Protein Engineering, Design & Selection vol. 18 no. 5 pp. 209–219, 2005
Published online April 21, 2005 doi:10.1093/protein/gzi026

Fold recognition aided by constraints from small angle X-ray scattering data

Wenjun Zheng[1,2,3] and Sebastian Doniach[1]

[1]Departments of Physics and Applied Physics and Laboratory for Advanced Materials, Stanford University, CA 94305 and [2]Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA

[3]To whom correspondence should be addressed.
E-mail: zhengwj@helix.nih.gov

We performed a systematic exploration of the use of structural information derived from small angle X-ray scattering (SAXS) measurements to improve fold recognition. SAXS data provide the Fourier transform of the histogram of atomic pair distances (pair distribution function) for a given protein and hence can serve as a structural constraint on methods used to determine the native conformational fold of the protein. Here we used it to construct a similarity-based fitness score with which to evaluate candidate structures generated by a threading procedure. In order to combine the SAXS scores with the standard energy scores and other 1D profile-based scores used in threading, we made use both of a linear regression method and of a neural network-based technique to obtain optimal combined fitness scores and applied them to the ranking of candidate structures. Our results show that the use of SAXS data with gapless threading significantly improves the performance of fold recognition.

Keywords: fold recognition/linear regression/neural network/small angle X-ray scattering

Introduction

With the explosive increase in DNA and protein sequences resulting from the fast progress of large-scale gene sequencing projects (The Genome International Sequencing Consortium, 2001; Venter *et al*., 2001), the gap between known protein sequences and known structures is widening dramatically. This has led to the establishment of a number of large-scale structural genomics projects (Burley, 2000) for the determination of protein structures with high throughput under the support of the Protein Structure Initiative (PSI; see Stevens *et al*., 2001). The initiative is targeted at the determination of structures of a minimal set of proteins which could putatively exhaust the universe of all protein folds. Once this goal is achieved, it is believed that the task of protein structure prediction given an unknown sequence would be reduced to the selection of the correct fold from a complete fold library, where a generalized fold recognition strategy which exploits maximal information (both sequence-based and structure-based) might be expected to provide an ultimate solution to the sequence–structure mapping problem for soluble proteins.

Fold recognition (see review by Marchler-Bauer and Bryant, 1999) has been a reasonably effective method by which to identify a probable fold from a fold library for an unknown target protein sequence which has no sequence homologue with a known structure. The standard procedure used is to thread the given sequence on to each candidate fold and evaluate the conformational potential energy which is expected to be minimal for the correct fold (potential based threading). Threading may be done either in gapless mode, where all possible gapless alignments of the target sequence with a given candidate fold are examined, or by making use of multiple sequence alignment using gap penalties, to create an optimal alignment (or alignments) for subsequent energy testing (Jones, 1999). Recently, attempts have been made to incorporate more sequence-based structural predictions into the fold recognition protocol (David *et al*., 2000). As an example, a 1D profile consisting of predicted secondary structural assignments and solvent accessibility is employed to do 'prediction based' threading (Rost *et al*., 1997). Sequential information derived from multiple sequence alignment is also helpful in improving the performance of fold recognition (Rykunov *et al*., 2000; Williams *et al*., 2001).

Besides using sequence-based predictions of structural information to supplement potential-based threading, an alternative approach by which to improve standard threading procedures is to exploit additional structural information derived from experiments such as circular dichroism spectroscopy, which are relatively easy to do in comparison with full-scale structural determination (i.e. based on X-ray crystallography or NMR). In this paper we report on the application of small angle X-ray scattering (SAXS) data as a way to impose physical constraints on threading-based protein structure prediction.

SAXS measures X-ray scattering from a protein in a relatively dilute solution. Thus the measurement of SAXS profiles avoids the need to crystallize the protein. SAXS yields physical information about the internal pair distribution of a molecule in its native state. Svergun *et al*. (2002) have shown that, given a SAXS profile that extends to 5 Å resolution, it is possible to reconstruct a map giving approximate 3D locations of all the residues in the protein. Hence, despite limitations in resolution resulting from the orientational averaging of the molecules in solution and from practical signal to-noise ratio limitations resulting from radiation damage effects, we believe this physical information has the potential to reduce false positives which naturally occur in fold identification processes based purely on sequence-based information. Recently, we have for the first time explored the application of SAXS-based physical constraints in improving *ab initio* protein structure prediction (Zheng and Doniach, 2002) and have obtained encouraging results. The present work was motivated by the above preliminary work and was aimed at providing a more comprehensive and in-depth study of this novel method in the context of fold recognition. The following improvements were made compared with the previous work (Zheng and Doniach, 2002): first, instead of an empirical combination of the SAXS-based fitness scores with the other scores, we

attempted more systematic optimizations of the combined scores; second, we tested this method on a significantly larger set of proteins (see Materials and methods).

Following our previous study, we used SAXS-derived structural information to compute a fitness score which evaluates the similarity in SAXS profile between that of the candidate fold (derived computationally from the Cα representation of the protein) and of the target protein (measured experimentally or simulated computationally). Because SAXS measurements are made on an intact protein (or protein fragment), gapped sequence alignments would not be expected to lead to a strong SAXS similarity (since extra or missing residues in the candidate structure would distort the SAXS profile). Therefore, in this paper we use this score as a supplementary constraint for fold identification that is based on a gapless version of the standard potential energy-based threading procedure. We use both a linear regression-based method (LR) and a neural network-based method (NN) to find optimized combinations of a set of fitness scores. Use of explicit optimization allows us to quantify the performance of the fold identification procedure. We find that the use of an optimized score which includes SAXS information leads to results which are significantly better than those obtained by using each individual fitness score separately and are also significantly better than results obtained by using an optimized combined score without including the SAXS information.

Besides providing an improved fold identification method, the present approach can also be used directly to identify domains which are structurally similar to the target. This is achieved by combining a fold library for fold recognition and a domain library for structural similarity identification. This approach potentially has the capability of recognizing structural homologues or analogues for proteins which are not related by significant sequence similarity.

Materials and methods

A flow chart is shown in Figure 1 to summarize the procedure with each step discussed in this section.

Selection of training and test sets of sequences

The protein sequences studied were selected from the list in our previous paper (Zheng and Doniach, 2002) and from the Rosetta test set from Baker's group at the University of Washington (Simons *et al*., 1999a), after excluding those irregular targets without well-defined secondary structures. These lists cover a variety of fold classes (α, β, α/β) with sequence lengths that vary between 31 and 172. In total we use 11 proteins in our training set and 62 proteins in our test set, which marks a significant extension to the set of sequences studied in our previous work (Zheng and Doniach, 2002).

Generating candidate structures by threading to the Dali domain library

In the Dali Domain Classification (Holm and Sander, 1998), each domain is assigned a Domain Classification number DC_lmnp representing the fold space attractor region (l), globular folding topology (m), functional family (n) and sequence family (p). We used the 'Dali Domain Definitions' (v3.01) published by Structural Genomics Group at EMBL-EBI in October 2000, which contains 3689 domains with different numbers of DC_lmnp. Given a target protein, we first exclude all domain entries that share the same DC_lmnp number with it because these sequences bear a ⩾25% sequence identity with the target.

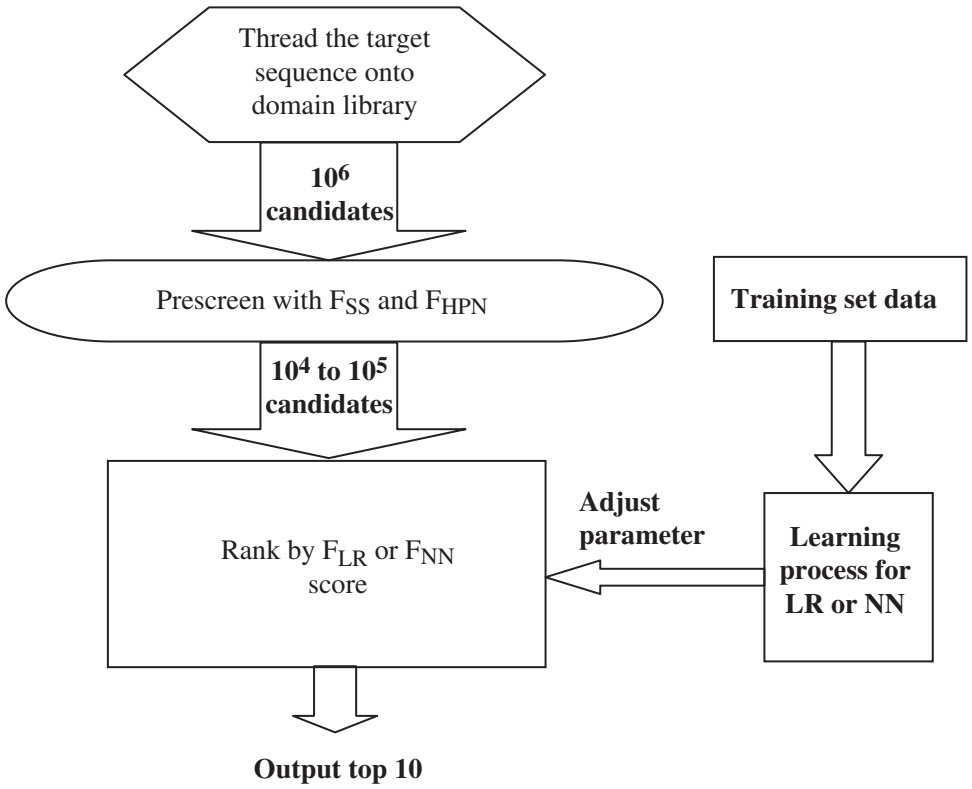

Fig. 1. Flow chart that shows the algorithm of SAXS-aided fold recognition. Each step is described in detail in the text.

Then we continuously thread the target sequence on to each domain which has longer sequence length and discard residues which do not overlap the target. Thus for a domain with length L1 and a target sequence with length L0 (L1 > L0), L1 – L0 + 1 structural candidates are obtained by threading. A continuous (gapless) threading is not expected to give good residue-wise alignment compared with the dynamic programming-based gapped threading but is much more efficient and sufficient to detect the globally correct folds for most targets we study.

Definition of native-like structures

In order to define a measure of the closeness of a candidate structure to the native structure of a target protein, we define a 'native-like' structure as lying in one of three classes, depending on the overall quality of the set of all generated candidates:

(A) A structure with $cRMS_1$ (cRMS of all Cα atoms with respect to the experimental structure, same below) less than 6 Å from the true structure if such structure exists.
(B) A structure with with $cRMS_{0.8}$ (cRMS of 80% of Cα atoms with respect to the experimental structure, same below) less than 5 Å but which fails to satisfy the criterion for (A), if no structure satisfying (A) exists.
(C) A structure with LGA_Q score >1.9 (LGA is a structural comparison tool capable of detecting partial structural similarity which simple cRMS fails to capture; see the subsection Structural alignment for details), but fails to satisfy the criteria for both (A) and (B), if no structure satisfying (A) or (B) exists.

Prescreening

Before doing full-scale structural evaluation, we perform a simple prescreening using the 1D profile consisting of secondary structural assignments (H for α-helix, E for β-strand and X for loop) and HPN-3 letter translation of the sequence (H for hydrophobic, P for polar, N for neutral), where the classification of hydrophobicity follows Huang *et al*. (1995). The secondary structural assignment of both target and candidate fold is obtained by the DSSP program (available at http://www.sander.ebi.ac.uk/dssp/).

The alignment of 1D profile between profile A and profile B is done as follows, where A and B are two sequences of either H/E/X or H/P/N:

Given a residue position i, the score $\mathrm{Align}_{AB}(i, i)$ is 1 (a match) if there exist $j \in [i-1, i+1]$ and $k \in [i-1, i+1]$ so that $A_j = B_k$; otherwise $\mathrm{Align}_{AB}(i, i)$ is 0.

To define F_{SS} and F_{HPN}, we compute the fraction (F) of matches for the whole alignment of 1D profile. We keep structures which satisfy the following criteria: $F_{SS} > 0.6$ and $F_{HPN} > 0.8$.

After prescreening, about 10^4–10^5 candidate structures are kept for further evaluation.

Fitness scores evaluation

We use the following fitness scores to evaluate the candidate structures:

1. Combined hydrophobicity and burial score F_{hpb}. First we define F_{hp} (HP fitness score; see Huang *et al*., 1995) based on the hydrophobic-polar (HP) model which counts pairs of contacts between hydrophobic residues. We define two residues to be in contact if the distance between their Cα atoms is <7 Å and they are not sequential neighbors.

Then we define F_{burial} (burial score; see Huang *et al*., 1995), which measures the extent to which hydrophobic residues are buried inside the core. It is computed by summing the number of residues within a 10 Å distance cutoff from every hydrophobic residue.

Finally, we combine the above two scores as

$$F_{hpb} = \left(F_{hp} - \langle F_{hp} \rangle\right) \times F_{burial} \quad (1)$$

where $\langle F_{hp} \rangle$ is the HP fitness score averaged by sequence permutation.

2. Statistical contact energy F_{stat}. We define the statistical energy as the sum of statistical pairwise contact energy between any two residues in contact based on the 20 × 20 matrix. The pairwise residue–residue interaction energy is calculated based on the frequencies of tertiary contacts in a given PDB structure database. We use the table given in Dima *et al*. (2000), which we have found to work better than the table used in our previous paper (Zheng *et al*., 2002).

3. Radius of gyration F_{Rg}. We define F_{Rg} as the root mean square distance from the center of mass of all Cα atoms along the Cα backbones. This is a useful fitness score for selecting compact structures. Since Rg can be reliably derived from the SAXS data, it is partially overlapping the SAXS score defined later.

4. SAXS fitness score F_{SAXS}. This is defined in the next subsection.

5. 1D profile alignment score: F_{SS}, F_{HPN}. This was defined in the previous subsection.

We make further use of these parameters to construct a combined fitness score in addition to the use in prescreening.

SAXS fitness score evaluation

We adopt the score function used by Walther *et al*. (2000). The profile of scattering intensity associated with a bead model is given as follows using the Debye equation in its pair-distance histogram form:

$$I(s) = N + 2\sum_{i=1}^{n_{bins}} g(r_i)\frac{\sin(2\pi|r_i|s)}{2\pi|r_i|s} \quad (2)$$

where N is the number of beads, s is the scattering vector with $s = k/2\pi$, $g(r_i)$ is the pair-distance histogram of all singly counted pairwise distances and the number of bins is n_{bins}. To represent the $I(s)$ profile, we discretize s with $ds = 0.002$ Å^{-1} and the maximal s is set to 0.12 Å^{-1}. Profiles are normalized to yield $I(0) = 1$. The score function or fitness was computed from

$$F = w(1.0 - r) + RMS \quad (3)$$

with

$$RMS = \sqrt{\sum_i (s_i/s_{max})^m [I_M(s_i) - I_E(s_i)]^2} \quad (4)$$

where r is the cross-correlation coefficient between the two scattering intensity curves (I_M and I_E are the two SAXS profiles computed for the structural model and obtained experimentally, respectively) and w is the weighting factor, chosen to be 10. The term $(s_i/s_{max})^m$ ($m = 3$) adds more weight to differences in the tail of the profile (at higher s values). Smaller value of F corresponds to better fits between the experimental and predicted profiles.

Here we simulate I_E with all-atom bead model whereas I_M is computed based on a Cα atoms only model without explicit consideration of side chain coordinates, assuming side chain atoms sitting at the same coordinate as the Cα atom. This approximation in computing I_M may reduce the performance of the SAXS score; however, it also increases the robustness of our approach, which may tolerate some extent of measurement errors.

Structural alignment

$CRMS_1$ and $CRMS_{0.8}$. We use the standard coordinate RMSD (cRMS) to do structural comparisons between our predicted backbone and the corresponding native Cα backbone (McLachlan, 1971). This is done by superimposing the above two structures on to each other and minimizing the RMS deviation between 100% or 80% of all the residues. We try both the given Cα backbone and its mirror image in the computation of cRMSD and keep the minimum value of cRMS.

LGA. The LGA program was developed by Zemla for structural comparative analysis of two protein structures (Zemla, 2003). We use LGA to search for the largest (not necessarily continuous) set of equivalent residues between a candidate structure and its native structure deviating by no more than DIST = 5 Å. We use the quality score LGA_Q (Zemla, 2003) to assess the structure comparison.

Linear regression

Given a set of N fitness scores F_i ($i = 1, 2, \ldots, N$), we determine a linearly weighted sum of them (F_{LR}) by fitting the following linear regression model of the form (Simons *et al.*, 1999b):

$$g(\text{cRMS}) = w(t) + \sum_{i=1}^{N} w_i F_i \tag{5}$$

where w_i are fitting constants independent of targets and $w(t)$ depends on target t.

$$g(\text{cRMS}) = \begin{cases} 4 & \text{if cRMS} < 4 \\ \text{cRMS} & \text{if } 4 < \text{cRMS} < 8 \\ 8 & \text{if cRMS} > 8 \end{cases} \tag{6}$$

We construct a training set of structures: $\{S(t, j) \mid 0 \leqslant t < T, 0 \leqslant j < N\}$ for T targets and N structures per target, then we minimize the following squared error:

$$\sum_{t,i} \left\{ g[\text{cRMS}(t, i)] - w(t) - \sum_{j=1}^{N} w_j F_j(t, j) \right\}^2 \tag{7}$$

Then w_j is obtained by solving the following equation:

$$\sum_n A_{kn} w_n = B_k \tag{8}$$

where

$$A_{kn} = \sum_{ij} F_k(i, j) \left[F_n(i, j) - \frac{1}{N} \sum_l F_n(i, l) \right] \tag{9}$$

and

$$B_k = \sum_{ij} g[\text{cRMS}(i, j)] \left[F_k(i, j) - \frac{1}{N} \sum_l F_k(i, l) \right] \tag{10}$$

and $w(t)$ is given by

$$w(t) = \frac{1}{N} \left\{ \sum_j g[\text{cRMS}(t, j)] - \sum_{jn} w_n F_n(t, j) \right\} \tag{11}$$

The A matrix is properly regulated so that it is non-singular and the above linear equation is uniquely solvable.

Multi-layer feed forward neural network

We use a typical three-layer feed-forward neural network (Figure 2) to do fold recognition: the input layer consists of six neurons corresponding to six fitness scores to be compiled for evaluation. The scores are rescaled by a sigmoid function $f(x) = 1/(1 + e^{-x})$ to values between 0 and 1 at the input layer. The hidden layer has five neurons which is sufficient for six input variables and the output layers has two corresponding to 'positive' and 'negative', respectively. Then we compute the ratio between them and rank the candidates with this ratio *P/N*: the higher it is, the more favorable is the candidate.

The computation at each neuron is done as follows: first compute the weighted sum of all input values from the

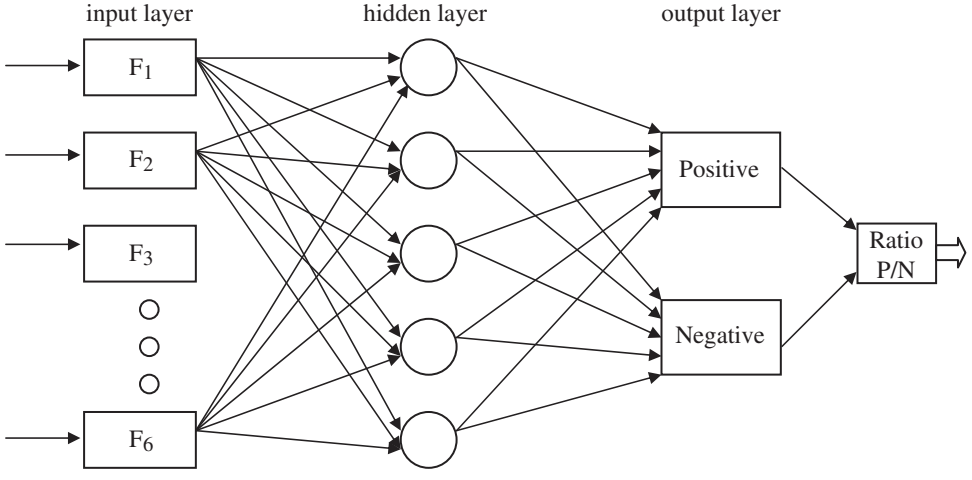

Fig. 2. A three-layer feed-forward neural network used for fold recognition. There are six input variables: $F_1 = F_{Rg}$, $F_2 = F_{SAXS}$, $F_3 = F_{hpb}$, $F_4 = F_{stat}$, $F_5 = F_{SS}$, $F_6 = F_{HPN}$. The hidden layer has five neurons and the output layers has two nodes corresponding to 'positive' and 'negative', respectively.

upstream layer (each link is associated with a weight), then apply the sigmoid function and output the result to the downstream layer.

The training is performed using the standard back-propagation algorithm and all link-associated weights are adjusted as a result of the learning process. The training set is composed of 11 proteins from Set (A). For each protein from the training set, 5000 candidate structures are extracted from its set of all candidates as ranked by their $cRMS_1$, which includes all the native-like candidates with $cRMS_1 < 6$ Å. The choice of 5000 results from a tradeoff between computing efficiency and the diversity of training data. The target values for both outputs are functions of $cRMS_1$: Positive output is set to 1 if $cRMS_1 <$ 4 Å, 0 if $cRMS_1 > 6$ Å and linearly interpolated in between; the negative output is set to 1 minus the positive output.

The learning process goes through the training set multiple times until 90% of the training targets have at least one native-like candidate ranking in top 10 by the ratio P/N. This choice of learning termination criteria ensures sufficient training without over-learning.

The validation of performance is done by running the neural network on a test set of 32 proteins from Set (A).

Results and discussion

Overview

The method used in this paper consists of the definition of a number of fitness scores with which to assess an alignment of a target sequence with a set of 10^4–10^5 candidate structures generated by gapless threading against the set of folds in the Dali domain library. An optimized combination of these fitness scores is then developed by use of two optimization methods, linear regression and neural network based, on a training set of 11 target sequences.

Once these optimized combinations of fitness scores have been generated, we apply them to a set of 62 test sequences for which we generate 10^4–10^5 candidates per target sequence. We then assess the performance of the fitness scores taken individually and of their optimized combinations, by computing their average Z-score for the native-like subset of candidate structures (see Materials and methods for the definition of 'native-like') and by finding the best Z-score rank of the native-like candidate structures.

As another measure of the effectiveness of the optimized fitness scores, we also determine if at least one of the structures with the top 10 Z-scores are structural neighbors to the target structure, as measured by the Dali structure alignment tool.

Generating candidate structures via gapless threading

To generate a set of candidate structures for training and evaluation of our fitness scores, we perform gapless threading of each of the target sequences against the Dali domain library (Holm and Sander, 1998), by a procedure which is described in detail in Materials and methods. The results are collected in Table I. This procedure generates sets of 10^4–10^5 candidates for each sequence. For small proteins with sequence length <80 these candidate sets are found to contain native-like structures of type (A) ($cRMS_1 < 6$ Å; see Materials and methods). For longer sequences (>90) the candidate sets contain structures with partially good structural alignments of type (C) (detectable by the LGA structure alignment tool with LGA_Q > 1.9, meaning significant structural similarity; see Materials and methods).

We divide the targets into three sets according to the quality of 'native-like' structures found in the sets generated by our threading protocol, for which the criteria of native-like structures are defined by (A), (B) and (C) as given in Materials and methods. Roughly, the (A) set is relatively easy for selecting native-like structures satisfying $cRMS_1 < 6$ Å which possess complete structural similarity to the native conformation, whereas the (C) set is more difficult as its structural similarity to the native is at most partially good with LGA_Q > 1.9. The (B) set is somewhere in between.

We select 11 proteins from the (A) set to serve as a training set for both the linear regression and neural network procedures. The rest of the targets are used as a test set for evaluating the performance of our SAXS-aided fold recognition protocol. Efforts are made to ensure that no protein in the test set is sequence homologue (with >25% sequence identity) of any protein in the training set.

Z-score evaluation of individual scores

In order to select native-like structures from the set of candidates, we need to define fitness scores (see Materials and methods) that are capable of discriminating them against non-native ones. It is then desirable to combine these scores to optimize the overall performance.

Before exploring the combination of multiple score functions we first study them individually. In total six fitness scores (F_{hpb}, F_{Rg}, F_{SAXS}, F_{stat}, F_{SS} and F_{HPN}) are used, which are described in detail in Materials and methods. They can be classified into three types: energy based (F_{hpb} which essentially evaluates how good the hydrophobic resides are buried inside a compact core and F_{stat} which is a statistical pair-wise contact energy derived from a protein structure database), 1D profile based (secondary structure assignment profile score F_{SS} and hydrophobic-polar profile score F_{HPN}) and SAXS based (F_{Rg} and F_{SAXS}). The main purpose of this study is to focus on the evaluation of the SAXS-based scores and their ability, in combination with the other scores, to improve the overall discrimination power of fold recognition.

For a given fitness score F and a given native-like structure s, we can define the following Z-score:

$$Z\text{-score} = \frac{F - \langle F \rangle}{\sigma_F} \tag{12}$$

where $\langle F \rangle$ is the average of F over the whole set of candidate structures and σ_F is its standard deviation. We use the Z-scores averaged over the set of native-like structures to evaluate the performance of a given score function: the more negative it is, the better is its ability to discriminate native-like from non-native structures.

In Table I of the Supplementary data (available at *PEDS* Online), we list the average and the optimal Z-scores and the best Z-score rank of the native-like structures for each individual fitness score F. One can see that F_{SAXS} (with average $Z_{avg} = -0.776$) and F_{Rg} (with average $Z_{avg} = -1.289$) do possess a good discrimination power to select native-like structures and that they are comparable to F_{hpb} (with average $Z_{avg} = -0.906$) and F_{stat} (with average $Z_{avg} = -0.739$). Therefore, SAXS-based scores indeed have the potential to help to

Table I. Summary of the results of generating candidate structures with gapless threading

Protein[a]	Sequence length	Fold class	All	Number of structures[b]			Optimal value		
				$cRMS_1 < 6$ Å	$cRMS_{0.8} < 5$ Å	LGA_Q > 1.9	$cRMS_1$ (Å)	$cRMS_{0.8}$ (Å)	LGA_Q
1ctf	68	α/β	36 589	11	8	44	5.331	4.507	2.452
1fwp	69	α/β	36 109	16	48	313	3.396	3.134	2.678
1nkl	78	α	39 592	9	22	115	5.353	4.107	3.241
1r69	63	α	56 548	31	49	28	3.306	2.153	4.346
2gb1	56	α/β	47 115	47	183	14	4.331	2.845	2.272
4icb	75	α	57 557	21	2	194	4.030	4.068	3.451
1csp	67	β	36 848	6	8	141	4.947	4.272	2.633
1leb	72	α	49 911	9	10	130	5.512	4.471	3.250
2ezh	65	α	74 561	54	134	87	4.319	3.749	2.131
2hp8	68	α	66 830	101	176	57	4.507	3.905	2.304
1ubi	76	β	62 277	3	12	116	3.718	1.548	6.043
1aa3	63	α	80 943	13	0	0	5.625	5.027	1.797
1apf	49	β	59 378	6	9	0	5.746	4.565	N/A
1c5a	65	α	82 817	7	11	0	5.448	4.577	1.870
1pou	71	α	51 259	39	9	35	5.077	4.667	2.061
1shg	57	β	37 376	13	12	69	0.825	0.729	6.452
1ag2	103	α	43 029	3	3	968	4.813	3.929	11.477
1aj3	98	α	42 883	41	32	4154	3.011	2.839	3.750
1svq	94	β	40 280	3	4	206	3.880	2.952	3.660
1wiu	93	β	47 165	16	11	1578	2.907	2.768	4.414
1erv	105	α/β	29 561	5	5	2730	2.869	2.589	15.186
1tit	89	β	62 090	4	17	1709	5.490	2.934	4.675
1afp	51	β	57 026	31	102	0	4.626	3.729	1.553
1orc	71	α/β	43 927	9	36	0	5.290	3.880	1.873
1ail	70	α	63 477	30	139	1532	4.900	3.513	3.166
1ajj	37	α/β	97 448	222	1240	0	4.561	3.303	N/A
1ayj	50	α/β	53 064	106	134	6	3.644	2.662	1.986
1bgk	37	α/β	30 276	2997	10 307	0	4.354	3.043	N/A
1cc5	83	α	49 479	10	5	133	4.381	4.034	3.160
1cmr	31	α/β	76 725	2998	9079	0	1.910	1.539	1.660
1dec	39	β	98 204	5	114	0	5.091	3.212	N/A
1erd	40	α	87 786	1181	6989	0	3.943	1.657	N/A
1gpt	47	α/β	43 054	130	418	5	3.916	2.838	2.300
1hev	43	α/β	62 569	10	51	2	3.064	1.564	3.615
1pft	50	β	1 30 518	1	32	0	5.824	4.357	N/A
1ptq	50	α/β	78 227	1	9	2	5.499	4.248	2.552
1qyp	57	β	95 499	2	16	0	5.637	3.938	1.813
1roo	35	α/β	49 836	2588	12 750	0	4.189	3.258	N/A
1utg	70	α	83 705	8	42	31	5.347	4.127	2.516
1vtx	42	α/β	84 850	50	246	0	3.519	2.427	1.565
2bds	43	β	53 775	29	159	0	4.593	3.678	1.402
2erl	40	α	72 604	4736	12 235	0	3.476	2.643	1.584
2fdn	55	α/β	66 785	29	86	76	2.503	1.337	3.861
1sro	76	β	53 992	0	2	190	7.253	4.763	2.507
2ncm	99	β	41 476	0	1	1489	6.526	4.463	5.185
1tig	94	α/β	27 554	0	71	2254	6.117	3.934	3.032
1aho	64	α/β	46 493	0	4	0	6.495	4.724	1.798
1bor	56	α/β	35 939	0	2	0	6.175	4.755	N/A
1lfb	77	α	78 466	0	1	15	6.433	4.974	2.218
2vgh	55	β	1 43 910	0	4	0	6.773	4.526	N/A
1bdo	80	β	58 711	0	0	53	6.467	5.754	3.201
1btb	89	α/β	34 539	0	0	235	7.069	5.676	2.401
1fbr 93	β	75 287	0	0	0	9.184	7.239	1.579	
1ris	97	α/β	15 841	0	0	275	6.746	5.066	3.232
1who	94	β	42 635	0	0	1282	8.795	5.455	2.980
2ezk	93	α	58 793	0	0	140	7.738	5.223	2.240
1ksr	100	β	48 704	0	0	1445	6.996	5.636	3.182
1pal	107	α	48 228	0	0	235	8.249	6.421	20.054
1tul	102	β	41 072	0	0	1194	7.330	5.982	17.004
2acy	98	α/β	22 683	0	0	163	8.499	5.052	2.687
1gvp	87	β	68 709	0	0	101	7.307	5.968	2.413
1aca	86	α	47 329	0	0	75	6.930	5.917	2.641
1aba	87	α/β	57 105	0	0	246	7.213	6.214	3.323
2ptl	78	α/β	71 314	0	0	24	8.855	5.337	2.172
1ddf	127	α	45 443	0	0	372	11.20	7.180	2.547
1hlb	157	α	31 500	0	0	732	10.93	8.898	24.87
1jvr	136	α	40 763	0	0	59	16.066	10.174	18.295
1kte	105	α/β	43 626	0	0	802	7.491	6.778	9.002
1lis	131	α	35 177	0	0	1397	9.324	6.027	74.87

Table I *Continued*

Protein[a]	Sequence length	Fold class	All	Number of structures[b]			Optimal value		
				$cRMS_1 < 6$ Å	$cRMS_{0.8} < 5$ Å	LGA_Q > 1.9	$cRMS_1$ (Å)	$cRMS_{0.8}$ (Å)	LGA_Q
1pdo	129	α/β	27 622	0	0	9493	9.074	6.336	44.10
1vls	146	α	25 205	0	0	3405	8.106	5.607	24.69
2fha	172	α	18 781	0	0	2445	10.22	7.378	43.56
2gdm	153	α	24 621	0	0	1466	9.990	8.331	20.27

This table summarizes the target proteins and for each of them candidate structures generated by our threading protocol for three sets of targets, (A), (B) and (C), which are separated by spaces. The top 11 proteins of Set (A) are used as training set for both LR and NN procedures while the rest are in the test set.
[a]Protein Data Bank (PDB) ID (Bernstein *et al*., 1977).
[b]Numbers of all and native-like structures for each target; here we use three ways to define native-like: $cRMS_1 < 6$ Å or $cRMS_{0.8} < 5$ Å or LGA_Q > 1.9, which emphasize different extents of structural similarity to the native conformation from complete to partial.

improve the selection of native-like structures in combination with the other more standard score functions.

Linear regression: performance evaluation of F_{LR}

To find an optimal linear combination of the individual scores that we have just evaluated, we use the linear regression (LR) method (Simons *et al*., 1999b), which is a simple and effective way of optimizing linear decision making. The motivation is to minimize the overall square deviation between a linear combination of all scores and a prediction quality function (see Materials and methods).

The coefficients for the optimal linear combination are evaluated for the training set of 11 target proteins by minimizing this function when averaged over those 5000 candidate structures closest in cRMS to each of the targets in the training set as explained in Materials and methods.

To evaluate the significance of SAXS scores in addition to other standard score functions, we run an LR for all the score functions excluding SAXS scores (F_{SAXS} and F_{Rg}) and then compare it with the LR results obtained when all score functions are included. Here is a summary of the results.

On average, the addition of the SAXS scores improves the *Z*-scores of F_{LR} from −2.066 to −2.319. Assuming that F_{LR} follows a Gaussian distribution approximately, then this improvement corresponds to a reduction of the *p*-value from 0.019 to 0.01 (or roughly by a factor of 2), which is fairly significant.

Out of 11 targets in the training set, 11 (100%) show better F_{LR} performance than any individual score *F* and 10 (90.9%) show better performance for F_{LR} with SAXS information than without it.

Out of 32 targets in the test set [also from Set (A)], 19 (59.4%) show better F_{LR} performance than any individual *F* and 24 (75%) show better performance for F_{LR} with SAXS information than without it. Therefore, LR provides a reasonably optimal way of combining multiple fitness scores into one score and manages to get the 'best of all' performance in most cases. Furthermore the incorporation of SAXS information improves LR's performance further with high probability (75%). Notably, in most of the cases where F_{SAXS} fails to improve the performance further, F_{LR} has already achieved a good *Z*-score without SAXS data.

In the light of the significantly better performance of F_{LR}, it is natural to ask how much each individual score contributes to this improvement. To shed some light on this issue, we also show the linear correlation coefficient between each individual score *F* and F_{LR} which measures the relevance of each *F* to F_{LR} (Table II). It is evident that F_{SAXS} [average correlation coefficient (c.c.) = 0.367] and F_{Rg} (average c.c. = 0.584) correlate better with F_{LR} than the other energy-based scores such as F_{hpb} (average c.c. = 0.104) and F_{stat} (average c.c. = 0.100). This suggests that F_{LR}'s significant improvement in discrimination of native-like structures is to a substantial extent due to the contribution of SAXS information.

We comment that the particularly large contribution of R_g to F_{LR} is largely a consequence of the gapless-threading-based protocol of candidate structures generation, which can easily produce many non-compact structures. We expect R_g to be less discriminating if applied to a set of more compact structures. Meanwhile, the weak contribution of F_{hpb} and F_{stat} is probably due to the prescreening which requires a significant matching of the HPN profile between the target sequence and the template sequence.

Neural network: performance evaluation of F_{NN}

Neural networks (NNs) have found extensive application in bioinformatics for their well-known capability of learning complicated patterns of relationships among multiple variables characteristic of biological knowledge of gene sequences and structures. There has been some application of NNs in fold recognition (Jones, 1999; Ding and Dubchak, 2001). Here we use a typical three-layer feed-forward NN to explore an optimal exploitation of the same six fitness scores used in LR (including SAXS scores). In comparison with LR, which is a typical linear decision procedure, non-linearity is introduced in NNs with the use of the sigma function (see Materials and methods), therefore it is not limited simply to producing a weighted linear combination of the original variables and is thus potentially more flexible in capturing complex patterns. The NN in use has six input variables corresponding to six scores: F_{hpb}, F_{Rg}, F_{SAXS}, F_{stat}, F_{SS} and F_{HPN}; each is normalized by subtracting the statistical average and then dividing by the standard deviation. There are two outputs, one corresponding to 'positive' and the other 'negative'. To make comparisons with LR's combined score function F_{LR}, we introduce a new score function which is the ratio between the 'positive' output and the 'negative' one and rank structure candidates with this ratio F_{NN}. Similarly to the evaluation procedure used in F_{LR}, we run the NN training and test with and without SAXS scores for comparison. In Table 1 of the Supplementary data, we list the *Z*-scores of F_{NN}. The training set for NN is the same as that used for LR. Here is a summary of the results.

On average, the addition of the SAXS scores improves the *Z*-scores of F_{NN} from −1.550 to −2.033. Again assuming that

Table II. Linear correlation coefficient of F_{hpb}, F_{stat}, F_{SAXS} and F_{Rg} with F_{LR}

Protein	Correlation coefficient			
	F_{hpb}	F_{stat}	F_{SAXS}	F_{Rg}
1ctf	0.091	0.066	0.494	0.653
1fwp	0.229	0.161	0.444	0.584
1nkl	0.217	0.277	0.368	0.534
1r69	0.172	0.088	0.410	0.527
2gb1	−0.019	0.023	0.386	0.549
4icb	0.292	0.116	0.474	0.599
1csp	0.208	0.211	0.474	0.554
1leb	0.103	0.068	0.534	0.665
2ezh	0.159	0.078	0.408	0.691
2hp8	0.097	0.070	0.416	0.680
1ubi	0.187	0.168	0.583	0.710
1aa3	0.129	0.099	0.434	0.604
1apf	0.131	0.089	0.268	0.412
1c5a	−0.013	0.021	0.399	0.560
1pou	0.176	0.084	0.429	0.604
1shg	0.261	0.061	0.356	0.423
1ag2	0.146	−0.023	0.655	0.789
1aj3	−0.015	−0.065	0.234	0.808
1svq	0.212	0.118	0.592	0.733
1wiu	0.125	0.102	0.554	0.737
1erv	0.277	0.183	0.611	0.715
1tit	0.182	0.119	0.516	0.697
1afp	0.130	0.157	0.227	0.404
1orc	0.070	0.146	0.449	0.611
1ail	0.118	0.130	0.307	0.683
1ajj	0.054	0.043	0.294	0.447
1ayj	0.050	0.182	0.328	0.502
1bgk	−0.128	0.145	0.119	0.295
1cc5	0.084	0.054	0.477	0.579
1cmr	0.027	0.085	0.121	0.384
1dec	0.130	0.064	0.183	0.418
1erd	0.146	0.210	0.322	0.491
1gpt	0.010	0.094	0.287	0.489
1hev	−0.060	−0.066	0.236	0.428
1pft	−0.107	0.032	−0.156	0.785
1ptq	0.043	0.126	0.356	0.504
1qyp	0.056	0.007	0.339	0.750
1roo	−0.029	−0.056	0.136	0.298
1utg	−0.031	0.054	0.217	0.697
1vtx	0.041	0.131	0.003	0.629
2bds	0.170	0.137	0.293	0.418
2erl	0.032	0.228	0.282	0.487
2fdn	0.096	0.081	0.336	0.450
Average	0.104	0.100	0.367	0.584

This table shows the linear correlation coefficients between four individual fitness scores and F_{LR}, which measure its relevance to the improved performance of F_{LR}. The results show that SAXS scores correlate more with F_{LR} than the energy scores, suggesting the importance of SAXS scores in F_{LR}.

F_{NN} follows a Gaussian distribution approximately, then this improvement corresponds to a reduction of the *p*-value from 0.0606 to 0.0212 (or roughly by a factor of 3), which is fairly significant.

Out of 11 targets in the training set, 11 (100%) show better F_{NN} performance than any individual score *F* and 10 (90.9%) show better performance with SAXS information than without it.

Out of 32 targets in the test set [also from Set (A)], 21 (65.6%) show better F_{NN} performance than any individual *F* and 24 (75%) shows better performance with SAXS information than without it. Therefore, NN shows a similar improvement to that found for LR and again SAXS is shown to be valuable in helping to improve the performance of the NN.

Testing F_{LR} and F_{NN} in native-like structure selection

After obtaining the optimal compilation of our fitness scores, we tested their performance in discriminating native-like structures from the candidate sets generated by our threading protocol. We list the best *Z*-score rank of native-like structures in Table III.

The results show that we have achieved reasonable success with the selection of native-like structures ($cRMS_1 < 6$ Å): in 8 (8) out of all 11 targets from the training set, at least one native-like structure is ranked in the top 10 by F_{LR} (F_{NN}). In 15 (14) out of all 32 targets from the test set [the rest of set (A)], at least one native-like structure is ranked in the top 10 by F_{LR} (F_{NN}). This suggests a success rate of good prediction to be between 40 and 50% for this protocol. We believe there is still ample room for improvement by using more accurate models that include side chains and other backbone atoms.

Testing F_{LR} and F_{NN} in structural neighbor identification

As an alternative test of the effectiveness of the performance of F_{LR} and F_{NN}, we measured which of the candidate structures in the top 10 of the *Z*-score ranked structures is also a structural neighbor (SN) of the actual protein as measured by the Dali structure alignment tool (alignment *Z*-score >2). This is a more challenging task than finding structures with low cRMS because the SNs are more remotely related to the target structure and the simple $cRMS_1$ does not detect the partial structural similarities that are detected by the Dali structural alignment. Since our scores are based mostly on the structure as a whole and are sensitive to possible fragmentation of the structure, their ability to discriminate native-like partial structural features is expected to be weaker.

In spite of this, the results in Table III still show that we have achieved a moderate success with the identification of correct SN's in the top 10 *Z*-sore candidates: in seven (six) out of all 11 targets from the training set, at least one candidate from a correct SN is ranked in top 10 by F_{LR} (F_{NN}). In 11 (11) out of all the 16 targets for which there exist correct SNs in the set of all candidates from the test set (the rest of set A), at least one native-like structure is ranked in top 10 by F_{LR} (F_{NN}). In 10 (11) out of all 26 targets from the harder test set [Sets (B) and (C)], at least one native-like structure is ranked in the top 10 by F_{LR} (F_{NN}). This suggests a success rate of SN identification to be between 60 and 70% for relatively easy targets, whereas for harder targets it drops to ~40%, which is still reasonably good.

We also give the *p*-values for the successful cases in Table III to assess the statistical significance of selecting an SN in the top 10. For some of the target proteins, the *p*-value is relatively high because of the large number of SNs for those proteins; for most others, the *p*-value is fairly low and suggests high statistical significance.

Compared with the previous test on native-like structure selection, this test is more relevant in the context of functional genomics based on structural homology relations. As is well known, a specific biological function of proteins is in general executed by a limited number of specific structural features (such as an enzyme's binding site) which are only part of the native structure as a whole. Therefore, the conservation of such partial structural features rather than the whole structure is more relevant to the conservation of function. In this context the present SN selection protocol seems to be fairly promising.

Table III. Performance evaluation of F_{LR} and F_{NN} in selecting native-like structures and correct structural neighbors (SNs)

Protein	Native-like prediction		Correct SN		*p*-Value	Protein	Native-like prediction		Correct SN		*p*-Value
	Best rank with F_{LR}	Best rank with F_{NN}	Best rank with F_{LR}	Best rank with F_{NN}			Best rank with F_{LR}	Best rank with F_{NN}	Best rank with F_{LR}	Best rank with F_{NN}	
1ctf	3	3	6	3	0.001	1qyp	8988	2962	–	–	–
1fwp	0	1	0	0	0.132	1roo	0	6	–	–	–
1nkl	942	457	83	45	–	1utg	1924	126	1965	9622	–
1r69	0	2	0	2	0.006	1vtx	18	679	–	–	–
2gb1	4	0	4	0	0.026	2bds	152	243	–	–	–
4icb	27	8	22	91	–	2erl	0	3	–	–	–
1csp	0	0	0	0	0.137	2fdn	0	6	0	6	0.060
1leb	1	28	1	27	0.035	1sro	831	175	9	21	0.098
2ezh	8	23	25	708	–	2ncm	15 685	36 644	6	0	0.542
2hp8	39	4	2347	1165	–	1tig	92	14	1	2	0.021
1ubi	0	0	0	0	0.024	1aho	929	12 724	–	–	–
1aa3	3282	2161	–	–	–	1bor	5417	2342	–	–	–
1apf	196	72	–	–	–	1lfb	1610	8854	1706	2791	–
1c5a	301	164	–	–	–	2vgh	18 183	3495	–	–	–
1pou	286	24	7	194	0.001	1bdo			154	366	–
1shg	0	0	0	0	0.043	1btb			593	394	–
1ag2	0	31	0	8	0.009	1ris			34	55	–
1aj3	1095	129	25	1	0.211	1who			35	1	0.557
1svq	0	0	0	0	0.005	2ezk			103	34	–
1wiu	0	2	0	2	0.568	1ksr			55	3	0.430
1erv	0	0	0	0	0.059	1pal			2268	2171	–
1tit	10	1	10	1	0.511	1tul			112	55	–
1afp	692	20	–	–	–	2acy			0	0	0.068
1orc	1	2	–	–	–	1gvp			40	119	–
1ail	844	269	47	269	–	1aca			121	48	–
1ajj	27	0	–	–	–	1aba			186	60	–
1ayj	0	0	0	0	0.002	2ptl			101	1286	–
1bgk	1	0	–	–	–	1ddf			109	32	–
1cc5	0	3	0	1	0.048	1hlb			7	6	0.007
1cmr	1	2	–	–	–	1jvr			3306	868	–
1dec	10	8029	–	–	–	1kte			2	0	0.045
1erd	19	20	–	–	–	1lis			63	64	–
1gpt	0	2	0	2	0.004	1pdo			0	1	1.870
1hev	0	67	0	365	0.001	1vls			1	0	0.110
1pft	65 823	67 582	–	–	–	2fha			9	0	0.480
1ptq	103	9766	85	1615	–	2gdm			3	7	0.012

This table summarizes the performance of F_{LR} and F_{NN} in selecting native-like structures and correct structural neighbors. For the latter we show the *p*-value, which is the probability of selecting a correct structural neighbor in the top 10 by chance. Structural neighbors are defined by the Dali structural alignment tool with a cutoff score $Z > 2$.

Applications of structural neighbor identification

The identification of correct SNs can provide clues to the functional study of a target protein. To illustrate this, we now discuss several such examples for targets we have studied for which correct SNs are selected and where we see interesting functional connections:

1. In a number of cases, the selected SN is in precisely the same family by sequence homology, for example:
   (a) 1r69 (a 434 repressor) and its SN 1b0nA (sinr protein) both belong to helix–turn–helix motif and fulfil DNA binding function;
   (b) 1shg (α-spectrin) and its SN 1griA (growth factor-bound protein) both belong to SH3 domain and are involved in signal transduction;
   (c) 1svq (severin) and its SN 1d0znA (horse plasma gelsolin) both belong to gelsolin and are involved in actin binding.

   In all of the above cases, the sequence identity is around 20–30% and so falls into the 'twilight zone' where sequence alignment does not give clear results.
2. In several cases, the selected SN is functionally related to the target:
   (a) 1csp (cold shock protein) is involved in DNA binding, whereas its SN 1ah9 (initiation factor) has RNA binding property; this suggests that they may both be derived from an ancient nucleic acid-binding protein;
   (b) 1leb (lexa repressor DNA binding domain) is involved in the DNA binding function of DNA repair regulation and transcription regulation process, whereas its SN 1ecl (*Escherichia coli* topoisomerase) participates in the process of DNA topological change and DNA unwinding, so they both share the function of DNA binding;
   (c) 1pou (pou-specific domain) is involved in binding to specific DNA sequences to cause temporal and spatial regulation of the expression of genes, whereas its SN 1knyA (kanamycin nucleotidyltransferase) binds to some RNA primer and has a significant homology to the family X of polymerases, so they both share the function of DNA binding.
3. In a few cases, there is no obvious functional relation between the target and the selected SN but there may

exist some undiscovered evolutionary relationship suggesting that it could be worth more effort to clarify such relationships:

(a) 1ctf (ribosomal protein) is involved in protein biosynthesis and its SN 1mla (an acyl carrier protein transacylase) functions as a multifunctional enzyme which participates in fatty acid biosynthesis, but it is not clear how they are related to each other;
(b) 1sro (pnpase fragment) is involved in RNA binding whereas its SN 1a62 (ATPase) is involved in ATP binding; it is noted that 1a62 contains a nucleotide-binding site for ATP and ADP which may be the common sub-structure for both of them;
(c) 2ncm (neural cell adhesion molecule fragment) belongs to immunoglobulin superfamily and may be involved in protein–protein and protein–ligand interactions, whereas its SN 1aac (amicyanin) is involved in copper binding and electron transport; this suggests the possibility of 2ncm binding metallic ions.

In summary, the above examples demonstrate that conservation in protein structures may imply evolutionary relationships and that structurally similar proteins may possibly share similar or related functions. Therefore, by identifying SNs which are structurally similar to a given target, we may gain some insight regarding the biochemical function of the target. Work in this direction is expected to be very fruitful.

Conclusion

We have carried out a systematic study of the use of structural information derived from SAXS measurements to improve fold recognition. The SAXS data for a target protein can serve as a structural fingerprint of its native conformation and can therefore be used to construct a similarity-based fitness score to evaluate candidate structures generated by threading. To combine the SAXS scores with the standard energy scores and other 1D profile-based scores, we have used both a linear regression method and a neural network approach from which we obtain optimal combined fitness scores and apply them to the ranking of candidate structures. Our results show that the use of SAXS scores combined with gapless threading significantly improves the performance of fold recognition. We also demonstrate the effectiveness of this protocol in selecting structural neighbors of target proteins, which can potentially aid the study of their biochemical functions.

The above results support the idea that SAXS-based fitness scores should contain newer structural information than the energy-based scores since the energy scores only take into account of spatially 'short range' native contacts (with inter-residue distance <7 Å) whereas the SAXS profile contains distance distribution information up to the size of the protein (although residue identities are not resolved). Indeed, at the angle cutoff of $S_{max} = 0.12$ Å^{-1}, the SAXS measurement is able to resolve the shape information (but not the detailed secondary structures). Therefore, besides the compactness information from R_g, the additional filtering capacity of F_{SAXS} is mostly due to the shape information encoded in the SAXS data. Therefore, the performance of F_{SAXS} for a given target protein may depend on the uniqueness of its shape.

To improve the SAXS-aided fold recognition further, it is desirable to replace gapless threading with more sophisticated gapped threading algorithms with inputs from the multiple sequence alignments (e.g. by PsiBlast; see Altschul *et al*., 1997). This will significantly enrich the native-like structures in the generated set of candidate structures compared with those obtained by gapless threading. We note that the threading-derived sequence–structure alignments must be further used to build a set of complete structural models before the SAXS scores can be assessed. This is not a straightforward task and may need *ab initio* modeling for those parts of the target protein for which no significant alignment with known structures is found.

In addition to the obvious application of this approach in the post-structural genomics age to help in the identification of the structures of specific genome sequences, it also has potential applications in the implementation of structural genomics projects. Given a set of proteins which have been shown by sequence alignment search to lack sequence homology to proteins of known structure, the use of SAXS data as an input, together with a fold recognition protocol, may be applied to identify a significant number of targets with structural similarity to known proteins even though they lack sequence homology. This approach will then help in target prioritization, either by confirming the putative structural homologues or analogues identified by the SAXS-based threading procedure or by suggesting target sequences with hitherto unknown folds. The SAXS-based technique may therefore help in reducing bottlenecks in high-throughput genomics projects by focusing attention on targets of specific biological or structural interest.

For future work, we plan to improve the SAXS-based protocol by using more accurate models which include side chains and other backbone atoms, in combination with experimentally obtained SAXS data, which may be complicated by measurement errors and the effects of hydration.

Acknowledgements

We thank D.Walther for his seminal contributions to the use of the SAXS fitness score. We are grateful to D.Hinds for providing valuable information about the simulation software that he had developed, to A.Zemla for providing the LGA software and to David Baker's group at the University of Washington for providing the Rosetta decoy set. This work is supported by NSF-PHY98. A hardware gift from INTEL is gratefully acknowledged.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.
Burley,S.K. (2000) *Nat. Struct. Biol.*, **7**, Suppl., 932–934.
David,R., Korenberg,M.J. and Hunter,I.W. (2000) *Pharmacogenomics*, **1**, 445–455.
Dima,R., Settanni,G., Micheletti,C., Banavar,J. and Maritan,A. (2000) *J. Chem. Phys.*, **112**, 9151–9166.
Ding,C.H. and Dubchak,I. (2001) *Bioinformatics*, **17**, 349–358.
Holm,L. and Sander,C. (1998) *Proteins*, **33**, 88–96.
Huang,E.S., Subbiah,S. and Levitt,M. (1995) *J. Mol. Biol.*, **252**, 709–720.
Jones,D.T. (1999) *J. Mol. Biol.*, **287**, 797–815.
Marchler-Bauer,A. and Bryant,S.H. (1999) *Proteins*, **37**, 218–225.
McLachlan,A.D. (1971) *J. Mol. Biol.*, **61**, 409–424.
Rost,B., Schneider,R. and Sander,C. (1997) *J. Mol. Biol.*, **270**, 471–480.
Rykunov,D.S., Lobanov,M.Y. and Finkelstein,A.V. (2000) *Proteins*, **40**, 494–501.
Simons,K.T., Bonneau,R., Ruczinski,I. and Baker,D. (1999a) *Proteins*, **3**, 171–176.
Simons,K.T., Ruczinski,I., Kooperberg,C., Fox,B.A., Bystroff,C. and Baker D. (1999b) *Proteins*, **34**, 82–95.
Stevens,R.C., Yokoyama,S. and Wilson,I.A. (2001) *Science*, **294**, 89–92.

Svergun,D.I., Petoukhov,M.V. and Koch,M.H. (2001) *Biophys. J.*, **80**, 2946–2953.
The Genome International Sequencing Consortium (2001) *Nat. Biotechnol.*, **409**, 860–921.
Venter,J.C. *et al*. (2001) *Science*, **29**, 1304–1351.
Walther,D., Cohen,F.E. and Doniach,S. (2000) *J. Appl. Crystallogr.*, **33**, 350–363.
Williams,M.G. *et al*. (2001) *Proteins*, **45**, Suppl. 5, 92–97.
Zemla,A. (2003) *Nucleic Acids Res.*, **31**, 3370–3374.
Zheng,W.J. and Doniach,S. (2002) *J. Mol. Biol.*, **316**, 173–187.

Received November 10, 2004; revised March 7, 2005; accepted March 25, 2005

Edited by Fred Cohen