Meta-analysis and Combining Information in Genetics

Rudy Guerra
&
David Allison

Contents

CHAPTER 1

Combining genomic data in human studies

Debashis Ghosh, Dan Rhodes and Arul Chinnaiyan
University of Michigan

1.1 Introduction

With the development of technology that has allowed for the high-throughput miniaturization of standard biochemical assays, it has become possible to globally monitor the biochemical activity of populations of cells. This has led to the emergence of cDNA microarrays in medical and scientific research and has allowed for large-scale trasncriptional characterization. It should also be noted that the microarray technology would have limited ability without the existence of large-scale genome sequencing projects, such as the Human Genome Project (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). Having such sequence data available allows for the characterization of the probes on the microarray. In this chapter, we will be using the term ”genomic data” to generically refer to any genetic data that is generated using large scale technologies.

While transcript mRNA microarrays have received much attention in the literature, there has been work on other types of microarrays. Examples include chromatin-immunoprecipitation (ChIP) microarrays, which measure transcription factor-DNA binding expression (Lee et al., 2002) and methylation microarrays (Yan et al., 2001), which assess DNA methylation on a global scale. In addition, there has also been much attention on high-throughput assays that measure protein-protein interactions, such as yeast two-hybrid systems (Uetz et al., 2001). Because of all the large-scale data that is being generated, there is much interest in attempting to integrate the data to provide a more complete understanding of the biological mechanisms that are at play. This type of analysis has been given the name ”systems biology” in the bioinformatics literature (Ideker et al., 2001).

For the statistician, this area brings many interesting and challenging problems. While the term ”meta-analysis” is familiar among most statisticians (Normand, 1999), the term here takes a very different meaning. The situation statisticians are familiar with

involves attempting to combine information from relatively homogeneous data structures from multiple similar experiments. However, in much of the genomic area, the issue is one of trying to combine relatively inhomogeneous data structures from multiple experiments that may or may not be similar.

Another complication is that data availability depends on the type of organism studied. In this chapter, we focus on data from human studies. Thus, protein-protein interaction data from two-hybrid experiments are not currently available for humans. We will talk about approaches for combining genomic data in human studies, primarily focusing on methods developed in the cancer setting. Some familiarity with microarray technologies is assumed; the reader is referred to the first and second volumes of *The Chipping Forecast*, a supplement to the journal *Nature Genetics* that has been made publicly available online (Chipping Forecast, 1999, 2002). Our goal here is to seek to outline the major issues involved in such analyses and describe some solutions that have been proposed. It is not our intent to provide an up-to-the date listing of all methodologies that have been used, as the literature is constantly changing. Given the dynamic nature of the field, an important component will be benchmarking of methods to see which should be used in practice.

1.2 Genomic data integration in cancer

1.2.1 Goals

Our group has focused primarily on the analysis of genomic data in cancer studies. There are two broad goals of this research. One is the discovery of new biomarkers that might be used potentially as screening tests or to better predict patient prognosis. Examples of potential promising biomarkers found using gene expression technology include enhancer of zeste homolog 2 (EZH2) in prostate cancer (Varambally et al., 2002). In this study, the transcript mRNA expression EZH2 gene transcript was found to be highly expressed in metastatic prostate cancer. A key point to make at this stage, which we will address later, is that mRNA expression does not necessarily perfectly correlate with protein expression. In terms of diseases, the action is happening at the protein level. In protein validation studies done by Varambally et al. (2002), the EZH2 protein was also found to be highly expressed in metastatic prostate cancer. Another example of a potential biomarker found using genomic data technologies is prostasin in ovarian cancer (Mok et al., 2001). In that study, the authors reported a sensitivity of 92% and a specificity of 94% for discriminating ovarian cancer cases from controls using validation by ELISA of serum. Thus, prostasin might serve as a potential biomarker for early detection of ovarian cancer.

The second is to better understand the biology of the disease. In the past, cancer was thought of as a heterogeneous collection of diseases. However, a more integrative view of the disease is currently being put forward by many researchers; this view was summarized eloquently in a review article by Hanahan and Weinberg (2000). According to their paradigm, there are six principles that underlie tumorigenesis (the

initiation and development of a tumor); equivalently, for a cancer to develop, it must acquire six ”hallmark capabilities”:

- Self-sufficiency in growth signals;
- Insensitivity to anti-growth signals;
- Evading apoptosis (cell death);
- Limitless replicative potential;
- Sustained angiogenesis;
- Tissue invasion and metastasis.

With the current availability of large-scale genomic data, we can address the Hanahan and Weinberg model in two ways. First, we can analyze the data to see the relative contributions of the six ”hallmark capabilities.” Second, we can use genomic data to further refine and identify the pathways that comprise each of the individual hallmark capabilities described above.

1.3 Combining data from related technologies: cDNA microarrays

The statistical problem closest in spirit to classical meta-analysis involves trying to combine multiple datasets in which the same type of cellular activity was assessed. As an example here, we consider multiple microarray studies in which the same comparison was considered, namely cancer versus normal.

There are several issues that must be considered when attempting such an analysis. First, one must consider the problem of study-specific artifacts, such as sampling bias, variations in experimental protocols and differences in laser scanners. However, there are two bigger issues in the analysis of such data. The first is that of matching genes from two studies. This is where the availability of large-scale genomic data figures in hugely. Each spot on a microarray corresponds to a DNA sequence. What one can do is to match up each spot to a putative gene in the Unigene Database, which is a collection of clusters of orthologous genes. The Unigene link can then be used to identify common genes across multiple datasets. Such a task can be done for Affymetrix chips from their website (http://www.netaffx.com/) or for two-color cDNA microarrays using the SOURCE tool at Stanford (Diehn et al., 2003).

While such a mapping is useful, there still might be errors that remain. A more challenging issue involves the fact that the numbers from different microarray platforms represent different things. That is, an expression value of 20 from a cDNA two-color microarray is much different from an expression value of 20 measured on an Affymetrix array. Another technique that has proven to be useful as a filtering device to enhance comparability across arrays of different platforms is known as the integrative correlation coefficient or correlation of correlation coefficients (Lee et al., 2002; Parmigiani et al., 2004). The idea underlying this method is that while raw expression values vary from study to study, the intergene correlations do not vary as much.

Thus, one would consider combining genes that have similar intergene correlations across the studies.

In terms of meta-analysis methods put forward, many have been based on the fact that the standardized effect size is combinable across studies. This is the approach advocated by Parmigiani et al. (2004) after filtering based on the integrative correlation coefficient. In Rhodes et al. (2002), the t-statistic was transformed into a p-value, a transformation of which was combined across multiple studies. By contrast, in Ghosh et al. (2003), the t-statistic was combined directly. An approach that was more Bayesian in nature was taken by Wang et al. (2004), in which expression values from one study were used to develop a prior distribution for the standardized effect size; data from the remaining studies were used to generate posterior distributions. A fully hierachical approach was taken by Choi et al. (2003), who then used Markov Chain Monte Carlo methods to sample from the posterior distributions. It should be noted that all of these methods make the assumption that a standardized effect size can be estimated directly for each individual study.

Another approach more in line with classification or supervised learning analyses is to build a classifier or find a gene expression signature on one dataset and to see how well it predicts in an independent microarray dataset. Such approaches were taken by Beer et al. (2002), Wright et al. (2004) and Jiang et al. (2004). An alternative method using hierarchical clustering, which is an unsupervised learning procedure, was taken by Sorlie et al. (2003). They found a gene expression signature that defined molecular subtypes in breast cancer; they found through interrogation of other datasets that the subtypes were present there as well. Given the increasing availability of publicly available large-scale gene expression datasets, it is increasingly important that results found by one investigator on a particular dataset be validated using other datasets as well.

A large-scale comprehensive meta-analysis was performed by Rhodes et al. (2004). They performed a meta-analysis of 40 independent datasets ($>$3,700 array experiments) across ??? tissue sites. They found a universal profile of 67 genes that could differentiate cancer versus noncancer tissue for a variety of cancers. In addition, they determined 36 cancer-specific signatures for determining a tissue-specific cancer. The signatures also demonstrated good discrimination performance on three independent datasets.

A more sophisticated method for meta-analysis was put forward by Shen et al. (2004), based on an idea of Parmigiani et al. (2002). Namely, the idea is that for a given gene from a given sample in a given study, it is either over-,under- or non-differentially expressed with respect to a baseline cohort of genes. Each of the three states defines a latent category, which induces a mixture model for gene expression values. The latent states of over-, under- or non-differentially expressed are inferred using a Markov Chain Monte Carlo sampling algorithm. The estimated probabilities of the latent states are then transformed to define a “probability of expression,” which is then used as input for a meta-analysis.

Much of the meta-analysis methods have studied differential expression across mul-

tiple studies. A notable exception is the study by Lee et al. (2004), in which inter-gene correlations across multiple studies was considered. The authors sought pairs of genes that were consistently coexpressed across several datasets. As will be described in the next section, such coexpression is the first step needed in building gene regulatory networks.

Because of the fact that information on thousands of genes are typically considered, there is an inherent multiple testing problem. A popular method for calibrating results in this setting has been through use of the false discovery rate (Benjamini and Hochberg, 1995). The false discovery rate, or FDR, is roughly defined as the expected proportion of falsely rejected null hypotheses among the set of rejected null hypotheses. A smaller FDR indicates that there are more "real" discoveries found by the investigator. This can be visualized by considering the cross-classification of n single-gene hypotheses by whether they are rejected based on the data and their true status (i.e. null hypothesis is true or alternative hypothesis is true). Such a table is given here:

Table 1.1 *Outcomes of n tests of hypotheses*

	Accept	Reject	Total
True Null	U	V	n_0
True Alternative	T	S	n_1
	W	Q	n

The definition of false discovery rate (FDR) as put forward by Benjamini and Hochberg (1995) is

$$FDR \equiv E\left[\frac{V}{Q} \mid Q > 0\right] P(Q > 0).$$

The conditioning on the event $[Q > 0]$ is needed because the fraction V/Q is not well-defined when $Q = 0$. Methods for controlling the false discovery rate have been proposed by several authors (Benjamini and Hochberg, 1995; Benjamini and Liu, 1999; Benjamini and Yekutieli, 2001, Sarkar, 2002). In addition, methods for directly estimating the false discovery rate (Storey, 2002) are also available.

A more recent innovation put forward by Storey and Tibshirani (2003) has been estimation of a quantity known as the q-value, which represents the minimum positive FDR rate at which significance is attained. It represents an analog of the p-value that takes multiple testing into account. It is quite commonplace for investigators to rank genes based on a q-value threshold.

Another technique that is done is to adjust p-values for multiple testing; a variety of methods for doing so is found in Westfall and Young (1993). The p-value corresponds to the minimum significance level at which significance is attained. For multiple testing as described in Table **??**, the an analog of the significance level is the familywide error rate (FWER), defined as $P(V \geq 1)$. Further dicussion for FWER-controlling procedures can be found in Ge et al. (2002) and in a collection of papers by Van der

Laan and colleagues (Dudoit et al., 2004; van der Laan et al., 2004a,b). One unintended result of the development of high-throughput genomic data technologies has been the development of new statistical methodologies for addressing the multiple testing problem.

1.3.1 Functional and Pathway Analyses

Once these meta-analyses are performed and a calibrated list of genes are generated, the gene lists can be entered into databases representing functional processes. A simple visualization exercise, done in Rhodes et al. (2002), is to find metabolic pathways in which multiple genes exist. One example of such a database is the Kyoto Encyclopedia of Genes and Genomes (KEGG). Based on a list of genes that were consistently dysregulated across multiple studies comparing prostate cancer to non-prostate cancer, pathways such as the purine biosynthesis were found to have multiple genes. This leads to the hypothesis that the purine biosynthesis pathway is dysregulated in prostate cancer. While the study is only generating a hypothesis and not confirming it, such a computational prediction can help to inform investigators as to the next series of experiments to perform. Also, a visual display such as that given by KEGG does not allow for any formal statistical assessment of significance.

More formal statistical analyses for enrichment of functional terms can be done using the hypergeometric distribution. This requires a database of functional annotation terms such as Gene Ontology (GO) (Ashburner et al., 2000). The idea behind this procedure is to see if the frequency of certain Gene Ontology terms in a list of genes is similar to or significantly larger than that in an external database. If it is determined that there is statistically significant enrichment of functional annotation terms in a list, then again this generates the hypotheses that certain pathways are dysregulated in the disease process. This can be easily seen with the following 2×2 table: There

Table 1.2 *Fisher's test example*

	Gene List	Non-Gene List	Total
GO term X	a	b	G
Non GO term X	c	d	N-G
	l	N-l	N

are l genes in the list and N genes total, i.e. on the chip. The null hypothesis is that there is no association between the rows and columns of the table; no association means that there is no functional enrichment of GO term X in the list of genes. This is tested for by calculating a p-value based on the hypergeometric distribution, which conditions on the row and column totals. An exact test is known as Fisher's exact test.

There are now many publicly available tools for performing such a test (Draghici et

al., 2003; Al-Sharour et al., 2004; Beissbarth and Speed, 2004). Note that the methods discussed in the last two paragraphs are post-hoc types of procedures in that the pathway analysis is done conditional on selecting a list of genes. An alternative is to directly model the information contained in the Gene Ontology databases with gene expression data. However, this raises the problem of what constitutes a proper metric by which the heterogeneous information from the two diverse databases can be related; this currently remains an open question. We later discuss the use of graphical models later in this chapter as well.

A resource initiated by our group is a database known as ONCOMINE (Rhodes et al., 2004), located at the URL http://www.oncomine.org/. The database represents an effort to systematically curate, analyze and make available all public cancer microarray data via a web-based database and data-mining platform. Within the database, one can perform over 100 types of differential expression analyses based on disease/non-diseased, stage of disease, subtype, etc., reported with study-specific q-values. These analyses are based on standard differential expression analysis with correction for multiple testing using the q-value. In addition, one can query individual genes for known available genetic and proteomic information that is stored at other databases (e.g., GenBank, Swiss-Prot, etc.). There are links with pathway databases for visualization and assessing functional enrichment of the gene lists that are found. One can also search for individual genes of interest to see their expression patterns across multiple cancer studies.

1.4 Combining Data from Different Technologies

In the traditional statistical view of meta-analysis, one thinks of attempting to combine information from multiple similar experiments. However, the challenge of bioinformatics is that high-throughput functional genomics data are being generated on a variety of platforms and stored in different databases. The challenge then becomes how to integrate diverse data. This leads to a new definition of "meta-analysis."

1.4.1 Bayesian networks

One tool that has been utilized quite heavily for this type of problem has been graphical models (Lauritzen, 1996; Jensen, 2001). These are also referred to as Bayesian networks and belief networks as well. The idea of graphical models is to estimate dependencies between random variables through calculation of measures of covariation between them. As a simple example, let us consider three random variables, A, B and C. If we assume that the joint distribution of (A, B, C) is multivariate normal, then assuming the random variables have mean zero, the distribution is summarized by the pairwise correlation coefficients between them. Thus, if we can estimate the correlations, then we have "learnt" about the system characterized by A, B and C. There was a lot of interest in attempting to construct regulatory networks by fitting

graphical models to gene expression data only. However, given the amount of experimental variability in such data, this turned out not to be a major direction, so the focus has been on building networks with multiple sources of data.

One major goal of Bayesian networks has been to predict protein-protein interactions. While much of the genomic data is measured at non-protein levels, actual cellular activity and disease occurs at a protein level. Thus, it is of interest to figure out how well functional genomic correlations predict protein-protein interactions. This was first studied in yeast by Jansen et al. (2003). However, they had the advantage of having high-throughput protein-protein interaction data available from yeast two-hybrid experiments. Such experiments currently do not exist for humans.

In a recent application (Rhodes et al., 2005), we used Bayesian networks to predict protein-protein human interactions using functional genomic data. We used several different types of information in order to develop the graphical model:

1. interactions between orthologs of human proteins;
2. intergene correlations from gene expression profiles;
3. shared functional annotations from Gene Ontology;
4. shared enrichment domains.

The idea was to develop a graphical model using known positive and negative protein-protein interactions in order to develop a scale of evidence for predicting a protein-protein interaction. To define the positives, we used the Human Protein Reference Database (HPRD) (Peri et al., 2003) , a bioinformatics resource that contains known protein-protein interactions manually curated from the literature by expert biologists. We queried 11,678 distinct literature-referenced protein-protein interactions among 5,505 proteins. For the negatives, we identified all protein pairs in which one protein was assigned to the plasma membrane cellular component and the other to the nuclear cellular component based on Gene Ontology. Based on fitting model, we predicted approximately 10,000 interactions with a false positive rate of 20% and about 40,000 interactions with a false positive rate of 50%. Several of the predicted protein-protein interactions were verified by subsequent experimentation, while other predictions mimicked what was found in the reported experimental literature. This model has been integrated into ONCOMINE and is available at the URL http://www.himapp.org/.

While there have been some successes with the graphical models approach, this area definitely remains in its infancy. One limitation of the graphical model is that it only uses pairwise covariation information. Furthermore, the graphical models used by Jansen et al. (2003) and Rhodes et al. (2005) involve a binning procedure that seems somewhat *ad hoc*. One interesting alternative to has been proposed by Balasubramaniam et al. (2004), who propose using a graph-theoretic approach to combining functional genomics data from diverse platforms and test for significance of the nodal connections using permutation testing. Interestingly, the appear to be similarities with the use of graph-theoretic ideas in this area with those in social network literature (Wasserman and Faust, 1994). This suggests that there may exist techniques from that field that may be of use here.

Another point of the Bayesian networks is that they are bidirectional and do not attempt to impose any directionality. However, we know that activity in biological systems consists of a series of ordered steps. Thus, there might be some advantage to incorporating directionality into the system. Let us take the transcription process as an example. First, the must be binding of DNA to the upstream promoter regions in the genome so that transcription is "turned on." Thus, one could imagine a model for expression as a function of upstream promoter sequence for this scenario. Models like this have been proposed for lower-level eukaryotes (Bussemaker et al., 2001; Conlon et al., 2003) and are referred to as "dictionary models." They take a view that the expression value is a function of a score computed using the sequence data, which is a conditional model. It remains to be seen whether such models could work for human genomic data.

1.4.2 Towards an understanding of regulatory mechanisms

In the previous sections, we have described methods for combining information in order to derive improved gene signatures and to make protein-protein interactions. Another goal of interest is to derive "regulatory" modules. It is likely that some gene expression patterns observed from microarray data represent a downstream readout of a small number of genetic aberrations (e.g., mutations, amplifications, deletions, translocations) that led to the activation or inactivation of a small number of transcription factors. In some cases, cancer-causing genetic aberrations may not be directly apparent from these downstream gene expression readouts. Recently, approaches to developing gene expression regulatory modules in human studies have been taken by Elkon et al. (2003), Segal et al. (2004) and by Rhodes et al. (2005).

The general approach requires a predefined list of genes. The list of genes can come from an external database, such as Gene Ontology (e.g. set of genes involved in a known process), or it may come from a differential expression analysis. Based on the gene list, the Segal et al. (2004) approach is to determine which arrays are commonly induced by multiple gene lists; the gene lists are then combined to form a "core" gene cluster. One then determines which arrays show significant differential expression based on the core gene cluster. One then determines if there is enrichment of clinical annotation in the set of arrays found at the previous step. Through this procedure, Segal et al. (2004) are able to find 456 regulatory modules from gene expression data consisting of measurements of 14,145 genes in 1917 samples across 22 tissue sites.

The approach taken by Elkon et al. (2003), while similar in spirit, involves a major difference. The difference is that sequence data are integrated with the gene expression profiling data. For the study by Elkon et al. (2003), approximately 13,000 putative promoter start sites were identified based on the NCBI Reference Sequence Database (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens). Next, a set of genes that were determined to be cell-cycle regulated from a human cell cycle gene expression profiling study () were used; of the 874 putative cell-cycle genes in that paper, promoter start sites were available for 568 of them. The authors searched for significantly

enriched position weight matrices in the entire set of the 568 cell cycle-regulated promoters using the original 13K set as the background set and found enrichment of six binding sets. Thus, this provides a set of candidate transcription factors which may play a role in cell-cycle progression.

The study of Rhodes et al. (2005) is similar to that of Elkon et al. (2003). They derive 265 gene lists from various differential expression analyses using a q-value cutoff of 0.10. Next, they identify putative transcription factor binding sites in the promoter sequences of human genes and come up with a database of 361 transcription factors. Next, enrichment of each transcription factor in each of the gene lists is done; again an adjustment for multiple testing based on false discovery rate calibration is performed. From this analysis, they defined 311 regulatory programs that displayed highly significant overlap ($P < 0.00033$) between a gene expression signature and a regulatory signature; these will serve as candidate regulatory modules that can be tested experimentally.

The crux of the analyses described in this section is that based on defined lists of genes, one calculates overlap measures of enrichment of a certain biological property (here binding sites) with the lists. It is fairly easy to see how other types of biological sequence information (e.g., protein structure information, etc.) might be used here as well. In addition, there are many ways of defining "interesting." It could be differential expression from a two-group comparison, or cell-cycle regulated (i.e., periodic expression) in a microarray time-course study. The overlap statistic is a very simple, and again, many other approaches are possible. This area will be a popular one for further study.

1.5 In vivo/in vitro genomic data integration

An area that is beginning to be considered more frequently in functional genomic studies in cancer is the integration of *in vitro*, i.e. experimental studies, with human gene expression studies, termed *in vivo* data. Integrating results from such experiments with in vivo cancer signatures holds the potential both to infer activity of specific oncogenic pathways in vivo and to identify relevant effectors of oncogenic pathways. For example, Huang et al. (2003) developed distinct in vitro oncogenic signatures for three transcription factors, Myc, Ras and E2F1-3. These signatures were able to predict Myc and Ras state in mammary tumors that developed in transgenic mice expressing either Myc or Ras, suggesting that specific oncogenic events are encoded in global gene-expression profiles.

To begin to understand the mechanisms by which oncogenes cause cancer, studies have used gene-expression profiling to identify downstream targets of oncogenic pathways in cell-culture systems. Conceptually, this involves manipulating a gene in an in vitro system and measuring a global profile using gene expression technology and then trying to relate the in vitro gene expression profile to an in vivo gene expression profile. Such an approach was taken by Lamb et al. (2004) to determine the

direct transcriptional effects of oncogene Cyclin D1. In vitro experiments were performed in which the Cyclin D1 was both over and underexpressed, and global gene expression profiles were determined. Lists of differentially expressed genes were then generated. To correlate the lists with in vivo gene expression data, a two-step process was utilized in which genes were first ordered based on correlation with Cyclin D1. Then, a Kolmogorov-Smirnov statistic was used to determine if the lists clustered within the ordered list based on correlation. Since there was significant evidence of clustering, Lamb et al. (2004) found that the in vitro-defined targets of Cyclin D1 were correlated with Cyclin D1 levels in vivo. This suggests that the direct regulatory effects of Cyclin D1 may play an important role in tumorigenesis. The statistical problem brought up this type of analysis is determining clustering of a list of genes within an ordered list of genes. While a Kolmogorov-Smirnov statistic has the advantage of being a nonparametric statistic, the potential disadvantage to the use of such a method will be a loss of efficiency. Determining alternative methodologies for this type of problem will be important.

Another setting that leads to consideration of in vitro and in vivo genomic data is when the in vitro experiment is performed in a model organism system. For example, Sweet-Cordero et al. (2005) defined a signature by comparing lung tumors generated from a spontaneous KRAS mutation mouse model to normal mouse lung and correlating it with gene expression profiles in human lung cancer studies. The major issue in such an analysis is mapping mouse genes to orthologous human genes. Sweet-Cordero et al. (2005) found that the mouse signature shared significant similarity with human lung adenocarcinoma but not with other lung cancer types. Next, they looked for evidence of the KRAS signature in human tumors carrying activating KRAS mutations relative to wild-type tumors. Although no individual genes were significantly associated with KRAS mutation status in human tumors, the mouse KRAS signature was significantly enriched among genes rank-ordered by differential expression in human tumors with a KRAS mutation.

It is expected that experiments such as those described in the previous two paragraphs will become much more commonplace in the future. Thus, it will be critical to address issues and to develop methods for integrating in vivo and in vitro genomic data so that inferences regarding transcriptional regulatory pathways in cancer can be generated.

1.6 Software availability

Due to the recent innovations previously described, public use software for implementing these methods is still in their infancy. As mentioned earlier, our group has developed a database, ONCOMINE, located at the following URL:

http://www.oncomine.org/.

The database is geared towards biologists and does automated data analyses. Examples include differential expression analyses, analyses for functional enrichment of

GO terms and Kolmogorov-Smirnov analyses in the spirit of Lamb et al. (2003). In addition, links to the protein-protein prediction project of Rhodes et al. (2005), are available. The website for this is located at the following URL:

http://www.himapp.org/.

Many genomic data analysts primarily use software languages such as MATLAB and R (R Development Core Team, 2005) for the analysis of genomic data. In particular, there has been a project towards the development of bioinformatics software packages in R, known as Bioconductor (Gentleman et al., 2004). The goals of the Bioconductor project are threefold: goals of the project include:

1. foster collaborative development and widespread use of innovative software;
2. reduce barriers to entry into interdisciplinary scientific research,
3. promote the achievement of remote reproducibility of research results.

One benefit of R is that it is a high-level interpretable language that allows for relatively fast development of methods. In addition, it has a nice ability for packaging related components.

Another language that is of great use in this type of bioinformatics research is Perl. Given that many of the databases are text databases, it is very important to be able to manipulate such databases relatively easily. Perl is a very useful language for such text manipulations.

1.7 Discussion

In this chapter, we have attempted to describe the current state of knowledge in the area of functional genomic analyses. Because of the different types of functional genomic datasets that are being generated, this has led to an extension of the statistical concept of meta-analysis. Now, analysts are faced with the prospect of combining different sources of information from different types of platforms.

One of the techniques described earlier, graphical models, is a tool from the area of machine learning. Machine learning algorithms tend to be black-box algorithms that are useful for predictive inference. While the application of machine learning algorithms to high-dimensional genomic datasets will lead to some predictions that will be bourne out, it is also important to attempt to build in biological information as much as possible into the analyses. As an example, a central tenet of biology is that binding of DNA to the binding sites transcription factors leads to activation of gene expression. It would seem sensible that a model in which transcription factor information is the independent factor and gene expression is the dependent variable should be a better model for the system than a graphical model that assumes no directionality.

Finally, an important non-statistical issue that needs to be addressed is how to store

information from these types of analyses such that they themselves can be combined. One can imagine that lists of genes from different analyses can be used to make inferences about various biological aspects in cancer studies. It then may be of interest to compare the lists themselves in another type of meta-analysis so that higher-order inferences about the biological network can be made. However, to do this will require work to develop database requirements and standardization, much as was done in the case of microarrays (Brazma et al., 2001).

1.8 Acknowledgments

The first author would like to acknowledge the support of grant GM72007 from the Joint NSF/NIGMS Biological Mathematics Program.

1.9 References

Al-Sharour, F., Diaz-Uriarte, R. and Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics, 20: 578 – 580.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 25: 25 – 29.

Balasubramanian, R., LaFramboise, T., Scholtens, D. and Gentleman, R. (2004). A graph-theoretic approach to testing associations between disparate sources of functional genomics data. Bioinformatics, 20: 3353-62.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B. and Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nature Medicine, 8: 816-24.

Beissbarth, T. and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics, 20: 1464-5.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B, 57: 289–300.

Benjamini, Y. and Liu, W. (1999). A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence. Journal of Statistical Planning and Inference, 82: 163 - 170.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of Statistics, 29: 1165–1188.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature Genetics, 29: 365-71.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2001). Regulatory element detection using correlation with expression. Nature Genetics, 27: 167 – 171.

Chipping Forecast (1999). Special Supplement, Nature Genetics. http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v21/n1s/index.html

Chipping Forecast (2002). Special Supplement, Nature Genetics. http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v32/n4s/index.html

Choi, J. K., Yu, U., Kim, S. and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. Bioinformatics 19: 84 - 90.

Conlon, E. M., Liu, X. S., Lieb, J. and Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci U S A, 100: 3339 – 3344.

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J. C., Hernandez-Boussard, T., Rees, C. A., Cherry, J. M., Botstein, D., Brown, P. O. and Alizadeh, A. A. (2003). SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Research, 31: 219-23.

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A. and Tainsky, M. A. (2003). Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucleic Acids Research, 31: 3775-81.

Dudoit, S., van der Laan, M. J. and Pollard, K. S. (2004). Multiple testing. Part I. Single-step procedures for control of general Type I error rates. *Statistical Applications in Genetics and Molecular Biology*, **3**, Article 13.

Elkon, R., Linhart, C., Sharan, R., Shamir, R. and Shiloh, Y. (2003). Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Research, 13: 773 – 780.

Ge, Y., Dudoit, S. and Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis (with discussion). *Test* **12**, 1 – 77.

Gentleman, R. C., Carey, V. J., Bates, D. M., 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, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. and Zhang, J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.

Ghosh, D., Barette, T. R., Rhodes, D. and Chinnaiyan, A. M. (2003). Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer. Functional and Integrative Genomics, 3: 180-8.

Hanahan D., and Weinberg, R. A. (2000). The hallmarks of cancer. Cell, 100: 57 – 70.

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., D'Amico, M., Pestell, R. G., West, M. and Nevins, J. R. (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. Nature Genetics, 34: 226 – 230.

Ideker, T., Galitski, T. and Hood, L. (2001). A new approach to decoding life: systems biology. Annual Review of Genomics and Human Genetics, 2: 343-72.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature, 409: 860 - 921.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 302: 449-53.

Jensen, F. V. (2001). Bayesian Networks and Decision Graphs. New York: Springer-Verlag.

Jiang, H., Deng, Y., Chen, H. S., Tao, L., Sha, Q., Chen, J., Tsai, C. J. and Zhang, S. (2004). Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics, 5: 81.

Lamb, J., Ramaswamy, S., Ford, H. L., Contreras, B., Martinez, R. V., Kittrell, F. S., Zahnow,

C. A., Patterson, N., Golub, T. R. and Ewen, M. E. (2003). A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. Cell, 114: 323 – 334.

Lauritzen, S. Graphical Models. Oxford University Press, 1996.

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. and Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. Genome Research, 14: 1085 - 1094.

Lee, J. K., Bussey, K. J., Gwadry, F. G., Reinhold, W., Riddick, G., Pelletier, S. L., Nishizuka, S., Szakacs, G., Annereau, J. P., Shankavaram, U., Lababidi, S., Smith, L. H., Gottesman, M. M. and Weinstein, J. N. (2003). Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. Genome Biology, 4: R82.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K, Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K. and Young RA. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science, 298:799-804.

Mok, S. C., Chao, J., Skates, S., Wong, K., Yiu, G. K., Muto, M. G., Berkowitz, R. S. and Cramer, D. W. (2001). Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology. J Natl Cancer Inst, 93: 1458-64.

Normand, S. L. (1999). Meta-analysis: formulating, evaluating, combining and reporting. Statistics in Medicine 18, 321 – 359.

Parmigiani, G., Garrett, E. S., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for molecular-based classification in cancer. J. Roy. Stat. Soc. Ser. B, 64: 717 – 736.

Parmigiani, G., Garrett-Mayer, E. S., Anbazhagan, R. and Gabrielson, E. (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clinical Cancer Research, 10: 2922-7.

Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., Chandrika, K.N., Padma, N., Harsha, H.C., Yatish, A.J., Kavitha, M.P., Menezes, M., Choudhury, D.R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S.K., Madavan, V., Joseph, A., Wong, G.W., Schiemann, W.P., Constantinescu, S.N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G.C., Dang, C.V., Garcia, J.G., Pevsner, J., Jensen, O.N., Roepstorff, P., Deshpande, K.S., Chinnaiyan, A.M., Hamosh, A., Chakravarti, A. and Pandey, A. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Research, 13: 2363-71.

R Development Core Team, R: A language and environment for statistical computing. (2005). R Foundation for Statistical Computing, Vienna, Austria.

Rhodes, D., Barrette, T. R., Rubin, M. A., Ghosh, D. and Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Research, 62: 4427 – 4433.

Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Barrette, T. R., Ghosh, D. and Chinnaiyan, A. M. (2005). Mining for regulatory programs in the cancer transcriptome. Nature Genetics, Epub 26 May 2005.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette T. R., Kalyana-Sundaram, S., Ghosh, D., Pandey, A. and Chinnaiyan, A. M. (2005). Integrative Prediction of the Human Protein Interaction Network and Application to Cancer Biology. Nature Biotechnology, to appear.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T.,

Pandey, A. and Chinnaiyan, A. M. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia, 6: 1-6.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A. M. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proceedings of the National Academy of Sciences U S A, 101: 9309-14.

Sarkar, S. K. (2002). Some results on false discovery rates in multiple testing procedures. Annals of Statistics, 30: 239 - 257.

Segal, E., Friedman, N., Koller, D. and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nature Genetics, 36: 1090 – 1098.

Shen, R., Ghosh, D. and Chinnaiyan, A. M. (2005). Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. BMC Genomics, 5: 94.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C.M., Lonning, P.E., Brown, P.O., Borresen-Dale, A.L. and Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A, 100: 8418-23.

Storey, J. D. (2002). A direct approach to false discovery rates. J. Roy. Statist. Soc. B, 64: 479 – 498.

Storey, J. D. and Tibshirani R. (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA, 100: 9440 – 9445.

Sweet-Cordero, A., Mukherjee, S., Subramanian, A., You, H., Roix, J. J., Ladd-Acosta, C., Mesirov, J., Golub, T. R. and Jacks T. (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. Nature Genetics, 37: 48 – 55.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, 403:623-7.

van der Laan, M. J., Dudoit, S. and Pollard, K. S. (2004a). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. Statistical Applications in Genetics and Molecular Biology, 3: Article 14.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004b). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Statistical Applications in Genetics and Molecular Biology, 3: Article 15.

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., Ghosh, D., Pienta, K. J., Sewalt, R. G., Otte, A. P., Rubin, M. A., and Chinnaiyan, A. M. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature, 419: 624 – 629.

Venter, J. C., Adams, M., Myers, E. W., Li, P. W. et al. (2001). The sequence of the human genome. Science, 291: 1304 – 1351.

Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV. (2004). Differences in gene expression between B-cell chronic lymphocytic leukemia and normal B cells: a meta-analysis of three microarray studies. Bioinformatics, 20: 3166 - 3178.

Wasserman, S., and Faust, K. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, Staudt LM. (2003). A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proc Natl Acad Sci U S A., 100: 9991-6.

Yan PS, Chen CM, Shi H, Rahmatpanah F, Wei SH, Caldwell CW, Huang TH. (2001). Dissect-

ing complex epigenetic alterations in breast cancer using CpG island microarrays. Cancer Res. 2001 Dec 1;61(23):8375-80.