# SALAMBOMINER: A BIOMEDICAL LITERATURE MINING TOOL FOR INFERRING THE GENETICS OF COMPLEX DISEASES

Leonor Rib
*Unit of Genomics of Complex Diseases, Institut de Recerca del HSCSP, Barcelona, Spain*
*mrib@santpau.cat*

Ricard Gavaldà
*Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain*
*gavalda@lsi.upc.edu*

Jose Manuel Soria, Alfonso Buil
*Unit of Genomics of Complex Diseases, Institut de Recerca del HSCSP, Barcelona, Spain*
*jsoria@santpau.cat, abuil@santpau.cat*

Keywords: Literature mining. Knowledge discovery. Bayesian Networks.

Abstract: We present SalamboMiner, a tool designed to discover relationships among genes, proteins and diseases from abstracts in the Pubmed database. SalamboMiner identifies relevant concepts included in biomedical articles by means of Biological Entities Recognition. The co-citation of these concepts gives a significance of the relationship between pairs of concepts. Moreover, by using a Bayesian network we can infer relationships between those concepts that are not co-cited.

## 1 INTRODUCTION

Databases of biomedical literature are information mines that are hard to excavate manually, due to their large dimensions and the variety of information that they store. Datamining and Machine Learning methods allow us to access these databases in an automated manner and to extract relationships between biomedical concepts. Here we present a method for extracting relationships between genes and diseases from the articles contained in the Pubmed database.

The co-citation of concepts in the same text has become a useful tool to look for the role of genetics in complex diseases (Weeber, 2005) (Zweigenbaum, 2007). The main idea of our strategy is that the co-citation of concepts in the same text gives information related to the degree of relationship between themselves.

## 2 METHODS

SalamboMiner has three consecutive modules:

- the first aim is to collect the desired articles,
- the second aim is to extract the desired concepts from the articles, and
- the third aim is to organize the extracted concepts so that allows to infer relationships among the concepts.

### 2.1 Collecting the Articles

This aim is critical because it must construct an impartial and exhaustive base of information that will ultimately allow to obtain results as accurate as possible with respect to the information published as of the time the inquiry is initiated. There are some difficulties to overcome. First, a text can include many different topics. This can hinder the extraction of real relationships between concepts. In addition, the topics can refer to affirmations and to negations, which may lead to completely opposite relationships. Fortunately, the biomedical literature is usually restricted to a specific subject. But, we have used only the title and abstract of the articles because, first, it is difficult to work with the whole texts of an article since their automated access is not allowed, due to copyright restrictions. In addition, processing the whole text of an article requires a great deal of time. On the other hand, the title and the abstract are much more useful than the complete article because they are usually specific, succinct and writen in an affirmative way.

There is a large variety of articles in PubMed but not all of them are useful for our purpose. The following are desirable characteristics:

- they must be writen in English, because the tools

to process them use English,

- we excluded reviews. Although the information contained in them is of good quality, it must be processed using natural language methods to recognize the structure and understand the contents of the text. As a first approach we have decided to work without them, and
- the articles should not have been retracted, which means that their contents are reliable.

We have implemented it by accesing Entrez from R.

## 2.2 Text Processing

Once the base of articles is collected, SalamboMiner provides lots of information, but it is necessary to extract the interesting elements. We must look forward to the moment when SalamboMiner will provide knowledge. The elements that will allow it must be extracted. These elements are those concepts that pertain to the biological pathways that connect genes with the causative diseases. There are lots of elements that cause diseases. To start, we extracted the more significative elements which are genes, proteins and diseases, since the elements that affect directly to the occurrence of diseases are proteins which are produced by genes.

The tools used in this module are:

- MeSH terms were assigned to the articles in the PubMed database by NCBI curators.
- Unified Medical Language System (UMLS), a set of sources and associated software developed in the US National Library of Medicine was useful to manage our biomedical data. An important source of information in SalamboMiner is the Metatheaurus, which is a unified set of biomedical thesaurus that includes terms identified in a conceptual manner. Another source used in our project is the Semantic Network, which gives semantics to the concepts and to the relationships among them. In addition, we have used the software MMTx that maps biomedical texts in order to identify the UMLS concepts included in them.
- Biotagger, a biological Entity Recognizer (McDonald and Pereira, 2005).

We have divided the text processing into two parts:

### 2.2.1 Extraction of Terms

It aims to filter, from all the terms in an article, those which refer to genes, proteins and diseases. For that, we have used MeSH terms assigned to articles. The original purpose of the MeSH terms was to index the articles so that they are easy to find in the Pubmed database. But, it is also useful for our purpose because the MeSH terms have also the ability to describe the contents of the articles. We used the MeSH terms to identify genes, proteins and diseases. But, we must improve the coverage of genes and proteins because the MeSH database does not cover completely the known genes (Mary, 2004). To do that, we use the Biotagger, which is specialized for recognizing genes and proteins (Yeh, 2005).

### 2.2.2 Translation of Terms into Concepts

It is necessary to deal with the synonyms and acronyms that are commonly used to refer to genes, proteins and diseases. For this purpose we used the UMLS, that allowed us to obtain the concepts that unify the terms in refering to the same biological entity. The UMLS thesaurus have a high coverage of concepts of genes and proteins (Mary, 2004), since the UMLS includes thesaurus like HUGO, OMIM, SwisProt, GO, GOA, GeneBank, MeSH and MeSH SR. And the UMLS also covers properly the diseases, because it includes medical databases such as MeSH, SNOMED, MEDCIN and ICD-9-CM among others. We used the MMTx software to parse the filtered input terms and assign them an UMLS concept identifier.

## 2.3 Creating and Querying the Knowledge Base

The knowledge base consists of the co-citations, as it contains the concepts and their connections. The method that we will use to obtain relationships between genes and diseases has two main components: one is a probabilistic measure of the association between the genes and diseases. This is useful to extract the degree of relationship between concepts that are co-cited. And the second component is a mechanism that facilitates the spread of the association probability, that allows the discovery of relationships between concepts that do not appear co-cited in the literature, but there is at least one intermediate element that connects them.

Association measure: The co-ocurrence of concepts in the same biomedical article provides information about the degree of relationship between these concepts. There exist numerous measures of association between two concepts based on co-citations (Lenca, 2008). We have selected the Bayes Factor. In order to conclude that there exists a relationship between two concepts, we require a Bayes Factor mini-

mum threshold of 3.

Spread of the association: Bayesian Networks offer a natural mechanism for the spread of probabilistic information. Given two concepts (C1 and C2) that are not co-cited directly, we create a Bayesian Network (directed acyclic graph) that includes all of the concepts that connect C1 with C2 in a path of length two. Each edge of the network contains a table with the conditional probability distribution of the states of a node in respect to their parents. From that information we can estimate the joint probability distribution between C1 and C2. And, with this, we can estimate the Bayes Factor between C1 and C2.

# 3 RESULTS

We present a user friendly tool implemented in Java and R. Given a gene, SalamboMiner provides a sorted list of the resulting relationships with diseases, based on the co-citations observed in the set of articles. Similarly it identifies genes if the input is a disease.

We have tested our tool using the set of Pubmed articles that appeared in the OMIM database (Online Mendelian Inheritance in Man) and that are related to diabetes and thrombosis. It collected a total of 14,212 articles. After processing them we obtain 4,044 relevant concepts that appear 56,386 times.

To validate the extraction of relationships, we have taken as reference of accuracy the information about the relationships between genes and diseases that appear in the OMIM articles. In addition, we have evaluated the results of the related concepts in two steps with the help of an expert in the area of these diseases.

## 3.1 The Query: the Gene Factor VIII

Given the gene for Factor VIII, SalamboMiner searches the diseases in which Factor VIII is involved.

## 3.2 One-Step Relationships

Table 1: Relationships that appear in OMIM.

| Disease |
|---|
| Combined Factor V and VIII deficiency |
| Hemophilia A |
| von Willebrand disease |

Two of the three associations are annotated in the OMIM database, “Hemophilia A” and “von Willebrand Disease”. Also, there is an OMIM annotation not found by SalamboMiner, “Combined Factor V and VIII deficiency”. This is due to the fact that it is a combined deficiency and not a disease in itself. This finding does not appear in any thesaurus. It is notable that the results contain a new concept, that is “Telangiectasia”. It is not annotated in OMIM, but it can be found associated with von Willebrand’s disease in the literature along with an inhibitor of Factor VIII activity (Conlon, 1978) (Hanna, 1984) (Sudarshan, 1985) (Gola, 1977).

Table 2: Extracted relationships.

| Concept | Bayes Factor |
|---|---|
| Telangiectasia | 67.52 |
| Hemophilia A | 62.16 |
| von Willebrand Disease | 38.49 |

## 3.3 Two-Steps Relationships

Some of the relationships we obtained from concepts that were not co-cited in the articles but have an intermediary element that connect them, are as follows:

Table 3: Discovered relationships.

| Concept | Bayes Factor |
|---|---|
| Hemophilia A | 125.93 |
| Facial Dermatoses | 52.60 |
| Telangiectasia | 51.30 |
| Factor VII Deficiency | 23.54 |
| Photosensitivity Disorders | 23.03 |
| Thrombus | 20.16 |
| Factor V Deficiency | 16.07 |
| Cerebellar Diseases | 12.59 |
| Puerperal Disorders | 11.54 |
| Budd-Chiari Syndrome | 11.53 |
| Coagulation Protein Disorders | 11.40 |
| Dysgammaglobulinemia | 10.64 |
| Mesenteric Vascular Occlusion | 10.01 |
| Dog Diseases | 9.48 |
| Von Willebrand Disease | 9.26 |
| Color Vision Defect | 9.05 |
| Blood Platelet Disorders | 8.91 |
| Irritable Bowel Syndrome | 8.47 |
| Virus Diseases | 8.42 |
| Factor XII Deficiency | 8.31 |
| Carotid Artery Diseases | 8.20 |
| Gastrointestinal Diseases | 8.20 |
| Bernard-Soulier Syndrome | 8.10 |
| Thrombophilia | 7.37 |

The coefficient for the “Hemophilia A” is now more consistent with reality, because Hemophilia A is caused directly by a deficiency of Factor VIII.

We asked a doctor who specializes in coagulation diseases, to fill in a questionnaire in which we listed the 61 concepts obtained. We asked him wether the relationships that we found were "very strong", "strong", "weak" or if there were "no relationship". The questionnaire had two stages: First, the doctor completed the questionnaire without using any auxiliary material. And second, the doctor used the Pubmed database to fill in the questionnaire.

Table 4: Contingency table of the questionnaire answers.

| | Without PubMed | With PubMed |
|---|---|---|
| No relationship | 32 (52.5%) | 24 (39.3%) |
| Relationship | 29 (47.5%) | 37 (60.7%) |

The 60.7% relationships among the diseases are actually related to the gene Factor VIII. Also a 39.3% of the concepts are false positives. There are two main groups of concepts with these results. One is composed by diseases associated to hemorrhages. Since the deficiency of Factor VIII causes the characteristic hemorrhages in Hemophilia A, any other manifestation associated with them will appear in our results, even though they do not have a true relationship with Factor VIII. The group contains diseases or syndromes caused by genes located on the Chromosome X. These are all false positives. These results illustrate a generalized problem in using datamining tools. However, it is more desirable to have false positives than false negatives because they can be checked in the literature.

Importantly, the answers given by the doctor did not correlate very well with the scores of the two-steps relationships. This can be due to the selected measure, the Bayes Factor, and to the fact that we were working with 15,000 articles.

The results that we obtained with SalamboMiner have provided new knowledge. 8 concepts (13%) have been related by the expert in the second stage of the questionnaire. These results contrast the information that the doctor did not know but SalamboMiner has automatically provided.

# 4 CONCLUSIONS

The strategy that we used to process the data in the articles (Mesh terms + Biotagger + UMLS + MetaMapTx) was suitable to reveal the essential concepts in the articles. Thus, it is feasible to annotate all of the articles in the Pubmed database including their relevant concepts. The more articles used the more objective and exhaustive will be the relationships.

The strategy of using a Bayesian Network to estimate the level of relationship between two non co-cited concepts is useful, but we have to continue analyzing the association measures.

Our study analyzed over 15,000 articles. Our results are informative and support the strategies that we used: First, it obtained the expected concepts of the OMIM database. And, second, it obtained relationships among concepts that were not co-cited in our sample of articles, but this relationships are exposed in the Pubmed database.

As a general conclusion, we feel that SalamboMiner is a promising tool that can be very useful for biomedical researchers to help them see relationships in the literature that are not otherwise obvious.

# REFERENCES


Conlon, C. (1978). Telangiectasia and von willebrand's disease in two families. *Annals of Internal Medicine*, 89:921–924.

Gola, A. (1977). Ein fall von morbus osler mit gleichzeitig bestehender thrombozytopenie und einem faktor viii-inhibitor. *Folia Haematologica*, 104:102–108.

Hanna, W. (1984). A study of a caucasian family with variant von willebrand's disease in association with vascular telangiectasia and haemoglobinopathy. *Thrombosis and Haemostasis*, 51:275–278.

Lenca, P. (2008). On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European journal of operational research*, 98(5):1031–9.

Mary, V. (2004). Mesh and specialized terminologies : coverage in the field of molecular biology. *Studies in Health Technologies and Informatics*, 107(Pt 1):530–4.

McDonald, R. and Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6(Suppl 1):S6.

Sudarshan, A. (1985). Hereditary hemorrhagic telangiectasia and factor viii inhibitor. *Southern Medical Journal*, 78:623–624.

Weeber, M. (2005). Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics*, 6(3):277–286.

Yeh, A. (2005). Biocreative task 1a: gene mention finding evaluation. *BMC Bioinformatics*, 6(Suppl 1):S2.

Zweigenbaum, P. (2007). Frontiers of biomedical text mining: current progress. *Briefings in Bioinformatics*, 8(5):358–375.