Recent trends in Remote homology detection: an Indian Medley

Venkataraman S. Gowri1 and Sankaran Sandhya2*
1Molecular Biophysics Unit, Indian Institute of Science, Bangalore -560 012;
2National Centre for Biological Sciences, TIFR, GKVK campus, Bellary Road, Bangalore -560 065;
Both authors contributed equally to this review. Sankaran Sandhya* - Email: sandhya@ncbs.res.in;
* Corresponding author received February 05, 2006; accepted February 15, 2006; published online February 21, 2006

Abstract:
The development of remote homology detection methods is a challenging area in Bioinformatics. Sequence analysis-based approaches that address this problem have employed the use of profiles, templates and Hidden Markov Models (HMMs). These methods often face limitations due to poor sequence similarities and non-uniform sequence dispersion in protein sequence space. Search procedures are often asymmetrical due to over or under-representation of some protein families and outliers often remain undetected. Intermediate sequences that share high similarities with more than one protein can help overcome such problems. Methods such as MulPSSM and Cascade PSI-BLAST that employ intermediate sequences achieve better coverage of members in searches. Others employ peptide modules or conserved patterns of motifs or residues and are effective in overcoming dependencies on high sequence similarity to establish homology by using conserved patterns in searches. We review some of these recent methods developed in India in the recent past.

Keywords: Sequence analysis; Remote homology detection; PSI-BLAST; Protein Evolution

Background:
The last few decades have seen spectacular developments in the growth of protein sequence and structure data and in tools to analyse them. Large-scale experimental determination of the biochemical roles of proteins is a phenomenal exercise. Bioinformatics shows exciting promise in the development of methods that allow quick resolution of newly sequenced proteins to their closest experimentally verified relatives. Sequence search procedures such as PSI-BLAST [1], can detect protein relationships effectively when sequence similarities are high. However, when sequence similarities become poor (20-30%), such detection becomes a non-trivial task.

Proteins dissimilar in sequence can adopt similar structure, perform similar functions and also be homologous. This is exemplified in structural databases, which classify proteins with high similarity in sequence, structure and function into families and group families of similar structure and function into superfamilies. For superfamily members, which show poor sequence similarity, it is difficult to determine evolutionary relationship in the absence of structure. Such proteins, related despite low sequence similarity, lie in the ‘twilight zone’ and are termed ‘remote homologues’.

Structure-based approaches can detect such relationships since protein structures are less perturbed by sequence changes. Some approaches consider overall structural similarity in defining relatedness while others such as Bhaduri et al., [2] have shown that conserved spatial interactions in a protein superfamily are excellent constraints in the identification of more members in the superfamily.

Description:
Sequence analysis based approaches for remote homology detection:
Methods like PSI-BLAST build profiles iteratively during searches in sequence databases. [1] The use of ‘intermediate protein sequences’ that share sequence features of more than one protein is quite effective in detecting distant protein similarities. [3] As seen in Figure 1, such sequences populate sequence space and relate proteins, traditionally difficult to relate, due to poor sequence similarities. ‘Intermediate sequences’ that share high similarities with more than one protein, if detected and employed in profile generation can improve effectiveness of sequence analysis-based approaches. Sandhya et al., [4] discuss an application of such sequences in PSI-BLAST searches in fold-specific databases to detect relationships not evident through simple sequence searches.

Several exciting and interesting efforts in remote homology search methods have been made in the Indian sub-continent in the last few years. These methods (Figure 1) show promise and effectively detect such deep relationships in proteins. In this review, we highlight salient features of these approaches.

Cascade PSI-BLAST: Hops through intermediates detect remote homologues:
Sandhya et al., [5] have recently reported the development of a method called Cascade PSI-BLAST
that propagates PSI-BLAST searches in a non-directed manner to detect distant similarities between proteins. In this method, that extensively employs intermediate sequences, a PSI-BLAST search termed “first generation” is initiated for a query in a database. Hits detected are allowed to propagate independent searches in the same database to detect more new hits (“second generation” search etc.). Typically, the authors recommend up to three generations of search for better coverage and detection of relationships. An assessment of the approach on the detection of existing relationships from the PALI database [6] shows that the coverage in detecting relationships in protein families is improved by 15% and in protein superfamilies by ~35% over traditional use of PSI-BLAST. This method is being made available for use in public domain through a web server (http://crick.mbu.iisc.ernet.in/~CASCADE) (manuscript communicated).

**Figure 1:** A superfamily of proteins whose members share poor sequence similarity (<20%). ‘Intermediate sequences’ (in yellow) populate protein space and owing to their high similarities with more than one protein (40-50%) can effectively detect such remote homologues. Methods developed recently in India address the problem of remote homology detection effectively with patterns/intermediate sequences.

**MulPSSM: A database of multiple family profiles corresponding to a constant alignment:**
Searches against sequence databases using PSSMs or profiles are effective in identifying distant relationships compared to searches involving pair wise sequence alignments. [1] The effectiveness of any profile-based search depends on the quality and diversity of sequences in the multiple sequence alignment used in profile generation.

Anand et al., [7] demonstrate that generating multiple family profiles enables the reliable detection of distant relationships in protein family and superfamily. In order to assess the effectiveness of multiple family profiles, they have generated a database of multiple and single-family profiles and profile-HMMs for all structural families from integrated sequence-structure database from PALI. [6] Searches against these three
sequences in profile-based searches or motif-based

India. These methods apply either intermediate
developed remote homology detection methods from
In this review, we have highlighted the recently
Conclusion:

Brahmachari and Dash

PLHOST: Peptide Library based Homology search
tool:

Motif-based approaches use patterns or features
unique to a class of proteins to detect remote
similarities. Gaurav et al., [9] have developed a neural
network-based method called ‘FASSM’ that employs
the positional conservation of sequences and ordering
and positions of family specific inter-motif lengths, in
a scoring scheme to guide family associations. Their
assessments of the approach in detecting over 30
superfamily level relationships indicate an improved
performance over other methods. An exciting
application of this method is its ability to detect
relationships in proteins despite discontinuous
domains and circular permutations.

Acknowledgement:
The authors wish to gratefully acknowledge R.
Sowdhamini and N. Srinivasan for discussions. The
authors also thank R. Sowdhamini for critical reading
of the manuscript. S.S thanks CSIR and NCBS for
financial and infrastructural support.

References:
[1] S. F. Altschul et al., Nucleic Acids Res. 25:3389
(1997) [PMID: 9254694]
[2] A. Bhaduri et al., Proteins. 54:657 (2004)
[PMID: 14997562]
[3] J. Park et al., J. Mol. Biol, 273:349 (1997)
[PMID: 9367767]
[4] S. Sandhya et al., FEBS Lett. 552:225 (2003)
[PMID: 14527691]
[5] S. Sandhya et al., J Biomol Struct Dyn. 23:283
(2005) [PMID: 16218755]
[6] V. S. Gowri et al., Nucleic Acids Res., 31:486
(2003) [PMID: 12520058]
[7] B. Anand et al., Bioinformatics 21:2821 (2005)
[PMID: 15817691 ]
[8] V. S. Gowri et al., Nucleic Acids Res., 34: D243
(2006) [PMID: 15371755]
[9] K. Gaurav et al., In Silico Biol. 5:0040 (2005)
[PMID: 16268788]
[10] S. Chakrabarti et al., Nucleic Acids Res. 33:
W274 (2005) [PMID: 15980468]
[11] S. K. Brahmachari & D. Dash, PCT International
Patent Publication WO 01/74130 A2 (2001)

Edited by N. Srinivasan

Citation: Gowri & Sandhya, Bioinformation 1(3): 94-96 (2006)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any
medium, for non-commercial purposes, provided the original author and source are credited.