Original Article

Journal of the Tsuruma Health Science Society Kanazawa University Vol. 33 (1)
17～23 2009

Exploiting residue location information to identify homologous proteins

Hiroshi Nakashima, Hiromi Sato, Madoka Hisazumi

Abstract

There are quite a few cases that two amino acid sequences have around 30% identity over 100 residues even though they are not homologous. A method to incorporate structural information of a protein along the sequence, such as simple location information expressed as interior or surface residues was introduced to increase the sensitivity of homology search. The amino acid residues of a globular protein were classified into two types, namely interior and surface residues based on solvent accessibility of known tertiary structures of proteins. By our definition of interior and surface residues, about half residues were defined as interior and the remaining half residues were classified as surface residues in homologous proteins. The interior residues show higher coincidence of residues than the surface residues in an alignment of homologous proteins. This is consistent with the empirical knowledge that the interior residues are more conserved than the surface residues in homologous proteins. However, this trend was not observed in alignments of non-homologous proteins. Therefore, the comparison of residues coincidence between interior and surface is effective for distinguishing homologous / non-homologous proteins.

Key words

homologous sequences, interior/surface residues, solvent accessibility, sequence identity, expected value

Introduction

To obtain information on a query sequence, the first step we do is a homology search. We can get the character of the query sequence from the hit sequences in a homology search. It is recognized that homologous sequences have similar folds and they usually have similar function although there are some exceptions known. It is important to select homologous sequences from the hit sequences. The sequence identity and expected value (e-value) are used to determine whether the two sequences are homologous or not. The two amino acid sequences which have more than 30% identity over 100 residues are considered to be homologous[1]. If the two sequences have less than 20% amino acid identity, it is impossible to identify homologous proteins using only sequence information. The sequence identity is dependent on the alignment, and the alignment becomes difficult with the decrease of sequence identity. When the three-dimensional (3D) structures of proteins are available, we can align two sequences of less than 20 % identity by structure comparison if they have similar folds[2].

Recently, we found that some false homologous pairs have around 30% amino acid identity over 100 residues. We thought the pairs we found were exceptional cases. However, we noticed that there were quite a few false homologous pairs from a homology search using structure known sequences. This is the motivation of our study. We thought we need a method to distinguish true/false homologous

Department of Clinical Laboratory Science, Graduate Course of Medical Science and Technology, Division of Health Science, Kanazawa University

proteins which have around 30% identity over 100 residues.

To increase the sensitivity in the homology search method, profile analysis using the position-specific scoring table[3], an attempt by combination and modification of existing methods[4] and improved basic local alignment search tool (BLAST)[5] using the position-specific scoring table[6] were developed. Recently, methods for detecting distant homologous relationships between proteins are reported by sequence-structure comparison[7] or by a profile hidden Markov model[8]. Empirically, it is known that 3D structures of proteins are more conserved than sequences. It is said that the interior residues which favor hydrophobic character in a protein are more conserved than the surface ones which favor hydrophilic character. If we could express the 3D structure information along the sequence, it is expected to raise the sensitivity of homology search. We intended to develop a method to distinguish true/false homologous proteins which have around 30% identity over 100 residues, and proteins in the twilight zone which have less than 25% identity are not our targets. Because detecting homologous proteins in the twilight zone is very difficult. We adopted location information of a residue in a protein, i.e. interior or surface residue, for the classification of amino acid residues into interior or surface residues is simple and it seems to be predictable. It is reported that two-state (interior/surface) prediction can be obtained at the accuracy of about 80%[9-11]. The study of interior/surface residues is carried out mainly in terms of protein-protein interfaces, and no attempt has been made to increase the sensitivity of homology search. We examined to what extent the location information would be useful to recognize homologous proteins. To classify interior and surface residues precisely, solvent accessibility of known 3D structures was employed.

Materials and Methods

1. Amino acid sequences

The 7,897 amino acid sequences of known 3D structures were obtained from the structural classification of proteins (SCOP)[12] sequence data 1.71 release at web site http//:scop.mrc-lmb.cam.ac.uk/scop/. Each sequence has a Protein Data Bank (PDB)[13] entry code and a SCOP structural classification code, which represents structural class, fold, superfamily and family. The amino acid sequences from SCOP were analyzed using the BLAST program[5] in an all-against-all configuration. The pairwise alignments which have around 30% sequence identity over 100 residues were selected. The selected alignments were divided into two groups. In the first group, the sequence pairs have identical SCOP structural codes. In the second group, the sequence pairs have different SCOP structural codes. The former group was regarded as a dataset of true homologous proteins and the latter group as a dataset of false homologous proteins.

2. Definition of interior and surface residues

The amino acid residues along a sequence were classified into two types, interior and surface residues based on their relative solvent accessibility. Relative solvent accessibility was obtained from the PDBFinder2 database[14], which has a summary of PDB information with additional annotations (http://ftp.cmbi.ru.nl/pub/molbio/data/pdbfinder2/). Relative solvent accessibility of a residue is expressed as a number from 0 (buried) to 9 (exposed). In this study, the residues with relative solvent accessibility from 0 to 2 were defined as interior residues, and those with that from 3 to 9 were defined as surface ones. This definition of interior/surface residues corresponds to the criteria of Fukuchi and Nishikawa[15]. The number of identical residues in the interior residues was counted and the percentage of identity was calculated. Similarly, the percentage of identity in the surface residues was obtained. Then, the identities between interior and surface residues were compared.

Results and Discussion

1. True and false homologous protein pairs

Ten pairs of true and false homologous proteins are listed in Table 1. The columns PDB, SCOP, length, identity and e-value in Table 1 represent the PDB entry code, SCOP structural classification

Table 1. List of true and false homologous protein pairs. The columns PDB, SCOP, length, identity and e-value represent the PDB entry code, SCOP structural classification code, length of alignment, amino acid identity and expected value, respectively.

No.	protein name	PDB	SCOP	length	identity	e-value
	true homologous protein pairs					
1	protozoan/bacterial hemoglobin	1idra	a.1.1.1	115	35.7	1E-19
	protozoan/bacterial hemoglobin	1dlwa	a.1.1.1			
2	deoxyuridine 5'-triphosphate nucleotidohydrolase	1q5ha	b.85.4.1	123	33.3	7E-11
	deoxyuridine 5'-triphosphate nucleotidohydrolase	1euwa	b.85.4.1			
3	mitochondrial cytochrome *c* (Baker's yeast)	1ycc	a.3.1.1	113	31.9	3E-10
	cytochrome *c* 2 (*Rhodopseudomonas palustris*)	1i8oa	a.3.1.1			
4	sarcoplasmic calcium-binding protein	2scpa	a.39.1.5	158	31.7	2E-99
	calerythrin (*Saccharopolyspora erythraea*)	1nyaa	a.39.1.5			
5	oxidoreductase, MrsD	1p3y1	c.34.1.1	124	31.5	1E-12
	4'-phosphopantothenoylcysteine decarboxylase	1mvla	c.34.1.1			
6	53BP2, complex (anti-oncogene/ankyrin repeats)	1ycsb	d.211.1.1	119	31.1	9E-06
	cell cycle inhibitor p19ink4D	1bd8	d.211.1.1			
7	D-ala carboxypeptidase/transpeptidase	1es5a	e.3.1.1	135	31.1	6E-07
	pencillin bindibg protein 4(PbpD)	1tvfa	e.3.1.1			
8	thymine-DNA glycosylase	1keaa	a.96.1.2	184	29.9	3E-21
	endonuclease III (*Escherichia coli*)	1orna	a.96.1.1			
9	EAT/MLC-1(Myeloid cell leukemia sequence)	1wsxa	f.1.4.1	104	29.8	4E-08
	proapoptotic molecule Bax	1f16a	f.1.4.1			
10	aggrecan core protein	1tdqb	d.169.1.1	131	27.5	2E-12
	galactose-specific C-type lectin	1jzna	d.169.1.1			
	false homologous protein pairs					
1	diol dehydratase, alpha subunit	1eexa	c.1.19.3	107	32.7	0.022
	phosphoserine phosphatase	1j97a	c.108.1.4			
2	hypothetical protein C14orf106	1wgxa	a.4.1.3	103	31.1	0.005
	poly(A)-specific ribonuclease PARN	1whva	d.58.7.1			
3	alpha-ribazole-5'-phosphate phosphatase	1v37a	c.60.1.1	130	30.8	0.006
	2-keto-3-deoxygluconate kinase	1v1aa	c.72.1.1			
4	glutamyl tRNA-reductase middle	1gpja	c.2.1.7	101	30.7	0.006
	spermidine synthase	1uira	c.66.1.17			
5	putative acetyltransferase EF0945	1u6ma	d.108.1.1	114	30.7	0.049
	1-aminocyclopropane-1-carboxylate deaminase	1tyza	c.79.1.1			
6	signal recognition particle 54 kDa	1wgwa	a.24.13.1	111	30.6	0.043
	NEDD8 ultimate buster-1, NUB1	1wjua	d.15.1.1			
7	thioredoxin-like protein 2 (Mouse)	1wika	c.47.1.1	113	30.1	0.014
	probable RNA-binding protein 19, Rbm19 (Mouse)	1whxa	d.58.7.1			
8	Fe superoxide dismutase (FeSOD)	1coja	d.44.1.1	121	29.8	0.087
	Peridinin-chlorophyll protein	1pprm	a.131.1.1			
9	(Apo)ferritin	1lb3a	a.25.1.1	119	29.4	0.012
	tRNA pseudouridine synthase TruD	1szwa	d.265.1.4			
10	DNA-binding protein SATB2	1wiza	a.35.1.7	108	28.7	0.032
	splicing factor, arginine/serine-rich 9	1wg4a	d.58.7.1			

code, length of alignment, amino acid identity and e-value, respectively. PDB codes are represented by four characters and the fifth character of the code specifies the chain. The SCOP codes of thymine-DNA glycosylase and endonuclease III in true homologous protein pairs were different in their family codes. The e-values of true homologous protein pairs were much lower than those of false homologous protein pairs. The protein pairs which showed the e-values less than e^{-5} were all true homologous proteins in Table 1. One might say that e-value alone is a good indicator, however, it is reported that e-values can be very unreliable[8]. The length of alignments ranged from 101 to 184 residues, and the sequence identities were in the range from 27.5% to 35.7%.

2. An example of the false homologous pairs

Alignment of diol dehydratase, alpha subunit form *Klebsiella oxytocal* (PDB: 1eex) versus phosphoserine phosphatase from *Methanococcus jannaschii* (PDB: 1j97) is shown in Figure 1. The interior and surface residues were indicated by 1 and 0 along the sequences. The residues which correspond to gaps were indicated by hyphens.

```
     00100100001101------1101100010110111111000100110111011111111
1eex  KDMPERNIVEDIKF------AQEIINKNRNGLEVVKALAQGGFTDVAQDMLNIQKAKLTG  60
      :: :       ::       : : :    :   :: :   ::: : :    :  : ::
1j97  KDLPIEKVEKAIKRITPTEGAEETIKELKNRGYVV-AVVSGGF-DIA---VNKIKEKLGL  60
      00100001001101100100100110010010111-1111001-001---1001000101

      11111111100000111111010101011001111100001001101 00
1eex  DYLHTSAIIVGDGQVLSAVNDVNDYAGPATGYRLQGERWEEIKNIPG             107
      ::      :: ::         :  :        ::  : :  : :
1j97  DYAFANRLIVKDGKL------TGDVEGEVLKENAKGEILEKIAKIEG             107
      0111110100000001------10010001100011001100110010
```

Figure 1. Amino acid sequence alignment of diol dehydratase, alpha subunit form *Klebsiella oxytocal* (PDB: 1eex) versus phosphoserine phosphatase from *Methanococcus jannaschii* (PDB: 1j97). The interior and surface residues are indicated by 1 and 0 along the sequences. Colons denote identical residues and hyphens denote gaps.

The identical residues in the alignment were described with colons. To count the interior and surface residues, gaps were not taken into consideration. So, 101 and 96 residues were considered in the sequence of diol dehydratase and phosphoserine phosphatase, respectively. The percentages of interior and surface residues were 58.4% (59/101) and 41.6% (42/101) in the sequence of diol dehydratase. In the case of phosphoserine phosphatase, corresponding percentages were 44.8% (43/96) and 55.2% (53/96). Therefore, the deviation of interior/surface residues was 13.6% between the two proteins. The sequence identity of the alignment was 32.7% (35/107) and the e-value was 0.022. The two proteins have different SCOP codes, therefore, they are considered as being non-homologous proteins.

The percentages of identical residues in the interior and surface residues were 28.8% (17/59) and 42.9% (18/42), respectively based on the defined interior and surface residues of diol dehydratase. The corresponding percentages were 25.6% (11/43) and 45.3% (24/53) based on phosphoserine phosphatase. Consequently, identities of interior residues in diol dehydratase and phosphoserine phosphatase were 14.1% and 19.7% lower than those of surface residues. This indicated that the interior residues were less conserved than surface residues in non-homologous proteins.

3. Interior residues were more conserved than surface residues in homologous proteins

Figure 2 shows the plot of identities deviations between interior and surface residues. Two deviations obtained from the identical alignment were plotted at the same coordinate of the horizontal axis. The plots from the true homologous protein pairs were represented by filled circles while those from the false homologous protein pairs were represented by open circles. Number of data from one to ten in Fig. 2 corresponds to the number of true homologous pairs in Table 1 and that from eleven to twenty does to the number of false pairs. So, diol dehydratase and phosphoserine phosphatase is plotted as data number eleven. All the deviations from the true homologous protein pairs were positive, indicating the identities of interior residues were higher than those of surface residues. This is consistent with the empirical knowledge that inside parts of proteins are more conserved than exterior parts, when the two proteins are homologous. On the other hand, most of the deviations from the false homologous protein pairs were negative. Only one pair from the false homologous protein group showed two positive

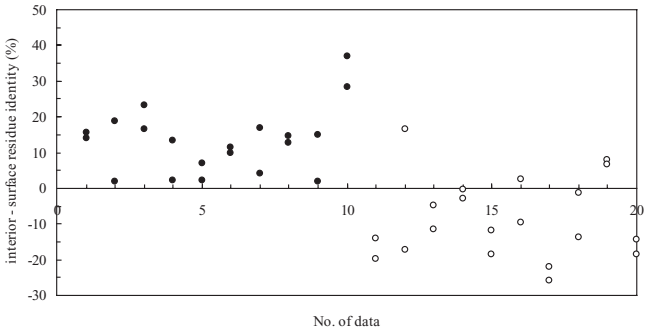

Figure 2. Plot of identities deviations between interior and surface residues. Filled circles indicate the plots from the true homologous proteins pairs and open circles from the false homologous protein pairs.

deviations. This result indicated that the comparison of the residues identities between interior and surface is an effective approach in the assessment of homologous proteins.

The averages of interior and surface residues for the 20 true homologous proteins were 48.2% and 51.8%, respectively. Therefore, about half of the residues of true homologous proteins were defined as interior and remaining half residues were classified as surface by our definition. For the 20 false homologous proteins, the averages were 42.2% and 57.8%, respectively. It indicated a 7.8% deviation which was larger compared to that of true homologous proteins.

4. Effect of protein forms on the determination of interior and surface residues

In this study, interior and surface residues were determined by relative solvent accessibility. It is reported that relative solvent accessibility varies between 4% and 14% in different complexes depending on the protein-protein interface[16]. The interior and surface residues were not directly comparable with the variation of relative solvent accessibility due to the different definitions. Four trypsin monomer proteins (PDB: 1g3e, 1qb1, 2fx4, 2fx6) and four trypsin complexes with inhibitor proteins (PDB: 1eb2, 1k1i, 1tpa, 1y3v) were employed to evaluate the effect of protein form i.e. monomer or complex, in the determination of interior and surface residues. The four trypsin monomer sequences and four trypsin complex sequences were all obtained from bovine and their amino acid sequences were identical. The defined interior and surface residues of trypsin monomer sequence and trypsin complex sequence were compared and the coincidence was calculated. Similarly, the coincidence of interior and surface residues in the trypsin monomer sequence and in the trypsin complex sequence was investigated. It was expected that solvent accessibility of some residues in trypsin complex would be reduced by protein-protein interface, and this reduction might affect the determination of interior and surface residues. As an example, the defined interior and surface residues were compared for the trypsin monomer (PDB: 1g3e) and (PDB: 1qb1). The coincidence of interior/surface residues was 96.9% (216/223). The average coincidence of interior and surface residues for the six possible combinations from the four trypsin of monomer sequences and complex sequences were 91.6% and 91.9%, respectively. Similarly, the average coincidence for the 16 cases between the four trypsin of monomer sequences and complex sequences was 92.5%. From this result, we concluded that the determination of interior and surface residues is independent on the protein form monomer or complex in the case of trypsin.

5. Application for proteins with less than 30% sequence identities

The alignments of approximately 30% sequence identities over 100 residues of structure known proteins were analyzed. We considered that the alignments of true homologous protein pairs were reliable as multiple sequence alignments including sequences of more than 30% identities indicated same alignments.

We applied this method to the proteins of low sequence identities. It is well known that proteins of the globin family have identical globin folds from the X-ray crystallographic structures though they have sequence identities of less than 30%. Amino acid sequences of myoglobin from sperm whale (PDB: 5mbn), erythrocruorin from *Chironomous thummi thummi* (PDB: 1eco) and hemoglobin alpha and beta chains from human (PDB: 3hhba, 3hhbb) were employed. The alignments based on their 3D structure comparison[2] and those from BLAST homology search were compared in terms of sequence identities for total, interior and surface residues. The alignment of sperm whale myoglobin and human hemoglobin alpha chains based on their 3D structure comparison was almost identical with that of the BLAST alignment except gap positions. Table 2 lists the sequence identities for alignments from 3D structure comparison and those from the BLAST homology search. Let's look at the No. 1 comparison between myoglobin and hemoglobin alpha chain in Table 2. The identity of whole sequences was 24.7% in the alignment based on 3D structure and that was 25.2% in the alignment of BLAST search.

Table 2. Comparison between residues identities in alignments based on 3D structure and BLAST search.

No.	alignment	residues	identities		e-value
1	myoglobin (PDB: 5mbn) vs. hemoglobin alpha (PDB: 3hhba)				
	3D structure	total residues	24.7% (38/154)		
		interior residues	38.5% (25/65)	35.1% (26/74)	
		surface residues	14.8% (13/88)	17.9% (12/67)	
	BLAST	total residues	25.2% (37/147)		1E-09
		interior residues	35.4% (23/65)	32.4% (24/74)	
		surface residues	17.1% (14/82)	19.4% (13/67)	
2	myoglobin (PDB: 5mbn) vs. erythrocruorin (PDB: 1eco)				
	3D structure	total residues	19.4% (30/155)		
		interior residues	30.8% (20/65)	30.2% (19/63)	
		surface residues	11.4% (10/88)	15.1% (11/73)	
	BLAST	total residues	26.1% (24/92)		1E-04
		interior residues	33.3% (12/36)	26.3% (10/38)	
		surface residues	18.5% (10/54)	25.5% (12/47)	
3	erythrocruorin (PDB: 1eco) vs. hemoglobin beta (PDB: 3hhbb)				
	3D structure	total residues	17.5% (26/149)		
		interior residues	23.8% (15/63)	21.9% (16/73)	
		surface residues	15.1% (11/73)	13.7% (10/73)	
	BLAST	total residues	19.2% (20/104)		0.006
		interior residues	17.4% (8/46)	27.0% (10/37)	
		surface residues	20.7% (12/58)	24.4% (10/41)	

The e-value of BLAST search was e^{-9}. The sequence identities of interior and surface residues in the alignment based on 3D structure comparison were 38.5% and 14.8% using the defined interior and surface residues of myoglobin, and the corresponding ones from BLAST alignment were 35.4% and 17.1%. Similar identities using the defined interior and surface residues of hemoglobin alpha chain were 35.1% and 17.9%, and the corresponding ones from BLAST alignments were 32.4% and 19.4%. Consequently, the identities of interior residues were 23.7% and 17.2% higher than those of surface ones in the 3D structural alignments. On the other hand, the identities of interior residues were 18.3% and 13.0% higher than those of surface ones in the BLAST alignments. This result indicated that the sequence coincidence of interior residues was higher in the correct alignment based on 3D structural comparison. This trend holds in the comparisons for myoglobin vs. erythrocruorin and for erythrocruorin vs. hemoglobin beta chain.

Rererences

1) Doolittle RF: Searching through sequence databases. Methods Enzymol 183: 99-110, 1990

2) Bashfold D, Chothia C, Lesk AM: Determinants of a protein fold unique features of the globin amino acid sequences. J Mol Biol 196: 199-216, 1987

3) Gribskov M, McLachlan AD, Eisenberg D: Profile analysis : detection of distantly related proteins. Proc Natl Acad Sci USA 84: 4355-4358, 1987

4) Nishikawa K, Nakashima H, Kanehisa M, et al: Detection of weak sequence homology of proteins for tertiary structure prediction. Protein Seq Data Anal 1: 107-116, 1987

5) Altschul SF, Gish W, Miller W, et al: Basic local alignment search tool. J Mol Biol 215: 403-410, 1990

6) Altschul SF, Madden TL, Schäffer AA, et al: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402, 1997

7) Shi J, Blundell TL, Mizuguchi K: FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310: 243-257, 2001

8) Söding J: Protein homology detection by HMM-HMM comparison. Bioinformatics 21: 951-960, 2005

9) Adamczak R, Porollo A, Meller J: Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 56: 753-767, 2004

10) Chen H, Zhou H-X: Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 33: 3193-3199, 2005

11) Wang J-Y, Lee H-M, Ahmad S: SVM-cabins: prediction of solvent accessibility using accumulation cutoff set and support vector machine. Proteins 68: 82-91, 2007

12) Murzin AG, Brenner SE, Hubbard T, et al: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540, 1995

13) Berman HM, Westbrook J, Feng Z, et al: The protein data bank. Nucleic Acids Res 28: 235-242, 2000

14) Hooft RWW, Sander C, Scharf M, et al: The PDBFINDER database : a summary of PDB, DSSP and HSSP information with added value. CABIOS 12: 525-529, 1996

15) Fukuchi S, Nishikawa K: Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol 309: 835-843, 2001

16) Conte LL, Chothia C, Janin J: The atomic structure of protein-protein recognition sites. J Mol Biol 285: 2177-2198, 1999

アミノ酸残基をタンパク質の内部、表面部分に分けた相同タンパク質検出方法

中島　広志，佐藤　裕美，久澄まどか

要　　旨

アミノ酸配列のホモロジー検索において、２つのアミノ酸配列が100残基以上で30％の一致割合を示すとき、それらは相同とみなされる。しかし、非相同タンパク質がこの程度の一致割合を示すことがある。よって、タンパク質が相同か非相同かを判別する方法が必要である。タンパク質の立体構造はアミノ酸配列より保存されており、内部残基が表面残基よりよく保存されていることが経験的に知られている。我々は、立体構造既知の球状タンパク質のアミノ酸残基を水との接触可能面積より、タンパク質内部、表面残基と２つのタイプに分け、内部残基及び表面残基それぞれの一致割合を比較した。相同タンパク質では内部残基の一致割合が表面残基の割合より高かった。非相同タンパク質では表面残基の割合が高い場合が多かった。よって、アミノ酸残基をタンパク質内部、表面残基と分けそれぞれの一致割合を比較すれば相同か非相同かの判別が可能と思われた。アミノ酸配列より内部残基か表面残基の予測精度は80％といわれているので、アミノ酸配列のみからもある程度の相同か非相同かの判別が可能と思われる。