Luis M. Rocha
Los Alamos National Laboratory, MS B256
Los Alamos, NM 87545, USA
E-Mail: rocha@lanl.gov
WWW: http://www.c3.lanl.gov/~rocha

Integrative Technology for Bioinformatics

This document outlines the research directions in Information Retrieval for Bioinformatics being pursued by Luis M. Rocha and the project teams of the *Active Recommendation Project* and of the LDRD-ER project FY02-CSSE-012: *Identification of interests, trends and dynamics in Document Networks*.

The production of larger and larger databases in molecular biology, particularly those containing genomic data, have lead to a strong interest in Bioinformatics and Computational Biology due to the obvious need to analyze and understand such large collections of data. In particular, microarray technology, with its ability to measure the expression patterns of thousands of genes simultaneously, presents researchers with formidable data analysis difficulties.

The first wave of methods employed to analyze gene expression data were brought in from the fields of data-mining, machine learning, and statistics [e.g. Eisen et al, 1998; Alter et al, 2000; Hastie et al, 2000]. These methods are typically used to discover patterns of expression behavior associated with subsets of genes, which are thus identified. But this analysis is pursued using exclusively the numerical expression values obtained from microarray experiments. Therefore, they cannot directly help us in deriving functional knowledge. The biological reasons for the patterns identified by these techniques must ultimately be ascertained by biologists who need to be able to integrate knowledge about a large number of possible underlying biological mechanisms. Given the large number of genes in microarrays and the myriad possible networks of cellular interaction, this is a daunting task indeed.

Recent renewed interest in Systems Biology has lead researchers in Bioinformatics to the idea that in general, no single set of measurements, data analysis method, or single research team will be sufficient to understand complex biological networks of vast size [e.g. Kanehisa, 2000; Kitano, 2000; Eckardt, 2001]. Instead, this research needs to be carried out by interdisciplinary teams empowered with Informatics technology capable of automatically integrating the results of pattern recognition analysis of microarray data, with available sources of functional knowledge. Clearly, such integrative technology does not aim to replace biologists, but rather to assist them by reducing the number of possible explanations of functional behavior.

One of the most promising avenues to develop such integrative technology, lies in the application of modern *Information Retrieval* (IR) and *Knowledge Management* (KM) algorithms to databases with biomedical publications and data [Masys, 2001]. Modern information resources can be thought of as networks of documents. The prime example of a Document Network is the World Wide Web (WWW). But many other

types of such networks exist: bibliographic databases containing scientific publications (e.g. MEDLINE: http://www.nlm.nih.gov), preprints (e.g. the e-Print Arxiv @ LANL http://xxx.lanl.gov/), as well as databases of datasets used in scientific endeavors (e.g. GenBank: http://www.ncbi.nlm.nih.gov/Genbank/and PROSITE: http://www.expasy.org/prosite/). Each of these databases possesses several distinct relationships among documents and between documents and semantic tags or indices that classify documents appropriately. For instance, documents in the WWW are related via a hyperlink network, while documents in bibliographic databases are related by citation and collaboration networks [Newman, 2000].

Furthermore, documents can be related by semantic information[1] about their content, including keywords and other types of annotations. In bioinformatics, we have access to many types of semantic annotations, for instance: The HUGO Nomenclature for human genes (http://www.gene.ucl.ac.uk/nomenclature), GenBank accession numbers for gene sequences, Medical Subject Headings (MeSH: http://www.nlm.nih.gov/mesh/meshhome.html), etc. Many research groups are beginning to exploit this semantic information, such as the three examples below:

1. Masys et al [2001] created a linking utility for interpreting gene groupings obtained from gene expression analysis. This proof-of-concept utility (http://www.array.ucsd.edu) allows users to obtain the conceptual similarity of groups of genes by generating concept hierarchies of keywords defined in MeSH. The ability to characterize genes with the MeSH hierarchy of terms is achieved by the automatic analysis of publications retrieved from PubMed/MEDLINE. First, the publications indexed by the GenBank accession numbers of the input group of genes are retrieved, then the frequency of keywords from a vocabulary which are found in these publications is collected, displayed, and further refined by the MeSH concept hierarchy.
2. Jenssen et al [2001] created a more comprehensive annotated gene network based on gene co-occurrence, available as a web utility (http://www.PubGene.org). The PubGene network was constructed from the automatic analysis of 10 million documents published in MEDLINE. They first identified a set of acceptable terms for 13,712 human genes as defined by HUGO, LocusLink, GENATLAS, and the Genome Database. An associative network for this set of genes was created by establishing a link between pairs of genes when they co-occur in the titles and abstracts of the set of documents retrieved[2]. The strength of the link is the number of documents where the co-occurrence is observed. Finally, the genes on this network are semantically annotated with the MeSH keyterms of the publications where they occur. PubGene thus allows users to find the similarities of groups of genes according to the "publication space", and it can furthermore offer a semantic characterization of the genes based on MeSH keyterms.
3. Shatkay et al [2000] pursued a different approach to compute gene similarity from publications. They tried to avoid the problem of detecting acceptable gene names used in documents, which affects the previous two approaches, as we discuss in more detail below. Instead, they consider two genes to be similar (in functionality) if the sets of documents retrieved for each are similar. Their similarity measure is based on the probability of pairs of genes (and associated keyterms)

---

[1] The term "semantic" is used in IR to refer to any auxiliary information about meaning or functionality, usually in the form of keywords, but generally in some form of meta-data that can be as complicated as knowledge representations such as conceptual graphs, frames, etc..

[2] Notice that only 7,512 genes have any neighbors in this network, 710 have associated publications but no neighbors, and 5,490 were not found in any publications, 5,202 of which have the status of 'reserved' or 'provisional' in the vocabularies used.

co-occurring in a document, rather than absolute co-occurrence as in PubGene of Jenssen et al [2001]. This approach is very useful to both identify literature associated with a cluster of genes, but also to predict associations between genes without experimental microarray measurements. However, instead of choosing a specific gene name nomenclature, currently, their approach depends on the definition by experts of kernel documents associated with each gene.

The first two approaches described above [Masys et al 2001; Jenssen et al 2001] faced the known problems of synonymy and polysemy plaguing keyterm analysis in IR [Masys, 2001]. Synonymy means that several keyterms can refer to the same item (e.g. gene), and polysemy means that the same keyterm can refer to several items. To evaluate their network, Jenssen et al [2001] manually studied the validity of a set of gene associations. Of 500 randomly chosen pairs of genes with more than 5 co-occurrences, 29% were incorrect, mostly because the same keyterm is used to identify more than one gene, or a gene keyterm is also used to refer to some other entirely different concept.

The approach of Shatkay et al [2000], tries to avoid these issues by not using ambiguous gene nomenclatures, but rather expert-defined kernel documents for individual genes. The computation of similarity measures from document vectors defined by occurrence probabilities is also much more in line with IR's methodology for dealing with linguistic ambiguity. Indeed, at LANL we have been using analogous measures of similarity[3] in our Active Recommendation Project [Rocha, 1999a, 1999b] for recommending scientific documents to users interested in sets of keyterms. We have furthermore used IR techniques such as Latent Semantic Analysis [Berry et al, 1995; Landauer, Foltz, and Laham, 1998] which are particularly good at disambiguating different usages of the same keyterm by accounting for indirect relationships. For instance, the different senses of the word 'Java' in a collection of web pages can be discerned because Java tends to occur with words such as 'programming' and 'computer' in a certain set of documents (about Java as a computer language) and with other words such as 'coffee' and 'Starbucks' in another set of documents. Thus, the automatic analysis of several orders of indirect relationships allows us to discern the several senses of a keyterm.

Our research in this area aims to improve the linguistic ambiguity errors found in the type of approaches pursued by Masys et al [2001] and Jenssen et al [2001], as well as reducing the dependence on human experts in the type of approach pursued by Shatkay et al [2000]. To achieve this, we follow two interacting lines of research.

1. ***Analysis of latent associations in networks of keyterms extracted from corpora of biomedical publications***. The methods described above for the automatic discovery of gene associations from published literature tend to rely exclusively on keyterms describing genes (e.g. PubGene). But to automatically disambiguate many of the synonymy and polysemy errors inherent in gene nomenclatures, one needs to include in the analysis methods all keyterms used in publications, as well as other relational information available in publications, such as citations. We are pursuing the following three avenues along these lines:
   a. *Latent Semantic Analysis* (LSA) on an index of documents to a much more comprehensive set of keyterms than just gene keyterms, should reduce the errors due to linguistic ambiguity inherent in the methodologies of Masys et al [2001] and [Jenssen et al [2001]. This analysis

[3] Which we refer to as proximity measures for reasons detailed in [Rocha, 2001a].

discovers the chains of indirect associations between gene keyterms and other keyterms, thus discovering the several usages of ambiguous gene keyterms. We are working on creating gene networks whose strength of association of gene pairs does not depend solely on gene keyterm co-occurrence, but also on indirect co-occurrence with other terms found in the text of publicationsas identified by LSA.

b. *Citation and Collaboration Network Analysis*. Citation networks among articles have been shown to be very useful in determining document relevance as well as discerning different topics in corpora of documents [Kleinberg, 1998]. The same applies to Collaboration Networks (who writes papers with whom) [Newman,2000]. This kind of analysis of the networks of documents associated with particular sets of genes should allow us to identify automatically both kernel documents needed for the methodology of Shatkay et al [2000], as well as distinct usages of gene keyterms (extracted from the clusters of documents found in citation and collaboration networks) useful to reduce the errors due to linguistic ambiguity inherent in the methodologies of Masys et al [2001] and [Jenssen et al [2001].

c. *Metric Behavior of Networks*. Another methodology we have been developing [Rocha, 2001b] at LANL aims at discovering strong indirect associations among keyterms and documents in associative networks of the type derived by Shatkey et al [2000]. In this approach, we produce a distance function from the similarity measure of a network and then extract pairs of nodes which observe very high semi-metric behavior. High semi-metric behavior for a pair of nodes, means that there exists an indirect pathway between the pair of nodes, with a distance value much smaller than the direct distance. We have collected evidence that semi-metric pairs are correlated with the subject matter of a collection of documents. Thus, we expect this metric analysis to also help reduce the errors due to linguistic ambiguity described above.

2. ***Collaborative Environments and Recommendation Systems***. The methodology of Shatkey et al [2000] for automated analysis of biomedical databases is very attractive since it does not rely on ambiguous subject nomenclature. In this sense, it can be more easily automated and potentially more error free. However, it still relies on expert definition of kernel documents for every single gene. One way to loosen the dependence on human experts is to integrate the behavior of communities of users as they retrieve publications from the databases under analysis. This is the fundamental idea behind collaborative environments and recommendation systems, such as those we have been developing at LANL [Rocha and Bollen, 2001]. Indeed, in our approach, we use a network of similarity values between keyterms derived from scientific publication databases, very much like the one used by Shatkey et al [2000]. But in our system, this network is just the first iteration of an integrative and adaptive recommendation process [Rocha, 1999b, 2001a]. By integrating the retrieval behavior of a community of users and adapting the similarity values in the network, we can discover (without access to experts) important (kernel) documents given a set of keyterms, as well as related keyterms, documents and even groups of users. We can also disambiguate different senses of keyterms appropriate for a subset of users or another. Given the existence of several digital libraries of biomedical publications at LANL[4], we have access to the retrieval logs of the community of scientists at LANL engaged in biomedical research. We are currently working on tapping this source of information with our recommendation algorithms to improve the state of the art in automated analysis of biomedical databases.

[4] For instance: BIOSIS, Science Server, as well as access to copies of PubMed.

It should be emphasized that the IR techniques here detailed are not meant as a substitute for pattern analysis of microarray expression experiments, nor for human expertise in Biology. Rather, we propose the development of these IR techniques for Bioinformatics as a complement to both these two sources of knowledge. Clearly, pattern recognition methods can discover expression relationships amongst groups of genes, but cannot by themselves reveal underlying biological causes or function. Furthermore, expert biologists are easily overwhelmed trying to grasp the biological causes of the groupings discovered by pattern recognition methods, due to the sheer volume of genes and potential biological mechanisms involved. Therefore, techniques that recommend possible functional mechanisms and associated literature, can only help biologists by mediating between the results of pattern recognition and scientific explanation available in the literature.

References

Alter, O., P.O Brown, and D. Botstein [2000]."Singular value decomposition for genome-wide expression data processing and modeling." *Proc. Natl. Acad. Sci. USA.* Vol. 97, No. 18, pp. 10101-10106.

Berry, M.W., S.T. Dumais, and G.W. O'Brien [1995]."Using linear algebra for intelligent information retrieval." *SIAM Review*. Vol. 37, no. 4, pp. 573-595.

Eckardt, N.A. [2001]."The New Biology: Genomics fosters a 'Systems Approach and Collaborations between Academic, Government, and Industry Scientists ." *Plant Cell, Vol. 13, 725-734.* .

Eisen, M.B., P.T. Spellman, P.O. Brown, and D. Botstein [1998]."Cluster abalysis and display of geneome-wide expression patterns." *Proc. Natl. Acad. Sci. USA.* Vol. 95, pp. 14863-14868.

Hastie, T., et al [2000]."'Gene Shaving' as a method for identifying distinct sets of genes with similar expression patterns." *Genome Biology.* Vol. 1, No. 2: 0003.1-0003.21, http://genomebiology.com/2000/1/2/research/0003/.

Jenssen, T.K., A. Lægreid, J. Komorowski, and E. Hovig [2001]."A literature network of human genes for high-throughput analysis of gene expression." *Nature Genetics.* V. 28, No. 1, pp. 21 - 28.

Kitano, H. [2000]."Perspectives on Systems Biology." *New Generation Computing.* Vol. 18, 199-216.

Kleinberg, J.M. [1998]."Authoritative sources in a hyperlinked environment." In: *Proc. of the the 9th ACM-SIAM Symposium on Discrete Algorithms*. . pp. 668-677.

Landauer, T.K., P.W. Foltz, and D. Laham [1998]."Introduction to Latent Semantic Analysis." *Discourse Processes.* Vol. 25, pp. 259-284.

Masys, D.R. [2001]."Linking microarray data to the literature." *Nature Genetics.* V. 28, No. 1, pp. 9-10.

Masys, D.R. et al [2001]."Use of keywords hierarchies to interpret gene expression patterns." *Bioinformatics.* Vol. 17, no. 4, pp. 319-326.

Newman, MJ [2000]."The structure of scientific collaboration networks." *Proc.Nat.Acad.Sci.*. No. 98, pp. 404-409.

Rocha, Luis M. [1999a]."TalkMine and the Adaptive Recommendation Project." In: *the Proceedings of the Association for Computing Machinery (ACM) - Digital Libraries 99. U.C. Berkely, August 1999*. . pp. 242-243.

Rocha, Luis M. [1999b]."Evidence sets: modeling subjective categories." *International Journal of General Systems.* Vol. 27, pp. 457-494.

Rocha, Luis M. [2001a]."TalkMine: A Soft Computing Approach to Adaptive Knowledge Recommendation." In: *Soft Computing Agents: New Trends for Designing Autonomous Systems*. V. Loia and S. Sessa (Eds.). Springer-Verlag.

Rocha, Luis M. [2001b]."Identification of interests, trends and dynamics in Document Networks." *Los Alamos National Laboratory Internal Report for LDRD-ER FY02-CSSE-012*. LAUR 01-2380.

Rocha, Luis M. and Johan Bollen [2001]."Biologically motivated distributed designs for adaptive knowledge management." In: *Design Principles for the Immune System and Other Distributed Autonomous Systems*. Cohen I. And L. Segel (Eds.). Santa Fe Institute Series in the Sciences of Complexity. Oxford University Press, pp. 305-334.

Shatkay, H., S. Edwards, W. Wilbur, and M. Boguski [2000]."Genes, themes, and microarrays: using information retrieval for large-scale gene analysis." In: *Intelligent Systems for Molecular Biology*. . AAAI Press, pp. 317-328.