MIODRAG ŽIVKOVIĆ[1]
SAŠA MALKOV[1*]
SNEŽANA ZARIĆ[2]
MILENA VUJOŠEVIĆ-JANIČIĆ[1]
JELENA TOMAŠEVIĆ[1]
GORAN PREDOVIĆ[1]
NOVICA BLAŽIĆ[1]
MILOŠ V. BELJANSKI[3]

[1]Department of Mathematics, University of Belgrade, Beograd, Serbia and Montenegro

[2]Department of Chemistry, University of Belgrade, Beograd, Serbia and Montenegro

[3]Institute of General and Physical Chemistry, Beograd, Serbia and Montenegro

SCIENTIFIC PAPER

547.96:547.466:519.23

# STATISTICAL DEPENDENCE OF PROTEIN SECONDARY STRUCTURE ON AMINO ACID BIGRAMS

*The statistical dependence of protein secondary structure on amino acid bigram frequencies was studied.*
*Proteins in the PDBSELECT subset of the Protein Data Bank database were investigated. Protein secondary structures were determined using DSSP software. The conditional probabilities of protein secondary structures were calculated and presented. The results on bigrams show the frequencies of all the possible bigrams in all secondary structure types. These results elucidate some factors important for the prediction of the secondary structures of proteins based on the amino acid sequence.*

*Key words: Amino–acid pairs, Protein secondary structure, Bigram, Bigram frequencies.*

Pharmacology and biotechnologies using protein engineering strongly depend on the prediction of the generic protein function and structure, since the protein function is related to its structure. Because detailed protein 3D structure determination is a very costly process, there is extensive development of methods for the prediction of protein structure based on the amino acid sequence. In many of these methods, the first step is the prediction of the protein secondary structure.

There are two approaches to the prediction of protein secondary structure. One is based on the simulation of the natural process of protein folding. The second approach is based on the different methods using known protein structures. The four basic methods [1] are: empirical statistical methods, methods based on physicochemical properties of amino acids, methods based on prediction algorithms and methods using modeling based on force field parameters. It has been known for long time that different amino acids have distinct propensities for the adoption of specific secondary structures [1–6]. Local structural information is often contained in local parts of the sequences [7,8]. It is estimated that local information contains roughly 65% of the secondary structure information [9]. There are also approaches that consider non–local interactions in the sequence [10–13]. Many prediction methods use the frequencies of amino acids in proteins with the known secondary structure [1]. Thefrequencies of amino acid bigrams can be used as a tool for efficient protein comparison and classification [14].

The occurrence of amino acids and amino acid pairs in different positions in α–helices has been studied by Goliaei et al. [15] and Engel et al. [16]. Exceptional amino acid pairs are identified as pairs, the frequencies of which substantially differ from the expected values. Statistical data on the amino acid pair compatibility of both spatially nearest neighbors and adjacent residues was presented by Sen [17]. It was shown that the compatibility of amino acid pairs is quite different in α–helices and β–strands. It was also shown that the propensities of amino acids for certain positions in the helix depend on the physico–chemical properties [16].

In this paper we describe the statistical dependence of protein secondary structure on amino acid bigram frequencies. The possible application of these results is in the refinement of protein secondary structure prediction.

## METHOD

Secondary structure types were assigned by DSSP [18]. They were denoted using the letters: H for α–helix, B for isolated β–bridge, E for extended strand, G for $3_{10}$–helix, I for π–helix, T for hydrogen–bonded turn and S for bend. All other structural elements, not belonging to these secondary structure types, were considered as a coil and denoted by C. Secondary structure types are often reduced to only three; H, E, and C [19,20]. Here we consider all eight secondary structure types, including coils.

Consider a set P of n protein chains. The primary structures of these protein chains are described by the sequences $a_1,\ldots a_n$. If len($a_i$) denotes the length of the sequence $a_i$, then the residues of the sequence $a_i$ are $a_{i,1},\ldots a_{i,len(i)}$, $1 \le i \le n$. The corresponding assigned secondary structures are described by the sequences $b_1,\ldots, b_n$, where $b_i$ is the sequence $b_{i,1},\ldots, b_{i,len(i)}$, $1 \le i \le n$.

*The paper was presented at the 1st South–East European Congress of Chemical Engineering, 25–28 September, Belgrade, Serbia and Montenegro

Author address: S. Malkov, Faculty of Mathematics, University of Belgrade, Studentski Trg 16, 11000 Beograd, Serbia and Montenegro
E–mail: smalkov@matf.bg.ac.yu
Paper received: June 10, 2005
Paper accepted: October 20, 2005

Let $N_{PS2}(p_1,p_2,s_1,s_2)$ denote the number of occurrences of secondary structure bigram $(s_1,s_2)$ built by the amino acid bigram $(p_1,p_2)$; $N_{P2}(p_1,p_2)$ the number of occurrences of amino acid bigram $(p_1,p_2)$; $N_2$ the total number of bigrams; $N_{PS1}(p,s)$ the number of occurrences of secondary structure type $s$ built by the amino acid $p$; $N_{P1}$ the number of occurrences of amino acid $p$; $N_1$ denote the total number of bigrams.

The conditional probabilities of structures $P(s|p)$ and $P(s_1,s_2|p_1,p_2)$ are calculated as

$$P(s|p) = P(sp)/P(p) = (N_{PS1}(p,s)/N_1)/(N_{P1}(p)/N_1) = \\ = N_{PS1}(p,s)/N_{P1}(p) \quad (1)$$

$$P(s_1s_2|p_1p_2) = P(s_1s_2p_1p_2)/P(p_1p_2) = \\ = (N_{PS2}(p_1,p_2,s_1,s_2)/N_2)/(N_{P2}(p_1,p_2)/N_2) = \\ = N_{PS2}(p_1,p_2,s_1,s_2)/N_{P2}(p_1,p_2) \quad (2)$$

## DATA

Our research was based on the Protein Data Bank (PDB) database, release #103 from January 2003, containing 18482 proteins [21]. The secondary structure assignment was performed by the program DSSP [18] (CMBI version by E. Krieger, from April 1, 2000). There are many families of proteins that are overrepresented in the PDB. The full set of protein sequences was filtered to eliminate redundant data – we used the PDBSELECT list of non–redundant protein chains [22], with the threshold 25%. The data set contains 1737 sequences with $N_1$ = 282,329 amino acid residues and $N_2$ = 280592 bigrams.

## RESULTS AND DISCUSSION

The conditional probabilities $P(s|p)$ are computed using (1). The obtained values (multiplied by 100) are listed in Table 1. The sums of the columns are not always 100% because of rounding. The most probable structure(s) for each amino acid (AA) are given in the last row. These results indicate tendency of AA to be part of a certain secondary structure. The data for some AA are in agreement with previous results, while for some AA the data reveal new features that differ significantly from previously published results [23].

The conditional structure probabilities for AA bigrams are given in Table 2. The most probable structure, and its probability multiplied by 100 are presented. Each cell contains the largest conditional probability (multiplied by 100), the most probable structure assignment and the difference between the largest and the second largest conditional probability (also multiplied by 100). The results show that most of the amino acid bigrams have the largest conditional probabilities for helices (HH).

In order to visually emphasize structure pairs different from HH, another presentation of the content is given in Table 3. Pairs of identical letters were replaced by a single letter, and the HH pairs were completely omitted. The values of the conditional probabilities were also omitted. The order of the amino acids was changed, according to their secondary structure preference. Amino acids supporting the formation of $\alpha$–helices were moved to the beginning. The second significant group contains amino acids supporting strands. In this way we obtained two groups of AA. AA in the same group have the common feature that their bigrams tend to support the same secondary structure type. Based on the results obtained, it appears that the occurrence of amino acid bigrams in secondary structures is far from being random. This is in accordance with previous results [15,17], but applied here to all secondary structure bigrams.

For each AA there are at least two different secondary structure bigrams, which are most probable for different primary structure bigrams including the AA. For 69.5% (122 of 400) of the AA bigrams the most probable secondary structure (SS) bigram is HH. For all bigrams of amino acids A, R, Q, E, L, M, K, D, H, W and S, the most probable SS bigram is HH (the corresponding cells in Table 3 are dark gray). Most of these AA are well known as AA that have a large tendency to be part of the helix secondary structure. However, it is interesting that even AA like S, that as single AA have a tendency to build strand structures, prefer to build a helix structure in pairs. For 17% (68 of 400) of the AA bigrams the most probable SS bigram is EE. For all bigrams of amino acids I, Y, T and V, the most probable secondary structure is EE. In this case all of

*Table 1. Conditional structure probabilities for AA monograms. The probabilities are presented as percentage values. The most probable structure asignments for each AA are given in the last row.*

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| H | 45 | 38 | 24 | 26 | 29 | 42 | 43 | 14 | 29 | 35 | 44 | 36 | 40 | 32 | 13 | 25 | 25 | 35 | 30 | 29 |
| E | 16 | 19 | 13 | 11 | 25 | 16 | 15 | 14 | 21 | 37 | 23 | 17 | 21 | 30 | 9 | 18 | 26 | 26 | 31 | 40 |
| T | 10 | 10 | 20 | 16 | 9 | 11 | 12 | 27 | 12 | 4 | 7 | 12 | 7 | 8 | 18 | 12 | 9 | 8 | 8 | 5 |
| B | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 |
| G | 4 | 3 | 4 | 5 | 3 | 4 | 4 | 3 | 4 | 2 | 3 | 4 | 3 | 3 | 5 | 5 | 2 | 5 | 4 | 2 |
| I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| S | 7 | 9 | 14 | 13 | 8 | 9 | 9 | 19 | 11 | 5 | 6 | 10 | 6 | 7 | 11 | 12 | 11 | 7 | 7 | 6 |
| C | 18 | 19 | 25 | 28 | 25 | 18 | 16 | 22 | 22 | 16 | 16 | 20 | 21 | 18 | 43 | 27 | 26 | 18 | 17 | 17 |
| | H | H | CH | CH | H | H | H | T | H | EH | H | H | H | HE | C | CH | ECH | H | EH | E |

*Table 2. Conditional structure probabilities for AA bigrams. The row of the cell is labeled by the first, and the cell column is labeled by the second AA, constituting the bigram, corresponding to the cell. The largest conditional probability, multiplied by 100, and the most probable structure assignment is given in each cell. The second row in each cell shows the difference between the largest and the second largest conditional probability, also multiplied by 100.*

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 55 HH [46] | 48 HH [36] | 34 HH [27] | 35 HH [29] | 33 HH [15] | 52 HH [44] | 56 HH [47] | 19 HH [3] | 33 HH [17] | 48 HH [21] | 55 HH [42] | 49 HH [38] | 49 HH [37] | 42 HH [20] | 19 CC [1] | 32 HH [20] | 31 HH [15] | 49 HH [33] | 40 HH [17] | 42 HH [14] |
| R | 48 HH [38] | 40 HH [27] | 27 HH [20] | 28 HH [20] | 36 HH [18] | 48 HH [38] | 49 HH [40] | 12 HH [1] | 31 HH [19] | 40 HH [11] | 47 HH [30] | 38 HH [27] | 42 HH [28] | 34 HH [11] | 22 CC [7] | 24 HH [13] | 28 HH [9] | 35 HH [9] | 33 HH [7] | 36 HH [4] |
| N | 32 HH [18] | 28 HH [13] | 20 TT [7] | 16 TT [1] | 18 HH [0] | 28 HH [16] | 29 HH [14] | 13 TT [1] | 16 HH [4] | 29 HH [8] | 32 HH [17] | 26 HH [11] | 24 HH [7] | 23 HH [5] | 34 TT [13] | 16 HH [1] | 14 HH [0] | 25 HH [8] | 22 HH [7] | 26 HH [6] |
| D | 32 HH [20] | 35 HH [26] | 16 HH [5] | 20 HH [7] | 19 HH [4] | 32 HH [22] | 33 HH [19] | 16 TC [5] | 27 HH [14] | 26 HH [4] | 35 HH [21] | 27 HH [9] | 29 HH [8] | 21 HH [3] | 29 TT [10] | 18 HH [7] | 16 CC [1] | 25 HH [8] | 23 HH [3] | 23 EE [1] |
| C | 39 HH [23] | 31 HH [14] | 17 HH [4] | 25 HH [15] | 28 EE [5] | 32 HH [22] | 36 HH [16] | 20 HH [5] | 23 HH [8] | 35 EE [1] | 43 HH [21] | 22 HH [4] | 33 HH [9] | 33 EE [9] | 22 CC [11] | 24 HH [10] | 21 HH [0] | 33 EE [2] | 27 HH [2] | 38 EE [12] |
| Q | 50 HH [41] | 46 HH [36] | 29 HH [20] | 34 HH [26] | 31 HH [19] | 54 HH [46] | 53 HH [46] | 15 HH [2] | 34 HH [20] | 46 HH [22] | 47 HH [31] | 48 HH [41] | 42 HH [30] | 38 HH [14] | 20 CC [2] | 29 HH [19] | 31 HH [15] | 43 HH [27] | 39 HH [18] | 36 HH [7] |
| E | 51 HH [42] | 53 HH [45] | 29 HH [22] | 32 HH [24] | 36 HH [19] | 52 HH [43] | 54 HH [47] | 17 HH [5] | 38 HH [29] | 43 HH [17] | 51 HH [37] | 47 HH [38] | 50 HH [39] | 39 HH [15] | 23 HH [4] | 29 HH [21] | 28 HH [16] | 33 HH [12] | 34 HH [11] | 39 HH [10] |
| G | 25 TT [7] | 24 TT [11] | 28 TT [17] | 23 TT [9] | 18 EE [5] | 25 TT [8] | 29 TT [15] | 14 TT [1] | 22 TT [11] | 25 EE [6] | 20 HH [6] | 25 TT [13] | 20 HH [3] | 15 EE [1] | 44 TT [31] | 16 TT [6] | 12 TT [1] | 17 HH [1] | 16 EE [2] | 24 EE [7] |
| H | 34 HH [19] | 34 HH [22] | 25 HH [17] | 23 HH [11] | 26 HH [12] | 39 HH [28] | 34 HH [23] | 15 TC [2] | 26 HH [12] | 31 EE [6] | 37 HH [13] | 32 HH [24] | 34 HH [12] | 29 HH [1] | 21 TT [5] | 20 HH [7] | 23 HH [5] | 30 HH [4] | 31 EE [12] | 39 EE [16] |
| I | 49 HH [24] | 36 HH [6] | 24 HH [4] | 29 HH [14] | 34 EE [4] | 41 HH [13] | 43 HH [18] | 19 EE [2] | 30 HH [0] | 47 EE [13] | 42 HH [7] | 33 HH [4] | 41 HH [11] | 39 EE [4] | 22 CC [6] | 29 HH [2] | 41 EE [15] | 42 HH [17] | 41 EE [12] | 53 EE [28] |
| L | 57 HH [44] | 45 HH [25] | 29 HH [17] | 34 HH [24] | 40 HH [15] | 49 HH [35] | 51 HH [36] | 22 HH [7] | 38 HH [17] | 41 HH [5] | 51 HH [29] | 43 HH [25] | 54 HH [36] | 47 HH [23] | 20 CC [4] | 34 HH [17] | 33 HH [6] | 49 HH [23] | 45 HH [21] | 39 EE [1] |
| K | 48 HH [38] | 42 HH [31] | 27 HH [17] | 26 HH [17] | 33 HH [18] | 45 HH [36] | 47 HH [39] | 10 SC [0] | 28 HH [15] | 41 HH [17] | 46 HH [31] | 36 HH [24] | 41 HH [27] | 28 HH [5] | 22 CC [8] | 24 HH [14] | 26 HH [13] | 35 HH [18] | 28 HH [9] | 32 HH [3] |
| M | 52 HH [38] | 45 HH [33] | 27 HH [17] | 37 HH [29] | 41 HH [20] | 49 HH [38] | 48 HH [33] | 19 HH [1] | 31 HH [18] | 43 HH [8] | 53 HH [33] | 38 HH [22] | 49 HH [33] | 44 HH [23] | 18 HH [0] | 34 HH [16] | 34 HH [12] | 47 HH [23] | 35 HH [9] | 38 HH [3] |
| F | 45 HH [25] | 31 HH [6] | 23 HH [7] | 28 HH [16] | 28 HH [2] | 34 HH [12] | 40 HH [19] | 17 EE [1] | 28 HH [4] | 42 EE [11] | 38 HH [10] | 32 HH [10] | 35 HH [9] | 33 EE [1] | 21 CC [7] | 26 EE [2] | 34 EE [11] | 38 HH [8] | 34 EE [1] | 44 EE [17] |
| P | 32 CC [20] | 32 CC [19] | 20 CT [5] | 28 CT [13] | 25 CT [5] | 28 CC [11] | 28 CC [12] | 19 CC [4] | 21 CT [4] | 34 CC [19] | 32 CC [15] | 32 CC [17] | 29 CC [15] | 27 CC [13] | 46 CC [34] | 20 CC [6] | 27 CC [13] | 23 CC [8] | 19 CC [5] | 30 CC [12] |
| S | 32 HH [20] | 29 HH [16] | 17 HH [7] | 20 HH [9] | 18 HH [2] | 32 HH [19] | 35 HH [25] | 13 CC [3] | 20 HH [5] | 30 EE [2] | 31 HH [14] | 31 HH [20] | 28 HH [9] | 23 EE [3] | 22 CC [5] | 16 HH [3] | 17 HH [0] | 25 HH [4] | 25 EE [4] | 31 EE [11] |
| T | 31 HH [11] | 26 HH [9] | 19 HH [7] | 24 HH [11] | 24 HH [7] | 31 HH [16] | 37 HH [24] | 15 EE [2] | 21 EE [3] | 39 EE [16] | 29 HH [5] | 30 HH [14] | 26 HH [5] | 32 EE [12] | 25 CC [10] | 19 HH [5] | 21 EE [3] | 28 EE [7] | 35 EE [16] | 43 EE [26] |
| W | 46 HH [29] | 35 HH [17] | 16 HH [4] | 27 HH [14] | 33 HH [15] | 39 HH [18] | 41 HH [27] | 18 EE [2] | 32 HH [11] | 38 EE [4] | 45 HH [21] | 35 HH [18] | 39 HH [8] | 39 HH [10] | 16 TT [2] | 26 HH [2] | 27 EE [3] | 28 HH [0] | 34 HH [2] | 46 EE [19] |
| Y | 41 HH [20] | 31 HH [10] | 21 HH [5] | 24 HH [13] | 32 HH [5] | 36 HH [17] | 37 HH [17] | 19 EE [4] | 29 HH [4] | 45 EE [16] | 38 HH [5] | 32 HH [16] | 43 HH [21] | 32 HH [1] | 17 CC [1] | 22 EE [2] | 31 EE [10] | 37 EE [9] | 36 EE [12] | 50 EE [27] |
| V | 43 HH [16] | 36 EE [7] | 22 EE [4] | 25 HH [9] | 36 EE [7] | 32 HH [2] | 34 HH [4] | 19 EE [3] | 32 EE [9] | 49 EE [20] | 39 EE [4] | 36 EE [9] | 38 EE [7] | 42 EE [13] | 21 CC [6] | 29 EE [7] | 44 EE [24] | 36 HH [2] | 43 EE [16] | 55 EE [34] |

*Table 3. Conditional secondary structure inclinations for AA bigrams. For each cell (1) the values of the conditional probabilities are omitted, (2) the pair of identical letters is replaced by a single letter, (3) cells containing HH are left blank. Amino acids supporting the formation of α–helices form the first group (dark gray cells). The second significant group contains amino acids supporting strands (light gray cells).*

| | A | R | Q | E | L | M | K | D | H | W | S | N | C | F | I | Y | T | V | G | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | | | | | | | | | | | C |
| R | | | | | | | | | | | | | | | | | | | | C |
| Q | | | | | | | | | | | | | | | | | | | | C |
| E | | | | | | | | | | | | | | | | | | | | |
| L | | | | | | | | | | | | | | | | | | E | | C |
| M | | | | | | | | | | | | | | | | | | | | |
| K | | | | | | | | | | | | | | | | | | | SC | C |
| D | | | | | | | | | | | | | | | | | C | E | TC | T |
| H | | | | | | | | | | | | | | | E | E | | E | TC | T |
| W | | | | | | | | | | | | | | | E | | E | E | E | T |
| S | | | | | | | | | | | | | | E | E | E | | E | C | C |
| N | | | | | | | | T | | | | T | | | | | | | T | T |
| C | | | | | | | | | | E | | | E | E | E | | | E | | C |
| F | | | | | | | | | | | E | | | E | E | E | E | E | E | C |
| I | | | | | | | | | | | | | E | E | E | E | E | E | E | C |
| Y | | | | | | | | | | E | E | | | | E | E | E | E | E | C |
| T | | | | | | | | | E | E | | | | E | E | E | E | E | E | C |
| V | | E | | | E | E | E | | E | | E | E | E | E | E | E | E | E | E | C |
| G | T | T | T | T | | | T | T | T | | T | T | E | E | E | E | T | E | T | T |
| P | C | C | C | C | C | C | C | CT | CT | C | C | CT | CT | C | C | C | C | C | C | C |

these AA have a tendency to be part of a strand. For most of the bigrams of amino acids C, F, I, Y, T, V and G, the most probable SS bigram is EE (the corresponding cells in Table 3 are light gray). Only for 1.75% (7 of 400) of the AA bigrams the most probable SS bigram is heterogeneous. For 75.8% (288 of 380) of the heterogeneous AA bigrams, the most probable SS bigram is independent of the order of amino acids in the bigram, i.e. the 75.8% of the content of Table 3 is symmetrical (not considering the diagonal elements representing homogenous AA bigrams).

These results show a dependence of all secondary structures on all AA bigrams and elucidate some factors important for protein secondary structure prediction based on the amino acid sequence.

**ACKNOWLEDGEMENT**

This study was supported under projects No 1858 and No 1795 by the Ministry of Science and Environmental Protection of the Republic of Serbia.

**REFERENCES**

[1] A.M. Lesk, Introduction to Protein Science, Oxford University Press, New York (2004) 204–211.

[2] C.A. Kim, J.M. Berg, Nature **362** (1990) 267–270.

[3] D.L. Minor, P.S. Kim, Nature **367** (1994) 660–663.

[4] K.T. O'Neil, W.F. DeGrado, Science **250** (1990) 646–651.

[5] S. Padmanabhan, S. Marqusee, T. Ridgeway, T.M. Laue, R.L. Baldwin, Nature **344** (1990) 268–270.

[6] C.K. Smith, J.M. Withka, L. Regan, Biochem. **33** (1994) 5510–5517.

[7] R.L. Baldwin, G.D. Rose, Trends Biochem Sci **24** (1999) 26–33.

[8] R.L. Baldwin, G.D. Rose, Trends Biochem Sci **24** (1999) 77–83.

[9] B. Rost, in Structural Bioinformatics, P.E. Bourne Ed., Wiley–Liss, Hoboken (2003) 559–587.

[10] D. Frishman, P. Argos, Protein Eng **9:2** (1996) 133–142.

[11] D. Frishman, P. Argos, Proteins **27** (1997) 329–335.

[12] M.M. Gromiha, S. Selvaraj, Progress in Biophysics & Molecular Biology **86:2** (2004) 235–277

[13] A.A. Salamov, V.V. Solovyev, J Mol Biol **268** (1997) 31–36.

[14] C.S. Tsai, An Introduction to Computational Biochemistry, Wiley–Liss, New York (2002) 233–236.

[15] B. Goliaei, M. Zarrin, FEBS Lett. **537** (2003) 121–127.

[16] D.E. Engel, W.F. DeGrado, J.Mol.Biol. **337:5** (2004) 1195–1205.

[17] S. Sen, Biophys. Chem. **103** (2003) 35–49.

[18] W. Kabsch, C. Sander, Biopolymers **22:12** (1983) 2577–2637.

[19] B. Rost, J.Struc. Biol. **134** (2001) 204–218.

[20] A. Kloczkowski, K.L. Ting, R.L. Jernigan, J. Garnier, Proteins **49** (2002) 154–166.

[21] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucl. Acids. Res. **28:1** (2000) 235–242

[22] U. Hobohm, C. Sander, Protein Sci. **3** (1994) 522–524.

[23] P.Y. Chou, G.D. Fasman, Biochemistry **13** (1974) 211–222.

## IZVOD

STATISTIČKA ZAVISNOST SEKUNDARNE STRUKTURE PROTEINA OD FREKVENCIJE BIGRAMA AMINOKISELINA

(Naučni rad)

Miodrag Živković[1], Saša Malkov[1*], Snežana Zarić[2], Milena Vujošević–Janičić[1], Jelena Tomašević[1], Goran Predović[1], Novica Blažić[1], Miloš V. Beljanski[3]
[1]Matematički fakultet Univerziteta u Beogradu
[2]Hemijski fakultet Univerziteta u Beogradu
[3]Institut za opštu i fizičku hemiju, Beograd

U radu je opisana statistička zavisnost sekundarne strukture proteina od bigrama amino kiselina. Istraživanje je uradjeno na proteinima koji se nalaze u podskupu PDBSELECT baze podataka o proteinima Protein Data Bank (PDB). Sekundarne strukture proteina su određene primenom programa DSSP. Izračunate su i prikazane uslovne verovatnoće pojavljivanja različitih sekundarnih struktura. Rezultati za bigrame su prikazani za sve tipove sekundarnih struktura. Rezultati ukazuju na faktore koji su značajni za predviđanje sekundarne strukture proteina koje je zasnovano na redosledu amino kiselina u proteinu.

Ključne reči: Parovi amino kiselina, Sekundarena struktura proteina, Bigram, Učestalost bigrama.