Binary Coding, mRNA Information and Protein Structure

Nikola Štambuk[1], Paško Konjevoda[1] and Nikola Gotovac[2]

[1] Ruđer Bošković Institute, Zagreb, Croatia
[2] Department of Radiology, General Hospital Požega, Croatia

We describe new binary algorithm for the prediction of α and β protein folding types from RNA, DNA and amino acid sequences. The method enables quick, simple and accurate prediction of α and β protein folds on a personal computer by means of a few binary patterns of coded amino acid and nucleotide physicochemical properties. The algorithm was tested with machine learning SMO (sequential minimal optimization) classifier for the support vector machines and classification trees, on a dataset of 140 dissimilar protein folds. Depending on the method of testing, the overall classification accuracy was 91.43% – 100% and the tenfold cross-validation result of the procedure was 83.57% – >90%.

Genetic code randomization analysis based on 100,000 different codes tested for the protein fold prediction quality indicated that: a) there is a very low chance of $p = 2.7 \times 10^{-4}$ that a better code than the natural one specified by the binary coding algorithm is randomly produced, b) dipeptides represent basic protein units with respect to the natural genetic code defining of the secondary protein structure.

Keywords: RNA, DNA, amino acid, genetic code, protein, structure, prediction, selection, evolution.

1. Introduction

Protein folding prediction from the nucleotide strings is important because the Genome Project has resulted in a large number of gene sequences that code for different, often unknown, protein sequences. Although a large body of literature relates physical and chemical properties of the amino acids and protein folding, relatively little is known about the relationships of the codon composition and secondary protein structure.

The genetic code defines how base triplets are assigned to the amino acids in order to code protein molecules [1-5]. As the protein chain is constructed on the ribosome, it folds up in the way that the free energy is minimized, i.e. the most comfortable configuration is achieved [2, 3]. Recent results indicated that simple binary coding patterns of amino acid and nucleotide physicochemical properties might be used for design, modeling and prediction of the basic protein folding types [6-11].

The aim of the paper is to present simple, quick and accurate algorithm for the prediction of secondary protein structure from the nucleotide sequences of newly sequenced exon regions [8-10]. The precision of the model is within the range of experimental error of the secondary protein structure determination [9]. The method enables data compression, digitalization of the RNA/DNA sequences and provides a basis for new heuristic algorithms. The procedure may be applied to database search and data structure analyses (e.g. of GenBank or PDB). Some other popular and often used methods, e.g. artificial neural networks, do not provide information on the data structure [12].

We investigated the algorithm with respect to mRNA coding of α (helix) and β (strand) protein fold structures. The prediction efficacy of the algorithm was also compared to a large number of randomly produced codes, in order to obtain better insight into the processes of code selection and optimization [3, 13].

2. Results and Discussion

2.1. Nucleotide and Amino Acid Coding

Sixty-four nucleotide triplets, i.e. codons, define 61 triplet and 3 stop codons for the amino acid and protein synthesis [1-5, 8-14]. Each triplet consists of 3 bases, selected out of 4 possible ones: uracil (U) or thymine (T), cytosine (C), adenine (A) and guanine (G) [8-10]. Binary algorithm investigated in this study is based on the representation of 4 nucleotide bases, according to the notation U or T $=$ 00, C $=$ 01, G $=$ 10 and A $=$ 11 [5, 8-10]. It enables the construction of a linear block code array that reconstructs the genetic code table and accurately predicts α and β protein folds [5, 8-10].

The first digit of the binary algorithm defines type of the base ring (pyrimidine is coded by 0 and purine by 1), while the second digit defines keto group (0) or amino group (1) of the ring. Complementarity is achieved by $0 \leftrightarrow 1$ digit changes. Partition of complementary base pairs into the weak (A, U or T) and strong (G, C) hydrogen bonding groups is also defined [5, 8-10].

Quantum chemical electron-donor and electron-acceptor base properties measured by Pullman are also in agreement with the presented binary notation [9]. If the bases that are bad donors and acceptors are denoted by 0 and good donors and acceptors by 1, the notation corresponds to Fig. 2 and Fig. 3 of the classic Pullman results published in 1965 (Hückel-type calculations) [9, 15].

2.2. Binary Algorithms of the Code

Binary notation of 64 codons and 20 amino acids is presented in Table 1. The information content of each codon sequence is "weighted" according to the classic method of a toss of a fair coin, i.e. by calculating the probabilities p of each element on [0, 1] interval P [9, 10, 16-20]. The coin is tossed n times to define the position of each binary address of length n over the alphabet $A = \{0, 1\}$, as follows: $p = \sum j_n/2^n$, $j_n = 0$ for coin tossing outcome 0 and $j_n = 1$ for the outcome 1. This binary algorithm is often applied in information theory for similar purposes [9, 10, 16, 20].

According to the Grantham's scale of molecular polarity, we performed binary defining of amino acid physicochemical properties by partitioning amino acids into nonpolar group 0 (Y, M, V, L, F, I, W, C) and polar group 1 (H, R, Q, K, N, E, D, P, A, T, S, G) [9, 21]. The groups were extracted by means of partitioning around medoids procedure of clustering, with S-Plus software [9, 22, 23].

This method enabled extraction of efficient binary algorithm for the protein fold prediction [9], which confirms experimental results of Kamtekar et al. [6]. Binary algorithm of amino acid polarity reduces the number of analyzed elements within the protein motif (or sliding block) of length n by the factor of 10^n (from 20^n to 2^n) [9].

2.3. Amino Acid-Nucleotide Relationships

Groups that share information content with respect to both nucleotide and amino acid binary coding of physicochemical properties were extracted by means of the classification tree [8, 9, 23]. Due to the fact that messenger RNA consists of codons, and serves as a template for amino acid based protein synthesis on the ribosome, codons were treated as independent and related amino acids as dependent variables (Table 1, Fig. 1).

Classification tree extracted, with 100% accuracy, 8 groups of codons that exhibit close statistical relationship based on the physicochemical coding of nucleotide properties and molecular polarity coding of amino acids (Table 1, Fig. 1).

The classification of 8 groups of codons in Fig. 1 is the result of the series of classification rules obtained by means of a procedure of recursive

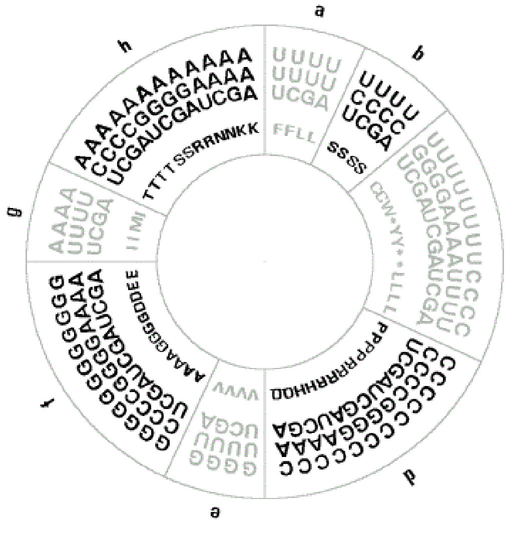

Fig. 1. Coding patterns of 64 nucleotide triplets and 20 amino acids define codon ring identical to the genetic code table. Codon groups a, c, e, g represent nonpolar amino acids and codon groups b, d, f, h polar amino acids.

aa	**codon**	**position**	**notation**	**aa**	**codon**	**position**	**notation**
F	***UUU***	0.0000	00 00 00	**K**	***AAA***	0.9843	11 11 11
F	***UUC***	0.0156	00 00 01	**K**	***AAG***	0.9688	11 11 10
L	***UUG***	0.0313	00 00 10	**N**	***AAC***	0.9531	11 11 01
L	***UUA***	0.0469	00 00 11	**N**	***AAU***	0.9375	11 11 00
S	***UCU***	0.0625	00 01 00	**R**	***AGA***	0.9219	11 10 11
S	***UCC***	0.0782	00 01 01	**R**	***AGG***	0.9063	11 10 10
S	***UCG***	0.0938	00 01 10	**S**	***AGC***	0.8906	11 10 01
S	***UCA***	0.1094	00 01 11	**S**	***AGU***	0.8750	11 10 00
C	***UGU***	0.1250	00 10 00	**T**	***ACA***	0.8594	11 01 11
C	***UGC***	0.1410	00 10 01	**T**	***ACG***	0.8438	11 01 10
W	***UGG***	0.1563	00 10 10	**T**	***ACC***	0.8281	11 01 01
st	***UGA***	0.1719	00 10 11	**T**	***ACU***	0.8125	11 01 00
Y	***UAU***	0.1875	00 11 00	**I**	***AUA***	0.7969	11 00 11
Y	***UAC***	0.2031	00 11 01	**M**	***AUG***	0.7813	11 00 10
st	***UAG***	0.2188	00 11 10	**I**	***AUC***	0.7656	11 00 01
st	***UAA***	0.2344	00 11 11	**I**	***AUU***	0.7500	11 00 00
L	***CUU***	0.2500	01 00 00	**E**	***GAA***	0.7344	10 11 11
L	***CUC***	0.2656	01 00 01	**E**	***GAG***	0.7188	10 11 10
L	***CUG***	0.2813	01 00 10	**D**	***GAC***	0.7031	10 11 01
L	***CUA***	0.2969	01 00 11	**D**	***GAU***	0.6875	10 11 00
P	***CCU***	0.3125	01 01 00	**G**	***GGA***	0.6719	10 10 11
P	***CCC***	0.3281	01 01 01	**G**	***GGG***	0.6563	10 10 10
P	***CCG***	0.3438	01 01 10	**G**	***GGC***	0.6401	10 10 01
P	***CCA***	0.3594	01 01 11	**G**	***GGU***	0.6250	10 10 00
R	***CGU***	0.3750	01 10 00	**A**	***GCA***	0.6094	10 01 11
R	***CGC***	0.3906	01 10 01	**A**	***GCG***	0.5938	10 01 10
R	***CGG***	0.4063	01 10 10	**A**	***GCC***	0.5781	10 01 01
R	***CGA***	0.4219	01 10 11	**A**	***GCU***	0.5625	10 01 00
H	***CAU***	0.4375	01 11 00	**V**	***GUA***	0.5469	10 00 11
H	***CAC***	0.4531	01 11 01	**V**	***GUG***	0.5313	10 00 10
Q	***CAG***	0.4688	01 11 10	**V**	***GUC***	0.5156	10 00 01
Q	***CAA***	0.4844	01 11 11	**V**	***GUU***	0.5000	10 00 00

aa = amino acids, st = stop codon

Table 1. Binary coding of nucleotide triplets (codons) and related amino acids.

partitioning [23, 24]. Since this exploratory statistical technique uncovers structure in data, it is not surprising that codon ring in Fig. 1, obtained by means of the algorithm, reconstructs standard genetic code arrangement presented in Table 1.

Classification rules for 8 groups of codons are defined by breakpoints between different polar and nonpolar amino acid groups according to the codon positions on the unit interval, as shown in Table 1. Seven breakpoints that separate 8 codon groups into the ones that code for nonpolar and polar amino acids are: 0.055, 0.117, 0.305, 0.492, 0.555, 0.742, and 0.805. More details on the procedure can be found in Štambuk and Konjevoda [9].

2.4. Protein Fold Prediction

The prediction method is based on the analyses of relative frequencies of 64 *dipeptide* patterns, i.e. 8 × 8 groups, of coded nucleotide and amino acid physicochemical properties.

The relative frequencies of 64 coding patterns were first analyzed (counted) within the protein sliding block of the length 2, which defines dipeptides [2, 8, 9]. Then, 64 relative frequency patterns of α and β protein folds were compared by means of SMO machine learning algorithm and classification tree, in order to provide two different and accurate learning algorithms for the protein fold prediction [8, 9, 23-25].

Data were analyzed by means of two software packages. Classification tree was obtained by means of S-Plus 2000 software [8, 9, 22-24]. Weka (Waikato Environment for Knowledge Analysis) software, version 3.1.7, was used for the classifications with machine learning Sequential Minimal Optimization (SMO) algorithm for the Support Vector Machines [8, 9, 25].

Eight groups of codons, specified by means of the nucleotide and amino acid physicochemical properties, differ significantly in α and β protein folds (Table 2) and enable the construction of 64 *dipeptides*. Based on 8 letter alphabet permutation, those 64 elementary *dipeptide patterns* carry over the information on the *symbolic* coding characteristics of basic protein units that define peptide bonds within protein and exon sliding block [2, 8, 9].

The term *dipeptide* is given in italics since it is not a standard dipeptide consisting of 2 amino acid elements [8, 9]. *Dipeptide* represents a coding pattern for 2 amino acids and their related nucleotide groups based on the physicochemical characteristics defined by means of the coding algorithm presented in Fig. 1 [9]. Relative frequencies of 28 out of 64 possible *dipeptides* differ significantly in α and β protein folds, and exhibit distinct patterns relevant for the pro-

group	mean α	mean β	t-value	p-level
a	0.0941	0.1133	-3.123	0.0022
b	0.0654	0.0684	-0.523	0.6019
c	0.1205	0.0945	3.603	0.0004
d	0.1022	0.0796	3.310	0.0012
e	0.1263	0.1526	-4.087	0.0001
f	0.1588	0.1574	0.177	0.8594
g	0.1324	0.1468	-1.956	0.0525
h	0.2002	0.1875	1.887	0.0612

$p < 0.05$ (Hotelling's T = 53.4, $p < 0.000001$)

Table 2. Relative frequencies of amino acid and codon groups a-h in α and β protein folds.

dipeptide	mean α	mean β	t-value	p-level
aa	0.0037	0.0043	-0.516	0.6069
ab	0.0025	0.0042	-1.222	0.2236
ac	0.0075	0.0061	1.067	0.2879
ad	0.0096	0.0091	0.278	0.7811
ae	0.0023	0.0036	-1.725	0.0867
af	0.0189	0.0134	2.570	0.0112
ag	0.0049	0.0040	0.788	0.4320
ah	0.0152	0.0155	-0.094	0.9249
ba	0.0015	0.0046	-3.147	0.0020
bb	0.0005	0.0026	-3.591	0.0005
bc	0.0060	0.0069	-0.633	0.5279
bd	0.0043	0.0059	-1.369	0.1733
be	0.0013	0.0038	-3.084	0.0025
bf	0.0096	0.0150	-2.650	0.0090
bg	0.0014	0.0039	-3.081	0.0025
bh	0.0055	0.0108	-3.298	0.0012
ca	0.0081	0.0051	2.415	0.0171
cb	0.0047	0.0076	-2.114	0.0363
cc	0.0209	0.0147	2.170	0.0317
cd	0.0240	0.0149	3.194	0.0017
ce	0.0053	0.0084	-2.171	0.0317
cf	0.0426	0.0262	5.510	0.0001
cg	0.0098	0.0082	0.949	0.3444
ch	0.0314	0.0300	0.455	0.6497
da	0.0105	0.0087	1.114	0.2673
db	0.0031	0.0067	-3.040	0.0028
dc	0.0203	0.0152	1.970	0.0508
dd	0.0205	0.0178	0.850	0.3969
de	0.0090	0.0105	-0.908	0.3653
df	0.0385	0.0323	1.679	0.0955
dg	0.0096	0.0098	-0.159	0.8741
dh	0.0301	0.0257	1.491	0.1383
ea	0.0029	0.0045	-1.808	0.0728
eb	0.0024	0.0040	-1.843	0.0675
ec	0.0046	0.0088	-3.163	0.0019
ed	0.0063	0.0088	-1.473	0.1431
ee	0.0019	0.0064	-3.788	0.0002
ef	0.0140	0.0206	-3.187	0.0018
eg	0.0031	0.0066	-2.903	0.0043
eh	0.0127	0.0197	-3.307	0.0012
fa	0.0195	0.0140	2.476	0.0145
fb	0.0073	0.0137	-4.072	0.0001
fc	0.0471	0.0277	5.662	0.0001
fd	0.0364	0.0335	0.814	0.4168
fe	0.0141	0.0198	-2.472	0.0146
ff	0.0813	0.0666	2.503	0.0135
fg	0.0261	0.0174	3.160	0.0019
fh	0.0528	0.0628	-2.521	0.0128
ga	0.0044	0.0031	1.166	0.2456
gb	0.0028	0.0024	0.414	0.6799

gc	0.0098	0.0075	1.531	0.1280
gd	0.0118	0.0108	0.518	0.6050
ge	0.0047	0.0060	-0.741	0.4602
gf	0.0201	0.0176	1.067	0.2877
gg	0.0051	0.0047	0.390	0.6970
gh	0.0178	0.0208	-1.093	0.2762
ha	0.0138	0.0160	-1.079	0.2823
hb	0.0061	0.0109	-3.293	0.0013
hc	0.0318	0.0290	0.962	0.3376
hd	0.0286	0.0268	0.626	0.5320
he	0.0094	0.0205	-5.060	0.0001
hf	0.0591	0.0640	-1.095	0.2752
hg	0.0164	0.0173	-0.450	0.6532
hh	0.0432	0.0522	-1.894	0.0603

Table 3. Relative frequencies of dipeptides.

tein fold prediction (Hotelling's T test = 253, $p < 0.0008$; Table 3).

Classification of 140 dissimilar α and β protein folds [8, 9] obtained by means of relative frequencies of 64 *dipeptides*, with sequential minimal optimization (SMO) machine learning algorithm [8, 9, 25], results in 91.43% overall prediction accuracy and 83.57% correct prediction under tenfold cross-validation of the procedure. Classification tree confirms the result of SMO by 100% accurate classification of α and β protein folds (Fig. 2).

At the level of a single amino acid element, overall protein fold prediction from the primary sequence is 90% (83.57% under tenfold cross-validation). However, at the level of dipeptide, it is not possible to analyze 400 patterns arising from the 20 × 20 dipeptide patterns, since it is well known that direct inclusion of such number of input parameters requires data reduction or filtration [8, 26]. Consequently, presented algorithm of *dipeptide*-based *data reduction* avoids such problems and enables accurate, quick and simple protein fold analysis and/or prediction at the level of basic protein units (dipeptides) that define peptide bond between two neighboring amino acids.

The test set of 140 nucleotide sequences used for the computation [8, 9] is, to our knowledge, the largest available mRNA dataset of 70 α and 70 β dissimilar (nonhomologous) protein folds with known three-dimensional structure, i.e., with NMR and those X-ray crystal structure at the resolutions of < 2.5 Angstroms [8, 9, 27-29]. To ensure the best possible accuracy,

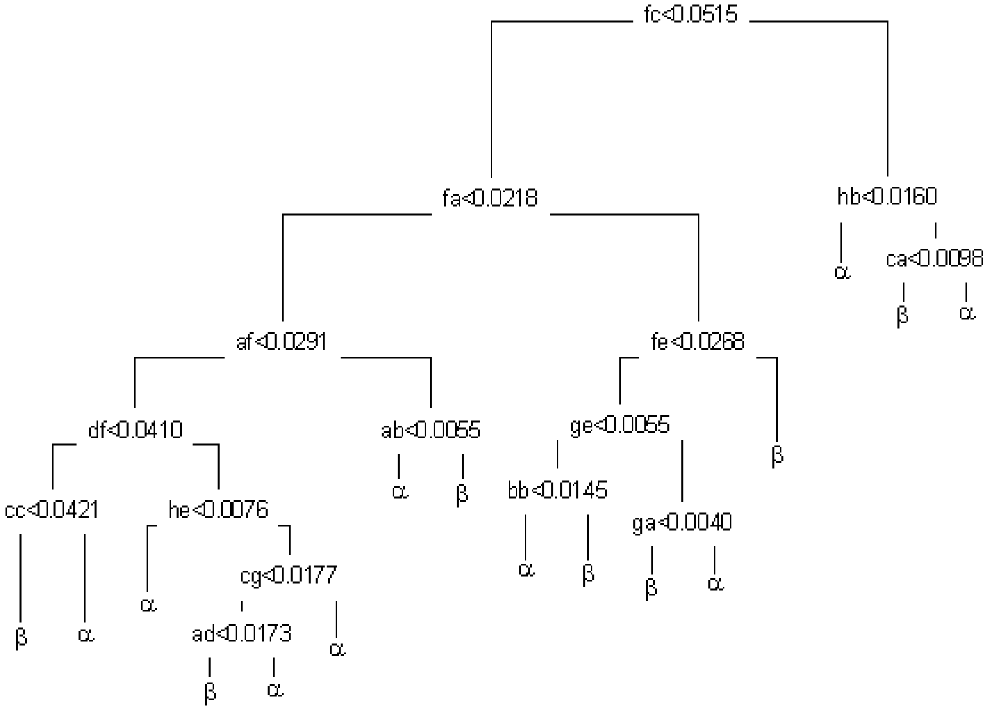

Fig. 2. Classification tree for the prediction of α and β protein folds from *dipeptides* (Table 3).

we analyzed the prediction results based on the tenfold cross-validation tests [25, 27].

Presented results may reflect the situation when physicochemical information of the codons becomes a functional component of the secondary protein structure following the translation process of ribosome mRNA. During the translation process nucleotide information stored within DNA strings, and subsequently transcribed into RNA message, is converted into the polypeptide structure through a series of peptide bonds of the globular proteins [2]. Therefore, it is not surprising that SMO algorithm prediction of α and β protein folds based on 8 symbolic elements a-h does not contain enough information (and elements) to accurately predict secondary protein structure. Consequently, 8-element protein fold prediction based on the coding of monomers is significantly lower (70.17% overall and 68.57% under tenfold cross-validation) than the fold prediction based on *dipeptides*.

2.5. Alphabet Reduction and Fold Prediction

The main influence on the protein folding comes from the side chain properties [2]. Polarity parameter of the Grantham scale is a typical amino acid side chain property that is correlated to the relative substitution frequency between protein residues [21, 30]. In globular proteins, polypeptides of specific stable shapes are folded up due to the interactions of the residual side chains and the interactions with molecules of the medium [2, 31].

The model we described defines protein fold prediction from the nucleotide sequence. When the fold prediction from the amino acid sequence is needed, the number of groups in Fig. 1 may be reduced until codon-amino acid bias is corrected. The sum of codon groups a, c, e and g defines the group of nonpolar amino acids (0), while the sum of codon groups b, d, f and h defines a group of polar amino acids (1). We also evaluated binary classification of α and β protein folds by means of SMO machine learning algorithm. The prediction results were obtained from the relative frequencies of the binary patterns of polar and nonpolar amino acids within protein sliding blocks of different lengths [32]. The best protein structure prediction result of 100% (tenfold cross-validation: 85%) was obtained with 128 binary heptapeptide patterns [32]. Tree model for heptapeptide-based classification confirmed the results of SMO machine learning procedure and extracted 12 patterns of amino acid polarity relevant for 100% accurate α and β protein fold prediction [32]. This result is in agreement with *dipeptide*-based tree classification presented in Fig. 2. It shows that a total number of basic protein units relevant for the secondary structure description is well below the finite set of basic folding types recently estimated to be between 500 and 1000 [9, 32, 33]. Similar results with respect to fold prediction and procedure cross-validations were also obtained for hexapeptides and octapeptides [9, 32].

2.6. Comparison of Prediction Methods

It is worth mentioning that alphabet reduction, e.g. from 8 letters algorithm of nucleotide-amino acid relationships into binary nucleotide-amino acid alphabet, reduces the information content of the sliding block during the protein fold prediction.

α and β protein fold prediction with 64 *dipeptides* of the 8 letter alphabet is very close to the prediction of hexapeptide sliding block based on binary patterns of amino acid polarity, i.e. the 2 letter alphabet [9]. Both classification procedures were done on the same dataset [9] and with identical machine learning classifier.

Described nucleotide-amino acid alphabet reductions, with respect to physicochemical parameter coding, may be useful in situations when specific length of the protein or nucleotide sliding block is requested for the fold prediction and modeling purposes [9, 32].

2.7. Genetic Code Randomization Analysis

Presented model of nucleotide and amino acid coding of physicochemical parameters was also tested on α and β protein fold dataset [9] by means of the permutation distribution of randomly produced models, i.e. codes [34].

The 99,999 artificial models were constructed by a random allocation (assignment) of 64

Table 1 codons within 8 groups in Fig. 1. The eight groups a-h consisted of 4, 4, 12, 12, 4, 12, 4 and 12 elements, respectively. Those groups were, as previously discussed, extracted by means of the binary algorithm for physicochemical coding of nucleotide-amino acid relationships (Table 1, Fig. 1). Another 99,999 artificial models were constructed by a random allocation (assignment) of codons within all possible 64 *dipeptides* that arise from the 8 Fig. 1 groups. In this way, the codes could be observed with respect to a protein fold prediction quality of: a) amino acid monomers, b) peptide bond possessing units, i.e. dipeptides consisting of 2 amino acids.

The 99,999 different random permutations of 8 and 64 groups of nucleotide triplets (codons) were chosen in accordance with the Algorithm P of Knuth [35-37]. The pseudorandom generator required as input to the algorithm was that provided by Delphi 5 of Borland Inter Inc [36-38]. The result of the models based on the binary patterns of nucleotide and amino acid coding of physicochemical properties (Fig. 1, Table 1) was added to the results of 99,999 randomly produced codes to obtain a total of 100,000 codes of both groups.

Prediction quality of the codes belonging to both distributions was determined with SMO classifier under tenfold cross-validation, from relative frequency patterns [8, 9, 25].

Distribution of the results in Fig. 3 reflects the permutation distribution of random allocation of 64 codons within 8 groups presented in Fig. 1. Eight group (letter)-based protein fold prediction of 68.57% is identical to the random distribution median (68.57%) and almost identical to the random distribution arithmetic mean (68.14%).

Completely different result is obtained for a random distribution of 64 *dipeptides* in Fig. 4. This distribution has a mean at 64.95% (median 65%) with a standard deviation of 6.19%. Out of 100,000 codes tested for the secondary protein structure prediction, only 27 randomly generated codes gave better structure prediction than the binary one based on 64 *dipeptides* (83.57% prediction under tenfold cross-validation, Table 3). The binary code position within a cumulative distribution (99.973%, Z = 3.008) implies that, with respect to the secondary protein structure, natural genetic code defining of codon and amino acid physicochemical properties is far away from the random organization. When *dipeptides* are observed, there is only 2.7×10^{-4} chance to produce better code than the natural one.

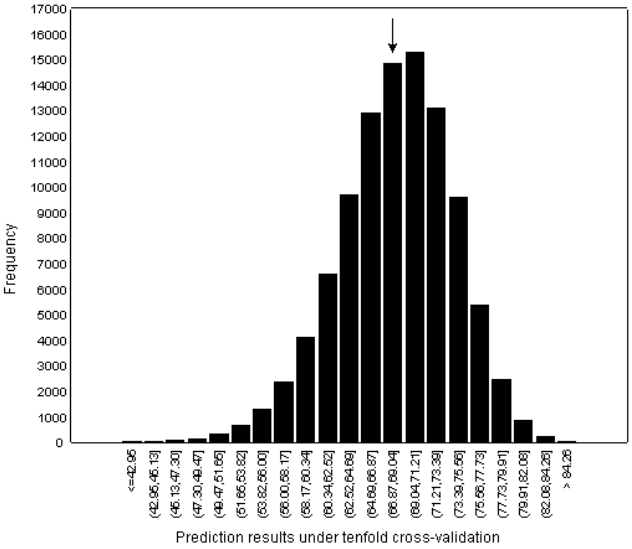

Fig. 3. Protein fold prediction quality of the natural code and 99,999 randomly produced codes. The procedure is based on 8 codon-amino acid groups (a-h). Arrow indicates the position of the natural code.

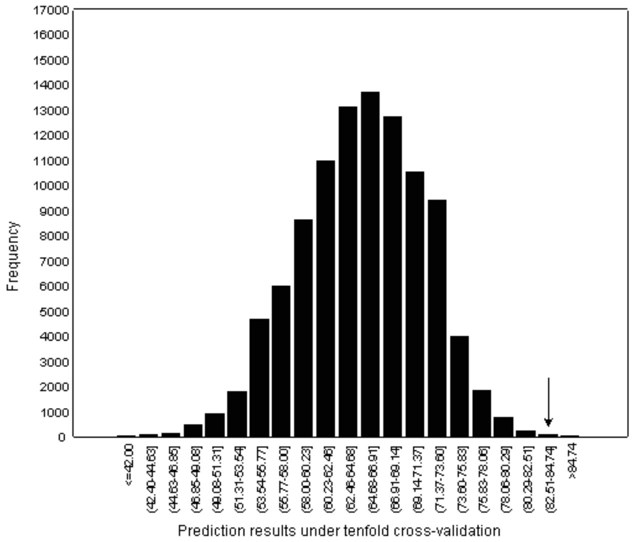

Fig. 4. Protein fold prediction quality of the natural code and 99,999 randomly produced codes. The procedure is based on 64 *dipeptides*, obtained by the permutation of 8 codon-amino acid groups (a-h). Arrow indicates the position of the natural code.

The result of our computation experiment is close to the prediction of Freeland et al. [13] that a code as good as, or better than that chosen by nature, would evolve without selection within the probability interval $2 \times 10^{-4} < p < 10^{-6}$. Consequently, *dipeptide*-based organization of the nucleotide patterns could be essential element of the secondary protein structure, buried within the genetic code.

3. Conclusion

Binary algorithms presented in this study enable simple, quick and accurate prediction of the secondary protein structure from the primary RNA, DNA and amino acid sequences.

Algorithmic information theory, and related cryptographic methods will provide useful tools for further analysis and modeling of the protein structure and genetic code patterns.

References

[1] CRICK F. H. C., The Origin of the Genetic Code, *Journal of Molecular Biology* 1968; 38:367–379.

[2] DOOLITTLE R. F., Proteins, *Scientific American* 1985; 253(4):74–83.

[3] KNIGHT R. D., FREELAND S. J., LANDWEBER L. F., Selection, history and chemistry: the three faces of the genetic code, *Trends in Biochemical Sciences* 1999; 24:241–247.

[4] SWANSON R., A Unifying Concept for the Amino Acid Code, *Bulletin of Mathematical Biology* 1984; 46(2):187–203.

[5] ŠTAMBUK N., On the Genetic Origin of Complementary Protein Coding, *Croatica Chemica Acta* 1998; 71(3):573–589.

[6] KAMTEKAR S., SCHIFFER J. M., XIONG H., BABIK J. M., HECHT M. H., Protein Design by Binary Patterning of Polar and Nonpolar Amino Acids, *Science* 1993; 262:1680–1685.

[7] LI H., HELLING R., TANG C., WINGREEN N., Emergence of Preferred Structures in a Simple Model of Protein Folding, *Science* 1996; 273:666–669.

[8] ŠTAMBUK N., KONJEVODA P., New Computational Algorithm for the Prediction of Protein Folding Types, *International Journal of Quantum Chemistry* 2001; 84(1):13–22.

[9] ŠTAMBUK N., KONJEVODA P., Prediction of Secondary Protein Structure with Binary Coding Patterns of Amino Acid and Nucleotide Physicochemical Properties, *International Journal of Quantum Chemistry* 2003; 92(2):123–134.

[10] ŠTAMBUK N., Universal Metric Properties of the Genetic Code, *Croatica Chemica Acta* 2000; 73(4):1123–1139.

[11] JIMENEZ-MONTANO M. A., DE LA MORA-BASÁNEZ C. R., PÖSCHEL T., The Hypercube Structure of the Genetic Code Explains Conservative and Nonconservative Aminoacid Substitutions In Vivo and In Vitro, *BioSystems* 1996; 39:117–125.

[12] BERGERON B., *Bioinformatics Computing*, New Jersey: Prentice Hall, 2003.

[13] FREELAND S. J., KNIGHT R. D., LANDWEBER L. F., Measuring adaptation within the genetic code, *Trends in Biochemical Sciences* 2000; 25:44–45.

[14] WOESE C. R., Evolution of the Genetic Code, *Naturwissenschaften* 1973; 60:447–459.

[15] PULLMAN B., Some Recent Developments in the Quantum-Mechanical Studies on the Electronic Structure of the Nucleic Acids, *Journal of Chemical Physics* 1965; 43(10):S233–S243.

[16] CALUDE C., *Information and Randomness*, New York: Springer; 1994.

[17] CHAITIN G. J., A Theory of Program Size Formally Identical to Information Theory, *Journal of the ACM* 1975; 22:329–340.

[18] CHAITIN G. J., Information-Theoretic Incompleteness, *Applied Mathematics and Computation* 1992; 52:83–101.

[19] CHAITIN G. J., On the Length of Programs for Computing Finite Binary Sequences, *Journal of the Association for Computing Machinery* 1966; 13(4):547–569.

[20] CHAITIN G. J., *The Unknowable*, Singapore: Springer; 1999.

[21] GRANTHAM R., Amino Acid Difference Formula to Help Explain Protein Evolution, *Science* 1974; 185:862–864.

[22] Data Analysis Products Division, MathSoft, *S-PLUS 2000 Guide to Statistics*, vol 2. Seattle: MathSoft; 1999.

[23] VENABLES V. N., RIPLEY B. D., *Modern Applied Statistics With S-PLUS*, New York: Springer; 1997.

[24] EVERITT B. S., DUNN G., *Applied Multivariate Data Analysis*, London: Arnold; 2001.

[25] WITTEN I. H., FRANK E., *Data Mining*, San Francisco: Morgan Kaufmann; 2000.

[26] EDLER L., GRASSMANN J., SUHAI S., Role and Results of Statistical Methods in Protein Fold Class Prediction, *Mathematical and Computer Modeling* 2001; 33:1401–1417.

[27] CUFF J A., BARTON G. J., Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction, *PROTEINS: Structure, Function and Genetics* 1999; 34:508–519.

[28] CUFF J. A., CLAMP M., SIDDIQUI A. S., FINLAY M., BARTON G. J., JPred: A Consensus Secondary Structure Prediction Server, *Bioinformatics* 1998; 14:892–893.

[29] CHOU K. C., MAGGIORA G. M., Domain Structural Class Prediction, *Protein Engineering* 1998; 11(7):523–538.

[30] MCLACHLAN A. D., Repeating Sequences and Gene Duplication in Proteins, *Journal of Molecular Biology* 1972; 64:417–437.

[31] ROSE G. D., GESELOWITZ A. R., LESSER G. J., LEE R. H., ZEHFUS M. H., Hydrophobicity of Amino Acid Residues in Globular Proteins, *Science* 1985; 229:834–838.

[32] ŠTAMBUK N., KONJEVODA P., GOTOVAC N., Nucleotide Coding of Amino Acid Polarity and Protein Structure, *Acta Physica et Chimica Debrecina* 2002; 34–35:171–188.

[33] DENTON M., MARSHALL C., Laws of the Form Revisited, *Nature* 2001; 410:417.

[34] RAMSEY F. L., SCHAFER D. W., *The Statistical Sleuth. A Course in Methods of Data Analysis*, Belmont CA: Duxbury Press; 1997.

[35] KNUTH D. E., *The Art of Computer Programming*, vol 2. Reading: Addison-Wesley; 1998.

[36] WITZTUM D., RIPS E., ROSENBERG Y., Equidistant Letter Sequences in the Book of Genesis, *Statistical Science* 1994; 9(3):429–438.

[37] DROSNIN M., *The Bible Code* London: Orion; 2001.

[38] LISCHNER R., *Delphi in a Nutshell – A Desktop Quick Reference*, Sebastopol CA: O'Reilly & Associates; 2000.

Received: June, 2004
Accepted: June, 2004

Contact address:

Nikola Štambuk
Paško Konjevoda
Ruđer Bošković Institute
Bijenička cesta 54
HR-10002 Zagreb
Croatia
e-mail: stambuk@irb.hr
pasko@kbsm.hr

Nikola Gotovac
Department of Radiology
General Hospital Požega
Osječka bb
HR-34000 Požega
Croatia
e-mail: nikola.gotovac@zg.htnet.hr

NIKOLA ŠTAMBUK was born in 1959, in Varaždin, Croatia. He received his MD from the Zagreb University School of Medicine in 1984 and completed his internship at Sisters of Mercy Clinical Hospital in Zagreb (1985–87). He spent the period between 1985 and 1987 as resident at the Railway Health Center in Zagreb and got his MS degree from the Zagreb University School of Medicine in 1987. In 1991 he completed doctoral studies and received his PhD from the Institute for Medical Research and Occupational Health, University of Zagreb, where he stayed as a research associate for another year. From 1994 to 2003 he worked at Ruđer Bošković Institute as senior scientific associate. NikolaŠtambuk has published 59 scientific and professional papers. His scientific interests include immunology and immunochemistry, mathematical and artificial intelligence-based modeling in biomedicine.

PAŠKO KONJEVODA was born in 1965, inŠibenik, Croatia. He received his MD from the Zagreb University School of Medicine in 1993 and his MS degree in pharmacology and toxicology from the same faculty in 2003. Since 2000 he has been a statistical editor of the Croatian Medical Journal. Paško Konjevoda has published 34 scientific and professional papers. His scientific interests are focused on computer modeling of the protein structure and function, and on statistical analysis of medical diagnostic and prognostic tests.

NIKOLA GOTOVAC was born in 1973, in Zagreb, Croatia. He got his MD from the Zagreb University School of Medicine in 1999 and completed his internship at General Hospital Požega between 1999 and 2000. Since 2000 he has been a resident at Radiology Department, General Hospital Požega. Nikola Gotovac has published 6 scientific and professional papers. His scientific interests include biocomputing and programming related to protein prediction, pattern recognition and medical imaging.