COMBINED APPROACH TO PROTEIN SECONDARY STRUCTURE PREDICTION

Amirova S.R.*, Machavariani M.A., Filatov I.V., Milchevsky Ju.V., Esipova N.G.,
Tumanyan V.G.

Engelhardt Institute of Molecular Biology RAS, Moscow, Russia
* Corresponding author: e-mail: amirova_s@hotmail.com

Keywords: protein secondary structure, rotamers library, conformational calculations, discriminant analysis

Summary

Motivation: The most important achievements in protein secondary structure prediction are based on two different approaches. The first one is the statistical approach and the second one is the physical-chemical approach. In the first approach we analyze appearance of different types of amino acids in given conformations. The second approach use conformational calculations and physical-chemical properties of a given molecule. Presently these approaches are developed independently. The creation of the method, which will contain advantages of the statistical and the physical-chemical approaches, is very important task.

Results: We have developed the new approach for secondary structure prediction. Using our combined approach one can obtain the secondary structure of a given protein from its primary structure only. The base of our method is a joint using advantages provided by conformational calculations, data on primary structure and physical-chemical properties of proteins. For the combined approach to be demonstrated, we have predicted the protein secondary structure of four basic types: α-helix, helix 3/10, coil, and turn using only sequences of given proteins from Protein Data Bank.

Introduction

Presently is reached the essential progress in the protein spatial structure prediction (Bonneau, Baker, 2001; Lim, 1974). The most important purposes of current researches are the fundamental understanding of physical-chemical principles, which define stability of a given protein. For a spatial model of protein to be created, it is necessary to take into account internal and external molecular interactions in a given protein. Molecular mechanical and molecular dynamical methods are widely used for these tasks. These methods allow to reproduce the structure of a given molecule by means of the certain force field and the finding of the global energy minimum (Milchevsky et al., 2001).

The base of our method is a joint using results provided by conformational calculations, data on primary structure and physical-chemical properties of model protein structures. These results are necessary for further statistical calculation by stepwise discriminant analysis.

Discriminant analysis is used with the aim to find discriminant functions. These functions divide (or discriminate) two or more category to the best advantage. Also one can use discriminant analysis with the aim to find variables, which provide the statistical significant contribution in the division on categories.

Model

Our method has been developed specially for simultaneous use of molecular-mechanics calculations, data on primary structure and the physical-chemical properties of proteins. Conformational computations for polypeptide chains of some proteins are carried out for different backbone
conformations. These conformations correspond to basic types of secondary structure. Optimization procedure is implemented using the rotamers library (Dunbrack, Cohen, 1997) with the aim to find optimal conformations of side chains. The best values of observed energy and its components are a subset of predictor variables for further statistical analysis by stepwise discriminant analysis. All variables with the exclusion of predictor variables represent sequence and physical-chemical properties of proteins. During discriminant analysis we predict classification into a given category as a function of the predictor variables. Developed program enables to handle more than 1000 given categories and 10000 predictor variables. The preliminary analysis of Protein Data Bank has shown that a bigger part of real peptides are contained in the smaller amount of categories. The method investigates the statistical significance for every predictor. As a result, statistically insignificant variables do not influence on category classification. The method provides quantitative measure for comparison of models based on different sets of predictor variables. The final purpose of calculations is the finding discriminant functions. Then on their base we obtain the probability for every amino acid to be in the conformation of four basics secondary structure types.

From Protein Data Bank we took 48 proteins with the following codes: 1A6N,1aon,1AYE,1baz,1bni,1BRS,1c8c,1c90,1cei,1CSE,1CSP,1divm,1eal,1fkb,1FN,1HNG,1hz6,1MQ,1mb,1OP,1mje,1O,1php,1pin,1PNJ,1psf,1QOR,1RA9,1ris,1shf,1SHG,1isl,1ten,1TT,1urn,1wit,256b,2A5E,2ABD,2CI2,2LZM2PDD,2RN2,2VIK,3CHY

For each primary structure corresponding to real protein we have executed starting-up procedure with the aim to get the initial backbone conformation. To find the optimal conformation we were varying dihedral angles of side chain in accordance with the limited selection algorithm. Then we optimized obtained structures using secondary structure dependent rotamer library (SSDEP). After that, we have done energy optimisation using Internal Coordinates Mechanical (ICM) method (Abagyan et al., 1994). From all obtained structures we have chosen the structure with the best conformational energy and considered it as final result. Thereby we have four categories in accordance with four basics secondary structures types. As for input variables, we have chosen all energy compounds, which we used at minimization: Van-der-Waals energy, hydrogen bonding energy, torsion energy, electrostatic energy and full conformational energy. For discriminant functions to be sensitive to not only linear dependences from energy, we included in the set of input variables squares and cubes of all used energy terms. Finally we had fifteen input variables. The most statistically important variable was the square of Van-der-Waals energy term.

Results and Discussion

After discriminant functions are known our program calculates the probability for every amino acid to be in the conformation of four basics secondary structure types. These probabilities for certain amino acids from a several of 48 accepted in attention real proteins to be in conformations of four basics secondary structure types are shown in Table 1. So Table 1 consists only the part of our results.

In our method the amino acid is distributed correctly on secondary structure type, if it has the maximal probability to be in the conformation of the certain secondary structure type in comparison with probabilities to be in conformations of the others secondary structure types. Using such approach of distribution we obtain the Table 2, which gives results of protein secondary structure prediction.

The every line in Table 2 corresponds to the certain secondary structure type in a given protein. For example considering the secondary structure type \( \alpha \)-helix in all accepted in attention proteins with known secondary structure we have 1125 + 693 + 351 + 225 = 2394 amino acids in \( \alpha \)-helix conformation. According to our calculations 1125 of them are in \( \alpha \)-helix conformation, 693 are in helix 3/10 conformation, 351 are in coil conformation and 225 are in turn conformation.
From the Table 2 one can evaluate the accuracy of our method using the percent of correct prediction secondary structure type. For helix 76% of amino acids in all accepted in attention proteins were predicted correctly, for coil the percent of corrected prediction is 47%, and for turn the percent is–12%.

The main limitation in the application of our method is an ambiguity in the classification on secondary structure types for a given backbone, because there are many different classifications on secondary structure types presently.

Using the Table 2 one can see the main advantages and disadvantages of our method. One of the most important advantages is the possibility of simultaneous prediction of several secondary structure types being kept in a given sequence.

As for disadvantages, one of them is the comparatively low percent of correct prediction secondary structure type presently.

The main trend of the improvement of the method is the using more detailed definitions of the secondary structure types and the assignment larger number of categories for them. For example using the DSSP classification (Kabsch, Sander, 1983) one can find a several groups of ten of b-turns. In this classification many conformations, which formally correspond to different types of b-turns, have absolutely different set of backbone dihedral angles. Preliminary it is necessary to distribute different secondary structure types on intelligent subcategories. After distribution we can attribute several secondary structure types in one category coming from physical considerations.

Acknowledgements

This work was supported by grants from the Russian Foundation for Basic Research (No. 03-04-49017 and No. 02-04-49114), grant for Ministry of Industry, Science and Technologies (No. 43.071.1.1.1517) and grant on Molecular and Cellular Biology RAS (Program No. 10).
References


