BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 12 2005, pages 2821–2826
doi:10.1093/bioinformatics/bti432

Sequence analysis

Use of multiple profiles corresponding to a sequence alignment enables effective detection of remote homologues

B. Anand†, V.S. Gowri and N. Srinivasan*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

Received on December 26, 2004; revised and accepted on April 4, 2005
Advance Access publication April 7, 2005

ABSTRACT

Motivation: Position specific scoring matrices (PSSMs) corresponding to aligned sequences of homologous proteins are commonly used in homology detection. A PSSM is generated on the basis of one of the homologues as a reference sequence, which is the query in the case of PSI-BLAST searches. The reference sequence is chosen arbitrarily while generating PSSMs for reverse BLAST searches. In this work we demonstrate that the use of multiple PSSMs corresponding to a given alignment and variable reference sequences is more effective than using traditional single PSSMs and hidden Markov models.

Results: Searches for proteins with known 3-D structures have been made against three databases of protein family profiles corresponding to known structures: (1) One PSSM per family; (2) multiple PSSMs corresponding to an alignment and variable reference sequences for every family; and (3) hidden Markov models. A comparison of the performances of these three approaches suggests that the use of multiple PSSMs is most effective.

Contact: ns@mbu.iisc.ernet.in

INTRODUCTION

Searches in the sequence databases using position specific scoring matrices (PSSMs or profiles), corresponding to multiple sequence alignments of protein domain families, are shown to be more effective in detecting distantly related homologues compared to searches involving pairwise sequence alignments (Gribskov *et al*., 1987; Altschul *et al*., 1997; Park *et al*., 1998). The PSI-BLAST procedure (Altschul *et al*., 1997) which iteratively builds PSSMs of a family is commonly used. A reverse procedure with sequence as a query in a search against a database of PSSMs has also been proved to be effective in detecting distantly related protein sequences (Schaffer *et al*., 1999). A PSSM is a weight matrix that, for each position in a group of aligned sequences, assigns a score for each of the 20 amino acid residues. A more sophisticated mathematical model of sequence alignments is the hidden Markov model (HMM) (Krogh *et al*., 1994; Karplus *et al*., 1998; Baldi *et al*., 1994; Eddy, 1998; Lindahl and Elofsson, 2000). The quality of the multiple sequence alignment used and the divergence of sequences are of high importance for the effectiveness of HMMs and PSSMs in detecting distant homologues.

*To whom correspondence should be addressed.

†Present address: Department of Biological Sciences and Bioengineering, Indian Institute of Technology, Kanpur 208 016, India.

A benchmarking of PSI-BLAST suggests that it could successfully identify 40% of the remote homologues among distantly related protein domains of a known three-dimensional (3-D) structure (Muller *et al*., 1999). Another evaluation on the performance of profile-based search procedures such as PSI-BLAST, IMPALA and HMMER suggests that PSI-BLAST and IMPALA yield similar coverages. However, HMMER has a better coverage by a few percentage points (Schaffer *et al*., 1999). A performance comparison of various search procedures suggests that multiple sequence information helps to detect related protein sequences at the family and superfamily level (Lindahl and Elofsson, 2000).

The reference sequence is chosen arbitrarily from the multiple sequence alignment in the case of search using RPS-BLAST or IMPALA (Schaffer *et al*., 1999) on a database of PSSMs of various protein families. While we used the longest of the homologues as the reference sequence in our earlier studies (Gowri *et al*., 2003; Pandit *et al*., 2004), there are advantages and disadvantages in using either the shortest or the longest of the homologues as the reference sequence. In the PSSM matching procedures such as PSI-BLAST, the PSSM is generated, at the end of every iteration, with the query as the reference sequence. However, the query may be considered as an arbitrarily chosen sequence in the family as far as PSSM generation is concerned.

The PSSM of a family encodes two distinct sets of information: (1) extent of occurrence of each of the 20 amino acid types in every position in the multiple sequence alignment; (2) extent of substitution of each one of the residues in the reference sequence by any of the 20 residue types. Hence, the PSSMs generated for a multiple sequence alignment with different homologues as reference sequences will be different. It is our contention that generation of multiple PSSMs for a given alignment using diverse homologues as reference sequences could increase the sensitivity and effectiveness of remote homology detection. Here we propose that RPS-BLAST searches against a database of PSSMs with multiple PSSMs representing every family improve the detection of remotely related protein sequences compared to a search made on a database of PSSMs with every family represented by only one PSSM.

MATERIALS AND METHODS

In the present study we compare the performances of three profile matching approaches.

(i) Match of HMMs of protein families using HMMER2 (Eddy, 1998).

(ii) RPS-BLAST (Schaffer *et al*., 1999) search on a database of PSSMs with every family represented by only one PSSM. Here we have chosen the homologue with the longest length as the reference sequence for that family. This search process is henceforth referred to as 'single PSSM approach' (SPA).

(iii) RPS-BLAST search on a database of PSSMs with more than one PSSM representing every family. The PSSMs of every family have been generated with many homologues as the reference sequence. An identical multiple sequence alignment of a family has been used to generate different PSSMs with different reference sequences. This search process is henceforth referred to as 'multiple PSSMs approach' (MPA).

Database

For the purposes of assessment and comparison of performance of different PSSM matching approaches, we chose to search using protein domains of known 3-D structures as queries on a database of protein domain families of known 3-D structure. If the fold of a hit is the same as that of the query then the hit is considered as correct, else incorrect. The PALI database (Balaji *et al*., 2001) comprises of families of proteins of known 3-D structures primarily derived from the structural classification of proteins (SCOP) database (Murzin *et al*., 1995). Structure-based sequence alignments are provided for all the multi-member families in PALI. An integrated sequence-structure (ISS) database (Gowri *et al*., 2003) has been generated by integrating the structural family from PALI with homologous sequences from either the PFAM database (Sonnhammer *et al*., 1997; Bateman *et al*., 2002) or the non-redundant database (NRDB). The dataset used for the current analysis has been generated from the ISS database by employing the following conditions:

(i) In a family, alignment among the members sharing <60% sequence identity is only considered. This minimizes the bias of the PSSMs generated towards closely related proteins of that family.

(ii) Multi-member families having at least three members are only considered.

The final dataset used contains 286 multi-member families. There are 1325 protein domains of known 3-D structure in these families integrated to about 28 000 sequences of homologues without experimentally derived 3-D structures.

Generation of HMMs

The HMMs are generated for the 286 families using the HMMER2 package (Eddy, 1998). The multiple sequence alignments of the ISS database derived from PALI (version 2.2) have been used for HMM profile generation. This database of HMM profiles will be referred hereafter as HMM_db.

Generation of PSSMs

The PSSMs for the 286 multi-member families are generated from the structure-based sequence alignments of homologues with known 3-D structures integrated with the sequence homologues. Two types of PSSMs are generated.

(1) Single family PSSMs are generated using the protein sequence having the longest length as the reference sequence in the family for a given multiple sequence alignment. Hence, there are 286 single family profiles generated. This database of 286 single family PSSMs will be referred hereafter as SFP_db.

(2) Multiple family PSSMs are generated using every protein sequence in the family with a known 3-D structure as the reference sequence. Hence, the number of PSSMs generated for a family is equal to the number of members of known structure in that family. This database of 1325 multiple family PSSMs will be referred hereafter as MFP_db.

Having identified a reference sequence present in a multiple sequence alignment, the following procedure is used for the generation of a PSSM: The multiple sequence alignment and the reference sequence are given as inputs to PSI-BLAST to 'iterate' against a database of sequences present in the input multiple sequence alignment. Since any 'hit' in such a 'search' corresponds to a sequence already fed as an entry in the input multiple sequence alignment, the multiple sequence alignment that results at the end of the PSI_BLAST run is the same as the input multiple sequence alignment. The −C option in PSI-BLAST is then used to generate the corresponding PSSM output.

Comparative analysis

Searches against the three databases of profiles (HMM_db, SFP_db and MFP_db) were performed using HMMER2 and RPS-BLAST. The searches were made using every sequence of known 3-D structure as the query, selected from the database itself, against the database of profiles. These searches were made using an E-value cut-off of 1 and the assessments were made at various E-value thresholds. In the RPS_BLAST searches, those hits of PSSMs which have their reference sequence to be the same as the query sequence are ignored in further analyses as these are trivial hits. The performance of these three approaches, HMMER (HMM approach), SPA and the MPA are evaluated using the following three parameters.

(i) Specificity = TP/(TP + FP);

(ii) Sensitivity = TP/(TP + FN); and

(iii) Error rate = (FP + FN)/TP

Here TP is the number of true positive profiles, FP is the number of false positive profiles and FN is the number of false negative profiles.

An estimate of difference between two PSSMs corresponding to an alignment

We have estimated the difference between two PSSMs corresponding to two different reference sequences present in the corresponding multiple sequence alignment. The extent of dissimilarity between any two profiles, corresponding to proteins A and B in the multiple sequence alignment, is calculated using the following formula:

$$\Delta D_{\mathrm{A,B}} = \sqrt{(D_{\mathrm{A}}^2 + D_{\mathrm{B}}^2 + D_{\mathrm{A,B}}^2)/3}$$

where,

D_{A} = 100 – mean of the percent sequence identities of A with the other homologous sequences.

D_{B} = 100 – mean of the percent sequence identities of B with the other homologous sequences.

$D_{\mathrm{A,B}}$ = 100 – percentage sequence identity between A and B.

RESULTS AND DISCUSSION

All the sequences with a known structure from the database are queried against the three profile databases HMM_db, SFP_db and MFP_db using HMMER2 (in the case of HMM_db) and RPS-BLAST (for searching against SFP_db and MFP_db). The hits were analyzed at various E-value thresholds ranging from 10^{-5} to 1 with an interval of 10^{-1}. A hit will be considered as a true positive if the query sequence and the protein domains in the hit profile share the same fold. In these searches, if a query sequence identifies even one of the family profiles as related, the protein sequence is associated with the family and the suggested relationship is evaluated for correctness. An analysis of the family and superfamily detection using various approaches is presented here; i.e. it is evaluated if the query identifies its own family profile with another member of the family as the reference sequence. It is also evaluated if the query identifies profiles of other families within the superfamily and the fold.

Specificity

Specificity is a measure of the ability of the profile matching approaches to identify the true hits among all the hits. The plots of specificity (%) versus log(E) for the HMM approach, SPA and MPA are shown (Fig. 1a). The specificity values in all the three approaches perform comparably well at the E-values between 10^{-5} and 10^{-2}. With E-values $>10^{-2}$ the specificities of the HMM approach and SPA drop drastically to ∼70 and ∼55% respectively. However, the percentage specificity in MPA drops smoothly to ∼80%. This suggests that at stringent E-value ranges of 10^{-5}–10^{-2} the three approaches could be used reliably. However, at very relaxed E-value ranges of 10^{-2} –1, MPA would be the approach of choice for effective identification of distantly related sequences.

Sensitivity

Sensitivity is a measure of the ability of the profile matching approaches to identify the true hits among all the correct hits. The plot of sensitivity (%) versus log(E) for all the three approaches are shown (Fig. 1b). The sensitivity of the HMM approach increases gently from ∼78 to ∼83% and for SPA it increases from ∼85 to ∼90%. However, the sensitivity of MPA remains at ∼98% for an E-value between 10^{-5} and 1. This suggests that MPA is powerful in identifying most of the correct family profiles even at stringent E-values.

Error rate

Error rate is a measure of the accuracy of the profile matching approaches to identify the bonafide members over the number of false hits identified and the number of profiles of true homologues missed. The error rate (%) versus log(E) is plotted for all the three approaches (Fig. 1c). The error rate drops drastically in case of SPA from ∼90 to ∼28% as the E-value threshold varies from 1 to 10^{-1}. There is a gradual decrease in the error rate to ∼20% when the E-value goes from 10^{-1} to 10^{-2} and the error rate stays as 20% when the E-value becomes more stringent. In the case of an HMM approach, the error rate decreases drastically from ∼68 to ∼24% as the E-value threshold varies from 1 to 10^{-1}. Further, the error rate gently increases to ∼26% as the E-value becomes more and more stringent. The error rate decreases gradually from ∼20 to ∼3% as the E-value threshold becomes stringent in case of MPA compared to the SPA and HMM approaches. The error rate curves for the three approaches suggest that the searches by the MPA approach are most effective even at relaxed E-values.

Comparison of SPA and MPA for remote homology detection

The ability of the two profile matching approaches (SPA and MPA) to identify the closely related protein sequences (family level) and the distantly related protein sequences showing probable evolutionary origin (superfamily level) was compared.

Superfamily level, inter family connections The 286 protein families in the database belong to 40 multi-member superfamilies. For each superfamily, the sequences belonging to different families within the superfamily are searched against the SFP_db and MFP_db databases. Two protein families can be related by these database searches if any one of the query protein sequences from one family picks up, as a hit, another family profile within the superfamily. Using the information from the SCOP database, on the relationships among the protein sequences across the families as standard, we assess the correctness of identification of superfamilies using sequence–profile matches. We have compared the efficiency of establishing such relationships by the SPA and MPA approaches.

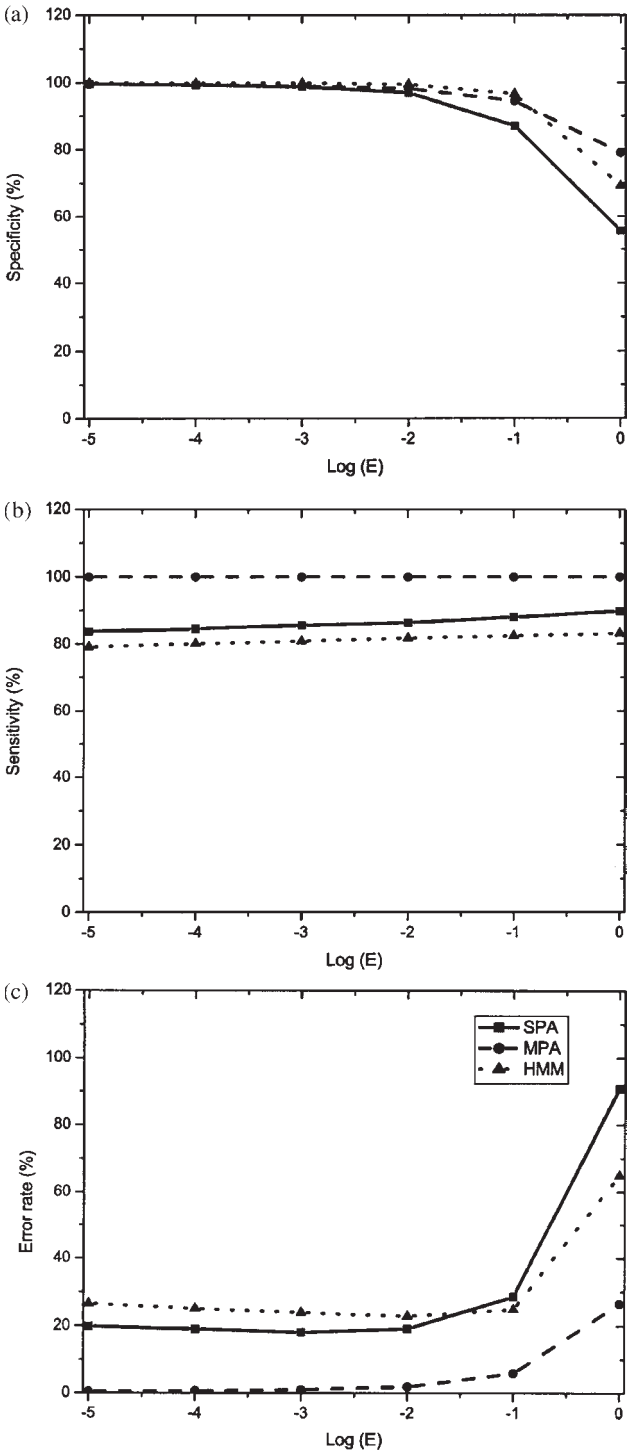

Fig. 1. Percentage of (**a**) specificity, (**b**) sensitivity and (**c**) error rate plotted as functions of the logarithm of E-values for searches involving HMMs (dotted line), SPA (solid line) and MPA (dashed line).

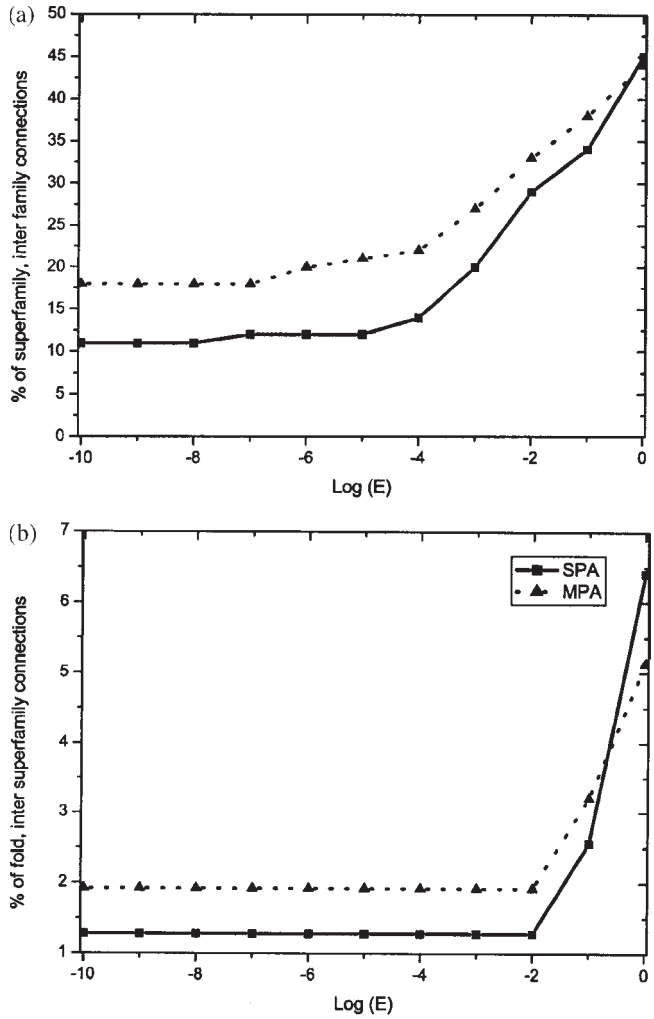

Fig. 2. Percentage of (**a**) inter-family connections within a superfamily and (**b**) inter-superfamily connections within a fold, identified using SPA (solid line) and MPA (dotted line) at various E-value thresholds.

The percentage of the number of across-family connections (within a superfamily) as a function of the logarithm of E-values was plotted (Fig. 2a). The plot clearly shows that MPA is capable of identifying more across-family connections, within a superfamily, compared to SPA even at very stringent E-values. This has been exemplified by the FAD/NAD binding domain superfamily. The protein sequence 1f8ra1 from the family of FAD/NAD-linked reductases is searched against the database of multiple family profiles (MPA_db). This search identified as hits the PSSMs of the query family, succinate dehydrogenase/fumarate reductase flavoprotein family (E-value $= 2 \times 10^{-9}$) with 1qlaa2 as the reference sequence and the C-terminal domain of adrenodoxin reductase-like family (E-value $= 7 \times 10^{-9}$) with 1gtea3 and 1o94a2 as reference sequences. All these families are in the same superfamily as that of the query. However when the same query is searched against the database of single profiles, the profile corresponding to the C-terminal domain of the adrenodoxin reductase-like family with 1cjca1 as the reference sequence was not identified as a hit. The distance between the family profiles of 1gtea3 and 1cjca1 is $\Delta D = 18.5$ and the distance between the PSSMs with reference sequences 1o94a2 and 1cjca1 is $\Delta D = 17.6$.

Fold level, inter superfamily connections The database of the 286 families has 57 folds having at least two superfamilies in each fold. Within a fold, sequences belonging to various superfamilies are searched against the profiles of other superfamilies. In these searches, if a query sequence from a superfamily identifies another sequence profile, belonging to a different superfamily within the same fold, then these two superfamilies are identified as related to each other by the profile matching methods. We present below a comparison between the SPA and MPA approaches in identifying the distantly related protein sequences in different superfamilies within a fold.

The number of across-superfamily connections identified within a fold using SPA and MPA is plotted as a function of the logarithm of E-values (Fig. 2b). At most significant E-values, the number of across-superfamily connections remains constant for both SPA and MPA. However, there is a marginally better performance of MPA compared to SPA. This marginal increase in the identification of distantly related protein sequence families could become important when such searches are made against genome databases. There is a slight increase in the performance of SPA over MPA at $\log(E)$ value close to zero. However, the error rate plot suggests that at $\log(E)$ close to zero, MPA identifies the closely related and distantly related homologues more reliably compared to SPA.

This can be exemplified by taking the example of the triose phosphate isomerase (TIM) fold. When searches were made using a sequence (1gox__) belonging to the FMN-linked oxidoreductase family in MPA, it identified the PSSMs of the family of tryptophan biosynthesis enzymes (E-value $= 2 \times 10^{-34}$) with 1a53__, 1pii__ as reference sequences. These two families belong to different superfamilies in the TIM fold. When the same query was searched in the SPA database, the family of tryptophan biosynthesis enzymes PSSM with a reference sequence (1ttqa_) was not identified as a hit. The distance between the PSSMs with 1a53__ and 1ttqa_as reference sequences is $\Delta D = 19.8$ and that with 1pii__ and 1ttqa_ as reference sequences is $\Delta D = 20.3$.

False positive connections If the searches against the databases of profiles relate the query protein sequence with a profile belonging to different folds, then such connections are labeled as false positives. Of the 1325 searches made, there are 72 such false positive connections identified using these profile-based searches. A plot of the alignment length (%) as a function of the logarithm of the E-value suggests that most of the false positive connections occur at low alignment lengths usually $<$70% of the query length (Fig. 3a). Out of the 72 false positive connections identified within various E-value thresholds, 65 of them do not have $>$70% alignment length. The true positives identified were also examined for query coverage. A plot of $\log(E)$ versus alignment length (%) for the true positive hits (Fig. 3b) suggests that the query coverage for the true positives is almost always $>$70%. Hence, if we include a suitable alignment length condition to filter the hits, we can minimize the number of false positive connections. Further analyses of the region of alignments of false positives suggest that there are structural similarities in many of those short regions which has brought about the across-fold and across-class relationships. This could be explained by an example of the connection between the DNA/RNA binding 3-helical bundle fold and RNA polymerase sigma subunit fold. The alignment

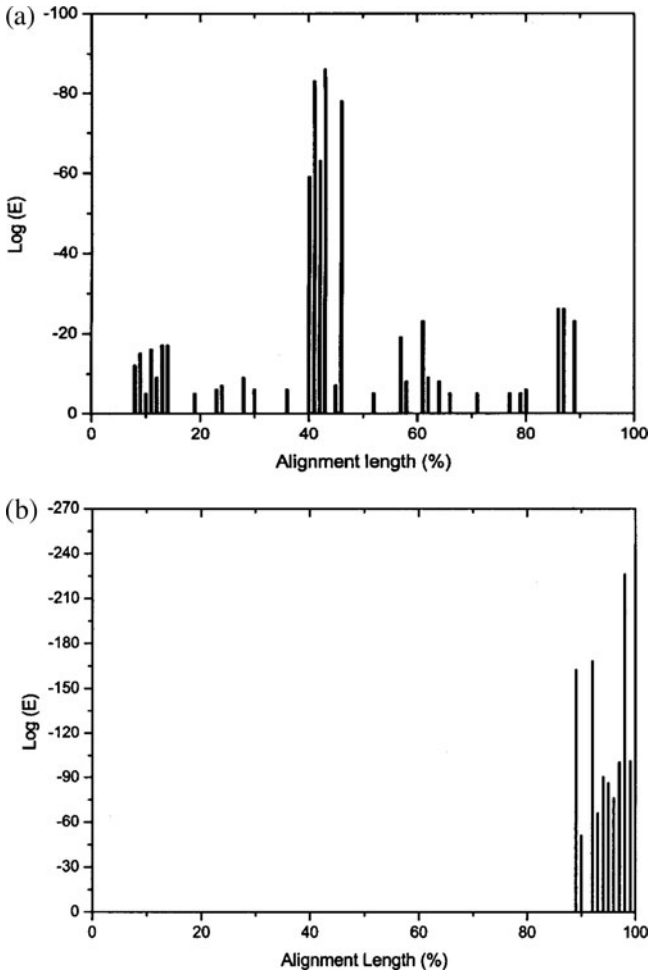

Fig. 3. Plot of the logarithm of E-value as a function of the alignment length (in percentage) for (**a**) false positive and (**b**) true positive connections.

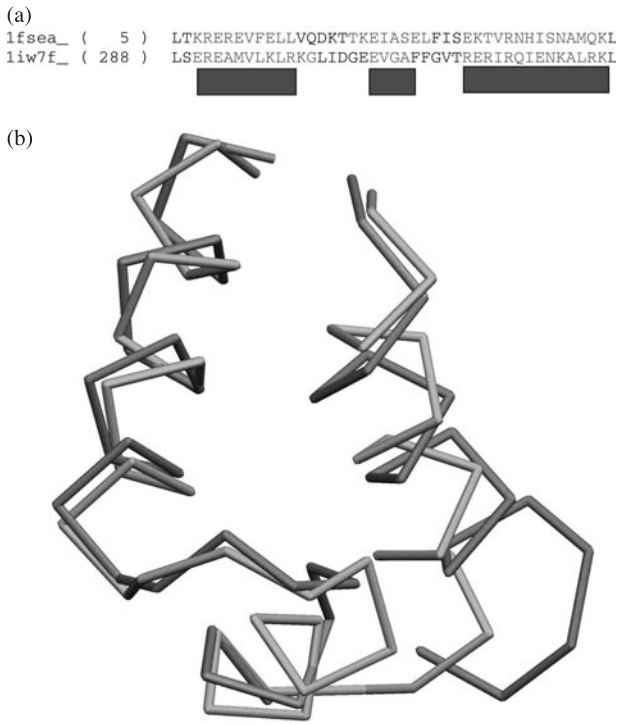

Fig. 4. (**a**) Pairwise alignment of sequences of DNA/RNA binding 3-helical bundle and RNA polymerase sigma subunit which adopt different gross folds. The residues in the alpha helical regions are highlighted in blue color. The blue bars indicate the 'equivalent' helical regions in the two sequences. The colors are visible only in the on-line version of this paper. (**b**) Structural superposition of the Cα trace for the regions of suggested similarity between protein domains from the DNA/RNA binding 3-helical bundle (1fsea_) and the RNA polymerase sigma subunit (1iw7f) folds. In the on-line version of this paper, the segments of 1fsea and 1iw7f are shown in brown and blue, respectively. The RMSD for this superposition is 1.48Å for 36 topologically equivalent $C\alpha$ atoms. This figure has been generated using SETOR (Evans, 1993).

of the protein sequences (Fig. 4a) with the structural feature represented suggests that there are three helices that align very well. This is further supported by the structural superposition of a segment from these two proteins (Fig. 4b).

In general, the extent of occurrence of false positive connections is as low as 3% in the case of MPA as compared to 20% in the case of SPA.

CONCLUSIONS

Hidden Markov models and PSSMs were generated for various protein families. Both single (SPA) and multiple family profiles (MPA) were generated in the case of PSSMs. The multiple PSSMs corresponding to a family arise out of identical sequence alignments, but by using different homologues as reference sequences. The distance between two PSSMs corresponding to a multiple sequence alignment ranges from $\Delta D = 8$ to $\Delta D = 50$. These family profiles were compared for their efficiency in identifying the closely related and remotely related protein sequences. The performance of HMMER was compared with SPA and MPA. In the case of HMM profiles, there is no notion of a reference sequence. The sensitivity, specificity and error rate values for the three approaches suggest that MPA approaches perform better than SPA as well as than HMM approaches. The bias of sequence profiles towards the reference sequence was completely removed in the case of the MPA approach. All the runs were carried out on an Intel P4 1500 MHz processor with 256 MB RAM. The CPU time taken for these 1325 searches using the three approaches suggests that the MPA approach is computationally economical compared to the HMM approach (Table 1).

Table 1. Comparison of CPU time for various approaches

Methods	Computational time[c]
HMM[a]	3.24 hr
Single PSSM[b]	17.34 min
Mulitple PSSM[b]	1.48 hr

[a] Computational time involves building the HMMs, calibration and search.
[b] Computational time includes generation of PSSMs using PSI-BLAST and searching the PSSMs using RPS-BLAST.
[c] All runs were carried out on an Intel P4 1500 MHz processor with 256 MB RAM running on Linux operating system.

This difference in the computational times is even more significant while handling large databases such as genome databases. Hence, the searches using multiple family profiles are economical in terms of computational run time as well as efficiency in the identification of distantly related protein sequences.

ACKNOWLEDGEMENTS

We thank Ms. Natasha Mhatre for her early analysis on the profile matching methods and Dr. Cyrus Chothia for useful comments on the multiple PSSM approach. This research is supported by the award of Senior Fellowship in Biomedical Sciences to N.S. by the Wellcome Trust, UK as well as by the computational genomics project supported by the Department of Biotechnology, Government of India.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Balaji,S. *et al*. (2001) PALI—a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res.*, **29**, 61–65.

Baldi,P. *et al*. (1994) Hidden Markov models of biological primary sequence information. *Proc. Natl Acad. Sci.*, **91**, 1059–1063.

Bateman,A. *et al*. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Evans,S.V. (1993) SETOR: hardware lighted three-dimensional solid model representations of macromolecules. *J. Mol. Graph*, **11**, 134–138.

Gowri,V.S. *et al*. (2003) Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res.*, **31**, 486–488.

Gribskov,M. *et al*. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci.*, **84**, 4355–4358.

Karplus,K. *et al*. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Krogh,A. *et al*. (1994) Hidden Markov models in computational biology. Applications to protein modelling. *J. Mol. Biol.*, **235**, 1501–1531.

Lindahl,E. and Elofsson,A. (2000) Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.*, **295**, 613–625.

Muller,A. *et al*. (1999) Benchmarking PSI-BLAST in genome annotation. *J. Mol. Biol.*, **293**, 1257–1271.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Park,J. *et al*. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pair-wise methods. *J. Mol. Biol.*, **284**, 1201–1210.

Pandit,S.B *et al*. (2004) SUPFAM: a database of sequence superfamilies of protein domains. *BMC Bioinform.*, **5**, 28.

Schaffer,A.A. *et al*. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **12**, 1000–1011.

Sonnhammer,E.L. *et al*. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.