Use of β–strand Interaction Pseudo-Potentials in Protein Structure Prediction and Modelling

Tim J.P. Hubbard

Centre for Protein Engineering, Medical Research Council Centre,
Hills Road, Cambridge, CB2 2QH, UK.

Abstract

A residue-residue interaction pseudo-potential has been developed specific for protein β–sheets. The potential is derived by scoring the occurrence of all i–(j-2...j+2) residue-residue pairs between any two interacting β–strands, subdividing according to 4 classes of hydrogen bond pattern.

The potential can be useful in distinguishing between correct and incorrect alignments between β–strands in a predicted protein structure. It is not unduly influenced by the quality or completeness of a model, because it is not based on atom-atom distances.

The potential can also be used directly to predict a β–strand contact map for any protein sequence. Maps generated when considering a single protein sequence have a poor signal to noise ratio, however when a prediction is smoothed by considering a family of homologous sequences, native and non-native β–strand pairings can be partially distinguished.

1: Introduction

Although the database of experimentally determined protein structures is now expanding quite rapidly [1, 2], the ratio of protein sequences to known structures is still increasing due in part to the increasing activity of genome sequencing projects worldwide. The prediction of protein structure from sequence remains an unsolved problem. There is now however a consensus that the number of different protein 'folds' is both finite and quite small [3], so the chance of a new protein sequence having a similar structure to one that is already known is probably increasing. As a result there has been a shift in research from 'fold prediction' to 'fold recognition'.

1.1: Fold Recognition

Fold recognition can be thought of as a development of homologous modelling, whose basic assumption is that where there is a significant sequence homology between sequences there is structural homology. This assumption has not so far been contradicted by nature. Automatic protein modelling programs have made it easy to generate useful 3-D structural models for any sequence with a strong homology (>50% residue identity) to a sequence of known structure [4, 5]. However, structural homology has also been seen when there is no apparent sequence homology. A recent example of this is the structures of the proteins Actin, the N-terminal ATPase fragment of 70-kDa Heat Shock Cognate (Hsc70) and Hexokinase which have turned out to be very similar (particularly Actin and Hsc70), despite having no significant sequence homology by standard methods [6].

Such examples have led to two developments. Firstly a number of speculative homology models have been made based on barely significant sequence homologies to proteins of known structures. A successful example of such a prediction was that of the HIV protease model based on structures of the aspartic proteases family [7]. A more speculative prediction which had not yet been confirmed or refuted by experiment is the model of the C-terminal of Hsc70 based on the HLA binding fold [8]. Secondly there has been an explosion of methods to extend the detection of structural similarity beyond what can be inferred from normal sequence alignment [9-11]. These methods are all ways of increasing the amount of information used in the matching 1-D information (sequence of unknown structure) with 1-D and 3-D information (sequence and structure) and will subsequently be referred to as 'threading' algorithms.

Therefore at present, with or without threading, the *accuracy* of the folds predicted when there is no clear sequence homology will vary greatly. The *quality* of any structural model will also vary considerably since although modelling programs build rough structures very quickly, refining models is difficult (particularly with the uncertainty of a prediction) and time consuming. Structural models of uncertain reliability will not all be of natural sequences. A number of protein designs have been made where a designed sequence has been modelled onto an example of the fold they are expected to adopt [12-14]. In these cases there is no certainty that the sequence concerned will fold into any stable structure let alone the one modelled.

How can such models be tested in a way which assesses the *accuracy* of a fold without being biased by the *quality* of the model? A number of authors have developed methods to test protein structures. In general these have relied mainly on the distribution of solvent accessible and inaccessible residues [15-18] or the quality of the residue packing [19]. A more recent derivation of one of the threading methods tests the compatibility of a residue with its local environment [20]. In all these cases the results will at least partly reflect the *quality* of the model.

A number of authors have derived residue-residue preference tables based on either a single cut-off distance defining a contact or deriving preferences for ranges of residue separation distances [21-23]. More recently contact preferences have been subdivided according to the number of residues separating the contacting residues in the chain [24]. In at least one case such contact preferences have been used to evaluate models [19].

In the work described here much more specific preference tables are derived to describe each of the local pairwise residue positions between two adjacent β−strands. The approach is similar to that adopted previously [25, 26] but takes advantage of the considerably larger structural database. The pseudo-potentials derived from these tables do not depend directly on arbitrary distance cut-offs and can be applied to the β−sheet parts of a structural model in a test which is relatively independent of model quality. Such pseudo-potentials are not limited to model testing but may be incorporated directly into threading algorithms.

1.2: Fold Prediction

The philosophy of fold recognition and threading is based on the assumption that we have already observed a substantial percentage of possible folds experimentally. If there are ~1000 folds [3] in nature and only ~100 have currently be determined experimentally, then, even if these are the most popular folds for proteins, fold recognition will still only be of use in the structure prediction of a sub-set of all sequences. The test case analysis of the 182 proteins identified in Yeast chromosome III [27, 28] showed that without threading only 14% of sequences could be associated with a protein of known structure. However there is now a substantial group of families of homologous sequences for which no structure is known and 24% of chromosome III sequences aligned with one of these families. If prediction of a protein structure from a single sequence is still impossible, prediction of a structure consistent with a family of sequences looks much more tractable. Multiple sequence alignments have been used in prediction for a long time [29] and have recently been used to improve secondary structure prediction [30, 31] (1-D prediction).

Multiple sequence information can be used to improve the predictive power of the β−interaction potential derived here too. When applied directly to a single protein sequence to predict a β-strand contact map the poor signal to noise ratio makes the prediction of little use. If however the prediction is smoothed by considering a family of homologous sequences native and non-native β-strand pairings can be partially distinguished, particularly for short protein chains. Since a β-strand pair interaction implies a large number of main chain atomic distances this effective 2-D prediction may provide sufficient unambiguous constraints to allow a 3-D prediction of the protein structure of short protein chains with high β-strand content using distance geometry techniques.

2: Analysis and results

2.1: Data analysed

Raw frequencies of occurrence of residue pairs in interacting β−strands were derived from a combination of the Brookhaven Protein Data Bank (PDB) [1, 2] and Homology derived secondary structure of proteins (HSSP) databases [32] . To compensate for the small number of β−strands in the structural database (PDB) relative to the number of parameters being derived, homologous sequence information (HSSP) was included where possible, weighted to correct for degeneracy.

Dictionary of protein secondary structure (DSSP) files [33] were generated for a sub-set of the protein structure files contained in release 63 (January 1993) of the (PDB) and pre-release protein structure files available at that time. The sub-set was defined by grouping all chains with residue identity (Rid) of >90%. Only the protein chain with the best crystallographic resolution in each group remained a member of the sub-set. All non-protein PDB entries, non-experimental entries, entries with poor geometry [34] and α-carbon only datasets were excluded. Of the 375 chains selected by this criteria, 293 had at least one β-strands pair and these formed the 90% Rid sub-set.

In order to further correct for over representation of similar chains this sub-set was clustered according to a structural homology threshold [32] corresponding to roughly 30% Rid. Each member of a cluster was given a weight such that the sum of the weights for each cluster was 1. 179 clusters were found with between 1 and 15 members.

The database was enlarged beyond the sequences found in PDB alone by the use of the HSSP database. For most PDB files there is an HSSP file which contains an alignment of all homologous sequences found in the SWISSPROT protein sequence database [35]. HSSP sequences with unknown structure were added to each cluster with appropriate weighting, taking the DSSP secondary structure summary from the closest structural member. HSSP sequences with >90% Rid to the closest DSSP member were not used.

The list of clusters at 30% Rid were used to generate lists for jack-knife cross validation. 10 parameter and 10 evaluation lists of chains were made: each parameter list contained 90% of the clusters and the corresponding evaluation list the remaining 10%. Evaluation statistics reported are the sum of the results for each evaluation list of chains, using parameters derived from the corresponding mutually exclusive parameter list of chains.

2.2: Generation of β–strand residue-residue propensity tables

The DSSP file for each selected protein chain was processed to sub-classify each β–strand amino acid residue pair into one of 4 classes and according to N-terminal, C-terminal order as shown in Fig□1. DSSP files specify the conformation and main chain hydrogen-bonding of each residue. A total of 10,600 antiparallel and 3,150 antiparallel residue pairs were counted from the DSSP files alone.

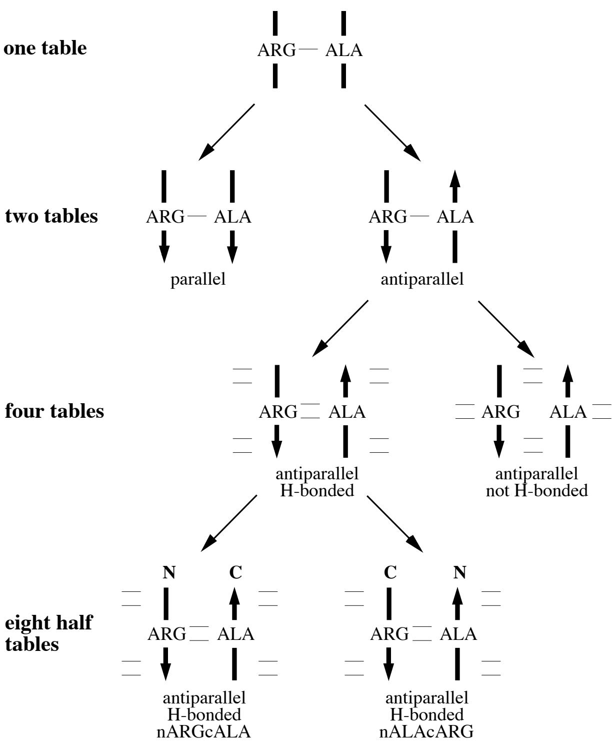

Figure 1 Subdivisions of β–strand residue pairs by parallel/antiparallel; hydrogen-bonding pattern and with respect to N and C termini. Some subdivisions have been omitted from the figure for clarity.

The use of HSSP sequence information relies on the alignment to the sequence of known structure being both accurate and meaningful. Wherever the local sequence homology was low or there were nearby insertions/deletions the extrapolation of structural information about that residue was regarded as unreliable and it was rejected. 59% of possible HSSP residues passed this test resulting in a further 149,650 antiparallel and 29,720 parallel residue pairs being counted.

For each accepted β–strand residue-residue pair (**i–j** in Fig□2), the occurrence of each residue-residue combination was counted for each of the 5 pair positions between adjacent β–strands shown.

The 20 tables of raw residue-residue counts created were converted into propensities according to (1) where, for the PDB files considered, **n(a,b)** is the number of cases where residues **i** and **j** in Fig 2 were amino acids **a** and **b**; **n(a)** is the total number of amino acids **a**. The sums of **n(x,y)** and the **n(x)** are the total number of **i,j** cases considered and the total number of amino acids in the PDB respectively.

$$prop(a,b) \;=\; -\log\left(\frac{n(a,b)\left(\sum_{x=all\,res} n(x)\right)^{2}}{\left(\sum_{x=all\,res}\sum_{y=all\,res} n(x,y)\right)\left(n(a)n(b)\right)} \right) \qquad (1)$$

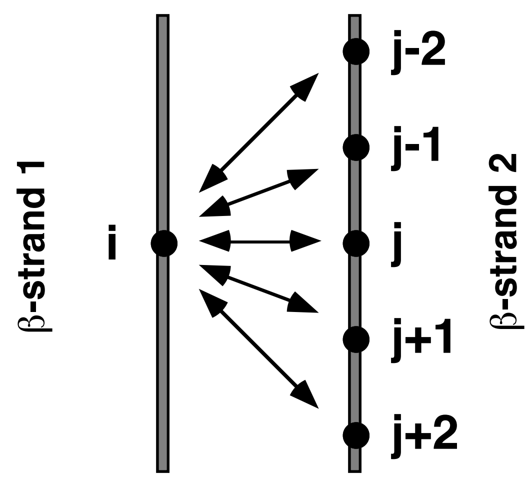

Figure 2 For each β–strand residue pair **ij**, the occurrences of pairs **ij-2, ij-1, ij, ij+1, ij+2** are counted in separate tables.

2.3: Evaluation of propensity tables

A qualitative measure of similarity between any two tables can be defined as the correlation coefficient (CC) between the 400 equivalent residue-residue elements in each table (fig 3). The average of the CCs calculated for all possible pairs of tables (190 pairs) is about 0.3. This suggests that there are significant differences in residue-residue propensities for the different β−strand structural sub-classes defined in Fig 1. However, since the total number of β−strand residue pairs is statistically small, it is important to establish that this difference between the tables is not just due to noise.

To test this the protein chains in the original list were randomly segregated into two half lists and a separate set of 20 tables calculated from each. CCs were then calculated between all pairs of tables generated from each half list. If the difference between tables was just noise then CCs between pairs of tables of identical structural sub-classes generated from each list might be expected to be similar to those CCs for pairs of different tables. The average of the CCs between different tables is 0.3 (fig 3a) whereas the average of the CCs between identical tables is 0.55 (fig 3b). This shows the large amount of noise in the tables (if there was no noise CC=1.0 for fig 3b), but the difference between these averages indicates that the propensities do contain useful information about protein structure.

Fig 3c shows CCs calculated within each table where the equivalent elements are pairs of **ij** and **ji** residue pairs (i.e. Ala-Val:Val-Ala). Since the only distinction between such pairs is which residue is N-terminal to the other and β−sheets are symmetrical from this point of view it is surprising that such large differences are observed, but justifies the sub-classification in this way.

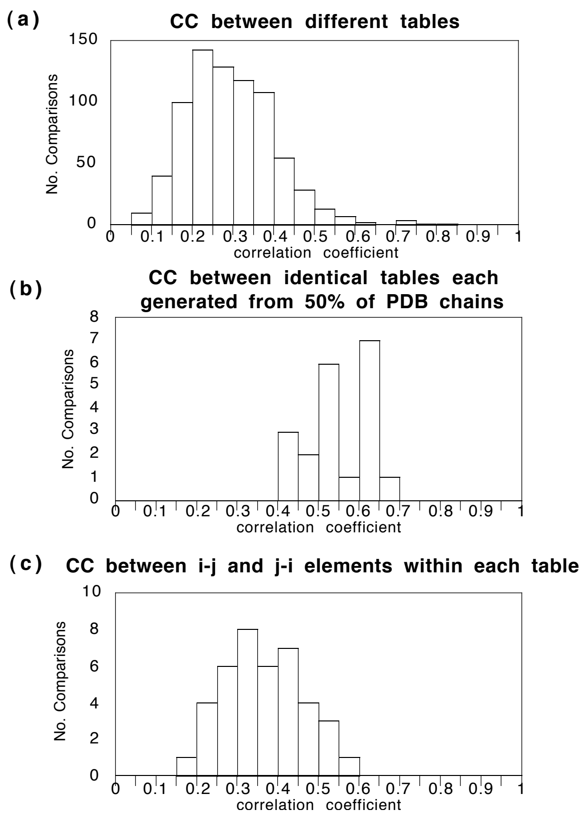

Figure 3 Correlation coefficients (CC) were calculated between all 400 elements in a pair of tables (a,b) or between 190 elements where **ij≠ji** in an individual table (c). Distribution of CCs are shown (a) for all pairwise comparisons between tables derived from structurally distinct data; (b) for all pairwise comparisons between tables derived from structural identical, but independent data; (c) for each table.

2.4: Construction of an interaction pseudo-potential from propensities tables

For any two segments of a protein sequence, a pseudo-potential for their interaction in β−sheet conformation can be defined as the sum of the propensities from the relevant tables for all pair interactions in Fig 2, for each residue in each segment, divided by the total number of interactions summed. For a pair of segments there are four possible β−strand-strand conformations as shown in Fig 1 (either parallel or antiparallel, each with two alternative hydrogen-bonding arrangements) and so four alternative pseudo-potential scores can be calculated and compared.

2.5: Selectivity of pseudo-potential when applied to known β−strands

Scores were calculated for all β−strand pairs in the DSSP database for each of the 4 alternative conformations shown in Fig 1. If the potentials are selective, the conformation observed in the protein would be expected to have the lowest score. Fig 4a and 4b show the distribution of differences in score between correct and incorrect conformations. Fig 4b shows differences in score between the two conformations with correct sheet type but correct and incorrect hydrogen-bonding. Fig 4a shows differences in score between the correct conformation and the best of the two conformations of incorrect sheet type. It can be seen that the potential's predictions for sheet type are on average about 75% accurate and between alternate hydrogen-bonding are about 85% accurate. The reliability of the prediction is however related to the difference in score. A greater than average difference in score gives a higher predictive accuracy for both types of classification.

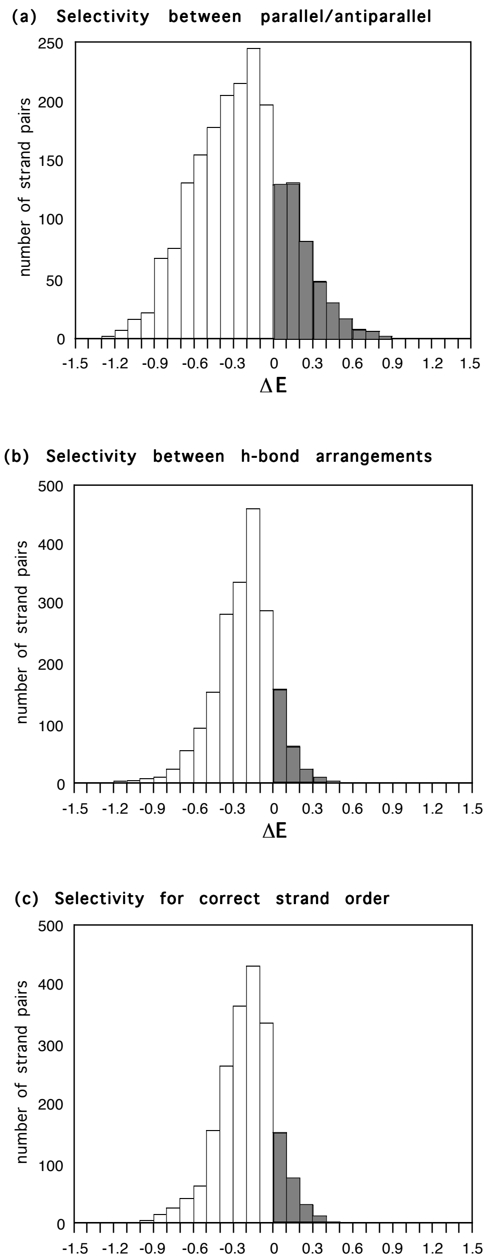

Figure 4 The distribution of differences in score between (a) correct and incorrect sheet type (parallel or antiparallel), (b) correct and incorrect hydrogen-bonding arrangements and (c) natural and reverse strand order (with respect to N and C termini), for all strand pairs in the database used. A -ve difference indicates the observed conformation has the lower score.

As an experiment, the score was also calculated for a reversed strand order with respect to N and C termini (fig 4c), to test the effect of the **ij**, **ji** sub-division discussed in 2.3. Remarkably, in 85% of cases, the naturally observed strand order in the sequence has the lower score.

2.6: Ability of pseudo-potential to identify correct β–strand alignments

The above test shows that the pseudo-potentials contain structurally distinct information, but not whether they can be usefully applied to real problems. A much more significant ability would be to distinguish between correct and incorrect strand alignments, without knowing the length of the strand-strand interaction in advance. To test this, for each known pair of interacting strands, the

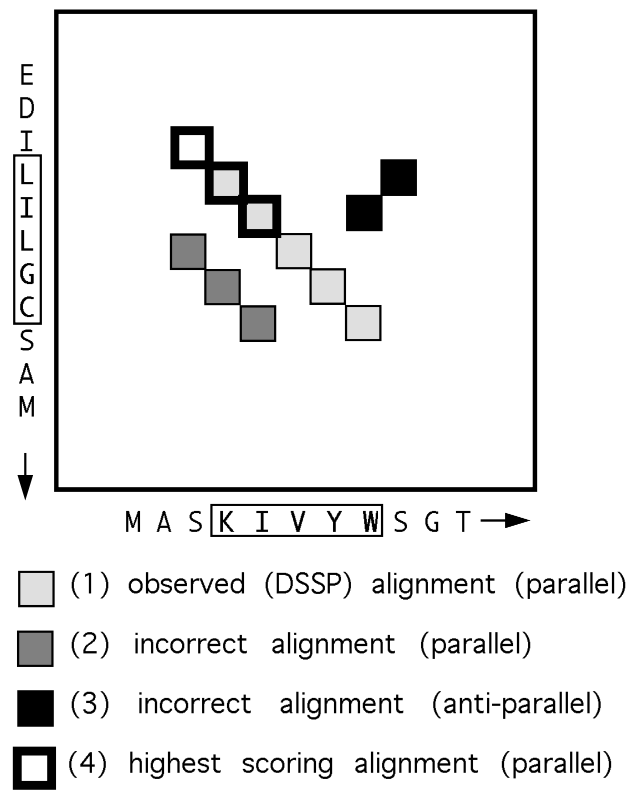

Figure 5 Method for searching local alignment space around a β–strand pair (KIVYW and LILGC) which interact to form a parallel β–sheet. Arrows following the sequences point towards the C-terminus of the protein. Each box indicates an **ij** alignment between corresponding residues on each axis. (1) indicates the correct alignment. (2) indicates an misalignment of the strands by 3 residues. (3) indicates an alignment of the wrong sheet type. (4) indicates an alignment which aligns the correct residues, but which is a different length and overlap.

scores for all possible interactions of a length of 2-8 residues were compared and the one with the best score recorded. The sequence-sequence area searched included an additional 3 or 6 residues beyond the ends of each strand. Fig 5 outlines the classification of the types of alignment that can result from such a search. The offset of the alignment with the best score from the correct alignment was then recorded and the distribution across the database is shown in Fig 6. In Fig 5, alignments of type (1) and (4) would give an offset of 0 in Fig 6b (correct parallel alignments); type (2) would have an offset of 3; type (3) would be scored in the last (incorrect sheet type) column.

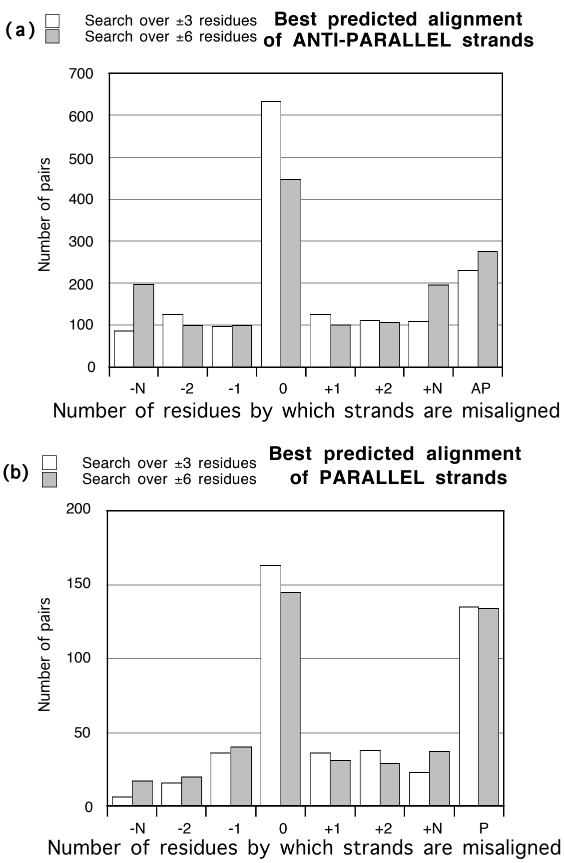

Figure 6 The difference between the alignment of a pair of sequences (fig 4) giving the best score using the β–strand interaction potential and the real alignment was calculated. The distribution for known antiparallel (a) and parallel (b) β–strand interactions is shown. The final column shows the number of pairs predicted with the wrong β–interaction type (parallel when real interaction is antiparallel and *visa versa*.). The relatively high frequency of mispredictions of parallel as antiparallel is due to the relative occurrence of antiparallel/parallel interactions in the database of 4:1.

It is important to appreciate that a prediction of type (4) gives a correct alignment despite being of a different length and despite overlapping the ends of the true β–strands.

The results from Fig 6 show that for searches over an additional 3 residues the correct alignment has the best score in ~45% of cases where the sheet type has been predicted correctly. The results are similar for parallel and antiparallel sheet classes and offsets of ±1, ±2, ±3 etc. all occurring with roughly equal frequency. If an additional 6 residues at the end of each strand are considered instead of 3, despite the increase in the number of possible alignments, in ~35% of cases the correct alignment is predicted.

This shows that there is significant recognition information in a substantial fraction of β–strand-interactions even when considering only one sequence. This type of analysis can therefore be usefully applied to real modelling problems: β–sheets in protein models can be analysed by comparing the potential for the modelled β–strand alignments with other possible surrounding alignments.

2.7: Use of pseudo-potential to predict a β–sheet contact map directly

Rather than just applying the potential to a local

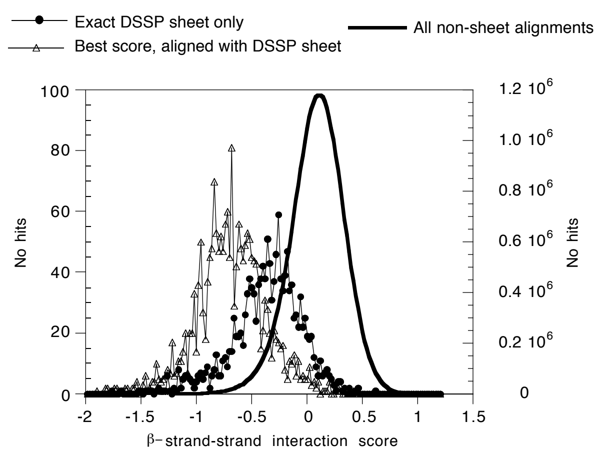

Figure 7 The distribution of scores generated from the potential for all possible pairwise combinations of fragments in a database sequence. The scores are classified as (•) An exact length match with an observed DSSP β–interaction; (Δ) the best score aligning with a DSSP β–interaction, but not necessarily of the same length; (–) all scores for non-β sequence and misaligned β–interactions. Note that (•) and (Δ) use the scale on the left whereas (–) uses the scale on the right.

sequence window around a known β–interaction it can of course be applied globally. Figure 7 shows the distribution of scores for β–interactions and non-β–interactions. Whereas it is clear that the peaks are separated, since there are 10^4 more non-β than β–interactions, the tail of the former swamps the peak of the latter. It should be noted that better scores are frequently found for interactions that align with real β–interactions but have a different length (i.e. as Fig 5, type (4) alignments). The distribution of these best alignments (Δ) has a peak better separated from the non-β peak (–) than the (•) peak however this does not eliminate the tail overlap problem. It would therefore seem that this potential cannot by itself predict a β–interaction contact map, however the use of multiple sequence information as shown in the following example substantially improves the situation.

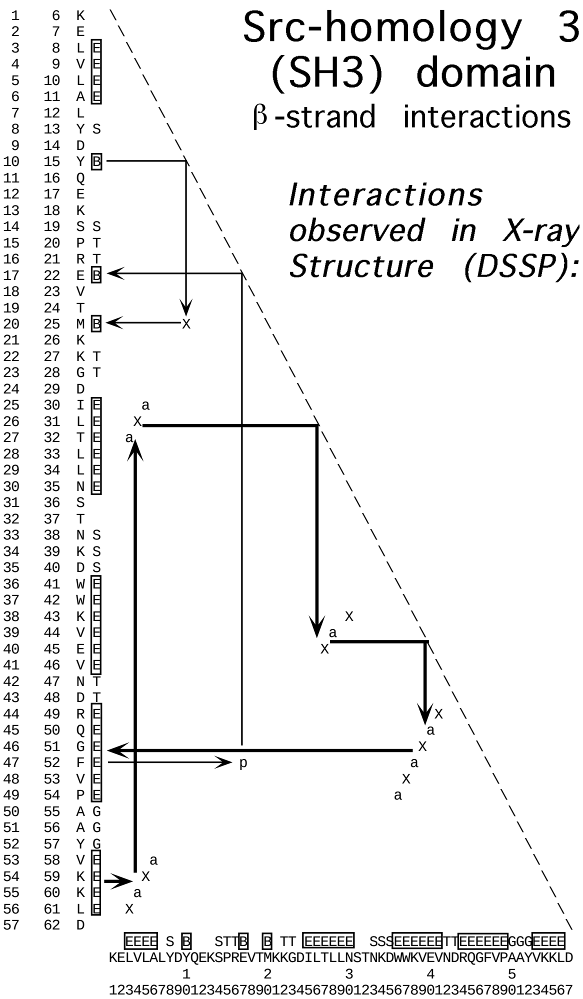

Figure 8 The β–strand contact map for Src-homology 3 (SH3) domain protein derived from the known structure also showing the connectivity of the sheet. Contacts containing 'a' are antiparallel and 'p' parallel. The 4 columns on the left are sequence index (1-n), residue number (from PDB file), aminoacid, DSSP summary information [B=isolated β–bridge, E=strand, G=3_{10} turn, H=helix, S = bend, T=turn].

The example shown is for the Src-homology 3 (SH3) structure. This was chosen here since it is a small, recently solved [36] mainly β–sheet protein that has been the subject of a number of secondary structure predictions [37, 38] and recent debate [31, 39-41].

Fig 8 shows the β–strand contact map derived from the known structure. It can be seen that there are 5 main

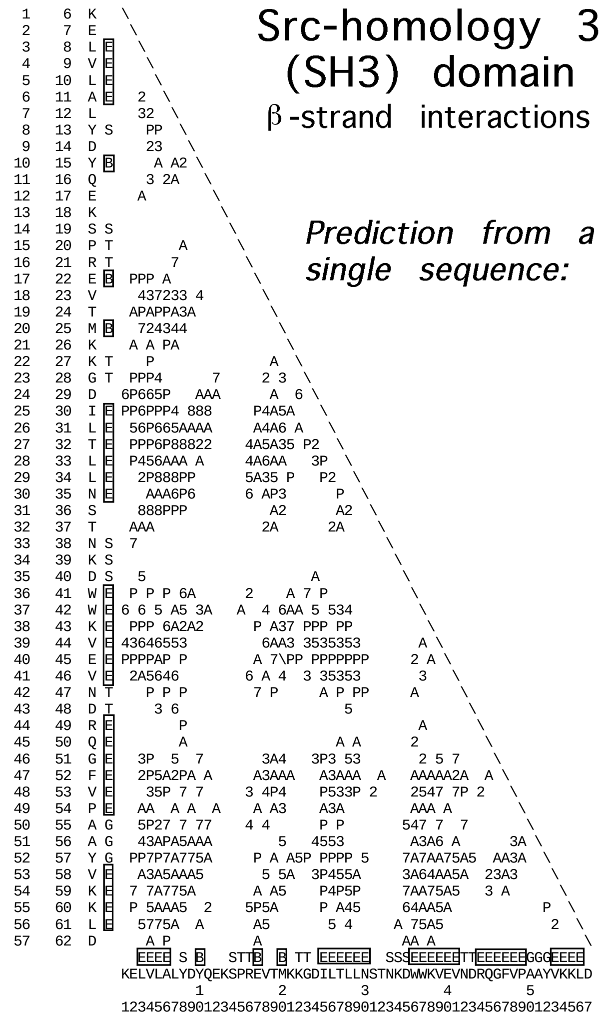

Figure 9 Predicted β–strand contact map for SH3 from a single sequence. The predicted contacts are identified as A5A5A and P4P4 for a length 5 antiparallel interaction and a length 4 parallel interaction respectively. No information about the score of the predicted contact is shown.

β−strands and 4 major antiparallel interactions between them constructing a single β−sheet. Two isolated β−bridges are also seen.

Fig 9 shows the result of a prediction from a single SH3 sequence, recording all scores above an arbitrary threshold. It can be seen that contacts are predicted almost everywhere and certainly for almost all possible β−strand interactions. A prediction of this sort contains no more information than a secondary structure prediction.

Figure 10 shows the result of smoothing with the scores calculated for 13 other sequences with sequence identities of between 90% and 36% to the sequence of the known structure. The smoothing procedure is that for any position on the contact map, the score for *all* sequences

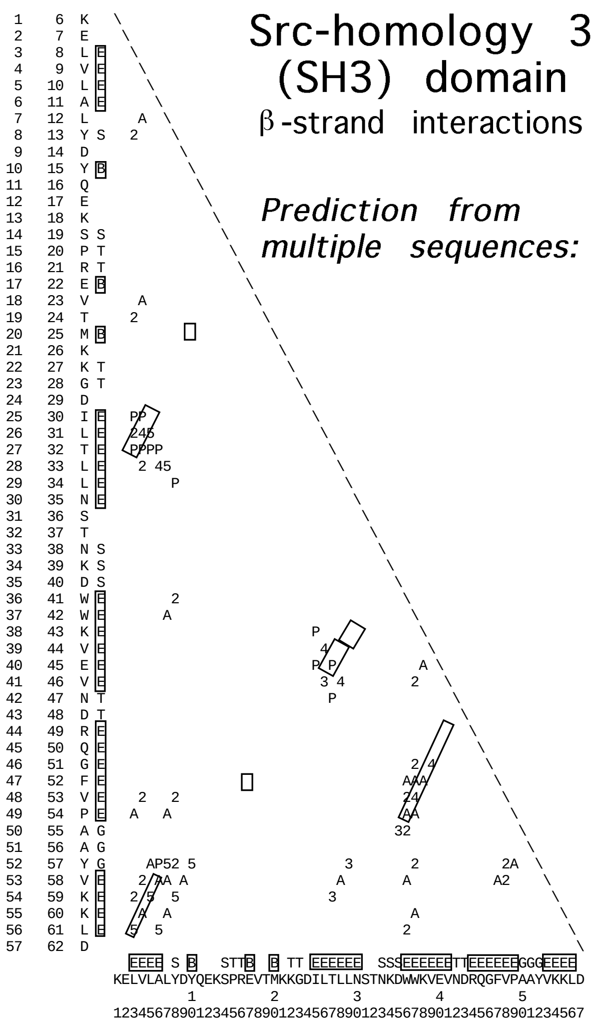

Figure 10 Predicted β−strand contact map for SH3 smoothed by considering multiple sequences. The Diagonal boxes show the location of the true β−interactions (from fig 7). Predicted contacts are identified as in figure 9.

must be above the threshold. The effects are quite dramatic, leaving some noise (mainly scores for length 2) but longer hits only in regions where a β−interaction is naturally observed. For two of these regions there is an exact alignment of one of the matches to the native alignment (3-6:56-53) and (36-41:49-44). The 3rd major interaction (27-30:40-38) (which includes a β−bulge) is predicted by the potential but incorrectly as a parallel interaction. However, since β−strands joined by a loop of <10 residues only ever form antiparallel interactions, this prediction can be resolved correctly.

The last major interaction (3-5:27-25) is also predicted but once more as parallel rather than antiparallel. If an attempt was made to model this protein on the basis of the predicted interactions, this misprediction might be resolved correctly since a single parallel interaction in a sheet which is otherwise antiparallel might lead to topological inconsistencies.

At present this contact map prediction is more qualitative than quantitative. On all small β−sheet proteins tested so far it works reasonably well, however a way of assessing the predictive power statistically is still being developed. It should be emphasised that no method can be proved by a single example such as is given here [41].

3: Discussion

The results presented show that residue-residue propensities are strongly affected by the local conformation of the interacting chains and that for β−sheet interactions useful pseudo-potentials can be generated.

Surprisingly these pseudo-potentials show an ability to predict the order of strands in the sequence (fig 4c). Since β−sheets are essentially symmetrical with respect to chain connectivity this difference is difficult to explain and perhaps implies the existence of some previously unknown signal in protein sequences.

The set of potentials described here only considers β−strand interactions. Further potentials for helix-helix and helix-strand interactions can be imagined and are being developed although, without the strong hydrogen-bond distance constraint of sheets, it is anticipated that the predictive power of the potential will be much weaker. These potentials are being applied to fold recognition by threading (data not shown) and it is hoped that they will prove more sensitive than those used in existing methods. There are also a large number of improvements to the existing potential in progress, most notably the inclusion of i–(i+1,i+2) information, which should improve the selectivity between parallel and antiparallel, and attempts to treat ends of strands and edges of strands separately.

As has been shown, the application of these pseudo-potentials is not limited to testing existing protein models or in fold recognition. The existing potential can been applied to a set of homologous sequences to predict a β-strand contact map and gives encouraging results. Predicting helices and sheets rather than helices and strands may be a better approach to protein structure prediction.

Acknowledgements

Thanks are due to John Moult and Chris Sander for useful discussions, to Chris Sander and colleagues for use of the DSSP program and access to the HSSP database and to all those who have deposited protein structure co-ordinates with PDB. T.J.P.H is grateful to MRC and ZENECA for financial support.

References

1. F. C. Bernstein, et al., *J Mol Biol* **112**, 535-42 (1977).
2. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, J. Weng, in *Crystallographic Databases - Information Content, Software Systems, Scientific Applications* F. H. Allen, G. Bergerhoff, R. Sievers, Eds. (Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987) pp. 107-132.
3. C. Chothia, *Nature* **357**, 543-4 (1992).
4. M. J. Sutcliffe, F. R. Hayes, T. L. Blundell, *Protein Eng* **1**, 385-92 (1987).
5. M. J. Sutcliffe, I. Haneef, D. Carney, T. L. Blundell, *Protein Eng* **1**, 377-84 (1987).
6. K. M. Flaherty, D. B. McKay, W. Kabsch, K. C. Holmes, *Proc. Natl Acad. Sci. USA* **88**, 5041-5045 (1991).
7. L. H. Pearl, W. R. Taylor, *Nature* **329**, 351-4 (1987).
8. F. Rippmann, W. R. Taylor, J. B. Rothbard, N. M. Green, *EMBO J* **10**, 1053-9 (1991).
9. J. Overington, M. S. Johnson, A. Sali, T. L. Blundell, *Proc R Soc Lond Biol* **241**, 132-45 (1990).
10. J. U. Bowie, R. Luthy, D. Eisenberg, *Science* **253**, 164-70 (1991).
11. D. T. Jones, W. R. Taylor, J. M. Thornton, *Nature* **358**, 86-9 (1992).
12. C. Sander, Eds., *Protein Design Exercises 1986, EMBL BIOcomputing Technical Doc. 1.* , vol. 1 (BIOcomputing, EMBL, Heidelberg, FRG., 1987).
13. T. J. Hubbard, T. L. Blundell, in *Computer simulations of Biomolecular systems: Theoretical and experimental applications*. W. F. van Gunsteren, P. K. Weiner, Eds. (ESCOM, Leiden, Holland., 1989) pp. 168-82.
14. C. Sander, et al., *Proteins* **12**, 105-10 (1992).
15. J. Novotny, A. A. Rashin, R. E. Bruccoleri, *Proteins* **4**, 19-30 (1988).
16. G. Baumann, C. Frommel, C. Sander, *Protein Eng* **2**, 329-34 (1989).
17. L. Chiche, L. M. Gregoret, F. E. Cohen, P. A. Kollman, *Proc. Natl. Acad. Sci. USA* **87**, 3240-3 (1990).
18. D. Eisenberg, A. D. McLachlan, *Nature* **319**, 199-203 (1986).
19. L. M. Gregoret, F. E. Cohen, *J Mol Biol* **211**, 959-74 (1990).
20. R. Luethy, J. U. Bowie, D. Eisenberg, *Nature* **356**, 83-5 (1992).
21. S. Tanaka, H. A. Scheraga, *Macromolecules* **9**, 945-50 (1976).
22. G. M. Crippen, V. N. Viswanadhan, *Int J Pept Protein Res* **25**, 487-509 (1985).
23. C. Wilson, S. Doniach, *Proteins* **6**, 193-209 (1989).
24. M. J. Sippl, *J Mol Biol* **213**, 859-83 (1990).
25. S. Lifson, C. Sander, in *Molecular Mechanisms of Biological Recognition* M. Balaban, Eds. (Elsevier, Amsterdam, 1979) pp. 145-156.
26. S. Lifson, C. Sander, *J Mol Biol* **139**, 627-39 (1980).
27. P. Bork, et al., *Nature* **358**, 287 (1992).
28. P. Bork, et al., *Protein Sci.* **1**, 1677-1690 (1992).
29. M. J. Zvelebil, G. J. Barton, W. R. Taylor, M. J. Sternberg, *J Mol Biol* **195**, 957-61 (1987).
30. S. A. Benner, D. Gerloff, *Adv. Enz. Regul.* **31**, 121-181 (1990).
31. B. Rost, R. Schneider, C. Sander, *TIBS* **18**, 120-123 (1993).
32. C. Sander, R. Schneider, *Proteins* **9**, 56-68 (1991).
33. W. Kabsch, C. Sander, *Biopolymers* **22**, 2577-637 (1983).
34. A. L. Morris, M. W. MacArthur, J. M. Thornton, *Proteins* **12**, 345-364 (1992).
35. A. Bairoch, B. Boeckmann, *Nuc. Acids Res.* **19**, 2247-2250 (1991).
36. A. Musacchio, M. Noble, R. Pauptit, R. Wierenga, M. Saraste, *Nature* **359**, 851-5 (1992).
37. S. A. Benner, M. A. Cohen, D. Gerloff, *J. Mol. Biol.* **229**, 295-305 (1993).
38. B. Rost, C. Sander, *Nature* **360**, (1992).
39. S. A. Benner, M. A. Cohen, D. Gerloff, *Nature* **359**, 781 (1992).
40. G. J. Barton, R. B. Russel, *Nature* **361**, 505-506 (1993).
41. B. Robson, J. Garnier, *Nature* **361**, 506 (1993).