Workshop on Data and Text Mining for Integrative Biology

In conjunction with ECML/PKDD 2006

17th European Conference on Machine Learning
10th European Conference on Principles and Practice of Knowledge Discovery in Databases

Berlin, Germany, September 18-22, 2006

Edited by

Mélanie Hilario
Claire Nédellec

Workshop Chairs

Mélanie Hilario
Artificial Intelligence Lab, CUI, University of Geneva
Melanie.Hilario@cui.unige.ch

Claire Nédellec
MIG, Institut National de la Recherche Agronomique (INRA)
Claire.Nedellec@jouy.inra.fr

Program Committee

Florence d'Alché-Buc	University of Evry (France)
Sophia Ananiadou	University of Manchester & NaCTeM (UK)
Christian Blaschke	Bioalma (Spain)
Nigel Collier	National Institute of Informatics (Japan)
James Cussens	University of York (UK)
A. Fazel Famili	NRC Institute for IT (Canada)
Lynette Hirschman	The MITRE Corp. (USA)
Alexandros Kalousis	University of Geneva (Switzerland)
Stefan Kramer	University of Freiburg (Germany)
Maria Liakata	University of Wales, Aberystwyth (UK)
Adeline Nazarenko	Université Paris-Nord (France)
See-Kiong Ng	Institute for Infocomm Research (Singapore)
Srinivasan Parthasarathy	Ohio State University (USA)
Céline Rouveirol	LRI Orsay (France)
Jasmin Saric	EML Research gGmbH (Germany)
Jude Shavlik	University of Wisconsin-Madison (USA)
Jaak Vilo	University of Tartu (Estonia)

Committee of Biology Experts

Terri K. Attwood	University of Manchester (UK)
Philippe Bessières	MIG-INRA (France)
François Radvanyi	Institut Curie (France)
Jean-Charles Sanchez	University Hospital of Geneva (Switzerland)
Antonia Vlahou	BioAcademy of Athens (Greece)

Additional Reviewers

Mark Goadrich	University of Wisconsin-Madison (USA)
Ai Kawazoe	National Institute of Informatics (Japan)

Sponsors

The workshop organizers gratefully acknowledge the support of our generous sponsors:

Hoffmann-LaRoche AG

Serono International SA

Preface

Increasing use of high-throughput methods in molecular biology has spawned unprecedented masses as well as novel types of data. Gene and protein microarrays, mass spectrometry, and SNP chips are a few examples of technologies that allow biologists to line up hundreds of experiments while studying thousands of genes or proteins in a single experiment. Thus high volume and high dimensionality are hallmarks of biological data that data miners must cope with.

The availability of comprehensive datasets on key biological entities has led life scientists from a reductionist, component-centred approach to a more holistic or systemic approach. They can now examine interactions among proteins or between DNA and proteins to build models of molecular pathways and networks in an effort to understand the functioning of cells, tissues and organisms. The trend toward systems biology compounds problems of scale and high dimensionality with that of increasing complexity: analysis must be pursued at multiple levels of organization in order to achieve a comprehensive and coherent view of a system's structure and dynamics. Systems biologists expect data miners to provide them with the computational tools for representing, integrating and modeling heterogeneous data as well as deciphering complex patterns and systems.

Notwithstanding the massive amounts of biological data accessible in around a thousand databases, the ultimate source of up-to-date information in the field remains the biological literature. Despite the efforts of curators who tirelessly scan known document servers, existing databases cannot keep up with the current pace of scientific publishing; text documents contain critical information that is not and may never be found in structured databases. There is a need for efficient methods of retrieving relevant documents and extracting information needed by both curators and biological practitioners. Text mining research has made significant strides in recent years, yet the stack of well-known unsolved problems (e.g., anaphora resolution, word sense disambiguation) is crumbling under new challenges such as analyzing multiple text grain levels to extract and formalize complex information concerning pathways, networks and systems.

The emergence of systems biology poses a new set of challenges for both data mining and text mining research. The ECML/PKDD-2006 Workshop on Data and Text Mining for Integrative Biology (Berlin, September 18, 2006) had a twofold goal. The first was to bring together researchers in data and text mining to discuss latest insights and innovations related to biological knowledge discovery from data and text. The technical papers presented in this volume cover issues ranging from consecutive support in genomic profiling to recognition of named biological entities in text and extraction of their properties, functions or relations. The second, more exploratory goal was to initiate a dialogue between biologists and data/text miners on data-analytical research tracks opened by the current trend toward systems biology. Position papers were solicited from both

the data mining and the biology communities to serve as a basis for discussion. Biologists' position papers describe ongoing work on protein family annotation, organogenesis, bacteria characterization and multiparameter cancer analysis, all of which could reap significant benefits from the use of automated knowledge discovery techniques. In return, data and text miners can draw fresh inspiration to devise new models and methods from the computational hurdles faced by biology as it evolves from a descriptive to a more quantitative, systems-oriented science.

Particular thanks go to our biologist friends who bravely accepted to participate in this cross-discplinary encounter within a highly specialized computer science conference such as ECML/PKDD. It is our hope that this workshop will take us a step closer to an interdiscplinary discovery science geared to an integrative understanding of the complex mechanisms of life.

Melanie Hilario and Claire Nédellec
Workshop Chairs

Table of Contents

Invited Talks

Linking Text with Knowledge – Challenges in Text Mining for Biology 1
Junichi Tsujii

Mining Large-Scale Data Sets on the Eukaryotic Cell Cycle 3
Lars Juhl Jensen

Technical Papers

Automated Information Extraction from Gene Summaries 4
Thierry Charnois, Nicolas Durand, Jirí Kléma

Using Consecutive Support for Genomic Profiling 16
Edgar H. de Graaf, Jeannette de Graaf, Walter A. Kosters

Identifying Heterogeneous and Complex Named Entities in Biology Text 28
Julien Lorec, Gérard Ramstein, Yannick Jacques

Annotation Guidelines for Machine Learning-Based Named Entity Recognition in Microbiology 40
Claire Nédellec, Philippe Bessières, Robert Bossy, Alain Kotoujansky, Alain-Pierre Manine

Position Papers: The Data/Text Mining Perspective

On the Feasibility of Heterogeneous Analysis of Large Scale Biological Data.. 55
Ivan G. Costa, Alexander Schliep

Marker Analysis with APRIORI-Based Algorithms 61
Giacomo Gamberoni, Evelina Lamma, Fabrizio Riguzzi, Sergio Storari, Chiara Scapoli

Position Papers: The Biological Perspective

Getting the Unknown from the Known in Bacteria, and the Role of Text Mining 67
Philippe Bessières, Robert Bossy, Alain-Pierre Manine, Erick Alphonse, Claire Nédellec

Towards an Integrated Bioinformatics System: Using Integrins as Case Study to Dissect the Molecular Basis of Cell Adhesion 73
Sophia Kossida, Charikleia Falkou, Ioannis Vasileiou, Christos Zervas

Challenges for Protein Family Annotation 81
Alex Mitchell, Ioannis Selimas, Teresa K. Attwood

Multiparameter Analysis of Cancer: How can Data and Text Mining Help? .. 88
François Radvanyi, Nicolas Stransky, Céline Rouveirol

Meeting the Challenge: Towards a Data and Text Mining Infrastructure for Biological Research 89
Alexandros Kalousis, Mélanie Hilario

Invited Talk

Linking Text with Knowledge - Challenges in Text Mining for Biology

Junichi Tsujii

Department of Computer Science
Faculty of Information Science and Technology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan

University of Manchester and National Centre for Text Mining (NaCTeM)
Manchester Interdisciplinary Biocentre, Oxford Road, Manchester, M13 9PL,UK

Extended Abstract

With an overwhelming amount of biomedical knowledge recorded in texts, it is not surprising that there is so much interest in techniques which can identify, extract, manage, integrate and exploit this knowledge, moreover discover new, hidden or unsuspected knowledge. It is noteworthy that the number of MEDLINE searches in January 1997 was 0.163 million compared to 82.027 millions in March 2006. MEDLINE contains approximately 15 million bibliographic units and its size is increasing at a rate of more than 10 % each year. With popularity of open access journal publishing, full text articles are becoming more and more available. The demand for tools dealing with ever-increasing knowledge embedded in text is real.

While there are a few Text Mining tools on market, they hardly satisfy actual requirements of biologists. Simple application of data mining techniques to text does not work. Since language and text have their own inherent structures, it is essential for TM tools to be able to recognize and exploit their structures to reveal information encoded in them. However, the major difficulties in treating information encoded in language are caused by the nature of the mapping between surface linguistic forms and information conveyed by them. It is hugely ambiguous. Furthermore, the same information can be conveyed by using many different surface forms. Before more ambitious goals such as discovering new, hidden knowledge, we have to resolve these essential properties of the mapping between language and information.

Although techniques have been developed in natural language processing (NLP) research to resolve the difficulties, they were considered, until very recently, non-deployable for large scale text mining. However, due to recent technological

development in corpus-based NLP techniques, many of NLP techniques have become robust and efficient enough for large scale text mining applications. The progresses in the field have been enormous, which will open up many possible applications of NLP-based Text Mining in the near future.

As examples, I will talk about the following new developments that my group at NaCTeM at Manchester and the University of Tokyo are jointly carrying out.

1. Full deep parsing by GRID and its application in Information extraction of protein-protein interaction and disease-gene association
2. Semantic Annotation of biological events and GENIA event ontology
3. Lexical Resource Building and Named Entity Recognition

Invited Talk

Mining Large-Scale Data Sets on the Eukaryotic Cell Cycle

Lars Juhl Jensen

European Molecular Biology Laboratory
Meyerhofstrasse 1
69117 Heidelberg, Germany
Lars.Jensen@embl-heidelberg.de

Abstract. Recent advances in high-throughput technologies have resulted in a flood of large-scale biological data sets. These are a challenge to the data mining community for two reasons primarily: first, most of the data sets contain a lot of noise and can thus not be considered as "facts", and second, knowledge about the underlying biology is needed to make sense of them. In this talk, I will focus on the available data relevant to the eukaryotic cell cycle. I will try to illustrate how careful reanalysis of the data can drastically improve the signal-to-noise ratio. how integration of heterogeneous data sets can be used in an explorative fashion to make new biological hypotheses, and finally how these can sometimes be proven through comparison of data from multiple organisms.

Automated Information Extraction from Gene Summaries

Thierry Charnois, Nicolas Durand, and Jiří Kléma

GREYC, CNRS - UMR 6072, Université de Caen
Campus Côte de Nacre, F-14032 Caen Cédex France
{Forename.Surname}@info.unicaen.fr

Abstract. Automated extraction of links among biological entities from free biological texts has proven to be a difficult task. In this paper we propose and solve a modified task in which we extract the links from short textual gene summaries collected automatically from NCBI website. The main simplification lies in the fact that each summary is unambiguously attached to a single gene. The agent part of binary biological interactions is thus known by default, the goal is to identify meaningful target parts from the summary. The outcome is a structured representation of each summary that can be used as background knowledge in consequent mining of gene expression data. As the gene summaries highly interact with the other structural information resources provided by NCBI website, these resources can be used as an annotation tool and/or a feedback for performance optimization of the system being developed. In particular we use the gene ontology terms in order to evaluate and improve the information extraction process.

Keywords: genomics, text mining, biological information extraction.

1 Introduction

As availability of textual information related to biology increases, research on information extraction (IE) is rapidly becoming an essential component of various bio-applications. It is expected that text mining in general, and IE in particular, will provide tools to facilitate the annotation of a large amount of genetic information, including gene sequences, transcription profiles and biological pathways. The biological function of cells, tissues and organisms can be understood by examination of interactions among proteins or between DNA and proteins.

The main interest has been devoted to MedLine abstracts, however there is also a vast effort to exploit full-text journal articles [20]. Applying IE to genomics and more generally to biology is not an easy task because IE systems require deep analysis methods to extract the relevant pieces of information. That is why we propose a modified task in which we extract the links from short textual gene summaries collected automatically from NCBI website. The main simplification lies in the fact that each summary is unambiguously attached to a single gene. The agent part of binary biological interactions is thus known by default, the goal is to identify meaningful target parts from the summary.

This work has started with the intention to develop a meaningful measure of interaction inside a closed set of genes in order to support consequent mining of gene expression data. Such a measure can be used in many ways. The measure can complement the gene distance measure based immediately on the expression data when the genes are clustered [15]. It can be used to select biologically meaningful patterns from the overwhelming pattern sets that technically appear in the expression data [17] or it can help in feature extraction and selection when a classification task is solved [24].

Public databases contain vast amount of rich data that can be used to create and evaluate both direct and indirect interactions among biological entities. Of course, the most straightforward way is to utilize the structured information such as gene ontology (GO) or Entrez's link files. The rationale sustaining the GO based measure is that the more GO terms the genes share, and the more specific the terms are, the more likely the genes are to be functionally related. [19] defines a distance based on the Czekanowski-Dice formula, the methodology is implemented within the GOProxy tool of GOToolBox [1]. [22] uses Entrez's link files in order to create a general entity graph. The authors also provide a measure that assesses the strength of a link between an arbitrary pair of vertices.

Nevertheless, the structured databases can hardly summarize all the available knowledge and text mining outcomes can reasonably complement the information gained from the knowledge sources mentioned in the previous paragraph. Tagging gene and protein names in free biomedical text has proven to be a difficult task [23]. Automated extraction of direct links among biological entities is even more difficult [10]. In this paper we restrict to a corpus of gene textual summaries. Possible interaction among a closed set of genes is studied indirectly. The main aim of the paper is to develop a structured and tagged representation of gene summaries. This structured representation can later serve to assume on interaction or similarity among the genes from multiple points of view. The proposed structured representation also seems to be promising with respect to its further generalization. The developed system provides an insight into relations among biological entities and it can be adjusted to extract arbitrary interactions among biological entities from abstracts or whole texts.

This paper is structured as follows. Section 2 briefly introduces the data we worked with. Section 3 gives an overview of related methods. Section 4 describes a tool LinguaStream that we have used, discusses the developed extraction rules and gives examples of real outputs. Section 5 provides two ways of evaluation – the first is based on a limited corpus of human annotated summaries, the second evaluates the full corpus with respect to GO terms.

2 Entrez Gene Summaries

Entrez Gene is the gene-specific database at the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM). Entrez Gene provides unique integer identifiers for genes and other loci for a subset of model organisms. It tracks those identifiers, and is integrated with the Entrez system for interactive query, LinkOuts, and access by E-utilities [25].

The information that is maintained includes nomenclature, chromosomal localization, gene products and their attributes (e.g. protein interactions), associated markers, phenotypes, interactions, and a wealth of links to citations, sequences, variation details, maps, expression reports, homologs, protein domain content and external databases.

As mentioned in Section 1 the long term goal is to develop a meaningful measure of interaction inside a closed set of genes. In our experiments we deal with the SAGE human gene expression dataset downloaded from [4]. Only the unambiguous tags (corresponding to genes) identified with RefSeq were selected, leaving a set of 11082 tags (expressed in 207 biological situations).

To access the gene annotation data for every tag considered, RefSeq identifiers were translated into EntrezGene identifiers [3]. The mapping approached 1 to 1 relationship. There were only 11 unidentified RefSeqs, 24 RefSeqs mapped to more than 1 id and 203 ids still appeared more than once. Knowing the gene identifiers, the annotations were automatically accessed through hypertext queries to the Entrez Gene database [4] and sequentially parsed by the method stemming from [28]. The non-trivial textual records were obtained for 6,302 ids which makes 58% of the total amount of 10,858 unique ids. 3,926 genes had a short summary, 5,109 had one abstract attached at least. 6,824 genes had at least a single GO term attached, which makes 63% of the total amount of genes.

3 Information Extraction and Methodology

Many approaches have been proposed for extraction of biological information from scientific texts. These approaches can be classified into two broad categories [8]: machine learning based and linguistic analysis based. That is the latter one, and more precisely IE technique, which is used in this paper. Some of the IE systems use similar approaches with the Natural Language Processing (NLP) understanding systems of the seventies/eighties and IE is often seen as a NLP understanding system. In fact, they do not share the same goal. IE aims at extracting very precise information from a restricted domain while the goal of the NLP systems was the whole understanding of all aspects of the text. For this purpose, extensive knowledge and linguistic resources were needed, and deep analysis was necessary (syntactic, semantic and pragmatic analysis).

Taking advantage of the restricted domain, some biomedical IE systems adopt this NLP based architecture [11, 14]. Nevertheless, the syntactic analysis still remains a difficult task. Actually the accuracy of the complete parsing can be estimated roughly about 50% of the analysed sentences (see [11]). Other works attempt to use "shallow parsing", a robust method, although less precise performing a partial decomposition of a sentence structure to identify phrasal chunks or entities of interest and relations between these entities. Generally, these kinds of systems are designed for extracting protein-protein relations, such as protein-location relations, binding relations, gene-gene interactions, etc. [21]. A common point of the IE systems is that they utilize resources, biological databases, ontologies, such as UMLS, LocusLink ...

Some other papers are devoted to a preliminary task: the recognition of gene/protein names and families. Difficulties are well known: multi-sense words, no formal criterion, multi-word terms, variations in gene/protein names. Different NLP methods are used for this like rule-based approach [12], or/and dictionary/knowledge approach [16, 18].

Our system differs from the approaches previously mentioned in several ways. For example, [14] carries out a terminological parsing, using a biological knowledge database, syntactic, semantic and discoursive analysis, using a domain model (ontology) to get a predicate argument representation and to fill an extraction template. Instead of those kinds of classical NLP techniques, we design simple declarative extraction rules, making the implementation process "light and quick". Let us note that they are domain-specific, but by no means corpus-specific. One of the aims is to reach similar results as the "heavy" methods published in the literature.

The design of the rules can be seen as a simplification of the "contextual exploration method" [26]. This approach aims at locating contexts in a corpus (i.e., linguistic indicators) from which some rules for identifying relevant textual segments are triggered. For example, linguistic indicators as "our conclusion", "consequently" or "so" found in a corpus can allow the extraction of conclusive sentences. That is the idea of our system: triggering extraction rules only if a context is located while avoiding the whole-corpus analysis. Another similarity with our work is that no syntactic analysis is processed. However, unlike our approach, this method is domain-independent, so linguistic indicators and extracted informations are general (causality, thematic announcement, conclusive sentences, ...).

Another important point is that our method is endogenous: no resources such as knowledge base or dictionary are needed at the beginning. The resources are constructed on the fly – the system learns new terms (which can be new terms in the domain or missing in the databases) to be used later or in other biological corpora and/or in other text mining applications.

Finally, our system is not designed to focus only on a specific aspect of gene/protein description but it is designed to identify protein/family/name and general biological function about the gene/protein involved. Actually four *types* of information are distinguished and annotated in the corpus: gene/protein name, family name, location and biological function.

4 Method

The presented approach consists in definition of extraction rules, and has been implemented using the LinguaStream platform.

4.1 LinguaStream

LinguaStream [2, 7] is an integrated experimental environment targeted to NLP researchers. It allows complex experiments on corpora to be realised conveniently, using various declarative formalisms.

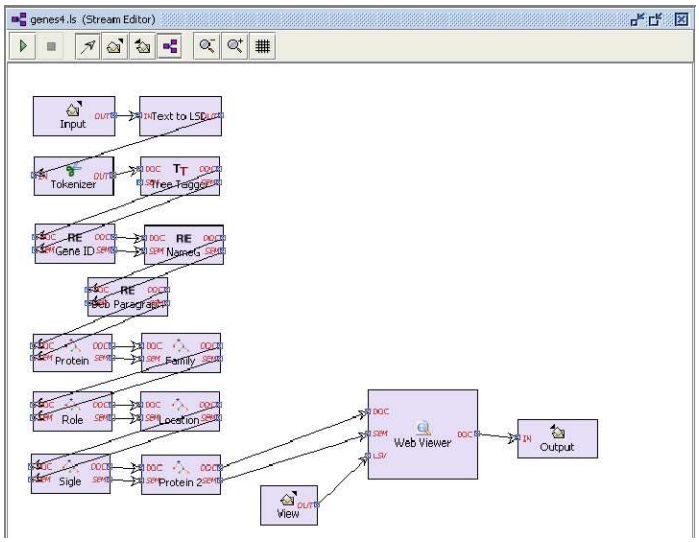


Fig. 1. Processing stream of the implemented rules in LinguaStream.

Its integrated environment allows processing streams to be assembled visually (see Figure 1), picking individual components from a "palette". Some components are specifically targeted to NLP, while others solve various issues related to document engineering (especially to XML processing). Annotations made on a single document are organized in independent layers and may overlap. Thus, concurrent and ambiguous annotations may be represented in order to be solved afterwards, by subsequent analysers. The platform is systematically based on XML recommendations and tools, and is able to process any file in this format while preserving its original structure. When running a processing stream, the platform takes care of the scheduling of sub-tasks, and various tools allow the results to be visualised conveniently.

IE from a raw text is composed of tokenization, POS tagging (using TreeTagger [5]), extraction and output generation which adds the final XML wrapper. Among fundamental principles, the platform allows the **declarative representations** to be used. Furthermore, the **modularity** of processing streams promotes the **reusability** of components in various contexts: a given module, developed for a first processing stream may be used in other ones. Section 4.2 demonstrates their utility.

4.2 Extraction rules

We have defined a set of rules to identify, extract and annotate relevant multiword terms from gene summaries. The results are given in a form of XML file containing the whole text where the recognized areas are highlighted and clickable (see Figure 2), and another XML file with the extracted information only (see Figure 3). Let us refer to these files as the *interactive* and *extracted* output.

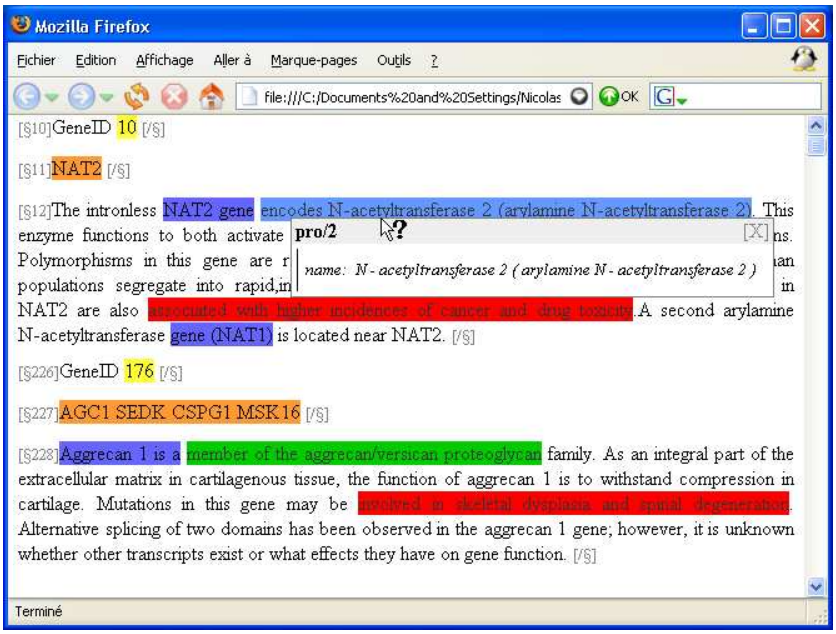


Fig. 2. Example of XML result.

The rule definition is decomposed into the following steps:

- Observe the corpus (in fact the training corpus) in order to get regularities and identify some contexts. For example, the expression "this gene encodes the X protein" has numerous occurrences in the corpus, so "encode" is a good context – a trigger word – to identify a protein name;
- Design rules from the contexts previously identified (a particular example is shown later).
- Implement the rules. It is a straightforward process using DCG Prolog and unification on feature structures thanks to GULP [9].
- Review the results after processing the rules and backtrack if necessary. This can be changing rules, or adding rules while possibly reusing the terms/knowledge already recognized/learned.

We have defined 4 sets of rules allowing the system to recognize 4 types of information: protein names, family names of proteins, roles / biological functions (including diseases, interactions, ...), and location (components, ...).

General structure of the rules We do not use patterns in the sense of the IE, that is without an a priori on the form of the expressions. Figure 4 presents the structure of the rules. From a "context", an expression (generally a multi-word term, a nominal phrase) is recognized until a stop phrase is encountered. The context is a set of "trigger" words. The stop phrases can be words, symbols, verbs, punctuation, ... They depend on the rule type.

```
<gene>
<id>10</id>
<name>NAT2</name>
<protein>encodes N-acetyltransferase 2 (arylamine N-acetyltransferase 2)</protein>
<role>associated with higher incidences of cancer and drug toxicity</role>
</gene>
<gene>
<id>176</id>
<name>AGC1 SEDK CSPG1 MSK16</name>
<protein>Aggrecan 1 is a</protein>
<family>member of the aggrecan/versican proteoglycan</family>
<role>involved in skeletal dysplasia and spinal degeneration</role>
</gene>
```

Fig. 3. The extracted results.

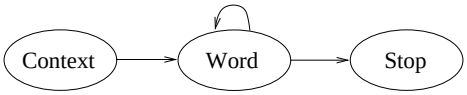


Fig. 4. Structure of the rules.

Let us take an example. The term "encodes N-acetyltransferase 2 (arylamine N-acetyltransferase 2)." is extracted by using the following set of rules:

```
protein(type:pro..name:N) --> @lemma:encode, np(N).
np(N) --> @tag:dt, namepro(N).
np(N) --> namepro(N).
namepro(N) --> ls_token(N,_), end.
namepro(N) --> ls_token(N,_), namepro(N2), {concat(N1,N2,N)}.
end --> punctuation ; verb ; relative_pronoun ; trigger_word.
```

In this example, "namepro" stands for the name of the protein to be extracted, "ls_token" a terminal symbol (a token), "end" the indicator for cutting out the recognition of a multi-word term. The trigger phrase is "encode" (i.e., also encoded, encodes atc.). Let us remark that "dt" corresponds to a possible determiner just before the name of the protein. The rules "namepro" allow the system to recognize the multi-word terms. The end phrase is a punctuation symbol here.

Context and stop phrases Currently, the identification of the context has been done manually, however an automatic learning of the context can also be considered. We have manually detected special phrases for each type of information (proteins, families, ...) on an excerpt of the corpus. We have noted the corresponding trigger words and the stop phrases. Table 1 presents some examples of trigger phrases for each type. The common stop phrases are the trigger words, some punctuation symbols and the relative pronouns.

The final rulebase consists of 186 rules – 74 for proteins, 46 for families, 27 for roles, 39 for locations.

Special processing Another set of rules benefits from the reusability principle of LinguaStream. It enables us to use the information (tokens, trigger

phrases, ...) recognized earlier within the current processing stream. Some protein names/families are recognized using entities already identified. These entities are considered as lexical units (tokens) in the rules. For instance, "`X are class of FAMILY`" where `FAMILY` is the entity previously recognized as a protein family, and `X` is the new extracted information: here an other protein family.

A special process for recognizing protein names expressed by an acronym is done. All acronyms are marked by a few special rules that extract words with upper cases, numerical figures and/or special symbols as in [12]. Then, the acronym context is used to decide whether the acronym corresponds to a protein name. For example, "protein CEBP-alpha" is detected using these specific rules. Here the rule is: the word "protein" followed by an acronym. We also use particular rules to filter out false or misleading expressions. For instance, the term "the secreted protein" must not be extracted as a protein name.

proteins	encodes an ...	@lemma:encode, @tag:dt
	the product of this gene is ...	@lemma:product, $'of', $'this', $'gene', @lemma:be
families	belongs to the ...	@lemma:belong, $'to'
	is a member of the ...	@lemma:member, $'of'
roles	an important role in ...	$'role', $'in'
	is involved in ...	$'involved', $'in'
locations	found in ...	$'found', $'in'
	located in ...	$'located', $'in'

Table 1. Examples of detected contexts.

4.3 Outcome

The corpus is about 2.33MB and contains 64,308 lines. There are 10,858 genes. In order to learn the contexts and the stop phrases, we have looked over 200 genes (1.8% of the corpus).

Type	No. text areas
proteins	3,058
families	3,056
roles	4,303
locations	1,023

Table 2. The number of recognized terms (text areas) according to their type.

The number of marked text areas (i.e., multi-word terms or pieces of extracted information) is presented in Table 2. Let us note that in a gene summary, the information about proteins, families, ... is not always present. We observe that the number of "roles" is larger than the number of "proteins". As a matter of fact, a single gene may have several "roles" because the type role contains biological functions, diseases and also different interactions. As regards families, we capture families and also subfamilies and superfamilies, if they are indicated.

The system has recognized 3,058 protein names. By rule of thumb, this is a relatively good result since there are 6,932 genes without summaries (i.e. without any chance to extract information). On average, there is nearly one protein name

extracted per existing summary. A more detailed evaluation of the performance is given in the next Section.

5 Evaluation

Two types of experiments have been carried out. First, we have evaluated the precision and the recall of the method using an excerpt of the data. The second experiment is a direct comparison between our extracted terms and the GO terms annotating the individual genes.

5.1 Evaluation on a human annotated corpus

We have evaluated our approach using 100 genes (and the corresponding summaries) randomly chosen. This excerpt has been annotated by two local experts to form the reference. We have computed the classical measures of precision and recall [8] to assess the performance of our system.

Table 3 presents the results and relates them to the results obtained by other existing methods published in literature. The comparison is illustrative only as the methods were not applied to the same corpus. The precision and recall values cannot be compared directly, but they may give an estimation of the performance. As we can see, the results of our system are comparable to the existing scores, without using a "heavy method" nor resources.

method		recall	precision
existing methods: [12, 18, 13, 27]		73-99%	73-95%
our approach:	proteins	73,6%	78,8%
	families	71,6%	93,4%

Table 3. Results.

Distinguishing various term types, a good precision and recall has been reached for families. Actually, for the family names and the locations, the implemented rules are appropriate to extract information from summaries. Moreover, we have recently improved the rules by using the results of this evaluation and by observing the information not recognized to define new contexts.

As for biological functions, the important point is to have a relatively complete list of the commonly used verbs. Our "list" is good enough, and it is easy to add new verbs to capture more cases. For this, ideas from specific works like [6] can be used.

The main problem concerns the recognition of proteins names. The current rules are able to capture the majority of the names, however some particular linguistic problems are not treated yet: anaphoras and coordination. For instance, in the phrase "the related proteins CEBP-alpha, CEBP-delta, and CEBP-gamma" all the three acronyms are recognized but only the first one ("CEBP-alpha") is identified as a protein name (the rule given above). The others would need to take the coordination problem into account.

5.2 Comparison with GO

A great part of GO terms associated with genes also appear in their summaries. In other words, if a gene is annotated with a GO term, this term (or its semantically equivalent phrase) often appears in the summary of the given gene too. This significant overlap between GO terms and summaries gives us a chance to utilize GO terms as an annotation tool for gene summaries. The quality of information extraction can be tested with respect to recall of the known GO terms. The main advantage of such an experiment is that it enables us to automatically evaluate the system over whole the corpus of gene summaries.

The basic assumption of this evaluation is that all the GO terms represent meaningful terms to be extracted. Then, the recall is estimated as the ratio between the number of GO terms identified within the extracted XML annotation and the number of GO terms that appear within the original summaries. The ideal case occurs when all the GO terms that appear within the gene summary of the given gene remain also in its XML record – recall would be 1 here.

The main and difficult problem is to identify the GO terms within free text of gene summaries. First, let us see what is the percentage of GO terms that appear in gene summaries immediately – as exactly the same term or phrase. The simple search for substrings suggests that only 7% of GO terms associated with the given gene co-appear in its summary immediately. These are mainly one word terms since for longer phrases the exact match is less likely – e.g., the GO term "amino acid metabolism" appears in the summary as an expression "function in the catabolism and salvage of acylated amino acids". That is why we have also applied a simple form of approximate match for longer phrases. If at least one of the stemmed words from the GO phrase appears in the gene summary exactly, we search for an approximate match of the other words in the same summary sentence. We use the bigram approximate string comparison for this purpose. The phrase is found if and only if the average of best-match values – we search for the nearest counterpart for all the words from the GO phrase – reaches a certain threshold. This simple approximate match reveals that 18% of GO terms associated with the given gene co-appear in the respective summary.

matching	original summaries	extracted XML	recall
exact	7%	3.9%	56%
approximate	18%	8.2%	46%

Table 4. Recall of GO terms – the exact and approximate match.

Table 4 gives an overview of the recall for exact and approximate GO terms. Precision of the IE cannot be revealed in this way as we do not search for GO terms only. Let us remind that we are interested in any biological terms and the goal is not to confine ourselves to the limited dictionary of GO terms. Nevertheless, the recall presented in Table 4 should be evaluated with respect to the condensation that the extraction process brings. The content of the extracted output makes 29.2% of the content of the original gene summary files. It also has to be considered that the sentences in gene summaries can be quite long while the extracted tags are quite compact. The chance that the GO phrase is scattered by tag tokens is thus increased.

6 Conclusion

In this paper, we presented an original approach for extracting and exploiting information from biological domain. The approach gives promising results on a specific but wide corpus. It is suitable for extracting biological information as well as to acquire knowledge. It also seems to be promising with respect to its further generalization. The developed grammar provides an insight into relations among biological entities and it can be adjusted to extract arbitrary interactions among biological entities from abstracts or whole texts using the acquired information such as terminological resources. The second investigation will consist in enhancing the grammar by an automated learning of context [6]. The process has not been designed yet but the terms already learned can guide and accelerate the learning process (to annotate the other corpus, etc.).

Another work is to propose and test a gene similarity measure based on the developed structured representation. While the measure itself is more or less obvious – the more overlap two genes show in their corresponding type fields the more they interact – the main effort will be to show a possible difference in comparison with the measures that are immediately based on the GO annotations [19] or a vector representation of whole summaries [17].

Acknowledgements. The authors thank A. Widlöcher and F. Bilhaut (the Linguastream team), and the CGMC Laboratory (CNRS UMR 5534, Lyon, France) for providing the gene expression database. This work has been partially funded by the ACI "masse de données" (French Ministry of research), Bingo project (MD 46, 2004-2007).

References

[1] GOToolBox website: http://crfb.univ-mrs.fr/gotoolbox/.
[2] LinguaStream website: http://www.linguastream.org/.
[3] MatchMiner website: http://discover.nci.nih.gov/matchminer/.
[4] NCBI website: http://www.ncbi.nlm.nih.gov/.
[5] TreeTagger website: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/.
[6] P. Bessières, G. Bisson, A. Nazarenko, C. Nédellec, M. Ould Abdel Vetah, and T. Poibeau. Ontology Learning for Information Extraction in Genomics Bibliography - the Caderige Project. In *Journées IMPG Ontologie et Extraction d'Information en Génomique*, Grenoble, France, May 2001.
[7] F. Bilhaut and A. Widlöcher. LinguaStream: An Integrated Environment for Computational Linguistics Experimentation. In *the European Chapter of the Association of Computational Linguistics (Companion Volume)*, Trento, Italy, 2006.
[8] K. B. Cohen and L. Hunter. *Artificial Intelligence Methods and Tools for Systems Biology*, volume 5, chapter Natural Language Processing and Systems Biology. Springer Verlag, 2004.
[9] M. A. Covington. GULP 3.1: An Extension of Prolog for Unification-Based Grammar, 1994.
[10] J. Cussens and C. Nédellec, editors. *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, Bonn, August 2005.
[11] N. Daraselia, S. Egorov, A. Yazhuk, S. Novichkova, A. Yuryev, and I. Mazo. Extracting Protein Function Information from MEDLINE Using a Full-Sentence

Parser. In *ECML/PKDD Workshop on Data Mining and Text Mining for Bioinformatics*, Pisa, Italy, Sept. 2004.

[12] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi. Toward Information Extraction: Identifying Protein Names from Biological Papers. In *Pacific Symposium Biocomputing (PSB'98)*, pages 362–373, Hawaii, Jan. 1998.

[13] K. Fundel, D. Güttler, R. Zimmer, and J. Apostolakis. A Simple Approach for Protein Name Identification: Prospects and Limits. *BMC Bioinformatics*, 6(Suppl 1), 2005.

[14] R. J. Gaizauskas, G. Demetriou, P. J. Artymiuk, and P. Willett. Protein Structures and Information Extraction from Biological Texts: The PASTA System. *Bioinformatics*, 19(1):135–143, 2003.

[15] P. Glenisson, J. Mathys, and B. D. Moor. Meta-Clustering of Gene Expression Data and Literature-Based Information. *SIGKDD Explor. Newsl.*, 5(2):101–112, 2003.

[16] K. Humphreys, G. Demetriou, and R. Gaizauskas. Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. In *Pacific Symposium Biocomputing*, pages 505–516, Hawaii, Jan. 2000.

[17] J. Kléma, A. Soulet, B. Crémilleux, S. Blachon, and O. Gandrillon. Mining Plausible Patterns from Genomic Data. In *the 19th IEEE International Symposium on Computer-Based Medical Systems*, pages 183–188, Salt Lake City, Utah, 2006.

[18] A. Koike and T. Takagi. Gene/Protein/Family Name Recognition in Biomedical Literature. In *Linking Biological Literature, Ontologies and Databases: Tools for Users, Workshop in conjunction with NAACL / HLT 2004*, pages 9–16, 2004.

[19] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq. GOToolBox: Functional investigation of gene datasets based on gene ontology. *Genome Biology*, 5(12):R101, 26 Nov. 2004.

[20] S. K. Parantu, P.-I. Carolina, B. Peer, and A. A. Miguel. Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4:20, 2003.

[21] J. Pustejovsky, J. Castano, J. Zhang, M. Kotecki, and B. Cochran. Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations. In *Pacific Symposium on Biocomputing (PSB'02)*, pages 362–373, Hawaii, Jan. 2002.

[22] P. Sevon, L. Eronen, P. Hintsanen, K. Kulovesi, and H. Toivonen. Link Discovery in Graphs Derived from Biological Databases. In *3rd International Workshop on Data Integration in the Life Sciences 2006 (DILS'06)*, Hinxton, UK, July 2006.

[23] L. Tanabe and W. J. Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18:1124–1132, 2002.

[24] J.-P. Vert and M. Kanehisa. Graph-Driven Feature Extraction From Microarray Data Using Diffusion Kernels and Kernel CCA. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS*, pages 1425–1432. MIT Press, 2002.

[25] D. Wheeler, D. Benson, and S. Bryant. Database Resources of the National Center for Biotechnology Information: Update. *Nucleic Acids Res.*, 33:D39–D45, 2005.

[26] D. Wonsever and J.-L. Minel. Contextual Rules for Text Analysis. In *CICLing '01: Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, pages 509–523, London, UK, 2001. Springer.

[27] H. Yu, V. Hatzivassiloglou, C. Friedman, A. Rzhetsky, and W. Wilbur. Automatic Identifying Gene/Protein Terms in MEDLINE Abstracts. *Journal of Biomedical Informatics*, 35(5-6), 2002.

[28] F. Zelezny, J. Tolar, N. Lavrac, and O. Stepankova. Relational Subgroup Discovery for Gene Expression Data Mining. In *EMBEC: 3rd IFMBE European Medical & Biological Engineering Conf.*, November 2005.

Using Consecutive Support for Genomic Profiling*

Edgar H. de Graaf, Jeannette de Graaf, and Walter A. Kosters

Leiden Institute of Advanced Computer Science, Leiden University, The Netherlands
{edegraaf,graaf,kosters}@liacs.nl

Abstract. We propose a new measure of support (the number of occurrences of a pattern), in which instances are more important if they occur with a certain frequency and close after each other in the stream of records. We will explain this new consecutive support and show how consecutiveness and the notion of hypercliques can be incorporated into the ECLAT algorithm.
Synthetic examples show how interesting phenomena can now be discovered in the datasets. The new measure can be applied in many areas, ranging from bio-informatics to trade, supermarkets, and even law enforcement. We will use it in genomic profiling, where it is important to find patterns contained in many individuals: patterns close together in one chromosome are more significant.

1 Introduction

In earlier research we explored the use of frequent itemsets to visualize deviations in chromosome data concerning people with a certain illness, genomic profiling [4]. During our exploration of this problem it became apparent that patterns are more important when the areas (transactions) in which they occur are close together. The consecutiveness of transactions containing the pattern plays an important role in other applications too. Patterns are frequent sets of items, where frequent means that their support, that can be defined in different ways, is more than a pre-given threshold. In the biological problem the items are individuals and the transactions are "clones", pieces of the chromosome that might occur more or less often than in a healthy individual. Patterns in close transactions are better because they are close together in the chromosome and are biologically more significant than patterns that are far apart and in different chromosomes.

Consecutive support informally is the support or the number of occurrences of patterns where we take into account the distance between transactions that contain the pattern: the consecutive support should be higher when occurrences are close together. Here distance is the number of in-between transactions that do not contain the pattern. Of course, this only makes sense if the transactions

* This research is carried out within the Netherlands Organization for Scientific Research (NWO) MISTA Project (grant no. 612.066.304).

are given in some logical order. We will use consecutive support for genomic profiling, however this type of support can be applied in a number of other domains:

- **Supermarket**. E.g., big supermarkets receive large quantities of goods every day. Knowing which goods will be sold in large quantities close in time helps the supermarket decide when to refill these goods.
- **Trading**. E.g., a combination of stock being sold once may lead to waves of these stocks being sold close after each other while other combinations might not.
- **Law enforcement**. E.g., when police investigates telephone calls, subjects that are discussed during a longer period might be more interesting than subjects (word combinations) that are mentioned often at separate moments.

In this paper we define consecutive support, having two parameters: the reward factor ρ and the punishment factor σ. Existing pruning methods can be easily incorporated. In particular h-confidence and hypercliques enable us to amplify consecutive behaviour. With our experimental results we show how consecutive support, compared to the results in [4], gives new and interesting patterns when applied to the biological problem of finding patterns in chromosomes.

This research is related to work done on the (re)definition of support and gap constraint. The notion of support was first introduced by Agrawal et al. in [1] in 1993. Much later Steinbach et al. in [8] generalized the notion of support providing a framework for different definitions of support in the future. Our notion of consecutive support is not easily fitted in the eval-function provided there. (Next to this framework Steinbach also provides a couple of example functions.) Frequent itemset mining on similar data was done by Rouveirol et al. in [7]. Our work is related to this work because of the minimal frequency constraint also used in consecutive support.

If we take the database of clones as an example, we have a database where the clones (or transactions) are itemsets of patients with gains or losses in the clones. We could transpose this database so that transactions correspond to the patients, and are itemsets of clones that showed gains or losses. Now we can search for patterns and with techniques like the time window constraint as defined in [5] or the gap constraint as defined in [2], we can search for clones that are close together in the chromosomes. However, the combination of patients with equal clones will be lost.

Finally this work is related to some of our earlier work. Primarily the work done in [4] already stated that the biological problem could profit from incorporating consecutiveness into frequent itemset mining. Secondly in [3] it was mentioned that support is just another measure of saying how good a pattern fits with the data. There we defined different variations of this measure, and consecutive support can been seen as such a variation.

The formal definitions concerning consecutive support are given in Section 2. A particular pruning method is discussed in Section 3. In Section 4 we present experimental results, and we conclude in Section 5.

2 Consecutive Support

2.1 Definition

The definition of association rules relies on that of support: the number of transactions that contain a given itemset. In this paper we propose a more general definition, that takes the consecutiveness of the transactions into account.

Suppose items are from the set $\mathcal{I} = \{1, 2, \ldots, n\}$, where $n \geq 1$ is a fixed integer constant. A *transaction* is an *itemset*, which is a subset of $\mathcal{I}$. A *database* is an *ordered series* of m transactions, where $m \geq 1$ is a fixed integer constant. If an itemset is an element of a database, it is usually referred to as a transaction.

The *traditional support* of an itemset I with respect to a database $\mathcal{D}$, denoted by $\text{TradSupp}(I, \mathcal{D})$, is the number of transactions from $\mathcal{D}$ that contain I. Clearly, $0 \leq \text{TradSupp}(I, \mathcal{D}) \leq m$.

An important property of the traditional support is the so-called Apriori property [1] or anti-monotonicity constraint: if itemset I is contained in itemset I', the support of I is larger than or equal to the support of I'. We want the new measure to satisfy this constraint also.

The support measure we propose is a generalization of the traditional support. In order to take into account the consecutiveness of a pattern we use two real parameters $\rho \geq 0$ and $0 \leq \sigma \leq 1$. With ρ we reward the pattern if it occurs in consecutive transactions, with σ we punish for the gaps between the consecutive occurrences of the pattern in the database.

Suppose we have an itemset I and let $O_j \in \{0, 1\}$ $(j = 1, 2, \ldots, m)$ denote whether or not the j^{th} transaction in the database $\mathcal{D}$ contains I (O_j is 1 if it does contain I, and 0 otherwise; the O_j's are referred to as the *O-series*). The following algorithm computes a real value t in one linear sweep through the database and the resulting t is defined as the *consecutive support* of I with respect to $\mathcal{D}$ (denoted by $\text{Supp}(I, \mathcal{D}, \rho, \sigma)$):

$t := 0$; $j := 1$; $\mathit{reward} := 0$;
while ($j \leq m$) **do**
  if ($O_j = 1$) **then**
    $t := t + 1 + \mathit{reward}$; $\mathit{reward} := \mathit{reward} + \rho$;
  else
    $\mathit{reward} := \mathit{reward} \cdot \sigma$;
  fi
  $j := j + 1$;
od

The consecutive support t can become very large, and one could for example use $\sqrt{t}$ instead. In our examples we will always employ just t.

Example 1. Assume the O-series of a certain pattern I equals 101101, $\rho = 1$ and $\sigma = 0.1$. The consecutive support t will then be 5.41:

O	1	0	1	1	0	1
reward	0	1	0.1	1.1	2.1	0.21
t	1	1	2.1	4.2	4.2	**5.41**

2.2 Formal Discussion

During the loop the value of *reward*, which "rewards" the occurrence of a 1, is always at least 0. If *reward* would never be adapted, i.e., it would remain 0 all the time, independent of the itemset I, the algorithm would compute $\mathrm{TradSupp}(I, \mathcal{D})$. This is the case when $\rho = 0$: $\mathrm{Supp}(I, \mathcal{D}, 0, \sigma) = \mathrm{TradSupp}(I, \mathcal{D})$ for any $0 \leq \sigma \leq 1$. So the consecutive support is indeed a generalization of the traditional support. Furthermore we have: for all $\rho \geq 0$ and $0 \leq \sigma \leq 1$, $\mathrm{Supp}(I, \mathcal{D}, \rho, \sigma) \geq \mathrm{TradSupp}(I, \mathcal{D})$.

It is clear that the APRIORI property is satisfied: for all $\rho \geq 0$ and $0 \leq \sigma \leq 1$, $\mathrm{Supp}(I, \mathcal{D}, \rho, \sigma) \geq \mathrm{Supp}(I', \mathcal{D}, \rho, \sigma)$ if the itemset I' contains the itemset I. This follows from the observation that the *reward*-values in the I'-case are never larger than those in the I-case.

Finally, we easily see that $0 \leq \mathrm{Supp}(I, \mathcal{D}, \rho, \sigma) \leq m + m(m-1)\rho/2$. The maximum value is obtained if and only if all transactions from the database $\mathcal{D}$ contain I, i.e., an O-series entirely consisting of 1s. Only the all 0s series gives the minimum value 0.

It is not hard to show that for the O-series $1^{a_1}0^{b_1}1^{a_2}0^{b_2}\ldots 0^{b_{n-1}}1^{a_n}$ (a series of a_1 1s, b_1 0s, a_2 1s, b_2 0s, …, b_{n-1} 0s, a_n 1s) the consecutive support equals

$$\sum_{i=1}^{n} a_i + \rho \sum_{i=1}^{n} a_i(a_i-1)/2 + \rho \sum_{1\leq i<j\leq n} a_i a_j \sigma^{b_i+b_{i+1}+\cdots+b_{j-1}} =$$
$$(1-\rho/2)S + \rho S^2/2 - \rho \sum_{1\leq i<j\leq n} a_i a_j (1-\sigma^{b_i+b_{i+1}+\cdots+b_{j-1}}),$$

where $S = \sum_{i=1}^{n} a_i$ (i.e., the traditional support); here 0^0 must be interpreted as 1 (an exponent 0 can be avoided by demanding all b_i's to be non-zero; if we also demand all a_i's to be > 0, both the number n and the numbers a_i and b_i are unique, given an O-series). The formula follows from the fact that if *reward* equals ε, then the series 1^k0^ℓ changes this into $(\varepsilon + k\rho)\cdot\sigma^\ell$, meanwhile giving a contribution of $k + k\varepsilon + k(k-1)\rho/2$ to the consecutive support. An extra series 0^ℓ at the beginning or end has no influence on the consecutive support.

The second part of the equation, $\rho \sum_{i=1}^{n} a_i(a_i-1)/2$, consists of the ρs added for a subset of consecutive 1s in the O-series. The last part of the equation is the addition of the rewards from the previous consecutive 1s decreased by σ, because of the number of 0s between the groups of consecutive 1s. Also note that when we choose $\rho = 2$ we get $S^2 - \rho \sum_{1\leq i<j\leq n} a_i a_j (1-\sigma^{b_i+b_{i+1}+\cdots+b_{j-1}})$. This shows that consecutive support is at most S^2 if $\rho = 2$.

Example 2. Take $\rho = 2$. Then the O-series $1^5 0^\ell 1^4$ has consecutive support $81 - 40(1-\sigma^\ell)$. Note that this is the same for the reverse $1^4 0^\ell 1^5$. As $\ell \to \infty$ this value approaches $41 = 5^2 + 4^2$, whereas for small ℓ and $\sigma \approx 1$ it is near $81 = (5+4)^2$.

It can be observed that the consecutive support as defined above only depends on the lengths of the "runs" and the lengths of the intermediate "non-runs": the a_i's and b_i's above. Here a *run* is defined as a maximal consecutive series of 1s in

a 0/1 sequence. Indeed, the sum $\sum_{k=i}^{j-1} b_k$ equals the number of 0s between run i and run j. This also implies that the definition is *symmetric*, in the sense that the support is unchanged if the order of the O-series is reversed — a property that is certainly required.

The reason why we add ρ and multiply by σ instead of, for example, add ρ and subtract σ, lies in the observation that in the latter case the symmetry property would not hold. Subtracting σ leads to different consecutive support values for an O-series and its reverse. E.g., if $\rho = 2$ and $\sigma = 0.5$, $1^5 0^3$ would give 25, whereas $0^3 1^5$ has 17.5 (the definition from Section 2.1 gives 25 in both cases). One should also take care that the support remains positive in that case.

Instead of this way of calculating consecutive support it is also possible to augment the O-series with *time stamps*. Then one is able to use the real time between two transactions in calculating the consecutive support. In the previous definition each transaction was assumed to take the same amount of time. Another improvement might be to reinitialize *reward* to 0 at suitable moments, for instance at chromosome boundaries or at "closing hours".

We consider algorithms that find all frequent itemsets, given a database. A *frequent* itemset is an itemset with support at least equal to some pre-given threshold, the so-called *minsup*. Thanks to the APRIORI property many efficient algorithms exist. However, the really fast ones rely upon the concept of FP-TREE or something similar, which does not keep track of consecutivity. This makes these algorithms hard to adapt for consecutive support.

One fast algorithm that does not make use of FP-TREEs is called ECLAT [10]. ECLAT grows patterns recursively while remembering which transactions contained the pattern, making it very suitable for consecutive support. In the next recursive step only these transactions are considered when counting the occurrence of a pattern. All counting is done by using a matrix and patterns are extended with new items using the order in the matrix. It is straightforward to adapt ECLAT to incorporate consecutiveness, the counting of traditional support is simply replaced by the $\mathrm{Supp}(I, \mathcal{D}, \rho, \sigma)$ function as proposed earlier. The overhead of extra calculations is minimal and the runtime complexity is expected to be equal to that of ECLAT as described in [10].

3 Hyperclique Patterns and h-confidence

Many pruning principles used for traditional support calculation can still be applied for consecutive support. We consider one method in particular. In the case of our major example, the database of clones, we wanted to visualize patterns with a certain minimal consecutive support. Unfortunately there are many patterns with this support. In order to speed up the search and to filter out uninteresting patterns we can search for *hyperclique patterns* as described in [9]. Because of space limitations we explain hyperclique patterns via an example:

Example 3. First a *minimal confidence threshold* h_c is defined, say $h_c = 0.6$. We want to know if $\{A, B, C\}$ is a hyperclique pattern. We calculate the confidence of $A \to \{B, C\}$, $B \to \{A, C\}$ and $C \to \{A, B\}$. The lowest of these

confidences is the *h-confidence*, which must be higher then h_c. Assume that $conf(A \rightarrow B, C) = \text{Supp}(\{A, B, C\}, \mathcal{D}, \rho, \sigma)$ / $\text{Supp}(\{A\}, \mathcal{D}, \rho, \sigma) = 0.58$. Then $\{A, B, C\}$ is no hyperclique pattern.

When we combine the concept of consecutive support with hyperclique patterns we get patterns that occur frequent but in the flow of transactions close after each other and there is a *strong affinity* between items: the presence of $x \in P$, where P is an itemset or clone, in a transaction strongly implies the presence of the other items or patients in P.

It is clear that hyperclique patterns possess the *cross-support property*. This means that we will not get *cross-support patterns*. These are patterns containing items of substantially different support levels. If one item has a high support and another item has a low support, then h-confidence will be low if the denominator is the item with the high support.

Example 4. Say A is an item with a consecutive support of 200 and B has a consecutive support of 50. The support of $\{A, B, C\}$ will at most be 50 because of the APRIORI property. So $conf(A \rightarrow B, C)$ can at most be $50/200 = 0.25$. As a consequence the h-confidence of $\{A, B, C\}$ will also be at most 0.25. So if $h_c = 0.6$, then $\{A, B, C\}$ and all the patterns that are grown from it can be pruned.

The combination of hyperclique patterns and consecutive support allows us to find patterns that occur in clones (transactions) that follow each other close, yet minimal support can be relatively low. This property is especially handy for our motivating example, because a low minimal consecutive support will generate many cross-support patterns, which are pruned if we search only for hyperclique patterns. Hyperclique patterns also posses the anti-monotonicity property, because as patterns grow the numerator of the confidence calculation stays the same or declines. The denominator stays fixed and so h-confidence will decrease or stay the same:

Example 5. Say $conf(A \rightarrow B, C) = 0.58$. The superset $\{A, B, C, D\}$ will at most have the same consecutive support as $\{A, B, C\}$. Also the denominator $\text{Supp}(\{A\}, \mathcal{D}, \rho, \sigma)$ stays the same, so the h-confidence of $\{A, B, C, D\}$ can at most be 0.58.

4 Results and Performance

The experiments were done for three main reasons. First of all we want to show that consecutive support can enable one to find new patterns that one does not find with the traditional support. Secondly we want to show how using the principle of h-confidence one can filter the data. Finally we want to give an indication how the reward factor ρ and punishment factor σ should be chosen.

All experiments were done on a Pentium 4 2.8 GHz with 512MB RAM. For our experiments we used five datasets. One biological dataset, referred to as the

Nakao dataset, was also used in [4]. This data set originates from Nakao et al. who used the dataset in [6]. This publicly available dataset contains normalized $\log_2$-ratios for 2124 clones, located on chromosomes 1–22 and the X-chromosome. Each clone is a transaction with 2 to 1020 real numbers corresponding to patients. We can look at gains and/or losses. If we consider gains, a patient is present in a transaction (clone) if his value is at least 0.225 higher than that of a healthy person (for losses at least 0.225 lower). The work in this paper reported losses and gains in chromosomes 1, 8, 17, 18 and 20. Two datasets are synthetic databases, but structured like the dataset of clones. One of these datasets, the *noisy dataset*, contains more noise then the other, the *ideal dataset*. The precise structure of these datasets is described in [4]. The remaining datasets are synthetic datasets made to show how consecutive support can be used to find patterns that could not be found before. The third synthetic data set, referred to as the *food+drink dataset*, describes a cafe-restaurant where in the middle of a day a lot of people buy bread and orange juice; it has 1000 transactions (customers) and 100 items (products). The fourth synthetic data set will be explained later.

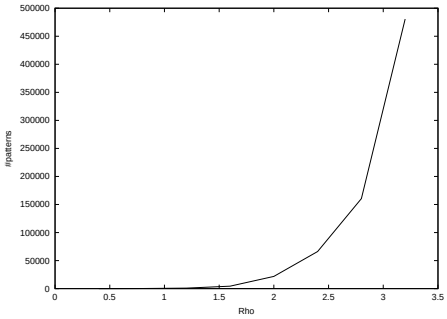


Fig. 1. Number of patterns from the Nakao dataset as ρ increases (gains, $minsup = 625$, $\sigma = 0.5$)

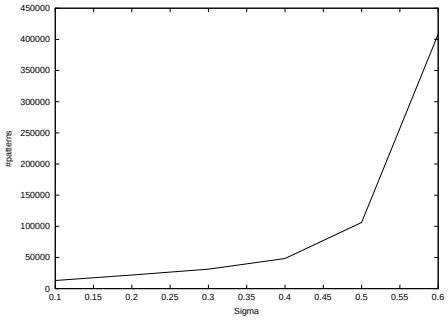


Fig. 2. Number of patterns from the Nakao dataset as σ increases (gains, $minsup = 625$, $\rho = 2.0$)

4.1 Consecutive Support

Figure 1 and Figure 2 show how the number of patterns increases with ρ and σ. Each setting therefore requires another *minsup*. In some cases it is best to select the *minsup* such that one gets a fixed number of patterns, e.g., 1000, in order to compare the results.

In the experiments of Figure 3–6 we tried to find approximately 1000 patterns with the highest traditional or consecutive support. After this we count for each transaction how many patterns it contains, allowing us to see how active areas are. For the Nakao dataset more active means that many clones (gains) in the same area are present in many groups of patients.

Figure 3 and Figure 5 show where patterns occur when we use traditional support, giving results similar to those in [4]. For each transaction the number

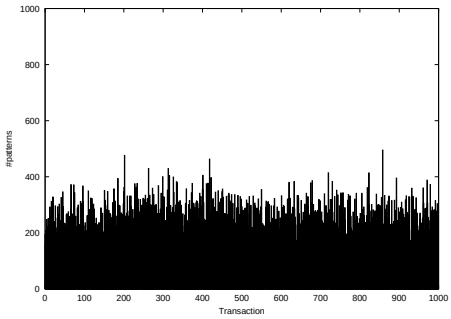


Fig. 3. Occurrence graph of food+drink using traditional support ($minsup = 257$)

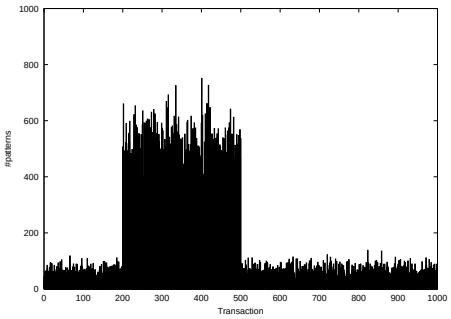


Fig. 4. Occurrence graph of food+drink using consecutive support ($minsup = 467$, $\rho = 1.0$ and $\sigma = 0.5$)

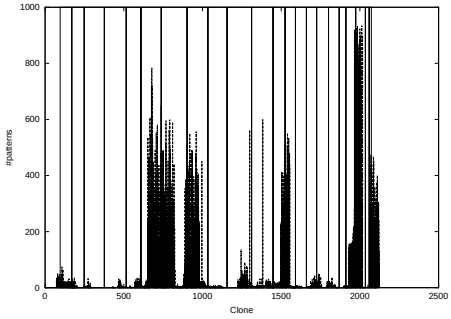


Fig. 5. Occurrence graph of Nakao using traditional support (gains, $minsup = 129$)

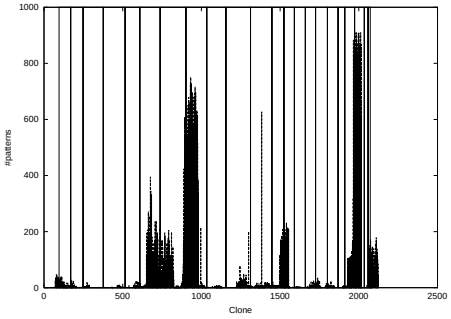


Fig. 6. Occurrence graph of Nakao using consecutive support (gains, $minsup = 827$, $\rho = 1.0$ and $\sigma = 0.5$)

of patterns that it occurs in is plotted in a so-called *occurrence graph*. In each of these graphs we will indicate chromosome borders when the Nakao dataset is visualized. In the food+drink dataset it is very clear that consecutive support enables us to see new patterns. Figure 4 shows that in certain areas patterns are more consecutive. Figure 6 shows that certain areas are less active if we use consecutive support instead of traditional support (chromosomes 7 and 8) and some areas contain more patterns (chromosome 9), hence providing patterns that occur together in one part of the chromosome instead of far apart. This shows additional activity compared to results reported by Nakao et al. in [6].

In order to evaluate the effect of more or less noise on consecutive support we used the ideal and noisy dataset. These datasets are generated with properties similar to the Nakao dataset with real patient information (see [4] for details). The results for the ideal dataset are plotted in Figure 7 and 8.

Figure 7 shows that some interesting areas are less clear when using traditional support. However they become more apparent when we apply consecutive support. The results for the noisy dataset are displayed in Figure 9 and 10. Because of the noise the middle peak becomes less clear. However overall the results seem hardly to be affected by noise.

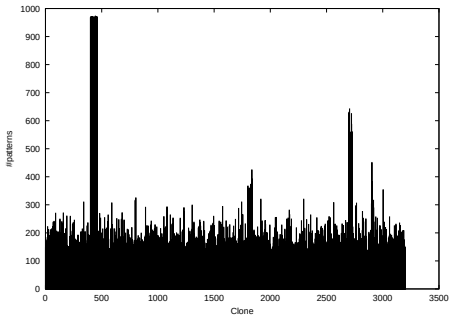


Fig. 7. Occurrence graph of the ideal dataset using traditional support (gains, $minsup = 479$)

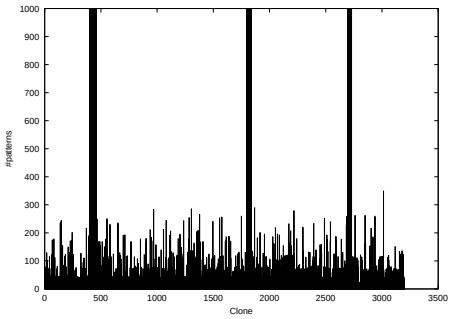


Fig. 8. Occurrence graph of the ideal dataset using consecutive support (gains, $minsup = 6180$, $\rho = 2.0$ and $\sigma = 0.7$)

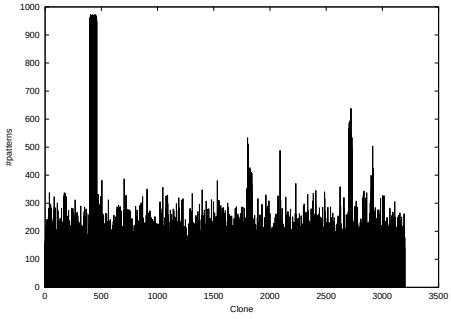


Fig. 9. Occurrence graph of the noisy dataset using traditional support (gains, $minsup = 617$)

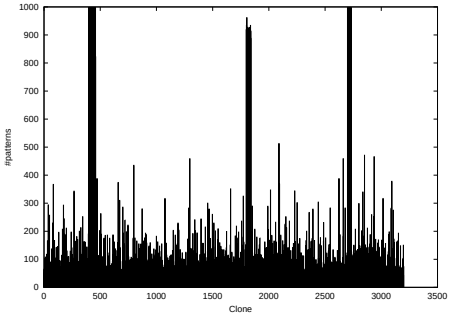


Fig. 10. Occurrence graph of the noisy dataset using consecutive support (gains, $minsup = 6039$, $\rho = 2.0$, $\sigma = 0.7$)

4.2 Selection of ρ and σ

The goal of the following experiments was to give some guidance in the selection of reward factor ρ and punishment factor σ. The right parameters should result in many patterns of which the O-series has large groups of consecutive 1s.

Figure 11 plots the average number of consecutive groups of 1s and 0s for all patterns. The plot gives an indication of consecutiveness of patterns found using different settings of ρ and σ (less groups indicate more consecutiveness). The plot seems to stabilize around $\rho = 2$. Figure 12 and 13 show that only if we choose σ very close to 1.0 we get results more like those for traditional support. However, Figure 13 still shows some influence of ρ. For the Nakao dataset it seems that if $\rho \approx 2$, then the influence of σ is minimalized as long as σ is not too close to 1.0. Also similar experiments showed significant changes in the occurrence graph only if ρ was chosen very small. Different datasets might require different settings depending on how much one wants to amplify consecutiveness. However results in this section indicate that $\rho = 2.0$ and $0.2 \leq \sigma \leq 0.8$ seem to be good choices. However, a lot of experimental work is necessary to settle this issue.

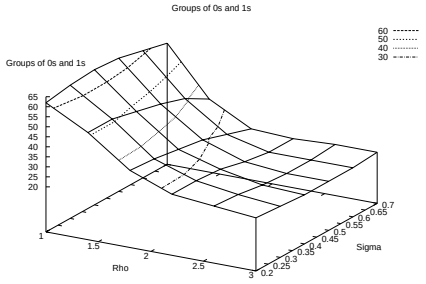


Fig. 11. Effect of ρ and σ on the O-series for the Nakao dataset (gains, $minsup = 625 \cdot (\rho/2)$, chosen to guarantee a reasonable amount of patterns)

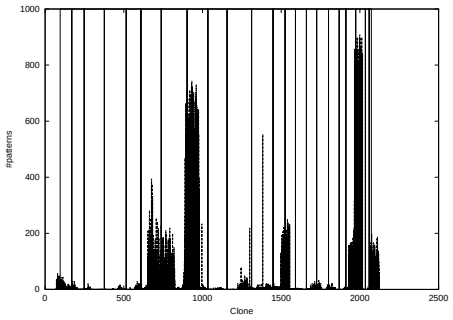


Fig. 12. Occurrence graph of Nakao using consecutive support (gains, $minsup = 2498$, $\rho = 2.0$ and $\sigma = 0.8$)

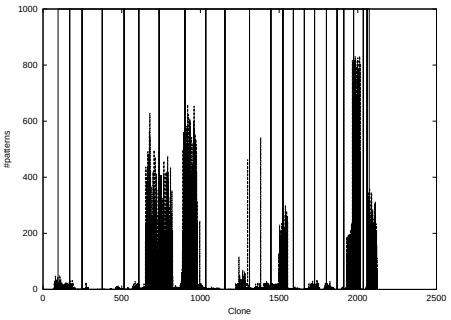


Fig. 13. Occurrence graph of Nakao using consecutive support (gains, $minsup = 6157$, $\rho = 2.0$ and $\sigma = 0.99$)

4.3 Combination with h-confidence

In the following experiments the goal was to show that combining hyperclique patterns with consecutive support enables us to see patterns occurring in bursts. In order to show this we created a new synthetic dataset, referred to as the *coffee+cookie dataset*, where in the cafe-restaurant small bursts of people buy coffee and a cookie, during the day in the coffee breaks.

Figure 14 does not show the small groups buying the same products: just hyperclique patterns do not reveal the bursts. Figure 15 shows that with only consecutive support we are also unable to discover these patterns. Figure 16 shows people buying the products in bursts. Consecutive support stresses patterns that are consecutive and the principle of h-confidence filters out the noise caused by cross-support patterns.

When we apply these techniques to the Nakao dataset (losses), in Figure 18, we can see, e.g., on chromosomes 14 and 15 (near transaction 1600) that certain areas become more active compared to not using h-confidence in Figure 17.

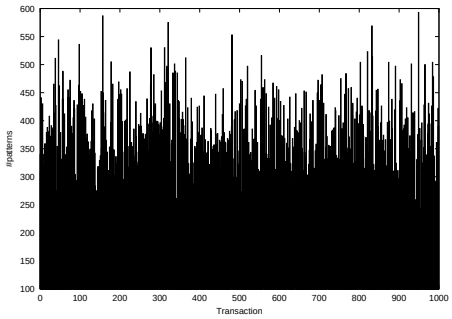


Fig. 14. Occurrence graph of coffee+cookie using only h-confidence ($minsup = 64$, $h_c = 0.5$)

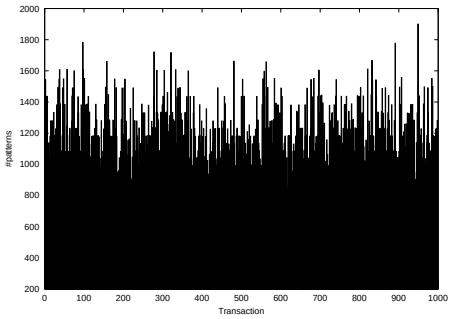


Fig. 15. Occurrence graph of coffee+cookie using only consecutive support ($minsup = 225$, $\rho = 1.0$, $\sigma = 0.5$, $h_c = 0$)

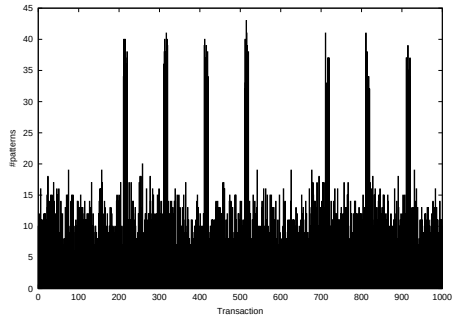


Fig. 16. Occurrence graph of coffee+cookie using both consecutive support and h-confidence ($minsup = 64$, $\rho = 1.0$, $\sigma = 0.5$, $h_c = 0.31$)

5 Conclusions and Future Work

Consecutive support enables us to find new and useful patterns in the chromosomes compared to methods using only traditional support. Principles applicable to traditional support can still be used with consecutive support. For instance the combination of consecutive support and the h-confidence threshold enables us to find small bursts of patterns. In this case h-confidence filters out noise and consecutive support amplifies the bursts.

Using the distance between transactions like it is done in this paper is an interesting area of research. In the future we want to examine if consecutive support enables us to visualize even more types of pattern occurrence, perhaps even detecting them automatically. Also we want to see if we can speed up the search for consecutive patterns. Finally we want to extend consecutive support by using distance between transactions in different ways, which might give us even more biological relevant patterns.

Acknowledgments We would like to thank Joost Broekens, Joost Kok, Siegfried Nijssen and Wim Pijls.

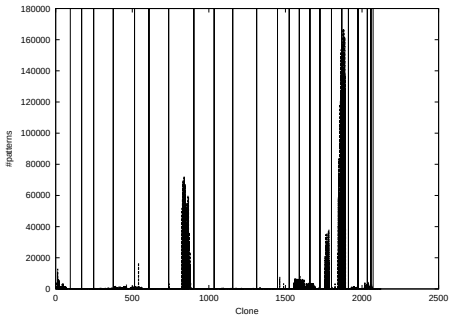


Fig. 17. Occurrence graph of Nakao: consecutive support (losses, $minsup = 400$, $\rho = 1.0$, $\sigma = 0.9$, $h_c = 0$)

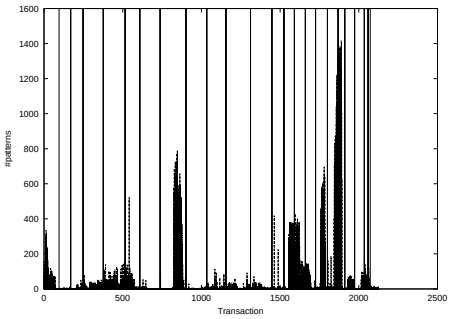


Fig. 18. Occurrence graph of Nakao: consecutive support and h-confidence (losses, $minsup = 25$, $\rho = 1.0$, $\sigma = 0.9$, $h_c = 0.15$)

References

1. Agrawal, R., Imielinski, T., Srikant, R.: *Mining Association Rules between Sets of Items in Large Databases*. In Proc. of ACM SIGMOD Conference on Management of Data (1993), pp. 207–216.
2. Antunes, C., Oliveira, A.L.: *Generalization of Pattern-Growth Methods for Sequential Pattern Mining with Gap Constraints*. In Machine Learning and Data Mining in Pattern Recognition (MLDM 2003), LNCS 2734, Springer, pp. 239–251.
3. Graaf, E.H. de, Kosters, W.A.: *Using a Probable Time Window for Efficient Pattern Mining in a Receptor Database*. In Proc. of 3rd Int. ECML/PKDD Workshop on Mining Graphs, Trees and Sequences (MGTS'05), pp. 13–24.
4. Graaf, J.M. de, Menezes, R.X. de, Boer, J.M., Kosters, W.A.: *Frequent Itemsets for Genomic Profiling*. In Proc. 1st International Symposium on Computational Life Sciences (CompLife 2005), LNCS 3695, Springer, pp. 104–116.
5. Leleu, M., Rigotti, C., Boulicaut, J.F., Euvrard, G.: *Constraint-Based Mining of Sequential Patterns over Datasets with Consecutive Repetitions*. In Proc. 7th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2003), LNAI 2838, Springer, pp. 303–314.
6. Nakao, K., Mehta, K.R., Fridlyand, J., Moore, D.H., Jain, A.N., Lafuente, A., Wiencke, J.W., Terdiman, J.P., Waldman, F.M.: *High-Resolution Analysis of DNA Copy Number Alterations in Colorectal Cancer by Array-Based Comparative Genomic Hybridization*. Carcinogenesis 25 (2004), pp. 1345–1357.
7. Rouveirol, C., Stransky, N., Hupé, Ph., La Rosa, Ph., Viara, E., Barillot, E., Radvanyi, F.: *Computation of Recurrent Minimal Genomic Alterations from Array-CGH Data*. Bioinformatics 22 (2006), pp. 849–856.
8. Steinbach, M., Tan, P., Xiong, H., Kumar, V.: *Generalizing the Notion of Support*. In Proc. 10th Int. Conf. on Knowledge Discovery and Data Mining (KDD '04), pp. 689–694.
9. Xiong, H., Tan, P., Kumar, V.: *Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution*. In Proc. Int. Conf. on Data Mining (ICDM'03), pp. 387–394.
10. Zaki, M., Parthasarathy, S., Ogihara, M., Li, W.: *New Algorithms for Fast Discovery of Association Rules*. In Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD '97), pp. 283–296.

Identifying heterogeneous and complex named entities in biology texts using controlled dictionaries

Julien Lorec[1,2], Gérard Ramstein[2], and Yannick Jacques[1]

[1] Équipe 3: cytokines et récepteurs, Département de Cancérologie, Inserm U601, Institut de Biologie/CHR, 9 quai Moncousu, 44093 Cedex 1, Nantes, France
{julien.lorec,yjacques}@nantes.inserm.fr
[2] Équipe C.O.D, LINA, École polytechnique de l'université de Nantes, La Chantrerie, rue Christian Pauc, 44306, Nantes, France
gerard.ramstein@polytech.univ-nantes.fr

Abstract. This work presents a method for biomedical named entity (NE) recognition and identification. NEs may be of different classes and we try not to limit ourselves to gene or protein name recognition only. The method combines rule-based identification of noun phrases as candidate NEs with matching against manually cleaned dictionaries from public sources. The paper discusses some techniques to overcome or restrict the problems of synonymies, term variability and ambiguous names and focuses on name normalization as well as context-based disambiguation. We first describe the construction and composition of specific dictionaries we use to identify the textual representations of various biological objects. Then we detail a generic methodology to extract potential NEs from text. Finally we comment on the disambiguation techniques used to help classifying the true nature of an identified NE. The NE extraction performance is evaluated by comparing with BioNLP/NLPBA 2004 contenders' results. The NE identification achievement is measured using an enriched subset of the same benchmark corpus.

1 Introduction

NE extraction and identification precede the discovery and the organisation of relations between entities gathered from biomedical publications in a well defined, machine-readable form. A biomedical NE is here defined as a textual representation of a biomedical-related object of interest. To this date, several methods for biological NE tagging have been proposed. Some depend on linguistic rule constructions [1] while others are based on machine learning techniques [2]. Nevertheless, the simple detection of NEs cannot properly answer their identification with or association to specific biological objects. Coupling NE extraction methods to dictionary resources has proven to be an efficient solution to this problem [3]. Extracting and identifying NEs require to overcome three main difficulties. First, synonymy, abbreviation and acronym resolution. Second, term variability at both the orthographic, morphologic, syntactic levels, and the lexico-semantic,

insertions/deletions and permutations levels. And third, homonymy between entities from similar or different types, and with common English words. Those difficulties are not specific to the domain but are worse in biology. Term ambiguities in biology texts are well described in [4].
The vast majority of NE recognition systems do not attempt to recognize more than a few classes of biological objects. Proteins and genes detection has been well studied over the past few years [5] whereas other biological descriptors have only been recently under scrutiny. This extension is a main issue that is yet to be addressed.

We have implemented a methodology for the extraction and the identification of complex NEs dedicated to biomedical corpora. It uses dictionaries as well as hand-made rules to identify biomedical objects, of different origins, in human biology texts.

2 Methods

2.1 Construction of dictionaries

Resources Several dictionaries gathering assorted classes of biological objects are at our disposal: human transcription factors binding sites, cell lines, tissues and organs, experimental protocols and techniques, human genes and proteins; from these last two, we isolate a final class: human transcription factors. The definition of these seven classes should be sufficient to cover a preparatory study of a whole biological process, such as human gene transcription. We selected a short but complete and relevant number of public databases in order to construct such dictionaries, respectively: TRRDSITE[3] for transcription factors binding sites, various sources[4] of the MetaThesaurus UMLS[6] for cell lines, tissues and organs and experimental protocols and techniques, LocusLink[5], HUGO[6], GDB[7] and OMIM[8] for genes and proteins, and TFD[9], COMPEL[10], TRRDFACTORS and TFFACTOR[11] for transcription factors.
The databases introduce one or more complete names and one or more symbols, acronyms and abbreviations for the same biological entity in each entry. Those namings are either official, frequent or casual. They may also be in use by publication authors nowadays or they may have been abrogated and only be part of old papers. We are interested in entity names related to human only.

[3] http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd
[4] CRISP Thesaurus 2003, Gene Ontology 2002_12_16, Medical Subject Headings 2004_2003_08_08, NCBI Taxonomy 2003, NCI Thesaurus 2001a, Standard Product Nomenclature 2002 and UMDNS: product category thesaurus 2003
[5] http://www.ncbi.nlm.nih.gov/LocusLink
[6] http://www.gene.ucl.ac.uk/nomenclature
[7] http://www.gdb.org
[8] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
[9] gopher://gopher.nih.gov/77/gopherlib/indices/tfd/index
[10] http://compel.bionet.nsc.ru/new/index.html
[11] http://www.gene-regulation.com/pub/databases.html#transfac

Generation of variant forms Two types of variant forms can be encountered: the biologically valid variant forms (spelling, morphologic, lexico-semantic, ...) established from *a priori* information and the ones (orthographic, morphologic and specific syntactic variations) that can be derived from plain English. For the first type, data retrieved from each database specific nomenclature is used to build up fresh aliases for an entity. These new forms may only be retrieved in scientific publications because they are absent in databases. For instance, a significant effort has been produced to generate: the mixed acronyms/full names combinations for the same entity ("chemokin like receptor 1", "CMKLR1", "CMKL receptor 1", "CMK like receptor 1", "chemokin like R 1" and "chemokin L R 1"). Several lexicons of character sequences with corresponding definitions reserved for specific use in biology are used to develop symbols and condense full names (e.g. "FAM" stands for "Family", "RE" symbolizes either "Responsive element", "Response element", "Regulatory element" or "Repressing element", etc). The set of combinations are generated if a character sequence (e.g. "L") from a lexicon and its correspondant definition ("like") are found in a symbol ("CMKLR1") and its associated full name ("chemokin like receptor 1"), respectively. We also take care of shifts of descriptive terms chunks ("class III alcohol dehydrogenase", "alcohol dehydrogenase class III" and "soluble aconitase 1", "aconitase 1 soluble") that may happen in texts. We produce such alternate forms using assessed data from experts only. This *a priori* information is gathered from HUGO naming recommendations that genes (and by extension proteins) databases authors must obey. The other databases we use either do not allow or highlight the presence of descriptive terms or do not enforce any specific appellation guideline. As a result, we only interpret as-is HUGO nomenclature for our genes and proteins databases.

Normalization For the second type, the variant form we elect must, on its own, be able to represent each and every writing convention variant for a given NE that may be encountered in publications. Several rules are used sequentially to keep only one of these derivative variants for an entity. We keep NEs as basic compound nouns and stems, we also resolve uppercase usage and punctuation discrepancies. First, NEs are only kept as basic compound nouns in our dictionaries and transformed if needed. For example, from the next three expressions symbolyzing the same entity: "linker for activation of T cells", "activation of T cells linker" and "T cells activation linker", only the last one is stored in our dictionaries. Uppercase usage and punctuation discrepancies are commonplace difficulties that are resolved simultaneously. Every word of an entity name is broken down into smaller pieces, each piece must either contains lowercase, uppercase characters or numeric and special characters only. These pieces are then separated from each other by a space character if needed. As a noted exception to this rule, a single uppercase located at the beginning of the original expression, both preceded by a non-alphanumeric character and followed by a lowercase, is not separated out from the next block. Then, the whole new expression is converted to lowercase. Finally, non-alphanumeric characters are deleted. For

example, "cAMP", "c-Amp" and "c Amp" which symbolize the same entity are transformed and stored as "c amp" in our dictionary. "CAMP", a different entity, is in turn present in our dictionary as "camp". A few exceptions remain and must be considered separately: first of all, common abreviations for Dalton measure unit and its multiples (i.e "kD", "kDa" or "kd"). Then, "-" and "." in combination with numbers. And finally, basic genetic and chemistry domain dependant usage of special characters (e.g. "-/-", "H+"). Roman numerals are replaced by their numeric counterparts and Greek symbols are converted in full text. As always, there are still some exceptions that we need to handle. For example, the "X" character, when used in combination with the words "chromosom" or "ray" is not transformed. Finally, each variant which is not an acronym is stemmed using a modified version of Porter's algorithm. Passive and active forms of verbs are kept as-is. For example, "cAMP regulated protein" shall remain different from "cAMP regulating protein".

Dictionaries Content Most selected databases include a lot of nomenclature mistakes and inappropriate namings. Along with automated cleanup procedures, manual curation is required for guaranteeing maximal reliability. We estimate that about 4% of the variants were erroneous or needed editing before manual validation. The dictionaries for genes and proteins, transcription factors binding sites, transcription factors, cell lines, tissues and organs, and experimental protocols and techniques contain respectively 183148, 6517, 11411, 508, 1768, 1284 variants and 43271, 2266, 1763, 312, 1202, 769 unique entities.

2.2 Recognition of NEs

Each document is first broken down into sentences using rules adapted from [7] to the biomedical domain. To each word in a sentence we associate its part of speech thanks to GENIA POS Tagger [8]. We use no shallow parser.

Extraction Our approach goes beyond direct matching of dictionary entries with raw text. We use simple grammatical rules to consider syntactically valid noun compounds in the molecular biology domain only. The shape of those noun compounds are in line with the most elaborate NEs one can come across in such specialized publications.
We first retrieve every noun-based syntagms from sentences. Sequentially, such syntagms are reconstructed using top to bottom rules described below. The goal is to gather the largest expression around a noun group that may represent the most complex NE.

1. Simple noun groups with possible symbols, numerals and adjectives directly connected to are extracted. For example, "DNA array", "Interleukin 2", "β adrenergic receptor".
2. Gerunds and past participles located at the end of one of these noun groups, or at the beginning if the preceding word is not a modal, a pronoun or an

adverb, are part of the same noun syntagm. Two blocs are then brought together if one of the verbs described above are in between. For example, "Interferon regulating factor 8".

3. Two of the syntagms at this level of complexity are aggregated if prepositions or conjunctions as of "of", "in", "at", "on", "by", "for", "to" or "with" separate them. For example, "regulator of G-protein signalling 4" or "cell adhesion molecule regulated by oncogenes".
4. Two syntagms are finally linked if separated by a coordinating conjunction "and", "but", "or". For example, "Signal transducer and activator of transcription 3 interacting protein 1".

We *a priori* consider that each syntagm retrieved at this stage represents zero or one NE. We seek out its mention in our dictionaries after normalizing it (following the same steps as described in the section Normalization) and proceeding to further NE block boundaries corrections if needed. Text portions we get through the various NE extraction phases may be connected with *satellite* nouns on their right-hand side. Such nouns either describe an action whose object is the NE (e.g. "assimilation", "transcription", "screening") or characterise it (e.g. "gene", "protein", "experiment"). Nouns located at the end of the expression are sequentially removed until a match could be found in our dictionaries. Adjective, numeral or symbol presence on the left-hand side are also taken into consideration. Nevertheless, some rare and difficult constructions such as "interleukin protein 2", which is represented in our dictionaries as "interleukin 2", can not be properly processed at the moment.
In case of successful look up we might have identified a true NE but we may still have to carry on a disambiguation process (see next section Disambiguation) before going any further. If the NE was not found in our dictionaries, we now allow the syntagm to either contain zero, one or more than one NE. To verify this new assumption, the current expression is broken down in reverse order of construction. Each resulting piece of the split syntagm is then evaluated independently against the dictionaries content, again. If several separators of the same kind are part of a syntagm, we generate the different combinations of text blocks on both of one side and on the other. The more elaborated syntagms are then tested first. In case of equal complexity, the ones located to the end of the original syntagm take precedence over the others. NEs are indeed preferentially found on the right side of a separator.
For example, using the syntagm "modulator of G-protein signalling 4 down-regulated by oncogenes", which contains the real NE "G-protein signalling 4", we check against our dictionaries content the text blocks in the following order:

- first "modulator of G-protein signalling 4 down-regulated by oncogenes" then
  - on one hand "G-protein signalling 4 down-regulated by oncogenes" then
    - "G-protein signalling 4" and "oncogenes" aside
  - and on the other "modulator".

Disambiguation We distinguish here two kinds of ambiguities: the ones closely related to NE discovery in text, which are resolved before or during NE extraction, and those connected with alternate NE identity questions, handled once NEs have been identified.
Authors of scientific articles usually redefine their own original aliases for long or complex NEs as soon as possible in their document. Some of them are unique to the article and thus absent from our dictionaries. Noun-based syntagms in between brackets and preceding known NEs are automatically associated as an alias of the latter. With a view to also enhance detection of implicit NE references in basic enumerations, we reformulate them if numerals or symbols/identifiers are involved. For example, "interleukin 1, 2 or 3 receptors" expression is transcribed as "interleukin 1 receptor or interleukin 2 receptor or interleukin 3 receptor". Coreference and anaphora resolution is not investigated here.
In addition, we need to evaluate the true nature of an identified NE for several reasons:

- First, an entity may be associated with different classes in our dictionaries (NE class level disambiguation). Three distinct techniques are used to classify the NE category, namely local contextual word environment analysis, connected verb categorization and "experimental trace" finding. Local contextual information surrounding an entity is determined by analyzing the word content alongside a NE. For this task, we created a lexicon of mono/bi/tri-words specifically associated to one or more classes. For the moment, the lexicon only includes nouns and noun clauses. The presence of these terms, when connected with an occurrence of a NE whose nature is to be checked, is used to corroborate or to undermine a class initially assigned to a NE by the dictionaries (e.g. "neuropeptide" solely characterises a protein, "transcription" a gene and "assay" an experimental protocol in our study). We look for such words into the most complex syntagm, transformed into a compound noun, incorporating the NE. The class, or collection of classes, found associated with the right-most lexicon term in such syntagm, is used as the contextual desambiguating class, or classes. Along with this method, we try to get help from the analysis of the verbs whose subjects and/or objects are NEs to disambiguate. We possess a pre-defined list of verbs, categorized by classes. A specific class of verb is used to corroborate or undermine the connected NEs, in a manner similar to the method seen above. Connected NEs are found using simple pattern matching rules. NEs that have been first disambiguated as experimental protocols or techniques, are also used to clarify the class of the other NEs found in the same sentence. Experimental protocols and techniques only involve one or few classes of other NEs in texts (e.g. Polymerase Chain Reaction is an assay specific to genes or small DNA portions and do not manipulate proteins). As such, we associated one or more classes to each experiment entry in our dictionary, using textual descriptions found in the Metathesaurus UMLS when available. Simple pattern matching techniques are still used to determine which NEs are connected to an experimental evidence in texts. NE class evaluation

process is identical to the methods seen above.
If only one class is common between both context and dictionary pre-associated classes for an entity, then the entity is desambiguated as being of this same class. If more than one class are common or if the context is non-existent, then the entity is left with ambiguous class definition. As a noted exception to this rule, if an ambiguity remains between the gene and protein classes or between the transcription factor and transcription factors binding site classes, then the NE now belongs to the protein class. Empirically, we observed that authors most often refer to proteins when no contextual information is given. If no class is common between context and dictionary, then we tag the entity as false positive.

- Then we have to deal with terms from the English dictionary that are incorrectly assimilated to biological objects of interest (e.g. is "Aim" the name for "absent in melanoma" gene or a synonym for goal?). For this purpose, each NE is considered as a potential source of ambiguity with standard English if each of its words is part of a derivative of the Webster 1913 dictionary [12]. This dictionary is composed of 91840 non-biologically related English words. Lower and upper case differentiation is herein retained during evaluation. An entity, previously tagged with an ambiguous class definition, and not solely composed of common English words, is desambiguated as being of the class or classes associated with the latest same disambiguated variant form encountered in the current article. If none has already been encountered, and if the entity is only associated with one class in our dictionaries, then we disambiguate it as being of the dictionary related class. In any other cases, we consider this entity as a false positive.
- Finally, two entities from our dictionaries may share the same name and be associated to the same class (individual NE level disambiguation) (e.g. "CARP" may either designate the "carbonic anhydrase VIII" or the "ankyrin repeat domain 1" gene). We do not try to handle this kind of ambiguity in the study because it is a relative rarity in human nomenclatures.

3 Results and Discussion

We first benchmarked our NE extraction system against the BioNLP/NLPBA 2004 protocol[13]. Only real NE extraction and assignation of correct classes were measured herein. NE mapping to identifiers in databases is not evaluated in this first part. The data used in the task come from the GENIA version 3.02 corpus [9]. This is formed from a controlled search on MEDLINE using the MeSH terms 'human', 'blood cells' and 'transcription factors'. It is composed of 404 abstracts, hand annotated according to a small taxonomy of 5 classes ('protein', 'DNA', 'RNA', 'cell line' and 'cell type'). The publication year of the selected publications ranges over 1978-2001.
While studying the behaviour of our system over the training data provided,

[12] http://humanities.uchicago.edu/orgs/ARTFL/forms_unrest/webster.form.html
[13] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/ERtask/report.html

we meet a few limitations inherent to our original methodology. We were only interested in detecting complete biological objects and not parts or generic occurences of NEs (e.g. "hormone binding domain" has been tagged as 'protein' in the test set whereas we considered it too be too unspecific to be accepted as such). As a result, we have adapted our method to the task. Unfortunately, compromises had to be made on class correspondences between GENIA and our dictionaries. GENIA's generic 'protein' and 'DNA' classes are mapped with our own narrowly-defined 'protein' + 'transcription factor' and 'gene' + 'transcription factor binding site' categories, respectively. The original 'cell line' and 'cell type' classes are also merged. We were not able to extend the number of classes registered in our ditionaries on time to do a comparison over the complete BioNLP/NLPBA 2004 protocol. Dictionary extension and curation is a time consuming task. At the same time and in order to to extend the coverage of our methods, we consider that syntagms containing specific terms from the local context word lexicon used to disambiguate NEs (see Disambiguation sub-section) are valid instances of a class. Such a simple technique greatly improve the F-Score, by almost 20%, globally, and allows fair compliance with the BioNLP/NLPBA 2004 protocol. Hence family names and implicit references of biological entities were transparently taken into account. Only the 'complete match' score provided by the official BioNLP/NLPBA evaluation tool is shown in table 1.

Table 1. NE extraction erformance test over BioNLP/NLPBA 2004 test set

Author/Method	Task	Recall	Precision	F-score
Zho04 [10]	protein	79.24%	69.01%	73.77%
	DNA	73.11%	66.84%	69.83%
	cell	79.98%	79.06%	79.52%
Our work	protein	53.03%	76.35%	62.58%
	DNA	32.84%	64.21%	43.45%
	cell	67.56%	81.02%	73.68%
Lee04 [11]	protein	62.13%	46.14%	52.96%
	DNA	28.88%	44.79%	35.12%
	cell	64.77%	46.55%	54.17%

In a second part of the study, we both measured the NE extraction and NE identification performances of our system. We randomly selected 100 abstracts from the 404 contained in the BioNLP/NLPBA GENIA corpus, using an additional MeSH term 'cytokine' during search. Those abstracts were re-annotated to fit our dictionary classes. No family name, group, substructure, subunit and complex for a biological object of interest were retained. Only valid members of our dictionary classes were marked up along with their correctly associated dictionary identifier. We consider as true positives the right entities associated with the right biological class, as the publication authors intended for both. If a NE is not extracted, or not disambiguated, or if the dictionaries lack its reference, then

the NE is tagged as a false positive. A NE wrongly disambiguated, or associated to an improper dictionary identifier, is, at the opposite, considered as both a false negative and a false positive. See table 2 for the results. In this study, entities correctly identified but belonging to mammals and not only to human are considered as true positives. As already said, no implicit NE reference resolution in texts (i.e. co-reference and anaphora resolution) is investigated here.

Table 2. NE extraction and identification performance test over a GENIA enriched subset

	G	P	F	S	C	O	E	TOTAL
Precision	0.78	0.74	0.72	0.65	0.93	0.87	0.85	0.77
Recall	0.75	0.71	0.68	0.63	0.71	0.87	0.71	0.71
Number of true positives (TP)	476	630	561	123	530	7	92	2419
Number of false positives (FP)	36	74	72	27	16	1	10	236
Number of false negatives (FN)	64	103	109	32	189	1	31	529
Number of both FN and FP	92	146	144	39	23	0	5	449

G = genes, P = proteins (minus transcription factors), F = transcription factor proteins, S = transcription factors binding sites, C = cell lines and types, O = tissues and organs, E = experimental protocols and techniques

Error Analysis While succinctly analyzing the main sources of errors in the BioNLP/NLPBA 2004 task, we observed that most of the mistakes made are related to discrepancies between how we structured the extraction process and the GENIA tagging method. First, NEs are sometimes embedded into larger ones and are annotated separately (e.g. "NF Kappa B binding sites" is correctly assigned to 'DNA' whereas the "NF Kappa B" portion within is sometimes, but not always, tagged as 'protein'). Second, other kinds of annotation mis-interpretations are related to optional and descriptive left-hand terms (e.g. "affinity-enriched NF-A2" versus "NF-A2") that may or may not be part of NEs, depending on the annotator. Other great sources of errors are due to the presence of family names, groups, substructures, subunits and complexes in the annotation. Our system does not handle correctly such biological objects, and especially the boundaries of the NEs. We tried to get rid of such limitations while re-annotating the GENIA subset to benchmark the identification process. During the analysis of the whole system performance (extraction + identification), we noticed that the errors were of few types and could be easily classified. The main source of false positives is both related to disambiguation problems and dictionary inconsistencies. Regarding problems with our dictionaries, it is not unfrequent that some generic names were registered by mistake in the dictionaries (e.g. "immediate early gene" is the name of a family of genes). Some entities herein wrongly identified correspond to a part of the name of the real entity (e.g. "pseudo tumor necrosis factor receptor" is absent from our dictionaries and

has been associated by default to "tumor necrosis factor receptor"). The main source of false negatives is first due to the absence of the entity in the dictionaries (to the most extent, it is usually seen with dictionaries made from the Metathesaurus UMLS), then to the lack of variant forms in our dictionaries, and finally to insufficient contextual information (to help disambiguating the true entity class). It is worth noting that a lot of errors that both count as false positives and false negatives are related to inaccurate disambiguation. For example, in phrases such as "gene activity of NF-Kappa-B" or "the expression of the gene encoding NF-Kappa-B", NF-Kappa-B is disambiguated as a gene but should definitely be a protein. It also usually happens that variants from our dictionaries belong to several identifiers, whereas most of these identifiers refer to the same entity. Curation of the dictionaries is an essential step that must not be overlooked. While analyzing the general results content, we did detect really few errors related to the variant generation and normalization methods, back from the dictionaries creation step. No additional ambiguity seems to be brought on by these techniques. The relatively bad results related to transcription factors binding sites are mainly the consequences of two major difficulties. On the one hand, their names are relatively short and ambiguous (e.g. "A box", "C"). Their nomenclature is the poorest, the less structured and the less supported of all our classes'. On the other hand, transcription factors binding sites references in text are usually implicit. For example, authors may use the nucleic sequence "TATAAA" to symbolize the entity "TATA box". Such nucleic sequences are extremely variable from one article to another.

While promising, there is a clear bias in the results obtained from the 100 abstracts corpus. The class representation is quite disproportionate and the corpus content is fairly homogeneous. NEs encountered in this subset are most often identical.

4 Related Work

Three works [12,13,14] are most similar to ours even if exclusively centered on gene/protein names. We were not able to directly compare the performances of our work to theirs because of different evaluation data sources. Tsuruoka *and al.* [12] present a method to alleviate the problem of spelling variation using an approximate string matching technique. Such approach appears to be relatively greedy in calculations. We have chosen to address this issue by carefully normalizing both dictionaries entries beforehand and extracting NEs on-the-fly. We seem to reach similar results and less overhead. Nevertheless direct comparison of both results remains difficult. They also try to handle short names problem with a machine-learning approach. Our strategy regarding this difficulty is, at the opposite, based on expert rules and encapsulated in a wider disambiguation process. Koike *and al.* [13] introduce an interesting feature we disregard in this study: the various objects manipulated in their dictionaries support composition and belonging relationships. They also try to detect if an entity name represents the protein itself or not (e.g. "IL-1 receptor expression" refers to the IL1 recep-

tor whereas "IL-1 receptor antagonist" does not). For this purpose, they seem to have manually constructed a list of contextual modifier terms. Egorov *and al.* [14] do only handle simple protein names but their system seems to perform very well on their test corpus. New valid spelling, morphologic and lexico-semantic variant forms of the protein names are not generated. They rely exclusively on the completeness of their dictionary content. They also do not try to check if an extracted NE is a syntactically valid noun compound in the molecular biology domain. Another interesting paper [15] proposes to combine an uncurated protein dictionary with Hidden Markov Models in order to identify NEs. They use hand-coded rules based on surface clues as a mean to pre-process the words or as features for the classifier. They achieved great results for protein name identification. However, all these approaches do not try to distinguish between the genic or proteic nature of the entity. Nor do they attempt to resolve forms of homonymy. Disambiguation methods exposed in [16] are able to discriminate between gene, protein and RNA forms with relatively great accuracy. It applies a machine learning approach as opposed to the hand-crafted techniques developed here. In their work, the original contextual features are the words and their relative positions around the current NE location. The rules we use could easily serve as contextual features for machine learning algorithms. The ProMiner tool [17] addresses two problems we eluded: species belonging of a NE and homonymy within a same class. It resolves these issues by respectively detecting organism names and unambiguous synonyms for a NE in the text. Implementing such methods will benefit our methodology, simply. It also avoids assigning an identifier to a partial NE match using an "aproximate search" technique. Using no other disambiguation methods, apart from common English clash detection, ProMiner achieves notable results on a large scale test corpus used during the BioCreative 2004 competition[14] and solely dedicated to genes and proteins.

5 Conclusion

We presented a simple method for extracting and identifying complex NEs in biology texts using controlled dictionaries. The main limitation of dictionary approaches is the inability to detect unknown entities. On the other hand, dictionaries are needed to recognize objects manipulated in texts. The main advantage of the techniques herein described are their relative genericity in the biomedical domain. We plan to re-implement the whole disambiguation process using machine-learning techniques. We hope to make it more reliable and easier to adapt to large corpora.
Nevertheless, several difficulties remain and need to be addressed. First, the coverage of the disambiguation process is still too limited and must be refined. Then, fully automatic updates of our dictionaries are not feasible yet. The databases used to create the dictionaries are far from being error-free and manual curation is needed. Finally, our text exploration is still fragmentary. Molecular biology

[14] http://www.mitre.org/public/biocreative/

publications exhibit a large number of NE family names and anaphora and coreferences that are not handled at the moment.

References

1. Fukuda, K., Tsunoda, T., Tamura, A., Takagi, T.: Toward information extraction: Identifying protein names from biological papers. In: Proc. of the Pacific Symposium on Biocomputing '98. (1998)
2. Collier, N., No, C., Tsujii, J.: Extracting the names of genes and gene products with a hidden markov model. In: Proc. COLING 2000. (2000) 201–207
3. Koike, A., Kobayashi, Y., Takagi, T.: Kinase pathway database: An integrated protein-kinase and nlp-based protein-interaction resource. Genome Res. **13**(6A) (2003) 1231–43
4. Tuason, O., Chen, L., Liu, H., Blake, J., Friedman, C.: Biological nomenclatures: Source of lexical knowledge and ambiguity. In: Proceedings of the Pacific Symposium of Biocomputing. Number 9 (2004) 238–249
5. Krauthammer, M., Nenadic, G.: Term identification in the biomedical literature. J. Biomed. Inform. **37**(6) (2004) 512–26
6. Lindberg, D., Humphreys, B., McCray, A.: The unified medical language system. Methods Inf Med. **32**(4) (1993) 281–91
7. Mikheev, A.: Periods, capitalized words, etc. Computational Linguistics (1999) 25
8. Tsuruoka, Y., Tateishi, Y., Jin-Dong, K., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a robust part-of-speech tagger for biomedical text. In: Proceedings of the 10th Panhellenic Conference on Informatics. (2005)
9. Jin-Dong, K., Ohta, T., Teteisi, Y., Tsujii, J.: Genia corpus - a semantically annotated corpus for bio-textmining. Bioinformatics **19(suppl. 1)** (2003) i180–i182
10. Zhou, G., Su, J.: Exploring deep knowledge resources in biomedical name recognition. In: Joint Workshop on Natural Language Processing in Biomedicine and its Applications. (2004)
11. Lee, C., Hou, W., Chen, H.: Annotating multiple types of biomedical entities: A single word classificication approach. In: Joint Workshop on Natural Language Processing in Biomedicine and its Applications. (2004)
12. Tsuruoka, Y., Tsujii, J.: Boosting precision and recall of dictionary-based protein name recognition. In: Proceedings of the ACL-03 Workshop on Natural Language Processing in Biomedicine. (2004)
13. Koike, A., Takagi, T.: Gene/protein/family name recognition in biomedical literature. In: Proceedings of HLT/NAACL BioLINK workshop. (2004) 9–16
14. Egorov, S., Yuryev, A., Daraselia, N.: A simple and practical dictionary-based approach for identification of proteins in medline abstracts. J Am Med Inform Assoc **11**(3) (2004) 174–178
15. Kou, Z., Cohen, W., Murphy, R.: High-recall protein entity recognition using a dictionary. Bioinformatics **21**(1) (2005) 266–273
16. Hatzivassiloglou, V., Duboue, P., Rzhetsky, A.: Disambiguating proteins, genes, and rna in text: A machine learning approach. In: Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology. (2001)
17. Hanisch, D., Fundel, K., Mevissen, H., Zimmer, R., Fluck, J.: Prominer: rule-based protein and gene entity recognition. BMC Bioinformatics **6 Suppl 1** (2005)

Annotation Guidelines for Machine Learning-Based Named Entity Recognition in Microbiology

Claire Nédellec, Philippe Bessières, Robert Bossy, Alain Kotoujansky, Alain-Pierre Manine

MIG-INRA, Domaine de Vilvert, F-78352 Jouy-en-Josas
email: forename.name@jouy.inra.fr

Abstract. Recent challenges on machine learning application to named-entity recognition in biology trigger discussions on the manual annotation guidelines for annotating the learning corpora. Some sources of potential inconsistency have been identified by corpus annotators and challenge participants. We go one step further by proposing specific annotation guidelines for biology and evaluating their effect on performances of machine learning methods. We show that a significant improvement can be achieved by this way that is not due to the feature set neither to the ML methods.

Keywords: Named-entity recognition, annotation guidelines, machine learning, biology.

1. Introduction

Named entities (NE) and terms represent the linguistic expressions that denote the objects and concepts in documents. As such their automatic annotation in document collections is a preliminary but crucial step for the semantic annotation and further document processing. Information Retrieval, Information Extraction (IE) and Question/Answering among others, rely on a proper identification of the objects and concepts in the documents. The NE dictionaries and terminologies that are needed for document annotation are available in some specific domains such as biology, but they often suffer from various limitations:

- they are incomplete with respect to the information processing tasks,
- additional disambiguation patterns are needed to handle the ambiguity and polysemy issue,
- variants of canonical terms and named entities that are needed to handle the synonymy issue are missing.

Automatic corpus-based acquisition of new NE and terms, disambiguation patterns, synonyms and variants has been considered as an attractive solution since the beginning of the nineties.

More recently the recognition of biological entities in scientific papers has been popularized by challenges such as NLPBA [Kim *et al*., 2004 ; *Collier et al*., 2005]

and BioCreative Task1a [Tanabe *et al.*, 2005 ; Yeh *et al.*, 2005]. As for MUC in the news wire domain, publicly available datasets and evaluation reports in biology have a very positive effect on research development in Machine Learning. However, as pointed out by [Tanabe *et al.*, 2005], [Dingare *et al.*, 2004] and [Alex *et al.*, 2006], it is difficult to build a consistent annotation of the training corpus in biology and this negatively affects the reliability of the method evaluation and comparison. Available corpora suffer from various inconsistencies. They are revealed by the analysis of the errors done by the learned NE recognition (NER) patterns when applied on test sets. The sources of potential errors are mainly the fuzzy frontier between entities denoted by proper nouns and entities denoted by terms (compound nouns), the lack of specification of the generality level of the objects to be recognized (entities *vs.* concepts) and the well-known problem of name boundaries. We have thus specified strict guidelines that make the manual annotations easier and more consistent and the NER patterns more learnable. Our strategy consists of splitting the NER task into two separate recognition subtasks, the recognition of the entities themselves and the recognition of their types (*e.g. GerE* and *protein* in *GerE protein*). The experiments reported here have been done on the classical problem of the recognition of new gene and protein names in the microbiology domain. We get much better results on the first subtask (*i.e.* entity recognition) than similar methods applied on biology corpora where the distinction between the annotation of entities and types is not so clear.
Section 2 motivates our annotation strategy as derived from the analysis of annotations inconsistencies in available corpora and from previous work on annotation guidelines. Section 3 reports on the experimental results and discuss them with respect to previous results in biology.

2. Annotation guidelines

2.1 NE *versus* terms

The distinction between entities and terms is recent and not fully linguistically relevant but it is operationally useful in IE where NER is one of the main tasks. The acquisition methods differ because of their morphological differences. Named-entities are proper nouns that often have upper case initials. Their variations are mainly typographic (*e.g. sigma K/ sigma(K)*). Terms are common nouns, often compound nouns, which follow traditional inflexion rules and their variations are mainly morpho-syntactic. The following four biological terms illustrate this:

ResD protein, either phosphorylated or unphosphorylated /
both unphosphorylated and phosphorylated ResD /
the phosphorylated form of ResD /
ResD~P.

In NER, the morphology usually determines the conditions that a given name should verify to be recognized as a NE rather than a term: NE recognition is mainly based on typographic criteria. Syntactic criteria have few effects on the NER performance. In biology, this usual morphology-based distinction does not apply. Terms often include proper nouns (Figure 1). Their role is generally to specialize the term meaning by

denoting specific identifiers as in *Streptococcus agalactiae NEM318 serotype III* where *NEM318* and *III* denote the reference to a *Streptococcus agalactiae* strain.
Moreover, the morphology-based distinction does not always fit the semantics; NE as proper nouns can denote concepts or types as well as instances of the concepts. Proper nouns and common terms can even be synonymous and then occur in similar contexts in corpus. Sense disambiguation (attaching the correct type to a given name) and new name recognition cannot rely on the morphology only but also on the context analysis in corpus. Therefore, NE (proper nouns) and term recognition patterns share similar contextual conditions. The example of acronyms and abbreviations clearly falls into this category. *glucose-specific enzyme II (EIIGlc)* where *glucose-specific enzyme II* is a term and *EIIGl* is its synonymous acronym is a representative example of this phenomenon. The NE and the term will be both recognized as enzymes. Typographic criteria are then not sufficient in biology for recognizing named entities.

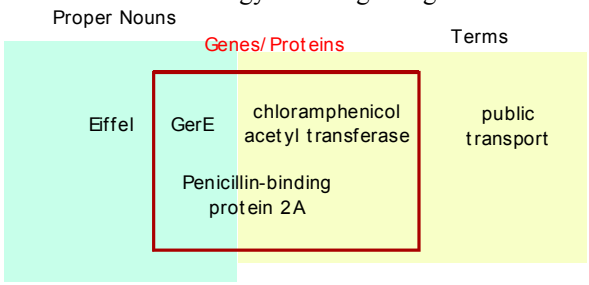


Fig. 1. Named-entities, Terms or Proper nouns?

In reality, the lexical frontier between the two kinds of knowledge is fuzzy and difficult to formalize. This affects the performances of the Machine Learning methods that are used for learning NER patterns because it is difficult to specify strict annotation guidelines so that the annotation can be reproducible and the NER patterns are learnable.
On the one hand, from the domain expert point of view the guidelines should refer to a consistent semantic category, for instance, *all company names* or *all gene/protein names* independently of their morphology. Such guidelines can make the learning task difficult because the morphologic constraints to be learned are different for the two classes of positive examples, NE and terms.
On the other hand, typographic conditions in recognition patterns are obviously much easier to learn if the guidelines are strict on the morphology - only proper nouns should be annotated as positive examples - but then, the contextual clues required for disambiguating the sense will be more difficult to learn, since terms considered as negative examples share the same contexts as positive examples.
The first strategy has been chosen in previous challenges and evaluations. In BioCreative, for instance, *SAA* and *serum response factor* (respectively proper noun and term) are both annotated as NE proteins. In BioLNLP, *PuB1* and *purine-rich binding sites* are both annotated as DNA (genes). It is natural from an application point of view: what one wants to acquire is a dictionary of a complete semantic category, independently of the morphology. However, the best scores in BioCreative are around 80% recall and precision and 76% recall and 69,4 % precision in NLPBA.

These relatively low scores compared to NE task in MUC can be explained by the morphologic difference of the names to be recognized.
We have thus explored the second strategy, *i.e.* learning proper noun recognition rules. Our hypothesis is that the different types of names, proper nouns and terms should be learned separately from different training corpus and with different methods. The target named entity dictionary would be then built by merging the results of the different learning tasks.
Since some terms include proper nouns, we have specified detailed guidelines so that the annotators can take consistent decisions. Terms that include proper nouns are annotated as named entities, when they denote specific objects and not general categories or types as detailed in the next section.

2.2 Entities versus concepts

The lack of clear distinction between entities (instances) and concepts (types, categories) is another source of inconsistent annotation and machine learning errors. General categories of biological objects are denoted by terms that occur in different contexts than the terms denoting the entities. They are very often in a coreference relation, mainly apposition as pointed out by [Vlachos *et al.,* 2006]. For instance in

> *[...] two alkaline phosphatases (APases) (PhoA and PhoB), an APase-alkaline phosphodiesterase (PhoD), a glycerophosphoryl diester phosphodiesterase (GlpQ), and the lipoproteinYdhF were identified […]*

the entity name *PhoA* is in apposition relation with the concept name *alkaline phosphatase, APase-alkaline phosphodiesterase* with *PhoD* and *glycerophosphoryl diester phosphodiesterase* with *GlpQ*.
Then learning relevant contextual conditions from mixed annotations of concepts and entities at different level of generality is difficult. Moreover, the frontier of the semantic category is much harder to specify in the annotation guidelines, if concepts are included. In biology, concepts are often denoted by general properties as it is in *binding-protein-dependent transport systems* and potentially not useful from an application point of view (*e.g. DNA-binding protein*). The decision to annotate a given term as a relevant concept or not is then difficult to take and very annotator-dependent. What is the limit between entities and concepts in the list *heat-shock sigma factor sigma 32*, *heat-shock sigma factor*, *heat-shock transcription factor*, *stress transcription factor*, *transcription factor*, *factor*? The usual strategies in previous work include both objects and concepts (*e.g. purine-rich binding sites* in NLPBA and *mouse synaptophysin gene* in BioCreative[1]).
We have followed another approach. Only specific objects are considered. For instance, *penicillin-binding protein 2A* is a positive example of protein while *penicillin-binding protein* is not, because it is too general and denotes a *family* of proteins. Following our guidelines, only the first element of the *factor* list above is considered as an entity (*i.e. heat-shock sigma factor sigma 32*). Note that this strategy partly resolves to the problem of the annotation of coordinated noun phrases pointed

[1] The task description mentions explicitly that *human gene* is too general. This illustrates how the limit is hard to specify.

out by [Alex *et al.*, 2006]. In *anhRad54 and hRad54B proteins* we annotate separately *anhRad54* and *hRad54B* and not *proteins*. The problem of annotating intersecting and non-contiguous noun phrases is then overcome. Some coordination problems still remain unsolved as in *interleukin 1 and 2*. Correct annotation of both *interleukin 1* and *interleukin 2* supposed that noncontiguous and intersecting annotations could be made. Note that it is not an inconsistency problem but a problem of specifying an appropriate syntax for the annotation.
Moreover, we have experimentally observed that specific objects (genes, proteins and species) are usually *not* denoted by common terms but either by proper nouns or by *mixed* terms that include proper nouns as identifiers. The morphology distinction looks then consistent with the entity/concept distinction.

2.3 Setting boundaries

The determination of the boundaries is a well-known source of errors. The most prevailing problem in biology is due to the term that denotes the semantic category in the context of the name to be recognized. It can occur before, as a modifier, or after, as a head (*e.g. GerE protein*, *protein GerE*). In most of previous works including NLPBA, the category has been considered as part of the entity name when the name is not an apposition in parentheses, or preceded by a comma. *cAMP regulatory element binding protein* is annotated as a unique name, as well as *a protein kinase A*. The two names in apposition are distinctly annotated in *monoclonal antibodies (mAb)* and in *GATA-1, an erythroid transcription factor*. This results in inconsistent NER where the type of the named entity can belong or not to the recognized name, depending on the punctuation marks of its context.
On the other hand, in BioCreative, the whole noun phrases are annotated even when commas or parentheses indicates chunk boundaries as in *Varicella-zoster virus (VZV) glycoprotein gI* that is annotated as a single named-entity. [Yeh *et al.*, 2005] hypothesized that the lower BioCreative results compared to similar tasks from MUC news wire domain could be explained by longer names in biology. The boundaries would be then more difficult to identify.
To overcome this problem, we follow a different strategy. As stated in the previous section, the expert does not annotate the general terms in apposition relation, such as *monoclonal antibodies* in *monoclonal antibodies (mAb)* but just the entity *mAb*.
Then two cases are considered, either the term denoting the semantic category is the head of the term containing the name, or it is a modifier. In the first case the head is not annotated as part of the entity name. For example, in *cAMP regulatory element binding protein,* only *cAMP* is annotated, as well as in, *Crp/Fnr family, the NtrB/C two-component system, P78 ABC transporter* (the entity names are in yellow). The short name is considered here as sufficient for naming the object.
In the second case where the semantic category is a modifier as in *cytochrome P450 102* and *penicillin-binding protein 2A*, the semantic category is annotated as part of the name only if it is required for the meaning, as it is the case in the second example but not in the first. *2A* is indeed not sufficient for denoting the protein, while *cytochrome* is redundant. The decision is based on biology expertise: is the category part of the name or not? In fact, the category is usually needed when the name is local

to the abstract (as *2A*). Then the name is generally very short and either a simple acronym or mostly composed of digits. Typographic criterion can then help in their identification. To summarize, the name denoting the entity should be annotated without its semantic type except when it is needed for comprehensibility reason. This guideline simplifies the annotation boundary problem and appears as intuitive for most of the biologist annotators in our experiments.

2.4 Semantic type

The last source of error is domain-dependent. The frontier of the semantic category to be annotated is often fuzzy as gene and protein categories are. We have decided to annotate the gene and protein category in their broad sense, including the following objects:

- the objects composed of protein and genes: *loci, alelles, operons*, *gene families*, *regulons*, *clusters, group, regions* and *fusion*
- the subpart of protein and genes: *promoters*, ORFs, *terminators,residues, motifs, boxes,* and *domains*
- part of the experimental material: *reporter genes*, *restriction enzymes*, *restriction sites*, *insertion elements*

A more detailed subtyping is left to further tasks.
The complete guidelines are available at genome.jouy.inra.fr/texte with more examples. The application of these guidelines to a corpus in microbiology is described in section 4. Section 3 presents the machine learning approach and the example representation language.

3. Machine-learning for NER

Our purpose is not to improve ML methods but to measure the effect of the guidelines on the NER performances. In our experiments, we have then selected the most successful approaches as reported in the related work. Previous works differ by the example feature sets, the use of external resources (dictionaries) and the ML method.

3.1 Related work in NER in biology

The main approach in NER in biology until the recent Machine Learning challenges was based on hand-coded pattern design. It relies on multiple sources of information: existing dictionaries and lexica such as UNIPROT, TREMBL, HUGO, UMLS among others [Rindflesh *et al.*, 2000; Cohen *et al.*, 2002; Leonard *et al.*, 2002], character and word-based approaches, linguistic processing [Proux *et al.* 1998], contextual disambiguation and domain knowledge [Humphreys *et al.* 2000; Fukuda *et al*. 1998; Hishiki *et al.* 1998; Franzen, 2002; Narayanaswamy *et al.*, 2003].
Until recently, the ML approach tended to use the linguistic information from the text but only few external resources. It was mainly achieved by the group of the GENIA project [Collier *et al.*, 2000; Nobata *et al.*, 1999; Takeuchi and Collier, 2002; Kazawa

et al., 2002]. Recent work agrees on the importance of example representation richness and the central role of the typographic features (see NLPBA and BioCreative conclusions). Among the most relevant features, the case and the non-alphabetic characters (*e.g.* hyphen, digits, symbols) and to a lesser extent, the neighborhood are determinant compared to syntactic categories [Collier and Takeuchi, 2004]. Syntactic dependencies are useful when semantic relations can be derived from them as described in [Wattarujeekrit and Collier, 2005].
Various ML and statistics-based methods have been tested, mainly Markov models, SVM, Maximum Entropy, naïve Bayes and decision tree algorithms. The best scores of the NLPBA challenge [Kim *et al.*, 2004] on the GENIA corpus have been obtained by [Zhou *et al.*, 2004a]. The method reaches 76,0 precision and 69,4 recall. It uses a rich example representation feature set and combines successively HMM and SVM. The best scores of Task1a at BioCreative were obtained by [Zhou *et al.*, 2004b] with a combined approach of HMM and SVM and by [Dingare et al., 2004] with a conditional Markov Model. Both yield around 82-83% recall and precision.
We have designed a similar feature set and selected SVM, C4.5 decision tree method and naïve Bayes (NB) as ML algorithms. We have applied the versions available in the WEKA library with the default parameters.

3.2 Dataset

Our training dataset is a subpart of an initial PubMed corpus on *Bacillus subtilis* (*Bs*) and transcription[2]. *Bacillus subtilis* is a model bacterium that has been extensively studied. The available knowledge on *Bs* genes, functions and metabolism can be usefully exploited for validating information extraction from text. We have chosen this domain because of our deep expertise on microbiology and on this specific *Bs* corpus. Therefore, we have been able to finely control the types of the biological objects to be annotated as well as the level of expert agreement on the annotation. The focus on the transcription issue increases the density of gene and protein names. With respect to the specific issue of transcription, we did not distinguish between genes and proteins as in BioCreative because they often cannot be automatically discriminated by their context because biologists consider the distinction as irrelevant and often use metonymies. A careful analysis did not reveal any obvious complexity difference between the names of our microbiology corpus and those of eukaryote corpora.
431 abstracts have been randomly selected among the 22397 references of the *Bacillus subtilis transcription* corpus. Among them, nine have been manually removed because of their heterogeneity. Their main topic was not microbiology but eukaryotic biology (*e.g. mycobacterium in tumor necrosis mice*). The remaining training corpus then contained 422 abstracts.

3.3 Corpus preparation

For saving manual annotation time, the corpus was first automatically pre-annotated

[2] The query was "Bacillus subtillis AND (transcription OR promoter OR sigma factor)"

by mapping a dictionary of gene and protein names. It was then manually corrected by biologist experts. This strategy is usual in NER. It globally improves the annotation quality but biases the annotation by preferring dictionary names which has a positive effect in our case. We have automatically designed the dictionary in order to limit the number of corrections to be done by the experts. The dictionary contains GenBank gene names of the only species mentioned in the corpus. We have assumed that no gene/protein name would occur in the corpus without a link to its species, except some experimental material such as *lacZ*. This limits the number of potential ambiguities and errors. As such, the dictionary still contained incorrect names because the format and guidelines for entering new references in GenBank are not strictly followed by the contributors. The dictionary was filtered by an anti-dictionary that contained the most frequent ambiguous names, such as *the* and *has* which are actually correct names but also highly ambiguous. It has been completed by six regular expressions that exclude the names represented by one or two letters or digits and long compound terms. The direct mapping of the dictionary to the corpus was completed by typographic variations. The anti-dictionary plus the regular expressions matched 25 014 occurrences in the corpus while the filtered dictionary matches 9 051 occurrences of species and gene/protein names. The number of potentially noisy occurrences was then more than twice the number of the potentially correct ones.

Table 1. Dictionary size.

Number of species names (including variations)	857 451
Number of protein/gene names (including variations)	401 790
Anti-dictionary size	289
Number of names removed by pattern matching	433

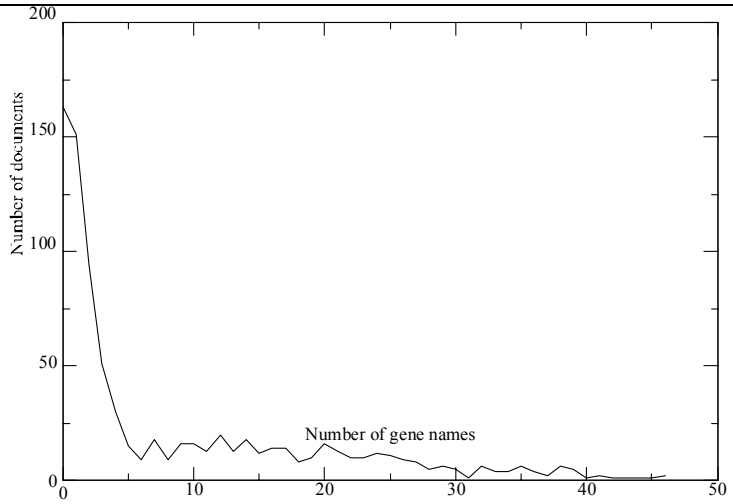


Figure 2. Number of abstracts (Y) containing X *gene/protein names*.

The annotation density varied among the abstracts. Figure 2 shows the distribution of the occurrences of the gene/protein names. Table 2 reports the number of protein/gene names automatically tagged in the corpus.

Table 2. Number of gene/protein names in *Bs transcription* corpus.

# *protein/gene* names occurrences	7 049
# *protein/gene* distinct names	1286

3.4 Manual annotation

The manual annotation was done with the Cadixe XML editor[3]. At the first stage, the corpus has been split into ten disjoint subparts and ten expert biologists corrected the automatic annotation of the dictionary mapping. Then one expert biologist carefully checked the annotation. In case of disagreement, a group of three biologists took a final decision. This way, a full agreement on all annotations was reached. This protocol has been applied for practical reasons only. Independent annotations should be done for measuring the expert agreement.

Three types of corrections of the automatic tagging were performed (Table 3):

- The annotations of the irrelevant homonyms were removed (for instance, *map*)
- The relevant anti-dictionary names (including regular expressions) were annotated (for instance, *has* gene). The length of most of them was one to two characters. The fourth column records those that are more than 2 characters long.
- The relevant names that were not in the dictionary were annotated (referred to as *new names*).
- The boundaries of the names have been modified.

Table 3 reports the number of manual corrections performed by category of error. These numbers are particularly important since they represent the goal of the learning approach: learning rules able to correct as much as possible the annotation done by a direct dictionary mapping.

Table 3. Manual corrections.

	Remove of irrelevant homonyms	Total additions	Addition of anti dictionary names	Addition of anti-dictionary names > 2 char.	Fully new names	Incorrect boundaries
# occ. (1st stage)	1057	1065	123	5	942	714
# occ. (2nd stage)	95	390	177	15	213	154
Total # occ.	956	1276	186	13	1090	781

[3] http://caderige.imag.fr/Cadixe

The number of ambiguities (false positives) was rather high (first column): 13 % (956/7049) of the annotations despite of the use of the anti-dictionary, which has been designed for reducing the ambiguities. The missing annotations were also close to 17 % of the total number of annotations and only a few of them (3 %) were present in the original dictionary and filtered by error. The other errors were due to fully new names, not present in the dictionary. This suggests that the anti-dictionary was not too strict. Incorrect boundaries represented a large part of the errors, around one quarter.
Table 4 reports the final numbers after manual correction. The figures in parentheses represent the name additions compared to automatic annotation. Additions represent the total of the name additions minus the deletions.

Table 4. Manual annotations of the *Bs transcription* corpus.

Total # protein/gene names occurrences	7 185 (+ 137)
Total # protein/gene distinct names	1647 (+ 361)
Total # species names occurrences	2 219 (+ 217)
Total # species distinct names	442 (+139)
Total number of occurrences of NE	9405 (+354)

Table 5 gives the recall and precision measures for the automatic filtered dictionary mapping compared to the manually annotated corpus. The measures were computed as a baseline for further comparison with the ML approach. We counted incorrect boundaries as two errors when an automatic annotation was replaced by one (one false positive, one false negative), three errors when the automatic annotation was replaced by two manual annotations (one false positive, two false negatives) and three errors when two automatic annotations were replaced by one manual annotations (two false positives and one false negative).

Table 5. Precision and recall of the filtered dictionary mapping.

Precision	Recall
76,1	78,1

The performances were surprisingly good compared to previous results by other authors, including the results obtained by hand-coded patterns. The way the dictionary has been filtered by choosing the names related to the relevant species and then filtered by the anti-dictionary was clearly very efficient.
The role of Machine Learning at this point is then double: disambiguating the homonyms and improving the coverage by recognizing new names.

3.5 Example representation

As other authors before, we hypothesized that typographic, linguistic and domain-specific features of the NE and their neighborhood are relevant for designing discriminant NER patterns. Table 6 describes the feature set.

Table 6. Features set

Features
Document structure
– **In_title**: the example belongs to the title.
Typographic features (boolean except length)
– **First_upper**: the example is capitalized (^[A-Z])
– **Middle_upper**: the example contains a non-initial uppercase letter (^.+[A-Z])
– **Only_upper**: all letters of the example are uppercase? (^[A-Z]*$)
– **Last_digit**: the last character of the example is a digit? ([0-9]$)
– **First_dash**: the example starts with an hyphen ('-')? (^-)
– **Middle_dash**: the example contains a non-initial hyphen? (^.+-)
– **Paren**: the example contains a paired set of parentheses? (\(.*\))
– **Space**: the example contains a space character (*ie* is the example is compound? ([ ])
– **Length**: number of characters of the example
– **Between_paren**: the example is enclosed between parentheses without any other word (not a regexp)
Dictionary features (boolean)
– **Eq_dict**: the example is a dictionary entry
– **In_dict**: the example is a strict subword of a dictionary entry
– **Eq_anti**: the example is an anti-dictionary entry?
– **In_anti**: the example is a strict subword of an anti-dictionary entry.
Linguistic features
– **Pos_following_X**: morpho-syntactic category of the Xth word following the example. X ∈ [1 .. 5]. Possible values: J (adjective), N (noun), PP (pronoun), RB (adverb), V (verb), O (other).
– **pos_preceding_X**: morph-syntactic category of the Xth word preceding the example.
Domain specific feature
– **Signal_in_following context:** word X from the signal list belongs to the following context of the example (window [+1 .. +5])
– **Signal_in_preceeding context:** word X from the signal list belongs to the preceding context of the example (window [-1 .. -5])

The role of the signal feature was to represent relevant signal words in the close context of the candidate named-entity. In order to define its value domain from the training corpus, we applied feature selection (based on information gain as implemented in WEKA) to the lemma of the predecessor and successor nouns, adjective and verbs of the positive and negative examples. The negative examples for computing feature selection were all nouns, non positive examples, and followed by a word from the signal list (Table 7), manually built for bootstrapping the process.

Table 7. Bootstrapping signal words acquisition.

activation box dependent enzyme expression fusion gene operon polymerase protease protein regulator regulon replication transcription

The size of the window varies from [-1 .. +1] to [-5 .. +5]. We retained the top 50 words for each window size. The most discriminant words differ depending on the position. For preliminary experiments, we did not want to consider exact position of signal words but an unordered set. In order to select the most popular words among

the five lists, we retained the words that belonged to at least 2 lists (*e.g.* it must be top 50 in 2-words window AND top 50 in 3-words window). The lists were then manually filtered by two ways: removing the spurious words such as auxiliary verbs (*be, do, have)* the semantics of which is not clear and removing too specific named entities with the exception of "*lacZ*" and "*Pho*" which are known to be within near context of gene names because they are part of the experiment material. The resulting filtered lists of signal words are given in Tables 8 and 9.

Table 8. List of signal words preceding the NE.

RNAse accumulate bacterial call collision contrary electrophoretic enable enzyme estimate expression genome include intracellular likely phosphorylation probe protein-mediated quantitative relative release respond result role second sequence-selective site-directed summary technique three-dimensional variety Pho activate activation analysis bind box dependent domain electrophoresis encode enzyme expression factor fusion homologue hybridization inhibit lacZ leader mRNA mutagenesis null phosphatase phosphorylated play polymerase protease protein regulator regulons replication reporter repress require responsible site strain substitution subunit synthetase transcript transcription transcriptional two-component

Table 9. List of signal words following the NE.

Pho activate activation analysis bind box dependent domain electrophoresis encode enzyme expression factor fusion homologue hybridization inhibit lacZ leader mRNA mutagenesis null phosphatase phosphorylated play polymerase protease protein regulator regulons replication reporter repress require responsible site strain substitution subunit synthetase transcript transcription transcriptional two-component

Most of the terms looked relevant as belonging to the candidate named-entity context while some others like *null* or *likely* looked more suspicious.
The positive examples were the examples of NE as tagged in the training corpus. Their description was based on their local context. We have considered fixed size windows within sentences boundaries. The negative examples were automatically derived from the annotated corpus as all noun phrases of one, two or three words in the corpus as analyzed by a basic chunker and non positive examples.

3.6 Experiments

Various combinations of example features were evaluated with the three ML methods, C4.5, SVM and NB. We report here the most significant features namely the typography, the signal words, the syntactic category and the dictionary (Table 10). The first three lines report the results computed with the whole feature set. C4.5 significatively yielded the best results. The most discriminant features of the resulting tree were typographic features (the root was the uppercase initial) and equality of a context word to a dictionary entry or inclusion. The rest of the table reports the results obtained by C4.5. As already pointed out in related work, the most discriminant features seemed to be the typographic ones (- 16 % precision and recall as shown in the last table line). The role of the features related to the dictionary was also important since their deletion yielded 5,5 % lack of precision and 2,1 % lack of recall. The POS

tag of the neighbor words of the candidate NE seemed to have no effect on the performances.

Table 10. Experiments with 3 ML algorithms and various feature sets.

	Precision	Recall
C4.5	93,6	93,4
SVM	86,2	89,9
NB	82,8	88,1
C4.5 no signal words	92 (-1,6)	93,3 (-0,1)
C4.5 no dictionary	88,1 (-5,5)	91,5 (-2,1)
C4.5 no POS tag	92,3 (-1,3)	93,9 (+0,5)
C4.5 no typography	77,4 (-16,2)	77,0 (-16,4)

The signal words lack of effect was surprising. Further experiments should be done with different sets of signal words on fixed position, since the lists obtained by the procedure of section 3 generated clearly different sets depending on the distance to the NE. At this stage our conclusion on the design of the feature set is very similar to those of previous works. The typography is very determinant while the POS tags seem to be useless.

Apart from the feature set, we evaluated the effect on the performance of the way the negative examples were generated. As such, the two negative and positive example sets were very unbalanced, the negative set being ten times larger. In order to assess the effect of the negative set size on learning, we trained C4.5 with a subset of randomly selected negative examples, such that this subset was of the same size as the positive set. The results did not improve as opposed to what was expected. It strongly affected the precision (77,6) and increased the recall (98,5). Further experiments should be done on intermediate negative example set sizes in order to evaluate the optimal size according to the corpus redundancy. We did other experiments with various near miss generation strategies that did not yield better results.

4. Discussion and conclusion

As expected, our experiments yielded higher performances than those reported by other authors on a similar NER task and on other corpora. They improve the precision of NLPBA best result by 17,6 % while the recall is 24 % better. Compared to BioCreative, the improvement is more than 10 % precision and recall. The main difference is the domain of the corpora (bacteria *vs.* eukaryotes) and the manual annotation rules. The sets of features are very similar. The ML algorithms are WEKA versions with default parameters and they are less sophisticated than the methods applied by previous challenges winners. We hypothesize that such a performance improvement is mostly due to the respect of consistent and strict annotation guidelines by the biologist annotators. The corpora on bacteria and eukaryotes do not

look so different with respect to the NER task that it would explain such different performances. In fact, our results reach similar rates as MUC ones on NER of proper noun such as location and person where the guidelines are comparable to ours: only proper nouns are annotated as NE and not general categories (*e.g.* not *town* in *town of Paris* or not *lake region* in *spring in lake region*). Further experiments with the same feature set and ML algorithm should be done on other corpora in order to confirm it.

We defend here the opinion that different types of knowledge, NER patterns for entities and categories should be separately acquired from corpus. It makes the manual annotation easier and the recognition patterns more learnable. We have demonstrated it here for NER pattern learning in microbiology. We have proposed relevant annotation guidelines with respect to this hypothesis. They are specific to biology and remove most of the inconsistencies observed by previous authors, namely, related to boundaries and granularity.

As specified, the NER learning task does not include more general category learning but only specific entities. We believe that it should be done by a separate learning task with more appropriate techniques that NER pattern learning, including ontology learning (Hearst's patterns and semantic distributional analysis) [Nedellec and Nazarenko, 2005] and term extraction methods that take into account morpho-syntactic variations instead of typographic features. Additionally to these acquisition considerations, it is more relevant from a knowledge modeling point of view to isolate the two tasks so that the two different kinds of knowledge, entities and types are formally represented and linked.

References

1. Kim J.-D, Ohta T. Tsuruoka Y., Tateisi Y. and Collier N. (2004). "Introduction to the Bio-Entity Recognition Task at JNLPBA", Collier et al. (eds), *Proceedings of NLPBA workshop* joint to Coling.
2. Zhou, G., Zhang, J., Su, J., Shen, D., Tan, C., 2004. Recognizing names in biomedical texts: a machine learning approach. Bioinformatics 20, 1178–1190.
3. Zhou G., Dan S., Jie Z., Jian S., Heng T. S. and Lim T. C. (2005) "Recognition of Protein/Gene Names from Text using an Ensemble of Classifiers", *BMC Bioinformatics Volume 6, Suppl 1*.
4. Collier N., Nazarenko A., Baud R. and Ruch P. (2005). Recent advances in natural language processing for biomedical applications. *Int J Med Inform*.
5. Yeh A., Morgan A., Colosimo M., Hirschman L. (2005). " BioCreAtIvE Task 1A: gene mention finding evaluation", *BMC Bioinformatics* 2005, 6(Suppl 1).
6. Tanabe L., Xie N., Thom L. H., Matten W., Wilbur W. J. (2005). "GENETAG: a tagged corpus for gene/protein named entity recognition". *BMC Bioinformatics 2005*, 6(Suppl 1).
7. Dingare S., Nissim M., Finkel J., Grover C., and Manning C. (2005) "A System For Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations". *Comparative and Functional Genomics*.
8. Alex B., Nissim M. and Grover C. (2006). The Impact of Annotation on the Performance of Protein Tagging in Biomedical Text. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006).* Genoa, Italy.
9. Vlachos A., Gasperin, C., Lewin I., Yamada C., Briscoe T.,"Bootstrapping the Recognition and Anaphoric Linking of Named Entities in Drosophila Articles", *Pacific Symposium on Biocomputing* 11:100-111, 2006.

10. Rindflesch T. C., Tanabe L., Weinstein J. N., Hunter L. (2000). EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. *Proceedings of PSB'2000,* vol 5:514-525.
11. Cohen K. B., Dolbey A. E., Acquaah-Mensah G. K. and Hunter L. (2002). Contrast and variability in gene names. *Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain*. pp. 14-20.
12. Leonard J. E., Colombe J. B., Levy J. L. (2002). Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics*, 18:1515-1522.
13. Proux D., Rechenmann F., Julliard L., Pillet V. and Jacq B. (1998). Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Informatics*. 9:72-80.
14. Humphreys K., Demetriou G., Gaizauskas R. (2000). Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. *PSB'2000*, 5:502-513.
15. Fukuda K., Tsunoda T., Tamura A., Takagi T. Toward information extraction : identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on biocomputing (PSB'1998),* 1998.
16. Hishiki T., Collier N., Nobata C., Ohta T., Ogata N., Sekimizu T., Steiner R., Park H. S., Tsujii J. (1998). Developping NLP tools for Genome Informatics: An Information Extraction Perspective. *Genome Informatics*. Universal Academy Press Inc., Tokyo, Japan.
17. Franzen K., Eriksson G., Olsson F., Asker L., Liden P. and Coster J. (2002). Protein names and how to find them. *Int J Med Inf.* 67(1-3): pp 49-61.
18. Narayanaswamy M., Ravikumar K. E., Vi jay-Shanker K. (2003). A Biological Named Entity Recognizer. *Pacific Symposium on Biocomputing* 8.
19. Collier N., Nobata C., Tsujii J. (2000). Extracting the Names of Genes and Gene Products with a Hidden Markov Model. *Proceedings of COLING-2000*, Sarrebrück.
20. Nobata C., Collier N. and Tsujii J. (1999). Automatic Term Identification and Classification in Biology Texts. In *Proceedings of the fifth Natural Language Processing Pacific Rim Symposium (NLPRS).* Beijin, China. pp. 369-374.
21. Takeuchi K. and Collier N. (2002). Use of Support Vector Machines in Extended Named Entity Recognition. *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002),* Taipei, Taiwan, August.
22. Kazawa J., Makino T., Ohta Y. and Tsujii Y. (2002). Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the Workshop of the Natural Language Processing in the Biomedical Domain in ACL '02*, Philadelphia, PA, USA, July.
23. Collier N.and Takeuchi K. (2004). "Comparison of character-level and part of speech features for name recognition in biomedical texts". *Journal of Biomedical Informatics* 37, 423-435.
24. Wattarujeekrit T. and Collier N. (2005), "Exploring Predicate-Argument Relations for Named Entity Recognition in the Molecular Biology Domain", *proceedings of the Eighth International Conference on Discovery Science (DS'05).*
25. Nédellec C. and Nazarenko A., (2005). "Ontology and Information Extraction: A Necessary Symbiosis", *Ontology Learning from Text: Methods, Evaluation and Applications.* Volume 123 Frontiers in Artificial Intelligence and Application, P. Buitelaar, P. Cimiano, B. Magnini (eds.), IOS Press, 2005.

On the Feasibility of Heterogeneous Analysis of Large Scale Biological Data

Ivan G. Costa, Alexander Schliep

Department of Computational Molecular Biology
Max Planck Institute for Molecular Genetics, Berlin, Germany
{ivan.filho,alexander.schliep}@molgen.mpg.de

Abstract. Secondary information such as Gene Ontology (GO) annotations or location analysis of transcription factor binding is often relied upon to demonstrate validity of clusters, by considering whether individual terms or factors are significantly enriched in clusters. If such an enrichment indeed supports validity, it should be helpful in finding biologically meaningful clusters in the first place. One simple framework which allows to do so and which does not rely on strong assumptions about the data is semi-supervised learning. A primary data source, gene expression time-courses, is clustered and GO annotation or transcription factor binding information, the secondary data, is used to define soft pair-wise constraints for pairs of genes for the computation of clusters. We show that this approach improves performance when high quality labels are available, but naive use of the heterogeneous data routinely used for cluster validation will actually decrease performance in clustering.

1 Introduction

A fundamental task in the analysis of gene expression time-courses is to find groups of genes undergoing the same transcriptional program or sharing similar functions. The numerous clustering methods proposed in the literature [2] are often validated by showing a statistically significant enrichment of individual Gene Ontology (GO) terms or transcription-factor binding information in some or all clusters. If the validity of a cluster is concluded from secondary data shared by its elements, a clustering procedure which prefers such clusters in the computation should yield superior results.

A natural, simple and mostly assumption-free framework is semi-supervised learning. Methods make use of labels which are available for a subset of objects in a combination of supervised and unsupervised learning. One particular type of methods is called clustering with constraints and it makes weaker assumptions about the labels by encoding secondary information as pair-wise constraints. We use either Gene Ontology annotation (`GO`) [1] or data from location analysis of transcription regulators bindings (`TR`) [6] as secondary information. For this data, the use of a clustering with constraints method, instead of a joint analysis approach [10, 11], has two advantages: (1) `GO` and `TR` is not available for all genes from expression experiments; and (2) gene expression time-courses provide

one view of the biological process under investigation, which is very unlikely to provide the same level of details as GO or TR data. Using such data as secondary information we can limit the results to biologically more plausible solutions.

One challenge of using GO or TR data as secondary knowledge is their complex and overlapping structure. GO, for example, consists of three directed acyclic graphs (DAG), composed of terms describing either molecular functions, processes or components. Most genes are directly annotated with several terms. Furthermore, if a gene is annotated with one term, it is also associated with all parent nodes of this term. Even though the structure of TR is simpler, genes are often associated with more than one transcription regulator and vice-versa. This work makes use of soft pair-wise constraints to model the secondary information, following the approaches of [5] and [7], and extending the semi-supervised approach applied for gene expression proposed in [9]. The challenge in this method is the formulation of the constraints between pairs of genes, which ideally should extract as much information from the secondary data as possible.

2 Mixture Model Estimation with Constraints

A standard mixture model can be defined as $\mathbf{P}[x_i|\Theta] = \sum_{k=1}^{K} \alpha_k \mathbf{P}[x_i|\theta_k]$, where $X = \{x_i\}_{i=1}^{N}$ is the set of observed vectors and $\Theta = (\alpha_1, ..., \alpha_K, \theta_1, ..., \theta_K)$ are the model parameters. By including a set of hidden labels $Y = \{y_i\}_{i=1}^{N}$, where $y_i \in \{1, .., K\}$ defines the component generating the x_i, we obtain the complete data likelihood, which can be estimated with the EM method.

$$\mathbf{P}[X, Y|\Theta] = \mathbf{P}[X|Y, \Theta]\mathbf{P}[Y|\Theta]$$

The constraints are incorporated in the estimation by extending the prior probability of the hidden variable to $\mathbf{P}[Y|\Theta, W] = \mathbf{P}[Y|\Theta]\mathbf{P}[W|Y, \Theta]$, where W is the set of positive constraints $w_{ij}^{+} \in [0, 1]$ and negative constraints $w_{ij}^{-} \in [0, 1]$, for all $1 \leq i < j \leq N$. As in Lu and Leen 2005, we use the following distribution from the exponential family to model $\mathbf{P}[W|Y, \Theta]$.

$$\mathbf{P}[W|\Theta, Y] = \frac{1}{Z} \exp^{\sum_i \sum_{j \neq i} -\lambda^{+} w_{ij}^{+} 1\{y_j \neq y_i\} - \lambda^{-} w_{ij}^{-} 1\{y_j = y_i\}}$$

Lange *et al.* [5] showed that this distribution follows the Maxent principle, where λ^{+} and λ^{-} are Lagrange parameters defining the penalty weights of positive and negative constraints violations. In this formulation, however, one cannot assume independence between elements in Y in the estimation step. Exact inference of the posterior $\mathbf{P}[y_i = k|x_i, \Theta]$ involves the marginalization over all objects with some non-zero constraint with the ith object. Such computation is only feasible when the constraints are highly decoupled, which is not the case of the structures in this study. One way to approximate the posterior distribution is to use a mean field approximation [5]. More formally, the posterior assignments will take the form

$$\mathbf{P}[y_i = k|Y_i', X, \Theta, W] = \frac{\alpha_k \mathbf{P}[x_i, \theta_k]}{Z} \exp\left(\sum_{j \neq i} -\lambda^{+} w_{ij}^{+}(1 - r_{j,k}) - \lambda^{-} w_{ij}^{-} r_{j,k}\right),$$

where $r_{j,k} = \mathbf{P}[y_j = k|Y'_j, X, \Theta, W]$. When there is no overlap in the annotations—more exactly, $w^+_{ij} \in \{0,1\}$, $w^-_{ij} \in \{0,1\}$, $w^+_{ij}w^-_{ij} = 0$, and $\lambda^+ = \lambda^- \sim \infty$—we obtain hard constraints as the ones used in [9], or as implicitly performed in [8].

2.1 Constraints Definitions

Each DAG of gene ontology is composed of a set of terms $T = \{t_1, ..., t_p\}$ and a set of parent child relations between pairs of terms $P(t_l, t_m) \in \mathcal{P}$. The annotation of a set of genes $G = \{g_1, ..., g_N\}$ can be defined as $A(t_l, g_i) \in \mathcal{A}$. Furthermore, we also have the property that genes annotated with a term are also annotated with the whole set of parents of this term, or $(P(t_l, t_m) \in \mathcal{P}) \wedge (A(t_m, g_i) \in \mathrm{A}) \rightarrow A(t_l, g_i) \in \mathcal{A}$. The main idea for calculating the constraint is to account for the similarity of the sub-dags $D(g_i) = \{t_m | A(t_m, g_i) \in \mathcal{A}, t_m \in \mathcal{T}\}$ associated with the gene pairs. More formally, for all pair of genes g_i and g_j, we define the constraints as (non-annotated genes have constraints equal to zero):

$$w^+_{ij} = \frac{\#\{t_m | t_m \in D(g_i) \cap D(g_j)\}}{\#\{t_m | t_m \in D(g_i) \cup D(g_j)\}}, \text{and } w^-_{ij} = \frac{\#\{t_m | t_m \in D(g_i) \uplus D(g_j)\}}{\#\{t_m | t_m \in D(g_i) \cup D(g_j)\}}.$$

Similarly, the formula above can be used for a set of of transcription factors $F = \{f_1, ..., f_q\}$, where $A'(f_l, g_i) \in \mathcal{A}'$ indicates that factor f_l bounds to g_i and $D'(g_i)$ is the set of factors associated with g_i.

3 Results

We use the expression profiles of 384 genes during Yeast mitotic cell division assigned to one of the five cell cycle phases classes [4], which we refer to as `YC5`. Even though this data set is biased towards profiles showing periodic behavior, and some of the class assignments are ambiguous, it is one of the few with a complete expert labeling of genes. The relation between regulators and target genes where obtained from large scale location analysis [6], comprising data from 142 candidate regulators. Relations were obtained after thresholding the confidence that the factor binds to a particular gene as in the source literature. In relation to GO, the SGD Saccharomyces cerevisiae annotation was used and for simplicity, we only included the DAG molecular process in our analysis.

Multivariate normal distributions with diagonal covariance matrix are used as models for the expression profiles. Each parameter estimation is performed 15 times and the best model is chosen, to lessen effects of random initialization. For all experiments we varied values of λ^+ and λ^-. We use the class labels to compute sensitivity (`Sens`), specificity (`Spec`) and corrected Rand (`CR`).

As a proof of concept, we use the class labels from `YC5` to generate pair-wise constraints for 5% of all pairs of genes—positive if the genes belong to the same class, negative else—and observe the performance of the method with distinct penalizing settings (Fig. 1 top). In all cases, `CR`, `Spec` and `Sens` tend to one for λ near ten, with the exception of the experiments with positive constraints. In

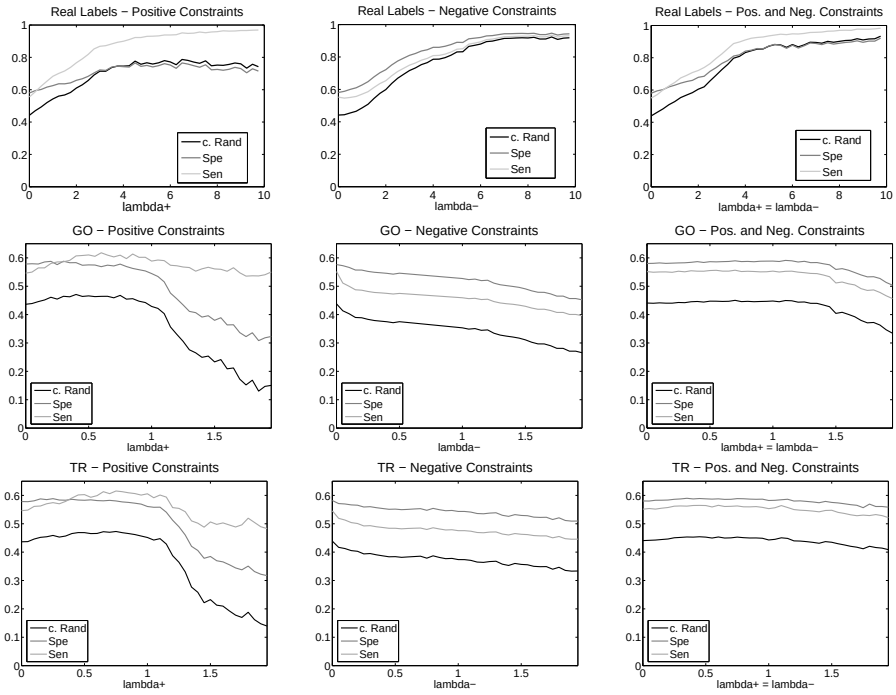


Fig. 1. We depict the `CR`, `Sens` and `Spec` after clustering `YC5` with positive (left), negative (middle) and positive and negative (right) constraints. We used either real class labels (top), `GO` (middle) or `TR` (bottom) as secondary information.

this case, one of the five components always remained empty, and two classes were joined. Furthermore, the use of positive constraints only had a stronger effect on the sensitivity, while the negative constraints affect the specificity. This is expected since these constraints penalize false negatives and false positives, respectively. It also explains the joined classes in the experiments with positive constraints, since the secondary data gives no penalty for those solutions (and the models for gene expression makes the decision).

We observe similar results if we use `GO` and `TR` as secondary data. There is a slight but significant increase of `CR` and `Sens` for the methods with positive constraints (t-test indicates an increase at $\lambda^+ = 0.5$ with p-value of $2.38e-10$), followed by a decrease in `CR`, `Sens` and `Spec` . No improvements were obtained with the use of positive and negative constraints, and the negative constraints alone only deteriorated the results. To better understand the results above, we repeated the experiments with real labels, but this time including random labels (also with 5% of pairs constrained). As seen in Fig. 2, the addition of random labels have a great impact on the recovery of the clusters. The inclusion of 20% of random labels worsen the results considerably, and for 60% of random labels the corrected Rand displays a behavior similar to `TR` and `GO`. This indicates that (1) the method is not robust in with respect to noise in the data, and (2) presence of noise or non-relevant information in `TR` and `GO`.

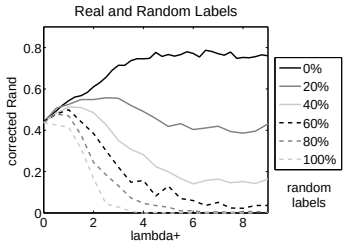


Fig. 2. We depict the `CR` obtained by clustering `YC5` with positive constraints from 5% of real labels with the inclusion of 0%, 20%, 40%, 60% and 100% random labels.

This however is not too surprising, so we attempt to estimate the maximal positive effect one can obtain from this secondary data. We perform the computation [3] for `GO` term and `TR` site enrichment used in cluster validation to obtain informative terms from the *true classes*. We repeat the experiments above with those most informative terms only. However, we observe only a slight improvement for the negative constraints and a marginal improvement with the use both positive and negative constraints in the `TR` data set (a `CR` from 0.454 to 0.472). On the other hand, no improvement was obtained after filtering terms in `GO` (data not shown).

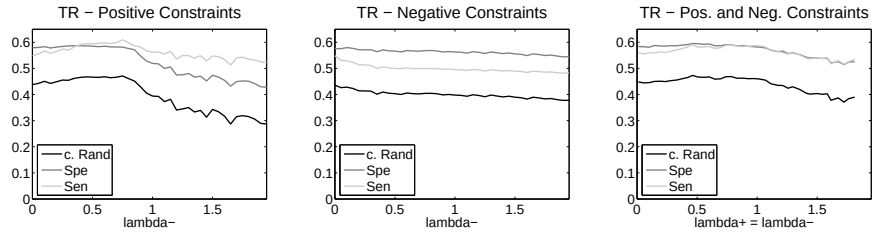


Fig. 3. We depict `CR`, `Sens` and `Spec` after clustering `YC5` with positive (left), negative (middle) and positive and negative (right) constraints after filtering of relevant `TR`.

4 Discussion

Semi-supervised learning is clearly an effective framework for joint analysis of heterogeneous data if high-quality secondary data is available as our experiments using class labels show. Surprisingly, using the very data routinely considered to support cluster validity—significantly enriched GO terms and location data—as secondary data can deteriorate cluster quality drastically. While there are parameter choices to explore, further theoretical questions to address and more data sets to repeat experiments on, the main point remains valid and clear: secondary data has little power for clustering, unless it is of very high quality, free of errors and ambiguities. Less than a percent of high-quality labels [9] have

a larger positive effect than 5% of labels of which 20% are incorrect. On one hand, this puts the economy of large-scale experiments into question. On the other hand, it stresses the importance of theoretical progress on how to reduce noise, assess reliability of individual data and how to incorporate per object quality indicators into methods.

Acknowledgments. The first author would like to acknowledge funding from the CNPq(Brazil)/DAAD.

References

1. M. Ashburner. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000.
2. Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–503, Nov 2004.
3. T. Beissbarth and T. P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.
4. R. Cho. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.
5. T. Lange, M. H. Law, A. K. Jain, and J. M. Buhmann. Learning with constrained and unlabelled data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 731–738, 2005.
6. T. Lee. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.
7. Z. Lu and T. Leen. Semi-supervised learning with penalized probabilistic clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 849–856. MIT Press, Cambridge, MA, 2005.
8. W. Pan. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, 22(7):795–801, 2006.
9. A. Schliep, I. G. Costa, C. Steinhoff, and A. A. Schnhuth. Analyzing gene expression time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3):179–193, 2005.
10. E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19(90001):273i–282, 2003.
11. C.-H. Yeang and T. Jaakkola. Time series analysis of gene expression and location data. In *Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)*, number 305, 2003.

Marker Analysis with APRIORI-based Algorithms

Giacomo Gamberoni[1], Evelina Lamma[1], Fabrizio Riguzzi[1]
, Sergio Storari[1], and Chiara Scapoli[2]

[1] ENDIF, University of Ferrara, Italy
{giacomo.gamberoni, evelina.lamma, fabrizio.riguzzi,
sergio.storari}@unife.it
[2] Department of Biology, University of Ferrara, Italy, scc@unife.it

Abstract. In genetic studies, complex diseases are often analyzed searching for marker patterns that play a significant role in the susceptibility to the disease. In this paper we consider a dataset regarding periodontitis, that includes the analysis of nine genetic markers for 148 individuals. We analyze these data by using two APRIORI-based algorithms: APRIORI-SD and APRIORI with filtering. The discovered rules (especially those found by APRIORI with filtering) confirmed the results published on periodontitis.

1 Introduction

In classical genetics [1], diseases are divided into Mendelian disorders and complex traits. While the former are attributed to single gene mutation with a strong effect on phenotype and a simple mode of inheritance, the latter are thought to result from interaction of a polygenic system, governed by the simultaneous action of many genes and an environmental component. The main task in the study of these polygenic systems is obviously to find the genetic patterns that increase susceptibility to the diseases.

In this paper we focus on marker analysis, and in particular we consider the problem of determining the relation between nine genetic polymorphisms and periodontitis. To this purpose we applied machine learning techniques based on association rules to a dataset regarding 148 subjects. In particular we applied APRIORI-SD and APRIORI with filtering from the Orange Suite [2].

The paper is organized as follows: Section 2 illustrates the case study: the group of subjects genetically characterized and the set of genetic markers analyzed. Section 3 presents subgroup discovery algorithms. Section 4 reports the results of applying these algorithms to the genetic dataset.

2 Marker Analysis

Most common diseases are complex genetic traits [1], where a multiple gene system and environmental variables contribute to the observed phenotype. Because

of the multi-factorial nature of complex traits, each individual genetic variant (susceptibility allele) generally has only a modest effect, and the interaction of genetic variants with each other or with environmental factors can potentially be quite important in determining the observed phenotype. Genetic association studies, in which the allele or genotype frequencies at markers are determined in affected individuals and compared with those of controls (case-control study design), may be an effective approach to detecting the effects of common susceptibility variants.

Periodontitis Dataset. As an example of a complex genetic trait, we choose Generalized Aggressive Periodontitis (GAP) as a case study. Periodonditis is a dental disorder that results from the progression of gingivitis, involving inflammation and infection of the ligaments and bones that support the teeth.

The dataset, provided by Prof. L. Trombelli of the Research Center for the Study of Periodontal Diseases, University of Ferrara (Italy), collects blood samples from 46 GAP patients (16 males and 30 females) and 102 periodontally healthy control subjects. All subjects were chosen amongst current and permanent residents in the area of the Ferrara district (Italy). Systemically healthy GAP patients were selected for the study among those undergoing periodontal supportive therapy at the Research Center for the Study of Periodontal Diseases, and the diagnoses were confirmed by the same clinician. The clinical diagnosis at the time of the initial visit was based on recent international classification [3]. The periodontally healthy control subjects were selected among voluntary people if they showed no interproximal attachment loss greater than 2 mm at any of the fully erupted teeth. Controls were matched by age and sex with GAP patients. All GAP patients and controls were Caucasian Italian. The study design was approved by the local ethical committee and written informed consent was provided by all participants in line with the Helsinki Declaration before inclusion in the study.

The following variants in the IL-1 gene cluster have been tested: IL-1α^{+4845} (recorded as $M1$), IL-1β^{+3953} ($M3$), IL-1β^{-511} ($M2$) and also the minisatellite of IL-1RN intron 2 ($M5$). Furthermore, it has been tested a new marker variant at the IL-1F5 ($M6$) gene as described in Scapoli et al. [4]. Besides polymorphisms at IL-1 cluster, other markers have been tested in different pro-inflammatory cytokine genes such as IL-6 (variant IL-6^{-174} ($M8$) and IL-6^{-622} ($M7$)) and TNF-A (variant TNF-α^{-308} ($M4$)). Finally also a polymorphism at the TNF-α receptor has been tested (TNFRSF1β^{+196} ($M9$)).

Related Studies. Several studies have shown a role for the involvement of interleukin-1 (IL) gene cluster polymorphisms in the risk of periodontal diseases. In [5] the authors tested polymorphisms, derived from genes of the IL1 cluster, for association with generalised aggressive periodontitis (GAP) through both allelic association and by constructing a Linkage Disequilibrium map of the 2q13-14 disease candidate region. For the IL-1RN, a statistically significant difference was found between patients and controls in the genotypic distribution, but no significant difference was found for allelic distribution. Authors also

observed some evidence for an association between GAP and the IL-1β^{+3953} polymorphism.

For the other IL-1 Cluster polymorphisms and cytokine genes, no significant differences were found between patients and controls for both genotypic and allelic frequencies. Moreover, other studies [6] identified the IL-1β^{+3953} polymorphism as implicated in GAP.

3 Algorithms

In order to find the attributes (genetic mutations) most related to the class (case or control group), we exploited APRIORI-SD for subgroup discovery and APRIORI with filtering for generating classification rules.

The goal of subgroup discovery [7] is to find subgroups, represented by rules, which describe subsets of the population that are sufficiently large and statistically unusual with respect to a target attribute. For example, we may look for groups that are as large as possible and on which the property of interest has a distribution that is as different as possible with respect to the distribution over the whole population.

3.1 Association and classification rules

Association rules Consider a table D having only discrete attributes. If D has also numeric attributes, they are discretized. An *item* is a literal of the form $A = v$ where A is an attribute of D and v is a value in the domain of A. Let M be the set of all the possible items. An *itemset* X is a set of items, i.e. it is such that $X \subseteq M$. A k-itemset is an itemset with k elements. We say that a record r of D *contains* an itemset X if $X \subseteq r$ or, alternatively, if r satisfies all the items in X. Let $n(X)$ be the number of records of D that contain X. Let $n(\overline{X})$ be the number of records of D that do not contain X. Let N be the number of records of D. The *support* of an itemset X (indicated by $Sup(X)$) is the fraction of records in D that contain X. i.e., $Sup(X) = n(X)/N$. It is also equal to the probability of a record of D of satisfying X, i.e. $p(X) = Sup(X)$. When X and Y are two itemsets we use the shorthand notation $n(XY)$, $Sup(XY)$ and $p(XY)$ to mean, respectively, $n(X \cup Y)$, $Sup(X \cup Y)$ and $p(X \cup Y)$.

Association rules are of the form $B \rightarrow H$ where B and H are itemsets such that $B \cap H = \emptyset$. B and H are respectively called *body* and *head*. For association rules a number of quality metrics can be defined.

Given an association rule $R = B \rightarrow H$, we define the following metrics:

- Support: $Sup(R) = p(BH) = Sup(BH) = \frac{n(BH)}{N}$
- Confidence: $Conf(R) = p(H|B) = \frac{Sup(BH)}{Sup(B)} = \frac{n(BH)}{n(B)}$
- Novelty: $Nov(R) = p(HB) - p(H)p(B)$
- Weighted Relative Accuracy: $WRAcc(R) = Nov(R)$

The definition of novelty [8] states that we are only interested in high support if that could not be expected from the marginal probabilities, i.e., when $p(H)$ and/or $p(B)$ are relatively low. It can be demonstrated that $-0.25 \leq Nov(R) \leq 0.25$: a strongly positive value indicates a strong association between H and B.

Classification rules are association rules whose head is of the form $Class = c$ where $Class$ is a special attribute of D. In this case, the records of D are also called *examples* and a rule $B \rightarrow Class = c$ *covers* a record r if $B \subseteq r$ and correctly covers a record if $B \cup \{Class = c\} \subseteq r$.

3.2 APRIORI

The task of discovering association rules consists in finding all the association rules having a minimum support $minsup$ and a minimum confidence $minconf$. In order to discover such rules, the approach proposed in [9] first discovers all the itemsets with support higher than $minsup$ and then finds the rules from them. The itemset with support above $minsup$ are called *large*.

APRIORI is based on the fact that $X \supseteq Y \rightarrow Sup(X) \leq Sup(Y)$. Therefore if $Sup(X) < minsup$ then $\forall Y \supseteq X,\ Sup(Y) < minsup$. So we can discard every itemset that has a non large subset.

APRIORI can also be used to produce classification rules, by filtering its output, keeping only the rules with a class assignment as the head.

3.3 APRIORI-SD

APRIORI-SD [10] is an algorithm for performing subgroup discovery that is based on APRIORI-C [11]. APRIORI-C runs the APRIORI algorithm, and takes into consideration only classification rules.

APRIORI-C also performs a post-processing step, in one of two ways: Select N best rules and Select N best rules for each class. In the first scheme, the algorithm first selects the best rule (the rule having the highest support), then eliminates all the covered examples, recomputes the support for the remaining rules and repeats the procedure until N rules are selected or there are no more rules to select or there are no more examples to cover. In the second scheme, the first scheme is repeated for each class in turn.

APRIORI-SD modifies the post-processing step of APRIORI-C by adopting a weighting scheme for the coverage of examples and by using a measure for evaluating rule different from support.

In the weighting scheme, example are not immediately removed when they are covered by a best rule, but instead their weight is reduced. Initially all the examples have weight 1, when an example has been covered i times its weight is reduced to $\frac{1}{i+1}$. In this way we increase the chance of returning rules covering every part of the training database.

As regards the evaluation measure, APRIORI-SD uses Weighted Relative Accuracy with Example Weights: it is the same formula of Weighted Relative Accuracy where each example count is replaced by the sum of the weights of the examples.

4 Experiments

Dataset preparation. The application of the algorithms to the dataset was performed by an examiner who was blinded as to the correspondence of the M1, M2,. . . ,M9 variables and the related polymorphisms, so that the examiner had not information on previous statistical analyses and on the expected results about IL-1β^{+3953} (M3), IL-1RN (M5) and TNFRSF1β^{+196} (M9) markers and the disease status.

Starting from the blinded dataset originated from the GAP study, we obtained a new dataset on which we ran the experiments. In the original dataset, each marker can assume three possible values: 11, 12 and 22. 11 and 22 are homozygote genotypes while 12 define the heterozygote status. As an example, if there are two markers $(M1, M2)$ a possible record of the dataset is $(11, 12)$. In our analysis we consider the configuration of a single chromosome and we want to test, for each marker, whether the allele on that chromosome is 1 or 2. For heterozygote individuals, we do not know on which chromosomes lies the 1: in other words, the allelic configuration for the marker on the two chromosomes could be 12 or 21 with equal probability. The new dataset will contain, for each record from the original dataset all possible configurations of a single chromosome (haplotype) compatible with the record. Therefore, for each record in the original dataset, we generate 2^k tuples in the new dataset, where k is the number of marker analyzed. For example, in the case of the record above, the new dataset will contain the four tuples: $(1, 1)$, $(1, 2)$, $(1, 1)$ and $(1, 2)$.

Results. Applying APRIORI-SD with settings $minsup = 0.05$, $minconf = 0$ (other parameters set as default), for the target class `status = GAP` (for `status = normal` the algorithm did not return any rule) we obtain:

```
1. M1=2 M9=2 → status=GAP
2. M3=2 M9=2 → status=GAP
3. M3=2 M4=1 M7=1 M9=1 → status=GAP
4. M3=2 M4=1 M6=1 M7=1 → status=GAP
5. M3=2 M6=1 M7=1 M8=1 → status=GAP
```

These rules involve almost all the markers, so they are not useful to discriminate the most involved ones. Moreover they do not include M5, known in literature to influence GAP.

For APRIORI, we used $minsup = 0.05$, $minconf = 0$ (other parameters set as default). As an example, we report the five rules with the best novelty value:

```
1. M9=1 → status=normal
2. M5=1 M9=1 → status=normal
3. M3=1 M5=1 M9=1 → status=normal
4. M3=1 M9=1 → status=normal
5. M1=1 M3=1 M9=1 → status=normal
```

Several rules show the three markers that have been reported in literature as involved in the pathology: M3, M5 and M9. The correlation found between them (i.e. the third rule) is interesting and will be object of further biological investigations.

5 Conclusion

In this work we applied APRIORI based algorithms to a marker analysis task. In particular we used two methods: a subgroup discovery algorithm (APRIORI-SD) and APRIORI with classification rules filtering. The subgroup discovery approach provided less general rules that involve almost all the markers. These rules are quite useless for marker analysis. Considering the second approach, ordering by novelty the learned rules, we found results that are both coherent with the literature and interesting for further studies.

6 Acknowledgments

We would like to thank N. Lavrac and P. Kralj for providing us the APRIORI-SD code.

Part of the research was funded by PRIN 2005 project "Specification and verification of agent interaction protocols".

References

1. Lewin, B.: Genes VII. Oxford University Press, Oxford (2000)
2. Demsar, J., Zupan, B., Leban, G.: Orange: From experimental machine learning to interactive data mining (2004)
3. Tonetti, M.S., Mombelli, A.: Early-onset periodontitis. Ann Periodontol **4** (1999) 39–53
4. Scapoli, C., Tatakis, D., Mamolini, E., Trombelli, L.: Modulation of clinical expression of plaque-induced gingivitis: interleukin-1 gene cluster polymorphisms. J Periodontol **76** (2005) 49–56
5. Scapoli, C., Trombelli, L., Mamolini, E., Collins, A.: Linkage disequilibrium analysis of case-control data: an application to generalized aggressive periodontitis. Genes Immun **6** (2005) 44–52
6. Diehl, S., Wang, Y., Brooks, C.e.a.: Linkage disequilibrium of interleukin-1 genetic polymorphisms with early-onset periodontitis. J Periodontol **70** (1999) 418–430
7. Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and Data Mining. AAAI/MIT (1996) 249–271
8. Lavrac, N., Flach, P., Zupan, B.: Rule evaluation measures: A unifying view. In Dzeroski, S., Flach, P.A., eds.: ILP. Volume 1634 of Lecture Notes in Computer Science., Springer (1999) 174–185
9. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In Bocca, J., Jarke, M., Zaniolo, C., eds.: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Morgan Kaufmann (1994) 487–499
10. Kavsek, B., Lavrac, N., Jovanoski, V.: Apriori-sd: Adapting association rule learning to subgroup discovery. In Berthold, M.R., Lenz, H., Bradley, E., Kruse, R., Borgelt, C., eds.: IDA. Volume 2810 of Lecture Notes in Computer Science., Springer (2003) 230–241
11. Jovanoski, V., Lavrac, N.: Classification rule learning with apriori-c. In P.Brazdil, Jorge, A., eds.: EPIA. Volume 2258 of Lecture Notes in Computer Science., Springer (2001) 44–51

Getting the Unknown from the Known in Bacteria, and the Role of Text Mining

Philippe Bessières, Robert Bossy, Alain-Pierre Manine, Erick Alphonse, and Claire Nédellec

Mathématique, Informatique et Génome (MIG), INRA,
78352 Jouy-en-Josas cedex, France
erick.alphonse@lipn.univ-paris13.fr
{philippe.bessieres,robert.bossy,alain-pierre.manine,claire.nedellec}@jouy.inra.fr

Abstract. A lot of biological knowledge is only described in scientific texts, such as regulations or rich functional descriptions of genes and proteins, while it is wanted to contribute to the decyphering of biological processes by large scale analyses applied in genomics. Then text mining is an important challenge in bioinformatics, as new techniques allow global approaches of organisms that call for mining large sets of gene expression and sequence data with knowledge from the literature.

Key words: Text Mining, Information Extraction, Bioinformatics, Computational Biology, Genomics, Bacteria

1 A Deluge of Experimental Data

Molecular biology has engaged into a permanent revolution from the second half of the last century, while DNA was, at least, identified as the support of genes. Several great leaps forward happened with the access to nucleic acid and protein sequences and structures, and the cloning of genes. The launching of the human genome project, for reading the complete sequence of human chromosomes, and from where was coined the term "genomics", is the most noticeable consequence of this evolution [1]. Similar projects addressing "model organisms" immediatly followed, such as the yeast *Saccharomyces cerevisiae*, and the bacterium *Bacillus subtilis*. While these large consortia have been implying hundreds of persons for around a decade, twelve years after the first publications of complete genomes in 1995, we will get one thousand of them, and the sequencing of a bacterium is now a matter of months, within a single laboratory's reach.

One very important and surprising result coming from the systematic sequencing projects was the unexpected discovery of important amounts of genes of unknown function [2], and this observation called in the nineties for new ambitious programs, termed "functional genomics" [3]. Their main goals are to unravel the function of unknown genes, and to understand how the components of the cells, tissues, and organisms, cooperate to realize biological functions.

2 Integrative Biology, Bioinformatics, and Text Mining

Such projects try to link the sequence to the "phenotype", formally an observable character, going through the hierarchy of biochemical and physiological functions into which the genes are implied. Therefore, they promote approaches at the global scale of living cells, calling for new techniques that give complementary views of the ways they work: expression of genes using "DNA microarrays" (transcriptome), production of proteins and their physical interactions (proteome), or metabolic pathways (metabolome). Still dramatically improving, with continuous refinements, scaling up, and large spreading, they become common features of the life of laboratories, and profoundly shape what is the biology of today. Then, we have entered into the era of "integrative biology", and the deluge of data coming from large scale experiments imposed bioinformatics and computational biology [4] as an essential tool for everyone, dealing with two large categories of tasks: analysis, and integration of biological data.

Sequence, structure, expression data analysis may be envisioned in the perspective of the annotation of genomes [5], basically meaning to identify the genes on the new chromosomal sequences, and then to assign them a function, firstly by comparisons of sequences with those of experimentally known genes. The assignment of functions to genes is a source of challenging data mining problems for computer scientists and mathematicians, as new approaches recently appeared that have been proven fruitful. Entitled "gene context analysis", the concept is gathering a variety of methods that try to functionally link unknown genes to known ones, for example by identifying coexpressed genes in one species [6], or conserved neighborhood of genes between different species [7]. Interestingly, several inventors of these methods also found of the greatest importance to extract information about interactions and functional links between genes and proteins from the text of publications [8]. They set to work with text mining on the statement that thousands of links resulting from small scale experiments are already published in the scientific literature, and therefore may usefully complement large scale experiments to feed interaction databases [9].

Combining heterogeneous data in computer analyses, and trying to make sense of them, is calling for their integration, the second great category of challenge for bioinformatics. International databanks dedicated to structures and sequences were created as soon as they were obtainable, explaining why the community of biologists promptly adopted the communication tools provided by the computer networks. Hence, when online specialized collections began to flourish from the genome projects, a first system was early proposed to link them through the Internet for retrieval purposes [10]. At this time, data collections upgraded from flat files of indexed records to real database structures, and looking for their interoperability by using available technologies, the bioinformaticians realized that this aim was seriously impeded by problems of semantics [11]. So they were introduced to the ontologies [12], as a way to solve the discrepancies between the textual annotations and the concepts used by the databases, and this finally was the reason of the creation of the Gene Ontology consortium [13].

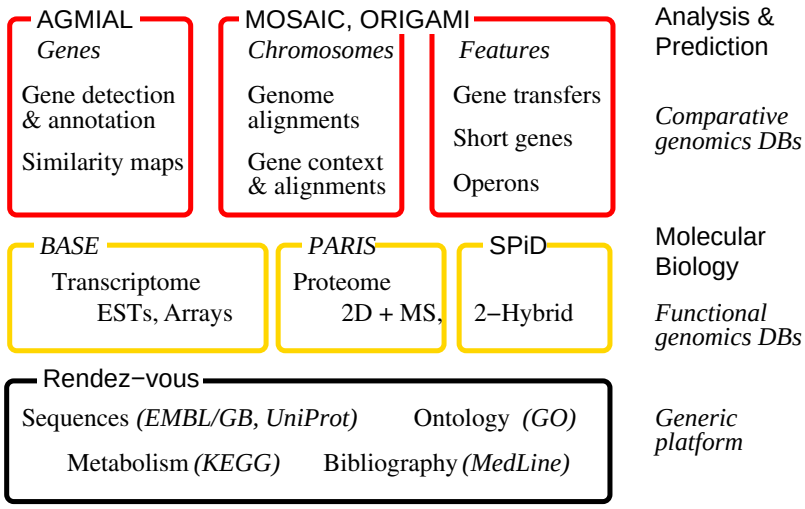


Fig. 1. An environment of relational databases used in MIG for the analysis of bacterial genomes. Three levels are considered, the first one, called Rendez-vous, integrates annotations of DNA (EMBL/GenBank) and protein sequences (UniProt), metabolism (KEGG), bibliographic notices (MedLine), and gene ontology from GO [13]. The second level is managing results of large scale experiments, with DNA microarrays (BASE), protein two-dimensional gel electrophoresis and mass spectrometry (PARIS) and physical interactions by the two-hybrid method (SPiD) [14]. Only the third level is specific to bacteria, and is also using the lower levels. AGMIAL is a genome annotation platform [5], while MOSAIC and ORIGAMI are dedicated to the analysis of special chromosomal features by comparative genomics and statistical methods.

3 Bacterial Gene and Protein Interactions from Text

We see that the general and growing interest of biologists and bioinformaticians for textual information and text mining is relying on several motivations. This should allow to collect sparse knowledge from publications into database environments for the analysis of genomes, such as the one presented in Fig. 1, but also to contribute to a better control of the integration of the databases between themselves, and to improve the efficiency of queries to their textual annotations.

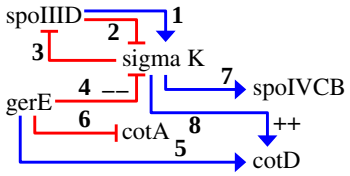
A lot of elaborate and fine knowledge is only described in scientific texts, like regulations, or rich functional descriptions of genes and proteins, and this is especially true for well studied and highly tractable model bacteria such as *Bacillus subtilis* and *Escherichia coli*. Being the best known biological species, they act as references beyond the scope of bacteria, so there is a considerable interest about extracting and structuring this knowledge from literature, and to mine with it large sets of gene expression and sequence data in integrated approaches. A first demand comes from genome annotation, where going back to the text of publications is the most limiting step, when trying to guess a function

1. SpoIIID is needed to produce sigma K

2. SpoIIID is capable of altering the specificity of RNAP–sigma K

3. production of sigma K leads to a decrease in the level of SpoIIID

4. GerE profoundly inhibits in vitro transcription of sigK encoding sigma K



5. GerE stimulates cotD transcription

6. ... and inhibits cotA transcription

7. sigma K has been found that causes weak transcription of spoIVCB

8. ... and strong transcription of cotD

Fig. 2. Example of a regulatory network in *Bacillus subtilis*, from the text of MedLine abstracts. The arrows represent activations, and the bars inhibitions.

for genes with uncertain sequence homologies. Concerning large scale expression data, text mining is needed to build and to validate the experiments, in preamble of using it for extracting some biological meaning. Finally, in a time where we are engaging into "systems biology" projects, our ability to extract knowledge on regulations from text is a critical point, in attempt to build realistic and predictive dynamic models for analyzing and decyphering biological functions.

Therefore, we want to automatically extract and represent regulatory networks from descriptions in scientific texts, and Fig. 2 shows an example, here implied in the formation of spores in *B. subtilis*, from MedLine abstracts. Ideally, we want to identify the agents and the targets (the actors), with their types (gene, protein), the types of regulations (transcription, translation, etc.), their roles (activate, inhibit), and their strength (strong, weak). For clarity, we selected simple examples for the figure, but this must not making us underestimating the complexity of the information extraction task.

Identifying Interaction Actors. As a convention for naming bacterial genes exists, and is followed in the case of a model species like *B. subtilis*, their identification in texts seems easy. They begin with an acronym of their function of three lower-case letters, ending by an upper-case one, to distinguish between several genes contributing to the function. *cotA*, and *cotD*, in Fig. 2 are coding for proteins composing the *coat* of the spore. So why *spoIIID* or *spoIVCB*? Because to date, far more than 26 genes have been found implied into the formation of the spore (*spo*), and latin numbers correspond here to different stages of its development. Nevertheless, these remain easy to match in texts, while first troubles occur with anciently discovered, therefore very important, genes using only three letters, and that may be homonyms to everyday words (*map*, *gap*).

Then, bacterial genes have synonyms, and there are a lot in the case of *B. subtilis*, mainly because of their renaming. One reason is that they usually were identified first from a mutant phenotype, getting a name derived from a physiological function, and a second name was prefered latter, issued from the determination of the biochemical function. But the most important source of synonyms for this bacterium comes from a massive renaming, to align their names to those of their homologues in *E. coli*, with the worst possible situations when the new name correspond to an ancient one.

From the point of view of the genes, protein names are derived by capitalizing their first letter, such as for GerE in Fig. 2, but following this rule, the protein "sigma K" should be named "SigK". Here sigma K stands for "RNA polymerase sigma K subunit", because many authors are using alternative naming according to their biochemical functions, and we shift from proper nouns to complex terms. They are very important in our case, as the actor of an interaction may be not only a gene or a protein, but also a cellular mechanism (*SpoIIQ* is *involved* in *engulfment* of the forespore), or a variation of the environmental conditions (*YxiE* was *induced* by *phosphate* starvation), and it is essential to consider them, for connecting the molecular regulatory network to the physiology of the cell. We get there to the limits of text mining actually used for biology, as illustrated by the recently published results of the BioCreAtIvE challenge [15], where the competitors efficiently identified gene and protein names from texts on human, but failed to link them to their functional annotations, according to the categories of GO [13]. This poor result is due to the great variety by which a function may be defined in natural language, variations not covered by GO, nor by the methods used by the competitors.

Applying Linguistics, and Machine Learning Approaches. Taking an example from bacteria, some genes are activated in response to an agent like a "heat stress", that can also be evocated as a "temperature elevation", and this is calling for highly adaptative approaches, and for deep analyses of the text, both at the syntactic and semantic levels. Beside the difficulty to map the expression of the concepts into semantic categories, the sentences may have complex syntactic structures, making them difficult to understand. Also, interactions include various features, such as descriptions of genetic regulations showed in Fig. 2, physical interactions, or binding, that underly them, coregulations of genes with the bacterial operons and the regulons, or formation of protein complexes. Moreover, we are interested in other descriptions, like other functions than regulations, or mentions in the texts of homologies between genes and proteins, that in this case are highly significant, since explicitly evocated in the publications.

That is why we develop a platform based on linguistics and machine learning, that we already validated for the extraction of simple interactions [16]. Briefly, this is relying on the identification of the actors of the interactions, and on the analysis of their syntactic relationships in the sentences [17]. These linguistic informations are then combined in logical rules applied to information extraction, and the rules are learned by Propal [18], a relational learning algorithm.

References

1. Watson, J.D.: The Human Genome Project: Past, Present, and Future. Science **248** (1990) 44–49
2. Dujon, B.: The Yeast Genome Project: What Did We Learn? Trends in Genetics **12** (1996) 263–270
3. Hieter, P., Boguski, M.: Functional Genomics: It's All How You Read It. Science **278** (1997) 601–602
4. Benton, D.: Bioinformatics — Principles and Potential of a New Multidisciplinary Tool. Trends in Biotechnology **14** (1996) 261–272
5. Bryson, K., Loux, V., Bossy, R., Nicolas, P., Chaillou, S., van de Guchte, M., Penaud, S., Maguin, E., Hoebeke, M., Bessières, P., Gibrat, J.-F.: AGMIAL: Implementing an Annotation Strategy for Prokaryote Genomes as a Distributed System. Nucleic Acids Research **34** (2006) 3533–3345
6. Brazma, A., Vilo, J.: Gene Expression Data Analysis. FEBS Letters **480** (2000) 17–24
7. Huynen, M.A., Snel, B., von Mering, C., Bork, P.: Function Prediction and Protein Networks. Current Opinion in Cellular Biology **15** (2003) 191–198
8. Marcotte, E.M., Xenarios, I., Eisenberg, D.: Mining Literature for Protein-Protein Interactions. Bioinformatics **17** (2001) 359–363
9. Xenarios, I., Eisenberg, D.: Protein Interaction Databases. Current Opinion in Biotechnology **12** (2001) 334–339
10. Etzold, T., Argos, P.: SRS — an Indexing and Retrieval Tool for Flat File Data Libraries. Computer Applications in the Biosciences **9** (1993) 49–57
11. Karp, P.D.: Database Links Are a Foundation for Interoperability. Trends in Biotechnology **14** (1996) 273–279
12. Schulze-Kremer, S.: Ontologies for Molecular Biology. Pacific Symposium on Biocomputing **3** (1998) 695–706
13. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. Nature Genetics **25** (2000) 25–29
14. Hoebeke, M., Chiapello, H., Noirot, P., Bessières, P.: SPiD: a Subtilis Protein Interaction Database. Bioinformatics **17** (2001) 1209–1212
15. Hirschman, L., Yeh, A., Blaschke, C., Valencia, A.: Overview of BioCreAtivE: Critical Assessment of Information Extraction for Biology. BMC Bioinformatics **6** (2005) S1
16. Alphonse, E., Aubin, S., Bessières, P., Bisson, G., Hamon, T., Laguarrigue, S., Manine, A.-P., Nazarenko, A., Nédellec, C., Ould Abdel Vetah, M., Poibeau, T., Weissenbacher, D.: Event-Based Information Extraction for the Biomedical Domain: the Caderige Project. In: Collier, N., Ruch, P., Nazarenko, A. (eds.): Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. University of Geneva (2004) 43–49
17. Aubin, S., Nazarenko, A., Nédellec C.: Adapting a General Parser to a Sublanguage. In: Theodoulidis, B., Tsalidis, C. (eds.): Proceedings of the International Conference on Recent Advances in Natural Language Processing. Borovets (2005) 89–93
18. Alphonse, E., Rouveirol, C.: Lazy Propositionalisation for Relational Learning. In: Horn W. (ed.): Proceedings of the 14th European Conference on Artificial Intelligence. IOS Press (2000) 256–260.

Towards an Integrated Bioinformatics System: Using Integrins as Case Study to Dissect the Molecular Basis of Cell Adhesion

Sophia Kossida, Charikleia Falkou, Ioannis Vasileiou, and Christos Zervas

Foundation for Biomedical Research of the Academy of Athens
Soranou Efesiou 4, 11521 Athens, Greece

skossida@bioacademy.gr, xfalkou@biol.uoa.gr, ivasil@med.uoa.gr,
czervas@bioacademy.gr

Abstract. Organogenesis is largely relying on specific cell-cell and cell-matrix interactions that determine the formation of particular tissues. Discrete families of cell surface molecules cooperate with elements of the extracellular matrix and cytoskeleton to form the cell adhesion machinery utilized in a cell-specific manner during development. Integrins is an ancient family of adhesion molecules highly conserved among metazoans. The role of integrins in health and disease is well established, supporting the need for a global understanding of the molecular mechanisms that control precisely their functions as well as other functional related families. Our goal is to combine initial bioinformatics work on *Drosophila melanogaster* with subsequent use of the powerful tools and technologies available to identify new components of the conserved molecular machinery that function in concert with integrins in vivo.

1 Introduction

A fundamental question in biology is to understand how individual cells assemble into tissues and how the different tissues interact to form an organism.

Cell adhesion molecules are essential during embryonic development for the assembly of cells in complex patterns of specialized tissues.

Cell adhesion is controlled by modulating the binding properties of cell surface receptors and their ligands. One family of cell adhesion receptors undergoing modulation of activity is the integrins, which play a pivotal role in cell-cell contact and in interactions between cells and the extracellular matrix. The importance of integrins during embryonic development and adult life is well documented [1].

The importance of the function of integrins for a normal development, as well as their involvement in several human diseases, including tumor metastasis, thrombosis, congenital myopathies, autoimmunity and angiogenesis is well known and of a great medical interest for the development of new therapeutic approaches [2].

The integrin family of cell surface receptors is strongly conserved in metazoans, making simple invertebrate genetic systems valuable contributors to understanding integrin function. In this respect, the model organism *Drosophila melanogaster*

quickly became a paradigm for genetic studies of integrin biology. The *Drosophila melanogaster* genome contains in total seven integrin genes, while many of the integrin-associated proteins are highly conserved in flies as well [3]. A number of these genes have already been analysed in detail, however a genetic determined pathway of integrin-mediated functions is still not well described, further emphasizing the importance to identify new integrin-related genes.

2 Biological goal and technological needs

Our goal is to combine initial bioinformatics work on *Drosophila melanogaster* with subsequent use of the powerful tools and technologies available to identify new components of the conserved molecular machinery that function in concert with integrins in vivo.

The identification of new target genes co-functioning with integrins will advance the understanding of basic molecular mechanisms implicated in several of the above mentioned human diseases.

For the purpose of identifying new target genes, two main phases are presented:

2.1 Data mining and data retrieval from publicly available databases accessible on line.

A great deal of information has and will be collected. Namely, the points which were and will be addressed are the following:

a) identify the orthologs of integrins as well as their associated proteins in all the genetic model organisms (extraction of up to 2-3 levels of interactions from starting point).

Orthologous proteins are proteins with common evolutionary history in different species which quite frequently have similar functions and can be identified from high degree of primary sequence similarity (proteins with low similarity can be shown to be orthologous too, based on 3D structural motifs and alignment of them).

b) record the network of the protein-protein interactions of integrins as well as their associated proteins (extraction of up to 2-3 levels of interactions from starting point).

Currently, with the usage of several different methodologies, it has been shown that more than 60 proteins have the potential to interact with the integrin adhesion complex [4]. This number is constantly increasing given that new interactions are being identified with either biochemical or genetic methodologies. Two high throughput protein interaction experiments were carried out using the yeast 2-hybrid system[1] for the organisms *Caenorhabditis elegans* and *Drosophila melanogaster*

[1] The yeast 2 hybrid experiments study protein-protein interactions in a semi in vivo system. Large scale screen have been produced to test entire yeast genomes for instance, however comparison of lab's screen of the same yeast strain usually only overlap ~10%. Protein localization to organelles, cytoplasm or membrane should be taken into account to remove all impossible in vivo interactions. It is very difficult

covering a big part of the whole genome [5,6,7]. Recently a smaller scale similar screen was conducted in humans and generated a first draft of the human proteome network [8,9]. The comparison of all these interactions in these 3 different organisms will show us the interactions which can be considered strongly conserved, presenting in that way the central nodes (hubs) of the complex. Figure 1 below depicts an interaction network for 2 of the 7 fly integrin subunits. Similar interaction networks are under construction for the remaining integrin subunits.

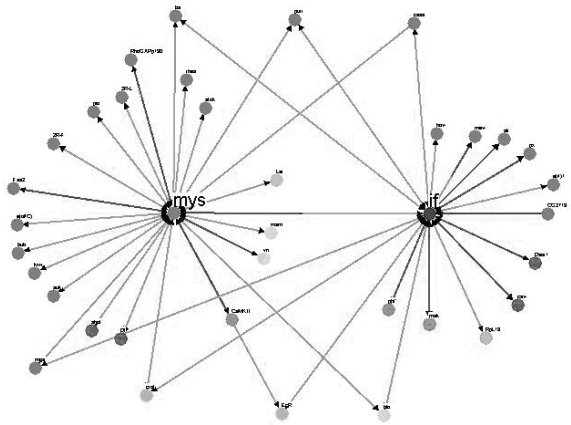


Fig. 1. Interaction network for the integrin subunits αPS2 (if) and βPS (mys) respectively. Different shades of grey indicate potential different functions of the proteins in question.

Ideally, the network of the protein-protein interactions would be reconstructed for the set of all orthologous proteins of integrins from different species.
Limiting factors for undertaking this task are the lack of high throughput analyses equivalent to the ones performed for the organisms *Caenorhabditis elegans* and *Drosophila melanogaster* as well as the fact that the experiments carried out in isolation for different organisms have not been gathered in a centralized easily queryable place.

IntAct (http://www.ebi.ac.uk/intact/index.jsp),
BIND (http://www.bind.ca/Action),
DIP (http://dip.doe-mbi.ucla.edu/),
Reactome (http://www.reactome.org/),
Cytoscape tool (http://www.cytoscape.org)

are some of the resources we have and intend to incorporate into the meta-base. It would be of tremendous help to automate the retrieval of info from the above mentioned and other sources into the currently described meta-base and similar meta-bases.

---

to use this system to look at transcription factor interactions. In conclusion this systems works great for proteins we speculate will interact and we know their sequences. It should be performed in conjunction with other interaction assays before the interactions are accepted as real.

We suggest that maybe some sophisticated text mining in connection with info deposited within the above mentioned databases could increase the sensitivity and reliability of information retrieved to serve construction of meta-bases like the one described here.

c) compare the structural organization of all identified proteins in yeast-two hybrid studies based on prediction models and correlation with known motifs, in order to find out possible similarities among proteins the sequence conservation of which is low. This information will be particular useful in the cross-species analysis of the various binding partners for a particular protein. The Simple Modular Architecture Research Tool (SMART) tool: http://pfam.cgb.ki.se/ , the protein families database of alignments and HMMs (PFAM): http://pfam.cgb.ki.se/ as well as the MyHitts: http://myhits.isb-sib.ch/cgi-bin/index could be used.

d) record the network of genetic interactions of the integrin complex from different organisms. This network will complement the network described in (a) [10].

e) extract genes with a tissue specific expression pattern during development in various organisms that resembles integrin expression pattern (flies, worms, zebrafish). Information was retrieved mainly from: http://flybase.bio.indiana.edu/, http://flymap.lab.nig.ac.jp/~dclust/getdb.html, http://www.wormbase.org and http://zfin.org .

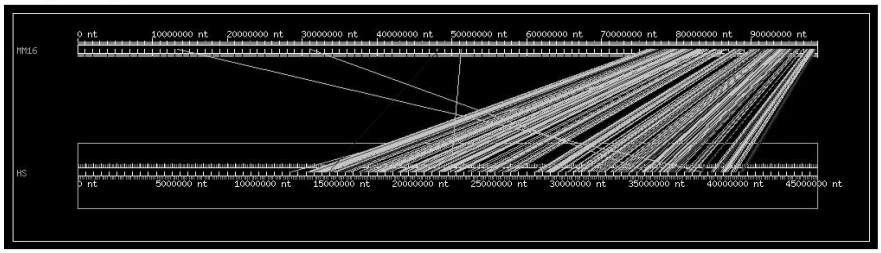
f) extract genes with a discrete sub-cellular localization either in the whole organism or at the single cell level. Two sources of information were used: http://flytrap.med.yale.edu/ and http://gfp-cdna.embl.de/ [11].

g) extract data from RNAi/morpholino screens in Drosophila, *C. elegans*, zebrafish and mammalian cells in order to associate the reported functions for a given gene among different species. One valuable source of information specific for the extraction of the Drosophila data is the FLIGHT data-base (http://flight.licr.org/).

h) study the regulatory elements of the gene targets, that is retrieval of well known regulatory sites, binding sites, promoters, enhancers as well as predicted ones. Taking advantage of the assumption that genes functioning together might be co-regulated [12], it would be of interest to have this information within the meta-base and correlate it with other sources of information from within the meta-base to guide our selection process. For instance a query which could be answered from the database: retrieve genes expressed in muscle and which have got a specific regulatory site at the 5'region, 100bp upstream of the transcription start site (TSS). Within reference 13 several tools are presented and compared whereas systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals was recently published [14] and we certainly anticipate to see more of these studies.

i) study the synteny of the potential target genes. Syntenic genes, that is genes co-localized between two chromosomes have higher probability of being co-regulated than distantly located genes. Figure 2 graphically depicts syntenic regions. It is impressive to note the degree of conservation between human chr21 versus mouse chr16 and the lack of conservation between *D. melanogaster* chr2R and *D. pseudobscura* chr3 [15].

Synteny Human Chr21 versus Mouse Chr16



Synteny *D. melanogaster* Chr2R versus *D. psedobscura* Chr3

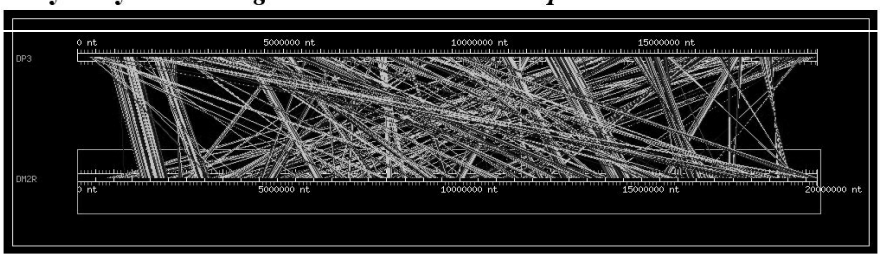


Fig. 2. Conservation of syntenic regions between human chr21 and mouse chr16. Lack of conservation between syntenic regions of *D. melanogaster* chr2R and *D. psedobscura* chr3. In the top figure a chromosomal region of human chromosome 21 is aligned with a chromosomal region of mouse chromosome 16. Lines have been drawn to link orthologous genes between these two regions. The majority of the lines are parallel to each other demonstrating in that way that the order of the ortholgous genes has been conserved. On the other hand, for the comparison of the syntenic areas of chromosome 2R of *D. melanogaster* and chromosome 3 of *D. psedobscura* we note that the order of the orthologous genes is not well conserved as the lines linking them are interconnected.

All above mentioned features within (1) will be stored within a meta-base, the purpose of which is to filter and correlate together a wealth of information. As shown in Figure 3, for each gene there are several different sources of information. The issue becomes multi-dimensional as for each one of these genes there are many orthologs available (Figure 4). A simplified scheme of the database is presented within Figure 5.

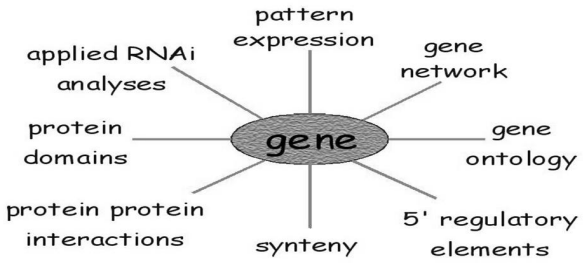


Fig. 3. A gene with several of the different pieces of info it comes with. All this gene accompanying info would ideally be deposited within the meta-base.

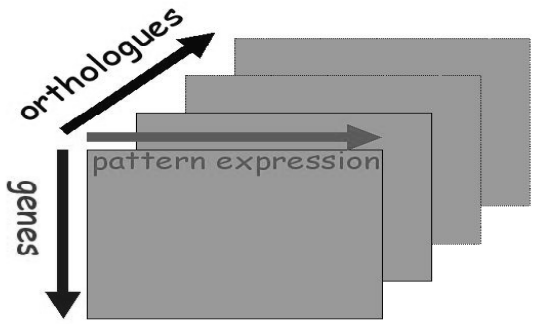


Fig. 4. For each one of the chosen *D. melanogaster* genes with their accompanying information as described within figure 3, the orthologs in various organisms will be identified, rendering in that way our system a multifactorial system. The x-axis which is shown here representing the pattern expression could equally well be any of the other info sources as given within figure 3.

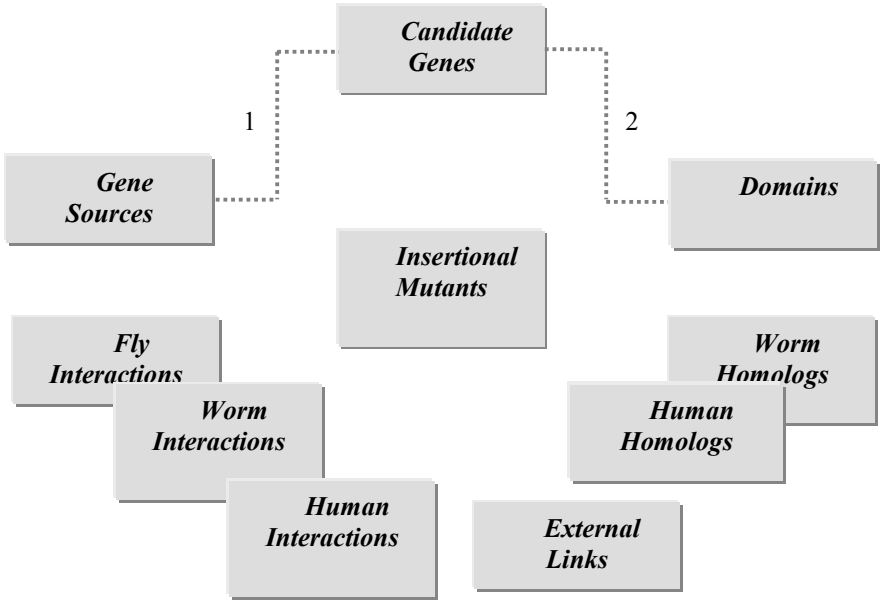


Fig. 5. Simplified scheme of the database, where (1) describes an N:M (many to many) relationship and (2) describes an (1:N) (one to many) relationship. The selection of a fly candidate gene could be based on several different gene sources and a specific gene source could be relevant to many fly candidate genes. In a later stage, the N:M relationship will turn into an 1:N relationship so each candidate gene could give rise to several proteins (isoforms).

After the meta-base has been created and put into place, queries can vary from simple ones like:

1. What are the domains for a particular gene?

2. Which genes have a particular domain?
3. What is the expression pattern for a particular gene?

and more complicated ones like:

1. Which fly genes expressed in the muscle have human and worm orthologs?
2. What are the genes expressed in all species in a particular organ or a functional related group of cells (e.g muscle, lymph gland)?
3. Which fly genes have got a specific domain, are associated with cytoskeleton, have at least 5 protein-protein interactions, and have human orthologs sharing at least 2 common protein-protein interactions with them?

It is obvious that the latter, that is the highly specific questions are the selection driven questions and hence the most informative and interesting ones. Taking into account the cost and benefit law, we have limited and focused our queries on the simpler ones, for the moment. We know several sources where from the answers to these questions could be found and we envisage three scenarios: (1) the answer is explicitly in our database although possibly not straightforward to collect, (2) the answer is not explicitly given, but it can be inferred from the info within the database (a rule-based or other deductive system could be plugged into it), or (3) the answer is neither given nor deducible, but it will have to be induced from other data present in the database, in which case the task becomes one of mining the data in the database in order to retrieve the answers.

2.2 In vivo validation of all identified target genes

Perform an in vivo functional analysis of candidate genes by RNAi [16] using the Drosophila embryo as a model system. Drosophila strains expressing various GFP-reporter genes in specific embryonic tissues will be used to monitor specific phenotypes caused by the elimination of the tested gene by RNAi using high-resolution confocal microscopy. Those genes that, when mutated, result in integrin-like phenotypes will be selected for further analysis. Experiments of functional genomics with RNAi have already been performed at the organism level (*C. elegans, D. melanogaster*). In particular in *C. elegans*, RNAi experiments have been conducted covering almost the entire genome using as a main phenotypic criterion the survival or not of the organism or focused on the early stages of embryonic development [17]. It remains to be shown however, the function of the genes in correlation with specific morphological and cellular processes [18]. Relevant, successful example is the recent functional analysis with RNAi of the genes affecting the heart [19] and the nervous system development in Drosophila [20].

3 Conclusion

There is a wealth of information out there derived from experimental work such as the genome sequencing projects, the yeast 2hybrid system assays, the RNAi assays. Text and data mining techniques are required to combine and manipulate all this information. The processed information will feed further experimental work and the process continuous. Such a procedure was described within this position paper

targeting mainly the integrin family,while it has a great potential to include novel genes that likely function in cell adhesion and tissue morphogenesis.

References

1. Hynes, R.O.: Integrins: bidirectional, allosteric signaling machines. Cell (2002), 110, 673-687.
2. Shimaoka, M., Springer, T. A.: Therapeutic antagonists and conformational regulation of integrin function. Nature Reviews Drug Discovery (2003), 2, 703-716.
3. Narasimha, M., Brown N.H.; Integrins and associated proteins in Drosophila development. In: Integrins and development Danen E, editor Landes Bioscience (2005).
4. Zamir, E., Geiger, B.: Molecular complexity and dynamics of cell-matrix adhesions. Journal of Cell Science (2001), 114, 3583-3590.
5. Giot, L. et al.: A protein interaction map of Drosophila melanogaster. Science (2003), 302, 1727-1736.
6. Stanyon, C. A., Liu, G., Mangiola, B. A., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J., Finley, R. L. Jr.: A Drosophila protein-interaction map centered on cell-cycle regulators. Genome Biology (2004), 5, R96.
7. Kafatos, F., Eisner, T.: Unification in the Century of Biology. Science (2004), 303, 540-543.
8. Stelzl, U et al.: A human protein-protein interaction network: a resource for annotating the proteome. Cell (2005), 122, 957-968.
9. Rual, JF et al.: Towards a proteome-scale map of the human protein–protein interaction network Nature (2005), 437, 1173-1178.
10. The FlyBase consortium: The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res. (2003), 31, 172-175.
11. Kelso, R. J., et al..: Flytrap, a database documenting a GFP protein-trap insertion screen in Drosophila melanogaster. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D418-20
12. Kim, S.K., et al.: A gene expression map for Caenorhabditis elegans. Science (2001), 293, 2087-2091.
13. Tompa, M. et al.: Assessing computational tools for the discovery of transcription factor binding sites Nature Biotechnology (2005) 23, 137-144
14. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M.: Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature (2005), 434, 338-345.
15. Richards, S et al..: Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. Genome Res. (2005) 15, 1-18.
16. Carthew, R.: Gene silencing by double-stranded RNA. Current Opinion in Cell Biology (2001), 13, 244-248.
17. Kamath, R., et al..: Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature (2003), 421, 231-237.
18. Friedman, A., Perrimon, N.: Genome-wide high-throughput screens in functional genomics. Current Opinion in Genetics & Development (2004), 14, 470-476.
19. Kim, Y-O., Park, S-J., Balaban, R.S., Nirenberg, M., and Kim, Y.: A functional genomic screen for cardiogenic genes using RNA interference in developing Drosophila embryos. PNAS (2004), 101, 159-164.
20. Ivanov, A.I et al.: Genes required for Drosophila nervous system development identified by RNA interference. PNAS (2004), 101, 16216-16221.

Challenges for protein family annotation

Alex Mitchell[1,2], Ioannis Selimas[1] and Teresa Attwood.[1,2]

[1]Faculty of Life Sciences & School of Computer Science,
University of Manchester, Manchester M13 9PT UK
[2]European Bioinformatics Institute, Cambridge CB10 1SD UK
mitchell@ebi.ac.uk

Abstract. In the wake of the many fruitful genome projects, tools to aid the annotation of proteomic data are sorely needed. Building from relatively simplistic approaches to an integrated system developed in collaboration with text mining experts, we have created tools capable of producing core annotation for protein families; but many challenges remain. Extending and improving these tools should have wide application, in biomedical research and beyond.

1 Introduction

As a variety of genome projects continue to bear fruit, the number of uncharacterised sequences deposited in the public domain grows ever larger. Consequently, the need for methods with which to garner information on these sequences is now pressing. PRINTS [1] is a protein 'fingerprint' database that uses conserved motifs in sequence alignments to characterise newly-determined sequences by assigning them to known families. PRINTS provides comprehensive hand-crafted annotation documenting its constituent sequence families (Fig.1*)*. Writing such annotation is laborious, time-consuming and rate-limiting for database growth, so the resource remains small by comparison with related family databases that provide no annotation.

To address this problem, we have created several assistant tools to help the annotation process, as described below. Building on these foundations, we believe that further collaborative work with text-mining experts would allow development of more sophisticated tools. This would have many potential benefits, as novel technologies that can fulfil the advanced requirements of annotators should have wide application, both in the biological sciences and in scientific research in general.

2 Towards semi-automated annotation

2.1 PRECIS

To create semi-automated fingerprint annotation, we first developed a pipeline to distil protein reports from information stored in the Swiss-Prot subsection of UniProt [2]. As each fingerprint contains links to the Swiss-Prot proteins it characterises, we could effectively exploit Swiss-Prot by recycling the annotation already produced by its curators. The pipeline, termed PRECIS (Protein Reports Engineered from Concise Information in Swiss-Prot) [3], is based on hand-crafted heuristics developed in

collaboration with PRINTS curators. It takes as input a 'raw' unannotated fingerprint, extracts the Swiss-Prot identifiers therein, and retrieves the full database entries. It then performs fingerprint "type" analysis (to determine whether it is dealing with a family, super-family or domain family). This is a crucial step, as the relationship between family members will determine the nature of the information to be recorded. The pipeline achieves this by scrutinising both the Swiss-Prot identifiers themselves, and the information extracted from certain database fields, looking for common text strings, such as "belongs to the family of X" or "contains Y domains", X and Y being the names of super-families and domains respectively.

Once the fingerprint type has been assigned, filters are applied to information culled from various Swiss-Prot fields (*e.g.*, Disease, Domain, Function, Similarity). For family fingerprints, these fields are analysed and distilled to create a report detailing the family's function, structure and associated diseases, the super-family to which it belongs and any domains it might contain. The family name, database cross-references (where further information might be found), some literature references, and keywords are also generated, together with a technical description of the fingerprint (the number of motifs it contains, the version of Swiss-Prot/TrEMBL it was scanned against, *etc.*) that mirrors the technical paragraph added by PRINTS annotators.

```
gc; VDCCGAMMA
gx; PR01792
gn; COMPOUND(3)
ga; 25-JUL-2002
gt; Voltage-dependent calcium channel gamma subunit signature
gp; PRINTS; PR01601 VDCCGAMMA1; PR01602 VDCCGAMMA2; PR01603 VDCCGAMMA4
bb;
gr; 1. BITO, H., DEISSEROTH, K. AND TSIEN, R.
gr; Ca2+-dependent regulation in neuronal gene expression.
gr; CURR.OPIN.NEUROBIOL. 7 419-429 (1997).
gr;
gr; 2. DUNLAP, K., LEUBKE, J. AND TURNER, T.
gr; Exocytotic calcium channels in mammalian central neurons.
gr; TRENDS NEUROSCI. 18 89-98 (1995).
...
gd; Voltage-dependent calcium channels are a diverse family of proteins that  encompass a
gd; variety of biological functions, including presynaptic neurotransmitter release and
gd; protein signalling within the cell [1,2]. The high voltage-activated (L-, N-, P-, Q-
gd; and R-type) channels comprise the alpha-1 subunit, which creates the pore for the
gd; import of extracellular calcium ions [2].
gd;
gd; The voltage-dependent calcium channel gamma (VDCCG) subunit family consists of at
gd; least 8 members, which share a number of common structural features [3-5]. Each member is
gd; predicted to possess 4 TM domains, with intracellular N- and C-termini. The first
gd; extracellular loop contains a highly conserved N-glycosylation site and a pair of conserved
gd; cysteine residues. The C-terminal 7 residues of VDCCG-2, -3, -4 and -8 are also conserved
gd; and  contain a consensus site for phosphorylation by cAMP and cGMP-dependent protein
gd; kinases, and a target site for binding by PDZ domain proteins [5].
gd;
gd; VDCCGAMMA is a 3-element fingerprint that provides a signature for the voltage-
gd; dependent calcium channel gamma subunit proteins. The fingerprint was derived
gd; from an initial alignment of 3 sequences: the motifs were drawn from conserved
gd; regions spanning virtually the full alignment length - motif 1 and 2 span the
gd; C-terminus of TM domain 1; and motif 3 lies partly within the second extracellular loop and
gd; partly within TM; domain 4. Three iterations on SPTR40_20f were required to reach
gd; convergence, at which point a true set comprising 28 sequences was identified.
```

Fig. 1. Manual annotation for VDCCGAMMA, a fingerprint for voltage-dependent calcium channel gamma subunits (for convenience, some literature references have been ablated).

In the case of domain- or super-family fingerprints, the same principles apply, but subsets of the Swiss-Prot fields that are used to produce family annotation are utilised, and modified weightings are applied during the information-filtering step. The output

is also formatted differently, with family-specific information being arranged into individual blocks, and generic information (such as structure, in the case of super-families) displayed in a common annotation field.

Overall, the system works well (see Fig.2). The reports it outputs are directly useful and represent good starting points for PRINTS annotators to expand into full annotation suitable for database deposition. Nevertheless it does have limitations: *e.g.*, (i) the reports are English-like, as they re-use existing annotation, but they exhibit the note-like style typical of Swiss-Prot; (ii) the approach is limited by the currentness, quality and extent of annotation available – information stored in databases often lags behind that published in the literature, meaning the PRECIS reports may not fully reflect prevailing scientific knowledge; and, if there is little or no information on a particular protein family in Swiss-Prot, the output of PRECIS will be minimal; (iii) the system performs well for protein family fingerprints, but can struggle to annotate those for domain- or super-families because Swiss-Prot stores little domain- or super-family-specific information – although we use the data it does supply (augmenting this with information on representative sub-family members, or proteins that share the same domain), this is not always directly useful to curators.

```
gc; DISHEVELLED
gx; PP00110
gn; COMPOUND(3)
ga; 01-AUG-2002
gt; Segment polarity protein dishevelled signature
gp; PFAM; PF02377 Dishevelled
gp; INTERPRO; IPR000591; IPR001158; IPR001478; IPR003351
gp; MIM; 601225; 601365; 601368; 602151
bb;
gr; 1. SEMENOV, M.V. AND SNYDER, M.
gr; Human dishevelled genes constitute a DHR-containing multigene family.
gr; GENOMICS 42 302-310 (1997).
gr;
gr; 2. BUI, T.D., BEIER, D.R., JONSSEN, M., SMITH, K., DORRINGTON, S.M.,
gr; KAKLAMANIS, L., KEARNEY, L., REGAN, R., SUSSMAN, D.J. AND HARRIS, A.L.
gr; cDNA cloning of a human dishevelled DVL-3 gene, mapping to 3q27, and
gr; expression in human breast and colon carcinomas.
gr; BIOCHEM.BIOPHYS.RES.COMMUN. 239 510-516 (1997).
bb;
gd; Function:
gd; May play a role in the signal transduction pathway mediated by multiple wnt genes.
gd;
gd; Disease:
gd; May be partly responsible for catch22 syndromes. This denomination includes
gd; developmental defects which associate cardiac defect, abnormal facies, thymic
gd; hypoplasia, cleft palate, hypocalcemia, and chromosome 22 deletions. (DVLL_HUMAN).
gd;
gd; Family and structural information:
gd; Belongs to the dsh family.
gd;
gd; Contains pdz/dhr domains.
gd;
gd; Keywords: Developmental protein; Phosphorylation; Alternative splicing.
gd;
gd; DISHEVELLED is a 3-element fingerprint that provides a signature for the segment
gd; polarity protein dishevelled. The fingerprint was derived from an initial alignment of
gd; 10 sequences: the motifs were drawn from conserved regions spanning virtually the full
gd; alignment length. Three iterations on SPTR40_20f were required to reach convergence, at
gd;  which point a true set comprising 10 sequences was identified.
```

Fig. 2. Example PRECIS output for the dishevelled protein family fingerprint.

2.2 METIS

To extend PRECIS, we attempted to mine the biomedical literature. This involved the addition of two sentence classification components capable of excising informative sentences from free text. The aim was to use PRECIS as an information retrieval (IR) element, either to supply relevant literature to the sentence classifiers directly, or to find search terms with which to seek it out. Sentences extracted by the sentence-classification step could then be used by annotators to extend the core PRECIS report.

The software, termed METIS (Multiple Extraction Techniques for Informative Sentences) [4], takes as input raw fingerprints and generates a protein report using PRECIS. In addition to the information normally collected from Swiss-Prot, METIS gathers the PubMed identifiers cited in each of the literature reference lines. The corresponding abstracts are then retrieved and passed to the sentence classifiers, which attempt to identify sentences that relate to protein structure, function and disease. Refineable PubMed query terms are also produced through analysis of the Swiss-Prot entries, employing the same heuristics used by PRECIS to determine protein family, super-family or domain names. These can be used to perform wider literature searches and the sentence classifiers can be run on the output.

The first sentence classification component is a set of Support Vector Machines (SVMs), developed in collaboration with text-mining experts. The SVMs were trained on 3 specialised corpora for structure, function and disease. Extensive sentence classification experiments were performed involving different feature representations, learning algorithms, and different SVM kernels and hyper-parameter values. The best performing models (linear SVMs with a C parameter value of 0.1) were selected.

The second classification component, BioIE [5], uses manually pre-defined templates and rules to identify sentences relating to the categories of interest. Annotators may extract all of the sentences from each category, or specify keywords to refine the extraction. The templates and user-specified keywords are marked up on the selected sentences, which are in turn ranked according to the number and type/complexity of templates found in them.

METIS is another solid step forward. Using Swiss-Prot data to supply, or to generate, search terms with which to amass relevant literature works well. The tool vastly reduces the time required to find and read suitable papers; it is versatile and easy to use; and its outputs (both from PRECIS and the sentences classifiers) are immediately useful in the annotation process. Moreover, the computer science input suggested that further development, with ‘proper’ IR and additional customisation of the sentence classifiers, would likely yield a more powerful system.

2.3 BioMinT: a collaborative approach

The next step towards generating PRINTS annotation automatically was pursued in the framework of a pan-European project called BioMinT. The goal of BioMinT was to bring together computer scientists and biologists to create curators’ assistants for the Swiss-Prot and PRINTS databases, and a generic researcher's assistant, capable of creating protein reports for biologists in academia and industry.

For PRINTS, this meant taking advantage of the computer science input to re-design our annotation pipeline and develop new elements. A fundamental change was the implementation of a specialised module to determine fingerprint type [6]. This replaced PRECIS' manually-derived heuristics with an SVM-based classifier, and brought with it a significant increase in accuracy compared to the hand-crafted rules.

METIS' pre-existing heuristics for automatically determining appropriate query terms from raw fingerprints were also refined, and several new rules were added. In addition, a bespoke document-ranking algorithm, tailored to PRINTS' needs, was developed. The sentence classifiers used in METIS were extended to recognise two further topics of information (subcellular localisation and tissue specificity). They were also retrained using additional data, and expanded from a single classifier per topic to a panel consisting of five classifiers for each. The differential bias of these classifiers towards precision and recall allows annotators to choose an appropriate recall/precision trade-off depending on their preferences, or to suit the amount or richness of information being processed.

A sentence-selection module was also added, which presents users with sentences pre-selected as relevant by the sentence classifiers, and allows them to choose which to include in their annotation. PRINTS-specific formatting of the results was also implemented, drawing on some of the techniques used in PRECIS (such as those for database cross-reference handling and generation of the technical paragraph). Finally, the whole pipeline was wrapped in an intuitive GUI for ease of use.

```
gc; MAJSPERMPROT
gx; PP00281
gn; COMPOUND(5)
ga; 22-OCT-2002
gt; Major sperm protein signature
gp; INTERPRO; IPR000535
gp; PROSITE; PS50202 MSP
gp; PFAM; PF00635 Motile_Sperm
gp; PDB; 1GRW; 3MSP; 1MSP; 2MSP
bb;
gr; 1. MANSIR A. AND JUSTINE J.L.
gr; Actin and major sperm protein in spermatids and spermatozoa of the
gr; parasitic nematode Heligmosomoides polygyrus.
gr; MOL REPROD DEV 45 332-341 (1996).
gr;
gr; 2. KLASS M., AMMONS D. AND WARD S.
gr; Conservation in the 5' flanking sequences of transcribed members of the
gr; Caenorhabditis elegans major sperm protein gene family.
gr; J MOL BIOL 199 15-22 (1988).
bb;
gd; In Caenorhabditis elegans and Ascaris suum , previous studies have reported
gd; that sperm motility does not involve actin , but , instead , requires a
gd; specific cytoskeletal protein , namely major-sperm-protein ( MSP ) [1].
gd;
gd; All MSP genes contained a consensus ribosome binding site , a consensus
gd; TATA homology 27 nucleotides distal to the site of mRNA initiation , and
gd; ten highly conserved nucleotides adjacent to the site of initiation [2].
gd;
gd; MAJSPERMPROT is a 5-element fingerprint that provides a signature for major
gd; sperm proteins. The fingerprint was derived from an initial alignment
gd; of 115 sequences: the motifs were drawn from conserved regions within the
gd; alignment. 2 iterations on SPTR40_20f were required to reach convergence, at
gd; which point a true set comprising 24 sequences was identified.
```

Fig. 3. An annotated fingerprint for the major sperm protein super-family constructed automatically using the BioMinT PRINTS annotation assistant.

The BioMinT PRINTS annotation assistant is impressive (Fig.3). The tool immediately benefits database annotators: in the same time it would take an annotator

to manually evaluate a fingerprint and devise appropriate PubMed search terms, it can take a raw fingerprint, automatically formulate and submit a PubMed query, retrieve relevant documents, extract informative sentences, and output these, along with literature and database cross-references, and so on, all in a PRINTS-specific format.

The system draws on both PRECIS and METIS, having inherited core heuristics from each, but is a clear improvement on both; it handles its information sources and the data derived from them in a more intelligent manner, seeking out the most pertinent abstracts and allowing the annotator fine control of the output. It also works well for all types of fingerprint: families, super-families and domains.

3 Further work and conclusions

BioMinT is a significant step forward in the development of annotation assistant tools, and addresses PRINTS annotators' basic needs. However, it is not a panacea. Its output contains discrete sentences grouped according to topic, but the tool doesn't combine them into a cohesive, cogent report. Moreover, the information content of the sentences is not scrutinised; hence redundant or contradictory sentences can be output by the system. Addressing these issues would bring clear benefits.

The most valuable annotation tools are those which are generically useful. For this reason we implemented BLAST-based interfaces for PRECIS and METIS, so they can take as input protein sequences rather than fingerprints; similarly, we gave BioMinT a generic entry-point so it can be used for general biological queries. With our core annotation needs addressed, our future objectives should have a much more broad-based appeal: *e.g.*, the concepts underpinning a tool capable of detecting redundancy in sentences by automatically determining their information content could be applied to a range of other tasks: these might include provenance tracing (where the percolation of facts through the literature is tracked to discover where they were first discovered and how they have spread), or evidence weighting (where corroborating and conflicting statements can be collected and examined). Similarly, a text-summarisation tool designed to distil individual pieces of information into joined-up reports would have numerous uses in diverse research areas.

Generating annotation automatically is clearly a non-trivial task, not least because biomedical text-mining is hard: there are still major obstacles to progress (because the literature is noisy and chaotic) and major people-dependent bottlenecks (because humans must train machine-learning methods and examine their outputs). Clearly, many challenges lie ahead. But our experience suggests that, although difficult to get right, the reward for stimulating meaningful dialogue between computer scientists and biologists is a collaborative environment in which different skills and perspectives can work in concert to produce practical solutions to difficult problems.

Acknowledgements

We thank the EC and EPSRC for grants QLRI-CT-2002-02770 GR/R80810/01. We thank the BioMinT consortium for their help and hard work.

References

1. Attwood, T.K., Bradley, P., Flower, D.R., *et al.,* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*. **31**, 400-402.
2. Bairoch A., Apweiler R., Wu C.H., *et al.,* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*. **33**, D154-159.
3. Mitchell, A.L., Reich, J.R. and Attwood, T.K. (2003) PRECIS - An automatic tool for generating Protein Reports Engineered from Concise Information in SWISS-PROT. *Bioinformatics*, **19**, 1664-1671.
4. Mitchell, A.L., Divoli, A., Kim, J-H, *et al.,* (2005) METIS: multiple extraction techniques for informative sentences. *Bioinformatics*, **21**, 4196-4197.
5. Divoli, A. and Attwood, T.K. (2005) BioIE: extracting informative sentences from the biomedical literature *Bioinformatics*, **21**, 2138-2139.
6. Hilario, M., Mitchell, A.L., Kim, J.-H., *et al.,*. (2004) Classifying Protein Fingerprints. *PKDD 2004*: 197-208.

Multiparameter Analysis of Cancer: How Can Data and Text Mining Help?

François Radvanyi[1], Nicolas Stransky[1], and Céline Rouveirol[2]

[1] UMR144, CNRS – Institut Curie
75248 Paris Cedex 05, France
Francois.Radvanyi@curie.fr
Nicolas.Stransky@curie.fr

[2] LRI, UMR CNRS 8623, Université Paris Sud, bat 490
91405 Orsay Cedex, France
Celine.Rouveirol@lri.fr

Abstract. Cancer research involves many aspects of biology as well as other disciplines. New techniques now make it possible to collect a huge quantity of information for the same tumour sample: for example the expression levels of almost every gene, all DNA copy number changes and the methylation status of all CpG islands. Progress is also made in the field of proteomics and it is now possible to have access to the levels of hundreds of proteins simultaneously. The analysis and interpretation of all these biological data lead to new bioinformatics challenges, such as their integration with clinical data and general/state-of-the-art biological knowledge (molecular pathways, different gene ontologies, known mutated genes, kwown oncogenes and tumour suppressor genes). As biology is a rapidly evolving field, no single source can contain all the necessary biological knowledge and it is necessary to systematically refer to the scientific literature. As a consequence, there is a need to tightly integrate structured data mining with text mining in cancer research. The final goals are to be able to predict the molecular pathways which are disrupted in a given tumour, the aggressiveness of this tumour and the response to therapy of individual patients.

Meeting the Challenge: Towards a Data and Text Mining Infrastructure for Biological Research

Alexandros Kalousis and Mélanie Hilario

CUI - University of Geneva
CH-1211 Geneva 4, Switzerland
{Alexandros.Kalousis|Melanie.Hilario}@cui.unige.ch

Extended Abstract

The goal of this position paper is to show how data mining and text mining can be incorporated into a comprehensive framework which would allow users to cope with large-scale and complex biological problems and data. To illustrate the issues involved we will give a rough schema of the scientific workflow for disease studies (Fig.1). Investigating a particular disease involves answering two distinct questions: *what* factors, e.g. genes or proteins, are related to the disease, and *how* these factors interact with other, possibly unobserved, factors.

A researcher typically starts with a thorough study of the relevant literature, then sets up an experimental design to uncover some of the biological factors implicated in the disease. This can be done using a variety of -omics technologies. Experimental results are then analysed to pinpoint the most plausible factors-biomarkers, thus answering the *what* question. At this stage, information retrieved from biological data bases and document collections can be brought to bear on and integrated with the experimental results to improve data-analytical findings and gain new insights that would otherwise go unnoticed. To unravel the underlying biological mechanisms, the identified biomarkers must be examined in a broader context to uncover interactions among them or with other factors that were not necessarily observed or measured during the experimental process. Consequently the researcher tries to reconstitute the molecular pathways governing expression of these biomarkers, in view of answering the *how* question. This work also relies heavily on access to heterogeneous sources, biological databases and literature, that deliver information about the individual biomarkers.

There is a need for a general computational infrastructure that will provide support to biologists at various phases of the research workflow. Independently of individual researchers' scientific objectives, such an infrastructure should address generic technological objectives such as :

- Provide comprehensive toolkits for the preprocessing as well as quality assurance and control of experimental data generated by multiple technologies;
- Explore and propose machine learning approaches adapted to the idiosyncrasies of genomic and proteomic data, addressing problems such as high dimensionality and reproducibility of learned models;

– Design methods for integrative knowledge discovery from multiple heterogeneous sources (structured databases, experimental data, the scientific literature and other domain knowledge sources such as ontologies), taking into account the complex representational requirements of biological processes and networks at multiple hierarchical levels (e.g., cells, tissues, organs).

Each of the above objectives has been the focus of intensive research that has already yielded significant results. The thrust toward systems biology provides an ideal testbed for efforts to integrate these diverse technological achievements.

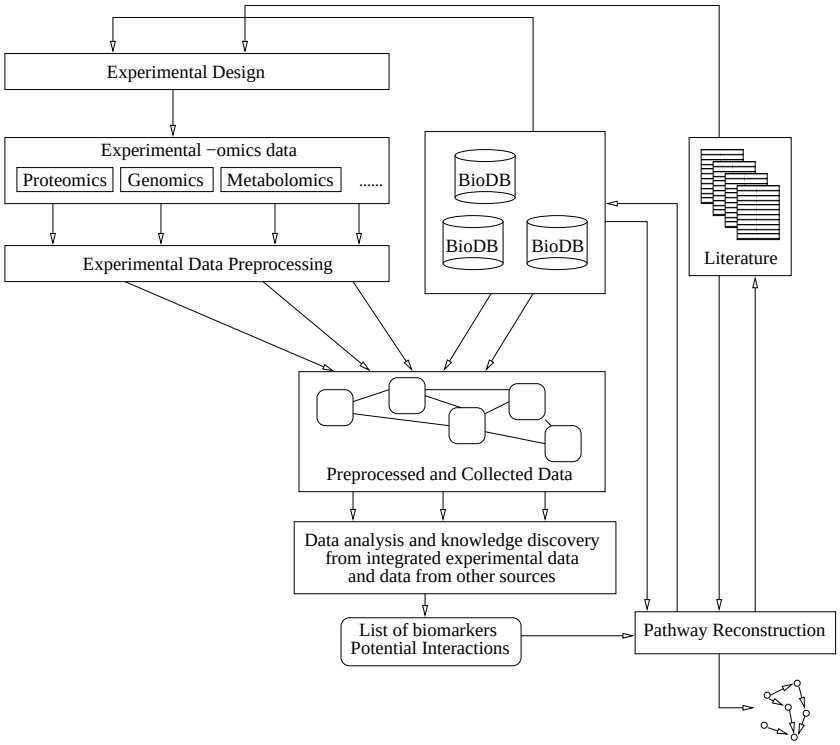


Fig. 1. Rough schema of a biological research workflow for disease study

In this presentation, we will present software developed by our team, which can be incorporated into the data and text mining infrastructure described above. This includes: an ontology-driven information retrieval system, an ILP-based information extraction system that generates case frames in the absence of predefined templates and pre-annotated corpora, a comprehensive toolkit for preprocessing and classifying protein mass spectra, and a set of kernel-based relational learning tools tailored to complex biological data structures such as sequences, trees, and graphs.