Recognition of remote protein homologies using three-dimensional information to generate a position specific scoring matrix in the program 3D-PSSM

Lawrence A Kelley, Robert M MacCallum & Michael J E Sternberg (*)

Biomolecular Modelling Laboratory
Imperial Cancer Research Fund
44 Lincoln's Inn Fields
London WC2A 3PX, England

(*) Corresponding author
e-mail m.sternberg@icrf.icnet.uk
FAX +44-171-269-3534
Tel +44-171-269-3565

Abstract

A method (3D-PSSM) to recognise remote protein sequence homologues is described. The method uses homologous proteins of similar three-dimensional structure in the SCOP database (Murzin, A. G. et al., 1995, J. Mol. Biol. 247, 536-540) to obtain a structural equivalence of residues. These equivalences are used to extend multiply-aligned sequences obtained by standard sequence searches (i.e. 1D-profiles). The resultant 3D profile is converted into a position specific scoring matrix (a 3D-PSSM). The approach is benchmarked on recognising remote homologues in the SCOP database and comparing the hit and error rates. 3D-PSSMs are compared with 1D-PSSMs and with two widely-used sensitive search approaches - PSI-BLAST (Altschul, S. F., et al 1997, Nucleic Acids Res. 25, 3389-3402) and global dynamic programming using the BLOSUM62 matrix (Henikoff, S. & Henikoff, J. G.,1992, Proc. Natl. Acad. Sci. USA 89, 10915-10919). In a cross-validated benchmark, 3D-PSSMs and 1D-PSSMs achieved similar results and both have lower error rates compared to the other two methods when recognising remote homologues. The combination of 1D- and 3D-PSSMs provide improved performance over either individual method and thus can identify remote homologies that would not be detected by PSI-BLAST. It is envisaged that 3D-PSSM can complement current homology searches in a two-stage approach in which 3D-PSSMs will follow an initial search using PSI-BLAST or dynamic programming.

RECOMB '99 Lyon France

Abbreviations used

SCOP - Structural classification of proteins
PSSM- Position specific scoring matrix
3D- three-dimensional
PSI-BLAST - Position-specific iterated BLAST
BLAST - Basic local alignment tool

1 - Introduction

With more than 275,000 protein sequences in the databases and with many more sequences being determined in genome projects, computational methods are urgently required suggest structures and functions for the gene products (e.g. Durbin et al., 1998). The standard strategy to examine a gene protein sequence is to establish if it is homologous to a protein with an assigned function and/or structure. Homologous proteins, with their related sequences, will have evolved from a common ancestor and may well have a similar function although divergence of activity remains a possibility. Additionally, homologous proteins nearly always have similar three-dimensional structures.

Homologies can be recognised by pairwise searches that start with the single sequence of the unknown and scanning it against each database entry using programs such as BLAST or FASTA or dynamic programming (for review see Durbin et al., 1998). Recently, marked improvements in detecting remoter homologies (i.e. with less similar sequences) have resulted using PSI-BLAST (Altschul et al., 1997). In this iterative search method, the unknown sequence identifies homologues that are then aligned to generate a weighted profile formalised as a PSSM (position specific score matrix) (Henikoff and Henikoff, 1994). This profile is then used to recognise remoter homologues in the database and the procedure iterated until no further significant hits are found. An alternate sensitive approach is to develop a hidden Markov model from a series of aligned sequences to recognise remote homologues (Durbin et al., 1998; Eddy, 1996; Krogh et al., 1994). In trials based on recognising remote homologues identified on the basis of similar 3D structure, PSI-BLAST and hidden Markov models gave comparable results which were superior to single sequence search methods (Park et al., 1998).

With the increase in the number of experimentally determined protein structures, there are now many families of protein folds with homologous proteins whose relationship could not have been detected by the above, purely sequence based methods. We report the use of known three-dimensional structures to extend the sequence coverage in a PSSM by generating a 3D-PSSM. The library of known structures is from the classification of proteins into homologous superfamilies in the SCOP (Structural Classification of Proteins) database that is based on visual examination of structure/function relations of proteins with similar folds (Murzin et al., 1995).

In outline the 3D-PSSM method (see Figure 1) starts with the sequence of a representative parent protein in the library (the master A) and uses PSI-BLAST (Altschul et al., 1997) to generate a sequence based alignment of homologues (a 1D-profile). Then for each master 1D-profile a search is made for other homologous proteins structures (say B and C) not included in the 1D-profile that can be reliably equivalenced structurally. The residue-based structural equivalence is then used to align the 1D-profile of B and the 1D-profile of C to the master profile of A and thereby generate the 3D-profile for A. From the 3D-profile a 3D-PSSM is generated using a standard approach (Altschul et al., 1997). An unknown probe is then scanned against the 3D-PSSM library.

The power of 3D-PSSM was evaluated in leave-one-out study that evaluated the coverage of detection of true homologues in the SCOP library versus the error rate for probes when only remote homologues were available to be identified. These results were compared to those from 1D-PSSMs, from PSI-BLAST

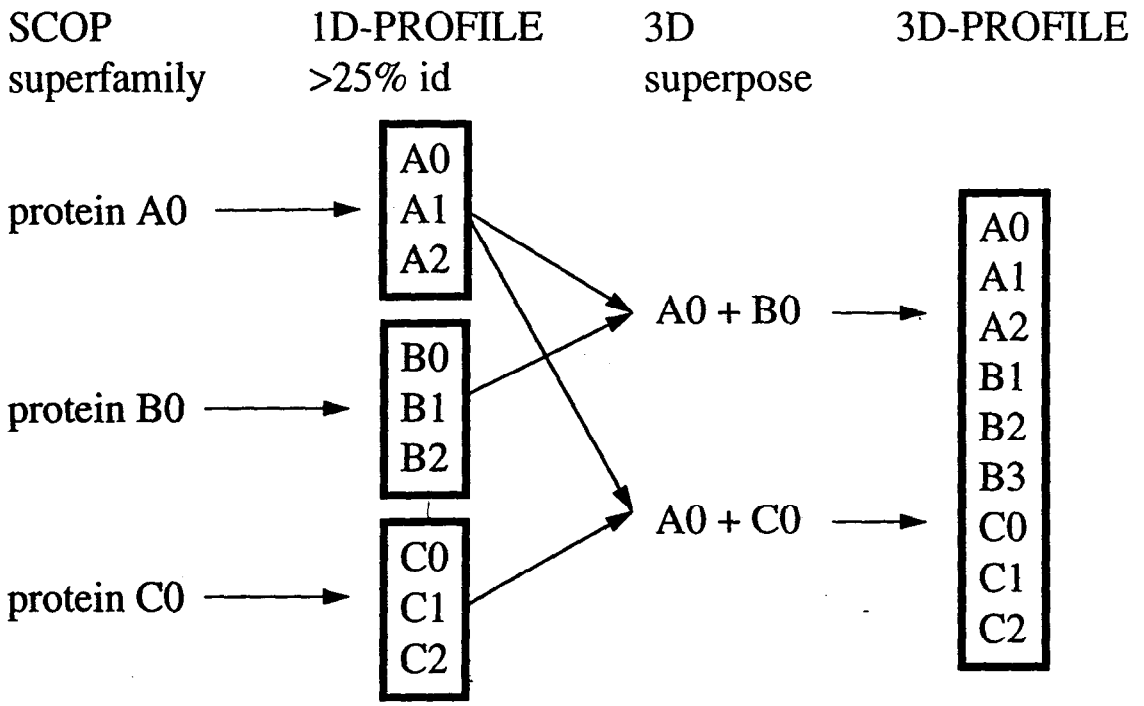

Figure 1 - Outline of the method to generate a 3D-PSSM

(Altschul et al., 1997) and using a pairwise dynamic programming with the BLOSUM62 matrix (Henikoff and Henikoff, 1992). In addition 1D-and 3D-PSSMs were combined. The results show that for this twilight zone of remote homologues, the combination of 1D- and 3D-PSSMs achieved a higher coverage for a given error rate than the alternative methods. It is envisaged that the combined 1D- and 3D-PSSMs approach would be used to complement other sensitive search algorithms and recognise some relationships that sequence methods alone they may fail to identify.

2 - Method

2.1 - Structural and sequence databases

Structural information was obtained from release 1.37 of the SCOP database (http://scop.mrc-lmb.cam.ac.uk/scop/) (Murzin et al., 1995). In SCOP, protein structural domains are grouped into folds with similar three-dimensional structures and these folds then divided into superfamilies. All proteins in a superfamily are homologous, having diverged from a common ancestor. Protein of the same fold type but from different superfamilies are analogues that have converged to a common structure. The SCOP assignment is based on the expertise of the authors who generally assign remote homology on the basis of conservation of function or of unusual structural features. Out of SCOP domains, we excluded certain problem entries: literature references without coordinates; errors in residue numbering, length < 20 residues, X-ray resolution of >3.5Å or

undefined , C-α trace only, more than 15 Cα-Cα separations of > 4.0 Å and more than five undefined residues leading to 11,373 domains. Then a representative library of 1560 domains was generated so that no pair of protein domains shared > 40% identity. These 1560 domains formed the known template library (SCOP-1560)

The benchmark for the recognition of remote homology was established by at random selecting one protein probe for each superfamily in the template library. A second member from the superfamily was then selected provided it was not from the same family and was not related to the first probe by a PSI-BLAST score (E_T) < 0.1 (see below). There were 656 proteins for which there was at least one other protein in the SCOP-1560 library in the same superfamily (i.e. a homologue with < 40% identity).

A non-redundant protein sequence database (NRPROT) was generated at the Imperial Cancer Research Fund. Sequences are progressively taken from the Protein Data Bank (Abola et al., 1997), TREMBL-NEW, TREMBL, SWISSPROT-NEW, SWISSPROT (Bairoch and Apweiler, 1998) and PIR (Barker et al., 1998) but excluding any sequences 100% identity to any previously included sequence. NRPROT was generated on 12 June 1998 and contained 276,289 entries.

2.2 - Generation of 3D-PSSM

Step 1- Generate a 1D-profile for each of the SCOP-1560 master proteins in the template library

i) Start with the sequence of the domain from the master protein (A_0) of known structure in a superfamily.

ii) Search this master sequence against NRPROT using 20 iterations of PSI-BLAST with a theoretical expectation value (E_T) of a hit < 0.1 and an expectation for including a sequence in the iteration (H) of 0.0001. Collect all sequences with >25% identity to A_0.

iii) Sequences identified as homologous by (ii) were then aligned using CLUSTALW (Thompson et al., 1994) provided that there was an overlap of >75% with the master and there was >25% sequence identity with the master. This ensured that high quality sequence alignments were generated. Using the alignment from PSI-BLAST directly would exclude the parts of sequences that could not be equivalenced to the parent sequence.

iv) Remove all redundant sequences with >99% identity that were in the alignment to obtain a representative set. Let there be nA sequences (denoted A_i, i= 1,nA) aligned to A_0.

v) Repeat for all 1560 master proteins in the SCOP-1560 library.

Step 2 - Generate a 3D-profile

For master protein A_0

i) Perform a three-dimensional structural superposition using the SAP program (Orengo et al., 1992; Taylor and Orengo, 1989) between the master structure A_0 and all other proteins within the same superfamily. Identify those pairs (B_0, C_0...) with a good structural match that can be expected to yield a good equivalence of the residues. The criteria were that (a) the weighted root mean square deviation < 6.0Å and (b) at least 60 residues equivalenced or at least 70% of the sequence of A is equivalenced. An equivalence is defined as a SAP equivalence score > 0. The program SAP was obtained from http://www.nimr.mrc.ac.uk/~mathbio.

ii) Use the residue equivalences from the structural alignment to augment the 1D-profile of A_0 with 1D-profiles from B_0, C_0... . Note that this is at a residue by residue level. This yields a profile with sequences (A_0, A_1, A_2,..A_{nA}, B_0, B_1, B_2,..B_{nB}, C_0, C_1,C_2,..C_{nC}) .
Repeat for each master protein in the SCOP-1560 library

Step 3 - Cross validation and removing easy matches

For each of the 656 probes generate a test library of 3D-PSSMs

i) If probe P_0 is included in the profile, remove sequences P_0,P_1..P_{nP}

ii) Create a list of homologies that can be identified with reliability by PSI-BLAST. The expectation value used is E_T< 0.1 with 20 iterations. Remove from the 3D-profile the 1D-profiles of any master sequence that would be identified. This procedure focuses the evaluation of our approach on targets that would not readily be found by PSI-BLAST.

iii) Remove any probe which did not have at least one homologue in a 3D-profile (as opposed to a 1D-profile) as using this probe would not evaluate the approach.

iv) Steps (i) to (iii) led to 165 probes for which there was an homology to a protein with a 3D-profile and for which this homology could not confidently be identified by PSI-BLAST.

v) From the resultant 3D-profile, generate a 3D-PSSM using the method implemented in PSI-BLAST. Denote the value in the 3D-PSSM for residue type r at position k as 3DPSSM(r,k).

2.3 - Searching the probe against the 3D-PSSM library

For each probe, the SCOP-1560 3D-PSSM library is scanned using the global dynamic programming algorithm which was developed for our fold recognition algorithm FOLDFIT (Russell et al., 1998). The score value for residue type r in the probe being aligned to position k in the 3D-PSSM is the PSSM value 3DPSSM(r,k). An afine gap penalty of 10 to open and 1 per gap extension was used based on preliminary trials. End gaps were also penalised.

For a search, the significance of a match was evaluated by fitting a linear relationship between log(number of hits up to a score) against log(total score). Only the top end of the distribution was used and the possibility that the correct hit forming contributing to the tail of the distribution considered by removed the top scoring hit and all entries belonging to the same superfamily. The probability of obtaining a match with that score by chance was converted to a theoretical error rate per query (E_T).

Evaluation followed the approached used in Park et al., (1998). The results for 165 probes were pooled and ordered in increasing error rate per query (E_T). At a given E_T value the cumulative number of probes correctly assigned to a true homologue divided by the total number of probes gave the coverage. For this E_T value, the observed error rate per query was the cumulative number of incorrect assignments of a probe to a superfamily divided by the number of queries (165). In this evaluation, only the first match to a member of a superfamily was considered in evaluating errors and correct matches. This is necessary as the 3DPSSM contains information from several superfamily members in the same template.

2.4 - Comparison with other approaches

i) PSI-BLAST - The utility of 3D-PSSMs was compared to that of PSI-BLAST to find the remaining homologues (i.e. with E_T>0.1). 20 iterations were run and the hits rank ordered by their expectation value. Evaluation was as for 3D-PSSMs but the list was ranked in order of increasing E_T-value.

ii) BLOSUM62 - The 3D-PSSM uses a global dynamic programming search against the template library and thus must be compared against a similar search strategy. We chose the BLOSUM62 (Henikoff and Henikoff, 1992) matrix with dynamic programming as a comparison as it performed well in our test on homologous fold recognition (Russell et al., 1998). The same search procedures and evaluation methodologies as for 3D-PSSMs were used.

iii) 1D-PSSMs - The 1D-profiles for each master sequence were used to construct PSSMs (called 1D-PSSMs) and these were used as the template library and searched in the same approach as for the 3D-PSSMs.

iv) COMBINED 1D- AND 3D-PSSMs - The results from searches against the 3D- and 1D-PSSMs libraries were combined. Matches were scored by their E_T values separately for the 1D- and 3D-PSSMs and then the results pooled.

3 - Results

Figure 2 plots the fraction of the 165 probes that had a correct homologue identified against the error rate per query for the different methods. Note that all these homologies are in the twilight zone as the easy homologies readily identifiable by PSI-BLAST (E_T<0.1) have been excluded. The most important region of the plot corresponds to low error rates per query (<0.2) as this is suggesting homologues without too many erroneous suggestions. In this region, BLOSUM62 with dynamic programming finds more correct homologues that PSI-BLAST. This probably reflects that to detect some remote homologies there is an advantage in a global search of a domain based library where similarity of length is an aid to correct identification compared to the local search procedure of PSI-BLAST. Thus this result does not contradict the general observation by Park et al., (1998) that multiple sequence methods such as PSI-BLAST are superior to pairwise searching such used in our results with BLOSUM62.

1D-PSSMs and 3D-PSSMs gave similar results, both being superior to the pairwise method with BLOSUM62. The 1D-PSSMs represents a multiple sequence search approach and thus the 3D-profiles are not on average superior. This may result from some of the structural alignments being in error and leading to incorrect sequence alignments. In addition, for some superfamilies inlcuding too many members may dilute the sequence signal. However when the 1D- and 3D-PSSMs results were pooled, the results were superior. This suggests that some of the 3D-PSSM templates do encode additional information beyond that extractable from sequence alone.

4 Discussion and Conclusion

Remote protein homologues of known structure can be recognised with a lower error rate of false positives using a combination of 1D- and 3D-PSSMs compared to PSI-BLAST or pairwise searching with BLOSUM62. A widely used alternative approach to recognise remote protein homologues is to augment sequence searching

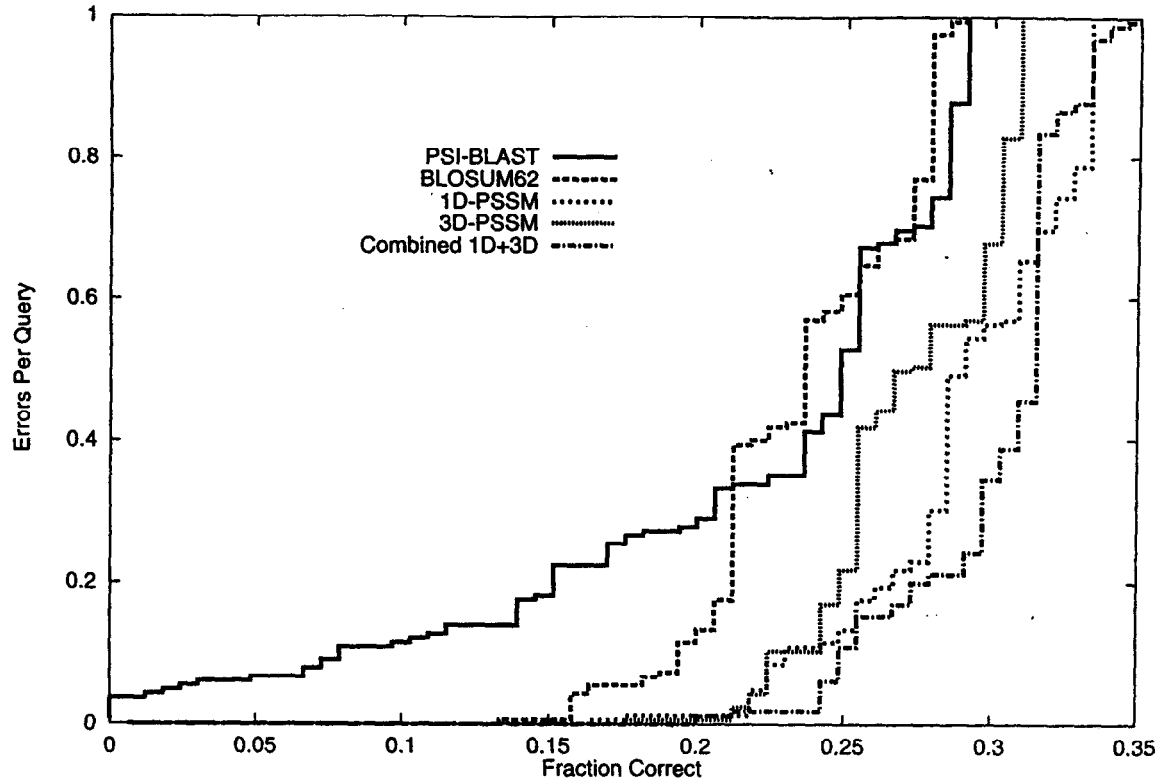

Figure 2 - Graph of errors per query against the fraction of correctly identified homologies for the different methods. Note that the relationships to be identified are those remote homologies that are not readily identifiable by PSI-BLAST.

with matching the predicted secondary structure of the probe against the experimental secondary structure of the templates (Fischer and Eisenberg, 1997; Russell et al., 1998). We are investigating combining secondary structure matching with the 1D- and 3D-PSSMs. However, we envisage that 3D-PSSMs should be used alongside multiple sequenced based methods and other fold recognition algorithms to maximise the number of confidently identified homologues.

The strategy of using 1D-PSSMs and 3D-PSSMs for fold recognition is being compared to the results from other automatic fold recognition servers using the targets from the blind trial of protein structure prediction CASP3 (see http://predictioncenter.llnl.gov/casp3/Casp3.html) . Our plan is to use 3D-PSSMs to suggest structures and functions for the open reading frames in the newly sequenced genomes.

Acknowledgements

L Kelley is supported by GlaxoWellcome. We thank Dr Mansoor Saqi (GlaxoWellcome, Stevenage, UK) for helpful discussion.

References

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). Protein data bank archives of three-dimensional macromolecular structures. Methods in Enzymology **277**, 556-571.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein data base search programs. Nucleic Acids Res. **25**, 3389-3402.

Bairoch, A. & Apweiler, R. (1998). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res. **26**, 38-42.

Barker, W. C., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., Orcutt, B. C., Srinivasarao, G. Y., Yeh, L.-S. L., Ledley, R. S., Mewes, H.-W., Pfeiffer, F. & Tsugita, A. (1998). The PIR-International Protein Sequence Database. Nucleic Acids Research **26**, 27-32.

Durbin, R., Eddy, S., Krogh, A. & Mitchinson, G. (1998). Biological Sequence Analysis. Cambrifge, UK. Cambridge University Press

Eddy, S. R. (1996). Hidden Markov models. Current Opinion in Structural Biology **6**, 361-365.

Fischer, D. & Eisenberg, D. (1997). Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. Proc Nat Acad Sci, USA **94**,

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA **89**, 10915-10919.

Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. J. Mol. Biol. **243**, 574-578.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modelling. J. Mol. Biol. **235**, 1501-1531.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: A structural classification of protein databases for the investigation of sequences and structures. J. Mol. Biol. **247**, 536-540.

Orengo, C. A., Brown, N. P. & Taylor, W. R. (1992). Fast structure alignment for protein databank searching. Proteins: Structure, Function, and Genetics **14**, 139-167.

Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisions using multiple sequences detect three times as many remote homologues as pairwise methods. J. Mol. Biol. **284,** 1201-1210.

Russell, R. B., Saqi, M. A. S., Bates, P. A., Sayle, R. A. & Sternberg, M. J. E. (1998). Recognition of analgous and homologous folds - Assessment of prediction success and associated alignment accuracy using empirical matrices. Prot Eng **11**, 1-9.

Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. J Mol Biol **208**, 1-22.

Thompson, J. D., Higgins, D. G. & Gibson, J. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**, 4673-4680.