introduction to Discrete Mathematics*, Moscow, Mir, 1989, pp. 298–336.
21. C. Roudier, I. Auger, and J. Roudier, *Immunol. Today* **17** (1996) 357–358.
22. N. Štambuk, K. Šentija, M. Rudolf, R. Mažuran, T. Marotti, V. Šverko, I. Svoboda-Beusan, S. Rabatić, M. Trbojević-Čepe, S. Seiwerth, V. Garaj-Vrhovac, M. Banović, and M. Č. Pešić, *Int. J. Thymology* **3** (1995) 322–327.
23. N. Štambuk, K. Šentija, B. Marušić-DellaMarina, M. Trbojević-Čepe, M. Rudolf, V. Garaj-Vrhovac, and B. Pokrić, *Ocular Immunol. Inflamm.* **5** (1997) S52–S53.
24. N. Štambuk, V. Brinar, V. Štambuk, I. Svoboda-Beusan, R. Mažuran, S. Rabatić, B. Marušić-DellaMarina, N. Zurak, Z. Brzović, T. Marotti, V. Šverko, M. Rudolf, M. Trbojević-Čepe, R. Martinić, B. Malenica, N. Mašić, A. Gagro, K. Karaman, Z. Sučić, I. Dujmov, and B. Pokrić, *Peptid-M (LUPEX®) Effects on the Immune Response and Clinical Status in Uveitis, Optic Neuritis and Multiple Sclerosis*, in: S. Ohno, K. Aoki, M. Usui and E. Uchio (Eds.), *Uveitis Today*, Elsevier, Amsterdam, 1998, pp. 319–322.
25. K. K. W. Wang and P. W. Yuen, *Trends. Pharmacol. Sci.* **15** (1994) 412–419.
26. R. S. Rodin and S. Ohno, *Proc. Natl. Acad. Sci. USA* **94** (1997) 5183–5188.
27. D. Grafstein, *J. Theor. Biol.* **105** (1983) 157–174.
28. C. J. Michel, *Math. Biosci.* **97** (1989) 161–177.
29. N. Štambuk, *Klin. Monatsbl. Augenheilkd.* **211 suppl 5** (1997) 4.
30. D. G. Arques and C. J. Michel, *J. Theor. Biol.* **182** (1996) 45–58.
31. D. G. Arques, J. P. Fallot, and C. J. Michel, *J. Theor. Biol.* **185** (1997) 241–253.
32. M. A. Jiménez-Montano, C. R. de la Mora Basanez and T. Pöschel, *BioSystems* **39** (1996) 117–125.
33. R. Swanson, *Bull. Math. Biol.* **46** (1984) 187–203.
34. C. T. Zhang, *J. Theor. Biol.* **187** (1997) 297–306.
35. D. Halitsky, *Math. Biosci.* **121** (1994) 227–234.
36. N. B. Tufillaro, T. Abbott, and J. Reilly, *An Experimental Approach to Nonlinear Dynamics and Chaos*, Addison-Wesley, Redwood City, 1992, pp 79–109.
37. C. R. Wall, *Fibonacci. Quart.* **28** (1990) 98–101.
38. R. J. Dohmen, K. Madura, B. Bartel, and A. Varshavsky, *Proc. Natl. Acad. Sci. USA* **88** (1991) 7351–7355.
39. A. Varshavsky, *Proc. Natl. Acad. Sci. USA* **93** (1996) 12142–12149.
40. J. Bruck and C. T. Ho, *IEEE Trans. Inform. Theory* **42** (1996) 2217–2221.

SAŽETAK

O genetičkom porijeklu komplementarnog proteinskog kodiranja

Nikola Štambuk

U radu su proučavani odnosi proteinskog kodiranja i hidropatskih osobina aminokiselina, u skladu s načelima teorije molekulskog prepoznavanja i Grafsteinove hipoteze o stereokemijskom porijeklu genetičkog koda. Pokazano je da se postupak kodiranja RNA i DNA zasniva na 14 skupina parova kodon-antikodon, kojima su kodirani komplementarni parovi aminokiselina i zaustavni kodoni. Teorija molekulskog prepoznavanja razmotrena je s obzirom na odnose kodon–antikodon u *m*RNA, DNA, *t*RNA te Siemionovljev aminokiselinski mutacijski prsten koji predočuje strukturu genetičkog koda. Prikazan je i model DNA, RNA i proteinskog kodiranja (i dekodiranja) zasnovan na načelima stacionarnosti i komplementarnosti, pri čemu *stacionarnost* označuje osobinu kodiranja vlastite uzvojnice dok *komplementarnost* odlikuje postupak kodiranja suprotne DNA/RNA. Kombinacije dviju brojki, u binarnoj notaciji, označju primarnu i sekundarnu osobinu pojedinih baza te se njihovim permutacijama definiraju kodoni. Odnosi pojedinih kodona definirani su s pomoću modela hiperkocke i binarnog drveta, s obzirom na Cantorov skup, Feigenbaumovu dinamiku i analizu signala. Kombinacije različitih čvorova hiperkocke i binarnog drveta definiraju riječi, rečenice i sintaksu DNA, RNA i proteinskog jezika. Moguće su primjene spomenute metode u analizi mreža, u nacrtu gena i proteina te u modeliranju lijekova.