ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/5841778

Context-specific amino acid substitution matrices and their use in the detection of protein homologs

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · MAY 2008

Impact Factor: 2.63 · DOI: 10.1002/prot.21775 · Source: PubMed

CITATIONS
14

READS
26

2 AUTHORS:

Nalin C. W. Goonesekere
University of Northern Iowa
15 PUBLICATIONS 705 CITATIONS

SEE PROFILE

BK Lee
National Institutes of Health
147 PUBLICATIONS 9,835 CITATIONS

SEE PROFILE

Available from: BK Lee
Retrieved on: 11 January 2016

Context-specific amino acid substitution matrices and their use in the detection of protein homologs

Nalin C. W. Goonesekere[1,2] and Byungkook Lee[1*]

[1] Laboratory of Molecular Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-4264

[2] Department of Chemistry and Biochemistry, University of Northern Iowa, Cedar Falls, Iowa 50613-0423

ABSTRACT

The sequence homology detection relies on score matrices, which reflect the frequency of amino acid substitutions observed in a dataset of homologous sequences. The substitution matrices in popular use today are usually constructed without consideration of the structural context in which the substitution takes place. Here, we present amino acid substitution matrices specific for particular polar–nonpolar environment of the amino acid. As expected, these matrices [context-specific substitution matrices (CSSMs)] show striking differences from the popular BLOSUM62 matrix, which does not include structural information. When incorporated into BLAST and PSI-BLAST, CSSM outperformed BLOSUM matrices as assessed by ROC curve analyses of the number of true and false hits and by the accuracy of the sequence alignments to the hit sequences. These findings are also of relevance to profile–profile-based methods of homology detection, since CSSMs may help build a better profile. Profiles generated for protein sequences in PDB using CSSM-PSI-BLAST will be made available for searching via RPSBLAST through our web site http://lmbbi.nci.nih.gov/.

Proteins 2008; 71:910–919.
Published 2007 Wiley-Liss, Inc.[†]

Key words: CSSM; score matrix; homology detection; sequence alignment; environment polarity; fold recognition.

INTRODUCTION

Whole genome sequencing projects have yielded sequences of a large number of proteins for which no direct experimental information is available. To begin to understand the biological importance of these proteins, they need to be related to other proteins for which more information is available. The usual first step for finding such related proteins is a sequence homology search using programs such as BLAST[1] and PSI-BLAST.[2]

A sequence homology search program measures the degree of sequence similarity between two sequences by using a score matrix, which assigns a score for each aligned residue pair. Many score matrices have been devised over the years.[3–6] These matrices are all based on average substitution frequency for each pair of amino acid types observed in a given database. For example, the popular BLOSUM62 matrix[6] is built from multiply aligned sequence segments or "blocks" clustered at 62% sequence identity.

When many homologous sequences are already known for a given sequence, one can design matrices tailored to the particular set of sequences by constructing them specific for each position of the sequence. Use of such position-specific score matrices (PSSMs or the "profile"), as is done in PSI-BLAST, has improved the ability to detect more remotely related sequences.[2] However, a PSSM is built in an iterative process, which begins by collecting homologous sequences identified by using a pairwise comparison score matrix. Therefore, it is still important to have a good score matrix for individual pair comparisons, to obtain a better PSSM or in cases when no or only few homologous proteins are identifiable.

Altschul and coworkers recently introduced methods for improving the score matrix by considering amino acid compositional biases in the background and target sequences.[7–10] Here, we show that the score matrices can also be improved by considering the polar/nonpolar environment of the amino acids. For example, when the residue is in a nonpolar environment, the substitution score should be high for a hydrophobic, and low or negative for a hydrophilic, residue.

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.
Grant sponsors: National Cancer Institute, NIH (Intramural Research Program), Battelle Applied Research Grant.
*Correspondence to: Byungkook Lee, Rm. 5120, Bldg. 37, National Institutes of Health, 37 Convent Drive, MSC 4264, Bethesda, MD 20892-4264. E-mail: bk@nih.gov
Received 25 September 2006; Revised 28 July 2007; Accepted 13 August 2007
Published online 14 November 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21775

Published 2007 WILEY-LISS, INC.

[†]This article is a US Government work and, as such, is in the public domain in the United States of America.

Many authors have built score matrices[4,11–17] of various kinds, which include structural information for use in protein-fold recognition. Here, we consider only the narrow problem of improving the amino acid substitution matrix by using structural information. The objective is to find as many homologous proteins as possible and not directly to predict the structure for a given protein sequence. Finding many homologous proteins increases the chances of gathering more information about the function of a given protein. It can also indirectly aid in finding the correct template for structure prediction, since one may be able to build a more powerful profile for every known structure.

Structural environment-specific amino acid substitution tables have been previously computed[11–14] and shown to significantly improve detection of homologous sequences. In the present study, we focus on the effect of the polarity of the environment of residues only, expanding the binary categorization (buried vs. solvent-exposed) to a description of residue environment polarity. Specifically, we construct a set of four matrices, each of which is specific for a specified range of the polarity of the environment of the residue being substituted. These matrices (CSSM for context-specific substitution matrix) were computed from a database of structurally aligned protein domain pairs, selected from an all-against-all pairwise structural alignment of 3992 SCOP[18] protein domains of low sequence homology. We find that these matrices are indeed strikingly different from one another and from the BLOSUM62 matrix.

Since we are primarily interested in extending the list of homologous sequences to a given sequence, we implemented CSSMs in BLAST and PSI-BLAST and evaluated their effectiveness in finding homologs against conventional programs that used the BLOSUM series matrices. We find that CSSMs indeed outperform BLOSUM matrices in both implementations.

MATERIALS AND METHODS

Structurally homologous protein domain pair database

Pairwise structural superposition of protein domains

A nonredundant set of 3992 protein domains with less than 40% sequence identity among them, which excluded structures determined by NMR, were selected from the ASTRAL SCOP v1.59 database.[19] Domain selection to this set was based on the SPACI score,[19] which is a measure of structure quality. These domains were subjected to an all-against-all pairwise structural superposition using the structure comparison program SHEBA,[20] to generate a structurally superposed domain pair dataset. The computations were performed on the NIH biowulf cluster (a Beowulf parallel processing system).

Selection of structurally homologous domain pairs

The structurally homologous protein domain pair (SHoPP) database consists of a subset of these all-against-all structurally aligned protein domain pairs. The subset was selected using the criteria given below. These criteria are based on the number of aligned residue pairs, m. SHEBA determines the aligned residue pairs by using a dynamic programming algorithm on two structurally superposed structures. A necessary condition for a pair of residues to be "aligned" is that the distance between the α carbons of the pair is less than 3.5 Å after superposition of the domains. The criteria for inclusion in the SHoPP database were

1. $m \geq 40$
2. $m \geq 0.6 \times$ number of residues in the larger domain in domain pair
3. The root-mean-square deviation of superposed residues ≤ 2.0 Å

These criteria were developed, in part, by manual examination of a sample of superposed domain pairs, to ensure the correct superposition of particularly the β-sheet residues.

The domain pairs selected in this manner also had z-scores greater than 3 for each domain, when the z-score was defined as

$$z = \frac{m_\mathrm{f} - \langle m_\mathrm{f} \rangle}{\sigma}$$

where

m_f = m/number of residues in the larger domain

$\langle m_\mathrm{f} \rangle$ = mean of m_f over all pairs involving the given domain

σ = standard deviation of m_f over all pairs involving the given domain

There were 9806 domain pairs and 2,164,492 aligned residue pairs in the SHoPP database.

Context-specific substitution matrices

Grouping by environmental polarity

The amino acid substitutions observed in the SHoPP database were categorized according to the polarity of the environment of the residues. The environmental polarity (EP) of a residue is defined as the fraction of the solvent-accessible surface area[21] of the residue that is either covered by a polar protein atom or exposed to the solvent. This was computed using the program SHEBA. An amino acid substitution matrix was computed separately for each EP range of 0%–25%, 25%–50%, 50%–75%, and 75%–100% (see the next section). These four matrices are referred to as the CSSMs. We will use the variable α to denote the four ranges of EP corresponding to the four CSSMs.

Computation of the log-odds scoring matrix

The CSSM for each EP range α was computed using the following log-odds (lod) formula:

$$s_{ij}^{\alpha} = \left(\ln \frac{q_{ij}^{\alpha}}{(q_{ij}^{\alpha})^R} \right) \frac{1}{c}$$

where s_{ij}^{α} is the score given when a residue of type i of the query sequence in the EP range α is aligned with a residue of type j of the subject sequence. The frequency q_{ij}^{α} was computed as

$$q_{ij}^{\alpha} = \frac{N_{ij}^{\alpha}}{\sum_{k=1}^{20} N_{ik}^{\alpha}}$$

where $N_{ij}{}^{\alpha}$ is the number of residue pairs of types i and j in the SHoPP database for which the EP of one of the residues is in the range α and that of the second residue is within ±30 of the first residue. The liberal (fuzzy) condition used for the EP of the second residue was (1) to ensure an adequate sampling of rare substitutions, for example, Glu → Arg in CSSM$_1$, and (2) to enable a smoother transition between the corresponding elements of different CSSMs. The corresponding frequency expected for a randomly aligned protein pair was calculated by

$$(q_{ij}^{\alpha})^R = \frac{\sum_{ab} N_i^{\alpha}(a) N_j(b)}{\sum_{k=1}^{20} \sum_{ab} N_i^{\alpha}(a) N_k(b)}$$

where $N_i{}^{\alpha}(a)$ is the number of residues of type i with EP in the range α in protein a, and $N_j(b)$ and $N_k(b)$ are the number of residues of types j and k, respectively, in protein b, which is aligned to protein a, and the summation labeled ab is over all aligned domain pairs, a–b, in the SHoPP database. Note that residues in all EP ranges are counted for the target protein b in computing the random alignment frequencies.

The constant factor $1/c$ is set to $2/\ln 2$, to express the score in half-bit units.

Evaluating the performance of CSSM in BLAST and PSI-BLAST

Implementation of CSSM in BLAST

Modifications were made to the BLAST source code (NCBI Toolkit v 2.2.4) to incorporate CSSMs. During a search for homologous sequences to a given query sequence, the appropriate CSSM was selected based on information provided on the environment polarity of each residue of the query sequence. In the case of PSI-BLAST, CSSMs were used to compute substitution scores only during the first cycle of PSI-BLAST. The initial PSSM was generated by using hits from this cycle. For the second and subsequent cycles of PSI-BLAST, the default scoring matrix of PSI-BLAST, BLOSUM62, was utilized in constructing the PSSM from hits. The statistical parameters λ, κ, α, and β, which are required to calculate the normalized score and the *E*-value of a hit, were computed by the computer program obtained from Steven Altschul. Input to this program was a matrix constructed in an analogous manner to CSSM by using the SHoPP database, but disregarding the environment polarity. PSI-BLAST experiments were performed for 20 cycles (or to convergence) using a threshold *E*-value of 0.001.[7,22] The default (11,1) affine gap penalty scheme was used for all experiments.

Evaluation of Hits

Domains that belong to the same SCOP[18] superfamily were considered to be homologous (true hits), whereas domains belonging to different folds were considered nonhomologous (false hits).[22–24] The domains that belong to the same fold, but different superfamilies, were not counted as either true or false hits.

Datasets used

For all experiments, the target sequence set was a subset of the ASTRAL-SCOP[19] v1.65 database, which contained 6442 protein domain sequences, each with less than 50% sequence identity to any other sequence in the database. This is somewhat different from the dataset used to set up the SHoPP database and the CSSMs (see above).

The query sequence set consisted of 92 sequences, each of which had at least 10 SCOP family members (including self) in the target sequence set. This was to ensure that a large number of true positives exist in the target dataset, since the success of PSI-BLAST depends on the creation of a robust position-specific scoring matrix (PSSM) from the initial hits. The 92 query sequences represented 6 classes and 64 folds in the SCOP database.

Unless otherwise stated, BLAST and PSI-BLAST were run using the BLOSUM62 matrix.[24]

Evaluation of alignment quality

Pairwise structure-based sequence alignments from the SABmark superfamily dataset[25] were used as the reference dataset to evaluate alignment quality. These sequences have, at most, 50% sequence identity to each other and are grouped by SCOP superfamily, representing 425 SCOP superfamilies.[25] This is a dataset that was set up entirely independently from the SHoPP dataset, which was used to construct the CSSM.

Table I
Amino Acid Substitution Scores for Leucine in $CSSM_1$ and $CSSM_4$ (Half-Bits)

		$CSSM_1$	$CSSM_4$	STRUCT	BLOSUM62
L	A	−1	−1	−1	−1
L	C	0	−5	−1	−1
L	D	−12	0	−3	−4
L	E	−11	**1**	−2	−3
L	F	**2**	−2	1	0
L	G	−7	−1	−3	−4
L	H	−6	0	−2	−3
L	I	**3**	−1	2	2
L	K	−13	**2**	−2	−2
L	L	**4**	0	3	4
L	M	**2**	0	1	2
L	N	−9	0	−2	−3
L	P	−6	0	−2	−3
L	Q	−8	**1**	−1	−2
L	R	−11	**1**	−2	−2
L	S	−6	0	−2	−2
L	T	−4	0	−1	−1
L	V	**2**	−2	1	1
L	W	0	−2	0	−2
L	Y	−2	−1	0	−1

The positive elements in $CSSM_1$ and $CSSM_4$ are shown in bold.

BLAST runs were performed for each sequence in a group, with the target database comprising all sequences in the group. Two metrics, f_D and f_M, were used to evaluate alignment quality.[26] f_D and f_M are the number of correctly aligned residues divided by the length of the reference and test alignments, respectively.

Construction of PSSM profile database using CSSM-PSI-BLAST

Thirty-nine thousand and nine-hundred and fifty-two (39,952) pdb files were obtained from the RCSB protein structure database[27] (February 13, 2007). They consisted of 18,484 single chain files and 21,468 multichain files. The environment polarity of each residue in each chain was calculated using the program SHEBA.[20] All single chains, and all chains with more than 50 amino acids in multichain files were used as input to CSSM-PSI-BLAST program to generate PSSM profiles (up to 20 cycles). The target database used in each case was the NCBI non-redundant protein (nr) database containing 4,565,699 sequences. A total of 75,794 PSSM profiles were generated this way. These profiles are searchable by the program RPS-BLAST included in the NCBI BLAST suit of programs (http://www.ncbi.nlm.nih.gov/).

RESULTS

Structurally homologous protein domain pair database

The ASTRAL SCOP v1.59[19] database contained 3992 protein domains with less than 40% sequence identity among them. An all-against-all structure comparison of these protein domains, using the structure comparison program SHEBA,[20] resulted in 9806 structurally aligned domain pairs that were judged to be structurally homologous to each other according to the criteria detailed in the Materials and Methods section. The SHoPP database consists of these protein domain pairs. It contained 2,164,492 structurally aligned residue pairs (Ca distance <3.5 Å).

Analysis of CSSMs

The raw counts of amino acid substitutions in the SHoPP database and the calculated CSSMs are given in the Supplementary Material of this article. The total number of residue pairs used to construct these matrices were 455,252, 1,176,586, 1,401,074, and 931,480, respectively, for the $CSSM_1$, $CSSM_2$, $CSSM_3$, and $CSSM_4$ matrices. The matrix elements in $CSSM_1$ ($0\% \leq EP < 25\%$) and $CSSM_4$ ($75\% \leq EP < 100\%$) that represent substitution scores for Leu and Arg are given in Tables I and II as examples.

For both Leu and Arg, the corresponding elements in the two matrices differ significantly from each other. For example, the positive elements ($\geq$1, shown in bold letters) in one matrix do not overlap with the positive elements in the other. As expected, for $CSSM_1$, the favored substitutions for Leu are all hydrophobic. This trend also holds for Arg, where even the self-substitution (Arg–Arg) is disfavored. For $CSSM_4$, all favored substitutions are hydrophilic both for Arg and for Leu. For BLOSUM62, in contrast, favorable substitutions for Leu are generally

Table II
Amino Acid Substitution Scores for Arginine in $CSSM_1$ and $CSSM_4$ (Half-Bits)

		$CSSM_1$	$CSSM_4$	STRUCT	BLOSUM62
R	A	**3**	−1	−1	−1
R	C	**2**	−5	−2	−3
R	D	−7	**0**	0	−2
R	E	−10	**1**	1	0
R	F	**2**	−3	−2	−3
R	G	−2	−1	−1	−2
R	H	−3	**0**	0	0
R	I	**2**	−4	−2	−3
R	K	−8	**3**	2	2
R	L	**2**	−3	−2	−2
R	M	**2**	−2	−1	−1
R	N	−4	**1**	0	0
R	P	−5	0	−1	−2
R	Q	−5	**2**	1	1
R	R	−4	**3**	4	5
R	S	−3	0	0	−1
R	T	−2	0	0	−1
R	V	**2**	−3	−2	−3
R	W	−1	−3	−1	−3
R	Y	0	−2	−1	−2

The positive elements in $CSSM_1$ and $CSSM_4$ are shown in bold.

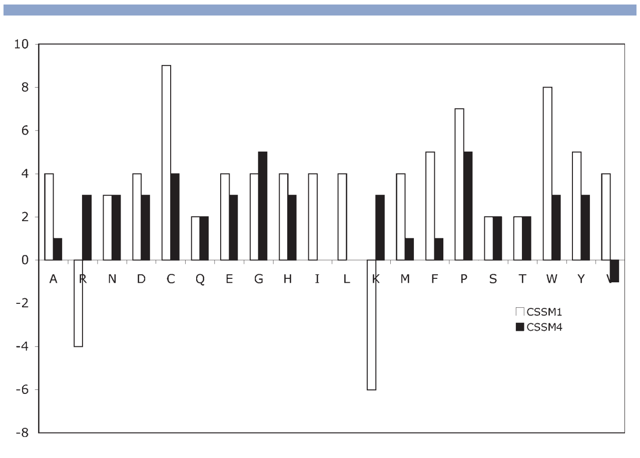

Figure 1

The self substitution scores for amino acids in CSSM$_1$ and CSSM$_4$. These values also represent the degree of conservation observed for each amino acid.

hydrophobic, whereas for Arg they are hydrophilic. Figure 1 shows that self-substitution is disfavored in CSSM$_1$ for both Arg and Lys.

Similarly, in the CSSM$_4$, the self-substitution score is either 0 (for Leu and Ile) or negative (for Val) for the strongly hydrophobic residues.

Examination of the full matrices (Supplementary Material) confirms this trend. The average score for substitution by a nonpolar residue (X replaced by L, I, V, M, or F) is positive (1.38) for CSSM$_1$, and negative (−2.71) for CSSM$_4$, whereas BLOSUM62 has an intermediate value (−1.12).

One can also notice that the elements of the CSSM$_1$ show a large variation, indicating that there are strong preferences and avoidances for residue types in the nonpolar environment. In comparison, the matrix elements of the CSSM$_4$ vary little, indicating that amino acid-type preferences are less strong when the residue is on the surface. As expected, the degree of variation of the BLOSUM62 matrix elements is between those of the CSSM$_1$ and CSSM$_4$ matrix elements. A quantitative measure of these features is the relative entropy.[28] The values of the relative entropy for the matrices computed in this study are given in Table III.

Analysis of hits from BLAST

To evaluate the effectiveness of CSSMs in detecting homologous sequences, they were implemented in BLAST (CSSM-BLAST) (see Materials and Methods section). A subset of the ASTRAL-SCOP v1.65 database, containing 6442 sequences having no more than 50% sequence identity to each other, served as the target database. A set of 92 query sequences were chosen from this database such that each sequence had at least 10 SCOP family members in the target database. The results were evaluated with respect to the default version of BLAST, which employed the scoring matrix BLOSUM62. Hits were the sequences with the E-value less than a cut-off value. A hit was considered true if it belonged to the same SCOP superfamily as the query.[22,23]

Table III

The Relative Entropy[20] of Matrices (Half-Bits)

Matrix	Relative entropy
CSSM$_1$	1.8785
CSSM$_2$	1.3331
CSSM$_3$	0.5682
CSSM$_4$	0.5852
STRUCT	0.6389
BLOSUM62	0.6979

When the cut-off *E*-value was set to 0.001, 47 of the 92 queries had more true hits by CSSM-BLAST than by BLAST, whereas the opposite was true only for eight queries (see Fig. 2). There were no hits to the "uncertain" region of same fold but different superfamily by either program. The false positive rate was very low, with one observation for CSSM-BLAST, and none for BLAST. The ROC$_{50}$ curves calculated for this data (see Fig. 3) indicate that the CSSM-BLAST performs better than BLAST at all *E*-value cut-off levels.

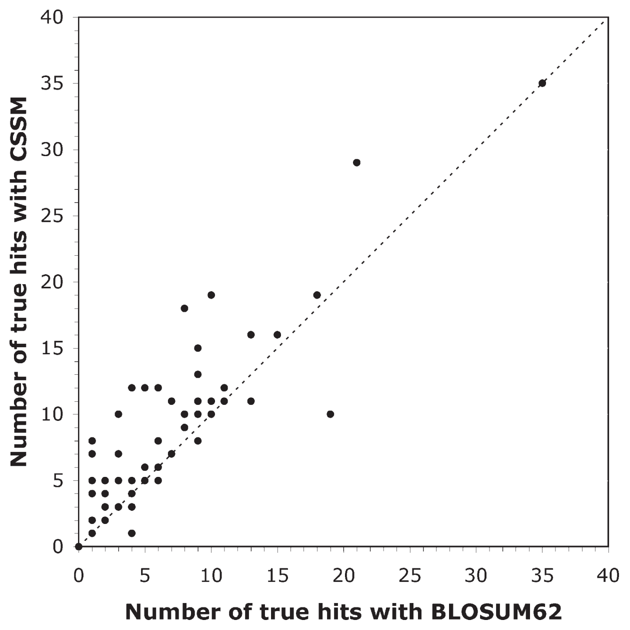

Figure 2

Number of true hits using CSSM-BLAST (y-axis) plotted against that using default BLAST with BLOSUM62 (x-axis). Each point represents one of 92 query sequences. The target sequences were ASTRAL-SCOP v1.65 database (50% sequence identity) in each case. True hits were those with E-values <0.001 and which belong to the same SCOP superfamily as the query. If the CSSM-BLAST performed the same as the default BLAST, all points would fall on the diagonal line, which is indicated by a dotted line.

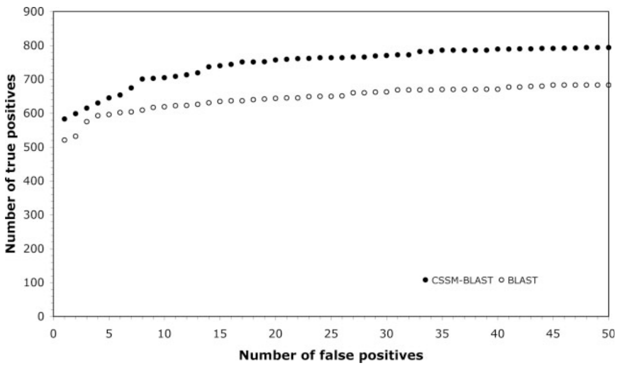

Figure 3

ROC$_{50}$ curves computed from pooled results for a set of 92 query sequences with ASTRAL-SCOP v1.65 (50% sequence identity) as target database, using CSSM-BLAST (solid circles) and regular BLAST with BLOSUM62 (open circles). Hits to the same SCOP superfamily were considered true positives. Hits to different SCOP folds were considered false positives.[11,12]

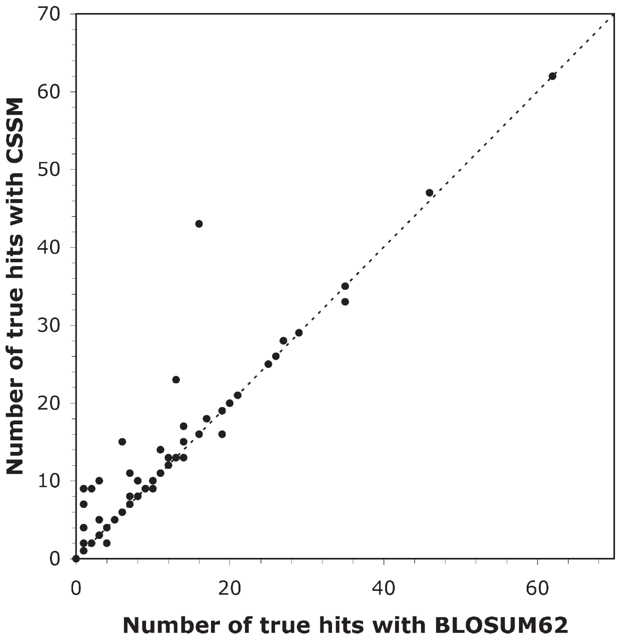

Figure 4

Number of true hits using CSSM-PSI-BLAST (y-axis) plotted against that using default PSI-BLAST with BLOSUM62 (x-axis). Each point represents one of 92 query sequences. The target sequences were ASTRAL-SCOP v1.65 database (50% sequence identity) in each case. True hits were those with E-values <0.001 and which belong to the same SCOP superfamily as the query. If the CSSM-PSI-BLAST performed the same as the default PSI-BLAST, all points would fall on the diagonal line, which is indicated by a dotted line.

Analysis of hits from PSI-BLAST

We were motivated to implement a CSSM version of PSI-BLAST (CSSM-PSI-BLAST), since the CSSM-BLAST found more true hits than regular BLAST (see above). For those query sequences for which many hits can be found by BLAST, a good PSSM can be set up from the initial round of PSI-BLAST, and CSSM is probably unnecessary. On the other hand, for those query sequences for which there are no or only few initial hits, CSSM-PSI-BLAST could help by finding more initial hits. Accordingly, in our initial implementation of CSSM-PSI-BLAST, CSSM was used to generate hits in the first round of PSI-BLAST. For the second and subsequent rounds, PSSM was generated by the default formula, which uses hits from the previous round and BLOSUM62. CSSM-PSI-BLAST was tested with the same query and target database as for CSSM-BLAST (see above), and the results were evaluated with respect to a default version of PSI-BLAST using BLOSUM62. In this case, the number of query sequences for which more true hits were found by CSSM-PSI-BLAST than by PSI-BLAST was 22, whereas the latter found more true hits than the former for only five query sequences (see Fig. 4). The ROC$_{50}$ curves (see Fig. 5) show that the CSSM-PSI_BLAST finds more true hits at all *E*-value cut-off levels except at the most stringent level.

A comparison of BLOSUM62 with a "structure-only" version of CSSM (STRUCT)

Since CSSM was derived from a database (SHoPP) that was different from the database used to construct BLOSUM62, we investigated the effects of the database alone, by constructing a single matrix analogous to BLOSUM62 from SHoPP, without considering environment polarity. This matrix, called STRUCT, was tested against BLOSUM62 by evaluating its performance in BLAST,

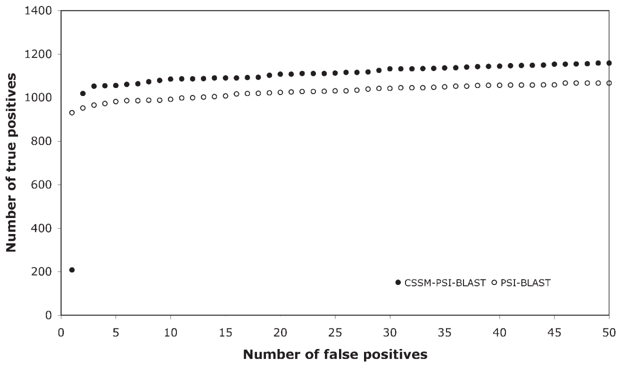

Figure 5

ROC$_{50}$ curves computed from pooled results for a set of 92 query sequences with ASTRAL-SCOP v1.65 (50% sequence identity) as target database. Results are shown for the CSSM implementation of PSI-BLAST (solid circles) and a default version of PSI-BLAST using BLOSUM62 (open circles).

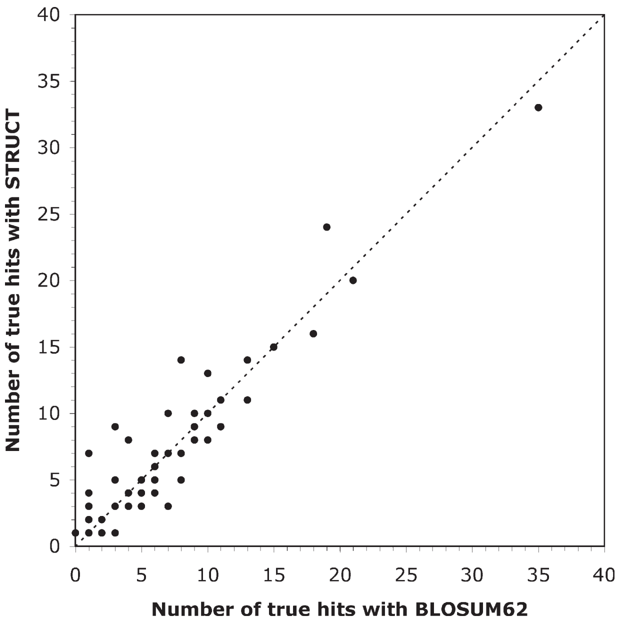

Figure 6

Number of true hits for BLAST using STRUCT (y-axis) plotted against that using BLOSUM62 (x-axis). Each point represents one of 92 query sequences. The target sequences were ASTRAL-SCOP v1.65 database (50% sequence identity) in each case. True hits were those with E-values <0.001 and which belong to the same SCOP superfamily as the query. If STRUCT performed the same as BLOSUM62, all points would fall on the diagonal line, which is indicated by a dotted line.

using the query dataset of 92 sequences under conditions described previously for evaluating CSSMs.

The results (see Fig. 6) indicate that the performances of the two matrices are comparable. The calculated entropy for the "structure-only" matrix is similar to the value reported[6] for BLOSUM62 (see Table III).

Analysis of alignment quality

The effectiveness of CSSM versus BLOSUM62 was also tested in terms of the accuracy of the sequence alignment between the query and the hit proteins. The structure-based sequence alignments in the SABmark superfamily database[25] was used to evaluate the alignment quality (see Materials and Methods section).

CSSM-BLAST yielded 32,979 BLAST hits while there were 29,445 hits for the default BLAST with BLOSUM62. The 28,880 common hits were analyzed by $\langle f_D \rangle$ and $\langle f_M \rangle$ (see Fig. 7) where the averaging was performed after binning the results with respect to % sequence identity. The significantly higher values obtained for $\langle f_D \rangle$ [Fig. 7(a)] using CSSM-BLAST over BLAST indicate that the former gives more correctly aligned residues than the latter, while the similar or slightly better $\langle f_M \rangle$ values [Fig. 7(b)] obtained indicate that there was no increase in the fraction of falsely aligned residues when CSSM-BLAST is used over BLAST.

DISCUSSION

Residue environment information has been used as one of the key descriptors of tertiary structure from the beginning of the fold recognition methods,[29] and its power in increasing the detection of homologous sequences was demonstrated early, when Overington *et al.*[11] constructed amino acid substitution matrices based on solvent accessibility and secondary structure attributes of the amino acids. These substitution matrices have been subsequently implemented in the fold recognition program FUGUE.[13] Exploiting the large set of solved protein structures now available, we have expanded the binary (buried vs. solvent exposed) description of residue environment used in this earlier study and constructed a set of four amino acid substitution matrices (CSSM) based on the polarity of the environment around an amino acid. These matrices, and the raw counts of each

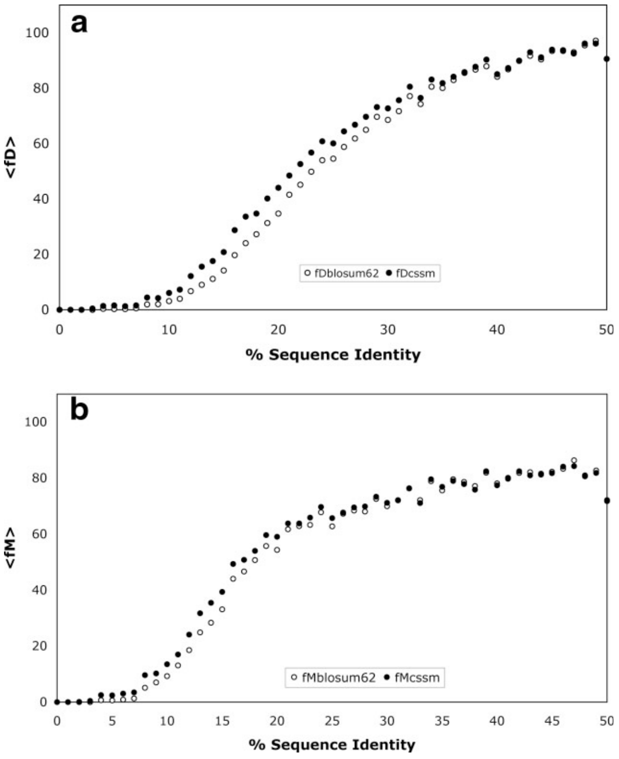

Figure 7

Twenty-eight thousand and eight-hundred and eighty (28,880) pairwise sequence alignments obtained from BLAST and CSSM-BLAST were evaluated using structure-based alignments from the SABmark database[14] as reference. (a) A plot of the averaged value of f_D as a function of % sequence identity for CSSM-BLAST (solid circles) and BLAST (open circles). (b) An analogous plot for f_M. The f_M and f_D values were binned by % sequence identity of the aligned pair in the reference alignment, prior to averaging.

amino acid substitutions in each polarity range, are available in the Supplementary Material. An examination of these matrices shows that there are striking differences between, for example, the $CSSM_1$ and $CSSM_4$ matrices and that both, in turn, differ from the BLOSUM62 matrix. Although there are exceptions, CSSM matrices show that nonpolar residues are generally strongly favored to replace any residue type in a nonpolar environment and polar residues tend to be favored in a polar environment (see Tables I and II and Fig. 1). In contrast, BLOSUM62 matrix shows the like-to-like substitution pattern, that is, polar residues are favored to substitute for a polar residue and a nonpolar residue for a nonpolar residue. We also observe that amino acid preferences are noticeably more pronounced in the nonpolar environment ($CSSM_1$) than in the polar environment ($CSSM_4$) and that the degree of preferences reflected in the BLOSUM62 matrix is in-between those in the $CSSM_1$ and $CSSM_4$ matrices (see Table III).

To test the effectiveness of the new matrices, we modified the BLAST and PSI-BLAST programs so that they can use these matrices. These modified programs (CSSM-BLAST and CSSM-PSI-BLAST) utilize the same statistical framework employed in the original BLAST program (see Materials and Methods section), although it is not yet clear if the statistical parameters we computed are entirely valid for these conditional matrices. In any case, we find that the modification enhances the sensitivity of both the BLAST and PSI-BLAST searches without loss of selectivity and that the modified BLAST produces a more accurate alignment. We thus confirm the importance of residue environment[11–14] both in finding more homologs and in improving the alignment accuracy.

The procedure we used above for testing the performance of the CSSM-PSI-BLAST is different from that used by Altschul for testing the performance of PSI-BLAST.[10] To take advantage of the large number of homologous sequences that are usually present in a large database, Altschul first runs PSI-BLAST on a comprehensive database, obtains the profile, and then uses this profile to see how many of the homologous sequences the program finds that are known to be present in a smaller, better known dataset. For the purpose of testing the CSSMs, however, we felt that this procedure might dilute the result because, for many sequences with a large number of homologs, our implementation of the CSSM-PSI-BLAST was expected to perform like the original PSI-BLAST. The applications we have in mind for using CSSMs are cases when few homologs can be found using the conventional substitution matrix. Our test results indicate that more homologs may indeed be found in such cases by using CSSMs. Since, in addition, CSSMs produce a more accurate alignment, they will help build a more powerful profile. Therefore, use of CSSMs may also help the more recent profile–profile-based homology detection methods[30–32] in cases when few homologs can be found using the conventional score matrices.

There are many methods, broadly classified as "fold recognition," which compare a sequence against a library of structures.[13,15,16,33–37] The use of structural information by fold recognition methods has generally resulted in an improved ability to detect remote homologs, when compared to purely sequence-based methods.[38] In the best-case scenario, these methods will yield a homologous protein as the best hit. More often, these methods identify the correct fold as one of their top hits, but fail to correctly score the best solution as best.[39] Thus, the enhanced sensitivity of fold recognition methods comes at the cost of lower selectivity. It has also been pointed out[13] that a fold recognition technique finds more proteins with the same fold than methods that compare sequences, but that the latter finds more homologs in the same family and superfamily than the former. For this reason, and for the significant computational resources required,[40] fold recognition remains an expert's tool.[39] In contrast, sequence-based methods such as BLAST[1] and PSI-BLAST[2] have become extremely popular because of their ease of use,[36,41] speed, and the existence of a rigorous statistical framework for analysis of the output.[24] The use of CSSM-PSI-BLAST, and the profiles that they generate (see below), bridges the structure- and the sequence-based methods in that it uses the structural information but has all the characteristics of the sequence-based methods listed above.

The structurally aligned domain pairs in the SHoPP database were selected by strictly structural criteria and had less than 40% sequence identity. Nevertheless, when the CSSMs generated from this database were implemented in BLAST, all but one hit were in the same superfamily. There were no hits to the "uncertain region"[22] of the same fold, but not of the same superfamily. Thus, CSSM-BLAST and CSSM-PSI-BLAST are homology detection tools, like the parent BLAST and PSI-BLAST programs. In this respect, these programs differ from many fold recognition tools, which, in principle, may also detect structural analogs.[17,42]

We envision that the CSSMs can be used in two different ways in practice. If the structure of the protein of interest is known, the CSSM-BLAST and CSSM-PSI-BLAST can be used directly against a sequence database to extend the list of homologous proteins. This can be important when the structure of the protein is known, but its function is not, as in the case of many protein structures produced by the structural genomic project. In the second application, the CSSMs can be used to generate more powerful profiles for each known structure. We have generated a set of PSSM profiles, one for each chain of every protein in pdb, using the CSSM-PSI-BLAST (see Materials and Methods section). These profiles can be used in detecting homologs of known structure, a process akin to fold recognition, using RPS-BLAST (NCBI BLAST suite of programs).

A web server to search the PSSM profiles using RPS-BLAST has been set up and will be made available to the public (http://lmbbi.nci.nih.gov). The program SHEBA,[20] the perl script to construct environment polarity from the output of SHEBA, and the modified BLAST suite of programs to run CSSM-BLAST and CSSM-PSI-BLAST are or will be made available for download from the same site. In a preliminary study with the ~2000 families classified as domains of unknown function (DUFs) in Pfam (version 22.0, July 2007), we found that CSSM-derived profiles could be used to associate 486 DUF families with a structure (*E*-value <0.001 for at least one sequence in the family), which previously have not been associated with a structure. (There are 205 other DUF families, which previously have been associated with a structure.) In contrast, when the sequence with the lowest *E*-value from each of these 486 families was used as the query sequence in BLOSOM62 PSI-BLAST runs against the nr database, 298 sequences did not produce any hits (*E*-value < 0.001) to sequences in pdb (data not shown).

The method we used to include structural information in sequence alignment/search procedure can be improved in at least two different ways. First, in our current implementation of CSSM in PSI-BLAST (CSSM-PSI-BLAST), CSSMs are used only to compute hits for the first round of PSI-BLAST (BLAST). For the second and subsequent rounds, the default BLOSUM62 matrix is used, in conjunction with the hits of the previous round. An alternate implementation would be to replace BLOSUM62 by CSSMs for all rounds of PSI-BLAST. Since environment polarity information would then be propagated throughout all cycles of PSI-BLAST, this modification may further enhance the sensitivity of CSSM-PSI-BLAST.

Secondly, the approach we have taken to introduce environment polarity information in computing substitution matrices can be extended to include the secondary structure information as Overington *et al.*[11] and Shi *et al.*[13] have done. Using a program that automatically finds the best set of structural features for optimal sequence alignment, Gelly and Gracy[14] found that the secondary structure and the solvent accessibility information were the most useful. If a simplified three-state description of the secondary structure (helix, strand, and coil) is adopted, the computation of 12 (4×3) matrices is necessary, to categorize each environment polarity matrix by secondary structure. This involves determination of 4800 parameters ($4 \times 3 \times 20 \times 20$ matrix elements), which requires a large amount of data. On the other hand, the SHoPP database can be greatly expanded by including domain pairs with >40% identity. It is possible that the resulting matrices could reflect substitution frequencies that are optimal for detecting sequences with higher similarity,[43] but this is by no means certain. For example, we found that the performance of BLOSUM45 was inferior to BLOSUM62 under conditions used to evaluate CSSMs (data not shown). It is possible that using sequences with somewhat higher sequence identity will produce matrices with generally improved sensitivity because of higher information content of such sequence pairs.[43]

In summary, we confirm the findings of Overington *et al.*[11] and others that the inclusion of the polar–nonpolar environment information strongly affects the amino acid substitution matrix. We show that this information can be incorporated into the popular BLAST and PSI-BLAST programs to effectively increase the sensitivity of homology detection and point to a couple of directions in which the method can be further improved.

ACKNOWLEDGMENTS

We thank Dr. Stephen Altschul for providing the program that calculates the statistical parameter values and for critically reviewing an earlier version of the manuscript. We thank Chin-Hsien Tai for setting up the RPS-BLAST web server. We are also indebted to the referees who made very valuable comments and suggestions.

REFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
2. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
3. Vogt G, Etzold T, Argos P. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J Mol Biol 1995;249:816–831.
4. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. J Mol Biol 2001;307:721–735.
5. Henikoff S, Henikoff JG. Performance evaluation of amino acid substitution matrices. Proteins 1993;17:49–61.
6. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
7. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
8. Yu YK, Wootton JC, Altschul SF. The compositional adjustment of amino acid substitution matrices. Proc Natl Acad Sci USA 2003; 100:15688–15693.
9. Yu YK, Altschul SF. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics 2005;21:902–911.
10. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schaffer AA, Yu YK. Protein database searches using compositionally adjusted substitution matrices. FEBS J 2005;272:5101–5109.
11. Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. Protein Sci 1992;1:216–26.
12. Johnson MS, May AC, Rodionov MA, Overington JP. Discrimination of common protein folds: application of protein structure to sequence/structure comparisons. Methods Enzymol 1996;266:575–598.

13. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310: 243–257.
14. Gelly JC, Chiche L, Gracy J. EvDTree: structure-dependent substitution profiles based on decision tree classification of 3D environments. BMC Bioinformatics 2005;6:4.
15. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
16. Teodorescu O, Galor T, Pillardy J, Elber R. Enriching the sequence substitution matrix by structural information. Proteins 2004;54:41–48.
17. Tan YH, Huang H, Kihara D. Statistical potential-based amino acid similarity matrices for aligning distantly related protein sequences. Proteins 2006;64:587–600.
18. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
19. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28: 254–256.
20. Jung J, Lee B. Protein structure alignment using environmental profiles. Protein Eng 2000;13:535–543.
21. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400.
22. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
23. Kinch LN, Grishin NV. Evolution of protein structures and functions. Curr Opin Struct Biol 2002;12:400–408.
24. Green RE, Brenner SE. Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. Proc IEEE 2002;9:1834–1847.
25. Van Walle I, Lasters I, Wyns L. SABmark–a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics 2005;21:1267–1268.
26. Sauder JM, Arthur JW, Dunbrack RLJ. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 2000;40:6–22.
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
28. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 1991;219:555–565.
29. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
30. von Ohsen N, Sommer I, Zimmer R. Profile-profile alignment: a powerful tool for protein structure prediction. Pac Symp Biocomput 2003;8:252–263.
31. Mittelman D, Sadreyev R, Grishin N. Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. Bioinformatics 2003;19:1531–1539.
32. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 2002;315:1257–1275.
33. Jones DT. Protein structure prediction in the postgenomic era. Curr Opin Struct Biol 2000;10:371–379.
34. Friedberg I, Jaroszewski L, Ye Y, Godzik A. The interplay of fold recognition and experimental structure determination in structural genomics. Curr Opin Struct Biol 2004;14:307–312.
35. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
36. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
37. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000;119–130.
38. Kihara D, Skolnick J. Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR-Q. Proteins 2004;55:464–473.
39. Przybylski D, Rost B. Improving fold recognition without folds. J Mol Biol 2004;341:255–269.
40. Thiele R, Zimmer R, Lengauer T. Protein threading by recursive dynamic programming. J Mol Biol 1999;290:757–779.
41. Jones DT, Swindells MB. Getting the most from PSI-BLAST. Trends Biochem Sci 2002;27:161–164.
42. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 2004;56:502–518.
43. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 1990;87:2264–2268.