Advances and Applications in Bioinformatics and Chemistry

Dovepress

open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Evaluating the efficacy of a structure-derived amino acid substitution matrix in detecting protein homologs by BLAST and PSI-BLAST

This article was published in the following Dove Press journal:
Advances and Applications in Bioinformatics and Chemistry
4 June 2009
Number of times this article has been viewed

Nalin CW Goonesekere

Department of Chemistry and Biochemistry, University of Northern Iowa, Cedar Falls, IA, USA

Abstract: The large numbers of protein sequences generated by whole genome sequencing projects require rapid and accurate methods of annotation. The detection of homology through computational sequence analysis is a powerful tool in determining the complex evolutionary and functional relationships that exist between proteins. Homology search algorithms employ amino acid substitution matrices to detect similarity between proteins sequences. The substitution matrices in common use today are constructed using sequences aligned without reference to protein structure. Here we present amino acid substitution matrices constructed from the alignment of a large number of protein domain structures from the structural classification of proteins (SCOP) database. We show that when incorporated into the homology search algorithms BLAST and PSI-BLAST, the structure-based substitution matrices enhance the efficacy of detecting remote homologs.

Keywords: computational biology, protein homology, amino acid substitution matrix, protein structure

Introduction

Whole genome sequencing projects have yielded sequences of a large number of proteins for which no direct experimental information is available. Homology detection is a widely used tool for the structural and functional annotation of such proteins, since two related proteins with a common ancestor may retain the same ancestral function. A number of sequence homology search algorithms[1–5] have been developed for this purpose, and are in wide use. Despite these developments, identification of distant homologs in the twilight zone (sequence identity $<25\%$) has remained a challenge.

Pairwise sequence homology search programs[1,3] evaluate alignments between sequences by using a scoring scheme that includes a 20×20 amino acid substitution matrix and a penalty function for gaps. The substitution matrix assigns a scaled log-odds score for each aligned residue pair. Substitution matrices are also important in programs that utilize "sequence-profile"[2] and "profile–profile"[5,6] alignments, since the initial profiles are built by collecting homologous sequences identified by using a pairwise comparison score matrix.

Many amino acid substitution matrices have been devised over the years, utilizing a variety of methods.[7–11] The popular BLOSUM series is built from multiply aligned sequence segments or 'blocks' that represent the most conserved regions in aligned families.[12] The accuracy of these alignments will obviously impact the success of these matrices in detecting homologs. In the case of *sequence-based* matrices such as the BLOSUM series, the sequence alignments tend to become less reliable at large evolutionary distances.[13]

Correspondence: Nalin CW Goonesekere
Department of Chemistry and Biochemistry, University of Northern Iowa, 1227 W. 27th Street, Cedar Falls, IA 50614-0423, USA
Tel +1 319 273 3949
Fax +1 319 273 7127
Email nalin.goonesekere@uni.edu

Structure-based matrices obtain amino acid substitution data directly from structural alignments and hence largely avoid the issues relating to sequence alignment at large evolutionary distances. Thus, sequence alignments obtained from structure alignments have long been considered to be the gold standard, only being superceded by human curated alignments.[14] When compared with sequence-based matrices, though, structure-based matrices[13,15–17] have suffered from a paucity of data in the form of homologous protein structures required to construct a substitution matrix.[13,18] More recent applications of structure-based matrices have focused primarily on the quality of sequence alignments.[13,17] In some evaluations,[19] structure-based matrices have not performed as well as sequence-based matrices in the detection of remote homologues, and in the most recent comparison of substitution matrices performed by Brenner and colleagues,[20] structure-based matrices were omitted from the comparison.

A large number of protein structures have become available in the past decade, fueled in part by the structural genomics initiative.[21] In this paper, we investigate whether structure-based amino acid substitution matrices that exploit this resource could improve the detection of remote homologs. Access to larger datasets of remote homology significantly improved the performance of sequence-based matrices[20] and we hypothesized that the same may hold true for structure-based matrices as well.

The structurally aligned substitution matrices (SASM) that we describe here were computed using structurally aligned protein domain pairs. These pairs were selected from an all-against-all pairwise structural superposition of protein domains obtained from the ASTRAL SCOP[22] protein domain database. The SASMs that we computed were implemented in BLAST and PSI-BLAST, and their effectiveness in detecting remote homologs was compared against BLOSUM62.

Materials and methods

Pairwise structural superposition of protein domains

Nonredundant data sets of protein domains with less than 40% and 50% sequence identity to each other, which excluded structures determined by NMR, were selected from the ASTRAL SCOP v1.67 database.[22] Domain selections to these sets were based on the SPACI score,[22] which is a measure of structure quality. The 0%–40% dataset contained 6551 domains, and the 0%–50% dataset contained 7444 domains. Domains in each set were subjected to an all-against-all pairwise structural superposition using the structure comparison program SHEBA,[23] in order to generate a structurally superposed domain pair dataset. The total number of pairwise structural superpositions for the two datasets were 42,915,601 (0%–40%) and 55,413,136 (0%–50%). For each domain pair *ab*, the best superposition (*ab* vs. *ba*) was selected. Self superpositions (*aa*) were removed from further consideration.

Selection of structurally aligned domain pairs

Structurally aligned domain pairs were selected from among the structurally superposed pairs, using the following criteria. 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 superposed structures. A necessary condition for a pair of residues to be "aligned" is that the distance between the alpha carbons of the pair is less than 3.5Å after superposition of the domains.

(1) $m \geq 40$
(2) $m \geq 0.6 \times$ number of residues in the larger domain in domain pair
(3) The domain pairs had z-scores greater than a cut-off value (filter)

where the z-score was defined as

$$z = \frac{m_f - < m_f >}{\sigma}$$

with

m_f = m/number of residues in the larger domain
$< m_f >$ = mean of m_f over all pairs involving the given domain
s = standard deviation of m_f over all pairs involving the given domain

Computation of the log-odds scoring matrix

The log-odds scoring matrix was obtained as follows:[24]

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

where S_{ij} is the score given when a residue of type i is aligned with a residue of type j. The frequency q_{ij} was computed as

$$q_{ij} = \frac{N_{ij}}{\sum_{i=1}^{20} \sum_{j=1}^{i} N_{ij}}$$

where N_{ij} is the number of aligned residue pairs of types i and j. The corresponding frequency expected for a randomly aligned protein pair was calculated by

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

where $N_i(a)$ is the number of residues of type i in protein a and $N_j(b)$ is the number of residues of type j in protein b which is aligned to protein a, and the summation labeled ab is over all aligned domain pairs, a–b. The constant factor 1/c is set to 2/ln2, to express the score in half-bit units. The matrices generated using structurally aligned protein domain pairs from the 0%–40%, 0%–50% and 0%–60% sequence identity sets were labeled SASM40, SASM50, and SASM60, respectively. Ten matrices were generated for each sequence identity set, by varying the z-score filter used.

The relative entropy H of a matrix was computed according to Altschul[24] as follows:

$$H = \sum_{i=1}^{20}\sum_{j=1}^{i} q_{ij}s_{ij}$$

Evaluating the performance of structurally aligned substitution matrices (SASMs)

Implementation of SASMs in BLAST and PSI_BLAST

SASMs were implemented in BLAST and PSI_BLAST to evaluate its performance in pairwise and profile homology search algorithms. The statistical parameters for the appropriate extreme value distribution, 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. PSI-BLAST experiments were performed for 20 cycles (or to convergence) using a threshold E-value of 0.001.[25,26] The default (11,1) affine gap penalty scheme was used for all experiments.

Evaluation of hits

True hits were determined based on the human-curated SCOP[27] database assignments, rather than a pure structure-alignment based score.[28] Domains that belong to the same SCOP superfamily were considered to be homologous (true hits), while domains belonging to different folds were considered non-homologous (false hits).[19,25,29] 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[22] v1.65 database, which contained 6,442 protein domain sequences, each with less than 50% sequence identity to any other sequence in the database. The query sequence set consisted of 92 protein domain 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. Also, the most difficult pairwise relations to detect tend to be those between members of larger families and superfamilies.[20] The 92 query sequences represented six classes and 64 folds in the SCOP database. The BLOSUM62 substitution matrix was chosen as a benchmark for two reasons. First, it continues to be a popular choice for detecting homologs, and is the default matrix employed in both BLAST and PSI-BLAST. Second, in recent tests conducted on the performance of amino acid substitution matrices, the BLOSUM matrices have faired well against sequence- and structure-based matrices.[19,20] The (11,1) affine gap penalty function, which is optimal for BLOSUM62,[30] was used for both BLOSUM62 as well as for the SASMs.

Results

Generation of structurally aligned substitution matrices (SASM)

Protein domain datasets in ASTRAL SCOP v1.67,[22] each containing protein domains selected by pairwise sequence identity, were downloaded, and all protein domains within each dataset were structurally superposed in a pair-wise manner, using the program SHEBA.[23]

The selection criteria described previously (see Methods) were used to identify structurally aligned protein domain pairs from among the structurally superposed domain pairs, for each ASTRAL SCOP dataset. Increasing the z-score filter resulted in increasing the stringency, and a concomitant reduction in the number of protein domain pairs selected (Table 1).

According to Table 1, the SASM40 matrix that is isentropic with BLOSUM62 (0.7) was constructed using a z-score filter of 6.5, and the corresponding amino acid substitution scores are given in Table 2. The coefficient of determination (R^2 value) for the two matrices is 0.91. The single substitution score that shows a sign inversion is the cystein/valine substitution, which has a negative score in BLOSUM62.

Analysis of hits from BLAST

In order to evaluate their effectiveness in detecting homologous sequences, the SASMs were implemented in BLAST. A subset

Table 1 Some statistics for SASM matrices: the numbers[a] of structurally aligned protein pairs and the entropy values[b]

z-score	SASM40	SASM50
3.5	26,166 (0.512)	35,492 (0.608)
4.0	23,637 (0.541)	32,500 (0.639)
4.5	20,444 (0.582)	28,702 (0.682)
5.0	17,652 (0.617)	25,161 (0.723)
5.5	15,009 (0.649)	21,529 (0.768)
6.0	12,754 (0.678)	18,549 (0.805)
6.5	11,206 (0.698)	16,415 (0.800)
7.0	9,988 (0.714)	14,689 (0.856)
7.5	8,829 (0.730)	13,217 (0.873)
8.0	7,847 (0.743)	11,741 (0.889)

Notes: [a]The number of structurally aligned pairs is given as a function of the z-score filter used in the selection (see Methods). [b]For each matrix, the entropy value[24] is given within parenthesis.

of the ASTRAL SCOP v1.65 database, containing 6,442 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 matrix in BLAST, BLOSUM62. A hit (E-value < 0.001) was considered true if it belonged to the same SCOP superfamily as the query.[25,29]

SASM50 and SASM40 matrices performed better than BLOSUM62 in returning more true hits for a greater number of queries when the z-score filter was between 3.5 and 8.0 (Figure 1). SASM60, which is formally the SASM matrix most akin to BLOSUM62, performed the poorest in this comparison (data not shown). There were no false positives observed for SASM or BLOSUM62 in BLAST.

Analysis of hits from PSI-BLAST

SASMs was tested in PSI-BLAST using the same query sequences and target database as for BLAST (see above) and the results were evaluated with respect to BLOSUM62 (Figure 2). As in the case of BLAST, SASM50 and SASM40 matrices performed better than BLOSUM62 in returning more true hits for a greater number of queries when the z-score filter was between 3.5 and 8.0. This effect was more pronounced for PSI-BLAST, when compared to BLAST (Figures 1 and 2). The SASM40 matrix that returned more true hits for the greatest number of queries when compared with BLOSUM62 (23 vs 7) was constructed using a z-score filter of 6.5 (Table 2), and is isentropic with BLOSUM62.

Table 2 The structure-derived substitution matrix SASM40 (z-score filter = 6.5) in half bits[a]

A	2^{4}																			
R	-1	4^{5}																		
N	-1^{-2}	0	4^{6}																	
D	-1^{-2}	0^{-2}	1	5^{6}																
C	1^{0}	-2^{-3}	-1^{-3}	-2^{-3}	8^{9}															
Q	0^{-1}	1	1^{0}	1^{0}	-2^{-3}	3^{5}														
E	-1	1^{0}	0	2	-3^{-4}	2	4^{5}													
G	0	-1^{-2}	0	-1	-1^{-3}	-1^{-2}	-1^{-2}	5^{6}												
H	-1^{-2}	0	1	0^{-1}	-1^{-3}	1^{0}	0	-1^{-2}	6^{8}											
I	-1	-2^{-3}	-3	-4^{-3}	0^{-1}	-2^{-3}	-3	-4	-2^{-3}	3^{4}										
L	-1	-2	-2^{-3}	-3^{-4}	0^{-1}	-2^{-2}	-2^{-3}	-3^{-4}	-2^{-3}	2	3^{4}									
K	0^{-1}	2	1^{0}	0^{-1}	-2^{-3}	2^{1}	1	-1^{-2}	0^{-1}	-2^{-3}	-2	4^{5}								
M	0^{-1}	-1	-2	-3	0^{-1}	-1^{0}	-2	-2^{-3}	-1^{-2}	1	2	-1	4^{5}							
F	-1^{-2}	-2^{-3}	-2^{-3}	-3	0^{-2}	-2^{-3}	-3	-3	0^{-1}	1^{0}	1^{0}	-2^{-3}	1^{0}	4^{6}						
P	-1	-1^{-2}	-1^{-2}	0^{-1}	-2^{-3}	-1	0^{1}	-1^{-2}	-1^{-2}	-2^{-3}	-2^{-3}	0^{-1}	-2	-2^{-4}	6^{7}					
S	0^{1}	0^{-1}	1	0	0^{-1}	0	0	0	0^{-1}	-2	-2	0	-1	-2	0^{-1}	3^{4}				
T	0	0^{-1}	0	0^{-1}	0^{-1}	0^{-1}	0^{-1}	-1^{-2}	0^{-2}	-1	-1	0^{-1}	-1	-1^{-2}	-1	1	3^{5}			
W	-2^{-3}	-1^{-3}	-2^{-4}	-3^{-4}	-2	-2	-2^{-3}	-3^{-2}	-1^{-2}	-1^{-3}	0^{-2}	-2^{-3}	0^{-1}	2^{1}	-2^{-4}	-2^{-3}	-2	8^{11}		
Y	-1^{-2}	-1^{-2}	-1^{-2}	-2^{-3}	0^{-2}	-1	-2	-2^{-3}	1^{2}	-1	0^{-1}	-1^{-2}	0^{-1}	2^{3}	-2^{-3}	-1^{-2}	-1^{-2}	2	5^{7}	
V	0	-2^{-3}	-2^{-3}	-3	1^{-1}	-2	-2	-3	-2^{-3}	2^{3}	1	-2	1	0^{-1}	-2	-2	0	-1^{-3}	-1	3^{4}
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Notes: [a]When different, the corresponding matrix element for BLOSUM62 is shown as a superscript.

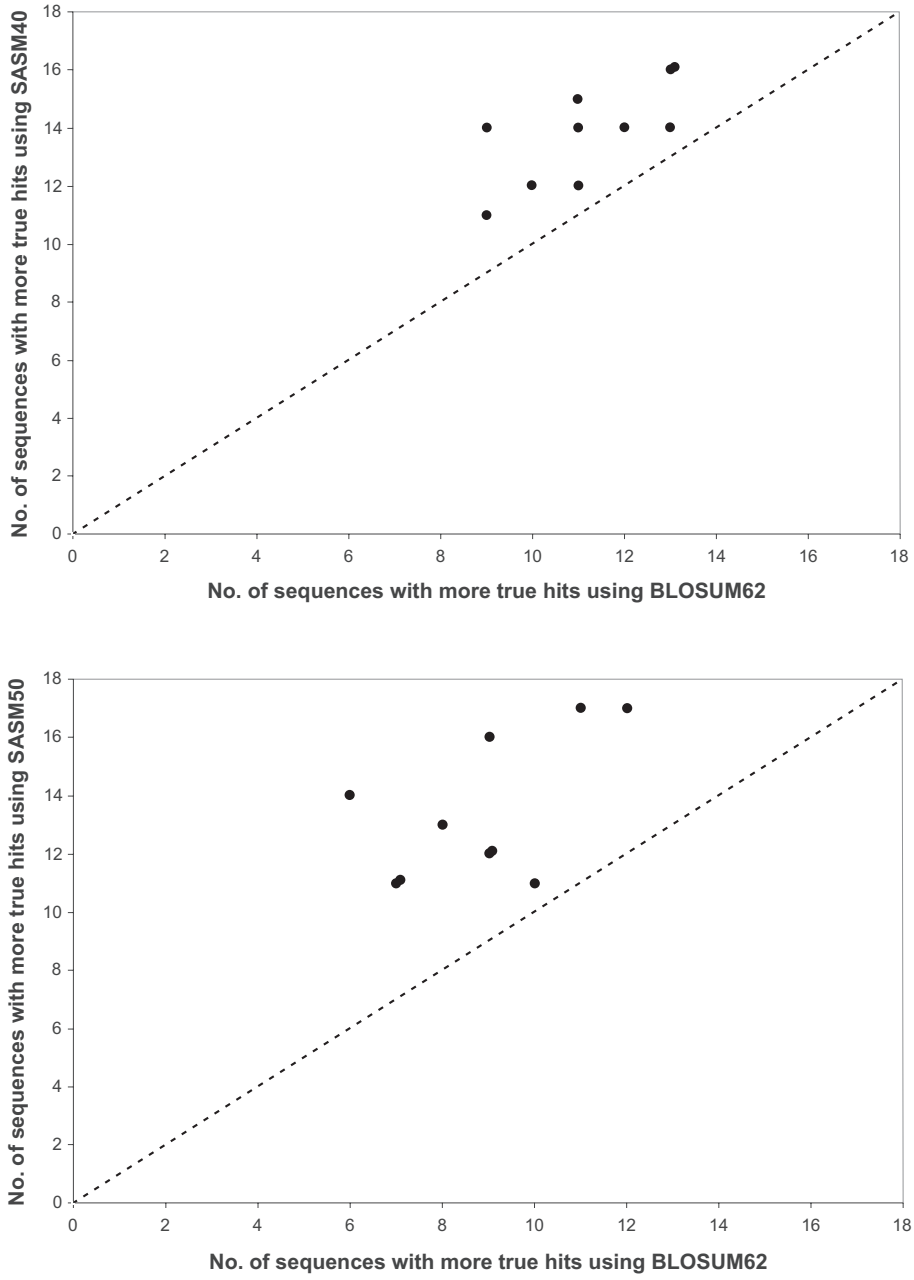

Figure 1 A comparison of the performance of SASM and BLOSUM62 in BLAST. Each data point represents results for 92 query sequences with ASTRAL-SCOP v1.65 (50% sequence identity) as target database, and depicts the numbers of query sequences that had more success (more true hits) with SASM (y-axis) vs.BLOSUM62 (x-axis). For each of the ten data points, the corresponding SASM matrix was computed using a different z-score filter (ten values in the range 3.5–8.0 (Table 1)). If a SASM matrix performed the same as BLOSUM62 the data point would fall on the diagonal, which is indicated by a dashed line. A hit was considered true if it belonged to the same SCOP superfamily as the query.

The ability of a similarity detection method to report homologous sequences (sensitivity) must be balanced against the spurious detection of nonhomologs or false hits (specificity). Receiver operating characteristic (ROC) curves[31] provide a convenient method of indicating the number of true hits for a given number of false hits.[19,26] A comparison of ROC_{50} curves generated from results of PSI-BLAST using BLOSUM62 and the SASM40 matrix isentropic with BLOSUM62 (Table 2) is given in Figure 3. The results show that the latter finds more true hits at all E-value cutoff levels.

Discussion

Accurate alignment of protein sequences is critical in obtaining reliable amino acid substitution frequencies required for computing substitution matrices. This task becomes more challenging as the evolutionary distances between proteins

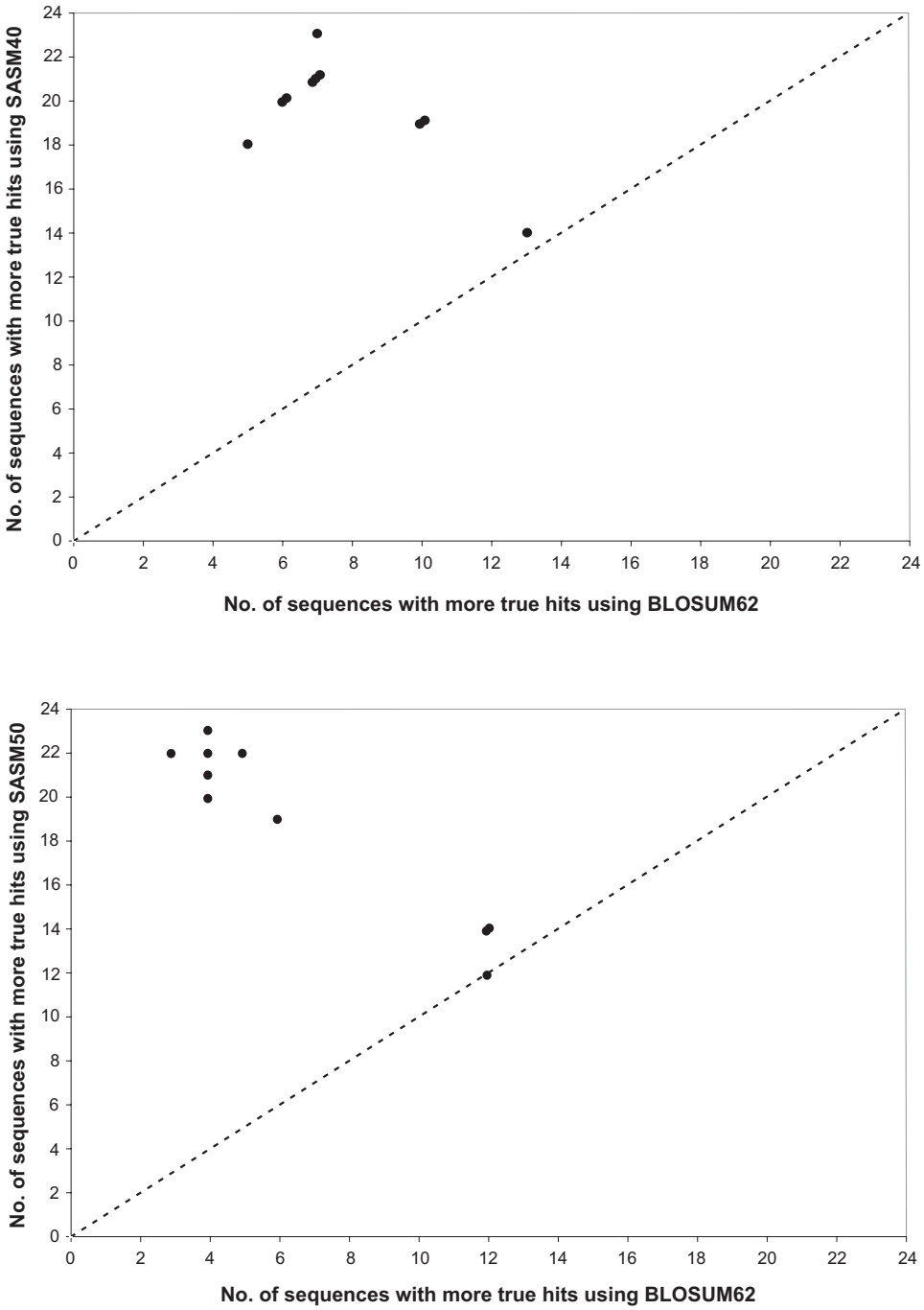

Figure 2 A comparison of the performance of SASM and BLOSUM62 in PSI-BLAST. Each data point represents results for 92 query sequences with ASTRAL-SCOP v1.65 (50% sequence identity) as target database, and depicts the numbers of query sequences that had more success (more true hits) with SASM(yaxis) vs BLOSUM62(x-axis). For each of the ten data points, the corresponding SASM matrix was computed using a different z-score filter (ten values in the range 3.5–8.0 (Table 1)). If a SASM matrix performed the same as BLOSUM62 the data point would fall on the diagonal, which is indicated by a dashed line. A hit was considered true if it belonged to the same SCOP superfamily as the query.

increase. Since sequence alignments based on structural alignments have been long considered to be the gold standard, it is reasonable to expect that substitution matrices based on structural alignments can lead to better performance in detecting remote homologs. Efforts at obtaining structure-based substitution matrices have been constrained by the limited amount of solved structures available, when compared with sequence data. For example, two of the more recent structure-based matrices, BC[17] and SDM[13] were computed from the alignments of protein pairs that numbered in the hundreds. These matrices have been useful in generating improved sequence alignments. However, the BLOSUM matrices have

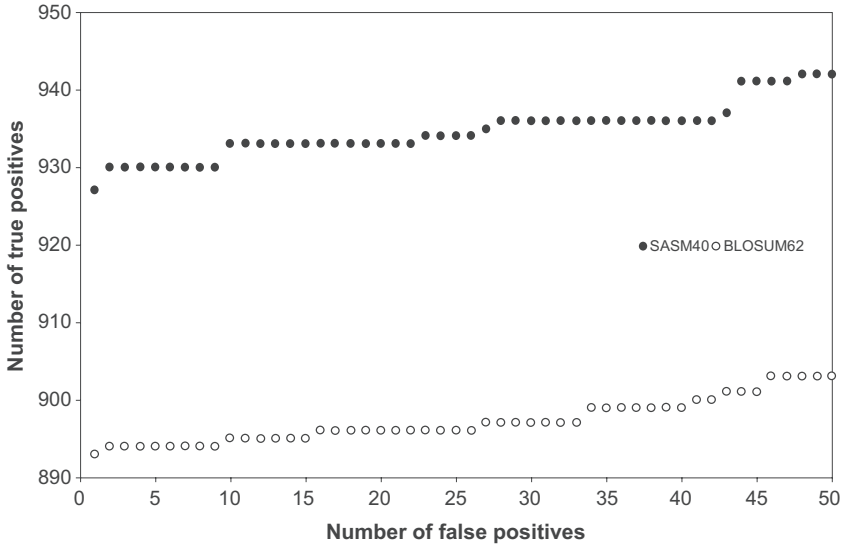

Figure 3 ROC_{50} curves computed from pooled results for a set of 92 query sequences with ASTRAL-SCOP v1.65 (50% sequence identity) as target database. Results are shown for PSI-BLAST with SASM40 (z score filter 6.5) (solid circles) and BLOSUM62 (open circles).

been shown to be superior in detecting remote homologs.[19] It has therefore remained an open question as to whether structure-based matrices could prove useful in the detection of remote homologs. In this study, we have shown that structure-based matrices computed using the expanded set of protein structures now available can detect a greater number of homologs in the popular homology detection programs BLAST and PSI-BLAST, when compared to BLOSUM62.

The large set of structurally aligned protein pairs used in this study (Table 1) were selected from an even larger set of structurally superposed proteins pairs (see Methods), and the amount of data clearly precluded a manual examination of structural superpositions as a basis for selection. The criteria that were developed to automate the selection process includes the use of a z-score filter (see Methods). The superiority of the structure-based SASM matrices in detecting remote homologs is relatively insensitive to the value of the z-score filter used, in the range between 3.5 and 8.0 (Figures 1 and 2). This is somewhat remarkable, given that the number of protein domain pairs selected changes by a factor of three in this range (Table 1). When the z-score filter value is further decreased (<3.5) we have anecdotal evidence showing that structural alignments in the beta sheet regions may, in some cases, be poor, which will lead to errors in pair-wise frequency counts. These results also suggest that the matrix elements themselves may be relatively insensitive to future increases in the size of the protein structure database.

The selection of structurally aligned protein domain pairs, which are presumed to be homologous, was based solely on the application of the selection criteria described in methods. This selection procedure is also supported by the observation that most protein domain pairs thus selected were related by SCOP classification.

There is an expectation that the optimal set of frequencies utilized to compute a substitution matrix, are in the words of Karlin and colleagues, "simply those found in the sort of region we seek to identify."[32] Our results are consistent with this expectation, since SASM50 performs better than SASM60 (or BLOSUM62) when using a set of query sequences which had, at most, 50% sequence identity to target sequences. In the case of sequence-based BLOSUM matrices, BLOSUM62 is often preferred to BLOSUM45 in detecting remote homologs. This discrepancy may be due to the difficulties associated with aligning sequences at greater evolutionary distances, in the absence of structural information.

Acknowledgments

We thank Dr. Stephen Altschul for providing the program that calculates the statistical parameter values for amino acid substitution matrices. Nalin CW Goonesekere received funding through a Summer Fellowship Award from UNI. The author reports no conflicts of interest in this work.

References

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–410.
2. Altschul SF Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25(17):3389–3402.
3. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*. 1988;85(8):2444–2448.

4. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*. 1998;14(10):846–856.
5. Yona G, Levitt M. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*. 2002;315(5):1257–1275.
6. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile – profile sequence alignments. *Nucleic Acids Res*. 2005;33(Web Server issue):W284–W288.
7. Fitch WM. An improved method of testing for evolutionary homology. *J Mol Biol*. 1966;16(1):9–16.
8. Grantham R. Amino acid difference formula to help explain protein evolution. *Science*. 1974;185(4154):862–864.
9. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. *Science*. 1992;256(5062):1443–1445.
10. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. Matrices for detecting distant relationships. *Atlas of Protein Sequence and Structure.* 1978;5:345–358.
11. 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(4):816–831.
12. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992;89(22):10915–10919.
13. Prlic A, Domingues FS, Sippl MJ. Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng*. 2000;13(8):545–550.
14. Kim C, Lee B. Accuracy of structure-based sequence alignment of automatic methods. *BMC Bioinformatics*. 2007;8:355.
15. Risler JL, Delorme MO, Delacroix H, Henaut A. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J Mol Biol*. 1988;204(4):1019–1029.
16. Johnson MS, Overington JP. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J Mol Biol*. 1993;233(4):716–738.
17. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. *J Mol Biol*. 2001;307(2):721–735.
18. Liu X, Zheng WM. An amino acid substitution matrix for protein conformation identification. *J Bioinform Comput Biol*. 2006;4(3):769–782.
19. Green RE, Brenner SE. Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. *Proceedings of the IEEE.* 2002;9:1834–1847.
20. Price GA, Crooks GE, Green RE, Brenner SE. Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap. *Bioinformatics*. 2005;21(20):3824–3831.
21. Burley SK, Joachimiak A, Montelione GT, Wilson IA. Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers. *Structure*. 2008;16(1):5–11.
22. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*. 2000;28(1):254–256.
23. Jung J, Lee B. Protein structure alignment using environmental profiles. *Protein Eng*. 2000;13(8):535–543.
24. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol*. 1991;219(3):555–565.
25. Park J, Karplus K, Barrett C, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol*. 1998;284(4):1201–1210.
26. Schaffer AA, Aravind L, Madden TL, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* 2001;29(14):2994–3005.
27. 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(4):536–540.
28. Gurler A, Knapp EW. Evaluation of sequence alignments of distantly related sequence pairs with respect to structural similarity. *Genome Inform*. 2007;18:183–191.
29. Kinch LN, Grishin NV. Evolution of protein structures and functions. *Curr Opin Struct Biol*. 2002;12(3):400–408.
30. Altschul SF. Generalized affine gap costs for protein sequence alignment. *Proteins*. 1998;32(1):88–96.
31. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput Chem*. 1996;20:25–33.
32. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*. 1990;87(6):2264–2268.