

Analysis of PRINTS annotation and BioMinT requirements.

A) The PRINTS annotation process

The PRINTS database is a compendium of protein "fingerprints", which are groups of conserved motifs drawn from sequence alignments that are used to characterise protein families. PRINTS provides large amounts of hand-crafted annotation aiming to document the constituent protein families and rationalise the conserved regions in structural and functional terms. The process of annotating a fingerprint is largely dependent on the component Swiss-Prot and TrEMBL (SPTr) sequences; the amount and type of annotation to be added will differ according to the relationship of the underlying sequences, how much is known of their biological function, and so on. However, a broad overview of the PRINTS annotation process is illustrated in Figure 1. A more in depth discussion of the individual fingerprint fields and their annotation is provided in Section B.

As illustrated in Figure 1, the first task in the process of annotation is identification of the fingerprint type. For human annotators this is usually a trivial process performed by examining the respective SPTr entries of the individual sequences that constitute a fingerprint to see how they relate. Broadly speaking, the sequences under investigation may constitute a gene family or super-family (united by a common function), or a domain family (united by a common structural motif). For examples of different fingerprint types see Table 1. The fingerprint type has important implications on what is reported within the annotation and which resources are therefore used to gather information (see below).

The fingerprint is given a unique identifier, accession number, and title that describes the relationship of the component sequences (*e.g.*, 5-hydroxytryptamine 2A receptor signature, Rhodopsin-like GPCR super-family signature, SH3 domain signature, *etc.*). Database cross-references are sought by searching a common core of family, structure and disease databases with representative sequences from the fingerprint. The bulk of annotation is then researched by reading relevant literature retrieved from a range of sources such as PubMed, specialised Websites, and online databases (see Figure 1). This tends to be an iterative process, whereby annotation is refined and further literature is sought as information is assimilated. The information is then added to the fingerprint as a block of free text representing the core of the bio-medical annotation.

As mentioned above, the information reported in the main annotation block depends largely on the fingerprint type. In the case of a super-family level fingerprint, the aim is to provide a broad overview of the structure, high-level function, and diseases associated with the super-family (see Figure 2 – blue text). For a family-level fingerprint, annotation from the PRINTS parent (*i.e.*, the super-family to which the family belongs) is usually inherited (see Figure 1). If no such parent exists, annotation is created as if processing a super-family. Family-specific information is then generated. This focuses on the individual family members, their specific function, expression pattern and relationship to other protein families (see Figure 2 – green text). Depending on the amount and relevance of information available, further annotation may be constructed pertaining to details such as the subcellular location of the proteins, their regulation, the history of their discovery, and so on. In the case of domain families, the annotation within the main block relates neither to individual

family members, nor to the families to which they belong, but to the domain itself (for an example of domain family fingerprint annotation, see Figure 3). Information of interest includes the structure and function of the domain (if known), and a representative set of protein families in which the domain is found.

The main resource used in fingerprint annotation is published literature. Differing additional resources are also often consulted, depending on the type of fingerprint undergoing annotation (see Figure 1). For example, in the case of a protein family such as the 5-HT_{2A} receptors, information found in Swiss-Prot relating to the structure and function of individual family members is highly relevant (each family member is a G protein-coupled receptor (GPCR), activation of which stimulates the phosphatidylinositol-calcium second messenger system). In the case of a protein super-family, such as the GPCRs as a whole, functional information from individual Swiss-Prot entries is likely to be too specific (GPCRs couple to a variety of second messenger systems, so the function of individual family members is unlikely to be representative). Here, resources such as review papers, or protein family ('pattern') databases such as PROSITE, are more likely to provide useful overviews of super-family function, and are more convenient to process than digesting and summarising Swiss-Prot functional information for an entire super-family. However, some elements of Swiss-Prot information are still likely to be relevant to super-family annotation. For example, in the case outlined above, since each family member is a GPCR, information on the structure of one these proteins may well pertain to the entire super-family.

With domain-family annotation, resources that provide information at the level of individual sequences are even less likely to be relevant, since the aim here is to provide information that is specific to the structure and function of the domain itself and not of the protein that contains it. Consequently, the type of published literature sought will also differ from that used in the annotation of family or super-family fingerprints; papers describing the function of individual proteins or super-families will be largely irrelevant compared to those specifically concerned with the domain.

In summary, the main annotation block contains annotation tailored to whether a family, super-family or domain is being processed and may be derived from different resources accordingly. For more discussion of information sources that may be used in the annotation of fingerprints, see section C.

Once the appropriate report has been constructed, literature references cited in the annotation block are then added to the fingerprint (see Figures 2 and 3). Finally, a paragraph is constructed, using a standard template, that details technical information about the fingerprint – the number of motifs, their location, and so on (see Figure 2 – red text). The fingerprint then undergoes a final review step to check formatting and spelling before submission to the database.

B) Automatic annotation of individual database fields: current and future strategies

'Naked' (unannotated) fingerprints are generated as a series of individual flat files, termed feature format (ffmt) files. These are composed of a series of two-character tags that indicate the type of data found on each line. A typical 'naked' ffmt file is

illustrated in Figure 4. The fields to be filled automatically are: 'gc;' general code or identifier; 'gt;' general title or signature name; 'gp;' general pointers to other resources, *i.e.* database cross-links; 'gr;' general literature references; and 'gd;' general description or annotation. An expanded description of each of these fields, the current automated annotation methods, and the expected difficulty in filling these fields is described below. All other field relate to technical aspects of the fingerprint and are pre-filled, or require some element of human interaction.

gc; identifier

This field holds the fingerprint identifier, which is a single text string with a maximum length of 12 characters. Ideally, the identifier is 'human readable' and encapsulates the relationship of the sequences within the fingerprint. Examples include: OCTAMER, TEADOMAIN, LUCIFERASE, TRPCHANNEL1, EXPANSNFAMILY, NUDIXFAMILY, *etc.* (for a complete list of identifiers see - http://www.bioinf.man.ac.uk/dbbrowser/sprint/printss_lis.html).

Annotation of this field is difficult to replicate automatically due to a lack of resources that hold concise gene family and super-family abbreviations. The PRECIS annotation system attempts to fill this field through processing the first elements of Swiss-Prot identifiers. For example, the 'naked' janus kinase 1 family fingerprint illustrated in Figure 4 contains the Swiss-Prot identifiers JAK1_HUMAN, JAK1_MOUSE and JAK1_CYPICA. The software would therefore return "JAK1" as an identifier, which in this case is appropriate. However, in the case of super-family or domain-family fingerprints, which are characterised by variable first element IDs, (*e.g.*, MUP_RAT, LACB_BOVIN, RETB_HUMAN, *etc.*, all belong to the lipocalin super-family) this approach is less than ideal (here PRECIS would return MUPLACBRETB as a lipocalin super-family fingerprint identifier). Intelligent application of word stemming tools to information gathered from wider sources may help resolve this issue.

gt; signature name

This field holds the title of the fingerprint as a single line of text, a maximum of 80 characters in length, terminating with the suffix 'signature '. Examples include: 6-phosphogluconate dehydrogenase signature; H⁺-transporting ATPase (proton pump) signature; Bombesin receptor signature; MAM domain signature; D-Ala-D-Ala carboxypeptidase 3 (S13) family signature; *etc.* (see http://www.bioinf.man.ac.uk/dbbrowser/sprint/printss_lis.html for a complete list of signature names).

The difficulty of generating this field automatically depends on the constancy of the biological nomenclature, and the complexity of the relationship of the sequences under investigation. For example, the fingerprint for the janus kinase 1 family of proteins (Figure 4) contains three sequences from Swiss-Prot, from human mouse and carp. Examining the Swiss-Prot entries for these sequences we find the following description (DE) lines:

```
Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1) (JAK-1) - Homo sapiens (Human).
Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1) (JAK-1) - Mus musculus (Mouse).
Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1) (JAK-1) - Cyprinus carpio (Common carp).
```

Here, it is quite straightforward to use a statistical approach, simply inheriting the most common string of text as the title ("Tyrosine-protein kinase JAK1 (EC

2.7.1.112) (Janus kinase 1) (JAK-1)”). This method is used by the current automated annotation system (see Figure 5 for PRECIS annotation of this fingerprint). However, if we examine a fingerprint for the entire janus kinase family, we find it contains 11 Swiss-Prot entries, which have the following DE lines:

```
Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1) (JAK-1) - Homo sapiens (Human).
Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1) (JAK-1) - Mus musculus (Mouse).
Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1) (JAK-1) - Cyprinus carpio (Common
carp).
Non-receptor tyrosine-protein kinase TYK2 (EC 2.7.1.112) - Homo sapiens (Human).
Non-receptor tyrosine-protein kinase TYK2 (EC 2.7.1.112) - Mus musculus (Mouse).
Tyrosine-protein kinase JAK2 (EC 2.7.1.112) (Janus kinase 2) (JAK-2) - Homo sapiens (Human).
Tyrosine-protein kinase JAK2 (EC 2.7.1.112) (Janus kinase 2) (JAK-2) - Rattus norvegicus (Rat).
Tyrosine-protein kinase JAK2 (EC 2.7.1.112) (Janus kinase 2) (JAK-2) - Mus musculus (Mouse).
Tyrosine-protein kinase JAK3 (EC 2.7.1.112) (Janus kinase 3) (JAK-3) - Mus musculus (Mouse).
Tyrosine-protein kinase JAK3 (EC 2.7.1.112) (Janus kinase 3) (JAK-3) - Rattus norvegicus (Rat).
Tyrosine-protein kinase JAK3 (EC 2.7.1.112) (Janus kinase 3) (JAK-3) (Leukocyte janus kinase) (L-
JAK) - Homo sapiens (Human).
```

In this case, the most common string is “tyrosine-protein kinase”. Whilst this is an accurate description of the super-family to which the janus kinases belong, it is not a specific title for the janus kinase family itself. Ideally, an automated annotation system would examine the relationship of the TYK2 proteins to the janus kinases. In this case, analysis of the literature abstracts referenced in Swiss-Prot would reveal the TYK2 proteins are in fact janus kinase family members, although their nomenclature does not follow that of other family members (*e.g.*, “The janus kinases (JAK) JAK1, JAK2, and TYK2 are protein tyrosine kinases, which play a pivotal role in the signal transduction process mediated by cytokines...” <PubMed=7518579>). However, such information may not always be available from the abstracts cited within Swiss-Prot, necessitating information retrieval and extraction from wider sources, such as recent review articles.

For super-families and domain families, it is unlikely that the Swiss-Prot DE lines will contain common annotation that may serve as a fingerprint title. In these cases, the current automated annotation system examines the Swiss-Prot CC Similarity sub-field lines, since these often contain strings relating to the super-family or domain (*e.g.*, “belongs to family 1 of G protein-coupled receptors”, or “contains 2 SH2 domains”). However, the same considerations for naming family-level fingerprints also apply.

gp; database cross-references

This field provides links to a common core of databases (specifically, PRINTS, PROSITE, Pfam, InterPro, PDB, SCOP, CATH, HSSP and MIM) that may provide additional structural, functional, family and disease-related information. Providing some level of automatic completion of these fields is relatively trivial, since the database hits of individual sequences are usually stored in Swiss-Prot and this information can be directly inherited. However, there is currently a distinction between the database cross-reference information in PRINTS, where only equivalent database links (*i.e.*, representing the same family) are reported, and annotation created automatically with PRECIS, where any database cross-references found in Swiss-Prot (which also belong to the common core of databases of interest) are reported (compare Figures 2 and 5, which represent hand- and PRECIS-generated annotation of the same fingerprint). The rationale behind this is that, in the case of automatic annotation, we aim to provide the end user with as many sources of information as possible and it is relatively easy for a PRINTS annotator to follow up and remove any non-equivalent links. Ideally, an automated annotation system would include an

option to report all or equivalent cross-links. The latter option could be introduced by having the software perform individual database searches with representative sequences and digest the results.

gr; literature references

This field contains a set of literature references that are cited within the main annotation block. The number is usually restricted, with 4-6 references typically sufficient to cover the salient annotation of a protein family. The references are formatted in a PRINTS-specific manner (see Figures 2 and 3). This process is currently replicated automatically by analysing literature references found within the Swiss-Prot entries that constitute a fingerprint. The analysis selects up to four most recent shared articles, but if common references are not found, then the most recent non-shared publications are included. In addition, papers containing details of structure determinations over-ride the requirement for the reference to be shared. Swiss-Prot RP lines are used to ascertain whether a structure is available. If these lines contain strings such as, “X-ray crystallography” or “structure by NMR”, the corresponding reference is included in the report.

Automatic annotation of this field is straightforward, provided literature references are reported in Swiss-Prot and that they are relevant. However, automatic completion of this field with the most relevant, up-to-date and information-rich references will probably represent more of a challenge, as these may not be available within Swiss-Prot, because the annotators cannot keep pace with the entire biomedical literature!

gd; annotation

The *gd*; annotation field represents the core of the bio-medical information. The annotation is hand-written and reflects the current literature for a given family. Typically, the information source is published literature, although in some cases this is augmented with additional information from a variety of sources, such as text books, protein family or disease databases and specialised Websites. The content of the *gd*; field varies in a gene family or domain family-specific manner (see section A), but in all cases comprises a section dedicated to the background biology of the family or domain; it also includes a final technical paragraph detailing the location of motifs, the number of sequences comprising the fingerprint, the number of iterations required to reach convergence and the sequence database version used to build the fingerprint.

Fingerprints for members of the same gene family will usually include some common annotation to reflect their shared ancestry, or homology. This shared annotation typically pertains to the super-family to which the proteins belong, discussing high level function, structural information and diseases associated with the super-family. Following this, family-specific information is added describing, where appropriate, tissue expression, history, cytological location, species distribution, *etc.*. Domain family annotation, on the other hand, summarises details such as the structure and function of the domain itself, the range of proteins in which it is found, other domains with which it is commonly associated, its physicochemical properties, and so on.

The current approach adopted by PRECIS is to take Swiss-Prot fields of interest and apply statistical filters, mapping them to slightly broader categories relevant to fingerprint annotation. The particular fields and filters to be used are determined by analysis of the component Swiss-Prot entries to determine what type of fingerprint

they represent. This system works well overall, with the significant advantage that reports are English-like, in the sense that they largely re-use existing human annotation, but consequently exhibit the rather clipped, note-like style typical of Swiss-Prot. However, PRECIS fails where Swiss-Prot annotation is poor or inconsistent. Application of advanced natural language processing techniques could bring immediate benefits in resolving fingerprint type, which PRECIS occasionally misdiagnoses, and in filtering inconsistent Swiss-Prot annotation. Consider, for example, the following Swiss-Prot function statements:

```
The muscarinic acetylcholine receptor mediates various cellular responses, including inhibition of adenylate cyclase, breakdown of phosphoinositides & modulation of potassium channels through the action of G proteins. Primary transducing effect is inhibition of adenylate cyclase.
```

```
The muscarinic acetylcholine receptor mediates various cellular responses, including inhibition of adenylate cyclase, breakdown of phosphoinositides & modulation of potassium channels through the action of G proteins. Primary transducing effect is adenylate cyclase inhibition.
```

Although identical in biological meaning, the statements have syntactical differences that disrupt simple statistical filtering systems. Extraction and filtering of the biological meaning, rather than individual text blocks, would help resolve this issue.

Ideally, an automated annotation system would also augment the core annotation with information extracted from published literature. This would increase the scope of the resource and also help ensure annotation was up to date. For example, analysis of the voltage-dependent calcium channel gamma-2 subunits in Swiss-Prot (CCG2_MOUSE, CCG2_HUMAN) reveals that the proteins are "thought to stabilize the calcium channel in an inactivated (closed) state". However, new evidence suggests that the subunits are also implicated in cellular trafficking. They interact with ionotropic glutamate AMPA receptor subunits, a process that has been shown to be essential in delivering functional AMPA receptors to the surface membranes of cerebellar granule cells. Since this function has been reported fairly recently, it is not mentioned in Swiss-Prot annotation, nor covered in the literature references Swiss-Prot provides for these entries. Nevertheless, this aspect of their function is thought to be equally important as their channel modulatory properties, if not more so.

Although Swiss-Prot may not always carry the most up to date information about a particular entry, it nevertheless represents a considerably useful resource to gather not only annotation, but a host of additional information, such as protein synonyms, gene names, database cross references and so on. Although the focus of an extended annotation system would be published literature, the PRECIS annotation tool may represent a useful first pass system from which to gather information, such as search terms, as part of an iterative annotation process.

One of the features of hand-crafted PRINTS annotation is that it contains pointers to the appropriate literature references. In the case of structural literature references, this process is currently replicated by PRECIS. A sentence is synthesised using a standard template that states, "The structure has been determined, *e.g.* "Title 1" [*i*] and "Title 2" [*j*], where Titles 1 and 2 are the titles of crystallographic and NMR structure determination papers extracted from the Swiss-Prot reference lines. The bracketed numbers indicate that the papers have been added as *i*th and *j*th articles to the shared

papers generated from the earlier reference-gathering process. Ideally, this system would be extended to cover all information extracted from published literature reported in the annotation. Not only would this make information instantly traceable, but it could also allow users to control the weight of evidence required for annotation to be reported (*e.g.*, display only information that is listed in many publications, or, report the information even if it is listed in only one publication, *etc.*).

The final component of the *gd*; annotation field is a paragraph that covers the technical details of a fingerprint (see Figure 2 – red text). Completing some aspects of this paragraph automatically is trivial and can be performed through parsing of certain fields in the *ffmt* file into a standard template. However, other details, such as the locations of the motifs with respect to domains and other structural features, require an in-depth analysis of the literature and frequently require inferences to be made that are not explicitly stated in reports, *e.g.* by reference to graphical figures. Currently, analysis of motif locations is ignored. One method of addressing this issue would be to run sequence analysis software on and/or parse the Feature Table of representative sequences from a fingerprint and digest the results.

C) Extensive list of Web resources that may be exploited for automatic annotation of PRINTS entries

PubMed

Published literature is the most important and information-rich resource for fingerprint annotation. A service of the National Library of Medicine (NLM), PubMed provides access to over 12 million MEDLINE citations dating back to the 1960s.

Swiss-Prot and TrEMBL (SPTR)

Swiss-Prot is a protein sequence database that strives to provide high-level annotation, and TrEMBL is its computer-annotated supplement, generated from translation of EMBL nucleotide sequences. SPTR is a composite of the two resources and is directly relevant for PRINTS/prePRINTS annotation (although perhaps less so than for general annotation). Refer to SIB information for further details. Also GenBank and the DNA Data Bank of Japan (DDBJ).

Protein family/pattern databases

The most commonly used family databases include PRINTS, PROSITE, Pfam, SMART, TIGRFAMs, Blocks and ProDom (an exhaustive description of PRINTS annotation is provided elsewhere). These exploit multiple sequence alignments of protein families and describe individual family members by creating diagnostic signatures, or patterns. Besides providing a potent means of annotating existing sequences and characterising unknown sequences, resources such as PROSITE and PRINTS are a valuable source of hand-written annotation (Figure 6).

As an integrated documentation resource of protein families, domains and functional sites, InterPro was created in 1999 from the PROSITE, PRINTS and Pfam databases, but now includes signatures from ProDom, SMART and TIGRFAMs. Currently, the resource comprises 6725 entries, representing 1453 domains, 5121 families, 136 repeats, and 15 post-translational modification sites. Each InterPro entry is referenced by a scientific name and details the function, and in many cases structure and related diseases, of each family (Figure 7). Furthermore, a list of relevant references is

provided, together with related Gene Ontology (GO) terms. InterPro often lags behind its member databases, highlighting the need to search the annotation of individual databases, particularly PRINTS and PROSITE.

Online Mendelian Inheritance in Man (OMIM)

OMIM is a catalogue of human genes and genetic disorders developed for the Web by the NCBI. It contains copious amounts of textual information, pictures and references. Individual entries are split, where appropriate, into fields such as description, gene function, gene structure, mapping (chromosomal location) and information on animal models (Figure 8).

Other resources

A full A-Z listing of databases is available at <http://www3.oup.co.uk/nar/database/a/>. For example, the ABCdb is a database devoted to ATP-binding cassette (ABC) protein domains. As well as protein sequences, the database contains useful annotation on functional domains, sequence motifs, predicted transmembrane segments and signal peptides. It also includes a classification of subfamilies.

Gene and domain family homepages

A number of gene family and domain family-specific Websites are available, and listed in the A-Z list described above. For example, the Wnt family homepage has been created by the Howard Hughes Medical Institute. Here, information is available for members of the Wnt gene family, as well as other key components of the Wnt signalling pathway. The resource includes an extensive amount of annotation, and pertinent references.

World Wide Web

A vast source of largely unstructured documents.

D) Annotation environment

Annotation may be performed using any software package capable of editing plain text files. Typically PRINTS annotators use the xemacs or nedit packages running under linux, but flat files are also occasionally edited using notepad or similar packages running under Microsoft XP. PubMed is currently accessed via the Web, although there are plans to install a local copy on the Manchester network.

E) List of requirements

Fields to be completed

The PRECIS annotation software currently completes all blank fields in the naked ffmt file, except the accession numbers. This would also be the aim with respect to the BioMinT annotation tool. The greatest challenge resides in completion of the gd; annotation lines (see section B). This represents the most labour-intensive aspect of fingerprint generation and the most challenging for the current annotation pipeline.

Accuracy, traceability and refinement requirements

From the perspective of PRINTS annotation, the principal requirement is for accuracy, *i.e.* to generate annotation that is correct. Traceability (the source of each piece of information) is also important, as is the ability to refine annotation and

update it easily. In order to address these issues, ideally the end product of the automated annotation gathering and processing stages would be a data-structure rather than a text format report itself. A bespoke GUI would then handle the manipulation of the data-structure, allowing fine control of the output.

A preliminary sketch of one potential interface is given in Figure 9. We propose that the annotation be broken down into a series of categories, as a logical extension of the current PRECIS annotation system. Some of these categories would be fixed: structure, high level function and associated diseases, for example, would be processed if dealing with a super-family fingerprint (for a list of suggested primary and secondary categories for gene family, super-family and domain fingerprints, see Table 2). Further annotation categories would be user-defined, and also extracted automatically from the literature. For example, in Figure 9, which shows a mock-up of 5-HT_{2A} receptor fingerprint processing, information has been found pertaining to several super-family and family categories, and the user has requested information on 'genomic structure' and 'LSD'. Information on 'psychotropism' and 'schizophrenia' has also been extracted, since these terms were found to occur in the analysed literature in relation to the 5-HT_{2A} receptors with some frequency. In the main interface window, we find three statements of consensus biological meaning relating to the family level function. Each of these has a confidence score showing how many supporting statements have been found, and each one is selectable for inclusion in the final report. We envisage utilising the confidence score in a system whereby automatic annotation is generated using a default score for each category in order to select which statements are included initially. This information could then be refined by human annotators. The consensus biological meaning statements are also editable, allowing a human annotator full control of the output. Each consensus statement is linked to its respective supporting statements, each of which may be selected as the template consensus statement, allowing granular control of the look and feel of the final report. Finally, each supporting statement is linked directly to the abstract or Swiss-Prot entry from whence it came, allowing the annotator to instantly trace the information, and assess the context in which it was originally reported.

We envisage the GUI to have its own requirements, distinct from the annotation generation process, but nevertheless important in facilitating fingerprint annotation. Obvious requirements include a function whereby the report can be viewed as its construction progresses, a set of editing tools (spell checking, find and replace, *etc.*), and the ability to export the final report into a PRINTS-ffmt file. A particularly useful feature would be a literature reference path finder, so the report could be exported with a concise set of citations (*e.g.*, given a choice of citations for one consensus statement, the software would choose one that has already been cited for a different statement, rather than adding a fresh reference). It would also be important to store not only the original underlying data-structure, but also a 'snapshot' of the annotation in progress (*i.e.*, which categories had been selected, which consensus statements had been chosen for output, *etc.*). This would allow users to return to previous reports to refine them, and also allow annotation to be updated. A useful function of the 'snapshot' would be to 'bless' fields that had been hand-checked or corrected by a human annotator, where the consensus information was confusing or contained errors. A suitable meta-data layer in the underlying data-structure should facilitate this.

Efficiency

Hand-crafted PRINTS entries are deposited at a rate of 50 every three months. The

prePRINTS pipeline has a capability to produce fingerprints at the rate of around 30 per day (although approximately only 25% of these are useable once non-specific “noisy” prePRINTS have been removed). Any reasonably efficient automatic annotation system with run times of minutes to hours is therefore unlikely to represent a significant bottleneck in the generation of PRINTS or prePRINTS entries. Biologists submitting queries over a Web interface may have higher demands in terms of speed, especially if a real-time response is required. This point will require further discussion with potential end users.

'Understandability' of results

One particular advantage of the PRECIS annotation system is that its results are English-like in that they largely re-use Swiss-Prot information, which has in turn been hand-crafted by a team of human curators. This feature of the annotation makes it much more useful to biological researchers than a mere list of keywords and associated statistical values, which some so-called annotation packages provide. It is particularly important that this principle is maintained in any extended annotation system. The process described above, whereby representative statements encapsulating consensus biological meanings may be selected and refined, would help to ensure annotation results are instantly readable.

F) PRINTS annotation versus general annotation

Although much of the discussion so far has centred on the automatic annotation of PRINTS entries, many of the principles underlying completion of the PRINTS *gd* annotation field can be applied to queries relating to more general biological annotation. Inevitably, the level of focus may be somewhat different, moving beyond gene family, super-family or domain family annotation, with more emphasis on user-defined categories of interest, or those automatically extracted from the literature. However, the main difference between PRINTS and general biological annotation is the starting point. A fingerprint is defined by its component Swiss-Prot entries, so the pertinence of information is predetermined: it is clear to which sequences the annotation should relate, and hence an automated system should have at least some indication of which information sources to query and digest.

In the case of more general biological annotation, the starting point is likely to be more nebulous. For example, a biologist researching the function of a particular protein family is unlikely to know their Swiss-Prot identifiers. One method of returning this information would be to query Swiss-Prot directly using the sequence retrieval system (SRS), which allows multiple queries to be linked using Boolean operators. However, we would suggest that this approach is less than ideal. Our own experience of SRS suggests that it does not allow adequate control of free text searching, often inundating the user with false positives, nor provide appropriate ranking of hits, so that true positives may languish towards the bottom of extensive and largely irrelevant hit lists. Application of an alternative indexing system may alleviate some of these problems. Nevertheless, we consider it likely that published literature will be the information source of choice for general biological queries, since it represents a much larger, more up to date source of annotation than Swiss-Prot, and general queries may relate to more abstract concepts not represented in the database.

Figure legends

Figure 1: Overview of the PRINTS annotation process.

Figure 2: A hand-annotated fingerprint for the janus kinase 1 family. Super-family level annotation is displayed in blue, family level in green, and technical details in red text. Three literature references have been removed for reasons of space.

Figure 3: A hand-annotated fingerprint for SH2 domains.

Figure 4: A 'naked' ffmt file for the janus kinase 1 family.

Figure 5: A fingerprint for the janus kinase 1 family annotated automatically by PRECIS.

Figure 6: Part of the PROSITE documentation for family 10 glycosyl hydrolases, PDOC00510. The annotation provides a breakdown of the enzymes that are known to belong to the family, illustrating the range of taxa that express this enzyme. The signature for this family is provided and functionally important residues are distinguished. The entry also contains a number of literature references (not shown).

Figure 7: Part of InterPro entry IPR006201, the neurotransmitter-gated ion-channel family. The entry is represented by signatures from PRINTS, PROSITE and TIGRFAMs, and features two domains that are conserved in all neurotransmitter-gated ion-channels. Family relationships are expressed in terms of a single 'parent' (not present in this entry) and one or more 'children', although only those relations that have a corresponding signature are described. Where possible, the molecular function, biological process and cellular component applicable to the protein are described using gene ontology (GO) terminology. An extensive amount of hand-written annotation is included, together with example proteins and relevant literature references.

Figure 8: Part of OMIM entry 137160, the gamma-aminobutyric acid A receptor, alpha 1 subunit (GABRA1). The entry contains detailed annotation outlining function, family, chromosomal location, molecular genetics and animal model data, as well as allelic variants and pertinent references (not shown).

Figure 9: Screen shot showing the proposed functionality of one potential graphical user interface (GUI). A set of information categories are listed in the left-hand window, some of which are predetermined, some requested by the user and others extracted from the literature. Any one of these categories can be clicked on to reveal a list of consensus biological meanings extracted from the literature, together with supporting statements and literature references used to derive each statement.

Table 1: Characteristics of the three main fingerprint types: gene family, super-family and domain family.

Table 2: Primary and secondary annotation requirements for gene family, super-family and domain family fingerprints. Primary annotation describes that which it is necessary to report if information is available. Secondary annotation describes additional desirable categories of information.

Figure 1

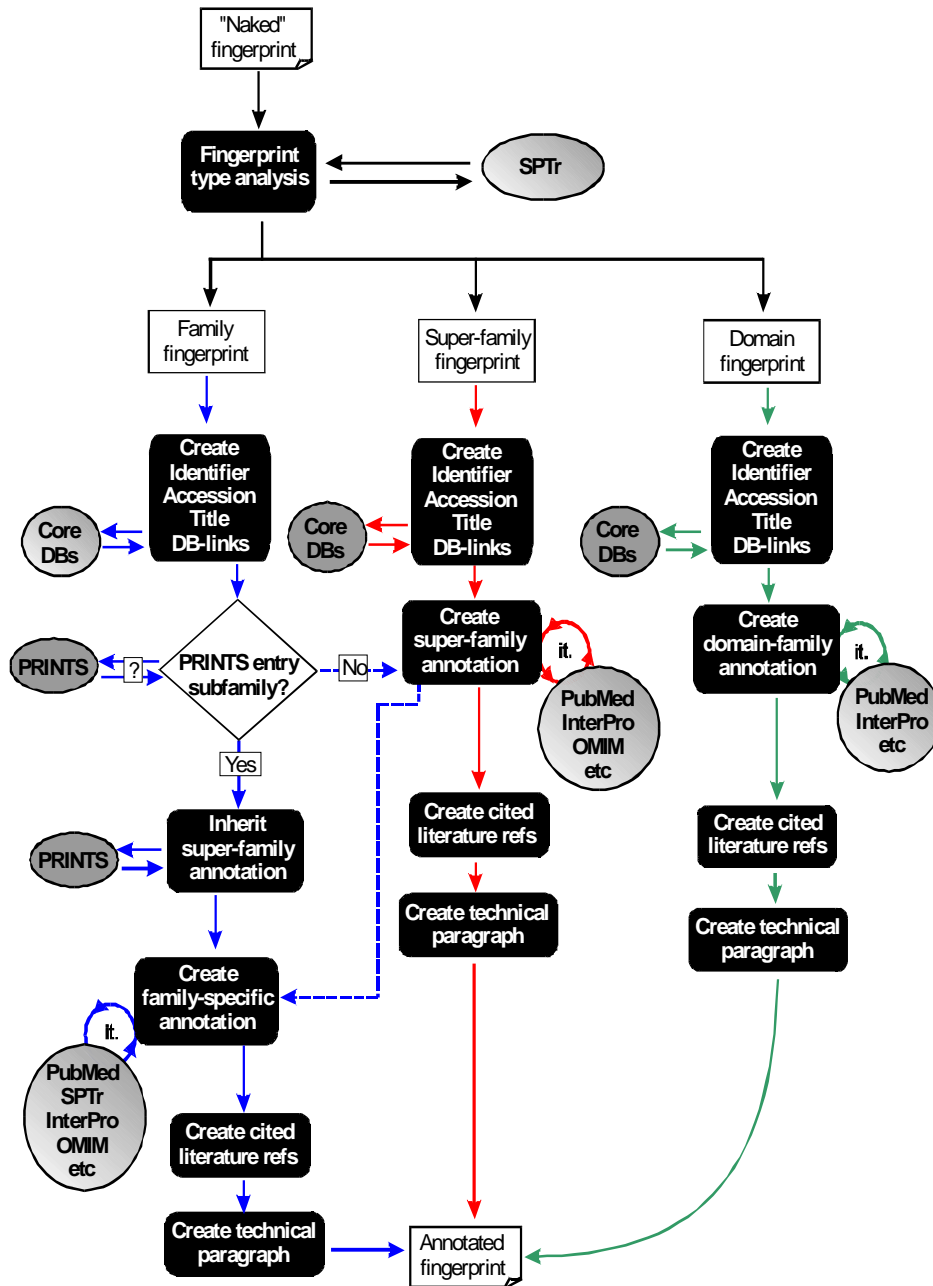


Figure 2

gc; JANUSKINASE1
gx; PR017
gn; COMPOUND(4)
ga; 09-OCT-2002
gt; Janus kinase 1 (JAK1) signature
gp; PRINTS; PR00109 TYRKINASE; PR01700 JANUSKINASE; PR01701 JANUSKINASE2
gp; PRINTS; PR01702 JANUSKINASE3; PRINTS PR01703 TYK2
gp; MIM; 147795
bb;
gr; 1. IHLE, J.N.
gr; Cytokine receptor signalling.
gr; NATURE 377 591-594 (1995).
gr;
gr; 2. LEONARD, W.J. AND O'SHEA, J.J.
gr; JAKS AND STATS: Biological Implications.
gr; ANNU.REV.IMMUNOL. 16 293-322 (1998).
gr;
gr; 3. IMADA K. AND LEONARD, W.J.
gr; The Jak-STAT pathway.
gr; MOL.IMMUNOL. 37 1-11 (2000).
gr;
gr; 4. HARPUR, A.G, ANDRES, A.C., ZIEMIECKI, A., ASTON, R.R. AND WILKS, A.F.
gr; JAK2, a third member of the JAK family of protein tyrosine kinases.
gr; ONCOGENE 7 1347-1353 (1992).
bb;
bb;
gd; The janus kinase (JAK) proteins are a family of tyrosine kinases that
gd; function in membrane-proximal signalling events initiated by a variety of
gd; extracellular factors binding to cell surface receptors. In particular, many
gd; type I and II cytokine receptors lack a protein tyrosine kinase domain and
gd; rely on JAKs to initiate the cytoplasmic signal transduction cascade [1,2].
gd; Ligand binding induces oligomerisation of the receptors which then activates
gd; the cytoplasmic receptor-associated JAK proteins. These subsequently
gd; phosphorylate tyrosine residues along the receptor chains they are
gd; associated with. The phosphotyrosine residues are a target for a variety of
gd; SH2 domain-containing transducer proteins. Amongst these are the signal
gd; transducers and activators of transcription (STAT) proteins which, after
gd; binding to the receptor chains, are phosphorylated by the JAK proteins.
gd; Phosphorylation enables the STAT proteins to dimerise and translocate into
gd; the nucleus where they alter the expression of cytokine-regulated genes.
gd; This process is known as the JAK-STAT pathway [3]
gd;
gd; Four mammalian JAK family members have been identified: JAK1, JAK2, JAK3,
gd; and TYK2. They are relatively large kinases of approximately 1150 amino
gd; acids with molecular weights of around 120-130 kDa. Their amino acid
gd; sequences are characterised by the presence of seven highly conserved
gd; domains referred to as JAK homology (JH) domains [4]. The C-terminal domain
gd; (JH1) is responsible for the tyrosine kinase function. The next domain along
gd; the sequence (JH2) is known as the tyrosine kinase-like domain, since its
gd; sequence shows high similarity to functional kinases but it does not possess
gd; any catalytic activity [5]. Although the function of this domain is not well
gd; established there is some evidence for a regulatory role on the JH1 domain,
gd; thus modulating catalytic activity [6]. The N-terminal portion of the JAKs
gd; (spanning JH7 to JH3) is important for receptor association and
gd; non-catalytic activity [7].
gd;
gd; JAK1 was initially cloned using a PCR-based strategy utilising degenerate
gd; primers corresponding to conserved motifs within the catalytic domain of
gd; protein-tyrosine kinases [5]. In common with JAK2 and TYK2, and in contrast
gd; to JAK3, JAK1 appears to be ubiquitously expressed [2].
gd;
gd; JANUSKINASE1 is a 4-element fingerprint that provides a signature for the
gd; janus kinase 1 (JAK1) proteins. The fingerprint was derived from an initial
gd; alignment of 3 sequences: the motifs were drawn from conserved regions
gd; spanning virtually the full alignment length, focusing on those sections
gd; that characterise JAK1 but distinguish it from other family members - motif
gd; 1 lies in the JH6 domain; motif 2 resides between the JH4 and JH3 domains;
gd; motif 3 lies partially between the JH4 and JH3 domains and partially within
gd; the JH4 domain; and motif 4 resides within the JH2 tyrosine kinase-like
gd; domain. Three iterations on SPTR40_20f were required to reach convergence,
gd; at which point a true set comprising 7 sequences was identified.

Figure 3

```
gc; SH2DOMAIN
gx; PR00401
gn; COMPOUND(5)
ga; 01-NOV-1995; UPDATE 22-JUN-1999
gt; SH2 domain signature
gp; PRINTS; PR00452 SH3DOMAIN
gp; INTERPRO; IPR000980
gp; PROSITE; PS50001 SH2
gp; BLOCKS; BL50001
gp; PFAM; PF00017 SH2
gp; PDB; 1AB2
gp; SCOP; 1AB2
gp; CATH; 1AB2
bb;
gr; 1. WAKSMAN, G., KOMINOS, D., ROBERTSON, S.C., PANT, N., BALTIMORE, D.,
gr; BIRGE, R.B., COWBURN, D., HANAFUSA, H., MAYER, B.J., OVERDUIN, M.,
gr; RESH, M.D., RIOS, C.B., SILVERMAN, L. AND KURIYAN, J.
gr; Crystal structure of the phosphotyrosine recognition domain SH2 of v-src
gr; complexed with tyrosine-phosphorylated peptides.
gr; NATURE 358 646-653 (1992).
gr;
gr; 2. MAYER, B.J. AND BALTIMORE, D.
gr; Signalling through SH2 and SH3 domains.
gr; TRENDS CELL BIOL. 3 8-13 (1993).
gr;
gr; 3. OVERDUIN, M., RIOS, C.B., MAYER, B.J., BALTIMORE, D. AND COWBURN, D.
gr; The 3D solution structure of the src homology 2 domain of c-abl.
gr; CELL 70(4) 697-704 (1992).
bb;
bb;
gd; The SH2 (src Homology-2) domains are small protein modules containing
gd; approximately 100 amino acid residues. They are found in a wide variety of
gd; protein contexts: e.g., in association with catalytic domains of phospho-
gd; lipase Cy (PLCy) and the nonreceptor protein tyrosine kinases; within
gd; structural proteins such as fodrin and tensin; and in a group of small
gd; adaptor molecules, i.e Crk and Nck. In many cases, when an SH2 domain is
gd; present so too is an SH3 domain, suggesting that their functions are
gd; inter-related. The domains are frequently found as repeats in a single
gd; protein sequence.
gd;
gd; The structure of the SH2 domain belongs to the alpha+beta class, its
gd; overall shape forming a compact flattened hemisphere. The core structural
gd; elements comprise a central hydrophobic anti-parallel beta-sheet, flanked
gd; by 2 short alpha-helices. In the v-src oncogene product SH2 domain, the
gd; loop between strands 2 and 3 provides many of the binding interactions
gd; with the phosphate group of its phosphopeptide ligand, and is hence
gd; designated the phosphate binding loop.
gd;
gd; SH2DOMAIN is a 5-element fingerprint that provides a signature for SH2
gd; domains. The fingerprint was derived from an initial alignment of 21
gd; sequences: the motifs were drawn from short conserved regions spanning the
gd; full alignment length, and largely correspond to the core structural
gd; elements (i.e., the N-terminal helix, each of the central 3 strands, and
gd; the C-terminal helix respectively) - motif 2 contains a highly conserved
gd; FLVRES sequence involved in phosphate binding (cf. PROSITE profile SH2
gd; (PS50001)). Five iterations on OWL26.2 were required to reach convergence,
gd; at which point a true set comprising 187 sequences was identified. Thirty-
gd; nine partial matches were also found (the fingerprint does not perform
gd; perfectly because of the low level of sequence similarity between SH2
gd; domains (the result of their disparate functions), and is further
gd; complicated by their multiple repeat nature).
```

Figure 4

```

gc; JAK1
gx; PR
gn; COMPOUND(4)
ga; 21-FEB-2003
gt; SIGNATURE NAME
gp; DATABASE LINKS
bb;
gr; LITERATURE REFERENCES
bb;
bb;
gd; ANNOTATION
bb;
bb;
si; SUMMARY INFORMATION
si; -----
sd; 7 codes involving 4 elements
sd; 0 codes involving 3 elements
sd; 0 codes involving 2 elements
bb;
bb;
ci; COMPOSITE FINGERPRINT INDEX
ci; -----
cr;
cd; 4 | 7 7 7 7
cd; 3 | 0 0 0 0
cd; 2 | 0 0 0 0
cd; ---+-----
cd; | 1 2 3 4
bb;
bb;
tp; JAK1_HUMAN Q9TTJ1 JAK1_MOUSE Q9PWM9
KA; P23458 M1 Q9TTJ1 M1 P52332 D1 Q9PWM9 M1
tp; JAK1_CYPCA O12990 O57612
KA; Q09178 M1 O12990 M1 O57612 M1
bb;
KF; FALSE_PARTIAL_POSITIVES
bb;
tt; JAK1_HUMAN Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1)
(JAK-1) - Homo sapiens (Human).
tt; Q9TTJ1 JANUS KINASE 1 - Sus scrofa (Pig).
tt; JAK1_MOUSE Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1)
(JAK-1) - Mus musculus (Mouse).
tt; Q9PWM9 TYROSINE KINASE JAK1 - Gallus gallus (Chicken).
tt; JAK1_CYPCA Tyrosine-protein kinase JAK1 (EC 2.7.1.112) (Janus kinase 1)
(JAK-1) - Cyprinus carpio (Common carp).
tt; O12990 TYROSINE-PROTEIN KINASE JAK1 (EC 2.7.1.112) (JANUS KINASE 1)
(JAK-1) (JAK1 KINASE) - Brachydanio rerio (Zebrafish) (Zebra danio).
tt; O57612 JAK1 TYROSINE KINASE - Tetraodon fluviatilis (Puffer fish).
bb;
bb;
sh; SCAN HISTORY
sh; -----
dn; SPTR40_20f 3 300 NSINGLE
bb;
bb;
im; INITIAL MOTIF-SETS
im; -----
ic; JAK11
il; 11
it; MOTIF_NAME I - 1
id; ETLNKTIRQRN Q9PWM9 223 223
id; ETLNKSIRQRN Q9TTJ1 211 211
id; ETLNKSIRQRN JAK1_MOUSE 223 223
id; DSLNRTIKQRN JAK1_CYPCA 222 222
id; DSLNRTIKQRN O12990 221 221
id; ETLNKSIRQRN JAK1_HUMAN 211 211
id; ETLNRSIKQRS O57612 219 219
bb;

```

Figure 5

```
gc; JAK1
gx; PP
gn; COMPOUND(4)
ga; 09-OCT-2002
gt; Tyrosine-protein kinase jak1 (ec 2.7.1.112) (janus kinase 1) (jak-1) signature
gp; PROSITE; PS00107 PROTEIN_KINASE_ATP; PS00109 PROTEIN_KINASE_TYR; PS50001 SH2
gp; PROSITE; PS50011 PROTEIN_KINASE_DOM
gp; PFAM; PF00069 pkinase
gp; INTERPRO; IPR000299; IPR000719; IPR000980; IPR001245
gp; HSSP; 1FGK
gp; MIM; 147795
bb;
gr; 1. CHANG, M.S., CHANG, G.D., LEU, J.H., HUANG, F.L., CHOU, C.K., HUANG,
gr; C.J. AND LO, T.B.
gr; Expression, characterization, and genomic structure of carp JAK1 kinase
gr; gene.
gr; DNA CELL BIOL. 15 827-844 (1996).
gr;
gr; 2. LEE, S.-T., STRUNK, K.M. AND SPRITZ, R.A.
gr; A survey of protein tyrosine kinase mRNAs expressed in normal human
gr; melanocytes.
gr; ONCOGENE 8 3403-3410 (1993).
gr;
gr; 3. YANG, X., CHUNG, D. AND CEPKO, C.L.
gr; Molecular cloning of the murine JAK1 protein tyrosine kinase and its
gr; expression in the mouse central nervous system.
gr; J.NEUROSCI. 13 3006-3017 (1993).
gr;
gr; 4. WILKS, A.F., HARPUR, A.G., KURBAN, R.R., RALPH, S.J., ZUERCHER, G.
gr; AND ZIEMIECKI, A.
gr; Two novel protein-tyrosine kinases, each with a second
gr; phosphotransferase-related catalytic domain, define a new class of
gr; protein kinase.
gr; MOL.CELL.BIOL. 11 2057-2065 (1991).
bb;
bb;
gd; Function:
gd; Tyrosine kinase of the non-receptor type, involved in the
gd; ifn-alpha/beta/gamma signal pathway. Kinase partner for the interleukin
gd; (il)-2 receptor.
gd;
gd; Atp + a protein tyrosine = adp + protein tyrosine phosphate.
gd;
gd; Additional Info:
gd; Wholly intracellular, possibly membrane associated.
gd;
gd; Family and structural information:
gd; Possesses two phosphotransferase domains. The second one probably
gd; contains the catalytic domain (by similarity), while the presence of
gd; slight differences suggest a different role for domain 1.
gd;
gd; Belongs to the tyr family of protein kinases. Jak subfamily.
gd;
gd; Contains sh2 domains.
gd;
gd; Keywords: Transferase; Tyrosine-protein kinase; ATP-binding;
gd; Phosphorylation; SH2 domain; Repeat.
gd;
gd; JAK1 is a 4-element fingerprint that provides a signature for the
gd; tyrosine-protein kinase jak1 (ec 2.7.1.112) (janus kinase 1) (jak-1)
gd; proteins. The fingerprint was derived from an initial alignment of 7
gd; sequences: the motifs were drawn from conserved regions spanning
gd; virtually the full alignment length. Three iterations on SPTR40_20f were
gd; required to reach convergence, at which point a true set comprising 7
gd; sequences was identified.
bb;
bb;
si; SUMMARY INFORMATION
si; -----
sd; 7 codes involving 4 elements
sd; 0 codes involving 3 elements
sd; 0 codes involving 2 elements
```

Figure 6

Glycosyl hydrolases family 10 active site

PROSITE cross-reference(s)	
PS00591: GLYCOSYL_HYDROL_F10	
Documentation	
<p>The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family F [3] or as the glycosyl hydrolases family 10 [4,E1]. The enzymes which are currently known to belong to this family are listed below.</p> <ul style="list-style-type: none"> - Aspergillus awamori xylanase A (xynA). - Bacillus sp. strain 125 xylanase (xynA). - Bacillus stearothermophilus xylanase. - Butyrivibrio fibrisolvens xylanases A (xynA) and B (xynB). - Caldocellum saccharolyticum bifunctional endoglucanase/exoglucanase (celB). This protein consists of two domains; it is the N-terminal domain, which has exoglucanase activity, which belongs to this family. - Caldocellum saccharolyticum xylanase A (xynA). - Caldocellum saccharolyticum ORF4. This hypothetical protein is encoded in the xynABC operon and is probably a xylanase. - Cellulomonas fimi exoglucanase/xylanase (cex). - Clostridium stercoararium thermostable celloxylanase. - Clostridium thermoecellum xylanases Y (xynY) and Z (xynZ). - Cryptococcus albidus xylanase. - Penicillium chrysogenum xylanase (gene xylP). - Pseudomonas fluorescens xylanases A (xynA) and B (xynB). - Ruminococcus flavefaciens bifunctional xylanase XYL1 (xynA). This protein consists of three domains: a N-terminal xylanase catalytic domain that belongs to family 11 of glycosyl hydrolases; a central domain composed of short repeats of Gln, Asn and Trp, and a C-terminal xylanase catalytic domain that belongs to family 10 of glycosyl hydrolases. - Streptomyces lividans xylanase A (xlnA). - Thermoanaerobacter saccharolyticum endoxylanase A (xynA). - Thermoascus aurantiacus xylanase. - Thermophilic bacterium Rt8.B4 xylanase (xynA). <p>One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the exoglucanase from Cellulomonas fimi, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. We have used this region as a signature pattern.</p>	
Description of pattern(s) and/or profile(s)	
Consensus pattern	[GTA]-x(2)-[LIVN]-x-[IVMF]-[ST]-E-[LIY]-[DN]-[LIVMF] [E is the active site residue]
Sequences known to belong to this class detected by the pattern	ALL, except for Thermoascus aurantiacus xylanase whose sequence seems to be incorrect.
Other sequence(s) detected in SWISS-PROT	16.
Expert(s) to contact by email	
Hennissat B.	bernie@afmb.cnrs-nrs.fr

Figure 7

Neurotransmitter-gated ion-channel

Database	InterPro
Accession	IPR006201; Neur_channel (matches 454 proteins)
Name	Neurotransmitter-gated ion-channel
Type	Family i
Dates	28-OCT-2002 (created) 28-OCT-2002 (last modified)
Signatures	PR00252; NRIONCHANNEL (411 proteins) PS00236; NEUROTR_ION_CHANNEL (445 proteins) TIGR00860; LIC (318 proteins)
Secondary no.	IPR001175
Contains i	IPR006029; Neurotransmitter-gated ion-channel transmembrane region (480 proteins) IPR006202; Neurotransmitter-gated ion-channel ligand binding domain (497 proteins)
Children i [tree]	IPR002394; Nicotinic acetylcholine receptor (176 proteins) IPR006028; Gamma-aminobutyric acid A receptor (221 proteins)
Process i	transport (GO:0006810)
Function i	extracellular ligand-gated ion channel (GO:0005230)
Component i	membrane (GO:0016020)
Abstract i	<p>Neurotransmitter ligand-gated ion channels are transmembrane receptor-ion channel complexes that open transiently upon binding of specific ligands, allowing rapid transmission of signals at chemical synapses [1, 2].</p> <p>Of the five families known, four have been shown to form a sequence-related superfamily. These are the gamma-aminobutyric acid type A (GABA-A), nicotinic acetylcholine, glycine and the serotonin 5HT3 receptors. The ionotropic glutamate receptors (IPR001320) have a distinct primary structure.</p> <p>However, all these receptors possess a pentameric structure (made up of varying subunits), surrounding a central pore. Each of these subunits contains a large extracellular N-terminal ligand-binding region; 3 hydrophobic transmembrane domains; a large intracellular region; and a fourth hydrophobic domain [1, 2].</p> <p>This InterPro entry represents the GABA-A, nicotinic, glycine, and 5HT3 receptors.</p>
Examples	<ul style="list-style-type: none"> ● O00591 GAAP_HUMAN ● O09028 GAAP_RAT ● O70212 5HT3_CAVPO <p>View examples</p>
References	1. Wagner K., Edson K., Heginbotham L., Post M., Haganir R.L., Czernik A.J.

Figure 8

CLONING

The GABA-A receptor is the receptor for the major inhibitory neurotransmitter in the vertebrate brain. It is known to contain alpha and beta subunits; 3 isoforms of the alpha subunit and 1 beta subunit have been identified by cDNA cloning from bovine brain. [Garrett et al. \(1988\)](#) isolated a cDNA clone of an alpha subunit of the human GABA-A receptor. The 1,055-bp clone contained an open reading frame and 260 nucleotides in the 5-prime noncoding region. The 351-amino acid sequence shows 97% homology with its bovine counterpart. Hybridization of the clone to Northern blots showed an RNA doublet in human cortex and in rat whole brain, cortex, hippocampus, midbrain, olfactory bulb, and cerebellum. 🧠

MAPPING

By in situ hybridization, [Buckle et al. \(1989\)](#) mapped 2 of the isoforms of the alpha subunit, GABRA1 and GABRA2 ([137140](#)), to 5q34-q35 and 4p13-p12, respectively. The gene for the beta subunit (GABRB1; [137190](#)) also was mapped to 4p13-p12, where it may be located in tandem to the GABRA2 gene.

By study of segregation in interspecies backcrosses, [Keir et al. \(1991\)](#) demonstrated that in the mouse Gabra1 is on chromosome 11 between I1-3 ([147740](#)) and Rel ([164910](#)). By linkage analysis using a highly polymorphic (CA)_n repeat within the GABRA1 gene, [Johnson et al. \(1992\)](#) refined the assignment of the human gene on distal 5q. 🧠

[Russek \(1999\)](#) determined that GABRA1 is a member of a gene cluster spanning approximately 480 kb of chromosome 5q34. The order of the genes is GABRB2 ([600232](#))--GABRA6 ([137143](#))--GABRA1--GABRG2 ([137164](#)).

MOLECULAR GENETICS

[Cossette et al. \(2002\)](#) studied a French Canadian family in which all affected family members with epilepsy, in an autosomal dominant pedigree pattern, had a similar phenotype that fulfilled the criteria for juvenile myoclonic epilepsy (JME; [606904](#)). The only exception was an individual in the most recent of 4 affected generations who had an earlier onset of disease but clinical features that were otherwise indistinguishable from those of other members in the family. A genome scan provided evidence of linkage to 5q34, with a maximum lod score of 3.1 at theta = 0 for marker D5S414. The fine mapping showed that the candidate region included a cluster of GABA-A receptor subunits. Screening for mutations in these GABA receptor genes revealed an ala322-to-asn amino acid substitution ([137160.0001](#)) in all 8 of the affected members available for study and in none of the unaffected members of the family. 🧠

ANIMAL MODEL

[Rudolph et al. \(1999\)](#) introduced a histidine-to-arginine point mutation at codon 101 (H101R) of the murine alpha-1 GABA-A receptor. Alpha-1 H101R mice showed no overt distinctive phenotype and bred normally. Immunoblotting confirmed that the mutant alpha-1 subunit and other major GABA-A receptor subunits were expressed in the alpha-1 H101R mice at normal levels. The immunohistochemical distribution of these subunits in mutant mice was indistinguishable from that of wildtype. The alpha-1 subunit gene is expressed mainly in cortical areas and thalamus and is rendered insensitive to allosteric modulation by benzodiazepine-site ligands in mutant mice, while regulation by GABA is preserved. Alpha-1 H101R mice failed to show the sedative, amnesic, and partly the anticonvulsant action of diazepam. In contrast, the anxiolytic-like, myorelaxant, motor-impairing, and ethanol-potentiating effects were fully retained, and are attributed to the nonmutated GABA-A receptors found in the limbic system (alpha-2; alpha-5, [137142](#)), in monoaminergic neurons (alpha-3, [305660](#)), and in motoneurons (alpha-2, alpha-5). Thus, benzodiazepine-induced behavioral responses are regulated by specific GABA-A receptor subtypes that contribute to distinct neuronal circuits. 🧠

Figure 9

The screenshot displays the Vapour-ware 2003 application window. The interface is divided into several sections:

- Information Category:** A sidebar on the left with expandable sections:
 - Super-family: Structure, High level function, Sub-families
 - Family: Function (selected), Disease
 - User-specified categories: Genomic structure, LSD
 - Extracted categories: Psychotropism, Schizophrenia
- Annotation:** A table with columns for 'Consensus biological meaning', 'Supporting statements', 'Select', and 'Reference'.

Consensus biological meaning:	Supporting statements:	Select:	Reference:
<input checked="" type="checkbox"/> Receptor for 5-HT <small>[edit] (4/5) [show]</small>	5-HT2A is a receptor for 5-HT	<input checked="" type="checkbox"/>	PubMed 283832
<input type="checkbox"/> Activates a phosphatidylinositol-calcium second messenger system <small>[edit] (4/5) [show]</small>	The serotonin receptor 5-HT2A is expressed in...	<input type="checkbox"/>	PubMed 240712
<input type="checkbox"/> Involved in smooth muscle contraction <small>[edit] (1/5) [show]</small>	Members of the 5-HT2A receptor subfamily display a low affinity for their endogenous ligand serotonin.	<input type="checkbox"/>	PubMed 657826
	This is one of the several different receptors for 5-hydroxytyptamine (serotonin) ..	<input type="checkbox"/>	5HT2A_HUMAN
- Bottom Panel:** Three icons with labels: 'New...' (book icon), 'Update' (globe icon), and 'Delete' (foot icon).

Table 1

Relationship		Common Features	Example
Gene family	Species orthologues (sub-family) Group of paralogues (family)	Gene sub-families perform the same function in different organisms High level of similarity in primary sequence Gene families perform different but related functions in an organism Families found in different species Greater level of diversity in primary structure than sub-family level	<u>Sub-family:</u> 5-HT _{2A} receptor Somatostatin receptor type 1 GABA(A) receptor alpha 1 subunit <u>Family:</u> 5-HT ₂ receptors GABA(A) receptor subunits Somatostatin receptors
Super-family	Group of related gene families	Homology sometimes difficult to prove through sequence similarity Encompass a wide variety of functions May share a common fold or structural framework	Rhodopsin-like GPCRs Neurotransmitter-gated ion channels
Domain family	Related by the presence of a common independent structural unit	Encompass a wide variety of functions Similarity between family members restricted to the domain Domain may confer shared function or activity to members	Zinc-finger containing DNA binding proteins Disparity of function in SH3 domain-containing proteins

Table 2

	Family	Super-family	Domain
Primary annotation	Super-family level common annotation Specific function Specific structure Associated diseases / alleles	High level function Sub-family overview Structure General super-family associated diseases	Domain function Length / amino acid composition Domain structure Found in' relationships & ancestry Associated diseases
Secondary annotation	Species orthologues Pattern of expression History of discovery Cytological location Gene structure / splice variants Other physicochemical properties	Species distribution General expression pattern General history of discovery Other physicochemical properties	Species distribution History of discovery Domain context Functionally important residues