VOL4 NO. 1 March 2015 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.ius.edu.ba

IUS Soft Computing Research Group

Historical Criteria for Structural Classes of Proteins in Percentages: After 20 Years

Betul Akcesme, and Mehmet Can
International University of Sarajevo, Faculty of Engineering and Natural Sciences, Hrasnicka Cesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
mcan@ius.edu.ba
bcicek@ius.edu.ba

Article Info

Article history:
Article received on march 2015
Received in revised form April 2015

Keywords:
Amino acid conformations; Structural classes; Protein structural class prediction; Secondarystructure

Abstract

Two decades ago scientists proposed some criteria for the structural classes in percentages. Today experts at SCOP classified hundreds of thousands of proteins into one of the four structural classes manually by inspection, and observation. Nakashima et al. gave a classification criteria. P.Y. Chou also proposed another method to classify proteins according their residue contents in three conformations, helix, sheet, and coil. Later P.Y. Chou revised his method. Today SCOP listed around 100.000 proteins with their structural classes. In this paper two datasets will be used to reveal the percentages of residues in α-Helices, β-sheets, and coils in proteins of classes all α, all β, α+β, and α/β, in the classifications made by experts in SCOP. The first of the data bases is PDBselect25 which contains 1670 twilight zone proteins whose similarity is less than 25%. The second data base BF30 consists of 10294 proteins picked from PDB database with the similarity threshold of 30%. Structural classes of these proteins are taken from SCOP database. It is seen that there is a very poor correlation between historical criteria, and SCOP's scientists' intuition in classification of proteins into structural classes.

1. INTRODUCTION

In SCOP database, a protein is mainly classified into one of the following five structural classes: all α, all β, α+β, α/β. Experts at SCOP have an intuition to decide about the structural class of a new protein. However no explicit formulation of this intuition revealed to the time. The structural class of a protein is hypnotized that it is somehowcorrelated with its amino acid composition. Various efforts have been made in finding coherent criteria that will help one to find out the structural class of a given protein. This article addresses the start and progress in this field.

There are two theoretical approaches to predict the structure of a protein. One is the free-energy minimization method, which is based on the empirical atomic potentials (see, e.g., Scheraga, 1968, 1987; Weiner and Kollman, 1981; Levitt, 1983; Gilson and Honig, 1988; Mackay et al., 1989; Rogers, 1989; McCammon et al., 1989; Chou et al., 1990; Karplus and Shakhnovich, 1992). The other is the statistical method, which was developed based on various statistical data extracted from structure-known proteins (see, e.g., Chou and Fasman, 1974, 1978; Lim, 1974; Garnier et al., 1978; Cid et al., 1982; Fasman, 1989; Jones et al., 1994; Orengo et al., 1994). Various physical theories of protein structures at different levels have alsobeen proposed for improving the prediction of protein structure (Ptitsyn and Rashin, 1975; Ptitsyn and Finkelstein, 1979, 1980, 1989; Ptitsyn et al., 1985;

Finkelstein and Ptitsyn, 1987; Chothia and Finkelstein, 1990; Kuwajima et al., 1993; Kolinskiand Skolnick, 1994; Vieth et al., 1994; Mitchell et al., 1994; McDonald and Thornton, 1994).

Secondary structures of proteins are obtained in the form of the x-ray analyses in three conformations helix "h", sheet "s" , and others ".". Others are interpreted as coils "c".

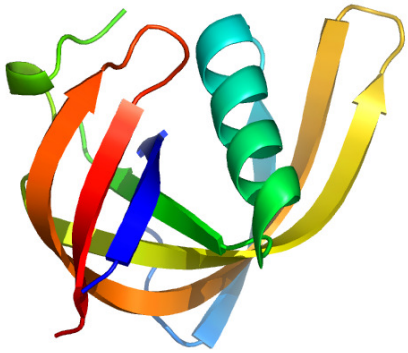


Figure 1. α-helices, β-sheets, and coils on the same picture

2. HISTORICAL CRITERIA FOR STRUCTURAL CLASSES IN PERCENTAGES

Proteins of known structures are generally classified into one of the following four structural classes: all α, all β, α+β, α/β, and irregular proteins (Levitt and Chothia, 1976; Richardson and Richardson, 1989). The earlier ideas for classification were based on the percentage of secondary structurecomponents, although there was no unified quantitative criterion. Today scientists at SCOP have some consensus which helps them to classify proteins manually by inspection. In the sequel it will be shown that their common sense fits previous percentage based criteria very poorly.

Percentages of amino acid residues atα-helices and β-sheets in a protein are abbreviated by α and β, respectively. The classification by Nakashima et al. (1986) was made according to the following criterion:

All α-proteins,α> 15%, β< 10%;
All β-proteins, α< 15%, β> 10%;
α+β-proteins, α> 15%, β> 10% with dominantly antiparallel β-sheets;
α/ β-proteins, α> 15%, β> 10% with dominantly parallelβ-sheets;and
ζ (irregular)-proteins, α< 15%, β< 10%.

Theclassification by Nakashima et al. (1986) covered 135 proteins, of which 31 were all α, 34 allβ, 27 α+β, 39α/β, and 4 ζ-proteins. It is shown in tabulated form in Table 1.

Table 1. Nakashima et al. (1986) made classification according to the following tabulated criterion:

□	α	β	p β	antip β
α	> 15	< 10	-	-
β	< 15	> 10	-	-
α + β	> 15	> 10	-	Domn
α / β	> 15	> 10	Domn	-
ζ	< 15	< 10	-	-

According to P.Y. Chou (1989), however, proteins were classified as follows:

All α-proteins, α> 45%, β< 5%;
All β-proteins, α< 5%, β> 45%;
α+β-proteins, α> 30%, β> 20% with dominantly antiparallel β-sheets;
α/ β-proteins, α> 30%, β> 20% with dominantly parallel β-sheets.

It is shown in tabulated form in Table 2.

Table 2. P.Y. Chou (1989) made classification according to the following tabulated criterion:

□	α	β	p β	antip β
α	> 45	< 5	-	-
β	< 5	> 45	-	-
α + β	> 30	> 20	-	Domn
α / β	> 30	> 20	Domn	-

The classification by P.Y. Chou (1989) covered 64 proteins: 19 all α, 15 all β, 14 α +β, and 16 α/ βproteins, but no irregular proteins.

Chouand Zhang concluded that (Chou, and Zhang, 1995) the relevant percentages given by Nakashima et al. (1986) for all α-proteins (α> 15%) and, all β-proteins (β> 10%) are not largeenough to reflect the real features of the two structural classes, and an all α- or, all β-protein should have at least α≥ 40% or β≥ 40%, respectively. Also no quantitative definitionwas given for the term "dominantly" mentioned in both of the two classification methods by Nakashima et al.(1986) , and Chou (1989), and this would certainly cause ambiguity in distinguishing α + βand α/βproteins.

In view of the above, a new classification has been proposed by P.Y. Chou (Chou, 1 995) that categorizes proteins according to the following quantitative criterion:

All α-proteins,α≥40%, β≤5%;
All β-proteins, α ≤ 5%, β≥40%;
α+β-proteins, α≥ 15%, β≥ 15% with more than 60% antiparallel β-sheets;
α/ β-proteins, α≥ 15%, β≥ 15% with more than 60% parallelβ-sheets;and
ζ (irregular)-proteins, α≤ 10%, β≤ 10%.

It is shown in tabulated form in Table 3.

Table 3. P.Y. Chou (1995)corrected his (1989) criterion as follows:

□	α	β	p β	antip β
α	≥ 40	≤ 5	–	–
β	≤ 5	≥ 40	–	–
α + β	≥ 15	≥ 15	–	≥ 60
α / β	≥ 15	≥ 15	≥ 60	–
ζ	≤ 10	≤ 10	–	–

3.SCOP EXPERTS' INTUITIVE CRITERIA

Today SCOP listed around 100.000 proteins with their structural classes. In this paper two datasets will be used to reveal the percentages of residues in α-Helices, β-sheets,parallel and antiparallel β-sheets in proteins of classes all α, allβ,α + β,and α/ β, in the classifications made by experts in SCOP.

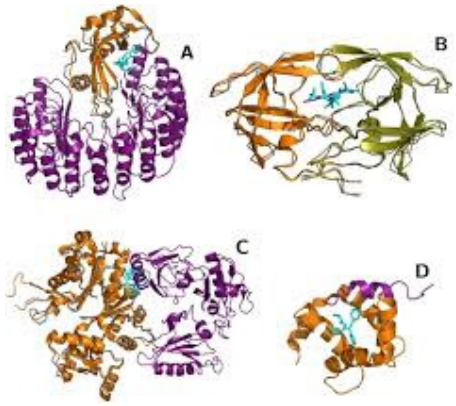


Figure 1. The four classes of protein structure

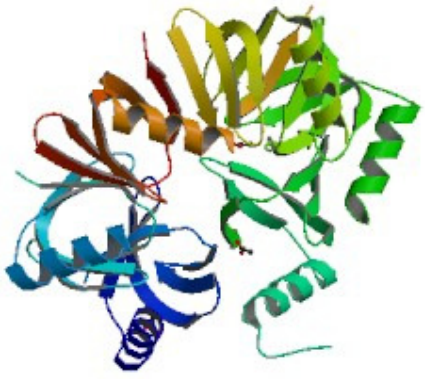
The first of the data bases is PDBselect25 which contains 1670 twilight zone proteins whose similarity is less than 25% (Hohohm, Sander 1994; Kurgan and Homaeian, 2006).). The second data base BF30 consists of 10294 proteins picked from PDB database with the similarity threshold of 30%.Structural classesof these proteins are taken from SCOP database.

3.1 Features to Specify the Percentages

Since the first proposed standard for protein structure classificationis the content of the secondary structural elements (Chou,2005), ConH and ConE were proposed to reflect the contents of H and E residues, respectively (Kurgan et al., 2008a, 2008b).

β-strands in α/β proteins are usually composedof parallel β-sheets, while the β-strands in α+β proteins are usuallycomposed of anti-parallel β-sheets, the second and the third featuresare based on the number of residues in β-strands that form parallel β-sheets(Pr) and the number of residues inβ-strands that form anti-parallel (Apr)β-sheets, respectively (Fig. 2a).

The features of the secondary structure were proposedon the basis of the structural characteristics of proteins from α/βand α+β classes.β-strands are usually separated by α-helices forming parallel β-sheets in α/β proteins,but are usually joined only by coils forming anti-parallel β-sheets in α+β proteins. Consider that the β-strands in α/β proteins are usually composedof parallel β-sheets, while the β-strands in α+β proteins are usuallycomposed of anti-parallel β-sheets, the third and the fourth featuresare based on the number of residues in β-strands that form parallel β-sheets(Pr) and the number of residues in β-strands that form anti-parallel (Apr)β-sheets, respectively (Fig. 2a). We proposed that if two β-strands(segments of E) are separated by α-helix (segments of H), these twoβ-strands would form parallel β-sheets. Otherwise, they would formanti-parallel β-sheets. Take the secondary structure sequence in example (Fig. 2b), β-strand 1 and β-strand 2 are supposed to form parallel β-sheets, and β-strand 3 and β-strand 4 are supposed to formanti-parallel β-sheets (Fig. 2c). So there are three β-strands thatform parallel β-sheets (Pr), and two β-strands that form antiparallelβ-sheets in the secondary structure sequence (Apr).



(a) A sample protein.

CCEEEECCCCHHHHHHHHHHCCCCCEEEECCCC
1 2

(b) parallel β-sheets

CCCCCCCEEEEEECCCCCEEEECCCHHHHHHHHHCCC
3 4

(c)Antiparallel β-sheets

Fig. 2. Graphical representation of the proposed determination of β-strands composing parallel β-sheets or anti-parallel β-sheets directly from protein secondary structural sequences. In the secondary sequence of the protein in the example, β-strands are labeled from 1 to 4 (Liu, and Jia2010).

3.2. Percentages of Residues in α-Helices, β-sheets,Parallel and Antiparallel β-sheets in 25PDB Data

25PDB data covers 1670 proteins, of which 443 are all α, 442all β,440α + β, and 345 α/ β proteins. Average percentages of residues in α-Helices, β-sheets,parallel and antiparallel β-sheets in proteins of classes all α, allβ,α + β,and α/ β of25PDB Data is shown in Table 4.

Table 4. Average percentages of number of residues in α-Helices, β-sheets,parallel and antiparallel β-sheets in proteins of classes all α, allβ,α + β,and α/ β of25PDB Data.

"o"	"α-Helices"	"β-Sheets"	"pr-β"	"apr-β"
"α"	62	3	0	1
"β"	10	38	8	57
"α+β"	31	23	11	23
"α/β"	40	17	15	9

It is clearly seen that SCOP experts' manual class estimations for 25PDB data do not support any of the historical sets of criteria especially in ,parallel and antiparallel β-sheets to define the structural classes of proteins.

3.3. Percentages of Residues in α-Helices, β-sheets,Parallel and Antiparallel β-sheets in BF30 data

BF30 data covers10294 proteins, of which 2438are all α, 2160all β,2887α + β, and2809 α/ β proteins. Average percentages of number of residues in α-Helices, β-sheets,parallel and antiparallel β-sheets in proteins of classes all α, allβ,α + β,and α/ β ofBF30 data is shown in Table 5.

Table 5. Average percentages of number of residues in α-Helices, β-sheets,parallel and antiparallel β-sheets in proteins of classes all α, allβ,α + β,and α/ β ofBF30 data.

"o"	"α-Helices"	"β-Sheets"	"pr-β"	"apr-β"
"α"	51	8	3	7
"β"	16	34	11	46
"α+β"	31	24	12	24
"α/β"	38	19	15	13

It is clearly seen that SCOP experts' manual class estimations for BF30 data do not support any of the historical sets of criteria to define the structural classes of proteins as well. Apparently SCOP experts' manual class estimations are conformal for both datasets.

4. Conclusion

The percentages of four features in the historical classification criteria is proven to be neither conformal with the percentages obtained from SCOP structural class predictions, nor so effective in protein structural class prediction. The number of residues inβ-strands that form anti-parallel (Apr)β-sheets, which are designed toimprove the prediction accuracy of proteins from α + βclass does not do so, instead causes confusion, since there are more anti-parallel (Apr)β-sheets in all βclass than α + β class. In addition to helix, sheet, and coin contents of proteins, recently developed features, especially inclusion of features derived from PSSM matrix, improvedthe prediction accuracy of protein classes enormously (Zhang, et al. 2011; Zhang, et al. 2014; Zhang, 2015; Ding, et al. 2014;Liu, and Jia, 2010;Liu, et. al., 2012).

REFERENCES

Chothia, C., and Finkelstein, A.V. (1990) "The Classification and Origins of Protein Folding Patterns." Annual Review of Biochemistry, Vol. 59: 1007-1035

Chou, P.Y. (1989) "Prediction of protein structural classes from amino acid composition." in Prediction of Protein Structure and the Principles of Protein Conformation, Fasman, G.D., Ed., Plenum Press, New York, 549-586

Chou, K.C. (1995) "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space." Proteins Struct. Funct. Genet., 21, 319-344.

Chou, K.C., and Zhang C. T. (1995) "Prediction of Protein Structural Classes." Critical Reviews in Biochemistry and Molecular Biology, 30(4): 275-349.

Chou, K.C., Nemethy, G. and Scheraga, H.A. (1990) "Energetics of interactions of regular structural elements in proteins." Acc. Chem. Res., 23, 134-141.

Chou, K.C., (2005). "Progress in protein structural class prediction and its impact to bioinformatics and proteomics". Curr. Protein Pept. Sci. 6, 423–436.

Chou, P.Y. and Fasman, G.D. (1974) "Prediction of protein conformation." Biochemistry 13, 222-245.

Chou, P.Y. and Fasman, G.D. (1978) "Prediction of secondary structure of proteins from amino acid sequence." Adv. Enzymol. Relat. Subj. Biochem. 47, 45-148.

Cid, H., Bunster, M., Canales, M., and Gazitua, F. (1992) "Hydrophobicity and structural classes in proteins." Proteins Eng., 5, 373-375.

Ding, S., Li Y., Shi Z., and Yan S., (2014) "A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile." Biochimie 97, 60-65.

Fasman, G.D. (1989) "The development of the prediction of protein structure." in Prediction of Protein Structure and the Principles of Protein Conformation, Fasman, G.D., Ed., Plenum Press, New York, 317-358.

Finkelstein, A.V., and Ptitsyn, O.B. (1987) "Why do globular proteins fit the limited set of folding patterns?" Prog. Biophys. M ol. Biol., 50, 171-190.

Garnier, J., Osguthorpe, DJ. and Robson, B. (1978) "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins." J. Mol. Biol. 120, 97-120.

Gilson, M.K. and Honig, B. (1988) "Energetics of charge-charge interactions in proteins, Proteins Struct. Funct. and Genet., 3, 32-52.

Hobohm, U. and Sander, C. (1994) "Enlarged representative set of protein structures." Protein Sci. 3, 522-524.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1994) "A model recognition approach to the prediction of all-helical membrane protein structure and topology." Biochemistry, 33, 3038-3049.

Karplus, M. and Shakhnovich, E. (1992) "Theoretical studies of thermodynamics and dynamics." in Protein Folding, Creighton, T.E., Ed., Freeman, New York, 127-195.

Kolinski, A. and Skolnick, J. (1994) "Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme." Proteins Struct. Funct. Genect, 18, 338-352.

Kurgan, L.A., and Homaeian, L. (2006) "Prediction of structural classes for protein sequences and domains-Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy", Pattern Recognition 39, 2323 – 2343.

Kurgan, L.A., Zhang, T., Zhang, H., Shen, S., Ruan, J., (2008a) "Secondary structure based assignment of the protein structural classes". Amino Acids 35, 551–564.

Kurgan, L.A., Cios, K., Chen, K., (2008b) "SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences". BMC Bioinformat. 9, 226-241.

Kuwajima, K, Semisotnov, G.V., Finkelstein, A.V., Sugai, S. and Ptitsyn, O.B. (1993) "Secondary structure of globular proteins at the early and the final stages in protein folding." FEBS Lett., 334, 265-268. (Published erratum appears in FEBS Lett., 1993, 336, 190.)

Levitt, M. (1983) "Protein folding by restrained energy minimization and molecular dynamics." J. M ol. Biol., 170, 723-764.

Levitt, M. and Chothia, C. (1976) "Structural patterns in globular proteins." Nature, 261, 552-557.

Lim, V.I. (1974) "Structural principles of globular protein secondary structure." J. Mol. Biol. 88, 857-872.

Liu, T., and Jia, C. (2010). "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information". Journal of Theoretical Biology, 267, 272–275.

Liu, T.G., Geng, X.B., Zheng, X.Q., Li, R., Wang, J. (2012) "Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles", Amino Acids 42, 2243–2249.

Mackay, D.H.J., Cross, A.J., and Hagler, A.T. (1989) "The role of energy minimization in simulation strategies of biomolecular systems." in Prediction of Protein Structure and the Principles of Protein Conformation, Fasman, G.D., Ed., Plenum Press, New York, 317-358.

McCammon, J.A., Wong, C.F., and Lybrand, T.P. (1989) "Protein stability and function." in Prediction of Protein Structure and the Principles of Protein Conformation, Fasman, G.D., Ed., Plenum Press, New York, 149-159.

McDonald, I.K. and Thornton, J.M. (1994) "Satisfying hydrogen bonding potential in proteins." J. Mol. Biol. 238, 777-793.

Mitchell, J.B., Nandi, C.L., McDonald, I.K., Thornton, J.M., Price, S.L. (1994) "Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding?" J. Mol. Biol. 239, 315-331.

Nakashima, H., Nishikawa, K., and Ooi, T. (1986) "The folding type of a protein is relevant to the amino acid composition." J. Biochem., 99, 152-162.

Orengo, C.A., Jones, D.T., and Thornton, J.M. (1994) "Protein superfamilies and domain superfolds." Nature, 372, 631-634.

Ptitsyn, O.B. and Finkelstein, A.V. (1979) "Mechanism of protein folding." Int. J. Quantum Chem., 16, 407-418.

Ptitsyn, O.B. and Finkelstein, A.V. (1980) "Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding?" Q. Rev. Biophys., 13, 339-386.

Ptitsyn, O.B., and Finkelstein, A.V. (1989) "Prediction of protein secondary structure based on physical theory. Histones." Protein Eng., 2, 443-447.

Ptitsyn, O.B. and Rashin, A.A. (1975) "A model of myoglobin self-organization." Bio- phys. Chem., 3, 1-20.

Ptitsyn, O.B., Finkelstein, A.V., and Murzin, A.G. (1985) "Structural model for interactions." FEBS Lett., 186, 143-148.

Richardson, J.S. and Richardson, D.C. (1989) "Principles and patterns of protein conformation." in Prediction of Protein Structure and the Principles of Protein Conformation, Fasman, G.D., Ed., Plenum Press, New York, 1-98.

Rogers, N.K. (1989) "The role of electrostatic interactions in the structure of globular proteins." in Prediction of Protein Structure and the Principles of Protein Conformation, Fasman, G.D., Ed., Plenum Press, New York, 359-389.

Scheraga, H.A. (1968) "Calculations of conformations of polypeptides." Adv. Phys. Org. Chem. 6, I 03-184.

Scheraga, H.A. (1987) "Conformational analysis of polypeptides and proteins for the study of protein folding, molecular recognition, and molecular design." J. Prot. Chem. 6, 61-80.

Vieth, M., Kolinski, A, Brooks, C.L. and Skolnick, J. (1994) "Prediction of the folding pathways and structure of the GCN4 leucine zipper." J. Mol. Biol. 237, 361-367.

Weiner, P.K. and Kollman, P.A. (1981) "AMBER: Assisted model building with energy refinement. a general program for modeling molecules and their interactions." J. Comp. Chem. 2, 287-303.

Zhang S. (2015) "Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC", Chemometrics and Intelligent Laboratory Systems 142, 28–35

Zhang, S.L., Ding, S.Y. , and Wang, T.M. (2011) "High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure", Biochimie 93, 710–714.

Zhang, L.C., Zhao, X.Q. , and Kong L. (2014) "Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition", J. Theor. Biol. 355 105–110.