Residues environment study - Solvent accessibility in globins protein family

Chia-le Wu

Ling Tung College

Department of Information Management

1 Ling Tung Road, Nantun, Tai-chung, Taiwan 408

(O) 04-23895624 ext. 421

(M) 0916-706900

(Fax) 0946-103192

albert@mail.ltc.edu.tw

Abstract

Attempts to match protein sequences to folds by describing the fold in terms of the environment of each residue in the structure could be a more efficient way to detect structural similarities in divergent proteins. In this report, the environment of each residue is described in terms of solvent accessibility. In an aqueous environment, hydrophobic residues tend to segregate into the core of the protein whereas polar residues remain exposed on the protein surface. In order to quantify hydrophobic residues burial, the residue solvent accessible area (SAA) is introduced. We computed the environment score of a particular residue resides in a buried, partly buried or exposed environment. Our study used the globins protein family, which belongs to the all-alpha class in the SCOP database. The SAA data are retrieved from the DSSP database. We obtained a 20x3 scoring matrix that describes the preference of a particular residue resides in one of the three buried states. This matrix allows us to construct a 3D profile for the globins protein and used as a template for the threading method; that is for each position in the sequence one can determine its environment class and the score of a particular residue in this position.

Keywords : solvent accessible surface, hydrophobicity, threading method, SCOP, DSSP

1. Introduction

In recent years, the numbers of biological DNA and protein sequences are producing at an accelerating rate. These sequences can provide a first step in understanding the biological functions of these genetic sequences. Knowing the amino acid sequence it is an intensive and challenging area of research in predicating the corresponding 3D structure. In the PDB databank [Berman et al. 2000], the 3D structures of 18188 proteins are known as of July 9, 2002.

One of the many approaches to predict the protein structure is the sequence-structure alignment approach. Although this approach has some successes in predicating the 3D structure, however, this approach is non-physical method, that is, it does not take into account of structure information. Recently, there were attempts to match sequences to folds by describing the fold in terms of the environment of each residue in the structure. The environment was described in terms of solvent accessibility, local secondary structures and the degree of burial by polar rather than non-polar atoms. Several studies indicated that the environment approach could perform better than the purely sequence-based method [Zhang and Kim 2000, Chang et al. 2001].

In section 2, we give an introduction on the solvent accessibility definition. In section 3, we will give a brief description of the 3D profile method, and define the environment score value. In section 4, we present our results on assigning an environment score to each residue according to its solvent accessible environment. In section 5, we summarize and discuss our results.

2. Solvent Accessibility

In an aqueous environment, hydrophobic residues tend to segregate into the core of the protein whereas polar residues remain exposed on the protein surface. The patterns of hydrophobic and polar residues in a sequence appear to be a strong driving force of protein folding. This is because the folded conformation is found by identifying the structure exhibiting the global minimum of the free energy. The free energy of the protein folding process in solution received many types of interactions, including those of intramoleular origin or resulting from solvent effects. In the all atom model, the solvation energy contributions are generally believed to be a significant

force in stabilizing the native conformations of a protein [Chothia 1976, Dill 1990]. The solvation energy of a molecule can be estimated by considering the free energy of transfer, ΔG_{sol} , of amino acid side chains from nonpolar to aqueous solution, and it is given by [Chothia 1974],

$$\Delta G_{sol} = \Sigma A_i \sigma_i$$

where A_i is the ith atomic solvent accessible surface area, and σ_i is the ith atomic solvation parameter (ASP) also called hydrophobicity. Each Angstrom square surface area contributes about 105 J of free energy of stabilization. The hydrophobicity parameters can be determined from molecule crystal geometry [Rees & Wolfe 1993] and from the distribution function of the areas of hydrophobic solvent accessible surface region in protein structures [Eisenhaber 1996].

In order to quantify hydrophobic burial (Chothia, 1974), B.K. Lee and Fred Richards introduced the solvent accessible surface (Lee and Richards, 1971, Richards 1977). The accessible surface is traced out by the probe sphere center as it rolls over the protein. It is a kind of expanded van der Waals surface.

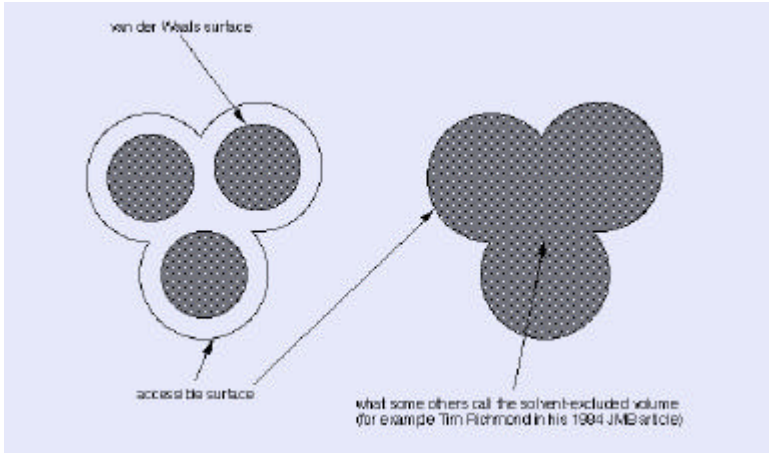

Fig. 1. Solvent accessible surface definition of a protein.

The solvent accessible surface (SAS) of a molecule is defined as the surface area of the molecule exposed to solvent. More specifically, our definition follows that proposed by Lee and Richards (1971), which is described by the center of a solvent sphere (called a probe) rolled over the surface of the molecule. The use of the center of the probe means that the surface area is directly proportional to the number of solvent molecules that could fit on the surface.

3. 3D profile method and the probability distributions of residues surface areas

3.1 The 3D profile method

The 3D profile method uses structural information [Bowie et al. 1990]. Instead of doing sequences alignment, the 3D profile method aligning 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:

- how buried it is (buried, partly buried or exposed)
- the local secondary structure (α-helix, β-sheet and coil)
- the fraction of surrounding environment that is polar

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

3.2 Score matrix of residues solvent accessible area (SAA)

In this report, we compute the probability of a particular residue's SAA resides in a buried, partly buried or exposed environment. Our study used the myoglobin domain, a globins protein family, which belongs to the all-alpha class in the SCOP database [Murzin et al. 1995]. The SAA data are retrieved from the DSSP database (Kabsch and Sander 1983), where DSSP is a database of secondary structure assignments for all protein entries in the Protein Data Bank (PDB). The DSSP program defines secondary structure, geometrical features and solvent exposure of proteins, given atomic coordinates in PDB format.

A sample of 136 proteins, with a total of 20713 resides are analyzed. The cutoffs used in defining buried, partly buried and exposed are taken to be 0-10%, 11-40% and >40% (Bowie et al. 1990) of the maximal accessibility for the residues (Rost and Sander 1994). The probability, P(i,m), associated with residue i in an environment m (in our study it is the solvent accessible area A) is given by

$$P(i,m) = n(i,m) / N(i)$$

where n(i,m) is the number of residue i with solvent accessible area A, and N(i) is the total number

of residue i.

For each position in the protein structure, the values of the SAA were divided into three categories of buried, partly buried and exposed. For all proteins in DSSP, we can tabulate the number of times we observed a particular residue in a particular environment class, and use this to compute a score, S_{ij}, for each environment class and each residue pair;

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

4. Environment Score Value Results

The environment scores, S_{ij}, for each residue is depicted in Table 1. A cross x denotes that the probability of finding a particular residue in that environment class is zero (the logarithm of zero is ill define).

Table 1. Environment class and score value, S_{ij},of the residues

Residue	Buried (B)	Partly buried (PB)	Exposed (E)
I	-2.27	-1.99	4.27
V	-2.29	-1.95	-0.88
L	-1.68	-0.93	0.39
F	-2.7	-2.33	3.48
C	x	x	x
M	-3.84	-2.42	-0.73
A	-1.15	-0.43	-1.46
G	-1.63	-1.8	-1.05
T	-1.82	0.6	-3.18
S	-1.46	-2.47	-2.24
W	-4.31	0.62	x
Y	-2.85	-3.4	-1.2

P	x	-2.25	-3.36
H	-1.46	-1.04	-1.77
N	x	-4.49	-1.82
D	x	-2.46	-2.51
E	x	-1.18	-2.04
Q	0.7	-2.54	-2.85
K	5.07	-0.87	-1.74
R	1.97	-3.29	-2.39

A large negative score value indicates a strong preference for the particular environment whereas large positive score value indicates an aversion. It is evident from the Table 1 that residues P, N, D, and E were not found at the buried state. Furthermore, residues Q, K and R all have a positive score, which is an indication of an aversion. These are the polar residues. Similarly, residues I, V, L, F and M are found to prefer reside in the buried state by examining the exposed column score values (a large positive or small negative value) in Table 1. These are the hydrophobic residues. These two conclusions are well consistent with the experimental determined hydrophobicity scale results (Kyte & Doolittle, 1982). It is interesting to note that the C (Cystine) residue does not occur in the whole globins protein family.

Given the score matrix one can build a 3D profile for a particular structure using this score matrix. For each position in the structure we determine its environment class and the score of a particular residue in this position depends on the score matrix we built. For instance, if the first position in our structure has the environment class buried, the score of having residue K in that position is 5.07. Thus, if there are n residues in the structure, we can build a profile for the target structure. Then to align a sequence s with a structure, we align the sequence with the descriptors of the 3D environment of the target structure. To find the best alignment, on can use the dynamic programming algorithm to find the optimal alignment.

5. Summary and Discussion

In this report, we compute the probability of finding a particular residue's SAA resides in a buried, partly buried or exposed environment. Our study used the globins protein family, which belongs to the all-alpha class in the SCOP database. The SAA data are retrieved from the DSSP database. We obtained a 20x3 scoring matrix (Table 1) that describes the preference of a particular residue resides in one of the three buried states.

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 pp.235-242.

Bowie J., Clarke N. D., Pabo C. O. and Sauer R.T. 1990. Identification of Protein Folds: Matching Hydrophobicity Patterns of Sequence Sets with Solvent Accessibility Patterns of Known Structures. Proteins, 7, pp.257-264.

Chang I., Cieplak M., Dima R., Maritan A. & Banavar J. 2001 Protein threading by learning. Proc. Natl. Acad. Sci, USA 987 pp.14350-14355.

Chothia C. 1974. Hydrophobic bonding and accessible surface areas in proteins. Nature 248 : pp. 338-339.

Chothia C. 1976. The nature of the accessible and buried surface in proteins. J. Mol. Biol. 105 pp. 1-14.

Dill K.A. 1990. Dominant forces in protein folding. Biochemistry 29, pp.7133-7155.

Eisenhaber F. 1996. Hydrophobic regions on protein surfaces. Deviation of the solvation energy from their area distribution in crystallographic protein structures, Protein Science, 5, pp.1676-1686.

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

Kyte & Doolittle, 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol., 157, pp.105-132.Lee B, and Richards F.M. 1971. "The Interpretation of Protein

Structures: Estimation of Static Accessibility", J. Mol. Biol., 55, pp.379-400.

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

Rees D.C., Wolfe G.M. 1993. Macromolecular solvation energies derived from small molecule crystal morphology. Protein Sci 2, pp.1882-1889.

Richards F.M. 1977. Areas, volumes, packing, and protein structure, Ann. Rev. Biophys. Bioeng., 6, pp.151-176.

Rost B., and Sander C., 1994. Conservation and Prediction of Solvent Accessibility in Protein Families. Proteins, 20, pp. 216-226.

Zhang C. and Kim Sung-Hou 2000. Environment-dependent residue contact energies for proteins. Proc. Natl. Acad. Sci, USA 987 pp.2550-2555.