Protein Structural Classes Prediction via Residues Environment Profile

KUO-CHING HSIAO[1], CHIEN-HUNG HUANG[2] AND KA-LOK NG[1,*]
[1]*Department of Biotechnology and Bioinformatics, Asia University, Taiwan*
[2]*Department of Computer Science and Information Engineering, National Formosa University, Taiwan*

ABSTRACT

We investigate how residue structural and physicochemical environment information, such as the protein secondary structure and residue solvent accessibility could be used for protein structural classes (all-alpha, all-beta, alpha/beta and alpha+beta) prediction. The residue environment information is described by the residue environment profiles which are derived from a relative small set of 500 protein sequences having a sequence identity less than 25%. It was demonstrated that this method is able to obtain an accuracy of 49.2% for a 4-type class prediction of monomeric and non-disulphide-bonded proteins, given the fact that none of the nonclassified protein sequences has a sequence identity higher than 25%. This result is comparable to the amino acid composition method which obtains an accuracy of 48% for a set of sequences having sequence similarity of less than 30%. The current approach has several advantages: (1) it is a physical approach, (2) there is no adjustable parameter, and (3) it is simple and efficient.

Key words: protein classes, solvent accessible area, residue environment, protein structural profile.

1. INTRODUCTION

Some of the common approaches used to predict the structure of a protein are homology modeling, knowledge-based fold recognition (threading) and the *ab initio* method. It was noticed that homology modeling fails if the query protein sequence and the target sequence have a sequence similarity below 25%. The *ab initio* method is difficult to implement because of the enormous number of protein conformations needed to search, and the inadequacies of the potential energy function to measure the free energy of the protein solvent system (Fisher, Rice, Bowie, & Eisenberg, 1996). A knowledge-based fold recognition method relies on the extraction of statistical parameters from an experimentally determined protein structure database and it has demonstrated some successes (Frishman & Argos, 1995; Sippl, 1995).

Although the sequence homology approach has some successes in predicting the 3D structure, this approach is a non-physical method, that is, it does not take into account the residues' structural information. It is known that the protein structure is more conserved than the sequence (Chothia & Lesk, 1986); therefore, prediction of protein structures based on structural similarity is less sensitive to specific sequence information. There have been attempts to match sequences to folds by describing folds in terms of the environment of each residue in the structure. For instance, the environment was described in terms of local secondary

* Corresponding author. E-mail: klng@asia.edu.tw

structures, residue solvent accessibility, and the degree of burial by polar rather than non-polar atoms (Bowie, Luthy, & Eisenberg, 1991). Several studies have indicated that the environment approach could perform better than the purely sequence-based method (Zhang & Kim, 2000; Chang, Cieplak, Dima, Maritan, & Banavar, 2001).

In this paper, we propose to extract residue environment information from an experimentally determined protein secondary structure database, such as DSSP (Kabsch & Sander, 1983). As a first approximation, we neglect intra and inter-chain interactions and consider monomeric and non-disulphide-bonded proteins. The same approximation has been employed in other works (Chou & Maggiora, 1998; Wang & Yuan, 2000). Then, five environment structure profiles are computed, one for each individual structural class (that is all-α, all-β, α/β and $\alpha+\beta$ according to the SCOP classification (Murzin, Brenner, Hubbard, & Chothia, 1995) plus the 4-classes structural profile where all four classes are included. These profiles are used to score the query protein sequence to be modeled for compatibility with the known structural classes. To demonstrate that the 3D structure profile method is able to detect sequences compatible with a known structural class, we align the query sequences with the environment of known protein structural classes. This will establish the fact that the structure profile approach is able to classify structural classes for distant sequences well below the twilight zone (sequence similarity lower than 25%). Our investigation shows that the residue environment information approach obtains a slightly better level of accuracy than the amino acid composition approach (Wang & Yuan, 2000) for the 4-classes prediction.

2. METHOD

2.1 Non-redundant Date Set

A set of monomeric and non-disulphide-bonded protein sequences were selected from the DSSP database by referring to the SCOP classification as our input data set. The DSSP database utilizes the DSSP program to define secondary structures, geometrical features and solvent accessible areas of proteins given atomic coordinates in the PDB (Berman et al., 2000). In the SCOP database, proteins are classified in a hierarchy according to their evolutionary origin and structural similarity. According to the public available information, there were 20,619 PDB entries and 54,745 domains annotated in SCOP release 1.65. In our study we considered four structural classes of proteins: all-α proteins, all-β proteins, $\alpha+\beta$ proteins and α/β proteins (multi-domain proteins, membrane and cell surface proteins, and small proteins could also be considered if needed).

We extracted the environment information, that is the residue solvent accessible surface areas (buried (B), partly buried (PB) and exposed (E)) and secondary structure (such as α-helix, β-sheet and coil) data from the DSSP database. For instance, we used the following nine environment classes, that is, (B,

P, E)$_{\alpha,\beta,C}$, in our study, where c stands for the coiled region. The cut-offs used in defining buried (B), partly buried (P) and exposed (E) were taken to be 0~8%, >8~38% and >38% of the maximal accessibility for the residues (Bowie, Clarke, Pabo, & Sauer, 1990; Bowie et al., 1991; Rost & Sander, 1994).

It is known that the protein sequences deposited in the PDB could possibly have a sequence similarity higher than 25%. In order to count the protein contribution once, we filtered out protein sequences which had a sequence identity higher than 25% because sequences with high similarity scores will be over-represented in the profile calculation. The PDB_SELECT database (Hobohm, Scharf, Schneider, & Sander, 1992; Holm & Sander, 1998) provides files of protein sequences with less than 25% sequence similarity. We compared the PDB_SELECT file and the DSSP sequences file. If there were more than one sequence from the DSSP set that had a sequence identity greater than 25%, we kept only one representative sequence and repeated this filtering process for all the four structural classes.

2.2 Residue Environment Structure Profile

The 3D profile method uses structural information (Bowie et al., 1990; Bowie et al., 1991). Instead of doing sequence alignment, the 3D profile method aligns a sequence to a string of descriptors that describe the 3D environment of the target structure. That is, for each residue position in the structure we determine:

1. solvent accessible surface areas (Lesk, 2001),
2. the local secondary structure (α-helix, β-sheet and coil), and
3. the fraction of surrounding environment that is buried, partly buried or exposed.

The basic assumption of this method is that the environment of a particular residue is expected to be more conserved than the actual residue itself, and so the method is able to detect more distant sequence-structure relationships than a purely sequence-based method.

The probability, $P(i, j)$, associated with residue j in an environment i (in our study it is the solvent accessible area A) is given by

$$P(i, j) = n(i, j) / N(j) \tag{1}$$

where $n(i, j)$ is the number of residues j with solvent accessible area A, and $N(j)$ is the total number of residues j.

For instance, one can compute the probability of having residue j's solvent accessible surface area in a buried, partly buried or exposed environment with one of the three secondary structures. Thus, for each residue position the 3D protein structure is assigned to one of the nine environment classes. The residue solvent accessible surface area and secondary structure data are retrieved from the DSSP database.

The score matrix element of the 3D residue structure profile, M_{ij}, for environment class i and residue j is given by;

$$M_{ij} = \ln\left(\frac{P(\textit{residue j in environment i})}{P(\textit{residue j in any environment})}\right) \tag{2}$$

The denominator in Eq. (2) is obtained from the residue's frequency in the DSSP database, where *j* is one of the nine environment classes; $(B, P, E)_{\alpha,\beta,C}$.

Given the scoring matrix for a class of proteins, we built a 3D profile for a particular structural class using this matrix. That is, for each position in the known protein structure we determined its environment class. The score of a particular residue in this position is given by the score matrix value. For example, if the first position in our structure has the environment class 'buried,' the score of having residue *j* in that position is the corresponding score matrix value M_{Bj}. Thus, if there are *n* residues in the structure, we could build a profile for the known protein structure. To align a sequence with a structural class, we align the sequence with the descriptors of the 3D environment of the known protein structure. The optimal alignment is defined by the following score function,

$$S_{QT} = \min\left(\sum_{i=1}^{m}\sum_{j=1,\ldots,|n-m+1|}^{m\ldots n} |\, q_i - t_j \,|\right) \tag{3}$$

where Q and T denote the sets of query and target sequences with lengths m and n respectively, $m<n$ and $0.5m<|m-n|<1.5m$ (if $m>n$ one simply interchanges m and n in Eq. (3)), where $Q = \{q_1, q_2, \ldots, q_m\}$ and $T = \{t_1, t_2, \ldots, t_n\}$ where q_i and t_j denote the score matrix values, respectively. This alignment method takes into account the residue's neighborhood effect. Since the length of the protein sequences in the alignment could be very different, we defined the normalized score function Δ_{QT} by

$$\Delta_{QT} = \frac{S_{QT}}{L} \tag{4}$$

where L denotes the overlapping length of the alignment. The value Δ_{QT} lies between 0 and 0.9. A score value close to zero and 0.9 indicates a perfect match and no match, respectively.

3. RESULTS

Among the 24,039 protein sequences from the DSSP, 5,483 are monomeric and non-disulphide-bonded sequences. Also, among the 1,171 monomeric, non-disulphide-bonded protein sequences, 578 sequences have less than 30% sequence identity.

In Figure 1 we plotted the sequence identity versus the normalized score value Δ_{QT} for the set of 578 sequences. It is evident from Figure 1 that all the protein sequences have a sequence identity of less than 25%, except three outliers. Most of the sequences' Δ_{QT} values lie between 0.50 and 0.70.

These 578 sequences were chosen to be our target set *T*, and were used to construct the environment structure profile and the four structural classes (4-classes) structure profiles. After the monomeric, non-disulphide-bonded filtering process, we were then left with 4,905 sequences (i.e. 5,483 minus 578), which serve as our query set *Q*.

In Figure 2 we plotted the score matrix value M_{ij} for buried, partially buried and exposed environments against the 20 residue types (from hydrophoblic to hydrophyllic) for the all-α proteins.

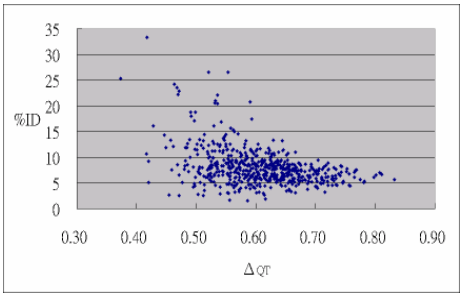
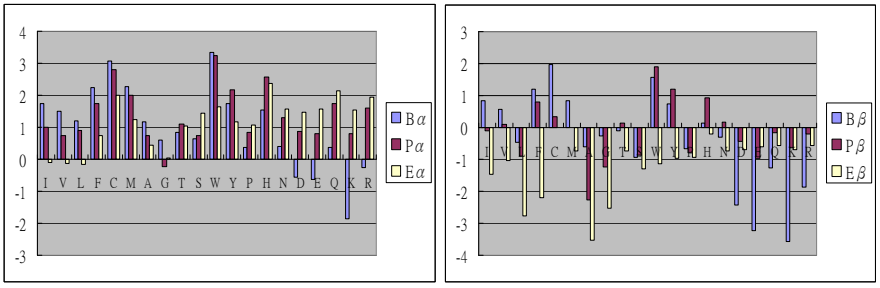
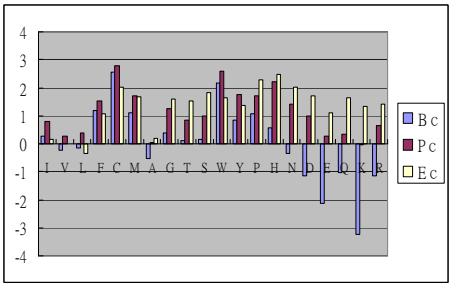


Figure 1. A plot of sequences identity verse the normalized score value Δ_{QT} for a set of 578 monomeric, non-disulphide-bonded proteins.



(a) (b)



(c)

Figure 2. The structural class structure profile of all-α proteins for (a) α-helix structure, (b) β-sheet structure and (c) coiled structure.

A large score value indicates a strong preference for the particular environment whereas a small or negative score value indicates an aversion. It is evident from Figures 2(a) through (c), the hydrophobic residues (I, V, L, F, C, M, A, G and Y) have a large score for the buried state, that is, they were found to prefer to reside in the buried state, whereas hydrophilic residues (acidic: D, E, basic: K, R and polar N, Q) have a large score for the exposed state. These results are consistent with the experimental determined hydrophobicity result (Kyte & Doolittle, 1982). Due to space limitations, the all-β, α/β, α+β structural profiles and the 4-classes structural profile were omitted.

In Table 1, we summarize the values of our predicted results for the four structural classes. The second column lists the different type of structural classes, the third column displays the total number of sequences in the query and target sets, and the fourth, fifth and sixth columns summarize our results using different Δ_{QT} cut-off score values, that is, 0.5, 0.6 and 0.9. The matching column indicates how many sequences were assigned to a structural class correctly, and the ratio column shows the prediction accuracy percentage.

To test the validity of our approach, we divided the 578 sequences into two parts randomly, one contained 78 sequences and the other contained 500 sequences. The 500 sequences were chosen to construct the environment structure profile as well as the 4-classes structure profiles. At the $\Delta_{QT} < 0.90$ level it was found that 30 out of 61 sequences were assigned to the correct structural class, which is a prediction accuracy of 49.2%. The other 17 sequences (from the set of 78 sequences) were not counted since they belong to other SCOP classes (such as multi-domains or membrane proteins). From the column of Δ_{QT}<0.6 in Table 1, we achieved a slightly lower prediction accuracy of 47.8%. This means that the result did not degrade as Δ_{QT} increased, which indicates that the current approach is robust against sequence alignment.

The third row summarizes the structural classes prediction accuracy using the 578 sequences as our target set and the 4,905 sequences as our query set, and it was found that we achieved a 44.6% prediction accuracy at the $\Delta_{QT} < 0.90$ level.

Table 1. *A summary of the structural classes prediction for sequences of less than 25% identity, all the monomeric, non-disulphide-bonded sequences, the four structural classes and the 4-classes*

1		*Q : T*	$\Delta_{QT} < 0.50$		$\Delta_{QT} < 0.60$		$\Delta_{QT} < 0.90$	
			matching	ratio	matching	ratio	matching	ratio
2	<25%	78: 500	4/7	57.1%	11/23	47.8%	30/61	49.2%
3	All	4,905 : 578	418/528	79.2%	616/954	64.6%	1607/3603	44.6%
4	all-α	4,905 : 160	136/181	75.1%	221/372	59.4%	1414/4816	29.4%
5	all-β	4,905 : 126	86/109	78.9%	140/274	51.1%	1762/4785	36.8%
6	α/β	4,905 : 91	34/34	100.0%	42/51	82.4%	1862/4749	39.2%
7	α+β	4,905 : 127	97/104	93.3%	132/235	56.2%	2086/4835	43.1%
8	4-classes	4,905: 504	349/410	85.1%	479/695	68.9%	1950/4892	39.9%

In Figure 3 we plotted the sequence identity versus the normalized score value Δ_{QT} for the set of 4905 sequences. It is evident from Figure 3 that most of the protein sequences have a sequence identity of less than 25% and the sequences' Δ_{QT} values lie between 0.50 and 0.80.

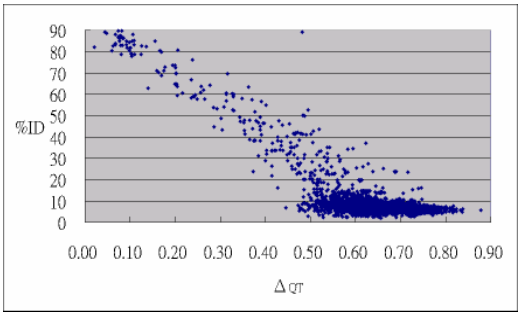


Figure 3. Sequences identity versus the normalized score value Δ_{QT} for a set of 4,905 monomeric, non-disulphide-bonded proteins.

The fourth, fifth, sixth and seventh rows summarize the structural classes prediction accuracy using the individual score matrix. It was found that our approach could achieved a 29.4% to 43.1% prediction accuracy. This prediction accuracy could possibly increase if a larger training set were used.

The eighth row summarizes the 4-classes prediction accuracy using the 504 sequences as our target set (only all-α, all-β, α/β and α+β proteins were selected and trained), and the 4,905 sequences as our query set. It was found that a 39.9% prediction accuracy was achieved, which is the same level of accuracy as the all-α, all-β, α/β and α+β prediction. This result suggests that using the 4-classes structural profile gives the same level of prediction accuracy as the individual structural class profile.

At a first glance, it seems that the prediction accuracy is a bit lower since there were reported accuracies as high as 70% by using the amino acid composition approach (Chou & Maggiora, 1998). However, it was pointed out (Wang & Yuan, 2000) that knowledge of amino acid composition alone cannot lead to a success rate higher than 60% for a 4-type class prediction. The main reason for this was due to preselection of query sets in the previous works.

From Table 6 of Wang and Yuan (2000), the authors had obtained an average accuracy of 48% for a 4-type class prediction, whereas a slightly better prediction accuracy of 49.2% was obtained in the present study. We notice that there is a major difference between our training set and the set used in Wang and Yuan (2000), in which the authors used single domain proteins for their training. It is known that proteins can have more than one domain, hence such an assumption is not valid in general. We do not make such an assumption in this work.

Furthermore, a comparison of our results with those reported by Li, Wang, Fan, & Wang (2003), indicated that the present work achieved a better level of

accuracy (49.2% versus 34.9%; see Table 2 in Li et al. (2003)) than the amino-acid-grouping scheme assuming an eight (N=8) amino acid groups calculation. For the case study of nine (N=9) amino acid groups, no data was reported by Li et al. (2003). In the case of ten (N=10) amino acid groups calculation, the work by Li et al. (2003) achieved a better level of accuracy than our work. However, the number of protein sequences that could be assigned to the α/β class is rather low. For instance, among the 30 testing protein sequences only two are identified as belonging to the α/β class for N=10, 11 and 12.

6. DISCUSSION AND CONCLUSIONS

We investigate how residue structural and physico-chemical environment information, such as the protein secondary structures and residue solvent accessibility, could possibly be used for protein structural classes (all-α, all-β, α/β and α+β) prediction. The residue environment information is described by the 3D residue environment profiles which are derived from protein sequences having a sequence identity of less than 25%.

The score values of the environment profiles for all-alpha, all-beta, alpha/beta, alpha plus beta proteins and the 4-classes are computed. We apply this approach to monomeric and non-disulphide-bonded proteins, and demonstrate that this approach is able to predict structural classes with an accuracy of 49.2%, given the fact that none of the nonclassified protein sequences have a sequence identity greater than 25%. Our result is slightly better than the amino acid composition method and the amino-acid-grouping method for the eight amino acid groups calculation.

The current approach has several advantages: (1) it is a physical approach, (2) there is no adjustable parameter, whereas composition models have to decide the optimal length of the subsequences, and (3) it is simple and efficient. There are several areas where we could extend our analysis: (1) extend the size of the training set from 578 to about a thousand, (ii) improve the alignment scoring function by considering a sliding window of size three, that is, replace $|q_i-t_j|$ in Eq. (3) with $(|q_{i-1}-t_{j-1}| + |q_i-t_j| + |q_{i+1}-t_{j+1}|)/3$, taking into account the residue's neighborhood effect, and (3) introduce affine penalty in the alignment.

ACKNOWLEDGMENTS

The authors would like to thank the National Science Council (NSC) for financial support. The work of Ka-Lok Ng is supported by NSC 95-2745-E-468-006-URD. The work of Chien-Hung Huang is supported by NSC 95-2221-E-150-025-MY2.

REFERENCES

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235-242.

Bowie, J., Clarke, N. D., Pabo, C. O., & Sauer, R. T. (1990). Identification of Protein Folds: Matching Hydrophobicity Patterns of Sequence Sets with Solvent Accessibility Patterns of Known Structures. *Proteins: Structure, Function, and Genetics, 7*(3), 257-264.

Bowie, J. U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science, 253*(5016), 164-170.

Chang, I., Cieplak, M., Dima, R., Maritan, A., & Banavar, J. (2001). Protein threading by learning. *Proceedings of the National Academy of Sciences*, USA 987, 14350-14355.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal, 5*(4), 823-826.

Chou, K. C., & Maggiora, G. M. (1998). Domain structural class prediction. *Protein Engineering, 11*, 523-538.

Fisher, D., Rice, D., Bowie, J. U., & Eisenberg, D. (1996). Assigning amino acid sequences to 3-dimensional protein folds. *The FASEB Journal, 10*(1), 126-136.

Frishman, D., & Argos, P. (1995). Knowledge-based secondary structure assignment. *Proteins: structure, function and genetics, 23*(4), 556-579.

Hobohm, U., Scharf, M., Schneider, R., & Sander, C. (1992). Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Science, 1*, 409-417.

Holm, L., & Sander, C. (1998). Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics, 14*(5), 423-429.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*(12), 2577-2637.

Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157*, 105-132.

Lesk, A. M. (2001). *Introduction to protein architecture*. Oxford, UK: Oxford University Press.

Li, T., Wang, J., Fan, K., & Wang, W. (2003). How simple can the proteins be: from the prediction of the classes of protein structures. *Modern Physics Letters B, 17*(5), 1-8.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: a structural classification of protein database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247, 536-540.

Rost, B., & Sander, C. (1994). Conservation and predication of solvent accessibility in protein families. *Proteins: Structure. Function and Genetics 20*(3), 216-226.

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, *5*(2), 229-235.

Wang, Z. X., Yuan, Z. (2000). How Good Is Prediction of Protein Structural Class by the Component-Coupled Method? *Proteins: Structure. Function and Genetics*, *38*, 165-175.

Zhang, C., & Kim, S. H. (2000). Environment-dependent residue contact energies for proteins. *Proceedings of the National Academy of Sciences*, USA 987, 2550-2555.

Kuo-ching Hsiao received a B.Eng. degree in electrical engineering from Chung Chou Institute of Technology, Taipei, Taiwan in 2003, and an M.Sc. in bioinformatics from Asia University (the former Taichung Healthcare and Management University) in 2005.

He is working at the Chin-fon Bank as a Financial Assistant Section-Chief. His interests include studying herb health food, analysis of protein structures and information technology.

Chien-Hung Huang received a B.S. degree in computer science from Tatung University, Taipei, Taiwan in 1991, and a Ph.D in computer and information engineering from National Tsing Hua University, Hsinchu, Taiwan in 1999.

He has been an Assistant Professor at the Department of Computer and Information Engineering, National Formosa University since August 2004. He has research funding and has published articles in the areas of Linux/FreeBSD Squid proxy server, graph embedding algorithms and biological networks. His research interests include parallel computing, graph algorithms and open source distribution.

Ka-Lok Ng received an Honours diploma from Hong Kong Baptist College in 1983, and a Ph.D. degree in theoretical physics from the Vanderbilt University in the USA in 1990. He has been an Associate Professor in the Department of Bioinformatics, Asia University since August 2003. He has research funding and has published articles in the areas of protein-protein interaction networks, topological and robustness study of biological networks, modeling of DNA sequences, cosmic microwave background radiation and neutrino study. His research interests include protein-protein interaction networks, microRNA target prediction, microarray time series, biochemical network simulation, and cancer biology.