HIDDEN MARKOV MODELS AND LARGE-SCALE GENOME ANALYSIS

Sean R. Eddy
Dept. of Genetics, Washington University School of Medicine
4566 Scott Ave., St. Louis MO 63110
eddy@genetics.wustl.edu

ABSTRACT

PFAM is a database of multiple alignments and hidden Markov models (HMMs) of common, conserved protein domains. PFAM HMMs complement BLAST analysis in the annotation of the *C. elegans* and human genome sequencing projects at Washington University and the Sanger Centre. PFAM2, based on full, gapped multiple alignments of structural and/or functional protein domains, currently contains 527 models. PFAM/HMM analysis hits at least one domain in 24% of the predicted proteins in the *C. elegans* genome project. 8% of *C. elegans* proteins are annotated as multidomain proteins by PFAM, with up to 5 different kinds of recognized domains per protein and up to 44 total recognized domains per protein.

INTRODUCTION

Automated, large-scale prediction of the functions and structures of predicted protein sequences is one of the most pressing problems faced by genome bioinformatics groups [1-3]. These computational predictions largely rely on database similarity searches; primarily, fast pairwise local alignment methods like BLAST [4] and FASTA [5]. In the *Caenorhabditis elegans* genome sequencing project at the Washington University Genome Sequencing Center and the Sanger Centre [6-8], informative BLAST hits are obtained for about 45% of predicted nematode proteins. There is great interest in increasing the fraction of protein sequences for which we can infer structural or functional properties accurately, automatically, and efficiently.

Especially in the higher eukaryotes, many proteins have evolved by extensive re-use and shuffling of domains [9]; for example, fibronectin type III domains, or protein kinase catalytic domains. This is both bad news and good news for genome-scale bioinformatics. The bad news is that the sheer ratio of hits to common protein domains can overwhelm a sequence analyst, causing missed or erroneous predictions that simply result from confusion. A protein that contains one or more common protein domains may produce hundreds or thousands of BLAST hits. The top BLAST hit may not correspond to a homologous gene, but rather to a homologous domain in an otherwise non-homologous sequence. Furthermore, there may be so many strong hits to the conserved domain(s) that other weak but more informative BLAST hits are missed. Specialized BLAST post-processing programs have been developed to help with these problems [10-12].

The good news is two-fold. First, the number of common protein domain families is relatively limited and tractable. Several estimates indicate that on the order of a thousand protein domain families account for a significant fraction of all proteins. Available protein

family databases ("second generation" databases that organize the primary Swissprot, PIR, and GenPept databases into evolutionary families) now include Prosite [13], PRINTS [14], BLOCKS [15], Prodom [9], ProClass [16], and SBASE [17], as well as the PFAM database discussed in this paper [18]. Second, when multiple sequence alignments, consensus patterns, and/or structures are available for a protein family, potentially powerful alternative search and detection methods can be utilized. These methods range in complexity from Prosite's motif patterns, to multiple alignment based "profile" methods [19-21] and structure based "inverse protein folding" methods such as threading or 3D/1D profiles [22].

HMM-PROFILES

A *profile* is defined here as a linear model of the consensus primary structure of a sequence family, containing position-specific scores for amino acids and insertions/deletions at each profile position. The main difference between profile alignment and standard pairwise sequence alignment methods is that the pairwise methods use *position-independent* scores (e.g. a PAM or BLOSUM substitution matrix [23]), whereas a profile uses *position-specific* scores. Position-specific scores allow one to model the fact that certain positions in a protein are crucial to its folding and function, whereas other residues do not matter as much.

A profile is usually built from a multiple alignment of a family of related sequences. Profiles can also be built from structural data (e.g. '3D-1D profiles' [24]). For each consensus primary structure position, twenty amino acid scores and two or more insertion/deletion scoring parameters are calculated. For example, if one is building a profile from a multiple alignment, most of the columns of the alignment correspond to the consensus primary structure positions. If a column appears to correspond to a strongly conserved cysteine, the assigned score for C may be strongly positive, and negative for the other 19 amino acids; if the column is apparently random and unconserved, all 20 scores may be close to zero. Ungapped profiles (called 'blocks' or 'weight matrices') are also used in some approaches [25, 26]. The term "position-specific scoring matrix" (PSSM) introduced by the Henikoffs is synonymous with a profile [25].

Position-specific scoring greatly increases the number of parameters which must be determined. A PAM or BLOSUM substitution matrix used by BLAST or FASTA contains 190 scores, and is determined by counting amino acid pairs over a large database of different trusted pairwise alignments. A profile contains about 22N scores (20 residue scores and 2 gap penalties per position) where N is the consensus length of the family (generally in the 100-500 range). These scores must be determined from a single multiple alignment (or one or a few 3D structures). Managing a model with thousands of poorly determined parameters is a challenge.

Probabilistic modeling is a nice approach to complicated inference problems. A class of probabilistic models called hidden Markov models (HMMs) have been used extensively in the speech recognition community for making speaker-independent linear profiles of digitized acoustic signatures of spoken words [27]. David Haussler's group at UC Santa Cruz introduced HMMs as a useful probabilistic modeling framework for biological sequence profiles (Figure 1) [28]. The parameters of an HMM-profile are probabilities, not arbitrary scores. Bayesian and/or maximum likelihood approaches are used to determine all the probability parameters of the HMM-profile. There is a recent review of HMM-profile methods [29].

The technical details of HMM-profile methods are beyond the scope of this paper, but the following general points are relevant. 1) HMM-profiles are merely an improved formalization

Multiple sequence alignment:

```
seq1   A  F  -  V
seq2   A  F  -  V
seq3   A  Y  d  V
seq4   G  F  -  V
```

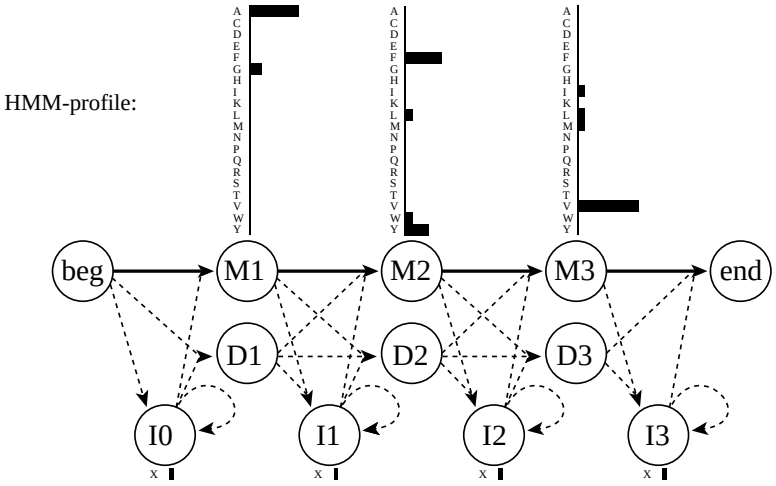

Figure 1: Top: A tiny example alignment of four sequences with three aligned "consensus" columns (upper case amino acid codes), and one insertion relative to the consensus (lower case amino acid code). Bottom: A cartoon view of an HMM-profile built from the same alignment. An HMM is composed of three things: *states*, *state transition probabilities*, and *symbol emission probabilities*. HMM "states" (circles) each align to one residue (or no residue). Match states (labeled M) align to a particular residue with some "emission probability" (illustrated schematically above each state, with a black bar for each residue of size proportional to the probability of that residue). Emission probabilities are usually calculated from the observed counts of residues in the corresponding column of the alignment. Delete (D) and insert (I) states allow for deletions and insertions relative to the consensus; a product of state transition probabilities (arrows) defines the probability of any given path (alignment) through the HMM.) Though the terminology may seem obscure, the HMM-profile is quite close to the standard formulation of sequence profiles, but with probabilities replacing arbitrary scores. For example, the state transition probability into an insert state (for a one-residue insertion) and then from the insert state back to itself (for each successive residue in an insertion) is a probabilistic version of the "gap-open" and "gap-extend" penalties commonly used in biological sequence alignment.

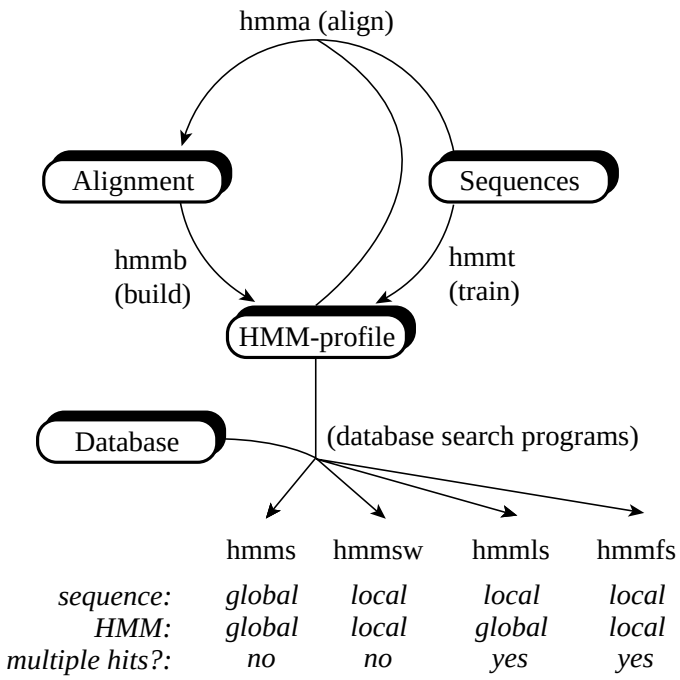

Figure 2: Main programs in the HMMER software package.

of the scoring scheme of previous profile methods. Essentially the same alignment algorithms are used. 2) The advantage of HMM-profiles over previous profile methods is that the large number of scoring parameters, including gap scores, may be automatically and consistently determined. 3) Because the statistics are consistent across different models and no manual tweaking of parameters is needed (in theory), one can automatically apply hundreds of different HMM-profiles of common protein domains to a complete genome, while still being able to efficiently interpret the results. HMM-profiles are well suited to large scale analysis problems.

HMMER

HMMER is a freely available software package that implements HMM-profiles for protein and nucleic acid sequence analysis. A flowchart of the programs in the HMMER package is shown in Figure 2.

The program *hmmb* ("HMM build") builds an HMM-profile from an existing multiple alignment. This is a quick process, usually taking one or two seconds. The resulting HMM-profile can be thought of as a compiled statistical representation of the multiple alignment. Options in *hmmb* allow the user to choose amongst different sequence weighting options, different probability parameter optimization strategies, and different ways to contribute "prior" information about sequence alignments (via substitution matrices or Dirichlet priors [30]).

The program *hmma* ("HMM align") aligns any number of sequences to an existing

HMM-profile. This allows one to build an HMM-profile of a "seed" alignment of a small number of carefully aligned representative sequences of a large sequence family, then use this HMM-profile to automatically create a high-quality alignment of the rest of the family. The number of sequences that can be handled is effectively unlimited. To date, the largest HMM-managed alignments contain tens of thousands of sequences, starting from manageable "seeds" of tens of sequences. The seed alignment strategy is central to the maintenance of the PFAM HMM database (see below).

Four database search programs are in the package. The program *hmms* ("HMM search") looks for global alignments of the entire HMM to the entire query sequence. Because HMM-profiles are usually models of domains rather than complete protein sequences, *hmms* is rarely used. The program *hmmsw* ("HMM Smith/Waterman") is more useful for database searches. It is an HMM version of the standard Smith/Waterman algorithm [31], allowing local alignments that match any fragment of the HMM to any fragment of the query sequence. Unpublished dynamic programming algorithms are used in two more powerful search programs. The program *hmmls* ("HMM local search") looks for one or more non-overlapping matches of the complete HMM to parts of the query sequence. The program *hmmfs* ("HMM fragment search") is similar, but looks for one or more non-overlapping matches of any fragment of the HMM to parts of the query sequence. If complete domains are expected, *hmmls* is typically the most sensitive and useful search program in the package; on the other hand, *hmmfs* can find fragmentary hits that *hmmls* cannot.

The program *hmmt* ("HMM training") trains an HMM from initially unaligned sequences, resulting in both a multiple alignment and a model. The algorithms used are of theoretical interest and the alignments are sometimes superior to those produced by more conventional multiple alignment programs. However, in practice, we have found that *CLUSTALW* [32] produces superior alignments about two-thirds of the time.

Other accessory programs in the package include *hmme*, which emits sequences consistent with a given HMM-profile (useful for simulation experiments or debugging); and *hmmconvert*, which converts HMMER format to other model formats. em hmm-convert can convert a HMMER HMM into GCG Profile format (with some loss of information), which allows the use of fast hardware implementations of GCG profile search (e.g. the Compugen Biocellerator).

HMMER source code, executables for various UNIX platforms, and documentation are available at `http://genome.wustl.edu/eddy/hmmer.html`. Other freely available HMM-profile implementations include SAM from the Haussler group at UC Santa Cruz (`http://www.cse.ucsc.edu/research/compbio/sam.html`), and PFTOOLS from Phillip Bucher (`http://ulrec3.unil.ch:80/ftp-server/pftools/`).

THE PFAM DATABASE

Using profile methods, one can readily build a profile of one's favorite sequence family and search a genome or sequence database for more members of the family: a one profile, many sequences problem. Systematic genome analysis presents a more complicated problem. For each new sequence, we wish to use profile analysis to identify what known domains it contains: a many profiles, one sequence problem. For the second problem, we need a large database of profiles of known domains. Shortly after the completion and release of the HMMER software, a group of collaborators including myself, Erik Sonnhammer, and Richard Durbin began creating

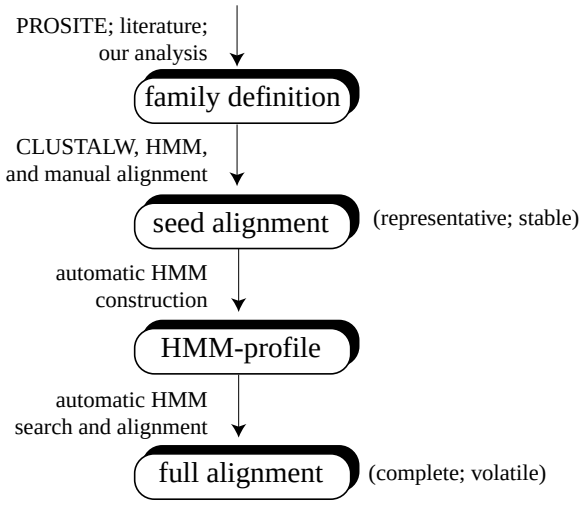

Figure 3: Flowchart of the construction of PFAM seed alignments, HMM-profiles, and full alignments.

an HMM-profile library that we could apply to *C. elegans* genome analysis.

There are a number of 'second generation' protein sequence databases which organize proteins into families, consensus models, and multiple alignments. Premiere amongst these is probably the PROSITE database developed by Amos Bairoch and collaborators [13]. PROSITE defines which known sequences belong to a particular family, gives careful and extensive documentation of the family's structure and function, and gives a motif pattern (regular expression) that recognizes the family members. Other protein family databases include PRINTS, BLOCKS, PRODOM, and SBASE. None of these were entirely suited as the basis for the development of a large HMM-profile database. We wanted comprehensive, manually curated, gapped multiple alignments of whole protein domains. The closest to our needs is the PRINTS database, which provides manually curated multiple alignments of representative members of each sequence family [14]; however, because the underlying search/alignment method of PRINTS is an ungapped 'fingerprint' method (akin to BLOCKS), the PRINTS database warns that the alignments have only been checked to be valid under the short regions that correspond to the ungapped fingerprints. Though we decided that none of these databases were perfectly suited to our purposes, we took extensive advantage of them (especially PROSITE) in creating our own alignment and HMM-profile database, which we call PFAM ('Protein FAMilies') [18].

There are three important files for each protein family in PFAM. The *seed alignment* (.seed file) is a multiple alignment of a representative subset of domain sequences. The *HMM-profile* (.HMM file) is built from the seed alignment. The *full alignment* is generated automatically by searching Swissprot with the HMM-profile and using HMMER to automatically align all the significant hits into a new alignment. A rough flowchart of PFAM model construction is shown in Figure 3.

The definition of the family is usually taken from PROSITE, and less frequently from the literature or our own domain identification research. Though we use independent structural and biochemical data wherever possible, it is important to keep in mind that family definitions in

PFAM and other databases is always subjective and operational, heavily dependent on the search method being used. If a known structural family cannot be recognized by a single HMM-profile, we split the family into two or more PFAM families that can be adequately recognized. Some PROSITE families were split in this way, because even though they share a short PROSITE active site motif, subfamilies have very different sequence consensus and sometimes even different structural folds. On the other hand, PFAM contains some families which PROSITE does not model, because even though the sequences are clearly related, no positions are conserved enough to make a discriminative PROSITE pattern (globins, for instance, which are now modeled in PROSITE by an HMM-profile developed by Bucher and collaborators.)

The definition of the bounds of the domain to be modeled is subjective. Again, we use 3D structural data if it is available. In some cases, for certain highly repetitive domains, intron position in eukaryotic genes is somewhat informative. Otherwise, domain definition is operational, based on the recognizable limits of sequence similarity among a group of sequences.

The seed multiple alignments are also subjective. We select a group of representative domain sequences and use CLUSTALW and HMM training, followed by manual editing, to produce the alignment. If an HMM built of the alignment fails to recognize one or more trusted members of the family, one or more new representative sequences is added, and the alignment is revised.

Thus, the process of generating a seed alignment is tedious and relies heavily on 'expert' subjective input, though we are computer-assisted at each point. An important feature of PFAM is that the seed alignment is considered to be a stable, reasonably permanent resource. After the seed is generated and documented, HMM software takes over, and the rest of the database, in particular the generation of the full alignments, is maintained fully automatically. This update strategy is a necessary feature if PFAM is to survive subsequent releases of the primary databases such as Swissprot. If all of PFAM had to be regenerated each time Swissprot is updated, it could not be maintained.

The current release of PFAM is PFAM 2.0, containing 527 families. Some relevant statistics about the database are given in Table 1. A detailed paper on the PFAM database was recently published [18]. Since we began the development of PFAM, at least two other profile databases have begun to be developed independently by other groups: the StrProf database from the Kanehisa group [33], and a growing HMM-profile collection in PROSITE [13].

PFAM WEB SITES AND ON-LINE SEARCHING

PFAM 2.0 is freely available via FTP and the Web. The U.S. home page is `http://genome.wustl.edu/Pfam/`. The U.K. home page is `http://www.sanger.ac.uk/Pfam/`. The Web pages are separately maintained and differ slightly in surface functionality, but the same database underlies them both. Both servers allow downloads of all or any one of the PFAM alignments and models, or browsing of the documentation for each family. The documentation is brief and relies heavily on links to PROSITE, PRINTS, and other Web resources for more detail.

Both servers allow on-line analysis. A Web user can cut and paste a query sequence into their browser and have it searched against one, a few, or all PFAM HMMs. The server returns the results in text form, tabular summary form, and in a Java applet that gives a color cartoon of the domain structure of the query protein. The PFAM alignments are also viewable

Table 1
Summary of protein families in PFAM 2.0

Largest families:	# seqs in seed	# seqs in full
C2H2 zinc fingers	165	1826
Ig superfamily	65	1351
Protein kinase catalytic domains	67	928
EGF domains	74	854
EF-hand domains	86	790
globins	61	699
7-TM receptors	64	597
Sequence lengths in full alignments:		
mean:	200	
shortest:	9	N-term neurohypophysial hormones
longest:	807	7-TM receptors, family 3
Pairwise sequence identities in full alignments:		
mean:	41%	
highest:	89%	Influenza virus nucleoprotein
lowest:	17%	PH domains
Number of sequences in full alignments:		
mean:	74	
most:	1826	C2H2 zinc fingers
least:	11	N-term of laminins (domain VI)

in a Java applet. An example of the search results from the U.S. server for the receptor tyrosine kinase Sevenless, from *Drosophila melanogaster*, is shown in Figure 4.

A typical search with a 350 residue query takes about three minutes. HMM searches are computationally intensive. The U.S. server uses a distributed processing system, written in Java, that parallelizes the load across a number of different processors on the network at Washington University. As more processors are recruited, search speed on the PFAM server will increase.

GENOME ANALYSIS USING PFAM

We are currently integrating PFAM/HMM analysis with BLAST analysis in the bioinformatics groups at the Sanger Centre (Hinxton, U.K.) and the Washington University Genome Sequencing Center. The domain-based PFAM analysis simplifies complicated BLAST outputs in some cases, especially in higher eukaryotic genomes, and HMM-profiles are sometimes more sensitive than BLAST at identifying informative similarities. PFAM/HMM analysis is also now used in 'production mode' in other academic and industrial bioinformatics groups.

PFAM/HMM analysis hits at least one domain in 24% of the predicted proteins in the *C. elegans* genome project. 8% of *C. elegans* proteins are annotated as multidomain proteins by PFAM, with up to 5 different kinds of recognized domains per protein, and up to 44 total recognized domains per protein.

One interesting analysis that PFAM simplifies is to rapidly classify the predicted proteins

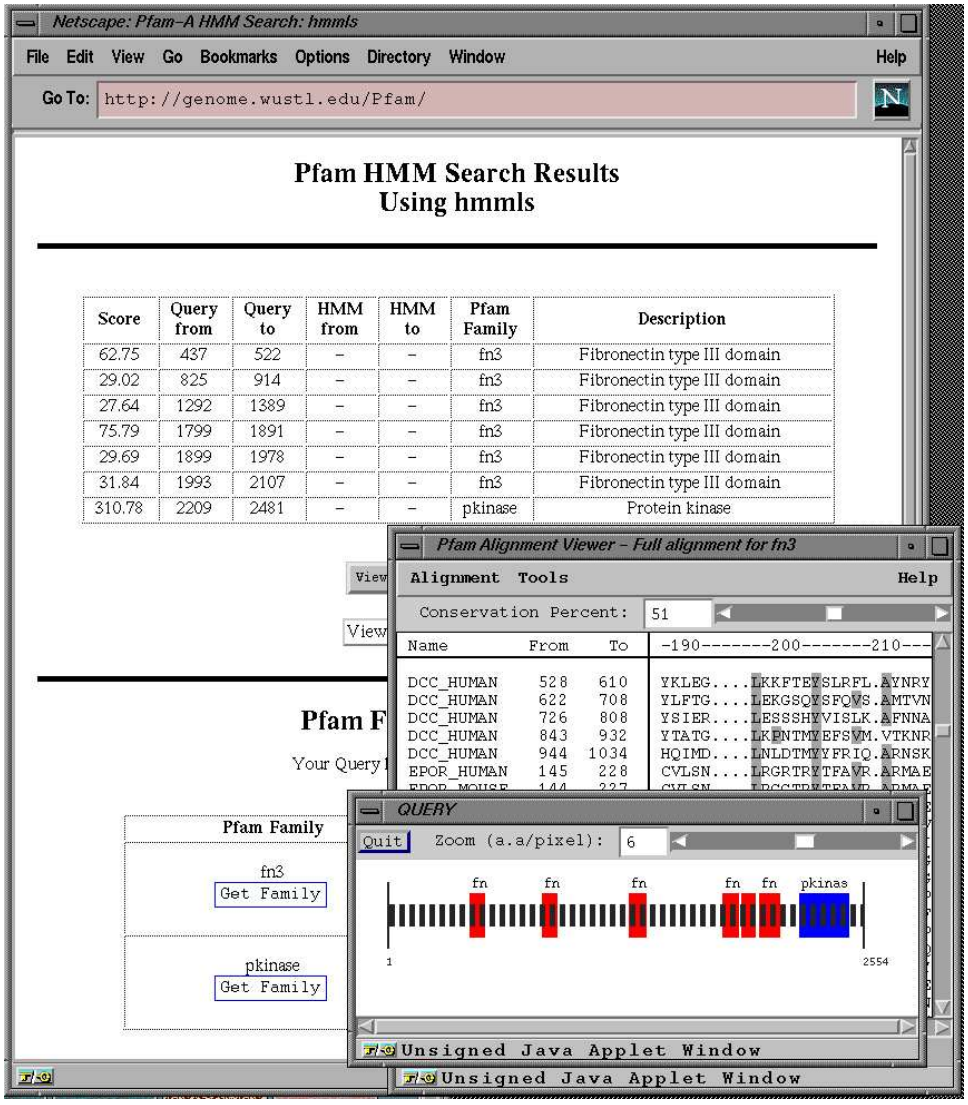

Figure 4: Screen dump of the results of submitting the *Drosophila Sevenless* protein sequence to the U.S. PFAM server. Top: tabular output of the positions and alignment scores of various fibronectin type III domains and a protein kinase domain. Middle: A Java alignment viewer applet, showing here the PFAM seed alignment for fibronectin type III domains. Bottom: A Java applet showing a cartoon of the predicted domain structure of *Sevenless*.

Table 2
Top ten protein families in *C. elegans* based on both PFAM 2.0 and BLAST analysis (protein counts based on analysis of 7299 predicted genes, 50% of the genome)

G-protein coupled receptors	179
Protein kinases	169
Collagens	97
C4-type zinc finger proteins (nuclear hormone receptors)	54
GTPase superfamily	52
Homeobox transcription factors	45
RNA recognition motif proteins	43
EGF domain containing proteins	42
short chain dehydrogenases (ADH-like)	34
ankyrin domain containing proteins	34

in a genome into families. A 'top ten' list of protein families in *C. elegans* according to PFAM analysis and some subsequent manual work is shown in Table 2.

CONCLUSIONS AND FUTURE PLANS

PFAM is now maintained by a consortium of researchers. The database is being actively developed and maintained for the use of the genome and EST bioinformatics groups that we are associated with. A number of collaborations with other databases and researchers have been initiated, as the project is too large for us to maintain by ourselves, and we welcome other contributions. To contact the PFAM consortium, email `pfam@genetics.wustl.edu` or `pfam@sanger.ac.uk`.

In addition to providing a resource for HMM-profile construction, the PFAM multiple alignment database is useful for other purposes. We are exploring its use for constructing new substitution scoring matrices, for large-scale phylogenetic studies of gene duplication and diversification, and for tuning and testing the next release of the HMMER software.

The next release of PFAM is anticipated in Fall 1997. It will be primarily a 'bugfix' release that fixes minor problems we have found in PFAM 2.0.

ACKNOWLEDGMENTS

The other members of the PFAM consortium are Erik Sonnhammer (NCBI, Bethesda, U.S.A.), Ewan Birney (Sanger Centre, Hinxton U.K.), Richard Durbin (Sanger Centre), and Alex Bateman (MRC Laboratory of Molecular Biology, Cambridge, U.K.). I thank Robert Finn for producing many of the seed alignments in PFAM 2, and Jose Aguilar for the development of the WashU PFAM server. A continuing collaboration with Graeme Mitchison (MRC-LMB, Cambridge), generously supported by NATO Collaborative Research Grant 961168, has been instrumental in the development of the theory behind HMMER. Work at WashU on HMMER and PFAM is supported by grant R01-HG01363 from the NIH National Institute for Human Genome Research, and a gift from Eli Lilly & Co., for whom I also consult.

REFERENCES

[1] T.J. Hubbard, *Curr. Opin. Struct. Biol.* **7** (1997), 190.

[2] E.V. Koonin, R.L. Tatusov, and K.E. Rudd, *Meth. Enzymol.* **26** (1996), 295.

[3] R.F. Smith, *Genome Res.* **6** (1996), 653.

[4] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, *J. Mol. Biol.* **215** (1990), 403.

[5] W.R. Pearson and D.J. Lipman, *Proc. Natl. Acad. Sci. USA* **85** (1988), 2444.

[6] R. Waterston, C. Martin, M. Craxton, C. Hunyh, A. Coulson, L. Hillier, R. Durbin, P. Green, R. Shownkeen, N. Halloran, M. Metzstein, T. Hawkins, R. Wilson, M. Berks, Z. Du, K. Thomas, J. Thierry-Mieg, and J. Sulston, *Nature Genet.* **1** (1992), 114.

[7] R. Wilson, R. Ainscough, K. Anderson, C. Baynes, M. Berks, et al., *Nature* **368** (1994), 32.

[8] R. H. Waterston, J. E. Sulston, and A. R. Coulson, in: *C. elegans II*, eds. D. L. Riddle, T. Blumenthal, B. J. Meyer, and J. R. Priess (Cold Spring Harbor Laboratory Press, 1997), p 23.

[9] E.L.L. Sonnhammer and D. Kahn, *Prot. Sci.* **3** (1994), 482.

[10] G.S. Miller and R. Fuchs, *Comput. Applic. Biosci.* **13** (1997), 81.

[11] E.L.L. Sonnhammer and R. Durbin, *Comput. Applic. Biosci.* **10** (1994), 301.

[12] J. Zhang and T.L. Madden, *Genome Res.*, **7** (1997), 649.

[13] A. Bairoch, P. Bucher, and K. Hofmann, *Nucl. Acids Res.* **25** (1997), 217.

[14] T.K. Attwood, M.E. Beck, A.J. Bleasby, K. Degtyarenko, A.D. Michie, and D.J. Parry-Smith, *Nucl. Acids Res.* **25** (1997), 212.

[15] J.G. Henikoff, S. Pietrokovski, and S. Henikoff, *Nucl. Acids Res.* **25** (1997), 222.

[16] C.H. Wu, S. Zhao, and H.L. Chen, *J. Comput. Biol.* **3** (1996), 547.

[17] P. Fabian, J. Murvai, K. Vlahovicek, H. Hegyi, and S. Pongor, *Nucl. Acids Res.* **25** (1997), 240.

[18] E. L.L. Sonnhammer, S. R. Eddy, and R. Durbin, *Proteins* **28** (1997), 405.

[19] G. J. Barton, *Meth. Enzymol.* **183** (1997), 403.

[20] M. Gribskov, R. Luthy, and D. Eisenberg, *Meth. Enzymol.* **183** (1990), 146.

[21] W. R. Taylor, *J. Mol. Biol.* **188** (1986), 233.

[22] B. Rost and C. Sander, *Annu. Rev. Biophys. Biomol. Struct.* **2** (1996), 113.

[23] S. Henikoff and J. G. Henikoff, *Proc. Natl. Acad. Sci. USA* **89** (1992), 10915.

[24] J. U. Bowie, R. Luthy, and D. Eisenberg, *Science* **253** (1991), 164.

[25] S. Henikoff, *Curr. Opin. Struct. Biol.* **6** (1996), 353.

[26] R.L. Tatusov, S.F. Altschul, and E.V. Koonin, *Proc. Natl. Acad. Sci. USA* **91** (1994), 12091.

[27] L. R. Rabiner, *Proc. IEEE* **77** (1989), 257.

[28] A. Krogh, M. Brown, I.S. Mian, K. Sjolander, and D. Haussler, *J. Mol. Biol.* **235** (1994), 1501.

[29] S. R. Eddy, *Curr. Opin. Struct. Biol.* **6** (1996), 361.

[30] K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, and D. Haussler, *Comput. Applic. Biosci.* **12** (1996), 327.

[31] T.F. Smith and M.S. Waterman, *J. Mol. Biol.* **147** (1997), 195.

[32] J.D. Thompson, D.G. Higgins, and T.J. Gibson, *Nucl. Acids Res.* **22** (1994), 4673.

[33] A. Ogiwara, I. Uchiyama, T. Takagi, and M. Kanehisa, *Protein Sci.* **5** (1996), 1991.