Discovery and Assessment of Gene-Disease Associations by Integrated Analysis of Scientific Literature and Microarray Data

Alberto Faro, Daniela Giordano, Concetto Spampinato

Abstract—The paper outlines a methodology and presents a tool to help biomedical researchers in interpreting complex experiments by automatically discovering gene networks and underlying biological processes (revealed by gene-expression patterns) that usually are extracted manually using existing tools. The proposed method, first, starts by mining specialized medical literature available on the web to discover possible associations between genes and diseases. Discovered gene-disease associations are subsequently explored by analyzing abnormally expressed genes using microarray data analysis. Afterwards, relevant gene networks are built by clustering these genes on the basis of the similarity of their profile expressions in microarrays data. Finally, molecular, biological processes, cellular components and molecular functions, which may have a role in the disease, are pointed out by querying the Gene Ontology (GO) database. The methodology is illustrated by a case study on neuromuscular disorders.

I. INTRODUCTION

In the post-genomic era a major challenge is to develop tools and methodologies to support, on the one hand, high-throughput analysis of biological data, and on the other hand, to support richer analysis through the integration of biological and scientific data coming from multiple sources. Although biological data are increasingly available on the Web, in the form of annotated gene databases and repositories of microarray experiments, and bioinformatics analysis tools are also available as open source software or web services, still there is a lack of integrated and usable tools that effectively support biomedical researchers in conducting their investigations. A fairly well-established approach to discover gene-disease association is based on text-mining of the scientific literature [1], [2], [3]. The number of associations that can be derived from text mining of the literature can be daunting, and it is important to devise methods, e.g. [4], for reliable identification of the most promising relationships that deserve future targeted investigation. In the meanwhile advances in molecular biology have led to the conclusion that in order to reveal the mechanisms of cell functioning, it is necessary to consider sets of interacting genes, e.g. a gene network, or “relevance network” [5], that is a group of genes whose expression levels are highly predictive of others genes in the group. Existing databases to research and build networks of genes (e.g. [6], [7], [8]) try to make available the results of the analysis of gene networks in such a way as to highlight genes’ functions and evolution; in particular, they record information about functional connections within the gene network and the experimental results from which these connections were deducted. Examples of these databases are: 1) GeNet [6], a database provided with a set of Java applets to visualize the gene networks; it is also used as a web publishing tool by molecular biologists who study the mechanisms of gene interaction; 2) STRING [7], a database that, for a given a protein, extracts the genes’ interactions using various sources such as genomic context, high-throughput experiments, conserved co-expressions and publications; and 3) INGENEU [8], capable of synthesizing molecular genetics data in models of networks of genes. However, these tools are not provided with reliable validation and explanation mechanisms to avoid incompatibility and false deductions from a biological point of view. Notable exceptions are BioCAD [9], GEN-CLIP [10] and GEM-TREND [11]: softwares for the construction and analysis of bio networks, which resort to text mining and annotation techniques for the validation of these processes and integrate different information sources (mainly a genetic regulatory network with its corresponding protein-protein interaction map). Approaches such as BioCAD, GEN-CLIP and GEM-TREND are useful to aid interpretation of complex experiments; in this paper we propose a novel system and a methodology that implements a complementary approach, i.e., instead of starting from the interpretation of an experiment, the starting point is represented by the hidden and novel associations between genes and diseases that are discovered from mining the literature and that might turn out crucial in developing new diagnostic and therapeutic procedures. These hidden associations are investigated by resorting to microarray data from which gene network analysis is carried out. Finally, a correlation between the identified network and biological processes, molecular functions and cellular components is obtained. The most important difference between our approach and the existing methods is that we do not only identify the gene networks in complex experiments, but also investigate the role that these genes have in human organisms. The paper is organized as follows: sect. II illustrates the workflow of the proposed method of analysis and the components of the software tool; sect. III further demonstrates the method by a case study on neuromuscular diseases. Sect. IV summarizes the key features of the method, and outlines the future developments in the tool capabilities.

Alberto Faro, Daniela Giordano and Concetto Spampinato are with Department of Informatics and Telecommunication Engineering, University of Catania, 95125 Catania, Italy {afaro,dgiordan, cspampin}@diit.unict.it
The proposed analysis methodology has the following workflow:

- Retrieval of the data needed to search and discover associations. By consulting biological databases a set of genes, a set of diseases, and a set of biomedical scientific articles are identified;
- Text mining of the scientific articles to build the associations according to the methodology proposed in [4];
- Scrutiny and validation of each association through the analysis of specific Microarray datasets available in public repositories of gene profile expressions;
- Analysis of selected gene expression data through specific software tools. Gene Relevance Networks (GRN) are generated for each gene-disease association;
- Finally, the genes involved in the relevance networks are matched to specific biological processes, molecular functions and cellular components.

In the following we detail each step and its mapping to the components of the developed software, whose functional architecture is shown in fig. 1.

![Image of software and web resources](http://www.ncbi.nlm.nih.gov/geo/)

**Fig. 1.** Outline of the software modules and web resources used in the software to support the proposed analysis methodology.

### A. Software Architecture

The software has been developed as a Web-based tool in Java, and integrates modules to access available external web-based resources (data sources and analysis tools) and tools developed by the authors research group to extract the gene-disease associations and to assist the user in the subsequent analysis. The first block of external resources to be accessed to start the analysis are the Entrez web services to 1) retrieve the abstracts of scientific articles from PubMed, and 2) query the biological, chemical and medical databases available through Entrez, in particular, Entrez Gene and 3) the MESH database. MESH [12] is used as a dictionary for diseases. The data gathered from Pubmed (papers’ abstracts), and Entrez Gene and Mesh (dictionaries) are input to Publimed Discoverer\(^1\), which performs text mining and build “gene-diseases associations” candidates.

1) **Publimed Discoverer**: Publimed Discoverer works according to the text mining methodology outlined by the authors in [4] to build gene-diseases associations. Publimed Discoverer uses a Vector Space Model (VSM) to represent documents as vectors of inverse document index frequencies, where a VSM space derives from genes, and one from diseases; then 3 subsets are created (documents indexed on gene terms with non null components on the disease space, documents indexed on disease terms with null components on the gene space, documents with non null components on both the gene and diseases spaces). Then document similarity matrices are built for each set, and for the third set two similarity matrices are computed, one taking into account genes components, and one taking into account diseases components. Each of the four similarity matrices is clustered either by k-means or hierarchical clustering, then the resulting classes are analyzed to extract the gene-diseases associations. For each cluster the set of its positive features, i.e., the terms occurring in a cluster with a frequency above a set threshold is computed; then associations between features are inferred by selecting the clusters that share more documents. This method promotes implicit associations over explicit ones, and a method to evaluate them is needed. These associations will be evaluated and further explored by resorting to the microarray data available from the GEO database\(^2\).

2) **The GEO databases**: The GEO database, is a public repository of experimental data; including single and double channel microarrays, and other high throughput non array techniques such as SAGE (serial analysis of gene expression) [13] and proteomic data from mass spectrography; in this paper we focus on microarray data. The Geo dataset is queried using the terms from the current association, i.e., a disease and a gene term, which usually refers to a family of genes; the query is expanded by retrieving from GEO Profile, a database that stores individual gene expression profiles from curated dataset of GEO database, all the terms of the gene family that have been used in microarray experiments. Each result from the query identifies the reference to the dataset, the microarray platform used, and the link to download the data set. Our software currently works only with the Affymetrix microarray and the SMF (Series Matrix File) data format.

3) **Microarray analysis**: Once a microarray has been selected from the query results, its analysis can start. The microarray analysis module reuses some key Java classes from the MEV (MultiExperiment Viewer) software [14]. The first step is to cluster, by K-means, the microarray data to obtain homogeneous gene sets. If the number of genes in the microarray is huge, a first filtering is done based on the variance of the gene expression levels. The software will propose to the investigator to identify in the given

\(^1\)[http://www.i3s-lab.ing.unict.it/PubMedDiscoverer](http://www.i3s-lab.ing.unict.it/PubMedDiscoverer)

microarray all the genes families identified by PubliMed Discoverer, with the assistance of another call to the GEO Profiles to match the gene identifiers (AFF-ID) used in the experiment to the gene family. Once the AFF-ID of each gene is identified, the relevant rows are selected. For each row (and therefore for each gene $G_x$), the cluster containing $G_x$ is analyzed to build the Gene Relevance Network, which represents a set of genes whose expression levels are highly predictive of others genes. Each couple of genes is featured by a correlation coefficient, which is computed by comparing their expression levels. The software provides two ways to compute gene relevance networks that are further filtered to obtain a sub relevance networks.

4) Gene Ontology and David Gene: The final step is to resort to the Gene Ontology database to investigate the biological processes associated to the genes belonging to the identified gene relevance networks. The Gene Ontology (GO) database provides a controlled vocabulary to describe gene product characteristics and gene product annotation data across species. GO allows to define “rules” to functionally link the transcriptional profile of the genes with respect to molecular functions, biological processes and cellular components. GO is subdivided in three categories: 1) Molecular Function (MF): describes cellular activity without details about who carries it out, when, and how; 2) Biological Process (BP): denotes a specific mechanism carried out until completion by an ordered set of molecular functions, and 3) Cellular Component (CC): terms in this category refer to the specific zone in the cell where a gene product is active. The David Gene [15] is a database for annotation, visualization, and integrated discovery and provides a set of tools for investigating biological meaning behind large list of genes.

B. Implementation notes

Since the overall model of the analysis is a predefined workflow, the application has been built based on Sun Wizard API. The Java classes reused from MEV are TMEV, MultipleArrayViewer, ISubmitData, Experiment, IViewer, AbstractAlgorithm, AlgorithmFactory, AlgorithmData and GEOSeriesMatrixLoader. Other important resources were the Entrez Programming Utilities and their SOAP interface.

III. CASE STUDY: BUILDING AND VALIDATING ASSOCIATIONS FOR NEURAL DISEASES

The case study focuses on diseases of the nervous system.

A. Building gene-disease associations

On MESH, we select “Diseases Category”, and restrict to the Nervous System Diseases. This leads to the identification of 114 pathologies, and the list of identified diseases is transferred as an XML file in our software, and included in its dictionary. A similar procedure is applied to build the gene dictionary, based on querying Entrez-Gene database with the expression “Nervous System Diseases”, from which about 50 symbols referring to specific gene families are retrieved.

Through the PubMed Web services, the software downloads the scientific abstracts that are input to Publimed Discoverer. 700 articles are retrieved as a result of the query with nervous system diseases. The computed set of gene-diseases associations is shown in fig. 2.

After obtaining the list of associations, these are closely analyzed for validation by searching for relevant microarrays and representative Gene Networks. Let’s focus on the gene-diseases associations concerning Muscular Diseases involving the $Tnf$ (Tumor Necrosis Factor) gene family.

B. Microarray Data Analysis to derive gene networks

The analysis of the entire genetic dataset of the microarray matrix relevant to the selected disease (in this case Muscular Diseases) focuses on the genes that are correlated with the symbol of the gene involved in the association. Two methodologies are supported by our tool. Both methodologies start from the k-means clustering of the entire dataset, then they follow an independent path. Method N.1 further clusters the data based on the Cluster Affinity Search Technique (CAST) [16], which is an iterative approach that inserts high affinity elements in a given cluster. Affinity is a similarity measure between a gene and all the genes in a cluster, based on the expression profile. The clustered genes are chosen to build the relevance network. In method N.2 the CAST module is substituted by a Principal Component Analysis (PCA), followed by a Significance Analysis for Microarray (SAM) [17], which is a technique available in MEV able to identify significant genes based on a statistical analysis of their differential expression between sample groups. The two methods are used to cross validate the associations, as shown in the following. Analysis of the microarray matrix highlights the genes strongly correlated (according to criteria chosen by the experimenter) with $Tnf$. The first step is clustering the whole microarray matrix (22000x121) by K-Means (parameters: 50 initial clusters, 50

3 The significance is a score assigned to each gene based on change in gene expression relative to the standard deviation of repeated measurements.
iterations, and Euclidean distance). As intermediate result, the individual genes selected after querying the GEO Profile are listed in fig. 3. The clusters where they have been inserted, also shown in fig. 3, will be analyzed to extract the Relevance Network.

Let’s focus on the cluster containing gene 202307_s_at; the cluster contains 62 elements (Fig. 3). Following methodology N.1, the next step is a more accurate clustering by using the CAST module that requires as input an affinity parameter between [0-1] and that must be exceeded by all the genes within the same cluster. By setting this value equal to 0.7 we obtain a further division in 17 clusters. The gene being analyzed shows up in a cluster that contains 52% of the data (32 genes). These genes are therefore selected to build the gene network, shown in fig. 4, consisting of 22 genes and 49 links with a correlation range [0.8 - 1].

The whole set of gene-networks for the genes of the Tnf family for muscular diseases (fig. 3) created by the two above methodologies is shown in fig. 5. Let’s compare the gene networks obtained, respectively, from applying methodology N.1 and N.2, recalling that the networks in method N.2 consist of genes that have been deemed relevant on a statistical basis.

By comparing the networks obtained by the genes under examination, we note that both have the same correlation range [0.8 - 1], that is, the same relevance level. The second network has less genes; however, a closer analysis reveals that it is contained in the first network, because the links are the same. Therefore we can imagine this network as being obtained from the first one by eliminating the non significant genes. This property has been verified for all the cases analyzed for the given pathology. Nonetheless, it was possible also for the other pathologies to build gene networks to accompany the gene-disease association found in the first phase of the analysis.

C. Gene Functional Annotation

To assist biomedical researchers in the process of data interpretation and of understanding of the functional meaning to the analyzed genes we use the Gene Ontology database. The goal is to identify what biological processes, molecular functions and cellular components are more recurrent amongst the genes involved in the relevance networks. This step is conducted by resorting to the DAVID Gene within GO. To follow up on the gene network previously mentioned (fig. 4) with 22 genes with a (correlation) range [0.8-1], fig. 6 shows the percentage of genes with features belonging to each of the three classes in GO.

For 18 out of 22 genes it is possible to attribute terms from the Biological Process category, whereas for 19 out of 22 it is possible to associate cellular components and for all the set of genes it is possible to associate molecular
functions. Still it is necessary to evaluate these terms, by means of statistical tests, namely a $p$ – value computed by Fisher exact test. Fig. 7 shows the results: in particular, it can be noted that “immune response” is a term with a very low $p$ – value, and it occurs in with a frequency of 52% in the list.

In other words, in the list of genes that make up the network, more than 50% are genes involved in the immune response biological process. Therefore, the analysis highlights a relationship between neuromuscular diseases and immune response to be further investigated by biomedical researches.

IV. CONCLUDING REMARKS

We have presented a methodology supported by an integrated web-based environment to assess gene-diseases associations discovered through text mining, by means of gene networks extracted from related microarray experiments, and enriched by functional annotations. The proposed system is a step forward in tools for helping biomedical researches in analyzing complex experiments; in fact existing tools associate diseases to gene networks but do not provide mechanisms for understanding these associations. In fact, our approach not only identifies the gene networks involved in specific diseases, but also investigates the role that these genes play in human organisms. Finally, since the software has a flexible and modular architecture it can be easily integrated with other systems/modules. For instance, the literature text mining module can be substituted with a module that provides known gene-disease associations and aims at analyzing biological processes related to these associations. We are also working on creating common terminologies using semantic web technologies (RDF and OWL) for organizing and storing information in RDF repositories (SESAME [18]). This would allow us to extend the tool in order to include other biological database that uses RDF or OWL or exposes data using XML interfaces.

V. ACKNOWLEDGMENTS

We acknowledge Eng. Davide Palermo for implementing the system, and Dr. Manuela Pennisi (specialist in Neurology) for the support in developing the neurological diseases use case. This paper has been also partially supported by the project “Catania Territorio d’Eccellenza-ICT E1” of the Catania municipality.

REFERENCES


