MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010
© 2010 IMIA and SAHIA. All rights reserved.
doi:10.3233/978-1-60750-588-4-954

Text Mining approaches for Automated Literature Knowledge Extraction and Representation

Angelo Nuzzo[a], Francesca Mulas[a], Matteo Gabetta[b], Eloisa Arbustini[c], Blaž Zupan[d,e], Cristiana Larizza[b], Riccardo Bellazzi[b]

[a] *Center for Tissue Engineering, University of Pavia, Italy*
[b] *Dipartimento di Informatica e Sistemistica, Università di Pavia, Italy*
[c] *IRCCS Policlinico San Matteo, Pavia, Italy*
[d] *Faculty of Computer and Information Science, University of Ljubljana, Slovenia*
[e] *Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX,USA*

Abstract

Due to the overwhelming volume of published scientific papers, information tools for automated literature analysis are essential to support current biomedical research. We have developed a knowledge extraction tool to help researcher in discovering useful information which can support their reasoning process. The tool is composed of a search engine based on Text Mining and Natural Language Processing techniques, and an analysis module which process the search results in order to build annotation similarity networks. We tested our approach on the available knowledge about the genetic mechanism of cardiac diseases, where the target is to find both known and possible hypothetical relations between specific candidate genes and the trait of interest. We show that the system i) is able to effectively retrieve medical concepts and genes and ii) plays a relevant role assisting researchers in the formulation and evaluation of novel literature-based hypotheses.

Keywords:

Text Mining, Annotation networks, Gene ranking, Candidate gene study

Introduction

Current biomedical research is starting to increasingly rely on automated literature analysis. Text Mining (TM) and Natural Language Processing (NLP) provide algorithms and techniques for automated summarization and analysis of textual content, so that it is possible to extract and interpret the information contained in literature databases and repositories. This task is particularly important in the early stage of any study, when gathering the available knowledge about the problem of interest is crucial in formulation of initial hypotheses and planning of the next research tasks. Several TM systems exists which can help us expose relevant biomedical literature [1] in order to retrieve available knowledge which is relevant to user's interest, like finding all publications about a disease candidate gene. The challenge is to broaden and deepen this search to expose possible other useful information for proposal of novel hypotheses [1 - 3]. For instance, an added value could be the suggestion that the candidate gene is often related to another gene, which has not been previously considered.

We describe the tools that we are developing to provide such kind of support. In particular, we focused on genetic studies, in which a set of initial hypotheses of genes-disease association is made on some candidate genes, so that the first step is to explore the recent literature to confirm their possible role in the disease mechanism. We extracted the concepts of interest (genes and medical terms, like pathologies) using a structured knowledge base like Unified Medical Language System (UMLS), by which we derived genes/disease annotation. Then we implemented a similarity metric that is based on a relevance measure of the terms for each gene. In this way the approach identifies which terms are shared between genes. The results of such analysis can be summarized as a graph in which the proximity of the nodes reflect how tightly related are these terms according to the available literature.

We tested our tools for the INHERITANCE research project, which aims to translate basic knowledge of the aetiology and pathophysiology of genetic dilated cardiomyopathies (DCM) into routine clinical practice and to identify novel therapeutic strategies. We show how our automated literature analysis strategy was able to both correctly reconstruct the available knowledge and support researcher's new hypotheses formulation and evaluation.

Materials and Methods

The analysis method we propose aims at derive a literature-based gene annotation by extracting UMLS terms related to diseases from the abstracts of the publications referencing each gene. The overall analysis consists of 3 main steps:

- querying PubMed via Web Services to retrieve the most recent literature about specific genes/diseases
- automated extraction of concepts (genes/disease) from PubMed abstracts based on NLP techniques
- construction of annotation/co-citation networks to interpret available knowledge and suggest new hypotheses that can be tested.

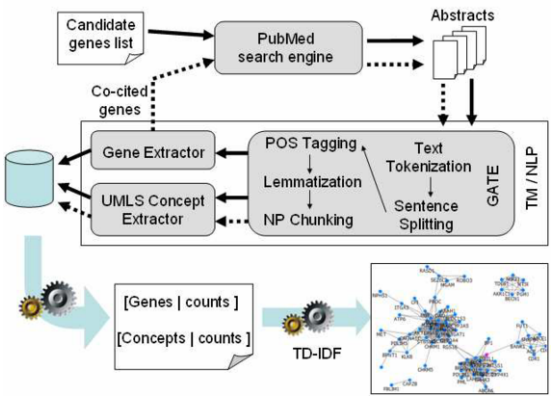

Figure 1 – Schematic representation of the analysis pipeline

The literature searching engine

The first module we developed is a searching engine which exploits the NCBI Web Service implementation of the Entrez Programming Utilities (EUtils) [4]. EUtils are software tools that provide access to Entrez data outside of the regular web query interface, in order to retrieve search results in another environments. In particular, the Web Service implementation, which enables developers to access Entrez Utilities via the Simple Object Access Protocol (SOAP). The modules query PubMed to retrieve scientific papers dealing with the genes of interest in XML format, so that their title and abstracts can be automatically processed and analyzed .

Abstracts text mining

The system is aimed at extracting both medical concepts (diseases, in particular) and genes, from titles and abstracts of articles obtained from PubMed databases. To work out this task, we have implemented a text extraction system that relies on a language processing environment called GATE [5]. The text analysis is handled through seven different steps, scheduled in a pipeline-like architecture. The first five steps are typical of many text mining systems and are in order: Text Tokenization, Sentence Splitting, Part Of Speech (POS) Tagging, Lemmatization and Noun-Phrase (NP) Chunking. The last two modules of the pipeline, i.e. the UMLS Concept Extractor and the Gene Extractor, have been designed specifically for our purposes.

The *Text Tokenizer* operates in two stages: the simple identification of parts of text separated by blank spaces and the management of the language-dependent exception to the basic rule.

The *Sentence Splitter* separates the sentences within the text. The so-called "ANNIE POS *Tagger*" module assigns each previously identified token to the grammatical class (POS) that it belongs to. The *Lemmatizer* derives the lemma belonging to every token. The *Noun-Phrase-Chunker* is aimed at identifying noun phrases (NP) within the text. The NP identification has proven its usefulness in many text mining tasks, because most searched concepts are contained in NPs [6].

The *UMLS Concept Extractor* module accomplishes the task of extracting medical concepts from the text. It relies on the resources available within the Unified Medical Language System (UMLS) [8], one of the National Library of Medicine (NLM) projects. In particular we exploited the UMLS Metathesaurus, a large database containing health-related concepts coming from many different source vocabularies. Among the different sources, our systems relies on the main ones: MTH (the official vocabulary of the Metathesaurus) and MSH (Medical Subject Headings), both provided by the NLM [9]. The module starts its analysis generating a set with all the possible substrings of each noun phrase in the document. For each token contained in the noun phrase those substrings are generated considering the token itself and its lemma. At this point the module sends a query to the database storing the concepts for each string. When a positive matching arises, the system makes another query aimed at identifying the official name of the found concept and creates a new annotation on the document.

Finally, the *Gene Extractor* finds the genes names present in the analyzed text. The identification is related to the whole abstract document, as the exact position of the name in the phrase is not what we are seeking. As well as the UMLS Concept Extractor, also the Gene Extractor relies on a database, which is directly derived from the Entrez Gene NCBI's database [10] (we consider human genes only). The module starts evaluating each token found by the Text Tokenizer in order to discard strings which do not fit with a standard gene representation (for example a string candidate to represent a gene must have almost one capitalized letter and can't be composed only by digits). Once the set of candidate strings is defined the Gene Extractor sends a first query to the database for each string, trying to find the string between all official genes names and their synonyms. When a match with an entry is found, a second query is sent in order to find out the official name of the gene previously identified.

Once that articles titles and abstracts have been retrieved for each candidate gene contained in the initial set (11 and 18 genes for Hypertrophic Cardiomyopathy and Dilated Cardiomyopathy respectively, see Results section for more details), these textual resources pass through the text mining pipeline just described. The analysis is therefore the same for titles and abstracts, but the results are kept separate in order to preserve the possibility of a further separate management of the results. From the first set of abstracts analysis, the system exploits the output of the Gene Extractor in order to build a new, larger, set of co-cited genes which are cited together with the first candidate ones. Then, for each gene of this augmented set, we retrieve articles titles and abstract from PubMed; all of these textual resources pass again through the analysis pipeline, now

ending with the UMLS Concept Extractor. By now the system is able to associate an array to each gene belonging to the larger set. The array contains the name of the diseases cited in the articles titles and abstract relative to the gene and, for every disease, the number of citation occurrences. At this point all the arrays are passed to the annotation measure and analysis module described in the next section.

Annotation networks

Our final aim is to derive a literature-based gene annotation by extracting UMLS terms related to diseases from the abstracts and the titles of the publications referencing each gene. Information about each gene in our study can be therefore represented by a gene annotation profile *A*, composed by a set of UMLS terms indicating diseases or symptoms and the counts of their occurrences in PubMed entries.

Formally, for the g^{th} gene, A_g is a feature vector which contains a set of tuples $a_{gj} = \{t_{gj}, f_{gj}\}$, with t_{gj} being the annotation term (i.e., a UMLS term) and f_{gj} the frequency the term appears in the annotation (i.e., the total number of times that the UMLS term t_{gj} is included in the papers citing gene *g*, with *j* ranging from 1 to the number of annotation terms found to be relevant for gene *g*). To expose the important terms for each gene, we applied a TF-IDF (Term-Frequency Inverse Document Frequency) transformation. TF-IDF is a popular technique used in Vector Space Model approaches [11] for the preprocessing of textual documents. Within this model, documents are represented by vector of features where each term is weighted according to its importance in the document. For a term *t* found in a document *d* of a document corpus *D*, the TF-IDF weight is computed as:

$$TF - IDF(t,d) = TF(t,d) * IDF(t) \tag{1}$$

where TF(*t*,*d*) is the term frequency in the document *d*, and IDF(*t*) is defined as:

$$IDF(t) = \log\left(\frac{|D|}{DF(t)}\right) \tag{2}$$

where |*D*| is the number of documents in the document corpus and DF(*t*) the number of times a term *t* appears in all documents.

Since we represent a gene by a vector of UMLS terms, the documents correspond in our case to genes and |*D*| is the total number of genes. As a result of this weighting scheme, a term indicating a disease is considered as an important annotation for a gene if it occurs frequently in the publications related to that gene. On the other hand, terms about diseases or symptoms that are not specific for a gene are rated as less important due to their low IDF. The weighted annotation profiles were then normalized.

One of our purposes was to test if the proposed gene annotation method is able to find groups of similarly annotated genes that are in fact known to be similar as they play an important role in the same disease. In addition, we aimed to find other genes related to cardiomyopathies. For these reasons, we needed a similarity measure that reflected the degree of gene similarity in terms of their feature terms and weights. To compute the similarity between the annotation profiles of two genes *g1* and *g2*, we resorted to the cosine similarity between the TF-IDF vectors W_g as follows:

$$sim(g_1, g_2) = \frac{W_{g1} \cdot W_{g2}}{\|W_{g1}\| * \|W_{g2}\|} \tag{3}$$

A similarity of 1 means that g1 and g2 have exactly the same terms and weights, while a cosine value of zero means that the two annotation vectors are orthogonal and had no match. To have a graphical visualization of the groups of similarly annotated genes we created gene association networks, where the genes are the nodes of a network and they are linked if their similarity is greater then a threshold (in our case set to 0.7).

The information coming from our literature-based annotation was used to build three types of gene annotation profiles: i) *Titles*, including for each gene only the terms extracted from the titles of the gene-related PubMed entries; ii) *Abstracts*, including only the terms extracted from the abstracts of the publications. iii) *Titles+Abstracts*, obtained by including all the terms from Titles and Abstracts and for each of them adding the corresponding weights resulting from TF-IDF.

Finally, we developed a Python procedure to process the gene annotation profiles, to evaluate the pair wise similarities and to create network files in a standard format. A network has been created for each of the annotations (Titles, Abstracts, Titles+Abstracts). The networks were visualized using an interactive network exploration module provided by the software Orange [12]. Giving information about gene names and attributes to this software, it is possible to show gene symbols next to each node, underline the genes of interest and make a zoom on a specific region. The software allows the selection of some nodes of interest (in our case the genes related to cardiomyopathies and their annotation-similar genes) in order to visualize their attributes. Giving as input data the weighted annotation vectors to Orange, we were also able to visualize in a table (see Figure 2 in Results section) the terms rated by TF-IDF scheme as the most important terms for our genes of interest and for their annotation-neighbors.

Results

We tested the strategy and tools described above to analyze data concerning Hypertrophic Cardiomiopathy (HCM) and Dilated Cardiomiopathy (DCM), which are the pathologies being studied in the Inheritance Project. HCM and inherited DCM are most commonly transmitted as an autosomal dominant traits is and they have been associated to mutations in a number of genes that encode for one of the sarcomere proteins. 11 loci (relative to genes *TNNT2*, *TTN*, *MYBPC3*, *ACTC*, *TPM1*, *MYH7*, *MYH6*, *MYL2*, *MYL3*, *TNNC1*, *TNNI3*) are known to be associated to HCM and 18 associated to DCM (*TNNT2*, *TTN*, *MYBPC3*, *ACTC*, *TPM1*, *MYH7*, which are the same as for HCM, together with *ABCC9*, *CLP*, *CTF1*, *DES*, *DMD*, *DSP*, *LDB3*, *LMNA*, *MVCL*, *PLN*, *SGCD*, *TAZ*) [13]. We used the two lists as starting set of candidate genes that we

want to investigate[1]. Through the search engine and the extractors modules we retrieved 1409 concepts and 866 other genes that are co-cited in the 30 most recent abstracts dealing with each of those initial genes and the 15 most recent abstracts for each of the other genes. For brevity, here we present and discuss the results of the Titles+Abstracts analysis, as it is the more informative combination we obtained[2]. The computed networks involve a great number of genes and other medical terms which are cited together with the initial ones, as well as all genes/medical terms related to the derived genes. Thus, we identified the spatial distribution of the initial sets, which have been highlighted in different colors: blue nodes are the six common associated genes, the orange ones are HMC related genes, the pink ones are the DCM related genes and the green filled ones are genes not mentioned in the known list but with an high co-citation index with the term "cardiomyopathy" (Figure 2).

TNNT2	**Hypertrophic Cardiomyopathy** - Hyperostosis, Diffuse Idiopathic Skeletal - **Cardiomyopathy, Dilated**
TTN	Respiratory Distress Syndrome - **Hypertrophic Cardiomyopathy** - **Cardiomyopathy, Dilated**
MYBPC3	**Hypertrophic Cardiomyopathy** - **Cardiomyopathies** - Heart failure
ACTC	**Hypertrophic Cardiomyopathy** - **Cardiomyopathies** - **Cardiomyopathy, Dilated**
TPM1	**Hypertrophic Cardiomyopathy** - **Cardiomyopathy, Dilated** - Exanthema
MYH7	**Hypertrophic Cardiomyopathy** - Myopathy – **Cardiomyopathy, Hypertrophic, Familial**
PRKAG2	**Cardiomyopathies** - Ischemia - **Hypertrophic Cardiomyopathy**
ANKRD1	Myopathy - Muscular Dystrophy - **Cardiomyopathy, Dilated**
TCAP	Muscular Dystrophy – Myocarditis - **Cardiomyopathy, Dilated**
PPIF	Muscular Dystrophy – Swelling - **Cardiomyopathy, Dilated**
...	...

Figure 2 – Most important terms for some of the genes of interest and for their annotation-neighbors

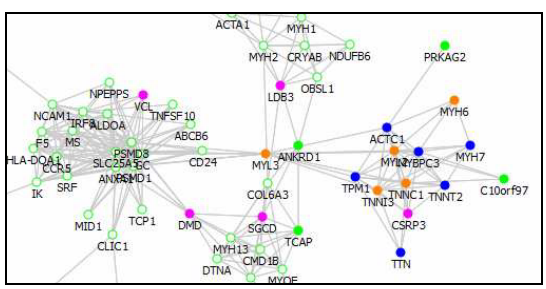

Figure 3 – Zoom on the HCM network . Blue nodes are the six common associated genes, orange nodes are HMC-related genes, pink-nodes are DCM related genes and the green-filled nodes are genes highly co-cited with the term cardiomyopathy

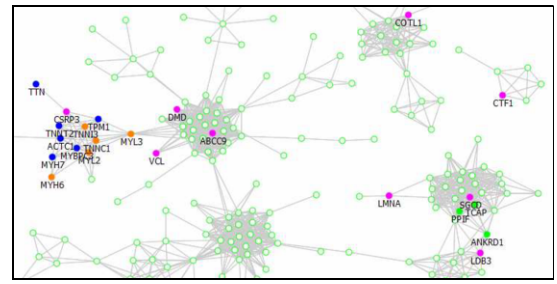

Figure 4 - Zoom on the DCM network

Figure 3 shows a snapshot of the region including the labeled genes for the network computed starting from the HCM related genes, while Figure 4 shows labeled genes for the network computed starting from the DCM related genes. Both of them are discussed in the next session.

Discussion

In this study we exploited the available knowledge about the genes which are known to be associated with the trait i) in order to perform an internal validation of the knowledge extraction capability of our tools, and ii) as a "background" to which new hypotheses of association can be compared.

In the first case, the results shown in the previous section confirm that the annotation networks built with our approach are able to reconstruct the existing knowledge, as the genes involved in cardiomyopathy are always clustered together. In particular, the six core genes (labeled with blue nodes in the figures) are always tightly connected, and the other genes related to HCM (highlighted in orange) always belong to the same clusters. This result gives a sufficient confidence that the algorithms have been properly set up to be able to extract evidence that we already know to be true. Similarly, networks show how the genes specifically related to DCM are on the contrary not so strictly linked to the HMC genes cluster, but they tend to be more scattered away from it. This also reflects the fact that current knowledge does not allow differentiating treatments on the basis of the different subtypes of DCM. Medical and interventional treatment strategies for DCM coincide with those optimised and indicated in the guidelines of scientific societies for HCM, congestive heart failure and atrial or ventricular arrhythmias or conduction disease. On the other end, the spatial distribution confirms the researcher hypotheses that DCM genetic mechanisms should be quite different from HCM ones.

Finally there are also some genes (green-labeled in figures) which are closely connected to the other known genes. Then, as the consideration made above gives us a good confidence

[1] Our elaboration took into account also synonyms of the gene names reported in [13]. In particular, our figures use ACTC1 for ACTC, COTL1 for CLP and VCL for MVCL.

[2] Titles are too shorts, and lead to very low connected network; however we kept the importance of the presence of a gene name or medical term in the title to strengthen Abstract-only results

that the literature analysis can properly extract knowledge, the new genes (green-filled nodes) related to the known ones may be interpreted as new candidate genes which could be further investigated.

Conclusion

In this paper we described a method for automated analysis of scientific literature. The tool is based on presentation of concepts in the association network, where concepts are related with respect to their similarity manuscript annotations in the most recent publications. We show how we can effectively retrieve medical concepts and a list of problem-related genes using this tool, which can play a substantial role in assisting researchers in the formulation and evaluation of literature-based novel hypotheses.

Our analysis investigates the data on the most recent published manuscripts, because the main interest is on the new findings, which have not yet been included in secondary databases that keep track of validated biological associations (e.g. OMIM, GAD, etc). The results depend on the abstracts content specificity. Our case study showed that the graphical representation can greatly facilitate results inspection, and the network tool is completely integrated with a data mining suite which provide a variety of other modules that can be used for further investigation of the subset of interesting genes.

The text mining process is comparable with the many already available tools, and a performance evaluation and comparison in term of precision and recall over a significant dataset is still ongoing. Nevertheless, our systems presents some peculiarities: i) the knowledge base on which the term recognition relies consists in whole terminological databases, while in general other systems use ad-hoc and manually cured collection of terms (moreover, this gives the system a greater modularity and scalability); ii) the similarity networks are based on a *weighted* measure of term's importance for a gene, and iii) the networks are able both to represent the current knowledge and to guide the hypothesis generation process.

The main assumption underlying the interpretation of the annotation similarity networks is that they associate concepts that are often referred to in the same manuscript and for this reason may have similar manuscript-based annotations. Notice that this does not (necessarily) correspond to a statistical or biological association between genes and diseases, but may nevertheless suggest a possible relation.

Acknowledgments

This work is a part of the "Bioinformatics for Tissue Engineering: Creation of an International Research Group" project, funded by the "Fondazione CARIPLO", the "ITALBIONET - Rete Italiana di Bioinformatica" FIRB project, and the "INHERITANCE - INtegrated HEart Research In TrANslational genetics of dilated Cardiomyopathies in Europe" EU project.

References

[1] Weeber M, Kors JA, Mons B: Online tools to support literature-based discovery in the life sciences. Brief Bioinform 2005, 6(3):277-286.

[2] Swanson, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge., Perspect. Biol. Med., 1986, 30(1): 7–18.

[3] Roos M, Marshall MS, Gibson AP, Schuemie M, Meij E, Katrenko S, van Hage WR, Krommydas K, Adriaans PW. Structuring and extracting knowledge for the support of hypothesis generation in molecular biology. BMC Bioinformatics. 2009 Oct 1;10 Suppl 10:S9.

[4] NCBI Entrez Utilities Web Services. http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esoap_help.html

[5] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proc.of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[6] Hepple M. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000), Hong Kong, October 2000.

[7] Ananiadou S. et al. Automatic Terminology Management in Biomedicine. Text Mining for Biology and Biomedicine, Cap. 4, 2006, pp. 67-98

[8] Lindberg DA, Humphreys BL and McCray AT. The Unified Medical Language System. Methods of information in medicine 1993,32:281–291.

[9] "UMLS – Metathesaurus Release Source Vocabularies", retrieved: 10/12/2009, http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/source_vocabularies.html.

[10] Maglott D, Ostell J, Pruitt KD, Tatusova T: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2005, 33 (Database Issue):D54-8.

[11] Salton G, Wong A, Yang CS. A Vector Space Model for Automatic Indexing. Commun ACM 1975:11(18):613-20.

[12] Curk T, Demsar J, Xu Q, Leban G, Petrovic U, Bratko I, Shaulsky G, Zupan B. Microarray data mining with visual programming. Bioinformatics 2005, Feb 1;21(3):396-8.

[13] Ahamad F, Seidman JG, Seidman CE. The genetic basis for cardiac remodeling. Ann Rev Genomics. Hum Genet 2005;6:185-216

Address for correspondence

angelo.nuzzo@unipv.it