Discrete modeling of structure preserving mutations in proteins [*]

Pavel S. Demenkov, Evgeny Y. Kharlamov
demps@math.nsc.ru, kharlamov@math.nsc.ru
Novosibirsk State University

Abstract

We constructed logical-probabilistic models determined structure preserving mutations in proteins. For model generation we used Neural Networks and modified method KRAB. As a data source for the algorithms we used physical-chemical and others proteins' characteristics as well as proteins free energy's changing under mutation. Algorithms were implemented. Designed software is able to predict allowable protein's mutations. In the paper it is also presented a comparison between the used methods. It is shown that algorithm KRAB makes more accurate predictions with respect to used data source.

Introduction

Artificial proteins mutation is a relatively old problem in proteinomics[1].

Researches in this area aimed to make new proteins with certain given characteristics. At present the problem is solved experimentally. That means there are experts who make substitutions and then look at the result. If experts are not satisfied with the result (e.g. protein changes its structure after mutation), they will make a new experiment. The main disadvantage of the approach is high cost of experiments. So, there is a need to develop mathematical methods and to design software, whose are able to predict results of the concrete experiment. Moreover, the software should be able to predict experiments, those lead to proteins with the given characteristics.

In the current paper we designed mathematical models and software to predict the best protein mutations, i.e. mutations those lead to required protein with high probability and preserve protein's structure.

As a data source we used Protein Data Bank (PDB). In the data base there were presented functions and structure of each protein.

[*] Работа поддержана грантом НШ-2112.2003.1

[1] Under the term mutation we understand substitutions of some amino acids in linear protein's structure.

We accepted a hypothesis that the main fluency on amino acid's type in a certain protein's position make amino acids those are in the nearest position's neighborhood. To implement the hypothesis we made software transforms information from PDB into special data tables. The main idea of the designed algorithm was to construct a sphere round each protein's position. The sphere's radius is equal to ten angstrom. Then we listed characteristics of amino acids captured by sphere and characteristics of the amino acid in the sphere's center. Each line in the table corresponds to the list of characteristics of amino acids those are in one sphere. Another data source for modeling was information about free energy' changes in proteins [14].

Mathematical description of the problem

We describe each atom in a protein as an element of a set $M \subseteq V \times A \times St \times \mathbb{R}^3$. Where V is a three dimensional vector space, A is a set of amino acids' types, St is a set of protein secondary structures' types, $\mathbb{R}$ is a set of real numbers. So, we describe a protein as a structure $\mathbb{S} =< M, Nb >$, where Nb is a relation defines connection between atoms. We define it as following. Let x,y be atoms of the same protein

$$Nb(x,y) = \begin{cases} t, & \text{if x and y are chemical bonded;} \\ f, & \text{else.} \end{cases}$$

We transform structure $\mathbb{S}$ into structure

$$\mathbb{S}_1 =< M_1, Nb_1 >,$$

where M_1 is homeomorphic image of M, M_1 contains only C_α aatoms of all protein's amino acids ($M_1 \subseteq M$). Relation Nb is always false if we consider any couple of elements from M_1. Therefore, we transform it into relation Nb_1. The Nb_1 differs from Nb in the way that $Nb_1(x,y)$ is true for elements from connected amino acids. Now we will define relation $Sur(x,y)$. The relation helps us to determine elements whose are close.

Let x, y - atoms of the same protein.

$$Sur(x,y) = \begin{cases} t, & \text{if the distance between } x \text{ and } y \leq 10\text{Å;} \\ f, & \text{else.} \end{cases}$$

We will add this relation to the structure $\mathbb{S}_1$ and obtain new structure $\mathbb{S}_2 =< M_1, Nb_1, Sur >$. Now we are ready to define the notion 'neighborhood' or 'surroundings'.

Let $S_x = \{y \mid Sur(x,y)\}$ be a sphere for the element $x \in M_1$. Let $N_x = S_x \setminus \{x\}$ be surroundings of the element $x \in M_1$. Let x be a center for S_x and N_x. We reduced the initial problem to the problem of determining regularities between a center x and its surroundings N_x. For getting a solution for the problem we will consider a structure

$$\mathbb{S}_3 =< M_2, Center(x, N_y) > .$$

Relation $Center(x, N_y)$ is equal to true, when the center corresponds to the surroundings and equal to false otherwise. We assume that the center x corresponds to the surroundings N_x if there is exist a sphere S_x, such that x is sphere's center and $N_x = S_x \setminus \{x\}$. $M_2 = \{M_1, \{N_x \mid x \in M_1\}\}$ So, we reduced the problem to the problem of constructing of the relation $Center$.

Therefore, using data about spheres and their centers in existence collected in PDB we are to find correlation between centers and their surroundings, i.e. to find relation $Center$. In order to do it we represent elements from the basic set of the structure M_2 as lists. The reason is that lists are convenient for processing using data analysis methods. In PDB each element from M_2 is represented in inconvenient way. For this purpose we designed software that prepares data from PDB for the future analysis.

Using data obtaining by PDB transforming program and amino acids' properties we construct five tables. In the tables there are different combinations of the following properties: amino acids' type; distance between amino acids from surroundings and central one; physical-chemical characteristics of amino acids from surroundings; set of independent Kidera's physical-chemical characteristics [12]; set of independent physical-chemical characteristics [12]; average values of physical-chemical characteristics and type of central amino acid, those are coded in a specific way (for getting average we decompose sphere in 3 parts). Different lines in tables correspond to different spheres.

Obtained tables are divided in two parts, i.e. learning sample and control sample. We will find relation $Center$ in the following form:

$$Center(x, N_y) = \begin{cases} t, & x \simeq F(N_y); \\ f, & \text{else.} \end{cases}$$

Thus, the last contraction of the initial problem is to find the function F. The function is able to determine allowable type of central amino acid using surroundings' properties. It is possible that F is many-valued function. Expression $x \simeq F(N_y)$ is true if x is equal to at least one of the values of $F(N_y)$. The function was successfully found using data analysis methods those were applied to the tables described above.

Results of the modeling

Modeling based on Neural Networks. We used four types of networks with different numbers of internal layers and neurons [9, 10]. When we used protein's characteristics as the data source the best result was 24.83%. The result was obtained for tables contained average values of physical-chemical properties. The network contained three internal layers and 10 neurons in each layer. When we used changing of proteins' free energy as the data source the best result was 72.58%. The network contained four internal layers and 10 neurons in each layer.

Modeling based on method KRAB[1, 3, 2]. When we used protein's characteristics as a data source the best result (among all method's modifications) was 65.86%. When the source was changing of proteins' free energy the best result was 73.08%.

We emphasize that training of neural network takes too much time. On the other hand further network usage, i.e. mutation prediction, does not take much time. Modified method KRAB can be quickly trained but further usage is costly with respect to time and computer's resources.

The next interesting point is that during neural network's training we often obtain strange combination of weights. Combination is strange, for the network with such weights predicts all mutations in the test sample with a probability a beat bigger than 0.5. That can lead to ambiguity. Ambiguity arises, for we don't have a restriction that there should be the unique answer (prediction). Our threshold value is 0.5, therefore the network's answer is 'all mutations are possible'. Accuracy of prediction obtained using coincidence of values in positions of the output-test-vector that are not equal to zero. Therefore, it is possible to get high accuracy in the test sample but the results are useless.

Modified method KRAB is a better tool for prediction of an allowable list of a protein's mutations. Method has an attractive feature; it allows choosing a value of the parameter responsible for the number of amino acids in the output list. Moreover, it is possible to output the list of allowable amino acids with the corresponding list of probabilities. Probability can be considered as a degree of assurance that the mutation made with this amino acid will not deconstruct the protein.

Developed approach gives an opportunity to solve problems in design of genetic engineering experiments in the area of molecular protein design. It also allows to research structural-functional proteins' organization, i.e. to discover amino acids whose are important for protein's structure and functionality.

More detailed description of the methods and result mentioned above is available in Demenkov's MS thesis (2005 year). The thesis is in Russian.

References

[1] Zagoruyko N.G. Prikladnye metody analiza dannyh i znanii. - Novosibirsk: Sobolev Institute of Mathematics Press, 1999.

[2] Arkad'ev A.G., Braverman E.M. Obuchenie mashiny raspoznavaniyu obrazov. M.: Nauka, 1964

[3] Prim Z.L. Kratchaishie svyazyvayuschie seti i nekotorye obobscheniya // Kiberniticheskii sb. 1961. №2. C. 95-107.

[4] Cavadore J.G. Polycondensations da-amino acides en milieu aqueux.: These doctorat es sciences physiques. Montpellier: Universsite des Sciences et Techniques du Languedoc. 1971. 350P.

[5] Zuckerkandl E. The appearence of new structures and functions in proteins during evolution. // J. Mol. Evol. 1987. V.7. P.1-15.

[6] Afonnikov D.A, Komp'yuternyi analiz koordinirovannyh zamen aminokislot v semeistvah gomologichnyh belkovyh posledovatel'nostei. - PhD thesis. Institute of Cytology and Genetics SBRAS. 2002.

[7] Dill K.A. and Chan H.S. From Levinthal to pathways to funnels. - Nat. Struct. Biol. 1997. V. 4. P.10-19.

[8] Dobson C.M., Sali A. andKarplus M. Protein folding: a perspective from theory and experiment. - Angew. Chem. Int. Ed. 1998. V. 37, 868-893.

[9] R. Kallan Osnovnye kontseptsi neironnyh setei. - Moskva, Sankt-Peterburg, Kiev: Izdatel'skmi dom Vil'yams, 2001r.

[10] Grossberg S. 1974. Classical and instrumental learning by neural networks. Progress in theoretical biology, vol. 3, pp. 51-141. New York: Academic Press.

[11] Dennis M.S., Carter P., Lazarus R.A. Binding interactions of kistrin with platelet glycoprotein IIb-IIIa: analysis by site-directed mutagenesis. // Proteins. 1993. V.15. P.312-321

[12] Kidera et al. ,1985,J.Prot.Chem.,4,265

[13] E. Capriotti, P. Fariselli, R. Casadio A neural-network-based method for predicting protein stability changes upon single point mutations // Bioinformatics. 2004. Vol. 20. P. i63-i68.

[14] Gromiha M.M., An J., Kono H. and others (2000) Pro Therm, version 2.0: thermodynamic database for proteins and mutants. Nucleic Acids Res.,28, 283-285.