KASKAD : a plat-form to extract temporal and interaction relations for genes in texts

Nicolas Turenne, Branislav Meszaros

International Workshop on NanoBioTechnology, Saint-Petersburg 2006

KASKAD : a plat-form to extract temporal and interaction relations for genes in texts.

Nicolas Turenne, Branislav Meszaros
INRA MIG 78352 Jouy-en-Josas, France
nicolas.turenne@jouy.inra.fr

Abstract
The new era of genome sequencing allows systematic identification of genes for any organism. But such biomolecular approach lacks of holistic knowledge about the gene function and its activation over time. As an integrative knowledge grabber, a tool scanning biomedical literature can bring pieces of information to understand part of gene networks over time. The KASKAD system tries to answer this question.

Introduction
Nowadays two main approaches attempts to extract knowledge in an integrative way: microarrays and biomedical literature analysis. The first one uses intensively data analysis and system biology models. The second one uses also data analysis and natural language processing. The advantage of the second approach is strongly related to compilation of facts by biologists over years. Of course uncertainty can surround these facts especially for old documents. And methods used for scanning collections of documents have to take it into account. We chose a "weak analysis" with co-occurrences of expressions in abstracts of documents to avoid constrained grammar dependencies difficult to manage (Blaschke & Valencia, 2002). Nevertheless as far as vocabulary used by authors is sensitive to variation of terms our method is based not only on term co-occurrences but co-occurrences of concepts defined by regular expressions. In our case a gene and a stage of development will be considered as a concept and associated to a regular expression.

Method
Our goal is to develop a corpus analysis tool which aims to extract contextual knowledge (i.e. temporal) about genes.
Firstly we have driven our study with a well know model in developmental biology, the coat formation in sporulation of Bacillus Subtilis species. This process implies around 5 regulators among 51 genes, and it occurs in stage 3, 4 and 5 of sporulation described into 7 stages. We have manually found all interactions (133) and their occurrence during the stages we have formalized into a final representation (Turenne & Schwer, 2006).
Secondly our tool would have to model the variation of terminology which is the main problem on which the studies leading to ambiguity solving focuse. We compiled a dictionary of 350 temporal markers and we have tagged a working corpus (1430 documents with titles and abstracts) to localize all temporal phrases. Hence we have associated a number of expressions to each stage of our biological frame (i.e. ontology of the domain). Ideally we could imagine a basic string matching co-occurrence between a gene name such as sigma K and stage 1. But as some of the expressions show in our corpus it is not efficient: *at about t1.5, at different times (t-1, to, t1, and t2), attaining maximum expression at t1 and t3, beginning at t0.5-1 of sporulation, between 0 and 3 h after the beginning of sporulation (t0 to t3), between T0 and T2 and decreased after T3, even at T1.5, stage I, The first stage*. If we imagine a regular pattern which symbolize the concept "stage 1" we can better reach our goal of association. Such pattern should be defined as follows: S = *between T0 and T[2-9]* | *t0 to t[2-9]* | *t1** | *T1** | *stage I* | *first stage*. And should gather all occurrences of previous cited expressions. We can do the same for the concept representing a gene. For instance the pattern : G = *sigK* | *SigK* | *sigma 27* | *sigma K* | *Sigma K*, should represent the concept "sigma K".

After crossing the pattern G and S we obtain the co-occurrence of the concept “sigma K” with the concept “stage 1”. In the same way we obtain the co-occurrence between 2 gene concepts.

Tool

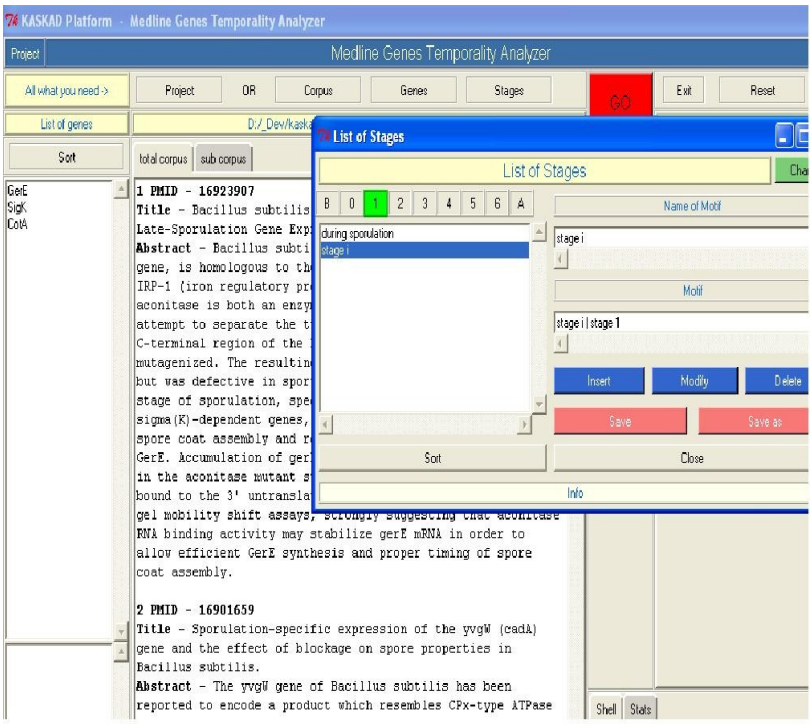

Fig.1 General Interface and window of stage patterns definition.

The program has been written in Perl / Tk and is freely available on the web at **http://genome.jouy.inra.fr /~turenne/kaskad_download.html**. Main starting features are selection of the corpus (online from medline, or an existing one), selection of a gene list with their patterns, selection of a stage list with their patterns. Patterns has to be defined by the user. The three files can be associated and saved into a project (see figure 1)

Main processing features are *gene-gene* matrix and *gene-stage* matrix computing executed with the same command. At present only strict pattern / pattern co-occurrence is processed without any use of extern linguistic resource for context analysis. Results are displayed on a notebook with two TabSheets (see figure 2). On the *gene-gene* matrix, cells mean numbers of documents in which co-occur the line and column labels (i.e. more exactly their associated patterns). About the *gene-stage* matrix, computing is the same but cells show if a gene (pointed out at a given line) occurs or not at a stage (pointed out by a given column. The user can also display documents in which information has been found by clicking on a cell.

Genes\Genes	CotA	GerE	SigK
CotA	5	3	2
GerE	0	12	3
SigK	0	0	16

Genes\Stages	-b-	-0-	-1-	-2-	-3-	-4-	-5-	-6-	-a-
CotA	0	1	1	0	0	0	0	0	0
GerE	0	1	1	0	0	0	0	0	0
SigK	0	1	1	0	0	0	0	0	0

Fig.2 Display of results by KASKAD (left, *gene-gene* matrix; right, *gene-stage* matrix).

We need to make evaluation of extraction with standard parameters (precision and recall), and to study the co-occurrence property (same sentence, same abstract, around a given verb).

Acknowledgments

This work was supported by the INRA-1077-SE grant from the French Institute for Agricultural Research (agriculture, food & nutrition, environment and basic biology).

References

Turenne N., Schwer S.R. (2006) “Temporal Representation for Gene Networks: towards a Qualitative Data Mining”, to appear in *International Journal of Data Mining and Bioinformatics*.

Blaschke C., Hirschman L., and Valencia A. (2002) “'Information extraction in molecular biology“, *Brief. Bioinform*. 3, 154-165.