



## Local structure-based sequence profile database for local and global protein structure predictions

An-Suei Yang\* and Lu-yong Wang

Department of Pharmacology and Columbia Genome Center, Columbia University,  
630 West 168th street, PH 7 W Room 318, New York, NY 10032, USA

Received on January 2, 2002; revised on April 10, 2002; accepted on May 24, 2002

### ABSTRACT

**Motivation:** A large body of evidence suggests that protein structural information is frequently encoded in local sequences—sequence–structure relationships derived from local structure/sequence analyses could significantly enhance the capacities of protein structure prediction methods. In this paper, the prediction capacity of a database (LSBSP2) that organizes local sequence–structure relationships encoded in local structures with two consecutive secondary structure elements is tested with two computational procedures for protein structure prediction. The goal is twofold: to test the folding hypothesis that local structures are determined by local sequences, and to enhance our capacity in predicting protein structures from their amino acid sequences.

**Results:** The LSBSP2 database contains a large set of sequence profiles derived from exhaustive pair-wise structural alignments for local structures with two consecutive secondary structure elements. One computational procedure makes use of the PSI-BLAST alignment program to predict local structures for testing sequence fragments by matching the testing sequence fragments onto the sequence profiles in the LSBSP2 database. The results show that 54% of the test sequence fragments were predicted with local structures that match closely with their native local structures. The other computational procedure is a filter system that is capable of removing false positives as possible from a set of PSI-BLAST hits. An assessment with a large set of non-redundant protein structures shows that the PSI-BLAST + filter system improves the prediction specificity by up to two-fold over the prediction specificity of the PSI-BLAST program for distantly related protein pairs. Tests with the two computational procedures above demonstrate that local sequence–structure relationships can indeed enhance our capacity in protein structure prediction. The results also indicate that local sequences encoded with strong local structure propensities play an important role in determining the native state folding topology.

**Availability:** All the computational and assessment procedures have been implemented in the integrated computational system PrISM.1 (Protein Informatics System for Modeling). The system and associated databases for LINUX systems can be downloaded from the website: [www.columbia.edu/~ay1/](http://www.columbia.edu/~ay1/).

**Contact:** [ay1@columbia.edu](mailto:ay1@columbia.edu)

### INTRODUCTION

The hypothesis that local structural information is frequently encoded in local sequences independent of global folding topology is supported by a large body of experimental and theoretical evidence (Baldwin and Rose, 1999a,b). These observations are in agreement with the hierarchical folding model, which suggests that a protein must fold through a sequential process by assembling pre-formed local structures into its native structure. According to this folding model, the structural determinants should reside in and on the surface of the local structures that remain as an integral part of the native structure, and thus, at least some of the local structures can be recognized from local sequence fragments, independent of the global folding topology. In this paper, computational procedures that predict protein structures by matching a query sequence through a local structure-based sequence profile database (LSBSP2) were devised to test this hypothesis. The goal is to enhance the capacities of computational procedures in protein structure prediction.

The basic principle in constructing the LSBSP2 database follows the line of evidence that some connecting loops flanked by two secondary structure elements are very specific in sequence–structure relationship (Aurora and Rose, 1998; Aurora *et al.*, 1994; Efimov, 1993; Hutchinson and Thornton, 1994; Rooman *et al.*, 1989; Rose *et al.*, 1985; Sibanda *et al.*, 1989; Sibanda and Thornton, 1985; Wilmot and Thornton, 1988; Yang *et al.*, 1996). The LSBSP2 database is comprised of a large number of local structure-based sequence profiles, which were derived from exhaustive pair-wise structural alignments for all local structures with two consecutive

\*To whom correspondence should be addressed.

secondary structure elements. The following section shows in detail the construction of the LSBSP2 database and the two computational procedures devised to test the prediction capacity of the LSBSP2 database.

## METHODS

### Local structure-based sequence profile database LSBSP2

To construct the local structure-based sequence profiles (LSBSP) for a seed protein, the seed protein structure is first parsed into local structures, each of which contains a loop region flanked by two consecutive secondary structure elements. The local structures are then used as probes to search for structural analogues in the local structure database (LSD2). The LSD2 in PrISM.1 was constructed by collecting contiguous protein sub-structures, each of which contains two consecutive secondary structure elements connected by a loop region, from a set of non-redundant structures in the protein data bank (PDB). DSSP (Kabsch and Sander, 1983) was used to define the secondary structure elements. The list of the non-redundant protein structures was obtained from PDB\_SELECT25 [ftp://ftp.embl-heidelberg.de/pub/databases/pdb\\_select](ftp://ftp.embl-heidelberg.de/pub/databases/pdb_select) (Feb/2001) with pair-wise sequence identities less than 25%. PDB entries with only C $\alpha$  traces or with less than 40 residues in sequence length were eliminated from the protein list. The LSD2 contains a total of 15 192 overlapping local structures.

For each of the local structures from the seed protein structure, the one-against-all pair-wise structural alignments over the LSD2 database (15 192 total) are sorted based on the PSD measure. PSD (protein structural distance) is a normalized structural similarity score based on similarities in secondary structure elements and the 3D arrangements of the amino acid residues; PSD for a pair of identical structures is zero, and the PSD increases with decreasing structural similarity (Yang and Honig, 2000a). In this work, the top 50 ranked local structures from the LSD2 with the lowest PSD are used to produce the sequence profile based on the pair-wise structural alignments with the computational procedures implemented in PrISM.1 (Yang and Honig, 2000b). It can be shown that, in general, the information content in the LSBSP begins to saturate with more than 50 local structures.

The structure-based sequence profiles from each of the seed local structures are then spliced to form a LSBSP for the seed protein structure. For the secondary structure element where two local structures overlap, the sequence profile is linked with two halves of the local sequence profiles from each of the two local structures for the overlapped secondary structure element. The reason to splice the local sequence profiles is to form a sequence profile for the contiguous sequence of the seed protein

structure such that sequence fragments of any sequence length can be used as a query sequence to match to the sequence profile. The procedure continues until the LSBSP of the seed structure is completed. A figure that summarizes the procedure above is available from the authors' ftp site (<ftp://ps7ayang.cpmc.columbia.edu/pub/LSBSP.pdf>).

The procedure above was applied to all the structures in a set of non-redundant structural domains (1779 total, see below) to form a database of 1779 LSBSP. This is the LSBSP2 database in PrISM.1.

### Protein structural domain assignment procedure

Structural domain boundaries of the structures in the non-redundant protein list PDB\_SELECT25 (version Feb/2001) were defined with a novel method implemented in PrISM.1 to ensure that all the structural domains are contiguous in sequence—Protein structural domains with non-sequential parts of the proteins are problematic for constructing the LSBSP2 database.

Figure 1 summarizes the PrISM.1 automatic structural domain assignment procedure. Overall, the parsing procedure assigns structural domains with the constraints that (1) the domain boundaries are located at loop regions, (2) the sequence in a domain is contiguous, (3) the  $\beta$ -sheets in a domain are intact, and (4) mutually packed  $\beta$ -sheets stay in the same domain.  $\alpha$ -helices between two domains are included in both domains, such that, in a few cases, the termini of the domains could overlap for a short piece of contiguous sequence.

The results of the automated domain assignment show that most of the proteins in the PDB\_SELECT25 list (total 1520 chains) are one-domain proteins; only 261 proteins were divided into more than one structural domain. The PrISM.1 domain assignment procedure produced a total of 1779 structural domains from the protein structures in the list of PDB\_SELECT25 (version Feb/2001, 1520 chains). These protein domains do not include PDB entries with only backbone coordinates or with sequence length less than 40 residues. Protein domains with one or zero secondary structure element were also excluded from the structural domain set. Overall, PrISM.1's domain assignments for multi-domain proteins are less likely to be over-parsed comparing with other automatic procedures; the PrISM.1's structural domains are larger in size on average. The distribution of the size of the PrISM.1's structural domains peaks around 120–140 residues and 76.6% of the domains are comprised of less than 200 residues. In comparison, structural domain size distribution generally peaks around 100 residues and 80–90% of the domains are shorter than 200 residues (Jones *et al.*, 1998).

The test results of the automated domain assignment procedure on a test set of 229 proteins compiled in the work of Siddiqui and Barton (Siddiqui and Barton,

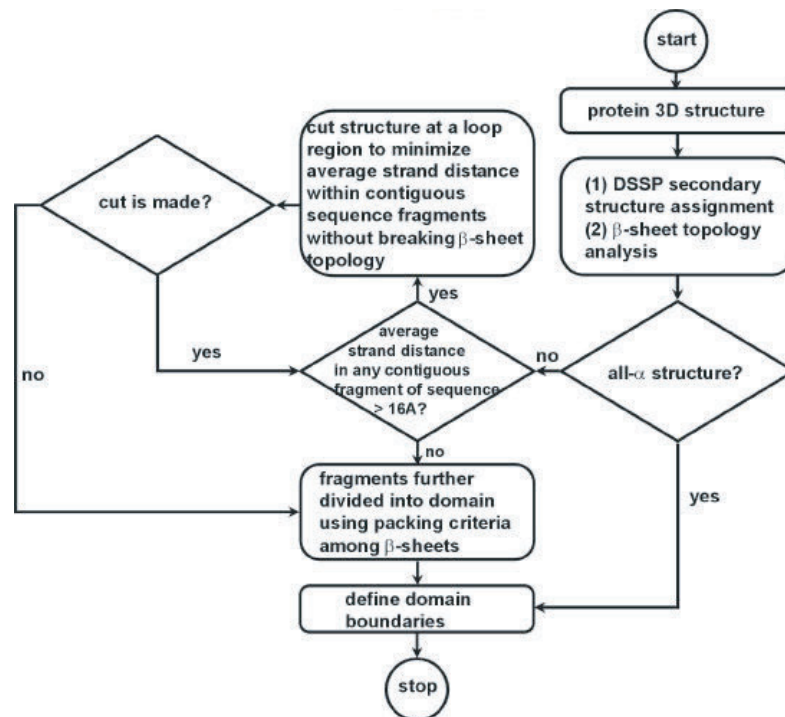


Fig. 1. Flow chart for the PrIMS.1 domain assignment procedure.

1995) have demonstrated that 76% of the proteins in the test set were parsed into domains that are in good agreement with the references based on the standards and assessing procedures in the work of Siddiqui and Barton (Siddiqui and Barton, 1995), and 12% of the proteins in the test set were parsed into domains that are completely in disagreement with the references. The detailed benchmark results and the comparison of the results from several published domain assignment procedures for the test set of 229 proteins are available from the authors' ftp site (<ftp://ps7ayang.cpmc.columbia.edu/pub/domainAssignment.pdf>). This accuracy rate is in the comparable range of the accuracy rates from several automatic procedures that are not restricted to assigning structural domains with contiguous sequences (Jones *et al.*, 1998).

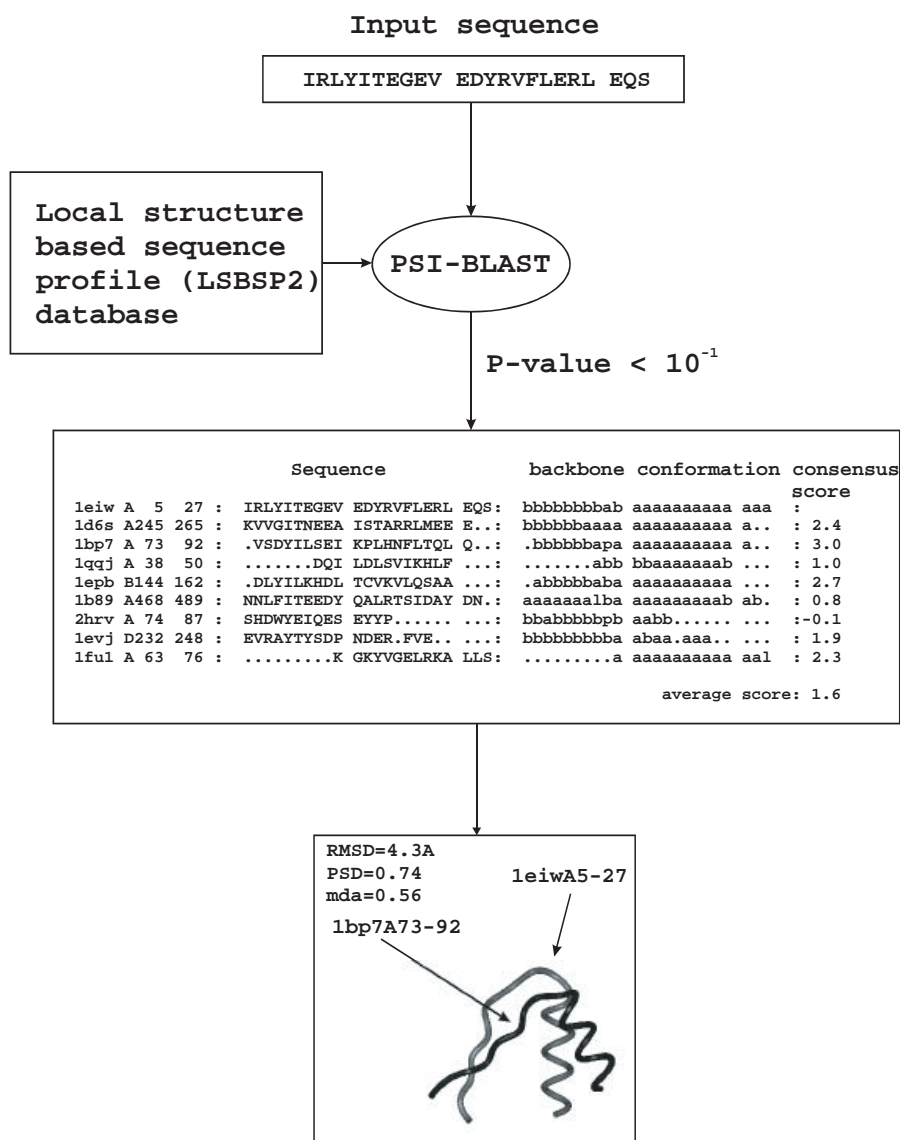
### Local structure prediction with LSBSP2 and PSI-BLAST

In principle, the LSBSP2 database contains the sequence–structure relationships for all local structures with two consecutive secondary structure elements. One important application is to predict the structure of a local sequence with the sequence–structure relationships in the database.

PSI-BLAST (Altschul *et al.*, 1997) makes use of position-specific scoring matrix (PSSM) derived from

a sequence profile to align a sequence to the sequence profile. An option to run the PSI-BLAST program is to 'jump start' the alignment procedure with a multiple alignment computed outside PSI-BLAST (-B option in the PSI-BLAST program) (Aravind and Ponting, 1998); PSI-BLAST calculates a PSSM from the multiple alignment and aligns the query sequence to the multiple alignment based on the PSSM.

In the local structure prediction procedure outlined in Figure 2, each of the local structure-based sequence profiles in the LSBSP2 database is used to jump start PSI-BLAST to align the sequence profile to the query sequence and to calculate the normalized  $E$ -value (and thus,  $p$ -value, i.e. the  $E$ -value for matching the query to a sequence in a database that contains only one sequence is equivalent to the  $p$ -value of the match—the probability for a random match to have normalized alignment score above a threshold) for the alignment. BLOSUM45 was used as the substitution matrix and the default parameters 11( $G$ ) and 3( $E$ ) were used for gap penalties. All the PSI-BLAST alignments with  $p$ -value  $< 10^{-1}$  are assembled in the multi-alignment as shown in Figure 2, where the first row of the multi-alignment is the input query sequence and the rows following the query sequence are fragments of structural domain sequences (PDB code names and residue ranges are shown), of which the sequence profiles



**Fig. 2.** Flow chart for the local structure prediction method with the LSBSP2 database. See text for discussion.

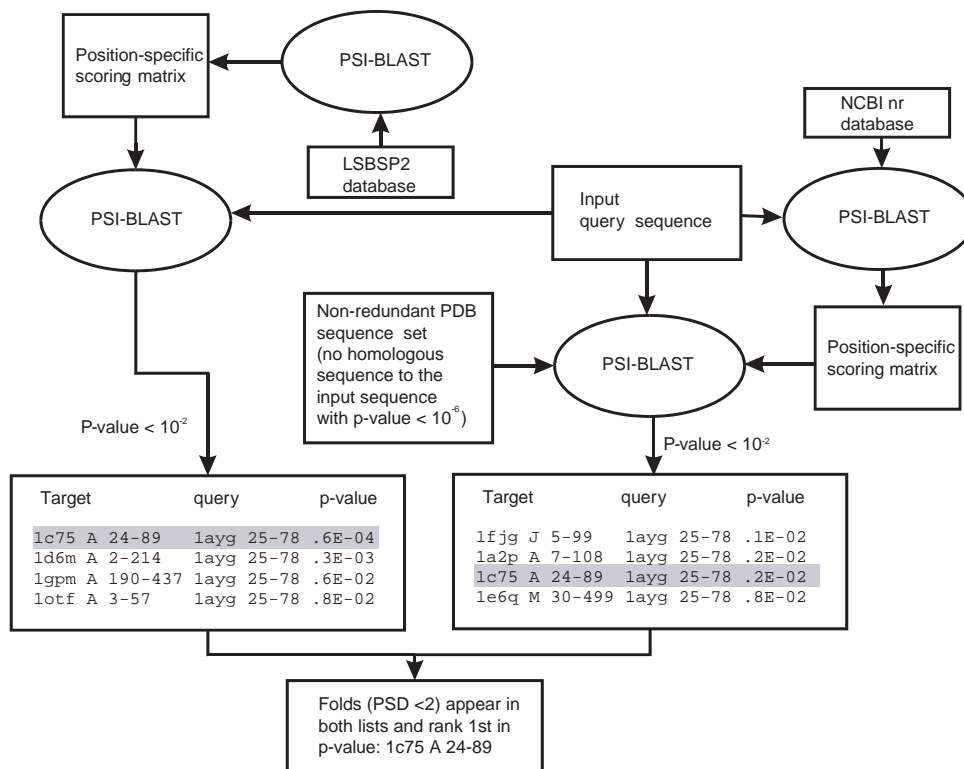
in the LSBSP2 database match to the query sequence with the  $p$ -value below the threshold  $10^{-1}$ . This is an empirical threshold to include some of the true-positive hits.

As shown in the example, not every structural fragment below the  $p$ -value threshold is true positive; a second tier scoring procedure has been implemented to single out the predicted structure. This is done by comparing the backbone conformation of each of the hits to the average backbone conformation. As shown in Figure 2, the backbone conformation of each of the residues in the multi-alignment is shown in the right-hand part of the multi-alignment. The characters (a, b, p, l...) representing the ranges of  $\phi$ - and  $\psi$ -angles have been

defined by Oliva *et al.* (1997). For example, ‘a’ represents the  $\alpha$ -region with  $-140^\circ < \phi < -20^\circ$ ,  $-100^\circ < \psi < 20^\circ$  plus  $-180^\circ < \phi < -60^\circ$ ,  $20^\circ < \psi < 60^\circ$ , and ‘b’ represents the  $\beta$ -region with  $-180^\circ < \phi < -60^\circ$ ,  $60^\circ < \psi < 180^\circ$ . For the sequence  $k$  ( $k = 2 \sim m$ , where  $m$  is number of the sequences in the multi-alignment), the consensus score is calculated with the following equation:

$$\text{consensus score}(k) = \frac{1}{n} \sum_{i=1, n} \sum_{j=2-m, j \neq k} M(q_{k,j}, q_{j,i}) \quad (1)$$

where  $n$  is the length of the multi-alignment, and  $q_{i,j}$  is the backbone conformation, as defined by the ranges of



**Fig. 3.** Flow chart depicting the filter system that singles out the predicted structural template from a list of PSI-BLAST hits. See text for discussion.

$\phi$ - and  $\psi$ -angle (see above), of residue  $j$  in sequence  $i$ . Note that  $q_{i,j}$  includes the empty spaces designated by dots in the multi-alignment.  $M(q_{k,i}, q_{j,i})$  in the above equation equals to 1 only when  $q_{k,i}$  and  $q_{j,i}$  are the same conformation, otherwise,  $M(q_{k,i}, q_{j,i}) = -1$ . In Figure 2, the consensus scores are shown on the right-hand side of the multi-alignment. The hit structure with the largest consensus score is the closest to the consensus structure, and thus is the predicted structure for the input sequence (1bp7A73–92, in the example shown in Figure 2). The average score reflects the level of convergence of the hit structures to the consensus structure. Empirically, predictions with hits below the  $p$ -value threshold (0.1) and with the consensus scores at least 1.5 and average consensus score above 0 show the best balance in sensitivity and specificity.

### The PSI-BLAST+filter system

The other computation procedure that makes use of the LSBSP2 database for protein structure prediction was devised to identify distantly related protein pairs that share a similar structure (see Figure 3). The key process of the procedure, as shown in the left-hand side of Figure 3, is to match the query sequence throughout all the LSBSP

(local structure-based sequence profiles) in the LSBSP2 database. The alignment of the query sequence to a LSBSP is carried out by feeding the LSBSP to PSI-BLAST with the -B option in BLASTPGP program to calculate the position-specific scoring matrix (PSSM), which is then used to calculate the alignment and the  $p$ -value of the match. Given the  $p$ -value cutoff of  $10^{-2}$ , as shown in the left-hand side of Figure 3, a list of protein structural domains for which the LSBSP match the query sequence with the  $p$ -values below the cutoff are generated (see the left-hand side list in Figure 3). As shown in the previous section, these protein structural domains could contain many of the local structures that are similar to the local structures predicted for the query sequence. The rationale that underlies the filter procedure is that a hit protein sequence from the PDB database selected with a sequence comparison method must be a false positive if the structure of the hit protein is not similar in structure to any of the proteins that are expected to contain some of the local structures of the query sequence. This rationale is assessed by the procedures in the following paragraph.

A total of 1779 sequences of the non-redundant protein structural domains with contiguous sequences are used as the testing set of query sequences. For each of the test

sequences, the search space is the sequence database that contains all the proteins in the PDB\_SELECT25 (version Feb/2001) list except for the sequences that are related to the test sequence with the  $p$ -value less than  $10^{-6}$  (or sequence identity above 18% on average). The goal is to search for a true structural template from a set of protein structures for which the sequences are not related to the test sequence based on the  $p$ -value cutoff. The sequence search procedure is a standard PSI-BLAST run using the test sequence as the seed to construct a position-specific scoring matrix (PSSM) with four iterations over the NCBI nr sequence database with default parameters. The PSSM is then used to align the test sequence to each of the sequences in the search space (see the right-hand side of Figure 3). The hit sequences with the  $p$ -value below the threshold  $10^{-2}$  as shown in Figure 3 (right-hand side list) are compared with the structures (left-hand side list) that are predicted to contain some of the local structures of the test sequence. The structure in the hit sequence list (highlighted in gray in the right-hand side list) that has the lowest  $p$ -value and is similar in structure to at least one of the structures in the left-hand side list in Figure 3 with  $\text{PSD} < 2$  is singled out as the structural template for the test sequence.  $\text{PSD} < 2$  is an empirical threshold that distinguishes protein structure pairs at the SCOP superfamily level (Yang and Honig, 2000a). The benchmark results for this and the computational procedure above are shown in the following section.

## RESULTS AND DISCUSSION

### Local structure predictions with the LSBSP2 database

In an assessment of the prediction method shown in Figure 2, a testing set of local sequence fragments were derived from a newer set of non-redundant proteins from the PDB\_SELECT25 (version Sep/2001) list: 250 new non-redundant protein structures that have not been included in the PDB\_SELECT25 (version Feb/2001) were parsed into 2381 sequence fragments with the same procedure in deriving the LSD2 database (see Methods). The testing proteins are not related to any seed proteins in the LSBSP2 database with the  $p$ -value threshold of  $10^{-6}$  (pair-wise sequence identity  $\sim 18\%$ ). Each of the local sequences in the testing set was used as the input query sequence to evaluate the specificity and sensitivity of the prediction procedure with the LSBSP2 database.

Out of a total of 2381 input sequences, 1151 local sequences had predicted structures based on the criteria:  $p$ -value  $< 0.1$ , consensus scores  $> 1.5$ , and average consensus score  $> 0$  (see the Method Section). The predicted structures were then compared with the known local structures of the input sequences to assess the prediction method. Two different structure similarity measures were

used: protein structure distance (PSD) calculated with the PrISM.1 structural alignment procedure (Yang and Honig, 2000a), and the Bystroff–Baker mda measure (Bystroff and Baker, 1998) where a window of eight residues are considered to be correctly predicted when none of the backbone torsion angles in the eight-residue fragment of the predicted structure differ by more than  $120^\circ$  from the corresponding region of the target structure based on the predicted alignment. In the bottom of Figure 2, the predicted structure (colored in black) and the target structure (colored in gray) are superimposed and the structure similarity measures are also shown. Following Bystroff and Baker (Bystroff and Baker, 1998), the mda value is the fraction of residues that are considered to be correctly predicted in at least one of the overlapping eight-residue windows.

The results show that 63% of the target-prediction pairs are similar in structure with  $\text{PSD} < 1$ ; that is, the optimized RMSD is less than  $\sim 4 \text{ \AA}$  and the secondary structure element type is correctly predicted, and that 59% of the target structures are correctly predicted with mda value greater than 0.4. Both results indicate that more than half of the predicted local structures were reasonably similar to the target structures. This accuracy level is comparable to the local structure prediction method based on the I-sites library (Bystroff and Baker, 1998), which is the only local structure prediction method that has been tested with a large set of testing sequences.

In a testing procedure for the local structure prediction method based on the I-sites library, local structures of 75% of the total 15 919 residues in a testing set of 55 proteins were predicted at confidence level above 40%; for the local sequences with predictions, 54% residues were predicted correctly based on the mda measure (see above) (Bystroff and Baker, 1998). In the benchmark procedure above, the local structures of 17 848 residues were predicted from a total 31 604 residues, and 53.5% of the 17 848 residues with predictions were correctly predicted based on the same Bystroff–Baker mda measure. This accuracy level (53.5%) is comparable to the accuracy level (54%) of the I-sites library method at  $>40\%$  level of confidence. However, the sensitivity level for the benchmark results above is 56.5% (17 848 out of total 31 604 residues), which is less than the 75% sensitivity level reported for the Bystroff–Baker method at  $>40\%$  confidence level (Bystroff and Baker, 1998).

The interpretation of the comparison in the previous paragraph is not straightforward because the testing sets of proteins and the prediction procedures are not at the same level of difficulty in local structure prediction. Specifically, the prediction method in this work is aimed at predicting longer local sequence fragments with two consecutive secondary structure elements, for which the average sequence length are about threefold longer in

**Table 1.** Comparison of the predictions of PSI-BLAST with the predictions of the PSI-BLAST+filter system

	PSI-BLAST			PSI-BLAST+filter			
	Number of predictions	Correct predictions (rank1 in $p$ -value with PSD<2)	Possible correct predictions (PSD<2 in any rank)	Number of predictions	Correct predictions (rank1 in $p$ -value with PSD<2)	Number of negative predictions for no hit	Correct negative predictions (PSD > 2 in any rank)
$p$ -value <10 <sup>-5</sup>	141	89	92	121	92		
$p$ -value <10 <sup>-4</sup>	376	194	215	240	171		
$p$ -value <10 <sup>-3</sup>	1022	283	364	441	249		
$p$ -value <10 <sup>-2</sup>	1629	330	600	617	289	1012	728

sequence than the predictions based on the I-sites library (Byströff and Baker, 1998). Because the mda measure is more tolerant for shorter sequences than for longer ones, it is difficult to compare the two local structure prediction methods in sensitivity and specificity at an even level. Moreover, because of the nature that the local sequence fragments in the I-sites library were determined by comparing the sequence profiles from various protein families, the boundaries of the I-sites library fragments are not correlated with any structural boundaries of the local structures and the I-sites library covered only the local structures that share conserved sequence features in different protein families. By contrast, the LSBSP2 database was constructed with structure-based alignments for all local structures with two consecutive secondary structure elements, and thus, the LSBSP2 database is expected to contain more general sequence–structure relationships for local structures.

### The PSI-BLAST+filter system

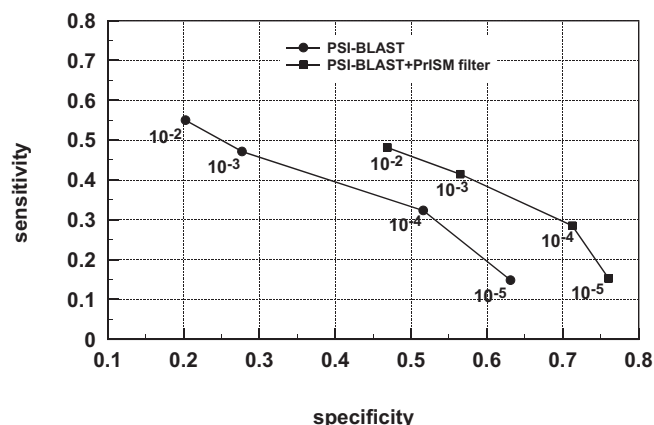
Table 1 summarizes the assessment results of the PSI-BLAST+filter procedure. With the  $p$ -value cutoff of 10<sup>-2</sup> (see the first column of Table 1), 1629 test sequences out of the total of 1779 have at least one hit from the PSI-BLAST sequence search procedure (see the second column of Table 1). Only 330 test sequences (specificity = 20%) have a true structural template that appears at the top of the hits with the lowest  $p$ -value (see the third column of Table 1). A true structural template is similar in structure to that of the testing sequence with PSD <2. Overall, a total of 600 test sequences have at least one true structural template with  $p$ -value less than 10<sup>-2</sup> (see the fourth column of the Table). In combination with the filter procedure, 617 test sequences are predicted to have a true structural template (see the fifth column of Table 1), and 289 of these test sequences (specificity = 48%) have a true structural template at the top of the hits with the lowest  $p$ -value (see the sixth column). This shows that the filter procedure has improved the specificity of the sequence search method by more than twofold. The reason for the

improvement is that most of the test sequences that do not have any true structural template in the database below the  $p$ -value cutoff are predicted to have no hit based on the filter procedure. Indeed, as shown in the seventh column of Table 1, a total of 1012 test sequences are predicted to have no true structural template and 728 of these sequences (72%) do not have any true structural template in the sequence database with  $p$ -value cutoff of 10<sup>-2</sup> (see the eighth column). Results with different  $p$ -value threshold are also shown in Table 1.

Results shown in Table 1 are further summarized in the specificity–sensitivity plot (Lindahl and Elofsson, 2000) in Figure 4. The sensitivity is the ratio of the correct predictions over the maximum possible correct predictions, and the specificity is ratio of the correct predictions over the total predictions (Lindahl and Elofsson, 2000). The numbers of total predictions are shown in the second and fifth column in Table 1 for PSI-BLAST and PSI-BLAST+filter procedure respectively, while the numbers of correct predictions are shown in the third and the sixth column respectively for the two procedures. The specificity–sensitivity plot shows that the PSI-BLAST+filter procedure has higher specificity by 15–20% than that of the standard PSI-BLAST procedure at the same level of sensitivity. The improvement is increasingly significant as the  $p$ -value threshold increases—the specificity is increased by more than twofold at the high end of the  $p$ -value threshold.

### CONCLUSIONS

Two benchmark results from the computational procedures using the LSBSP2 database to predict local and global structures from protein sequences have supported the novel applications of the LSBSP2 database in protein structure predictions—the implication is that protein structural information is encoded to an extent in the local sequences independent to the long-range interactions in the native state folding topology.



**Fig. 4.** The specificity–sensitivity plot for the comparison of the assessment results in detecting true structural templates with PSI-BLAST (solid circles) and PSI-BLAST+filter (solid squares). The  $p$ -value thresholds are shown by the data points.

## ACKNOWLEDGEMENT

This work was supported by the William J. Matheson foundation.

## REFERENCES

- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Aravind,L. and Ponting,C.P. (1998) Homologues of 26S proteasome subunits are regulators of transcription and translation. *Protein Sci.*, **7**, 1250–1254.
- Aurora,R. and Rose,G.D. (1998) Helix capping. *Protein Sci.*, **7**, 21–38.
- Aurora,R., Srinivasan,R. and Rose,G.D. (1994) Rules for alpha-helix termination by glycine [published erratum appears in *Science* 1994 Jun 24;264(5167):1831] [see comments]. *Science*, **264**, 1126–1130.
- Baldwin,R.L. and Rose,G.D. (1999a) Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.*, **24**, 26–33.
- Baldwin,R.L. and Rose,G.D. (1999b) Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.*, **24**, 77–83.
- Bystroff,C. and Baker,D. (1998) Prediction of local structure in proteins using a library of sequence–structure motifs. *J. Mol. Biol.*, **281**, 565–577.
- Efimov,A.V. (1993) Standard structures in proteins. *Progress in Biophysics and Molecular Biology*, **60**, 201–239.
- Hutchinson,E.G. and Thornton,J.M. (1994) A revised set of potentials for beta-turn formation in proteins. *Protein Sci.*, **3**, 2207–2216.
- Jones,S., Stewart,M., Michie,A., Swindells,M.B., Orengo,C. and Thornton,J.M. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.*, **7**, 233–242.
- Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
- Lindahl,E. and Elofsson,A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.*, **295**, 613–625.
- Oliva,B., Bates,P.A., Querol,E., Aviles,F.X. and Sternberg,M.J. (1997) An automated classification of the structure of protein loops. *J. Mol. Biol.*, **266**, 814–830.
- Rooman,M.J., Wodak,S.J. and Thornton,J.M. (1989) Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. *Protein Eng.*, **3**, 23–27.
- Rose,G.D., Gierasch,L.M. and Smith,J.A. (1985) Turns in peptides and proteins. *Adv. Protein Chem.*, **37**, 1–109.
- Sibanda,B.L., Blundell,T.L. and Thornton,J.M. (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.*, **206**, 759–777.
- Sibanda,B.L. and Thornton,J.M. (1985) Beta-hairpin families in globular proteins. *Nature*, **316**, 170–174.
- Siddiqui,A.S. and Barton,G.J. (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.*, **4**, 872–884.
- Wilmot,C.M. and Thornton,J.M. (1988) Analysis and prediction of the different types of beta-turn in proteins. *J. Mol. Biol.*, **203**, 221–232.
- Yang,A.S., Hitz,B. and Honig,B. (1996) Free energy determinants of secondary structure formation: III. beta-turns and their role in protein folding. *J. Mol. Biol.*, **259**, 873–882.
- Yang,A.S. and Honig,B. (2000a) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structure alignment and quantitative measure for protein structural distance. *J. Mol. Biol.*, **301**, 665–678.
- Yang,A.S. and Honig,B. (2000b) An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. *J. Mol. Biol.*, **301**, 691–712.