Available online at www.sciencedirect.com
ScienceDirect
Biochemical and Biophysical Research Communications 351 (2006) 1031–1036
BBRC
www.elsevier.com/locate/ybbrc

A new method for quantifying residue conservation and its applications to the protein folding nucleus

Xinsheng Liu [a,b,*], Jing Li [b], Wanlin Guo [a], Wei Wang [b]

[a] Institute of Nanoscience, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[b] National Lab of Solid State Microstructure, Department of Physics and Institute of Biophysics, Nanjing University, Nanjing 210093, China

Received 26 October 2006
Available online 7 November 2006

Abstract

The conservation of residues in columns of a multiple sequence alignment (MSA) reflects the importance of these residues for maintaining the structure and function of a protein. To date, many scores have been suggested for quantifying residue conservation, but none has achieved the full rigor both in biology and statistics. In this paper, we present a new approach for measuring the evolutionary conservation at aligned positions. Our conservation measure is related to the logarithmic probabilities for aligned positions, and combines the physicochemical properties and the frequencies of amino acids. Such a measure is both biologically and statistically meaningful. For testing the relationship between an amino acid's evolutionary conservation and its role in the Φ-value defined protein folding kinetics, our results indicate that the folding nucleus residues may not be significantly more conserved than other residues by using the biological-relevance weighted statistical scoring method suggested in this paper as an alternative to entropy-based procedures.

Keywords: Evolutionary conservation; Physicochemical property; Multiple sequence alignment (MSA); Folding nucleus

It is well known that the conserved patterns of amino acids in columns of a multiple sequence alignment (MSA) indicate the degree of the stability constraint and functional constraint of those positions. 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 [1–5], for guiding both analysis and prediction of protein–protein interfaces [6], and for designating biologically relevant crystal contacts [7]. In 1970, Wu and Kabat introduced the first widely applied measure of conservation [8]. Their score cares only about the frequencies of the most commonly occurring symbols at the aligned positions. Since then, a number of scoring methods have been proposed to quantify residue conservation. In general, these scores can be divided into several kinds. The first kind of scores reflected only amino acid frequencies [8–13], and some of these scores are related to Shannon's information theoretic entropy. These scores are a specialization of symbol frequency scores. The second kind of scores considered only the physicochemical properties of the amino acids in a column [14,15], or both the physicochemical properties and the amino acid frequencies [3,16]. In the latter case, to some extent 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. 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 [17–21].

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 [22]. In this paper, we present a new approach for measuring the evolutionary conservation at an aligned

* Corresponding author. Fax: +86 25 84895827.
E-mail address: xsliu@nuaa.edu.cn (X. Liu).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.10.157

position. We group all the columns of the underlying alignment into 20 sets, each containing the columns that are dominated by one of 20 kinds of amino acids. Based on the widely accepted BLOSUM62 substitution matrix, we estimate the probability that each kind of amino acid appears in a certain kind of columns by using the similarity of physicochemical property between amino acids. Then the conservation is measured by the logarithmic probability for this aligned position to take place. Our conservation measure can be statistically and biologically meaningful, by which one can compare quantitatively the conservation between any pair of columns. The suggested score has been used for determining the relationship between an amino acid's evolutionary conservation and its role in protein folding kinetics. By using our scoring method we found that folding nucleus residues are not significantly more conserved than the rest of the residues of the whole protein.

Models and methods

Grouping the universe of columns. We consider amino acids as symbols in an alphabet. The first widely accepted conservation score introduced by Wu and Kabat [8], which cares only about the frequency of the most commonly occurring symbol in a column, implicates that the most commonly occurring symbol is most important for scoring the evolutionary conservation of that position. Based on this sense we group all the columns into 20 sets, each of which contains the columns that are dominated by one of 20 kinds of amino acids. For convenience sake, if the residue D (aspartic acid) dominates in a column, we shall call this column the D-dominated column (it is reasonable that any symbol present in a column could be regarded as the dominated one if no symbol dominates in this 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 D, and the mutations are independent of each other (the evolutionary correlations between sequences are then considered 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.

How can we quantify this similarity? A possible method for measuring the similarity of physicochemical property between symbol D and each symbol type is to use the Venn diagram [14,15], or to use a hierarchical clustering of amino acids [23]. However, these methods have difficulties to describe the physicochemical properties quantificationally.

Because substitution matrices provide a quantitative and reasonably objective assessment of amino acid substitution and similarity, we use mutation data from a substitution matrix to measure the similarity of physicochemical property between, for example, residue D and each symbol type in a D-dominated column. Here, we use the widely accepted BLOSUM62 substitution matrix [24].

We first add a number to each line of the BLOSUM62 substitution matrix so that the minimum number in each line of revised matrix is 2 or 3 (3 only for the 5th and 18th line, since the ranges for values on line 5 and 18 of original matrix are relatively larger). Then the numbers in each line are divided by the maximum number in this line (the number at the diagonal of this matrix) and multiplied by 10, and then rounded to the nearest integer values. We thus obtain a similarity matrix **S** (Fig. 1). The normalization above ensures that $S(a,a) = 10$, and $2 \leqslant S(a,b) \leqslant 10$ for any symbols a and b. Each line of the similarity matrix represents *the similarity scores* between a certain residue and each of 20 kinds of amino acids. In addition, we set 1 as the similarity score for the gap heuristically, as done similarly by previous methods [18,19], and then it contributes 0 to the following conservation score.

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. Then, we define the conservation score for this column to be that

$$\pi = \sum_{i=1}^{20} n_i \log a_i. \tag{1}$$

Statistical meaning. Assume that N sequences exist in the underlying MSA (then $\sum_{j=1}^{20} n_j = N$). Let $p_i = a_i / \sum_{j=1}^{20} a_j$. It is easily seen that the probability for the ith amino acid type to appear in a D-dominated column is directly proportional to the similarity score a_i, that is to say, the higher the similarity of the ith amino acid type in this column, the larger the probability for the ith amino acid type to appear in this column. Therefore, the normalized similarity score of the ith amino acid type p_i just describes the probability that the ith amino acid type appears in this D-dominated column. We now consider the quantity that the conservation score for a D-dominated column minus a constant, that is, let

$$\theta = \pi - \sum_{i=1}^{20} n_i \log\left(\sum_{i=1}^{20} a_i\right) = \sum_{i=1}^{20} n_i \log a_i - \sum_{i=1}^{20} n_i \log\left(\sum_{i=1}^{20} a_i\right) = \sum_{i=1}^{20} n_i \log p_i. \tag{2}$$

Because n_i and p_i ($i = 1,\ldots,20$) are the frequency and probability for the ith amino acid type in this D-dominated column respectively, θ in (2) is just the logarithm of the probability for this column to take place. 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 $\pi_{\max} = N$ log10, when objects of only one type are present, and its values increase with increasing conservation.

Sequence weighting. Statistically, the process of selecting sequence examples may not be ideally random. In a typical alignment, there are often some sequences that are very closely related to each other. Intuitively, some of the information from these sequences is shared, so we should not give them each the same influence in the underlying issue as a single sequence that is more highly diverged from all the others. A large number of methods for sequence weighting have been suggested in the literature [25–27]. A widely applied formulation weights sequences at individual positions in an alignment and then combines position weights to give sequence weights [26]. 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 the present model, we use 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 ($i = 1,\ldots,20$) 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

	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

Fig. 1. The similarity matrix **S**, made from the BLOSUM62 substitution matrix. We first add a number to each line of the BLOSUM62 substitution matrix so that the minimum number in each line of revised matrix is 2 or 3 (3 only for the 5th and 18th line). Then the numbers in each line are divided by the maximum number in this line (the number at the diagonal of this matrix) and multiplied by 10, and then rounded to the nearest integer values. The normalization above ensures that $S(a,a) = 10$, and $2 \leqslant S(a,b) \leqslant 10$ for any 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.

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

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, $i = 1,\ldots,20$. For example, a column has 15 D residues and 5 S residues (i.e., $n_1 = 15$, $n_2 = 5$, and the other frequencies are zeros). Among 15 sequences whose residue in this column is D residue, 10 sequences each have a weight 0.7 and 5 sequences each have a weight 0.8, and 5 sequences whose residue in this column is S each have a weight 0.9. Then the net number of 15 D residues is $m_1 = 10\times 0.7 + 5\times 0.8 = 11$, and similarly $m_2 = 5\times 0.9 = 4.5$ (the others are zeros). The dominated symbol after sequence weighting is still the residue D. 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 log10, when objects of only one type are present.

Discussion

A testing example

Fig. 2 shows columns of amino acids taken from a hypothetical multiple-sequence alignment presented by Valdar [22]. For simplicity, assume that the weight of each sequence is equal to 1. Applying basic biochemical knowledge, the following order seems reasonable: (a) > (b) > (c) > (d) > (e) > (f), then (g) > (h) > (i), and (j) > (k), from most conserved to least conserved [22]. We think that the comparison between columns (j) and (k) in Valdar's review has little meaning in many applied situations because of the lack of data (with six sequences missing). So in this study we add one D residue and five gaps to the column (j). Here, columns (j) and (k) illustrate the effect of gaps. Column (k) should be more variable than column (j) (one more gap in column (k)), though D has least similarity with L residue. Fig. 2 could be used as a testing ground for a conservation score.

Comparison with previous scores

Sander and Schneider defined their score as a normalized Shannon's entropy:

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

where $p_i = n_i/N$, the fractional frequency of type i, and $K = 20$, representing the 20 amino acid types [12]. Mirny and Shakhnovich [3] gave the reduced entropy score

		Columns										
		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
Sequences	1	D	D	D	D	D	D	I	P	D	L	L
	2	D	D	D	D	D	D	I	P	V	L	L
	3	D	D	D	D	D	D	I	P	Y	L	L
	4	D	D	D	D	D	D	I	P	A	L	L
	5	D	D	D	D	D	D	L	W	T	D	—
	6	D	D	E	D	E	E	L	W	K	—	—
	7	D	D	E	D	E	E	L	W	P	—	—
	8	D	D	E	D	E	E	L	W	C	—	—
	9	D	D	E	D	E	F	V	S	R	—	—
	10	D	E	E	F	F	F	V	S	H	—	—

Fig. 2. Some example columns from a hypothetical multiple-sequence alignment used by Valdar [22]. Each labeled column represents a residue position in this multiple-sequence alignment. The rows denote the sequence numbers. Amino acids are identified by their one-letter code, and gaps by a dash ("—").

$$D = \sum_{i=1}^{K} p_i \ln p_i, \tag{5}$$

where $K = 6$. The set of K partitions is: aliphatic [AVLIMC], aromatic [FWYH], polar [STNQ], positive [KR], negative [DE], and special conformations [GP]. Indeed, there is an improvement over pure entropy-based scores. However, in latter case 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 property.

It has been illustrated that the scores that considered only one aspect of the amino acid frequencies and the physicochemical properties of amino acids cannot make sense well in biochemistry. Though the SP scores seem better, they do not make sense in what the statistic means [22,27]. It is implausible that the diversity in a column arises from all the pairwise amino acid substitutions. Additionally, by calculating conservation scores for many real alignments Pei and Grishin showed that the usage of the entropy-based conservation measures is not inferior to that of the SP measure [28].

The aim of this article is to find a way that considers both the physicochemical properties and frequencies of amino acids. We consider the physicochemical properties of the amino acids in a column by presenting their similarity scores, which imply the probabilities that the amino acids appear in this column. Exactly, by replacing the fractional frequencies p_i of the residues by their similarity scores at the logarithmic positions in a pure entropy score, e.g., in Eq. (4), consider the amino acid frequencies by replacing the fractional frequencies p_i of the residues by their frequencies n_i at the coefficient positions. Unlike the SP scores, our conservation score is statistically meaningful, which is the log probability for the underlying column to take place. Thus our score may achieve both biological and statistical rigor. Also, our score considers the gaps and the sequence weighting, and fulfills the principle of simplicity. For the above simple testing example, the score scheme in the present study can correctly reproduce the ranks (a) > (b) > (c) > (e) > (f) (the scores are 23.03, 22.67, 21.24, 20.40, and 15.79, respectively), since it has well taken account of the amino acid frequencies in a column; the score scheme can recognize conservative replacements and that some substitutions incur more chemical and physical change than others, so it correctly reproduces the ranks (g) > (h) > (i) (the scores are 21.92, 14.76, and 13.34, respectively); the present score scheme can rank column (j) as more conserved than column (k) illustrating the effect of gaps. With biochemical intuition, one cannot give a conservation order for all columns in the above simple testing example. However, our score scheme produces the following conservation order for these columns: (a) > (b) > (g) > (d) > (c) > (e) > (f) > (h) > (i) > (j) > (k).

Recently, some methods evaluated site-specific evolutionary rates in proteins by taking into account the phylogenetic trees [29–33]. Some computations show that the results from our model are not perceptibly different by making use of the phylogenetic trees. Thus in our models we consider the evolutionary correlations between sequences and sequence redundancy by sequence weighting in the final formula (3).

Applications to the protein folding nucleus

It is now 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. One of the most intriguing aspects of the nucleation–condensation mechanism of protein folding is its relation to protein evolution. Several studies correlating experimental measurement of residue participation in folding nucleus and sequence conservation have reached different conclusions. The reported

discrepancy was partially attributed to the use of the entropy-based score [1–4]. Here, we report analysis of conservation of folding nucleus using our scoring method as an alternative to entropy-based procedures.

The studies in question use directly the highest quality sequence alignments available from Supplementary Material of the study by Larson and co-authors [1]. We use the same Φ-values for these proteins which were collected by Plaxco and co-authors from literatures and provided by Plaxco (personal communication). 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 [34–37].

As done in the study by Larson and co-authors [1], we discuss correlations between Φ-value and sequence conservation. For 14 protein examples, the results here do not follow the assumption that residues participating more strongly in the nucleus will be relatively better conserved (one example in Fig. 3, the other 13 examples are presented in Supplementary). We have also 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. 4). These results suggest that, by using the scoring method suggested in this paper as an alternative to entropy-based procedures, there is no indication that folding nucleus residues are significantly more conserved than other residues. The reported discrepancy about the conservation of the folding nucleus between the previous studies is not attributed to the use of the entropy-based scores. The disagreement may be due to the difference in other aspects, such as, the choice and processing of the data set, the quality of sequence alignments, the definition of folding nucleus, and so on [1,38]. Recently, by using a codon evolutionary model Tseng and Liang concluded that at the level of codon substitution, there is no indication that folding nucleus residues are significantly more conserved than the rest of the residues of the whole protein [36].

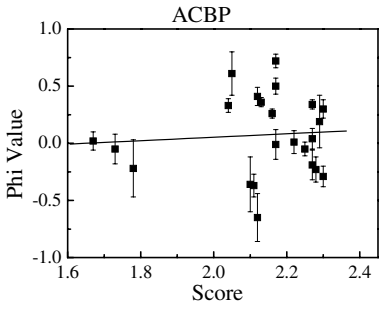

Fig. 3. Correlations between Φ-values and conservation scores calculated by our biological-relevance weighted statistical scoring function for the protein ACBP (the data for Φ-value analysis are provided by Plaxco and the alignments are available from Supplementary Material of the study by Larson and co-authors). We have that $r^2 = 0.0061$ (the other 13 proteins are presented in Supplementary).

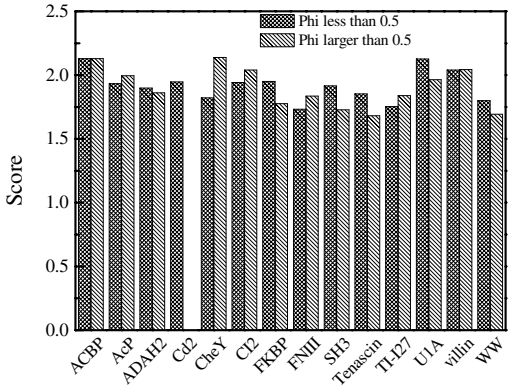

Fig. 4. The mean biological-relevance weighted statistical scores of high Φ residues ($\Phi > 0.5$) and low Φ residues for the 14 proteins. Little evidence is observed in favor of preferential conservation of the folding nucleus (as defined by $\Phi > 0.5$): for 5 proteins these residues are more conserved than low Φ residues, for 9 they are less well conserved.

Conclusion

In this study, we present a more elaborate conservation measure which could be used in the situations in which the high precision in calculations is needed, such as, for characterizing the functional sites of a protein. More importantly, this new measure considers both physicochemical properties and frequencies of amino acids, and then it can be statistically meaningful and make sense biochemically. The suggested scoring method has been used for determining the relationship between an amino acid's evolutionary conservation and its role in protein folding kinetics. Our results suggest that, there is no significant evidence in favor of the preferential conservation of the Φ-value defined folding nucleus, and the reported discrepancy about the conservation of the folding nucleus between the previous studies is not attributed to the use of entropy-based scores. These results are also consistent with those derived from a codon evolutionary model [36], which present a substantial support for our scoring method.

Acknowledgments

Our thanks to K.W. Plaxco for providing the data for Φ-value analysis, to W.S.J. Valdar for a testing multiple-sequence alignment, and to them for helpful suggestions. This work is supported by the National Natural Science Foundation of China (Nos. 10372044, 50275073; Nos. 90403120, 10021001, and 10474041). Jiangsu Province NSF, the Innovation Team Programme and the Cultivation Fund of the Key Scientific and Technical Innovation Project of the Ministry of Education of China (No. 705021).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.10.157.

References

[1] S.M. Larson, I. Ruczinski, A.R. Davidson, D. Baker, K.W. Plaxco, Residue participating in the protein folding nucleus do not exhibit preferential evolutionary conservation, J. Mol. Biol. 316 (2002) 225–233.

[2] K.W. Plaxco, S. Larson, I. Ruczinski, D.S. Riddle, E.C. Thayer, B. Buchwitz, A.R. Davidson, D. Baker, Evolutionary conservation in protein folding kinetics, J. Mol. Biol. 298 (2000) 303–312.

[3] L.A. Mirny, E.I. Shakhnovich, Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function, J. Mol. Biol. 291 (1999) 177–196.

[4] L. Mirny, E. Shakhnovich, Evolutionary conservation of the folding nucleus, J. Mol. Biol. 308 (2001) 123–129.

[5] O.B. Ptitsyn, Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? J. Mol. Biol. 278 (1998) 655–666.

[6] W.S.J. Valdar, J.M. Thornton, Protein–protein interfaces: analysis of amino acid conservation in homodimers, Proteins 42 (2001) 108–124.

[7] W.S.J. Valdar, J.M. Thomton, Conservation helps to identify biologically relevant crystal contacts, J. Mol. Biol. 313 (2001) 399–416.

[8] T.T. Wu, E.A. Kabat, 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 (1970) 211–249.

[9] R. Jores, P.M. Alzari, T. Meo, 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 (1990) 9138–9142.

[10] S.W. Lockless, R. Ranganathan, Evolutionarily conserved pathways of energetic connectivity in protein families, Science 286 (1999) 295–299.

[11] P.S. Shenkin, B. Erman, L.D. Mastrandrea, Information-theoretical entropy as a measure of sequence variability, Proteins 11 (1991) 297–313.

[12] C. Sander, R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment, Proteins 9 (1991) 56–68.

[13] M. Gerstein, R.B. Altman, Average core structures and variability measures for protein families: application to the immunoglobulins, J. Mol. Biol. 251 (1995) 161–175.

[14] W.R. Taylor, The classification of amino acid conservation, J. Theor. Biol. 119 (1986) 205–218.

[15] M.J. Zvelibil, G.J. Barton, W.R. Taylor, M.J. Sternberg, Prediction of protein secondary structure and active sites using the alignment of homologous sequences, J. Mol. Biol. 195 (1987) 957–961.

[16] R.M. Williamson, Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters, J. Theor. Biol. 174 (1995) 179–188.

[17] S. Karlin, L. Brocchieri, Evolutionary conservation of RecA genes in relation to protein structure and function, J. Bacteriol. 178 (1996) 1881–1894.

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

[19] J.D. Thompson, T.J. Gibson, F. Plewniak, F. Jeanmougin, D.G. Higgins, The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876–4882.

[20] Y. Pilpel, D. Lancet, The variable and conserved interfaces of modeled olfactory receptor proteins, Protein Sci. 8 (1999) 969–977.

[21] R. Landgraf, D. Fischer, D. Eisenberg, Analysis of heregulin symmetry by weighted evolutionary tracing, Protein Eng. 12 (1999) 943–951.

[22] W.S.J. Valdar, Scoring residue conservation, Proteins 48 (2002) 227–241.

[23] R.F. Smith, T.F. Smith, Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling, Protein Eng. 5 (1992) 35–41.

[24] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA 89 (1992) 10915–10919.

[25] M. Vingron, P. Argos, A fast and sensitive multiple sequence alignment algorithm, Comput. Appl. Biosci. 5 (1989) 115–121.

[26] S. Henikoff, J.G. Henikoff, Position-based sequence weights, J. Mol. Biol. 243 (1994) 574–578.

[27] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.

[28] J.M. Pei, N.V. Grishin, AL2CO: calculation of positional conservation in a protein sequence alignment, Bioinformatics 17 (2001) 700–712.

[29] T. Pupko, R.E. Bell, I. Mayrose, F. Glaser, N. Ben-Tal, Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues, Bioinformatics 18 (2002) 71–77.

[30] I. Mayrose, D. Graur, N. Ben-Tal, T. Pupko, Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior, Mol. Biol. Evol. 21 (2004) 1781–1791.

[31] I. Mayrose, A. Mitchell, T. Pupko, Site-specific evolutionary rate inference: taking phylogenetic uncertainty into account, J. Mol. Evol. 60 (2005) 345–353.

[32] Y. Suzuki, G.V. Glazko, M. Nei, Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics, Proc. Natl. Acad. Sci. USA 99 (2002) 16138–16143.

[33] O.S. Soyer, R.A. Goldstein, Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters, J. Mol. Biol. 339 (2004) 227–242.

[34] E. Shaknovich, V. Abkevich, O. Ptitsyn, Conserved residues and the mechanism of protein folding, Nature 379 (1996) 96–98.

[35] L.A. Mirny, V.I. Abkevich, E.I. Shakhnovich, How evolution makes proteins fold quickly, Proc. Natl. Acad. Sci. USA 95 (1998) 4976–4981.

[36] Y.Y. Tseng, J. Liang, Are residues in a protein folding nucleus evolutionarily conserved? J. Mol. Biol. 335 (2004) 869–880.

[37] A.R. Fersht, Nucleation mechanisms in protein folding, Curr. Opin. Struct. Biol. 7 (1997) 3–9.

[38] I.A. Hubner, J. Shimada, E.I. Shakhnovich, Commitment and nucleation in the protein G transition state, J. Mol. Biol. 336 (2004) 745–761.