doi:10.1006/jmbi.2001.5324 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) 316, 173–187

JMB

Protein Structure Prediction Constrained by Solution X-ray Scattering Data and Structural Homology Identification

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

[1]*Departments of Physics and*

[2]*Applied Physics, and Laboratory for Advanced Materials, Stanford University CA 94305, USA*

Here we perform a systematic exploration of the use of distance constraints derived from small angle X-ray scattering (SAXS) measurements to filter candidate protein structures for the purpose of protein structure prediction. This is an intrinsically more complex task than that of applying distance constraints derived from NMR data where the identity of the pair of amino acid residues subject to a given distance constraint is known. SAXS, on the other hand, yields a histogram of pair distances (pair distribution function), but the identities of the pairs contributing to a given bin of the histogram are not known. Our study is based on an extension of the Levitt-Hinds coarse grained approach to *ab initio* protein structure prediction to generate a candidate set of C^{α} backbones. In spite of the lack of specific residue information inherent in the SAXS data, our study shows that the implementation of a SAXS filter is capable of effectively purifying the set of native structure candidates and thus provides a substantial improvement in the reliability of protein structure prediction. We test the quality of our predicted C^{α} backbones by doing structural homology searches against the Dali domain library, and find that the results are very encouraging. In spite of the lack of local structural details and limited modeling accuracy at the C^{α} backbone level, we find that useful information about fold classification can be extracted from this procedure. This approach thus provides a way to use a SAXS data based structure prediction algorithm to generate potential structural homologies in cases where lack of sequence homology prevents identification of candidate folds for a given protein. Thus our approach has the potential to help in determination of the biological function of a protein based on structural homology instead of sequence homology.

Keywords: SAXS; protein structure prediction; diamond lattice model; structural homology

**Corresponding author*

Introduction

As a result of the tremendous progress in large-scale DNA sequencing projects,[1,2] the rapid growth in accumulation of biological sequence information has put strong pressure on the structural biology community to produce structural information for new genes with high throughput. Experimentally, large scale structural genomics projects have been initiated to streamline X-ray crystallography and NMR measurements at a factory scale, aiming to determine all (1000 to 10,000) available protein folds within a few decades or even years.[3] However, this progress by no means weakens the importance of *ab initio* protein structure prediction tools. On the contrary, difficulties of crystallizing many proteins, and limitations in the NMR approach increase the need for powerful prediction algorithms to offer maximal structural information for a large number of unknown proteins. It is well-known that the conservation of fold structures is more robust than the conservation of protein sequences, and that there exists significant correlation between structure and biological function which is lately being explored systematically.[4–6]

Abbreviations used: SAXS, small angle X-ray scattering; DLW, diamond lattice walk; dRMS, distance root mean square deviation; cRMS, C^{α} root mean square deviation.

E-mail address of the corresponding author: doniach@drizzle.stanford.edu

0022-2836/02/010173–15 $35.00/0

Structural homology is thus a very powerful tool by which we can assign functions *in silico* to new genes which bear only remote if any association with known genes in terms of sequence homology. For this purpose, a sufficiently high-quality prediction of the structure is crucial in order to capture the unique feature of the fold of a given protein against all other folds. There have been many discussions as to how good structure prediction must be in order to provide meaningful functional information.[7,8] Recently considerable progress has been made in *ab initio* protein structure prediction and fold recognition which are benefiting from the ever-increasing structure databases,[9,10] and it has been suggested that low-to-moderate resolution structural models produced by state-of-the-art structure prediction algorithms are sufficient to identify protein active sites and thereby provide pointers to function.[11,12]

The input of physical constraints from a variety of experiments can greatly aid both the efficiency and reliability of structure prediction algorithms. In particular, it has been shown that NMR-based distance constraints can be used to substantially improve structure prediction.[13–15] Here, we present an algorithmic approach which uses physical constraints derived from small angle X-ray scattering (SAXS) experiments to improve the quality of protein structure prediction. Because SAXS measures X-ray scattering from a protein in a relatively dilute solution it avoids the need to crystallize the protein and also allows measurements of protein conformations in nearly physiological conditions. Despite limitations in resolution resulting from the orientational averaging of the molecules in solution, SAXS yields physical information about the internal pair distribution of a molecule in its native state and is relatively easy to obtain. We show that use of this data can provide a substantial improvement in the reliability of protein structure prediction and that the resulting low resolution structures are capable of generating potential structural homologies in cases where lack of sequence homology prevents identification of candidate folds for a given protein. Thus the use of SAXS measurements as constraints on structural prediction algorithms may be expected to contribute to the effective operation of high throughput structural genomics and ultimately to its application in identifying the function of unknown genes.

Our approach to protein structure prediction incorporates SAXS data into the following general strategy: first a sampling procedure is used to generate a set of decoys with both native-like structural features and sufficient structural diversity for extensive sampling of protein folds, then several physical filters (including a filter based on SAXS data) are used to select a small number of promising native-like structures. This strategy has been shown to be successful in *ab initio* protein structure prediction.[16,17] Its advantages are threefold: First, it is very generic and flexible and both of its two stages, namely sampling and selection, are independently open to a variety of improvements (such as new models to represent protein conformations and new filters for selection etc); Secondly, the requirement on the energetic score function is less demanding compared with that needed for folding simulations where it must be sufficiently accurate to guide a miss-folded protein to its native state. This strategy is therefore much more tolerant to the use of simple models with significant coarse-graining and other approximations.[18] Thirdly, it can be easily implemented in parallel allowing for a huge set of candidate structures to be generated and evaluated at the same time.

Here, we implement this strategy at two levels of structural detail: diamond lattice walks (DLW) and off-lattice C^{α} backbones. At the first level, we perform an exhaustive enumeration of all self-avoiding walks on the diamond lattice within a given bounding volume, and then apply a combination of filters to select a limited set of walks for further refinement and selection; at the second level, we use the SAXS score to guide off-lattice relaxation and residue decoration in order to generate C^{α} backbones from the selected diamond lattice walks. Further filtering is then performed to select a small number of native-like structures.

The use of diamond lattice models to study protein conformations at low resolution has been well established by Hinds and Levitt.[18] Further hierarchical refinement toward all-atom models has been pursued recently by the Levitt Group.[16] Although following a similar strategy, our approach is novel in the following two aspects: (1). We incorporate new physical constraints in which information about the distribution of inter-atomic pairwise distances implicit in the one-dimensional SAXS data is used to help to guide the prediction algorithm. It is well-known that even a small number of specific inter-residue distance constraints (obtained by NMR experiments) can be very effective in filtering out non-native structures and greatly improve the performance of structure predictions.[13–15] The purpose of the present paper is to establish the effectiveness of the use of the ''non-specific'' inter-residue distance distribution provided by SAXS measurements as a physical constraint to protein structure enumeration. Reliable algorithms have been developed to reconstruct a low-resolution 3D electron density map of protein native conformation from one-dimensional SAXS data, which can reproduce the shape of the native structure with moderate accuracy, allowing for any topology of the target structure without prior estimation of its dimension.[20] With further improvement of the experimental resolution and refinement algorithms, Svergun and collaborators have recently shown that one can even extract the non-specific positional information at the residue level with relatively high accuracy with only the mapping of the polypeptide chain onto the residue positions undecided.[21]

Here, we go beyond these reconstructions of electron density to provide a physical map of the

polypeptide chain. To do this we adopt a straightforward way of employing SAXS data as a physical filter, where a SAXS score is defined as a linear combination of the root mean square (RMS) deviation and cross-correlation coefficient between the SAXS data of the native and candidate conformations[20] (see Methods). Our motivation is to establish the effectiveness of a SAXS filter in screening a decoy set and in particular to evaluate the level of improvement of its screening performance when combined with the widely used energetic filters. To the best of our knowledge, this work is the first to assess and employ a SAXS filter systematically in protein structure prediction.

(2). We further explore the effectiveness of structural homology searches based on the native-like C^{α} backbones predicted by our algorithm. Since the ultimate goal of protein structure prediction is to provide functional information for a target, which in turn may be closely associated with its structural features at a variety of fold hierarchy levels, it is of practical interest to evaluate the SAXS-guided predictions by using them to ascertain the fold family to which a target protein belongs. Here we adopt the Dali structural domain classification where a hierarchy of four levels (fold space attractor, fold topology, functional family and sequence family) is defined.[22] This structural comparison is implemented by using the LGA program developed by Zemla at Livermore National Laboratory,[23] which allows more flexible structural alignment than is provided by the standard coordinate RMSD. Our preliminary test of structural homology identification on a list of 12 target proteins shows that the predicted C^{α} backbones with the correct global topology are in most cases capable of providing fold classification information for the target in spite of its lack of local structural details and limited accuracy. This result is encouraging and suggests that we are after all not very far away from being able to turn the physical constraints-based structural prediction method into an effective application for deriving functional information. Recently Baker and collaborators have reached similarly optimistic conclusions with respect to *ab initio* protein structural genomics,[24] though they use a very different approach of structural prediction based on a fragment library derived from the known structural database[25] rather than the experimentally derived physical constraints used here. We are optimistic that high-throughput SAXS measurements combined with the present structural prediction algorithms (with future improvements) can potentially contribute significantly to large-scale structural genomics, and ultimately provide significant improvements in structure-based functional genomics.

Results and Discussion

Here, we have selected 16 target proteins which represent a variety of domain classes (eight all α, four all β, one α/β, three $\alpha+\beta$) with length ranging from 48 to 89. These targets were selected based on their use in previous studies.[18,25] They are not used in compilation of the statistical energy.[18] In selecting the targets, the only major criterion is that the sequence length should not be long (<90), which is a limitation resulting from our current implementation of the use of the diamond lattice model. See Table 1 for the list of targets and the relevant parameters.

Diamond lattice representation and exhaustive enumeration

The diamond lattice representation[18] used here is intended to capture the overall topology of the protein's native conformation at a low resolution. No secondary structure or side-chain orientation

Table 1. Target proteins and parameters

Protein[a]	Sequence length	DLW length[b]	Fold class	Dali domain ID no.[c]
1bdo	80	38	β	DC_4_12_1
1btb	89	38	α/β	DC_1_81_1
1ctf	68	34	α/β	DC_5_4_2
1nkl	78	38	α	DC_3_166_2
1r69	63	30	α	DC_3_257_1
1ubq	76	38	α/β	DC_2_65_3
4icb	76	38	α	DC_3_241_1
1csp	67	34	β	DC_4_8_12
1aa3	63	32	α	DC_3_254_1
1leb	72	36	α	DC_3_161_1
2ezh	65	32	α	DC_3_167_1
1c5a	66	34	α	DC_3_57_2
1pou	71	34	α	DC_3_257_1
1ubi	76	36	β	DC_2_65_3
1fwp	69	34	α/β	DC_5_2_37
1apf	48	24	β	DC_7_434_1

[a] Protein Data Bank (PDB) ID.[32]
[b] Length of DLW which is chosen to be an even number nearest to half the sequence length with the maximal cutoff at 38.
[c] Dali domain classification number DC_*l*_*m*_*n*_*p* representing fold space attractor region(*l*), globular folding topology(*m*), functional family(*n*) and sequence family(*p*).[22]

information is included in it. This representation has been proven to be a good starting point for hierarchical modeling of protein structure with more details added at different levels.[16] In most cases it can represent conformations of small proteins with a best dRMS fit of around 3 to 4 Å (see Table 2), regardless of the detailed architecture or secondary structures of the targets. The representation performs better for shorter sequences, which is because of more extensive sampling and better fit to the constraint volume.[16,18]

The size and shape of the bounding volume play a crucial role in determining the efficiency and accuracy of this algorithm: a larger volume results in exponentially longer searches while a smaller volume or wrong shape may cause the failure of exhaustive enumeration which may miss the relevant native-like diamond lattice walks. Therefore, we expect that the incorporation of a more detailed shape constraint derived from SAXS data could lead to considerable improvement of the present algorithm (work in progress).

Filtering at the diamond lattice walk level

After completing the exhaustive enumeration of 10^6-10^7 self-avoiding diamond lattice walks (DLW) we proceed to applications of multiple filters to select 1000 DLW for refinement at the next level. At the DLW level the following physical filters are used (they will also be used at the next level): radius of gyration (F_{Rg}), HP fitness score (F_{hp}), burial score (F_{burial}), statistical energy (F_{stat}) and SAXS score (F_{saxs}). The definitions of these filters are given in Methods. F_{Rg} has been extensively used to select compact structures;[18] F_{hp} and F_{burial} combined are able to optimize the prediction by favoring the formation of a hydrophobic core which has been shown to be essential to the establishing of a protein's native fold.[26] F_{stat} is the knowledge-based pairwise contact energy defined by a 20 by 20 score matrix derived from the statistical analysis of a selected set of representative protein structures.[18] All the above filters or similar ones are widely used in modeling and predicting protein structures. One should keep in mind that these filters overlap each other in capturing the features of native conformations, so their performances are not simply additive. Their performances are also limited by the resolution intrinsic to the model in use, since the diamond lattice model used at this level can only resolve different vertices rather than each individual residue (two to three residues per vertex). So we can expect additional filtering power by re-using these filters after refining the structural representation to higher resolution at the next level. We also introduce a new filter based on SAXS data: F_{saxs} (see Methods for detail of its definition and computation) which basically represents the similarity in SAXS intensity profiles between the given structure and the native structure.

Now we proceed to present the performance of each of the above filters at the DLW level (Table 2). We plot the dRMS distribution of DLW filtered by F_{Rg}, F_{hp}, F_{burial} and F_{stat}, respectively for 1ctf (Figure 1): In all cases, the distribution is pushed substantially (0.5 Å or larger) to the lower end of dRMS, which proves the effectiveness of these physical filters and is consistent with their relevance to protein folding principle and previous work[16,18]. We also tested F_{saxs} as a single filter (data not shown), which however shows no significant overall shift in its dRMS distribution: although native-like DLW (with small dRMS) tend to accumulate into the low F_{saxs} region, at the same time, a substantial fraction of non-native DLW

Table 2. Statistics of the enumeration and selection of DLW

	Top 1000 DLW dRMS (Å)[a]			Top 10,000 DLW dRMS (Å)[b]			All DLW dRMS (Å)[c]	
Protein	Min	Mean	Mwan-shift	Min	Mean	Mwan-shift	Min	Mean
1bdo	5.402	6.487	−0.5690	5.146	6.649	−0.4070	4.40	7.056
1btb	4.197	6.137	−0.5950	4.197	6.249	−0.4830	4.12	6.732
1ctf	4.099	5.397	−0.6200	4.070	5.697	−0.3200	3.41	6.017
1nkl	4.247	5.846	−0.2560	4.247	5.926	−0.1760	3.24	6.102
1r69	3.802	4.801	−0.4960	3.379	4.861	−0.4360	3.03	5.297
1ubq	4.055	5.625	−0.5260	4.055	5.895	−0.2560	3.82	6.151
4icb	4.337	5.291	−0.4190	4.059	5.506	−0.2040	3.51	5.710
1csp	4.311	5.606	−0.6860	4.311	5.816	−0.4760	3.73	6.292
1aa3	4.556	5.851	−0.1530	4.283	5.866	−0.1380	3.42	6.004
1leb	4.499	5.823	−0.2510	4.385	5.932	−0.2080	3.74	6.074
2ezh	4.824	6.070	−0.2330	4.731	6.205	−0.0980	3.15	6.303
1c5a	4.527	5.637	−0.4740	4.527	6.017	−0.0940	3.21	6.111
1pou	4.546	5.844	0.0230	4.386	5.823	0.0020	3.30	5.821
1ubi	3.816	5.572	−0.7630	3.816	5.864	−0.4710	3.58	6.335
1fwp	3.947	5.242	−0.7100	3.363	5.400	−0.5520	3.33	5.952
1apf	3.813	5.165	−0.7190	3.621	5.303	−0.5810	3.06	5.884

[a] dRMS distribution statistics: minimum, mean, and mean shift (relative to all DLW) of the 1000 DLW selected for the next step of C^{α} backbones refinement.

[b] dRMS distribution statistics: minimum, mean, and mean shift (relative to all DLW) of the 10,000 most compact DLW selected by F_{Rg}.

[c] dRMS distribution statistics: minimum, mean of all DLW enumerated within the given constraint volume.

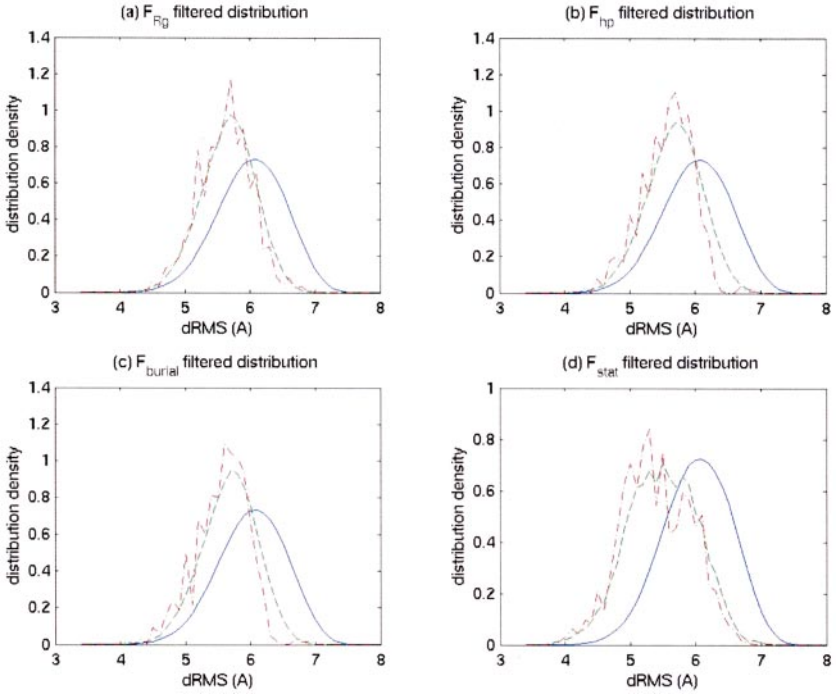

Figure 1. dRMS distribution of DLW for 1ctf filtered by F_{Rg} (a), F_{hp} (b), F_{burial} (c) and F_{stat} (d). The continuous lines are for all DLW (about 2.6×10^7), broken lines for top 1% DLW (2.6×10^5) and dot dashed lines for top 0.01% DLW (2.6×10^3). The distribution is shifted to the lower end of dRMS by about 0.5 Å or more after filtering.

(with large dRMS) also spread to this region, so the distribution becomes broader as we apply F_{saxs} progressively. This is because of the intrinsic structural degeneracy associated with the SAXS filter, namely that two different DLW can have a close F_{saxs} score as long as they share a similar density map, disregarding the order of vertices being traversed by the walks. The density map of the native structure is most likely highly degenerate in accommodating different walks. This requires us to make cautious use of the SAXS filter in combination with other filters, keeping in mind that a low SAXS score is only a necessary rather than sufficient condition to select a native-like structure. We will assess the performance of the F_{saxs} in detail at the next level.

The above result shows that in spite of the lack of local structural information (such as secondary structure) the simple low-resolution lattice modeling and the physical filters applied to it indeed provide an effective approach for the representation and selection of native conformations because it to some extent captures several important features characteristic of native structures, such as the formation of a hydrophobic core, its compactness etc. However, due to the limitation of low resolution and the coarse-grained nature of this simple model, it is unlikely to effectively eliminate most mis-folded structures (which tend to remain even after filtering and lead to wide spread distribution of dRMS of low energy structures) and to obtain a small set of high-quality native-like structures at this level. Thus we only perform a moderate filtering (down to 1000 candidates) at the DLW level and leave further refinement and discrimination to the next level where C^α backbones are generated.

Here is how we select the 1000 DLW to be used in the next stage:

As shown in Table 2, we first screen all DLW using F_{Rg} and keep the 10,000 most compact DLW. We then apply a combined filter of a heuristic average of F_{hp}, F_{burial}, F_{stat}, and F_{saxs} and keep the top 1000 DLW. After this screening process, the average dRMS is shifted (range: 0.15 to 0.76 Å) to the lower end (see Table 2). A heuristic combination of multiple filters is used so that they can complement each other and no filter needs to be too strict or over-used. We expect to achieve a more stable filtering performance in this way than by applying any individual filter alone. We also tried other possibilities of ordering and combining the above filters and did not find a filtering scheme with significantly better overall performance. Although we cannot exclude the existence of an optimal filtering scheme other than the heuristic one used here, considering the low-resolution of our lattice model and the moderate filtering strategy, strict optimization does not pose a crucial issue here.

Generation of C^α-backbones from 1000 diamond lattice walks

After the moderate filtering at the DLW level in the last step, we have 1000 DLW containing a much higher fraction of native-like conformations than the original huge set of all DLW.[18] Due to the limitation of resolution intrinsic to the lattice model, the coarse-grained representation must be augmented to the residue level in order to have a

structural representation which makes biological sense. Furthermore, a more detailed structural model allows the performance of the physical filters to be exploited more thoroughly so we can further purify the set of candidate structures. There is always a trade off between the extent of sophistication in structural representation and the feasibility of computation. As a result of this compromise, we choose to work with the C^{α} backbone representation. It is well known that the C^{α} backbone retains most of the structural features of native conformations, from global topology to local secondary structures. Given a sufficiently accurate model of C^{α} backbones, an all-atom model can be built in a relatively straightforward manner by side-chain packing.[27,28]

To generate a C^{α} backbone from a given DLW where the positions of residues assigned between neighboring vertices are ambiguous, we must work out a way to position all residues close to the given DLW while satisfying the constraints given by biochemistry. Due to the sharply increased number of degrees of freedom compared with that in the DLW representation, a sampling procedure is necessary to locally explore configurations near the starting DLW.

The C^{α} backbone generation procedure consists of the following two steps:

(1) Relax vertices from the diamond lattice by simulated annealing to optimize the SAXS score. This procedure is to remove the artifacts in the local geometry introduced by the lattice model and reshape the structure according to the known SAXS profile. In order to keep the global topology of the original DLW, each vertex is confined to within a distance cutoff from its starting position on the diamond lattice.

(2) Decorate inter-vertex residues onto the DLW, subject to the following geometry constraints characteristic of protein C^{α} backbones: (a) the nearest neighbor C^{α}-C^{α} distance is close to 3.8 Å, (b) the C^{α} bond angle is larger than 1rad[29], (c) no two residues are within 2 Å in distance. The sampling process is quite straightforward: local rotations of short fragments up to three residues long are attempted along the sequence accompanied with local optimization of the statistical energy (F_{stat}). This local optimization is justified under the proposition that the native pairs of contacting residues are already brought to proximity by the native topology at the DLW level so that a local adjustment is expected to push them finally into contact.

This procedure preserves the low resolution topology captured by the DLW without necessarily refining the secondary structure. This makes the approach generic and not susceptible to possible errors introduced by secondary structure predictions. Indeed, a significant fraction of secondary structures depend on their tertiary structure environment as well as their local sequence and cannot be predicted with confidence without knowing the tertiary structure in the first place. As shown by Levitt's group,[16] the secondary structure fitting procedure they used does not significantly change the cRMS distribution of decoy sets, though it does improve the number of native contacts due to hydrogen bonding within secondary structures. Thus it is reasonable to believe that it is the overall topology captured by the C^{α} backbones that plays the essential role in determining the global quality of a predicted structure.

Through this procedure, we can generate native-like C^{α} backbones with cRMS around 6 Å for most of our targets, regardless of their folds and secondary structure compositions. In particular, its performance on β-strand dominated targets is almost as good as α-helix dominated targets: out of the four all β targets, three have the best native-like C^{α} backbones with cRMS less than 6 Å. We attribute this to the extensive sampling at the DLW level which helps to increase the chance of representing conformations with significant number of non-local interactions characteristic of β-strands dominated structures. In Figure 2, we show the structural alignment between the native-like C^{α} backbones and the experimental structures for a number of targets. One can see the backbones indeed represent the native topology in a satisfactory way.

Assessment of the SAXS filter

In this subsection we present our assessment of the SAXS filter (F_{saxs}) at the C^{α} backbone level which is the crucial part of this work. Our purpose is to quantify its additional filtering power when used to supplement the action of the other available filters used here. It is in principle possible to combine all the above filters in an optimal way through supervised learning or other training processes. However for the purpose of filter assessment we prefer to separate them at different levels so that their individual filtering power can be analyzed separately. In particular, we apply the ''old'' energy filters (F_{hp}, F_{stat}) before using the new SAXS filter (though we did make use of the SAXS score to do off-lattice relaxation in the previous step), in order to assess the ''additional'' filtering power due to the new filter. This is an objective strategy: because of the possible correlation among these filters, a filter used earlier tends to be more effective than being used later. Therefore, saving the new SAXS filter to the last stage provides a reasonably stringent test of its filtering power.

Let us define the top ten C^{α} backbones with the smallest cRMS to the target as ''native-like'' backbones. Our aim is to capture at least one of them by applying our filters. We use a ''purifying factor'' (pF) to quantify the performance of a given filtering scheme. pF is defined as the ratio between the concentrations of native-like backbones after and before the filtering. A value of $pF > 4$ is considered very effective purifying, while a pF close to 1 means poor performance. Anything in between is seen as marginally effective.

For the purpose of SAXS filter assessment we perform a two-round filtering procedure:

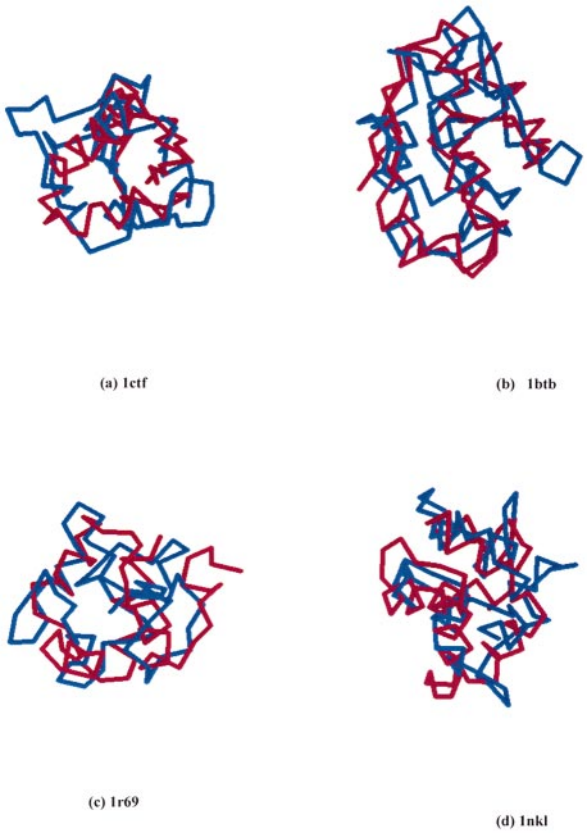

Figure 2. Comparison of C^{α} backbones between the predicted structures (blue) and the experimental structures (red) for 1ctf(a), 1btb(b), 1r69(c), 1nkl(d). The structural alignment is generated by LGA (GDT mode, DIST = 5 Å). The corresponding cRMS is 5.66 Å (a), 6.55 Å(b), 5.34 Å (c) and 5.81 Å (d). It is observed that the predicted C^{α} backbones capture the global topology of the native structures although there exist some local structural distortions.

First, the energy filters (F_{hp}, F_{stat}) are applied one by one to filtering 1000 C^{α} backbones down to N ($N = 200,100,50,20,10$), and record the corresponding purifying factor pF_1 as a function of N or $pF_1(N)$. Then select the maximum $pF_{1,max}$ which occurs at $N = N_{1,max}$, representing the maximal filtering power of the energy filters.

Second, we sort the resulting $N_{1,max}$ C^{α} backbones using the SAXS filter (F_{saxs}) and keep the top $N \times f$ (where $0 < f < 1$). We then study the corresponding purifying factor pF_2 at this round as a function of f and record its maximum $pF_{2,max}$ at $f = f_{2,max}$. We use $pF_{2,max}$ to represent the additional filtering power gained by using F_{saxs} subsequent to the use of the energy filters.

The results are shown in Table 3, where we also list the best ranks of the ten native-like backbones before and after using the SAXS filter. The results are summarized as follows:

(1) Out of all 16 targets, the new best rank is significantly improved in 13 of them, while in the remaining three, two of them have the old best rank already in the top ten. Only in one target (1fwp) does the SAXS filter fail, together with the energy filters (neither is able to select at least one native-like backbone in top ten).

(2) By examining the purifying factors, we find that out of all 16 targets, a $pF_2 > 4$ is achieved in 11of them, which suggests a significant purifying effect of F_{saxs}, while in the remaining five, three show marginal purifying effect ($2 \leqslant pF_2 < 4$) and two show none. We further notice that in the six targets where the energy filters perform poorly or marginally ($pF_1 < 4$), four of them are compensated by good F_{saxs} performance ($pF_2 > 4$).

The above results clearly show that the use of the SAXS filter significantly improves the selection of native-like backbones in combination with the energy filters. (Note: there are many other possible combinations for the first round of filtering without F_{saxs}, our motivation is to maximize its overall performance, so we pick only F_{stat} and F_{hp} which perform better than F_{Rg} and F_{burial} in terms of single filter performance. We also tried including F_{Rg} or F_{burial} but this mostly resulted in worse if not similar overall performance.)

In order to establish algorithms for structure prediction incorporating SAXS-based physical constraints, we have used simulated SAXS data obtained from the known structures of a test set of proteins in the Protein Data Bank. At the first stage of algorithm testing reported here, where side-chains are not included in the modeling, we use simulations done at a much simplified C^{α} backbone level using the Debye formula (see Methods). In order to relate the results of this simplified approach to experimental data, we also tested the use of simulations at the all-atom, solvent contrast level provided by the program CRYSOL.[19] In a preliminary study we find that, at the relatively low level of resolution implicit in the production of C^{α} backbones, the performance based on the more realistic CRYSOL simulations is somewhat reduced relative to that obtained with use of the simplified model, although still significantly better than results obtained without the use of a SAXS filter. At a later stage of algorithm development in which side-chains are included, we expect the use of realistic data simulation, or actual data, to be an essential step in improving filter performance.

Despite the positive results we just obtained, there are still concerns about the possibility that our target list may not be representative or that the result may depend on the approach we use to generate C^{α} backbones. In order to exclude such possibilities and establish our conclusion firmly, we proceed to test the filtering power of F_{saxs} on the selected decoy sets for 34 targets generated by Rosetta.[9] The result is shown in the lower part of Table 3. To select these decoy sets, we exclude fragmentary ones which may not preserve the SAXS profile of the whole native structure and those with low prediction quality (minimum

Table 3. Statistics of generated C^{α} backbones and performance assessment of SAXS filter

Protein	Top 10 backbonesc RMS (Å)[a]	All 1000 backbones cRMS (Å)[b]	Old best rank[c]	New best rank[d]	*pF* of 1st filtering[e]	*pF* of 2nd filtering[f]
1bdo	7.665-8.311	7.665-14.119	17(8.06)	3(8.31)	6.122	4.083
1btb	6.547-7.401	6.547-13.498	5(7.37)	5(7.37)	9.091	1.833
1ctf	5.658-6.306	5.658-11.613	8(6.15)	1(5.93)	11.11	4.500
1nkl	5.811-6.590	5.811-12.158	85(6.52)	5(6.59)	2.326	7.167
1r69	5.338-6.044	5.338-10.590	12(5.70)	3(5.70)	5.000	5.000
1ubq	5.615-6.683	5.615-13.001	17(6.15)	3(6.15)	4.762	5.250
4icb	6.570-7.257	6.570-12.064	67(7.24)	24(7.19)	1.980	2.020
1csp	5.748-6.551	5.748-12.187	46(6.18)	26(6.34)	4.008	1.064
1aa3	5.357-6.309	5.357-11.135	36(6.31)	1(6.31)	9.091	5.500
1leb	6.104-6.891	6.104-12.252	28(6.45)	0(6.45)	9.091	11.00
2ezh	6.166-6.433	6.166-11.872	49(6.17)	5(6.17)	2.000	8.333
1c5a	5.888-6.319	5.888-11.150	6(6.32)	7(6.32)	10.00	1.250
1pou	5.952-6.947	5.952-11.583	72(6.90)	0(6.90)	2.000	50.00
1ubi	5.737-7.074	5.737-13.046	133(7.04)	9(7.04)	2.020	4.950
1fwp	6.097-6.679	6.097-12.535	23(6.57)	28(6.40)	3.000	2.128
1apf	5.174-5.719	5.174-11.419	3(5.58)	0(5.58)	9.091	11.00
1bdo	6.578-7.726	6.578-16.655	36(7.61)	16(7.65)	3.996	2.500
1btb	6.273-7.849	6.273-18.272	25(7.58)	5(7.34)	3.996	4.167
1ctf	3.282-3.569	3.282-12.534	8(3.42)	6(3.37)	18.164	1.10
1csp	4.389-5.153	4.389-19.539	141(4.46)	71(4.46)	0.999	2.439
1aa3	1.858-3.689	1.858-16.049	39(3.45)	4(3.45)	3.996	5.000
2ezh	2.640-3.061	2.640-23.150	44(2.77)	2(3.04)	1.998	16.67
1fwp	5.788-6.634	5.788-15.995	89(5.79)	3(5.79)	0.999	25.00
1apf	4.809-5.752	4.809-17.139	26(5.729)	20(5.73)	1.998	2.381
1acf	6.441-8.356	6.441-22.448	93(8.36)	4(8.24)	1.498	13.33
1svq	5.524-7.208	5.524-20.792	98(6.98)	14(7.21)	0.999	6.667
1pal	6.946-7.926	6.946-18.961	74(7.45)	0(7.45)	0.999	100.0
2fha	6.946-7.927	5.908-27.686	17(8.24)	3(7.66)	3.996	6.250
1pdo	6.059-7.512	6.059-31.307	7(7.14)	5(7.14)	5.258	3.167
2ktx	2.137-3.181	2.137-7.608	9(3.03)	5(3.03)	5.258	3.167
1kte	5.036-7.139	5.036-21.502	12(5.68)	5(5.68)	5.258	3.167
2ncm	7.070-8.702	7.070-22.698	86(7.80)	23(7.80)	0.999	4.167
2pac	4.899-5.658	4.899-18.423	18(5.37)	3(5.37)	5.258	4.750
1ail	2.452-4.749	2.452-16.663	4(4.74)	8(4.75)	9.082	1.222
1lfb	3.545-4.804	3.545-15.805	75(4.31)	3(4.31)	0.999	25.00
1aj3	5.513-6.963	5.513-18.213	33(6.49)	29(6.49)	1.998	1.667
1eca	5.705-7.309	5.705-24.740	69(6.98)	51(6.98)	0.999	1.923
1erv	5.678-7.326	5.678-24.149	1(6.03)	0(6.03)	9.082	11.00
1ark	4.252-4.474	4.252-14.576	176(4.34)	17(4.45)	0.999	5.556
1msi	5.703-6.684	5.703-14.832	2(6.40)	1(6.41)	9.082	5.500
1ris	4.643-6.560	4.643-18.808	8(4.64)	9(4.64)	5.258	1.900
5icb	3.828-4.288	3.828-12.298	66(4.23)	12(4.24)	0.999	7.692
5pti	4.849-5.706	4.849-15.765	39(4.94)	3(5.62)	1.998	12.50
1gb1	1.953-2.768	1.953-15.525	33(2.67)	7(2.67)	3.996	3.125
1gpt	4.676-5.407	4.676-13.874	12(4.70)	7(4.70)	5.258	2.375
2ezk	4.611-6.233	4.611-23.446	33(5.86)	0(6.09)	1.998	50.00
1bor	4.980-5.617	4.980-11.492	96(5.29)	2(5.29)	0.999	33.33
2fdn	3.703-5.04	3.703-13.311	60(4.99)	11(4.60)	0.999	8.333
1orc	4.003-4.336	4.003-12.953	79(4.17)	6(4.08)	1.498	9.524
1tit	6.414-8.063	6.414-19.964	78(6.41)	8(6.41)	0.999	11.11

The upper part shows the results for our 16 targets while the lower part for the list of Rosetta decoys[9] for 34 targets selected from a complete list of 92 proteins based on the following criteria: (1) the decoy must be relatively complete with its length larger than 90 % of the sequence length of its target; (2) the best decoy has cRMS $\leqslant 7$ Å; (3) the 1st round of filtering does not dilute the density of native-like decoys, or $pF_1 \geqslant 1$. Among these 34 Rosetta targets, eight coincide with our target selection.

[a] cRMS range of the best ten C^{α} backbones defined as native-like backbones which we aim to select from the 1000 generated C^{α} backbones *via* filterings.

[b] cRMS range of all the 1000 generated C^{α} backbones.

[c] Best rank of the ten native-like backbones by the energetic filter (F_{stat}), where the corresponding cRMS is in the bracket.

[d] Best rank of the ten native-like backbones by the SAXS filter F_{saxs} after the first round of filtering by the energetic filters), where the corresponding cRMS is in the bracket.

[e] Purifying factor (*pF*) of the first round of filtering by the energetic filters (F_{hp} and F_{stat}).

[f] Purifying factor (*pF*) of the second round of filtering by the SAXS filter F_{saxs}.

cRMS > 7 Å). We also do not include decoy sets where the energy filters perform so poorly ($pF_1 < 1$) at the first round that it is not worth trying F_{saxs} at the second round.

We summarize the result as follows: (1) Out of all 34 targets, the new best rank is significantly improved in 26 of them, while in the remaining eight, seven of them have their old best rank

already in the top ten and only in one target (1aj3) does the SAXS filter fail, together with the energy filters (neither is able to select at least one native-like backbone in the top ten).

(2) By examining the purifying factors, we find that out of all 34 targets, a $pF_2 > 4$ is achieved in 21 of them, which suggests a significant purifying effect of F_{saxs}, while in the remaining 13, ten show marginal purifying effect ($2 \leqslant pF_2 < 4$) and three show none. We further notice that in the 19 targets where the energy filters perform marginally or poorly ($pF_1 < 4$), 15 of them are compensated by good F_{saxs} performance ($pF_2 > 4$).

In summary both results are consistent with each other and show that the SAXS filter performs effectively in purifying native-like structures. Therefore in general it is favorable to develop a multi-filter scheme including F_{saxs} in order to optimize the selection of native-like candidates from a decoy set of non-fragmentary structures. We will attempt this in the next subsection.

It is natural to ask to what extent does F_{saxs} contain new information relative to that already contained in the other filters. To answer this question we study the correlation coefficients (*c.c*) between the results of the SAXS filter (F_{saxs}) and those of the energy filters (Figure 3). It is found that there is no significant correlation between F_{saxs} and F_{stat} (average *c.c* is −0.033), compared with the positive correlation between F_{burial}/F_{hp} and F_{stat}. This is because the energy filters only take account of spatially ''short range'' native contacts (with inter-residue distance <7 Å) while the SAXS filter contains distance distribution information up to the size of the protein although the residue identity is not resolved. This explains why it can provide significant discrimination power on top of the energy filters.

Predictive filtering of C^{α} backbones

After establishing the effectiveness of the SAXS filter, we should construct a combined filter based on the above filters including F_{saxs} so that we can apply it to the selection of a few final predictions without prior information of the structure we try to predict. The optimal solution to this combinatorial problem is beyond the scope of this paper. At this stage we will rely on heuristics rather than a strict optimization machinery, in order to focus on the demonstration of feasibility of this approach while leaving further technical refinement to future work.

Let us first examine the discrimination power of each single filter (Table 4) by using the best rank of top ten native-like C^{α} backbones according to each of them. We notice that each single filter (except for F_{stat}) is at most moderately effective in discriminating native-like structures from other structures, partly because they were used before at the DLW level. However considering their being mutually-complementary in capturing different aspects of the native structure, their combination is potentially capable of performing a significantly better

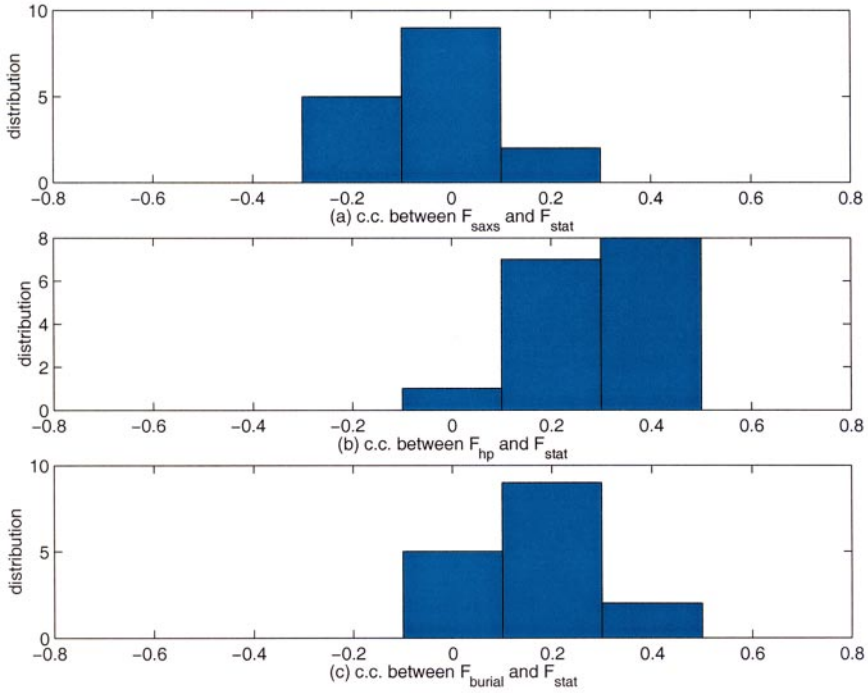

Figure 3. This Figure shows histograms for 16 targets of the correlation coefficient (*c.c.*) between F_{saxs} and F_{stat} (a), in comparison with the *c.c.* between F_{hp} and F_{stat} (b), and the *c.c.* between F_{burial} and F_{stat} (c). It is observed that there is at most unsignificant correlation between the SAXS filter and the energy filters, which explains why it provides significant discrimination power on top of the latter.

Table 4. Performance of single filters and the combined filter

Protein	F_{Rg} Best rank (cRMS)	F_{burial} Best rank (cRMS)	F_{hp} Best rank (cRMS)	F_{stat} Best rank (cRMS)	F_{saxs} Best rank (cRMS)	F_{comb}[a] Best rank (cRMS)
1bdo	3(8.065)	17(8.065)	25(8.065)	19(8.301)	281(8.311)	7(8.065)
1btb	161(7.401)	263(7.367)	123(7.321)	5(7.367)	149(6.915)	3(7.367)
1ctf	87(6.229)	64(5.934)	172(5.934)	8(6.153)	113(5.905)	2(5.934)
1nkl	26(6.397)	142(6.397)	168(6.524)	85(6.524)	10(5.811)	X
1r69	28(5.817)	59(5.817)	58(5.701)	12(5.701)	268(5.338)	6(5.701)
1ubq	26(5.872)	168(6.643)	90(6.332)	17(6.151)	67(5.992)	4(6.151)
4icb	58(7.147)	138(7.135)	207(7.192)	67(7.240)	137(6.570)	10(7.135)
1csp	54(6.184)	85(6.184)	96(6.474)	46(6.184)	224(6.474)	7(6.184)
1aa3	126(6.309)	58(5.645)	36(6.309)	38(6.309)	30(5.969)	12(6.309)
1leb	100(6.811)	97(6.811)	38(6.450)	28(6.450)	128(6.838)	15(6.450)
2ezh	306(6.433)	376(6.315)	140(6.388)	49(6.166)	36(6.201)	X
1c5a	35(5.888)	13(5.888)	6(6.316)	54(6.316)	38(6.255)	7(6.316)
1pou	126(6.898)	193(6.898)	103(6.898)	72(6.904)	87(6.904)	X
1ubi	60(6.462)	199(7.026)	133(7.037)	152(7.026)	27(6.897)	X
1fwp	1(6.430)	14(6.430)	38(6.430)	23(6.567)	103(6.679)	17(6.567)
1apf	239(5.175)	55(5.175)	24(5.175)	3(5.583)	45(5.420)	2(5.583)

[a] F_{comb} is a two-step filter with the first step pre-screening by F_{Rg}, F_{burial}, F_{hp} and F_{saxs} with a uniform fraction, and the second step using F_{stat} . X represents failure to select native-like C^{α} backbone in top 20.

discrimination than each individual filter, in particular than F_{stat} alone. The strategy we adopt is to first moderately prescreen the dataset with ancillary filters like F_{Rg}, F_{hp}, F_{burial}, F_{saxs} and then to apply the statistical energy filter (F_{stat}) to give a final ranking of all structures. In this way, we expect to eliminate false positive structures from high ranking that fail to be discriminated by F_{stat} alone.

Here is how we implement this strategy to generate a final set of ten C^{α} backbones for a given target: (1) Prescreen using F_{hp}, F_{burial}, F_{saxs} and F_{Rg} one by one with a uniform fraction f, namely sort all C^{α} backbones by each filter separately and record whether or not each C^{α} backbone is ranking in top $f \times 1000$. For better tolerance against filter error, one failure to meet the criterion of top $f \times 1000$ rank is allowed. The citerion for picking the value of f is to pick that value out of 0.5,0.4,0.3,0.2 which ensures that the number of backbones left after prescreening is in the range 100-200.

(2) Sort the dataset after prescreening by the statistical energy filter F_{stat}.

(3) Keep the top ten (or 20) as the final predicted set.

The criterion of success (moderate success) for the above multi-filtering scheme is to select at least one native-like C^{α} backbone within the final set of ten (20). The results are also shown in Table 4. Out of all 16 targets, nine are selected in top ten and 12 are selected in top 20. In comparison, using F_{stat} alone only three are selected in top ten and six are selected in top 20. The results on the list of targets suggest that the approach of multi-filter combination is very promising in achieving a good overall performance, although there is ample space for further improvement.

Structural homology search

In this subsection we report on results of using the structures predicted above in a structural homology search with the following two purposes: First, to evaluate the quality of our predicted C^{α} backbones selected in the last step *via* structural comparisons. Second, to explore in general whether a predicted C^{α} backbone with correct global topology but limited accuracy is able to correctly identify the folds structurally similar to it.

We do the structural alignment using the suite of LGA software developed by Zemla at Livermore National Laboratory[23] which was also successfully used in CASP4 for prediction evaluation. The predicted C^{α} backbones are the native-like ones selected among top 20 for 12 target proteins by using the above combined filtering scheme. The structural homology search is performed against the Dali domain library[22] prescreened by the SAXS filter, aiming to capture a domain which belongs to the same fold as the target (by sharing the same first three domain numbers) or is its structural neighbor (with the Z score of the structural alignment larger than 2). Either case should give clues for the determination of possible biological function.

The results are shown in Table 5. In most cases (eight out of 12) the predicted C^{α} backbone is able to rank one structural neighbor among the top three; five out of the eight pairs of target and its structural neighbor have sequence identity less than 15 % (1btb:8 %, 1ctf:8 %, 1ubq:11 %, 1aa3:6 %, 1fwp:4 %) which means their structural similarity is beyond the scope of sequence homology. Considering the limited accuracy of cRMS of around 6 Å and the lack of local structural details, this is quite encouraging and suggests the importance of global topology in deciding the protein fold.

Table 5. Summary of structural homology search using selected C^{α} backbones

Target protein prediction[a]			Domain hit by LGA[b]			LGA alignment result[c]		
PDB code	Dali domain ID no.	C^{α} backbone cRMS (Å)	PDB Code	Dali domain ID no.	Z	N_align	cRMS (Å)	Rank
1btb	1_81_1	7.37	1b9zA	3_234_1	3.2	38 (63)	2.99	1
1ctf	5_4_2	5.93	1dar	5_2_14	2.7	44 (75)	3.16	2
1r69	3_257_1	5.70	1b0n	3_257_1	11.4	41 (68)	3.55	2
1ubq	2_65_3	6.15	1alo	2_65_1	4.5	40 (79)	3.18	1
1csp	4_8_12	6.18	1ckmA	4_8_15	3.8	38 (63)	3.28	1
1aa3	3_254_1	6.31	1c3pA	7_216_1	2.7	38 (66)	3.24	1
1fwp	5_2_37	6.57	1qltA	5_9_1	2.5	41 (84)	2.76	1
1apf	7_434_1	5.18	1shl	7_434_1	3.9	26 (48)	3.13	2

[a] Information about the target protein predicted by our approach: its PDB code, Dali domain number,[22] and cRMS of the selected predicted C^{α} backbone (see Table 4) which is used to do LGA alignment.
[b] Information about the domain captured by LGA alignment: its PDB code (with chain specification), Dali domain number and the Z score ($Z > 2$ represents a significant structural similarity between the domain and the target protein).
[c] LGA structural alignment result: N_align is the number of residues aligned within the distance cutoff DIST = 5 Å, cRMS is the corresponding cRMS of these aligned residues, rank is the rank of the domain by LGA alignment score LGA_Q (see Methods).

Overall performance

Here is a brief summary of the overall performance of our SAXS-based structural prediction algorithm: Out of the total 16 target proteins: For 15 targets there is at least one native-like C^{α} backbone generated in the 1000 backbone samples with reasonably small cRMS (less than 7 Å). For nine (12) targets the multi-filtering scheme successfully selects at least one of the top ten native-like C^{α} backbones in the final set of ten (20) C^{α} backbones. For eight targets the selected native-like C^{α} backbone is able to capture one structural neighbor of its target protein among the rank of top three.

We note that our success in selecting native-like predicted C^{α} backbones and in the structural homology search is made for proteins spanning a variety of structural families and classes. This approach is also robust to the choice of native-like C^{α} backbones: we ran structural homology searches on all top ten native-like C^{α} backbones for each target, and found successful hits for most of them (data not shown).

The major limits to our approach are as follows:

(1) Size of the target protein: large proteins are more likely to be poorly sampled at DLW level and native-like C^{α} backbones may not be generated using our present implementation for proteins with more than 90 residues. (2) Discrimination power of multi-filters: the heuristic combination of multi-filters may not be optimal, it is desirable to explore more sophisticated ways to fully exploit the available filters such as clustering. (3) Overall quality of C^{α} backbone samples: though proven to be good for the purpose of structural homology search, a significant improvement in overall quality may be achievable through refinement and optimization, which should also considerably relieve the demand on filter power.

A few more comments on the structural homology search results: It is more or less surprising that a predicted structure with a limited 6 Å cRMS accuracy can be helpful in function analysis, contrary to the belief that a much better accuracy (about 2 Å cRMS) is required for such purpose.[7,8]

We comment as follows: First, we attribute our preliminary success to the use of very effective structural comparison tool, LGA, which is much more sensitive to shared global structural features than standard cRMS; in fact, our results show that LGA scoring system can select structural neighbors of a given target protein which are only poor alignments to the experimental target structure in terms of cRMS, suggesting the crucial role played by the effective structural comparison technique in addition to a high-quality prediction. However we must caution that even with the good tool, the alignments corresponding to a hit are still statistically weak in comparison to false positive domains with very close alignment scores. This calls for further efforts to sharpen the alignment tool and improve the prediction quality.

Second, we only aim to achieve the rather moderate goal of capturing at least one structural neighbor by structural alignment, which at most provides a rather partial clue to the biological function. Indeed the number of structural neighbors of a given protein can be large and it remains unclear how much functional information of the target is contained within a single neighbor of it.

Conclusion and future directions

Here, we have reported a systematic exploration of the use of distance constraints derived from SAXS measurement to filtering candidate protein structures for the purpose of protein structure predictions. This is intrinsically a more complex task than that of applying distance constraints derived from NMR where the identity of the pairs of amino acid residues subject to the given distance constraints is known, which is not the case for SAXS. Despite this complexity, our study shows that the implementation of SAXS filter is capable of

effectively purifying a set of native structure candidates. Furthermore, we have tested the quality of our predicted C^{α} backbones by doing structural homology searches against the Dali domain library and the results are very encouraging: in spite of its lack of local structural details and limited modeling accuracy useful information about fold classification can be extracted from this procedure which is potentially helpful to the ultimate determination of biological function.

For future work, we plan to improve the present approach along the following directions: First, we will refine the diamond lattice representation with a volume constraint so that larger proteins are also well represented and sampled without too much burden on enumeration, the incorporation of shape information from SAXS measurements is very promising for achieving this purpose. Second, we plan to add local structure and side-chain information to our prediction to further improve its quality. We expect that at this stage the use of experimental SAXS data will further improve the prediction quality. Third, we will explore better approaches to generating a small number of best candidate structures based on a larger set, either by optimizing the multi-filtering strategy or by using more sophisticated methods such as distance geometry.

Methods

To summarize our approach to SAXS-based protein structure prediction, a flow chart is given in Figure 4. The individual components are described in detail as follows.

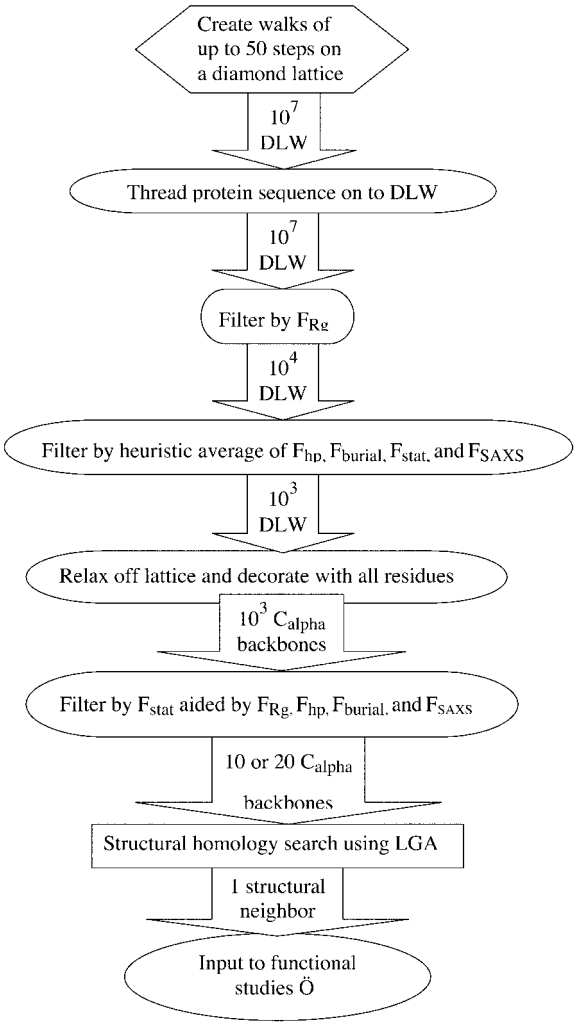

Figure 4. A flowchart describing our approach to SAXS-based protein structure prediction, where DLW indicates diamond lattice walks.

DLW enumeration

We enumerate all self-avoiding walks within the given ellipsoidal bounding volume containing 55 vertices on the diamond lattice, where the walk length is decided by the length of the target protein sequence (roughly two residues per vertex for small proteins with less than 80 residues). The length of an edge is 5.2 Å, and the major axes are 24 Å, 24 Å, 32.4 Å. In order to eliminate redundancy, only symmetrically unrelated starting vertices are used; also, by requiring the walk length to be even numbers, we only need to consider vertices with even coordinates as starting points, after the above preprocessing, only eight out of a total 55 vertices are eligible as starting vertices of walks. Depending on the walk length (up to 38 vertices), the total number of DLW varies in the range of 10^6 to 10^7 and can be enumerated by a Linux-based PC in a couple of hours.

Threading of a sequence onto DLW

For each DLW, we try to optimize the spacing of residues along it by using the statistical energy F_stat as a guide. Each vertex is mapped to a specific residue and from zero to two residues are assigned between each pair of neighboring vertices. Initially, we align the sequence uniformly onto the DLW, so that each vertex v is assigned a residue with sequence number $r(v)$, then we associate each vertex v with a set of candidate residues $R(v) = \{r | r(v) - 3 < r < r(v) + 3\}$; next, we implement a greedy algorithm to pick a specific residue r from $R(v)$ for each vertex v, so that the sum of pairwise contact energy for all pairs of vertices in contact spatially (but not sequentially) is optimized. This optimization method is very efficient and effective in selecting energetically favorable contacting pairs of residues which are promising to be utilized in the native conformation. For the same purpose Hinds and Levitt used a dynamic programming algorithm to optimize this threading process.[18]

Definitions of physical filters

For each DLW or C^{α} backbone, we evaluated the following filters:

F_{Rg}: (radius of gyration) RMS distance from the center of mass of all vertices (CA atoms) along the DLW (C^{α} backbones). This is a useful filter for selecting compact structures. Since *Rg* can be reliably derived from the

SAXS data, it is partially overlapping the SAXS filter defined later.

F_{hp}: HP fitness score[26] based on the hydrophobic-polar (HP) model which counts pairs of contacts between hydrophobic residues. At the diamond lattice level, we define contact between two vertices when they are nearest or next nearest neighbors on the lattice but not adjacent along the DLW. The contacts between inter-vertex residues are also included albeit with 0.5 weight (similar to what was done by Hinds & Levitt[18]). At the C^{α} backbone level, two residues are in contact if the distance between their CA atoms is less than 7 Å and they are not sequential neighbors.

F_{burial}: burial score[26] which measures the extent by which hydrophobic residues are buried inside the core, computed by summing the number of residues in contact with every hydrophobic residue.

F_{stat}: statistical energy which is the sum of statistical pairwise contact energy between any two residues in contact based on the 20 by 20 matrix constructed by Hinds & Levitt.[18] 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 by Hinds & Levitt.[18]

F_{saxs}: SAXS score, see later subsection for details of its definition and computation.

Structure comparison

dRMS at the DLW level: we use the distance RMS (dRMS)[30] to do a structural comparison between a given DLW and the corresponding native C^{α} backbone with given residue assignment.

The residue assignment is determined by optimizing the statistical energy (F_{stat}) using a greedy algorithm described above. Since its computation is based on distance only there is no discrimination between a DLW and its mirror image. This ambiguity is not resolved in the present algorithm owing to lack of a chirality measure.

cRMS at the C^{α} backbone level: we use standard coordinate RMS (cRMS) to do structural comparison between our predicted backbone and the corresponding native C^{α} backbone.[31] This is done by superimposing the above two structures onto each other and minimizing the RMS deviation between 90% of the residues (tolerating a small extent of errors at both terminals). We try both the given C^{α} backbone and its mirror image in the computation of cRMS and keep the minimum value of cRMS.

Structural homology search tool (LGA)

We use the LGA program developed by A. Zemla for structure comparative analysis of two protein structures or fragments of protein structures.[23] It has been successfully applied to the assessment of recent CASPs.[23] The program can be implemented in two general modes: sequence-dependent analysis and sequence-independent analysis. The first mode includes two analysis algorithms: LCS, which is to localize the longest continuous segments of residues that can fit under the selected RMSD cutoff, or GDT, which searches for the largest (not necessarily continuous) set of equivalent residues deviating by no more than a given distance cutoff (DIST). Since we are more interested in the global aspect of our modeling, we choose to use the GDT (DIST = 5 Å). The LGA generates the following scores as assessment of the structure comparison: N, number of residues superimposed under the distance cutoff; RMSD, RMSD computed on N residues superimposed under the distance cutoff; LGA_S, (0.00-100.00) calculated with reference to the number of residues in target protein; LGA_Q, quality score computed using formula:

$$Q = \frac{0.1N}{0.1 + RMSD}.$$

Among them we choose to use LGA_Q to rank the structure comparisons between our predicted C^{α} backbone and the Dali domain library.

SAXS score function computation

We adopt the score function used by Walther *et al.*[20] The profile of scattering intensity associated with a bead model (where the bead is a vertex or CA atom) is given as follows using the Debye formula in its pair-distance histogram form:

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

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:

$$F = w(1.0 - r) + RMS$$

with

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

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

Experimentally, measurement of SAXS profiles to a maximal value of $s_{\max} = 0.12$ Å^{-1} is in a range of intermediate scattering angles beyond the small angle region usually studied. Nevertheless reliable data in this range are readily accessible and in a recent paper, Svergun and collaborators[21] report data out to $s_{\max} = 0.27$ Å^{-1} for a number of proteins. We have done some tests of our algorithm for a reduced $s_{\max} = 0.06$ Å^{-1}. Although the performance is somewhat degraded, use of a SAXS filter is still found to be positive. Clearly, use of data out to the higher s values is beneficial.

Evaluation of structural homology using Dali domain classification and structural neighbors

In the Dali domain classification,[22] each domain is assigned a Domain classification number $DC_{l\ m\ n\ p}$ 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 $DC_{l\ m\ n\ p}$. Given the target protein, we first exclude all domain entries that share the same

$DC_{l\ m\ n\ p}$ number with it because these sequences bear a 25% or more sequence identity with the target. Then we prescreen the domain library with the SAXS filter and keep 10% of them. This filtering step effectively eliminates domains which differ significantly in length and shape from the target. Finally we undertake a one-to-all LGA structural comparison between the predicted C^{α} backbone and the post-screening domain library, and then compare the domains identified by the above comparison with high ranks (top 3) to the correct domain representative of the target. We call it a hit if both share the same first three Domain classification numbers (*l*, *m* and *n*) or are structural neighbors of each other. The definition of structural neighbors is based on the all to all Dali alignment[22] and its criterion is that the *z* score is no less than 2. This criterion is relatively strict in defining structural similarity so that it is biologically meaningful.

Acknowledgments

We thank Dirk Walther for his seminal contributions to the use of the SAXS filter. We are grateful to D. Hinds for his valuable information about the simulation software he had developed, and S. Subbiah for stimulating discussion and suggestions, to A. Zemla for providing the LGA software, and to David Baker's group at University of Washington for providing the Rosetta decoy set. W.J.Z. is supported by Stanford Graduate Fellowship. This work is supported by NSF-PHY98. A hardware gift from INTEL is gratefully acknowledged.

References

1. The Genome International Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature Biotechnol.* **409**, 860-921.
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. *et al*. (2001). The sequence of the human genome. *Science,* **291**, 1304-1351.
3. Burley, S. K. (2000). An overview of structural genomics. *Nature Struct. Biol.* **7(Suppl.)**, 932-934.
4. Turcotte, M., Muggleton, S. H. & Sternberg, M. J. (2001). Automated discovery of structural signatures of protein fold and function. *J. Mol. Biol.* **306**, 591-605.
5. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113-1143.
6. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. & Orengo, C. A. (2000). From structure to function: approaches and limitations. *Nature Struct. Biol.* **7(Suppl.)**, 991-994.
7. Koehl, P. & Levitt, M. (1999). A brighter future for protein structure prediction. *Nature Struct. Biol.* **6**, 108-111.
8. Wei, L., Huang, E. S. & Altman, R. B. (1999). Are predicted structures good enough to preserve functional sites? *Struct. Fold. Des.* **7**, 643-650.
9. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999). *Ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins: Struct. Funct. Genet.* **37(Suppl. 3)**, 171-176.
10. Sippl, M. J., Lackner, P., Domingues, F. S. & Koppensteiner, W. A. (1999). An attempt to analyse progress in fold recognition from CASP1 to CASP3. *Proteins: Struct. Funct. Genet.* **37(Suppl. 3)**, 226-230.
11. Fetrow, J. S. & Skolnick, J. (1998). Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T_1 ribonucleases. *J. Mol. Biol.* **281**, 949-968.
12. Skolnick, J., Fetrow, J. S. & Kolinski, A. (2000). Structural genomics and its importance for gene function analysis. *Nature Biotechnol.* **18**, 283-287.
13. Skolnick, J., Kolinski, A. & Ortiz, A. R. (1997). MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265**, 217-241.
14. Debe, D., Carlson, M., Sadanobu, J., Chan, S. & Goddard, W. (1999). Protein fold determination from sparse distance restraints. *J. Phys. Chem. B,* **103**, 3001-3008.
15. Bowers, P. M., Strauss, C. E. & Baker, D. (2000). De novo protein structure determination using sparse NMR data. *J. Biomol. NMR,* **18**, 311-318.
16. Xia, Y., Huang, E. S., Levitt, M. & Samudrala, R. (2000). *Ab initio* construction of protein tertiary structures using a hierarchical approach. *J. Mol. Biol.* **300**, 171-185.
17. Samudrala, R., Xia, Y., Huang, E. & Levitt, M. (1999). *Ab initio* protein structure prediction using a combined hierarchical approach. *Proteins: Struct. Funct. Genet.* **37(Suppl. 3)**, 194-198.
18. Hinds, D. A. & Levitt, M. (1994). Exploring conformational space with a simple lattice model for protein structure. *J. Mol. Biol.* **243**, 668-682.
19. Svergun, D. I., Barberato, C. & Koch, M. (1995). CRYSOL: a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallog.* **28**, 768-773.
20. Walther, D., Cohen, F. E. & Doniach, S. (2000). Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. *J. Appl. Crystallog.* **33**, 350-363.
21. Svergun, D., Pethoukov, M. & Koch, M. (2001). Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946-2953.
22. Holm, L. & Sander, C. (1998). Dictionary of recurrent domains in protein structures. *Proteins: Struct. Funct. Genet.* **33**, 88-96.
23. Zemla, A., Venclovas, C., Moult, J. & Fidelis, K. (1999). Processing and analysis of CASP3 protein structure predictions. *Proteins: Struct. Funct. Genet.* **37(Suppl. 3)**, 22-29.
24. Simons, K. T., Strauss, C. & Baker, D. (2001). Prospects for *ab initio* protein structural genomics. *J. Mol. Biol.* **306**, 1191-1199.
25. Simons, K. T., Kooperberg, C., Huang, E. & Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* **268**, 209-225.
26. Huang, E. S., Subbiah, S. & Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* **252**, 709-720.
27. Lee, C. & Subbiah, S. (1991). Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* **217**, 373-388.
28. Koehl, P. & Delarue, M. (1994). Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249-275.

29. Rackovsky, S. (1990). Quantitative organization of the known protein X-ray structures. I. Methods and short-length-scale results. *Proteins: Struct. Funct. Genet.* **7**, 378-402.
30. Cohen, F. E. & Sternberg, M. J. (1980). On the prediction of protein structure: The significance of the root-mean-square deviation. *J. Mol. Biol.* **138**, 321-333.
31. McLachlan, A. D. (1971). Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551. *J. Mol. Biol.* **61**, 409-424.
32. Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J. *et al*. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535-542.

Edited by F. E. Cohen

(Received 15 August 2001; received in revised form 6 November 2001; accepted 26 November 2001)