Blind Predictions of Local Protein Structure in CASP2 Targets Using the I-Sites Library

Christopher Bystroff and David Baker*
Department of Biochemistry, University of Washington, Seattle, Washington

ABSTRACT **Blind predictions of the local structure of nine CASP2 targets were made using the I-sites library of short sequence—structure motifs, revealing strengths and weaknesses in this new knowledge-based method. Many turns between secondary structural elements were accurately predicted. Estimates of the confidence of prediction correlated well with the accuracy over the whole set. Bias toward structures used to develop the library was minimal, probably because of the extensive use of cross-validation. However, helix positions were better predicted by the PHD program. The method is likely to be sensitive to the quality of the sequence alignment. A general measure for evaluating local structure predictions is suggested. Proteins, Suppl. 1:167–171, 1997.** © **1998 Wiley-Liss, Inc.**

Key words: sequence profiles building-blocks; secondary helix; strand turn knowledge-based

INTRODUCTION

Predicting a protein structure from its amino acid sequence is still best accomplished by looking for a homologue in the database of known structures. Unfortunately, there are still a great many protein sequences that have no such homologues. For these, we must rely on ab initio methods such as secondary structure prediction. The accuracy and precision of a secondary structure prediction is limited by the standard three-state model.[1–4] For example, the current turn prediction gives little specific structural information. Can one do better by defining more local structure motifs? One approach to this problem is to look for short sequence patterns, which recur in many protein families and ask whether any of them correlate with a particular structure besides helix or sheet. An affirmative answer to this question[5,6] led to the development of a library of sequence–structure motifs.

Our method uses multiple sequence information to predict the structure of protein fragments 3 to 15 residues in length. A library ("I-sites") of sequence–structure motifs was obtained by a procedure that identified sequence patterns that correlate with structure over a large, nonredundant database[7] of multiple sequence alignments and structures. Briefly, the library was generated as follows: Sequence segments were clustered,[8] using a measure of sequence similarity, and profiles[9] were generated from each cluster. All clusters that were found to have a single, predominant structure were identified and their profiles were iteratively optimized to correlate with that structure. The resulting profiles were weighted to best reflect the natural occurrence of each type of structure. The types of local structures that can be predicted by using this library are described elsewhere[10] and summarized in Table I. A list of predicted fragments is obtained by scanning a multiple sequence alignment for I-sites sequence patterns. The fragments can then be joined to generate predictions for longer segments.

Any prediction method that optimizes parameters against the data it proposes to predict is susceptible to database bias. In the development of the I-sites Library, cross-validation was used to avoid such bias. However, multiple usage of the same cross-validation dataset to optimize the method may introduce a bias in favor of that set. Blind predictions of data not included in any training data set are a necessary measure of the true predictive power of the method. Here we present blind predictions for nine targets, part of the Critical Assessment of Structure Prediction experiment (CASP2) to which this issue is dedicated.[11] The results revealed a number of strengths and weaknesses in the method with regard to different structural motifs, and a reassuring absence of database bias.

RESULTS AND DISCUSSION

How Predictions Were Made

The I-sites library (Table I) contains 82 sequence patterns, each with a corresponding local structure type. The clusters can be grouped into 13 sequence–structure motifs, describing 11 types of local structure.

For each target sequence, a set of aligned, homologous sequences was obtained from the PHD server,[12] and a profile was generated from the alignment. Each segment within the profile was scored against each of the 82 clusters in the library. Backbone

*Correspondence to: Dr. David Baker, Department of Biochemistry, University of Washington, Seattle, WA 98195-7350. E-mail: baker@ben.bchem.washington.edu

Received 9 May 1997; Accepted 9 August 1997

TABLE I. The I-Sites Library*

Motif	No. of clusters	Secondary structure	Summary of the sequence profile
1 Amphipathic α helix	13	HHHHHHH	π · ϕππϕ ·
2 Nonpolar α helix	6	HHHHHHH	[QA]Aϕ · AϕA
3 Gly αC-cap type 1	6	HHHHLLLLEEEE	· ϕ · π · G[AV]D · ϕϕϕ
4 Gly αC-cap type 2	10	HHHHLLLEEEE	· ϕππhGϕPϕϕ ·
5 Pro αC cap	10	HHHHLLL	ϕϕπ · [HnYF]P[DE]
6 Frayed α	2	HHHHHLLL	ϕ · π · ϕππ[HY]ϕ
7 Ser αN cap	10	LLHHHHHHHH	[TDS]Pπ[EQ]ϕ · πϕ · π
8 Amphipathic β strand	8	EEE	ϕ · ϕ
9 Hydrophobic β strand	5	EEE	ϕϕϕ
10 Asp β bend	2	LLEEEE	ϕD · ϕϕϕ
11 Ser β hairpin	4	EELLLLLEE	ϕ[DN]P · [ST]Geϕ ·
12 PDG hairpin	2	EELLLLLEE	ϕ · [DN]P[DN]Gπϕϕ
13 Diverging turn	4	ELLLLEE	ϕ[PK]PG[DQe] · ϕ

*A summary of the I-sites motifs is presented. Each motif represents a subset of the 82 clusters that make up the library. Secondary structure (3-state) and a simplified summary of the sequence profile are presented for a characteristic segment of each motif. In the sequence profile, each symbol represents an amino acid preference as follows: uppercase amino acid code = strong preference; lowercase amino acid code = weak preference; π = general preference for polar, ϕ = general preference for nonpolar; · = no preference. Symbols in brackets are multiple preferences at a single position.

angles from the highest-scoring, mutually compatible segments were chosen to generate the coordinates of the predicted structure. No attempt was made to make the structure compact or to avoid bad nonlocal contacts.

Fourteen predictions were submitted to CASP2: (targets 6, 11, 19, 20, 21, 22, 23, 26, 30, 31, 32, 37, 38, 42) Nine of those structures have been made available to the predictors (11, 20,[13] 22, 30, 31,[14] 32 [G. Boissy, unpublished data, 1997], 37, 38,[15] 42 [E. Liepinsch, M. Andersson, J.-M. Ruysschaert, G. Otting, unpublished data, 1997], although one (target 22) only as the α-carbon backbone. PHD secondary structure predictions were combined with I-sites predictions for targets 37 and 42 only.

How Predictions Were Evaluated

These local structure predictions are not adequately evaluated using a "Q3" score,[4] since the number of states is now more than three. But, since nonlocal interactions were not considered, the global root-mean-square deviation (RMSD) in alpha-carbon positions is also meaningless. One possible solution is to measure the local RMSD for an N-residue window around each residue and ask how many residues were included in a fragment of length N with an RMSD less than some cutoff value. A similar approach is to measure the maximal deviation in backbone torsion angles (MDA) over an N-residue window, again counting the number of residues that fall in fragments with MDA below some cutoff value. The local MDA method has several advantages over the local RMSD method. Conserved interresidue contact patterns were found to correlate better with MDA than with RMSD. For example, a type I versus a type II β hairpin may have a low RMSD, but the high MDA will correctly differentiate them. Conversely, RMSD for a pair of flexible β strands may be large while the low MDA will correctly place them in the same category. Therefore, we used the following measure to evaluate the CASP2 predictions:

$$\%\text{correct} = \frac{100\% \times \sum_{j=1}^{N} \left\{ \begin{matrix} 1 & \text{if } \min_{k=j-7,j}[\text{MDA}(8)_k] < 120^\circ \\ 0 & \text{otherwise} \end{matrix} \right\}}{N} \qquad (1)$$

where $\text{MDA}(8)_k$ is the maximum deviation in backbone torsion angles for a segment of length 8 starting at residue k. Other evaluation measures have been used for these predictions by the CASP2 assessors, as described elsewhere in this volume.

A summary of the results using the MDA(8) measure is presented in Table II. Of the 8 angle sets submitted for the experiment, only 2 targets used PHD information; the other 6 were purely I-sites predictions.

What Went Right

I-sites turn motifs complement secondary structure

This work was intended to improve the accuracy and precision of local structure prediction in loop regions, and in that respect it succeeded. In regions predicted to be loops by PHD, I-sites predictions succeeded more often in generating correct 8-mers than the alternative: assigning generic most-probable turn angles throughout the region (see Table II). Predictions of PHD loop positions using this alternative were about 22% correct, whereas for I-sites the

TABLE II. Prediction Summary for Eight Targets*

	All	%correct				PHD loop	%correct	
Confidence	positions	I-sites	PHD	Combined	Submitted	positions	Submitted	PHD
0.8–1.0	110	76.	72.	78.	76.	19	74.	18.
0.6–0.8	240	52.	51.	65.	53.	72	39.	23.
0.4–0.6	374	38.	47.	55.	46.	121	28.	22.
0.2–0.4	337	28.	36.	43.	33.	177	23.	16.
0.0–0.2	210	22.	31.	32.	30.	144	24.	23.
Totals	1271	39.	45.	52.	44.	533	29.	21.

*Evaluation of predictions for CASP2 targets 11, 20, 30, 31, 32, 37, 38, and 42. Each of the following were evaluated position by position as a function of the prediction confidence: I-sites predictions, PHD predictions converted to local structure predictions, an I-sites/PHD combination, and the submitted predictions (targets 37 and 42 had combined I-sites/PHD predictions, the others were I-sites only). In the last two columns are the statistics on the subset of the submitted predictions corresponding to the PHD-predicted loop positions. The percent correct refers to the MDA(8) (Eq. 1). PHD predictions were translated to backbone angles using idealized phi/psi angles for helix and sheet and generic turn angles (−75, −15) in the loop regions. For "Combined," the following formula was used to choose which method to use at each position:

$$\text{if}\left\{\begin{matrix}\text{H and }(0.2r - 0.30) > cf\\ \text{E and }(0.3r + 0.05) > cf\end{matrix}\right\}\text{use PHD}$$

where r is PHD's reliability (0–9), cf is I-sites' weighted confidence (0.0–1.8). Thus, most PHD predictions of helix (H) were used if the reliability was over 6 and most sheet (E) predictions were used if the reliability was over 3. PHD loop predictions were not used.

percent correct reached 74% in the highest confidence bin. This means that even the most simple-minded combination of the two methods—that is, combining PHD helix and sheet positions with I-sites loop positions—is better than the best secondary structure prediction. Unfortunately, due to prediction deadlines, only two of the nine targets used the combined methods. In these two targets (both helical proteins), adding I-sites to the PHD prediction increased the MDA(8) by contributing correctly predicted helix caps.

Figure 1 highlights a few of the successful predictions of motifs other than pure helix or sheet. Of these, the most novel is the "diverging" turn (Fig. 1f), found at around residue 57 of target 38 (*Pseudomonas cellulase*[15]). Although helix and sheet comprise a large portion of the local structures, the addition of specific chain reversals to the library of motifs that can be predicted from sequence has the potential to facilitate the assembly of secondary structure fragments into a global fold.

Accurate confidence values from cross-validation

An important part of structure prediction is a good estimate of the reliability of the prediction. For I-sites predictions, an estimate of the reliability ("confidence") was reported for each backbone angle (because the submission format did not allow this, these numbers appeared in the "REMARK" lines). The confidence values assigned to each of the positions in the target are derived from the sequence score of the best fragment prediction at that position. The confidence of a fragment prediction is the probability that a sequence segment with a given score has the predicted structure. These were determined using a jackknife procedure. Once the highest-confidence fragments are pieced together the confidence values are somewhat blurred, since predicted fragments of different lengths and different confidences often overlap. It was found that using the MDA(8) measure, the percent correct correlates well with the confidence, both in our training dataset and in the CASP2 blind predictions.

No significant dataset bias

The supervised learning procedure used to generate the scoring matrices is susceptible to database bias. We made an effort to avoid such bias by using a large, nonredundant set of structures and by extensive use of cross-validation. Cross-validation should prevent the circular logic inherent in a supervised learning approach. In fact, the percent correct for the CASP2 targets (39%), using the I-sites method alone (without PHD), was not much different from the percent correct for predictions of all proteins in the training set (41%).

What Went Wrong

Poor helix predictions

I-sites predictions of helix positions were not as good as published methods, particularly the PHD server.[12] Using the mda(8) measure, I-sites predictions placed about 52% of the helix positions in correct 8-mers, while PHD predictions placed 82%. Possibly this may be attributed to the fact that PHD and other methods use a longer window (i.e., 11–23 residues[2]) to predict secondary structure. Sequence

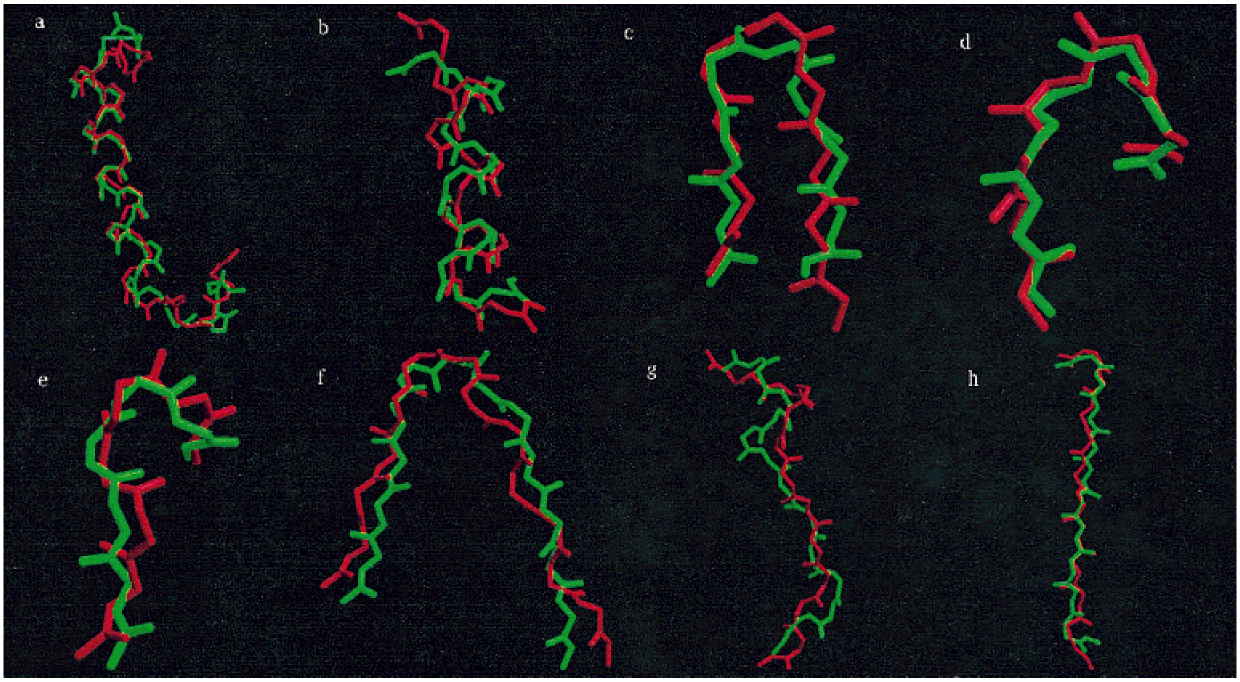

Fig. 1. Fragments of CASP2 predictions (*red*) overlaid on the true structure (*green*). Particularly interesting high-confidence fragments were chosen for display. The target segment; the I-sites motif(s) predicted; the RMSD between the predicted and the true segment; and the confidence(s), for each fragment are as follows. **a:** Target 20, 147–174; S αN-cap, NP helix, P αC-cap; RMSD = 2.5 Å; confidence = 0.53–0.75. **b:** Target 20, 233–248; S αN-cap, G αC-cap; RMSD = 2.3 Å; confidence = 0.50–0.77. **c:** Target 31, 199–2–7; PDG hairpin; RMSD = 2.2 Å; confidence = 0.65. **d:** Target 30, 22–28; S β hairpin; RMSD = 1.1 Å; confidence = 0.43. **e:** Target 37, 78–84; S αN-cap; RMSD = 1.2 Å; confidence = 1.00. **f:** Target 38, 51–64; diverging turn; RMSD = 1.7 Å; confidence = 0.74. **g:** Target 31, 78–91; S β turn, amphipathic β; RMSD = 3.0 Å; confidence = 1.00, 0.62. **h:** Target 11, 131–142; D β bend, amphipathic β; RMSD = 1.2 Å; confidence = 0.49, 0.65.

information outside of the window in which the helix occurs may contribute to the helix formation by signaling a cap or the continuation of a helix. For sheet positions the two methods were more or less comparable, with I-sites doing slightly better: 40% vs 35%. For both methods, the accuracy is generally higher for helical proteins.

Dependence on sequence alignment

Both the number of homologous sequences and the accuracy of their alignment to the target affected the accuracy of the I-sites predictions. Our sequence alignments were obtained by using the public-domain PHD server.[12] Only three of the nine multiple sequence alignments had more than four aligned sequences. Two of these three (targets 11 and 37) produced most of the high-confidence predictions. Upon inspection we observed some doubtful alignments for some of the other targets, especially when distant homologues were involved. Occasionally, removing a suspicious sequence improved the accuracy of the structure prediction; however, in general, removing sequences of low homology decreased the accuracy of I-sites predictions. Correctly aligned true homologues of the lowest percent identity add the most information to the sequence profile. Therefore, the success of this method hinges on the accuracy of alignment of those sequences that are often hardest to align.

Nonlocal interactions were ignored

The I-sites method does not consider the nonlocal context of the sequence fragment when predicting the local structure, nor does it predict nonlocal interactions. There is nonlocal information present in the sequence–structure motifs,[5] but it was not used here. At the level of local structure prediction, the neglect of nonlocal interactions probably compromises β strand prediction, as is true for conventional secondary structure prediction. At the nonlocal level, no attempt was made to maximize burial of hydrophobic surface area or optimize the packing of the predicted fragments, therefore our predictions contain bad contacts and are not compact. This is not really something that “went wrong,” but rather something that was deliberately ignored.

Next Steps

A useful way to view the results of the I-sites predictions are as a set of alternative conformational possibilities for each fragment of the sequence, each with a probability attached. Such a list of fragments can be used as a move set for an algorithm that searches global conformational space, similar to the use of nearest-neighbor fragments.[16] The usefulness of a move set can be measured by asking how much of the native structure, on average, is contained in the set of moves. If we consider all fragments predicted with confidence greater than zero for each target sequence, the list contains at least one true-positive fragment for 90% of all residues in the eight target proteins. On the other hand, only 51% of the residues in these target proteins occur in either helix or sheet. The CASP2 predictions suggest that I-sites fragments may be used successfully to greatly reduce conformational search space. For example, the backbone of target 38 (152 residues: PDB code 1ULO[15]) can be reproduced with an RMSD of 3.1 Å by using only 13 fragments from a submitted I-sites prediction. I-sites predictions may also be useful for differentiating possible threads of a sequence onto a structure in fold recognition applications.

Modifications and Resubmissions

A number of changes have been made in the I-sites method since submission of the CASP2 predictions. The database has been expanded from 392 sequence families to 471, and the length of the sequence profiles has been extended by 2 residues on either side. Both changes improve the prediction of local structure in an independent test set. Predictions for the five CASP2 targets that have not, to date, been solved (targets 6, 19, 21, 23, 26) have now been resubmitted by using the updated method, and are now available at the web site: http://PredictionCenter.llnl.gov/

CONCLUSIONS

This I-sites method was aimed primarily at predicting the parts of the protein left undefined by secondary structure predictions. In that respect it succeeded; positions predicted to be in loops by the PHD program are more accurately described by I-sites. By itself, I-sites failed to correctly predict as much local structure overall as the PHD server. This was due to its lackluster performance on helices. The values cited for the confidence of a prediction accurately reflect the probability of a residue being in a correctly predicted 8-residue fragment (Eq. 1). More information about the I-sites Library is available at the web site: http://ganesh.bchem.washington.edu/∼bystroff/Isites

REFERENCES

1. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
2. Salamov, A.A., Solovyev, V.V. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. J. Mol. Biol. 247:11–15, 1995.
3. Russell, R.B., Barton, G.J. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J. Mol. Biol. 234:951–957, 1993.
4. Rost, B., Sander, C., Schneider, R. Redefining the goals of protein secondary structure prediction. J. Mol. Biol. 235:13–26, 1994.
5. Han, K.F., Bystroff, C., Baker, D. Three dimensional structures and contexts associated with recurrent amino acid sequence patterns. Protein Sci. 6:1587–1590, 1997.
6. Han, K.F., Baker, D. Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc. Natl. Acad. Sci. U.S.A. 93:5814–5818, 1996.
7. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
8. Han, K.F., Baker, D. Recurring local sequence motifs in proteins. J. Mol. Biol. 251:176–187, 1995.
9. Gribskov, M., Luthy, R., Eisenberg, D. Profile analysis. Methods Enzymol. 183:146–159, 1990.
10. Bystroff, C., Baker, D. Improved local protein structure prediction using a library of sequence-structure motifs. J. Mol. Biol. (submitted), 1997.
11. Moult, J., Judson, R., Fidelis, K., Pedersen J.T. A large-scale experiment to assess protein structure prediction methods. Proteins 23:ii–iv, 1995.
12. Rost, B., Sander, C., Schneider, R. PHD: An automatic mail server for protein secondary structure prediction. Comput. Appl. Biosci. 10:53–60, 1994.
13. Al-Karadaghi, S., Hansson, M., Nikonov, S., Jonsson, B., Hederstedt, L. Crystal structure of ferrochelatase. EMBO J., in press 1997.
14. Vath, G.M., Earhart, C.A., Rago, J.V., et al. The structure of the superantigen exfoliative toxin A suggests a novel regulation as a serine protease. Biochemistry 36:1559–1566, 1997.
15. Johnson, P.E., Joshi, M.D., Tomme, P., Kilburn, D.G., McIntosh, L.P. Structure of the N-terminal cellulose-binding domain of cellulomonas fimi CenC determined by nuclear magnetic resonance spectroscopy. Biochemistry 35:14381–14394, 1996.
16. Simons, K.S., Kooperberg, C., Huang, E., Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J. Mol. Biol. 268:209–225, 1997.