Alex Bateman
leads the Pfam group at the Wellcome Trust Sanger Institute in the UK.

Daniel Haft
leads the TIGRFAMs group at The Institute for Genomic Research in Rockville, MD.

HMM-based databases in InterPro

Alex Bateman and Daniel H. Haft

Date received (in revised form): 14th June 2002

Abstract

Protein family databases are an important resource for protein annotation and understanding protein evolution and function. In recent years hidden Markov models (HMMs) have become one of the key technologies used for detection of members of these families. This paper reviews the Pfam, TIGRFAMs and SMART databases that use the profile-HMMs provided by the HMMER package.

Keywords: *protein family, hidden Markov model, SMART, Pfam, TIGRFAMs*

INTRODUCTION

Protein family databases have become a key tool in the armoury of the molecular biologist as well as the dedicated computational biologist. They provide a simplifying principle to the wealth of genomic data now available. Their classifications can aid the design of experiments and allow inferences of protein function. There are many databases that classify proteins into families. In this paper we will look at Pfam, TIGRFAMs and SMART, which share the HMMER package as their technology for classifying sequences. We will review the HMMER package and look each of these databases is turn to understand their commonalities as well as their differences.

PROFILE-HMMS

Pairwise sequence search methods such as BLAST[1] and FastA[2] are a rapid and sensitive way to identify similarities between proteins. In pairwise search methods a scoring matrix such as BLOSUM62 is used, which scores the alignment of each of the 20 amino acids against each other. For example a match of a glutamate to a chemically similar aspartate scores 2, whereas a glutamate matching a dissimilar valine scores −2. These methods treat all residues in an alignment in the same way using the scoring matrix. Pairwise search methods will always score a glutamate aligned to a valine the same anywhere in the sequence, whether the glutamate is a critical active site residue or a non-functional surface residue. Intuitively one would expect the active site residue to be mutated at a much lower frequency than the non-functional surface residue. However, pairwise methods cannot use this kind of information.

In the 1980s several researchers used information available in multiple sequence alignments to infer what kinds of residues were likely substitutions at each site in a protein. These tools that were developed became known as profiles. Profiles are able to catch information specific to a particular position in a protein. Therefore a conserved glutamate residue in an active site will score alignment to glutamates very highly and all other residues will score poorly. A non-functional surface residue with little conservation will allow any residue to match with a low score. These methods were found to be much more sensitive than pairwise search methods particularly for distant similarities. The original profile methods had rather arbitrary schemes for calculating what scores should be given to substitutions at each site. Although the methods worked well there was little statistical justification for the methodology.

Alex Bateman,
The Wellcome Trust Sanger Institute,
The Wellcome Trust Genome Campus,
Hinxton CB10 1SA, UK

HMMs are built from multiple sequence alignments

HMM interpretation of profiles

In 1994 Krogh *et al.*[3] introduced a hidden Markov model (HMM) interpretation of the profile. These models have become known as profile-HMMs to distinguish them from the more general term HMM.[4] HMMs are statistical models that are based on probabilities rather than scores. An HMM is represented by a series of states that can emit (or match) symbols and transitions between these states. At each state a symbol is emitted (or matched) with a certain probability. Each transition has a probability associated with it that describes how likely it is to move between any two states. The HMMER package uses an HMM architecture that is shown in Figure 1.

Profile-HMMs have a large number of parameters that need to be estimated. These can be derived in two different ways. In principle, a profile-HMM can be derived from unaligned sequences by training. A naive profile-HMM of an appropriate size can be refined by successive rounds of optimisation of its fit to the training sequence set until some form of model is created. The training procedure was an early focus of HMM work with protein sequences. In practice, however, the parameters for a profile-HMM are more accurately estimated from a multiple sequence alignment and this has become the method of choice.

THE HMMER PACKAGE

One of the most popular packages that allow users to make profile-HMMs is the HMMER package written by Sean Eddy. The HMMER package (pronounced 'hammer' for better comparison to BLAST) is freely available.[5] It is easy to use and requires no knowledge of how the profile-HMMs work. Here we provide a brief description of the major programs and their use. The HMMER package documentation provides tutorials and fuller explanations of all aspects of the package.

To build a profile-HMM one runs the *hmmbuild* program, which takes a multiple sequence alignment as input and produces a file representing the profile-HMM. This program will take about a second to run for all but the largest alignments. This profile-HMM can be used as it is to search sequences. However, the HMMER package also provides another program called *hmmcalibrate*. This program takes a profile-HMM and searches it against a simulated set of 5,000 proteins. This is used to calculate mu and lambda

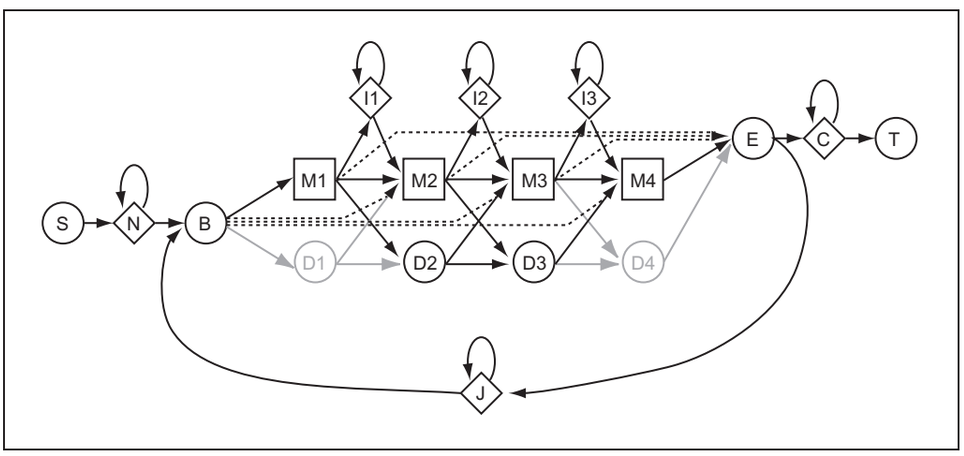

Figure 1: The 'plan 7' architecture of HMMER profile-HMMs. States are represented by circles and boxes. Arrows between states show transitions. This model has match states labelled M1 to M4. Each of these corresponds to one alignment column. The long looping transition from the E (end) to B (begin) state allows multiple repeating matches of the profile-HMM within a sequence

values that allow accurate estimation of *E*-values for the profile-HMM. Although this step can take several minutes to run it is strongly recommended. A calibrated profile-HMM is much more sensitive than an uncalibrated one.

The HMMER package contains a program called *hmmsearch* that is used to scan the profile-HMM against a database of sequences. A related program, *hmmpfam*, searches a single sequence against a library of HMMs. Both programs are computationally demanding and may be unacceptably slow on a single processor, but have options to engage multiple processors and to run on a parallel virtual machine.

Local and global models

The HMMER package provides the *hmmbuild* program for constructing profile-HMMs from protein multiple sequence alignments. When building a model the major decision is whether to build the model in ls (global) mode or fs (local) mode. The ls (default) mode enforces that the sequence matches to the whole profile-HMM from the first to the last match state. The fs (-f option) mode profile-HMMs allow only part of the profile-HMM to match to a target sequence (see Figure 2). Both of these types of models allow you to find matches to just part of the target sequence and are therefore local with respect to the target sequence. Both are able to find repeated hit regions if the target sequence has duplications; the seldom-used -g option allows only a single global match per target sequence. The fs mode may split a single match region into several local match regions and fail to cover the full length of the match. This may make it difficult to assign and compare the domain architectures of different proteins. It is the preferred mode for fragmentary sequences such as translated expressed sequence tags (ESTs). The ls mode profile-HMM is liable to miss partial sequences. In rare cases it may force spurious extension of or spanning between some strong local match regions, but in general it is the most sensitive for searching and produces the most accurate assignment of domain boundaries.

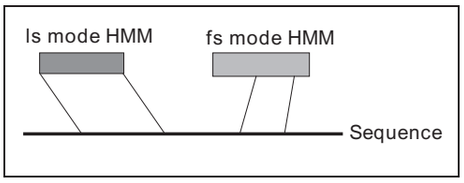

Figure 2: The two main profile-HMM construction modes available in HMMER. Both modes are local with respect to the sequence. The ls mode is global with respect to the profile-HMM whereas fs mode is local with respect to the profile-HMM

SEED alignments

When trying to keep track of all members of a protein family, researchers would tend to keep a multiple alignment of all known examples. As new sequences appeared they were appended onto the complete alignment. This strategy works well when curating one or two families. There are several problems with this approach, however, that cause it not to scale well. For very large families the complete alignment can become unmanageable and difficult to edit. As new sequences come into the database and old sequences are updated it is also difficult to edit the complete alignment of the family. Some homologous sequences may have undergone such extensive mutation, or have been mistranslated, that they contribute more noise than signal to the modelling of any family; forcing their alignment (or misalignment) will yield weaker, less useful mathematical signatures for the family than would leaving them out.

A good seed alignment leads to an HMM that identifies and aligns the entire family

To address such problems, the concept of SEED alignments was introduced by Sonnhammer *et al.*[6] A SEED alignment is a collection of sequences that are somehow representative of a protein family. This SEED alignment is then used to build an HMM that can be used to search a protein database with the aim to identify (and align) all examples of the protein family (see Figure 3). If the HMM

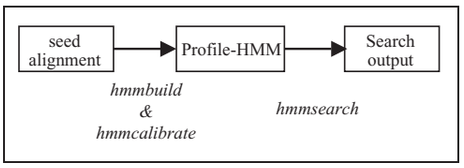

Figure 3: Scheme for using the programs in the HMMER package as used by Pfam, TIGRFAMs and SMART. Programs are in italic and files are in boxes

does not perform as well as expected, then more sequences can be added to the SEED alignment and a new HMM constructed.

The majority of proteins have at least one region identified by a Pfam-A model

Pfam

Pfam contains a large collection of profile-HMMs as well as alignments and annotation.[7] Pfam is available via the World-Wide Web.[8] Pfam contains well-known domain families such as the Ig domain and SH2 domain. Release 7.2 of Pfam contains 3,735 curated families. The main goal of Pfam is to classify novel protein sequences into their constituent functional protein domains. Although Pfam tries to classify proteins into domain similarities there are many families for which it is probable that the defined alignment only covers part of a domain or multiple domains. Pfam aims to correct these as more data become available, particularly from structural studies.

The basis of a Pfam family is the SEED alignment. SEED alignments contain a non-redundant representative set of examples of the family. In general SEED alignments are small allowing the Pfam curators to quickly assess their quality and manually edit them if needed. For most SEED alignments standard sequence alignment methods such as ClustalW[9] and T-Coffee[10] produce adequate quality alignments.

Pfam also provides FULL alignments that try to provide an alignment of all examples of a particular family. FULL alignments can be large, with the largest family, the HIV GP120 glycoprotein, containing more than 27,000 sequences.

Pfam-A and Pfam-B

The Pfam database is composed of two types of family. Firstly there are high-quality curated families that are called Pfam-A families. Over 70 per cent of proteins in SWISS-PROT[11] and TrEMBL have a match to at least one Pfam-A domain. This is useful for users whose proteins are in the 70 per cent. However what about the other 30 per cent of proteins? In an effort to be comprehensive Pfam contains a large number of automatically derived protein families, known as Pfam-B. These families are constructed by taking the PRODOM database[12] and removing the families that already are covered by Pfam-A families. In cases where a Pfam-A family partly overlaps a PRODOM family, the PRODOM family is trimmed to exclude the Pfam-A region.

Pfam-B families are similar to Pfam-A families, but they do not include profile-HMMs or FULL alignments or annotation. Pfam-B also changes at every release of Pfam and its identifiers are not stable. So although a useful resource for suggesting regions of similarity not found in Pfam-A, it is not curated in any way and is of much lower quality than Pfam-A families.

Global and local Pfam profile-HMMs

As was mentioned earlier profile-HMMs can be built in either ls (global) or fs (local) mode. A feature of the Pfam database is that it provides models and thresholds in both modes. The set of sequences found in the FULL alignment in Pfam is a combination of matches from both ls and fs mode searches. This means that fragmentary sequences are picked up well by this method. The downside to this increased sensitivity is a doubling in the number of profile-HMMs to search. This can mean a significant compute burden if you need to search a large number of sequences such as a complete genome.

SMART

The SMART (Simple Modular Architecture Research Tool) database[13,14] currently contains 639 entries in Release 3.3. Currently SMART concentrates on domains and repeats found in signalling proteins (both eukaryotic and bacterial), extracellular and nuclear domains. The scope of the families matched by SMART is similar to Pfam. The entries in SMART have been made as comprehensive as possible and in many cases are more sensitive than the corresponding Pfam family. Although there is considerable overlap between entries in SMART and Pfam there are many examples where one database finds matches that the other does not. The database can be accessed via the WWW.[15]

Some TIGRFAMs models distinguish among related proteins

SMART offers sequence searches and relational queries

The SMART database allows a large number of search configurations from the home page. One can do complex queries of the underlying data by domain architecture as well as taxonomic distribution. For example, one can find all *Arabidopsis thaliana* tyrosine protein kinases with a single query. A direct interface to the SMART relational database is provided where SQL commands can be used to query the *protein*, *taxonomy* and *domain* tables. To further increase search sensitivity SMART includes the option to search for outlier matches to a family using a BLAST search. For a given query sequence a BLAST search is carried out against all sequences of known domains, which may find similarities that profile-HMMs have missed. When searching sequences other features such as low complexity, coiled-coils and transmembrane helices are predicted and displayed on the search output.

TIGRFAMs

TIGRFAMs is a collection of profile-HMMs, the annotations connected to each, and the alignments from which the models are built.[16] TIGRFAMs was designed from the beginning to supplement, not replace, searches performed with Pfam profile-HMMs.

Orthologues with conserved function are 'equivalogs'

The focus of TIGRFAMs is families of proteins, especially from prokaryotes, for which function is conserved to the point that a well-informed annotator would assign the same protein name across different species with good confidence.

It has been shown many times that a single amino acid substitution can change the specificity of an enzyme or transporter. The assignment of protein function by homology is always a prediction. High confidence in a functional prediction is indicated, typically, by omitting the qualifier 'putative' from the protein name. The goal of TIGRFAMs is not to make unchallengeable assertions of protein function. It is to capture, in the profile-HMM itself and in the information attached to it, a set of criteria for identifying proteins with high confidence. Care is taken in building each model that the members of the seed alignment are appropriately chosen, that the alignment is substantially correct, and that the parameters used during profile-HMM construction lead to models that cause members of the family reliably to score higher than non-members. Examinations during the building of each model include the construction of phylogenetic trees, consideration of metabolic context, and following links into annotated protein databases[11,17] and the scientific literature. The value of the collection comes from the care with which mutually exclusive subsets of related proteins are represented in profile-HMMs which, by virtue of their construction and assigned cutoffs, produce non-overlapping hits.

The scope of a profile-HMM is the set of all proteins whose score against the profile-HMM is higher than its gathering threshold. TIGRFAMs models belong to several different 'isology types', according to the degree of functional similarity among the proteins in their scope. The term isology means 'type of relatedness' and generally refers to homology, since TIGRFAMs does not deliberately include other forms of sequence similarity. We have introduced the term 'equivalog'[16] to

describe homologous proteins whose function is equivalent, and has been equivalent continuously back to the last common ancestor. Equivalogs include orthologues, laterally transported genes (xenologues), and paralogues, as long as function is conserved from a common ancestor. Related proteins from a single ancestral sequence are equivalogs even if other lines from the same ancestral sequence have evolved different functions. A profile-HMM is called an equivalog model if all proteins within its scope are believed to share the same function.

Useful equivalog models cannot be built for some protein families. Pairs of proteins known to have different functions may be more closely related to each other than either is to an orthologue that shares function with one of the pair. In some cases, the member with altered function may have undergone enough mutations, including insertions, deletions, truncations and replacements of key residues, that setting a threshold for an equivalog profile-HMM to exclude the altered sequence is easy to do. But in other cases the sequence with altered function may score nearly as well as its paralogue, and an equivalog model may be impossible to build. In such a case, it is still useful to build a profile-HMM, but the type of the model will be 'subfamily' instead of 'equivalog'. Subfamily models identify regions of sequence space in which equivalog model construction is difficult, and in which annotation by homology to the most closely related sequence is particularly error-prone. See Table 1 for a listing of the homology types in use at TIGRFAMs and their meanings.

Validation is more problematic for TIGRFAMs than for Pfam or SMART because the models are intended to support specific inferences about function. This goes beyond the statistically tractable question of whether or not two sequences are homologous. Short of performing experiments to verify function, validation must be done by checks of the results for logical self-consistency.

First, no two equivalog models should hit the same region of the same protein. This rule is checked periodically as a new genomic sequence becomes available. Second, equivalog-based identifications should fit publicly available annotation that is based on experimental evidence. In

Table 1: Homology types in TIGRFAMs 2.0

Isology type	Definition
domain	Represents a sequence domain, or region of sequence similarity among sequences that otherwise tend to differ elsewhere along their sequences. It may coincide with an independently folding structural domain but is not guaranteed to do so. 'domain' families should encompass all homologous sequences out to the limits of detection.
equivalog	Subfamily of proteins whose function has been conserved continuously since the last common ancestral sequence. Regulatory mechanisms, alternative substrates and other features may differ among members. Equivalogs will tend to be orthologues, but may include laterally transferred genes.
equivalog_domain	Region of sequence associated with a conserved function, but found with some regularity in proteins with varied architecture, as in fusions with additional functional domains.
hypoth_equivalog	Subfamily for which the function is unknown (conserved hypothetical protein), but for which it is suggested that all may have the same function.
hypoth_equivalog_dom	Region that acts as a domain with respect to its length within the proteins that contain it and as 'hypoth_equivalog' with respect to presumed conserved function.
paralog	Set of proteins, restricted to a species or set of closely related species, that arose by a series of duplication events recent relative to the separation of the genus from other lineages.
subfamily	Branch of a superfamily. Generally, members of a subfamily are closely related enough to suggest that a substantial fraction have the same or similar functions. A subfamily model should exclude distant homologues, if any exist.
subfamily_domain	Region that acts as a domain with respect to its length within the proteins that contains it and acts a subfamily with respect to the relative degree of similarity among its members.
superfamily	Set of proteins with the same overall architecture, encompassing all homologous sequences out to the limits of detection. Generally, this type is assigned if the grouping is believed to contain at least two clades that differ from each other in function.

some cases, a model was built as equivalog rather than subfamily deliberately and a known counterexample was noted in the model's INFO file in a DR EXCEPTION. For example, SWISS-PROT:P29707 was noted as a Na^+-translocating exception to TIGR01039, proton-translocating ATP synthase F1 beta. Third, multiple hits to equivalog models in small genomes should be rare and should tend to suggest relatively recent duplication. Finally, functional assignments made by equivalog models should make sense in their metabolic context. Beta subunits should appear in the same genomes as alpha subunits, third steps in pathways should occur in the same genomes as second, etc.

TIGRFAMs equivalog models are similar in scope to many clusters of orthologous groups (COGs)[18] but are often smaller. The basis of COGs is transitive similarity by bi-directional best hit relationships. In the event of ever-increasing numbers of completed genomes, it may be expected that some lineage may be found in which the orthologous protein has developed an altered function. These would likely be retained within a COG but excluded from an equivalog family.

TIGRFAMs families may contain widely distributed Pfam and SMART domains

Uses of TIGRFAMs

Equivalog HMMs offer automated annotation of protein function

TIGRFAMs models are designed to assist in genomic annotation. Equivalog profile-HMMs may be used to make fully automatic assignments of protein names, gene symbols and EC numbers. Human review is recommended where a single Equivalog profile-HMM hits twice or more in a microbial genome and where the length of the protein differs substantially from the length of the profile-HMM. Curation is recommended also for cases of hits between the trusted cut-off (which is used as the gathering threshold) and noise cut-off, and for hits to models of type other than 'equivalog'. Hits below the noise cut-off should be ignored, even if the *E*-value is very low (highly significant), unless there happens to be no profile-HMM from Pfam or SMART hitting the same protein.

Comparison of TIGRFAMs, Pfam AND SMART

Annotation by homology consists of two processes. The first is to determine whether or not the observable level of sequence similarity indicates homology, a common evolutionary origin. The second is to determine what inferences about protein structure and function can be drawn from the evolutionary relationship. The Pfam and SMART profile-HMM libraries make robust assignments of homology regions, even in instances of small, highly diverged domains. These profile-HMM hits will often clarify whether a small region of sequence similarity seen in BLAST output represents homology or a fortuitous match. In contrast, most TIGRFAMs models do not extend the limits of search sensitivity. Typically, they exist to delineate equivalog families, or near-equivalog subfamilies, among sets of related proteins all of which would appear in a BLAST search for any of its members. TIGRFAMs give very accurate functional assignments compared to the often less specific Pfam and SMART annotations.

A good illustration of the differences between the primarily domain-oriented classification system of Pfam (or SMART) and the largely family-level classification system of TIGRFAMs is seen in the case of DEAD/DEAH box helicase. The model PF00270 from Pfam describes this domain of about 200 residues, found in over 1,000 proteins. Five equivalog models have been built so far (TIGR00580, TIGR00595, TIGR00631, TIGR00643, TIGR01389), representing five functionally distinct protein families with an average length of 700 residues, each containing a DEAD/DEAH box helicase domain. These equivalog models are strictly non-overlapping with each other and contain about 50 sequences each. Often several different domains are contained in a single TIGRFAMs entry; see Figure 4 for an example.

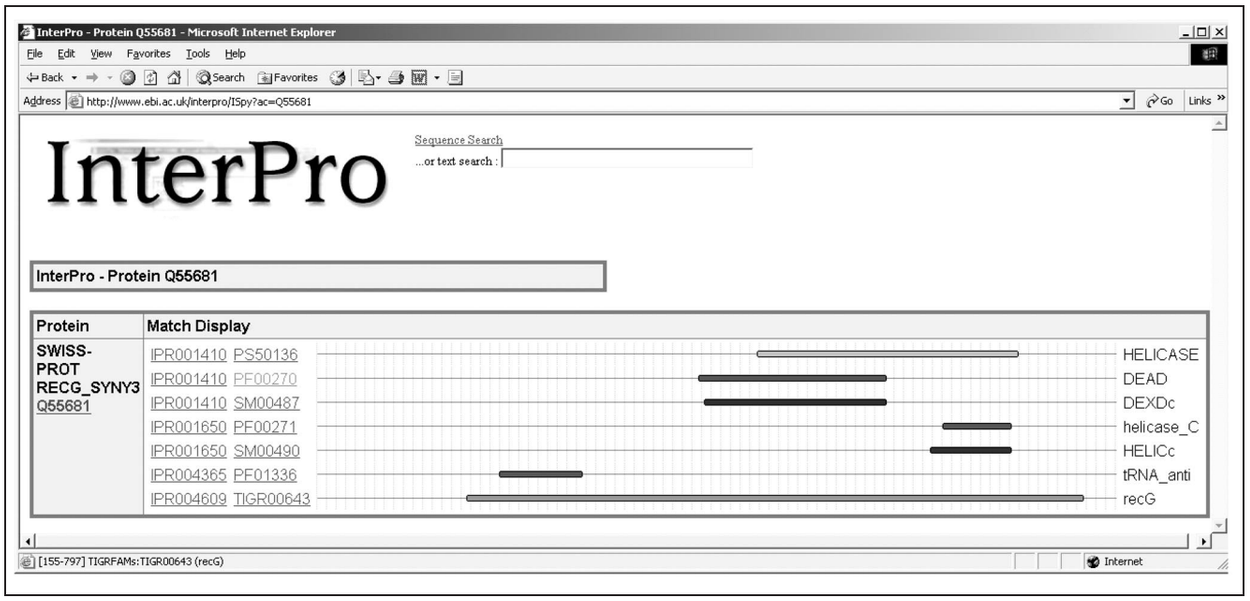

Figure 4: InterPro view of SWISS-PROT entry RECG_SYNY3 (Q55681) shows links to three Pfam domains and two SMART domains, and one TIGRFAMs family. This protein is the ATP-dependent DNA helicase RecG from *Synechocystis* sp. strain PCC6803

The three profile-HMM databases also support genomic-scale computational studies. An example of the principle of using profile-HMM results at the genomic level for data mining is seen in the work of Sprinzak and Margalit.[19] They found particular pairs of InterPro (including Pfam and SMART) signatures to be over-represented among 2,908 pairs of interacting *Saccharomyces cerevisiae* proteins, and that these correlations enable additional predictions of protein–protein interactions. For example, pleckstrin homology domain proteins (PF00169, SM00233) tend to interact with SH3 domain proteins (PF00018, SM00326).

The protein domain and family identifications made possible by Pfam, SMART and TIGRFAMs profile-HMMs support fairly detailed investigation into the available information about the each protein. A particular protein may show several distinct domains, with each connecting the protein to abstracted information about structure and function. The Pfam and SMART WWW interfaces provide extensive links to descriptive text, solved three-dimensional structures, lists of proteins sharing each domain, etc. However the interfaces provided by each also include features not available in the other database. For example SMART provides links to OMIM (Online Mendelian Inheritance in Man) for families that include human disease proteins as well as summaries about the cellular localisation of the members of the family. Within Pfam the multiple sequence alignment contains extra information about secondary structure, solvent accessibility and active site information.

InterPro provides a hierarchical classification of protein families; however neither TIGRFAMs nor Pfam include any hierarchical classification. SMART provides some subfamily relationships, for example, within the protein kinases subclasses are defined. Also different structural classes are defined for the Ig domain.

Each profile-HMM that matches a protein produces a bit score and an *E*-value that can be compared to cut-off scores. It produces N-terminal and

C-terminal coordinates, and therefore a domain. Performing complete profile-HMM searches on protein sets of interest, loading information on both hit regions and the profile-HMMs themselves into a sufficiently powerful database system to make precomputed results readily available is essential to many forms of analysis. Loading profile-HMM search results into a relational database creates opportunities for designing genomic analysis and display tools,[20] and allowing ad hoc queries for interesting classes of protein, and performing data mining.

The three databases SMART, TIGRFAMs and Pfam differ in their focus (eukaryotic, prokaryotic and general, respectively) and the scope of the typical model (domains, narrow subfamilies and general), and in the features provide by their respective web sites. Pfam and SMART provide exquisite search sensitivity, curated domain boundaries, carefully chosen cut-offs and rich annotation to support high-quality annotation. Many TIGRFAMs models focus on specific proteins, subfamilies from within broader superfamilies, often with some conserved function; such models should be well-suited to automated annotation of selected proteins, for phylogenetic profiling,[21] and for metabolic reconstruction. Through the mechanism of InterPro[22] the results from searches of these different resources can be easily integrated allowing a comprehensive view of protein architecture and function.

References

1. Altschul, S. F., Madden, T. L., Schaffer, A. A. *et al.* (1997), 'Gapped BLAST and PSI-BLAST: A new generation of protein database search programs', *Nucleic Acids Res.*, Vol. 25, pp. 3389–3402.
2. Pearson, W. R. and Lipman, D. J. (1988), 'Improved tools for biological sequence comparison', *Proc. Natl Acad. Sci. USA*, Vol. 85, pp. 2444–2448.
3. Krogh, A., Brown, M., Mian, I. S. *et al.* (1994), 'Hidden Markov models in computational biology', *J. Mol. Biol.*, Vol. 235, pp. 1501–1531.
4. Eddy, S. R. (1996), 'Hidden Markov models', *Curr. Opin. Struct. Biol.*, Vol. 6, pp. 361–365.
5. URL: http://hmmer.wustl.edu
6. Sonnhammer, E. L. L., Eddy, S. R. and Durbin, R. (1997), 'Pfam: A comprehensive database of protein domain families based on seed alignments', *Proteins*, Vol. 28, pp. 405–420.
7. Bateman, A., Birney, E., Cerruti, L. *et al.* (2002), 'The Pfam protein families database', *Nucleic Acids Res.*, Vol. 30, pp. 276–280.
8. URLs: www.sanger.ac.uk/Software/Pfam/ (UK), http://pfam.wustl.edu/(USA), http://www.cgr.ki.se/Pfam/(Sweden), http://pfam.jouy.inra.fr/(France).
9. Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994), 'CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice', *Nucleic Acids Res.*, Vol. 22, pp. 4673–4680.
10. Notredame, C., Higgins, D. G. and Heringa, J. (2000), 'T-Coffee: A novel method for fast and accurate multiple sequence alignment', *J. Mol. Biol.*, Vol. 302, pp. 205–217.
11. Bairoch, A. and Apweiler, R. (2000), 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Res.*, Vol. 28, pp. 45–48.
12. Corpet, F., Servant, F., Gouzy, J. and Kahn, D. (2000), 'ProDom and ProDom-CG: Tools for protein domain analysis and whole genome comparisons', *Nucleic Acids Res.*, Vol. 28, pp. 267–269.
13. Schultz, J., Milpetz, F., Bork, P. and Ponting, C. P. (1998), 'SMART, a simple modular architecture research tool: Identification of signaling domains', *Proc. Natl Acad. Sci. USA*, Vol. 95, pp. 5857–64.
14. Letunic, I., Goodstadt, L., Dickens, N. J. *et al.* (2002), 'Recent improvements to the SMART domain-based sequence annotation resource', *Nucleic Acids Res.*, Vol. 30, pp. 242–4.
15. URLs: http://smart.embl-heidelberg.de/ (Germany), http://smart.ox.ac.uk (UK).
16. Haft, D. H., Loftus, B. J., Richardson, D. L. *et al.* (2001), 'TIGRFAMs: A protein family resource for the functional identification of proteins', *Nucleic Acids Res.*, Vol. 29, pp. 41–43.
17. Wu, C. H., Huang, H., Arminski, L. *et al.* (2002), 'The Protein Information Resource: An integrated public resource of functional annotation of proteins', *Nucleic Acids Res.*, Vol. 30, pp. 35–37.
18. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V. *et al.* (2001), 'The COG database: New developments in phylogenetic classification of

proteins from complete genomes', *Nucleic Acids Res*., Vol. 29, pp. 22–28.

19. Sprinzak, E. and Margalit, H. (2001), 'Correlated sequence-signatures as markers of protein–protein interaction', *J. Mol. Biol.*, Vol. 311, pp. 681–692.
20. Peterson, J. D., Umayam, L. A., Dickinson, T. *et al.* (2001), 'The Comprehensive Microbial Resource', *Nucleic Acids Res.*, Vol. 29, pp. 123–125.
21. Pellegrini, M., Marcotte, E. M., Thompson, M. J. *et al.* (1999), 'Assigning protein functions by comparative genome analysis: protein phylogenetic profiles', *Proc. Natl Acad. Sci. USA*, Vol. 96, pp. 4285–4288.
22. Apweiler, R., Attwood, T. K., Bairoch, A. *et al.* (2000), 'InterPro – an integrated documentation resource for protein families, domains and functional sites', *Bioinformatics*, Vol. 16, pp. 1145–1150.