Prediction of protein functions from tertiary structure of proteins together with evolutionary perspective

Elizabeta Ilievska Mirchevska[1], Ilinka Ivanoska[1] and Slobodan Kalajdziski[1]

[1]Ss. Cyril and Methodious University, Faculty of Electrical Engineering and Information Technologies, Karpos 2 bb, 1000 Skopje, Macedonia

e.ilievska@vipoperator.mk,{ilinka.ivanoska, slobodan.kalajdziski}@finki.ukim.mk

Abstract. The analysis of the characteristics of two or more proteins can result in comparison of their mutual or totally different characteristics and as a result can lead in obtaining general conclusion about determining their function. Although comparative methods give substantial results, we would like to point out that including the so called, "logical" determination will mean that functional predictions can be greatly improved by focusing on how the genes became similar in sequence (i.e., evolution) rather than on the sequence similarity itself [22]. This evolutionary perspective method, based on phylogenetic analysis, will separate only the relevant data, thus simplifying the function prediction and making it more accurate. From the obtained results we can conclude that the combination of phylogenetic analysis with comparative methods results in better accuracy in prediction.

Keywords: Protein function prediction, Phylogenetic analysis, Gene Ontology, C4.5 Classification

1 Introduction

Fast and accurate prediction of protein functions is significant challenge in this era. Each day, more and more sequences are discovered, but the knowledge of their function is still not satisfactory. Prediction of protein function is based on conclusions from previously gathered data of proteins with known functions and their similarity with the protein of interest [1].

Analysis of characteristics of two or more proteins leads to comparison of their common or totally different characteristics to the resulting conclusion i.e. assigning function to the unknown protein. In order to find and determine the function of unknown protein it is necessary to know its structure and the similarity metrics that will be used for measuring similarity with other proteins. Protein structure can be described in many ways. The primary structure is the sequence of amino acids of the protein, while the secondary and tertiary structure describe the position of the protein in 3D

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

space. Spatial band determines the chemical characteristics of the protein and its function. Characteristics such as length of amino acid sequences, the radius of molecules or parts of polarity describe the building structure, and thus the function of the protein. By comparing these features with such characteristics of other proteins we can get information about a similarity with other proteins or we can infer the information that they share the same ancestor. Hence, knowledge of the protein structure and finding other similar protein structure leads to the discovering functional correlation between protein structures.

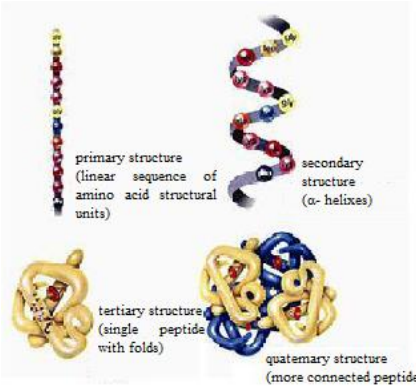


Fig. 1. Protein structure levels

Relatively new approach in predicting protein functions is the concept of phylogenetic analysis. Phylogenetic analysis is based on a relatively simple premise - because genetic features have changed as a result of evolution, the reconstruction of their evolutionary history should help in predicting the functions of uncharacterized genes [2]. This means that predicting the functions can be significantly improved if the methods focus on the evolution of genes / proteins i.e. how they became similar in sequence instead of just their similarity. This evolutionary perspective will separate only useful data and prediction will be simpler and more accurate.

Until now predicting the functions of uncharacterized proteins was based mainly on the examination of primary and secondary structure or a combination of these features. Here, we will also be make a combination of two descriptors of proteins - their primary, secondary and tertian structure described by descriptor and their evolutionary characteristics derived from the primary structure. As mentioned, the secondary and tertiary structure describe the position of the protein in 3D space as a consequence of inter atomic forces affecting this structure and affect the function of the protein. Hence, the secondary structure of proteins is an important feature associated with the function of the protein which is to be expected. The choice of the second feature that will be considered in this work i.e. evolutionary characteristic of the protein is taken because of the logical concept of such analysis and its connection with the function of the protein.

In this paper, we will explain the concepts of phylogenetic analysis, various methods for sequence alignment and building phylogenetic tree together with an analysis of advantages and disadvantages giving an explanation for the selected method. We will also describe the method for calculating the secondary structure of the protein and its biological properties, mainly introducing the concept of ontological

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

structure - Gene Ontology [3] and discussion of the final results for the prediction of protein functions obtained by using these methods.

In section 2 we present details about the proposed phylogenetic analysis approach, additionally, section 3 describes a protein descriptor used in the analysis. Section 4 presents the experimental results, while the section 5 concludes the paper.

2 Phylogenetic analysis

Analysis of the characteristics of two or more organisms by comparing their similarities or differences leads to conclusion of their connection. The methods used vary in the approach of prediction. If the methods perform analysis and make a final conclusion depending on the similarity of the entities examined, they belong to the group of homologous methods [4]. In the other group, in non - homologous methods other characteristics are examined rather than their similarity. The key determining factor in both homologous and non-homologous methods are comparison of the characteristics of unknown proteins with the characteristics of known protein. This comes to the conclusion that both methods have comparative nature [4] i.e. they simply focus on counting and characterizing the similarities and differences between organisms. However, with a better understanding of the biology, it is helpful to clarify how did these similarities and differences came. This concept is known as an evolutionary perspective of comparative biology.

Prediction of protein function with comparative method together with methods based on the evolutionary perspective leads to better results, i.e. more precise final prediction of protein functions. This can be considered as a combination of "numerical" and "logical" description of the characteristics. Numerical determination representing the results of comparative methods i.e. numbered similarities and differences of proteins. The practical logic behind this comparison is based on the fact that genes that have similar characteristics are likely functionally related. However, some of these characteristics are important for determining functional relationship, and some can even be completely irrelevant. Therefore, it is always good to use the advantage of the so-called "logical" choice especially in combination with such numerical comparative methods. Thus, by focusing on how genes occurred similar in sequence (evolution) predictions will be significantly improved. Methods of the evolutionary perspective will separate only the relevant data and determining the final protein function will be simpler and more accurate.

This method is based on a relatively simple assumption—because gene functions change as a result of evolution, reconstructing the evolutionary history of genes should help predict the functions of uncharacterized genes. The first step is the generation of a phylogenetic tree representing the evolutionary history of the gene of interest and its homologs. Such trees are distinct from clusters and other means of characterizing sequence similarity because they are inferred by special techniques that help convert patterns of similarity into evolutionary relationships [5]. The basic concepts of phylogenetic analysis are quite easy to understand, but understanding what the results of the analysis mean, and avoiding errors of analysis can be quite difficult.

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

The final product of phylogenetic methods is phylogenetic tree from which we can get information about evolution. Phylogenetic tree represents structure in which organisms are arranged in branches that are connected according their relationship and evolutionary distance. An example of such a tree with root (ancestor) and scaled branches is shown in Figure 2.

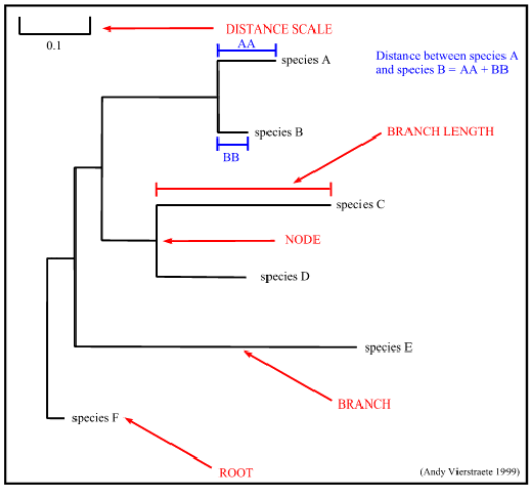


Fig. 2. Phylogenetic three with scaled branches

The phylogenetic tree is a binary tree with or without root, made of leaves and branches where the leaves can be organisms, genes, proteins etc. It is constructed of protein or DNA sequences. The leaves and branches reflect the evolutionary link between members of these sequences where the leaves are sequences and the branch length marks evolutionary distance (during evolution) among members.

Regardless of the sequence type the first step in building a phylogenetic tree is the sequence alignment of the data set. There are methods for global and local alignment, methods for pairwise alignment or alignment of group of sequences and each different method is more accurate than the other under different conditions and individual datasets. For example, global alignment gives better results for sequences that are similar to each other and local alignment is recommended for long sequences with regions of similarity. Therefore the choice of methods should be made taking in consideration all the parameters of the training and test set and the aim of the experiment and this choice makes great impact because the result of the alignment is actually the input for the method for building phylogenetic tree and thus significantly affect the accuracy of evolutionary relationships.

Methods to generate phylogenetic trees can be divided according to how they process the data or by the approach for building the tree [7]. The method of processing data are the methods that are based on the distance or the characters (discrete methods). The first measure is the distance / difference between two genes and construct the tree from the resulting matrix of distances. Others evaluate all possible trees and choose the final one that optimizes evolution. Another way of dividing methods for building phylogenetic trees is according the approach for building the tree.

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

Methods of clustering follow many steps (algorithm) when generating the tree. Methods for building phylogenetic trees from second class use so called optimality criterion to choose one of the many possible trees. This criterion is used to assign "points" or rank of each candidate tree. Points / rank are calculated according to the link between the tree and the input data. Examples of these methods are maximum parsimony or maximum likelihood.

In general, methods of distance are often a choice because of their speed. Also, they generate trees with values for the length of the branches. NJ compared with UPGMA method is faster and provides more precise results. It also does not work on the basis of the presumption of molecular clock. The method of maximum parsimony works faster than the maximum likelihood method and "weighted" schemes of parsimony method can cope with the different models used in the method of maximum likelihood [7][8][9]. Maximum likelihood method is the slowest, but most intense and in most cases provides the best result and tree with most information.

Selecting the method depends mostly on the data set, available memory and computational resources. For large datasets and small memory and computational resources recommended method is neighbor join [10][11], which is also our choice for this experiment. Also, for purposes of this work the phylogenetic tree needs to be expressed with values of the evolutionary distance between sequences, so the choice comes down to some of the methods based on distance. Neighbor join method is the best choice because of its speed and accuracy.

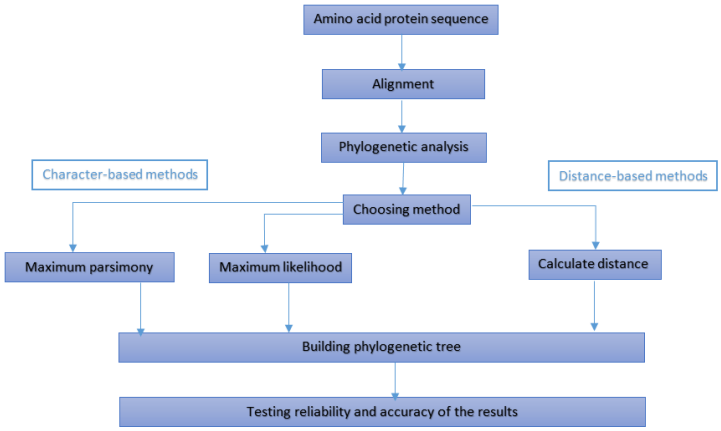


Fig. 3. Building phylogenetic tree

3 Methods

In this experiment we work with data sets from Protein Data Bank [6] and SCOP [5]. The training set is a subset of SCOP version 1.73 (November 2007) and a test set of SCOP version 1.75 (June 2009). The whole data set is divided into subsets according to the SCOP where all members of a subset belong to the same class in SCOP. SCOP is classified into 11 classes, but only some of them have a sufficient number of samples needed to perform the experiment.

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

The experiment consists of several steps:

1. Separation of the data set into subsets according to the SCOP classification where each subset consists of a protein belonging to a single SCOP class;
2. Alignment of each subset with progressive method (global sequence alignment method);
3. Creating a phylogenetic tree for each subset separately with the method of nearest neighbors (Neighbor Join) and clustering of proteins according to our custom algorithm that follows the structure of the phylogenetic tree. This clustering algorithm combines proteins from the first level of the phylogenetic tree and if in the first level a protein is alone in the cluster it joins with the second level of the tree. In the prediction descriptor we use information in which cluster every protein from the training and the test set belongs;
4. Calculation of 3D descriptors for the protein data set according to the method described - the descriptor [12] relies on the geometric 3D structure of the proteins. It consists of four phases: triangulation, normalization, voxelization of the 3D protein structures, and the Spherical Trace Transform applied. As a result, geometry-based descriptors are produced, which are completely rotation invariant. The training procedure is the descriptor extraction. Descriptors consisting of 450 features (416 of them describe the protein's geometry, while 34 of them give information for the primary and secondary protein structure) are generated for each protein forming a training set for a C4.5 decision tree algorithm.

5. Creating Gene Ontology (GO) descriptor for the protein data set – for each class in SCOP clustering in smaller segments according to the phylogenetic tree. For each protein belonging to a particular cluster a descriptor is made according to the percentage of how many GO molecular functions it contains compared to all molecular functions of proteins from all members of a given cluster. For example, we are working with 5165 proteins from the class of 46456 - alpha helixes. They are clustered in 132 clusters. Proteins from the first cluster have different annotated GO molecular functions whose union has a total of 11 molecular functions. The first protein cluster instance has 3 out of 11 functions and is described by descriptor 3/11 and so on for each protein from each cluster.
6. Creating a data classifier for the training and testing set. It is a classifier for predicting that uses the descriptor with the attributes of the 3D protein descriptor, GO and information in which cluster derived from phylogenetic analysis tree the protein belongs to. For the need of the research, experiments are made with a combination of any of the above attributes and each attribute separately;
7. Classification with C.45 algorithm;
8. Analysis of results;

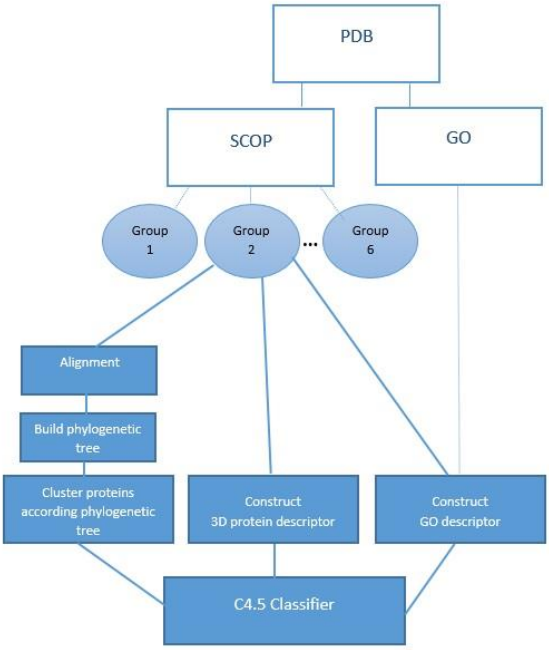
S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

For the purpose of the experiment we are working with several groups of datasets:

Group 1. Set of proteins with SCOP class 46 456 - α helices
Group 2. Set of proteins with SCOP class 48 724 - β helices
Group 3. Set of proteins with SCOP class 51 394 - α / β domains
Group 4. Set a group of proteins SCOP class 53 931 - α + β domains
Group 5. Mixed set with multiple protein classes (46456, 51394)
Group 6. Mixed set with multiple protein classes (56835, 56572, 57942, 58231 and 58788) [16]

The number of samples for all of the groups in the training and test sets is roughly the same, because that affects the accuracy in prediction. For all groups we make a prediction of a SCOP family using only the 3D protein descriptor, only phylogenetic analysis attribute and a combination of both, with and without the GO descriptor. As noted above the 3D descriptor so far gives good accuracy for prediction and from our experiment is expected that phylogenetic analysis will improve the precision of sets belonging to the same SCOP family. Additionally, we make addition of the GO descriptor to test the prediction accuracy in case of a combination of the Gene ontology with phylogenetic analysis.

The first four sets are proteins belonging to the same SCOP class because of the assumption that the use of phylogenetic information will yield to better results in sets of proteins that have similar properties. Additionally, the latter two groups are selected to verify this assumption. Selecting sets of proteins belonging to the same class is due to the SCOP classification logic where SCOP classes are determined according to the secondary structure of the protein which is closely related to the evolutionary history.



S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

Fig. 4. Graphical presentation of the experiment. All 6 groups go through all steps that are presented for group 2

4 Experimental Results

The results shown in the following tables confirm our expectations and the theoretical background explained in upper sections. Table 1 includes the results of the first four groups of similar proteins and table 2, the latter two groups mixed proteins. Therefore, in the first table we can notice the positive impact of phylogenetic information accuracy in predicting, and in the second table we can conclude that the phylogenetic analysis has no effect.

Table 1. Results for sets with similar proteins

Data set	Group 1	Group2	Group3	Group4
Number of proteins in training data set	6222	7144	6220	6211
Number of proteins in test data set	689	852	885	1127
Accuracy with 3D descriptor	80.8418	67.4883	58.3051	62.2893
Accuracy with 3D descriptor and phylogenetic analysis	80.9869	68.0751	58.6441	62.4667
Accuracy with 3D and GO descriptor	80.2612	71.1621	58.5311	61.4662
Accuracy with 3D and GO descriptor and phylogenetic analysis	80.4064	71.1621	58.6441	61.5602
Accuracy with phylogenetic analysis	3.9187	3.1023	1.9209	2.2496

Table 2. Results for sets with similar and different proteins

Data set	Group 5	Group 6
Number of proteins in training data set	5624	1269
Number of proteins in test data set	620	209
Accuracy with 3D descriptor	85.4839	77.9904
Accuracy with 3D descriptor and phylogenetic analysis	85.4839	77.9904
Accuracy with 3D and GO descriptor	85.6452	77.9904
Accuracy with 3D and GO descriptor and phylogenetic analysis	85.6452	77.9904
Accuracy with phylogenetic analysis	7.5806	2.8708

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

According to the results shown in Table 1, the accuracy of the prediction is improved by adding additional information from phylogenetic analysis of the first four groups as expected because proteins belonging to them are functionally and evolutionary related and belong to the same class. The accuracy in the last two groups containing related and unrelated proteins at the same with and without addition of phylogenetic information. The present high rate of accuracy comes due to the already good precision of the 3D descriptor, but also it is unavoidable to conclude that phylogenetic analysis does not affect the improvement of precision.

Further confirmation of the benefits of including the phylogenetic analysis is when in predicting experiments we involve a GO factor as an attribute for classification. In such cases, accuracy is increased when introducing phylogenetic analysis.

It is worth mentioning the poor prediction accuracy when using only information from phylogenetic analysis. Taking in consideration the data set we are working with and the type of information from phylogenetic analysis which is forwarded in predicting, the conclusion is that the phylogenetic data attribute is not sufficiently descriptive to independently give satisfactory results. Reason for further work is creation of a more extensive descriptor which would give information of the phylogenetic origin of each sample and which would include information also for the other levels of the phylogenetic tree that is generated.

5 Conclusion

The evolutionary perspective is closely related to the function of proteins. Proteins change with time and therefore change its features and functions. How did the change come is equally important as the change itself. Therefore, better results in predicting are obtained when you take into account information how certain proteins become similar together with data on their similarity.

In this paper we tried to present the difference in predicting including information about the evolutionary origin of proteins, which is a new perspective known as phylogenetic analysis. Predicting with 3D descriptor already provided excellent results but bioinformatics always strives to improve the precision so every new way that gives slightly better results than the previous is a great progress taking in consideration the benefits of the practical uses of prediction of protein functions in the world.

The prediction of protein functions by an inclusion of phylogenetic information gives better results in proteins that are similar to each other i.e. proteins with close evolutionary origins. In addition, from the experiments with the included descriptor for the functions of each protein by Gene Ontology and the accompanying results we can infer that although the inclusion of this descriptor sometimes gives worse results in general it does not affect the accuracy in predicting with phylogenetic information because under the same conditions the predicting is always improved when you introduce phylogenetic information.

A motivation for further work is expanding the descriptor using Gene Ontology striving not to deteriorate the accuracy in predicting under the same conditions and testing results when using information from two-level phylogenetic analysis.

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288

References

1. Eissen, Jonathan A. Phylogenomics: Improving Functional Prediction for uncharacterised Genes by Evoluationary Analysis (1998).
2. Matteo Pellegrini, Edward M. Marcotte, Michael J. Thomson, David Eisenberg, and Todd O. Yeates: Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, Los Angeles, CA (1999).
3. The Gene Ontology Consortium: http://www.geneontology.org/ [Accessed June 2014].
4. Eisen, Jonathan A. and Wu, Martin: Phylogenetic Analysis and Gene Functional Predictions, Phylogenomics in Action, Rockville, Maryland (2002).
5. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: Scop: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, vol. 247, pp. 536--540 (1995).
6. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res., vol.28, pp. 235--242 (2000).
7. Andreas D. Baxevanis and B.F. Francis Ouellette, A practical guide to the analysis of genes and proteins, Bioinformatics (2001).
8. P.Higgs and Manchester: Introduction to Phylogenetics Methods, ITP series on-line seminars (2001).
9. Atteson K.: The performance of neighbor-joining algorithms of phylogeny reconstruction, pp. 101–110, In Jiang, T., and Lee, D., eds., Lecture Notes in Computer Science, 1276, Springer-Verlag, Berlin. COCOON '97 (1997).
10. Mihaescu R, Levy D, Pachter L, Why neighbor-joining works, Algorithmica 54 (1): 1–24. doi:10.1007/s00453-007-9116-4 (2009).
11. Samworth R. J.: Optimal weighted nearest neighbour classifiers, Annals of Statistics 40 (5): 2733–2763. doi:10.1214/12-AOS1049 (2012).
12. Kalajdziski, S., Mirceva, G., Trivodaliev, K., Davcev D.: Protein Classification by Matching 3D Structures. In: Frontiers in the Convergence of Bioscience and Information Technologies 2007, pp. 147--152. Jeju Island, Korea (2007).

S. Loshkovska, S. Koceski (Editors): ICT Innovations 2015, Web Proceedings, ISSN 1857-7288