SalamboMiner
A Biomedical Literature Mining Tool for Inferring the Genetics of Complex Diseases

Leonor Rib[1], Ricard Gavaldà[2], Jose Manuel Soria[1] and Alfonso Buil[1]
[1]*Unit of Genomics of Complex Diseases, Institut de Recerca del HSCSP, Barcelona, Spain*
[2]*Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain*

Keywords: Literature mining, Knowledge discovery, Bayesian Networks.

Abstract: In the Era of Information researchers have utilized the Web to make their knowledge readily available. The Web is an important tool to improve the communication in the research community. But, the large amounts of information available makes it difficult to access the information that is needed.
We present SalamboMiner, a Text-Mining tool that helps biomedical researchers to obtain the information about the genetics of complex diseases which is in the published biomedical literature. The methodology is based in the idea of co-citation: the co-citation of two concepts gives the significance of the relationship between the pair of concepts. In addition, the co-citation allows to infer new relationships that are not explicitly said in the literature. By using a Bayesian network, we infer the significant relationships between those concepts that are co-cited in two steps.

1 INTRODUCTION

During the past several years, biomedical research has accumulated a large amount of knowledge. All this information is published in professional journals available to the entire research community. Thanks to the internet information is available online. Although online information provides an excellent oportunity for interaction among researchers, the amount of information is enormous and is increasing exponentially year by year (Zweigenbaum, 2007). As a result, it is increasingly difficult for researchers to utilize the information efficiently.

In the last two decades, many tools have been developed to help researchers obtain information from the literature databases. These tools are included in the field of Biomedical Text-Mining. Broadly defined, text-mining includes a range of applications, from the extraction of facts using Named Entity Recognition to thenew knowledge discovery performed with Literature-Based Discovery. All of these methods can be applied in quite different areas of biomedicine from drugs discovery to genomics of diseases. SalamboMiner is a text-mining tool that focuses on obtaining information about the genes that are involved in complex diseases by analyzing the co-citation of concepts in the same article. The co-citation of concepts is a useful tool to uncover the role of genetics in complex diseases (Weeber, 2005) (Zweigenbaum, 2007). The co-citation of concepts in the same biomedical article gives information about the degree of relationship among themselves in the real life.

Many aspects of text-mining tools must be defined very carefully, because the way they are defined affects the computing cost and the reliability of the results. Many questions arise when developing a text-mining tool. Is it good to use the full text of the articles? What elements should be extracted from the articles and what is the better way to extract them? How will the extracted elements be analyzed?

Is it good to use the full text of the articles? The articles are the source of information we need to do co-citation analysis but now there are restrictions to use the full text of all the articles due to copyright issues. The usage of the abstract and the title is now the more feasible option, because there are no limitations to access them. In addition, the abstracts provide very relevant information since they summarize the contents of the articles. Especially, the abstracts are the part of the articles that contain the more complete information about the genes, proteins and diseases in-

Rib L., Gavaldà R., Manuel Soria J. and Buil A..
SalamboMiner - A Biomedical Literature Mining Tool for Inferring the Genetics of Complex Diseases.
DOI: 10.5220/0003143201430148
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2011), pages 143-148
ISBN: 978-989-8425-36-2

cluded in the articles (Shah et al., 2003).

What elements should be extracted from the articles and what is the better way to extract them? There are different strategies to approach the extraction of information from articles. One possibility is to extract facts from the text by applying Biomedical Language Processing methods (Weeber, 2005). There are some important difficulties to be taken into account (Hunter, 2006). The words can be ambiguous, allowing for multiple interpretations for strings, like genes whose names are identical to prepositions, acronims and abbreviations. Another difficulty is to deal with the polysemy of gene symbols, when a single name or symbol can refer to more than one gene. And metonymy can also occur when some string can refer to more than one concept that are related, like the proteins of a gene. This is the case of the p53 described in Lakoff and Johnson article (Lakoff and Johnson, 1980). Named Entity Recognizers are tools that help the correct extraction of biomedical terms like genes and proteins. Now, there is an increasing attention paid to syntax analyzers (Zweigenbaum, 2007). Their task is computationally complex but the improvement of computing power and faster algorithms makes it more feasible every day.

In addition, from articles can also be retrieved metadata associated to them. Concretely, the MeSH Terms are useful information. They are biomedical terms and, originally, they aimed to index the articles in such a way that can be found easily in the databases. At the same time, the MeSH terms are useful in text-mining because they summarize the relevant contents, especially genes and diseases. Moreover, other information can be included to the analysis. For example, chromosome location (Hristovski et al., 2005), (Perez-Iratxeta, 2002) can be used to better identify the desired elements from the articles. Other biological information can also be used to improve the knowledge base like in the tool developed by Stein Aerts et al (Aerts, 2006).

Once the desired elements of the text have been retrieved, the synonimous terms must be unified into its concept (Bard and Rhee, 2004). There are initiatives called ontologies that build representation systems biomedical knowledge which unify terms into concepts. In addition, the concepts are organized in a semantic manner: they have associated a semantic category and they are related among themselves by a semantic relationship. Some extended ontologies are Gene Ontology (GO), Open Biological Ontologies (OBO) and Unified Medical Language System (UMLS).

How will the extracted elements be analyzed? The analysis of the co-citation of concepts can be done in many different ways. For example FACTA (Tsuruoka et al., 2008) uses Pointwise Mutual Information, PubGene uses a Network of Co-citations (Jenssen et al., 2001), PolySearch (Cheng et al., 2008) uses Association Rules, Anni 2.0 (Jelier et al., 2008) uses Concepts Profiles, the tool of Varun K. Gajendrana et al (Gajendrana et al., 2007) uses a model based in the Zipf law, Seki et al (Seki and Mostafa, 2007) use a probabilistic network, and Bitola (Hristovski et al., 2005) uses Association Rules. The last four use these methods to extract relationships and to discover new relationships that are not explicitly writen in the articles, while the others apply only to methods to extract implicit relationships.

SalamboMiner is a text-mining tool that aims to obtain a prioritized list of relationships among genes and diseases. The relationships obtained are twofold: based on explicit information and based on implicit relationships found by literature discovery. It extracts from the title, the abstract and MeSH terms of PubMed articles the relevant concepts within the articles. We have defined as relevant concepts those that represent genes, proteins and diseases. To do the extraction of genes and proteins, we use the BNER Biotagger (McDonald and Pereira, 2005) and we obtain the diseases from the MeSH terms. It is important to unify the elements into their respective concepts because it helps to manage of the extracted terms. In SalamboMiner we use UMLS because of it gives a relatively good coverage of genes and protein concepts (Mary, 2004) although it has some problems in the semantic assignment as described by Huanying Gu et al (Gu, 2007).

In our project, we use the Bayes Factor measure to extract relevant relationships among concepts. And, to discover new relationships, we construct a Bayesian Network with the concepts as nodes and their co-citations as edges, and we use the Bayes Factor again to infer new relationships.

We have tested SalamboMiner with data set that comprised articles that appeared in the database of Online Mendelian Inheritance in Man (OMIM) that are related to Thrombosis and Diabetes.

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 written 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 the statistical software R (www.r-project.org).

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. 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 extract the more relevant 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 has 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. 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 minimum threshold of 3.

$$\frac{P_{A/B}}{P_{A/\neg B}} \quad (1)$$

A Bayes Factor higher than one means that appearing both concepts A and B is more strongly supported than appearing A but not B. The Bayes Factor is an asymetric measure and we will obtain two Bayes Factors (BFA-¿B and BFB-¿A) associated to a pair of co-cited concepts (A and B). To conclude that there exists a relationship between two concepts, we require a Bayes Factor minimum threshold of 3 (Jeffreys, 1961).

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 the programming language Java and the statistical software 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

Table 2: Extracted relationships.

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

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).

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.

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.

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%)

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 only 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 these 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

Aerts, S. e. a. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24:537 – 544.

Bard, J. B. L. and Rhee, S. Y. (2004). Ontologies in biology: design, applications and future challenges. *Nature Reviews Genetics*, 5:213–222.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., and Wishart, D. S. (2008). Polysearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res*, 36.

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

Gajendrana, V. K., Linb, J.-R., and Fyhrie, D. P. (2007). An application of bioinformatics and text mining to the discovery of novel genes related to bone biology. *Bone*, 40(5):1378–1388.

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

Gu, H. (2007). Evaluation of a umls auditing process of semantic type assignments. *AMIA Annu Symp Proc*, page 294298.

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.

Hristovski, D., Peterlin, B., Mitchell, J. A., and Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, 74(2-4):289–98.

Hunter, L. (2006). Biomedical language processing: Perspective whats beyond pubmed? *Molecular cell*, 21(5):589594.

Jeffreys, H. (1961). *The Theory of Probability*. Oxford University Press.

Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G., and Kors, J. A. (2008). Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biology*, 9(6).

Jenssen, T. K., Laegreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21–8.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. IL: University of Chicago Press.

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.

Perez-Iratxeta, C. (2002). Association of genes to genetically inherited diseases using data mining. *Nature Genetics*, 31:316 – 319.

Seki, K. and Mostafa, J. (2007). Discovering implicit associations between gens and hereditary diseases. *Pacific Symposium on Biocomputing*, 12:316–327.

Shah, P. K., Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2003). Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4:20.

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

Tsuruoka, Y., Tsujii, J., and Ananiadou, S. (2008). Facta: a text search engine for finding associated biomedical concepts. *Bioinformatics*, 24:2559–2560.

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.