J. Mol. Biol. (1997) **273**, 349–354

JMB

Intermediate Sequences Increase the Detection of Homology Between Sequences

Jong Park[1], Sarah A. Teichmann[2], Tim Hubbard[3] and Cyrus Chothia[2]

[1]*Cambridge Centre for Protein Engineering and* [2]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

[3]*Sanger Centre, Wellcome Trust, Genome Campus Hinxton, Cambs, CB10 1SA UK*

Two homologous sequences, which have diverged beyond the point where their homology can be recognised by a simple direct comparison, can be related through a third sequence that is suitably intermediate between the two. High scores, for a sequence match between the first and third sequences and between the second and the third sequences, imply that the first and second sequences are related even though their own match score is low. We have tested the usefulness of this idea using a database that contains the sequences of 971 protein domains whose structures are known and whose residue identities with each other are some 40% or less (PDB40D). On the basis of sequence and structural information, 2143 pairs of these sequences are known to have an evolutionary relationship. FASTA, in an all-against-all comparison of the sequences in the database, detected 320 (15%) of these relationships as well as three false positives (i.e. 1% error rate). Using intermediate sequences found by FASTA matches of PDB40D sequences to those in the large non-redundant OWL database we could detect 550 evolutionary relationships with an error rate of 1%. This means the intermediate sequence procedure increases the ability to recognise the evolutionary relationships amongst the PDB40D sequences by 70%.

Keywords: sequence search; FASTA; OWL database; SCOP

The homology of protein sequences is usually found by matching pairs of sequences with programs such as BLASTP (Altschul *et al.*, 1990) or FASTA (Pearson, 1988). However, the sequences of related proteins can diverge beyond the point where their relationship can be detected by such procedures. In attempts to overcome this limitation, matching methods have been developed that use the features present in multiple aligned sequences of protein families. Examples of such work are sequence templates (Taylor, 1986; Bashford *et al.*, 1987; Tatusov *et al.*, 1994; Yi & Lander, 1994), profiles (Gribskov *et al.*, 1987; Luthy *et al.*, 1994; Thompson *et al.*, 1994) and hidden Markov models (Krogh *et al.*, 1994; Baldi *et al.*, 1994; Eddy, 1995; Eddy *et al.*, 1995). The problem with these procedures is that (i) multiple divergent sequences are required for significant improvements over what is found from single sequence searches, (ii) the accurate alignment of related sequences with low residue identities involves some expertise, and (iii) the scoring schemes for models based on multiple sequence alignments do not at present give thresholds that define high coverage and low error.

Here we describe what we call the 'intermediate sequence procedure'. The intermediate sequence procedure is straight forward, relatively fast and, in a test on related sequences with low residue identities, produces a 70% improvement over what can be found by direct sequence comparisons.

The intermediate sequence procedure

The essential idea is that two homologous sequences, which have diverged beyond the point where their relationship can be recognised by a simple direct comparison, can be related through a third sequence, if its sequence characteristics are intermediate between the two being matched. A high match score between the first and third

Abbreviation used: PDB40D, a database of sequences of known protein structures that have sequence identities of less than 40%.

sequences and between the second and the third sequences, implies that the first and second sequences are related even though their own match score is low. Thus, in suitable cases, it should be possible to demonstrate the relationship between the two distantly related sequences by collecting, from a large sequence databank, homologues which both match with high scores.

The use of more than one sequence to determine homology is not a novel idea. Indeed information on intermediate sequences is available from the Entrez database (Schuler *et al.*, 1996). It lists for each sequence the homologues found by the BLAST program and it is possible to move in the databases from homologues to homologues of homologues. What is novel in the work described here is the demonstration of the very significant extent to which this simple procedure can increase our ability to detect evolutionary relationships.

A database with sequences of low similarity and known evolutionary relationships

The Structural Classification of Proteins (SCOP) database contains a description of the evolutionary and structural relations of those proteins whose atomic structure has been determined (Murzin *et al.*, 1995). The unit of classification in the database is the protein domain. Small proteins, and most of those of medium size, have a single domain and are, therefore, treated as a whole. The domains that form large proteins are treated individually. Domains are clustered together into families if they have close evolutionary relationships. Superfamilies bring together families whose proteins have low sequence identities but whose structural details and, in many cases, functional features suggest that a common evolutionary origin is very probable; for example, the variable and constant domains of immunoglobulins. The fold classification brings together proteins that have the same secondary structures in the same arrangement but which are believed not to have evolutionary relationships (see Murzin *et al.*, 1995).

To test the intermediate sequence procedure we measured the extent to which it could help find the evolutionary relationships described in the SCOP database. As there are few problems in finding relationships between proteins that have more than 40% sequence identity, we use a set of sequences that have pairwise identities of some 40% or less. We call this set PDB40D. There were 971 sequences of proteins or, in the case of large proteins, domains in the version of PDB40D used here. According to the SCOP classification, 2143 different pairs of these sequences have an evolutionary relationship at the family or super-family level.

In previous work (Brenner, 1996; Brenner, Chothia & Hubbard, unpublished results), an all-against-all comparison was made of the PDB40D sequences using the program FASTA version 3.0 with *k*-tuple = 1 algorithm. Analysis of the results of these comparisons show that, of the 323 matches with the lowest E-values, 320 are between proteins known from the SCOP classification to be homologous and the remaining three matches are between proteins that are known not to have an evolutionary relationship. Thus, at an error rate of 1%, FASTA is able to detect 15 % of the PDB40D evolutionary relationships. Calculations of the same kind with the BLASTP algorithm show that it can detect 11% of the relationships at a 1% error level.

A test of the intermediate sequence-single search procedure

Each PDB40D sequence was matched against the 185,000 sequences in the OWL non-redundant protein sequence database version 29.2 (Bleasby *et al.*, 1994) using FASTA version 3.0 with *k*-tuple = 1. The OWL sequences that matched PDB40D sequences with an E-value equal to, or less than, five were collected. At this level the PDB40D sequences matched between one and 4976 OWL sequences; the average number was 74.

The matches were inspected to find those cases where a pair of PDB40D sequences matched the same region of one or more OWL sequence over a length of 30 residues or more. These cases were placed in a list whose order was determined by the larger of the E-values given by matches of the two PDB40D sequences. At the top of the list we found 607 different pairs of PDB40D sequences that match one or more OWL sequences with E-values of 0.15 or less. The pairs of PDB40D sequences were then examined to see whether or not they are homologous. Of the 607 pairs, 302 are known to be related and had significant scores when PDB40D sequences had been matched directly against each other using FASTA. A further 299 pairs are known from the SCOP classification to be homologous but had insignificant scores when PDB40D sequences had been matched directly against each other with FASTA. The remaining six pairs involved proteins that are known not to have an evolutionary relationship. Thus, with an error rate of 1%, the intermediate sequence procedure increases the ability to recognise the evolutionary relationships amongst PDB40D sequences by 88%.

(The threshold used here for the E-value and for the minimum size of the region matched by both query sequences were derived empirically by determining the combination that gives the highest number of true matches with an error rate of 1%. Alternative thresholds near those used here give results only a little less good. For a common match for regions of at least 30 residues we get, for $E \leqslant 0.1$, 584 true matches and four false matches and, for $E \leqslant 0.5$, 636 true matches and 28 false matches. For a common match for regions of at least 50 residues we get, for $E \leqslant 0.1$, 516 true matches and two false matches and, for $E \leqslant 0.5$, 563 true matches and 22 false matches.)

Errors produced by multidomain proteins in the intermediate search procedure

The results described in the last section are impressive. However, the errors found in the calculation are somewhat less than those that would probably occur in normal sequence searches. This is because in the PDB40D database the domains of large proteins are treated as separate entities. This avoids the errors that come from two unrelated multi-domain proteins matching different domains of a multi-domain intermediate sequence. If FASTA limited the matches accurately to the boundaries of the domains this would not be a problem: the requirement that the two query sequences match the same region of the intermediate would eliminate such errors. In practice, however, matches extend beyond domain boundaries because a region where the match is strong can carry the matches in adjacent regions where it is weak. This can produce a significant common match region.

To illustrate this problem consider two query sequences, protein 1, built from domains A and B, and protein 2, built from domains C and D; and the intermediate sequence I built from domains A′ and D′ which are homologues of A and D in proteins 1 and 2, respectively (Figure 1(a)). Correct FASTA alignment of the A and A′ domains in 1 and I is sometimes accompanied by a false weak match between parts of the B and D domains that are adjacent to A and A′ (Figure 1(b)). Similarly, the correct alignment of the D and D′ domains in 2 and I can also be accompanied by a false weak match between regions of the C and A′ domains adjacent to D and D′. These regions of the false matches can be long enough to satisfy reasonable criteria for proteins 1 and 2 matching a common region of the intermediate I (Figure 1(b)).

This problem with multiple domain proteins can be avoided to a large extent by the procedure described in the next section.

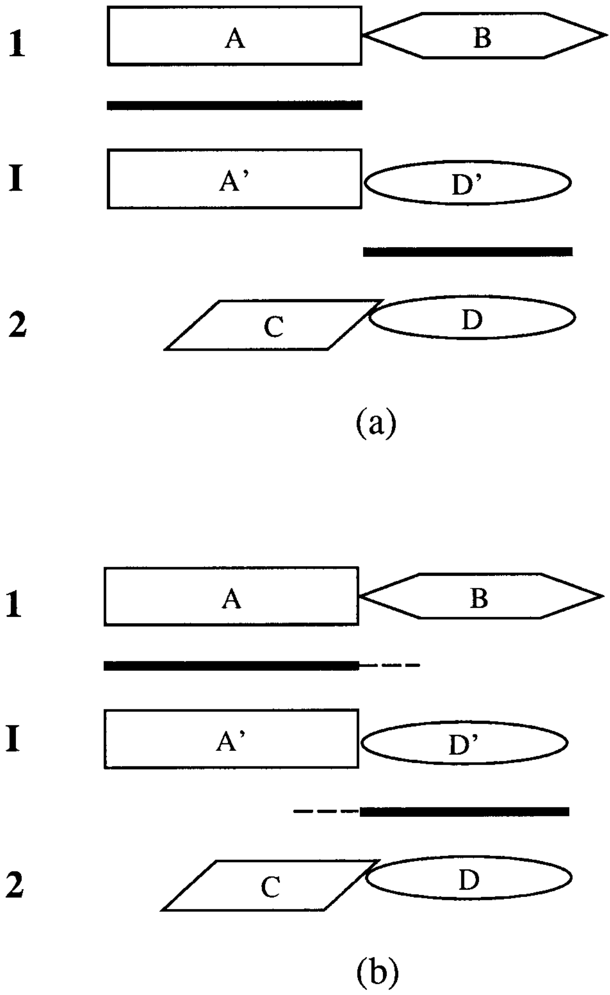

Figure 1. Errors introduced by multiple domain proteins in the intermediate sequence search procedure. Here we show two unrelated query proteins: 1 with domains A and B and 2 with domains C and D, and an intermediate sequence, I, with domains A′ and D′ that are homologous to A and D. The matches of domains A to A′ and D to D′ have low E-values. In (a), the match regions do not extend beyond the boundaries of the homologous domains and the absence of a common match region indicates 1 and 2 are not homologous. In (b) small regions of weak false matches are produced between B and D′ and between C and A′ . The total length of the region of common match can be sufficient to indicate 1 and 2 are homologues. In the intermediate sequence-double search procedure protein 1 is matched against I and the region of D not matched by 1 is removed. This is usually sufficient to prevent I matching 2 in the second match step (see the text).

A test of the intermediate sequence-double search procedure

In this version of the intermediate sequence search the procedure is carried out in three steps: (i) PDB40D sequences are matched against the OWL database and the matched pairs of PDB40D-OWL sequences collected; (ii) the region of the OWL sequence matched by the PDB40D sequence is saved and the rest of the OWL sequence discarded; (iii) the saved fragment of the OWL sequence is then matched against the PDB40D database and significant hits are collected.

To explain how this procedure reduces errors we can use the three sequences described in the last section: 1, 2 and I. Step (i) can be the match of 1 and I. This produces a true match of the A and A′ domains and a false weak match between some parts of the B and D′ domains. Step (ii) removes most of domain D′ sequence from I. Now in step (iii) the I sequence usually does not have enough of the D′ domain to match sequence 2 (Figure 1).

In this procedure, there are two different E-value thresholds: one for the search of PDB40D sequence against the OWL database and the other for the OWL intermediate fragment against the PDB40D database. (The E-values differ because the databases differ in size; see below.) We found from an examination of our results that optimum thresholds for a 1% error are E-values of 0.081 for the first search and 0.0006 for the second search. (As before, E-values close to these give results that are only a little worse.) At these thresholds 550 pairs of PDB40D sequences matched one or more OWL intermediates. Of the 550 pairs, 315 are known to be related and had significant scores when PDB40D sequences had been matched directly against each other using FASTA. A further 230 pairs are known from the SCOP classification to be homologous but had insignificant scores when PDB40D sequences had been matched against each other with FASTA. The remaining five pairs involved proteins that are known not to have an evolutionary relationship. Thus, with an error rate of 1%, the intermediate sequence-double search procedure increases the ability to recognise the evolutionary relationships in the PDB40D sequences by 70%.

(The E-value is dependent on the size of the database (Pearson, 1996) and it is theoretically possible to calculate the ratio of the threshold E-values when databases of different sizes are used.

(a)

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase

```
                          60        70        80        90       100
Amicyanin         MPHNVHFVAGVLGEAALKGPMMKKEQAYSLTFTEAGTYDYHCTPHPFMRGKVVVE
                                             :..:  . ::. ::
Ascorbate Oxidase ILQRGTPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGSLI
                       70        80        90       100       110       120
```

(b)

Alignment by FASTA of the sequences of amicyanin and domain 1 of ascorbate oxidase with the intermediate plastocyanin sequence

```
                                    10        20        30        40
Amicyanin                   DKATIPSESPFAAAEVADGAIVVDIAKMKYETPELHVKVGDTVTW-IN
                                  : .: .:  ::   . .: . . .:   .....:::: : .:
PLASTOCYANIN      SLFAVAAVLCVGSFFLSAAPASAQTVAI-KMGADNGMLAFEPSTIEIQAGDTVQW-VN
                               ..   ::   ....   ::  ::.  :   ::. .:::.:   ..
Ascorbate oxidase    SQIRHYKWEVEYMFWAPNCNENIV---MGI-NGQ--FPGPTIRANAGDSVVVELT


50        60        70        80        90         100           110
REAMPHNVHFVA-------GVLG----EAALKGPMMKKEQAYSLTFT--EAGTYDYHCT--PH--PFMRGKVVVE
 .  ::::  :        :       : . :   ..  ...  ::.  : ::: :.:   ::    : ::.::.
NKLAPHNV-VVE-------G-QP----ELSHKDLAFSPGETFEATFS--EPGTYTYYCE--PHRGAGMVGKIVVQ
:::  ..::...       :  :      : .. :..:::::  .:.  .:::. :. .   .:.::. :...:.
NKLHTEGV-VIHWHGILQRG-TPWADGTASISQCAINPGETFFYNFTVDNPGTFFYHGHLGMQRSAGLYGSLIVD
```

Figure 2. This shows (a) the alignment produced by the FASTA sequence match of amicyanin from *Paracoccus denitrificans* and the first domain of ascorbate oxidase from zucchini (*Cucurbita pepo medullosa*) and (b) the alignment of these two sequences with the intermediate sequence: the plastocyanin precursor from *Synechocystis sp*.

As the number of sequences in the OWL database is 191 times larger than that in PDB40D, the E-value threshold for the second search that should correspond to the 0.081 used in the first is, theoretically, 0.0004. This is close to the optimum we found empirically: 0.0006.)

The match of PDB40D sequences by intermediate sequences

An example of an intermediate sequence matching two PDB40D sequences that cannot be related directly by FASTA is shown in Figure 2. In the Figure, we give (i) the alignment produced by the direct sequence match of two PDB40D sequences that belong to the same super-family: amicyanin from *Paracoccus denitrificans* and the first domain of ascorbate oxidase from zucchini (*Cucurbita pepo medullosa*) and (ii) the alignment of these two sequences to an OWL intermediate: the plastocyanin precursor from *Synechocystis sp*. The E-value for a direct match of the two PDB40D sequences is 53 or 73 depending upon the direction of the match. These scores are insignificant statistically. Amicyanin and the first domain of ascorbate oxidase match plastocyanin with E-values of 3.6×10^{-8} and 7.1×10^{-5}, respectively.

As examples of what happens for individual families, we can compare the results for the globin and fibronectin type III superfamilies. The globin superfamily in PDB40D is represented by 12 globin sequences and two phycocyanin sequences. This means that, counting the match of sequence x to y and y to x as two relationships, there are within the PDB40D sequences 132 globin relationships, two phycocyanin relationships and 48 globin-phycocyanin relationships. The match of PDB40D against itself found 67 of the globin relationships and the two phycocyanin relationships. The intermediate sequence relationship found 118 of the globin relationships and the two phycocyanin relationships. Neither procedure found any match between a globin and a phycocyanin sequence. PDB40D contains seven fibronectin type III sequence with 42 relationships. The match of PDB40D against itself found two relationships; the intermediate sequence procedure found four relationships.

The relationships found by the intermediate sequence procedure can involve very distantly related sequences. If the 230 pairs it detected are matched against themselves, 97 pairs have E-values in the range 0.01 to 0.99; 40 pairs have E-values in the range 1.0 to 4.99 and 93 have E-values between 5.0 and 910. The two sequences whose match has an E-value of 910 are globin sequences from ascaris and midge. They match an insect globin with E-values of 0.06 (for the search of the ascaris globin against OWL) and 10^{-18} (for the search of the insect globin against PDB40D) respectively.

Conclusion

The details of the results described are to some extent dependant on programs and databases used here. Our E-value thresholds, for example, should be seen as good rough guides for other work. Our results do show, however, that the intermediate sequence procedure substantially increases our ability to detect homology amongst proteins with low sequence identities. It will be interesting to see how it compares with methods that use multiple sequences such as profiles and hidden Markov models.

Acknowledgements

We thank Alex Bateman for the discussion of his calculations with hidden Markov models that first suggested to us that the investigation of the intermediate sequence procedure might be worth while and for comments on the paper. We also thank Steven Brenner for his comments on the paper. S.A.T. thanks the Boehringer Ingelneim Fonds for support.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Baldi, P., Chauvin, Y., Hunkapiller, T. & McClure, M. A. (1994). Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci. USA,* **91**, 1059–1063.

Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold: unique features of the globin amino acid sequences. *J. Mol. Biol.* **196**, 199–216.

Bleasby, A. J., Akrigg, D. & Attwood, T. K. (1994). OWL –a non-redundant composite protein sequence database. *Nucl. Acids Res.* **22**, 3574–3577.

Brenner, S. E. (1996), University of Cambridge.

Eddy, S. R. (1995). *ISMB 95, Intelligent Systems in Molecular Biology conference, Multiple Alignment Using Hidden Markov Models.*

Eddy, S. R., Mitchison, G. & Durbin, R. (1995). Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.* **2**, 9–23.

Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA,* **84**, 4355–4358.

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

Luthy, R., Xenarios, I. & Bucher, P. (1994). Improving the sensitivity of the sequence of the sequence profile method. *Protein Sci.* **3**, 139–146.

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

Pearson, W. R. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444–2448.

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol,* **266**, 227–258.

Schuler, G. D., Epstein, J. A., Ohkawa, H. & Kans, J. A. (1996). Entrez-Molecular Biology Database and Retrieval System. *Methods Enzymol,* **266**, 141–162.

Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. USA,* **91**, 12091–12095.

Taylor, W. R. (1986). Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233–258.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput. Appl. Biosci.* **10**, 19–29.

Yi, T. M. & Lander, E. S. (1994). Recognition of related proteins by iterative template refinement (ITR). *Protein Sci.* **3**, 1315–1328.

Edited by J. Thornton

(Received 22 April 1997; received in revised form 24 June 1997; accepted 27 July 1997)