Gene Ontology (GO) Annotation in Biomedical Literature

Carmen Galvez[1]

cgalvez@ugr.es

[1] University of Granada, 18071, Granada, Spain

Abstract: In this paper, we propose an approach for doing Gene Ontology (GO) annotation on biomedical texts. The GO is an effort to create a controlled terminology for labelling gene functions in a more precise. Our system is based on the application of Parametrized Finite-State Graphs (P-FSG) for GO tagging. This process was implemented to the annotation of genes related with *Alzehimer disease*. This prototype is an undergoing work, in the future should be evaluated to verify its value.

Keywords: Gene Ontology Annotation (GOA); Biomedical Document Processing; Terminology

1. Introduction

In the post-genomic era, one of the major activities in molecular biology and biomedicine is to determine the precise functions of individual genes, which has been producing a large number of publications. To structure the information related to gene functions scattered over the literature, a great deal of efforts has been made to annotate articles by using the Gene Ontology (GO) terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2001). Ontologies describe and formalize the terminology and knowledge common to a domain.The GO project is a collaborative effort to address two aspects of information integration: providing consistent descriptors for gene products in different databases and standardizing classifications for sequences. The integration of genomic information is a valuable research issue. The project has three purposes: (*i*) to develop a set of controlled, structured vocabularies, or ontologies, to describe key domains of molecular biology, including gene product attributes and biological sequences; (*ii*) to apply GO terms in the annotation of sequences, genes or gene products in biological databases; and (*iii*) to provide a centralized public resource allowing universal access to the ontologies, annotation data sets and software tools developed for use with GO data (The Gene Ontology Consortium, 2001).

The GO project began as collaboration between different model organism databases, such as *FlyBase*, *Saccharomyces Genome Database* (SGD), and the *Mouse Genome Informatics* (MGI) Database. In GO, there are three kinds of structured controlled vocabularies, or sub-ontologies, to describe three semantic types of concepts, including Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). The sub-ontologies represents different categories of genomic characteristics which are described by GO terms. Each GO term is associated with a GO ID. The GO holds functional gene annotation hierarchically in a directed acyclic graph (DAG) structure that reflects the relationship between the biological terms and associated genes, and the GO database can be pictured as a tree structure, where nodes (or terms) are more general if closer to the root and more specific if closer to the leaf. Currently, the GO vocabulary consists of >13,000 terms. For example, a GO term *'molecular_function'* has the GO ID of *GO:0003674*. GO annotation is a major activity in most model organism databases projects and annotates genes functions using a controlled vocabulary.

Because of the large number of publications, and biological and medical data stored in different databases, annotating genes with these controlled vocabulary codes is a labor-intensive task. An expert inspects the literature associated with each gene to determine the appropriate function code. GO annotation requires extensive human efforts and domain knowledge, which is usually conducted by experts. Thus, there is a potential need to automate or semi-automate GO annotation, which could greatly alleviate the human curation. This was one of the primary objectives pursued at the Text Retrieval Conference (TREC) 2004 Genomics Track (Hersh et al., 2005). The emergence of powerful methods for analyzing text arise the possibility that gene annotation can be facilitated using Natural Language Processing (NLP) techniques. This paper focuses on an undergoing work for doing GO annotation in biomedical abstracts, based on association of gene names (in text) and GO terms (in ontology). Although the main problem of mapping is the term variation and term ambiguity, these issues are beyond the scope of the present paper.

2. Material and Methods

The system prototype presented involves NLP and relies on finite-state graphs compiled in transducers. The methodology presented here involves three stages: (1) Collection of gene terms using a biological database; (2) encoding of gene-naming terms in a binary matrix; (3) design of a *Parametrized Finite-State Graph* (P-FSG) that formalizes the gene name variants encoded in the binary matrix; and (4) annotation of gene names with GO terms.

We assembled publicly available gene information from the *Entrez Gene* database. The dataset is limited to the organism '*Homo sapiens'* and genes related to the '*Alzheimer Disease'*. For example, the search provided the following information for the gene name 'PSEN1': *Official Symbol*; *Other Aliases*; *Other Designations*; *Chromosome*; *GeneID*. On the other hand, the system was tested on a total of 1,543 records downloaded from *MEDLINE* searches supported by the *PubMed®* interface*;* the search was benchmarked to records with '*Alzheimer Disease*' and '*Humans*' in the Medical Subject Heading (MeSH) field.

With the material supplied by Entrez Gene and PubMed, we propose a procedure that permit to map gene names into GO terms or codes. The proposal is based on the application of an extension of finite automata, or mathematical models. Thus, Finite-State Transducers (FSTs) are of a system with **input** and **output**, and can be defined as a finite set of states and a set of transitions from one state to another. Transducers define **relations between languages**. To compute the relations, a transducer has transitions labeled with two symbols from two alphabets: input and output. Formally, an FST is referred to as a *5-tuple* quintuple (Roche & Schabes, 1997) expressed as:

$$FST = (\Sigma, Q, q_o, F, \delta)$$

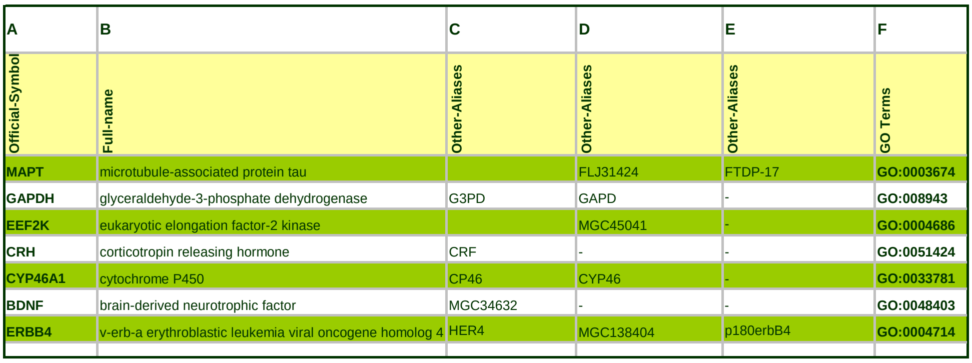
where Σ is the **input and output alphabet**, Q is a **finite set of states**, q_o is the **initial state**, $q_o \in Q$, F is the **final state**, $F \subseteq Q$, and δ is the **number of transitions between states**. FSTs can be represented as directed graphs, whose vertices denote states, while the transitions form the edges, or arcs, with arrows pointing from the initial state to the final state.

Using a graphic interface, *FSGraph* (Silberztein, 2000), we drew finite-state graphs that would represent the possible gene names, as input, and could produce as output the GO term. As the transducers are utilized utilized to represent relations of equivalence among languages, we consider that the problem of the annotation of terms would be able to resolve if was presented like a relation that maps gene names to concepts in ontology. To establish this relation we propose the utilization of Parametrized Finite-State Graphs (P-FSG), defined as graphics of state-finite compiled in FSTs, whose alphabet of input and output contains parameters with the values stored previously in a binary matrix.

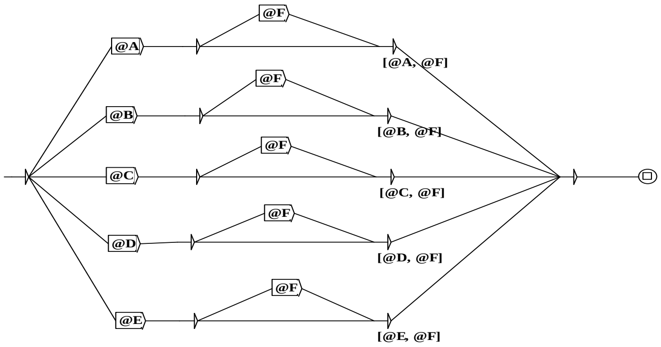
3. Results

On the biomedical abstracts downloaded from *MEDLINE*, we apply the parametrized graphic according to the data stored in the binary matrix (Table 1).Using the graphic application based on finite-state techniques, we draw a network of transition compiled in a transducer with parameters (Fig. 1), that are bound to the content of the matrix through variables (@A, @B, @C,..). The process of annotation would be carried out through replacement of variables, having as input the variables that refer to gene terms, and as output the variable that refer to GO terms. The application allows for the graph´s transformation into a table or transition matrix where the following components are specified: **number of states**, Q = 22; **number of alphabet symbols**, or vocabulary, $\Sigma = 13$, where the symbol <E> represents the empty string; **initial state**, $q_o = 0$; **final state**, $F = 1$; **number of transitions between states**, $\delta = 26$, where each transition is defined by a 3-tuple: *current state, symbol, outgoing state*.

Table 1 – Binary matrix



A	B	C	D	E	F
Official-Symbol	Full-name	Other-Aliases	Other-Aliases	Other-Aliases	GO Terms
MAPT	microtubule-associated protein tau		FLJ31424	FTDP-17	GO:0003674
GAPDH	glyceraldehyde-3-phosphate dehydrogenase	G3PD	GAPD	-	GO:008943
EEF2K	eukaryotic elongation factor-2 kinase		MGC45041	-	GO:0004686
CRH	corticotropin releasing hormone	CRF	-	-	GO:0051424
CYP46A1	cytochrome P450	CP46	CYP46	-	GO:0033781
BDNF	brain-derived neurotrophic factor	MGC34632	-	-	GO:0048403
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4	HER4	MGC138404	p180erbB4	GO:0004714



GO Annotation.grf
Fri Feb 29 13:54:23 2008

Figure 1 – Parametrized finite-state graph

The GOA has been verified in all abstracts obtained of the MEDLINE database (Table 2).

Table 2 – Mapping gene names to GO terms

Decrease of dehydrogenase activity of cerebral glyceraldehyde-3-phosphate dehydrogenase in different animal models of Alzheimer's disease.

Shalova,-I-N; Cechalova,-K; Rehakova,-Z; Dimitrova,-P; Ognibene,-E; Caprioli,-A; Schmalhausen,-E-V; Muronetz,-V-I; Saso,-L

Biochim-Biophys-Acta. 2007 May; 1770(5): 826-32

Recently, a relationship between glyceraldehyde-3-phosphate dehydrogenase [**GAPDH, GO:008943]** and the beta-amyloid precursor protein (betaAPP) in relationship with the pathogenesis of Alzheimer's disease (AD) has been suggested. Therefore, we studied the specific activity of [**GAPDH, GO:008943]** in the different animal models of AD: transgenic mice (Tg2576) and rats treated with beta-amyloid, or thiorphan, or lipopolysaccharides (LPS) and interferon gamma (INFgamma) [...].

4. Conclusions

This paper proposes the application of parametrized graphs for mapping genes names, in biomedical texts, to GO terms or codes. This approach is an undergoing work, in the near future should be enhanced and evaluated to verify its real usefulness.

References

Ashburner, M. et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25, 25-29.

Gene Ontology Consortium (2001). Creating the Gene Ontology Resource: Design and Implementation. *Genome Research*, *11*, 1425-1433.

Hersh, W. et al. (2005). Evaluation of Biomedical Text-Mining Systems: Lessons Learned from Information Retrieval. *Briefings in Bioinformatics, 6* (4,), 344-356.

Roche, E & Schabes, Y. (1997). Finite State Language Processing. Cambridge, Massachusetts: MIT Press.

Silberztein, M. (2000). INTEX: An FST Toolbox. *Theoretical Computer Science, 231*, 33-46.