Predicting Secondary Structure for Human Proteins based on Chou-Fasman Method

Fotios Kounelis[1], Andreas Kanavos[1,2], Ioannis E. Livieris[2], Gerasimos Vonitsanos[1] and Panagiotis Pintelas[3]

1. Computer Engineering and Informatics Department
University of Patras, Patras, Greece
{kounelis, kanavos, mvonitsanos}@ceid.upatras.gr
2. Computer and Informatics Engineering Department
Technological Educational Institute of Western Greece, Antirrion, Greece
livieris@teiwest.gr
3. Department of Mathematics
University of Patras, Patras, Greece
ppintelas@gmail.com

Abstract. Proteins are constructed by the combination of a different number of amino acids and thus, have a different structure and folding depending on chemical reactions and other aspects. The protein folding prediction can help in many healthcare scenarios to foretell and prevent diseases. The different elements that form a protein give the secondary structure. One of the most common algorithms used for secondary structure prediction constitutes the Chou-Fasman method. This technique divides and in following analyses each amino acid in three different elements, which are α-helices, β-sheets and turns based on already known protein structures. Its aim is to predict the probability for which each of these elements will be formed. In this paper, we have used Chou-Fasman algorithm for extracting the probabilities of a series of amino acids in FASTA format. We make an analysis given all probabilities for any length of a human protein without any restriction as other existing tools.

Keywords: Protein Structure Prediction, Computational Structural Biology, Human Proteins, Secondary Structure Prediction, Protein Folding

1 Introduction

Proteins form one or more determined conformations guided by a number of complicated and reversible non-covalent interactions in order to unfold their their biological functions. Although determining the structure of a protein can be obtained by time-consuming as well as relatively expensive techniques such as crystallography, dual polarization interferometry and nuclear-magnetic resonance spectroscopy, new techniques based on the area of bioinformatics have been developed. These techniques aim to compute and in following predict protein structures based on their amino acid sequences.

Amino acids are one of the main components that form proteins [26,27]. They are encoded directly from the genetic code through the transcript-RNA. These specific components contain functional groups of amine and carboxyl as well as side chains. Twenty different amino acids are encoded from the genetic code, although a lot more have naturally occurred [28]. Using different techniques, additional amino acids which generate proteins with newly enhanced properties [29] are genetically encoded.

Protein structure can be predicted with the use of structure analysis and prediction tools according to their amino-acid sequences. More specifically, solving the structure of a given protein can be considered as highly important in medicine (another example is the drug design) and biotechnology (an example constitutes the design of novel enzymes). Thus, the area of computational protein structure prediction is constantly evolving, by following the development of intelligent algorithms and more important, by taking advantage of the increase in computational power of machines.

Regarding the protein structure prediction, the primary structure is employed to predict secondary as well as tertiary structures. The primary structure consists of the linear sequence of amino acids [4,21]. There are twenty different amino acids, that their occurrence in the sequence and the position or positions they occur, will form a different protein.

Furthermore, the different elements that form a protein as well as their 3D structure give the secondary structure [19]. α-helices, β-sheets, turns, and omega loops are the elements that occur in the secondary structure of a protein and the ones that our paper predicts.

After the above elements are formed and in following are combined in the 3D space, the tertiary structure of the protein is shaped [2]. This structure has one or more secondary structures and may have interactions and bonds with amino acid side chains. The tertiary structure is considered as the highest structure of a protein, where all the protein subunits are arranged and numbered [11].

In the present manuscript, we have employed the Chou-Fasman method, an algorithm used for secondary structure prediction. With this method, we aim to divide and analyse each amino acid in three different elements, namely α-helices, β-sheets and turns based on already known protein structures to predict the probability for each one of these elements. There are also other tools to implement the Chou-Fasman method but they lack in different aspects that we overcome.

The first obstacle is the size limit of online tools. Our implementation predicts the secondary structure for any length of sequence. Furthermore, our implementation provides the probabilities for all three different elements that the Chou-Fasman algorithm can predict. This part of our paper overcomes the problem of other tools that don't predict the probabilities for turns. Even if turns have the lowest probabilities, it is yet a very useful part of the secondary structure of a protein.

The second contribution of the proposed work deals with the proteins we study. We aim at the human proteins and their probabilities so that we can

primarily understand what is the performance of this method in this specific type of proteins. This will allow us in the future to build a more robust map for human proteins that are produced from different chromosomes and check the performance of different algorithms on this kind of map.

The remainder of the paper is organized as follows. Section 2 presents a brief survey of recent studies concerning protein structure prediction, homology modeling and protein fold recognition. Section 3 presents a reference to the prediction tools and methods and introduces the Chou-Fasman method while in Section 4, we present our experimental results. Finally, Section 5 discusses our concluding remarks, the open problems as well as our future work.

2 Background Topics

Regarding protein structure prediction, several methods and techniques have been proposed with the aim of attempting to specify the structure and bonds of an amino acid sequence [12,24]. As a result, the knowledge of the protein structure is critical and thus, several determinants can be considered as important; the bond angle stresses, the electrostatic interactions, the hydrogen and covalent bonds, the hydrophobicity and hydrophilicity of residues, the van der Waals interactions, as well as the enthalpy and the entropy.

Regarding protein structure determinants, there are two important aspects as well; initially, the information about them, and as a result, the information regarding the structure of a protein is entirely contained within the sequence (in addition to knowledge of the solvent). In following, these determinants can also be considered as measurements of physical properties. Without the aid protein chaperones, we can assume that proteins can take their native conformation in a solvent; then enough information for predicting protein structures ab initio (from basic principles) can be considered. However, many of the determinants are not precisely known or may be too compute-intensive and computationally non-tractable.

Protein structure prediction has resorted to knowledge-based methods, since there is lack of feasible ab initio methods. Specifically, homology modeling [6] as well as protein fold recognition methods [13] constitute the two major and complementary approaches that were taken.

2.1 Homology Modeling

In terms of homology modeling, the amino acid sequence of a protein with unknown structure is aligned against sequences of proteins with known structures. Furthermore, high degrees of homology (very similar sequences across and between the proteins) can be used for determining the global structure of the corresponding protein and in turn for placing it into a certain fold category. On the other hand, lower degrees of homology may be still used in order to determine local structures, with an example being the Chou-Fasman method [8,16] for

predicting secondary structure. An advantage for homology modeling methods is the lack of dependence on the knowledge of physical determinants.

Every homology modeling process consists of four steps which are always the same [17]. Firstly, the homologous template proteins of already known proteins should be gathered. Secondly, the best set of templates should be selected. In following, this set will be used in order to optimize the multiple sequence alignment for the query. Finally, the model that will be as close as possible to the structure of the templates, will be built. These steps can be repeated until a satisfying model finally occurs [22].

More to the point, homology modeling constitutes a very important tool for 3D protein structure models. This is the reason why web-based integrated systems have been built. SWISS-MODEL [2,5,22] is considered as such a workspace, where for any queries, a library is searched for suitable templates. This tool is very reliable as it depends on sequence alignment for predicting the structure of the protein.

Homology modeling can be further utilized for the correct prediction of quaternary protein structure [1]. The researchers in this work used homology to fill the gap of non-structural experimental evidence for the 3D multimeric assemblies. Apart from that, homology modeling can be also used for large proteins too. In [23], an approach for the structure prediction of large proteins is described. In this work, an alignment method, which exploits bioinformatics algorithms to match values from a database to experimental data, is produced.

2.2 Protein Fold Recognition

Fold recognition methods take a complementary approach. In these methods, structures, not sequences, are aligned. If the sequence of a protein has an unknown structure, with the method called "threading", it undertakes the conformation of the backbone (protein sans side chains) of a known structure. For each attempt, higher physical determinants produce a more complete score for the alignment. These methods tend to be more computational-intensive than homology modeling methods, but they give more confidence in the physical viability of the results.

One of the major contributions is Phyre2 [15]; a suite of tools available on the web to predict and analyse protein structure, function and mutations. The focus of Phyre2 is to provide biologists with a simple and intuitive interface to state-of-the-art protein bioinformatics tools.

A recent work [7] has improved Chou-Fasman method in three aspects; authors replaced the nucleation regions with extreme values of coefficients calculated by the continuous wavelet transform, in following substituted the original secondary structure conformational parameters with folding type-specific secondary structure propensities and finally modified Chou-Fasman rules.

In a more recent work, authors in [20] present a detailed overview of the molecular, functional, and structural characteristics of collagenase from Pseudomonas aeruginosa which might be helpful for understanding the possible structure and functions of unknown proteins.

3 Prediction Tools and Methods

A great number of structure prediction software tools are developed for dedicated protein features, some among which are disorder prediction, dynamics prediction, structure conservation prediction, etc. Homology modeling, protein threading, ab initio methods, secondary structure prediction, and transmembrane helix and signal peptide prediction are approaches that help.

Each method uses a different level of the protein structure to predict another one. The primary structure is based on the amino acids linear sequence, which when combined in different places and varying numbers, form a different protein. In our paper, we deal with the secondary structure prediction. Regarding the tools used for predicting the secondary structure of a protein, they are based on the primary structure. In following, we analyze the secondary structure prediction and especially, the method for predicting the secondary structure on the human proteins in a more detailed way.

3.1 Secondary Structure Prediction

The prediction of local secondary structures is facilitated via these tools. These structures are solely based on the amino acid sequence of the protein and their prediction is afterwards compared to DSSP (Define Secondary Structure of Proteins) score [14], which is calculated based on the crystallographic structure of the protein [3,18]. It is the standard method used to assign secondary structure annotations to a protein structure. In addition, it calculates all DSSP entries from PDB (Protein Data Bank). DSSP utilizes the atomic coordinates of a structure to assign the secondary codes, but is not a secondary structure prediction tool.

Neural nets and support vector machines comprise modern machine learning methods on which prediction methods for secondary structure rely on. Taking into consideration databases of known protein structures along with these methods is an optimal combination.

More specifically, the secondary structure prediction problem for a protein sequence is to predict for each amino acid, whether it appears in α-helix, β-sheets or any other team. There are several prediction algorithms, which provide such probabilities to solve the problem [10,30,31]. One of the most well-known and pioneer method is the Chou-Fasman technique. This method provides probabilities of amino acids whether they appear in α-helix, β-sheets or turns.

3.2 Chou-Fasman Method

The Chou-Fasman method exports the probabilities for a block of amino acids to be shown in each one of the three elements. Based on the frequencies that are calculated from observation for each amino acid, the probability for the whole sequence is predicted. When the method predicts helices and sheets, it works on a similar way; it has a window sliding on the frequency and upon meeting a high probability, it gets extended until the probability is under the predefined threshold.

On the other hand, even if turns have the same length of windows as in helices, the procedure is different due to the fact that many turn regions may occur in both helices and sheets. In order to predict a turn, its probability should be greater than helix or sheet as well as the same probability should be greater than a predefined threshold.

$$p(t) = p_t(j) \times p_t(j+1) \times p_t(j+2) \times p_t(j+3) \quad (1)$$

In formula (1), we can see this probability which explains why it is lower than the probabilities of helices and turns in all our experiments, where j is the position of each amino acid in the positions of the window with length equal to four.

There have been found that different amino acids have a higher tendency to occur in a certain secondary structure class, for example, alanine is known to appear in a helix form. The Chou-Fasman technique is based on statistical data and rules to provide such probabilities. When the method was introduced not many things were known about protein structure and as a result, the accuracy of the method was lower than it is nowadays.

Lab experiments can clearly prove which of the predicted probabilities of this technique are true and which are not. This allows the methods that produce their probabilities from already known structures, such as Chou-Fasman, to improve. As a result, we examine this method on the latest prediction updates. This will allow us to compare these results in later research, with probabilities that either have been collected in the past with the same method or by other modern methods. In the following section, we will analyze the proteins that were used and their form, as well as explain the methodically exported data.

4 Experimental Methodology

In order to gain a deeper insight on the elements that are analyzed, we present the following results that portray the distributions for secondary structure prediction.

The Chou-Fasman algorithm was implemented and run for ten different human proteins. These ten proteins as depicted in Tables 1 and 2 as *Apolipoprotein C-I*, *Casein Kinase 2*, *Disulphide Isomerase*, *Glutaredoxin*, *Insuline like 3*, *Intereleukin6*, *Lysozyme C*, *Major prion Protein*, *Prion like Protein Dropel* and *S100-B*. Each protein was split in five continuous amino acid blocks. For each block, the three different probabilities for each structure that will be formed, is produced. In order to understand and explain these data in a more rigorous way, we have Figure 1 to illustrate them. The three different lines represent the structures; while in y axis, the percentage is represented and in x axis, the different blocks in a consecutive order for each protein, are represented.

In following, we analyze the results of the ten proteins that were used in the Chou-Fasman technique. All these proteins belong to the human organism (Homo Sapiens). The human organism has 23 pairs of chromosomes and in the first row, we can see the number of the chromosomes in which each protein is

Table 1. Human Proteins Details 1/2

	Apolipo-protein C-I	**Casein-Kinase 2**	**Disulphide Isomerase**	**Glutare-doxin**	**Insuline like 3**
Chromosome	19	20	16	5	19
Number of amino acids	83	391	525	106	131
Most frequently appearing amino acids	Leu(L)	Leu(L)	Leu(L)	Leu(L), Gln(Q)	Leu(L)
Number of most frequently appearing amino acids	12	33	63	12	22
Percentage of most frequently appearing amino acids	14.5%	8.4%	12%	11.3%	16.8%
Instability index	30.79	48.79	46.79	43.72	53.10
Stable/Unstable	Stable	Unstable	Unstable	Unstable	Unstable
Non appearing amino acids	Pyl(P), Sec(S)	Pyl(P), Sec(S)	Pyl(P), Sec(S)	Pyl(P), Sec(S)	Pyl(P), Sec(S)

deployed. The proteins are constructed from several amino acids. In the second row, we can see their number of amino acids. We can observe diversity in this specific characteristic as for example, the number of amino acids in protein S100-B is 92, whereas in Disulphide isomerase, the number of amino acids is 525. This diversity gives us a better insight to check the probability spectrum of the Chou-Fasman technique.

In the third, fourth and fifth lines, the most appeared amino acids as well as their percentage in the whole sequence are presented. We notice that Leucine (Leu) is the most common amino acid, which exists in almost all the proteins; specifically, this happens in seven out of the total ten proteins. However, even if Leucine is the most common amino acid in Casein Kinase 2 protein, the percentage remains low; this makes the deduction that all of the amino acids which occur share the same percentage, but this one is slightly larger.

Furthermore, the sixth line presents the instability index. This metric informs us whether a protein will be stable in a protein tube. In order to be stable, a protein should have an instability index of less than 40. As we can see, the four out of the total ten proteins are stable where Prion-like Protein Dropel is the most stable of all and the Casein Kinase 2 is the most unstable. Finally, in all of these proteins, two amino acids never occur inside the sequence, namely Pyrrolysine (Pyl) and Selenocysteine (Sec).

As we can see in Figure 1, both α-helices and β-sheets have for each block a greater probability than turns. Concretely, α-helices have in more than half the blocks a greater probability than β-sheets. Our first remark from the figure is

Table 2. Human Proteins Details 2/2

	Intere-leukin 6	**Lyso-zyme C**	**Major prion Protein**	**Prion like Protein Dropel**	**S100-B**
Chromosome	7	12	20	20	21
Number of amino acids	211	148	253	176	92
Most frequently appearing amino acids	Leu(L)	Ala(A)	Gly(G)	Leu(L)	Glu(E)
Number of most frequently appearing amino acids	28	15	45	20	16
Percentage of most frequently appearing amino acids	13.3%	10.11%	17.8%	11.4%	17.4%
Instability index	37.38	27.71	43.11	23.06	36.45
Stable/Unstable	Unstable	Stable	Unstable	Stable	Stable
Non appearing amino acids	Pyl(P), Sec(S)	Pyl(P), Sec(S)	Pyl(P), Sec(S)	Pyl(P), Sec(S)	Pyl(P), Sec(S)

that for all the blocks in these ten human proteins, these amino acids are more possible to form an α-helix structure.

A second observation that follows the previous one, is that there are whole proteins, like the third (*Casein Kinase 2*) and the sixth (*Prion like Protein Dropel*) for the amino acid blocks 11-15 and 26-30 respectively, that have bigger probability of forming an α-helix. This means that these two proteins, in their whole sequence, according to the Chou-Fasman algorithm, will have a structure of an α-helix. This is not usual regarding the corresponding proteins, as only the two out of ten have this probability.

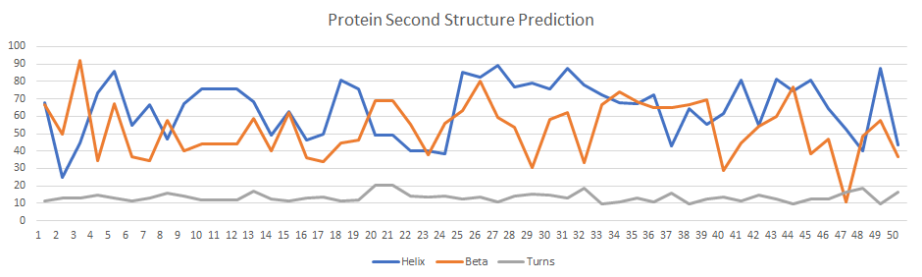

Fig. 1. The different probabilities for the three structures in all amino acid blocks

5 Conclusions and Future References

In this paper, we analysed the problem of the secondary structure prediction for human proteins. Chou-Fasman is a method which solves this problem by predicting the structure in which a block of amino acids will be formed. Depending on the probability of each amino acid that will appear in α-helices, β-sheets, or turns, this method produces the output. We used this method for ten human proteins to predict their secondary structure form.

Chou-Fasman method achieves success rate equal to 50%-60%, while other more modern secondary structure prediction methods, like GOR [9] or Porter 5 [25] achieve better results. As an open problem for research will be to implement other modern methods and make a comparative study between these methods and thus, discuss their differences in human proteins. Furthermore, it would be of great interest to implement those methods for proteins that are coded from different human chromosomes. This could potentially create a hybrid model for the use of different prediction methods depending on the chromosome from which a protein is derived of.

References

1. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T.: Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. Scientific Reports 7(1) (2017)
2. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L., Schwede, T.: SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research 42(Webserver-Issue), 252–258 (2014)
3. Bliven, S., Lafita, A., Parker, A., Capitani, G., Duarte, J.M.: Automated evaluation of quaternary structures from protein crystals. PLoS Computational Biology 14(4) (2018)
4. Bock, J.R., Gough, D.A.: Predicting protein-protein interactions from primary structure. Bioinformatics 17(5), 455–460 (2001)
5. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J.N.D., Schwede, T.: Protein structure homology modeling using swiss-model workspace. Nature Protocols 4(1), 1–13 (2008)
6. Cavasotto, C.N., Phatak, S.S.: Homology modeling in drug discovery: current trends and applications. Drug Discovery Today 14(13), 676–683 (2009)
7. Chen, H., Gu, F., Huang, Z.: Improved chou-fasman method for protein secondary structure prediction. BMC Bioinformatics 7(S-4) (2006)
8. Chou, P.Y., Fasman, G.D.: Empirical predictions of protein conformation. Annual Review of Biochemistry 47(1), 251–276 (1978)
9. Garnier, J., Gibrat, J.F., Robson, B.: Gor method for predicting protein secondary structure from amino acid sequence 266, 540 – 553 (1996)
10. Ito, M., Matsuo, Y., Nishikawa, K.: Prediction of protein secondary structure using the 3d-1d compatibility algorithm. Computer Applications in the Biosciences 13(4), 415–424 (1997)
11. Janin, J., Bahadur, R.P., Chakrabarti, P.: Protein-protein interaction and quaternary structure. Quarterly Reviews of Biophysics 41(2), 133–180 (2008)

12. Jiménez-Montaño, M.A., de la Mora-Basáñez, C.R., Pöschel, T.: The hypercube structure of the genetic code explains conservative and non-conservative aminoacid substitutions in vivo and in vitro. BioSystems 39(2), 117–125 (1996)
13. Jones, D., Taylor, W., Thornton, J.: A new approach to protein fold recognition. Nature 358, 86–89 (1992)
14. Joosten, R.P., te Beek, T.A.H., Krieger, E., Hekkelman, M.L., Hooft, R.W.W., Schneider, R., Sander, C., Vriend, G.: A series of PDB related databases for everyday needs. Nucleic Acids Research 39(Database Issue), 411–419 (2011)
15. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., Sternberg, M.J.: The phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols 10(6), 845 (2015)
16. Kumar, T.A.: Cfssp: Chou and fasman secondary structure prediction server. Wide Spectrum 1(9), 15–19 (2013)
17. Meier, A., Söding, J.: Automatic prediction of protein 3d structures by probabilistic multi-template homology modeling. PLoS Computational Biology 11(10) (2015)
18. Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Trong, I.L., Teller, D.C., Okada, T., Stenkamp, R.E., Yamamoto, M., Miyano, M.: Crystal structure of rhodopsin: A g proteincoupled receptor. Science 289(5480), 739–745 (2000)
19. Perticaroli, S., Nickels, J.D., Ehlers, G., O'Neill, H., Zhang, Q., Sokolov, A.P.: Secondary structure and rigidity in model proteins. Soft Matter 9(40), 9548–9556 (2013)
20. Rani, S., Pooja, K.: Elucidation of structural and functional characteristics of collagenase from pseudomonas aeruginosa. Process Biochemistry 64, 116–123 (2018)
21. Sanger, F.: The arrangement of amino acids in proteins. In: Advances in Protein Chemistry, vol. 7, pp. 1–67 (1952)
22. Schwede, T., Kopp, J., Guex, N., Peitsch, M.C.: SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31(13), 3381–3385 (2003)
23. Shen, Y., Bax, A.: Homology modeling of larger proteins guided by chemical shifts. Nature methods 12(8), 747–750 (2015)
24. Siman, R., Noszek, J.C.: Excitatory amino acids activate calpain i and induce structural protein breakdown in vivo. Neuron 1(4), 279–287 (1988)
25. Torrisi, M., Kaleel, M., Pollastri, G.: Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. bioRxiv (2018)
26. Vickery, H.B.: The history of the discovery of the amino acids ii. a review of amino acids described since 1931 as components of native proteins. In: Advances in Protein Chemistry, vol. 26, pp. 81–171 (1972)
27. Vickery, H.B., Schmidt, C.L.: The history of the discovery of the amino acids. Chemical Reviews 9(2), 169–318 (1931)
28. Wagner, I., Musso, H.: New naturally occurring amino acids. Angewandte Chemie International Edition in English 22(11), 816–828 (1983)
29. Xie, J., Schultz, P.G.: Adding amino acids to the genetic repertoire. Current Opinion in Chemical Biology 9(6), 548–554 (2005)
30. Yavuz, B.C., Yurtay, N., Özkan, Ö.: Prediction of protein secondary structure with clonal selection algorithm and multilayer perceptron. IEEE Access 6, 45256–45261 (2018)
31. Zhou, Z., Yang, B., Hou, W.: Association classification algorithm based on structure sequence in protein secondary structure prediction. Expert Systems with Applications 37(9), 6381–6389 (2010)