Semantic Relations for Interpreting DNA Microarray Data

Dimitar Hristovski, PhD[1], Andrej Kastrin[2], Borut Peterlin[2], MD
and Thomas C Rindflesch, PhD[3]
[1]Institute of Biomedical Informatics, Faculty of Medicine Ljubljana, Slovenia
[2]Institute of Medical Genetics, University Medical Centre, Ljubljana, Slovenia
[3]National Library of Medicine, NIH, Bethesda

Abstract

The results from microarray experiments, in the form of lists of over- and under-expressed genes, have great potential to support progress in biomedical research. However, results are not easy to interpret. Information about the function of the genes and their relation to other genes is needed, and this information is usually present in vast amounts of biomedical literature. Considerable effort is required to find, read and extract relevant information from the literature. A potential solution is to use computerized text analysis methods to extract relevant information. Our proposal enhances current methods in this regard and uses semantic relations extracted from biomedical text with the SemRep information extraction system. We describe an application that integrates microarray results with semantic relations and discuss its benefits in supporting enhanced access to the relevant literature for interpretation of results.

Introduction

With the decoding of the human genome and other eukaryotic organisms, molecular biology has entered a new era. High-throughput technologies, such as genomic microarrays can be used to measure the expression levels of essentially all the genes within an entire genome scale simultaneously in a single experiment and can provide information on gene functions and transcriptional networks [1]. However, the successful interpretation of this information for integration into research underpinning biomedical progress is impossible without comparison to the published literature.

The exponential growth of the life sciences literature makes it difficult even for experts to absorb all the relevant knowledge in their field of interest. Sophisticated technologies are needed for effective data acquisition. Automatic text mining techniques are increasingly used to help assimilate online textual resources; however, current approaches, based on term co-occurrence in text, do not provide an adequately expressive representation of content to satisfy the needs of biologists in elucidating microarray data.

In this paper we propose the use of semantic relations (or predications) for interpreting and exploiting microarray data. We suggest that semantic relations are able to convert textual content into structured information amenable to further computation and that such "executable knowledge" underpins the ability to quickly provide valuable information while interpreting microarray results, such as current knowledge about the disease of interest, genes known to be involved in the etiology of the disease, as well as their relationships (ASSOCIATED_WITH, PREDISPOSES or CAUSES). In addition, information can be determined about the relation between the information determined from the microarray and that in the research literature.

We describe the use of SemRep [2] for extracting semantic predications from MEDLINE citations and discuss a tool for accessing and manipulating a database of such predications, which we used to explore the possibility of interpreting the results of a microarray experiment from the GEO repository (GSE8397) [3] on Parkinson disease.

Background

Considerable research has focused on exploiting textual resources (usually MEDLINE citations) as a way of elucidating the results of microarray experiments while investigate the genetic component of specific diseases. A variety of statistical techniques, including document similarity algorithms [4], have been used to identify relevant information in text. Many systems manipulate co-occurring text features (e.g. [5,6]), or text features in conjunction with additional information such as MeSH indexing or structured information from related databases such as the Gene Ontology (e.g. [7]). Although some systems exploit a thesaurus to identify concepts in text [8] or calculate implicit information by identifying terms related through co-occurrence with shared, intermediate terms [9], semantic predications have not been used to manage the research literature relevant to the results of microarray experiments. The Chilibot system [10] extracts information about gene and protein interactions from text, but it is not integrated with microarray results.

Extracting semantic relations with SemRep

SemRep was originally developed to extract semantic predications relevant to clinical medicine from MEDLINE citations. Subsequently, the program was extended to address molecular genetics and pharmacogenomics [11,12,13]. The system is both rule based and symbolic; processing uses (underspecified) syntactic analysis and relies on structured domain knowledge in the Unified Medical Language System® (UMLS)® [14], augmented for molecular genetics and pharmacogenomics. SemRep depends on MetaMap [15] to identify Metathesaurus concepts and is augmented by ABGene [16] to identify gene names, which are mapped to Entrez Gene.

Semantic relations extracted by SemRep are represented as predications consisting of Metathesaurus concepts as arguments and Semantic Network relations as predicates. The relations that SemRep currently addresses are listed below in the form of semantic schemas, which define the well-formed semantic predications SemRep can produce. Arguments (in braces) in the schemas are defined as classes of UMLS semantic types. For example, "{Substance}" includes an array of semantic types, such as 'Amino Acid, Peptide, or Protein', 'Gene or Genome', 'Biologically Active Substance', and Pharmacologic Substance', among several others.

Genetic Etiology: {Substance} ASSOCIATED_WITH OR PREDISPOSES OR CAUSES {Pathology}

Substance Relations: {Substance} INTERACTS_WITH OR inhibits OR stimulates {Substance}

Pharmacological Effects: {Substance} AFFECTS OR DISRUPTS OR AUGMENTS {Anatomy OR Process}

Clinical Actions: {Substance} ADMINISTERED_TO {Living Being}; {Process} MANIFESTATION_OF {Process}; {Substance} TREATS {Living Being OR Pathology }

Organism Characteristics: {Anatomy OR Living Being} LOCATION_OF {Substance}; {Anatomy} PART_OF {Anatomy OR Living Being}; {Process} PROCESS_OF {Living Being}

Co-existence: {Substance} CO-EXISTS_WITH {Substance}; {Process} CO-EXISTS_WITH {Process}

For example the predication "MDB1 CAUSES Autistic Disorder" is extracted from the text … *the loss of Mbd1 could lead to autism-like behavioral phenotypes* … This interpretation is based on the following processing: *Mbd1* has semantic type 'Gene or Genome' and *autism* maps to the concept "Autistic Disorder." *Lead to* is an indicator for the semantic predicate CAUSES. Similarly the predication "MBD1 INTERACTS_WITH HTR2C" is extracted from … *Mbd1 can directly regulate the expression of Htr2c, one of the serotonin receptors*, … on the basis of the identification of the two genes in this text and the verb *regulate* indicating the INTERACTS_WITH.

Methods

Preparing the microarray experiments and results

For this report we selected microarray data for Parkinson disease (PD). A total of 47 Affymetrix HG-U133A CEL files (29 PD patients and 18 controls) were retrieved from the GEO repository (GSE8397) [3]. All computations were carried out in the R software environment for statistical computing using additional Bioconductor packages [17,18]. The normalization of the raw data was performed using the MAS5 algorithm as implemented in `affy` package. Hybridization probes were mapped to Entrez Gene IDs by annotation data in the `hgu133a.db` package. Analysis of differentially expressed genes (DEG) was performed using Welch's t-test from the `multtest` package. The Benjamini and Hochberg method was selected to adjust *p*-values for multiple testing [19]. As a confidence threshold we used an adjusted value of $p \leq 0.01$. A total of 567 DEGs were used for further processing.

Integrated database with semantic relations and microarray results

One of the advantages of our methodology is that we have built an integrated database with the semantic relations extracted by SemRep and the microarray results we processed. We used MySQL to organize and store the data in relational database format. The data is spread among several tables. We store data about the arguments (subject and object) and the semantic relations from the SemRep predications. For each argument we store names and synonyms as well as semantic types. Arguments are UMLS concepts; when an argument is a gene, in addition to the UMLS CUI (Concept Unique Identifier) we also store the Entrez Gene ID as the argument ID. The Entrez ID serves as a link to the microarray results, which are organized in two tables. One is for the general microarray experiment data and the other is for the expression of the genes within a particular experiment.

To allow fast and flexible searching of the integrated database we used Lucene and have built separate text indexes. Lucene is a well known open source information retrieval tool. This way we have a hybrid database. We use Lucene for fast searching and then used the data stored in MySQL when needed.

The tools for searching are web-based and are developed with the Ruby on Rails application development framework. These tools are still under development; later we plan to make them publicly available. The tools provide a flexible way to specify questions that frequently arise in microarray results interpretation. Questions can refer to both semantic relations from the literature and the microarray results.

Results

Numbers describing size of processing

First we provide some numbers to illustrate the size of the processing involved in building the integrated database. We extracted with SemRep semantic predications from 4,928,419 MEDLINE citations published between 2003 and the end of 2008. That is a considerable part of MEDLINE, especially when we consider that in the last few years the genetic research is much more intensive then earlier. 14,126,438 semantic predication instances were extracted, representing 5,212,540 distinct predication types.

Answering questions on Parkinson disease example

We illustrate the capabilities of our methodology on a microarray for Parkinson's disease (PD), a leading neurodegenerative disorder characterized by progressive movement impairments including tremor, muscle rigidity, postural abnormalities and slowness of volitional movements. Cognitive and mood disorders are also common. While PD is rarely due to monogenic genetic predisposition, both genetic and environmental factors likely account for the prevalence of 'sporadic' (idiopathic) Parkinson's disease. Nevertheless, the etiology of the sporadic disease is only partially understood and the relative importance of putative genes contributing to its pathobiology within populations of different ethnic origins remain obscure.

PD has so far been associated with more than 200 genes, predominantly on the basis of implication of these genes in known mechanisms of disease. On the other hand microarray experiments provide hypothesis-free results on differential expression of genes in PD patients compared to controls. Microarray expression studies are typically characterized by the low signal to noise ratio, which makes interpretation of results challenging. Typically, the relevance of the differentially expressed genes is evaluated by their function (GO) and participation in metabolic networks associated with disease pathobiology.

Our system provides the means for independent evaluation of microarray results based on text mining and additionally improves analysis by including the evaluation of relations among biomedical concepts.

In the following example we demonstrate the type of information the user gets by using our system. We were interested in the interpretation of a microarray experiment (GEO GSE8397) in PD [20]. First, we might search for any genes associated with PD from the literature. We get the well known genes such as *alpha-Synuclein* or *LRRK2* at the top of the list:

alpha-Synuclein	**CAUSES**	**Parkinson Disease**
Alpha-synuclein mutations that **cause** familial **Parkinson's disease** (PMID: 16959795)		
The new mutation, E46K, of **alpha-synuclein causes Parkinson** (PMID: 14755719)		
Alpha-synuclein is known to be the major **cause** of **Parkinson's disease** (PMID: 18282005)		

LRRK2	**CAUSES**	**Parkinson Disease**
Mutations in **LRRK2** are the single most common known **cause** of **Parkinson's disease** (PMID: 19006185)		
Mutations in five PARK genes (SNCA, PARKIN, DJ-1, PINK1, and **LRRK2**) are well-established genetic **causes** of **Parkinson disease** (PMID: 18704525)		

Next, we might want to investigate genes that were differentially expressed in the experiment and have been directly associated with PD already. For example:

SNCA	**ASSOCIATED_WITH**	**Parkinson Disease**
Genetic variability within **SNCA** has been implicated **in idiopathic PD** (PMID: 15455394)		
Variation in **SNCA** expression may be critical **in** common, genetically complex **PD** (PMID: 18669654)		

We can also search for relations between the up- or down-regulated genes on the microarray and other concepts. For example, the top 300 mostly up-regulated genes appear as arguments in over 20,000 semantic relations.

On the other hand we might be interested in genes, that have not yet been directly associated with PD, but are associated with PD via some biological concept. If we formulate an appropriate query we get information on intermediate concepts which are linked to genes that are differentially expressed on the one hand and to PD on the other. Moreover, our system delivers not only the intermediate concept, but also the type of relation extracted from the literature. For example, if one is interested in the TNF gene as the intermediate concept, our system provides the following information:

KRAS	**STIMULATES**	**TNF**
We found that HCV core and **NS3** proteins **induced** **TNF-alpha** (PMID: 17595379)		

TNF	**ASSOCIATED_WITH**	**Parkinson Disease**
Tumor necrosis factor alpha (TNFalpha) is toxic to dopamine neurons and increased levels of **TNFalpha** are observed **in Parkinson's disease** (PMID: 18482714)		
Tumor necrosis factor alpha (TNF-alpha) is a key inflammatory cytokine and several studies linked increased **TNF-alpha** to dopaminergic cell death **in PD** (PMID: 18930140)		

In other words our system provides two pieces of information: (1) NS3 protein (product of KRAS gene) induced expression of TNF-alpha gene and (2) Tumor necrosis factor alpha is toxic to dopamine neurons (the main neuron population affected in PD) and that increased levels of TNFalpha are observed in PD. While the KRAS gene has not yet been directly associated with PD, there is growing evidence that neurodegeneration and tumorigenesis processes are interconnected [21,22].

Similarly, the user can be informed about all intermediate concepts which might have been already directly associated to a given disease or could present no contribution to an understanding of the molecular pathology of disease.

Conclusion

In this paper we presented an application that integrates the results of microarray experiments with a large database of semantic predications representing the content of nearly 5 million MEDLINE citations. We discuss the value of this system with examples from microarray data on Parkinson disease, illustrating the way semantic relations elucidate the relationship between current knowledge and information gleaned from experiments.

References

1. Cordero F, Botta, M, Calogero RA. Microarray data analysis and mining approaches. Brief Funct Genomic Proteomic. 2007;6(4):265-281.
2. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462-77.
3. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: Mining tens of millions of expression profiles - database and tools update. Nucleic Acids Res. 2007;35(Database issue):D760–D765.
4. Shatkay H, Edwards S, Wilbur WJ, Boguski M.: Genes, themes and microarrays: using information retrieval for large-scale gene analysis. Proc Int Conf Intell Syst Mol Biol. 2000:317-28.
5. Liu Y, Navathe SB, Civera J, Dasigi V, Ram A, Ciliax BJ, Dingledine R. Text mining biomedical literature for discovering gene-to-gene relationships: a comparative study of algorithms. IEEE/ACM Trans Comput Biol Bioinform. 2005 Jan-Mar;2(1):62-76.
6. Yen YT, Chen B, Chiu HW, Lee YC, Li YC, Hsu CY. Developing an NLP and IR-based algorithm for analyzing gene-disease relationships. Methods Inf Med. 2006;45(3):321-9.
7. Yang J, Cohen AM, Hersh W. Automatic summarization of mouse gene information by clustering and sentence extraction from MEDLINE abstracts. AMIA Annu Symp Proc. 2007 Oct 11:831-5.
8. Jelier R, 't Hoen PA, Sterrenburg E, den Dunnen JT, van Ommen GJ, Kors JA, Mons B.

Literature-aided meta-analysis of microarray data: a compendium study on muscle development and disease. BMC Bioinformatics. 2008 Jun 24;9:291.

9. Burkart MF, Wren JD, Herschkowitz JI, Perou CM, Garner HR. Clustering microarray-derived gene lists through implicit literature relationships. Bioinformatics. 2007 Aug 1;23(15):1995-2003. Epub 2007 May 30.

10. Chen H and Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics. 2004 Oct 8;5:147

11. Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. AMIA Annu Symp Proc. 2003:554-8. 2003:554-8.

12. Masseroli M, Kilicoglu H, Lang FM, Rindflesch TC. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. BMC Bioinformatics. 2006 Jun 8;7:291.

13. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput. 2007:209-20.

14. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical lan-guage System: An informatics research collaboration. J Am Med Inform Assoc 1998 Jan-Feb;5(1):1-11.

15. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc AMIA Symp. 2001;17-21.

16. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002;18(8):1124-32.

17. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2008.

18. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J. Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. 2004, 5(10):R80.

19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc B. 1995;57(1): 289-300.

20. Moran LB, Duke DC, Deprez M, Dexter DT, Pearce RK, Graeber MB. Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics. 2006 Mar;7(1):1-11.

21. Inzelberg R, Jankovic J. Are Parkinson disease patients protected from some but not all cancers? Neurology. 2007 Oct 9;69(15):1542-50.

22. Staropoli JF. Tumorigenesis and neurodegeneration: two sides of the same coin? Bioessays. 2008 Aug;30(8):719-27.