Fitting the finite-state automata platform for mining gene functions from biological scientific literature

Inès Jilani **Natalia Grabar** **Marie-Christine Jaulent**
Université Paris Descartes, Faculté de Médecine, Inserm U729, Paris 75006 France
{first.second}@spim.jussieu.fr

Abstract

The large amount of information which has to be processed in the area of biology requires automated tools. Our work aims at creating a platform for the acquisition of gene functions from biological documents. The first step of our work consists in fitting an existing data-processing platform to this purpose.

1 Introduction

When biologists work with micro-arrays or compare gene sequences the most difficult and tedious task consists in the validation and the interpretation of the obtained results. Repressed or co-expressed genes distinguished with micro-arrays, or genes with similar sequences, have thus to be annotated by their known functions, these obtained results have to be compared with those described previously by other researchers. Biologists can then consult existing species-specific databases which annotate genes and gene products with functional terms. But the information contained in these databases is often not complete and does not correspond to the current state of the art in biology. Biologists must then consult the bibliographic databases, like *PubMed*, to be able to locate more recent and exhaustive literature about genes they are researching on. Working with hundreds and even thousands genes makes it difficult to analyze all the available literature manually. Even for the annotation of few genes a lot of time is required. For instance, generating 719 gene clusters on the basis of micro-arrays and sequences comparison between two species, as described in [Lefebvre et al.2005], needed about one month, while for the interpretation and validation of few clusters researchers spent more than four months. Automated methods designed for the analysis of available literature are then highly suitable. Such analysis should provide searched information to biologists and complete knowledge already registered into databases.

In the present work we aim at developing automated methods for the extraction of gene functions from scientific literature and for the functional annotation of genes and their products. Different methods can be used to achieve this task. [Blaschke et al.2001] have designed GEISHA system in which gene groups (for instance, clusters of co-expressed genes) are used as the framework for clustering the literature corresponding to each of the genes in the group. The meaningfulness of information is weighted by the significance of its frequency and specificity for each group of genes. Thus the significant terms extracted allow relating genes to its specific functions. A fuzzy relation system is proposed by [Perez-Iratxeta et al.2002] and allows associating genes with genetically inherited diseases. In this approach, the relationship between terms is strong (pathology condition and chemical terms, chemical terms and genes…) if they occur together in many abstracts. [Grabar et al.2005] apply the likelihood ratio algorithm to the co-occurring genes and *Gene Ontology* terms. This allows annotating genes functionally by associating them with *Gene Ontology* terms. The report on BioCreAtIvE contest [Blaschke et al.2005] presents some other methods used for mining gene functions from biomedical literature. During this contest, participants have applied information content of terms, manually crafted regular expressions, statistical terms frequencies, heuristic rules, machine learning algorithms, pattern matching and template extraction. The last approach, which consists in the application of pattern matching and template [Chiang and Yu2003] has shown the best precision although the recall has been very low. Patterns used in this system have been acquired automatically with learning algorithm. In the previous contest on text mining from biological texts, KDD Cup 2002 [Yeh et al.2003], information extraction patterns have as well shown a remarkable efficiency: actually, the winning system

[Regev et al.2002] used manually defined patterns for the detection of both gene names and experimental results.

2 Objectives

The fact that the patterns appear to be efficient for mining biological literature encourages us to use such patterns for the extraction of functional annotation of genes. The manual creation of them being long and tedious task, we apply automatic method for their acquisition. This method is inspired from the work described in [Morin 1999] where the author acquires patterns for the detection of hierarchical relations between terms. We suppose that the same approach can be used for the acquisition of more specific relations, i.e. relations between genes and their functions. This method is bootstrapped from raw text, already projected and recognized gene names and functional terms. This method requires, particularly, suitable data-processing platform. In this paper, we describe the fitting of finite-state automata platform Unitex for the extraction of functional information about genes. We describe first the material used (sec. 3) and methods (sec. 4). We present the obtained results (sec. 5) and then conclude (sec. 6).

3 Material

Our available material is consisted of gene names and their aliases, functional terms and their synonyms, and textual corpora of scientific paper abstracts collected from *PubMed*. The textual corpora allow acquiring patterns and then relating genes with their functions. This task is done through the Unitex system[1] .

Unitex consists of several programs (written in the language C and easily handled by a JAVA interface), one for each crucial state of a text mining pattern matching: the loaded corpus is first preprocessed by separating all sentences, then cut into tokens. After that the software applies an English dictionary to recognize POS categories (nouns, verbs, adjectives, etc) and to tag syntactically the corpus. Furthermore it offers the possibility to add and apply the user specific dictionaries, what we did creating resources suitable for the functional annotation of genes. This functionality was crucial for the use of Unitex according to our purposes.

We created three dictionaries: (1) Species names, i.e. *D. melanogaster*, *C. elegans* and their aliases. (2) Gene names: we have 3950 genes names (1925 for each species) grouped into 719 clusters, according to the previous biological experiments results [Lefebvre et al., 2005]. In order to make the text mining processing more complete, we extracted aliases of these genes (3950) from corresponding databases: *WormBase* [Stein et al.2001] and *FlyBase* [FlyBase Consortium 1994]. Among these 3950 genes, 24 are known in *WormBase* and 1023 in *FlyBase*, but only 4 (*C. elegans* genes) and 1009 (*D. melanogaster)* of them receive aliases. (3) Terms from *Gene Ontology* [Gene Ontology Consortium, 2000] for the description of functional profiles of genes. *Gene Ontology* provides 18,696 terms distributed into three separate hierarchies: molecular functions (MF), biological processes (BP) and cellular components (CC). About 20% of *Gene Ontology* terms, which are related to existent UMLS [NLM, 2005] concepts, have been completed with their synonyms.

We also use textual corpora built with abstracts of scientific literature from *PubMed*[2]. In this work, we analyze corpora of abstracts collected with keywords relative to gene names (and their aliases) and species names. These corpora consist of 5627 abstracts for *D. melanogaster* and 2113 for *C. elegans*.

4 Methods

In order to the fit Unitex both for the extraction of patterns and functional annotation of genes, we performed some modifications of the system and its interface. Notice that the same interface will be used for the integration of results obtained with other methods, like [Grabar et al.2005], and for the presentation of all obtained results to biologists during the validation step.

Unitex components are used as a sub-layer in a new interface which design is more user-friendly and suitable for the functional annotation of genes: the searched knowledge is highlighted. To ease the navigation in the text, different possibilities are proposed.

When working with Unitex we benefit of grammars (not exactly algebraic grammars because they integrate the concept of *transduction* which is a concept borrowed from the finite-state automata, meaning that a grammar can produce outputs) also called RTN (Recursive Transitions Network) which can be easily depicted graphically. The graphs we have constructed are formalisms very close to that of the finite-state automata used to implement Regular Expressions

[1] *http://www-igm.univ-mlv.fr/~unitex/*

[2] *www.ncbi.nlm.nih.gov/entrez*

defined by the RegExp grammar[3]. This allows us to use the domain specific dictionaries we have added to the system and to build our own graphs close to the kind of literature we are dealing with. Some of them, which correspond to functional patterns, are presented in the next section.

However the application and matching of patterns in a raw text can reach the correct knowledge about genes as well as incorrect one. *In Drosophila melanogaster the yolk protein (YP) genes are normally expressed only in the fat body and follicular epithelium of adult females* is a sentence obtained when applying one of our patterns concerning cellular components. This information appears to be reliable enough, whereas the sentence *A single dominant gene (H30) seems to determine the Hessian fly resistance in this introgression line* has a lower reliability because of the use of the verb *seem*. Finally the traditional patterns applied to the sentence *Interestingly, while previous studies have concluded that ectopic expression of the homeotic genes Dfd, Scr and Antp has no effect on the segmental identity of the abdominal segments, our results demonstrate that this is not true* would detect that *ectopic expression of the homeotic genes Dfd, Scr and Antp has no effect on the segmental identity of the abdominal segments* without any reference to the end of this sentence *it is not true*. To make the distinction between these different cases contextual information is required. In our work, we rely on markers of confidence such as verbs, adverbs, negation or adjectives. Thus the application of patterns is reinforced by the defined confidence markers. The patterns we are currently defining would have correctly dealt with this nuance.

5 Results and Discussion

The creation and addition of domain specific dictionaries, as described in the section 3, was the first step in fitting Unitex to our purposes. This allows first recognizing relevant biological entities (gene names, terms, etc.) and using them in functional patterns.

All programs of Unitex process have been hidden by a user friendly JAVA interface in order to have only the last top layer i.e. an interface where only the extracted desired knowledge appears underlined.

The interface (Fig. 1) displays the relevant extracted knowledge and excerpts of corpus according to the patterns instantiated with the genes chose by the user. In the right upper part, a list of all matched units from the corpus is displayed according to the patterns; the user can click on a result and the abstract (with the searched terms highlighted) from where it is extracted appears in the right bottom part. The left column of the interface is designed for navigation and selection of knowledge materials (genes, species, *Gene Ontology* terms…).

The use of tool such as Unitex in addition with our own interface gives various possibilities and a huge sphere of activity. For instance, for the acquisition of functional patterns we can use the following automata *<G><Verb><Prep>*<Det>*<BP>*, where *G* stands for gene names, *BP* for biological process (both contained in our dictionaries), and *Verb*, *Prep* and *DET* result from the POS tagging. This allows extracting the following sentence, among 1042 results: *Cad99C contributes to eggshell formation and female fertility and is expressed in follicle cells, which produce the eggshells*. It also allows relating indicated gene (*Cad99C)* to its function (*eggshell formation*) and then to store this knowledge discovered.

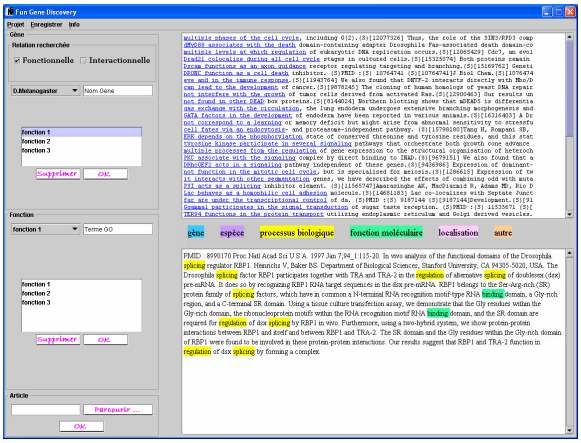

Figure 1: the interface for the extraction of gene functions and visualization of results

The possibility of adding and using such precise categories as *<G>* or <BP> or *<MF>* (Molecular Function) or <CC> (Cellular Component) allows extracting precise information about genes. In contrast, the application of biologically less specific pattern like *<Noun><Verb><Prep>*<Det>*<Noun>* where the first *<Noun>* is expected to match the gene name and the second *<Noun>* is expected to match functional term, gives rather noisy results: over 93,000 pattern matching. Up to now, we have designed a list of 20 patterns: 14 of them are able to retrieve information related to the functions of genes, and 6 interactions inter genes.

[3] *http://nltk.sourceforge.net/lite/doc/en/regexps.html*

The ongoing work concerns the acquisition of more patterns which rely on other POS categories, such as nouns and deverbal nouns.

As we have indicated the application of patterns requires often indications about the confidence of information they allow extracting. For this, we use contextual markers of their reliability. These confidence markers are mostly of the form of a negation (*no*, *not*, etc.) followed by an adverb (*necessarily*, *always*, *absolutely*, *completely*, *exclusively*, *generally*, etc.). Other adverbs may occur without negation (*seldom*, *never*, *always*, *certainly*, etc.). Sometimes adjectives (*sure*) or verbs (*induce*, *seem*, etc.) can also give indications about the confidence of putative functional annotation. However it is important to recall that all markers can be organized around the same axis, and that their confidence can either be positive or negative. These markers will be used for the definition of confidence score of functional annotation. A score under which the information extracted is not warranted, and over which it is sure so that the biologist can choose whether to take into account the information or not. In the further steps of the work, the user will be able to use the interface to ask for a gene name. All the known functional information about this gene will be then displayed. This information can be provided by functional patterns, by other text mining approaches or by annotations in existing species-specific databases.

6 Conclusion and Perspectives

The first results of our work are encouraging, many things still able to be improved especially the ambiguity and synonymy of gene names and of terms have to be handled. Indeed nowadays various genes naming remains problematic as they can change from a researcher to another or receive the same synonyms.

The ability of combining several knowledge extraction methods (likelihood ratio, association rules, syntactical analysis and lexico-syntactic patterns) is interesting to raise to the maximum the ability and reliability of text mining system.

Moreover designing a generic methodology to apply the system to other species as far as we have the primary information (gene names, biological process and molecular functions) is seen by most biologists to be a logical and necessary continuation to our project.

References

Christian Blaschke, Juan~C. Oliveros, and Alfonso Valencia. 2001. Mining functional information associated with expression arrays. *Functional & Integrative Genomics*, 1(4):256--268.

Christian Blaschke, Eduardo~Andres Leon, Martin Krallinger, and Alfonso Valencia. 2005. Evaluation of biocreative assessment of task 2. *BMC Bioinformatics* , 6(Suppl 1):S16.

Jung-Hsien Chiang and Hsu-Chun Yu. 2003. MeKE: discovering the functions of gene products from biomedical literature vis sentence alignment. *Bioinformatics*, 19(11):1417--1422.

FlyBase Consortium. 1994. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Research*, 22(17):3456--3458.

Gene Ontology Consortium . 2000. Gene Ontology : tool for the unification of biology. *Nature genetics*, 25:25--29.

Natalia Grabar, Magali Sillam, Marie-Christine Jaulent, Céline Lefebvre,Edouard Henrion, and Christian Néri. 2005. From likelihoodness between words to the finding of functional profile for ortholog genes. In *RANLP 2005 WS on Text Mining Research, Practice and Opportunities*, Borovets, Bulgaria.

Céline Lefebvre, Jean-Christophe Aude, Eric Clément, and Christian Néri. 2005. Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates. *Bioinformatics*, 21(8):1550--1558.

Emmanuel Morin. 1999. Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues (TAL)*, 40(1):143--166.

NLM, 2005. *UMLS Knowledge Sources Manual*. National Library of Medicine, Bethesda, Maryland. www.nlm.nih.gov/research/umls/.

Carolina Perez-Iratxeta, Peer Bork, and Miguel~A. Andrade. 2002. Association of genes to genetically inherited diseases using data mining. *Nature Gene*tics, 31:316--319.

Y Regev, M Findelstein-Landau, and R Feldman. 2002. Rule-based extraction of experimental evidence in the biomedical domain. the KDD Cup 2002 (task 1). In *SIGKDD Explorations newsletter*, pages 90--92.

Lincoln Stein, Paul Sternberg, Richard Durbin, Jean Thierry-Mieg, and John Spieth. 2001. WormBase: network access to the genome and biology of *caenorhabditis elegans*. *Nucleic Acids Research*, 29(1):82--86.

Alexander S Yeh, Lynette Hirschman, and Alexander A Morgan. 2003. Evaluation of text data mining for database curation: lessons learned from the kdd challenge cup. *Bioinformatics*, 19(Suppl):i331--i339.