Nucleic Acids Research, 2006, Vol. 34, Database issue ***D247–D251***
doi:10.1093/nar/gkj149

Pfam: clans, web tools and services

Robert D. Finn*, Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich[1], Timo Lassmann[1], Simon Moxon, Mhairi Marshall, Ajay Khanna[2], Richard Durbin, Sean R. Eddy[2], Erik L. L. Sonnhammer[1] and Alex Bateman

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK, [1]Center for Genomics and Bioinformatics, Karolinska Institutet, S-171 77 Stockholm, Sweden and [2]Department of Genetics, Howard Hughes Medical Institute, Washington University School of Medicine, St Louis, MO 63110, USA

Received September 15, 2005; Revised October 19, 2005; Accepted October 28, 2005

ABSTRACT

Pfam is a database of protein families that currently contains 7973 entries (release 18.0). A recent development in Pfam has enabled the grouping of related families into clans. Pfam clans are described in detail, together with the new associated web pages. Improvements to the range of Pfam web tools and the first set of Pfam web services that allow programmatic access to the database and associated tools are also presented. Pfam is available on the web in the UK (http://www.sanger.ac.uk/Software/Pfam/), the USA (http://pfam.wustl.edu/), France (http://pfam.jouy.inra.fr/) and Sweden (http://pfam.cgb.ki.se/).

INTRODUCTION

Pfam is a comprehensive database of protein families, containing 7973 families in the current release (18.0). Each family is manually curated and is represented by two multiple sequence alignments, two profile-Hidden Markov Models (profile-HMMs) and an annotation file. All data are available for download in flatfile format from the FTP sites linked from each Pfam website and also as a set of MySQL relational database files. Pfam families are periodically updated with each family having on average been modified four times since its creation. The data and additional features are accessible via the four websites (http://www.sanger.ac.uk/Software/Pfam/, http://pfam.wustl.edu, http://pfam.jouy.inra.fr/ and http://Pfam.cgb.ki.se/). The structure and use of Pfam are well established and are documented elsewhere (1,2).

Several new features have been added to Pfam in the past 2 years. The main focus of this paper will be to describe a change in Pfam philosophy that has allowed us to group protein families into a hierarchical classification of clans. In the latter sections, we will describe new web tools and Pfam web services. An additional feature, *i*Pfam (a sister database containing details of Pfam domain–domain interactions) has been described in a recent publication (3).

THE GROWTH OF PFAM

Pfam has increased by 1783 families since Pfam release 10.0 (1). Despite the near doubling of sequences in the underlying sequence database over the past 2 years, the fraction of sequences in UniProt (4) that match a Pfam family remains at 75%. One of the main uses of Pfam is genome annotation, thus an important measure is the coverage of the non-redundant set of proteins encoded by a genome, called proteome coverage. Table 1 shows the increase in Pfam coverage of a selected set of proteomes since Pfam began 9 years ago. The proteomes analysed were the bacteria *Escherichia coli* and *Rickettsia prowazekii*, the archaeon *Methanococcus jannaschii*, the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans* and *Homo sapiens*. Release 5.5 contained 2478 families/models, with an average protein coverage (the fraction of proteins with at least one hit to Pfam) of 53% and an average residue coverage (the fraction of residues matched to a Pfam family) of 34%. Despite a large increase in the number of models (an additional 3712 models) between releases 5.5 and 10.0, there was an average increase in protein coverage of 14% and residue coverage of 12%. A less than proportional increase in coverage is observed for the 1783 models added between Pfam 10.0 and 18.0, such that release 18.0 matches an average of 71 and 48% of sequences and residues, respectively. This illustrates a law of diminishing returns for adding new families.

Nevertheless, over the past 2 years there has been a steady increase in both measures of proteome coverage. Pfam now matches 60–84% of proteins in each proteome compared with 47–62% 5 years ago and 57–76%, just 2 years ago.

*To whom correspondence should be addressed. Tel: +44 1223 495330; Fax: +44 1223 494919; Email: rdf@sanger.ac.uk

Table 1. Increase in coverage of 6 representative proteomes over the past 9 years of Pfam

	Release/Date	No. Models	*E.coli K12*	*R.prowazekii*	*M.jannaschii*	*S.cerevisiae*	*C.elegans*	*H.sapiens*
Protein coverage (%)	18.0 (07/2005)	7973	84	81	72	64	60	64
	10.0 (07/2003)	6190	76	77	68	60	57	60
	5.5 (09/2000)	2478	55	62	52	47	47	52
Residue coverage (%)	18.0 (07/2005)	7973	65	61	55	36	35	37
	10.0 (07/2003)	6190	61	58	53	35	34	35
	5.5 (09/2000)	2478	42	44	40	25	26	29

The models from releases 5.5, 10.0 and 18.0 were searched against each proteome, downloaded from Integr8 (http://www.ebi.ac.uk/integr8) (16). Every protein domain satisfying the curated Pfam gathering threshold cut-off was scored as a hit. Two different coverage measures have been included, protein coverage and residue coverage.

PFAM CLANS

One of the fundamental philosophies of Pfam is that new protein families are not allowed to overlap with existing Pfam entries (2). Thus, any residue in a given sequence can only appear in one Pfam family. Building new Pfam families and/or revisiting existing families often highlights two important points. (i) Many Pfam families are related and may have artificially high thresholds to stop them from overlapping. (ii) For some large, divergent families we cannot build a single HMM that detects all examples of the family. To resolve these issues, we have introduced Pfam clans.

What are Pfam clans?

A clan contains two or more Pfam families that have arisen from a single evolutionary origin. We use up to four independent pieces of evidence to help assess whether families are related: related structure, related function, significant matching of the same sequence to HMMs from different families and profile–profile comparisons. To perform profile–profile comparisons we use PRC 1.5.2 (downloadable from http://supfam.mrc-lmb.cam.ac.uk/PRC/). Currently, the presence of related structures and significant profile–profile comparison scores are our primary indicators of a relationship between families. From an analysis comparing Pfam families with a known structure, we deem a significant profile–profile comparison score as one with an *E*-value of <0.001. Profile–profile comparison *E*-values in the range of 0.1–0.001 can indicate a true relationship, but we require additional evidence to include the family in the clan.

After identifying a set of related families, our first aim is to try and merge them to make a single, comprehensive model that detects all of the proteins detected by the individual models. If this cannot be achieved we create a clan, with the maximum coverage using the minimum number of models. However, as mentioned previously, having a set of related families has historically led to artificially high thresholds to prevent the families from overlapping. To remedy this situation, thresholds are redefined, to include the maximum number of significant matches, excluding all false positives. This can cause overlaps between the members of a clan. To maintain the 'no overlap' rule in Pfam, only the best scoring match is reported and presented in the set of full alignments. For example, the sequence Q5Z855 matches the ENTH domain (PF01417) with a score of 16.5 bits and the ANTH domain (PF07651) with a score of 327.5 bits, but the match only appears in the alignment of ANTH. We have updated the software that allows searching of Pfam models locally (pfam_scan.pl) so that it resolves overlaps between clan members in a similar fashion. The clans are annotated in an analogous way to Pfam entries, including a stable accession of the form CL0001, short identifier, one line description and a summary of the clan. Where appropriate, cross-references to other databases are included (Figure 1A).

As of Pfam 18.0, there were 172 clans, containing 1181 Pfam families. This represents 15% of Pfam families and as these tend to represent the larger protein families, account for 31% of the domain hits in Pfam. Clans help us to improve the annotation of families. For example, knowing the 3D structure of a domain is an essential part of understanding the biology of that domain. Pfam clans are helping to identify, previously undetected, structural homologues. Currently, 66% of all families in clans contain at least one sequence with a known 3D structure. A further 418 families (30%) where a structural homologue is not found in the family are in a clan where at least one family contains a known 3D structure. In addition to relating families with unknown structure to those with a known structure, we can also use Pfam clans to improve annotation. Currently, there are 81 domains of unknown function (DUFs/UPFs) in clans. We can assign a putative function to 78 of these DUFs, based on similarities of these DUFs to characterized families in a clan.

Pfam clans provide a hierarchical view of a diverse range of proteins families. How do Pfam clans relate to other classifications of protein families? There are many databases providing a hierarchical view of protein sequence space, using a variety of techniques, e.g. SCOP (5), CATH (6), SUPFAM (7), Protomap (8) and Superfamily (9). Below, we consider two of the more closely related databases.

SCOP is a classification of proteins of known structure (5). Many of the Pfam clans have a similar family membership to SCOP superfamilies, as both classification systems use structures to relate families. However, there is not a one-to-one correlation between Pfam clans and SCOP superfamilies. Profile–profile comparisons can detect significant similarities between families occurring in different SCOP superfamilies. For example, the TIM Barrel (CL0036) and NADP Rossmann (CL0063) clans cover multiple, related SCOP superfamilies (8 and 4, respectively). There are some Pfam clans, though, that are not as comprehensive as SCOP superfamilies, often owing to a lack of protein sequence coverage preventing the generation of effective seed alignments. The primary difference though, is that the Pfam classification is not confined to those families with a known 3D structure. Indeed, some Pfam clans contain groups of related families where none of the members have a determined 3D structure. For example,

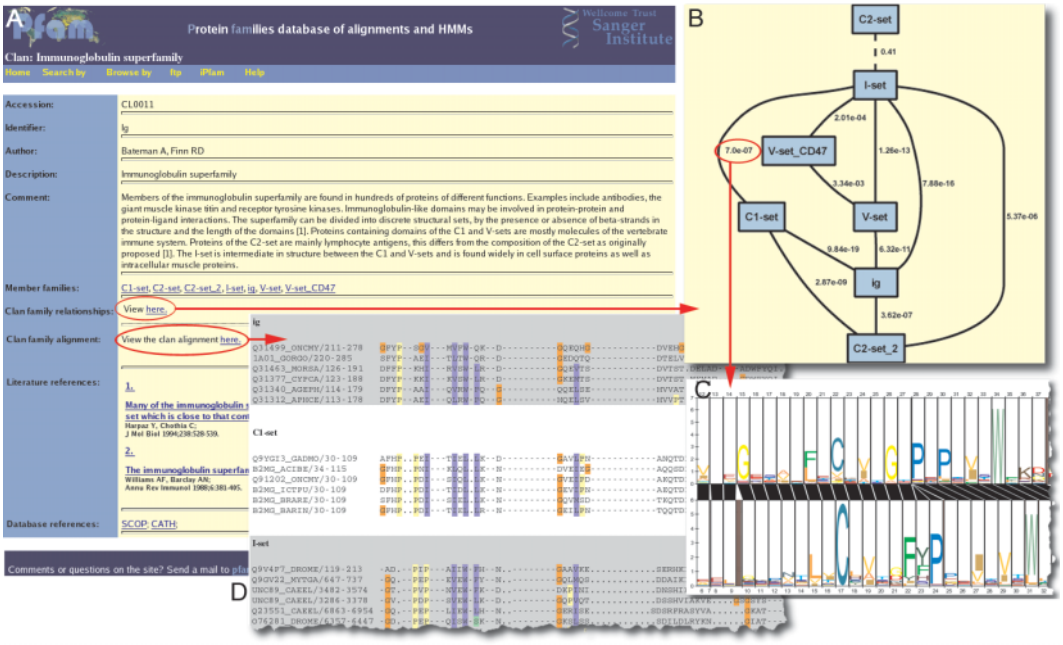

Figure 1. Clan pages in Pfam. (**A**) A screen shot of a clan summary page, containing the description, annotation and membership of the clan. From this page, the user can view the family relationship diagram (**B**). Each family in the clan is represented by a blue box and its relationship to other families is represented by solid lines (significant profile–profile comparison score) or dashed lines (non-significant profile-profile comparison score). Beside each line, the profile–profile comparison *E*-value score is presented. This score is also linked to a visualization of the profile–profile comparison alignment (**C**). The clan summary page also provides a link to the clan alignment (**D**) (for more details see text). The clan alignment is a multiple sequence alignment of all of the clan members seed alignments (each set of seed sequences are separated by the alternate background shading). The alignments are coloured using Jalview.

the Major Facilitator Superfamily is a clan of 19 Pfam families and should be a high priority for structural genomics.

The SUPFAM database (7) classifies Pfam families into superfamilies based on SCOP and RPS-BLAST profile comparisons. A brief comparison of SUPFAM superfamilies to Pfam clans is of interest. At the time of writing, SUPFAM was based on Pfam version 14.0, making this comparison less straightforward. The automated approach used by SUPFAM means that many more Pfam families have been classified into SUPFAM superfamilies. Many of these additional superfamilies contain a single Pfam family. In such cases a Pfam clan would not be created as Pfam clans are only created when there are two or more related Pfam families. Generally, corresponding clans/superfamilies have a similar membership. Where SUPFAM use SCOP for the classification of families, the differences described above for SCOP are paralleled. In addition, where there are differences in domain definitions between SCOP and Pfam, there has been some misclassification of Pfam domains. Interestingly, Pfam clans with no known structures tend to have a larger membership than the corresponding SUPFAM superfamilies, even accounting for the 524 families added since release 14.0.

ACCESSING CLAN DATA

There are three different ways of accessing the clan information. First, there is an additional release flatfile, Pfam-C, which contains all of the clan information and a list of the Pfam families that are members of the clan. Second, all of the information is contained in the Pfam MySQL database that we make available for download. Third, clan information can be accessed via the websites. There are two web entry points to the clan information. A user can 'browse by' a list of clans or follow links from clan member families (Table 2). For each clan, we display annotation and a list of Pfam families that constitute the clan (Figure 1A). In addition, there are links to two additional features; a clan relationship diagram and a clan alignment.

The clan relationship diagrams show how the individual families are related to each other (Figure 1B). To produce these diagrams, we perform an all-against-all profile–profile comparison between the clan members. In the relationship diagram Pfam families are graph nodes. Edges are added between nodes when a significant profile–profile score is observed between two nodes (represented by solid lines). After all edges have been added in this way, any nodes/domains that have no connecting edges are identified. Where possible, these detached nodes are connected by adding an edge between it and the node in the clan with the highest scoring profile–profile score that falls above the 0.001 threshold (i.e. *E*-values 0.001–10). A dashed line represents these edges in the final graph. Domains that have been brought into the clan based on a structural similarity may remain detached, indicating that profile–profile comparisons are not able to detect all distant relationships. The *E*-values used to construct the edges are displayed. These *E*-values are also clickable links to a visualization of the profile–profile alignment (10) (Figure 1C).

The clan alignment is an alignment of all the clan seed alignments (Figure 1D). These are produced by an option in MUSCLE (11) that aligns two input multiple sequence

Table 2. Summary of the new website features and web services, including server location

Feature	Mirror site	Specific URL
Clan summaries	UK, Sweden	Follow links from: http://www.sanger.ac.uk/Software/Pfam/browse/clans.shtml, http://pfam.cgb.ki.se/browse/clans.html
Clan alignments/relationship diagrams	UK	Example URLs: http://www.sanger.ac.uk//Software/Pfam/data/clans/alignment/CL0132.shtml, http://www.sanger.ac.uk/cgi-bin/Pfam/clanacc?CL0132
Coloured alignments	Sweden	Example: http://pfam.cgb.ki.se/cgi-bin/getalignment.pl?name=TAF&acc=PF02969&type=seed&format=msaviewer&size=10&color=0
Domain images/XML upload	UK	http://www.sanger.ac.uk/cgi-bin/Pfam/xml_upload.pl
HMM logo	UK	http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi?pfamid=PF02969
Domain query tools	Sweden	http://pfam.cgb.ki.se/cgi-bin/domainquery/DQL_sel_domains.pl
Core web services	UK	http://services.sanger.ac.uk/pfamWebService/services/pfamWebService?wsdl
Web service Perl client	UK	ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/pfamWSclient.pl
DQL web service	Sweden	http://www.cgb.ki.se:8080/pfam/WSFacadeServicePort?wsdl
PfamAlyzer	Sweden	http://pfam/cgb.ki.se/pfamalyzer

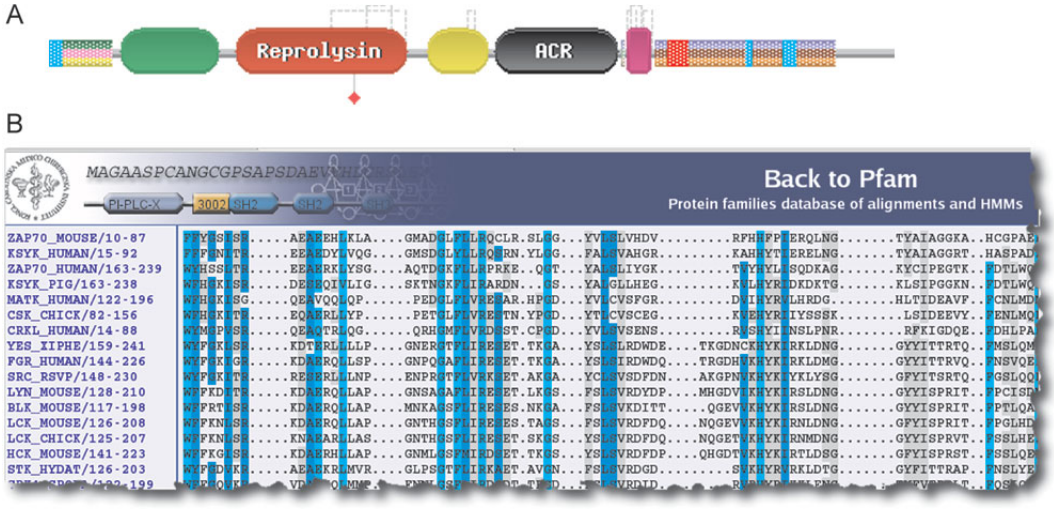

Figure 2. (**A**) Graphical representation of domains on the sequence ADA19_HUMAN. The sequence is represented as a grey bar. As of release 18.0, Pfam identifies four domains: Pep_M12B_propep (PF01562, coloured green), Reprolysin (PF01421, red), Disintegrin (PF00200, yellow) and EGF_2 (PF07974, magenta). The black domain is the ACR domain from SMART (15). The striped boxes represent PfamB families, while the small blue and red boxes represent low-complexity and transmembrane regions respectively. Above the domain images, the dashed lines represent disulphide bridges found within the sequence. The red diamond below the Reprolysin domain indicates an active site position. (**B**) The seed alignment of SH2 (PF00017) marked-up according to the Belvu colouring system, using the new multiple sequence alignment viewer on the Swedish site.

alignments without altering their local alignments. Where more than two seed alignments are being aligned, we use the profile–profile comparison scores to guide the progressive alignment procedure so that the most similar seed alignments are aligned first, before more divergent alignments. These alignments are pushing the boundaries of what is feasible to align by sequence alone, so the alignments must be treated with some caution.

NEW WEBSITE FEATURES

All of the Pfam mirror sites use the same underlying data and provide the same basic features. New tools and features based on the common dataset are being developed independently at the different sites. The new features that are available from the different mirror sites are described in Table 2.

Domain images

Influenced by SMART (12) and PROSITE (13), the graphical representation of domains has been updated on the website (Table 2). In addition to each domain having more of a 3D look, additional sequence features are now visualized. For example, we now include disulphide bonds and active sites (Figure 2A). The disulphide bond and active site data is derived from UniProt annotation (4). Pfam is often approached about the use of domain graphics in publications. To make our domain graphics more accessible and flexible, we have developed an interface so that users can customize a graphical view of a sequence. The user controls the style of a domain image using a simple XML file, enabling user-defined domains and sequence features to be added to the view (see XML schema at http://www.sanger.ac.uk/Software/Pfam/xml/pfamDomainGraphics.xsd). After uploading the XML file, the image is rendered and can be downloaded from the resulting page.

HMM logos

HMM logos are graphical representations of an HMM, which allow the visualization of its distinguishing features (14). HMM logos are provided for every Pfam family via the

LogoMat-M tool (Table 2). As a variation of classical sequence logos, LogoMat-M uses relative entropy to identify residues that are of particular interest. The HMM logos are related to the profile–profile alignment logos shown in Figure 1C.

Coloured alignments

A new HTML/javascript multiple alignment view has been added to the website (Table 2). The alignments are shown in their natural linear format to avoid splitting up conserved blocks, which may happen in wrap-around formats (Figure 2B). In the new view, the sequence name column stays fixed when scrolling horizontally. Residues are coloured using the same methods as in Belvu (http://www.cgb.ki.se/cgb/groups/sonnhammer/Belvu.html), either according to conservation based on the average BLOSUM62 score or by residue type.

Domain query tools

The domain query tools have undergone significant reconstruction, making them more powerful, flexible and user friendly. Domain query functionality is now offered as a web interface, in form of the PfamAlyzer Java applet and as a web service for automated searches (Table 2). The web interface allows the user to select a set of Pfam domains and arrange them in order with the possibility to define the gap size in between. The domain query can be asked to widen the results, where appropriate, by exchanging a specified domain for all domain(s) within the clan.

PfamAlyzer is an applet that combines and extends many functions from the Pfam sites and integrates them into one tool. The domain query of PfamAlyzer is more user friendly and more powerful than the web interface. A graphical query language using drag and drop formulates the query. PfamAlyzer adds taxonomic analysis functionality to the domain query. Queries can be limited to specific taxonomic groups such as *Chordata*, which is especially helpful when studying architectures with a large number of members. PfamAlyzer can also display the query results as a species distribution that shows the resulting proteins as leaves on the species tree.

ACCESSING PFAM USING WEB SERVICES

We have implemented recently a range of web services that allow machine interoperable access to Pfam. Currently, the web services cover the following basic operations: annotation of a UniProt sequence based on an accession or identifier and access to Pfam family annotation. In the coming months, we plan to add services that allow the automatic download of alignments and searching of sequences. A basic Perl client is available for accessing these web services (Table 2).

The Pfam domain query described in the new website features section has also been implemented as a web service. An example to run within the JBossIDE (http://www.jboss.org/products/jbosside) is available at http://pfam.cgb.ki.se/pfamalyzer/example.zip.

ACKNOWLEDGEMENTS

We would like to thank all users who have contributed new families and annotation to Pfam. In addition, we would like to thank Matthew Fenech, Song Choon Lee, Rafaella Rossi, Arthur Wuster and Corin Yeates who have added many new families and clans to Pfam over the past 2 years. We would also like to thank Lorenzo Cerutti for maintaining the French Pfam website. This work was funded by The Wellcome Trust and an MRC (UK) E-science grant (G0100305). Funding to pay the Open Access publication charges for this article was provided by The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
2. Sonnhammer,E.L.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.
3. Finn,R.D., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
4. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
5. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
6. Pearl,F., Todd,A., Sillitoe,I., Dibley,M., Redfern,O., Lewis,T., Bennett,C., Marsden,R., Grant,A., Lee,D. *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **33**, D247–D251.
7. Pandit,S.B., Bhadra,R., Gowri,V.S., Balaji,S., Anand,B. and Srinivasan,N. (2004) SUPFAM: a database of sequence superfamilies of protein domains. *BMC Bioinformatics*, **5**, 28.
8. Yona,G., Linial,N. and Linial,M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **28**, 49–55.
9. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235–D239.
10. Schuster-Bockler,B. and Bateman,A. (2005) Visualizing profile–profile alignment: pairwise HMM logos. *Bioinformatics*, **21**, 2912–2913.
11. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
12. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
13. Hulo,N., Sigrist,C.J., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
14. Schuster-Bockler,B., Schultz,J. and Rahmann,S. (2004) HMM Logos for visualization of protein families. *BMC Bioinformatics*, **5**, 7.
15. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
16. Pruess,M., Kersey,P. and Apweiler,R. (2005) The Integr8 project—a resource for genomic and proteomic data. *In Silico Biol.*, **5**, 179–185.