PRECIS: An automated pipeline for producing concise reports about proteins

P.W.Lord[1], J.R.Reich[1], A.Mitchell[2], R.D.Stevens[1],
T.K.Attwood[2] and C.A.Goble [1]
Department of Computer Science[1] and School of Biological Sciences[2]
University of Manchester
Oxford Road
Manchester
M13 9PL
UK
p.lord@russet.org.uk
[mitchell,attwood]@bioinf.man.ac.uk
[c.goble,r.stevens]@cs.man.ac.uk

Abstract

There have been several attempts at addressing the problem of annotating sequence data computationally. Annotation generation can be considered a pipeline of processes: first harvesting data from a variety of data sources, then distilling and transforming it into a form more appropriate for the end database. This task is usually performed by human annotators, a solution that is clearly not scaleable. There have been several attempts to mimic some of these pipelines in software. However, these have generally focused on low level annotation, such as database cross-references, or by harvesting data from computational techniques such as gene finding or similarity searches. Higher level annotation such as that seen in the PRINTS database is usually formed from data that is free text, or only partly structured. This presents a much greater computational challenge. Therefore we studied the pipeline that is used to generate annotation for the PRINTS database, and have developed prototype software that reflects and automates this pipeline. As this software operates primarily on data culled from the SWISS-PROT database, we have called it PRECIS (Protein Reports Engineered from Concise Information in SWISS-PROT). This software is currently being used to generate annotation for the prePRINTS database. As the output is a structured report detailing the function, structure and disease associations of a protein, and providing literature references and keywords we believe it will be of more generic use. The software is available on request from `mitchell@bioinf.man.ac.uk`.

1 Introduction

With the ability to sequence entire genomes, biological databases have become essential resources for storing, manipulating and extracting information about biological entities. These databases contain relatively large amounts of data, which is of a rich and varied set of data types. Most of the biological databases however are built on top of a single type of data: the biological sequence, either protein or DNA. On its own, biological sequences are not very informative or useful. They require large amounts of subsidiary information or "annotation" of the sequence to be useful to a biologist.

The word annotation covers a multitude of sins. Here, we classify annotation into two forms:

1. "low-level" systematised, prescriptive annotation, for example cross-links to other databases, information about the organism of origin and sequence/data submitter. These are typically expressed using some form of mark up and are computationally processable and generate-able.

2. "high-level" semi-structured text-based annotation, representing the accumulated knowledge of the biological community about the data entry, culled from other sources such as other database entries annotations and the literature. The annotation is intended to be human readable rather than machine processable.

In the primary databases, such as the DNA databases EMBL [21], or GENBANK, [8], annotation is commonly of the "low-level" kind. Secondary databases use both pri-

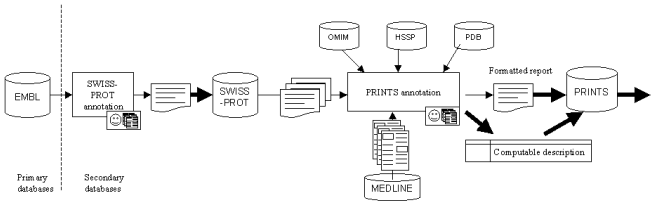

Figure 1. The flow of information through the annotation pipeline. A large variety of databases or other data sources are used, including database entries from for example EMBL, and primary literature. Human annotation takes place both in SWISS-PROT and PRINTS entry formation.

```
Prion protein signature

PROSITE; PS00291 PRION_1; PS00706 PRION_2
BLOCKS; BL00291
PFAM; PF00377 prion
INTERPRO; IPR000817

1. STAHL, N. AND PRUSINER, S.B.
Prions and prion proteins.
FASEB J. 5 2799-2807 (1991).


Prion protein (PrP) is a small glycoprotein found in high quantity
in the brain of animals infected with certain degenerative
neurological diseases, such as sheep scrapie and bovine spongiform
encephalopathy (BSE), and the human dementias Creutzfeldt-Jacob
disease (CJD) and Gerstmann-Straussler syndrome (GSS).
```

Figure 2. Example protein family annotation produced by a human annotator. The main body of the annotation comes in coherent paragraphs, and contains biological detail from a variety of different sources as well as SWISS-PROT. The entry has been abridged.

mary and secondary databases, drawing on their data, linking and analysing it and adding value to it (see Figure 1). These secondary databases contain both kinds of annotation. For example, the SWISS-PROT database (`http://www.expasy.org`)[7] carries information about the individual proteins which can be traced back to the DNA sequence found in EMBL, or GENBANK. SWISS-PROT annotation contains additional information about the protein, such as functions or structures. This is added manually by expert human annotators who make use of the large amounts of data present in the primary literature, distilling and condensing it into the form seen.

The PRINTS database [5] is in turn built on top of SWISS-PROT. PRINTS is a pattern database, gathering information about sets of related proteins (protein "families"). An example PRINTS entry can be seen in Figure 2. This database enables more refined, sensitive similarity searching because the target sequences are formed as a consensus of more than one protein sequences, a process known as "fingerprinting". PRINTS is a primary source of fingerprint family information for G-Protein Coupled Receptors (GPCRs), which currently account for 50% of drug targets by the pharmaceutical industry, where high quality annotation is particularly useful.

To create an entry for the PRINTS database, the sequence fingerprint is identified and a list of proteins that are potentially related (SWISS-PROT ID's) is generated. Database cross-references are followed and high level textual annotation is added describing the current state of biological knowledge about the protein families. This data comes from a variety of sources, notably SWISS-PROT annotation and the primary literature. This process involves both data gathering and distillation.

This flow of data from one database to another is called the "annotation pipeline" and is illustrated in Figure 1. As the sequence data flows through the pipeline the number of entries decreases but the quality of information accumulated increases. The annotation pipeline can be viewed as an annotation transformation process.

Secondary databases are often actively "curated". Forming, updating and ensuring the consistency of annotation is one of the primary tasks of the database curators and is extremely labour intensive. For example, forming a PRINTS entry takes several days. Combining this with advances in sequencing technology it becomes clear that annotating all of this data by hand is no longer feasible. At the same time, the increase in the amount of data and the number of tools with which to operate on the data has made analysis difficult for the non-specialist, resulting in a greater requirement for accurate curation. To resolve this contradiction, there is a pressing need for automation within the annotation process.

The most serious difficulty in creating tools that support the curator is that the sources of information are semi-structured database entries or the primary scientific literature. Neither of these sources are readily machine processable. The mining and generation of semi-structured textual annotations are thus the focal points of an annotators' assistant.

There have been a number of previous attempts to develop expert systems that address this requirement, including systems such as GeneQuiz [12], MAGPIE [11], PEDANT [10] and EditToTrembl [14]. These have generally focused on the lower level annotation. In many cases rather than distilling the results from a variety of techniques, the functional designation of a sequence solely rests with the top best match of similarity tools such as FASTA [15] or BLAST [1]. This approach has obvious drawbacks and may lead to erroneous annotation of data.

These tools all operate on the structured parts of the annotation found in the primary databases, or the results of the

KEYWORD (Score)	List of Proteins Containing the Keyword.
prion (3.7e-24)	PRIO_BOVIN, PRIO_CHICK, PRIO_HUMAN, PRIO_MESAU, PRIO_RAT, PRIO_SHEEP
kuru (4.3e-20)	PRIO_BOVIN, PRIO_HUMAN, PRIO_MESAU, PRIO_MOUSE, PRIO_RAT, PRIO_SHEEP
tme (4.3e-20)	PRIO_BOVIN, PRIO_HUMAN, PRIO_MESAU, PRIO_MOUSE, PRIO_RAT, PRIO_SHEEP
gerstmann straussler syndrome (4.3e-20)	PRIO_BOVIN, PRIO_HUMAN, PRIO_MESAU, PRIO_MOUSE, PRIO_RAT, PRIO_SHEEP
bse (4.3e-20)	PRIO_BOVIN, PRIO_HUMAN, PRIO_MESAU, PRIO_MOUSE, PRIO_RAT, PRIO_SHEEP
creutzfeldt disease	PRIO_BOVIN, PRIO_HUMAN, PRIO_MESAU, PRIO_MOUSE, PRIO_RAT, PRIO_SHEEP
spongiform encephalopathy (4.3e-20)	PRIO_BOVIN, PRIO_HUMAN, PRIO_MESAU, PRIO_MOUSE, PRIO_RAT, PRIO_SHEEP

Table 1. Example output from the "Protein Annotator Assistant" for the prion family. It contains a set of keywords with related E-values, but no context

analysis tools. There are also a number of tools that operate on free or semi-structured text data. For instance, AbXtract operates on literature abstracts [2], Easy operates on the output of several search tools [17] and The Protein Annotators Assistant make use of the semi-structured and controlled text of SWISS-PROT [22]. All three operate around a keyword approach although the former two attempt to retain some of the context of these words. Ultimately these tools provide limited information. Keyword extraction is important and useful but only forms a small part of the annotation pipeline (Figure 1), and therefore these tools are of limited use for the specific task of generating protein annotation. For example, Table 1 shows the results of the "Protein Annotator's Assistant" for the prion protein family. Albeit a useful list, it would be impossible to generate the report in Figure 2 from such a set of keywords.

In this paper we describe the PRECIS system (Protein Reports Engineered From Concise Information in SWISS-PROT), which has been designed to address many of these issues. We have examined the processes that are used by human annotators to generate PRINTS entries, and developed software which directly reflects this process. The results that it provides move beyond keyword lists in an attempt to offer more comprehensive reports on protein structure, function and association with disease, in a format that is English-like. PRECIS uses simple techniques, with mappings between the SWISS-PROT and PRINTS annotation markup. It also encodes certain heuristics used by the expert curator when filtering and deriving from these data sources. Although designed with PRINTS annotation in mind, we feel this approach has wider implications.

2 The annotation process

In implementing PRECIS we wanted to produce an concise report in a English-like style. As an ideal model we used existing PRINTS entries, as seen in Figure 2.

PRINTS entries contain a variety of high level annotation describing a protein family in terms of its function, disease-associations and so on. The main source of information from which to gather this data was SWISS-PROT. A SWISS-PROT entry carries much of the data required for PRINTS. A PRINTS entry, however, has a significantly different format, and organisation, and represents the combined data from a number of SWISS-PROT entries. There are other data sources which are routinely used during the production of a PRINTS entry. Mostly the primary literature is used, often by following the bibliographic references that are provided within SWISS-PROT, but a variety of other databases such as PDB [9] are also used.

We therefore have a variety of different data sources, and within that a variety of different data-types, including both structured, semi-structured, and free text based data. There are several ways in which we can extract information from this data.

- *Database structure* Many of the primary and secondary databases have a relatively sparse data model. They are most commonly represented using two letter codes at the beginning of each line, although there are increasing moves toward use of relational [5] or XML based schema [3]. There is often additional structure expressed in other syntaxes. For instance, database cross references (references to other databases) are often highly structured, whilst in SWISS-PROT the "comment" lines are themselves semi-structured, providing metadata about the free text. We discuss various methods in which this implicit metadata can be exploited.

- *Words.* Much of the available data is in the form of free text. Many databases also link to literature resources and if this is included there is a large amount of textual information. In some cases it should be possible to directly extract information from these resources. Tools such as Easy, the Protein Annotators Assistant, and AbXtract already attempt this. This task is simplified for those parts of the free text which use controlled vocabularies such as the Gene Ontology [4].

- *Domain knowledge.* One of the main advantages of manually annotated databases is that the annotation has been formed in light of a large amount of biological knowledge or "domain knowledge". Currently there are two automatable ways in which we exploit this. Firstly knowledge can be built implicitly into the software, when for instance defining the rules that are being used to operate on the data. Secondly we can explicitly use knowledge represented as an ontology[20], such as that used in TAMBIS [18] or the Gene Ontology [4]. We could also make use of a semi-automated

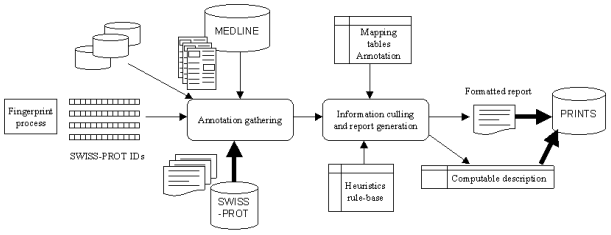

Figure 3. The phases of PRECIS activity. Given a list of ID's, full entries are retrieved from SWISS-PROT. Information from specific fields is culled for further processing, with rules and heuristics being applied in order to generate the final report.

annotation pipeline with interactive human intervention at certain defined points.

During its operation, PRECIS makes use of all of these different methods to varying degrees. The PRECIS software itself consists of several different modules that perform the work-flow shown in Figure 3. These phases reflect the process which is used by human annotators to generate PRINTS entries:

- *Data gathering phase.* PRECIS is given a series of SWISS-PROT ID codes, which are used to gather information from a variety of databases. Information from these provide the basic data on which PRECIS operates. Medline is used to extract abstracts and titles of references, that are scanned for the appearance of certain key phrases.

- *Annotation extraction and synthesis phase*. Annotation is extracted from these data sets by a series of heuristics and mappings from the metadata embedded in SWISS-PROT. Thus we are exploiting the implicit structure in the semi-structured SWISS-PROT annotations. This phase consists of a series of filters through which different parts of the gathered data pass during the synthesis of the final report.

During the extraction and synthesis phase, some of the filters are used repeatedly. These include:-

- *Ranking.* Where large amounts of data are available there is a need to rank this data and only include "the best". Frequency analysis is used to determine how often a piece of data is retrieved. As much of the data is text, basic Information Retrieval techniques are used to gather term frequencies [6]. Weights are also given to data depending on its type, source and rarity. Extra weighting is also given to text dealing with disease, or structural information, as defined by the SWISS-PROT comment field metadata or a "subject keyword heuristic" (see Section 2.1). Data weighted in this way may refer to a single protein rather than the family, so the SWISS-PROT accession number is attached to the annotation. Other data types are subjected to "majority voting", where the common terms will propagate to the PRECIS annotation report.

- *Redundancy checks.* As PRECIS gathers results from several SWISS-PROT entries, and these entries are not independent, much of the information is repeated. Redundancy checks are used for almost all of the sections. Clearly this process is much more straightforward when the data refers to a unique identifier and is more challenging when the data is free text (Section 3).

- *Heuristics.* As well as these generally useful filters, there are also some heuristics that are specific to a single data type. These are explained in further detail in Section 2.1.

The PRECIS pipeline consists of processes that are largely independent of each other. At the current time these processes run serially, but parallelisation would provide an obvious performance enhancement. Only a single "conditional" process exists, which is the decision as to whether the proteins being operated on form part of a domain, a family or a super-family. This has an impact on later parts of the pipeline, for instance in the formation of the PRECIS description line.

2.1 Heuristics

PRECIS uses a variety of heuristics during its operation. These are :-

- *SWISS-PROT ID heuristic.* One of the key decisions made by the PRECIS software is whether the input SWISS-PROT ID's are members of a family, superfamily or domain. The SWISS-PROT ID codes are meant to provide a "human readable" equivalent to accession numbers. The first half of the ID denotes the protein type, whilst the second half the species. For example PRIO_HUMAN, PRIO_BOVIN, PRIO_SHEEP, are all prion proteins in the different species. By analysing the first part of the ID it is possible to determine whether sequences are closely related or not. Within PRECIS, if more than 75% of the roots are identical, the proteins are assumed to be part of a family.

- *Subject keyword heuristic.* Information such as protein structure information is often associated with certain "key-phrases", such as "by NMR" or "X-ray crystallography". We use the appearance of these phrases in the subject line of Medline references to weight these references.

- *Preferred links.* Some database cross-references are considered to be more useful than others. For instance, PDB[9] links are preferred over HSSP [16]. This information is used to order the cross-references.

- *Date Priority* Newer references are preferred to older ones.

These heuristics can be considered as a rule base. At the current time these rules are implicit within the scripts that comprise PRECIS (see Section 2.2). We aim to abstract these rules away into a more explicit representation in later versions of PRECIS.

2.2 Implementation

Scripting languages allow fast and easy parsing of (semi)structured text-based data resources. We prototyped the software using a combination of perl and awk. The current software can rapidly parse and analyse hundreds of SWISS-PROT entries in several minutes on a desktop PC or SGI workstation.

2.3 An examination of PRECIS output

Example output from PRECIS is shown in Figure 4, which illustrates the report generated for the prion protein family. Thirty-two sequence ID's were provided to the program, and their entries retrieved from SWISS-PROT. The full data-set, occupying 40 pages of text was then distilled into a 1.5-page report.

The title of the report is "Major prion protein precursor (PRP)", which was the most frequent description (DE) occurring in the 32 SWISS-PROT entries. Database cross-references are provided for PRINTS, PROSITE, Pfam, InterPro, PDB, SCOP, CATH and MIM (from the SWISS-PROT "DR" lines). Five literature references are given, the last relating to structure determination.

Within the body of the annotation, we find information concerning the function of the protein (currently unknown!) ("CC" Function sub-field), and the diseases with which it is associated ("CC" Disease sub-field). Each paragraph reported is assigned its relevant protein ID and accession number, so that it is possible to trace the provenance of this information.

Following the disease information, we learn about the tendency of the protein to aggregate into polymeric rods ("CC" Subunit sub-field), and discover that the structure of the murine protein has been determined by means of NMR spectroscopy. Finally, the report indicates that the sequences belong to the prion family ("CC" Similarity sub-field), and 9 keywords are provided ("KW" line).

3 Discussion

For the PRINTS database curators, PRECIS has the potential to be a simple but effective tool. In particular, it significantly reduces the time and manual burdens inherent in the process of writing annotation for protein families, as this requires deriving consensus annotation from perhaps tens or hundreds of representatives from the matched set. Use of similar techniques will be more generally applicable.

The use of SWISS-PROT allows us to directly exploit both the in-built structure of its entries and the richness of information already incorporated by teams of annotators. It would be possible to extend PRECIS to incorporate other sources of data. The usefulness of this procedure would be dependant on the amount of structure within these databases. It would for instance perform poorly on data sources which are mostly unstructured free text. This approach is, however, clearly limited by the quality and extent of annotation available. If there is little existing annotation, PRECIS will at best provide some literature references, database cross-references and keywords, although even this is useful as links to literature references make the retrieval of further information relatively straightforward. Also if there are consistent errors in the SWISS-PROT annotation, PRECIS will inherit them, although random errors will be filtered out as data is drawn from multiple entries.

The reports are English-like, as they largely re-use existing human annotation, but consequently exhibit the rather clipped, note-like style typical of SWISS-PROT entries. Although highly informative, the result is inevitably not the same as would be produced by an annotator working from scratch and with resources other than SWISS-PROT, as illustrated in Figure 2. In comparison with the results in Figure 4, the text falls into coherent paragraphs and contains biological details that are either not available within, or not common to, the majority of SWISS-PROT entries. Nevertheless, the automatically-generated result contains not only more, but also more up-to-date, database cross-references, literature references and disease-association information. We therefore believe that PRECIS represents an important step towards the development of more intelligent knowledge-based automatic annotation tools.

PRECIS is a first step and there are many limitations. Although the example illustrated in Figure 4 is promising, many sets of data are less amenable to analysis with this pipeline. Some of the current pipeline tends to generate duplications for instance, particularly where the data

a

```
Major prion protein precursor (PRP)

PRINTS; PR00341 PRION
PROSITE; PS00291 PRION_1; PS00706 PRION_2
PFAM; PF00377 prion
INTERPRO; IPR000817
PDB; 1B10; 1AG2
SCOP; 1B10; 1AG2
CATH; 1B10; 1AG2
MIM; 176640; 123400; 137440; 245300; 600072
```

b

```
1. CERVENAKOVA, L., BROWN, P., GOLDFARB, L.G., NAGLE, J., PETTRONE, K., RUBENSTEIN, R., DUBNICK,
M., GIBBS, C.J. AND GAJDUSEK, D.C.
Infectious amyloid precursor gene sequences in primates used for experimental transmission of
human spongiform encephalopathy.
PROC.NATL.ACAD.SCI.USA 91 12159-12162 (1994).

2. LOWENSTEIN, D.H., BUTLER, D.A., WESTAWAY, D., MCKINLEY, M.P., DEARMOND, S.J. AND PRUSINER, S.B.
Three hamster species with different scrapie incubation times and neuropathological features
encode distinct prion proteins.
MOL.CELL.BIOL. 10 1153-1163 (1990).

3. KALUZ, S., KALUZOVA, M. AND FLINT, A.P.F.
Sequencing analysis of prion genes from red deer and camel.
GENE 199 283-286 (1997).

4. SCHATZL, H.M., DACOSTA, M., TAYLOR, L., COHEN, F.E. AND PRUSINER, S.B.
Prion protein gene variation among primates.
J.MOL.BIOL. 245 362-374 (1995).

5. RIEK, R., HORNEMANN, S., WIDER, G., GLOCKSHUBER, R. AND WUETHRICH, K.
NMR characterization of the full-length recombinant murine prion protein, mPrP(23-231).
FEBS LETT. 413 282-288 (1997).
```

c

```
The function of prp is not known. Prp is encoded in the host genome and is expressed both in
normal and infected cells.

Attached to the membrane by a gpi-anchor.
```

d

```
(PRIO_HUMAN; P04156): Prp is found in high quantity in the brain of humans and animals infected
with neurodegenerative diseases known as transmissible spongiform encephalopathies or prion
diseases, like: creutzfeldt-jakob disease (cjd), gerstmann-straussler syndrome (gss), fatal
familial insomnia (ffi) and kuru in humans; scrapie in sheep and goat; bovine spongiform
encephalopathy (bse) in cattle; transmissible mink encephalopathy (tme); chronic wasting disease
(cwd) of mule deer and elk; feline spongiform encephalopathy (fse) in cats and exotic ungulate
encephalopathy (eue) in nyala and greater kudu. The prion diseases illustrate three manifestations
of cns degeneration: (1) infectious (2) sporadic and (3) dominantly inherited forms. Tme, cwd,
bse, fse, eue are all thought to occur after consumption of prion-infected foodstuffs.

(PRIO_HUMAN; P04156): Kuru is transmitted during ritualistic cannibalism, among natives of the new
guinea highlands. Patients exhibit various movement disorders like cerebellar abnormalities,
rigidity of the limbs, and clonus. Emotional lability is present, and dementia is conspicuously
absent. Death usually occurs from 3 to 12 month after onset.

(PRIO_SHEEP; P23907): Polymorphism at position 171 may be related to the alleles of scrapie
incubation-control (sic) gene in this species.
```

e

```
Prp has a tendency to aggregate yielding polymers called "rods".
```

f

```
The structure has been determined, e.g. "NMR characterization of the full-length recombinant
murine prion protein, mPrP(23-231)" [5].
```

g

```
Belongs to the prion family.
```

h

```
Keywords: GPI-anchor; Repeat; Signal; Prion; Brain; Glycoprotein; Polymorphism; Disease mutation;
3D-structure.
```

Figure 4. Example PRECIS output for the PRION protein family. For conciseness 7 disease related descriptions have been removed. a) "low-level" annotation. b) Literature cross-references c) Descriptive information found in several entries d) Disease related information, with sequence provenance indicated e) as c f) Structural information g) Family membership h) Keywords

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 *pi turnover.*

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.*

Figure 5. The difficulty of redundancy detection within free text. Similar but not identical "function" comments for the various members of the muscarinic acetylcholine family.

is free text. For example, duplications from identical comment fields can be easily removed. Its more difficult where there are small variations, as illustrated in Figure 5. Such small differences can be detected but a lot of domain knowledge is required to differentiate between spelling corrections, word order alterations and biologically significant additions. For example, the sentences, "Primary transducing effect is pi turnover" and "Primary transducing effect is inhibition of adenylate cyclase" and "Primary transducing effect is adenylate cyclase inhibition" are all similar. It should be possible to identify the word order changes between two of these statements, but the difference between "pi turnover" and "adenylate cyclase inhibition" requires human interpretation. At the current time we have accepted duplication rather than throwing away potentially valuable data. Selective use of human intervention at this point could improve this situation.

Although PRECIS has been designed to reflect the pipeline used by human annotators of PRINTS, the most significant difference is that little or no direct use is made of primary literature resources. The decision not to use directly Medline abstracts was a pragmatic one, based on some early experiments. They are free text which makes extraction of the information difficult, although some progress has been made in this area [13]. However, typically the abstracts alone are too short to provide informative, generic reports on protein families.

3.1 Future work

At the current time, PRECIS provides a useful and informative report on the basis of a number of SWISS-PROT ID's, and is therefore already a useful tool. Various problems, however, remain to be addressed. For example, we are looking into ways to reduce the text duplications described earlier. Notwithstanding these issues, there are several important future developments that we plan to explore. One of the most important applications will be to exploit PRECIS to provide annotation for protein family databases, either fully automatically, or in a decision-support fashion, to assist human curators. Databases such as PROSITE and PRINTS, which add value to the information they contain by providing extensive, hand-crafted annotation, could benefit from such an approach. We are currently exploring the use of PRECIS to add annotation to an automatically-derived supplement to PRINTS (prePRINTS). This will ensure that automatically-generated fingerprints have at least some level of annotation associated with them. This application will require the development of additional tools to assist human annotators in the annotation-gathering process, both as a means to provide editorial/quality control, and to allow the addition of more extensive annotation during the migration process from prePRINTS to PRINTS itself.

Other important and ongoing developments include:- Firstly the introduction of a formally structured metadata layer to PRECIS output. PRECIS was developed as a proof-of-concept. If the tool is to be more widely useful, then it is important to generate information in a form that is not only human readable but is also machine processable. We hope that the introduction of a formal structured metadata layer within the PRECIS output will ease the third party use of the data generated. It would be an advantage to do this without the problems that we have faced in extracting information from sources which are essentially free text. Support for human readable output in English would be relatively easy to generate from such structured data. Our greater vision is the implementation of PRECIS as part of workflow, using a workflow language and enactment system (see [19]).

Secondly we wish to investigate methods to generate more informative reference lists:- the current method of frequency analysis is crude and will not always guarantee that the most relevant references are included. And thirdly we wish to provide greater support for provenance for the data presented. Currently very little information about the source of particular data is given. It is limited to references for SWISS-PROT ID's for disease association or structural information (see Figure 4).

Finally PRECIS reports represent a distillation of data from a variety of different sources. We feel that the results are of more general use beyond forming prePRINTS entries. We have therefore coupled PRECIS to the BLAST database search tool. For a given query sequence, BLAST returns a series of proteins ranked according to their similarity. In many cases, the top ranking matches are all to a single protein family. To discover more about the nature of this family, the individual database entries must be examined. PRECIS has now been configured to do this automatically. The system accepts as input a single query sequence in FastA format. A BLAST search is performed, and SWISS-PROT ID's found above a user-defined E-value

cut-off are extracted. PRECIS then provides the user with a formatted report describing the selected protein family, as illustrated previously in Figure 4.

4 Conclusions

Tools both to assist and to automate the process of sequence annotation are sorely needed. To address this need, we have developed PRECIS, which generates PRINTS-like reports from sets of related SWISS-PROT entries. The results, although dependent on SWISS-PROT, provide a useful first step in producing a) a fully-automatic annotation tool and b) a process flow framework that a human annotator can use to gather more detailed information. The tool has considerable potential to assist curators of protein family databases, where it would have the dual advantage of reducing current manual burdens, and of creating information in a format that is consistent, computer readable and readily update-able.

5 Acknowledgements

PWL and JRR were funded under the ESPRC/BBSRC Bioinformatic Programme (Grant number: BIF/10507). RDS was supported by BBSRC/EPSRC grant 4/B1012090.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, Oct 5 1990.

[2] M. A. Andrade and A. Valencia. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600–7, 1998.

[3] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist, and E. M. Zdobnov. InterPro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, 16(12):1145–50, Dec. 2000.

[4] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, May 2000.

[5] T. K. Attwood, M. D. Croning, D. R. Flower, A. P. Lewis, J. E. Mabey, P. Scordis, J. N. Selley, and W. Wright. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res*, 28(1):225–7, Jan 1 2000.

[6] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc, 1999.

[7] A. Bairoch and R. Apweiler. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res*, 28(1):45–8, Jan 1 2000.

[8] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. GenBank. *Nucleic Acids Res*, 28(1):15–8, Jan 1 2000.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42, Jan 1 2000.

[10] D. Frishman, K. Albermann, J. Hani, K. Heumann, A. Metanomski, A. Zollner, and H. W. Mewes. Functional and structural genomics using PEDANT. *Bioinformatics*, 17(1):44–57, Jan. 2001.

[11] T. Gaasterland and C. W. Sensen. Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. *Biochimie*, 78(5):302–10, 1996.

[12] S. Hoersch, C. Leroy, N. P. Brown, M. A. Andrade, and C. Sander. The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem Sci*, 25(1):33–5, Jan. 2000.

[13] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. *Pac Symp Biocomput*, pages 505–16., 2000.

[14] S. Moller, U. Leser, W. Fleischmann, and R. Apweiler. EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics*, 15(3):219–27, Mar. 1999.

[15] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–8, Apr. 1988.

[16] R. Schneider, A. de Daruvar, and C. Sander. The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res*, 25(1):226–30, Jan 1 1997.

[17] J. N. Selley, J. Swift, and T. K. Attwood. EASY–an Expert Analysis SYstem for interpreting database search outputs. *Bioinformatics*, 17(1):105–6, Jan. 2001.

[18] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. W. Paton, C. A. Goble, and A. Brass. TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16(2):184–5, Feb. 2000.

[19] R. Stevens, C. Goble, P. Baker, and A. Brass. A classification of tasks in bioinformatics. *Bioinformatics*, 17(2):180–8, Feb. 2001.

[20] R. Stevens, C. Goble, and S. Bechhofer. Ontology-based Knowledge Representation for Bioinformatics. *Briefings in Bioinformatics*, 1(4):398–416, November 2000.

[21] G. Stoesser, W. Baker, A. van den Broek, E. Camon, M. Garcia-Pastor, C. Kanz, T. Kulikova, V. Lombard, R. Lopez, H. Parkinson, N. Redaschi, P. Sterk, P. Stoehr, and M. A. Tuli. The EMBL nucleotide sequence database. *Nucleic Acids Res*, 29(1):17–21, Jan 1 2001.

[22] M. J. Wise. Protein Annotators' assistant. *Trends Biochem Sci*, 25(5):252–3, May 2000.