Enabling More Sophisticated Gene Expression Analysis for Understanding Diseases and Optimizing Treatments

Donny Soh[1], Difeng Dong[3], Yike Guo[2], Limsoon Wong[3]

[1] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
[2] Imperial College of Science Technology & Medicine, 180 Queen's Gate, London SW7 2BZ, UK
[3] National University of Singapore, 3 Science Drive 2, Singapore 117543

donnysoh@gmail.com, dong.difeng@gmail.com
yg@doc.ic.ac.uk, wongls@comp.nus.edu.sg

ABSTRACT

We survey the progress in the analysis of gene expression data for the purposes of disease subtype diagnosis, new subtype discovery, and understanding of diseases and treatment responses. We find existing works fall short on several issues: these works provide little information on the interplay between selected genes; the collection of pathways that can be used, evaluated, and ranked against the observed expression data is limited; and a comprehensive set of rules for reasoning about relevant molecular events has not been compiled and formalized. We thus envision an advanced integrated framework, and are developing a system based on it, to provide biologically inspired solutions. It comprises: (i) automated analysis and extraction of information from biomedical texts; (ii) targeted construction of known pathways; and (iii) direct hypothesis generation based on logical reasoning on, and tests for, consistencies and inconsistencies of observed data against known pathways.

1. INTRODUCTION

Classification of patient samples is a crucial aspect of cancer diagnosis and treatment, as treatment of this type of diseases is often stratified according to the specific disease subtype and the likely treatment response of the patient. For example, childhood acute lymphoblastic leukaemia (ALL), has as many as 6 different subtypes with differing treatment outcome. Under-treatment causes relapse and eventual death. Over-treatment causes severe long-term side effects. Thus accurate diagnostic subgroup must be assigned upfront to ensure correct intensity of therapy [33]. One of us previously demonstrated a very accurate platform based on gene expression profiling analysis to risk stratify childhood ALL patients [41]. We can see from Figure 1 the tremendous promise of this work—survival rates are increased, side effects are reduced, and significant cost savings are achieved. There is thus considerable excitement in the development of gene expression profiling analysis for the purposes of understanding diseases and optimizing treatment.

In this paper, we first provide in Section 2 a succinct but in-depth review of existing gene expression analysis methods. Our survey spans techniques for disease subtype diagnosis, disease subtype discovery, and treatment response understanding. In particular, we describe the progress in approaches to gene selection, paying special attention to more recent developments such as overlap methods [42; 8], direct group methods [40; 22], and biological network co-clustering methods [38; 15; 14; 37; 13; 18; 39; 35].

Then we discuss in Section 3 critical issues remaining in the effective analysis of gene expression data—viz., these works provide little information on the interplay between selected genes; the collection of pathways that can be analysed against the observed gene expression data is limited; and a comprehensive set of rules for reasoning about relevant molecular events has never been compiled and formalized. We outline our vision for an advanced integrated system that is required to directly address these issues.

Finally, we present in Section 4 our preliminary work to realise the vision. In particular, we are developing an integrated system having the capabilities to: (i) automate analysis and extraction of information from biomedical texts, (ii) automate targeted construction of known pathways and circuits, and (iii) reason logically and test for consistencies and inconsistencies of observed data with respect to known pathways and circuits. We hope to enable more biologically inspired interpretations of gene expression profiles, so as to better decipher the underlying causes of a disease, and the reasons for a drug to be effective or ineffective.

2. ACCOMPLISHMENTS OF THE PAST

We summarize here accomplishments of the past with respect to gene expression analysis for the purposes of diagnosing disease subtypes (Subsection 2.1), discovering new disease subtypes (Subsection 2.2), and understanding the genetic and molecular causes of a disease (Subsection 2.3).

2.1 Diagnosing Disease Subtypes

Each disease and its various subtypes have their underlying causes, which may have different down-stream effects that are useful as diagnostic indicators. These down-stream effects are often manifested as consistent gene expression profile differences in a large number of target genes over the different disease subtypes. The recognition of gene expression profile differences, and their use for diagnosing disease subtypes, has thus become an intensely researched topic.

Treatment	Cost–new cases	Cost–relapses	Total cost
Low-intensity treatment for everyone	$36K * 2000	$150K * 1000	$222M
Intermediate-intensity treatment for everyone	$60K * 2000	$150K * 200	$150M and 50% of patients have side effects
High-intensity treatment for everyone	$72K * 2000	$0	$144M and 90% of patients have side effects
Risk-stratified treatment; viz., low intensity to 50%, intermediate intensity to 40%, high intensity to 10%	$36K * 1000 + $60K * 800 + $72K * 200	$0	$98M

Figure 1: Contemporary approaches to the diagnosis of childhood ALL use an extensive range of procedures that require multi-specialist expertise, generally unavailable in developing countries. Thus, although childhood ALL is a great success story of modern cancer therapy with survival rates of 75–80% in major advanced hospitals, it is still a fatal disease in developing countries with dismal survival rates of 5–20%. About 2000 new cases of childhood ALL are diagnosed in ASEAN countries each year. About 50% of these cases need low-intensity therapy, 40% need intermediate intensity, and 10% need high intensity. Treatment for childhood ALL over 2 years for intermediate-risk costs USD 60k, good-risk costs USD 36k, and high-risk costs USD 72k. Treatment for relapse cases costs USD 150k. As the less developed ASEAN countries generally lack the ability to diagnose the subtypes of their childhood ALL patients, the treatment for intermediate risk case is conventionally applied for everyone, as it maximizes the expected benefit in such a situation; as shown in the table above. The single-test platform based on gene expression analysis developed by Yeoh and colleagues [41] has over 96% accuracy in risk stratification of childhood ALL patients. As shown in the table above, this can result in savings of USD 52M a year yet with better cure rates and much reduced side effects, as the correct intensity of therapy can be applied upfront.

The main approach to this problem is that of supervised learning, as illustrated by the classic paper of Golub and colleagues [12]. The gene expression profiles of patients are collected and labeled according to the disease subtype of the patients. The analysis then proceeds in two main steps. In the first step, those genes that are most differentially expressed or most associated to specific disease subtypes are identified. In the second step, a supervised learning algorithm is applied to the expression profiles of genes shortlisted in the first step to induce a classifier. The resulting classifier is then used for predicting the disease subtypes of future patients based on their gene expression profiles.

A wide variety of test statistics have been proposed for the first step to select relevant genes, which appears to be the more challenging of the two steps. Initially, classical test statistics such as t-statistics,[1] χ^2, and Wilcoxon rank sum test are used. As the number of genes far exceeds the number of samples in typical datasets, more elaborate gene selection test statistics are also developed, such as rank products [5] and sparse logistic regression [6], as well as techniques for assessing false discovery rates [34]. Integrated methods [28; 11], typically involving grouping genes with correlated expression profiles into bins and then selecting representatives

from each bin, have also been used. One of the more interesting recent developments in gene selection techniques is to look for gene pairs with expression values that are highly correlated [32], instead of considering a single gene at a time. This is a reasonable technique because genes and their products generally function as a group in a specific pathway, and thus they expression values should be correlated.

An excellent demonstration of this type of analyses is the diagnosis of childhood acute lymphoblastic leukemia (ALL) subtypes [41]. ALL is the most common form of childhood cancer. It has as many as 6 different subtypes. To avoid under-treatment, which causes relapse and eventual death, or over-treatment, which causes severe long-term side effects, accurate diagnostic subgroup must be assigned upfront so that the correct intensity of therapy can be delivered to ensure that the child is accorded the highest chance for cure [33]. Yeoh et. al. [41] first use χ^2 statistics to select genes that are most associated with each of the ALL subtypes, and then use a support vector machine to learn a classifier for the ALL subtypes from the expression profiles of the selected genes. Their classifier achieves an exceedingly accurate overall diagnostic accuracy of 96%.

2.2 Discovering Disease Subtypes

New diseases and subtypes may emerge over time. A clinician generally detects the emergence of such cases when he is unable to assign a known phenotype to a patient using traditional procedures such as morphology, immunophenotyping, cytogenetics, and molecular diagnostics. Such a new disease subtype may manifest itself at the gene expression level. The recognition of distinguishing gene expression profiles from untyped samples has thus become an important research topic in bioinformatics.

The main approach to this problem is that of unsupervised learning, as illustrated by the classic work of Cheng and

---

[1]The t-statistics of a gene g given two sets P and N of gene expression profiles of patients in two contrasting classes $\mathcal{P}$ and $\mathcal{N}$ is defined as:

$$t(g|P,N) = \frac{|\mu(g,P) - \mu(g,N)|}{\sqrt{\frac{\sigma^2(g,P)}{|P|} + \frac{\sigma^2(g,N)}{|N|}}}$$

where $\mu(g,P)$ and $\mu(g,N)$ are the mean expression values of gene g in P and N respectively, $\sigma^2(g,P)$ and $\sigma^2(g,N)$ are the variances of the expression values of gene g in P and N respectively. Typically those g where $t(g|P,N) > \alpha$, for some threshold α, are considered significantly differentially expressed and are used as feature vectors to train a classifier to distinguish $\mathcal{P}$ and $\mathcal{N}$ samples.

Church [7], although it was not initially used for detecting new disease subtypes. The gene expression profiles of patients are collected, including both patients of known subtypes and unknown subtypes. The analysis then proceeds in two steps. In the first step, those genes whose expression values do not exhibit sufficient variance are removed. In the second step, a biclustering algorithm is applied on the patients and the remaining genes to obtain clusters of patients that share similar expression profiles on the remaining genes. If a cluster mostly contains untyped patients, it is inferred as a new subtype of the disease, and the gene expression profile shared by these patients is proposed as a distinguishing marker for this new subtype.

A large number of clustering methods have been proposed for the second step of this problem [30; 7; 10; 29; 26], which appears to be the more challenging of the two steps. As gene expression data are typically arranged as a matrix, with rows as genes and columns as patient samples, the goal of biclustering is to identify "homogeneous" submatrices from such an input matrix. However, there are many ways to define homogeneity—e.g., biclusters with constant values, biclusters with coherent values, and so on. Most of the biclustering methods applied for this problem design their own measure of homogeneity. As such, they tend to only find clusters that are interesting according to that measure. A recent outstanding development [26] is to look for deficiency of randomness instead of homogeneity. The deficiency of randomness is an insightful generalization as it is a measure that can detect many more types of homogeneous submatrices. This follows because a homogeneous submatrix—regardless of the definition of homogeneity—is one that exhibits some interesting regularity, which implies the lack of randomless.

An excellent demonstration of this type of analysis is again the work of Yeoh et. al. on ALL [41] mentioned in the previous subsection. That study contains 327 patient samples, over 60 of which do not fit into existing ALL subtypes. A biclustering of the expression profiles of the genes selected by Yeoh et. al. is shown in Figure 2. The strong association of different groups of genes for different ALL subtypes is obvious. Moreover, 14 of the samples with unknown subtypes share a novel common distinguishing gene expression profile, as indicated in Figure 2. This novel subtype may be linked to lipoma-associated chromosomal translocation [41].

2.3 Understanding Disease Subtypes

The two subsections above describe the major approaches to infer differentially expressed genes that are useful for diagnosis and discovery of disease subtypes. However, there are a number of fundamental problems. Firstly, the number of patient samples is very small compared to the number of genes. Thus, the statistical significance of the selected genes and the accuracy of the resulting diagnosis system have a high degree of uncertainty. Secondly, the transition from the selected genes to the understanding of the sequences of causative molecular events is unclear.

Let us illustrate these issues using a set of *in vitro* gene expression data on three nasopharyngeal cancer (NPC) cell lines. A drug CYC202 is tested on the three cell lines, CNE1, CNE2 and HK1. Six time points are taken from the three individual cell lines. HK1 responds to CYC202. CNE1 does
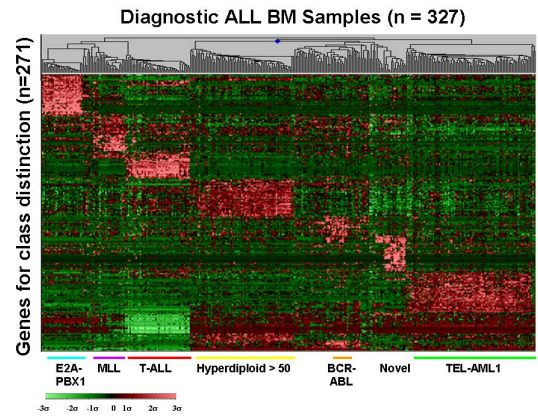

Figure 2: Gene expression profiles of childhood ALL. Each row is a gene. Each column is a patient. The group of patients labelled "novel" is the newly emerged subtype detected by the biclustering [41].

not respond. CNE2 responds in a limited way. Applying the unsupervised approach, or the supervised approach to select genes followed by biclustering on the selected genes, gives a figure similar to Figure 3. We can see clearly that three distinct gene expression profiles are associated with the three cell lines. Even if the selected genes and their expression profiles are truly reliable distinguishing markers for the three cell lines, one does not know how to explain the different ways the cell lines respond to CYC202!
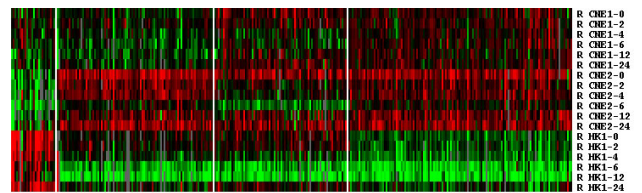

Figure 3: Gene expression profiles of NPC cell lines. Rows are samples. Columns are genes.

In order to qualitatively improve statistical power of the methods described earlier and reliability of the results, and to extend the reach of the predictions, additional dimensions present in the problem have to be brought into consideration. For example, each disease subtype usually has an underlying cause, and thus there should be a unifying biological theme for genes that are truly associated with a disease subtype. Hence the uncertainty in the reliability of the selected genes can be reduced by considering the molecular functions and the biological processes associated with the genes. Such a unifying biological theme is also a basis for inferring the underlying cause of the disease subtype. There are a number of existing approaches to analyze gene expression data with respect to biological context. They can be roughly categorized into three groups [38], viz. the overlap methods, the direct group analysis methods, and the biological network co-clustering methods.

The overlap methods determine what are the biological path-

ways that have a statistically significant overlap—typically the hypergeometric test[2] is used—with the list of differentially expressed genes [42; 8]. The significant pathways are then postulated as basis for inferring the underlying causes of the disease subtypes studied or treatment responses observed. At the same time, those differentially expressed genes that overlap with these significant pathways are used as more reliable markers. An important weakness of these methods is that the initial list of differentially expressed genes is generally defined using test statistics such as those mentioned in Subsection 2.1 with arbitrary thresholds. Different test statistics and different thresholds often result in a distressingly different list of differentially expressed genes. Consequently, the outcome of the whole procedure—i.e., the biological pathways that are significant—is usually not stable with respect to variations in the test statistics and thresholds used in selecting the differentially expressed genes.

The direct group analysis methods [40; 22] determine if a biological pathway is relevant by comparing the distributions of expression values of genes on the biological pathway with the distributions of expression values of all the other genes measured in the experiment. The methods in this family differ from each other primarily in the test statistics used for comparing the two sets of distributions. An outstanding example of this family of methods is gene set enrichment analysis, GSEA [40], which uses a weighted Kolmogorov-Smirnov statistics[3] to compare the two sets of distributions and also

---

[2]The hypergeometric test value for a set W of genes on a pathway $\mathcal{W}$ given two sets P and N of gene expression profiles of patients in two contrasting classes $\mathcal{P}$ and $\mathcal{N}$ is defined as:

$$h(W|P,N) = \frac{\binom{wa}{wd} * \binom{a-wa}{d-wd}}{\binom{a}{d}}$$

where A is the set of genes on the microarray used, D is the set of differentially expressed genes with respect to the contrasting sets P and N, $a = |A|$, $d = |D|$, $wa = |A \cap W|$, $wd = |D \cap W|$. Typically a pathway $\mathcal{W}$ such that $h(W|P,N) < \alpha$, for some threshold α, is considered significant. Those differentially expressed genes belonging to such pathways can then be analysed further to understand the underlying causes of $\mathcal{P}$ vs $\mathcal{N}$, as well as for use as feature vectors to train a classifier for distinguishing $\mathcal{P}$ and $\mathcal{N}$.

[3]The weighted Kolmogorov-Smirnov statistics of a set W of genes on pathway $\mathcal{W}$ given two sets P and N of gene expression profiles of patients in two contrasting classes $\mathcal{P}$ and $\mathcal{N}$ is defined as:

$$KS(W|P,N) = \max_i H(W,i|P,N) - M(W,i|P,N)$$

where

$$H(W,i|P,N) = \sum_{\substack{g \in W \cap A \\ r'(g|P,N) \leq i}} \frac{r(g|P,N)^q}{\sum_{h \in W \cap A} r(h|P,N)^q}$$

$$M(W,i|P,N) = \sum_{\substack{g \in A-W \\ r'(g|P,N) \leq i}} \frac{1}{|A-W|}$$

Here, A is the set of genes in the microarray used, $r(g|P,N)$ is a correlation statistics—e.g., t-statistics $t(g|P,N)$ of a gene g with respect to the contrast sets P and N—and $r'(g|P,N)$ is the ranking of g according to $r(g|P,N)$ among all the genes in A, while q is a control parameter. Note that $KS(W|P,N)$ reduces to the standard Kolmogorov-Smirnov statistics when $q = 0$. A pathway $\mathcal{W}$ is considered significant if the p-value for $KS(W|P,N)$ is significant, where the

uses resampling to estimate false discovery rates. These direct group analysis methods start with a set of genes that are determined *a priori*—namely, those that are on the biological pathway being considered. They are thus less vulnerable to the instability discussed earlier of the overlap methods, which start with a set of genes that are determined *a posteriori* using various test statistics and thresholds. Furthermore, these direct group analysis methods are able to detect more subtle changes in gene expression profiles [38]. For example, if the majority of genes on the biological pathway have small expression level changes, probably none of them can be detected and selected as differentially expressed. The whole group is then missed in the overlap analysis. On the other hand, the high correlation of the changes of expression values in this group of genes can result in high statistical significance of the biological pathway under a direct group analysis method like GSEA.

The biological network co-clustering methods [38; 15; 14; 37; 13; 18; 39; 35] integrate gene expression data with information on events underlying cellular response—e.g., protein interaction, promoter-binding, protein modification—to infer the relevant signaling and regulatory cascades that explain the disease subtypes studied and drug responses observed. Some of them use co-clustering techniques where the distance between genes depends both on the expression profile correlation between the genes and the number of "hops" between the genes in a given biological network [14; 37; 13]. However, the existence of "hubs"—i.e., those proteins or genes with very high connectivity in the biological network—can significantly distort such distance measures. So some methods in this family [39; 35] downplay connections through hubs. The most interesting recent developments along this approach are the network enrichment analysis algorithms [38]. These algorithms actually operate in a way similar to GSEA, but they do not consider the genes in a biological pathway as a whole. Instead, for each regulator in the pathway, all its targets are considered as a group, which is then evaluated in a GSEA-like manner. Not all genes in the biological pathway are expected to be differentially expressed due to the complexity of regulatory events. This splitting into separate regulatory groups can pinpoint the transcriptional regulators whose targets exhibit consistent and significant differential expression pattern, leading to sharper hypotheses that explain the disease subtypes studied or drug responses observed.

3. PLANS FOR THE FUTURE

We now present some of the issues (Subsection 3.1) that still remain in the effective analysis of gene expression data. Then we state our vision (Subsection 3.2) for an advanced system for more sophisticated analysis of gene expression data for understanding a disease and its treatment response. Then we discuss the challenges (Subsection 3.3) in realising the envisioned system.

3.1 Issues

We have seen in Section 2 the tremendous progress already

---

p-value is estimated by repeatedly swapping the members of P and N giving P' and N' and computing the fraction of $KS(W|P',N') \geq KS(W|P,N)$.

made in the analysis of gene expression data. Nevertheless, we still fall short—in at least three ways—of deciphering and understanding the causative events of disease subtypes and treatment responses. Firstly, the functional groups determined by methods such as GSEA are usually too general, contain little information on the interplay between their members, and do not tell us if the selected gene groups have expression values that are consistent or inconsistent with their known underlying pathways. Secondly, the collection of pathways and other information that can be analysed against the observed expression data is still limited. Thirdly, a comprehensive set of rules for reasoning about signaling cascades, regulatory interactions, and other molecular events has never been compiled and formalized.

The first issue calls for continued effort in research on selecting relevant genes that serve as starting points for analysis of disease subtypes and treatment responses. The data mining community has already been paying significant attention to this problem, and we should see continued progress. An advice that we would like to offer in this aspect is that the data mining community should direct their effort in this regard towards gene expression analysis methods that consider gene expression data along with additional information such as promoter-binding, protein modification, protein interaction, and other bio-molecular events forming the machinery that underlies cellular response.

The second issue calls for continued effort in the construction of high quality databases of biomolecular networks. Such databases [21] have traditionally been developed by dedicated curators in a mostly manual manner. Such developments should be continued and expanded. We suspect this may be most effectively carried out by countries or companies that have access to a large pool of scientifically competent curators at an acceptable cost—e.g., Molecular Connections Pvt Ltd in Bangalore, India.[4] More recently, there is also an increased interest in the text mining and natural language processing communities in applying their technologies to extract information on protein modification, protein interaction, and other bio-molecular events from published literature [17; 3; 19], which we hope will eventually lead to a more automated way of constructing and maintaining high quality databases of biomolecular networks.

The third issue calls for the development of a comprehensive set of reasoning rules for thinking about signaling, regulation, and other biomolecular events, as well as the development of inference systems for supporting the use of such rules. Although there is considerable effort in the systems biology arena in developing accurate simulations [16; 25], the development of systems for reasoning about biomolecular events at the logical level appears to be more limited. Such an inference system is likely to involve a significant level of abductive reasoning, as well as some level of deductive reasoning and perhaps some level of inductive reasoning.

3.2 Vision

For more sophisticated gene expression analysis for understanding diseases and optimizing treatments, we envision an integrated system with the following capabilities:

- Automated analysis and extraction of information from

biomedical texts pertinent to the disease being studied and the drug response being investigated.

- Automated construction of known pathways and circuits pertinent to the disease being studied and the drug response being investigated.

- Reliable tests for consistencies and inconsistencies of the observed gene expression data and other test data with respect to these pathways and circuits based on statistics and heuristics.

- Logical inference of the chain of causative events, possible breakages and rewiring of these pathways and circuits, leading to the disease being studied and the drug response being investigated.

3.3 Challenges

Many established repositories—e.g., GOPubMed[5]—focus on data integration and aggregation, where they collect and aggregate the latest abstracts, papers, and findings. SemRep [9; 36; 2] is a step closer to our vision. Nevertheless, it is still basically a repository collection of biological information, extracted from abstracts and papers using natural language processing techniques. In contrast to SemRep, rather than building a grand biological repository, our vision calls for an integration of (i) knowledge from known biological data repositories, (ii) microarray experiment gene expression data being studied, and (iii) domain-specific inference. Such an integration has to be performed in order to support discovery of consistencies and contradictions with known information. After discovering the consistencies and contradictions, we can then generate more informative hypothesis on the causative sequence of events underlying the disease subtypes and treatment responses being studied.

The least explored component of the envisioned system—at least within the bioinformatics community—is the development of the domain-specific inference system in the context of gene expression analysis for understanding disease subtypes and treatment responses. Such an inference system should contain (i) a logical framework for representing biomolecular interactions and events, (ii) a set of domain-specific reasoning rules or heuristics, and (iii) support for abductive, deductive, as well as inductive reasoning. In addition, there should preferably be some support for extracting relevant information from biomedical literature and converting them into the logical framework for representing biomolecular interactions and events. We outline here the challenges to realise our envisioned system.

Challenge 1

As mentioned earlier, current data silos are generally pure repositories of data. Their task is to make data publicly available. However, such databases are more suited towards document retrieval, making knowledge extraction and integration difficult and complicated. As most of the information stored is in text format, to effectively extract knowledge from such databases, we will need text mining and natural language processing techniques to extract the required information. Although the concept of applying text mining and

---

natural language processing to biological text is not novel, such techniques have been limited to establishing biomolecular interactions through co-occurrence or shallow parsing, and not detailed relationships between the genes and proteins or conditions underlying the interactions [17]. For instance, consider the geneRIF [31] entry below:

- *Chk2 phosphorylates and activates E2F1 in response to DNA damage, resulting in apoptosis.*

Conventional natural language processing techniques are able to tag both the entities Chk2 and E2F1 as proteins and interpret them as a protein-protein interaction pair due to their co-occurrence in the text. However, we wish to go more in-depth and obtain the following relationships:

- *activated(Chk2) causes activated(E2F1) provided DNA-damage*

- *activated(E2F1) causes apoptosis*

Challenge 2

As one of the objectives is understanding treatment response, specialized metrics, heuristics, and rules are needed to determine and differentiate responses between responsive patients and non-responsive patients. Such metrics, heuristics, and rules have to:

- locate similarities and differences in the expression values of genes across responsive and non-responsive patients;

- locate similarities and differences in gene relationships across responsive and non-responsive patients;

- locate similarities and differences in the expression values of genes within responsive and non-responsive patients;

- locate similarities and differences in gene relationships within responsive and non-responsive patients; and

- use the information above to pinpoint known individual gene regulatory relationships that are observed and those that are contradicted by the gene expression data studied.

Challenge 3

Knowing which individual gene regulatory relationships expected in normal pathways are observed in or contradicted by the gene expression data studied may still not be sufficient for one to identify the sequence of causative events underlying a disease or a drug response. We therefore need a logical reasoning framework to allow us to chain together the individual observed gene regulatory relationships so that the relevant activated or rewired signaling cascades can be traced, hypothesized, and constructed.

In addition, a significant portion of the known molecular circuits is at the protein level. In contrast, gene expression data primarily reflect actions at the RNA level and at the transcription factor-DNA interface. Unfortunately, much less information is known at the RNA level and at the

transcription factor-DNA interface. Hence we also need a logical reasoning framework that allows us to make inferences across all three levels, and especially to make cross inference between the known protein-level circuits and the observed gene expression data.

Such a logical reasoning framework should contain a knowledge base of known biological phenomena at all the three levels of molecular circuitries mentioned above. It should satisfy the following criteria:

- Biologically sound—no matter how strong a model is theoretically, it is difficult for strong relevant conclusions to be drawn unless the model is biologically relevant.

- Flexible—to allow the framework's knowledge base to be built from different and diverse sources. Regardless of the original format of the biological literature and whatever the microarray gene expression analysis techniques used, they must be representable in the framework.

- Rules discrimination—to choose between different, possibly conflicting, set of logical rules and observations. For instance, we might have two sets of rules which are biologically mutually exclusive. Yet they give us the same conclusion. The framework should be able recommend which set of rules best described the experiment given the conclusion.

4. EXPERIMENTS OF THE PRESENT

We describe in Subsection 4.1 our preliminary explorations toward an advanced integrated system for the analysis of gene expression profiles for understanding disease subtypes and treatment responses. Then we discuss in Subsection 4.2 our thoughts on the less explored area of developing a logical model and reasoning system for biomolecular events.

4.1 A Statistics/Heuristics-Based Framework

To address the issues discussed earlier, we are experimenting with an integrated system having the following capabilities: (i) automated analysis and extraction of information from biomedical texts, (ii) targeted construction of known pathways and circuits, and (iii) tests for consistencies and inconsistencies of observed data with respect to these pathways and circuits based on statistics and heuristics. Figure 4 summarizes the integrated system we are developing. Its four major components are represented by the red boxes on the left of the figure—viz., information gathering, processing engine, biological knowledge, and correlation combination. We describe them below:

Information Gathering

We currently grab data from one main source, NCBI,[6] pulling out the required geneRIFs [31] and paper titles. We concentrate on just using geneRIFs for the time being. As the data in NCBI is well structured, gathering of such information is not difficult. Scripts were written to query directly the NCBI database. The scripts consist mainly of two
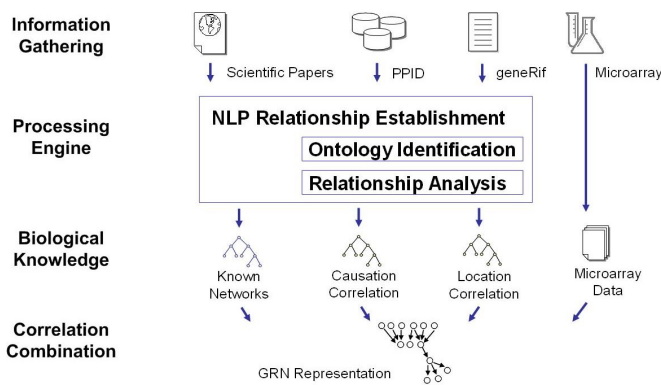
---

[6]http://www.ncbi.nlm.nih.gov

Figure 4: An integrated framework for analysis of molecular and genetic causative events in disease subtypes and treatment responses.

posts to the database. The first post[7] queries NCBI with the gene name, allowing us to obtain the NCBI GeneID for that particular gene. Another query will then be posted[8] to NCBI, querying it with the GeneID obtained earlier on. After which we use a series of regular expressions to obtain all the geneRIFs for further processing.

Processing Engine

We then process the geneRIFs by identifying the genes and their protein products. It is not difficult to map to the genes mentioned in the geneRIFs using the gene list from the microarray experiment being analysed. After we have successfully identified the genes and protein products in the geneRIFs, we use the snowball [1] algorithm—for now—to extract individual relationships between the genes and their protein products.

We constraint every relationship between two genes to take the form depicted in Figure 5. Specifically, a gene or its protein product can fall into the following four main groups in any reaction: "regulator", "co-regulator", "regulatee", and "co-regulatee"; and a gene or its protein product can participate in four main types of reactions: "activates", "inhibits", "non-activates", and "non-inhibits".

We used the snowball algorithm here because it has been shown to be effective in determining relationships between object pairs within sentences, and has been reported to reach an accuracy of 88% [1]. However the snowball algorithm ignores many semantics aspects of natural language, and thus may miss out on many gene relationships that are within the text. In particular, the algorithm can fail poorly when there are multiple genes within the same geneRIF. Wrong associations and relationships can be developed consequently. Hence, we plan to require that the text be processed using certain natural language processing themes, and we

---

[7]The post is of the form `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=search&term=`X, where X is the gene name.

[8]The post is of the form `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=full_report&list_uids=`Y, where Y is the GeneID.

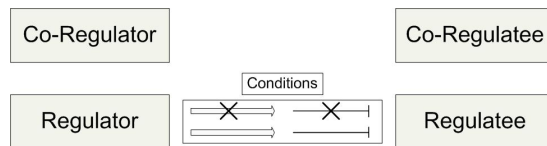are currently experimenting with the Stanford Parser.[9]

Figure 5: Relationships between genes and proteins to be extracted from geneRIFs.

Biological Knowledge

The individual relationships extracted earlier are fused into a form of Boolean gene regulatory network (GRN). We have thus created a small repository of biological information from available biological knowledge.

Correlation Combination

With this repository of information, we next process the microarray data. For example, to understand patients' responsiveness to a drug, we first group the gene expression data according to patients' responsiveness to the drug. With the repository of reaction pairs between a regulator g_i and a regulatee g_j in the GRN, we calculate three correlation metrics between g_i and g_j:

- The correlative relationship between g_i and g_j across the two groups of patients. In particular, we capture the situation where the expression values of g_i and g_j are correlated in the first group of patients but not in the second group; or the situation where the expression values of g_i and g_j are correlated in the second group but not in the first group.

- The correlative relationship in g_i (g_j) across the two groups of patients. In particular, we capture the situation where the expression values of g_i (g_j) in the first group are not correlated with the expression values of g_i (g_j) in the second group.

- The correlative relationship in g_i (g_j) within each group. In particular, we want to capture the situation where the expression values of g_i (g_j) are correlated to each other within each of the two group.

These three metrics allow us to find relationships between g_i and g_j that are distinctly different only across the two contrasting groups and show similarity in behaviour within groups.

Finally, we combine these three metrics with a Bayesian mixture model to yield a single score on g_i and g_j. A high score suggests that there may be a regulatory relationship between g_i and g_j that has undergone a change in the two contrasting groups of patients. By comparison with biological information captured earlier in our repository, we will uncover relationships that are consistent or contradictory to our repository of biological knowledge. For example, if "g_i

---

[9]`http://nlp.stanford.edu/downloads/lex-parser.shtml`. See also Klein and Manning [23; 24].

inhibits g_j" is in our repository, and a high score is given to the pair g_i and g_j by the Bayesian mixture model, then it is likely that "g_i inhibits g_j" is observed in one of the patient groups but is not observed in the other patient groups.

This proposed analysis procedure is similar in strategy to network enrichment analysis [38], where one starts with a known regulator and its regulatees as a group. However, network enrichment analysis methods generally test if the whole group share a common gene expression change pattern. In our case, we test if a regulatory pattern between a regulator and a regulatee is broken or changed. Hence our test can directly detect if a pathway is behaving normally and directly hypothesize possible breakages in a pathway. Such a test is useful for diseases such as cancers, because cancer cells can hijack normal pathways and rewire them around various checkpoints.

4.2 A Logic-Based Framework

The analysis technique described in Subsection 4.1 can pinpoint individual gene regulatory relationships that are observed and those that are contradicted by the gene expression data studied. However, it still leaves the researcher to chain together these individual relationships to trace, hypothesize, and construct the pertinent activated or rewired signal cascades and other causative events. We outline here a reasoning framework for making inferences across interactions between proteins, RNAs, and genes.

Our reasoning framework has the following components: observables, events, rules, and reasoning. These components are explained below.

Observables

We need to represent several types of "observables". An observable is some state or property that a researcher measures or determines at a system-wide level or individual molecule level. For example, apoptosis is an observable that can be determined by TUNEL assays, the expression level of a gene can be determined by a microarray, and so on.

The observables at a system-wide level—such as DNA damage, apoptosis, etc.—are represented as a formula of the form $state(system, X)$, where X is what is observed. For brevity, we write X to mean $state(system, X)$. For example:

- *DNA-damage*

- *apoptosis*

The observables on the protein complex, protein, RNA, and DNA levels are represented as a formula of the form $state(X, S)$, where X is a protein complex, protein, RNA, or DNA; and S is the observed state or property of the molecule, such as activated, inhibited, upregulated, downregulated, abundant, degraded, short half life, etc. For brevity, we write $S(X)$ to mean $state(X, S)$. For example:

- *activated(Chk2)*

- *short-half-life(Claspin)*

Events

We need to represent "events" at various levels of molecular circuits. An event is basically a direct or indirect interaction that causes the system or a protein complex, a protein, a RNA, or a DNA to adopt or enter a specific observable state. Thus events capture our knowledge of biology at the molecular level.

We consider two levels of events. A first-order event is represented as a formula of the form *E causes F*, meaning if E was observed then F would be observed. A second-order event is represented as a formula of the form *E causes F provided G*, meaning if G was observed then the first-order event *E causes F* would be observed. For example,

- *activated(E2F1) causes apoptosis*

- *activated(Chk2) causes activated(E2F1) provided DNA-damage*

We also permit multi-molecule version of these events. For example, the binding of cyclinA and cdk2 to form a complex:

- *activated(cyclinA), activated(cdk2) causes activated(cyclinA-cdk2)*

We deliberately use the term "causes" as we do not require the event to represent a direct interaction between molecules. In other words, if E_1 triggers E_2, and E_2 triggers E_3, it is acceptable to model the indirect triggering of E_3 by E_1 as E_1 *causes* E_3, and omit mentioning E_2 altogether. The granularity is left at the discretion of the modeler. This allowance for indirectness is necessary because the current state of knowledge of biology at the molecular level is incomplete. For example, the inhibition of polymerase II can lead to apoptosis, but we do not really know the precise step-by-step chain of interactions underlying it. This representation of events permits us the flexibility of modeling those events that we know in great detail in a more fine-grain way, and those that we do not know in detail in a more coarse-grain way.

Rules

The static knowledge that we capture as events are generally rich in actions at the protein complex and protein levels, but are usually poor in actions at the RNA and DNA levels. In contrast, the observables that we capture from our gene expression data are at the RNA level. So we also need some "rules" to capture some domain-specific reasoning that a biologist may use to make cross inference between the known protein-level circuits and the observed gene expression data. We consider two types of rules: those expressing normal expected behavior and those expressing normally mutually exclusive events.

A rule expressing normal expected behavior is represented in the same way as events—viz., *E causes F* and *E causes F provided G*—however, the events may contain variables or place holders for the actual molecules involved. For example,

- *upregulated(X) causes abundant(X),*

  capturing the general rule that upregulating the expression of gene X leads to abundance of the corresponding protein X.

- *downregulated(X) causes degraded(X) provided short-half-life(X)*,

  capturing the general rule that downregulating the expression of gene X leads to the disappearance of the corresponding protein X if it has a short half life.

A rule expressing observables that are normally mutually exclusive of each other is represented as a formula of the form E *conflicts* F, meaning that E and F are normally not observed together. For example,

- *abundant(X) conflicts degraded(X)*

- *activated(X) conflicts degraded(X)*

- *upregulated(X) conflicts downregulated(X)*

Reasoning

We are now ready to describe the desired "reasoning" to be done in our framework. At the core of the framework is a classical proof system, $\mathcal{D}, \mathcal{R} \vdash \sigma$, depicted in Figure 6. Here, σ is a ground formula to be proved; $\mathcal{R}$ is a set of domain-specific rules; and $\mathcal{D}$ is a set of observables and events.

Let us use the CYC202 responses of our three NPC cell lines from Subsection 2.3 to outline how we intend to use the system for reasoning. Recall that the HK1 cell line is the cell line that responds to CYC202. Suppose we have done a TUNEL assay and show that apoptosis is observed in HK1 after treatment by CYC202, and we desire to know if CYC202 is the cause of apoptosis. Let $\mathcal{R}^*$ be the set of all domain-specific rules, and $\mathcal{D}^*$ be the set of all observables actually observed in our gene expression profiling experiments and other experiments on the HK1 cell line. It is easy to see that CYC202 is probably the cause of apoptosis if there is a target inhibited by CYC202—e.g., cdk2—such that we can find $\mathcal{R}$ as small as possible, $\mathcal{R}'$ as large as possible, $\mathcal{D}$ as small as possible, and $\mathcal{D}'$ as large as possible, satisfying the conditions below:

(1) $\mathcal{R}' \subseteq \mathcal{R}^*$;

(2) $\mathcal{R}' \subseteq \mathcal{R}$;

(3) $\mathcal{D}' \subseteq \mathcal{D}^*$;

(4) $\mathcal{D}' \subseteq \mathcal{D}$;

(5) $\mathcal{D}, \mathcal{R} \not\vdash apoptosis$;

(6) $\mathcal{D} \cup \{inhibited(cdk2)\}, \mathcal{R} \vdash apoptosis$; and

(7) $\mathcal{D} \cup \{inhibited(cdk2)\}, \mathcal{R} \not\vdash E \; conflicts \; F$, for all E and F.

The logic behind the conditions above is as follows. Conditions (1) and (3) allow us to omit some rules, events, and observations that lead to contradictions, corresponding to breakages in the expected normal pathways. Conditions (2) and (4) allow us to insert some new rules and events, as well as hypothesized observations to be subsequently tested, corresponding to rewiring of pathways. Condition (5) shows that the breakages and rewirings do not by themselves lead

to apoptosis. Condition (6) shows that once CYC202 inhibited its target, cdk2, the system goes into apoptosis. Condition (7) shows that the resulting model contains no inconsistency. As apoptosis is only provable when the CYC202 target cdk2 is inhibited, this suggests CYC202 has acted through inhibiting cdk2 in HK1 cell lines. The exact cascade of signaling events can be reconstructed by tracing the proof tree of $\mathcal{D} \cup \{inhibited(cdk2)\}, \mathcal{R} \vdash apoptosis$.

Recall also that the CNE1 cell line does not respond to CYC202 treatment, and we desire to know how this cell line escapes the fate of apoptosis after CYC202 treatment. Let $\mathcal{R}^*$ be the set of all domain-specific rules, and $\mathcal{D}^*$ be the set of all observables actually observed in our gene expression profiling experiments and other experiments on the CNE1 cell line. Suppose cdk2, cdk7, and cdk9 are the only known targets of CYC202. Then the CNE1 cell line probably escapes apoptosis if we can find $\mathcal{R}$ as small as possible, $\mathcal{R}'$ as large as possible, $\mathcal{D}$ as small as possible, and $\mathcal{D}'$ as large as possible, satisfying the conditions below:

(1) $\mathcal{R}' \subseteq \mathcal{R}^*$;

(2) $\mathcal{R}' \subseteq \mathcal{R}$;

(3) $\mathcal{D}' \subseteq \mathcal{D}^*$;

(4) $\mathcal{D}' \subseteq \mathcal{D}$;

(5) $\mathcal{D} \cup \{inhibited(cdk2), inhibited(cdk7), inhibited(cdk9)\}$, $\mathcal{R} \not\vdash apoptosis$; and

(6) $\mathcal{D} \cup \{inhibited(cdk2), inhibited(cdk7), inhibited(cdk9)\}$, $\mathcal{R} \not\vdash E \; conflicts \; F$, for all E and F.

The logic for Conditions (1)–(4), and (6) is as before. Condition (5) says that even after CYC202 has inhibited all its known target, apoptosis is not observed. Thus the breakages and rewirings in Conditions (1)–(4) allow CNE1 to escape the effects of CYC202.

The two hypothetical theoretical scenarios above suggest we know how to search for which parts of which pathways to break, which parts of which pathways to rewire, and which CYC202 target to check. In practice, the search must be guided. This is where the integrated system presented earlier in Subsection 4.1 comes in. Specifically, the last component of the integrated system is designed to identify candidate regulatory relationships of normal pathways that are observed to be consistent or changed in the gene expression data.

5. CONCLUDING REMARKS

Personalized drugs are widely promised to revolutionize the face of medicine. Yet it is important to note that during its infancy now, it begins with simple stratification policies in drug prescription, albeit at a highly sophisticated and precise level. This approach of prescribing personalized drugs for patients is known collectively by the term pharmacogenetics. As defined by [27], we refer to the term pharmacogenetics as different individuals having differences in drug responses as a consequence of having different genetic makeup, hence the term "genetics". The term pharmacogenomics however, refers to the study of the effects of drug responses on the entire genome. Hence the term "genomics".

$$\frac{E \in \mathcal{D}}{\mathcal{D}, \mathcal{R} \vdash E} \qquad \frac{E \ causes \ F \ \in \mathcal{D}}{\mathcal{D}, \mathcal{R} \vdash E \ causes \ F} \qquad \frac{E \ causes \ F \ provided \ G \ \in \mathcal{D}}{\mathcal{D}, \mathcal{R} \vdash E \ causes \ F \ provided \ G}$$

$$\frac{\mathcal{D}, \mathcal{R} \vdash E \quad \mathcal{D}, \mathcal{R} \vdash E \ causes \ F}{\mathcal{D}, \mathcal{R} \vdash F} \qquad \frac{\mathcal{D}, \mathcal{R} \vdash G \quad \mathcal{D}, \mathcal{R} \vdash E \ causes \ F \ provided \ G}{\mathcal{D}, \mathcal{R} \vdash E \ causes \ F}$$

$$\frac{\mathcal{D}, \mathcal{R} \vdash E\theta \quad E \ causes \ F \in \mathcal{R}}{\mathcal{D}, \mathcal{R} \vdash F\theta} \qquad \frac{\mathcal{D}, \mathcal{R} \vdash G\theta \quad E \ causes \ F \ provided \ G \in \mathcal{R}}{\mathcal{D}, \mathcal{R} \vdash E\theta \ causes \ F\theta}$$

$$\frac{\mathcal{D}, \mathcal{R} \vdash E\theta \quad \mathcal{D}, \mathcal{R} \vdash F\theta \quad E \ conflicts \ F \ \in \mathcal{R}}{\mathcal{D}, \mathcal{R} \vdash E\theta \ conflicts \ F\theta}$$

Figure 6: A classical proof system at the core of the reasoning component of the logical framework. Here θ stands for some grounding substitution as usual.

The primary distinction is their intended use. One serves to study the differences of a single compound with a number of individuals, while the other serves to study the differences of a number of drugs with a single individual.

Gene expression analysis is a key tool for pharmacogenomics and pharmacogenetics [20]. The most common example cited in this category is that of the breast cancer treatment drug trastuzamab [4]. This is a specific antibody against breast cancer and happens to be ineffective against two thirds of the patients who do not over-express the drug's target, whereas it significantly improves the survival in the remaining one third where the target is over-expressed.

In this paper, we have provided a succinct but in-depth survey of progress in the analysis of gene expression data for the purposes of (i) disease subtype diagnosis, (ii) new subtype discovery, and (iii) understanding of disease subtypes and treatment responses. We have discussed the issues where existing works still fall short on. We have further envisioned, and are developing, an integrated system—comprising (i) automated analysis and extraction of information from biomedical texts, (ii) targeted construction of known pathways, and (iii) direct hypothesis generation based on logical reasoning on and tests for consistencies and inconsistencies of observed data with known pathways—to address the issues. Thus the system can provide a researcher possible biologically inspired interpretations and solutions to his questions, enabling him to better decipher what are the reasons that cause a drug to be effective or ineffective.

Acknowledgements

6. REFERENCES

[1] E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In *Proc. 5th ACM International Conference on Digital Libraries*, pages 85–94, 2000.

[2] C. Ahlers, M. Fiszman, D. Demner-Fushman, et. al. Extracting semantic predications from medline citations for pharmacogenomics. In *Proc. Pacific Symposium on Biocomputing*, pages 209–220, 2007.

[3] S. Ananiadou and J. McNaught, editors. *Text Mining for Biology and Biomedicine*. Artech House, Norwood, MA, 2006.

[4] J. Baselga, D. Tripathy, J. Mendelsohn, et. al. Phase II study of weekly intravenous recombinant humnaized anti-p185 (HER2) monoclonal antibody in patients with HIER2/neu-overexpressing metastatic breast cancer. *J Clin Oncol*, 14(3):737–744, 1996.

[5] R. Breitling and P. Herzyk. Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *Journal of Bioinformatics and Computational Biology*, 3(5):1171–1190, October 2005.

[6] G. C. Cawley and N. L. C. Talbot. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics*, 22(19):2348–2355, 2006.

[7] Y. Cheng and G. M. Church. Biclustering of expression data. In *Proc. 8th International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, 2000.

[8] S. W. Doniger, N. Salomonis, K. D. Dahlquist, et. al. MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biology*, 4(1):R7, 2003.

[9] C. Friedman, P. Kra, H. Yu, et. al. GENIES: A natural language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(Suppl. 1):S74–S82, 2001.

[10] M. E. Futschik and B. Carlisle. Noise-robust soft clustering of gene expression time-course data. *Journal of Bioinformatics and Computational Biology*, 3(4):965–988, 2005.

[11] L. Goh and N. Kasabov. An integrated feature selection and classification method to select minimum number of variables on the case study of gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(5):1107–1136, 2005.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, et. al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(15):531–537, 1999.

[13] K. C. Gunsalus, H. Ge, A. J. Schetter, et. al. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature*, 436(7052):861–865, 2005.

[14] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(Suppl. 1):S145–S154, 2002.

[15] C. Henegar, R. Cancello, S. Rome, et. al. Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. *Journal of Bioinformatics and Computational Biology*, 4(4):833–852, 2006.

[16] W. J. Heuett and H. Qian. Combining flux and energy balance analysis to model large-scale biochemical networks. *Journal of Bioinformatics and Computational Biology*, 4(6):1227–1244, 2006.

[17] L. Hirschman, J. C. Park, J. Tsujii, et. al. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18:1553–1561, 2002.

[18] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(Suppl. 1):S233–S240, 2002.

[19] J. L. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: From information retrieval to biological discovery. *Nature Reviews Genetics*, 7:119–129, 2006.

[20] W. Kalow. Pharmacogenomics: Historical perspective and current status. *Methods in Molecular Biology*, 311:3–15, 2005.

[21] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. The KEGG database at GenomeNet. *Nucleic Acids Research*, 30(1):42–46, 2002.

[22] S. Y. Kim and D. J. Volsky. PAGE: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 8(6):144, 2005.

[23] D. Klein and C. D. Manning. Fast exact inference with a factored model for natural language parsing. *Proc. 15th Neural Information Processing Systems Conference*, pages 3-10, 2002.

[24] D. Klein and C. D. Manning. Accurate unlexicalized parsing. *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

[25] C. Li, S. Suzuki, Q.-W. Ge, et. al. Structural modeling and analysis of signaling pathways based on petri nets. *Journal of Bioinformatics and Computational Biology*, 4(5):1119–1140, October 2006.

[26] H. Li, X. Chen, K. Zhang, and T. Jiang. A general framework for biclustering gene expression data. *Journal of Bioinformatics and Computational Biology*, 4(4):911–933, 2006.

[27] K. LindPaintner. Pharmacogenetics and the future of medical practice. *J Mol Med*, 81:141–153, 2003.

[28] H. Liu, J. Li, and L. Wong. Selection of patient samples and genes for outcome prediction. In *Proc. 3rd International Computational Systems Bioinformatics Conference*, pages 382–392, 2004.

[29] Z. Liu, D. Chen, H. Bensmail, and Y. Xu. Clustering gene expression data with kernel principal components. *Journal of Bioinformatics and Computational Biology*, 3(2):303–316, 2005.

[30] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:24–45, 2004.

[31] J. A. Mitchell, A. R. Aaronson, J. G. Mork, et. al. Gene indexing: Characterization and analysis of NLM's GeneRIFs. In *Proc. 27th AMIA Annual Symposium*, pages 460–464, Washington, D. C., 2003.

[32] V. Olman, C. Hicks, P. Wang, and Y. Xu. Gene expression data analysis in subtypes of ovarian cancer using covariance analysis. *Journal of Bioinformatics and Computational Biology*, 4(5):999–1014, 2006.

[33] C. H. Pui and W. E. Evans. Acute lymphoblastic leukemia. *New England Journal of Medicine*, 339:605–615, 1998.

[34] X. Qiu and A. Yakovlev. Some comments on instability of false discovery rate estimation. *Journal of Bioinformatics and Computational Biology*, 4(5):1057–1068, 2006.

[35] D. Rajagopalan and P. Agarwal. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, 21(6):788–793, 2005.

[36] A. Rzhetsky, I. Iossifov, T. Koike, et. al. GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1):43–53, 2004.

[37] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(Suppl. 1):i264–i271, 2003.

[38] A. Y. Sivachenko, A. Yuryev, N. Daraselia, and I. Mazo. Molecular networks in microarray analysis. *Journal of Bioinformatics and Computational Biology*, 5(2):???–???, 2007. In press.

[39] F. Sohler, D. Hanisch, and R. Zimmer. New methods for joint analysis of biological networks and expression data. *Bioinformatics*, 20(10):1517–1521, 2004.

[40] A. Subramanian, P. Tamayo, V. K. Mootha, et. al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci. USA*, 102(43):15545–15550, 2005.

[41] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, et. al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.

[42] B. R. Zeeberg, W. Feng, G. Wang, et. al. GoMiner: A resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4):R28, 2003.