Protein Science (1999), *8*:1127–1133. Cambridge University Press. Printed in the USA.

Enhanced protein fold recognition using secondary structure information from NMR

DANIEL J. AYERS,[1] PAUL R. GOOLEY,[2] ASAPH WIDMER-COOPER,[1] AND ANDREW E. TORDA[1]

[1]Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia
[2]The Russell Grimwade School of Biochemistry and Molecular Biology, The University of Melbourne, Melbourne VIC 3052, Australia

(RECEIVED November 2, 1998; ACCEPTED February 2, 1999)

Abstract

NMR offers the possibility of accurate secondary structure for proteins that would be too large for structure determination. In the absence of an X-ray crystal structure, this information should be useful as an adjunct to protein fold recognition methods based on low resolution force fields. The value of this information has been tested by adding varying amounts of artificial secondary structure data and threading a sequence through a library of candidate folds. Using a literature test set, the threading method alone has only a one-third chance of producing a correct answer among the top ten guesses. With realistic secondary structure information, one can expect a 60–80% chance of finding a homologous structure. The method has then been applied to examples with published estimates of secondary structure. This implementation is completely independent of sequence homology, and sequences are optimally aligned to candidate structures with gaps and insertions allowed. Unlike work using predicted secondary structure, we test the effect of differing amounts of relatively reliable data.

Keywords: chemical shift index; fold recognition; protein folding; protein structure prediction; protein threading; remote homology detection; secondary structure

There is no shortage of methods for predicting a protein's structure based only on its sequence (Böhm, 1996; Westhead & Thornton, 1998). Unfortunately, unless the sequence has significant sequence homology to something of known structure, one could not regard any of the methods as reliable (Lesk, 1997; Levitt, 1997; Marchler-Bauer et al., 1997). At the same time, they may be the only means of predicting structure in the absence of experimental data. A different problem arises for a protein sequence when a limited amount of experimental information is available. A typical case might come from a protein that yields a barely useful NMR spectrum. In this situation, one would like to use the available data, even if it is not suitable for conventional high resolution structure calculations. This has led to a series of approaches that have their roots in structure prediction, but attempt to incorporate very sparse experimental data such as a few intramolecular distance estimates (Smith-Brown et al., 1993; Aszódi et al., 1995; Lund et al., 1996; Skolnick et al., 1997). Typically, these methods produce low resolution structures and operate with the caveat that answers may sometimes be quite wrong.

Reprint requests to: Andrew E. Torda, Research School of Chemistry, Australian National University, Canberra ACT 0200, Australia; e-mail: Andrew.Torda@anu.edu.au.

Abbreviations: CSI, chemical shift index; DSSP, dictionary secondary structure of proteins; FSSP, families of structurally similar proteins.

Taking this theme further, NMR data may provide still more low resolution data. Even if a protein's structure will never be solved, its proton and heteronuclear NMR assignments may be largely determined. The relationship between chemical shift and structure has long been recognized (Pardi et al., 1983; Spera & Bax, 1991), but it can be better quantified. Given a fairly complete set of proton and heteronuclear chemical shifts, one can expect secondary structure assignments to be more than 92% accurate (Wishart et al., 1991, 1992; Wishart & Sykes, 1994a, 1994b). The aim of this work is to quantify the benefit this secondary structure information alone will have on a typical, unreliable, protein fold recognition method. This could be viewed as a way to improve the performance of a poor prediction method or it could be seen as exploiting experimental data that would not normally be sufficient to determine a structure. In either case, there are two reasons for this to be useful. First, there is a huge repository of protein chemical shift information (Seavey et al., 1991). Secondly, protein chemical shifts can be assigned in large proteins (more than 150 or 200 residues), even when relaxation effects would prevent acquiring reliable distance information.

In this work, secondary structure information was added to an existing protein sequence threading program (Huber & Torda, 1998). Threading means a protein sequence of interest is threaded through a library of known protein folds generating many trial structures (Jones et al., 1992; Sippl & Weitckus, 1992). These can be ranked by energy or score, and the most favorable ones taken as guesses

for the native structure. In practice, the method is complicated by having to allow for gaps and insertions to get the best sequence to structure alignment. Clearly, this methodology will fail when one encounters a new protein fold. One can also see that the potential accuracy is limited compared to ab initio methods. The best one can achieve with threading is limited by the most similar homologue in one's structure library. In practice, an alignment will rarely be perfect, and the modeled coordinates will be even worse. With these weaknesses, one should remember the reason for optimism is that new protein folds are found relatively infrequently (Holm & Sander, 1994a).

The sequence to structure alignment is calculated using a dynamic programming algorithm adapted from sequence to sequence alignment (Needleman & Wunsch, 1970). The score functions (force fields), however, are quite different to those that travel under the umbrella of knowledge-based force fields and that are commonly used for structure prediction (Sippl, 1995; Jernigan & Bahar, 1996; Jones & Thornton, 1996; Sippl & Flöckner, 1996; Torda, 1997). The score functions used here have no reliance on Boltzmann statistics and yield sequence to structure compatibility scores rather than energies. They have no explicit physical basis and are constructions built purely for protein fold recognition (Huber & Torda, 1998a).

To gauge the utility of secondary structure data, as might be obtained from NMR assignments, we have a series of calculations with synthetic data. We take a literature test set of protein sequences and folds, designed to test protein fold recognition methods (Rost et al., 1997). Since correct structures are known, one can generate secondary structure assignments as might have been obtained by NMR. More usefully, one can delete various fractions of the synthetic experimental data to create data sets spanning a range from better to worse than real data. As well as synthetic data, some examples of assigned secondary structures have been arbitrarily chosen from the literature and the fold recognition calculations performed with these real cases.

Results and discussion

Calculations using synthetic secondary structure data

Fold recognition was tested using the data set from Rost et al. (1997). This consists of 89 "probe" sequences, for which a structure is known. For each probe sequence, there is at least one protein of similar structure, but without significant sequence homology. Structural similarity was defined using R_{ali} of Equation 3 (Rost et al., 1997). For the tests here, we used the 26 probe sequences that had a structural homologue similar for at least 70% of their extent ($R_{ali} \geq 0.7$). For each probe sequence, there is at least one protein of similar structure, but without significant sequence homology. The structural homologue(s) are then hidden in a library of 723 mostly decoy structures. A fold recognition method should be able to take a probe sequence (and secondary structure information) and find a similar structure at first rank. Since methods do not work perfectly, one needs some way to estimate success. Here, we use $Q(R)$ (Equation 4 in Materials and methods) simply so one can directly compare with literature (Rost et al., 1997). This measures how often one finds the first correct homologue for a protein at a certain rank. If $Q(5)$ equaled 0.4, one would have a 40% chance of the first correct homologue being in the first 5 ranked places. Given the size of the data sets, a random ranking would give a $Q(10) < 2\%$.

The results of the calculations with synthetic data are shown in Figure 1. The better a set of results, the faster the curve rises toward 100%. The plot shows several curves, each with a different amount of data present. The best results are those with 100% correct secondary structure assignments. The worst results are with 0% secondary structure assignments (the prediction method alone). Each of the curves with simulated data has a different amount of secondary structure assignments randomly deleted/shortened (as described under Materials and methods).

In the absence of experimental data, the protein fold recognition method gives a correct answer in first place about 10% of the time. With ideal secondary structure, this improves to about 30%. One might consider how often the first correct guess is found within the top 10 places out of 723. With no experimental data, Figure 1 suggests one might expect to find a correct answer in the top 10 positions about one-third of the time. With perfect data, there is an 80% chance of the first correct prediction being in the top 10.

Experimental data, however, will not be ideal. One can observe the effect by considering results with secondary structure discarded. Surprisingly, it would appear from Figure 1 that with only 50% of the possible secondary structure data, the results are practically the same as with complete data. One could reasonably expect to find a correct answer about 60–70% of the time in the top 10 guesses. In fact, with only 25% of the possible secondary structure data, there is still about a 50% chance of finding a correct answer.

This work was geared to exploit sparse experimental data, and there is no intention of revisiting the territory of comparing sec-

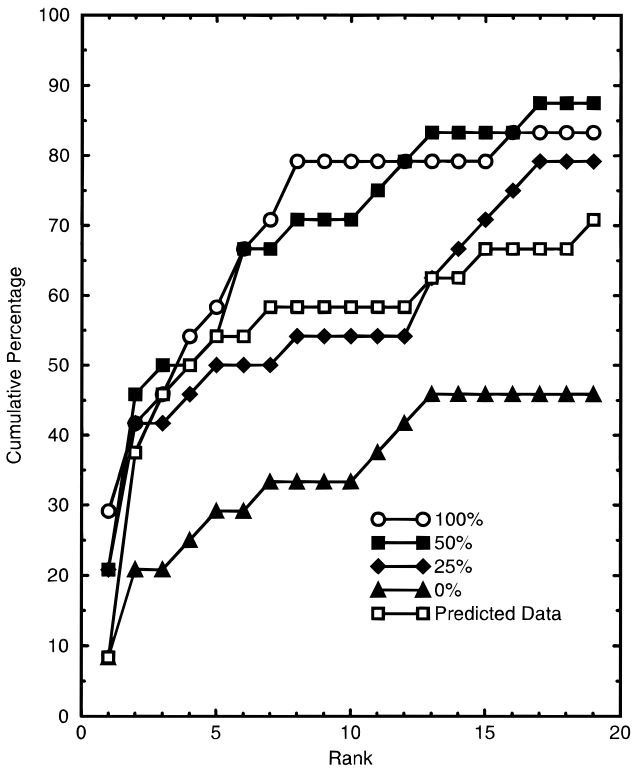

Fig. 1. Fold recognition, Q(rank), using synthetic data, of varying quality. Percentages in the legend give the amount of secondary structure information used; 100% is ideal data, whereas 0% is the protein score function alone. Predicted data are where the secondary structure data have come from secondary structure prediction.

ondary structure prediction methods. At the same time, it is easy to run just one calculation for the sake of quick comparison with methods using purely predicted secondary structure information. The probe sequences were sent to a popular secondary structure prediction server (Rost & Sander, 1993, 1994; Rost et al., 1994), and we accepted only high confidence predictions of α-helices and β-strand residues. This set of data performs slightly better than having about 25% of the simulated secondary structure.

Calculations with experimental data

The results described above are with artificial data. It is possible to do the calculations on real data from the literature, although this is less controlled than synthetic data. Secondary structure assignments for six proteins were taken directly from the literature as listed in Table 1. The sequences spanned a range from where one knew what to expect (from sequence homology) to those where very little was known about the sequence. Three of the examples had high sequence homology to a known structure, so they are listed as "known" in Table 1. Next, a sequence may have poor sequence identity to anything of known structure, but this is enough to identify the overall fold. This is the case with P450 reductase, which has 27% sequence identity to 5fx2, a flavodoxin of known structure. Oncostatin M is similar. It has only 25% sequence identity to anything of known structure, but with functional information it is assumed to be a cytokine. Finally, there is a protein in this set where almost nothing is known about the structure (staphylokinase) and one where the structure has just been published (gp41).

For each sequence with its secondary structure data, Table 1 shows the top 10 guesses from each list of 1,692 candidates. Where the fold can be reliably determined, a dagger is given next to the template structure. For each guess, a z-score is also quoted. This is a statistical measure, often used in fold recognition calculations, which measures how many standard deviations an observation is from the mean. In this application, a z-score of 1.8 for a structure prediction would mean that the score of this candidate was 1.8 standard deviations away from the average calculated over all 1,692 candidates.

In each of the three cases where the protein fold is known, Table 1 shows that the method has produced a correct guess in the first or second rank. These particular sequences adopt rather common folds, so one could point out that the same fold recognition could have been achieved using a simple sequence homology search. This is true, but not relevant. The methods here use the library folds as structural templates and will work in the absence of sequence homology. Quite impressively, Table 1 also shows a structural homologue for each of these sequences, which would probably not have been found using straightforward sequence searches.

The second category of protein sequences are those where one can be fairly certain of the overall fold, but with very little accuracy. Here, the method again seems rather successful. Oncostatin M is known to be a cytokine (Hoffman et al., 1996), and the sixth rank guess is 1lki, the cytokine leukaemia inhibitory factor. This is an example of how information from several sources can be complementary. Of all the known Protein Data Bank (PDB) structures, 1lki is the most homologous to Oncostatin M (25% sequence identity). This may be near the limit of significance. Combined with the functional similarity and the predictions from the secondary structure, the overall fold of the protein is not in doubt.

For P450 reductase it will not be possible to judge the results until the crystal structure is publicly released (Wang et al., 1996, 1997). It has 27% sequence identity with 2fx2 and other flavodoxins have been used as the starting point for homology modeling (Barsukov et al., 1997). The first rank guess, 2dri, is not actually a flavodoxin, but can be structurally aligned to one (1rcf) for 101 of its 169 residues with a root-mean-square difference of 3.3 Å. Of the top 10 postulated homologues, 5 structurally align to 1rcf for at least half their length, and 7 are nucleotide binding proteins involved in oxidative or reductive synthesis.

The structure for the ectodomain of gp41 consists of two long helices. Unfortunately, the list of predictions generated from the secondary structure data (Caffrey et al., 1997) does not contain any

Table 1. *Fold recognition using experimental data*[a]

	Known						Likely				Unknown		Solved	
	S100B		Flavodoxin		GRB2–SH2		Oncostatin M		C-P450 reductase		SAK		GP41	
Rank		Z score		Z score		Z score		Z score		Z score		Z score		Z score
1	1lpe	1.6	1rcf[b]	1.8	1ghu[b]	2.2	1vom	1.8	2dri	1.7	1prn	1.5	1bcfA	1.6
2	1symA[b]	1.5	1ovf[b]	1.7	1japA	1.8	1reqA	1.8	7aatA	1.7	1ikfL	1.5	1ier	1.6
3	2gdm	1.5	1pkn	1.7	1gsa	1.7	1reqB	1.8	1pfkA	1.7	1crl	1.5	1fha	1.6
4	2spcA	1.5	2fx2[b]	1.6	1ayaA[b]	1.6	1sly	1.8	1pea	1.7	1lte	1.5	2spcA	1.6
5	1pbxA	1.5	1nal1	1.6	1chd	1.6	1ribA	1.8	1xyzA	1.7	1bbdL	1.4	1lpd	1.6
6	2asr	1.5	2fcr[b]	1.6	1csyA[c]	1.6	1lki[b]	1.7	4xis	1.7	1cleA	1.4	1cpq	1.6
7	1babA	1.5	3rubL[c]	1.6	1cglA	1.6	1csc	1.7	1fcdA	1.7	1ovaA	1.4	1sctB	1.6
8	1outA	1.5	1lbiA	1.6	1iae	1.6	1ygp	1.7	1pkm	1.7	1loeA	1.4	1bucA	1.6
9	1ncx[c]	1.5	1fcmA	1.6	1lst	1.6	1derA	1.7	1ldnA	1.6	1lybB	1.4	1lht	1.5
10	1hdaA	1.5	1pkm	1.6	1slm	1.6	1fps	1.7	8atcA	1.6	1rinA	1.4	1emy	1.5

[a]For each test sequence with its experimental data, the PDB acquisition codes of the 10 top ranked predicted guesses are given.
[b]Correct guess, when known or predicted by sequence homology.
[c]Correct guess that is not predicted by sequence homology.

likely candidate folds. The gp41 column of Table 1 is dominated by helical proteins such as ferritins, globins, and four helix bundles but none are similar to the structure of gp41 (Caffrey et al., 1998). This is disappointing, but may not be surprising since the calculation was done on the monomer, but, in solution, the protein exists as a trimer. Finally, the predictions for staphylokinase are the most speculative. The list of guesses is dominated by all β proteins. Fortunately, one may eventually be able to judge these results since a modified version of this protein has been crystallized (Chattopadhyay et al., 1997).

One could say that the results show a remarkable improvement of the fold recognition capabilities of the SAUSAGE program. From a spectroscopist's point of view, one might say that the method provides a statistically useful method to use data that are acquired as part of the NMR assignment process. Realistically, there is about a 80% chance of finding a structural homologue in the first 10 guesses for a sequence of interest if there is a reasonable set of secondary structure assignments and if a homologue exists, which is similar for about 70% of the extent of the structures. This does mean that the results from SAUSAGE still require interpretation but much more confidence can be placed in the predictions.

In practice the results claimed here are probably not unrealistic. The method performs well even with very sparse data, and if Figure 1 presents any slant, it is pessimistic. For comparison to literature, we used the $Q(R)$ measure from Rost et al. (1997), but this only reflects the first correct homologue for a sequence. In practice, the first few guesses for a sequence may contain several examples of the same fold as shown by some of the examples of Table 1.

An interesting feature of this procedure is that a sequence to structure alignment implies a model for the sequence, based on the template structure. Obviously, this could be used as a starting point for homology modeling. More importantly, this means that one will often be able to make very quick judgments as to the reliability of the results. For example, there is a small amount of intramolecular distance information for the ectodomain of gp41. If one wanted to pursue this, the models here could be checked for compatibility with this or any other information.

The results also give some idea of the limits of the procedure. The most obvious restriction is that one can only hope to do well if the unknown structure is similar to a known structure. There are, however, more subtle limitations, which can be illustrated with the ectodomain of gp41. First, apparently simple patterns of secondary structure do not seem very informative. Complex patterns with mixed α-helices and β-strands of varying lengths seem quite effective at limiting the number of candidate folds. In contrast, simple patterns like the all α-helical data from the ectodomain of gp41 simply match to ferritins, globins, and α-helical bundles. Second, there is a limitation or error in our modelling. We took the sequence of the ectodomain of gp41 by itself, but in solution, this exists as a trimer (Caffrey et al., 1998). The physical consequences of this are clear. A hydrophobic residue in the sequence may be hidden from the solvent when such a residue is in the trimer, but the calculations will give the best results when such a residue is buried within some, probably incorrect, structure.

One might also note a disappointing aspect to the results in a numerical sense. The predictions for gp41 are wrong, but there is little indication of this. The z-scores quoted are no worse than for the other proteins. In fact, a z-score of 1.5 can be typical of correct fold recognition as shown by the first three columns of Table 1. Z-scores have achieved popularity in fold recognition calculations, but they assume a Gaussian distribution of scores. In practice, this is not the case for the aligned structures (data not shown), and the measure is not so useful. At the same time, it is clear from Table 1 that the rank order itself is useful.

In the more general field of protein structure prediction, there have been several efforts using predicted secondary structure and combining this with sequence similarity information. This is not really comparable to the work here, which blends secondary structure data with a fold recognition function operating on three-dimensional coordinates. Purely so as to give an example of what might be expected, there is the one calculation shown in Figure 1 using secondary structure predictions from the PHD server (Rost & Sander, 1993, 1994; Rost et al., 1994). The results cannot be compared with Rost et al. (1997) since we only used residues predicted with very high confidence and only considered pairs of proteins with somewhat higher similarity ($R_{ali} \geq 0.7$). Using these thresholds, there is a distinct improvement over the original fold recognition functions. Probably the only clear interpretation is that the local interaction terms in the structural score function are not as good as the predictions from the PHD server. If they were, one would not see the improvement when the secondary structure predictions are added.

One could experiment with different thresholds for prediction confidence, but this would run counter to the spirit of this work. The experimental data we have considered is sparse, but rarely incorrect. The methods here are not robust when given incorrect secondary structure (results not shown), and this is what one might expect (Dandekar & Argos, 1994). For some methods and certainly those here, incorrect data lead to worse results than missing data.

If one was dealing with unreliable data, the issue of weighting the experimental terms against the rest of the force field would be critical. For this work, k_{sec} of Equation 2 was chosen by gradually increasing it until no improvement could be seen on test data. It might be tempting to set k_{sec} higher since it is certainly more reliable than the score function, but this can harm the alignments. For example, 15 residues of a sequence may be correctly classified as helical, but in the best template, the corresponding helix is shortened by one turn (three or four residues). If k_{sec} is too high, 11 or 12 residues will be aligned to the correct helix, but a gap will be introduced and the remaining three or four residues forced to align to some other, incorrect helical region.

There is always the possibility that the experimental information is of varying quality with some reliable and some speculative. In that case, one could apply different weights (k_{sec}) to different sites. This is feature is implemented in the code, but not tested in this work.

From a technical point of view, it is most interesting to compare the methodology here with others who have used secondary structure (predicted) in fold recognition (Russell et al., 1996, 1998; Rice & Eisenberg, 1997; Rost et al., 1997). In these cases, a similarity matrix was used for physical properties such as secondary structure, much in the same manner as for sequence comparisons. In this work, similar discrete score methods were found to be quite sensitive to thresholds and definitions. For example, results would depend on when a structural motif was taken as a series of turns or when it was regarded as a fragment of α-helix (D.J. Ayers & A.E. Torda, unpubl. data). This problem has been seen by others (Di Francesco et al., 1997) and tackled by loosening the criteria for some secondary structure states. In this work, the

continuous formulation of Equation 1 was more robust than any function based on discrete secondary structure states.

One obvious question is what other data would easily fit into the framework of fold recognition. In the SAUSAGE implementation, properties that can be described as a structural characteristic of a site can be used for both alignment calculations and the ranking of guesses. This might include data such as the direct use of ^3J coupling or solvent accessibility. Properties that depend on specific pairs of amino acids (such as long-range nuclear Overhauser effects (NOEs) or knowledge of disulfide bonds) are more amenable to ranking generated sequence to structure alignments. Ultimately, if one keeps adding information, one should probably use something like a low resolution distance geometry or simulation approach (Smith-Brown et al., 1993; Aszódi et al., 1995; Skolnick et al., 1997). With enough experimental data, these considerations become irrelevant and conventional distance geometry or restrained molecular dynamics would be appropriate.

Finally, one can say that the score function/force field used here includes more signal than noise and is complementary to the information from recent sophisticated comparison matrices. If that is the case, there it should be possible to combine the methods, add sparse available data, and make even better use of experimental results.

Materials and methods

Score functions and calculations

Sequence to structure alignments and the ranking of trial structures were carried out using the SAUSAGE program, parameter sets, and two step approach of Huber and Torda (1998b). The experimental input consisted of only the protein sequence and secondary structure, regardless of their source. Secondary structure information was encoded in a term E_{sec} based on the backbone dihedral angle ψ of the template structure.

$$E_{sec} = \cos(\psi_0 - \psi) + 1 \tag{1}$$

where ψ_0 is the ideal backbone dihedral angle. For α-helices $\psi_0 = -47°$ (IUPAC-IUB, 1970) and for β-strands, $\psi_0 = 124°$, a value appropriate for either parallel or antiparallel sheets (IUPAC-IUB, 1970). The total score was given by

$$E_{tot} = E_{scr} + k_{sec} E_{sec} \tag{2}$$

where E_{scr} is the protein fold recognition score and k_{sec} gives a relative weight to the experimental secondary structure data. As implied by Equation 1, we adopt the convention, opposite to energy, that a more positive score is more satisfactory.

Sequence to structure alignments were carried out using a dynamic programming algorithm (Needleman & Wunsch, 1970), with E_{scr} given by the scoring function described as neighbor nonspecific (Huber & Torda, 1998b). Ranking of the generated alignments was done with E_{scr} given by the neighbor specific score function. Values of adjustable parameters are given in Table 2. Sequence similarity searches were done with the BLAST package (Altschul et al., 1990).

Measures of fold similarity and recognition

Calculations on synthetic data rely on having pairs of similar structures within a library of decoy proteins. One member of each pair is called the probe sequence, and the question is whether the fold recognition method can find the probe sequence's structural relative among the decoy proteins. This requires a criterion as to what constitutes a homologous pair. For this, the FSSP database with structural alignments was used (Holm & Sander, 1994a, 1994b, 1996, 1997, 1998). For a pair of aligned structures, one knows L_{ali}, the number of aligned residues, and L_1 and L_2, the number of residues in structures 1 and 2. Then a measure of similarity was taken from Rost et al. (1997).

$$R_{ali} = \frac{2L_{ali}}{L_1 + L_2}. \tag{3}$$

Successful of fold recognition was quantified following Rost et al. (1997) using the measure of the cumulative frequency of the first successful prediction, $Q(R)$.

$$Q(R) = 100 \sum_{r=1}^{R} \frac{N_{corr}(r)}{N_{probe}} \tag{4}$$

where $N_{corr}(r)$ is the number of first correct folds detected at rank r and N_{probe} is the number of probe sequences in the test set. A $Q(10)$ of 50 means that, considering all probe sequences, there is a 50% chance of finding the first correct homologue in the top 10 guesses. The measure is used for a comparison to literature, but one may note that it does not show when more than one correct homologue is detected.

Table 2. *Parameters used in the alignment and fold recognition calculations*[a]

	k_{gaps}	k_{ins}	k_{sec}
Alignment	1,000	1,000	100
Ranking	500	500	100

[a] k_{gaps} and k_{ins} are gap and insertion penalties used in sequence to structure alignments described in Huber and Torda (1998b); k_{sec} is given in Equation 2. All quantities are in arbitrary units.

Test data sets

Using the test set from Rost et al. (1997) requires an arbitrary decision as to what constitutes structural similarity. Using the R_{ali} measure given by Equation 3, and setting a threshold of 0.7 gave a set of 24 test sequences and 37 structural homologues hidden in the fold library as listed in Table 3.

For tests with real data, a library of 1,692 proteins was used as described in T. Huber and A.E. Torda (in prep.). This is essentially the whole PDB with only the most obvious sequence homologues removed.

For testing with real data, seven proteins (Table 4), each with recently published secondary structure assignments, were chosen. The authors' secondary structure assignments were used exactly as published with no further interpretation.

Input data

The input used by the SAUSAGE program consists of the sequence, as its one letter code, and the secondary structure data. The

Table 3. *Protein test sequences and structural homologues used for calculations with synthetic data*[a]

Sequence	Homologues						
1bct	1fos F	1pfi A	1lts C	1ifi			
1cpcA	1cpc B	2hbg	1bab A	1ash	3sdh A		
1cpt	1oxa	2hpd A					
1dvh	1cyj	451c	1ycc	1cc5			
1eaf	3cla						
1frpA	1imb A						
1fxd	1fca						
1gfc	1abo A	1pse					
1hryA	1hma						
1irk	1cdk A	1csn					
1lccA	1r69						
1mdyA	2dgc A	1ifi	1pfi A	1fos E	2ifo	1fos F	1lts C
1mjc	1bov A						
1pba	1ctf	2bop A	1ptf				
1plq	2pol A						
1pls	1btn	1dyn A					
1pyp	1ino						
1sso	1hum A	3il8					
1tie	2i1b	1bfg	1hce				
1xxaA	1urn A						
2cyp	1arv						
2pna	1lkk A						
3ebx	1cds						
3wrp	1fip A						

[a]The names are PDB acquisition codes with chain identifiers appended where necessary (Bernstein et al., 1977).

secondary structure data can be in one of three forms. It can be in either the very simple manual input, the output from the CSI program (Wishart & Sykes, 1994a) or the output from the PHD (Rost & Sander, 1993, 1994; Rost et al., 1994) server.

Synthetic secondary structure data

Ideal secondary structure data for the 22 test sequences was calculated using the DSSP program (Kabsch & Sander, 1983). To better simulate NMR data, only α-helical and β-strand secondary structure assignments were used. Other elements are less likely to be assigned from chemical shift data. To simulate poor data, secondary structure elements were randomly shortened until the desired fraction of data was discarded. Random deletion in this fashion meant that smaller secondary structure elements were likely to disappear completely as would be the case with real experimental data.

Predicted secondary structure data

The sequences for the test set of Rost et al. (1997) were submitted to the PhD server (Rost & Sander, 1993, 1994; Rost et al., 1994). From these secondary structure predictions, data were used for residues predicted to be in an α-helix or a β-strand with a confidence of eight or nine (on a scale of zero to nine).

The package is available as source code at ftp://ftp.rsc.anu.edu.au/pub/torda/sausage/README and documentation is at http://www.rsc.anu.edu.au/~torda/sausage.html

Note added in proof

We wish to note that the coordinates of Wang et al. (1996, 1997) for NADPH-cytochrome P450 reductase have now been released and are available from the Protein Data Bank as 1AMO.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Aszódi A, Gradwell MJ, Taylor WR. 1995. Global fold determination from a small number of distance restraints. *J Mol Biol 251*:308–326.

Barsukov I, Modi S, Lian L-Y, Sze KH, Paine MJI, Wolf CR, Roberts GCK. 1997. ^{1}H, ^{15}N and ^{13}C NMR resonance assignment, secondary structure and global fold of the FMN-binding domain of human cytochrome P450 reductase. *J Biomol NMR 10*:63–75.

Bernstein FC, Koetzle TF, Williams GJB, Myer EF Jr. 1977. The Protein Data Bank. *Eur J Biochem 80*:319–324.

Böhm G. 1996. New approaches in molecular structure prediction. *Biophys Chem 59*:1–32.

Caffrey M, Cai M, Kaufman J, Stahl SJ, Wingfield PT, Gronenborn AM, Clore GM. 1997. Determination of the secondary structure and global topology of the 44kDa ectodomain of gp41 of the simian immunodeficiency virus by multidimensional nuclear magnetic resonance spectroscopy. *J Mol Biol 271*:819–826.

Caffrey M, Cai M, Kaufman J, Stahl SJ, Wingfield PT, Gronenborn AM, Clore GM. 1998. Three-dimensional solution structure of the 44kDa ectodomain of SIV gp41. *EMBO J 17*:4572–4584.

Chattopadhyay D, Stewart JE, Smith CD, DeLucas LJ, Narayana SVL. 1997. Preliminary crystallographic study on a low molecular weight form of bacterial plasminogen activator staphylokinase. *Acta Cryst D53*:480–481.

Dandekar T, Argos P. 1994. Folding the main chain of small proteins with the genetic algorithm. *J Mol Biol 236*:844–861.

Table 4. *Proteins used for testing SAUSAGE with experimental data*

Protein	Abbreviation[a]	Size (residues)	Reference
Calcium bound S-100B	S100B	92	Smith and Shaw (1997)
Oxidised *Desulfovibrio desulfuricans* flavodoxin	Flavodoxin	148	Pollock et al. (1996)
Oncostatin M	Oncostatin M	203	Hoffman et al. (1996)
Grb2 SH2 domain	GRB2-SH2	112	Wang et al. (1996)
Plasminogen activator protein staphylokinase	SAK	136	Ohlenschläger et al. (1997)
Ectodomain of gp41	GP41	123	Caffrey et al. (1997, 1998)
FMN-binding domain of human cytochrome P450 reductase	P450 reductase	185	Barsukov et al. (1997)

[a]Abbreviated protein name, with chain identifiers appended where necessary, used in Table 4.

Di Francesco V, Garnier J, Munson PJ. 1997. Protein topology recognition from secondary structure sequences: Application of the Hidden Markov Models to the alpha class proteins. *J Mol Biol 267*:446–463.

Hoffman RC, Moy FJ, Price V, Richardson J, Kaubisch D, Frieden EA, Krakover JD, Castner BJ, King J, March CJ, Powers R. 1996. Resonance assignments for Oncostatin M, a 24-kDa α-helical protein. *J Biomol NMR 7*:273–282.

Holm L, Sander C. 1994a. Searching protein structure databases has come of age. *Proteins 19*:165–173.

Holm L, Sander C. 1994b. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res 22*:3600–3609.

Holm L, Sander C. 1996. The FSSP database of structurally aligned protein fold families. *Nucleic Acids Res 24*:206–210.

Holm L, Sander C. 1997. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res 25*:231–234.

Holm L, Sander C. 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res 26*:316–319.

Huber T, Torda AE. 1998. Protein fold recognition without Boltzmann statistics or explicit physical basis. *Protein Sci 7*:142–149.

IUPAC-IUB Commission on Biochemical Nomenclature. 1970. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969). *Biochemistry 9*:3471–3479.

Jernigan RL, Bahar I. 1996. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol 6*:195–209.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.

Jones DT, Thornton JM. 1996. Potential energy functions for threading. *Curr Opin Struct Biol 6*:210–216.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Lesk AM. 1997. CASP2: Report on ab initio predictions. *Proteins Suppl 1*:151–166.

Levitt M. 1997. Competitive assessment of protein fold recognition and alignment accuracy. *Proteins Suppl 1*:92–104.

Lund O, Hansen J, Brunak S, Bohr J. 1996. Relationship between protein structure and geometrical constraints. *Protein Sci 5*:2217–2225.

Marchler-Bauer A, Levitt M, Bryant SH. 1997. A retrospective analysis of CASP2 threading predictions. *Proteins Suppl 1*:83–91.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarity in the amino acid sequence of two proteins. *J Mol Biol 48*:443–453.

Ohlenschläger O, Ramachandran R, Flemming J, Gührs K-H, Schlott B, Brown LR. 1997. NMR secondary structure of the plasminogen activator protein staphylokinase. *J Biomol NMR 9*:273–286.

Pardi A, Wagner G, Wüthrich K. 1983. Protein conformation and proton nuclear-magnetic-resonance chemical shifts. *Eur J Biochem 137*:445–454.

Pollock JR, Swenson RP, Stockman BJ. 1996. ^{1}H and ^{15}N NMR resonance assignments and solution secondary structure of oxidised *Desulfibrio desulfuricans* flavodoxin. *J Biomol NMR 7*:225–235.

Rice DW, Eisenberg D. 1997. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol 267*:1026–1038.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins 19*:55–72.

Rost B, Sander C, Schneider R. 1994. PHD—An automated mail server for protein secondary structure prediction. *CABIOS 10*:53–60.

Rost B, Schneider R, Sander C. 1997. Protein fold recognition by prediction-based threading. *J Mol Biol 270*:471–480.

Russell RB, Copley RR, Barton GJ. 1996. Protein fold recognition by mapping predicted secondary structures. *J Mol Biol 259*:349–365.

Russell RB, Saqi MAS, Bates PA, Sayle RA, Sternberg MJE. 1998. Recognition of analogous and homologous protein folds—Assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng 11*:1–9.

Seavey BR, Farr EA, Westler WM, Markley JL. 1991. A relational database for sequence-specific protein NMR data. *J Biomol NMR 1*:217–236.

Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol 5*:229–235.

Sippl MJ, Flöckner H. 1996. Threading thrills and threats. *Structure 4*:15–19.

Sippl MJ, Weitckus S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins 13*:258–271.

Skolnick J, Kolinski A, Ortiz AR. 1997. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J Mol Biol 256*:217–241.

Smith SP, Shaw GS. 1997. Assignment and secondary structure of calcium-bound human S100B. *J Biomol NMR 10*:77–88.

Smith-Brown MJ, Kominos D, Levy RM. 1993. Global folding of proteins using a limited number of distance constraints. *Protein Eng 6*:605–614.

Spera S, Bax A. 1991. Empirical correlation between protein backbone conformation and Cα and Cβ ^{13}C nuclear magnetic resonance chemical shifts. *J Am Chem Soc 113*:5490–5492.

Torda AE. 1997. Perspectives in protein fold recognition. *Curr Opin Struct Biol 7*:200–205.

Wang M, Roberts DL, Paschke R, Shea TM, Masters BSS, Kim J-JP. 1997. Three-dimensional structure of NADPH-cytochrome P450 reductase: Prototype for FMN- and FAD-containing enzymes. *Proc Natl Acad Sci USA 94*:8411–8416.

Wang Y-S, Frederick AF, Senior MM, Loyns BA, Black S, Kirschmeier P, Perkins LM, Wilson O. 1996. Chemical shift assignments and secondary structure of the Grb2 SH2 domain by heteronuclear NMR spectroscopy. *J Biomol NMR 7*:89–98.

Westhead DR, Thornton JM. 1998. Protein structure prediction. *Curr Opin Biotechnology 9*:383–389.

Wishart DS, Sykes BD. 1994a. The ^{13}C chemical-shift index: A simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J Biomol NMR 4*:171–180.

Wishart DS, Sykes BD. 1994b. Chemical shifts as a tool for structure determination. *Methods Enzymol 239*:363–392.

Wishart DS, Sykes BD, Richards FM. 1991. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J Mol Biol 222*:311–333.

Wishart DS, Sykes BD, Richards FM. 1992. The chemical shift index: A fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry 31*:1647–1651.