

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA



ReBIL: Relating Biological Information through Literature

Francisco José Moreira Couto

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

2006

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA



ReBIL: Relating Biological Information through Literature

Francisco José Moreira Couto

DOUTORAMENTO EM INFORMÁTICA
ESPECIALIDADE BIOINFORMÁTICA

2006

Tese orientada pelo Prof. Doutor Mário Jorge Costa Gaspar da Silva
e pelo Prof. Doutor Pedro Maldonado Coutinho

Abstract

Bioinformatics is a new research field that aims at using computer technology to uncover biological knowledge of high relevance to the biotechnology community. An important research topic in Bioinformatics involves the exploration of vast amounts of biological and biomedical scientific literature (BioLiterature). Over the last few decades, text-mining systems have exploited this BioLiterature to reduce the time spent by researchers in its analysis. However, many of these systems rely on manually inserted domain knowledge, which is time-consuming.

This thesis proposes an approach where domain knowledge is automatically acquired from publicly available biological databases, instead of using manually inserted domain knowledge. Based on this approach, innovative methods for retrieval, extraction and validation of information published in BioLiterature were developed and evaluated. The results show that the proposed approach is an efficient alternative to domain knowledge explicitly provided by experts. The new methods were also integrated into a system for automatic annotation of genes and proteins, which was successfully demonstrated in several applications.

Keywords: Bioinformatics, Text mining, Data mining, Gene and protein automatic annotation.

Resumo

A Bioinformática é uma disciplina científica recente, que recorre a meios computacionais para revelar novo conhecimento de grande interesse para a Biotecnologia. Um tópico de investigação importante da Bioinformática é a exploração da vasta quantidade de literatura científica existente sobre Biologia e Biomedicina. Nas últimas décadas, têm sido propostos sistemas de prospecção de texto para identificar informação relevante nestas bases de informação por forma reduzir o tempo despendido a analisá-las. Contudo, grande parte dos métodos de prospecção utilizados baseiam-se em conhecimento específico do domínio inserido manualmente, requerendo por isso muito tempo para a sua criação.

Esta tese propõe uma abordagem que integra conhecimento adquirido automaticamente a partir de bases de dados biológicos, em vez de usar conhecimento específico do domínio inserido manualmente. Esta abordagem permitiu o desenvolvimento de métodos inovadores, cujos resultados de avaliação evidenciaram poder constituir uma alternativa eficiente às baseadas em conhecimento de domínio introduzido por peritos. Os métodos foram integrados num sistema para anotação automática de genes e proteínas, que foi utilizado com sucesso em várias aplicações.

Palavras-chave: Bioinformática, Prospecção de texto, Prospecção de dados, Anotação automática de genes e proteínas.

Resumo Alargado

A Biotecnologia tem como objectivo a produção e utilização aplicada de materiais de natureza biológica. A Bioinformática é por sua vez uma disciplina científica cujo principal objectivo é a produção de conhecimento de interesse para a biotecnologia. Estuda técnicas inovadoras de manipulação, gestão e análise de grandes quantidades de informação biológica, permitindo aos cientistas extrair conhecimento a partir dessa informação. As fronteiras que limitam a variedade das aplicações da Bioinformática são difíceis de identificar, pois esta integra conhecimentos de diversas áreas da ciência, como a Biologia, a Bioquímica, a Biofísica, a Estatística, a Matemática e, naturalmente, a Informática. O factor comum de todas as suas aplicações é o uso de sistemas computacionais no tratamento de informação biológica para a obtenção eficaz de importantes resultados científicos.

O grande interesse pela Bioinformática nos últimos anos deve-se sobretudo à explosão da informação disponível proveniente sobretudo dos esforços de sequenciação parcial ou completa dos genomas de diferentes organismos. Esta informação permitiu o estabelecimento de disciplinas da pós-genómica, associadas ao estudo de processos biológicos relacionados com o genoma, o que gerou ainda mais informação. Para a gerir têm sido criadas diversas bases de dados de grande dimensão. Por exemplo, a base de dados EMBL¹ arma-

¹<http://www.ebi.ac.uk/embl/>

zena as sequências biológicas da maior parte dos genes conhecidos, indicando quando possível informação estrutural e funcional adicional. A estrutura de um gene é representada por uma sequência de ácidos nucleotídicos, que podem ser de quatro tipos. Para os representar recorre-se a um alfabeto genético, constituído por quatro letras que correspondem às quatro alternativas possíveis. A EMBL disponibilizava através da Internet em Dezembro de 2005 cerca de 108 GB de informação representando os resíduos de ácidos nucleotídicos de 64.619.747 sequências, observando-se um crescimento exponencial desde a sua criação. Este valor não conta com a informação descritiva de cada sequência, que é ainda de maior dimensão e de enorme importância.

A gestão destas bases de dados afigurou-se desde cedo como um processo complexo. A ausência de recursos para caracterização dos dados armazenados foi infelizmente acompanhada pela utilização de métodos simplistas de anotação, causadores de muitas das incongruências encontradas presentemente. Em consequência, quando uma equipa de investigação precisa de obter informação sobre uma determinada entidade biológica, procura primeiro toda a informação existente nas bases de dados, e tenta depois, se esta for relevante para o seu estudo, encontrar uma prova que suporte a veracidade dessa mesma informação. Na maior parte dos casos a prova só pode ser encontrada na literatura, o modo tradicional de divulgação do conhecimento científico. Na literatura a informação é expressa de uma forma não estruturada, o que dificulta o seu tratamento automático. Desta forma, muitas das bases de dados são actualizadas directamente ou indirectamente por equipas de peritos cuja função é procurar informação relevante em artigos científicos para anotar os genes ou proteínas. Esta tarefa é bastante complexa devido à grande quantidade e diversidade da informação a analisar. Isto requer um esforço enorme às equipas de peritos, que por vezes geram anotações erradas

quando têm de analisar informação fora da sua especialidade. Para facilitar e normalizar o processo de anotação têm sido desenvolvidas BioOntologias que organizam e descrevem conceitos biológicos e as suas relações. Por exemplo, GO² (Gene Ontology) é uma BioOntologia que tem permitido uniformizar as anotações entre diferentes espécies.

Mesmo usando um elevado número de recursos humanos é quase impossível acompanhar o crescimento de informação gerada diariamente. Por exemplo, em 2003, cerca de 560.000 citações foram adicionadas na base de dados MEDLINE, um repositório de literatura relacionada com Biomedicina e Biologia. Este repositório está disponível através da PubMed³, que em 2005 continha mais de 15 milhões de citações. O PubMed é um dos recursos mais utilizados na área de ciências da vida. Desta forma, a exploração da imensa quantidade de literatura científica sobre Biologia e Biomedicina é uma questão de muito interesse para a comunidade científica, que motiva o desenvolvimento de ferramentas que possam extrair automaticamente informação da literatura, ou que, pelo menos, permitam uma melhor orientação no trabalho das equipas de investigação.

Nas últimas décadas, os sistemas de prospecção de texto têm sido aplicados na identificação de informação relevante contida na literatura, reduzindo desta forma o tempo despendido a analisá-la. Como a literatura científica tem vindo gradualmente a ser disponibilizada na Internet em formato electrónico, o estudo de métodos de prospecção de literatura constituiu-se recentemente como um tópico de investigação muito activo. Estes métodos têm por objectivo identificar e estruturar informação relevante expressa nos textos de publicações científicas para posterior inserção em bases de dados. Estão em curso um grande número de projectos que têm como objectivo o desenvol-

²<http://www.geneontology.org/>

³<http://www.pubmed.org/>

vimento de sistemas de extracção automática de informação da literatura científica para catalogação em bases de dados de informação biológica. O primeiro sistema de extracção de informação da literatura biológica foi desenvolvido por Andrade et al. em 1998. A partir daí tem sido desenvolvidos vários sistemas, mas sempre com níveis de precisão abaixo do que seria desejável. De forma a encontrar os métodos que mais se adequam a esta tarefa, algumas competições internacionais têm sido realizadas:

- *ACM KDD 2002 Cup challenge: bio-text task;*
- *BioCreAtIvE 2004: Critical Assessment of Information Extraction systems in Biology;*
- *TREC 2003 and 2004: genomics track.*

A principal conclusão retirada destas competições foi que os métodos de prospecção de texto utilizados com maior sucesso noutras áreas não obtêm resultados satisfatórios quando aplicados à literatura biológica. Os obstáculos mais difíceis de transpor têm sido: o uso de diferentes nomenclaturas; a heterogeneidade da informação; e existência de diferentes interpretações para o mesmo texto. A prospecção de texto biológico contrasta com o que hoje se alcança noutros domínios, como na identificação automática de entidades mencionadas em textos retirados de jornais noticiosos, onde é já possível alcançar níveis de qualidade equivalentes aos de um perito humano. Contudo, uma grande parte dos métodos de prospecção de literatura biológica baseiam-se em conhecimento do domínio introduzido manualmente sob a forma de regras e padrões que modelam todos os casos possíveis e conjuntos de treino demasiado específicos para serem estendidos a outros domínios. Este tipo de abordagem tem custos muito elevados, pois exige um grande esforço dos

peritos na introdução de conhecimento do domínio do problema a resolver, custos esses que muitas vezes não são compensados pelos resultados obtidos.

O trabalho descrito neste manuscrito propõe uma nova abordagem para a prospecção de literatura biológica que evita o complexo problema do uso de conhecimento de domínio inserido manualmente. Este trabalho advoga em alternativa a integração de conhecimento de domínio adquirido automaticamente a partir de bases de dados biológicos. Para validar a abordagem proposta, foram desenvolvidos métodos para recolha, extracção e validação de informação. Estes métodos foram também integrados no ProFAL (bioProducts Functional Annotation through Literature), um sistema desenvolvido para anotação automática de genes e proteínas.

Para seleccionar documentos relevantes foi desenvolvido o WeBTC (Web Biological Text Classification), um método inovador que permite a classificação de literatura de literatura relacionada com a Biomedicina e a Biologia. O WeBTC recorre a informação extraída a partir de fontes disponíveis na Web com métodos estatísticos de classificação de texto tradicionais. O WeBTC conseguiu aumentar significativamente a precisão (atingindo 100%) dos métodos de classificação de texto tradicionais. Foi submetido ao *ACM KDD 2002 Cup challenge*, onde mostrou ser uma alternativa eficaz para melhorar os resultados obtidos pelos métodos de classificação de texto tradicionais.

Para identificar anotações relevantes na literatura foi desenvolvido o FiGO (Finding Genomic Ontology), um método não supervisionado para identificação de propriedades biológicas organizadas numa BioOntologia em texto não estruturado através do conteúdo de informação de cada palavra. O FiGO não necessita de conjuntos de treino, já que calcula o conteúdo de informação de cada palavra a partir de uma BioOntologia que estrutura os termos. Desta

forma, a utilização do FiGO requer uma intervenção humana mínima. Apesar de ter sido criado para reconhecer termos e não para extrair anotações, FiGO obteve um bom desempenho no *BioCreAtIvE* quando comparado com os métodos utilizados pelos outros participantes. O FiGO demonstrou ser uma técnica eficiente de reconhecimento de termos na literatura, melhorando o desempenho de sistemas automáticos de anotação.

Para validar as anotações identificadas foi desenvolvido o CAC (Correlate the Annotations' Components), um método para eliminar anotações incorrectamente identificadas por sistemas de anotação, com recurso a anotações curadas de estrutura e função semelhantes. O CAC foi aplicado a um conjunto de anotações extraídas automaticamente da literatura. Os resultados mostram que o CAC pode ser usado eficazmente para remover anotações incorrectas geradas automaticamente. O CAC requer pouca ou nenhuma intervenção humana, pois recolhe o conhecimento do domínio a partir de bases de dados públicas.

O ProFAL está a ser usado para anotar enzimas activos sobre os glúcidos ("carbohydrate-active") com informação bibliográfica. Estes enzimas estão classificadas na base de dados CAZy⁴ em várias famílias segundo a sua estrutura modular.

O ProFAL foi igualmente usado para anotar funcionalmente um conjunto de genes relacionados com o desenvolvimento do pólen na *Arabidopsis*. Os genes e as suas descrições funcionais estão armazenados na bases de dados APEG⁵ (*Arabidopsis* Pollen Expressed Genes).

Em colaboração com investigadores do *European Bioinformatics Institute* foi criada a ferramenta GOAnnotator⁶, que usa os termos automaticamente

⁴<http://afmb.cnrs-mrs.fr/CAZY/>

⁵<http://xldb.fc.ul.pt/rebil/tools/apeg/>

⁶<http://xldb.fc.ul.pt/rebil/tools/goa/>

extraídos da literatura pelo ProFAL para facilitar a verificação automática de anotações não curadas de proteínas da bases de dados UniProt.

O sucesso da aplicação do ProFAL a estas três bases de dados para explorar e identificar informação relevante na literatura, demonstra a sua eficácia na anotação de genes e proteínas. Os resultados obtidos pelo ProFAL e pelos métodos desenvolvidos evidenciam que a abordagem proposta é uma alternativa eficiente para o conhecimento de domínio introduzido explicitamente por peritos.

Acknowledgments

I am very grateful to many people who helped and encouraged me while I was undertaking the work described in this thesis. They are inevitably too numerous to name all individually, but I will at least make an attempt.

I would like to express my gratitude to:

- Prof. Mário Silva for his deep commitment to this work, for his constant and exceptional advice and guidance, and for his unstinting and stupendous support and encouragement;
- Prof. Pedro Coutinho for his crucial insight of Bioinformatics, for his invaluable suggestions, and for his essential encouragement;
- Pooja Jain for her excellent evaluation of ProFAL and useful suggestions;
- All the other XLDB-LASIGE group members (Ana Paula Afonso, André Falcão, Bruno Martins, Daniel Gomes, Josefa Jul, Marcirio Chaves, Miguel Costa, Norman Noronha, Nuno Cardoso, Leonardo Andrade and Sérgio Freitas) for their assistance with system administration, and for many useful and wide ranging discussions;
- Everyone in the AFMB-CNRS (Architecture et Fonction des Macro-

molécules Biologiques, Marseille, France) group for an amiable gathering during my stays;

- Everyone in the EBI (European Bioinformatics Institute, Hinxton, UK) with whom I collaborated for their important suggestions and comments, in particular Dietrich Rebholz-Schuhmann for his valuable advice and Emily Dimmer for her assistance in reviewing the abstract of this thesis.
- Everyone in the Departamento de Informática da FCUL for making it such an interesting and stimulating place to work;
- Prof. Arlindo Oliveira for introducing me to Bioinformatics, and for inspiring me to work on Bioinformatics;
- Elsa Abranches for helping me to find researchers working on Bioinformatics before starting my thesis;
- Pedro Fernandes for organising high quality courses and seminars about Bioinformatics at IGC (Instituto Gulbenkian de Ciência, Oeiras, Portugal), which were very beneficial for my work;
- My Parents, Francisco de Oliveira Couto and Maria Fernanda dos Santos Moreira Couto, who bought me a Sinclair Spectrum home computer when I was a kid, and encouraged me to use it. Looking back, my parents inspired decision was the first step on the road that has culminated in this thesis. Throughout my life I have been extremely fortunate to have such an extraordinary parents, to whom I will be always grateful.

Lisbon, 11th January 2006

Francisco José Moreira Couto

To my parents

Contents

1	Introduction	1
1.1	Objectives and Contributions	5
1.2	Methodology	10
1.3	Results	11
1.4	Organisation of the Dissertation	12
2	Bioinformatics	15
2.1	The Basics of Molecular Biology	17
2.2	Biological Databases	25
2.3	BioOntologies	30
2.4	BioLiterature	33
2.5	Current Research Topics	35
2.6	Conclusions	36
3	Text Mining of BioLiterature	39
3.1	The Basics of Text Mining	40
3.2	Text Mining Approaches	43
3.3	State-of-the-art Systems	44
3.3.1	Rule-based Systems	45
3.3.2	Case-based Systems	46

3.3.3	Discussion	47
3.4	Evaluating Text Mining of BioLiterature	48
3.5	Conclusions	50
4	System: ProFAL	53
4.1	Architecture of ProFAL	54
4.2	Verification Use-Case	58
4.2.1	CAZy	59
4.2.2	APEG	61
4.2.3	UniProt	64
4.3	Conclusions	70
5	Retrieval: WeBTC	73
5.1	Method	74
5.2	Assessment	76
5.2.1	Setup	78
5.2.2	Results	79
5.3	Discussion	81
5.4	Conclusions	83
6	Extraction: FiGO	85
6.1	Method	86
6.2	Assessment	91
6.2.1	Results	93
6.3	Discussion	95
6.4	Conclusions	98
7	Validation: CAC	99
7.1	Method	100

7.2	Assessment	105
7.2.1	Results	107
7.3	Discussion	110
7.4	Conclusions	113
8	Conclusions	115
8.1	Research Contributions	116
8.1.1	WeBTC	117
8.1.2	FiGO	117
8.1.3	CAC	118
8.2	Limitations	118
8.3	Future Work	120
A	ProFAL Class Diagram	123
A.1	Retrieval	123
A.2	Extraction	124
B	Semantic Similarity Measures	127
B.1	Basic Concepts	127
B.2	State-of-the-art Measures	130
B.3	GraSM	131
B.3.1	Computational Aspect	136
B.4	FuSSiMeG	138
	Bibliography	141

List of Figures

1.1	MEDLINE growth.	3
1.2	Process flow diagram for Text Mining of BioLiterature.	5
1.3	ProFAL Architecture.	7
1.4	Methods and databases used by ProFAL.	8
2.1	Gene sequence.	16
2.2	Chromosome.	18
2.3	Gene expression.	19
2.4	Protein sequence	20
2.5	The central dogma of molecular biology.	20
2.6	Protein Folding.	22
2.7	Protein 3D structure.	23
2.8	Sub-graph of GO.	32
4.1	The primary use cases for ProFAL.	55
4.2	Bibliographic interface designed to be integrated in CAZy.	59
4.3	CAZy's bibliographic interface in May 2005.	60
4.4	Documents shown by APEG.	62
4.5	Annotations shown by APEG.	63
4.6	Documents shown by GOAnnotator.	65
4.7	Annotations shown by GOAnnotator.	66

5.1	Performance of WeBTC.	80
5.2	Scoring Results of the BioText Task of KDD2002 Cup.	80
6.1	Performance of all BioCreAtIvE submissions in task 2.	92
6.2	GO evaluation of FiGO predictions.	93
6.3	Performance of FiGO.	93
7.1	Performance of CAC.	107
A.1	Class diagram of the Retrieval use-case.	124
A.2	Class diagram of the Extraction use-case.	125
B.1	FuSSiMeG screenshot.	139

List of Tables

2.1	Comparative genome sizes of humans and other organisms. . .	24
2.2	BioOntologies, main advantage and disadvantage.	30
2.3	Data sources and their Web addresses.	37
3.1	Bag-of-words representation.	40
3.2	Text Mining approaches, main advantage and disadvantage. .	43
3.3	Categorisation of some recent text-mining systems.	47
3.4	Recent challenging evaluations.	48
4.1	Distribution of the GO terms.	67
4.2	Evaluation of the evidence text provided by GOAnnotator. . .	67
4.3	Evaluation of the extracted GO terms.	67
5.1	Performance of WeBTC.	79
7.1	Statistics of the three sets of annotations.	106
7.2	Results obtained by CAC.	109
7.3	Performance of CAC.	109
7.4	Overall results obtained by CAC.	112
B.1	Information content example.	134

1

Introduction

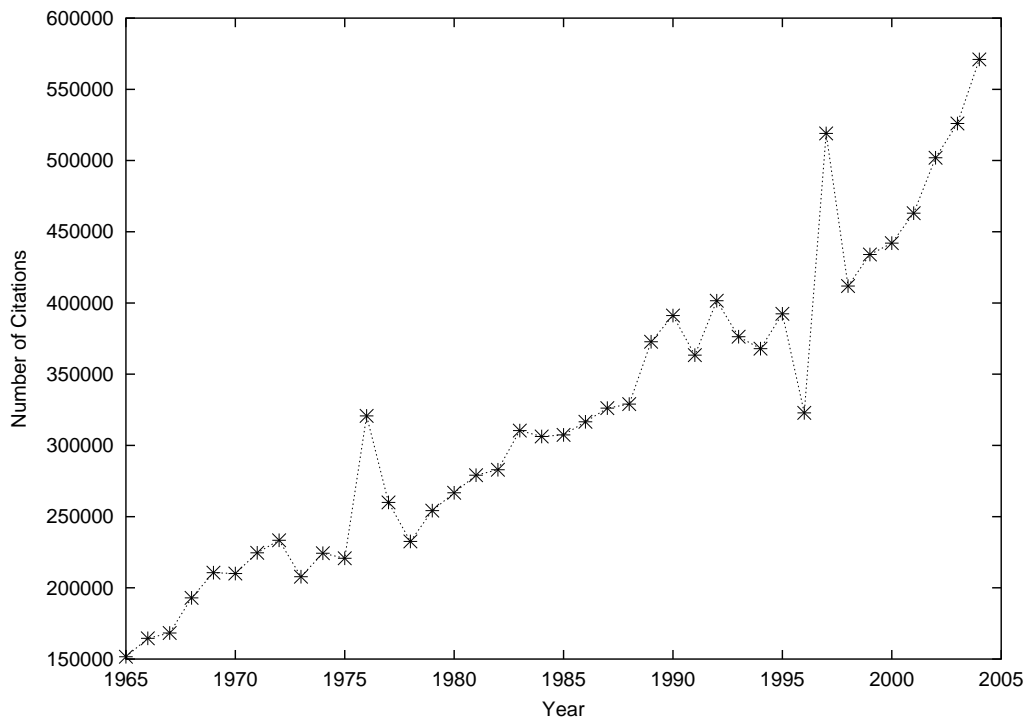
The large amounts of data produced by Molecular Biology projects spawned computational methods that analyse these data to discover novel insights into how living systems work. This field, known as Bioinformatics, is concerned with understanding living systems by exploring biological information using computer technology (Baldi and Brunak, 2001). In the beginning, Bioinformatics tools were basically used for handling, retrieving and analysing the large amounts of sequence data. Nowadays, Bioinformatics tools are already being used for predicting and modelling various aspects of living systems, and also for building and maintaining databases of biological data.

A large amount of the information discovered in Molecular Biology has been mainly published in BioLiterature (a shorter designation for the biological and biomedical scientific literature). Analysing and identifying information in a large collection of unstructured texts is a painful and hard task, even to an expert. To improve the access to the information, most researchers also deposit their findings in databases in a structured form. For instance, databases, such as UniProt (Universal Protein Knowledgebase), collect and distribute biological information (Apweiler et al., 2004). However, the management of these databases also became a complex problem, and most of them contain a significant number of errors (Devos and Valencia, 2001). Therefore, researchers cannot only rely in the facts available in these

databases. They also need the evidence substantiating them, which is normally present in the BioLiterature. The evidence can be the description of the biological setting where the experiment was conducted or the subsequent discussion of the results. Most facts are only valid in a specific biological setting, and should not be directly extrapolated to other cases. In addition, different research communities have different needs and requirements at a given period in time. As these constraints evolve, its management becomes harder to fulfil by databases, which have a static structure. Thus, researchers tend to use databases as an additional source to store and find facts, but the evidence substantiating them is still described as unstructured text, given its higher flexibility. As a consequence, a large amount of the knowledge acquired in Molecular Biology can only be found in the BioLiterature.

At present, most access to BioLiterature is done through PubMed, an interface that gives open access to over 15 million citations for documents related to life sciences (Wheeler et al., 2003). These citations are mainly issued from MEDLINE, a BioLiterature repository of abstracts and bibliographic information. Figure 1.1 presents the number of citations added to MEDLINE in the past decades, showing that there is an increasing large amount of documents that researchers have to deal with. As biological data and information continue to grow exponentially, the need for efficient access to BioLiterature is becoming critical to allow the researchers to conduct informed work, avoid repetition, and generate new hypotheses.

An approach to improve the access to the knowledge published in BioLiterature is to use Text Mining, which aims at automatically extracting knowledge from natural language text (Hearst, 1999). The application of text-mining tools to BioLiterature started just a few years ago (Andrade and Bork, 2000). Since then, the interest in the topic has been steadily in-



Source: http://www.nlm.nih.gov/bsd/index_stats_comp.html

Figure 1.1: Chronological listing of the number of citations present in MEDLINE from its beginning to the present day.

creasing, motivated by the vast amount of documents that curators have to read to update biological databases, or simply to help researchers keep up with progress in a specific area (Couto and Silva, 2005). Thus, Bioinformatics tools are increasingly using Text Mining to collect more information about the concepts they analyse. Text-mining tools have mainly been used to identify:

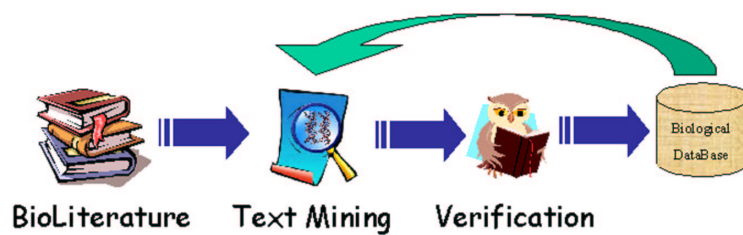
- entities, such as genes, proteins and cellular components;
- relationships, such as protein localisation or protein interactions;
- events, such as experimental methods used to discover protein interactions.

An important application of text-mining tools is the automatic annotation of genes and proteins. A gene or protein annotation consists of a pair composed by the gene or protein and a description of its biological role. The biological role is often a concept from a BioOntology, which organises and describes biological concepts and their relationships. Using a BioOntology to annotate genes or proteins avoids ambiguous statements that are domain specific and context dependent. For example, the Gene Ontology is a well-established structured vocabulary that for example has been successfully used for gene annotation of different species (GO-Consortium, 2004). To understand the activity of a gene or protein, it is also important to know the biological entities that interact with it. Thus, the annotation of a gene or protein also involves identifying interacting chemical substances, drugs, genes and proteins.

Nowadays, the performance of state-of-the-art text-mining tools for automatic annotation of genes or proteins is still not acceptable by curators. Gene or protein annotation is more subjective and requires more expertise than simply finding relevant documents and recognising biological entities in texts. To improve their performance, state-of-the-art text-mining tools use domain knowledge manually inserted by curators (Yeh et al., 2003). This knowledge consists of rules inferred from patterns identified in the text, or on predefined sets of previously annotated texts. Domain knowledge improves precision, but it cannot be easily extended to work on other domains and demands an extra effort to keep the knowledge updated as BioLiterature evolves. Since this approach is time-consuming and makes the systems too specific to be extended to new domains, a novel approach to avoid this process is much needed.



(a) State-of-the-art approach



(b) Proposed approach

Figure 1.2: Generic process flow diagram for Text Mining of BioLiterature. Text-mining systems extract relevant information from the BioLiterature. Experts verify the extracted information before adding it to a database. To obtain acceptable levels of performance state-of-the-art text-mining systems use domain knowledge explicitly inserted by experts. Besides being time-consuming and therefore expensive, this manually created information is generally useless in other domains. The proposed approach obtains the domain knowledge directly from publicly available biological databases.

1.1 Objectives and Contributions

Nowadays, publicly available biological databases already provide a significant amount of information covering almost all fields of Molecular Biology. I propose to use this information as domain knowledge for text-mining tools. Instead of asking experts to provide the domain knowledge, it can be acquired from biological databases (including BioOntologies) that already contain curated data.

Figure 1.2 outlines the approach proposed by this thesis. It requires mini-

mal human intervention, since it avoids the complexities of creating rules and patterns covering all possible cases or creating training sets that are too specific to be extended to new domains (Shatkay and Feldman, 2003). Besides avoiding direct human intervention, automatically collected domain knowledge is usually more extensive than manually generated domain knowledge and does not become outdated, since public databases can be automatically tracked for updates as they evolve. The effectiveness of the proposed approach depends on the following hypothesis.

Hypothesis: In the automatic annotation of biological databases, the use of domain knowledge automatically integrated from biological data resources represents a feasible alternative to the use of domain knowledge explicitly created by experts.

This hypothesis was successfully demonstrated in non-biological domains. For example, Basu et al. (1998) formalise movie recommendation as a classification problem, and show that classification performance can be improved using features extracted from the Web. Cohen (2000) proposed a method that produces new features from a collection of Web pages, which reduced the error rate of classifiers in a wide variety of situations. External data sources may not be available for all individual problems, but when it is available, their information is often useful.

Biological databases that distribute biological information on the Web are nowadays common, and automatic tools that integrate these data sources are a potential approach to correct and complete our knowledge about biological entities (Gerstein, 2000). As a result, Molecular Biology offers a promising scenario for successful application of text-mining tools based on the proposed approach.

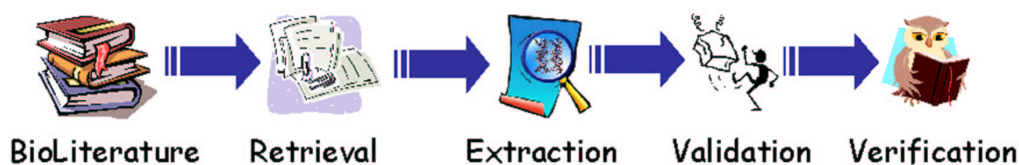


Figure 1.3: ProFAL starts by retrieving the relevant documents form BioLiterature. From this documents, it extracts relevant facts, which are then automatically validated. Before adding the facts to the database, they are manually verified by an expert.

To assess the above hypothesis, I developed ProFAL (bioProducts Functional Annotation through Literature), a modular system for automatic annotation of biological databases that integrates alternative text-mining methods. The diagram in Figure 1.3 shows the sequence of processing steps of ProFAL that are:

Retrieval: receives all the BioLiterature available as input and returns a set of documents containing relevant information.

Extraction: receives a document as input and returns relevant annotations reported in the document together with the pieces of text that substantiate them.

Validation: discards the incorrect annotations found in the previous step.

Verification: receives the annotations together with the evidence texts and displays this information to the user that takes the final decision about their accuracy.

Each step aims at reducing the information given as input. For implementing the retrieval, extraction and validation steps, I developed novel text-mining methods that achieved acceptable results and do not use manually inserted domain knowledge: The new proposed methods are:

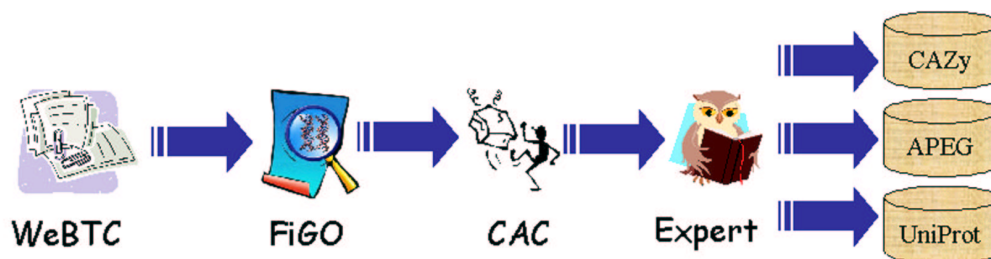


Figure 1.4: Methods and databases used by ProFAL. WeBTC retrieves the relevant documents. FiGO extracts relevant facts and their evidence texts. CAC validates the facts extracted. ProFAL was applied to CAZy, APEG and UniProt databases.

WeBTC (Web Biological Text Classification): a novel method for generating new features from biological text involving the integration of extracted information from biological Web resources (Couto et al., 2003a, 2004a). The new features are used to improve text classification on BioLiterature. WeBTC automatically extracts domain knowledge from publicly available databases that include relevant information about the scientific documents given as input.

FiGO (Finding Genomic Ontology): a novel unsupervised method for identifying biological terms organised in a BioOntology in unstructured text (Couto et al., 2004b, 2005a). FiGO automatically extracts domain knowledge from the information content of each word present in the nomenclature of the BioOntology.

CAC (Correlate the Annotations' Components): a novel method for discarding misannotations identified by automated systems (Couto et al., 2003b,d, 2005c). CAC automatically extracts domain knowledge from previously curated annotations with similar structure and function.

Figure 1.4 shows each method as an implementation of one of the processing steps of ProFAL. WeBTC selects relevant documents, FiGO identifies

relevant annotations on those documents, and CAC validates the annotations identified. The evaluation of these methods included the comparison of the effectiveness of the proposed methods with alternative state-of-the-art methods based on different approaches. To assess the viability of both the proposed methods and ProFAL, I developed novel tools that use ProFAL to automatically annotate the following biological databases:

CAZy (Carbohydrate Active enZymes): a database that describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds (Coutinho and Henrissat, 1999). ProFAL was applied to CAZy in collaboration with the AFMB-CNRS (Architecture et Fonction des Macromolécules Biologiques, Marseille, France) to automatically add bibliographic information about each CAZy entry.

APEG (Arabidopsis Pollen Expressed Gene): a database that provides functional information about *Arabidopsis thaliana* pollen selectively expressed genes (Jain, 2004). ProFAL was applied to APEG in collaboration with the IGC (Instituto Gulbenkian de Ciência, Oeiras, Portugal) to automatically annotate the genes in APEG (Jain et al., 2005).

UniProt: a generic database that aims at providing information about all known proteins. The new tool, named GOAnnotator, was designed in collaboration with the EBI (European Bioinformatics Institute, Hinxton, UK) to bring together evidence from the uncurated annotations and facts extracted from the BioLiterature linked to UniProt entries (Couto et al., 2005e).

The observation of ProFAL and the proposed methods operating in different realistic scenarios demonstrated that, with the proposed approach, it

is possible to produce tools that are both efficient and useful to database curators with much smaller update and maintenance costs than those requiring manually inserted domain knowledge.

1.2 Methodology

The hypothesis was validated by applying the scientific method described by Adrion (1993). I developed a theory describing each solution proposed, and tested it to verify the claims of the hypothesis. Each solution was a text-mining method integrating domain knowledge automatically collected from publicly databases. Each test evaluated the performance of the developed method in comparison to alternative methods.

The experimental data used in each test was collected using a controlled method (Zelkowitz and Wallace, 1998). WeBTC and FiGO were submitted to challenging evaluations. These evaluations compared the performance of different approaches in solving the same tasks using the same data at the same time. CAC was evaluated using the data from the challenging evaluation in which FiGO participated. Although not statistically significant, the datasets used by these challenging evaluations were carefully selected to be representative enough to allow meaningful and accurate conclusions.

In addition to the individual evaluation of each proposed method, I evaluated the effectiveness of the proposed approach in a real environment. For CAZy, APEG, and UniProt, I developed tools that integrated ProFAL in their curation process. The teams of each database used these tools to evaluate the information provided by ProFAL.

The performance was measured using the standard evaluation metrics: precision, recall, and F-measure (Manning and Schütze, 1999). Precision

measures the number of correct answers divided by the number of answers returned by a tool. Recall measures the number of correct answers divided by the maximum number of answers that can be correct. F-measure combines precision and recall by calculating the harmonic mean of the two measures:

$$\text{F-measure} = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} \quad (1.1)$$

1.3 Results

In all experiments, ProFAL has shown to be useful in finding new biologic annotations and providing a user-friendly interface for the curation process. ProFAL obtained the lowest performance in CAZy, but the evaluation took place with an early version of ProFAL. Since then ProFAL increased the levels of performance, reaching high levels of precision that meet the expectations of the curation process. These performance levels were only possible by integrating the novel methods, which also obtained positive results on their own:

- WeBTC was able to significantly increase the precision (reaching 100%) of a standard classification method. The performance of WeBTC was evaluated in the BioText Task of KDD2002 Cup versus state-of-the-art systems (Yeh et al., 2003). The evaluation indicated that WeBTC provided an effective alternative to enhance the performance of standard classification methods.
- FiGO obtained a good performance in BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) (Hirschman et al., 2005). Compared with the other submissions, FiGO demonstrated to

be an effective approach to recognize terms in BioLiterature, and to improve the performance of automatic annotation systems.

- CAC was applied to a set of annotations automatically extracted from BioLiterature. The method was able to improve the F-measure by achieving a large increase in precision for a low decrease in recall. This results show that CAC can effectively be used to discard incorrect annotations generated by automatic systems.

1.4 Organisation of the Dissertation

The remainder of this thesis is organised into four parts. The first part includes Chapters 2 and 3. They introduce the necessary background to comprehend the following Chapters. Chapter 4 represents the second part. It describes ProFAL in detail. The third part contains the description of the proposed methods and their evaluation in Chapters 5, 6 and 7. Chapter 8 represents the fourth part and presents the main conclusions of this thesis. The remainder of this Section summarises the contents of each Chapter.

Chapter 2 introduces basic concepts of Bioinformatics, describing how the enormous amount of biological data is being managed in public databases, the main problems of these databases and how they have been recently tackled by BioOntologies.

Chapter 3 introduces basics concepts of Text Mining and its application to BioLiterature, with a special focus on automatic gene and protein annotation. The Chapter describes the main approaches taken and presents, classifies and discusses some of state-of-the-art systems, which have been developed for automatically annotating genes or proteins.

ProFAL was designed to integrate the research contributions of this thesis

and demonstrate their viability in realistic scenarios. Chapter 4 presents the architecture of ProFAL, and describes the successful application of ProFAL to CAZy, APEG and UniProt databases. For each database, the Chapter describes the tool developed for the manual verification of the data, and discusses the obtained results.

Chapters 5, 6 and 7 describe WeBTC, FiGO and CAC in detail, respectively. These Chapters presents and discusses the results obtained by each method.

Finally, Chapter 8 summarises the contributions made by this dissertation and gives directions for future work building on this research.

2

Bioinformatics

Biotechnology aims at the transformation and application of biological materials. This technology can have profound impacts on human health, agriculture, the environment and energy. Well-known examples of biotechnology applications are:

- analysis of a person's genome and its expression to infer and measure susceptibility to different diseases and apply treatments at the earliest possible stages when they are more likely to be successful;
- production of stronger, more drought, disease and insect resistant crops and improvement of the quality of livestock, making them healthier, more disease resistant and more productive;
- replacement of pollutant material such as plastics and combustible by less pollutant biological materials.

Bioinformatics is a recent research area that aims at using computer technology for uncovering biological knowledge of high relevance to Biotechnology hidden in the vast amount of molecular biological data. Bioinformatics is an interdisciplinary research area at the interface among Biology, Biochemistry, Biophysics, Statistics, Mathematics, and Informatics.

This Chapter does not attempt to provide a complete overview of Bioinformatics. It only covers the topics related to this dissertation. Some of

```

ATGGGGCTCA GCGACGGGGA ATGGCAGTTG GTGCTGAACG TCTGGGGGAA GGTGGAGGCT
GACATCCCAG GCCATGGGCA GGAAGTCCTC ATCAGGCTCT TTAAGGGTCA CCCAGAGACT
CTGGAGAAGT TTGACAAGTT CAAGCACCTG AAGTCAGAGG ACGAGATGAA GGCATCTGAG
GACTTAAAGA AGCATGGTGC CACTGTGCTC ACCGCCCTGG GTGGCATCCT TAAGAAGAAG
GGGCATCATG AGGCAGAGAT TAAGCCCCTG GCACAGTCGC ATGCCACCAA GCACAAGATC
CCCGTGAAGT ACCTGGAGTT CATCTCGGAA TGCATCATCC AGGTTCTGCA GAGCAAGCAT
CCCGGGGACT TTGGTGCTGA TGCCCAGGGG GCCATGAACA AGGCCCTGGA GCTGTTCCGG
AAGGACATGG CCTCCAATA CAAGGAGCTG GGCTTCCAGG GCTAG

```

Figure 2.1: DNA coding sequence of the *Human Myoglobin* gene.
EMBL-Bank accession number = HSMG01.

the described facts are not applicable to all living systems, since in Molecular Biology no rule goes without an exception. The exceptions are usually rare, thus they were omitted to keep the text clear and easy to perceive. A more comprehensive overview of Bioinformatics is available in the articles of Kanehisa and Bork (2003); Chicurel (2002); Cohen (2004) or in the book of Attwood and Parry-Smith (1999).

The organisation of the Chapter is as follows. Section 2.1 gives a brief review of the basic concepts of Molecular Biology, focusing on the genetic information contained in each cell. Section 2.2 explains how this information is being maintained and made available in public databases. Section 2.3 describes the main sources of structured information used within the research community to characterise the genetic data. Section 2.4 presents the main sources of BioLiterature. Section 2.5 outlines current research topics. Finally, Section 2.6 provides the Web address of each data source and presents concluding remarks.

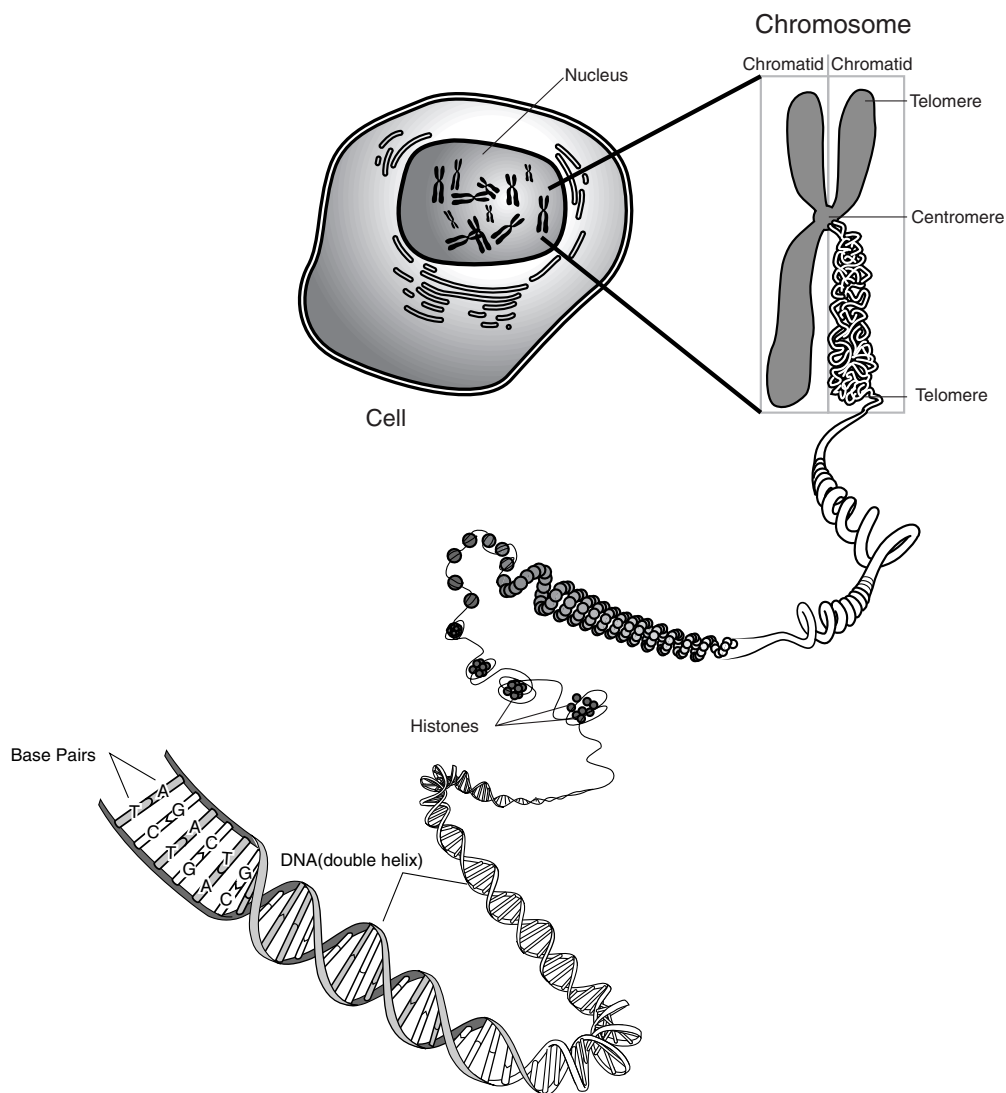
2.1 The Basics of Molecular Biology

Living organisms contain a complex system for storing and processing information essential for their survival. The information stored in each organism is transmitted from generation to generation through the process of reproduction, which causes their offspring to resemble their parents. The information is stored in DNA (deoxyribonucleic acid), a chain of phosphodiester-linked nucleotide residues inside a cell that contains the genetic instructions for creating and maintaining living systems (Alberts et al., 1989).

Each nucleotide residue consists of three parts: a deoxyribose sugar, a phosphate group and a nitrogenous base. There are four bases in DNA: adenine, thymine, guanine, and cytosine, which are usually denoted by A, T, G and C, respectively. A DNA molecule is normally represented as a sequence of characters, one for each base. For example, Figure 2.1 shows the DNA sequence of the *Human Myoglobin* gene from 5' to 3' (see below) as the usual convention.

One end of the DNA molecule is referred to as 5' (five prime) and the other end is referred to as 3' (three prime) according to the number of carbon atoms in the last deoxyribose sugar. A chromosome is composed of a continuous chain of DNA nucleotide residues stabilised by DNA-interacting proteins and is typically present in the cellular nucleus. The number of chromosomes inside a cell depends on the species. The genome is the entire DNA contained in the chromosomes. Figure 2.2 shows the generic structure of this genetic information.

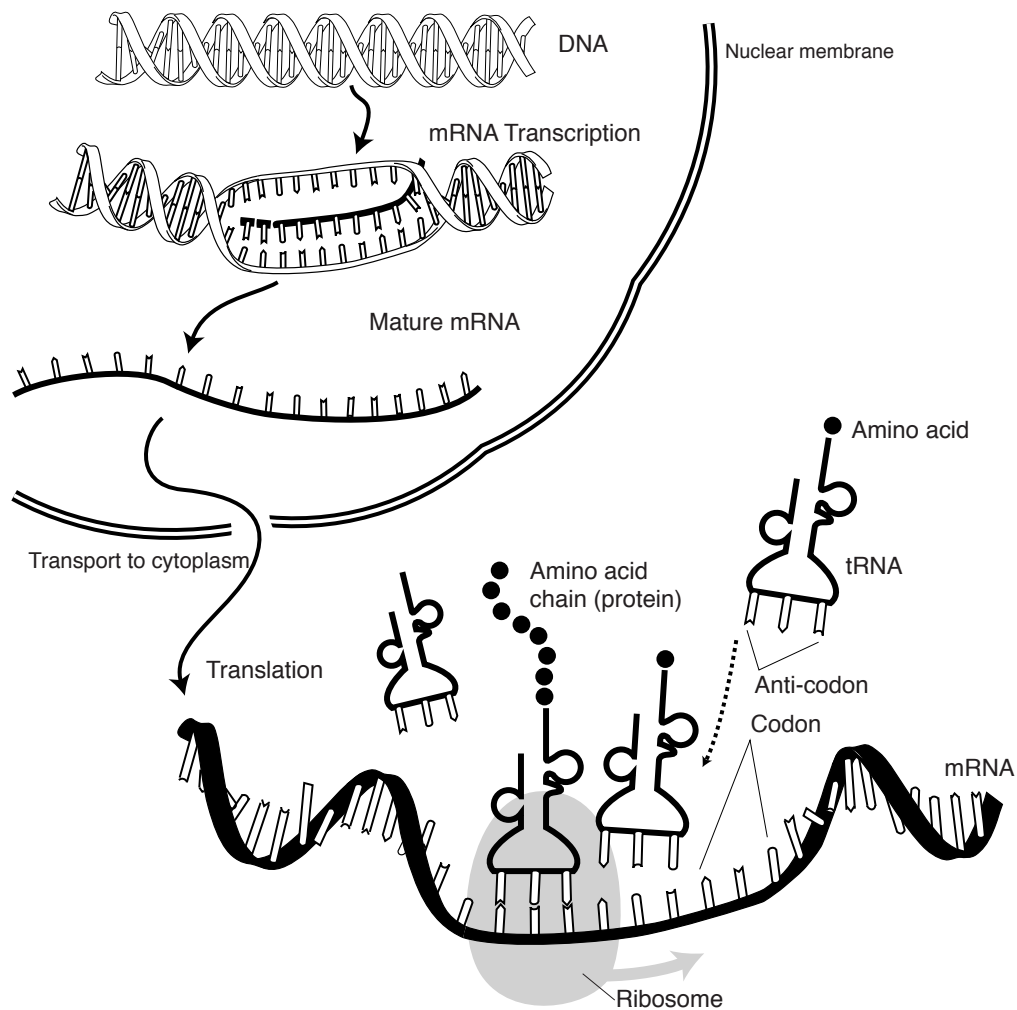
The genes are segments of DNA that encode genetic instructions for the protein synthesis. Figure 2.3 shows the transcription and translation mechanisms present in every cell, which are jointly responsible for the protein synthesis. These mechanisms read the active genes to produce proteins. In



Source: Talking Glossary of Genetics (<http://www.genome.gov/>)

Figure 2.2: A chromosome is a long strand of DNA in the nucleus of a cell. The DNA is composed by two anti-parallel chains of nucleotides twisted into a double helix and joined by hydrogen bonds.

simple organisms, a gene encodes a single protein, but in complex organisms a single gene can encode different proteins by alternative splicing (different combinations of the gene sequence). Independently of their type, all the cells in an organism contain identical DNA. However, on each individual cell only a



Source: Talking Glossary of Genetics (<http://www.genome.gov/>)

Figure 2.3: The segment of DNA that encodes a gene is transcribed into a molecule of messenger RNA (mRNA), which is then translated into a protein. This is the process by which proteins are made from the instructions encoded in DNA.

small fraction of the genes are active according to the cell cycle, environment and external signals.

The segment of DNA encoding a gene is first transcribed to a molecule of RNA (ribonucleic acid), which is then translated to a protein. RNA is a temporary intermediary in the transmission of information from the DNA to the protein, and it can be translated to proteins directly or in a processed

MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIIRLFKGGHPETLEKFDKFKHLKSEDEMKASEDLKKH
 GATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPGDFGADAQG
 AMNKALELFRKDMASNYKELGFQG

Figure 2.4: *Human Myoglobin* protein sequence.
 UniProt accession number = P02144.

Replication

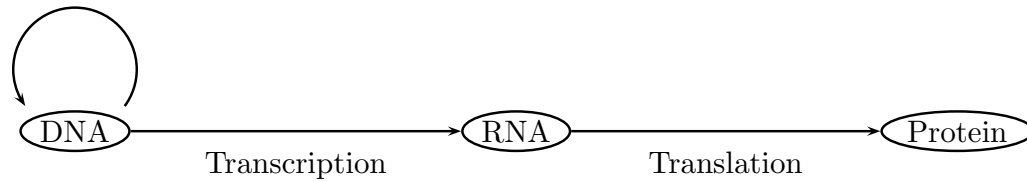


Figure 2.5: The central dogma of molecular biology.

form. The replication and transcription mechanisms read DNA in the 3' to 5' direction, while the translation mechanisms read RNA in the opposite direction.

Proteins are sequences of amino acid residues arranged in a specific order. Amino acids are a group of 20 different kinds of small molecules, which are also usually denoted by an alphabetic character. Each amino acid of a protein results from a set of three bases in the gene sequence. For example, Figure 2.4 shows *Human Myoglobin* protein sequence synthesised from the *Human Myoglobin* gene (sequence shown in Figure 2.1). The two ends of the amino acid chain are referred to as the amino terminus (N-terminus) and the carboxy terminus (C-terminus), corresponding to the 5' and 3' ends of the gene, respectively.

Replication is the mechanism responsible for producing identical copies of the original DNA so that it can be passed to new cells and offspring. Errors in the transmission of genetic material can have serious effects on the cell viability, but they are also at the basis of genome evolution and life diversity.

The transcription, translation and replication mechanisms form the central dogma of molecular biology as illustrated in Figure 2.5 (Crick, 1958).

The shape into which a protein naturally folds is known as its native state, which is determined by its sequence of amino acids residues. Figure 2.6 shows the four levels that define the structure of a protein.

Primary structure: sequence of amino acids residues that compose a polypeptide chain;

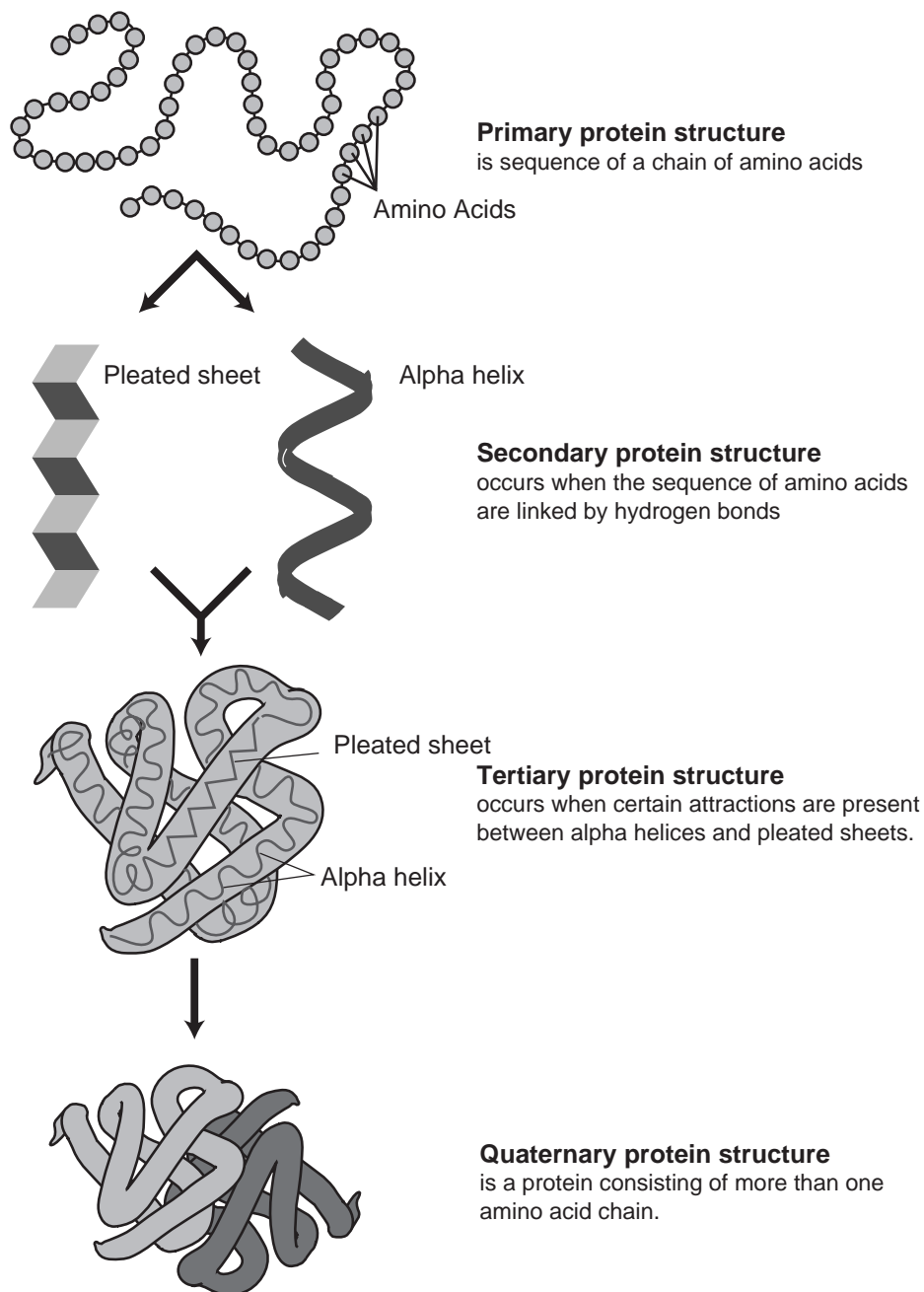
Secondary structure: highly patterned sub-structures, i.e. segments of the polypeptide chain that assume a stable and regular shapes (strands or helices);

Tertiary structure: the 3D (3-Dimensional) conformation assumed by a single polypeptide chain.

Quaternary structure: structure level assumed only by some proteins where individual polypeptide chains, forming protein subunits, are combined to get the final functional form.

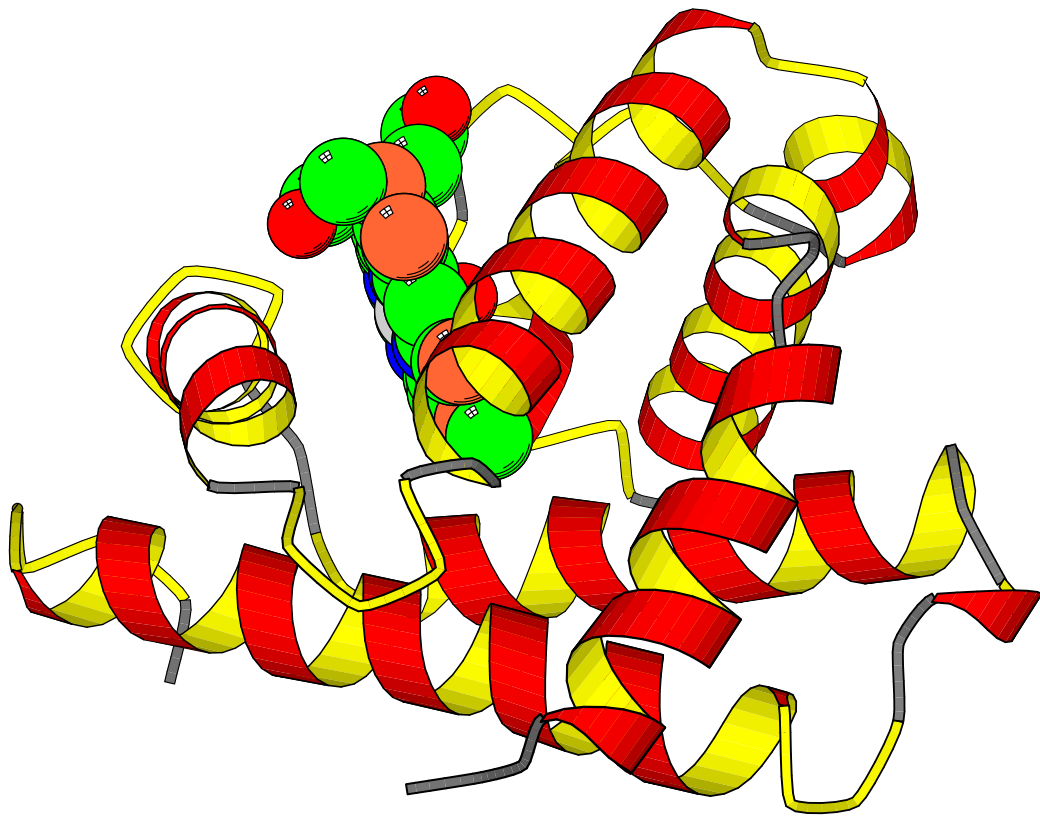
For example, Figure 2.7 shows a representation of the 3D structure of the *Human Myoglobin* protein.

Proteins are basic components of all living cells and control most of the functioning of living systems. They perform a wide variety of activities in the cell. Some proteins are enzymes and catalyse chemical reactions in the cell, but proteins can also perform structural and regulation activities. The activity of a gene is commonly considered the set of activities performed by the proteins it encodes, which can be indirectly found through the transcription levels or more directly through the quantification of expressed proteins. Only a reduced number of genes do not contain information for expressing proteins.



Source: Talking Glossary of Genetics (<http://www.genome.gov/>)

Figure 2.6: Proteins are amino acid chains that fold into unique 3D protein structures. The structure of a protein has four distinct aspects, all of them unequivocally determined by its amino acids sequence.



Source: Image Library of Biological Macromolecules

Figure 2.7: 3D structure of *Human Myoglobin* protein.
PDB accession number = 2MM1.

The activity or function of a protein can be defined at different levels: phenotypic (effect on the outward appearance of an organism), cellular, and molecular levels (Boork et al., 1998). Each protein has elementary molecular functions that are normally independent of the environment, such as catalytic (acceleration of a chemical reaction) or binding (capacity to hold or to attach to other molecules) activities. Sets of proteins interact to perform a variety of cellular functions, such as metabolism (transformation of molecules), signal transduction or RNA processing. A protein can act in different cellular localisations and be involved in different cellular functions while performing the same molecular function. At a higher level, the set

Organism	Bases	Genes	Chromosomes
Homo sapiens (human)	2,900 Mb	30,000	46
Rattus norvegicus (rat)	2,750 Mb	30,000	42
Mus musculus (mouse)	2,500 Mb	30,000	40
Drosophila melanogaster (fruit fly)	180 Mb	13,600	8
Arabidopsis thaliana (plant)	125 Mb	25,500	5
Caenorhabditis elegans (roundworm)	97 Mb	19,100	6
Saccharomyces cerevisiae (yeast)	12 Mb	6,300	16
Escherichia coli (bacteria)	4.7 Mb	3,200	1
H. influenzae (bacteria)	1.8 Mb	1,700	1

Source: U.S. Department of Energy Human Genome Program (<http://www.ornl.gov/hgmis/>)

Table 2.1: Comparative genome sizes of humans and other organisms. The table shows for each organism the estimated number of DNA bases and genes, and the exact number of chromosomes in the genome. The amount of DNA is not proportional to the number of genes.

of all cellular functions perform phenotypic functions, which determine the structure (morphology), functioning (physiology) or behaviour (psychology) of a living system.

Identifying the activities of genes and proteins will allow a better understanding of how living systems work (Nowak, 1995). However, identifying the activities of genes and proteins is a non-trivial task. The activity of a gene or protein is normally regulated by other genes or proteins that interact with it, and the number of genes and proteins in a given organism is enormous. For example, recent studies indicate that there are about 25,000 unique human genes, and at least 20% of them encode more than one protein (Larsson et al., 2005). Moreover, the behaviour of genes and proteins is not stable. Their activity can be dramatically affected by slight changes in the environment, such as different molecular signals or physiological conditions.

2.2 Biological Databases

The large amount of biological data available nowadays has transformed the traditional way of research and development in life sciences. For example, in December 2005 there were 332 complete genomes published and 1,766 genome projects in progress¹. Table 2.1 shows the large amount of information that is being produced by these projects.

The original focus of Bioinformatics was on the creation and management of databases that store the biological information being produced. Like biological data, the amount of public biological databases has grown at an exponential rate. Nowadays, there are virtually thousands of public databases available on the Web. However, most of them link their data to a few primary databases. Primary databases aim at storing the existing sequence and structural information of genes and proteins. Occasionally, they also include functional information. These databases add new information mainly through direct submission, and most of them use crosscheck tools for validating and updating their data. The primary databases used in this thesis are:

EMBL-Bank (EMBL Nucleotide Sequence Database): the Europe's primary nucleotide sequence resource (Kanz et al., 2005). Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications. In December 2005, EMBL-Bank contained 64,619,747 sequence entries comprising 116,719,333,361 nucleotides.

GenBank: a repository of genetic sequence data, containing an annotated collection of all publicly available DNA sequences (Benson et al., 2004).

¹<http://www.genomesonline.org/>

The information in GenBank is maintained uniformly across the collaborating databases: the EMBL-Bank Nucleotide Sequence Database, and the DNA Data Bank of Japan (DDBJ). The gene sequences are associated with protein information by using the PID, a unique protein identification number that is also used by the collaborating databases. In December 2004, GenBank contained approximately 56,037,734,462 nucleic acids bases in 52,016,762 sequence records.

UniProt (Universal Protein Resource): a database of protein sequence and functional data, created by joining the information in SwissProt, TrEMBL, and PIR databases (Apweiler et al., 2004). UniProt aims at removing redundant protein sequences and at curating the information based on experimental data. However, the level of annotation on a single entry can vary significantly. In December 2005, UniProt contained 2,710,972 sequence records.

PDB (Protein Data Bank): a repository for the processing and distribution of 3D biological macromolecular structure data (Berman et al., 2000). In December 2005, PDB contained 34,375 protein structures.

Additional biological databases can be found in the catalogue of biological resources maintained by the European Bioinformatics Institute², or in the Nucleic Acids Research Database Categories List³.

Secondary databases aim at organizing, integrating and classifying all, or more commonly a part of, the information stored in the primary databases. These databases can focus on specific species and/or on a specific set of activities. The secondary databases used in this thesis are:

²<http://www.ebi.ac.uk/biocat/>

³<http://www3.oup.co.uk/nar/database/c/>

FlyBase: a comprehensive database for information on the genetics and molecular biology of *Drosophila* (fruit fly) (Rubin, 1996). This database managed 57,216 genes in December 2005.

APEG (Arabidopsis Pollen Expressed Gene): a database that describes the function of *Arabidopsis thaliana* pollen selectively expressed genes (Jain, 2004). This database managed 147 genes in December 2005.

CAZy (Carbohydrate Active enZymes): a database that describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds (Coutinho and Henrissat, 1999). This database managed over 41,000 proteins in December 2005.

GOA (Gene Ontology Annotation): a database of protein annotations within the UniProt Knowledgebase (Camon et al., 2004). It contains high-quality curated annotations, but manual annotation tends to be slow and therefore covers less than 3% of UniProt. For better coverage, GOA also integrates uncurated annotations deduced from automatic mappings between UniProt and other manually curated databases. This database contained 8,146,303 protein annotations in December 2005.

Pfam: a database that provides a set of protein domains and families (Bateman et al., 2004). The Pfam families are constructed semi-automatically using hidden Markov models (HMMs). Each family describes a set of related proteins that can have identical molecular functions, are involved in the same process, or act in the same cellular location. This database contained 8183 families in December 2005.

GeneRIF: a database of functional gene annotations submitted by the research community (Mitchell et al., 2003). The submissions are reviewed for inappropriate content and typographical errors. The annotations aim at facilitating the access to documents reporting experiments relevant to understand a gene and its function. This database contained bibliographic annotations for more than 32,000 genes in December 2005.

Most primary databases also link their data to complementary information found in secondary databases. For example, UniProt links its proteins to Pfam families. Normally, secondary databases contain additional experimental data not present in primary databases. These additional experimental data are most of the times essential for understanding the biological roles of genes and proteins. Frequently, these data are identified by curators through careful analysis of the BioLiterature. Since the knowledge of how genes and proteins interact is fundamental to understand their activity, some additional experimental data provides information about the networks of interacting genes and proteins, instead of sequence or structural information. This information is generated by post-genome projects that are analysing metabolic, gene regulatory and protein-protein interaction networks. Given the transient structure of this information, these projects are producing even more data than genomic projects (Kitano, 2002).

Traditional functional characterisation of genes and proteins cannot cope with the large amount of sequences being produced. Therefore, automatic tools have been used to extrapolate functional annotations from similar functionally characterised sequences. However, these tools have also produced a significant number of misannotations that are now present in the databases (Devos and Valencia, 2001). For example, the modular nature of proteins is often disregarded, and therefore some functional annotations being extrapo-

lated between proteins based on similar parts of the sequence are unrelated to the functional annotation. Even more problematic, some of these tools are extrapolating new annotations from misannotations and therefore spreading the errors, because most annotation efforts do not distinguish between extrapolated and curated annotations.

Functional characterisation is not normally linked to the experimental evidence that substantiates it, which makes the judgement about what is correct difficult. The misannotations present in the databases includes under-predictions and over-predictions (Doerks et al., 1998):

Under-prediction: an annotation that is too generic. Even though being usually correct, these annotations are of little value for the researchers.

Over-prediction: an annotation that is too precise and restrictive. These annotations are normally the result of wrong extrapolations, which disregard small variations in the sequence that are enough to change the protein specificity.

The right equilibrium between under and over predictions is hard to establish. Under-predictions can sometimes provide useful outlines of many proteins to non-experts, and over-predictions can sometimes provide useful hints about a protein's biological role to experts, which are able to filter the errors.

The lack of a standard nomenclature across biological databases also makes the crosschecking normally ineffective in removing the errors. Often we can find different names (synonyms) for the same genes or proteins, or, even worse, different genes or proteins from different organisms sharing the same name (homonyms) (Rebholz-Schuhmann et al., 2005).

BioOntology	Advantage	Disadvantage
EC	deep specificity	ambiguity
MeSH	literature indexing	narrow scope
GO	broad scope and wide applicability	low specificity

Table 2.2: BioOntologies, main advantage and disadvantage.

2.3 BioOntologies

Biological databases annotate genes or proteins with statements that describe their biological role. Sometimes, these annotations are stored as ambiguous statements that are domain specific and context dependent. To cope with this, the research community is developing and using BioOntologies to annotate genes and proteins (Stevens et al., 2003). Ontology is defined as a specification of a conceptualisation that describes concepts and relationships used within a community (Gruber, 1993). This is a generic definition of ontology, which comprises different ways for describing the concepts and the relationships. For example, controlled vocabularies, taxonomies and thesaurus are considered to be ontologies. Controlled vocabulary is a list of terms that have been enumerated explicitly. Taxonomy is a collection of controlled vocabulary terms organised into a hierarchical structure. Thesaurus is a networked collection of controlled vocabulary terms. Ontologies enable knowledge sharing and reuse, but designing them is a complex task. They require a common agreement among the members of a community on concepts that change over time.

In enzymology, the IUBMB (International Union of Biochemistry and Molecular Biology) maintains the EC (Enzyme Commission) hierarchy, which provides a hierarchical classification schema for enzymes (NC-IUBMB, 1992). The classification is limited to only four numbers, the first of which defines the kind of reaction catalysed, the next two define the chemical nature of the

substrates, and the fourth is a catalogue number. In December 2005, there were 4,579 enzyme activities for which Enzyme Commission (EC) numbers have been assigned⁴. Updates and revisions of the hierarchy are rare, and therefore some biological reactions have no EC numbers. This kind of classification is also restrictive, which can explain some of its ambiguity. For example, there are EC numbers describing generic reactions that cover more precise reactions described by a more recent EC number.

In the health sciences, the National Library of Medicine provides the MeSH (Medical Subject Headings) BioOntology (Nelson et al., 2004). It consists of sets of clinical terms naming descriptors in a hierarchical structure. There are 22,997 descriptors in MeSH in December 2005. The MeSH thesaurus is mainly used for indexing scientific documents, and it is continually revised and updated. People having subject matter knowledge perform literature selection, thesaurus maintenance and indexing. This provides high accuracy and consistency, but also low coverage. Therefore, most biological features are out of the scope of MeSH, as it is often restricted to clinical terms.

The GO (Gene Ontology) project is one of the major efforts in Molecular Biology, for constructing a BioOntology of broad scope and wide applicability (Bada et al., 2004). GO provides a structured controlled vocabulary of gene and protein biological roles, which can be applied to different species (GO-Consortium, 2004). GO comprised 20,069 distinct terms in December 2005. Since the activity or function of a protein can be defined at different levels, GO has three different aspects: *molecular function*, *biological process* and *cellular component*. Each protein has elementary molecular functions that normally are independent of the environment, such as catalytic or binding

⁴<http://www.expasy.org/enzyme/>

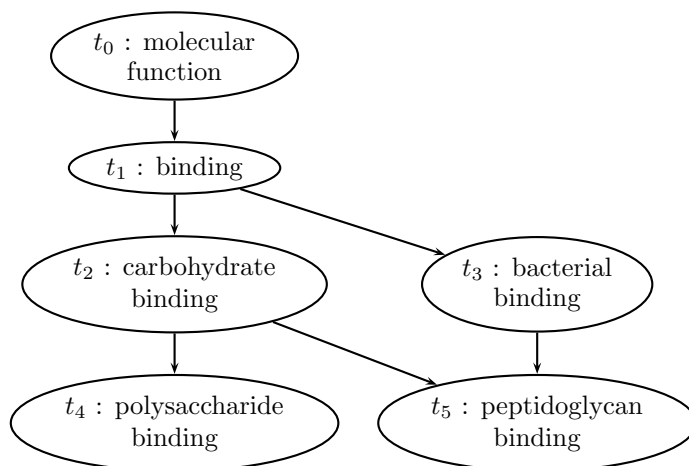


Figure 2.8: Sub-graph of GO.

activities. Sets of proteins interact and are involved in cellular processes, such as metabolism, signal transduction or RNA processing. Proteins can act in different cellular localisations, such as the nucleus or membrane.

GO organises the concepts as a DAG (Directed Acyclic Graph), one for each aspect. Each node of the graph represents a concept, and the edges represent the links between concepts (see example in Figure 2.8). Links can represent two relationship types: *is-a* and *part-of*. GO is a dynamic hierarchy: its content changes every month with the publication of a new release. Any user can request modifications to GO, which is maintained by a group of curators who add, remove and change terms and their relationships in response to modification requests. This prevents GO from becoming outdated and from providing incorrect information.

GO started by adding generic terms and simple relationships to provide a complete coverage of the Molecular Biology domain. Thus, the main limitation of GO is the lack of specific terms that, for example, represent precise biochemical reactions like EC numbers. However, as different research communities understand the importance of adding their domain knowledge to

GO, it will acquire more specific terms and relationships and therefore overcome this limitation.

2.4 BioLiterature

The notion of BioLiterature used in this thesis includes any type of scientific text related to Molecular Biology. The text is mainly available in the following formats:

Statement: a short piece of text that is normally a remark or an evidence for a fact stored in a database.

Abstract: a short summary of a scientific document.

Full-text: the full-text of a scientific document including scattered text such as figure labels and footnotes.

Statements contain more syntactic and semantic errors than abstracts, since they are not peer-reviewed, but they are directly linked to the facts stored in the databases. The main advantage of using statements or abstracts is the brief and succinct format on which the information is expressed. However, usually this brief description is insufficient to draw a solid conclusion, since the authors have to skip some important details given the text size constraint. These details can only be found in the full-text of a document, which contains a complete description of the results obtained. For example, important details are sometimes only present in figure labels. The main problem of full-text is its availability, since most of the full-text has restricted access. In addition, the structure of the full-text and the format on which is available varies according to the journal in where it was published. Having more information does not mean that it is all beneficial to text-mining tools. Some of

the information may even induce the tools in error. For example, the value of a fact reported in the Results Section would be different if the fact was reported in the Related Work Section. Therefore, the use of full-text will also create several problems regarding the quality of information extracted (Shah et al., 2004).

Most access to BioLiterature is done through PubMed, which in 2005 included over than 15 million citations from MEDLINE and other life science journals dating from the 1950s (Wheeler et al., 2003). PubMed aims at making it easier for the general public to search BioLiterature. The users can search for citations by author name, journal title or keywords. PubMed also includes links to full-text documents and other related resources. MEDLINE is a large repository of citations to the BioLiterature. It contains nearly 11 million citations from over 7,300 different publications from 1965 to the present day. Besides the bibliographic citations, MEDLINE also stores the abstracts of most documents, especially of the newer ones. The articles from 1950 through 1965 are in OLDMEDLINE, which contains approximately 1,760,000 citations (Demsey et al., 2003). These old citations do not contain the abstract and certain fields may contain outdated or erroneous data.

MEDLINE was designed to deal with printed documents, but nowadays many journals provide the electronic version of their documents. Moreover, some of them became Open Access Publications, which means that their documents are freely available with unrestricted use. These documents have been exploited by tools, such as Google Scholar⁵, Scirus⁶ or EBSCO⁷, which can be used to search and locate scientific documents. One of the major free digital archives of life sciences full-text documents is PMC (PubMed Central),

⁵<http://scholar.google.com/>

⁶<http://www.scirus.com/>

⁷<http://www.epnet.com/>

which aims at preserving and maintaining access to this new generation of electronic documents. Presently, PMC includes over 400,000 documents. The availability of full-text documents offers new opportunities to text-mining tools, which are most of times restricted to analysing only the abstracts of scientific documents.

2.5 Current Research Topics

Biologists are no longer capable of analysing the vast amount of genomic data being produced without using computational techniques. Thus, more than just managing and gathering the vast amount of genomic data, Bioinformatics is now focusing on exploring all these data to aid researchers in acquiring a better understanding of how living systems work (Couto et al., 2005d).

Typical tasks where Bioinformatics methods have been employed include:

Comparing Sequences: When a novel protein sequence is discovered, researchers attempt to find out as much as possible about it before laboratory testing. This involves comparing the novel sequence to other existing sequences, in particular to those already annotated (Altschul et al., 1997).

Constructing Evolutionary (Phylogenetic) Trees: Sequence data have been used to generate graphical representations of the evolutionary relationship among taxonomic groups or sequences (Nei, 1996).

Detecting Patterns in Sequences: Some functional features of proteins can be detected by recognizing patterns in their sequence (Rigoutsos et al., 2000). The patterns are like regular expressions that have been

identified in several proteins sharing the same feature. These patterns are useful to extrapolate the features of novel sequences. For example, they are used to automatically annotate UniProt/TrEMBL proteins (Bairoch and Apweiler, 2000).

Determining 3D Structures from Sequences: Most functional features of proteins depend on the protein 3D structure, but determining the structure in laboratory is expensive. Thus, a great effort has been devoted to the prediction of the protein structure from its sequence. Several approaches have been developed to tackle this problem, such as *ab initio* techniques, threading and homology modelling. Most of the systems participate in CASP (Critical Assessment of Techniques for Protein Structure Prediction), an annual international competition (Moult, 2005).

Inferring Cell Regulation: The activity of a protein is normally regulated by several molecular interactions that occur in the cell. Thus, discovering and modelling these interactions is an essential issue to understand what happens in the cell (Kitano, 2002).

2.6 Conclusions

This Chapter presented the topics of Molecular Biology related to this dissertation. It explained how the genetic information is being maintained and explored, which is a non-trivial problem due to the enormous amount of information available and to its high complexity. Table 2.3 presents the list of data sources described in this Chapter together with their Web addresses. Bioinformatics tools have been developed to cope with this problem. In a few

Primary Databases	
EMBL-Bank	http://www.ebi.ac.uk/embl/
GenBank	http://www.ncbi.nlm.nih.gov/Genbank/
DDBJ	http://www.ddbj.nig.ac.jp/
UniProt	http://www.ebi.ac.uk/uniprot/
PDB	http://www.rcsb.org/pdb/
Secondary Databases	
FlyBase	http://www.flybase.org/
APEG	http://xldb.fc.ul.pt/rebil/tools/apeg/
CAZy	http://afmb.cnrs-mrs.fr/CAZY/
GOA	http://www.ebi.ac.uk/goa/
Pfam	http://www.sanger.ac.uk/Software/Pfam/
GeneRIF	http://www.ncbi.nlm.nih.gov/projects/GeneRIF/GeneRIFhelp.html
BioOntologies	
GO	http://www.geneontology.org/
MeSH	http://www.nlm.nih.gov/mesh/meshhome.html
EC	http://www.chem.qmw.ac.uk/iubmb/enzyme/
BioLiterature	
MEDLINE	http://www.nlm.nih.gov/databases/databases_medline.html
OLDMEDLINE	http://www.nlm.nih.gov/databases/databases_oldmedline.html
PubMed	http://www.pubmed.org/
PMC	http://www.pubmedcentral.nih.gov/

Table 2.3: Data sources and their Web addresses.

cases the existing tools already achieve satisfactory results, but this problem is, in general, still far from being solved.

Now that we have most of the genomes available, the next big challenge is to understand the function of genes and proteins on concerted action, so the scientific community can understand how the living systems work. This constitutes a new discipline of Systems Biology (Ideker et al., 2001). This discipline will produce even more data that need to be processed, explored and disseminated. The fusion of biology and information sciences was therefore inevitable and we can anticipate that it will continue to expand in the future.

Molecular Biology has many exciting and hard problems to address. To solve them, we may need additional data not yet produced. Nevertheless, we will need novel Bioinformatics tools to reveal important information from the vast amount of data being produced. The solution to many complex problems may depend on facts already discovered by previous studies and published in BioLiterature. Thus, Molecular Biology will benefit from the development of text-mining tools providing friendly and efficient access to text reporting existing research achievements.

3

Text Mining of BioLiterature

Text Mining generally concerns the process of extracting relevant and non-trivial information and knowledge from unstructured text, usually a collection of documents. One target application of Text Mining is the BioLiterature, from where details of experimental results can be automatically extracted. However, the development of efficient text-mining techniques specific to BioLiterature is a recent research topic. As a result, the observed performance of text-mining tools in BioLiterature has been much lower than in other areas such as news text (Dickman, 2003).

The main problem in BioLiterature is coping with the lack of a standard nomenclature for describing biologic concepts and entities (Rebholz-Schuhmann et al., 2005). In BioLiterature, we can often find different terms referring to the same biological concept or entity (synonyms), or the same term meaning different biological concepts or entities (homonyms). Genes, whose name is a common English word, are frequent, which makes it difficult to recognize biological entities in the text. The information to extract is also more complex. It is almost impossible to derive a rule without having a significant number of exceptions.

This Chapter gives an overview of Text Mining and its application to BioLiterature, with a special focus on automatic gene and protein annota-

Sentence	<i>Prot</i> ₁	<i>Prot</i> ₂	protein	binds	enzyme	an	to
<i>s</i> ₁	1	1	2	1	0	0	1
<i>s</i> ₂	1	0	1	0	1	1	0

Table 3.1: Bag-of-words representation of the following sentences: *s*₁=*Protein p*₁ *binds to protein p*₂ and *s*₂=*Protein p*₁ *is an enzyme*.

tion. The organisation of the Chapter is as follows. Section 3.1 explains the basics of Text Mining. Section 3.2 describes two different approaches to extract knowledge. Section 3.3 presents examples of state-of-the-art text-mining systems and discusses their approaches. Section 3.4 describes recent assessments of text-mining systems. Finally, Section 3.5 presents concluding remarks.

3.1 The Basics of Text Mining

Text Mining draws from other areas such as Data Mining and NLP (Natural Language Processing).

$$TextMining = NLP + DataMining$$

Data Mining aims at automatically extracting knowledge from structured data (Hand et al., 2000). Thus, Text Mining is a special case of Data Mining, where input data is text instead of structured data. Normally, text-mining systems create structured representations of the text, which are then analysed by Data Mining tools. Table 3.1 presents the bag-of-words representation of two sentences, one of the simplest. The text is represented by a vector with the number of occurrences of each word in the sentences.

This representation can be easily created and manipulated, but ignores all the text structure. Text-mining systems may also use NLP techniques to

represent and process text more effectively. NLP is a broad research area that aims at analysing spoken, handwritten, printed, and electronic text for different purposes, such as speech recognition or translation (Manning and Schütze, 1999). The most popular NLP techniques used by text-mining systems include:

Tokenization: aims at identifying boundaries in the text to fragment the text into basic units called tokens. The first step in a text-mining system is to identify the tokens. The token most commonly used is the word. In most languages, the white-space character can be considered as accurate boundary to fragment the text into words. This problem is more complex in languages without explicitly delimiters, such as Chinese (Wu and Fung, 1994).

Morphology analysis: aims at grouping the words (tokens) that are variants of a common word, and therefore are normally used with a similar meaning (Spencer, 1991). This involves the study of the structure and formation of words. A common kind of inflectional variants results from the tense on verbs. For example, *binding* and *binds* are inflectional variants of *bind*. Some other word variants result from prefixing, suffixing, infixing or compounding.

Part-of-speech tagging: aims at labelling each word with its semantic role, such as article, noun, verb, adjective, preposition or pronoun (Baker, 1989). This involves the study of the structure and formation of sentences. The tagging is a classification of words, according to their semantic role and to their relations to each other in a sentence.

Sense disambiguation: selects the correct meaning of a word in a given piece of text. For example, *compound* has two different senses in the ex-

pressions *compound the ingredients* and *chemical compound*. Normally, the part-of-speech tags are used as a first step in sense disambiguation (Wilks and Stevenson, 1997)

Parsing: aims at identifying the syntactic structure of a sentence (Earley, 1970). The syntactic structure of a sequence of words is composed by a set of other syntactic structures related to smaller sequences, except for the part-of-speech tags that are syntactic structures directly linked to words. Normally, the syntactic structure of a sentence is represented by a syntax tree, where leafs represent the words and internal nodes the syntactic structures. Algorithms to identify the complete syntactic structure of a sentence are in general inaccurate and time-consuming, given the combinatorial explosion in long sentences. An alternative is shallow-parsing, which does not try to parse complex syntactic structures. Shallow-parsing only splits sentences into phrases, i.e. subsequences of words that represent a grammatical unit, such as noun phrase or verb phrase.

Anaphora (or co-reference) resolution: aims at determining different sequences of words referring to the same entity (Mitkov, 2002). For example, in the sentence *The enzyme has an intense activity, thus, this protein should be used*. The noun phrases *The enzyme* and *this protein* refer to same entity.

Some of the NLP techniques described above can be implemented using algorithms also used in Data Mining. For example, part-of-speech taggers can use Hidden Markov Models (HMMs) to estimate the probability of a sequence of part of speech assignments (Smith et al., 2004). Not all NLP techniques improve the performance of a given text-mining system. Thus,

Approach	Advantage	Disadvantage
rule-based	high precision	rules not found
case-based	less expertise	large training sets

Table 3.2: Text Mining approaches, main advantage and disadvantage.

designers of text-mining systems have to select which NLP techniques would be useful to achieve their main goal.

3.2 Text Mining Approaches

After creating a structured representation of texts, text-mining systems can use a rule-based or a case-based approach for extracting knowledge (Leake, 1996).

Rule-based approach: relies on rules inferred from patterns identified from the text by an expert. The rules represent, in a structured form, the knowledge acquired by experts when performing the same task. The expert analyses a subpart of the text and identifies common patterns in which the relevant information is expressed. These patterns are then converted to rules to identify the relevant information in the rest of the text. The main bottleneck of this approach is the manual process of creating rules and patterns. Besides being time-consuming, this manual process is, in most cases, unable to devise from a subpart of the text the set of rules that encompass all possible cases.

Case-based approach: relies on a predefined set of texts previously annotated by an expert, which is used to learn a model for the rest of the text. Cases contain knowledge in an unprocessed form, and they only describe the output expected by the users for a limited set of examples. The expert analyses a subpart of the text (training set) and provides

the output expected to be returned by the text-mining system for that text. The system uses the training set to create a probabilistic model that will be applied to the rest of the text. The main bottleneck of this approach is the selection and creation of a training set large enough to enable the creation of a model accurate for the rest of the text.

The manual analysis of text requires less expertise in the case-based approach than in the rule-based approach. In the rule-based approach, the expert has to identify how the relevant information is expressed in addition to the expected output. However, rule-based systems can use this expertise to achieve high precision by selecting the most reliable rules and patterns.

Example

Assume that we need to develop a system to recognise the gene names mentioned in a given set of documents.

If we decide to build a rule-based system, we have to ask the expert to read a subset of documents and provide us a set of rules that can be used to recognise the gene names in the text. For example, a rule might state that gene names are in most cases written in uppercase.

Otherwise, if we decide to build a case-based system, we have to ask the expert to read a subset of documents and highlight the pieces of text where the gene names occur.

3.3 State-of-the-art Systems

This thesis focuses on text-mining systems that automatically annotate genes or proteins, like ProFAL. This kind of systems can be categorised according

to: the mining approach taken (rule-based or case-based), the NLP techniques applied, and the amount of manual intervention required. The following Sections describe state-of-the-art text-mining systems designed for automatic annotation using this categorisation.

3.3.1 Rule-based Systems

Andrade and Valencia (1998) developed AbXtract, one of the first text-mining systems trying to characterise the function of genes and proteins based on information automatically extracted from BioLiterature . The system assigns relevant keywords to protein families based on a rule comprising the frequency of the keywords in the abstracts related to the family. AbXtract relies only on this rule and does not require human intervention.

Pérez et al. (2004) developed a system that annotates genes with keywords extracted from abstracts based on mappings among different BioOntologies. The system uses association rules that can be applied with all generality to any pair of linked databases.

Corney et al. (2004) developed BioRAT that given a query finds documents and highlights the most relevant facts in their abstracts or full-texts. BioRAT uses rules that are exclusively derived from patterns inserted by the user.

Müller et al. (2004) developed Textpresso, another rule-based system that finds documents and marks them up with terms from a built-in BioOntology. The system assigns to each entry of the BioOntology regular expressions that capture how the entry can be expressed in BioLiterature. Textpresso is less dependent on the user than BioRAT, since

many of the regular expressions are automatically generated to account for regular forms of verbs and nouns.

Kim and Park (2004) developed BioIE, a system that takes more advantage of NLP techniques. It extracts biological interactions from BioLiterature and annotates them with GO terms. The system uses morphology, sense disambiguation, and rules with syntactic dependencies to identify GO terms in the text. BioIE uses 1,312 patterns to match interactions in the sentences, thus it also requires substantial manual intervention.

Koike et al. (2005) developed a system similar to BioIE, which annotates gene, protein and families with GO terms extracted from texts. The system uses morphology, part-of-speech tagging, shallow parsing, and simple anaphora resolution. To extract the relationships, it uses both automatically generated and manually inserted rules.

3.3.2 Case-based Systems

Palakal et al. (2003) developed a case-based text-mining system, which extracts relationships between biological objects (e.g. protein, gene, cell cycle). The system uses sense disambiguation, and a probabilistic model to find directional relationships. The model is trained using examples of sentences expressing a relationship.

Chiang and Yu (2003) developed MeKE, another system that extracts protein functions from BioLiterature using sentence alignment. MeKE also uses sense disambiguation. The system uses a statistical classifier that identifies common patterns in examples of sentences expressing

System	Mining	NLP	Manual
Andrade and Valencia (1998)	rule-based	-	-
Pérez et al. (2004)	rule-based	-	-
Corney et al. (2004)	rule-based	Low	High
Müller et al. (2004)	rule-based	Low	Medium
Kim and Park (2004)	rule-based	Medium	Medium
Koike et al. (2005)	rule-based	High	Medium
Palakal et al. (2003)	case-based	Medium	Low
Chiang and Yu (2003)	case-based	Medium	Low
ProFAL	Hybrid	-	-

Table 3.3: Categorisation of some recent text-mining systems designed for automatic annotation of genes and proteins.

GO annotations. The classifier uses these patterns to decide if a given sentence expresses a GO annotation.

3.3.3 Discussion

The systems described above show how Text Mining can help curators in the annotation process. For each system, Table 3.3 indicates the mining approach taken, the proportion of NLP techniques used and the proportion of manual intervention needed to generate rules, patterns or training sets. Most systems rely on domain knowledge manually inserted by curators. Domain knowledge improves precision, but it cannot be easily extended to work on other domains and it demands an extra effort to keep the knowledge updated as BioLiterature evolves. This approach is time-consuming and makes the systems too specific to be extended to new domains.

Text-mining tools acquire domain knowledge from the curators as rules or cases. The identification of rules requires more effort from the curators than the evaluation of a limited set of cases. However, a single rule can express knowledge not contained in a large set of cases. None of the knowledge

Evaluation	Task	Submissions	F-measure	Precision	Recall
BioText Task of KDD2002 Cup	document selection	32	78%		
	gene identification	32	67%		
2004 TREC genomics track	document classification	47		41%	
	document selection	59	27%		
	gene annotation	36	56%		
BioCreAtIvE	evidence selection	21		78%	23%
	protein annotation	18		34%	12%

Table 3.4: Recent challenging evaluations of text-mining systems using BioLiterature. They compared the performance of different systems in solving a text-mining task using a given corpus. However, each evaluation evaluated a different task in a different corpus. Thus, the results cannot be compared among the three evaluations.

representation techniques subsumes the other: the knowledge enclosed in a rule is normally not fully expressed by a finite set of cases, and it is difficult to identify a set of rules encoding all the knowledge expressed by a set of cases.

ProFAL integrates a suite of methods that automatically collect the domain knowledge from databases. ProFAL uses a hybrid mining approach, since it uses different rules embedded in the proposed methods, which use the information stored in the databases as training sets. This represents a novel approach that has the advantages of being less time-consuming, easier to adapt to new or different conditions, and constantly updated as the information in the databases evolves.

3.4 Evaluating Text Mining of BioLiterature

Recent advances in text-mining tools using the BioLiterature achieve acceptable levels of accuracy in identifying gene and protein names in the text. However, performance of more complex tasks, such as the extraction of functional annotations, is still far from being satisfactory. Recent surveys report these advances by presenting text-mining tools that are run in different cor-

pora (collection of documents) to perform different tasks (Hirschman et al., 2002; Blaschke et al., 2002; Dickman, 2003; Shatkay and Feldman, 2003). On the other hand, recent challenging evaluations compared the performance of different approaches in solving the same tasks using the same corpus. Table 3.4 summarises these challenging evaluations. There were three different challenging evaluations:

BioText Task of KDD2002 Cup: consisted on identifying biomedical documents containing relevant experimental results about *Drosophila* (fruit fly), and the genes (transcripts and proteins) involved (Yeh et al., 2003). The best submission out of 32 obtained an F-measure of 78% in the document decision, and an F-measure of 67% in the gene decision.

2004 TREC genomics track: consisted of two tasks for identifying relevant documents and documents with relevant experimental results about the mouse (Hersh et al., 2004).

The first task was a typical Information Retrieval task. A list of documents and a list of topics were given as input. The goal was to identify the relevant documents for each topic. The best submission out of 47 obtained a precision of 41%.

The second task comprised the selection of documents with relevant experimental information. The best submission out of 59 obtained an F-measure of 27%. In addition to document selection, the task also comprised automatic annotations of genes. The best submission out of 36 obtained 56% F-measure.

BioCreAtIvE: was a critical assessment of information extraction systems in biology that comprised two tasks (Hirschman et al., 2005). The first aimed at identifying genes and proteins in BioLiterature. The best

submission out of 40 obtained 83% F-measure. The second task addressed the automatic annotation of human proteins, and involved two subtasks.

The first subtask required the identification of the texts that provided the evidence for extracting each annotation. From 21 submissions, the highest precision was 78% and the highest recall was 23% obtained by two different submissions.

The second subtask consisted on automatic annotation of proteins. From 18 submissions, the highest precision was 34% and the highest recall was 12% obtained by two different submissions.

3.5 Conclusions

Standard techniques have been used to cope with the problems found in BioLiterature, but text-mining systems are still far from reaching performance levels comparable to the obtained in other areas, such as in personal name recognition on news text (both precision and recall higher than 90%) (Kaufmann, 1995). Thus, novel techniques are required to reinforce and further improve the quality and impact of Text Mining of BioLiterature. The potential benefits to Molecular Biology are immeasurable, and therefore there is a growing interest in this exciting new research topic.

A useful application of text-mining tools to BioLiterature is in aiding curators to reduce the amount of information they have to manually process. Curators are in most cases unable to analyse directly all the information being published. Thus, text-mining tools can be used to find pieces of text that contain relevant experimental results. In this activity, the tools do not necessarily have to obtain high accuracy to be useful, since the curators later

verify the information obtained by the tool. It is less time-consuming to scan the provided information for errors than all the information that needs to be tracked. ProFAL provides annotations together with evidence text and lets the curator decide about their relevance and accuracy, which makes it a useful tool to assist curators in this activity.

4

System: ProFAL

The access to efficient text-mining systems using BioLiterature is crucial to researchers, enabling them to conduct informed work, avoid repetition, and generate new hypotheses. To make these systems more accessible and efficient, developers must understand the needs, requirements, and preferences of users and study how they would use them. This is an often neglected and non-trivial task. This thesis addresses this task by integrating the proposed methods in ProFAL (bioProducts Functional Annotation through Literature), a text-mining system developed to automatically annotate biological databases (Couto et al., 2003e). This system annotated genes and proteins of several databases with biological terms reported in BioLiterature. Curators of these databases evaluated the performance of ProFAL. Thus, besides individually validating the proposed methods, they were assessed as a whole in realistic biological settings.

Each evaluation used a quantitative measure to determine the accuracy of the information provided by ProFAL. These measures are not comparable to other studies since they evaluated a specific task performed in a restricted dataset. However, together with the feedback from the users, they were important to access how useful is ProFAL. User evaluation of a tool is most of the times subjective, but of extreme importance to achieve intended goals with effectiveness, efficiency and satisfaction in a realistic scenario. The num-

ber of curators that evaluated ProFAL was too small to provide a conclusive demonstration of its usability according to common standards (Jakob, 1994). Nevertheless, the evaluation shows that it is feasible to develop accessible and efficient text-mining systems based on the approach proposed by this thesis.

This Chapter describes ProFAL and how it was integrated and used to annotate three different databases. For each database, the Chapter presents the results obtained and describes the interface developed for presenting the information identified by ProFAL. The organisation of the Chapter is as follows. Section 4.1 describes the architecture of ProFAL. Section 4.2 describes the application of ProFAL to different databases. Finally, Section 4.3 presents the main conclusions derived from the experience of developing and using ProFAL in its application environments.

4.1 Architecture of ProFAL

ProFAL was designed as a software system to meet the needs of its target users. The main properties of ProFAL can be briefly described as:

For: Biological database curators.

Who: Require efficient access to BioLiterature.

The: ProFAL is an automated gene annotation system.

That: Provides the ability to retrieve relevant documents from the BioLiterature, to extract annotations together with their evidence texts from the documents retrieved, and to validate the annotations extracted.

Unlike: Replace human curation.

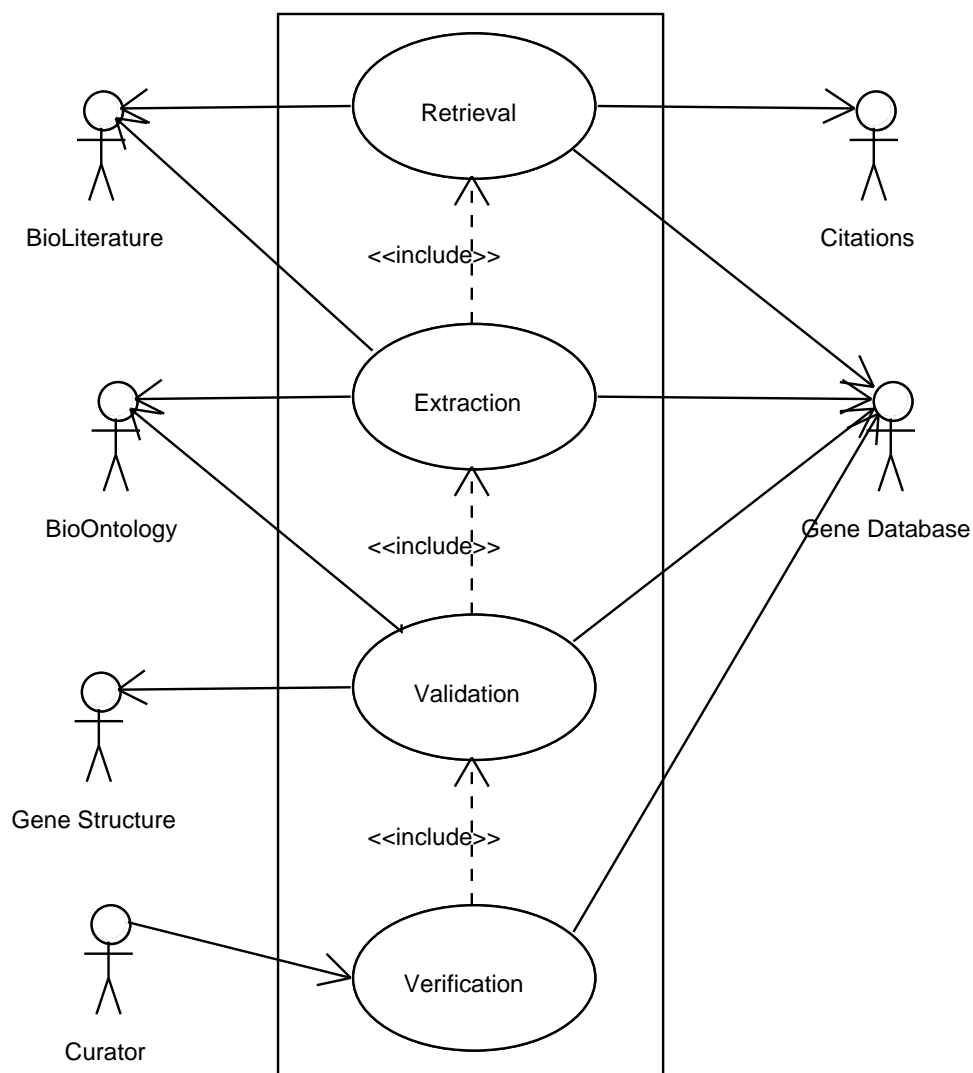


Figure 4.1: The primary use cases for ProFAL.

Novelty: Does not require domain knowledge explicitly created by experts, and can be easily applied to different biological databases.

ProFAL considers proteins as a special case of genes. Therefore, the discussion about gene annotation is also applicable to protein annotation, and vice versa. An annotation is a (gene, term) pair, associating the gene (or protein) to the term that describes a biological role of the gene.

It is not always possible or practical to design a product to have universal accessibility or usability. However, since different databases were interested in using ProFAL, it was designed to require minimal adjustments when applied to a new database. ProFAL uses a generic class diagram that can be easily integrated with different databases. Appendix A describes in detail this class diagram, which ProFAL uses to store the generated data and the data from the biological databases.

Figure 4.1 shows the primary use-case model of ProFAL describing its interaction with the users and its main processing steps. The model is represented in UML (Booch et al., 1998). The actors of ProFAL are:

BioLiterature: a collection of scientific documents.

Gene Database: a set of genes, which need to be annotated with terms from the BioOntology.

Citation: a collection of bibliographic references, which link the genes to BioLiterature.

Gene Structure: a gene classification scheme that organises the genes according to their structure or sequence.

BioOntology: a representation of a set of biological concepts and their relationships.

Curator: an individual who has the ability to manually verify gene annotations.

Each use-case in the model represents a processing step of ProFAL. The Verification use-case uses the annotations that are identified by the Extraction use-case and filtered by the Validation use-case. The Extraction use-case

uses the documents found by the Retrieval use-case to identify the annotations in their text.

The Verification use-case aims at using curators to evaluate all the information identified by the other use-cases. Thus, this use-case evaluates the performance of ProFAL as a whole. The following section describes implementations of this use-case in three different biological settings and discusses the results obtained. The Retrieval, Extraction and Validation use-cases are independent processing steps of ProFAL, which can also be individually assessed. Their implementation using the approach proposed by this thesis and their assessment is presented in the next three Chapters. The remainder of this section presents a brief description of each of these use-cases.

Retrieval:

Input: list of genes, a collection of scientific documents (e.g. all the documents available in PubMed), and a list of external sources (optional).

Output: for each gene a list of documents that express relevant information about the gene.

Process: from the given BioLiterature, the Retrieval use-case selects for each gene a set of documents that report relevant information about the gene.

Implementation: Chapter 5.

Extraction:

Input: list of genes, list of documents assigned to each gene, and a BioOntology.

Output: list of annotations of the input genes with terms from the given BioOntology, and evidence texts substantiating the annotations.

Process: in the documents assigned to each gene, the Extraction use-case identifies annotations that associate the gene with terms from the given BioOntology. The evidence texts are the sentences in the document that substantiate the identified annotations.

Implementation: Chapter 6.

Validation:

Input: list of annotations, a structural classification of genes and the BioOntology used in the Extraction use-case.

Output: confidence score of each annotation being correct.

Process: the Validation use-case automatically scores the annotations predicted by the Extraction use-case based on heuristics that measure the confidence degree on the annotation's correctness. The heuristics may use the structural information about the genes and the BioOntology to compare different annotations.

Implementation: Chapter 7.

4.2 Verification Use-Case

The Verification use-case provides an interface to assist the curator to analyse and judge the information identified by ProFAL. The curator verifies the annotations and the evidence texts, and modifies the confidence scores of the identified annotations. The following sections describe the interface implementations and the obtained results in three databases.

Publications								
PubMedID	MedlineID	Title	ISSN	Year	Classification	Note	#Authors	DB_ac
11435116	21345419	The kappa-carrageenase of <i>P. carrageenovora</i> features a tunnel-shaped active site: a novel insight in the evolution of Clan-B glycoside hydrolases.	0969-2126	2001	1	-3D-	7	
8112578	94156170	The gene encoding the kappa-carrageenase of <i>Alteromonas carrageenovora</i> is related to beta-1,3-1,4-glucanases.	0378-1119	1994	1		3	
Options	Search All		<input type="text"/>	Insert PMID		11435116 ▾	9 ▾	Alter Classification

Annotations					
TermsID	TermsName	Classification	Note	PubMedIDs	
GO:0016787	hydrolase	1		8112578	
GO:0008810	cellulase	1		11435116	
Options	<input type="text"/>	Insert Term	GO:0008810 ▾	9 ▾	Alter Classification

Figure 4.2: Bibliographic interface designed to be integrated in CAZy.

4.2.1 CAZy

CAZy is a database that describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds (Coutinho and Henrissat, 1999). ProFAL was integrated in CAZy to complement the information about the enzymes in CAZy with automatically extracted bibliographic information by taking advantage of the references to external databases, such as GenBank, UniProt, PDB and others.

Interface

Figure 4.2 presents an example of the bibliographic description of a specific enzyme, a bacterial *k-carrageenase*. The interface has two tables: the Publication table shows the bibliographic references, and the Annotation table shows the terms annotated with this enzyme. It has two bibliographic references. For each reference, the following information is presented: PubMed

Publications								
docAcc [9]		Title	ISSN	Year	Note	#Au	Same Reference	Value
11435116		The κ -carrageenase of <i>P. carrageenovora</i> features a tunnel-shaped active site: a novel insight in the evolution of Clan-B glycoside hydrolases.	0969-2126	2001	3D	7		100 A
8112578		The gene encoding the κ -carrageenase of <i>Alteromonas carrageenovora</i> is related to β -1,3-1,4-glucanases.	0378-1119	1994		3		100 A
33808		κ -Carrageenase from <i>Pseudomonas carrageenovora</i> .	0014-2956	1979	M EMP	2		100 A
Options	View all: PubMed HubMed XplorMed		docAcc: <input type="text"/>		NewValue: <input type="text"/>			
			<input type="button" value="Add document"/>		<input type="button" value="Delete document"/>		<input type="button" value="Modify value"/>	

Figure 4.3: CAZy's bibliographic interface in May 2005.

and MEDLINE accession numbers, title, journal (*ISSN*), year of publication, comments (*note*), authors and other referenced enzymes. The **-3D-** symbol is automatically added into the *note* field when the document is a PDB reference. This alerts the curator for the fact that this document may contain important structural information. The *authors* column presents the number of authors through a link to information about them. The *DB_ac* column presents accession numbers of other CAZy's enzymes that are also referenced by the document. The enzyme is annotated with 2 terms. For each reference the following information is presented: term's identifier number, term's type, term's name, comments (*note*), and the documents from where it was extracted. Both tables have a classification column for curating the entries. Its default value is 1, and its range goes from 0 to 9. An expert can replace the entry's classification according to its relevance. The last row of both tables has buttons to invoke administrative tools, for inserting new entries, and to reclassify the presented entries.

Given the great interest on having the proteins directly linked to scientific documents the Publication table was directly integrated in CAZy's interface. Figure 4.3 shows the actual look of the interface. The GO annotations were

not integrated until now, since GO has at present not enough specificity for providing valuable characterisations of the classes of enzymes in CAZy.

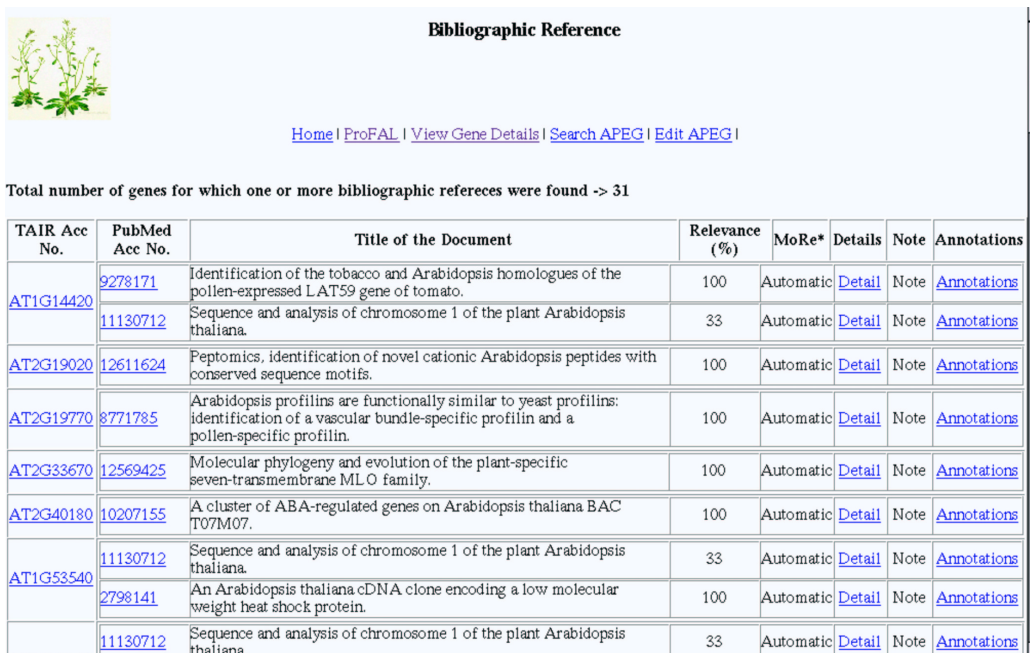
Results

In 2003, ProFAL assigned 6,377 distinct documents to 17,363 proteins, and identified 13,869 annotations in the documents, annotating 6,918 proteins with 1,342 GO terms. Only about 40% of the proteins were annotated because of the lack of bibliographic references for most proteins. This is not a limitation of ProFAL since it extracted, on average, 2.2 annotations per document. A CAZy curator manually verified 173 extracted annotations related to 5 distinct families. This curator classified their relevance to the characterisation of the functionality of the families as follows: 32 were classified as very important, 27 were classified as important, and 36 were classified as not so important. The remaining 78 annotations were classified as having no relevance. This gives a total of 95 correct annotations and 78 misannotations, representing a precision of 55%. However, some of the 78 misannotations could still be correct, since some proteins could also belong to other families that have not been considered.

ProFAL has been used to periodically add bibliographic references to proteins recently added to CAZy and to update the bibliographic references of the other proteins. Last time it was executed, in May 2005, ProFAL assigned 11,700 distinct documents to 18,345 proteins.

4.2.2 APEG

APEG is a database that describes the function of a collection of pollen selectively expressed genes of *Arabidopsis thaliana* (Jain, 2004). ProFAL was used by APEG to retrieve relevant documents from BioLiterature and



Bibliographic Reference

[Home](#) | [ProFAL](#) | [View Gene Details](#) | [Search APEG](#) | [Edit APEG](#)

Total number of genes for which one or more bibliographic refereces were found -> 31

TAIR Acc No.	PubMed Acc No.	Title of the Document	Relevance (%)	MoRe*	Details	Note	Annotations
AT1G14420	9278171	Identification of the tobacco and Arabidopsis homologues of the pollen-expressed LAT59 gene of tomato.	100	Automatic	Detail	Note	Annotations
	11130712	Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana.	33	Automatic	Detail	Note	Annotations
AT2G19020	12611624	Peptomics, identification of novel cationic Arabidopsis peptides with conserved sequence motifs.	100	Automatic	Detail	Note	Annotations
AT2G19770	8771785	Arabidopsis profilins are functionally similar to yeast profilins: identification of a vascular bundle-specific profilin and a pollen-specific profilin.	100	Automatic	Detail	Note	Annotations
AT2G33670	12569425	Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family.	100	Automatic	Detail	Note	Annotations
AT2G40180	10207155	A cluster of ABA-regulated genes on Arabidopsis thaliana BAC T07M07.	100	Automatic	Detail	Note	Annotations
AT1G53540	11130712	Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana.	33	Automatic	Detail	Note	Annotations
	2798141	An Arabidopsis thaliana cDNA clone encoding a low molecular weight heat shock protein.	100	Automatic	Detail	Note	Annotations
	11130712	Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana.	33	Automatic	Detail	Note	Annotations

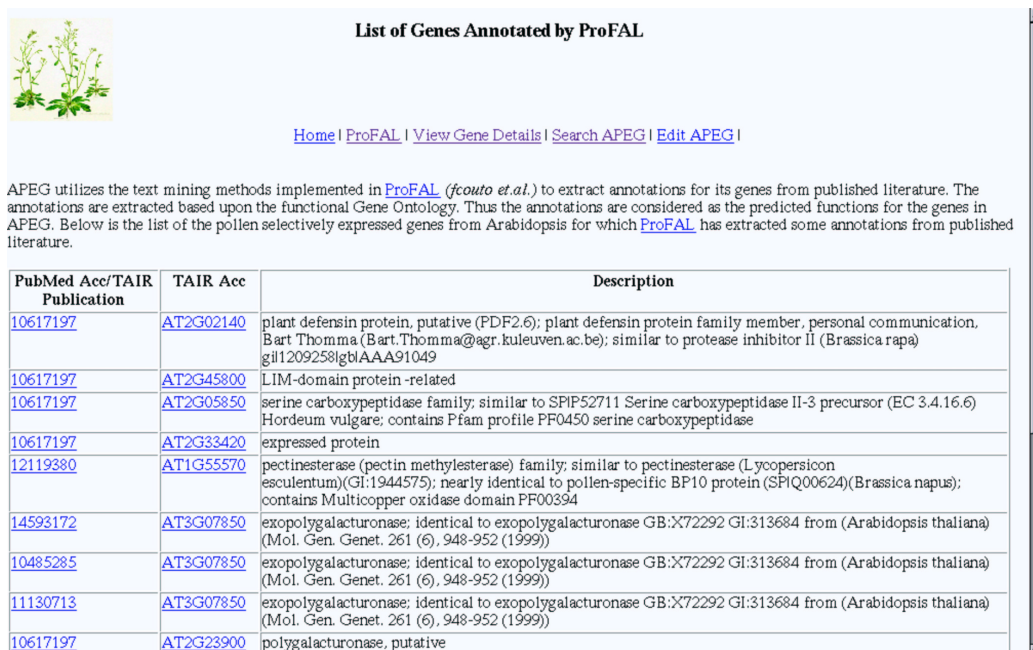
Figure 4.4: Documents shown by APEG.

for extracting functional annotations from them to add to the gene entries in APEG (Jain et al., 2005).

Interface

Figure 4.4 shows a list of bibliographic references returned by ProFAL. The information presented includes, for each document: the document identifier, its title, the confidence score returned by ProFAL, and a flag that indicates if it was manually or automatically identified.

Figure 4.5 shows a list of annotations returned by ProFAL. The information presented includes: the annotation, the document from which it was extracted, its evidence text, and the confidence score returned by ProFAL.



List of Genes Annotated by ProFAL

[Home](#) | [ProFAL](#) | [View Gene Details](#) | [Search APEG](#) | [Edit APEG](#) |

APEG utilizes the text mining methods implemented in [ProFAL](#) (*fcouto et al.*) to extract annotations for its genes from published literature. The annotations are extracted based upon the functional Gene Ontology. Thus the annotations are considered as the predicted functions for the genes in APEG. Below is the list of the pollen selectively expressed genes from Arabidopsis for which [ProFAL](#) has extracted some annotations from published literature.

PubMed Acc/TAIR Publication	TAIR Acc	Description
10617197	AT2G02140	plant defensin protein, putative (PDF2.6); plant defensin protein family member, personal communication. Bart Thomma (Bart.Thomma@agr.kuleuven.ac.be); similar to protease inhibitor II (Brassica rapa) gi12092581gb1AAA91049
10617197	AT2G45800	LIM-domain protein -related
10617197	AT2G05850	serine carboxypeptidase family; similar to SFP52711 Serine carboxypeptidase II-3 precursor (EC 3.4.16.6) Hordeum vulgare; contains Pfam profile PF0450 serine carboxypeptidase
10617197	AT2G33420	expressed protein
12119380	AT1G55570	pectinesterase (pectin methylesterase) family; similar to pectinesterase (Lycopersicon esculentum)(GI:1944575); nearly identical to pollen-specific BP10 protein (SPIQ00624)(Brassica napus); contains Multicopper oxidase domain PF00394
14593172	AT3G07850	exopolygalacturonase; identical to exopolygalacturonase GB:X72292 GI:313684 from (Arabidopsis thaliana) (Mol. Gen. Genet. 261 (6), 948-952 (1999))
10485285	AT3G07850	exopolygalacturonase; identical to exopolygalacturonase GB:X72292 GI:313684 from (Arabidopsis thaliana) (Mol. Gen. Genet. 261 (6), 948-952 (1999))
11130713	AT3G07850	exopolygalacturonase; identical to exopolygalacturonase GB:X72292 GI:313684 from (Arabidopsis thaliana) (Mol. Gen. Genet. 261 (6), 948-952 (1999))
10617197	AT2G23900	polygalacturonase, putative

Figure 4.5: Annotations shown by APEG.

Results

In 2004, ProFAL assigned 55 distinct documents to 71 genes. A curator of APEG evaluated the annotations extracted by ProFAL using different thresholds for the annotation confidence score. The annotations extracted by ProFAL were compared to annotation manually extracted from the same documents. The highest recall was 78% obtained with 199 annotations that achieved 61% precision. The highest precision was 76% obtained with 107 annotations that achieved 55% recall. ProFAL was also able to extract additional annotations not identified in a previous study on the functional characterisation of the genes in APEG (Becker et al., 2003).

4.2.3 UniProt

GOA curators use pre-existing uncurated annotations as a guide in their manually annotation process. These uncurated annotations can also be used to direct text-mining tools (Camon et al., 2004). Since GOA curators primarily require high precision in a text-mining solution, it is expected that the information from the uncurated annotations will support this goal without going through the complex issues of creating rules and patterns encompassing all possible cases, or creating training sets that are too specific. This Section illustrates how ProFAL was integrated in the GOA database curation process through GOAnnotator, a tool for assisting the GO annotation of UniProt proteins (Couto and Silva, 2005).

Interface

GOAnnotator uses ProFAL to link the GO terms present in the uncurated annotations with evidence text automatically extracted from the documents linked to UniProt proteins. GOAnnotator is available on the Web¹.

GOAnnotator aims at assisting the GO annotation of UniProt proteins by linking the GO terms present in the uncurated annotations with evidence text automatically extracted from the documents linked to UniProt proteins. Initially, the curator provides a UniProt accession number to GOAnnotator. GOAnnotator follows the bibliographic links found in the UniProt database and retrieves the documents. Additional documents are retrieved from the GeneRIF database (Mitchell et al., 2003). Curators can also provide any other text for mining. GOAnnotator then extracts from the documents GO terms similar to the GO terms present in the uncurated annotations. GO terms are similar if they are in the same lineage or if they share a common

¹<http://xldb.fc.ul.pt/rebil/tools/goa/>











PubMedId	Title	MostSimilarTermExtracted	Scope	Authors	Year	Extract	AddText
11594756(FullText)	Distinct phosphoinositide binding specificity of the GAP1 family proteins: characterization of the pleckstrin homology domains of MRASAL and KIAA0538.	100% GTPase activator activity (f)	GeneRIF	3	2001	 Pre-Processed	 Text
11448776(FullText)	CAPRI regulates Ca(2+)-dependent inactivation of the Ras-MAPK pathway.	100% GTPase activator activity (f)	SEQUENCE FROM N.A.	3	2001	 Pre-Processed	 Text
9628581(FullText)	Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro.	28% cell communication (p)	SEQUENCE FROM N.A.	7	1998	 Pre-Processed	 Text
14702039(FullText)	Complete sequencing and characterization of 21,243 full-length human cDNAs.	-	GeneRIF	154	2004	 Pre-Processed	 Text
12853948(FullText)	The DNA sequence of human chromosome 7.	-	SEQUENCE FROM N.A.	107	2003	 Pre-Processed	 Text

Figure 4.6: Some of the documents retrieved for the protein *Ras GTPase-activating protein 4*. The documents are sorted by the most similar term extracted from their text. The curator can use the *Extract* option to see the extracted terms together with the evidence text. By default, GOAnnotator only uses the abstract, but the curator can use the *AddText* option to replace or insert text.

parent in the GO hierarchy. A semantic similarity measure is used to determine the degree of similarity between two GO terms (see Appendix B). The extraction of GO terms is done by FiGO, which assigns a confidence value to each GO term that represents the terms' likelihood of being mentioned in the text (see Chapter 6).

GOAnnotator ranks the documents based on the extracted GO terms from the text and their similarity to the GO terms present in the uncurated annotations. Figure 4.6 shows the list of documents related to the protein *Ras GTPase-activating protein 4* provided by GOAnnotator. The list is sorted by the similarity of the most similar term extracted from each document. The curator can invoke the links in the *Extract* column to see the extracted terms together with the evidence text. By default, GOAnnotator only uses the abstracts of scientific documents, but the curator can replace or add text





Similar GO Terms Extracted	GOA Electronic Term: intracellular signaling cascade (p) [- ▾]	
inactivation of MAPK (p) [- ▾]	CAPRI regulates Ca ²⁺ -dependent inactivation of the Ras- MAPK pathway Ca ²⁺ is a universal second messenger that is critical for cell growth and is intimately associated with many Ras-dependent cellular processes such as proliferation and differentiation [1].	
protein kinase C activation (p) [- ▾]	A role for intracellular Ca ²⁺ in the activation of Ras has been previously demonstrated, e.g., via the nonreceptor tyrosine kinase PYK2 [3] and by Ca ²⁺ /calmodulin-dependent guanine nucleotide exchange factors (GEFs) such as Ras-GRF [4]; however, there is no Ca ²⁺ -dependent mechanism for direct inactivation .	
phosphoinositide-mediated signaling (p) [- ▾]	Previously, we have shown that these C2 domains do not regulate Ca ²⁺ - mediated membrane association; instead, membrane targeting is mediated by phosphoinositide binding PH domains [11, 12 and 13].	
Comment: <input type="text"/>	Evidence: [- ▾] <input type="button" value="--- Add ---"/>	New Terms: <input type="text"/> 

Figure 4.7: For each uncurated annotation, GOAnnotator shows the similar GO terms extracted from a sentence of the selected document. If any of the sentences provides correct evidence for the uncurated annotation, or if the evidence supports a GO term similar to that present in the uncurated annotation, the curator can use the *Add* option to store the annotation together with the document reference, the evidence codes and any comments.

(links in the *AddText* column). Any extracted GO term is an indication for the topic of the document, which is also taken from the UniProt entry.

GOAnnotator displays a table for each uncurated annotation with the GO terms that were extracted from a document and found similar to the GO term present in the uncurated annotation (see Figure 4.7). For each uncurated annotation, GOAnnotator shows the similar GO terms extracted from a sentence of the selected document. If any of the sentences provides correct evidence for the uncurated annotation, or if the evidence supports a GO term similar to that present in the uncurated annotation, the curator can use the *Add* option to store the annotation together with the document reference, the evidence codes and additional comments. The sentences from which the GO terms were extracted are also displayed. Words that have contributed to the extraction of the GO terms are highlighted. GOAnno-

GO Aspect	GO Terms
molecular function	54
biological process	18
cellular component	6
total	78

Table 4.1: Distribution of the GO terms from the selected uncurated annotations through the different aspects of GO.

Evidence Evaluation	Extracted Annotations
correct	83
incorrect	6
total	89

Table 4.2: Evaluation of the evidence text substantiating uncurated annotations provided by GOAnnotator.

GO Terms	Extracted Annotations
exact	65
same lineage	15
different lineage	3
total	83

Table 4.3: Comparison between the extracted GO terms with correct evidence text and the GO terms from the uncurated annotations.

tator gives the curators the opportunity to manipulate the confidence and similarity thresholds to modify the number of predictions.

Results

From the set of UniProt/SwissProt proteins with uncurated annotations and without manual annotations, GOAnnotator identified evidence texts with more than 40% similarity and 50% confidence for 66 proteins. For 80 uncurated annotations to these proteins, GOAnnotator extracted 89 similar annotations and their evidence text from 118 MEDLINE abstracts. The 80 uncurated annotations included 78 terms from different aspects of GO (see Table 4.1). After analysing the 89 evidence texts, GOA curators found

that 83 were valid to substantiate 77 distinct uncurated annotations (see Table 4.2), i.e. 93% precision.

In most cases, where the evidence text was correct, the GO term present in the extracted annotation was the same as the GO term present in the uncurated annotation (65 cases, see Table 4.3). Although the evidence text being correct, most of the times it did not exactly contain any of the known representations of the extracted GO term. In the other cases the extracted GO term was similar: in 15 cases the extracted GO term was in the same lineage of the GO term in the uncurated annotation; in 3 cases the extracted GO term was in a different lineage, but both terms were similar (share a parent).

In general, we can expect GOAnnotator to confirm the uncurated annotation using the findings from the BioLiterature, but it is obvious as well that GOAnnotator can propose new GO terms. In both cases, the curator profits from the integration of both approaches into a single interface. By comparing both results, the curator gets convenient support to take a decision for a curation item based on the evidence from the different data resources.

GOAnnotator ensures high accuracy, since all GO terms that did not have similar GO terms in the uncurated annotations were rejected. This meets the GOA team's need for tools with high precision in preference to those with high recall, and explains the strong restriction for the similarity of two GO terms: only those that were from the same lineage or had a shared parent were accepted. Thus, GOAnnotator not only predicted the exact uncurated annotation but also more specific GO annotations of strong interest to the curators. GOAnnotator takes advantage of uncurated annotations to avoid general terms by only extracting similar terms, i.e. popular proteins tend

to be annotated to specific terms and therefore GOAnnotator will extract specific annotations to them.

Examples

GOAnnotator provided correct evidence for the uncurated annotation of the protein *Human Complement factor B precursor* (P00751) with the term *complement activation, alternative pathway* (GO:0006957). The evidence is the following sentence from the document with the PubMed identifier 8225386:

The human complement factor B is a centrally important component of the alternative pathway activation of the complement system.

GOAnnotator provided a correct evidence for the uncurated annotation of the protein *U4/U6 small nuclear ribonucleoprotein Prp3* (O43395) with the term *nuclear mRNA splicing, via spliceosome* (GO:0000398). From the evidence, the tool extracted the child term *regulation of nuclear mRNA splicing, via spliceosome* (GO:0048024). The evidence is the following sentence from the document with the PubMed identifier 9328476:

Nuclear RNA splicing occurs in an RNA-protein complex, termed the spliceosome.

However, this sentence does not provide enough evidence on its own, the curator had to analyse other parts of the document to draw a conclusion.

GOAnnotator provided a correct evidence for the uncurated annotation of the protein *Agmatinase* (Q9BSE5) with the term *agmatinase activity* (GO:0008783). From the evidence, the tool extracted the term *arginase activity* (GO:0004053) that shares a common parent. The evidence was provided

by the following sentence from the document with the PubMed identifier 11804860:

Residues required for binding of Mn(2+) at the active site in bacterial agmatinase and other members of the arginase superfamily are fully conserved in human agmatinase.

Although, the annotation only received a NAS (Non-traceable author statement) evidence code, as the sentence does not provide direct experimental evidence of arginase activity. Papers containing direct experimental evidence for the function/subcellular location of a protein are more valuable to GO curators.

GOAnnotator provided a correct evidence for the uncurated annotation of the protein *3'-5' exonuclease ERI1* (Q8IV48) with the term *exonuclease activity* (GO:0004527). The evidence is the following sentence from the document with the PubMed identifier 14536070:

Using RNA affinity purification, we identified a second protein, designated 3'hExo, which contains a SAP and a 3' exonuclease domain and binds the same sequence.

However, the term *exonuclease activity* is too high level, and a more precise annotation should be *3'-5' exonuclease activity* (GO:0008408).

4.3 Conclusions

ProFAL was not designed to replace human curation, it only aims at assisting the curation process by reducing the amount of information that curators have to verify. The experiments described in this Chapter have shown that

ProFAL is useful in the process of gene annotation by providing a user-friendly interface that allowed fast verification of existing and novel annotations from evidence texts. More than facts, researchers need the source from which the facts derive. ProFAL provides not only facts but also their evidence, which curators found to be helpful to reduce their workload.

The evidence text taken from the abstract of a document is sometimes not enough to evaluate an annotation with a strong confidence. In addition to annotations, the curator needs additional information, such as the kind of experiments applied and the species from which the gene originates. Unfortunately, quite often this information is only available in the full-text of the scientific document, and ProFAL only retrieves the abstracts automatically, not the full-texts. Additionally, the list of documents cited in external sources is not enough for the curation process. In most cases, the curators found additional sources of information in PubMed. In the future, ProFAL should automatically query PubMed using the gene's names to provide a more complete list of documents.

ProFAL obtained the lowest performance in CAZy since its evaluation was performed at an early prototype stage. Since then, ProFAL increased its performance, as it is shown by the results obtained with APEG and UniProt. ProFAL reached 93% precision in UniProt, meeting the expectations of typical curators. More experiments are needed to obtain a complete demonstration of ProFAL's usability, but persuading curators to use a tool and provide some feedback is a non-trivial task. Nevertheless, the results clearly show that it is feasible to develop an efficient system incorporating methods based on the approach proposed by this thesis.

The observation of curators using ProFAL, in as realistic a situation as possible, was important to discover imperfections in the proposed methods

and to identify areas of improvement. Thus, besides demonstrating the feasibility of the proposed approach, the use of ProFAL also demonstrated the applicability of the proposed methods. The next three Chapters detail the methods that implemented each processing step of ProFAL.

5

Retrieval: WeBTC

The classification of BioLiterature is an important recent research topic, motivated by the large number of biological documents that curators have to read to update biological databases, or simply to be aware of progress in a specific area. Text classification applied to BioLiterature can minimize this effort by automatically selecting the relevant documents to a given task. Text classification systems are primarily designed to assign categories to documents to support information retrieval, or to provide an aid to human indexers in the assignment task. In the simplest form, binary classification, the system decides the relevant and irrelevant documents (or passages) from large corpora (Salton, 1989). Most text classification systems use the case-based approach (see Chapter 3). These systems require a training set of documents to build a statistical model, which is later applied to classify other documents. The differences among these systems differ are on the way they create the statistical representation of each document, and in the method used to create the classification model.

A successful system requires features in the document representation providing relevant information to the classification method (e.g., the relevant terms occurring in the documents). Since appropriate features are not always available, a common approach is constructing new ones, for instance, by combining old features in an efficient way (Pagallo and Hassler, 1990). An-

other common approach is to use domain specific knowledge to improve classification in restricted application domains. This Chapter presents WeBTC (Web Biological Text Classification), a new method for classifying biological text that generates new features from external data sources. WeBTC integrates the text with domain knowledge automatically extracted from biological Web resources, which are the source of new features. These new features are integrated in the document representation to improve the performance of classification methods. Given a collection of documents, WeBTC produces a richer representation of each document, based on related information extracted from external sources. If this information is valuable, classification will achieve a higher accuracy than simply using the text from the document.

The rest of this Chapter describes WeBTC, its experimental evaluation, the results obtained and analyses and discusses its performance.

5.1 Method

WeBTC relies on biological results stored in public databases available on the Web. It is motivated by the observation that most authors of recently published biomedical documents also submit their results to public databases. Therefore, these databases usually have their data associated with bibliographic information, which provides an important source for document classification. Since this information is stored in a structured form, it can be easily used in an automated system.

Input:

- A collection of documents with its content and its meta-data (e.g. title, authors, accession number in a bibliographic database);

- A biological database where information about the documents can be found.

Output:

- A statistical representation for the documents, where each document is represented by the number of occurrences of each term found in the database.

Procedure:

1. For each document, WeBTC identifies all the related database accession numbers. This information can be extracted by three different ways:
 - (a) Directly from the document content. Most authors present accession numbers in their documents, referencing the database where their results were submitted. It is not hard to find an accession number in the text, since they have a common format dependent on the database, e.g. two letters followed by 6 digits. Moreover, sentences with an accession number usually also reference the database common name.
 - (b) When the authors of a published document submit their results to a database, they often include the document identification. In this case, WeBTC only has to identify the database entries that cite the document, which is possible if the entered bibliographic information can be searched.
 - (c) When a database entry has no bibliographic information but mentions its source indirectly (e.g. the authors, the date, the laboratory, the technique), WeBTC matches these data against the

document's meta-data to infer that the document represents the information source of the database entry.

2. WeBTC retrieves the content of the database entries, and identifies the number of distinct terms mentioned on them.
3. For each document, WeBTC computes the occurrences of each term in its associated database entries.

Example

The document available in PubMed with the identifier *12803610*, contains the following sentence:

The sequence of the nramp cDNA was filed at the EMBL-Bank/GenBank/DDBJ Databases under the accession number AJ514946.

For this document, WeBTC's step 1 extracts the accession number *AJ514946* whose entry is available in the database GenBank. Besides other terms, this GenBank entry contains the term *Hordeum vulgare subsp. vulgare*, which is the name of the organism. Step 2 identifies this term, and step 3 counts at least one occurrence of the term. Therefore, the WeBTC output will contain a representation of the document where the feature representing the term *Hordeum vulgare subsp. vulgare* has at least one occurrence.

5.2 Assessment

WeBTC was experimentally evaluated for classifying biomedical documents on the BioText Task of KDD2002 Cup (Yeh et al., 2002). The task consisted on identifying which biomedical documents contained relevant experimental

results, and which were the gene products (transcripts and proteins) involved. This represents one stage of the curation process in FlyBase. FlyBase is a comprehensive database for information on the genetics and molecular biology of *Drosophila* (fruit fly). The curators take a set of documents and extract new relevant information reported on them. New relevant information means experimental results applicable to wild-type (non-mutated) fruit flies, which are not just merely citations of other documents. The complete curation process of FlyBase is a typical case of application of ProFAL, and the BioText Task of KDD2002 Cup represents the retrieval step of ProFAL.

The evaluation environment included a collection of documents about *Drosophila* genetics or molecular biology. For each document, the full content was provided as a raw text file, along with an XML template containing its identifiers and the list of the genes mentioned in it. The gene's names follow a standardised nomenclature, and a synonym list for each gene was provided. Other collections of data from biological databases publicly available on the Web could also be used, to better mimic real conditions.

Each participant had to submit the following items:

- For each document, a Boolean decision on whether there are relevant experimental results reported on it.
- For each document assumed to have relevant experimental results, the genes involved and the gene-product type (transcript, protein, or both).
- A ranked list of documents, sorted by the assurance degree of having relevant experimental results. The documents more likely to contain experimental results should be ranked higher than the documents with no experimental results.

Each output item was considered a sub-task that was evaluated sepa-

rately. The collection of documents was divided in two sets: the training set with 862 documents and the test set, with 213 documents. The training set was made available 6 weeks before the test set, and it was available for 2 weeks until the submission deadline. The expected output for each document in the training set was provided. Only 283 documents of the training set reported relevant experimental results. The output of these documents was extended with result evidences.

5.2.1 Setup

The implementation of WeBTC in this evaluation environment involved the retrieval of the meta-data of each document through its PubMed identifier. These identifiers were provided for each document. WeBTC used the following external biological databases:

MeSH: a collection of keywords for classifying documents;

GenBank: a repository of gene structure data.

The first step of WeBTC procedure, which associates each document with the MeSH terms, can be skipped, since PubMed already manually classifies each document with a set of MeSH terms. WeBTC retrieved the GenBank accession numbers in the documents' text and through the citations. The third approach was not implemented, i.e. WeBTC did not use the documents' meta-data to retrieve accession numbers. WeBTC was executed with MeSH terms and with the gene and protein information retrieved from GenBank. The result was three different statistical representations of each document. Their features were combined to integrate these representations into a single one, named WeBTC representation.

	Bag-of-words	WeBTC	Combined
TruePositives	41	19	15
FalsePositives	19	2	0
TrueNegatives	103	120	122
FalseNegatives	50	72	76

Table 5.1: Results achieved by WeBTC and the Bag-of-Words and Combined approaches.

A model was built from the WeBTC representations and another model from the bag-of-words representations. A combined model was implemented that only considers a document relevant if both models agree in doing so. The bag-of-words representation of each document was created by Bow, a toolkit for statistical language modelling, text retrieval, classification and clustering (McCallum, 1996). The stemming algorithm available in Bow increased the features quality in both WeBTC and bag-of-words representations. Given the statistical representations of each document, Bow built the models using the Naïve Bayes statistical classification method.

5.2.2 Results

WeBTC vs. Standard Approach

Table 5.1 presents the results obtained by the three models that predicted the classification of the 213 documents in the test set. TruePositives are the number of documents that a model correctly predicted to be relevant. TrueNegatives are the number of documents that a model correctly predicted to be irrelevant. FalsePositives are the number of documents that a model incorrectly predicted to be relevant. FalseNegatives are the number of documents that a model incorrectly predicted to be irrelevant.

Figure 5.1 compares the precision and recall obtained by the three models. Results show that WeBTC achieved a significantly higher precision. The

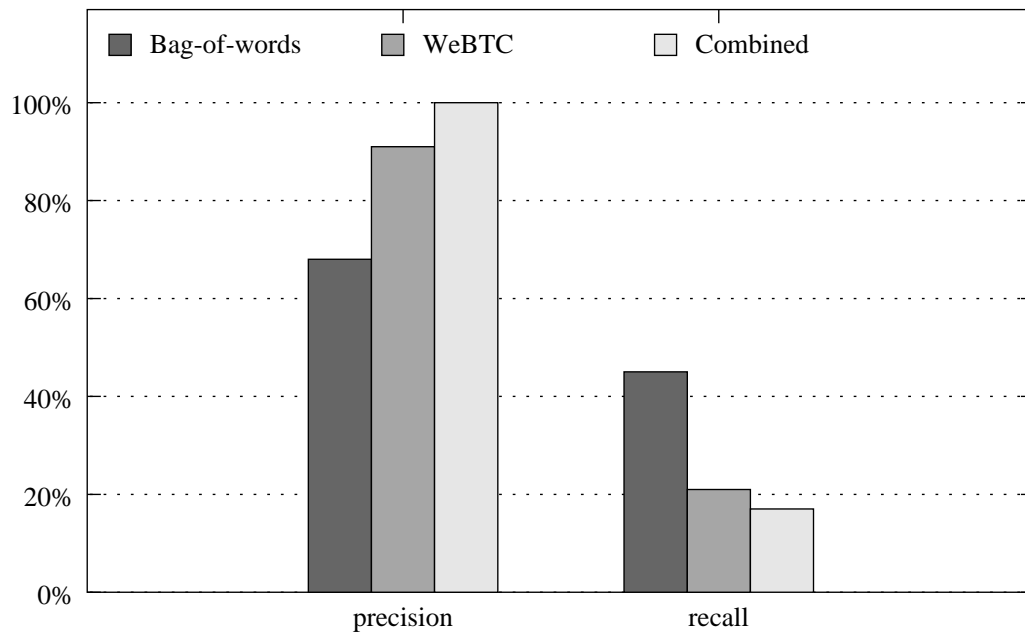


Figure 5.1: Precision achieved by WeBTC and the Bag-of-Words and Combined approaches.

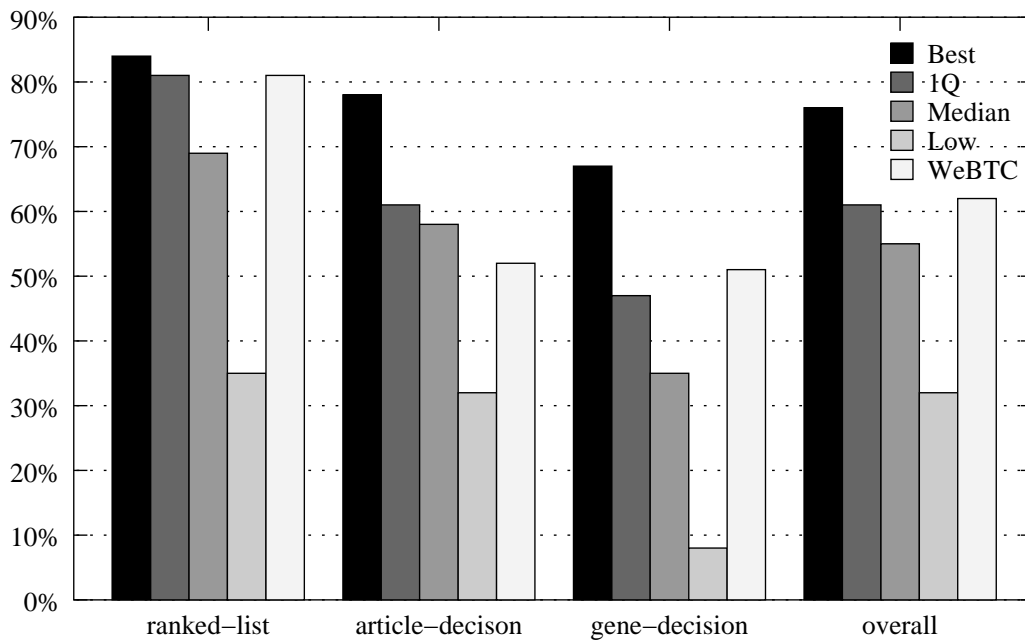


Figure 5.2: Scoring Results of the BioText Task of KDD2002 Cup.

combined model achieved 100% precision, however at the cost of a decrease in recall.

WeBTC vs. State-of-the-art Approaches

Since the combined model achieved a better performance, it was submitted to the BioText Task of KDD2002 Cup. The results of 32 state-of-the-art systems were provided by the BioText Task of KDD2002 Cup organisation committee, which applied a scoring method to evaluate each of the sub-tasks. They scored the ranked list by the ROC curve (Bradley, 1997), the document decision and the gene-product decision by the standard F-measure. The overall score was obtained by the sum of these three scores, normalised to a 0% to 100% range representing the efficiency of the systems.

Figure 5.2 shows the results for the three sub-tasks and the overall score. The *Best* values represent the highest score, which in this case was always obtained by the same team. The *1Q* values represent the score limit of the first quartile (Kenney and Keeping, 1962), i.e. the ninth highest score in this contest. The *Median* values represent the arithmetic average of all scores. The *Low* values represent the lowest score obtained. The *WeBTC* values represent the scores obtained by using WeBTC. The overall score of WeBTC was in the first quartile in two sub-tasks. The exception was in the document decision sub-task, where the score was even lower than the median. In this sub-task, WeBTC achieved a precision of 81% but a recall of only 38%.

5.3 Discussion

The main problem of WeBTC was the low recall. This happened because it was not able to retrieve information for all documents because of the small

number of external biological sources used. The period of time to implement WeBTC was small, otherwise WeBTC would have retrieved more information, since the databases would be more complete and more resources could be covered. Results obtained with information retrieved after the BioText Task of KDD2002 Cup deadline would have had a larger recall. This would not have constituted a fair evaluation, since database curators in real situations have also a deadline to classify documents. Thus, WeBTC needs to cover in due time a broader range of resources. On the other hand, for the documents with information available, WeBTC provided an accurate prediction, reaching 100% precision. The high levels of precision are useful for database curators, since they do not have to manually verify predictions of relevant documents.

The ClearForest and Celera team developed the winning system of the BioText Task of KDD2002 Cup (Regev et al., 2002). Their system was implemented through a rule-based approach (see Chapter 3). The rules were built specifically for the task with basis on domain knowledge, and were essentially sequences of terms to use in pattern matching. A team from Singapore obtained an honourable mention by developing a system based on feature extracting with a Naïve Bayes Classifier (Keerthi et al., 2002). However, their feature extraction was based on a set of keywords manually extracted from the training texts and on manual selection of positive examples. Another honourable mention was given to a team from UK (Ghanem et al., 2002). Their system was also based on feature selection and on statistical classification methods, but feature selection was also based on relevant keywords supplied by local domain experts.

All the systems described above use domain knowledge as a crucial component of their systems. The main conclusion retained from the BioText

Task of KDD2002 Cup was that statistical text classification systems reasoning without considering domain knowledge achieved poor results. WeBTC attempted to obtain domain-specific knowledge through information automatically extracted from external biological sources available on the Web. Its implementation, at an early prototypical stage, performed close to the best submissions, which were more mature and resorted to manually inserted domain knowledge.

5.4 Conclusions

This Chapter presented a novel approach for text classification involving automatic integration of extracted information from biological Web resources with common statistical text classification methods. WeBTC was developed based on this approach, and it was able to significantly increase the precision (reaching 100%) relative to standard classification methods. However, it obtained low levels of recall, because of the small number of documents for which information in the external databases was found. If more information had been retrieved, WeBTC would have achieved higher levels of recall maintaining its remarkable levels of precision.

The performance of WeBTC was also evaluated in the BioText Task of KDD2002 Cup versus state-of-the-art systems. Besides being developed in an early stage of this thesis, WeBTC achieved results close to well-established approaches using manually inserted domain knowledge. This substantiates the hypothesis of this thesis by showing that domain knowledge automatically acquired from external sources represents an efficient alternative to domain knowledge explicitly created by experts.

An annotation tool, such as ProFAL, can only perform well when it is

using the correct documents. WeBTC aims at selecting relevant documents from a large collection and therefore implements the first processing step of ProFAL. Finding the right documents is a crucial and time-consuming task. Therefore, an efficient text classification method, such as WeBTC, reduces both the time spend by curators and improves the performance of annotation methods, such as FiGO described in the following Chapter.

6

Extraction: FiGO

Text-mining systems for gene annotation are becoming an important tool for the development and curation of microarray, mass spectrometry and other biological databases. As an example, the GOA (Gene Ontology Annotation) project aims at identifying GO annotations to supplement the UniProt knowledgebase (Camon et al., 2004). They provide high-quality manual GO annotations, but manual curation is a time-consuming task that currently covers less than 5% of UniProt. Thus, the GOA database coverage mainly consists of uncurated annotations that are automatically generated and have a lower quality than manual annotations. Besides identifying novel annotations, text-mining systems can also be used to support the curation process by identifying evidence texts that substantiate the uncurated annotations.

This Chapter proposes FiGO (Finding Genomic Ontology), a novel unsupervised method to identify biological terms organised in a BioOntology in unstructured text. The method follows the approach proposed by this thesis, since it does not require any rules or training sets written by the user. It automatically acquires the domain knowledge from the nomenclature of a given BioOntology, by using the frequency of each word present in the nomenclature to calculate its relevance.

The rest of this Chapter describes FiGO, its implementation details, and presents and discusses the results achieved.

6.1 Method

FiGO assumes that the evidence content of a word measures its importance to identify a term in text. The evidence content is inversely proportional to the number of times the word appears in the names of all terms. The notion of evidence content derives from the definition of information content used by Resnik (1995). For instance, consider the GO term *punt binding*. If only the word *binding* is present in the text, the probability of the GO term being referenced is low, because *binding* is used in many other names. On the other hand, if only the word *punt* is present, then there is strong evidence that the GO term is mentioned in the text because this word is not part of any other name.

FiGO receives i) a BioOntology, *Ont*, and ii) a piece of text, *Txt*, as input. Each entry in *Ont* represents a biological term that can be assigned to genes. The output is the list of terms that FiGO detected in the given text. FiGO returns these terms ranked according to how strong is the evidence found in the text. For example, *Ont* can be the GO with each biological term representing a GO term, and *Txt* can be a sentence taken from a document.

The Words

FiGO derives a map between the terms and their names:

$$Names(term) = \{n_0, \dots, n_k\}, \quad (6.1)$$

where $term \in Ont$ and n_0, \dots, n_k are its name and synonyms in the BioOntology. If $term$ does not have synonyms, then $k = 0$ and $Names(term) = \{n_0\}$.

The set of words that compose a name n is given by:

$$Words(n) = \{w_0, \dots, w_l\}. \quad (6.2)$$

In addition, the set of words contained in a term $term$ is:

$$Words(term) = \{w \in Words(n) : n \in Names(term)\}. \quad (6.3)$$

Furthermore, the words of the BioOntology are

$$Words(Ont) = \{w \in Words(term) : term \in Ont\}. \quad (6.4)$$

Evidence Content

The evidence content of each word decreases with its frequency. The frequency of a word w is the number of terms that contain the word:

$$Freq(w) = \#\{term \in Ont : w \in Words(term)\}. \quad (6.5)$$

A word present in only one name has high evidence content. The maximum frequency is defined using the following equation:

$$MaxFreq = \max\{Freq(w) : w \in Words(Ont)\}. \quad (6.6)$$

$WordEC(w)$, the evidence content of a word w , is defined using the following equation:

$$WordEC(w) = -\log\left(\frac{Freq(w)}{MaxFreq}\right). \quad (6.7)$$

Since each name is composed of a set of words, the evidence content of a

name n is the sum of the evidence content of its words:

$$NameEC(n) = \sum_{w \in Words(n)} WordEC(w). \quad (6.8)$$

The evidence content of a term $term$ is defined as the highest evidence content of all its names:

$$EC(term) = \max\{NameEC(n) : n \in Names(term)\}. \quad (6.9)$$

Local Evidence Content

The input text is modelled as set of words:

$$Txt = \{w_0, \dots, w_l\}. \quad (6.10)$$

The local evidence content (LEC) measures how much of the name n is mentioned in the text Txt . LEC is the sum of the evidence content of those words, which are present in the text as well as in the name:

$$NameLEC(n, Txt) = \sum_{w \in (Txt \cap Words(n))} WordEC(w). \quad (6.11)$$

The LEC also measures how much the term $term$ is mentioned in the text Txt :

$$LEC(term, Txt) = \max\{NameLEC(n, Txt) : n \in Names(term)\}. \quad (6.12)$$

The LEC divided by the EC is a confidence level for the term $term$

occurring in the *Txt*:

$$Conf(term, Txt) = \frac{LEC(term, Txt)}{EC(term)}. \quad (6.13)$$

$Conf(term, Txt) \in [0, 1]$, since LEC is smaller than EC by definition.

If the confidence level is larger than a given threshold $\alpha \in [0, 1]$, then *term* is considered to occur in *Txt*:

$$Conf(term, Txt) \geq \alpha. \quad (6.14)$$

If $\alpha = 1$, the complete name has to appear in the text to be selected. Thus, the α parameter is used to tune recall and precision of FiGO. An increase in α increases precision, a decrease in α increases recall. $Conf(term, Txt)$ is used to rank the returned terms, and represents the evidence strength found in the text for each biological term.

Example

Given a term *t* with:

$$\begin{aligned} Names(t) &= \{punt\ binding, punt\ function\}, \\ Freq(punt) &= 1, \\ Freq(binding) &= 4, \\ Freq(function) &= 8, \\ MaxFreq &= 16. \end{aligned}$$

Then, we have:

$$\begin{aligned} WordEC(punt) &= -\log(1/16) = 4, \\ WordEC(binding) &= -\log(4/16) = 2, \\ WordEC(function) &= -\log(8/16) = 1, \end{aligned}$$

$$\begin{aligned} \text{WordEC}(\text{punt binding}) &= 4 + 2 = 6, \\ \text{WordEC}(\text{punt function}) &= 4 + 1 = 5, \\ \text{EC}(t) &= \max\{6, 5\} = 6. \end{aligned}$$

Consider the following pieces of text:

$$\begin{aligned} \text{Txt}_1 &= \text{The protein has a binding activity}, \\ \text{Txt}_2 &= \text{The protein has a punt activity}, \\ \text{Txt}_3 &= \text{The protein has a punt binding activity}. \end{aligned}$$

Since we have:

$$\begin{aligned} \text{LEC}(t, \text{Txt}_1) &= 2, \\ \text{LEC}(t, \text{Txt}_2) &= 4, \\ \text{LEC}(t, \text{Txt}_3) &= 6, \end{aligned}$$

then we have:

$$\begin{aligned} \text{Conf}(t, \text{Txt}_1) &= 1/3, \\ \text{Conf}(t, \text{Txt}_2) &= 2/3, \\ \text{Conf}(\text{term}, \text{Txt}_3) &= 1, \end{aligned}$$

which means that FiGO will decide that t occurs in:

$$\begin{aligned} \text{Txt}_1 &\text{ when } \alpha \leq 1/3, \\ \text{Txt}_2 &\text{ when } \alpha \leq 2/3, \\ \text{Txt}_3 &\text{ when } \alpha \leq 1. \end{aligned}$$

The cases of Txt_1 and Txt_2 show how FiGO gives more importance to rare words to identify the terms in a given text.

6.2 Assessment

This Section describes the FiGO implementation used when preparing the submission to BioCreAtIvE tasks 2.1 and 2.2 (Blaschke et al., 2005). Given a document and a GO annotation, task 2.1 consisted of identifying the text in the document that provided evidence for the annotation. Given a document and the number of GO annotations to find for each GO class, task 2.2 consisted of identifying the GO annotations and extracting an evidence text for each of them from the document.

GO pre-processing

FiGO used the GO BioOntology, considering its terms as the terms to identify. FiGO identified the set $Words(GO)$, and removed from this set all the stop words, such as *in* or *on*. FiGO then computed the evidence content of each word, name, and finally of each term. FiGO also computed the annotation frequency of each GO term as the number of times the term and its descendants in the GO hierarchy were annotated in GOA. The most frequently annotated terms represent general GO terms, such as *protein*, and *binding*. These terms were discarded in the extraction of annotations from text.

The Text

FiGO parsed the file given for each document and structured the text in sentences. Each sentence represented a piece of text from where FiGO identified GO terms.

In task 2.1, the submitted sentences were the ones in the ranked list returned by FiGO for the given term. In the case of having multiple sentences,

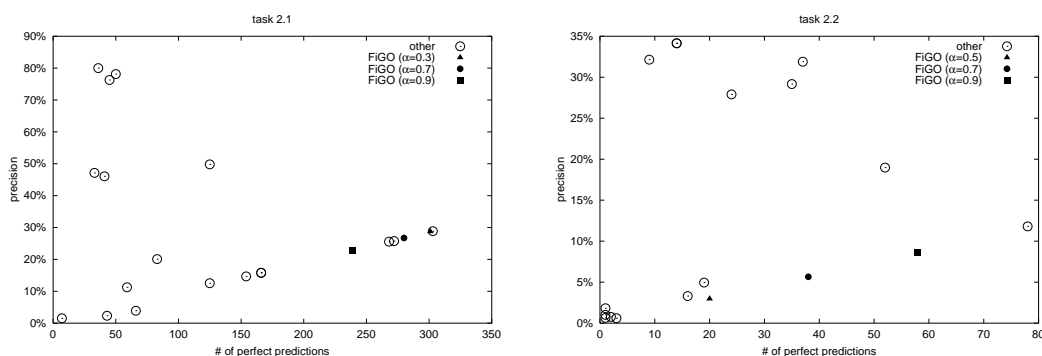


Figure 6.1: These charts compare the quality of the predictions returned by FiGO with all the other submissions to BioCreAtIvE tasks 2.1 and 2.2. For each submission, the charts show the precision versus the number of perfect predictions identified. The precision is the number of perfect predictions over the number of predictions submitted.

the submitted sentence was the one with the highest rank and mentioning the protein. In the case of not having any sentence, the submitted sentence was the one returned by FiGO for the most similar term. The similarity between terms was calculated using the semantic similarity measure proposed by Lin (see Appendix B). In this task, FiGO was executed three times with α assigned to 0.3, 0.7 and 0.9, resulting in three different submissions.

In task 2.2, the submitted sentences were the ones in the ranked list returned by FiGO that mentioned the protein. Then, generic terms were discarded by only submitting the sentences containing the rarest annotated terms. In this task, FiGO was executed three times with the α assigned to 0.5, 0.7 and 0.9, resulting in three different submissions.

A naïve method based on exact matching identified the proteins in the text. The method consider that a sentence mentions a protein if it contains all the words of its name or synonym. The name and synonyms of each protein were collected from the UniProt database.

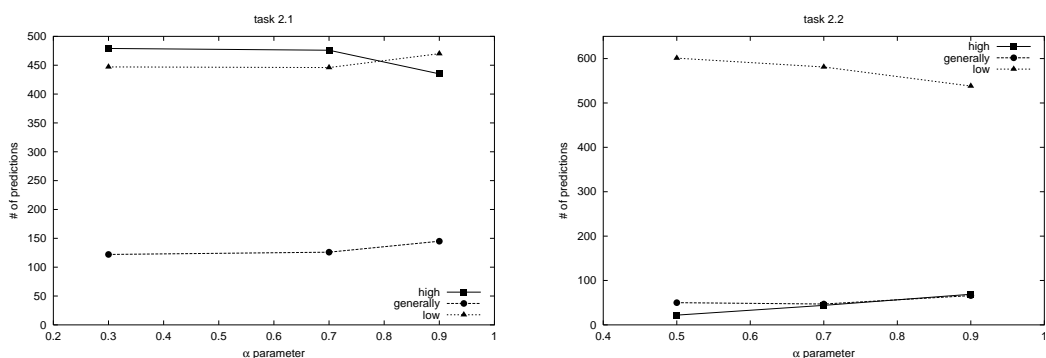


Figure 6.2: A prediction was considered perfect when it provided high evidence of both GO term and protein. Thus, a prediction could provide high evidence of the GO term without being perfect. These charts disregard the protein evaluation and show the number of predictions submitted by FiGO to BioCreAtIvE tasks 2.1 and 2.2, which provided a high, generally and low evidence of the GO term for each value of α used.

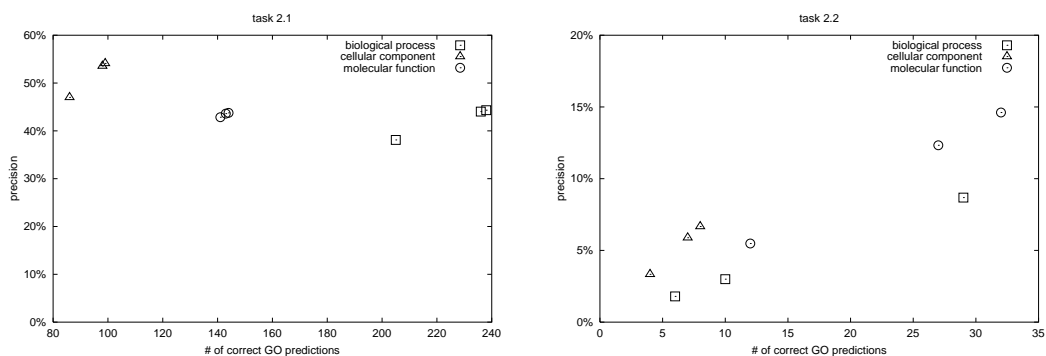


Figure 6.3: For each aspect of the GO hierarchy, these charts compare the performance of the three submissions of FiGO to BioCreAtIvE tasks 2.1 and 2.2. For each aspect, the charts show the precision versus the number of correct GO predictions identified by each submission. The precision is the number of correct GO predictions over the number of predictions submitted. In task 2.2, raising α increases precision in all aspects, while in task 2.1 has the opposite effect.

6.2.1 Results

In the BioCreAtIvE task 2, each submitted prediction had a GO term and a protein evaluation. Both evaluations assigned a high, generally or low score to the prediction. High score means that the predicted evidence supports a correct GO term or protein. A generally score means that the predicted

evidence supports a related GO term or protein. Low score means that the predicted evidence does not support a correct GO term or protein. A prediction was considered perfect when both the GO and protein evaluation assigned a high score to it.

Figure 6.1 shows the performance of FiGO in tasks 2.1 and 2.2. It compares its precision and number of perfect predictions with all the other submissions. For each submission, the charts show the precision versus the number of perfect predictions identified. The precision is the number of perfect predictions over the number of predictions submitted.

In task 2.1, the best performance of FiGO was obtained using $\alpha = 0.3$, which achieved a large number of perfect predictions and a precision of almost 30%. On the other hand, in task 2.2 the best performance of FiGO was obtained using $\alpha = 0.9$, which achieved a significant number of perfect predictions and precision of almost 10%.

A prediction was considered perfect when it provided high evidence of both the GO term and associated protein. Thus, a prediction could provide high evidence of the GO term without being perfect. Figure 6.2 shows the GO evaluation of FiGO predictions for the values of α used in tasks 2.1 and 2.2. The charts disregard the protein evaluation and show the number of predictions submitted by FiGO to BioCreAtIvE tasks 2.1 and 2.2, which provided a high, generally and low evidence of the GO term for each value of α used. The manipulation of the α parameter had a different effect on each task. In task 2.1, FiGO obtained better results using a smaller α value. On the other hand, in task 2.2 the increase of α implied a better performance of FiGO.

Figure 6.3 compares the performance of FiGO in each aspect of GO. For each aspect of the GO hierarchy, the charts compare the performance

of the three submissions of FiGO to BioCreAtIvE tasks 2.1 and 2.2. For each aspect, the charts show the precision versus the number of correct GO predictions identified by each submission. The precision is the number of correct GO predictions over the number of predictions submitted. In this Figure, a prediction was considered a correct GO prediction when the GO evaluation assigned a high score to it. In task 2.1, the best performance of FiGO was in the *biological process* aspect. On the other hand, in task 2.2 the best performance of FiGO was in the *molecular function* aspect.

6.3 Discussion

FiGO achieved a good performance when compared with the other submissions. In both tasks, FiGO almost defined the highest number of correct predictions, but its precision was far from the best results. However, the submissions with higher precision were composed by fewer predictions than requested. FiGO chose to always submit the requested number of predictions, even when they had a low confidence score.

Since the core of FiGO was the identification of GO terms, a significant part of the predictions was not considered perfect just because of the protein evaluation. For example, in task 2.1 with $\alpha = 0.3$, the GO evaluation assigned a high score to 479 predictions (see Figure 6.2). However, only 301 of them were considered perfect (see Figure 6.1). This means that 178 out of 479 predictions (37.2%) were not considered perfect because they did not provide high evidence of the protein. In addition to this major problem, FiGO also had the following limitations:

1. in task 2.1, it predicted about 20 obsolete GO terms;

2. it did not filter the GO terms that could not be annotated with Human proteins (e.g. germination);
3. it selected sentences from irrelevant Sections (e.g. *Material and Methods*);
4. sometimes just one sentence is not enough to support an annotation. For instance, the protein and the term are sometimes in the same paragraph, but not in the same sentence;
5. it did not account for the number of times a term occurs in the text;
6. it did not account for the word order in the name;
7. in task 2.2, it predicted GO terms out of context.

The first two limitations could be easily solved before BioCreAtIvE, but they were not identified before the submission. On the other hand, the last five limitations represent important topics of research that deserve further study. The performance in task 2.2 was lower than in task 2.1 mainly because of the last problem on the list. To discard terms out of context more domain knowledge about the proteins and the documents would be useful. An effective approach would be the integration of domain knowledge from publicly available resources.

In task 2.1, the GO terms with higher precision occurred in the BioLiterature exactly as described in GO, such as *cell proliferation*. This particular GO term had the highest precision with 11 high and 1 low scores assigned. The GO terms with lower precision were the ones whose name was composed by words with low evidence content, such as *regulation of transcription*. This particular GO term had the lowest precision with 1 high and 8 low scores assigned.

In task 2.2, the GO terms with higher precision were generic terms, such as *binding*, and those whose name had high evidence content, such as *galactose 3-O-sulfotransferase activity*. This last GO term had the second highest precision with 4 high and 2 low scores assigned. The GO term *binding* had the highest precision with 20 high and 3 low scores assigned. The GO terms with lower precision were the ones whose name was composed by words with low evidence content or multiple meanings, such as *receptor activity*. This particular GO term had the lowest precision with 1 high and 8 low scores assigned, because *activity* has low evidence content and *receptor* can be used to mention other protein. For example, in UniProt there are more than 20,000 proteins whose name contains the word *receptor*.

Figure 6.3 shows that in task 2.1 it was easier to find evidence for GO terms in the *biologic process* aspect. This can be explained because these terms use specific names. On the other hand, the same Figure shows that in task 2.2 it was easier to predict terms in the *molecular function* aspect. This can be explained because normally there are more occurrences of these terms in the documents.

The reason for having better results using a smaller α value in task 2.1 is that there were a large number of terms not explicitly mentioned in the text. Some sentences were correctly selected when only less than 70% of the term's name appeared in text. On the other hand, for smaller values of α , FiGO identified more terms out of context. Thus, in task 2.2, the selection of terms with a larger α turned up to be an effective approach to predict which relevant terms were mentioned.

6.4 Conclusions

This Chapter presented FiGO, a novel unsupervised method for recognizing biological terms in unstructured text, involving the evidence content of their names. FiGO does not require training data, since it computes the evidence content based on the nomenclature of a BioOntology that structures the terms. Therefore, the use of FiGO represents minimal human intervention.

FiGO was designed for recognizing terms, not for extracting annotations. However, it has still obtained a good performance in BioCreAtIvE when compared with other submissions. The evaluation raised a set of problems that should be addressed in further developments. The main limitation of FiGO in BioCreAtIvE was in the protein identification, since it used a naïve method for this task. If a more effective method was used, FiGO would likely have achieved an even higher performance (Fukuda et al., 1998).

The performance of FiGO demonstrated that it provides an effective method to recognise terms in BioLiterature and to improve the performance of automatic annotation systems, such as ProFAL. It also substantiated the approach proposed by this thesis, since FiGO was able to obtain good results without resorting to manually inserted domain knowledge.

Despite the good performance of FiGO when compared to other submissions, its accuracy is still not acceptable to curators. This explains the reluctance of many curators to the use of text-mining tools. Therefore, these tools would benefit from novel methods that could automatically discard misannotations, such as CAC, which is described in the next Chapter.

7

Validation: CAC

The large amount of biological data available nowadays has transformed the traditional way of conducting research and development in the life sciences. Traditional functional characterisation of genes and proteins cannot cope with the large amount of sequences being produced. Therefore, a significant number of genes and proteins have been functionally characterised by automated tools, which extrapolate functional annotations from similar sequences. However, these tools have also produced a significant number of misannotations that are now present in the databases (Devos and Valencia, 2001). Some of these tools have been extrapolating new annotations from misannotations and are therefore spreading the errors. This happens because most databases do not distinguish between extrapolated and curated annotations. Functional characterisation is not normally linked to the experimental evidence that substantiates it, which makes it difficult to judge if it is correct.

Many databases are using GO terms to annotate their proteins. For example, the GOA (Gene Ontology Annotation) database provides GO annotations to supplement the UniProt (Universal Protein Resource) (Camon et al., 2004). UniProt is a universal repository of protein sequence and functional data (Apweiler et al., 2004). GOA provides high-quality manual GO annotations, but manual curation is a time-consuming task that currently covers less than 5% of UniProt. The manual processing capacity for gene and

protein characterisation is overloaded by the increasingly larger amounts of literature to analyse. Thus, the GOA database mainly consists of automated annotations that have a lower quality than manual annotations.

This Chapter proposes a new approach to validate the automated annotations and therefore improve their accuracy. The approach uses the large amount of publicly available information to compare automated annotations to preexisting curated annotations. The manual annotation methodology adopted by curators, who also use preexisting annotations as a guide to evaluate automated annotations, inspired this approach. The underlying intuition is that automated annotations having similar curated annotations should also be correct. Similar annotations mean annotations with similar proteins and similar GO terms. This is supported by the dogma of Molecular Biology, which postulates that proteins with similar sequence should also have similar biological activities (Lord et al., 2003a). CAC (Correlate the Annotations' Components), was developed based on the proposed approach. It is a novel heuristic method to discard misannotations identified by automated systems. CAC requires minimal human intervention, since it takes advantage of publicly available domain knowledge to score each automated annotation according to previously curated annotations.

The remainder of this Chapter describes CAC in detail, presents the experimental evaluation of CAC and discusses the obtained results.

7.1 Method

Algorithm 1 outlines CAC, which assigns a confidence score to $a_{predicted}$, an annotation predicted by an automated system given as input. CAC also

Algorithm 1: CAC

Input: $a_{predicted}$, an annotation predicted by an automated system;
 $\mathcal{A}_{curated}$, set of previously curated annotations.
Output: $confidence \in [0, +\infty]$, confidence score of the predicted annotation.

- 1: $confidence(a_{predicted}) = 0$
- 2: $(g_{predicted}, p_{predicted}) = a_{predicted}$
- 3: $\mathcal{G}_{curated} = \{g : \exists p (g, p) \in \mathcal{A}_{curated}\}$
- 4: **for all** $g_{curated} \in \mathcal{G}_{curated}$ **do**
- 5: $\mathcal{P}_{curated} = \{p : (g_{curated}, p) \in \mathcal{A}_{curated}\}$
- 6: $geneSim = geneSim(g_{predicted}, g_{curated})$
- 7: $propSim = \sum_{p_{curated} \in \mathcal{P}_{curated}} propSim(p_{predicted}, p_{curated})$
- 8: $confidence(a_{predicted}) += geneSim \times propSim$
- 9: **end for**
- 10: $SG = similarGenes(g_{predicted}, \mathcal{G}_{curated})$
- 11: $confidence(a_{predicted}) = \frac{confidence(a_{predicted})}{SG}$

receives as input $\mathcal{A}_{curated}$, a set of preexisting curated annotations collected from public databases, e.g. GOA.

CAC starts by assigning a zero confidence score to the predicted annotation (line 1). Next, CAC collects all the genes in the set of curated annotations (line 3). For each curated gene, CAC collects the properties annotated to it (line 5). Next, CAC calculates the similarity between the curated and the predicted genes (line 6), and calculates the similarity between the predicted property and each property annotated to the curated gene (line 7). CAC increments the confidence of the predicted annotation by the product of the gene similarity and the sum of all property similarities (line 8). Thus, the confidence only increases if both the gene similarity and at least one property similarity are larger than zero, i.e., if they are similar genes and have been annotated with at least one similar property.

However, the $\mathcal{A}_{curated}$ set can contain groups of similar genes that are over-represented. In this case, the predicted annotations that contain genes with

a large number of similar curated genes will tend to have higher confidence scores. To overcome this problem, CAC calculates the number of curated genes similar to the predicted gene (line 10), and employs it as a damping factor (line 11). This factor reduces the effect of the amount of similar curated genes in the confidence score calculation.

CAC returns a confidence score of $a_{predicted}$ being correct. To filter the annotations predicted by an automated system, CAC scores each predicted annotation and discards those scored below a confidence threshold (CT). CAC is able to trade precision against recall by manipulating CT . Raising CT increases precision and decreases recall, lowering CT has the opposite effect.

CAC cannot score annotations without similar curated annotations. When the given predicted annotation has no similar curated genes ($SG = 0$), CAC assigns a confidence score of $+\infty$ to it. This means that the predicted annotation will never be filtered independently of the threshold used. Therefore, CAC does not discard new knowledge; instead, it gives the curators the opportunity to manually verify these potentially novel annotations.

Gene Similarity

The most popular way to calculate the similarity between two genes is by comparing their sequence (Attwood and Parry-Smith, 1999). However, sequence similarity is not the only kind of structural similarity that can be computed between two proteins. Family similarity is also a structural similarity of a higher level than sequence similarity. Each family describes a set of related proteins, which can have identical molecular functions, are involved in the same process, or act in the same cellular location. Classifying proteins in families has been a common technique to organise them according

to their biological role. For example, the most successful large-scale effort for increasing the coverage of GO annotations within the UniProt database is based on the exploitation of family annotations (Camon et al., 2004). Unlike standard sequence similarity methods, family categorisation is normally based on experimental results about protein domains, which represent some evolutionarily conserved structure and have implications on the protein's biological role.

geneSim was implemented as the number of shared Pfam families. Pfam is a structural classification scheme, which provides a set of protein domains and families, designed for well-established uses, including genome annotation (Bateman et al., 2004). The UniProt database provides family assignments, where each protein is assigned to a set of Pfam families. This implementation can be improved by taking in account the sequence related to each Pfam family. For example, the length of the sequence and the percentage of similarity may constitute important factors to calculate the *geneSim* function.

Property Similarity

CAC assumes that two properties are similar if one of them subsumes the other or if they have a common parent in the functional classification scheme, e.g. GO. To calculate the degree of similarity between properties, CAC can use a semantic similarity measure that combines the structure and content of a BioOntology with statistical information from corpus (Resnik, 1995). Recent projects investigated the use of semantic similarity measures over GO (Lord et al., 2003b; Couto et al., 2005b). Their results demonstrated the feasibility of a semantic similarity measure in a biological setting.

propSim was implemented using the measure proposed by Jiang&Conrath, which is one of the most efficient semantic similarity measures (Jiang and

Conrath, 1997; Budanitsky and Hirst, 2001). Jiang&Conrath defined the semantic distance of two concepts in a corpus as the difference between their information content and the information content of their most informative common ancestor. The information content of a concept is inversely proportional to its frequency in the corpus. Concepts that are frequent in the corpus have low information content. For example, the stop words (such as *the*) that occur almost everywhere in the text normally provide little semantic information. The information content of a GO term was calculated as the number of proteins annotated with it. The ancestor of two GO terms having the largest information content was considered the most informative common ancestor of both terms.

Example

In the subtask 2.2 of BioCreAtIvE, the participants annotated the protein *Lipid phosphate phosphohydrolase 1* to the GO terms *membrane* and *mRNA metabolism* (Blaschke et al., 2005). However, only the assignment of *membrane* is correct. Below the results obtained by CAC for these two annotations are described.

The protein *Lipid phosphate phosphohydrolase 1* belongs to the *PF01569* family. For the annotation of this protein to *membrane*, CAC found 91 curated proteins from the *PF01569* family ($geneSim = 1$) that were annotated to similar GO terms ($propSim > 0$) in GOA. From these 91 proteins, 21 were annotated to the same term. For example, the protein *Lipid phosphate phosphohydrolase 2* belongs to the *PF01569* family ($geneSim = 1$) and is annotated to *membrane* and *integral to membrane*, which results in $propSim = 1.445297776$. The confidence score resulted from these 91 pro-

teins is 53.09, but since the *PF01569* family contains 630 proteins ($SG = 629$), CAC returned $\frac{53.09}{639} \approx 0.08$.

On the other hand, for the annotation of the protein *Lipid phosphate phosphohydrolase 1* to *mRNA metabolism*, CAC only found one curated protein (*HH1165*) from the *PF01569* family ($geneSim = 1$) that was annotated to a similar GO term (*metabolism*) ($propSim = 0.1$) in GOA. Thus, in this case CAC returned $\frac{0.1}{639} \approx 0.0002$.

7.2 Assessment

CAC was tested to find how effectively it could discard the misannotations submitted to BioCreAtIvE independently of their evidence text. CAC scored each submitted annotation individually ($a_{predicted}$), using the GOA annotations as the curated set of annotations ($\mathcal{A}_{curated}$). The annotations submitted to BioCreAtIvE and the GOA annotations are both publicly available on the Web¹². However, in the publicly available information there is no reference to the author of each annotation submitted to BioCreAtIvE. It is not even possible to know which annotations were submitted by the same system.

It was decided not to increase the confidence of a predicted annotation based on curated annotations to the same protein, i.e., the protein $g_{predicted}$ was discarded from $\mathcal{G}_{curated}$. This way, CAC was restricted to score each predicted annotation based only on curated annotations to similar but distinct proteins. This restriction ensures a fair evaluation of CAC by checking if CAC copes with proteins having no previously curated annotations.

The restriction increased the number of proteins for which it was not

¹http://www.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/results/data/

²ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz

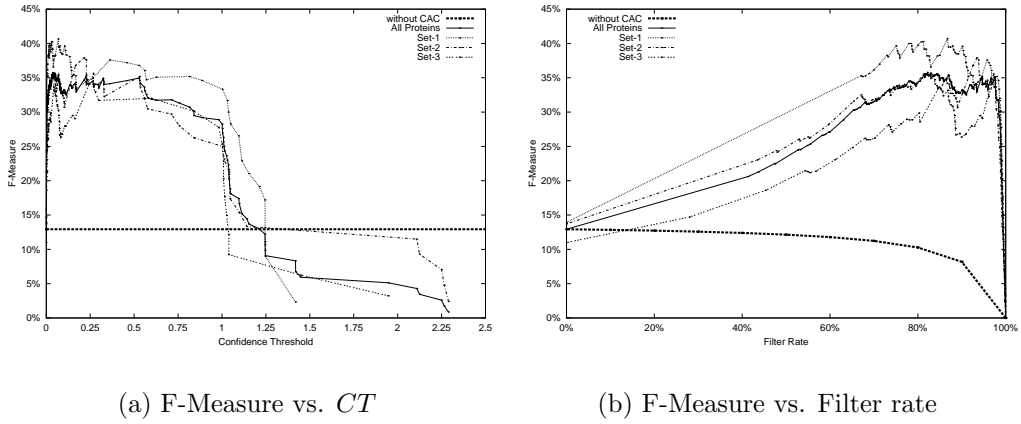
Set	$\#annotations$	$\#proteins$	$max(SG)$	$min(SG)$	\overline{SG}
Set-1	1135	30	583	5	223.7841
Set-2	1101	25	1762	613	1077.7221
Set-3	1049	22	11605	1855	3098.9790

Table 7.1: Statistics of the three sets of annotations created according to the number of similar curated proteins per annotation (SG). The statistics include the number of annotations, the number of distinct predicted proteins, and the maximum, minimum and average of SG for each set.

possible to obtain similar proteins, i.e., having $SG = 0$. However, only 455 out of the 3740 predicted annotations did not have a similar protein in the December 2004 release of GOA. These novel annotations have a precision of 7%, i.e., only 32 of them were correct. The assumption that supports CAC is not applicable to these novel annotations, thus scoring these annotations is out of CAC objectives. CAC does not discard these annotations, since it assigns an infinite score to them. Therefore, in the first part of the evaluation these annotations were disregarded, but they were included in the end to show the overall impact of CAC on the curation process.

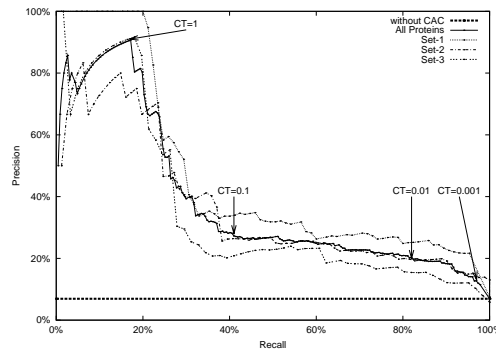
The 3285 annotations having $SG > 0$ assign 1239 distinct GO terms to 77 UniProt proteins. The 77 proteins were assigned to 87 distinct Pfam families with an average of 1.6 families per protein. These 87 families contained 64863 distinct proteins. Thus, each protein had $\frac{64863}{87} \times 1.6 = 1192.9$ similar curated proteins on average.

To compare the performance of CAC when applied to over-annotated or under-annotated proteins, the 3285 annotations were divided in three different sets (*Set-1*, *Set-2* and *Set-3*) according to the number of similar curated proteins (SG). Table 7.1 shows statistical information about each set.



(a) F-Measure vs. CT

(b) F-Measure vs. Filter rate



(c) Precision vs. Recall

Figure 7.1: Accuracy of the annotations retained by different confidence thresholds (CT) after running CAC. The *All Proteins* lines represent all the 3285 annotations. The *Set 1* and *Set 3* lines represent the annotations with the smallest and the largest number of similar curated proteins, respectively. The *Set-2* lines represent all the other annotations not present in *Set 1* and *Set 3*. The *without CAC* baselines represent the original annotations without using CAC. In chart (a), the baseline shows the F-Measure when none of annotations is filtered. In the other charts, the baselines assume a random model to filter the annotations, i.e., having a constant precision for any filter rate.

7.2.1 Results

Each distinct confidence score was used as a confidence threshold to obtain different subsets of the 3285 predicted annotations. For each confidence threshold, the resulting subset contains all the annotations with a confidence

score not below the threshold. For a zero confidence threshold, the subset contains all the predicted annotations, since none of them is discarded. As the confidence threshold increases, the size of the subset decreases. For each subset, it was calculated: the precision, representing the fraction of correct annotations in the subset; the recall, representing the number of correct annotations in the subset over the number of correct annotations in the original set; and the F-measure $= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$, representing the trade-off between precision and recall. Note that if we replace CAC by a random model to filter the annotations, the precision would remain constant. For instance, if we select at random 25% of the annotations in the original set, it is predictable that the selected annotations also contain 25% of the correct annotations in the original set.

Only 227 out of the 3285 annotations submitted to BioCreAtIvE were considered correct, a precision of 6.9%. The real recall is unknown, since the organisation of BioCreAtIvE did not measure it. Thus, we can assume a recall of 100% for the original set of annotations. Note that CAC cannot increase recall. As a filter, it does not generate new annotations.

Figure 7.1(a) shows the F-measure for different confidence thresholds. For confidence thresholds smaller than one, the chart shows that the use of CAC to discard annotations is beneficial by achieving a substantial improvement in F-measure. The F-measure achieves its maximum value when the confidence threshold is around 0.1. Figure 7.1(c) shows the precision and recall obtained for different confidence thresholds. With a few exceptions, we have a steadily increase in precision as we increase the confidence threshold.

Table 7.2 shows the accuracy of the predicted annotations when not using CAC ($CT = 0$), and the accuracy of the subsets of annotations retained by different confidence thresholds. Besides the precision, recall and F-measure,

CT	Filter Rate	#correct	#incorrect	Precision	Recall	F-measure	Misannotations Discarded
0	0%	227	3058	6.9%	100%	12.9%	0%
0.001	47.5%	219	1506	12.7%	96.5%	22.4%	50.8%
0.01	72%	186	733	20.2%	81.9%	32.5%	76%
0.1	90%	92	235	28.1%	40.5%	33.2%	92.3%
1	98.7%	39	4	90.7%	17.2%	28.9%	99.9%

Table 7.2: Results obtained by filtering the 3285 annotations using different confidence thresholds.

Filter Rate	All Proteins		
	Precision	Recall	CT
0%	6.9%	100%	0
70%	19.3%	84.6%	0.008
80%	22.6%	67%	0.025
90%	27.3%	41%	0.094
95%	40.6%	29.5%	0.235

Filter Rate	Set-1			Set-2			Set-3		
	Precision	Recall	CT	Precision	Recall	CT	Precision	Recall	CT
0%	7.5%	100%	0	7.4%	100%	0	5.8%	100%	0
70%	22.6%	90.6%	0.007	19.8%	81.5%	0.008	15.4%	82%	0.008
80%	27.3%	72.9%	0.028	23.3%	64.2%	0.028	18.4%	67.2%	0.018
90%	32.5%	47.1%	0.091	25.6%	38.3%	0.102	20.8%	36.1%	0.083
95%	40.6%	30.6%	0.263	40%	29.6%	0.243	30.4%	27.9%	0.162

Table 7.3: Results obtained by filtering the 3285 annotations using different filter rates.

the Table shows the number of correct and incorrect annotations that were not discarded by CAC, and the percentage of misannotations discarded by CAC from the original set. For example, by using $CT = 0.001$ CAC discarded 50.8% ($\frac{3058-1506}{3058}$) of the misannotations, maintaining 96.5% ($\frac{219}{227}$) of the correct annotations.

The confidence threshold has no biological meaning to curators. They simply would like to discard a given amount of annotations to speedup the curation process without losing a significant part of valuable information. This can be done by increasing CT until a defined filter rate is reached. The filter rate means the percentage of annotations that are discarded by CAC from the original set. For example, a filter rate of 90% means that only 10% of the original annotations were retained. Figure 7.1(b) shows the F-measure obtained by CAC for different filter rates. The chart shows that the use of CAC to discard annotations is beneficial by achieving a steady improvement

in F-measure as we increase the filter rate, except for filter rates larger than 99% ($CT > 1$). Table 7.3 shows the precision and the recall of the different sets of annotations over different filter rates, together with the selected CT in each set. The standard deviation of both recall and precision is always smaller than 5% for the same filter rate, even with a standard deviation of 0.8% in precision in the original sets. The selected CT is almost the same in all sets, except in the *Set-3* where in some cases CT is about 1/3 smaller.

7.3 Discussion

The increase in precision is already a positive result to GOA curators, since they primarily require high precision in an automated annotation system. In this experiment, CAC increased precision at the cost of a low decrease in recall. The trade-off between precision and recall is worth it, as it is shown by the increase in the F-measure. This is always true except for filter rates larger than 99% ($CT > 1$), because recall decreases and precision is not improved. For such high confidence thresholds, there are still some misannotations not discarded. For example, CAC assigned a high confidence score to the annotation that assigns the GO term *kinase activity* to the protein *Sulfate transporter 1.2*, but this annotation is not in GOA. However, the GO term *protein kinase activity* is annotated to the same protein in GOA. Since the term *kinase activity* is a generalisation of *protein kinase activity*, the predicted annotation is correct but still not of interest to curators.

From 3058 misannotations, four remain with a confidence threshold of one. These four annotations assign generic GO terms to proteins. These annotations were considered incorrect, because they are not defined in GOA. However, they are correct, but too generic to be of interest to curators. For

large confidence thresholds many of these generic annotations remain. Thus, by considering generic annotations as correct, the performance of CAC would increase, but this would not reflect the curators' interest for precise and specific annotations. Nevertheless, it is undesirable to discard these generic annotations, since the evidence substantiating them may be of interest to curators.

The participant of BioCreAtIvE who achieved the largest precision predicted 41 annotations, 14 of which were correct. Using a confidence threshold of 1, CAC selected 43 annotations, 39 of which were correct. On the other hand, the participant who achieved the largest recall predicted 661 annotations, 78 of which were correct. Using a confidence threshold of 0.1, CAC selected 327 annotations, 92 of which were correct. Therefore, by proper adjustment of the confidence threshold we can use CAC to outperform each individual submission to BioCreAtIvE.

For a small decrease in recall, CAC was able to obtain a large improvement in precision, since annotations that clearly do not satisfy the correlation between structure and function are normally incorrect. Unfortunately, there are exceptions. Using a confidence threshold of 0.001, CAC discarded 8 out of 227 correct annotations. For these eight annotations, CAC could not find similar annotations mainly because of the restriction that discarded curated annotations to similar but distinct proteins. When CAC was tested without this restriction, 47% of the misannotations were discarded maintaining all the correct annotations, i.e., a two-fold increase in precision maintaining 100% recall. This restriction was applied to ensure a fair evaluation of CAC. However, in a real application setting, this restriction would not be applied and therefore obtain a higher performance. It is expected that, as the scientific

CT	Filter Rate	#correct	#incorrect	Precision	Recall	F-measure	Misannotations Discarded
0	0%	259	3481	6.9%	100%	13%	0%
0.001	41.7%	251	1929	11.5%	96.9%	20.6%	44.6%
0.01	63.3%	218	1156	15.9%	84.2%	26.7%	66.8%
0.1	79.1%	124	658	15.9%	47.8%	23.8%	81.0%
1	86.7%	71	427	14.3%	27.4%	18.8%	87.7%

Table 7.4: Results obtained by filtering all the 3740 annotations using different confidence thresholds.

community produces better classification schemes, CAC will also improve its performance.

The results of the three different sets of annotations show that CAC is not biased toward proteins with a large number of similar curated proteins. In Figure 7.1, the results of these sets were uniform over all the confidence thresholds. The small differences are due to different precision values of each original set. The Set 1 of under-annotated proteins has the highest precision (7.5%) and the Set 3 of over-annotated proteins has the lowest precision (5.8%). The Set 1 achieves a precision of 100% for a recall larger than 20%, because any correct annotation to under-annotated proteins is of interest to curators, i.e., the problem of generic annotations described above is not applicable to these proteins.

The results show that the performance obtained by a given filter rate is preserved when applied to different sets of annotations. Therefore, curators can expect to obtain similar performances in different sets of annotations by using similar filter rates. Using different sets of curated and automated annotations may imply different *CT* for obtaining the same filter rate. For example, the automated annotations in *Set-3* have more similar curated annotations than in the other sets, thus it is also expected to have larger confidence scores. However, curators can easily adjust *CT* to obtain a required filter rate.

CAC does not discard new knowledge, but it does not discard the misan-

notations to under-annotated proteins either. To measure the real impact of using CAC on the curation process it should take into account the 455 novel annotations. CAC never discards these annotations, leaving the decision to the curator by assigning an infinite confidence score to them. Table 7.4 shows that including these novel annotations has a small effect on the performance of CAC. For example, by using a filter rate of 41.7% ($CT = 0.001$) the curator only has to verify 58.3% (100%-41.7%) of the original annotations only losing 3.1% (100%-96.9%) of the correct annotations. However, the precision for large filter rates is constrained by the precision of the novel annotations. Since CAC does not discard any of the 455 novel annotations, the precision converges to 7% (32 out of 455 annotations are correct) as CT increases. Nevertheless, CAC can overcome this limitation and contribute toward adding new knowledge. Nowadays, there are automated systems that predict generic annotations with high precision. If these generic annotations were considered, CAC would use them to score specific annotations, which is what curators really want. CAC can also be used to crosscheck annotations predicted by different automated systems. For example, CAC can score annotations predicted by a text-mining system based on annotations predicted by sequence similarity.

7.4 Conclusions

A significant number of genes and proteins have been functionally characterised by automatic tools, which have also produced a significant number of misannotations. This Chapter proposed a novel approach that uses curated annotations as domain knowledge for validating these automated annotations. To demonstrate its feasibility and efficiency, I developed and evalu-

ated CAC, a novel method to score automated annotations based on similar curated annotations. The results show that CAC can effectively be used to speed up the curation process by discarding a large amount of misannotations without losing a significant amount of correct annotations.

The precision/recall trade-off is tunable by a method's confidence threshold, which can be adjusted to obtain different filter rates according to the curator's requirements. The results obtained by similar filter rates were consistent for different subsets of the annotations, so the performance of CAC is predictable as we change a single tuning parameter.

CAC is an add-on data-mining tool that can be used by any automated annotation system to improve the accuracy and to require less effort to curators. Since CAC uses extensive domain knowledge automatically collected from public databases, it requires minimal human intervention. In addition, CAC can score relationships between other objects than genes and biological properties. All it requires is a similarity measure for each kind of object used and a set of curated relationships.

8

Conclusions

Different research communities produce different types of information whose relevance and structure change over the time. Biological databases store a large part of this information, but managing this wide, dynamic, and vast amount of data is a complex issue. Thus, researchers continue to mainly publish their findings in BioLiterature, which imposes fewer constraints to express their ideas than structured databases. In addition, researchers receive more credit for publishing their findings in BioLiterature than depositing the facts into databases. However, identifying information in BioLiterature is harder than in databases. Therefore, researchers are typically able to keep up with only a small part of BioLiterature related to their work. Even if authors submitted all the facts to databases, this would not solve the problem. Researchers do not only look for the facts but also the evidence substantiating them, since most facts are constrained to specific biological settings.

Text-mining systems have been used to minimize the effort spent on automatically identifying the facts and the evidence texts in BioLiterature. However, existing text-mining tools do not always provide what the curators want. On the contrary, they spend a large amount of their time finding the right documents. An annotation tool can only perform well when it is using the correct documents and entities. Errors in the retrieval of documents or

in the recognition of entities are propagated to the annotation process. This explains why many curators are sometimes reluctant on using text-mining tools in their work.

The performance of most state-of-the-art text-mining tools is still too much dependent on domain knowledge provided by experts, which is time-consuming and cannot be easily extended to other domains with different user requirements. This dissertation presented a set of text-mining methods that are effective and require minimal human intervention by integrating automatically acquired domain knowledge. In addition, a system for automatic annotation of biological databases integrating these methods was successfully applied to several databases. These contributions provide substantial evidence for validating the proposed hypothesis.

Hypothesis: In the automatic annotation of biological databases, the use of domain knowledge automatically integrated from biological data resources represents a feasible alternative to the use of domain knowledge explicitly created by experts.

8.1 Research Contributions

To obtain a sound evidence for substantiating the hypothesis, I developed ProFAL (bioProducts Functional Annotation through Literature), a system for automatic annotation of biological databases that can integrate novel methods based on the proposed approach. ProFAL was incorporated in the curation process of the CAZy, APEG, and UniProt databases. The teams of each database evaluated the tools by measuring the usefulness of the information provided by ProFAL. They found the tools useful in revealing new biologic annotations and providing a user-friendly interface for the curation

process. ProFAL obtained the lowest performance in CAZy since it was evaluated in a premature phase. However, this evaluation identified some problems that after tackled increased the accuracy, reaching 93% of precision in UniProt. ProFAL met the expectations of the curation process by reducing the workload of curators. The positive results were only possible because ProFAL integrates a set of novel methods developed according to the proposed hypothesis. These methods do not rely on manually inserted domain knowledge but on information automatically collected from biological databases.

8.1.1 WeBTC

I developed WeBTC, a novel approach for text classification that involves automatic integration of extracted information from biological Web resources with common statistical text classification methods. WeBTC was able to significantly increase the precision (reaching 100%) of a standard classification method.

WeBTC was evaluated in the BioText Task of KDD2002 Cup versus state-of-the-art systems. The evaluation indicated that WeBTC provided an effective alternative to enhance the performance of standard classification methods.

8.1.2 FiGO

I developed FiGO, a novel unsupervised method for recognizing biological terms in unstructured text involving the evidence content of their names. FiGO computes the evidence content based on the nomenclature of a BioOntology that structures the terms.

FiGO was designed for recognizing terms and not for extracting annotations. However, the method obtained a good performance in BioCreAtIvE when compared with other submissions. The performance of FiGO demonstrates that it provides an effective approach to recognise terms in BioLiterature, and to improve the performance of automatic annotation systems.

8.1.3 CAC

I developed CAC, a novel method to score biological annotations using the correlation between structure and function. CAC was applied to a set of annotations automatically extracted from BioLiterature. The results show that CAC can effectively be used to discard incorrect annotations generated by automatic systems.

The confidence threshold used by CAC can be manipulated to obtain high precision, high F-measure, and an increase in precision maintaining an acceptable recall. The results obtained by different thresholds were consistent for different subsets of the annotations, so the performance of CAC is predictable as we change a single tuning parameter.

8.2 Limitations

Despite its success, the approach proposed in this thesis has also its limitations. It is only effective when there is a substantial amount of accurate additional information available. In the experiments presented in this dissertation, the percentage of genes without available information was not significantly high. In the future, more genes will be characterised, especially for model organisms, whose characterisation has a great fundamental economical and social impact. However, the percentage of uncurated genes will tend

to grow, since the characterisation efforts are powerless to overcome the huge amount of data being generated by high-throughput analysis tools.

The retrieval of relevant documents from BioLiterature can be extended by using relevant keywords, such as names and alias of genes and proteins, accession numbers, notes and EC numbers. The inclusion of these keywords in queries to PubMed is expected to improve the retrieval of documents related to each gene and therefore enrich the pool of significant/relevant references.

Since PubMed only provides the abstracts of the documents, ProFAL only extracted annotations reported on the abstracts. Therefore, ProFAL is missing a large amount of information that is only reported on the full-text documents. Curators normally need additional information that is not present in the abstracts, such as the type of experiments applied and the species from which proteins originate. Nowadays, many scientific publishers are starting to provide open access to their documents. Thus, full-text documents will certainly be more accessible in the future, but their exploration will originate new problems, such as the selection of the document's sections that report information of interest to the users (Shah et al., 2004).

FiGO generated mispredictions in the instances where all the words of a GO term appeared in disparate locations of a sentence or in an unfortunate order. Improvements can result from the incorporation of better syntactical analysis into the identification of GO terms. For example, a reduction of the window size of FiGO, for example using noun phrases instead of sentences, can further increase precision. The handling of evidence text with numbers and abbreviations should also be improved to avoid confusion with the numbers and abbreviations used in the GO terms.

When FiGO identifies a term, it assumes that it represents an annotation.

This is clearly a strong assumption because most of the times the term is not related to the gene. Thus, there is a need for methods that can analyse the context of where the term occurs to decide if there is an annotation or not. These methods can be based on NLP techniques already applied in other areas that analyse the syntax and semantics of text. Moreover, FiGO only extracts GO terms. ProFAL will profit from extending FiGO to other BioOntologies (e.g. KEGG, MeSH, EC numbers) that provide also important sources of biochemical information.

8.3 Future Work

The vast prospects of strong fundamental economical and social impact stemming from Bioinformatics lead to the creation of large research consortia sponsored by both public and private efforts. These consortia run expensive research projects that maintain and generate many of the available biological data. Nevertheless, small institutions with limited resources are still important to complement these large projects. They usually exploit the available data to develop innovative approaches that could transform themselves into important trends. For example, the management of well-founded and broad BioOntologies is clearly an issue to be addressed by large research institutes, but smaller institutions can make important contributions on the development of useful tools to explore that information.

The work presented in this document constitutes a small and relatively early contribution to the advance of Text Mining of BioLiterature. The path I have chosen seems to be promising and the results encourage me to carry on. I will continue to improve the performance of the proposed methods and overcome the limitations mentioned in the previous Section.

The improvements will also expand the functionality of ProFAL contributing to a wider acceptance among curators of biological databases.

This research was only possible thanks to the research community efforts in developing accurate and valuable data resources and by making them available. These data resources are continually being updated with more information, enhancing the tools proposed by this thesis. The data resource most used by this thesis is GO, whose quality enabled the successful applications of the proposed approach. In my opinion, BioOntologies will have an important role, given the lack of a standard nomenclature in Molecular Biology. For example, they may support the integration of a wide variety of data sources, such as clinical and post-genomic data, providing new insights into how living systems operate.

This thesis also attracted my attention to other topics of Bioinformatics, where many problems remain unresolved. Despite rapid explosion of knowledge in the life sciences, its fully digitalisation, storage and curation are nowhere close to being completed. It is hard to ignore that Bioinformatics enabled increased scientific progress in Molecular Biology. Nevertheless, Bioinformatics is still a venture into an uncertain world that holds a great potential to benefit human health and society. For example, recent efforts to combine and coordinate diverse elements of Molecular Biology for understanding living systems as a whole are promising (Ideker et al., 2001). Many relevant biological discoveries in the future will certainly result from an efficient exploitation of the existing and newly generated data. This will require innovative and efficient data-management approaches, which I intend to develop based on the knowledge and skills that I acquired through this thesis.

A

ProFAL Class Diagram

ProFAL can be easily integrated with different biological databases. When these databases are implemented on relational database management systems, the integration involves little more than developing SQL queries to get data from the biological database and convert that data according to ProFAL's class diagram.

The ProFAL's class diagram models the data used by the retrieval and extraction use-cases, which were presented in Chapter 4. The other use-cases only use the data generated by these two use-cases. This Appendix describes the class diagram in two sections, one for each use-case.

A.1 Retrieval

Figure A.1 presents the UML class diagram of data involved in the Retrieval use-case (Booch et al., 1998). The *Author*, *DocAuthor*, *Document* and *Gene* classes represent information about the genes and documents automatically collected.

The *Citation* class represents the relationship between documents and genes. The *type* attribute indicates that the relationship was predicted or, alternatively, that it was manually assigned. The *value* attribute represents

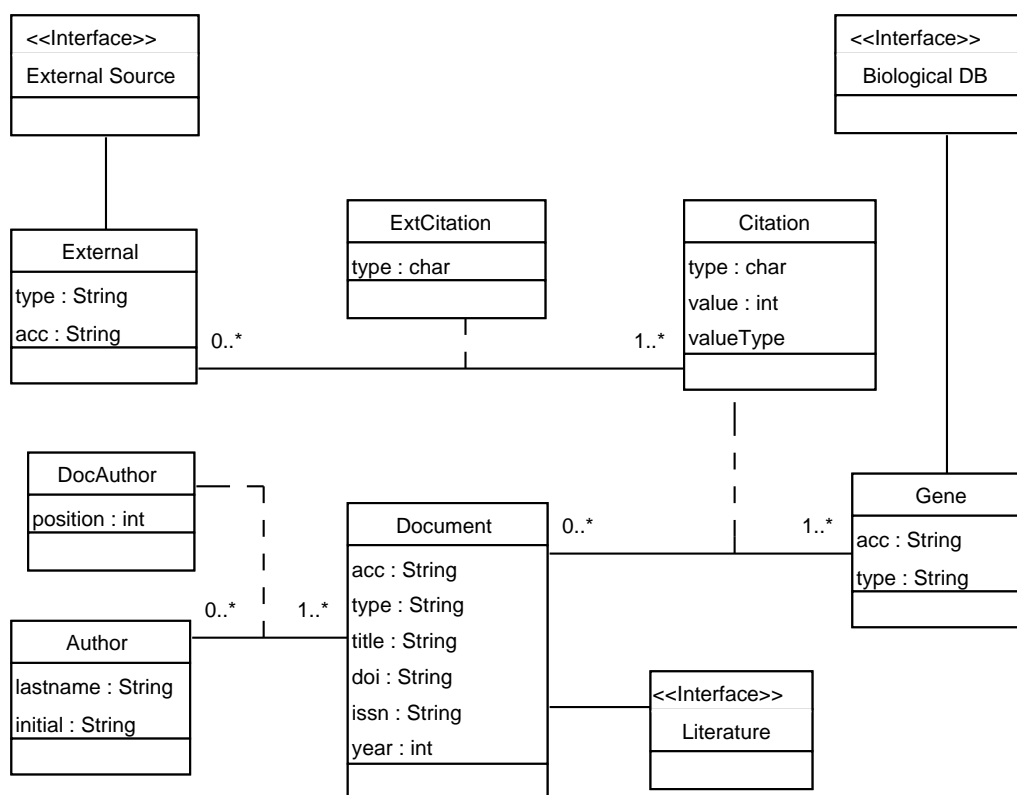


Figure A.1: Class diagram of the Retrieval use-case.

the confidence in each relation, and the *valueType* attribute specifies if this confidence was automatically or manually assigned.

The *External* and *ExtCitation* classes represent citations that were automatically collected from external sources. These identify the entries in the external sources linked to a given gene and associate them with the documents cited in their linked entries.

A.2 Extraction

Figure A.2 presents the UML class diagram of data involved in the Extraction use-case. The *Annotation* class represents the relation between genes

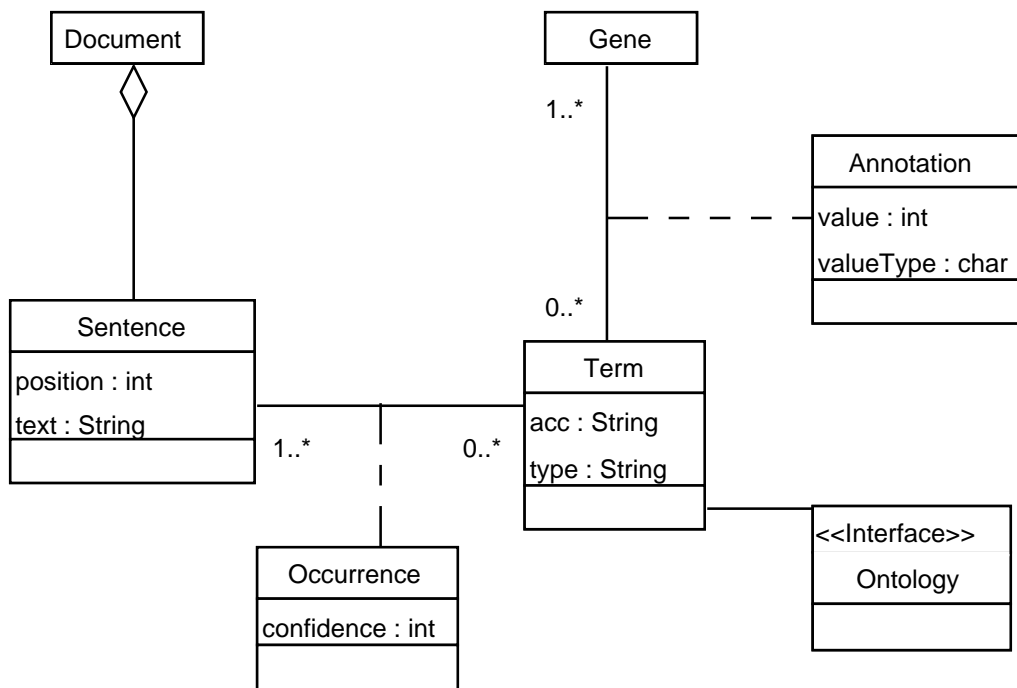


Figure A.2: Class diagram of the Extraction use-case.

and terms from the BioOntology. The *value* attribute represents the confidence score in each relation, and the *valueType* attribute specifies if this confidence was automatically or manually assigned. The Extraction process identifies occurrences of these terms in the sentences of each document. This information is regarded as evidence text and is represented by the *Occurrence* and *Sentence* classes.

In Figure A.2, the *value* attribute in the *Annotation* class represents the score assigned by the Validation use-case, which can be modified by the Verification use-case.

B

Semantic Similarity Measures

Most research on acquiring semantic properties of concepts has focused on semantic similarity, a research field that aims at calculating how similar two concepts are based on their semantic properties, normally acquired from corpora (Manning and Schütze, 1999). Research on Information Theory developed many semantic similarity measures. Some of them calculate maximum likelihood estimates for each concept using the corpora, and then calculate the similarity between probability distributions. Rada et al. (1989) emphasised the use of semantic similarity in ontologies by combining the structure and content of an ontology with statistical information from corpora. This Appendix presents state-of-the-art semantic similarity measures following this approach, including a measure developed on purpose for achieving a better performance in GO. The last section of this Appendix describes a tool that uses all the described measures in GO.

B.1 Basic Concepts

Semantic similarity measures can be used to calculate the similarity of two concepts organised in an ontology. The ontology structure defines the function $Parents(c)$ that, given a concept c , returns the set of more generic

concepts directly linked to c . In the case of an ontology organised as a tree, $Parents(c)$ always returns a single concept. On the other hand, in GO, $Parents(c)$ can return more than one term (concept), because each aspect of GO is composed by a set of terms organised as a DAG. Using the function $Parents(c)$ the set of paths between two concepts c_a and c_b can be defined as:

$$\begin{aligned}
 Paths(c_a, c_b) = & \tag{B.1} \\
 & \{\prec c_1, \dots, c_n \succ \mid (c_a = c_1) \wedge (c_b = c_n) \wedge \\
 & (\forall i : (1 \leq i < n) \wedge (c_i \in Parents(c_{i+1})))\}.
 \end{aligned}$$

A concept a is an ancestor of a concept c when there is at least one path from a to c :

$$Ancestors(c) = \{a \mid Paths(a, c) \neq \emptyset\}. \tag{B.2}$$

Note that since $\prec c \succ \in Paths(c, c)$, we have $c \in Ancestors(c)$.

The information content of a concept is inversely proportional to its frequency in a corpus. The frequency of a concept c , $Freq(c)$, can be defined as the number of times that c and all its descendants occur:

$$Freq(c) = \sum \{occur(c_i) \mid c \in Ancestors(c_i)\}. \tag{B.3}$$

Note that, for each ancestor a of a concept c , we have $Freq(a) \geq Freq(c)$, because the set of descendants of a contains all the descendants of c . An estimate for the occurrence of each GO term is the number of proteins annotated with it.

An estimate for the likelihood of observing an instance of a concept c is:

$$Prob(c) = \frac{Freq(c)}{maxFreq}, \quad (B.4)$$

where $maxFreq$ is the maximum frequency of all concepts. The $maxFreq$ of each aspect of GO is always equal to the frequency of the maximum (root) term in the DAG. For example, the GO term $t = molecular\ function$ has $Prob(t) = 1$, because all the GO terms in the *molecular function* aspect are descendant of t , and therefore $Freq(t) = maxFreq$.

The information content of a concept c can be defined as the negative logarithm of its probability:

$$IC(c) = -\log(Prob(c)). \quad (B.5)$$

Note that the information content is monotonic, since it is non-increasing as we descend in the hierarchy.

Semantic similarity measures assume that the similarity between two concepts is related to the extent to which they share information. The common ancestors of two concepts c_1 and c_2 are:

$$\begin{aligned} CommonAnc(c_1, c_2) = & \quad (B.6) \\ & Ancestors(c_1) \cap Ancestors(c_2). \end{aligned}$$

Given two concepts c_1 and c_2 , their shared information, $Share(c_1, c_2)$, can be defined as the information content of their most informative common ancestor:

$$Share(c_1, c_2) = \quad (B.7)$$

$$\max\{IC(a) \mid a \in CommonAnc(c_1, c_2)\}.$$

The most informative common ancestor is the one with the largest information content. Note that $Share(c, c) = IC(c)$, because $c \in Ancestors(c)$.

B.2 State-of-the-art Measures

Many semantic similarity measures applied to ontologies have been proposed. Resnik (1995) defined a semantic similarity measure based on the information content of the most informative common ancestor. The information content of a concept is inversely proportional to its frequency in the corpora. Concepts that are frequent in the corpora have low information content. For example, the stop words (such as *the*) that occur almost everywhere in the text normally provide little semantic information. Jiang and Conrath (1997) proposed a semantic distance measure based on the difference between the information content of the concepts and the information content of their most informative common ancestor. Lin (1998) proposed a semantic similarity measure based on the ratio between the information content of the most informative common ancestor and the information content of both concepts. The rest of this section details these three measures.

Given two concepts c_1 and c_2 , Resnik defined their semantic similarity as the information content of their most informative common ancestor:

$$Sim_{Resnik}(c_1, c_2) = Share(c_1, c_2). \quad (B.8)$$

Given two concepts c_1 and c_2 , Jiang&Conrath defined their semantic distance as the difference between their information content and the information

content of their most informative common ancestor:

$$\begin{aligned} dist_{JC}(c_1, c_2) = & \hspace{15em} (B.9) \\ & IC(c_1) + IC(c_2) - 2 \times Share(c_1, c_2). \end{aligned}$$

Note that Jiang&Conrath’s formula measures a distance, the inverse of similarity. A similarity measure based on Jiang&Conrath distance measure can be defined as:

$$Sim_{JC}(c_1, c_2) = \frac{1}{dist_{JC}(c_1, c_2) + 1}. \quad (B.10)$$

$dist_{JC} + 1$ is used to avoid infinity values, since $dist_{JC}(c, c) = 0$.

Given two concepts, c_1 and c_2 , Lin defined their similarity as the information content of their most informative common ancestor over their information content:

$$Sim_{Lin}(c_1, c_2) = \frac{2 \times Share(c_1, c_2)}{IC(c_1) + IC(c_2)}. \quad (B.11)$$

B.3 GraSM

GO is not organised as a tree-like hierarchy, but as directed acyclic graphs (DAG), one for each aspect. This enables a more complete and realistic annotation. The semantic similarity measures described above only use the most informative common ancestor of both concepts. Therefore, when applied to a DAG, these measures discard other common ancestors even if they are disjunctive ancestors. When all but the most informative common ancestor nodes are ignored, different possible interpretations of the biologic concepts are disregarded. To tackle this limitation, I developed GraSM (Graph-based Similarity Measure), a novel method for incorporating the semantic richness of a graph by selecting disjunctive common ancestors of two concepts.

GraSM selects and uses all the disjunctive common ancestors representing all interpretations.

GraSM assumes that two common ancestors are disjunctive if there are independent paths from both ancestors to the concept. Independent paths mean those that use at least one concept of the ontology not used by the other paths. Two disjunctive ancestors of a concept represent two distinct interpretations of a concept. For example, in Figure 2.8 the terms *carbohydrate binding* and *bacterial binding* are two disjunctive ancestors of *peptidoglycan binding*. Thus, the similarity between *peptidoglycan binding* and *polysaccharide binding* is smaller than if *peptidoglycan binding* only had the ancestor *carbohydrate binding*. The similarity is smaller because *peptidoglycan binding* can also be interpreted as *bacterial binding*, which is not an ancestor of *polysaccharide binding*.

Calculating the similarity between two concepts using just the most informative common ancestor only accounts for one of the interpretations. However, similarity measures should also account for other interpretations of both concepts. GraSM selects all the common disjunctive ancestors of two concepts in a DAG to calculate their similarity.

GraSM considers that a_1 and a_2 represent disjunctive ancestors of c if there is a path from a_1 to c not passing through a_2 and a path from a_2 to c not passing through a_1 :

$$\begin{aligned}
 DisjAnc(c) = & \tag{B.12} \\
 & \{(a_1, a_2) \mid \\
 & (\exists p : (p \in Paths(a_1, c)) \wedge (a_2 \notin p)) \wedge \\
 & (\exists p : (p \in Paths(a_2, c)) \wedge (a_1 \notin p))\}
 \end{aligned}$$

Note that if $a_1 \notin \text{Ancestors}(a_2)$ and $a_2 \notin \text{Ancestors}(a_1)$ then a_1 and a_2 are disjunctive ancestors of c . For example, in Figure 2.8 $(t_2, t_3) \in \text{DisjAnc}(t_5)$. Otherwise, if $a_1 \in \text{Ancestor}(a_2)$ it is still possible that a_1 and a_2 represent disjunctive ancestors of c . For example, in Figure 2.8 $(t_1, t_2) \in \text{DisjAnc}(t_5)$ because the path $\prec t_1, t_3, t_5 \succ$ does not pass through t_2 , and the path $\prec t_2, t_5 \succ$ does not pass through t_1 .

Given two concepts c_1 and c_2 , their common disjunctive ancestors are the most informative common ancestor of disjunctive ancestors of c_1 and c_2 , i.e., a_1 is a common disjunctive ancestor of c_1 and c_2 if for each ancestor a_2 more informative than a_1 , a_1 and a_2 are a disjunctive ancestor of c_1 or c_2 :

$$\begin{aligned}
\text{CommonDisjAnc}(c_1, c_2) = & \tag{B.13} \\
& \{a_1 \mid a_1 \in \text{CommonAnc}(c_1, c_2) \wedge \\
& \forall a_2 : [(a_2 \in \text{CommonAnc}(c_1, c_2)) \wedge \\
& (IC(a_1) \leq IC(a_2)) \wedge (a_1 \neq a_2)] \Rightarrow \\
& [(a_1, a_2) \in (\text{DisjAnc}(c_1) \cup \text{DisjAnc}(c_2))]\}
\end{aligned}$$

Note that $\text{CommonDisjAnc}(c, c) = \{c\}$ because all the ancestors of c are not disjunctive ancestors of c , i.e., $(c, a) \notin \text{DisjAnc}(c)$ for all $a \in \text{Ancestors}(c)$. In Figure 2.8, $\text{CommonDisjAnc}(t_4, t_5) = \{t_1, t_2\}$ because t_2 is the most informative common ancestor, and t_1 and t_2 are disjunctive ancestors of t_5 .

GraSM defines the shared information between c_1 and c_2 as the average of the information content of their common disjunctive ancestors:

$$\begin{aligned}
\text{Share}_{\text{GraSM}}(c_1, c_2) = & \tag{B.14} \\
& \overline{\{IC(a) \mid a \in \text{CommonDisjAnc}(c_1, c_2)\}}.
\end{aligned}$$

GO term	Annotations	<i>Freq</i>	<i>Prob</i>	<i>IC</i>
t_0	8	16	1	0
t_1	3	8	0.5	1
t_2	2	4	0.25	2
t_3	1	2	0.125	3
t_4	1	1	0.0625	4
t_5	1	1	0.0625	4

Table B.1: The information content of each term considering a certain number of annotations

Share can be replaced by $Share_{GraSM}$ yielding three new variants of the semantic similarity measures presented in the previous Section: $Sim_{ResnikGraSM}$, $Sim_{JCGraSM}$ and $Sim_{LinGraSM}$.

Example

Considering only the sub graph of GO represented in Figure 2.8 and the values in Table B.1, the set of common ancestors of t_4 and t_5 in a descendant order of *IC* is $\{t_2, t_1, t_0\}$, and the set of common disjunctive ancestors is $CommonDisjAnc(t_4, t_5) = \{t_2, t_1\}$, as described above. Thus, $Share_{GraSM}(t_4, t_5) = \overline{\{IC(t_2), IC(t_1)\}} = 1.5$. The similarity between t_4 and t_5 with and without using GraSM is:

$$\begin{aligned} & Sim_{Resnik}(t_4, t_5) \\ &= Share(t_4, t_5) = 2 \end{aligned}$$

$$\begin{aligned} & Sim_{ResnikGraSM}(t_4, t_5) \\ &= Share_{GraSM}(t_4, t_5) = 1.5 \end{aligned}$$

$$\begin{aligned}
& Sim_{JC}(t_4, t_5) \\
&= \frac{1}{IC(t_4) + IC(t_5) - 2 \times Share(t_4, t_5)} \\
&= \frac{1}{4 + 4 - 2 \times 2} = 0.25
\end{aligned}$$

$$\begin{aligned}
& Sim_{JCGraSM}(t_4, t_5) \\
&= \frac{1}{IC(t_4) + IC(t_5) - 2 \times Share_{GraSM}(t_4, t_5)} \\
&= \frac{1}{4 + 4 - 2 \times 1.5} = 0.2
\end{aligned}$$

$$\begin{aligned}
& Sim_{Lin}(t_4, t_5) \\
&= \frac{2 \times Share(t_4, t_5)}{IC(t_4) + IC(t_5)} \\
&= \frac{2 \times 2}{4 + 4} = 0.5
\end{aligned}$$

$$\begin{aligned}
& Sim_{LinGraSM}(t_4, t_5) = \\
&= \frac{2 \times Share_{GraSM}(t_4, t_5)}{IC(t_4) + IC(t_5)} \\
&= \frac{2 \times 1.5}{4 + 4} = 0.375
\end{aligned}$$

If the shared information of one ancestor is high, and then we find another disjunctive common ancestor with lower information content, it seems that finding the additional relationship should increase the similarity rather than lessen it. However, this is an incorrect intuition. Finding a disjunctive

Algorithm 2: $Share_{GraSM}(c_1, c_2)$

```
1:  $Anc = CommonAnc(c_1, c_2)$ 
2:  $CommonDisjAnc = \{\}$ 
3: for all  $a \in sortDescByIC(Anc)$  do
4:    $isDisj = true$ 
5:   for all  $cda \in CommonDisjAnc$  do
6:      $isDisj = isDisj \wedge$ 
        $(DisjAnc(c_1, (cda, a)) \vee DisjAnc(c_2, (cda, a)))$ 
7:   end for
8:   if  $isDisj$  then
9:      $addTo(CommonDisjAnc, a)$ 
10:  end if
11: end for
12:  $shared = 0$ 
13: for all  $cda \in CommonDisjAnc$  do
14:    $shared += IC(cda)$ 
15: end for
16: return  $shared/sizeof(CommonDisjAnc)$ 
```

Algorithm 3: $DisjAnc(c, (a_1, a_2))$

```
Input:  $IC(a_1) \leq IC(a_2)$ 
1:  $nPaths = \#Paths(a_1, a_2)$ 
2:  $nPaths_1 = \#Paths(a_1, c)$ 
3:  $nPaths_2 = \#Paths(a_2, c)$ 
4: return  $nPaths_1 \geq nPaths \times nPaths_2$ 
```

common ancestor means that at least one of the terms has a distinct and more distant interpretation to the other term, which makes the terms less similar. Thus, by taking in account the less informative common ancestor, GraSM provides lower similarities than the original measures.

B.3.1 Computational Aspect

Algorithm 2 describes a possible implementation of $Share_{GraSM}$. It starts by selecting the common ancestors of both concepts (line 1) and by initial-

using the list of common disjunctive ancestors as an empty list (line 2). The algorithm selects each common ancestor in descending order of information content (line 3). For each selected ancestor, the algorithm checks if the ancestor is disjunctive to all the common disjunctive ancestors already selected (lines 4 to 7). If the ancestor is disjunctive, it adds it to the list of common disjunctive ancestors (line 9). At the end, the algorithm calculates the average of the information content of all the ancestors in the common disjunctive ancestors list (lines 12 to 16).

Algorithm 3 describes an efficient technique to check if a pair of ancestors (a_1, a_2) are disjunctive ancestors of a given concept c . Since $IC(a_1) \leq IC(a_2)$, then there are no paths from a_2 to a_1 . Thus, from Definition B.13, we can conclude that a_1 and a_2 are disjunctive if and only if there is at least one more path from a_1 to c than from a_1 to c passing through a_2 . Thus, the algorithm only checks if the number of paths from a_1 to c is larger than the sum of the number of paths from a_1 to a_2 and from a_2 to c .

These implementations show that using GraSM is not prohibitively expensive. In addition to finding the common ancestors, as *Share*, *Share_{GraSM}* only has to check the list of common ancestors, which is normally smaller than the depth of the graph. Counting the number of paths is also not time-consuming. For example, in the GO distribution there is a table that stores each path between two GO terms.

The evaluations of FiGO, CAC and GOAnnotator, presented in this document, did not use GraSM to calculate the semantic similarity between GO terms because GraSM was not properly defined and evaluated at the time they were performed.

B.4 FuSSiMeG

All the semantic similarity measures described in this Appendix were implemented by FuSSiMeG (Functional Semantic Similarity Measure between Gene-Products), which measures the functional similarity between proteins based on the semantic similarity of the GO terms annotated to them (Couto et al., 2003c). FuSSiMeG is available on the Web¹, affording the similarity calculation on the fly.

Figure B.1 presents the results displayed by the tool for two UniProt proteins using the $Sim_{LinGraSM}$ measure. The tool displays the semantic similarity between the GO terms annotated to the proteins. Besides the similarity of the annotated GO terms, their specificity cannot be disregarded when comparing two proteins. For example, both proteins can be annotated with a generic GO term (100% similarity), but this does not mean that they are similar since many other proteins are also annotated to this term. Therefore, FuSSiMeG also displays the weighted similarity between the GO terms, which divides the semantic similarity by the information content of both terms.

FuSSiMeG has been used for distinct tasks:

- The USA national institutes of health and aging used FuSSiMeG to validate the extraction of networks of probes/genes that are coregulated in large-scale expression studies for helping building and testing hypotheses about neurodegeneration.
- The Samuel Lunenfeld Research Institute used FuSSiMeG to work on gene function analysis for whole mouse genome.
- The MPI for Molecular Genetics in Berlin is using FuSSiMeG to work

¹<http://xldb.fc.ul.pt/rebil/tools/ssm/>

ReBIL Project

FuSSiMeG: Functional Semantic Similarity Measure between Gene-Products



Similarity between the GO terms annotated with:
[P16403](#) (*H12_HUMAN*) and [Q8NHM5](#) (*FXLA_HUMAN*)

P16403's terms	Q8NHM5's terms	Terms' Similarity	Weighted Similarity
<i>function</i>			
GO:0003677 (DNA binding)	GO:0003677 (DNA binding)	100.0%	17.9%
GO:0003677 (DNA binding)	GO:0008270 (zinc ion binding)	11.2%	3.4%
<i>process</i>			
GO:0006334 (nucleosome assembly)	GO:0006355 (regulation of transcription, DNA-dependent)	6.0%	2.6%
GO:0007001 (chromosome organization and biogenesis (sensu Eukaryota))	GO:0006355 (regulation of transcription, DNA-dependent)	6.1%	2.1%
GO:0007001 (chromosome organization and biogenesis (sensu Eukaryota))	GO:0000004 (biological_process unknown)	5.6%	1.9%
GO:0007001 (chromosome organization and biogenesis (sensu Eukaryota))	GO:0006512 (ubiquitin cycle)	4.3%	1.6%
<i>component</i>			
GO:0005634 (nucleus)	GO:0005634 (nucleus)	100.0%	18.4%
GO:0000786 (nucleosome)	GO:0005634 (nucleus)	7.0%	2.9%
GO:0005694 (chromosome)	GO:0005634 (nucleus)	8.7%	2.8%

Results using JiangConrathGraSM measure.

Note: *Weighted Similarity* takes the information content of each term in consideration. The information content of a term is inversely proportional to the number of times that the term is annotated. For instance, the GO term 'protein' is a very frequent annotated term, so it is not very relevant for comparing the functionality of two proteins.

Figure B.1: Functional semantic similarity of P42973 and O85465 UniProt proteins.

on a multi genome annotation tool, which provides potential new GO annotations via sequence similarities.

- The Facultad de Farmacia of the University of San Pablo - CEU in Madrid is using FuSSiMeG to compare different similarity measures.
- The Universiti Teknologi Malaysia is using FuSSiMeG to work on applying GO for protein function prediction.
- The Iowa State University is using FuSSiMeG to carry out a clustering process using GO information.

Bibliography

- Adrion, W. (1993). Research methodology in software engineering: Summary of the Dagstuhl workshop on future directions in software engineering. *ACM SIGSOFT Software Engineering Notes*, 18(1):36–37.
- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. (1989). *The Molecular Biology of the Cell*. New York Garland Publishing.
- Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17).
- Andrade, M. and Bork, P. (2000). Automated extraction of information in Molecular Biology. *FEBS Letters*, 476:12–17.
- Andrade, M. and Valencia, A. (1998). Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics*, 14(7):600–607.
- Apweiler, R., Bairoch, A., Wu, C., Barker, W., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M., Natale, D., O’Donovan, C., Redaschi, N., and Yeh, L. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 32(Database issue):D115–D119.

- Attwood, T. and Parry-Smith, D. (1999). *Introduction to Bioinformatics*. Longman Higher Education.
- Bada, M., Stevens, R., Goble, C., Gil, Y., Ashburner, M., Blake, J., Cherry, J., Harris, M., and Lewis, S. (2004). A short study on the success of the gene ontology. *Journal of Web Semantics*, 1(1):235–240.
- Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28:45–48.
- Baker, C. (1989). *English syntax*. MIT Press.
- Baldi, P. and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach, Second Edition*. MIT Press.
- Basu, C., Hirsh, H., and Cohen, W. (1998). Recommendation as classification: Using social and content-based information in recommendation. In *Proc. of the 15th national/10th conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*.
- Bateman, A., Coin, L., Durbin, R., Finn, R., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E., Studholme, D., Yeats, C., and Eddy, S. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32(Database issue):D138–D141.
- Becker, J., Boavida, L., Carneiro, J., Haury, M., and Feijó, J. (2003). Transcriptional profiling of Arabidopsis tissues reveals the unique characteristics of the pollen transcriptome. *Plant Physiology*, 113(2):713–725.
- Benson, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler,

- D. (2004). GenBank: update. *Nucleic Acids Research*, 32(Database issue):D23–D26.
- Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28:235–242.
- Blaschke, C., Hirschman, L., and Valencia, A. (2002). Information extraction in Molecular Biology. *Briefings in Bioinformatics*, 3(2):154–165.
- Blaschke, C., Leon, E., Krallinger, M., and Valencia, A. (2005). Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, 6(Suppl 1):S16.
- Booch, G., Rumbaugh, J., and Jacobson, I. (1998). *The Unified Modeling Language User Guide*. Addison-Wesley.
- Boork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *Journal of Molecular Biology*, 283(4):707–725.
- Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.
- Budanitsky, A. and Hirst, G. (2001). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Proc. of the Workshop on WordNet and Other Lexical Resources co-located with the 2nd North American Chapter of the Association for Computational Linguistics*.
- 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

- Annotations (GOA) database: sharing knowledge in UniProt with Gene Ontology. *Nucleic Acids Research*, 32:262–166.
- Chiang, J. and Yu, H. (2003). MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*, 19(11):1417–1422.
- Chicurel, M. (2002). Bioinformatics: bringing it all together. *Nature*, 419(6908):751–757.
- Cohen, J. (2004). Bioinformatics - an introduction for computer scientists. *ACM Computing Surveys*, 36(2):122–158.
- Cohen, W. (2000). Automatically extracting features for concept learning from the web. In *Proc. of the 17th International Conference on Machine Learning*.
- Corney, D., Buxton, B., Langdon, W., and Jones, D. (2004). BioRAT: Extracting biological information from full-length papers. *Bioinformatics*, 20(17):3206–3213.
- Coutinho, P. and Henrissat, B. (1999). *Recent Advances in Carbohydrate Bioengineering*, chapter Carbohydrate-active enzymes: an integrated database approach. The Royal Society of Chemistry.
- Couto, F., Martins, B., and Silva, M. (2004a). Classifying biological articles using web resources. In *Proc. of the 2004 ACM Symposium on Applied Computing*.
- Couto, F., Martins, B., Silva, M., and Coutinho, P. (2003a). Classifying biomedical articles using web resources: application to KDD Cup 02. DI/FCUL TR 03–24, Department of Informatics, University of Lisbon.

- Couto, F. and Silva, M. (2005). *Advanced Data Mining Technologies in Bioinformatics*, chapter Mining the BioLiterature: towards automatic annotation of genes and proteins. Idea Group Inc. (in press).
- Couto, F., Silva, M., and Coutinho, P. (2003b). Curating extracted information through the correlation between structure and function. In *Proc. of the 3rd meeting of the special interest group on Text Data Mining co-located with 11th Internacional Conference on Intelligent Systems for Molecular Biology*.
- Couto, F., Silva, M., and Coutinho, P. (2003c). Implementation of a functional semantic similarity measure between gene-products. DI/FCUL TR 03–29, Department of Informatics, University of Lisbon.
- Couto, F., Silva, M., and Coutinho, P. (2003d). Improving information extraction through biological correlation. In *Proc. of the Data Mining and Text Mining for Bioinformatics Workshop co-located with 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*.
- Couto, F., Silva, M., and Coutinho, P. (2003e). ProFAL: Protein functional annotation through literature. In *Proc. of the 8th Conference on Software Engineering and Databases*.
- Couto, F., Silva, M., and Coutinho, P. (2004b). FiGO: Finding GO terms in unstructured text. In *Proc. of the BioCreAtIvE Challenge Evaluation Workshop*.
- Couto, F., Silva, M., and Coutinho, P. (2005a). Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6(S1):S21.

- Couto, F., Silva, M., and Coutinho, P. (2005b). Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. In *Proc. of the ACM Conference in Information and Knowledge Management as a short paper*.
- Couto, F., Silva, M., and Coutinho, P. (2005c). Validation of automated protein annotation. DI/FCUL TR 05–24, Department of Informatics, University of Lisbon.
- Couto, F., Silva, M., and Fernandes, P., editors (2005d). *BKDB2005 - Bioinformatics: Knowledge Discovery in Biology*. Faculdade de Ciências da Universidade de Lisboa.
- Couto, F., Silva, M., Lee, V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H., and Rebholz-Schuhmann, D. (2005e). GOAnnotator: linking electronic protein GO annotation to evidence text. DI/FCUL TR 05–25, Department of Informatics, University of Lisbon.
- Crick, F. (1958). On protein synthesis. In *Proc. of the 12th Symposium of the Society for Experimental Biology*.
- Demsey, A., Nahin, A., and Braunsberg, S. V. (2003). Oldmedline citations join pubmed. *NLM Technical Bulletin*, 334(e2).
- Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends Genetics*, 17(8):429–431.
- Dickman, S. (2003). Tough mining. *PLoS Biology*, 1(2):144–147.
- Doerks, T., Bairoch, A., and Bork, P. (1998). Protein annotation: detective work for function prediction. *Trends Genetics*, 14(6):248–250.

- Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of ACM*, 13(2):94–102.
- Fukuda, K., Tsunoda, T., Tamura, A., and Takagi, T. (1998). Toward information extraction: Identifying protein names from biological papers. In *Proc. of the 3rd Pacific Symposium on Biocomputing*.
- Gerstein, M. (2000). Integrative database analysis in structural genomics. *Nature Structural Biology*, 7(Supplement Structural genomics):960–963.
- Ghanem, M., Guo, Y., Lodhi, H., and Zhang, Y. (2002). Automatic scientific text classification using local patterns: KDD CUP 2002 (task 1). *SIGKDD Explorations*, 4(2):95–96.
- GO-Consortium (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32(Database issue):D258–D261.
- Gruber, T. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220.
- Hand, D., Mannila, H., and Smyth, P. (2000). *Principles of Data Mining*. MIT Press.
- Hearst, M. (1999). Untangling text data mining. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*.
- Hersh, W., Bhuptiraju, R., Ross, L., Johnson, P., Cohen, A., and Kraemer, D. (2004). TREC 2004 genomics track overview. In *Proc. of the 13th Text REtrieval Conference*.
- Hirschman, L., Park, J., Tsujii, J., Wong, L., and Wu, C. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18(12):1553–1561.

- Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1.
- Ideker, T., Galitskiand, T., and Hood, L. (2001). A new approach to decoding life: systems biology. *Annual Reviews of Genomics and Human Genetics*, 2:343–372.
- Jain, P. (2004). Gene function prediction by mining biomedical literature. Master’s thesis, Department of Informatics, University of Lisbon. DI/FCUL TR-04-12.
- Jain, P., Couto, F., Silva, M., and Becker, J. (2005). Literature based functional annotation of genes. In *Proc. of the BKDB2005 - Bioinformatics: Knowledge Discovery in Biology Workshop*.
- Jakob, N. (1994). *Usability engineering*. Academic Press Inc.
- Jiang, J. and Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the 10th International Conference on Research on Computational Linguistics*.
- Kanehisa, M. and Bork, P. (2003). Bioinformatics in the post-sequence era. *Nature Genetics*, 33(3):305–310.
- Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Stoehr, S. S. P., Tuli, M., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. (2005). The embl nucleotide sequence database. *Nucleic Acids Research*, 33(Database issue):D29–D33.

- Kaufmann, M., editor (1995). *Sixth Message Understanding Conference (MUC-6)*.
- Keerthi, S., Ong, C., Siah, K., Lim, D., Chu, W., Shi, M., Edwin, D., Menon, R., Shen, L., Lim, J., and Loh, H. (2002). A machine learning approach for the curation of biomedical literature - KDD Cup 2002 (task 1). *SIGKDD Explorations*, 4(2):93–94.
- Kenney, J. and Keeping, E. (1962). *Mathematics of Statistics*, chapter Quartiles. Princeton, NJ: Van Nostrand.
- Kim, J. and Park, J. (2004). BioIE: retargetable information extraction and ontological annotation of biological interactions from literature. *Journal of Bioinformatics and Computational Biology*, 2(3):551–568.
- Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5560):1662–1664.
- Koike, A., Niwa, Y., and Takagi, T. (2005). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, 21(7):1227–1236.
- Larsson, T., Murray, C., Hill, T., Fredriksson, R., and Schioth, H. (2005). Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery. *FEBS Letters*, 579(3):690–698.
- Leake, D. (1996). *Case-Based Reasoning: Experiences, Lessons, and Future Directions*. AAAI Press/MIT Press.
- Lin, D. (1998). An information-theoretic definition of similarity. In *Proc. of the 15th International Conference on Machine Learning*.

- Lord, P., Stevens, R., Brass, A., and Goble, C. (2003a). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283.
- Lord, P., Stevens, R., Brass, A., and Goble, C. (2003b). Semantic similarity measures as tools for exploring the Gene Ontology. In *Proc. of the 8th Pacific Symposium on Biocomputing*.
- Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.
- McCallum, A. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. <http://www.cs.cmu.edu/mccallum/bow>.
- Mitchell, J., Aronson, A., Mork, J., Folk, L., Humphrey, S., and Ward, J. (2003). Gene indexing: characterization and analysis of NLM’s GeneRIFs. In *Proc. of the AMIA 2003 Annual Symposium*.
- Mitkov, R. (2002). *Anaphora Resolution*. Longman.
- Moult, J. (2005). A decade of casp: progress, bottlenecks and prognosis in protein structure prediction. *Current opinion in structural biology*, 15(3).
- Müller, H., Kenny, E., and Sternberg, P. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLOS Biology*, 2(11):E309.
- NC-IUBMB (1992). *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press.

- Nei, M. (1996). Phylogenetic analysis in molecular evolutionary genetics. *Annual Review of Genetics*, 30.
- Nelson, S., Schopen, M., Savage, A., Schulman, J., and Arluk, N. (2004). The MeSH translation maintenance system: Structure, interface design, and implementation. In *Proc. of the 11th World Congress on Medical Informatics*.
- Nowak, R. (1995). Entering the postgenome era. *Science*, 270(5235):368–371.
- Pagallo, G. and Hassler, D. (1990). Boolean feature discovery in empirical learning. *Machine Learning*, 5(1):71–99.
- Palakal, M., Stephens, M., Mukhopadhyay, S., and Rajee, R. (2003). Identification of biological relationships from text documents using efficient computational methods. *Journal of Bioinformatics and Computational Biology*, 1(2):307–342.
- Pérez, A., Perez-Iratxeta, C., Bork, P., Thode, G., and Andrade, M. (2004). Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics*, 20(13):2084–2091.
- Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems*, 19(1):17–30.
- Rebholz-Schuhmann, D., Kirsch, H., and Couto, F. (2005). Facts from text - is text mining ready to deliver? *PLoS Biology*, 3(2):e65.
- Regev, Y., Finkelstein-Landau, M., and Feldman, R. (2002). Rule-based extraction of experimental evidence in the biomedical domain - the KDD Cup (task 1). *SIGKDD Explorations*, 4(2):90–92.

- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th International Joint Conference on Artificial Intelligence*.
- Rigoutsos, I., Floratos, A., Parida, L., Ga, Y., and Platt, D. (2000). Gapped blast and psi-blast: a new generation of protein database search programs. *Metabolic Engineering*, 2(3).
- Rubin, G. (1996). Around the genomes: The Drosophila genome project. *Genome Research*, 6(2):71–79.
- Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.
- Shah, P., Perez-Iratxeta, C., Bork, P., and Andrade, M. (2004). Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4(20).
- Shatkay, H. and Feldman, R. (2003). Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821–855.
- Smith, L., Rindfleisch, T., and Wilbur, W. (2004). MedPost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.
- Spencer, A. (1991). *Morphological theory*. Oxford: Blackwell.
- Stevens, R., Wroe, C., Lord, P., and Goble, C. (2003). *Handbook on Ontologies*, chapter Ontologies in Bioinformatics. Springer.
- Wheeler, D., Church, D., Federhen, S., Lash, A., Madden, T., Pontius, J., Schuler, G., Schriml, L., Sequeira, E., Tatusova, T., and Wagner, L. (2003).

- Database resources of the national center for biotechnology. *Nucleic Acids Research*, 31(1):28–33.
- Wilks, Y. and Stevenson, M. (1997). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Journal of Natural Language Engineering*, 4(1):135–143.
- Wu, D. and Fung, P. (1994). Improving chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proc. of the 4th Conference on Applied Natural Language Processing*.
- Yeh, A., Hirschman, L., and Morgan, A. (2002). Background and overview for KDD Cup 2002 task 1: Information extraction from biomedical articles. *SIGKDD Explorations*, 4(2):87–89.
- Yeh, A., Hirschman, L., and Morgan, A. (2003). Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. *Bioinformatics*, 19(1):i331–i339.
- Zelkowitz, M. and Wallace, D. (1998). Experimental models for validating technology. *Computer*, 31(5):23–31.

Index

- APEG, 8–10, 13, 27, 37, 61–63, 71, 116
- BioCreAtIvE, 11, 48, 49, 91–96, 98, 104, 105, 108, 111, 118
- Bioinformatics, 1, 3, 9, 12, 15, 16, 18, 25, 26, 35, 36, 38, 120, 121
- BioLiterature, 1, 2, 4, 5, 7–9, 12, 16, 28, 33, 34, 38, 39, 45–50, 53, 54, 56, 57, 61, 68, 73, 96, 98, 115, 118–120
- BioOntology, 4, 8, 30, 31, 45, 56–58, 85–87, 91, 98, 103, 117, 125
- BioText Task of KDD2002 Cup, 11, 49, 76, 77, 80–83, 117
- CAC, 8–10, 12, 13, 98, 100–114, 118, 137
- CAZy, 8–11, 13, 27, 37, 59–61, 71, 116, 117
- DAG, 32
- EMBL-Bank, 16, 25, 26, 37
- FiGO, 8, 10, 11, 13, 65, 84–86, 89–95, 97, 98, 117–120, 137
- FlyBase, 27, 37, 77
- FuSSiMeG, 138, 140
- GenBank, 25, 26, 37, 59, 76, 78
- GeneRIF, 28
- GO, 30–33, 37, 46, 47, 60, 61, 64–70, 85, 86, 91, 93–97, 99, 100, 103–106, 110, 119–121, 127–129, 131, 134, 137, 138, 140
- GOA, 27, 37, 64, 67, 68, 85, 91, 99–101, 104–106, 110
- GOAnnotator, 9, 64–70, 137
- GraSM, 131–134, 136–138
- MEDLINE, 2, 3, 34, 37, 60, 67
- MeSH, 30, 31, 37, 78, 120
- NLP, 40–43, 45–47, 120
- PDB, 23, 26, 37, 59, 60

Pfam, 27, 28, 37, 103, 106

PMC, 34, 35, 37

ProFAL, 7–13, 44, 47, 48, 51, 53–
59, 61–64, 70–72, 77, 83,
84, 98, 116, 117, 119–121,
123

PubMed, 2, 34, 37, 57, 59, 69–71,
76, 78, 119

UniProt, 1, 8–10, 13, 20, 26–28, 36,
37, 59, 64, 66, 67, 71, 85,
92, 97, 99, 103, 106, 116,
117, 138, 139

WeBTC, 8, 10, 11, 13, 74–76, 78–
84, 117