Data Article

Data on the application of the molecular vector machine model: A database of protein pentafragments and computer software for predicting and designing secondary protein structures

Vladimir Karasev

St. Petersburg State Electrotechnical University, Prof. Popov str. 5, 197376, St. Petersburg, Russia

ARTICLE INFO

Article history:
Received 15 March 2019
Received in revised form 7 November 2019
Accepted 7 November 2019
Available online 19 November 2019

Keywords:
Molecular vector machine
Database of protein pentafragments
Software for predicting and design the secondary protein structure

ABSTRACT

Based on ideas about the molecular vector machine of proteins [1], a database of protein pentafragments has been created and algorithms have been proposed for predicting the secondary structure of proteins according to their primary structure and for designing the primary protein structure for a given secondary structure that it takes on. A comprehensive software suite (Predicto @ Designer) has been developed using the pentafragments database and the said algorithms. For the proteins used to create the pentafragments database, a high accuracy (close to 100%) in predicting the secondary protein structure as well as good prospects for its use for designing secondary structures of proteins have been demonstrated.

© 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Data

In this paper, software is described based on the model [1]. The process of predicting secondary protein structure described in the patent [2]. An example of prediction result is given in Table 1, A (a fragment of porcine myoglobin [3]). This fragment illustrates that the whole fragment under consideration can be predicted as a sequence of 10-digit numbers. The comparison with structured experimental data [4], visualized with “Protein 3D” software [5], proved that the software predicts this structure correctly (Fig. 1).

Correction of prediction. Since our approach uses digital description of pentafragment conformations, replacement of a single amino acid has an impact on prediction accuracy, which is a disadvantage of this method. In this situation, if some pentafragment is missing in the database for any reason, a gap in the structure is predicted, which is clearly seen in Table 1, A on the example of alligator’s myoglobin fragment [5]. However, this disadvantage can be rectified by employing correction methods that we have developed [6]. A method for replacement of amino acids is the most interesting among them (See below).

The results given by this method are shown on the example of alligator myoglobin, whose primary structure was determined by Ref. [7]. Whereas the results in the middle column in Table 1, A on the example of alligator’s myoglobin fragment [5]. However, this disadvantage can be rectified by employing correction methods that we have developed [6]. A method for replacement of amino acids is the most interesting among them (See below).

The results given by this method are shown on the example of alligator myoglobin, whose primary structure was determined by Ref. [7]. Whereas the results in the middle column in Table 1, to which correction was not different amino acids in i-th position, then it is possible to replace the original pentafragment search with the search for pentafragment with similar structure but with amino acid changed in i-th position.

2. Experimental design, materials, and methods

2.1. Creating the database of protein pentafragments

Text files describing hydrogen bonds in the secondary structure of proteins were obtained on the basis of about 2333 PDB-files of the Protein Data Bank (subunits — 2446). The list of proteins is given in the appendix. With the help of the Protein 3D program developed by us [5] (the program is free to
Table 1
Predicting secondary myoglobin structure without correction (A) and with correction based on the replacement of amino acids in pentafragments (B).

| Pig (Pig without correction.dbkx) | Alligator (ALLIGAT without correction.dbkx) | Alligator (ALLIGAT amino acid correction.dbkx) |
|---------------------------------|-----------------------------------------------|-----------------------------------------------|
| 141 XXX D Asp 1111111111       | 142 XXX D Asp 1111111111                     | 142 XXX D Asp 1111111111                     |
| 140 XXX N Asn 1111111111       | 141 XXX N Asn 1111111111                     | 141 XXX N Asn 1111111111                     |
| 139 XXX R Arg 1111111111       | 140 XXX R Arg 1111111111                     | 140 XXX R Arg 1111111111                     |
| 138 XXX F Phe 1111111111       | 139 XXX F Phe 1111111111                     | 139 XXX F Phe 1111111111                     |
| 137 XXX L Leu 1111111111       | 138 XXX L Leu 111111111121                   | 138 XXX L Leu 1111111111                     |
| 136 XXX E Glu 1111111111       | 137 XXX E Glu 1111111111                     | 137 XXX E Glu 1111111111                     |
| 135 XXX L Leu 1111111111       | 136 XXX L Leu 1111111111                     | 136 XXX L Leu 1111111111                     |
| 134 XXX A Ala 1111111111       | 135 XXX A Ala 1111111111                     | 135 XXX A Ala 1111111111                     |
| 133 XXX K Lys 1111111111       | 134 XXX K Lys 1111111111                     | 134 XXX K Lys 1111111111                     |
| 132 XXX S Ser 11111111111      | 133 XXX R Arg 11111111111 ASN                | 133 XXX R Arg 11111111111 ASN                |
| 131 XXX M Met 11111111101      | 132 XXX M Met 11111111101                    | 132 XXX M Met 11111111101                    |
| 130 XXX A Ala 1111110101       | 131 XXX A Ala 1111110101                     | 131 XXX A Ala 1111110101                     |
| 129 XXX G Gly 11101010101      | 130 XXX A Ala 11101010101                   | 130 XXX A Ala 11101010101                   |
| 128 XXX Q Gin 110101010101     | 129 XXX Q Gin 110101010101                  | 129 XXX Q Gin 110101010101                  |
| 127 XXX A Ala 01010101010      | 128 XXX S Ser 01010101010                   | 128 XXX S Ser 01010101010                   |
| 126 XXX D Asp 010101010100     | 127 XXX D Asp 010101010100                  | 127 XXX D Asp 010101010100                  |
| 125 XXX A Ala 01010100000      | 126 XXX A Ala 01010100000                   | 126 XXX A Ala 01010100000                   |
| 124 XXX G Gly 01100000010      | 125 XXX G Gly 01100000010                    | 125 XXX G Gly 01100000010                    |
| 123 XXX F Phe 100000101111     | 124 XXX F Phe 100000101111                   | 124 XXX F Phe 100000101111                   |
| 122 XXX D Asp 000010111011     | 123 XXX D Asp 000010111011                   | 123 XXX D Asp 000010111011                   |
| 121 XXX G Gly 00101101010      | 122 XXX A Ala 02000000000                    | 122 XXX A Ala 00001210100 GLY               |
| 120 XXX P Pro 10110101011      | 121 XXX P Pro 00000000000                    | 121 XXX P Pro 00121010100                    |
| 119 XXX H His 11010101111      | 120 XXX Y Tyr 12101010111 HIS               | 120 XXX Y Tyr 12101010111 HIS               |
| 118 XXX K Lys 10101111111      | 119 XXX K Lys 00000000000                     | 119 XXX K Lys 1010101111 ARG                |
| 117 XXX S Ser 10111111111      | 118 XXX E Glu 10101111111 SER               | 118 XXX E Glu 10101111111 SER               |
| 116 XXX Q Gin 11111111111     | 117 XXX A Ala 0000000000                      | 117 XXX A Ala 1011111111 HIS                |
| 115 XXX L Leu 11111111111     | 116 XXX Ile 11111111111 LEU                 | 116 XXX Ile 11111111111 LEU                 |
| 114 XXX V Val 11111111111     | 115 XXX V Val 11111111111                    | 115 XXX V Val 11111111111                    |

Bold indicate substitutions of amino acids in the polypeptide chain at which the prediction in column B occurs. The substituted amino acids used are shown in this column to the right.

Fig. 1. Fragment 114–141 of the polypeptide chain of porcine myoglobin [4].
download), these files were processed in a step-by-step fashion using mini-programs with a view to obtaining and sorting pentafragments. The steps are listed below.

2.1.1. Obtaining text files
Open the source PDB file using the Protein 3D program. The Rendering icon in the CHBBS settings submenu will show us the type of protein with a specification of its hydrogen bond systems. Next, in the CHBBS icon, check the box against the line item named Trace in memory. Open the bond types table from the Select bond types line item using the dropdown arrow, check the boxes against the N…Oi–3 and Ni…Oi–4 bonds, and uncheck the Show all line item. Next, click on the Show selected bonds line item and click OK. This will open a window with information about the H-bonds of the protein. After clicking the Save links button, we will get a text file with a description of these links. Table 2, A shows a sample fragment from a 1MWC text file (Sus scrofa myoglobin).

2.1.2. Inverting text files
For the Predicto @ Designer program to work, the amino acid sequences contained in our pentafragments database need to be written from bottom to top. This pattern simulates the protein synthesis process, which evolves from the N-end to the C-end. The Invertor program takes the data written in the text file and rearranges them from the bottom up (Table 2, B).

2.1.3. Cutting text files into pentafragments
Using the cutter_u program, cut the inverted files into pentafragments that will store information about the arrangement of H-bonds. Cutting is done by shifting the frame by one amino acid. Table 2, C shows some examples of such pentafragments.

2.1.4. Sorting and simplifying pentafragments
Use the Selector program to sort the pentafragments obtained as shown above in accordance with the link encoding system we have adopted (see Tables 3 and 4). Use the Simplification program to simplify the files obtained (Table 2, D).

An identification system was developed to sort pentafragments in database folders based on the binary coding of H-bonds [8–11]. An example of describing the structure of pentafragments with the help of implemented coding is given in Table 3. In this case, the 10-digit numbers describing a conformation of pentafragments were transferred to the file names (Table 3, E).

Subsequently, this coding procedure became more complicated (Table 4). Additional figures to identify various types of secondary structures were introduced, but retained its binary principle [11]. The structure of the database organized in accordance with the link encoding system as per Table 4 is shown in Table 5. It consists of folders containing pentafragment files and designated by the i-th pair of variables (see the Folder numbering column, Table 5), of files enclosed in these folders and containing 10-digit numbers that describe the structure of the pentafragments (column 2), and of pentafragments contained in these files and associated to their specific positions in proteins (column 3). To speed up the search for pentafragments, the software has the database written in the form of strings (see Ref. [6] for an example).

2.2. Program layout

The computer program named PREDICTO @ DESIGNER The program is written in C++. It has been registered [12] as well as described in detail in Ref. [13]. For the program, a file of the.pdb format (Protein Data Bank) and.gen (Genbank) can be used, which are transformed by the program into the.dbk format (Table 6, A) in which the program predicts the secondary structure of the protein. The result of the program is written in.dbkx format (Table 6, B).

Fig. 2, a shows the startup screen of the PREDICTO @ DESIGNER program. Clicking on the word PREDICTO sets the program to the secondary protein structure prediction mode (Fig. 2, b shows the workspace where digital and structural information is displayed) and clicking on the word DESIGNER sets it to the design mode (Fig. 2, c shows the workspace, control panel, and icons used to display information required for the design).
2.3. The procedure for prediction

The method of predicting secondary protein structure described in the patent [2] consists in isolating pentafragments in a file with specially formatted primary structure of proteins (files.dbk) and their search in the Database. Since every pentafragment has a 10-digit identification number in the Database, the software reads the code number of the found pentafragment and displays it onto the numeric operating field in a bottom-up sequence progressively as pentafragments are selected in a protein chain from start to finish. This procedure consists of two stages: an initial pentafragment is

| A | B | C | D |
|---|---|---|---|
| Fragment from a text file (1MWD text file.txt) | Fragment from an inverted text file (inv_1MWD inverted text file.txt) | Examples of pentafragments obtained by cutting (rezfile cutting.txt) | Example of simplified file (sim_2111111211.txt) |
| 114 VAL | 125 ALA O - 129 GLY N | 1MWC | 1THB |
| 114 VAL N - 110 ALA O | 125 ALA | 120 PRO N - 116 GLN O | 136 GLY |
| 114 VAL O - 118 LYS N | 124 GLY O - 128 GLN N | 120 PRO | 135 ALA |
| 115 LEU | 124 GLY | 119 HIS O - 123 PHE N | 134 VAL |
| 115 LEU N - 111 ILE O | 123 PHE N - 119 HIS O | 119 HIS N - 115 LEU O | 133 VAL |
| 115 LEU O - 119 HIS N | 123 PHE | 119 HIS | 132 LYS |
| 116 GLN | 122 ASP | 118 LYS N - 114 VAL O | 1HDS |
| 116 GLN N - 112 ILE O | 121 GLY | 118 LYS | 134 ALA |
| 116 GLN O - 120 PRO N | 120 PRO N - 116 GLN O | 117 SER N - 113 GLN O | 133 VAL |
| 117 SER | 120 PRO | 117 SER | 132 VAL |
| 117 SER N - 113 GLN O | 119 HIS O - 123 PHE N | 116 GLN O - 120 PRO N | 131 LYS |
| 118 LYS | 119 HIS N - 115 LEU O | 116 GLN N - 112 ILE O | 130 GLN |
| 118 LYS N - 114 VAL O | 119 HIS | 116 GLN | 1THB |
| 119 HIS N - 115 LEU O | 118 LYS | 119 HIS O - 123 PHE N | 69 ALA |
| 119 HIS O - 123 PHE N | 117 SER N - 113 GLN O | 119 HIS N - 115 LEU O | 68 ASN |
| 120 PRO | 117 SER | 119 HIS | 67 THR |
| 120 PRO N - 116 GLN O | 116 GLN O - 120 PRO N | 119 HIS | 66 LEU |
| 121 GLY | 116 GLN N - 112 ILE O | 118 LYS N - 114 VAL O | 65 ALA |
| 122 ASP | 116 GLN | 118 LYS | 1AZI |
| 123 PHE | 115 LEU O - 119 HIS N | 117 SER N - 113 GLN O | 67 VAL |
| 123 PHE N - 119 HIS O | 115 LEU N - 111 ILE O | 117 SER | 66 THR |
| 124 GLY | 115 LEU | 116 GLN O - 120 PRO N | 67 VAL |
| 124 GLY O - 128 GLN N | 114 VAL O - 118 LYS N | 116 GLN N - 112 ILE O | 66 THR |
| 125 ALA | 114 VAL N - 110 ALA O | 116 GLN | 65 GLY |
| 125 ALA O - 129 GLY N | 114 VAL | 115 LEU O - 119 HIS N | 64 HIS |
| | | 115 LEU N - 111 ILE O | 63 LYS |
| | | 115 LEU | 115 LEU |

1MWC
118 LYS N - 114 VAL O
118 LYS
117 SER N - 113 GLN O
117 SER
116 GLN O - 120 PRO N
116 GLN N - 112 ILE O
116 GLN
115 LEU O - 119 HIS N
115 LEU N - 111 ILE O
115 LEU
114 VAL O - 118 LYS N
114 VAL N - 110 ALA O
114 VAL

}|
It has been found that when applying this approach, the secondary structure of all proteins used to develop the database is predicted with an accuracy close to 100%.

2.4. Prediction correction method by replacement of amino acids

The method consists in the following [6]. Let us assume that at some i-th stage the software has isolated a pentafragment to be searched for that has not been found under a code number defined on the basis of search algorithm. If this pentafragment could be found at the previous i-1-th stage, then it is all about the amino acid that appeared in the pentafragment at the i-th stage. It is well known that these changes (mutations) are frequently observed for the same type proteins but extracted from different kinds of organisms. Because the search for pentafragment with missing i-th amino acid should be conducted under the same folder number, as for the other pentafragments with similar structure but with applied, show quite low prediction accuracy, a region with amino acids from 115 to 138 (Table 1) was completely predictable as a result of applying this method. Comparison of the predicted structure of alligator myoglobin with porcine myoglobin (Table 1, left column) shows that in

| A. Notations in text PDB-files | B. Types of H-bonds | C. Coding | D. An example of pentafragment and its coding |
|-------------------------------|---------------------|-----------|---------------------------------------------|
| X1,X2 Abs                     | **No H-bonds**      | 00        | **51 Gln O - 55 Glu N 01**                  |
| X1,X2 Abs O–Y1,Y2 Deh N       | **H-bond only with C—O-group** | 01 | 50 Pro 00                                  |
| X1,X2 Abs                     | **H-bond only with NH-group** | 10 | 48 Asp 00                                 |
| X1,X2 Abs N–Y3,Y4 Ehf O       | **H-bonds both with C—O and with NH-group** | 11 | E. 10-digit descriptions of PFs and file names |
| X1,X2 Abs                     | **NH-group**        |           |                                             |

In cell D, the selected first two lines correspond to the highlighted designation 01 in cell E.

| N²  | Types of H-bonds | Binary Combinations |
|-----|------------------|---------------------|
|     | Bonds   | Code | Bonds   | Code | Bonds   | Code |
| a-helix |         |      |     |       |       |     |       |     |
| 1.   | N₄H ... O₁₄   | 0    | 00   | 01   | 1     | 10   | 1     | 11   |
|     | O₁₄ ... HN₄   | 0    | 1    | 0    | 1     |      |       |      |
| Inverted a-helix |       |      |     |       |       |     |       |      |
| 2.   | N₄H ... O₁₃   | 0    | 00   | 70   | 0     | 07   | 1     | 77   |
|     | O₁₃ ... HN₄   | 0    | 0    | 1    | 1     |      |       |      |
| helix 3₁₀ |       |      |     |       |       |     |       |      |
| 3.   | N₄H ... O₁₃   | 0    | 00   | 03   | 1     | 30   | 1     | 33   |
|     | O₁₃ ... HN₄   | 0    | 1    | 0    | 1     |      |       |      |
| Inverted helix 3₁₀ |       |      |     |       |       |     |       |      |
| 4.   | N₄H ... O₁₃   | 0    | 00   | 60   | 0     | 06   | 1     | 66   |
|     | O₁₃ ... HN₄   | 0    | 0    | 1    | 1     |      |       |      |
| Combination of a-helix and helix 3₁₀ |       |      |     |       |       |     |       |      |
| 5.   | N₄H ... O₁₃ ... O₁₃ | 0  | 00   | 02   | 2     | 20   | 2     | 22   |
|     | O₁₃ ... HN₄ ... HN₄ | 0  | 2    | 0    | 2     |      |       |      |
| Combination of Inverted a-helix and helix 3₁₀ |       |      |     |       |       |     |       |      |
| 6.   | N₄H ... O₁₃ ... O₁₃ | 0  | 00   | 40   | 0     | 04   | 2     | 44   |
|     | O₁₃ ... HN₄ ... HN₄ | 0  | 0    | 2    | 2     |      |       |      |

found at the first stage and if it is detected correctly then the remaining protein is predicted further at the second stage [2]. It has been found that when applying this approach, the secondary structure of all proteins used to develop the database is predicted with an accuracy close to 100%.

2.4. Prediction correction method by replacement of amino acids

The method consists in the following [6]. Let us assume that at some i-th stage the software has isolated a pentafragment to be searched for that has not been found under a code number defined on the basis of search algorithm. If this pentafragment could be found at the previous i-1-th stage, then it is all about the amino acid that appeared in the pentafragment at the i-th stage. It is well known that these changes (mutations) are frequently observed for the same type proteins but extracted from different kinds of organisms. Because the search for pentafragment with missing i-th amino acid should be conducted under the same folder number, as for the other pentafragments with similar structure but with applied, show quite low prediction accuracy, a region with amino acids from 115 to 138 (Table 1) was completely predictable as a result of applying this method. Comparison of the predicted structure of alligator myoglobin with porcine myoglobin (Table 1, left column) shows that in

Table 3

Notations of bonds in text PDB-files (A), types of H-bonds (B), their coding with Boolean pairs of variables (C), an example of pentafragment (D) and its 10-digit description (E).

Table 4

Coding of types of H-Bonds in the form of binary combinations for an improved database of pentafragments.
general both structures have similar position of α-helixes in this fragment. Thus, applying this correction method significantly improves prediction accuracy for secondary structure of proteins.

Table 5
Pentafragment database structure.

| No. | Folder | No. | Folder |
|-----|--------|-----|--------|
| 1   | 00-XX  | 20  | 30-XX  |
| 2   | 01-XX  | 21  | 31-XX  |
| 3   | 02-XX  | 22  | 32-XX  |
| 4   | 03-XX  | 23  | 33-XX  |
| 5   | 04-XX  | 24  | 34-XX  |
| 6   | 06-XX  | 25  | 36-XX  |
| 7   | 07-XX  | 26  | 37-XX  |

Folder 37-XX
(Pentafragment Files of Folder 37-00.JPG)

Pentafragments of the file 3730000373.txt
(Pentafragment of File 3730000373.JPG)

Table 6
 Formats used by the program PREDICTO @ DESIGNER.

| A                  | B                  |
|--------------------|--------------------|
| A fragment of the pig myoglobin protein | Recording the result of the program in.dbkx format |
| (1MWC file) in.dbk format | (1MWD_A.dbkx) |
| 15 XXX G GLY bbbbbbbbb | 15 XXX G GLY 1112121011 3K9Z 1DMR |
| 14 XXX W TRP bbbbbbbbb | 14 XXX W TRP 1212101111 3K9Z 1DMR |
| 13 XXX V VAL bbbbbbbbb | 13 XXX V VAL 1210111111 3K9Z 1MWC |
| 12 XXX N ASN bbbbbbbbb | 12 XXX N ASN 1011111111 3K9Z 1MWC |
| 11 XXX L LEU bbbbbbbbb | 11 XXX L LEU 1111111111 3K9Z 1MWC |
| 10 XXX V VAL bbbbbbbbb | 10 XXX V VAL 1111111101 3K9Z 1MWC |
| 9 XXX L LEU bbbbbbbbb | 9 XXX L LEU 1111110101 3K9Z 1MWC |
| 8 XXX Q GLN bbbbbbbbb | 8 XXX Q GLN 1111010101 3K9Z 1MWC |
| 7 XXX W TRP bbbbbbbbb | 7 XXX W TRP 1101010101 3K9Z 1DMR |
| 6 XXX E GLU bbbbbbbbb | 6 XXX E GLU 0101010100 3K9Z 1DMR |
| 5 XXX G GLY bbbbbbbbb | 5 XXX G GLY 0101001000 3K9Z 1DMR |
| 4 XXX D ASP bbbbbbbbb | 4 XXX D ASP bbbbbbbbb |
| 3 XXX S SER bbbbbbbbb | 3 XXX S SER bbbbbbbbb |
| 2 XXX L LEU bbbbbbbbb | 2 XXX L LEU bbbbbbbbb |
| 1 XXX G GLY bbbbbbbbb | 1 XXX G GLY bbbbbbbbb |
| 0 ATG M MET bbbbbbbbb | 0 ATG M MET bbbbbbbbb |
2.5. Further ways to develop the prediction method

Applying the described prediction correction method is convenient and relevant to use for the groups of proteins with similar structure but derived from different species (as in cases with myoglobin and other heme-containing proteins). Ideally, it would be better to have a universal database that could be used to predict secondary structure of any protein with high accuracy. We have shown a practical possibility for creating it [14]. However, a high increase in the number of pentafragments in the database significantly increases the number of alternative options for prediction of secondary structures. This, in its turn, sharply slows down software performance and deteriorates the prediction quality.

Due to the above-mentioned reasons, we believe it is more relevant to develop ad-hoc databases aimed at predicting structurally close proteins. In this case, a universal database can be built on the basis of hierarchical structure of specialized databases. A prediction algorithm will consist of two
stages: a) preliminary search of common elements being attributable to certain protein groups; b) final prediction based on a specialized database. There is a lot of work to be done in this respect, but the results of this work seem to be quite promising.

2.6. Developing a design method for secondary structures

Because the proposed approach can predict secondary structures of proteins quite accurately, it would be logical to apply the same approach to design secondary structures based on the predefined secondary structure. This method is detailed in the patent of [15]. It is implemented in the Designer section [13] of the Predicto @ Designer software. The initial protein pentafragment and its description in the form of 10-digit number in the binary numeral system is set using the control panel. The selected pentafragment is searched for in the database and, if it is found, then it is necessary to see one new amino acid and 10-digit description of a new pentafragment containing the previous four amino acids and one new and run a new search in the database. If the new pentafragment is found, then the procedure should be repeated.

The description presented in the patent is based on the data available in literature, and therefore, it confirms the feasibility of this design. However, before this method is recommended for a large-scale implementation, it must pass a more comprehensive experimental validation on the basis of up-to-date scientific and engineering know-how. The studies are being carried out in this respect.

Acknowledgments

We are grateful to V.V. Luchinin for useful discussion of the paper and S.B. Kalinin for preparing the program.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104815.

References

[1] V.A. Karasev, A model of molecular vector machine of proteins, BioSystems 180 (2019) 7–18. https://DOI10.1016/j.biosystems.2019.02.001.
[2] V.A. Karasev, V.V. Luchinin, A Method of Predicting Secondary Structure of Protein, R F Patent No.2425837 date of publ. 10.08.2011, Bull. No.22 (In Russian).
[3] E. Akaboshi, Cloning and sequence analysis of porcine myoglobin cDNA, Gene 40 (1985) 137–140. https://DOI10.1016/0378-1119(85)90033-2.
[4] S. Krzywda, G.N. Murshudov, A.M. Brzozowski, M. Jaskolski, E.E. Scott, S.A. Klizas, Q.H. Gibson, J.S. Olson, A.J. Wilkinson, Stabilizing bound O2 in myoglobin by valine68 (e11) to asparagine substitution, Biochemistry 37 (1998) 15896–15907. https://DOI10.1021/bi9812470.
[5] E.L. Demchenko, V.A. Karasev, Protein 3D – the Visualizer of Supramolecular Biotissues, 2017. http://protein-3d.ru/.
[6] V.A. Karasev, S.B. Kalinin, PREDICTO @ DESIGNER computer software for prediction and design of protein secondary structures: UPGRADE. III. Algorithms for searching pentafragments in databases and correction methods for predicting secondary structures of proteins, Biotechnosfera 2 (2016) 39–48 (In Russian).
[7] H. Dene, J. Sazy, M. Goodman, A.E. Romero-Herrera, The amino acid sequence of alligator (Alligator mississippiensis) myoglobin. Phylogenetic implications, Biochim. Biophys. Acta 624 (1980) 397–408. https://DOI10.1016/0005-2795(80)90081-1.
[8] V.A. Karasev, Principles of Topological Coding of Chain Polymers and Structure of Proteins, SPB ETU “LETI”, Saint-Petersburg, 2014 (In Russian).
[9] V.A. Karasev, V.E. Stefanov, 10-digits boolean system in description of protein pentafragments, Symmetry: Sci. Cult. 24 (2013) 275–293.
[10] V.A. Karasev, A.I. Belyaev, V.V. Luchinin, Database of Protein Pentafragments, Registered in ROSPAPENT No. 2010620364, 2010. (In Russian).

[11] V.A. Karasev, S.B. Kalinin, PREDICTO @ DESIGNER computer software for prediction and design of protein secondary structures: UPGRADE. I. Database of protein pentafragments considering N-H…O-3, N-H…O-4, and other types of H-bonds in secondary structures of proteins, Biotechnosfera 1 (2016) 49–55 (In Russian).

[12] S.B. Kalinin, V.A. Karasev, V.V. Luchinin, Software to Predict Secondary Protein Structure and Design Primary Protein Structure with Defined Secondary Structure (Predicto@Designer). Registered in ROSPATENT, No.201562295, dated 17.02.2015. (In Russian).

[13] V.A. Karasev, S.B. Kalinin, PREDICTO @ DESIGNER computer software for prediction and design of protein structures: theory. Design. Application, Biotechnosfera 3–4 (2016) 38–48 (In Russian).

[14] V.A. Karasev, S.B. Kalinin, PREDICTO @ DESIGNER computer software for prediction and design of protein secondary structures: UPGRADE. II. Principles of developing theoretical database of protein pentafragments, Biotechnosfera 2 (2016) 29–38 (In Russian).

[15] V.A. Karasev, V.V. Luchinin, A Method of Designing Primary Structure of Protein with Specified Secondary Structure, RF Patent No.2511002, date of publ. 10.04.2014, Bull. No.10 (In Russian).