Amino Acids (2008) 35: 495–501
DOI 10.1007/s00726-007-0586-0
Printed in The Netherlands

Amino Acids

Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach

S.-W. Zhang[1], **Y.-L. Zhang**[2], **Q. Pan**[1], **Y.-M. Cheng**[1], and **K.-C. Chou**[3]

[1] College of Automation, Northwestern Polytechnical University, Xi'an, China
[2] Department of Computer, First Aeronautical Institute of Air Force, Xinyang, Henan, China
[3] Gordon Life Science Institute, San Diego, California, USA

Received January 20, 2007
Accepted February 23, 2007
Published online August 21, 2007; © Springer-Verlag 2007

Summary. Evolutionary conservation derived from a multiple sequence alignment is a powerful indicator of the functional significance of a residue, and it can help to predict active sites, ligand-binding sites, and protein interaction interfaces. The results of the existing algorithms in identifying the residue's conservation strongly depend on the sequence alignment, making the results highly variable. Here, by introducing the amino acid similarity matrix, we propose a novel gap-treating approach by combining the evolutionary information and von Neumann entropies to compute the residue conservation scores. It is indicated through a series of tested results that the new approach is quite encouraging and promising and may become a useful tool in complementing the existing methods.

Keywords: Evolutionary conservation – Amino acid similarity matrix – von Neumann entropy – Functional residue – Sensitivity – Specificity

1. Introduction

Determining which amino acid residues in a protein are responsible for its function is very important in order to understand the molecular mechanism of protein and for drug discovery as well (Chou, 2004e; Kesel, 2005; Lubec et al., 2005; Clercq, 2006). The experimental characterization of the constituent residues in their function and role is very expensive, time-consuming, and difficult to automate. This kind of difficulties has challenged us to develop computational approaches.

It is assumed in all the existing alignment methods or the evolutionary residue-scoring methods that the importance of a residue is reflected by its evolutionary conservation, meaning that the more important the residue, the sooner it becomes fixed in different evolutionary branches and the more divergent are the branches between which it does vary (Mihalek et al., 2004). The evolutionary conservation varies among amino acid sites due to differing degrees of functional constraints on them (Holmquist et al., 1983; Troy et al., 1993; Zhou and Troy, 2005a). Sites that are important for the protein's tertiary structure and folding, enzymatic activity, ligand binding, or interaction with other proteins are generally more conserved (Zhou and Troy, 1995, 2003, 2005b; Brocchieri et al., 2002; Ran et al., 2004; Zhou et al., 2004; Schnell et al., 2005). Actually, knowledge of the conserved and functionally important residues, such as those involved in forming the binding pocket of a protein to its ligand, has been widely used to help the structure-based drug design and to guide the mutagenesis studies (see, e.g., Kang and Liang, 1997; Chou et al., 1999, 2000, 2003, 2006; Hu et al., 2003; Yu, 2003; Chou, 2004a–d; Du et al., 2004, 2005a, b; Zhang and Yap, 2004; Fan et al., 2005; Gan et al., 2006; Wu et al., 2006; Zhang et al., 2006; Liang and Li, 2007; Wang et al., 2007a, b; Wei et al., 2005, 2006a, 2006b, 2007), indicating the importance of finding conservative residues.

One of the representative methods in identifying the conservative residues is ConSurf (Armon et al., 2001), which used a maximum parsimony tree to calculate a site conservation score as the number of substitutions weighted by their physicochemical distance. Another algorithm was named evolutionary trace (ET) (Lichtarge and Sowa, 2002), which utilizes a phylogenetic tree to identify residues that are identically conserved in a subtree. The maximum tree depth at which a residue remains unchanged is used to rank the degree of conservation. This analysis was later modified to incorporate a quantitative model of residue substitutions. The ET method had been shown to be capable of detecting protein interaction sites and directing protein mutation studies. In order to make the ET method more robust against deviations from the ideal family tree picture occurring in the actual protein evolution, Lichtarge et al. (Lichtarge and Sowa, 2002) developed a class of hybrid methods (real-valued evolutionary trace method and zoom method) that combine evolutionary and Shannon entropies from multiple sequence alignments (Mihalek et al., 2006a, b). However, the hybrid methods do not account for the physicochemical similarities found between the different amino acids, and the gaps in the multi-sequences alignment are treated as the 21^{st} amino acid.

In the present study, we are to combine evolutionary and von Neumann entropies and propose a different gap-treating approach for estimating the residue's evolutionary conservation. It is demonstrated by using the insulin receptor kinase domains as an example to show that the current hybrid approach can enhance the prediction quality in comparison with the existing methods.

2. Materials and methods

2.1 Key residues

It is difficult to give a comprehensive and accurate definition for "functionally important residue", although everyone seems to have an intuitive concept of what these residues mean. A widely accepted definition is that the functionally important residues are those indispensable for the protein to perform its molecular function or to play its biological role in the sense that these residues cannot be freely changed (except for change to some compatible amino acids) without directly affecting its native function. In this study we call these "functionally important residues" the key residues, which are in a broad sense directly related to the active/catalytic sites, protein binding sites, small ligand binding sites, nucleic acids binding sites, and so on. For example, the key residues taken from the insulin receptor (1irk.pdb) are listed in Table 1.

2.2 Proteins in the testing dataset

Half of proteins in the testing dataset were taken from (Mihalek et al., 2004). Some proteins whose computational results could not be obtained from Lichtarge Computational Biology Lab web service at http://mammoth.bcm.tmc.edu/report_maker/index.html or Ben-Tal's ConSurf web service at http://consurf-hssp.tau.ac.il/cgi-bin/consurf-hsspNew.cgi were deleted. Other five proteins were from (Zhu et al., 2006). To construct a reasonable independent testing dataset, the key residues were defined according to the PDBsum database (http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/). In general, the key residues are the constituents of the following: active sites, catalytic residues, interaction with ligand, interaction with metal, and PROSITE pattern residues (mainly refer to red, red orange, and orange residues). The protein PDB code and the number of key residues in the testing dataset are given in Table 2.

2.3 Initial selections of sequences

Three sets of sequences from three different sources are considered. The raw sequence sets were created by using three iterations of PsiBlast (Altschul et al., 1997), with the 0.001 E-value cutoffs on the UniProt database of proteins. The PsiBlast resulting sets were aligned by a standard alignment method such as ClustalW 1.8 (Thompson et al., 1994). The pruned sequence sets were Lichtarge_HSSP and Consurf_HSSP. HSSP is a standard database of sequence selections/alignments obtained by carefully rethinking the similarity cutoffs for sequences of different lengths. The Lichtarge_HSSP alignments have the sequences that are selected according to Litcharge's criterion (Mihalek et al., 2004, 2006b). The Consurf_HSSP alignments have the sequences that are selected according to Ben-Tal's criterion (Glaser et al., 2003).

2.4 Residue ranking score

The residue ranking function assigns a score to each of the residues concerned, and according to the scores they can be sorted in the order of the presumably decreasing evolutionary pressure they experience. By combining the evolutionary and von Neumann entropies to estimate the residue evolutionary conservation, we propose a new approach to calculate the

Table 1. The 38 key residues of the insulin receptor (1irk.pdb)[a]

SER1006	VAL1010	LYS1030	LEU1038	ILE1042	LEU1045	GLU1047
PHE1054	ARG1061	GLU1077	MET1079	GLY1082	ASP1083	LYS1085
ARG1089	ARG1092	ARG1131	ARG1136	ASN1137	MET1139	ASP1150
PHE1151	GLY1152	ARG1155	TYR1158	TYR1162	TYR1163	ASP1132
ARG1164	GLY1166	LEU1171	PRO1172	VAL1173	MET1176	LEU1181
ASN1215	GLU1216	LEU1219				

[a] For a full discussion about these residues, see Hubbard (1997) and Mihalek et al. (2004)

Table 2. The proteins used for the testing dataset

PDB code	1au1A	1aulA	1bkx	1cqiA	1ctq	1exqA	1fha	1hzxA	1mxrA	3tmkA
Number of residues	166	243	339	286	166	147	172	340	339	216
Number of key residues	19	18	21	37	22	17	23	96	28	29

residue ranking score as formulated below. The first step is to calculate the Shannon entropy and von Neumann entropy of each alignment column. The Shannon entropy for a residue belonging to column i in an MSA (multiple sequence alignment) is given by

$$S_i^{\text{Shanon}} = -\sum_{a=1}^{20} f_{i,a}\log_{20} f_{i,a} + f_{i,\text{gap}} \tag{1}$$

where $f_{i,a}$ is the relative occurrence frequency of amino acid a at the alignment position i. The base of 20 ensures that all values are bounded between zero and one. The symbol $f_{i,\text{gap}}$ represents the number of non-standard amino acids (such as "–", "X", "Z", "B") at the alignment position i divided by the number of alignment sequences. The von Neumann entropy is given by

$$S_i^{\text{von}} = -\text{Tr}(\omega_i \log_{20}\omega_i) \tag{2}$$

where ω_i is a density matrix with trace = 1. Apart from normalization by the trace, the density matrix is given by the product of the relative frequencies of the amino acids in each alignment position $f_{i,a}$ and an appropriate similarity matrix, that is, $\omega_i = \text{diag}[f_{i,A}, f_{i,C}, \ldots f_{i,a}, \ldots, f_{i,Y}]\times$ similarity matrix. The calculation of Eq. (2) is facilitated by first calculating the eigenvalues $\lambda_{i,m}$ of ω_i, and hence it follows that

$$S_i^{\text{von}} = -\sum_{m}\lambda_{i,m}\log_{20}\lambda_{i,m} + f_{i,\text{gap}} \tag{3}$$

When the similarity matrix is the identity matrix, the von Neumann entropy (Eq. (3)) becomes identical to the Shannon entropy of Eq. (1). The second step is to divide an MSA into sub-alignments (that is n groups) that correspond to nodes in the tree. This subdivision of an MSA into smaller alignments reflects the tree topology, and hence the evolutionary variation information within it. The evolutionary score for a residue belong to column i in an MSA is given by the following series of equations

$$R_i = 1 + \sum_{n=1}^{N-1} w_{\text{node}}(n)\sum_{g=1}^{n} w_{\text{group}}(g)\left[-\sum_{a=1}^{20} f_{i,a}^{g}\log_{20} f_{i,a}^{g} + f_{i,\text{gap}}^{g}\right] \tag{4}$$

where $w_{\text{node}}(n)$, $w_{\text{group}}(g)$ are weights assigned to a node n and a group g, respectively.

$$w_{\text{node}}(n) = \begin{cases} 1, & \text{if } n \text{ on the path to the query protein} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$w_{\text{group}}(g) = \begin{cases} 1, & \text{if } g \text{ on the path to the query protein} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$f_{i,a}^{g}$ is the frequency of amino acid of type a within a sub-alignment corresponding to group g at the level in which the sequence similarity tree is divided into n groups. Namely, the nodes (labeled by n) are assumed to be numbered in the order of increasing distance from the root, and each one of them is associated with a division of the tree into n groups (subtrees). N is the number of alignment sequences, $f_{i,\text{gap}}^{g}$ the number of non-standard amino acids of g group in the alignment position i divided by the number of g group alignment sequences. Further details about division of tree nodes and groups can be found in literature (Mihalek et al., 2004). We call the scoring function by Eq. (4) the improved zoom (IZ) method. Considering the physicochemical similarities between the different amino acids, Eq. (4) can be further formulated as follows:

$$R_i = 1 + \sum_{n=1}^{N-1} w_{\text{node}}(n)\sum_{g=1}^{n} w_{\text{group}}(g)\left[-\sum_{m}\lambda_{i,m}^{g}\log_{20}\lambda_{i,m}^{g} + f_{i,\text{gap}}^{g}\right] \tag{7}$$

where $\lambda_{i,m}^{g}$ is the eigenvalues of the density matrix ω_i^{g} of g group in the alignment position i. Equation (7) is called the physicochemical similarity zoom (PSZ) method. If we take $w_{\text{node}}(n) = 1/n$, $w_{\text{group}}(g) = 1$, Eqs. (4) and (7) can be, respectively, expressed by

$$\rho_i = 1 + \sum_{n=1}^{N-1}\frac{1}{n}\sum_{g=1}^{n}\left[-\sum_{a=1}^{20} f_{i,a}^{g}\log_{20} f_{i,a}^{g} + f_{\text{gap}}^{g}\right] \tag{8}$$

$$\rho_i = 1 + \sum_{n=1}^{N-1}\frac{1}{n}\sum_{g=1}^{n}\left[-\sum_{m}\lambda_{i,m}^{g}\log_{20}\lambda_{i,m}^{g} + f_{i,\text{gap}}^{g}\right] \tag{9}$$

Equations (8) and (9) are called the improved real-valued (IRV) method and the physicochemical similarity real-valued (PSRV) method, respectively.

Through Eqs. (1), (3)–(4) and (7)–(9), each residue of the protein concerned can be assigned a score. The smaller the score is, the more conservation the residue is.

2.5 Sensitivity and specificity

To compare the performance of different methods, the parameters of sensitivity (S_n) and specificity (S_p) are introduced to estimate the quality of different approaches.

$$S_n = \frac{TP}{TP + FN} \tag{10}$$

$$S_p = \frac{TN}{TN + FP} \tag{11}$$

where TP is the number of important residues found correctly, TN is the number of unimportant residues predicted correctly, and FN and FP are numbers of unimportant and important residues predicted incorrectly, respectively. In fact, the sensitivity is the ratio of the number of important residues found correctly to the known total number of important residues (true identified positive/actual positive), while specificity is the number of unimportant residues predicted by the method divided by the number of residues known not to be important (true identified negative/actual negative).

3. Results and discussion

The insulin receptor is a transmembrane protein that binds to the insulin hormone. The binding leads to autophosphorylation of tyrosine residues in the activation loop of the protein, resulting in an enhancement of the catalytic activity and creation of binding sites for downstream signaling proteins. Its structure and residue enumeration can be found in the Protein Data Bank (http://www.rcsb.org/pdb) under the code 1irk. The four key parts of the 1irk machinery are (i) the residues involved in ATP binding, (ii) active site (peptide-binding site), (iii) rotational pivot points and other residues involved in lobe closure, which are important in conformational change between inactive and phosphorylated state, and (iv) activation loop, which occupies the ATP-binding site in the inactive form with the three key tyrosine residues involved in autophosphorylation highlighted. Sensitivity is the percentage of key residues that is found among the top n residues on the list (the score was aligned from small to large, i.e., the top *n* residues have small scores), and specificity is the percentage of residues that are not singled out as important and can be found below the *n*th position. Each point on the graph is the specificity-sensitivity pair for one particular choice of *n*. The sensitivities and specificities of six methods (Shannon, von Neumann, IZ, PSZ, IRV, and PSRV) with raw and Lichtarge_HSSP alignment sequence sets are shown as Figs. 1 and 2, from which we can see that IZ and PSZ top the other four methods in both sensitivity and specificity for almost any choice of *n* (the exception being of marginal size) indicating that the performances of IZ and PSZ are better than those of the other four methods. Furthermore, it can also be seen from Figs. 3 to 4, as well as Table 3, that the performance of our gap-treating approach is better than that of Mihalek gap handling approach (Mihalek et al., 2004).

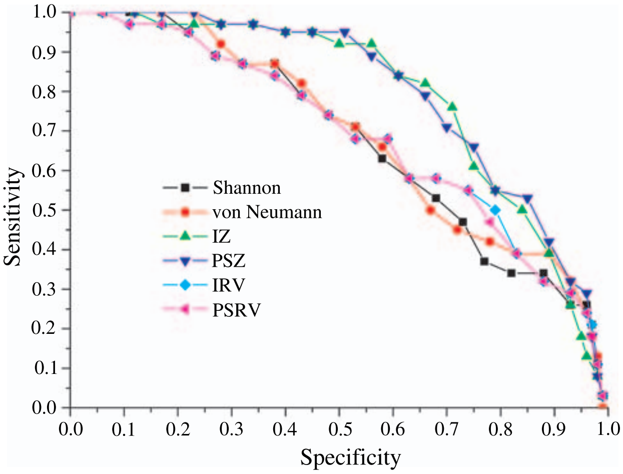

Fig. 1. The sensitivity versus the specificity of the prediction for the six methods using the raw sequence selection (for a color reproduction of this figure, the reader is referred to the online version of this paper under www.springerlink.com)

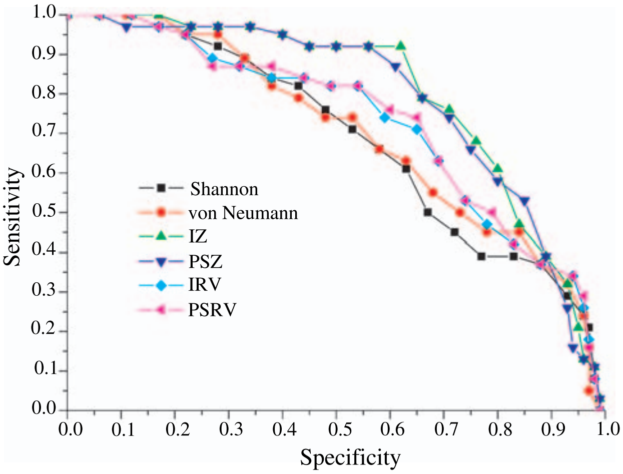

Fig. 2. The sensitivity versus the specificity of the prediction for the six methods using the Lichtarge-HSSP (for a color reproduction of this figure, the reader is referred to the online version of this paper under www.springerlink.com)

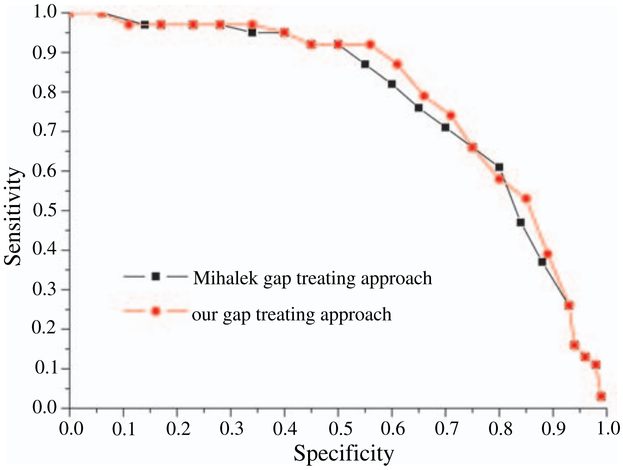

Fig. 3. The sensitivity versus the specificity of the prediction for ours and Mihalek gap treating approache with Lichtarge-HSSP using IZ (for a color reproduction of this figure, the reader is referred to the online version of this paper under www.springerlink.com)

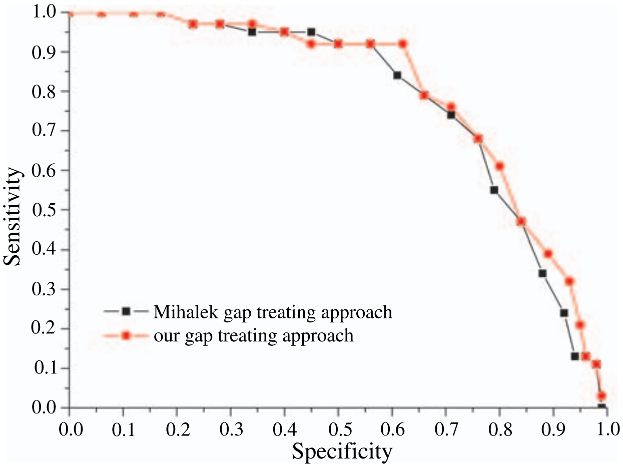

Fig. 4. The sensitivity versus the specificity of the prediction for ours and Mihalek gap treating approache with Lichtarge-HSSP using PSZ (for a color reproduction of this figure, the reader is referred to the online version of this paper under www.springerlink.com)

In addition to the insulin receptor 1irk, the comparison has been extended to cover more proteins. Shown in Table 3 are the predicted results by the four different methods on the 10 proteins listed in Table 2. As shown in Table 4,

Table 3. The top 25% (76 residues) of all residues in the ranking sequence according to the conservation score from small to big for the insulin receptor (lirk.pdb) by different methods[a]

PSZ (including 23 key residues)	IZ (including 22 key residues)	Lichtarge (including 18 key residues)	Ben-Tal (including 17 key residues)
1003, 1005, 1007, 1008, **1010**, 1013, 1028, **1030,** 1044, **1047**, 1051, **1054**, 1062, 1064, 1076, **1077**, **1079**, **1082**, **1083**, 1084, 1088, 1112, 1116, 1119, 1120, 1122, 1123, 1130, **1131**, **1132**, 1133, 1134, 1135, **1136**, **1137**, 1138, **1139**, 1147, 1148, 1149, **1150**, **1151**, **1152**, 1153, **1155**, **1162**, **1163**, **1172**, **1173**, 1174, 1175, **1176**, 1179, **1181**, 1184, 1186, 1190, 1191, 1193, 1194, 1196, 1197, 1198, 1199, 1200, 1201, 1209, 1210, 1212, 1218, 1234, 1242, 1245, 1246, 1253, 1256	1003, 1007, 1008, **1010**, 1013, 1028, **1030**, **1047**, 1051, 1060, 1062, 1064, 1065, 1071, **1077**, **1079**, **1082**, **1083**, 1084, 1088, 1116, 1119, 1120, 1122, 1123, 1129, 1130, **1131**, **1132**, 1133, 1134, 1135, **1136**, **1137**, 1138, **1139**, 1140, 1146, 1147, 1148, **1150**, **1151**, **1152**, 1153, **1155**, **1162**, **1163**, **1172**, **1173**, 1174, 1175, **1176**, 1178, 1179, 1180, **1181**, 1190, 1191, 1193, 1194, 1196, 1197, 1198, 1199, 1200, 1201, 1203, 1210, 1218, 1234, 1242, 1245, 1246, 1253, 1254, 1256	1003, 1005, 1007, 1008, **1010**, 1027, 1028, **1030**, 1044, **1047**, 1048, 1051, **1054**, 1056, 1060, 1062, 1064, 1074, **1077**, **1082**, **1083**, 1084, 1088, 1115, 1119, 1120, 1122, 1123, 1130, **1131**, **1132**, 1133, 1134, 1135, **1136**, **1137**, 1138, **1139**, 1140, 1146, 1147, 1148, 1149, **1150**, **1151**, **1152**, 1153, 1154, **1155**, **1172**, 1174, 1175, **1176**, 1177, 1178, 1179, 1186, 1190, 1191, 1192, 1193, 1194, 1196, 1198, 1200, 1201, 1206, 1209, 1210, 1228, 1231, 1242, 1245, 1246, 1253, 1256	1003, 1005, 1007, 1008, **1010**, 1027, 1028, 1029, **1030**, **1047**, 1048, 1051, 1056, 1058, 1060, 1062, 1064, 1074, 1076, **1077**, **1082**, **1083**, 1084, 1088, 1115, 1117, 1119, 1120, 1122, 1123, 1129, 1130, **1131**, **1132**, 1133, 1134, 1135, **1136**, **1137**, **1139**, 1147, 1148, 1149, **1150**, **1151**, **1152**, 1153, 1154, **1155**, **1172**, 1174, 1175, **1176**, 1177, 1178, 1179, 1180, 1187, 1190, 1191, 1192, 1193, 1194, 1196, 1197, 1201, 1204, 1209, 1228, 1231, 1242, 1245, 1246, 1253, 1254, 1256

[a] The key residues identified by the corresponding method are highlighted in bold-face type

Table 4. The key residue number found in the top 25% of all residues for the multi-proteins with different methods

PDB code	PSZ		IZ		Lichtarge		Ben-Tal	
	NKR	S_n^{25}	NKR	S_n^{25}	NKR	S_n^{25}	NKR	S_n^{25}
1au1A	7	7/19 = 0.37	9	9/19 = 0.47	9	9/19 = 0.47	8	8/19 = 0.42
1auLA	10	10/18 = 0.56	10	10/18 = 0.56	10	10/18 = 0.56	9	9/18 = 0.50
1bkx	16	16/21 = 0.76	15	15/21 = 0.71	16	16/21 = 0.76	17	17/21 = 0.81
1cqiA	25	25/37 = 0.68	30	30/37 = 0.81	28	28/37 = 0.76	29	29/37 = 0.78
1ctq	15	15/22 = 0.68	15	15/22 = 0.68	13	13/22 = 0.59	13	13/22 = 0.59
1exqA	6	6/17 = 0.35	6	6/17 = 0.35	6	6/17 = 0.35	5	5/17 = 0.29
1fha	12	12/23 = 0.52	13	13/23 = 0.57	12	12/23 = 0.52	8	8/23 = 0.35
1hzxA	27	27/96 = 0.28	27	27/96 = 0.28	22	22/96 = 0.23	25	25/96 = 0.26
1mxrA	14	14/28 = 0.50	13	13/28 = 0.46	13	13/28 = 0.46	8	8/28 = 0.29
3tmkA	19	19/29 = 0.66	18	18/29 = 0.62	18	18/29 = 0.62	18	18/29 = 0.62

[a] NKR represents the number of key residues found

performances of IZ and PSZ are better than that of Lichtarge' s hybrid method and Ben-Tal's ConSurf method.

4. Conclusion

Introducing von Neumann entropy and a novel gap-treating approach in the sequence alignment is quite promising for estimating residue evolutionary conservation. The novel approach can at least play a complementary role in the existing methods in this area.

Acknowledgements

This paper was supported in part by the National Natural Science Foundation of China (No. 60634030, No. 60372085), the Technological

Innovation Foundation of Northwestern Polytechnical University (No. KC02), and the Science Technology Research and Development Program of Shaanxi (2006k04-G14).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

Armon A, Graur D, Ben-Tal N (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 307: 447–463

Brocchieri L, Zhou GP, Jardetzky O (2002) Allostery and induced fit: NMR and molecular modeling study of the trp repressor-mtr DNA complex. In: Eaton GR, Wiley DC, Jardetzky O (eds) ACS Symposium Series 827: Structures and Mechanisms from Ashes to Enzymes. American Chemistry Society, Washington, DC, pp 340–366

Chou KC (2004a) Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. Biochem Biophys Res Commun 319: 433–438

Chou KC (2004b) Insights from modelling the tertiary structure of BACE2. J Proteome Res 3: 1069–1072

Chou KC (2004c) Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, 5. Biochem Biophys Res Commun 316: 636–642

Chou KC (2004d) Molecular therapeutic target for type-2 diabetes. J Proteome Res 3: 1284–1288

Chou KC (2004e) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11: 2105–2134

Chou KC, Tomasselli AG, Heinrikson RL (2000) Prediction of the tertiary structure of a caspase-9/inhibitor complex. FEBS Lett 470: 249–256

Chou KC, Watenpaugh KD, Heinrikson RL (1999) A model of the complex between cyclin-dependent kinase 5(Cdk5) and the activation domain of neuronal Cdk5 activator. Biochem Biophys Res Commun 259: 420–428

Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: ibid, 2003, Vol. 310, 675). Biochem Biophys Res Commun 308: 148–151

Chou KC, Wei DQ, Du QS, Sirois S, Zhong WZ (2006) Review: progress in computational approach to drug development against SARS. Curr Med Chem 13: 3263–3270

Clercq ED (2006) Potential antivirals and antiviral strategies against SARS coronavirus infections. Expert Rev Anti Infect Ther 4: 291–302

Du QS, Wang SQ, Jiang ZQ, Gao WN, Li YD, Wei DQ, Chou KC (2005b) Application of bioinformatics in search for cleavable peptides of SARS-CoV Mpro and chemical modification of octapeptides. Med Chem 1: 209–213

Du QS, Wang SQ, Wei DQ, Zhu Y, Guo H, Sirois S, Chou KC (2004) Polyprotein cleavage mechanism of SARS CoV Mpro and chemical modification of octapeptide. Peptides 25: 1857–1864

Du QS, Wang S, Wei DQ, Sirois S, Chou KC (2005a) Molecular modelling and chemical modification for finding peptide inhibitor against SARS CoV Mpro. Anal Biochem 337: 262–270

Fan K, Ma L, Han X, Liang H, Wei P, Liu Y, Lai L (2005) The substrate specificity of SARS coronavirus 3C-like proteinase. Biochem Biophys Res Commun 329: 934–940

Gan YR, Huang H, Huang YD, Rao CM, Zhao Y, Liu JS, Wu L, Wei DQ (2006) Synthesis and activity assess of an octapeptide inhibitor designed for SARS coronavirus main proteinase. Peptides 27: 622–625

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19: 163–164

Holmquist R, Goodman M, Conroy T, Czelusniak J (1983) The spatial distribution of fixed mutations within genes coding for proteins. J Mol Evol 19: 437–448

Hu LD, Zheng GY, Jlang HS, Xai Y, Zhang Y, Kong XY (2003) Mutation analysis of 20 SARS virus genome sequences: evidence for negative selection in replicase ORF1b and spike gene. Acta Pharmacol Sinica 24: 741–745

Hubbard SR (1997) Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. Embo J 16: 5572–5581

Kang TB, Liang NC (1997) Studies on the inhibitory effects of quercetin on the growth of HL-60 leukemia cells. Biochem Pharmacol 54: 1013–1018

Kesel AJ (2005) Synthesis of novel test compounds for antiviral chemotherapy of severe acute respiratory syndrome (SARS). Curr Med Chem 12: 2095–2162

Liang GZ, Li SZ (2007) A new sequence representation (FASGAI) as applied in better specificity elucidation for human immunodeficiency virus type 1 protease. Biopolymers 88: 401–412

Lichtarge O, Sowa ME (2002) Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 12: 21–27

Lubec G, Afjehi-Sadat L, Yang JW, John JP (2005) Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol 77: 90–127

Mihalek I, Res I, Lichtarge O (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. J Mol Biol 336: 1265–1282

Mihalek I, Res I, Lichtarge O (2006a) Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. Bioinformatics 22: 1656–1657

Mihalek I, Res I, Lichtarge O (2006b) A structure and evolution-guided Monte Carlo sequence selection strategy for multiple alignment-based analysis of proteins. Bioinformatics 22: 149–156

Ran RQ, Zhou GP, Lu AG, Zhang L, Tang Y, Zhu HY, Rigby AC, Sharp FR (2004) Hsp70 mutant proteins modulate additional apoptotic pathways and improve cell survival. Cell Stress Chaperones 9: 229–242

Schnell JR, Zhou GP, Zweckstetter M, Rigby AC, James J, Chou JJ (2005) Rapid and accurate structure determination of coiled-coil domains using NMR dipolar couplings:application to cGMP-dependent protein kinase Iα. Protein Sci 14: 142421–142428

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680

Troy FA, Cho J-W, Ye J (1993) Polysialic acid: from microbes to man. In: Roth J, Rutishauser U, Troy F (eds) A. Birkhauser, Basel, pp 93–111

Wang JF, Wei DQ, Li L, Zheng SY, Li YX, Chou KC (2007a) 3D structure modeling of cytochrome P450 2C19 and its implication for personalized drug design. Biochem Biophys Res Commun (Corrigendum: ibid, 2007, Vol. 357, 330) 355, 513–519

Wang SQ, Du QS, Chou KC (2007b) Study of drug resistance of chicken influenza a virus (H5N1) from homology-modeled 3D structures of neuraminidases. Biochem Biophys Res Commun 354: 634–640

Wei DQ, Du QS, Sun H, Chou KC (2006a) Insights from modeling the 3D structure of H5N1 influenza virus neuraminidase and its binding interactions with ligands. Biochem Biophys Res Commun 344: 1048–1055

Wei DQ, Sirois S, Du QS, Arias HR, Chou KC (2005) Theoretical studies of Alzheimer's disease drug candidate [(2,4-dimethoxy) benzylidene]-anabaseine dihydrochloride (GTS-21) and its derivatives. Biochem Biophys Res Commun 338: 1059–1064

Wei DQ, Zhang R, Du QS, Gao WN, Li Y, Gao H, Wang SQ, Zhang X, Li AX, Sirois S, Chou KC (2006b) Anti-SARS drug screening by molecular docking. Amino Acids 31: 73–80

Wei H, Zhang R, Wang C, Zheng H, Chou KC, Wei DQ (2007) Molecular insights of SAH enzyme catalysis and their implication for inhibitor design. J Theor Biol 244: 692–702

Wu YS, Lin WH, Hsu JT, Hsieh HP (2006) Antiviral drug discovery against SARS-CoV. Curr Med Chem 13: 2003–2020

Yu XJ (2003) Putative hAPN receptor binding sites in SARS-CoV spike protein. Acta Pharmacol Sin 24: 481–488

Zhang R, Wei DQ, Du QS, Chou KC (2006) Molecular modeling studies of peptide drug candidates against SARS. Med Chem 2: 309–314

Zhang XW, Yap YL (2004) Exploring the binding mechanism of the main proteinase in SARS-associated coronavirus and its implication to anti-SARS drug design. Bioorg Med Chem 12: 2219–2223

Zhou GP, Troy FA (1995) 2-D NMR analyses reveals a specific interaction between polyisoprenols (PIs) and the polyisoprenol recognition sequences (PIRS) in model membranes. Glycoconi J 12: 434

Zhou GP, Troy FA (2003) Characterization by NMR and molecular modeling of the binding of polyisoprenols (PI) and polyisoprenyl recognition sequence (PIRS) peptides: three-dimensional structure of the complexes reveals sites of specific interactions. Glycobiology 13: 51–71

Zhou GP, Troy FA (2005a) Invited review: NMR studies on how the binding complex of polyisoprenol recognition sequence peptides and polyisoprenols can modulate membrane structure. Curr Protein Peptide Sci 6: 399–411

Zhou GP, Troy FA (2005b) NMR study of the preferred membrane orientation of polyisoprenols (dolichol) and the impact of their complex with polyisoprenyl recognition sequence peptides on membrane structure. Glycobiology 15: 347–359

Zhou GP, Surks HK, Schnell JR, Chou JJ, Michael E, Mendelsohn ME, Rigby AC (2004) The three-dimensional structure of the cGMP-dependent protein kinase I-α leucine zipper domain and its interaction with the myosin binding subunit. Blood 104: 963a

Zhu H, Domingues FS, Sommer I, Lengauer T (2006) NOXclass: prediction of protein–protein interaction types. BMC Bioinformatics 7: 27

Authors' address: Shao-Wu Zhang, College of Automation, Northwestern Polytechnical University, Xi'an, 710072, China,
Fax: +86-29-88494352, E-mail: zhangsw@nwpu.edu.cn