Protein Family Databases: Text-mining & Annotation

Anna Divoli & Teresa K. Attwood
Faculty of Life Sciences and School of Computer Science, University of Manchester, Manchester M13 9PT, UK

Abstract
In recent years the text-mining and bioinformatics communities have worked together to address the problem of finding pertinent information within the vast volume of biomedical literature. In this paper, we focus on a specific bio text-mining issue: the annotation of protein family databases. We describe an evaluation performed on an in-house text-mining tool, we report how successful it was in reproducing several examples of human-created annotation and discuss the problems and challenges we encountered.

1 Protein Family Databases & Annotation Needs

Protein family databases are used by bioinformaticians and biologists to provide detailed knowledge about proteins, their inter-relationships, and their structures and functions. But they are not just static 'knowledgebases' – one of their most important applications is a diagnostic one, *i.e.,* in the characterisation of newly determined sequences. The starting point for creating entries to populate protein family databases is usually a multiple alignment of a set of related amino acid sequences. The closeness of the relationship between such sequences is evident from the patterns of residues, or residue groups, that are shared between them – the more residues shared, or conserved, the greater the confidence we may have that the sequences are evolutionarily related (*i.e.*, that they are homologous). These patterns of conservation can then be used to recognise and identify uncharacterised sequences from newly sequenced genomes. To facilitate this diagnostic process, each family-specific pattern needs to be annotated with biological information prior to deposition in a database – a match to a particular family may then be understood in terms of its evolution, possible function, its disease associations, and so on. In most family databases, such annotation is written in an unstructured free-text format: most of the cutting-edge information needed to derive this annotation is distilled manually from the biomedical literature; some of the more basic information is gleaned from standard text-books; and some is inferred directly from observation of the sequence alignment or from prior knowledge of the protein's structure. Overall, annotation is thus a complex, labour-intensive task, and today is still the major obstacle to the growth of protein family databases.

In recent years, the annotation bottleneck has driven the development of tools to help automate information retrieval (IR) and extraction from the literature. Much of this work has focused on identifying protein-protein interactions (Chen and Sharp, 2004; Jenssen *et al.,* 2001); some more recent approaches have concentrated on annotating databases of model organisms (Muller *et al.,* 2004) and of protein sequences (Camon *et al.,* 2004; Dobrokhotov *et al.,* 2005). But mining the literature for specific interactions and developing annotation tools for different databases are very different activities with very different needs, and it is vital to appreciate these differences from the outset. To give a trivial example, unlike other annotators, those of protein family databases often prefer review papers to experimental ones, as they provide more comprehensive family-related information. But no matter what the source, above all, the curator needs the information to be relevant and correct (high precision), and is not usually concerned about finding every published paper in the field (high recall). Even among different protein family databases, however, the requirements for, and standards of, annotation may differ considerably.

2 PRINTS & BioIE

PRINTS (Attwood *et al.,* 2003) is a protein fingerprint database that provides diagnostic signatures for protein families. This resource has been developed in a hierarchical manner, accommodating entries at the level of families, superfamilies and domains. Each of these different 'views' has specific annotation requirements. Wherever possible, each view includes information on the structure, function, family and disease relationships of the proteins under investigation (this information is provided in the form of human-readable, free-text paragraphs); however, there are subtle differences in the way this information relates to each of the constituent proteins. For example, in the context of a superfamily, each protein will have the same general architecture (*e.g.,* all G protein-coupled receptors (GPCRs) share a 7-transmembrane alpha-helical bundle fold); by contrast, in the context of a domain family, each domain will have the same fold, but the parent protein is likely to have a very different overall structure, of which the domain is just a part (*e.g.,* an 80-residue kringle domain embedded within an 800-residue plasminogen sequence). Any attempt to automate the process of annotation must therefore take on board these different views of the family hierarchy.

PRINTS is also a member of the InterPro (Mulder *et al.,* 2005) integrated family resource, where annotation is either inherited directly from its source databases (*e.g.*, PROSITE and PRINTS), or must be created from scratch when a source database provides little free-text annotation of its own (*e.g.*, Pfam). In undertaking this work, our hope was to help lift the burden of manual annotation for PRINTS curators, and hopefully also for those of InterPro.

BioIE (Divoli and Attwood, 2005) is a rule-based system that has been designed as a decision-support tool to help protein family database curators with the task of annotation. Its main role is to extract informative sentences (categorised according to protein function, structure, related diseases and therapeutic compounds, localisation, and familial relationships) from MEDLINE abstracts or from uploaded text, by using manually-defined templates and rules. It is highly interactive, providing the option to specify keywords prior to sentence extraction to allow queries to be tailored to family-specific user interests.

3 The Study

In order to evaluate BioIE, to identify its weaknesses, and hence to improve its usefulness as an annotation assistant, we performed an evaluation using a subset of PRINTS (termed miniPRINTS) as a test-set. miniPRINTS comprises a representative sample of the database selected specifically for the purpose of software evaluation. We analysed the annotation of its 20 families and endeavoured to identify particular sentences in the source text from which given statements came. This was done in 2 ways: first, using BioIE, and second, for the purpose of comparison, manually from electronically available sources – here, the idea was to mimic an ideal super tool that had cognition and domain knowledge. Two sources were used: (i) the full papers cited in the database entries; (ii) other relevant abstracts from PubMed (up to the date of the miniPRINTS entry).

As anticipated, not all the information in the annotation could be identified in the literature. We therefore looked for syntactic patterns that might further improve template selection, and we investigated the importance of these patterns for weight allocation and better ranking of extracted sentences; as these were responsible for only a few missed statements, we also investigated why some types of information were not returned from the literature and tried to evaluate the scale of the problem.

4 Results & Discussion

During source retrieval, finding some of the papers in electronic form (especially early ones) and their successful conversion to text was the main challenge (information from books remained inaccessible). To obtain the relevant abstracts, the challenge was to use the right query terms (names, combination of names, appropriate Boolean operators and search fields), which varied depending on the type of family. BioIE does not automate this IR part; it only provides an embedded PubMed query service, so the IR was done manually as a database curator would do normally, using PubMed. In doing this, some interesting tendencies were observed. For example, when retrieving domain names, we found that it was better to seek the name both in titles and abstracts, rather than just in titles. This is because in PRINTS, some domain annotation is an artefact of the annotation process: *i.e.*, representative sequences are used to find information first in Swiss-Prot and then from linked MEDLINE abstracts – much of the retrieved information therefore refers to the parent protein and not to the domain of interest; hence the domain name is less likely to occur in the title.

Once the relevant text was acquired, we used BioIE to extract pertinent sentences and to match them to the annotation statements. On average, 50% of the statements were matched (Table A). In the first instance, we manually investigated the "unmatched" annotation statements for clues of syntactic patterns that might help to improve our system. Indeed, some recurring patterns were observed in some of the originally missed information (*e.g.*, we found "abundant in" to occur frequently and hence included it in the localisation template-set, and we updated our regular expressions to include variations in the use of units (40kDa, 40 kDa, 40-kDa, *etc.*)), but these were only responsible for 3.6% of missed statements; most (85.2%) were a consequence of particular statements simply not being available in the source text; 11.2% resulted from other problems that the idiosyncrasies of biomedical text present.

As just mentioned, more than 85% of unmatched annotation statements resulted from them simply not being in the electronically available source. Part of this problem was probably because some sources were not cited, or indeed that some statements may not actually have originated from the literature. For example, in some cases, it appeared that the annotator had summarised/interpreted from a long textual description or from a table or figure (*e.g.,* the miniPRINTS statement *"The electron flow in the sulphite oxidase reaction is sulphite -> molybdopterin -> cytochrome b5 -> cytochrome c" was probably derived from a figure)*. In other cases, the statements could be considered "obvious" textbook knowledge that was used to augment the source-text, in order to create a human-readable story rather than just a string of facts (*e.g.,* in "*The Hb molecule exists as a tetramer, typically of two alpha- and two beta-globin chains, which form a well-defined*

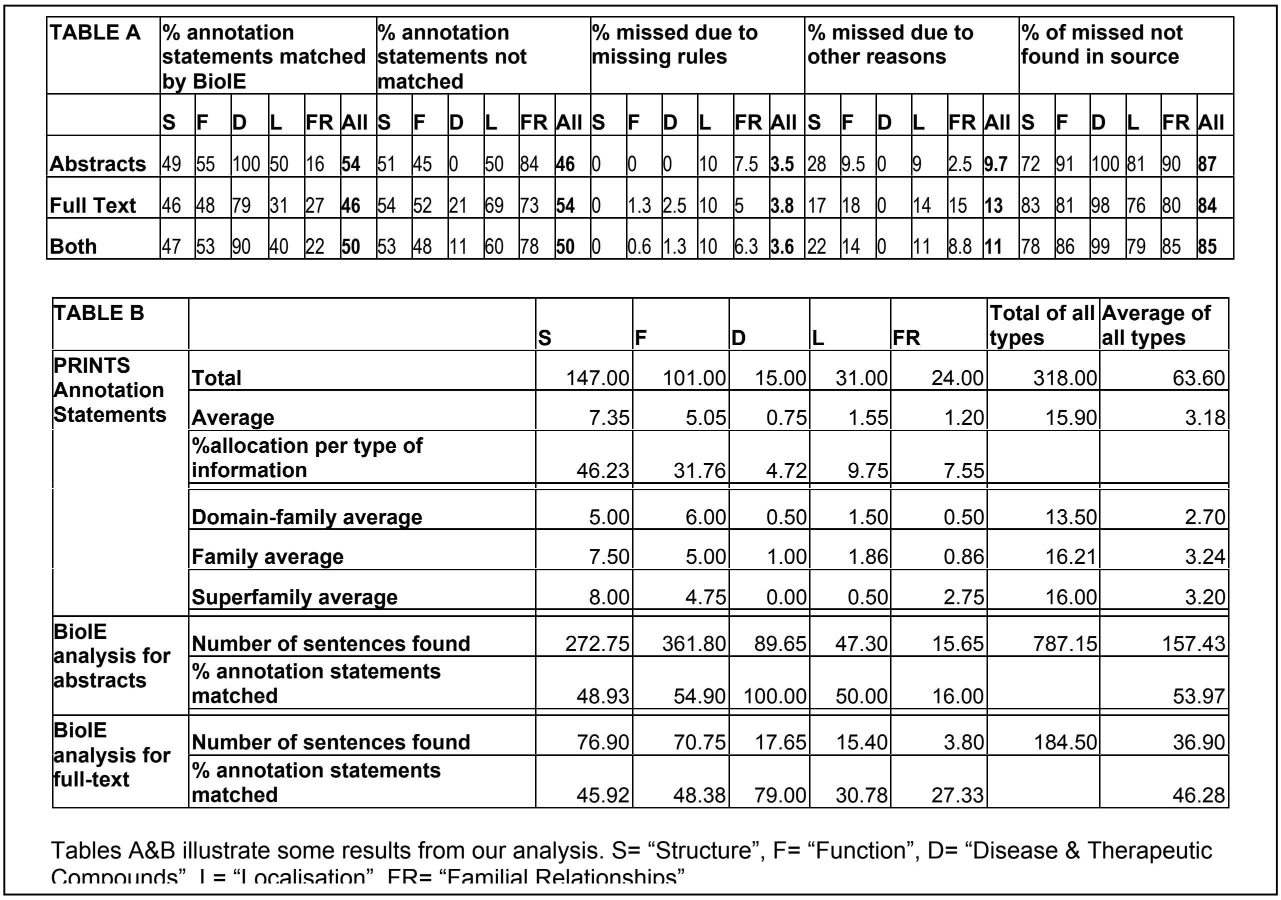

TABLE A	% annotation statements matched by BioIE						% annotation statements not matched						% missed due to missing rules						% missed due to other reasons						% of missed not found in source					
	S	F	D	L	FR	All	S	F	D	L	FR	All	S	F	D	L	FR	All	S	F	D	L	FR	All	S	F	D	L	FR	All
Abstracts	49	55	100	50	16	**54**	51	45	0	50	84	**46**	0	0	0	10	7.5	**3.5**	28	9.5	0	9	2.5	**9.7**	72	91	100	81	90	**87**
Full Text	46	48	79	31	27	**46**	54	52	21	69	73	**54**	0	1.3	2.5	10	5	**3.8**	17	18	0	14	15	**13**	83	81	98	76	80	**84**
Both	47	53	90	40	22	**50**	53	48	11	60	78	**50**	0	0.6	1.3	10	6.3	**3.6**	22	14	0	11	8.8	**11**	78	86	99	79	85	**85**

TABLE B		S	F	D	L	FR	Total of all types	Average of all types
PRINTS Annotation Statements	Total	147.00	101.00	15.00	31.00	24.00	318.00	63.60
	Average	7.35	5.05	0.75	1.55	1.20	15.90	3.18
	%allocation per type of information	46.23	31.76	4.72	9.75	7.55		
	Domain-family average	5.00	6.00	0.50	1.50	0.50	13.50	2.70
	Family average	7.50	5.00	1.00	1.86	0.86	16.21	3.24
	Superfamily average	8.00	4.75	0.00	0.50	2.75	16.00	3.20
BioIE analysis for abstracts	Number of sentences found	272.75	361.80	89.65	47.30	15.65	787.15	157.43
	% annotation statements matched	48.93	54.90	100.00	50.00	16.00		53.97
BioIE analysis for full-text	Number of sentences found	76.90	70.75	17.65	15.40	3.80	184.50	36.90
	% annotation statements matched	45.92	48.38	79.00	30.78	27.33		46.28

Tables A&B illustrate some results from our analysis. S= “Structure”, F= “Function”, D= “Disease & Therapeutic Compounds”, L= “Localisation”, FR= “Familial Relationships”

quaternary structure,” the concept “quaternary structure” does not appear anywhere in the retrieved literature). Similarly, finding structural descriptions of well-characterised motifs, such as the zinc-finger, was very difficult (in recent abstracts at least) as this is also basic “textbook” knowledge. Other types of information were missed because they were context related and did not include the name of the family being annotated (*e.g.,* in “*Vision is effected through the absorption of a photon by the chromo-phore, which is isomerised to the all-trans form, promoting a conformational change in the protein”* there is no direct reference to opsin, the subject of the annotation; similarly, in *“The activating ligands of the different superfamily members vary widely in structure and character”* there is no mention of the annotation subject, rhodopsin). BioIE has been designed to address such issues and to allow users to simultaneously investigate several entities (the protein name, potential synonyms, the ligands it binds, the domain it contains or whatever); however, in this study, for practical purposes we limited the investigation to information relating to the main entities (*i.e.*, to the protein name itself). Another important mechanism for missing statements is that some are derived from sources other than the literature (*e.g.,* the observation that *“The primary sequences of PrP's from different sources are highly similar”* probably originated from analysis of sequence alignments). Finally, owing to the hierarchical architecture of the database, some information is inherited directly from parent entries (*e.g.*, “*Opsins are the photoreceptors of animal retinas*” is inherited by the rhodopsin entry from its opsins parent). In such cases, it would be worth investigating whether the hierarchy of the database, based on sequence alignment, agrees with that derived from available ontologies.

Considering the 11.2% of unmatched annotation statements, some result from bad/failed text conversion from PDFs. Others are the result of typical anaphora problems (say, where a pronoun is used instead of the entity name) or more subtle anaphora (*e.g.*, “*The structure suggests plausible electron and proton transfer pathways,”* where “structure” is used in place of the entity name).

On analysing the sentence category allocation (Table B), we see that this is different in the annotation, referenced papers and retrieved abstracts – clearly, miniPRINTS annotators include more structure-related sequences. The categories “structure”, “localisation” and “familial relationships” have fewer matches in the literature, suggesting that much of the annotation is a consequence of hierarchical inheritance (from parent entries to their children) and, crucially, much is derived from the annotator’s ability to make interpretations based both on the literature and on observations of alignments. The “disease” category is the only one to be matched 100% in the literature (when the source is abstracts). Although abstracts are not as rich in detailed information as full text, they can nevertheless offer useful biological insights (*e.g.*, the haemoglobin annotation had no disease-related statements because the curator chose structure-related papers, but its involvement in haemolytic anaemia, thalassaemia and polycystic kidney disease was obvious from the abstracts collected in this study).

5. Conclusions

Before attempting to automate processes for augmenting database annotation, it is important to understand the nature of the annotation required for a given database: its content, its patterns and its idiosyncrasies. During this study, among vastly inconsistent standards of annotation, we discovered some consistent tendencies in the types of information selected by annotators for different types of family, and identified some problems/limitations when attempting to semi-automate the process.

In trying to develop a semi-automatic annotation tool, we should not prioritise sentences that are generic, that set the scene, that refer to parent-child or sibling relationships, or that are derived from interpretation of alignments, and so on; the human annotator will contribute these. We should instead focus on sentences that provide biological facts. Future software evaluations should then only be measured against such factual statements.

However, notwithstanding what a human annotator may prefer to include, there is no right way to create annotation, no single accepted set of facts. It can therefore be advantageous to use a decision-support tool to facilitate the process, as this can provide a more objective, up-to-date and comprehensive picture of what is known about a family in the literature.

We found that a rule-based system, given the right rules based on specific domain knowledge, will extract relevant sentences from a given source. The main reason for failing to be able to recreate manual annotation automatically was the difficulty of retrieving the appropriate sources of text.

6. Acknowledgements

This work was funded by BBSRC and Inpharmatica Ltd.

7. References

Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P., Uddin,A. and Zygouri,C. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.* Jan 1;31(1):400-2.

Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res. Jan 1;32(Database issue):D262-6.

Chen,H. and Sharp,B.M. (2004) Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics. Oct 8;5(1):147.

Divoli,A. and Attwood,T.K. February 2, 2005. BioIE: extracting informative sentences from the biomedical literature. Bioinformatics. May 1;21(9):2138-9.

Dobrokhotov,P.B., Goutte,C., Veuthey,A.L. and Gaussier,E. (2005) Assisting medical annotation in Swiss-Prot using statistical classifiers. Int J Med Inform. Mar;74(2-4):317-24.

Jenssen,T.K,, Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. May;28(1):21-8.

Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A. *et al.* (2005) InterPro, progress and status in 2005. Nucleic Acids Res. Jan 1;33(Database issue):D201-5.

Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. Nov;2(11):e309