Pak. J. Biotech. 1 (1) 67-77 (2004)

THE ROLE OF PATTERN DETABASES IN SEQUENCE ANALYSIS

Engr.Imran Anwar Ujan and *Mrs.Lubna Mustafa

*Department of Statistics and *Institute of Information Technology, University of Sindh, Jamshoro, Pakistan.*

ABSTRACT: In the wake of the numerous now-fruitful genome projects, we are entering an era rich in biological data. The field of bioinformatics is poised to exploit this information in increasingly powerful ways, but the abundance and growing complexity both of the data and of the tools and resources required to analyze them are threatening to overwhelm us. Databases and their search tools are now an essential part of the research environment. However, the rate of sequence generation and the haphazard proliferation of databases have made it difficult to keep pace with developments. In an age of information overload, researchers want rapid, easy to use, *reliable* tools for functional characterization of newly determined sequences. But what are those tools? How do we access them? Which should we use? This review focuses on a particular type of database that is increasingly used in the task of routine sequence analysis the so-called pattern database. The paper aims to provide an overview of the current status of pattern databases in common use, outlining the methods behind them and giving pointers on their diagnostic strengths and weaknesses.

Keywords:*Tools, Biotechnology, Databases, Sequence analysis, Bio-computation*

Biocomputation: The juxtaposition of computation and biology opens up a new world of science and technology. Richard Feynman characterizes the young and fast developing world of computer science as follows: "it is like engineering, it is all about getting something to do something [1]. Viewed from this perspective, the scope of research and development at this intersection is a vast, two-way street, what computer science has to offer to biological science and biotechnology and vice versa. Computational thinking helps characterize, predict, and influence the dynamics of biological processes from molecular to cellular to organ in a way that revolutionizes our understanding of health and drug design. In turn, understanding the architecture and principles of natural biological processes and organization might require new models of computation, which could lead to robustness in the design of large-scale software and hardware systems, a hitherto elusive goal [3,4].

Now days, two key areas drive this convergence between computation and biology. First, the post-genomic challenge is the creation of computational models that characterize a natural living cell's inner workings. We now know how to crack an organism's genetic code through sequencing technology. As we begin the 21st century, the next major challenge is to model the genetic program that is executed through gene–protein interactions in a way that characterizes the spatiotemporal dynamics of cellular events. Such models can assist in predicting and controlling responses to

external agents and in recognizing highly selective targets and drug design. The task of harnessing the open source community's power to develop models for intracellular dynamics, cell to cell signaling and even organism organization is a Herculean one, perhaps on a par with the recently completed national human genome project. A DARPA project called bio-computation (www.darpa. mil/ito/ research) is the latest attempt at this great challenge. A key question to address is the issue of whether the metaphor of circuits and networks is rich enough to deal with the amazing subtleties and complexities of biological systems [2]. Secondly we reach the limits of Moore's law and look beyond silicon for novel substrates to perform computations, information processing and storage, bio-molecular mechanisms present potentially revolutionary alternative. Code system in DNA fragments carry out complex information processing with nucleotide operations such as ligation, restriction, and hybridization in a potentially massively parallel fashion. Since Len Adleman's seminal work in 1994, which showed the potential of DNA computing for complex problems such as the traveling salesman problems, impressive ideas and developments have emerged such as the solution of 6 to 10 variable satisfiability problems and tagged DNA storage with thousands of elements. Although building a DNA Pentium chip is still an elusive goal, an area that holds much promise is the design of computationally driven, precisely engineered nano-structures that exploit DNA self-assembly. The essential idea is to produce arbitrary two and three-dimensional structures from many smaller, information rich structures that can carry the self-assembly code. Such structures can help us to build molecular cages for crystallography, layout molecular electronic devises, and create new materials. As with other new computational substrates such as quantum, spin, or molecular electronics, it is simply too early to tell where this work will lead us. However, there are enough signposts pointing to revolutionary capabilities to give us optimism about the serendipity in discovering new technologies [5,6].

INTRODUCTION: There are hundreds of databanks around the world housing information those floods from the genome projects. The endeavour to store and analyze these vast quantities of data has required increasing levels of automation. However, automation carries a price. For example, although software robots are essential to the process of functional annotation of newly determined sequences, they pose a threat to information quality because they can introduce and propagate misannotations. Although the curators strive to improve the quality of their resources, databases nevertheless carry the indelible scars of time and are far from perfect. To get the most current biological databases it is thus important to have an understanding both of their powers and of their pitfalls.

To characterize a new sequence, the first step usually involves traveling a sequence database with tools such as BLAST2 or FASTA.3. Such searches quickly reveal similarities between the query and a range of database sequences. The trick then lies in the reliable inference of homology (the verification of a divergent evolutionary relationship) and from this, the inference of function. Ideally, a search output will show unequivocal similarity to a well-characterized protein over the full length of the query, providing sufficient

information to make a sensible diagnosis. Sometimes, however, an output will reveal no significant hits or, more commonly, will furnish a list of partial matches to diverse proteins, many of which are un-characterized, or possess dubious or contradictory annotations [4].

There are several reasons why such searches might not give direct answers. For example, the growth of sequence databases and their population by greater numbers of poorer-quality partial sequences makes it increasingly likely that high-scoring matches will be made to a query simply by chance. Low-complexity matches, in particular, may swamp search outputs – these are parts of a sequence that have high densities of particular residues (eg poly-GxP, such as occurs in sequences like collagen, or poly-glutamine tracts that occur in Huntingdon's disease etc). Although it is possible to mask such sequences, this can also create complications. The modular and domain nature of many proteins also causes problems on different levels. When matching multi-domain proteins, it may not be clear which domain or domains correctly correspond to the query. Even if the right domain has been identified, it may not be appropriate to transfer the functional annotation to the query because the function of the matched domain may be different, depending on its biological context. Similar issues arise with the existence of multi-gene families, because database search techniques cannot differentiate between orthologues (usually the functional counterparts of a sequence in another species) and paralogues (homologues that perform different but related functions within the same organism). Given these complexities, correct functional assignment from searches of sequence databases alone can be difficult or impossible to achieve. As a result, it is now customary also to search a range of 'pattern' databases, so-called because they distil patterns of residue conservation within groups of related sequences into discriminators that aid family diagnosis. Searching pattern databases is thus more selective than sequence database searching because discriminators are designed to detect particular families. Different analytical approaches have been used to create a bewildering array of discriminators, which are variously termed regular expressions, profiles, fingerprints, blocks, etc.5 – these terms are summarized in Figure 1. The different descriptors have different diagnostic strengths and weak-nesses and different areas of optimum application, and have been used to generate different pattern databases, which also differ in content! The aim of this paper is to provide an overview of pattern databases in common use and to offer pointers on how best to use them. As this is a fast moving area, a list of web addresses is given in Table 1 to allow readers to obtain current information on the resources discussed [9-11].

Web addresses of pattern and alignment databases in common use given below. For a more exhaustive list, refer to the annual database issue of Nucleic Acids Research (http://www3.oup.co.uk/nar/) PROSITE http://www.expasy.ch/prosite/ Blocks http://www.blocks.fhcrc.org/ PRINTS http:// www.bioinf.man.ac.uk/ dbbrowser/PRINTS/IDENTIFY http:// dna.Stanford.EDU/identify/Profiles http:// www.isrec.isb-sib.ch/software /PFSCAN-form.html Pfam http://www.sanger.ac.uk/ Software/ Pfam/ ProDom http://www. toulouse.inra.fr/prodom.html SBASE http://www.icgeb.trieste.it/sbase/PIR-ALN

http://www-nbrf.georgetown.edu/pirwww/search/textpiraln.htmlPROT-FAM http://vms.mips.biochem.mpg.de/mips/programs/classification.htmlDOMOhttp://www.infobiogen.fr/~gracy/domo/ProClass http://pir.george-town.edu/gfserver/proclass. Html Proto Map http://www.protomap.cs.huji.ac.il/PIMA http://dot.imgen.bcm.tmc.edu: 9331/seq-search/protein-search.html InterPro http://www.ebi.ac.uk/interpro/

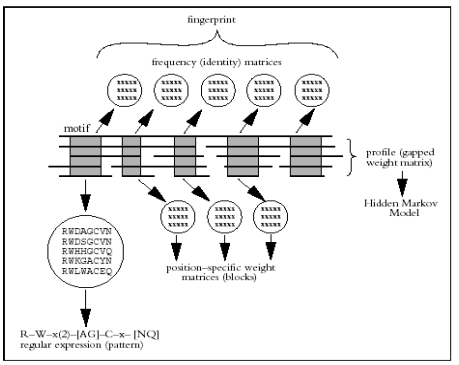

Figure 1: At the heart of sequence analysis methods is the multiple sequence alignment. Application of these methods involves the derivation of some kind of representation of conserved features of the alignment, which may be diagnostic of structure or function. Various terms are used to describe the different types of data representation, as shown. Within a single conserved region (motif), the sequence information may be reduced to consensus expression (regular expression), often simply referred to as a pattern. In this example, square brackets indicate residues that are allowed at this position of the motif and x denotes any residue, the (2) indicating that any residue can occupy consecutive positions in the motif. The term used to describe groups of motifs in which all the residue information is retained within a set of frequency (identity) matrices is a fingerprint, or signature. Adding a scoring scheme to such sets of frequency matrices results in position-specific weight matrices, or blocks. Using information from extended conserved regions that include gaps (usually referred to as domains) gives rise to profiles; and probabilistic models derived from alignment profiles are termed hidden Markov models [14].

THE METHODS BEHIND THE DATABASES

At the heart of the analysis methods that underpin pattern databases is the multiple sequence alignment. When building an alignment, as more distantly related sequences are included, insertions are often required to bring equivalent parts of adjacent sequences into the correct register, as illustrated schematically in Figure 2 As a result of this gap insertion process, islands of conservation emerge from a backdrop of mutational in change. These regions, usually termed motifs or blocks, are typically around 10– 20 residues in length and tend to correspond to the core structural or functional elements of the protein. The conserved nature of motifs effectively provides us with a set of familial blueprints, and different techniques have evolved to exploit this fact. As shown in Figure 2, the methods fall broadly into three categories, depending on whether they use single motifs, multiple motifs or full domain alignments. All of these methods involve the derivation of some kind of discriminatory representation of aspects of the alignment, providing characteristic signature for the family that can be used to diagnose future query sequences [7,8]. The diagnostic success of the different methods depends on how reliably true family members (true-positives) can be distinguished from

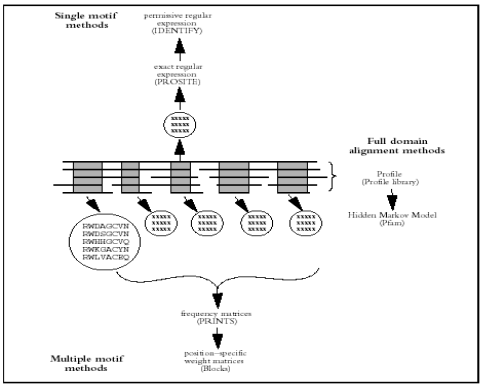

Figure 2: Illustration of the three principal methods for building pattern databases: ie using single motifs, multiple motifs and full domain alignments. Single-motif (regular expression pattern) approaches have given rise to the PROSITE and IDENTIFY databases; multiple-motif methods have spawned the Blocks and PRINTS databases; and domain alignment methods have resulted in the Profiles and Pfam resources [15] non-family members (true-negatives). In practice, there is a crucial balance between the number of incorrect matches that are made (falsepositives) and the number of correct matches that are missed (false-negatives) at a given scoring threshold. As shown in Figure 3, for a given search, this requires the distribution of true-positive matches to be resolved from that of the true negatives, such that the overlap between them is minimized or eliminated. This is important because, for matches in the overlapping area, it can be difficult or impossible to determine which are correct (statistical approaches are used to assign confidence levels to matches in this area, but mathematical significance does not give biological proof). The different analytical methods that have been designed to tackle these issues are outlined below.

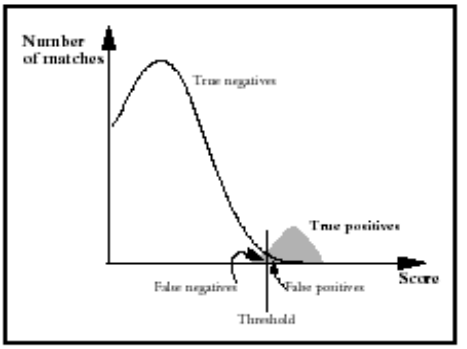

Figure 3: Resolving true and false matches. In a database search, the desire is to establish which sequences are related (true-positive) and which are unrelated (true-negative). At a given scoring threshold, it is likely that several unrelated sequences will match a search pattern erroneously (so-called false positives) and several correct matches will fail to be diagnosed (false-negatives). In sequence analysis, the challenge is to improve diagnostic performance by capturing all (or the majority) of true positive family members, including no (or few) false-positives, and minimizing or precluding false-negatives.

Single motif methods: Of the various approaches, single-motif (regular expression pattern) methods are easier to understand. The idea is that the single most conserved, often functionally important can characterize particular protein family, region (eg an enzyme active site) observed in a sequence alignment. The motif is reduced to a consensus expression in which all but the most significant residue information is discarded. For example, the expression D-x-{KR}-[NQ] means that a conserved aspartic acid (D) residue is followed by an arbitrary residue (x) and any residue except lysine (K) or arginine (R), and finally a polar residue, which may be asparagines (N) or glutamine (Q). No

other residues or residue combinations are tolerated by the expression; matches to it must therefore be exact, or will be disregarded. So rigid is this syntax that regular expression patterns do not perform well when used to represent highly divergent protein families [16,22,23]. For example, such patterns will fail to match significant sequences if they contain a single amino acid difference. The sequence DARN is thus a mis-match, in spite of matching the above expression in all but one position (it has a forbidden arginine as its third residue). Conversely, a pattern will match anything that corresponds to it exactly, regardless of whether it is a true family member. The problem is that match to single motifs lack biological context, match to a pattern is just a match to a pattern and may well only be fortuitous. To assess the likelihood of a match being 'real', it must be verified with corroborating evidence, whether via other database searches the literature or experiment [24].

An approach that addresses the strict nature of exact regular expression matching is to assign amino acid residues to distinct, but over-lapping, substitution groups corresponding to various biochemical properties (eg charge and size), as shown in Table 2. This is biologically sensible because each amino acid has several properties and can serve different functions, depending on its biochemical context. However, although the technique is more flexible, its inherent permissiveness has an inevitable signal-to-noise trade-off, ie resulting patterns not only have the potential to make more true-positive matches, but they will consequently also match more false-positives. For example, the sequence EVEN, which would be excluded by the exact regular expression above, would be matched by the permissive one (because Asp and Glu belong to the same group), even if aspartic acid were biologically mandatory at the first position of the motif [25,26].

Table 2: Overlapping sets of amino acids and their properties. These are used to create the permissive regular expressions used as the basis of the IDENTIFY resource

Residue property	Residue groups
Small	Ala, Gly
Small hydroxyl	Ser, Thr
Basic	Lys, Arg
Aromatic	Phe, Tyr, Trp
Basic	His, Lys, Arg
Small hydrophobic	Val, Leu, Ile
Medium hydrophobic	Val, Leu, Ile, Met
Acidic/amide	Asp, Glu, Asn, Gln
Small/polar	Ala, Gly, Ser, Thr, Pro

Multiple-motif methods in response to these problems, diagnostic techniques evolved to exploit multiple motifs. Within a sequence alignment, it is common to find several motifs that characterize the aligned family. Diagnostically, it makes sense to use many or all such regions to build a family signature. In a database search, there is then a greater chance of identifying a distant relative, whether or not all parts of the signature are matched. For example, a sequence that matches only four of seven motifs may still be diagnosed as a true match if the motifs are matched in the correct order in the sequence and the distances between them are consistent with those expected of true neighbouring motifs. The ability to tolerate mis-matches, both at the level of individual residues within motifs, and at

the level of motifs within the complete signature, makes multiple-motif matching a powerful diagnostic approach.

Different multiple-motif methods have arisen, depending on the technique used to detect the motifs and on the scoring method employed. Probably the simplest to understand is the technique of fingerprinting [8]. Here, groups of conserved motifs are existed from a sequence alignment and used to create a series of frequency (identity) matrices – no mutation or other similarity data are used to vet the results. The scoring scheme is thus based on the calculation of residue frequencies for each position in the motifs, summing the scores of identical residues for each position of a retrieved match. However, the simplicity of this approach is both its strength and its weakness. In other words, because the method exploits observed residue frequencies, the scoring matrices are sparse and thus perform cleanly (with little noise) and with high specificity; at the same time, their absolute scoring potential is limited by the nature of the observed data. For richly populated families, this is not a problem because the resulting matrices will reflect the constituent sequence diversity; but for poorly populated families, the matrices may be too sparse and may not encode sufficient variation to be able to detect distant relatives reliably 19,20,27].

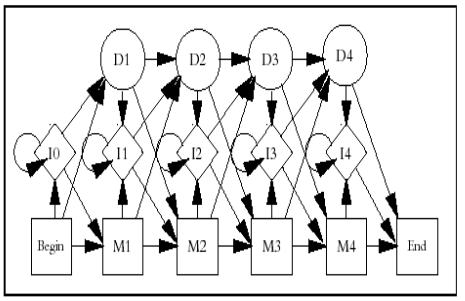

Figure 4: Linear hidden Markov model (HMM). Each position of an alignment is represented as a match (M), an insert (I), or a delete (D) state in the HMM. This allows a query sequence to be aligned by assigning the most probable state transition to each of its residues

Profile methods: An alternative philosophy to motif-based approaches takes into account the variable regions between conserved motifs, which also contain valuable information. Here, the complete conserved portion of the alignment (including gaps) effectively becomes the discriminator. The discriminator, termed a profile, defines, which residues are allowed at given positions, which positions are highly conserved and which degenerate, and which positions can tolerate insertions. The scoring system is intricate and may include evolutionary weights and results from structural studies, as well as data implicit in the alignment. In addition, variable penalties may be specified to weight against insertions and deletions occurring within core secondary structure elements [12,13]. Profiles (sometimes referred to as weight matrices) provide a sensitive means of detecting distant sequence relationships where only very few residues are well conserved.

PATTERN DATABASES: The different methods of analyzing sequences and encoding protein families have given rise to different pattern databases, as shown in Table 3. Despite their differences, pattern databases have arisen from the same principles, i.e. homologous sequences share conserved motifs, presumably

crucial to the structure or function of the protein, which provide a signature of family membership. A new sequence that matches these predefined characteristics may then be assigned to a family. If the structure and function of the family are known, searches of pattern databases thus theoretically offer a fast track to the inference of biological function. Because these resources are derived from multiple sequence information, searches of them are often better able to identify distant relationships than are searches of the sequence databases. However, none of the pattern databases is yet complete. They should therefore be used to augment sequence searches, rather than to replace them.

Table 3: Some of the major pattern databases in common use. In each case, the primary source is noted, together with the type of pattern stored (regular expression, fingerprint, HMM, etc.)

Pattern database	Data source	Stored information
PROSITE	SWISS-PROT	Regular expressions (patterns)
PRINTS	SWISS-PROT/TrEMBL	Raw aligned motifs (fingerprints)
Profiles	SWISS-PROT	Gapped weight matrices (profiles)
Pfam	SWISS-PROT/TrEMBL	Gapped domain alignments (HMMs)
Blocks	PROSITE/PRINTS	Weighted aligned motifs (blocks)
IDENTIFY	Block/PRINTS	Permissive regular expressions (patterns)

WHICH DATABASE IS BEST? The plethora of available databases presents bewildering choices to the sequence analyst. Which is diagnostically most reliable? Which has the most useful annotations? Which is the most comprehensive? Which should I use? It is difficult to assess the quality of particular resources: each has different diagnostic strengths and weaknesses, each offers different family coverage and different levels of annotation each has good points and bad. Nevertheless, some general points bear consideration. Initially, the clustered family resources appeal because they are so comprehensive, yet they suffer certain limitations. Automatic clustering is based on pre-set scores and the resulting clusters need not have precise biological correlations. Furthermore, the search tools tend to involve flavours of BLAST or FASTA, which are good at highlighting generic similarities but cannot pinpoint differences (eg such as between highly similar but functionally disparate receptor subtypes). Perhaps the biggest failing of automatically generated pattern and cluster databases is that they carry no annotations. The advantage of searching them is that they are more comprehensive than their manually derived counterparts. The disadvantage is that there may be no way to ascertain the biological significance of a match, if indeed it has any (that a match has been made does not mean an evolutionary relationship necessarily exists). This is important to understand – automatic methods can only detect similarities, but it is for the user to infer homology from supporting biological evidence [17].

Among pattern databases, single-motif methods that use exact regular expression pattern-matching have known diagnostic limitations. These methods tolerate no similarity, so will fail to diagnose sequences that contain subtle changes not catered for by the pattern. Moreover, single motifs offer no biological context within which to assess the significance of a match – each has therefore to be verified individually. Multiple motif approaches inherently offer improved diagnostic reliability by virtue of the mutual context provided by motif neighbours. Thus, if a query fails to match all the motifs in a signature, the pattern of matches formed by the

remaining motifs still allows the user to make a confident diagnosis. Pattern resources derived from existing databases have the limitation that they offer no further family coverage. Nevertheless, they have the advantage of implementing different analytical methods from their source databases, thus offering different scoring potentials on the same data, and furnishing important oppor-tunities to diagnose relationships missed by the original implementations. Finally, manually annotated databases are set apart from their automatically created counterparts by virtue of (i) attempting to provide validation of results, and (ii) offering detailed information that helps to place conserved sequence information in structural or functional contexts. This is vital for the user, who not only wants to discover whether a sequence has matched a pre-defined motif, but also needs to understand its biological significance.

A COMPOSITE PATTERN DATABASE

Today, comprehensive sequence analysis requires accessing a variety of disparate databases, gathering the range of different outputs and arriving at some sort of consensus view of the results. In the future, however, this process should become more straightforward. The curators of PROSITE, PRINTS, Pfam, Profiles and ProDom are creating a unified database of protein families, termed InterPro. The aim is to provide a single family annotation resource, based on existing documentations from PRINTS and PROSITE, and on the minimal annotations in Pfam, with each InterPro document linking back to the relevant entries in the satellite pattern databases. This will simplify sequence analysis for the user, who will thereby have access to a central resource for protein family diagnosis [15].

CONCLUSION: Creating and searching pattern databases are activities that lie at different ends of a fallible chain of events. We begin with a sequence alignment; we create some kind of scoring function to encode the conservation within the alignment (a scoring matrix, HMM, etc.), we store the discri-minators in a database, and we search them with different algorithms. Problems arise if unrelated sequences have crept into the alignment, which in turn lead to errors in the discriminators, which then give ambiguous or incorrect search results. Alternatively, the discriminators may be sound, but the search algorithms may not be sufficiently sensitive to allow un-equivocal diagnosis, leading the user to false conclusions of family ties. If the user has performed this experiment on a newly determined sequence and submits the results to one of the sequence databases, the annotation error becomes available for mass propagation.

Recently, there has been doom-mongering in the literature about the quality of our databases, some harbingers of misfortune predicting a future error catastrophe. At the same time, claims of success for some approaches to family classification and function prediction have been equally overdone. A more balanced view recognizes that our databases and search routines are not perfect, but with the right approach we can avoid the pitfalls of jumping to over-pessimistic or over-zealous conclusions. Until we have sufficient experimental data available, pattern and sequence databases are probably the best tools we have for accessing the functional and evolutionary clues latent in the sequences flooding from the genome projects. Pattern

databases offer several benefits: (i) by distilling multiple sequence information into family or domain descriptors, trivial errors in the underlying sequences may be diluted; (ii) annotation errors may be quickly spotted if the description of one sequence differs from that of its family; and (iii) they allow specific diagnoses, placing individual sequences in domain or family contexts for a more informed assessment of possible function. By contrast, searches of sequence databases tend to reveal only generic similarities, making precise pinpointing of a particular biological niche more difficult. While there is some overlap between them, the contents of the pattern databases differ. Together they encode about 2,200 families, including globular and membrane proteins, modular polypeptides, and so on. It has been estimated that the total number of families might be in the range 1,000 to 10,000, so there is a long way to go before any of the databases can be considered complete. Thus, in building a search strategy, it is good practice to include all available pattern resources, to ensure that the analysis is as comprehensive as possible and that it takes advantage of a variety of search methods. Where there is consensus, diagnoses can be made with greater confidence.

Unfortunately, creating and annotating family descriptors is time-consuming, so pattern databases have not kept pace with the deluge of sequence data and are conesquently still very small. Nevertheless, as they become more comprehensive, as the volume of sequence data expands and search outputs become more complex, their diagnostic potency ensures that pattern databases will play an increasingly important role as the post-genome quest to assign functional information to raw sequence data gains pace.

REFERENCES

1. Doerks, T., A.Bairoch and P.Bork, Protein annotation: detective work for function prediction, Trends Genetics **14**: 248–250 (1998).
2. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Res., **25**: 3389–3402 (1997).
3. Pearson, W. R., Empirical statistical estimates for sequence similarity searches, J.Mol.Biol.**276**:71–84 (1998)
4. Hofmann, K., Protein classification and functional assignment In: Trends Guide to Bioinformatics', Elsevier Science, New York, pp. 18–21(1998).
5. Attwood,T.K.,Exploring the language of bioinformatics', in 'Oxford Dictionary of Biochemistry and Molecular Biology, Stanburg, H., Ed., Oxford University Press, Oxford, pp. 715–723 (1997).
6. Attwood, T.K. and D.J.Parry-Smith, Introduction to Bioinformatics, Addison Wesley Longman, Harlow (1999).
7. Nevill-Manning, C. G., T.D.Wu and D.L.Brutlag, Highly specific protein sequence motifs for genome analysis, Proc. Natl. Acad. Sci., USA, **95**: 5865–5871(1998).
8. Parry-Smith, D. J. and T.K.Attwood, ADSP– a new package for computational sequence analysis, Comput. Appl. Biosci., **8**: 451–459 (1992).
9. Dayhoff, M. O., Schwartz, R. M. and Orcutt, B.C., A model of evolutionary change in proteins, in Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3, Dayhoff, M.O., Ed. National Biomedical Research Foundation, Washington, DC, pp.345–352 (1978).

10. Henikoff, J.G. and S.Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl Acad. Sci., USA, **89**: 10915–10919 (1992).
11. Doolittle, R.F., Of URFs and ORFs: A Primer On How to Analyse Derived Amino Acid Sequences, University Science Books, Mill Valley, CA (1986).
12. Gribskov, M., A. D. McLachlan and D. Eisenberg Profile analysis: Detection of distantly related proteins, Proc. Natl Acad. Sci., USA, **84**: 4355–4358 (1987).
13. Luthy, R., I. Xenarios, and P. Bucher, Improving the sensitivity of the sequence profile method, Protein Sci., **3**: pp. 139–146 (1994).
14. Hughey, R. and A. Krogh, Hidden Markov models for sequence analysis: extension and analysis of the basic method, Comput. Applic. Biosci. **12**: 95–107 (1996).
15. Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch, The PROSITE database, its status in 1999, Nucleic Acids Res. **27**: 215–219 (1999).
16. Bairoch, A. and R. Apweiler, The SWISS -PROT protein sequence data bank and its supplement TrEMBL Nucleic Acids Res.**27**: 49–54 (1999)
17. Henikoff, J. G., S. Henikoff and S. Pietrokovski, New features of the Blocks Database servers, Nucleic Acids Res. **27**: 226–228 (1999).
18. Attwood,T.K., D.R.Flower,A.P.Lewis, J.E. Mabey, S. R. Morgan, P. Scordis, J. Selley and W. Wright, PRINTS prepares for the new millennium, Nucleic Acids Res.**27**:220–225 (1999).
19. Bateman, A., E.Birney, R. Durbin, S. R. Eddy, R.D.Finn, and E.L.L Sonhammer, Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins, Nucleic Acids Res. 27(1): 260–262 (1999).
20. Gouzy, J., F.Corpet and D.Kahn, Recent improvements of the ProDom database of protein domain families, Nucleic Acids Res.**27**: 263–267 (1999)
21. Murvai, J., Kristian, K.Vlahovicek, E.Barta, C.Szepesvári, C.Acatrinei, and S. Pongor, The SBASE protein domain library, release 6.0: A collection of annotated protein sequence segments, Nucleic Acids Res. **27**: 257–259 (1999).
22. Yona, G., N.Linial, N. Tishby, and M. Linial, A map of the protein space an automatic hierarchical classification of all protein sequences, in Proceedings of 6th International Conference on ISMB, AAAI Press, Menlo Park, CA, pp. 212–221 (1998).
23. Srinivasarao, G.Y, L. S. L. Yeh, C.R. Marzec, B.C.Orcutt, W.C.Barker and F.Pfeiffer, Database of protein sequence alignments: PIR-ALN, Nucleic Acids Res.**27**: 284–285 (1999).
24. Mewes, H. W., K.Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker and D.Frishman, MIPS: a database for genomes and protein sequences, *Nucleic Acids Res.***27**: 44–48 (1999)
25. Wu, C. H., S. Shivakumar, and H. Huang, Pro-class protein family database, *Nucleic Acids Res.* **27**: 272–274 (1999)
26. Gracy, J. and P. Argos, DOMO: a new database of aligned protein domains', *Trends Biochem. Sci.* **23**: 495–497 (1998).
27. Smith, R. F. and T. F. Smith, Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling, *Protein Eng.* **5**: 35–41 (1992).