Hydropathy Conformational Letter and its Substitution Matrix
HP-CLESUM: an Application to Protein Structural Alignment

Sheng Wang1,

1Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China

Motivation: Protein sequence world is discrete as 20 amino acids (AA) while its structure world is continuous, though can be discretized into structural alphabets (SA). In order to reveal the relationship between sequence and structure, it is interesting to consider both AA and SA in a joint space. However, such space has too many parameters, so the reduction of AA is necessary to bring down the parameter numbers.

Result: We've developed a simple but effective approach called entropic clustering based on selecting the best mutual information between a given reduction of AAs and SAs. The optimized reduction of AA into two groups leads to hydrophobic and hydrophilic. Combined with our SA, namely conformational letter (CL) of 17 alphabets, we get a joint alphabet called hydropathy conformational letter (hp-CL). A joint substitution matrix with \((17 \times 2)^2\) indices is derived from FSSP. Moreover, we check the three coding systems, say AA, CL and hp-CL against a large database consisting proteins from family to fold, with their performance on the TopK accuracy of both similar fragment pair (SFP) and the neighbor of aligned fragment pair (AFP).
The TopK selection is according to the score calculated by the coding system's substitution matrix. Finally, embedding hp-CL in a pairwise alignment algorithm, say CLeFAPS, to replace the original CL, will get an improvement on the HOMSTRAD benchmark.

Contact*: wangsheng@itp.ac.cn

I. INTRODUCTION

Proteins fold into specific spatial conformations to perform their biological functions [1] and there are abundant evidences to show their amino acid (AA) sequences determining the structures [2]. The attempt to find the relationship between structure and sequence is a fundamental task in computational biology [3].

Compared to the sequence world which is discrete of 20 AAs, the structure world is continuous, though the local conformational space of a protein backbone fragment is rather limited [4]. The idea of representing the backbone with a string of discrete letters was first observed by Corey and Pauling [6, 7] and later refined into the concept of protein secondary structure elements (SSEs). However, segments of a single SSE may vary significantly in their 3D structures, especially for the state coil, which is not a true secondary structure but is a class of conformations that indicate the absence of regular SSEs, say alpha helix or beta strand [23]. Although the SSE can be predicted with high accuracy (\(\geq 80\%\)) [3], the description of a protein in terms of its SSEs is not sufficient to capture accurately its 3D geometry [40].

To overcome this limitation, several groups have proposed the idea that representing protein structures as a series of overlapping fragments, each labeled with a symbol, which defines a structural alphabet (SA) for proteins [13]. Such alphabet can be used to predict local structure [12, 17], to reconstruct the full-atom representation [18], to identify the structural motifs [19], to classify protein structures [20] and to search against a database [21, 22]. We've proposed our SA, namely conformational letter (CL) [14], which is composed of 17 alphabets and each with 4 residues in length. Our SA is focused on the fast pairwise [24], multiple [25] and flexible [26] structure alignment problems, combined with its substitution matrix CLESUM [14].

After we discretized the continuous structure world into SAs, it is the time to consider both AA and SA in a joint space. However, such space is too large for about \((20 \times 17)^2\) parameters when using the current popular SAs. It is necessary to employ the reduction of AAs [21]. Several groups have put forward their reduced AAs either experimentally or computationally. For example, Baker et.al found a five-letter alphabet for 38 out of 40 selected sites of SH3 chain [28]; Wang & Wang [29] introduced the minimal mismatch principle to reduce the alphabet based on Miyazawa-Jernigan's residue-residue statistical potential [50]. Murphy et.al [31] approached the same problem using the BLOSUM matrix [32]. Recently, de Brevem et.al proposed to use their SA, namely Protein Blocks [10] to analyze equivalences between the different kinds of amino acids, then obtained their reduced AAs [33].

Here we present a novel reduction method, called entropic clustering [34]. Briefly, given two discrete distributions A and B, merging \(a_i\) and \(a_j\) into one group \(a_{i,j}\) will result in a loss of mutual information of A and B. Thus, mutual information I can be naturally chosen as the objective function for optimized clustering. When grouping the 20 AAs into two categories, we've got a result of hydrophobic and hydrophilic, which agrees with HP-model [35] exactly. Then we construct a joint substitution matrix HP-CLESUM with \((17 \times 2)^2\) indices by the similar means as constructing CLESUM. The following tests are employed to check and compare different coding systems, namely AA, CL and hp-CL with their corresponding substitution matrix, say BLOSUM, CLESUM and HP-CLESUM. We first compare the TopK accuracy of SFPs (similar fragment pairs) and the neighbor of AFP (aligned fragment pairs) against a large dataset encompassing the protein homologues levels from family to fold according to SCOP [18]; then we embed hp-CL into CLeFAPS [20], replacing the original CL, to get an improvement against the popular benchmark HOMSTRAD [42].
2

\[ i' = \arg \max \mathcal{P}(C_i | x_i), \mbox{ where } x_i = (\theta, \theta', \tau) \]
\[ \mathcal{P}(C_i | x_i) \propto \mathcal{E}^{1/2} \exp \left[ -\frac{1}{2} (x_i - \mu_i) \cdot \Sigma^{-1} (x_i - \mu_i) \right] \]

II. MATERIALS AND METHOD

A. Datasets

We use PDB-SELECT database [36] to construct our CLs, and use FSSP database [38] to derive the substitution matrix CLESUM. Particularly, PDB-SELECT contains 1544 non-membrane proteins from PDB [37] with amino acid identity less than 25%. FSSP is based on exhaustive all-against-all structure comparison of the representative protein structures, where the representative set contains no pair which has more than 25% sequence identity. A tree for the fold classification of the 2,860 representative set is constructed by a hierarchical clustering method based on the structural similarities. Family indices of the FSSP are obtained by cutting the tree at levels of 2, 4, 8, 16, 32 and 64 standard deviations above the database average.

B. Conformational letter and its substitution matrix

Four contiguous \( C_\alpha \) atoms, say \( i-2, i-1, i \) and \( i+1 \), determine two bending angles \( \theta, \theta' \) and a torsion angle \( \tau \) which is the dihedral angle between the two planes of triangles \( i-2, i-1, i \) and \( i-1, i, i+1 \) (see Fig. 1). By using a mixture model for the density distribution of the three angles, the local structural states from PDB-SELECT have been clustered as 17 discrete states (see our previous work [14] for more details). To use our SAs directly for the structural comparison, a score matrix similar as BLOSUM [32] for AAs is desired. In details, we first convert the structures of the representative set from FSSP to their CL strings; then collect all the pair alignments with the same first three family indices (DALI Z-Score \( \geq 8 \)) (see Fig. 2); finally count all ungapped aligned pairs of CLs to generate the substitution matrix, say CLESUM (Table I). The total number of structures is 10,047 pairs, consisting of 175,723 fragment pairs and 1,284,750 code pairs.

C. Entropic clustering

From FSSP, which contains also the AA information, it is possible to construct a substitution matrix in the joint space of the structure and sequence. However, such matrix would have about \((17 \times 20) \times (17 \times 20) \) parameters (Fig. 3(A)). If we group the 20 AAs into two clusters, then the parameters of the matrix are reduced to \((17 \times 2) \times (17 \times 2) \).

Generally, the mutual information \( I \) of two discrete
distribution $(A \times B)$ is defined as,

$$I = \sum_{a,b} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}. \quad (1)$$

If we cluster $a_i$ and $a_j$ into $a_{i\&j}$ leads to

$$p(a_{i\&j}) = p(a_i) + p(a_j), \quad p(a_{i\&j}, b) = p(a_i, b) + p(a_j, b). \quad (2)$$

The difference between values of $I$ after and before clustering is given by

$$I = \left[ p(a_i, b) + p(a_j, b) \right] \log \frac{p(a_i, b) + p(a_j, b)}{p(a_i) p(b)} - p(a_i, b) \log \frac{p(a_i, b)}{p(a_i) p(b)} - p(a_j, b) \log \frac{p(a_j, b)}{p(a_j) p(b)}. \quad (3)$$
which, by introducing
\[
x_i = \frac{p(a_i,b)}{p(a_i) + p(a_j)},
\]
\[
\omega_i = \frac{p(a_i)}{p(a_i) + p(a_j)},
\]
(4)
and their analogs \(x_j\) and \(\omega_j\), then defining \(\langle F(x) \rangle = \omega_i F(x_i) + \omega_j F(x_j)\) and \(\langle F(x) \rangle = F(\omega_i x_i + \omega_j x_j)\) where \(\omega_i + \omega_j = 1\). We may now see that Eq. (3) is proportional to \(f(x) - \langle f(x) \rangle\) with \(f(x) = x \log x\). From the Jensen’s inequality, for the convex function \(x \log x\) here we have \(f(x) \leq \langle f(x) \rangle\), so \(I\) never increases after any step of clustering.

That is to say, merging any two members into one cluster will result in a loss of mutual information. To make the loss of mutual information as small as possible, \(I\) should be maximized, so it can be naturally chosen as the objective function during clustering. We call this approach entropic clustering \([24]\). If we partition \(n\) objects into \(m_1\) and \(m_2\) classes, where \(m_2 > m_1\), it is easy to prove that the maximal \(I\) at \(m_2\) is always greater than the maximal \(I\) at \(m_1\) \([25]\).

Now turning back to our substitution matrix in the joint space, we may define the average mutual information as follows like BLOSUM,
\[
I = \sum_{X,Y} P_{XY} \log \frac{P_{XY}}{P_X P_Y},
\]
(5)
where \(X\) and \(Y\) means a joint state of CL and AA (either reduced or not). Given a clustering group we may calculate its \(I\) based on Eq. (5) and according to entropic clustering we should get a categories which maximize \(I\) (Fig. 3F)).

### III. RESULT

#### A. Joint substitution matrix of conformational letters and reduced amino acids

For clustering the 20 AAs into two categories, the Monte Carlo finds AVCFWLMY and DEHKHPQRT as the groups, which is just the hydrophilic and hydrophobic cluster \([35]\). Such enlarged CLs are called hp-CLs.

### TABLE II: CLESUM-hh (lower left) and CLESUM-pp (upper right) (in units of 0.05 bit).

| J  | H | I  | K  | L  | M  | P  | Q  | R  | S  | T  | U  | V  | W  | X  | Y  | Z  |
|----|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| J  | 45| 33 | 24 | 19 | 3 | 41 | 27 | 13 | 21 | 40 | 32 | 21 | 22 | 37 | 31 | 62 |
| H  | 33| 30 | 24 | 14 | 6 | 19 | 17 | 40 | 32 | 17 | 20 | 11 | 16 | 12 | 11 | 20 |
| I  | 24| 20 | 14 | 11 | 5 | 10 | 5 | 17 | 30 | 23 | 12 | 8 | 15 | 12 | 13 | 22 |
| K  | 19| 16 | 11 | 8 | 4 | 6 | 8 | 6 | 10 | 13 | 12 | 9 | 11 | 10 | 11 | 14 |
| L  | 20| 17 | 12 | 9 | 5 | 7 | 9 | 7 | 10 | 13 | 12 | 9 | 11 | 10 | 11 | 14 |
| M  | 31| 28 | 22 | 18 | 13 | 14 | 16 | 14 | 17 | 20 | 18 | 14 | 16 | 14 | 17 | 20 |
| P  | 32| 30 | 25 | 21 | 17 | 19 | 21 | 20 | 24 | 27 | 25 | 22 | 24 | 22 | 25 | 27 |

### TABLE III: CLESUM-hp (row-column) (in units of 0.05 bit).

| J  | H | I  | K  | L  | M  | P  | Q  | R  | S  | T  | U  | V  | W  | X  | Y  | Z  |
|----|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| J  | 45| 33 | 24 | 19 | 3 | 41 | 27 | 13 | 21 | 40 | 32 | 21 | 22 | 37 | 31 | 62 |
| H  | 33| 30 | 24 | 14 | 6 | 19 | 17 | 40 | 32 | 17 | 20 | 11 | 16 | 12 | 11 | 20 |
| I  | 24| 20 | 14 | 11 | 5 | 10 | 5 | 17 | 30 | 23 | 12 | 8 | 15 | 12 | 13 | 22 |
| K  | 19| 16 | 11 | 8 | 4 | 6 | 8 | 6 | 10 | 13 | 12 | 9 | 11 | 10 | 11 | 14 |
| L  | 20| 17 | 12 | 9 | 5 | 7 | 9 | 7 | 10 | 13 | 12 | 9 | 11 | 10 | 11 | 14 |
| M  | 31| 28 | 22 | 18 | 13 | 14 | 16 | 14 | 17 | 20 | 18 | 14 | 16 | 14 | 17 | 20 |
| P  | 32| 30 | 25 | 21 | 17 | 19 | 21 | 20 | 24 | 27 | 25 | 22 | 24 | 22 | 25 | 27 |

The substitution matrix of hp-CLs is called HP-CLESUM, this symmetry matrix can be divided into three sub-matrices: CLESUM-hh, CLESUM-pp, and CLESUM-hp. The first two, shown in Table III correspond to the same hydrophobicity aligned amino acid types (i.e., h-h and p-p). The third, shown in Table III corresponds to the different hydrophathy types h-p. As expected, compared with the original CLESUM (Table II),
elements of CLESUM-hh and CLESUM-pp generally become larger in absolute values, and those of CLESUM-hp show the opposite tendency. The tendency is stronger for letters dominated by helices or sheets.

B. Comparison between different coding systems

1. Overview

FIG. 4: TopK accuracy check procedure. (A) Select a pair of structures and aligned by MATT, resulting a series of AFPs. (B) Encode the pairwise structures, the coding system may be AA or SA. (C) Search for any SFPs with length L according to a certain coding system. (D) Sort these SFPs in descending order, the score is calculated by the corresponding substitution matrix. (E) Check whether there exists a correct SFP within TopK. (F) Search for any possible neighbor of a given AFP whose length is over L against the other string. (G) Check these AFP’s neighbors. (H) Check their accuracy.

A coding system in protein structure is defined as an alphabet combined with its corresponding substitution matrix. Amino acids (AA) or all kinds of SAs can be treated as coding systems, so long as the alphabet has its substitution matrix. We’ll compare the performance of the following three ones, namely AA, CL and hip-CL, based on their TopK accuracy against a benchmark (Fig. 5). The difference between SFP and AFP is, we use SFPs as a subset of SFPs that each of them should be in the following three ones, namely AA, CL and hp-CL, its substitution matrix. We’ll compare the performance of any SFPs with length L, the accuracy of all coding systems grows better, while from family to fold level, the accuracy declines. It is surprising that at fold level, the accuracy of hip-CL outperforms AA more than 50% at the TopK SFP accuracy test and more than 100% at the TopK neighbor of AFP test, while in the latter, hp-CL got the average accuracy at about 71% given L = 18 from the Top-1 highest neighbor of an AFP. Such feature may be applied to construct the Highest Similarity Fragment Block (HSFB) during the multiple structure alignment. Given a seed structure and a certain position, if this position got many high score neighbors in other structures, we may say that this block (consisting the seed position and its neighbors) has a more probable chance in the final multiple alignment.

Moreover, we’ve shown the effectiveness of parameter self-adaptive strategy to create the SFP-list in [26]. At most cases self-adaptive strategy is compatible with fixed parameters, while the size of the SFP-list can be controlled empirically to about O(n^2/LEN_H/6) with the LEN_H=9 (Fig. 5). However, its hard to control the balance between the size of the SFP-list and the threshold of SFP generated with fixed parameters strategy. Actually, the data of fixed length used in Table IV is considered all O(n^2) SFPs, then to select TopK; we’ve tested different SFP thresholds, if it is set too high there’ll lead to blank or few SFP-list while if it is set too loose then the SFP-list will be too much (data not shown).

Finally, we may get the conclusion that, during structure comparison, the only consideration of the TopK highest SFPs to built the initial alignment is feasible from family level to fold, so long as the coding system is specific enough. Also the employment of parameter self-adaptive strategy to generate SFPs is effective and economic.
TABLE IV: TopK accuracy check with different strategies, from TopK SFP check (left part) to TopK AFP’s neighbor check (right part); different coding systems, from AA (A%), CL (C%) to hp-CL (H%); different homologous level, from family (Fam), superfamily (Sup) to fold; and different SFP length $L$, from 6, 12 to 18.

| Level  | Fam | Self(9-18)* |
|--------|-----|-------------|
| $L = 6$ | 53.1 | 74.3 | 71.1 |
| $L = 12$ | 74.4 | 94.3 | 75.5 |
| $L = 18$ | 85.7 | 96.6 | 95.2 |

| Level  | Sup | Self(9-18)* |
|--------|-----|-------------|
| $L = 6$ | 39.1 | 58.2 | 56.2 |
| $L = 12$ | 58.2 | 79.2 | 54.9 |
| $L = 18$ | 71.8 | 85.4 | 88.5 |

| Level  | Fold | Self(9-18)* |
|--------|------|-------------|
| $L = 6$ | 13.9 | 17.0 | 31.9 |
| $L = 12$ | 28.8 | 34.9 | 31.9 |
| $L = 18$ | 39.5 | 50.1 | 87.0 |

*: Self(9-18) means the application of self-adaptive strategy of the SFP’s length from 9 to 18.

C. Implement of hydropathy conformational letters to structural alignment

We embed hp-CL to the pairwise protein structural alignment problem under the framework of CLeFAPS [26]. Particularly, we first transform each structure to its hp-CL strings; then search for both highly specific SFPs (SFP_H) that have a high HP-CLESUM score to build an initial alignment from the best TM-score [39] SFP within TopK, and highly sensitive SFPs (SFP_L) that have a low HP-CLESUM score (must above 0) to refine the alignment through fuzzy-add strategy. These two SFP-lists can be generated simultaneously [24]; finally we apply an elongation based on Vect-score to collect local flexible fragments.

HOMSTRAD is a database of protein structural alignments for homologous families [42]. Its alignments were generated using structural alignment programs, then followed by a manual scrutiny of individual cases. There are totally 1033 families (633 at pairwise level). We’ll show the improvement based on hp-CL as the coding systems instead of CL under the same algorithm, say CLeFAPS, in Table V.

TABLE V: Alignment accuracy metric on HOMSTRAD

| Metric        | CLeFAPS(CL) | CLeFAPS(hp-CL) | MATT |
|---------------|-------------|----------------|------|
| C/LOA 1       | 0.929       | 0.939          | 0.948 |
| C/LOR 2       | 0.898       | 0.907          | 0.831 |

1: Correct/(Length of the algorithm).
2: Correct/(Length of the reference).

IV. DISCUSSION AND FUTURE WORK

To explore the joint space of both AAs and SAs, entropic clustering is a simple but effective approach. In this work, only the reduction of AAs is considered, while we may also reduce CLs and AAs simultaneously while balancing the accuracy and the parameter numbers. For example, if reducing the CL to 9 letters, (actually, from Fig. 1 there are 4 codes for helix which can be grouped to one cluster, the same as sheet.) we may then consider up to four AA cluster instead of two, while the total alphabet number is about the same as hp-CL.

It is interesting that, hp-CL can be applied during the situation we know only a little information about the AA sequence of the structure, i.e., the hydropathy features, or even none. That is true because, from the knowledge of protein design [40, 41], the hydropathy patterns from a 3D structure may probably be deduced. Then the usage of hp-CLs and HP-CLESUM that consider the hydropathy features during the alignment process may bring advantages.
thetic patterns will get a more accurate result than CLs and CLESUM that only consider the 3D structure.

We also verify a basic idea in CLeFAPS, i.e., self-adaptive strategy to generate SFPs, that we needn’t consider the parameters to deal with different purposes and different proteins. The result showed its accuracy is maintained well and the SFP-list size is controlled in O(n^2/LEN_H/6) while its hard to judge the balance between accuracy and size with fixed parameters.

TopK accuracy check has demonstrated the basic strategy of both CLePAPS and CLeFAPS efficient, which only considers TopK highest SFPs to build the initial alignment. Moreover, TopK accuracy check is an effective approach to measure the coding systems against a reference dataset, especially to judge the substitution matrix. If a coding system is good enough, it should rank those SFPs with highly specificities top enough among other SFPs. In a future work, we’ll use this approach to test the current available SAs based on their performance for finding specific SFPs. Also we can do the comparison between SAs and RMSD values or some p-values derived from RMSD. Such comparison between 1D coding systems with 3D expression will show the effectiveness of SAs because they contain the statistic information from the database [24].

Acknowledgments

This work is supported by ...

References

[1] Chothia, C. (1984). Principles that determine the structure of proteins. Annu. Rev. Biochem. 53:537-572.
[2] Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. Science 294:93-96.
[3] Jones, D. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292:195-202.
[4] Eishammer, I., Jonassen, I. & Taylor, W.R. (2000). Structure comparison and structure patterns. J. Comput. Biol. 7:685-716.
[5] Offmann, B., Tvigj, M. & de Brevern, A.G. (2007). Local protein structures. Current Bioinformatics 2:165-202.
[6] Pauling, L., Corey, R. & Branson, H. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl Acad. Sci. USA. 37:205-234.
[7] Pauling, L. & Corey, R. (1951). Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc. Natl Acad. Sci. USA. 37:729-740.
[8] Rooman, M., Rodriguez, J. & Wodak, S. (1990). Automatic definition of recurrent local structure motifs in proteins. J. Mol. Biol. 213:327-336.
[9] Fetrow, J., Palumbo, M. & Berg, C. (1997). Patterns, and amino acid frequency in structural building blocks, a protein secondary structure classification scheme. Proteins 27:249-271.
[10] de Brevern, A., Etchebest, C. & Hazout, S. (2001). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins 41:271-287.
[11] Kolednya, R., Koehi, P., Guibas, L. & Levitt, M. (2002). Small model derived structural alphabet and BLOSUM like substitution matrix derived structural alphabet for proteins. J. Mol. Biol. 319:591-605.
[12] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. 28:235-242.
[13] Laskowski, R.A., Moss, D.S., Thornton, J.M. & Panchenko, A.R. (2003). PDBsum: a comprehensive analysis server for PDB. Nucleic Acids Res. 31:363-366.
[14] Zheng, W.M. & Wang, S. (2009). CLeFAPS: fast pair alignment of protein structures based on conformational letters. J. Bioinform. Comput. Biol. 7:367-374.
[15] Liu, X., Zhao, Y.P. & Zheng, W.M. (2009). CLeMAPS: multiple alignment of protein structures based on conformational letters. Proteins. In press.
[16] Wang, S. (2009). CLeFAPS: fast flexible alignment of protein structures based on conformational letters.
[17] Chan, H.S. (1999). Folding alphabets. Nat. Struct. Biol. 6:994-996.
[18] Mulvey, R.E., Drosos, A., Levy, R. & Brown, C.J. (1992). Amino acid substitution matrices for protein blocks. Proc. Natl Acad. Sci. USA. 89:10915-10919.
[19] Wang, T. & Wang, W. (1999). A computational approach to simplifying the protein folding alphabet. Nat. Struct. Biol. 6:1033-1038.
[20] Miyazawa, S. & Jernigan, R.L. (1993). A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng. 6:267-276.
[21] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G. & Levy, R.M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng. 13:149-152.
[22] Yang, J.M. & Tung, C.H. (2006). Protein structure database search and evolutionary classification. Nucleic Acids Res. 34:3646-3659.
[23] Wang, S. (2009). CLeFAPS: fast flexible alignment of protein structures based on conformational letters. J. Mol. Biol. 388:1059-1069.
[24] Zheng, W.M. & Wang, S. (2009). CLePAPS: fast alignment of protein structures based on conformational letters. J. Bioinform. Comput. Biol. 7:685-716.
[25] Liu, X., Zhao, Y.P. & Zheng, W.M. (2009). CLeMAPS: multiple alignment of protein structures based on conformational letters. Proteins 71:728-736.
[26] Wang, S. (2009). CLeFAPS: fast flexible alignment of protein structures based on conformational letters.
[27] Chan, H.S. (1999). Folding alphabets. Nat. Struct. Biol. 6:994-996.
[28] Mulvey, R.E., Drosos, A., Levy, R. & Brown, C.J. (1992). Amino acid substitution matrices for protein blocks. Proc. Natl Acad. Sci. USA. 89:10915-10919.
[39] Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins 57:702-710.

[40] Hecht, M.H., Das, A., Go, A., Bradley, L.H. & Wei, Y. (2004). De novo proteins from designed combinatorial libraries. Protein Sci. 13:1711-1723.

[41] Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. & Hecht, M.H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. Science 262:1680-1685.

[42] Mizuguchi, K., Deane, C., Blundell, T.L. & Overington, J. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci. 11:2469-2471.

[43] VanWalle, I., Lasters, I. & Wyns, L. (2005). SABmark: a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics 21:1267-1268.

[44] Kolodny, R., Koehl, P. & Levitt, M. (2005). Comprehensive evaluation of protein structure alignment methods. Scoring by geometric measures. J. Mol. Biol. 346:1173-1188.

[45] Menke, M., Berger, B. & Cowen, L. (2008). Matt: local flexibility aids protein multiple structure alignment. PLoS Comput. Biol. 4(1): e10.

[46] Murzin, A.G., Brenner, S.E., 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.

[47] Chandonia, J.M., Hon, G., Walker, N.S., Conte, L.L., Koehl, P., Levitt, M. & Brenner, S.E. (2004). The ASTRAL compendium in 2004. Nucleic Acids Res. 32:D189-D192.

[48] Doolittle, R.F. (1981). Similar amino acid sequences: chance or common ancestry. Science 214:149-159.