Amino Acids (2008) 34:643–652
DOI 10.1007/s00726-007-0017-2

ORIGINAL ARTICLE

Robustness of the residue conservation score reflecting both frequencies and physicochemistries

X.-S. Liu · W.-L. Guo

Received: 22 October 2007 / Accepted: 7 December 2007 / Published online: 4 January 2008

Abstract Measuring residue conservation at aligned positions has many applications in biology. Recently, a new conservation score has been defined. Unlike the previous methods, the new approach considers both residue frequencies and physicochemistries. Specifically, it measures physicochemistries based on BLOSUM matrices disregarding the meaning of the entries in such matrices, which may involve the problem of log–log probability. In this paper we present a conservation measure that also reflects both frequencies and physicochemistries while considering the fact that the entries of BLOSUM matrices are already interpreted as log probability. When the supposed score is applied to 14 protein examples, the results show that these two conservation scores are equivalent aside from the different score ranges. The method is also used to score the functional sites of three protein families. Compared with the widely used entropy-based methods, the resulting scores are more robust and consistent in the sense that the functional sites are much more conserved because of functional constraints.

Keywords Conservation score · Physicochemistry · Multiple sequence alignment · Functional site · Protein folding nucleus

X.-S. Liu (✉) · W.-L. Guo
Institute of Nanoscience, Academy of Frontier Science, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
e-mail: xsliu@nuaa.edu.cn

W.-L. Guo
e-mail: wlguo@nuaa.edu.cn

Introduction

Proteins tend to form distinct families and superfamilies based on their homologies and similarities. Homologous proteins within a protein family usually share the same fold and possess related functions (Orengo et al. 1994; Murzin et al. 1995). By placing the sequence in the framework of the overall family, multiple sequence alignments can be used in the analysis of protein function and evolutionary relationships. Also, to get the correct sequence alignment (particularly the multiple sequence alignments) is vitally important for predicting the 3-dimentional structure of a query protein based on the homology principle and timely providing useful information for drug design (see, e.g. Chou 2004a, b, c, 2005a, b; as well as a comprehensive review article (Chou 2004d) and monograph (Chou 2006)). Specially, identifying conserved regions of proteins is extremely useful in many situations. For example, a certain conservation score can be used for reading evolutionary signals about stability, folding kinetics and function (Hannenhalli and Russell 2000; Mirny and Shakhnovich 1999; Plaxco et al. 2000; Soyer and Goldstein 2004), for guiding both analysis and prediction of protein–protein interfaces (Valdar and Thornton 2001a), and for designating biologically relevant crystal contacts (Valdar and Thornton 2001b). A number of methods of conservation analysis have been proposed over the last 30 years. The first kind of scores reflected only amino acid frequencies (Jores et al. 1990; Lockless and Ranganathan 1999; Wu and Kabat 1970), and some of them are represented according to Shannon's information entropy and von Neumann entropy (Gerstein and Altman 1995; Mirny and Shakhnovich 1999; Shenkin et al. 1991; Zhang et al. 2007). The second kind of scores considered only the physicochemical properties of the amino acids in a column (Taylor

1986; Zvelibil et al. 1987). Another kind of scores, which are so-called "sum-of pairs" (SP) scores, measures the conservation by calculating the sum of all possible pairwise similarities between residues in an aligned position (Armon et al. 2001; Pilpel and Lancet 1999; Thompson et al. 1997; Valdar and Thornton 2001a).

However, despite the applications of these scores in the analysis of conservation none of them has achieved both biological and statistical rigor and appeared as a generally accepted standard, as pointed out by Valdar in an instructive review (Valdar 2002). Recently, a new conservation score that tries to consider both the physicochemistries and the frequencies of amino acids has been proposed (Liu et al. 2006). Specifically, it measures residue physicochemistries based on BLOSUM62 substitution matrix disregarding the meaning of the entries in this matrix, which may involve the problem of log–log probability. In this paper we present a similar residue conservation measure that reflects both frequencies and physicochemistries while thinking of the fact that the entries of BLOSUM matrices are already interpreted as log probability. We illuminate the statistical meaning of the present score in this consideration. When the score supposed in this paper is applied to 14 protein examples, the consistent results shows that these two conservation scores are equivalent besides the different score ranges. The method is also used to measure the functional sites of three well known protein families. The results indicate that, compared with the widely used entropy-based methods, the resulting scores are more robust. The new approach produces significantly larger percents of conservation for functional sites, which may be more consistent with the conclusion that the functional sites are much more conserved because of functional constraints.

Materials and methods

Grouping the universe of columns

As in Liu et al. (2006), we consider amino acids as symbols in an alphabet. The universe of columns is grouped into 20 sets, each of which contains the columns that are dominated by one of 20 kinds of amino acids (for convenience sake, for example, if the residue D (aspartic acid) dominates in a column, we call this column the D-dominated column). Because of the constraints of physicochemical properties, the degree of conservation of an amino acid should be different in a different column type. For example, in a D-dominated column, obviously D has the highest degree of evolutionary conservation. E has higher degree of evolutionary conservation than F in this column, since F is large and nonpolar, whereas D and E are both smaller and polar. However, F has the highest degree of evolutionary conservation in an F-dominated column. We shall quantify the degrees of evolutionary conservation of 20 kinds of amino acids for 20 different column types in the following.

Quantifying the degree of symbol's conservation

In a D-dominated column, we consider that all the substitutions are those of the amino acids in this column for the symbol D, and the mutations are independent each other [we then consider the evolutionary correlations between sequences by sequence weighting in the final formula (3)]. Then the degree of evolutionary conservation of an amino acid in this column can be measured by the similarity of physicochemical property between D and this symbol. Because substitution matrices provide a quantitative and reasonably objective assessment of amino acid substitution and similarity, we use the widely accepted BLOSUM62 substitution matrix (Henikoff and Henikoff 1992), to measure the similarity of physicochemical property between, for example, residue D and each symbol type in a D-dominated column.

We normalize the BLOSUM62 substitution matrix to obtain a similarity matrix **S** (Fig. 1), so that $S(a,a) = 10$, and $2 \leq S(a,b) \leq 9$ for different symbols a and b (Liu et al. 2006). Each line of the similarity matrix represents the similarity scores between the residue at diagonal and each of 20 kinds of amino acids. In addition, we set 0 as the similarity score for the gap heuristically, as done similarly by previous methods (e.g., Armon et al. 2001; Thompson et al. 1997), and then it contributes the lowest similarity score. It should be noticed that a substitution matrix cannot be used directly in our scoring procedure [similarly for the related scoring methods reviewed by Valdar (2002)], since any symbol should be assigned an equally maximum score in the column dominated by this symbol.

Conservation score for a column

Without loss of generality, we consider a D-dominated column. Let a_i and n_i ($i = 1, \ldots, 20$) be the similarity score from the similarity matrix **S** and the frequency for the ith amino acid type in a D-dominated column respectively. Here a_i, $i = 1, \ldots, 20$, are located on the line 4 of Fig. 1 (generally, for each of 20 column types, a_i, $i = 1, \ldots, 20$, lie on the line of Fig. 1 on which the corresponding dominated amino acid type is evaluated by the highest score 10). To avoid the problem of log–log probability, rather than taking the logarithm of a_i as in Liu et al. (2006), we then define the conservation score for this column to be that

$$\pi = \sum_{i=1}^{20} n_i a_i \qquad (1)$$

Fig. 1 The similarity matrix **S**, normalized from the BLOSUM62 substitution matrix by the method as in Liu et al. (2006). The normalization ensures that $S(a,a) = 10$, and $2 \leq S(a,b) \leq 9$ for different symbols a and b. Each *line* of the similarity matrix represents the similarity scores between 20 amino acids and one of them, say, D, in a D-dominated column

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	10	4	3	3	6	4	4	6	3	4	4	4	4	3	4	7	6	2	3	6
R	4	10	5	3	2	6	5	3	5	2	3	7	4	2	3	4	4	2	3	2
N	3	5	10	6	3	5	5	5	6	3	3	5	3	3	3	6	5	2	3	3
D	3	3	6	10	3	5	7	4	4	3	2	4	3	3	4	5	4	2	3	3
C	4	3	3	3	10	3	2	3	3	4	4	3	4	3	3	4	4	3	3	4
Q	4	6	5	5	2	10	7	3	5	2	3	6	5	2	4	5	4	3	4	3
E	5	5	5	7	2	7	10	4	5	3	3	6	4	3	5	5	5	3	4	4
G	5	3	5	4	3	3	3	10	3	2	2	3	3	3	3	5	3	3	3	3
H	2	4	5	3	2	4	4	2	10	2	2	3	2	3	2	3	2	2	5	2
I	5	3	3	3	5	3	3	2	3	10	8	3	7	6	3	4	5	3	5	9
L	5	4	3	2	5	4	3	2	3	8	10	4	8	6	3	4	5	4	5	7
K	4	7	5	4	2	6	6	3	4	2	3	10	4	2	4	5	4	2	3	3
M	4	4	3	2	4	5	3	2	3	6	7	4	10	5	3	4	4	4	4	6
F	3	3	3	3	3	3	3	3	4	5	5	3	5	10	2	3	3	6	8	4
P	4	3	3	4	2	4	4	3	3	2	2	4	3	2	10	4	4	2	2	3
S	7	4	7	6	4	6	6	6	4	3	3	6	4	3	4	10	7	2	3	3
T	4	3	4	3	3	3	3	2	2	3	3	3	3	2	3	6	10	2	2	4
W	2	2	2	2	3	3	2	3	3	2	3	2	3	4	2	2	3	10	5	2
Y	3	3	3	2	3	3	3	2	6	3	3	3	3	7	2	3	3	6	10	3
V	6	2	2	2	4	3	3	2	2	9	7	3	7	4	3	3	6	2	4	10

Statistical meaning

We illuminate the statistical meaning of the present score when considering the fact that the entries of BLOSUM matrices are already interpreted as log probability. Assume that N sequences exist in the underlying MSA (then $\sum_{j=1}^{20} n_j = N$). Let B be a substitution matrix (here the BLOSUM62 substitution matrix) with elements b_{ij}, i, $j = 1, \ldots, 20$. We known that $b_{ij} = \mathrm{c}\ \log(p_{ij}/p_i p_j)$, where c is a constant, p_{ij} is the probability that a substitution occurs between the ith amino acid type and the jth amino acid type, p_i is the marginal probability meaning the expected probability of occurrence of the ith amino acid type, and $p_{ij}/p_i p_j$ is called the odds ratio for the substitution between the ith amino acid type and the jth amino acid type (Henikoff and Henikoff 1992). For a D-dominated column, we set $d_i = p_{Di}/(p_D p_i)$ for its convenience, the odds ratio for a substitution between D and the ith amino acid type. It is well understood that the value of the odds ratio d_i indicates the similarity degree in physicochemical property between D and the ith amino acid type. Also, the similarity degree between D and the ith amino acid type indicates the probability for the ith amino acid type to appear in a D-dominated column. Then the probability for the ith amino acid type to appear in a D-dominated column is directly proportional to the odds ratio d_i, that is to say, the bigger the value of the odds ratio d_i, the larger the probability for the ith amino acid type to appear in this column. Let $q_i = d_i / \sum_{j=1}^{20} d_j$, the normalization of d_i. Therefore, the normalized odds ratio q_i just describes the probability that the ith amino acid type appears in this D-dominated column (here we use the odds ratios rather than the transition probabilities for amino acid substitutions, because the expected probability distribution of occurrence of amino acids in nature is not uniform, while the odds ratios give a more appropriate quantity scale for applications). Because n_i and q_i ($i = 1, \ldots, 20$) are the frequency and probability for the ith amino acid type to appear in this D-dominated column, respectively, $\theta = \log \prod_{i=1}^{20} q_i^{n_i} = \sum_{i=1}^{20} n_i \log q_i$ describes the logarithm of the probability for this column to take place.

From the derivation of the suggested similarity matrices we have that the similarity score $a_i = \mathrm{m} \log d_i + n$, where m and n are constants and $m > 0$, which is an increasing function of d_i (Liu et al. 2006). Thus a_i can also describes the degree of similarity between D and the ith amino acid type as d_i does, and then it is well defined. Thus from Eq. (1) we have

$$\pi = \sum_{i=1}^{20} n_i a_i = \sum_{i=1}^{20} n_i (m \log d_i + n) = m \sum_{i=1}^{20} n_i \log d_i + nN.$$

Referring to that $\theta = \sum_{i=1}^{20} n_i \log q_i$ and $q_i = d_i / \sum_{j=1}^{20} d_j$, we further obtain that

$$\begin{aligned} \pi &= m \sum_{i=1}^{20} n_i \log q_i + m \sum_{i=1}^{20} n_i \log \left(\sum_{i=1}^{20} d_i \right) + nN \\ &= m\theta + h, \end{aligned} \tag{2}$$

where h is a constant for a derived MSA. The score π is a linear transformation of θ, so they are trivially different as a conservation score. We use π, rather than θ, since π purports to have conveniently bounded range: π ranges from zero, when all symbols in this column are gaps, to $\pi_{\max} = 10N$, when objects of only one type are present, and its values increase with increasing conservation. Hence the score π defined by Eq. (1) is related to the logarithmic probabilities for aligned positions to take place, indicating a definite statistical meaning for it.

Incorporating the sequence weighting method

We now consider the evolutionary correlations between sequences in our conservation scores by sequence weighting method that is a common concern both for scoring residue conservation in a column and for building sequence profiles. A large number of methods for sequence weighting have been suggested in literature (Durbin et al. 1998; Henikoff and Henikoff 1994). A widely applied formulation (Henikoff and Henikoff 1994) weights sequences at individual positions in an alignment and then combines position weights to give sequence weights. The weight of the ith sequence at position x is $\omega_{i_x} = 1/k_x n_{x_i}$, where k_x is the number of amino acid types presented in column x and n_{x_i} is the frequency of the ith sequence's amino acid at that position. By averaging along all positions in an alignment, each sequence then has weight

$$\omega_i = \frac{1}{L} \sum_{x}^{L} \omega_{i_x},$$

where L is the length of the alignment.

In our present model, the sequences in underlying MSA are weighted by the above weighting metric. For example, in a D-dominated column, assume that the weights of n_i sequences whose symbol in this column is just the ith amino acid type are respectively $\omega_{i1}, \ldots, \omega_{in_i}$, where n_i (i = 1, …, 20) is the frequency for the ith amino acid type in this column. Then by substituting n_i with $\sum_{j=1}^{n_i} \omega_{ij}$. we rewrite the conservation score for this column as

$$\pi = \sum_{i=1}^{20} \left(\sum_{j=1}^{n_i} \omega_{ij} \right) a_i. \tag{3}$$

That is, the net effect of n_i sequences whose symbol in this column is just the ith amino acid type is of having $\sum_{j=1}^{n_i} \omega_{ij}$ of them, and then the net number of n_i ith amino acid types in this column is $m_i = \sum_{j=1}^{n_i} \omega_{ij}$. In this case the dominated residue is determined by the maximum of all m_i, i = 1, …, 20. From the definition of the weights we can see that π in Eq. (3) is bounded from zero, when all symbols in this column are gaps, to 10, when objects of only one type are present.

The multiple sequence alignments

First, we use the scoring method suggested in this paper to observe the relationship between an amino acid's evolutionary conservation and its role in the Φ-value defined protein-folding nucleus. It is widely accepted that protein folding occurs via the formation of a small region of native-like structure that serves as a nucleus upon which further residues condense in a process analogous to a phase transition. Experimentally determined Φ-values provide a readily obtainable, objective means of quantifying participation in the native-like transition–state interactions that define the kinetic folding nucleus (Fersht 1997; Shaknovich et al. 1996; Tseng and Liang 2004). One of the most intriguing aspects of the protein-folding nucleus is its relation to protein evolution. As in Liu et al. (2006), the studies in question use directly the 14 highest quality sequence alignments available from supplementary material of the study by Larson and co-authors (Larson et al. 2002).

We then apply our scoring method to the functional sites and compare the resulting scores with the entropy-based scores. For this purpose we use directly the well-derived multiple sequence alignments available at http://www.gpcr.org/articles/ from supplementary material of the study by Oliveira and co-authors (2003). The sequences in these three sequence alignments are from three well-known sequence families: Globins, Ras-like proteins and Serine-proteases, respectively. The sequence alignment for Globins contains 753 sequences with 113 actual aligned positions, the sequence alignment for Ras-like proteins contains 562 sequences with 152 actual aligned positions, and the sequence alignment for Serine-proteases contains 301 sequences with 173 actual aligned positions. We use these three families because the role of almost every residue in these families is known from literature (Oliveira et al. 2003).

Results and discussion

Consistency with biochemical judgment

A number of methods of conservation analysis have been proposed in the literature. It has been illustrated by a hypothetical multiple-sequence alignment in the review by Valdar that the previous scores that considered only one aspect of the amino acid frequencies and the physicochemical properties of amino acids cannot make sense well in biochemistry (Valdar 2002). In applications the SP scores seem better, but they do not make sense in what the statistic means. It is implausible that the diversity in a column arises from all the pairwise amino acid substitutions (Valdar 2002; Durbin et al. 1998). Additionally, by calculating conservation scores for many real alignments Pei and Grishin (2001) showed that the usage of the entropy-based conservation measures are not inferior to that of the SP measure.

Mirny and Shakhnovich (1999) gave the reduced entropy score that attempted to incorporate physicochemical properties into the entropy score. Their score is

$$D = \sum_{i=1}^{K} p_i \ln p_i, \tag{4}$$

where $K = 6$. The set of K partitions is: aliphatic [AVLIMC], aromatic [FWYH], polar [STNQ], positive [KR], negative [DE], special conformations [GP]. There is an improvement over pure entropy-based scores. Meanwhile, for the reduced entropy one should make a subjective partitioning of the 20 kinds of amino acids that accounts for some physicochemical properties but ignores relative frequencies within a partition, and further the residues of different types belonging to a same division should be considered to have the same physicochemical properties, which may induce the disagreeable variableness of it as illustrated in the following. But, the present scoring method considers the similarity score of each amino acid, and then the derived conservation scores for the aligned positions are more consistent (see the following). With simple computations it is illustrated by the hypothetical multiple-sequence alignment in Valdar (2002) that the present score can be consistent with the biochemical judgment. For example, the score in present study can correctly reproduce the conservation ranks (a) > (b) > (c) > (e) > (f) by taking account of the amino acid frequencies in a column; the score can recognize that some substitutions incur more chemical and physical change than others by considering the physicochemistries of amino acids.

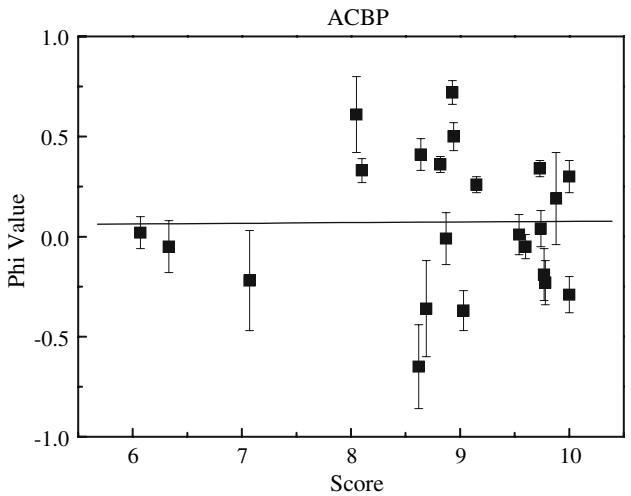

Fig. 2 Correlations between Φ-values and conservation scores calculated by the present scoring method for the protein ACBP (the data for Φ-value analysis are provided by Plaxco and the alignments are available from supplementary material of the study by Larson and co-authors). We have that $r^2 = 0.0001$ (the other 13 proteins are not presented to save space)

Conservation of the protein-folding nucleus

We discuss correlations between Φ-value and sequence conservation. For each of 14 protein examples available from supplementary material of the study by Larson and co-authors (2002), we obtain similar results as those in Liu et al. (2006), which shows that there is little such correlation and then that residues participating more strongly in the nucleus will not be relatively better conserved (one example in Fig. 2, the others are not presented to save space). We have examined the conservation of the folding nucleus by defining participation in it as coinciding with Φ-values >0.5. We also fail to observe any statistically significant conservation of residues for this definition of the folding nucleus (Fig. 3). The formation of the folding nucleus for one protein segment does not only depend on sequence conservation in this region but also the ability of forming certain secondary structure of this segment in solution, which might be one of reasons why there is little correlation between Φ-values and conservation scores. The present results are consistent with those in Liu et al. (2006) indicating that these two conservation scores are equivalent aside from the different score ranges. However, here we have considered the probabilistic meaning for the entries of BLOSUM matrices when we describe the present scores as the logarithmic probabilities for aligned positions to take place.

Conservation scores for functional sites

It is widely accepted that natural selection makes the functionally important residues to be more conserved in proteins in order to preserve biological activity (Mirny and

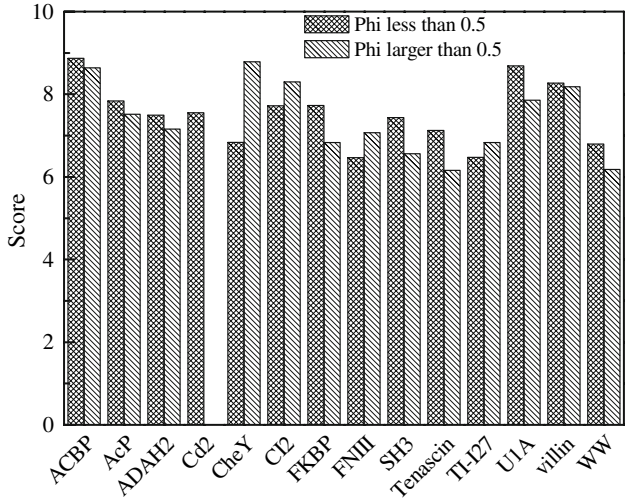

Fig. 3 The mean scores of high Φ residues ($\Phi > 0.5$) and low Φ residues for the 14 proteins calculated by the present scoring method. Little evidence is observed in favor of preferential conservation of the folding nucleus (as defined by $\Phi > 0.5$): for four proteins these residues are more conserved than low Φ residues, for nine they are less well conserved (note, none of the few residues that have been characterized in CD2.d1 exhibit a Φ-value of >0.5)

Shakhnovich 1999). The entropy-based scores, are widely used to characterize the function related sites in a MSA, and prove superior to the others (Hannenhalli and Russell 2000; Mirny and Shakhnovich 1999; Oliveira et al. 2003; Pei and Grishin 2001). Thus, as a test for the present scoring method we apply it to the functional sites and compare the resulting scores only with the pure entropy scores and the reduced entropy scores.

We consider the functional sites for three protein families, which are located at Box 11 and Box 12 of Fig. 3 in the study by Oliveira and co-authors. Box 11 contains residue positions form the main functional site. These residue positions are involved in catalysis or signaling mechanisms. Most positions of key structural residues (e.g., Cys–Cys bridges) are also found in this box. These positions should be highest conserved because of the main functional constraints. Box 12 contains positions of residues located in the core. They are adjacent to the positions of Box 11 and mainly form the first shell of positions around the main functional site. We calculate the scores of these positions for the three protein families, respectively, by the scoring method suggested in this paper and the reduced entropy in Eq. (4) in which the 20 amino acids are divided into 6 kinds for accounting for some physico-chemical properties, whereas the scores of these positions from the pure Shannon entropy with original 20 kinds of amino acids were provided in the supplementary material of the study by Oliveira and co-authors (Tables 1, 2, 3). These three kinds of scores have different ranges which are the intervals [0, 10], [0, log6] and [0, log20], respectively,

Table 1 The conservation scores of functional positions for Globins measured by the three kinds of scoring methods

Position	*V*	E(20)	E(6)	*B*	AA
Box 11					
20	3	0.09	0.12	9.70	GACV
24	3	0.08	0.04	9.87	LFM
26	3	0.12	0.11	9.79	RSK
32	1	0.00	0.03	9.82	P
34	2	0.08	0.04	9.66	TS
38	1	0.00	0.01	9.96	F
48	2	0.06	0.08	9.65	HQ
49	1	0.01	0.04	9.79	G
68	1	0.00	0.00	9.98	L
72	1	0.00	0.00	10	H
103	2	0.09	0.14	9.46	KNH
116	1	0.01	0.05	9.68	Y
Box 12					
3	3	0.67	0.06	8.53	EDN
7	3	0.72	0.01	9.52	VIL
11	3	0.41	0.10	8.95	WFL
28	3	0.75	0.69	7.85	LFI
42	3	0.73	0.10	8.47	SND
45	3	0.70	0.00	9.26	VLI
51	4	0.70	0.61	7.90	KTVR
52	2	0.24	0.04	9.83	VI
55	4	1.01	0.93	8.50	ASGK
56	4	1.04	0.63	8.73	LFIV
69	2	0.33	0.34	9.43	SA
73	4	1.00	0.32	8.28	ACSF
76	3	0.38	0.35	8.71	LHI
78	2	0.35	0.01	9.78	VI
79	2	0.33	0.38	8.91	DP
80	3	0.39	0.34	8.84	PVI
82	2	0.31	0.33	8.94	NY
83	2	0.30	0.31	9.44	FL
86	4	0.47	0.10	9.32	LIFP
98	3	0.67	0.76	7.65	QHA
112	3	0.44	0.01	9.46	LIM
117	4	1.00	0.73	5.92	HRKF

and our present score increases with increasing conservation while the others do the converse. For comparison sake, the intervals are equally divided into ten subintervals respectively. The extreme points of the subintervals are *i*, $i = 0,1, \ldots, 10$, for the interval [0, 10], are 0, 0.179, 0.385, 0.538, 0.717, 0.896, 1.075, 1.254, 1.433, 1.613 and 1.792 for the interval [0, log6], and 0, 0.300, 0.599, 0.899 1.198, 1.498, 1.797, 2.097, 2.397, 2.696 and 2.996 for the interval [0, log20]. Then we can know that the score subinterval of highest conservation measured by the present method is [9, 10], and they are [0, 0.179] and [0, 0.300] by the reduced

Table 2 The conservation scores of functional positions for Ras-like proteins measured by the three kinds of scoring methods

Position	*V*	E(20)	E(6)	*B*	AA
Box 11					
2	2	0.12	0.08	9.54	KRQ
7	1	0.00	0.03	9.84	G
12	1	0.00	0.01	9.91	G
13	1	0.00	0.01	9.90	K
55	1	0.01	0.03	9.87	D
56	3	0.13	0.08	9.60	TIS
57	3	0.12	0.10	9.72	ATG
58	1	0.01	0.04	9.90	G
59	4	0.20	0.12	9.41	QTSL
60	5	0.26	0.14	9.54	EDAHT
105	3	0.22	0.26	9.40	GAR
107	3	0.10	0.14	9.77	KQF
109	1	0.00	0.06	9.85	D
124	4	0.28	0.18	9.43	EDSF
126	1	0.02	0.09	9.66	ST
127	4	0.32	0.31	9.48	ASRVK
137	6	0.30	0.29	9.15	FSEIA
Box 12					
3	6	1.39	0.28	7.93	LVICF
4	5	1.10	0.05	8.82	VLICA
6	5	1.15	0.09	8.77	IVLFA
10	6	1.00	0.94	7.58	GASNC
11	5	0.60	0.26	9.27	VTCAS
14	2	0.69	0.00	8.45	ST
15	6	1.48	0.71	5.68	SCANT
16	5	0.92	0.30	9.24	LIFMV
17	6	1.34	0.47	8.19	LTVMI
19	6	1.12	0.94	7.31	RQSVC
20	4	1.00	0.29	8.68	FYLH
25	5	0.46	0.12	9.51	FYIVP
33	3	0.46	0.35	9.14	TIV
34	6	1.01	0.47	8.42	IVGLK
51	6	0.98	0.36	8.71	LAFMI
53	4	0.86	0.05	9.08	ILVC
54	6	0.59	0.52	8.08	WLVTF
62	4	0.76	0.15	8.53	FYLGS
69	4	0.80	0.56	7.71	YSFH
70	5	0.81	0.57	7.83	YMIFS
75	6	1.24	0.76	6.61	GVCAH
78	5	1.21	0.08	8.66	LIVCM
80	4	0.73	0.13	8.25	YFVT
81	4	0.85	0.91	7.44	DSAY
82	5	1.05	0.16	8.79	IVLDT
87	3	0.65	0.11	8.80	STR
88	4	0.73	0.28	8.91	FYLS
103	5	0.54	0.03	9.57	LVIMF
106	5	0.48	0.17	8.89	NTCLS

Table 3 The conservation scores of functional positions for Serine-proteases measured by the three kinds of scoring methods

Position	*V*	E(20)	E(6)	*B*	AA
Box 11					
4	5	0.17	0.17	9.57	GEAKS
13	4	0.32	0.31	9.50	PRAK
25	3	0.09	0.06	9.73	CGA
26	4	0.32	0.30	9.44	GSAM
27	3	0.32	0.31	9.45	GAC
39	2	0.06	0.00	9.95	AV
40	3	0.17	0.16	9.83	ATV
41	2	0.05	0.04	9.95	HR
42	3	0.08	0.04	9.75	CTA
47	3	0.24	0.19	9.72	GRQ
79	2	0.04	0.02	9.98	DN
82	4	0.35	0.02	9.73	LIVM
113	4	0.27	0.27	9.32	GDCK
114	4	0.26	0.06	9.38	WFYE
115	5	0.30	0.28	9.23	GTIEQ
123	1	0.00	0.02	9.94	C
130	1	0.00	0.00	10	C
141	2	0.07	0.00	9.89	DE
142	1	0.01	0.00	9.97	S
143	2	0.14	0.13	9.70	GC
144	4	0.25	0.20	9.57	GQSA
145	5	0.26	0.15	9.32	PGVAK
155	5	0.26	0.24	9.57	GVEHI
158	2	0.05	0.18	9.76	SA
168	5	0.33	0.43	8.64	PYRLS
180	1	0.00	0.02	9.96	W
Box 12					
29	4	0.64	0.07	9.36	LIVY
30	5	1.04	0.19	9.21	ILVYM
36	3	0.57	0.00	9.66	VIL
37	5	1.03	0.00	9.01	LVMIA
38	2	0.57	0.00	9.20	TS
80	5	0.77	0.12	9.54	ILVYF
83	4	1.05	0.04	8.88	LIVM
132	5	0.44	0.56	9.15	GAEYD
138	5	0.66	0.53	8.79	CFYGA
170	5	0.92	0.21	8.90	VIAFL
174	3	0.71	0.02	9.49	VLI
181	4	0.46	0.00	9.80	IVLM

The functional sites for three protein families, Globins, Ras-like proteins and Serine-proteases, are located at Box 11 and Box 12 of Fig. 3 in the study by Oliveira and co-authors (2003). Box 11 contains residue positions form the main functional site. Box 12 contains positions of residues located in the core, which mainly form the first shell of positions around the main functional sites. Here, position stands for the column number in the alignment of Fig. 2a in the study by Oliveira and co-authors, *V* is the variability at a position in the multiple sequence alignment, which is the number of different residue types observed at this position in at least 0.5% of all sequences. E(20) represents the pure Shannon entropy with original 20 kinds of amino acids by which the scores of the positions are obtained (the data are given in supplementary material of the study by Oliveira and co-authors), E(6) represents the reduced entropy with 6 kinds of amino acids given in Eq. (4), and *B* represents the our new scoring method with the similarity matrix from the BLOSUM62 (Fig. 1). AA stands for the more frequent amino acids found which are given in supplementary material of the study by Oliveira and co-authors, and the first symbol is the dominated residue by our algorithms

Table 4 The distributions of the numbers of functional positions over the percent of conservation measured by the three kinds of scoring methods

Percent of conservation			The number of positions					
			90–100	80–90	70–80	60–70	50–60	40–50
Box 11	Globins	E(20)	12	0	0	0	0	0
		E(6)	12	0	0	0	0	0
		B	12	0	0	0	0	0
	Ras-like proteins	E(20)	15	2	0	0	0	0
		E(6)	13	4	0	0	0	0
		B	17	0	0	0	0	0
	Serine-proteases	E(20)	20	6	0	0	0	0
		E(6)	16	9	1	0	0	0
		B	25	1	0	0	0	0
Box 12	Globins	E(20)	1	9	8	4	0	0
		E(6)	9	6	1	3	2	1
		B	8	10	3	0	1	0
	Ras-like proteins	E(20)	0	5	10	9	5	0
		E(6)	12	6	4	3	1	3
		B	6	15	6	1	1	0
	Serine-proteases	E(20)	0	4	4	4	0	0
		E(6)	8	2	1	1	0	0
		B	9	3	0	0	0	0

Here, Box 11, Box 12, E(20), E(6), and *B* have the same meanings as in Table 1, 2, 3. 90–100 represents a more than 90% of conservation, and similarly for the others. For example, the first line in Box 12 represents the distribution of the number of the positions in Box 12 for Globins over the percent of conservation measured by the pure Shannon entropy method, that is, among 22 positions 1 position has the conservation of more than 90%, 9 positions have the conservation of between 80 and 90%, 8 positions have the conservation of between 70 and 80%, and 4 positions have the conservation of between 60 and 70%

entropy and the pure Shannon entropy respectively, indicating the conservation of more than 90%. The rest is deduced by analogy. The numbers of the functional positions over the percent of conservation measured by the three kinds of scoring methods are listed in Table 4.

As mentioned above, the positions in Tables 1, 2, 3 should be well conserved because of the functional constraints. A good scoring method should give these functional positions larger percents of the conservation. For the main functional positions in Box 11, our new scoring method produces a significantly larger number of positions with the conservation of more than 90% than that the entropy-based methods do, and the pure Shannon entropy seems to be better than the reduced entropy in this sense. For example, among 26 main functional positions of Serine-proteases, the numbers of positions with the conservation of more than 90% are 20, 16, 25 and 26, measured, respectively, by the pure Shannon entropy, the reduced entropy, the scoring method suggested in this paper. For the function-related positions in Box 12, our new scoring method gives a significantly larger number of positions with the conservation of more than 90% than that the pure Shannon entropy does, and can yield a larger number of positions with the conservation of more than 80% than that the reduced entropy does (Table 4). Contrarily, compared with the pure Shannon entropy with original 20 kinds of amino acids, the reduced entropy seems much improved in the sense that it produces a remarkably larger number of positions with the conservation of more than 90% (for example, among 22 positions of Globins in Box 12, the numbers of positions with the conservation of more than 90% are 1 and 9, measured respectively by the Shannon entropy and the reduced entropy).

Remarkably, among all the scoring methods the reduced entropy gives the smallest percents of conservation for some positions, indicating the most variable distributions (Fig. 4). In fact, the variableness of the reduced entropy may be due to its definition that makes a subjective partitioning of the 20 kinds of amino acids for accounting for some physicochemical properties but ignores the differences between residues within a partition. For example, the 42nd position of Globins contains the residues S, N and D, and is dominated by the residue S (Table 1). Because the residues S and N stay within the same partition defined by the reduced entropy, and they are more frequent than D, the percent of conservation for this position measured by the reduced entropy is much larger than that measured by

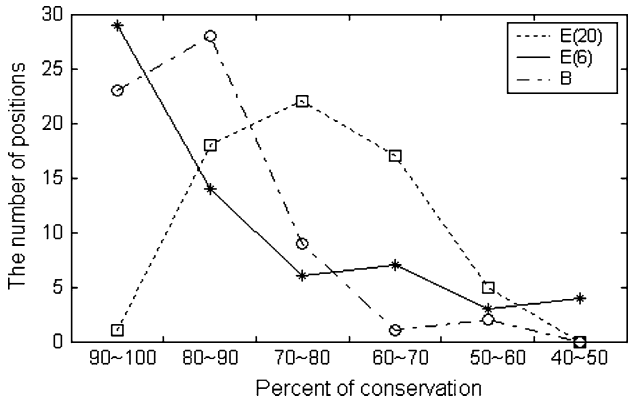

Fig. 4 The distributions of the numbers of function-related positions in Box 12 for all the three protein families over the percent of conservation measured by the three kinds of scoring methods, respectively. The distribution measured by the reduced entropy (E(6)) has a long tail in lower percent of conservation, indicating the variableness for this method

the pure Shannon entropy with original 20 kinds of amino acids (the former is between 90 and 100%, and the latter is between 70 and 80%). Inversely, because the 55th position of Globins contains the residues A, S, G and K (A is the dominated residue, and K is infrequent), and these residues belong to four different partitions, the percent of conservation for this position measured by the reduced entropy is even smaller than that measured by the pure Shannon entropy (the former is between 60 and 70%, and the latter is between 40 and 50%). However, since the similarity scores for N and D, 7 and 6, respectively, at the 42nd position (S is the dominated residue), are equal to those for S and G at the 55th position, respectively (Fig. 1), the present method yields a similar percent of conservation for these two positions.

In brief, the present scoring method definitely produces larger percents of conservation for functional sites, compared with the entropy-based methods. Unlike the reduced entropy that makes a subjective partitioning of the 20 kinds of amino acids for accounting for some physicochemical properties but ignores the differences between residues within a partition, the present scoring method considers the similarity score of each amino acid, and then the derived conservation scores for the aligned positions are more accurate and robust.

Conclusion

To summarize, in this study we present a conservation measure that reflects both frequencies and physicochemistries while considering the fact that the entries of BLOSUM matrices are already interpreted as log probability. It is illustrated by a hypothetical multiple-sequence alignment that the present score can be consistent with the biochemical judgment. The present score is more elaborate than the previous scores that considered only one aspect of the amino acid frequencies and the physicochemical properties of amino acids, and then it could be used in the situations in which the high precision in calculations is needed, such as, for characterizing functional sub-types of a protein (Hannenhalli and Russell 2000). The new residue score approach might also be effectively used to generate different kind of pseudo amino acid composition (Chou 2001, 2005c) so as to enhance the prediction quality for protein subcellular localization (Chou and Shen 2007a, b; Shen and Chou 2007a), enzyme functional class (Shen and Chou 2007b), membrane protein type (Chou and Shen 2007c), signal peptide (Chou and Shen 2007d; Shen and Chou 2007c), as well as protein structural class prediction (Chou and Zhang 1995; Chou 2005d). When the supposed score is applied to 14 protein examples, the results show that the present conservation score and the score suggested in Liu et al. (2006) are equivalent aside from the different score ranges. However, the present score considers the specific meaning of the entries of BLOSUM matrices and then avoids the problem of log–log probability when being described as the logarithmic probabilities for aligned positions to take place. The present score makes sense both statistically and biochemically, and one can obtain an exact conservation score for any position of a protein by using it. The scoring method has been used for measuring the conservation of the functional sites of three well known protein families. Compared with the widely used entropy-based methods, the resulting scores are more robust since the method considers the exact similarities between amino acids. Furthermore, the new approach produces more consistent conservation values for functional sites in the sense that the functional sites are much more conserved because of functional constraints.

Acknowledgments Our thanks to K. W. Plaxco for providing the data for Φ-value analysis, to G. Vriend for giving permission to use the alignments and some data from their article. The work is supported by 973 Program (2007CB936204), the Ministry of Education (No. 705021, IRT0534), National NSF and Jiangsu Province NSF of China.

References

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

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins (Erratum: ibid., 2001, Vol.44, 60) 43:246–255

Chou KC (2004a) Insights from modelling three-dimensional structures of the human potassium and sodium channels. J Proteome Res 3:856–861

Chou KC (2004b) Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. Biochem Biophys Res Commun 316:636–642
Chou KC (2004c) Molecular therapeutic target for type-2 diabetes. J Proteome Res 3:1284–1288
Chou KC (2004d) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134
Chou KC (2005a) Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. J Proteome Res 4:1681–1686
Chou KC (2005b) Modeling the tertiary structure of human cathepsin-E. Biochem Biophys Res Commun 331:56–60
Chou KC (2005c) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19
Chou KC (2005d) Review: progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Pept Sci 6:423–436
Chou KC (2006) Structural bioinformatics and its impact to biomedical science and drug discovery. In: Atta-ur-Rahman, Reitz AB (eds) Frontiers in medicinal chemistry, vol 3. Bentham Science Publishers, The Netherlands, pp 455–502
Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349
Chou KC, Shen HB (2007a) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370:1–16
Chou KC, Shen HB (2007b) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6:1728–1734
Chou KC, Shen HB (2007c) MemType-2L: a Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360:339–345
Chou KC, Shen HB (2007d) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357:633–640
Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge
Fersht AR (1997) Nucleation mechanisms in protein folding. Curr Opin Struct Biol 7:3–9
Gerstein M, Altman RB (1995) Average core structures and variability measures for protein families: application to the immunoglobulins. J Mol Biol 251:161–175
Hannenhalli SS, Russell RB (2000) Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol 303:61–76
Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919
Henikoff S, Henikoff JG (1994) Position-based sequence weights. J Mol Biol 243:574–578
Jores R, Alzari PM, Meo T (1990) Resolution of hypervariable regions in T-cell receptor chains by a modified Wu–Kabat index of amino acid diversity. Proc Natl Acad Sci USA 87:9138–9142
Larson SM, Ruczinski I, Davidson AR, Baker D, Plaxco KW (2002) Residue Participating in the protein folding nucleus do not exhibit preferential evolutionary conservation. J Mol Biol 316:225–233
Liu XS, Li J, Guo WL, Wang W (2006) A new method for quantifying residue conservation and its applications to the protein folding nucleus. Biochem Biophys Res Commun 351:1031–1036
Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286:295–299
Mirny LA, Shakhnovich EI (1999) Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 291:177–196
Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
Oliveira L, Paiva PB, Paiva ACM, Vriend G (2003) Identification of functionally conserved residues with the use of entropy-variability plots. Proteins 52:544–552
Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372:631–634
Pei JM, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17:700–712
Pilpel Y, Lancet D (1999) The variable and conserved interfaces of modeled olfactory receptor proteins. Protein Sci 8:969–977
Plaxco KW, Larson S, Ruczinski I, Riddle DS, Thayer EC, Buchwitz B, Davidson AR, Baker D (2000) Evolutionary conservation in protein folding kinetics. J Mol Biol 298:303–312
Shaknovich E, Abkevich V, Ptitsyn O (1996) Conserved residues and the mechanism of protein folding. Nature 379:96–98
Shen HB, Chou KC (2007a) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1011
Shen HB, Chou KC (2007b) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59
Shen HB, Chou KC (2007c) Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Commun 363:297–303
Shenkin PS, Erman B, Mastrandrea LD (1991) Information-theoretical entropy as a measure of sequence variability. Proteins 11:297–313
Soyer OS, Goldstein RA (2004) Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters. J Mol Biol 339:227–242
Taylor WR (1986) The classification of amino acid conservation. J Theor Biol 119:205–218
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882
Tseng YY, Liang J (2004) Are residues in a protein folding nucleus evolutionarily conserved? J Mol Biol 335:869–880
Valdar WSJ (2002) Scoring residue conservation. Proteins 48:227–241
Valdar WSJ, Thornton JM (2001a) Protein–protein interfaces: analysis of amino acid conservation in homodimers. Proteins 42:108–124
Valdar WSJ, Thomton JM (2001b) Conservation helps to identify biologically relevant crystal contacts. J Mol Biol 313:399–416
Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med 132:211–249
Zhang SW, Zhang YL, Pan Q, Cheng YM, Chou KC (2007) Estimating residue evolutionary conservation by introducing von Neumann entropy and a novel gap-treating approach. Amino Acids doi:10.1007/s00726-007-0586-0
Zvelibil MJ, Barton GJ, Taylor WR, Sternberg MJ (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 195:957–961