PublisherInfo		
PublisherName	:	BioMed Central
PublisherLocation	:	London
PublisherImprintName	:	BioMed Central

The family way

ArticleInfo		
ArticleID	:	3641
ArticleDOI	:	10.1186/gb-2000-1-2-reports2044
ArticleCitationID	:	reports2044
ArticleSequenceNumber	:	38
ArticleCategory	:	Web report
ArticleFirstPage	:	1
ArticleLastPage	:	5
ArticleHistory	:	RegistrationDate : 2000–4–28 Received : 2000–4–28 OnlineDate : 2000–6–20
ArticleCopyright	:	BioMed Central Ltd2000
ArticleGrants	:	

ArticleContext	:	130591122

Colin Semple

Abstract

Since its inception in 1996, Pfam has aimed to be a comprehensive database of protein families defined by the presence of shared domains.

Mirror site

Pfam at Washington University in St Louis, USA and Pfam at the Karolinska Institutet, Sweden

Content

Since its inception in 1996, Pfam has aimed to be a comprehensive database of protein families defined by the presence of shared domains. The database is partitioned into two sections: Pfam-A contains accurate multiple alignments that are curated manually, whereas Pfam-B is generated automatically by clustering and aligning protein sequences not already covered by Pfam-A. Pfam-A families have permanent accession numbers and contain functional annotation and cross-references to other protein domain and motif databases, whereas Pfam-B families are regenerated at each release and are less well annotated. Each domain in Pfam is represented by a position-specific scoring model or profile; the particular type of statistical model used to derive such profiles in Pfam is the hidden Markov model (HMM). Pfam consists of a library of HMMs that can be used as sensitive tools to find new members of the domains represented. It is reported that Pfam HMMs match around two thirds of proteins in SWISSPROT and TrEMBL, and that for complete genomes Pfam currently matches up to half of the proteins discovered. It is possible to browse Pfam online or to perform a search with a protein or DNA sequence of interest. Once a Pfam domain has been identified, information relating to its physical structure, typical position (for example, carboxy-terminal), functional annotation and species distribution can be retrieved.

Navigation

The site is well documented with links to related literature and software as well as other profile search sites. It is possible to bookmark individual Pfam domain pages.

Reporter's comments

Timeliness

Pfam is updated erratically every few months from the protein sequences deposited in SWISSPROT and TrEMBL.

Best feature

A single Pfam search can often be as sensitive as a detailed trawl through the databases with BLAST or FASTA, and can therefore save a lot of time in characterizing proteins with no strong similarities to known sequences. In addition, the HMM software on which Pfam is based is licensed under the Gnu Public License and is freely available. Using Sean Eddy's excellent software HMMER2 - profile hidden Markov models for biological analysis, one can create and manipulate HMMs based on protein sequence data. The Wise2 package creates other possibilities by allowing users to search genomic DNA sequence with HMMER2-generated models.

Worst feature

Those contributing to the annotation of Pfam domains may be rewarded with a Pfam T-shirt, but the Pfam-B section of the database is often many T-shirts away from useful annotation. This is of course a failing of the available literature rather than of Pfam, and will presumably be remedied by more data, for example from current structural genomics initiatives such as the US National Institute of General Medical Sciences initiative.

Wish list

Although E values (the expected probability of a match by chance alone) are given for the results of a protein search against Pfam, no analogous statistics are given for the significance of a match of a DNA sequence to Pfam HMMs. Some guidance on meaningful score thresholds is given in the Wise2 documentation (for example, scores greater than 35 are usually reliable evidence for a domain whereas

scores less than 15 are untrustworthy) but it would be nice to have a more rigorous assessment of significance. Unfortunately, the software has raced ahead of the statistics here and various problems have yet to be addressed. Hopefully the development of Wise2 will continue.

Related websites

There are various other databases that seek to describe conserved regions of proteins with profiles derived from multiple sequence alignments. The PRODOM database (which forms the basis of Pfam-B) is automatically constructed using recursive searches with the NCBI PSI-BLAST program. PROSITE contains profiles derived from manually edited alignments and probably leads the field in terms of quality of documentation for domains. The PRINTS and BLOCKS databases contain profiles representing smaller, subdomain sections and motifs in proteins. The current received wisdom is to search several profile databases with a protein of interest, as their contents do not completely overlap. The InterPro database is a recent attempt to integrate profile-based databases and combines data from Pfam, PROSITE, PRINTS and PRODOM. InterPro has been used by the European Bioinformatics Institute in the analysis of several proteomes, and has already established itself as an important sequence-annotation tool. Finally, the [Structural genomics initiative] [http://www.nigms.nih.gov/funding/psi.html] of the National Institute of General Medical Sciences funds a program of work in structural genomics.

Table of links

Pfam: a database of multiple alignments of protein domains and conserved regions

Pfam at Washington University in St Louis, USA

Pfam at the Karolinska Institutet, Sweden

HMMER2 - profile hidden Markov models for biological analysis

Wise2

PRODOM

PSI-BLAST

PROSITE

PRINTS

BLOCKS

InterPro

References

1. Pfam: a database of multiple alignments of protein domains and conserved regions

.