Chapter 19
THE GENOMIC DATA MINE

Lorraine Tanabe

National Center for Biotechnology Information, Computational Biology Branch, National Library of Medicine, Bethesda, MD 20894

Chapter Overview

The genomic data mine represents a fundamental shift from genetics to genomics, essentially from the study of one gene at a time to the study of entire genetic metabolic networks and whole genomes. Experimental laboratory data are deposited into large public repositories and a wealth of computational data mining algorithms and tools are applied to mine the data. The integration of different types of data in the genomic data mine will contribute towards an understanding of the systems biology of living organisms, contributing to improved diagnoses and individualized medicine. This chapter focuses on the genomic data mine consisting of text data, map data, sequence data, and expression data, and concludes with a case study of the Gene Expression Omnibus (GEO).

Keywords

genomics; text mining; data mining; gene expression data

> "… Medical schools, slow to recognize the profound implications of genomics for clinical medicine, have been lurching, if not stumbling, forward to embrace the genomification of medicine…"
>
> *Canadian Medical Association Journal* editorial, 2003

1. INTRODUCTION

The field of genomics began in the late 20th century with the physical and genetic mapping of genes, followed by the application of DNA sequencing technology to the genetic material of entire organisms to elucidate the blueprints of life. The main branches of genomics are distinguished as 1) structural genomics, including mapping and sequencing, 2) comparative genomics, including genetic diversity and evolutionary studies, and 3) functional genomics, the study of the roles of genes in biological systems. In addition to DNA sequencing, one of the most important technologies for genomics is DNA microarrays, which can measure the expression of thousands of genes simultaneously. Largely due to the generation of voluminous gene expression data from microarrays, genomics in the 21st century is evolving from its sequence-based origins towards a systems biology perspective which encompasses the molecular mechanisms as well as the emergent properties of a biological system.

Systems Biology is not a new research area, but it has been revitalized by genomic data. In January 2003, the Massachusetts Institute of Technology (MIT) started a Computational and Systems Biology Initiative, and Harvard and MIT's Broad Institute was specifically designed to bridge genomics and medicine. The NASA Ames Research Center currently funds the Computational Systems Biology Group, an association of statisticians, computer scientists, and biologists at Carnegie Mellon University, the University of Pittsburgh, and the University of West Florida. The Systems Biology Markup Language (SBML) is a computer-readable format for representing models of biochemical reaction networks (Hucka et al., 2003). Because human patients are biological systems, the systems biology approach has enormous potential to ease the transition of genomics knowledge from the laboratory to the clinical setting. Before this transfer of knowledge can happen, the large-scale genomics data need to be interpreted with a combination of hypothesis-driven research and data mining. This chapter will present some data mining techniques for genomic data.

Data Mining is the exploration of large datasets from many perspectives, under the assumption that there are relationships and patterns in the data that can be revealed. This can be a multi-step procedure with an automatic component followed by human investigation. It is a data-driven approach, exploratory in nature, which complements a more traditional hypothesis-driven methodology. Large-scale genetic sequence and expression data generated from high-throughput experimental techniques constitute a huge data mine from which new patterns can be discovered, contributing to a greater understanding of biological systems and their perturbations, leading to new therapeutics in medicine. The genomic data mine represents a

fundamental shift from genetics to genomics, essentially from the study of one gene at a time to the study of entire genetic and metabolic networks and whole genomes. Experimental laboratory data are deposited into large public repositories, and a wealth of computational data mining algorithms and tools are applied to mine the data.

Genomic databases are continually growing in depth and breadth. The Molecular Biology Database Collection lists many of these resources at the *Nucleic Acids Research* web site http://nar.oupjournals.org/. Each year, *Nucleic Acids Research* publishes a special database issue, including updates on many broad genomics databases like GenBank (Benson et al., 2004), the EMBL Nucleotide Sequence Database (Kulikova et al., 2004), the Gene Ontology (GO) database (Gene Ontology Consortium, 2004), the KEGG resource (Kanehisa et al, 2004), MetaCyc (Krieger et al., 2004), and UniProt(Apweiler et al., 2004), as well as more specialized databases like WormBase (Harris et al., 2004), the Database of Interacting Proteins (Salwinski et al., 2004), and the Mouse Genome Database (Bult et al., 2004).

In this chapter, the focus will be on genomic text data, map data, sequence data, and expression data. Protein 3-D structural data will not be covered. For brevity, the genomic data freely available at the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) will be highlighted. The chapter will conclude with a case study of NCBI's Gene Expression Omnibus (GEO) data mining tool.

2. OVERVIEW

The genomics data mine contains text data, map data, sequence data, and expression data.

Table 19-1. Genomics questions can be answered using different types of data

	Text Data	Map Data	Sequence Data	Expression Data
Where on the chromosome is this gene located?	X	X		
Is there a model organism with a related gene?	X	X	X	
How has this gene evolved?	X	X	X	X
What tissues is this gene expressed in?	X			X
How does a drug affect gene expression?	X			X
What is the function of this gene?	X	X	X	X

NCBI's Entrez system is a starting point for exploring these rich datasets (Schuler et al., 1996). LocusLink is a gene-centered interface to sequence

and curation data. RefSeq is a database supplying citations for transcripts, proteins, and entire genomic regions for 2000 organisms. RefSeq and LocusLink provide a non-redundant view of genes, to support research on genes and gene families, variation, gene expression, and genome annotation (Pruitt and Maglott, 2000). Unigene classifies GenBank sequences into about 108,000 gene-related groups (Schuler, 1997). Some basic questions that can be answered by mining genomics data types are summarized in Table 19-1.

2.1 Genomic Text Data

Text mining is an emerging field without a clear definition in the genomics community. Text mining can refer to automated searching of a sizeable set of text for specific facts. A more rigid definition of text mining requires the discovery of new or implicit knowledge hidden in a large text collection. In genomics, text mining can also refer to the creation of literature networks of related bimolecular entities. Text mining, like data mining, involves a data-driven approach and a search for patterns.

Scientific abstracts, full-text articles, and the internet all contain text data that can be mined for specific information or new facts. Because it contains the collective facts known about nearly all genes that have ever been studied, the genomics text data mine represents the entire genomics knowledge base. This knowledge is encoded in natural language and is a meta-level representation of the information gleaned from hypothesis-driven and numerical-data-driven experimentation.

Text mining research in genomics is a growing field of research comprising: 1) relationship mining, 2) literature networks, and 3) knowledge discovery in databases (KDD). Relationship mining refers to the extraction of facts regarding two or more biomedical entities. Literature networks are meaningful subsets of MEDLINE based on co-occurring gene names and/or functional keywords. Literature networks based on co-occurrence are motivated by the fact that functionally related genes are likely to occur in the same documents. Stapley and Benoit define *biobibliometric distance* as the reciprocal of the Dice coefficient of two genes *i* and *j*:

$$d_{ij} = \frac{|i| + |j|}{|i \cap j|} \tag{1}$$

The distance between all pairs of genes can be calculated for an entire genome and the results can be visualized as edges linking co-occurring genes (Stapley and Benoit, 2000). PubGene (Jenssen et al., 2001) adds annotation to pairs of genes using functional terminologies from MeSH and

GO. MedMiner (Tanabe et al., 1999) uses functional keywords like *inhibit, upregulate, activate,* etc. to filter the documents containing a pair of genes into subsets based on the co-occurrence of gene names in the same sentence as a functional keyword. Thematic analysis (Shatkay et al., 2000, Wilbur, 2002) finds themes in the literature, sets of related documents based on co-occurring terms. Table 17-2 summarizes some of the genomics research performed in these areas since 1998. KDD genomics tasks include prediction of gene function and location (Cheng et al., 2002) and automatic analysis of scientific papers (Yeh et al., 2003).

Table 19-2. A sample of genomic text mining. M = a MEDLINE corpus, J = biomedical journal articles, T = any biomedical text

	Date	Relation Mining	Literature Networks	Given	Returns
Sekimizu et al.	1998	X		M, verb list	Subjects, objects of verbs
BioNLP/BioJAKE Ng and Wong	1999	X		T, query term	Graphical pathways
ARBITER Rindflesch et al.	1999	X		M	Binding relationships
Blaschke et al.	1999	X		M, genes, verbs	Gene-gene relationships
MedMiner Tanabe et al.	1999	X	X	M query	Keyword summaries
Craven & Kumlien	1999	X		T, classes	Classes/ relationships
Thomas et al.	2000	X		M, verbs, frames	Filled frames
EDGAR Rindflesch et al.	2000	X		M	Gene/drug/cell relations
Stapley, Benoit	2000	X	X	M, gene list	Gene/gene networks
Shatkay et al.	2000		X	M, query term	Literature themes
Stephens et al.	2001	X	X	M, thesauri	Gene pair relationships
XplorMed Perez-Iratxeta et al.	2001	X	X	M query or T	Literature topics
MeSHmap Srinivasan P.	2001		X	M query	Searchable MeSH terms
PubGene Jenssen et al.	2001	X	X	M, gene list	Literature networks
Yakushiji et al.	2001	X		T, verb list	Predicate/arguments
PIES Wong, L.	2001	X	X	M, action verbs	Generated pathways

continued

	Date	Relation Mining	Literature Networks	Given	Returns
GENIES Friedman et al.	2001	X		T, grammar, lexicon	Gene/protein interactions
SUISEKI Blaschke, Valencia	2001	X		T, frames	Filled frames
MEDSTRACT Chang et al.	2002	X		M, relationships	Extracted relationships
Palakal et al.	2002	X		M, E/R model	Entities and relationships
Temkin et al.	2003	X		T, grammar, lexicon	Gene/protein interactions
PreBIND Donaldson et al.	2003	X		J, gene list	Protein/protein relations
MedGene Hu et al.	2003	X	X	M, disease/gene list	Gene/disease summaries
PASTA Gaizauskas et al.	2003	X		M, templates	Filled templates
MeKE Chiang, Yu	2003	X		M, gene list	Protein roles
MedScan Novichkova et al.	2004	X		M, protein list	Protein interactions
MedBlast Tu et al.	2004		X	Sequence	MEDLINE summary

2.1.1 Text Mining Methods

Methods for genomics text mining vary and can be classified into three main approaches: statistical, linguistic, and heuristic. Statistical and linguistic methods both require natural language processing (NLP), a broad expression covering computerized techniques to process human language. Statistical NLP ignores the syntactic structure of a sentence, hence it is often referred to as a "bag-of-words" approach, although it can also involve non-word features like co-occurrence, frequency, and ngrams to determine sentence and document relatedness. Often terms or other features are used in machine learning algorithms including Bayesian classification, decision trees, support vector machines (SVMs), and hidden markov models (HMMs) (Baldi and Brunak, 1998).

More linguistically-motivated approaches utilize part-of-speech (POS) tagging and/or partial or full parsing. One difference between statistical and linguistic methods is that statistical processing often discards common words like *and, or, become, when, where*, etc. (called a stop list), while linguistic techniques rely on these terms to help identify parts of speech and/or sentence syntax. In statistical NLP the resulting words or ngrams are

isolated from the larger discourse, making anaphoric reference resolution impossible. For example, in the following text, the *A2780 cells* mentioned in the first sentence are later referred to as *cells*, *oblimersen-pretreated cells*, and *these cells*. The relationships between *A2780 cells* and *temozolomide*, *PaTrin-2,* and *oblimerson* in the second sentence cannot be extracted unless *cells* are resolved to *A2780 cells*:

> Using a human ovarian cancer cell line (A2780) that expresses both Bcl-2 and MGMT, we show that cells treated with active dose levels of either oblimersen (but not control reverse sequence or mismatch oligonucleotides) or PaTrin-2 are substantially sensitized to temozolomide. Furthermore, the exposure of oblimersen-pretreated cells to PaTrin-2 leads to an even greater sensitization of these cells to temozolomide. Thus, growth of cells treated only with temozolomide (5 Î¼g/mL) was 91% of control growth, whereas additional exposure to PaTrin-2 alone (10 Î¼mol/L) or oblimersen alone (33 nmol/L) reduced this to 81% and 66%, respectively, and the combination of PaTrin-2 (10 Î¼mol/L) and oblimersen (33 nmol/L) reduced growth to 25% of control.

Linguistically-motivated methodologies adapt syntactic theory and semantic/discourse analysis to the biomedical domain. Although this is a difficult task due to the complexity of biomedical text, it is necessary for capturing the full meaning of the text.

Heuristic methods make use of biomedical domain knowledge. The manual effort required to translate expert knowledge into rules and patterns is often not prohibitive, and systems using this approach have been successful at extracting pertinent facts from text collections. However, heuristic methods often miss facts that appear in unpredicted contexts, are subject to human bias, and can have problems scaling up to large full text corpora.

Many text mining systems for genomics involve a combination of statistical, linguistic, and/or heuristic methods; for example, the PubMiner system (Eom and Zhang, 2004) uses an HMM-based POS tagger, an SVM-based named-entity tagger, a syntactic analyzer, and an event extractor that uses syntactic information, co-occurrence statistics, and verb patterns. More detail on text mining methods can be found in other chapters in Unit III of this book.

2.1.2 Knowledge Discovery

Text mining is an essential component of map, sequence, and expression data mining efforts in genomics, since no experimental results can be interpreted without reference to pre-existing knowledge. The multitude of facts stored in natural language text databases, like MEDLINE, constitute a rich source of potential new discoveries.

New information can be assembled from separate texts by *literature synthesis*, which involves finding implicit connections between facts. For example, Swanson found articles showing that fish oils cause blood and vascular changes, and connected these to separate articles revealing certain blood and vascular changes that might help patients with Raynaud's syndrome (Swanson, 1990). Two years later, a clinical trial reported the benefit of fish oil for Raynaud patients. Swanson found further examples of productive literature synthesis including connections between magnesium deficiency and migraine headaches and arginine intake and somatomedins in the blood, leading him to suppose that such connections are not rare. Weeber et al. simulated Swanson's fish oil discovery using the drug-ADR-disease (DAD) -system, a concept-based NLP system for processing PubMed documents (Weeber et al., 2000). The DAD-system uses the UMLS Metathesaurus (NLM, 2000) as a basis for text mining PubMed. Query terms are mapped to UMLS concepts using MetaMap (Aronson, 1996), and then the relevant PubMed abstracts are retrieved to a local database. The UMLS concepts contained in these abstracts are presented to the user, who selects concepts for further document retrievals. The ranking of concepts depends on their interconnection and the user formulates and checks hypotheses based on this ranking. The DAD-system has been used to mine biomedical literature on side effects and adverse drug reactions (ADR).

As an alternative to documents, words, or UMLS concepts, gene/protein relations can be used as the basic analytical unit for text mining. *Relational chaining* is the linkage of entities through their relations across multiple documents, facilitating the discovery of interesting combinations of relations that would be impossible to find in a single document. Blaschke and Valencia compared the interactions of yeast cell cycle genes/proteins before and after the year 2000 and found that recent discoveries often originated from entities near each other in previously networked relations, suggesting that initially extracted gene interaction data can be combined into a plan for knowledge discovery (Blaschke and Valencia, 2001). A different strategy for text mining with gene/protein relations involved: 1) establishment of a database of gene/protein relations extracted from MEDLINE and 2) a query mechanism to mine the database for implicit knowledge based on relational chaining. In a prototype system implementing this approach, typical

gene/protein queries resulted in PubMed documents automatically linked by gene/protein relations (*decreased_levels_of , associated_with, etc.*) (Tanabe, 2003).

2.2 Genomic Map Data

Genomic map data identify the position of a gene on a chromosome or on the DNA itself, vital information for identifying human disease genes and mutations. Chromosome maps are created by cytogenetic analysis (also known as karyotyping), linkage, or *in situ hybridization*, where a DNA probe is used to visualize the chromosomal position. Many disease-related genes are found by linkage to chromosomal regions. For example, chromosomal aberrations have been found to be associated with cancer (Mitelman et al., 1997). Physical maps show the location of a gene on the DNA itself, measured in basepairs, kilobasepairs, or megabasepairs. The Entrez Map Viewer presents genomic map data using sets of aligned chromosomal maps that can be explored at various levels of detail, including UniGene clusters. Map Viewer offers maps for a variety of organisms including mammals, plants, fungi, and protozoa (Wheeler et al., 2004). Graphical views show genes, markers, and disease phenotypes along each chromosome of an organism, as well as the genomic locations showing hits on all chromosomes. Cytogenetic map location is also available through LocusLink.

2.2.1 Finding Candidate Disease Genes

Cytogenetic map data was used for data mining by Perez-Iratxeta et al. to associate genes with genetically inherited diseases using a scoring system based on fuzzy set theory (Perez-Iratxeta, 2002). First, the system used MEDLINE to find disease and chemical terms with frequent co-occurrence in the literature. Next, the RefSeq database was mined for associations between function and chemical terms for annotated genes. Finally, the function terms, chemical terms, and disease terms were combined to get relations between diseases and protein functions. For 455 diseases with chromosomal maps, a score was assigned based on the relation of the RefSeq sequences to the disease given their functional annotation. The disease gene candidates on relevant chromosomal regions were sequence compared to the scored RefSeq sequences. Hits were scored based on the scores of RefSeq homologous sequences. In a test involving 100 known disease genes, the disease gene was among the best-scoring candidate genes with a 25% chance, and among the best 30 candidate genes with a 50% chance.

2.3 Genomic Sequence Data

DNA sequence data are made publicly available through GenBank. Nucleotide Basic Local Alignment Search Tool (BLAST) searches allow one to input nucleotide sequences and compare these against other sequences. Pairwise BLAST performs a comparison between two sequences using the BLAST algorithm. MegaBLAST allows for a sequence to be searched against a specific genome. Position-Specific Iterated (PSI)-BLAST is useful for finding very distantly related proteins. The basic BLAST algorithm looks for areas of high similarity to a query sequence in the sequence database, returning hits that are statistically significant. Non-gapped segments with maximal scores that cannot be extended or trimmed (high scoring segment pairs, HSP) represent local optimal alignments. HSPs above a score threshold are subject to gapped extensions and the best alignment is chosen. If the score of the chosen alignment is statistically significant, it is returned as a hit (Altschul et al., 1990). NCBI tools for sequence data mining include HomoloGene and TaxPlot. HomoloGene is an automated system for finding homologs among eukaryotic gene sets by comparing nucleotide sequences between pairs of organisms. Curated orthologs are incorporated from a variety of sources via LocusLink. TaxPlot is a tool for 3-way comparisons of genomes on the basis of the protein sequences they encode. A reference genome is compared to two additional genomes, resulting in a graphical display of BLAST results where each point for each predicted protein in the reference genome is based on the best alignment with proteins in each of the two genomes being compared. Generally, sequence similarity is associated with similar biological function (although this is not always the case), so mining sequence databases can lead to the discovery of new genes, regulatory elements, and retroviruses.

2.3.1 Predicting Protein Function

Sequence data can also be used to predict protein functional class. Using sequence data and other relevant features like annotation keywords (words used to describe protein function, for example, *apoptosis*), species, and molecular weight, King et al. predicted protein functional classes in *M. tuberculosis* using a combination of Inductive Logic Programming (ILP) and decision tree learning (King et al., 2000). ILP is a machine learning strategy that uses a set of positive and negative training examples to induce a theory that covers the positive but not the negative examples (Muggleton, 1991). ILP requires a set of features that can be used to construct the theory. Decision tree algorithms partition training data into a tree structure where each node denotes a feature in the training data that can be used to partition

the positive and negative examples. King et al. retrieved the sequence data with a PSI-BLAST search for homologous proteins to *M. tuberculosis* genes with known function. Relevant features were extracted including percent amino acid composition, PSI-BLAST similarity score, number of iterations, and amino acid pair frequency. ILP was used to mine for patterns in the sequence descriptions and the decision tree algorithm C4.5 (Quinlan, 1993) was used to learn rules predicting function. A simple rule example is: If the percentage composition of lysine in the gene is > 6.6%, then its functional class is "Macromolecule metabolism." This rule was 85% accurate on a test set, predicting proteins involved in protein translation. Overall, the system predicted the function of 65% of *M. tuberculosis* genes with unknown function with 60-80% accuracy.

Protein function can also be predicted using the Clusters of Orthologous Groups of proteins (COGs) database at NCBI (Tatusov et al., 1997, Tatusov et al., 2000). Orthologs are genes in different species that evolved from a common ancestor, as opposed to paralogs, which are genes within the same species that have diverged by gene duplication. COGs are determined by all-against-all sequence comparisons of genes from complete genomes using gapped BLAST. For each protein, the best hit in each of the other genomes is found and patterns of best hits determine the COGs. Each COG is assumed to have evolved from one common ancestral gene. Short stretches of DNA sequences called expressed sequence tags (ESTs) of unknown function can be mined for sequences likely to have protein function using COG information. Using more than 10,000 ESTs from dbEST (Boguski et al., 1993) and 77,114 protein sequences from COG, Faria-Campos et al. mined 4,093 ESTs for protein characterization based on homology to COG groups (Faria-Campos et al., 2003).

In addition to the global view of full genomes represented by COGs, a complementary approach detecting protein families by clustering smaller pieces of sequence space is possible. In a fully automated method, BLAST-scored sequences are clustered around a query protein using pairwise similarities, and then adjacent clusters are pooled to generate potential protein families that are similar to COGs based on a sample of 21 complete genomes (Abascal and Valencia, 2002). The clustering algorithm is a derivation of the minimum cut algorithm (Wu and Leahy, 1993). The merging algorithm pools two clusters if the relative entropy of the merged clusters decreases. Like COGs, the resulting groups can be used to predict the protein function of uncharacterized genes or ESTs.

2.4 Genomic Expression Data

Gene expression data generated from DNA microarrays, oligonucleotide chips, Digital Differential Display (DDD), and Serial Analysis of Gene Expression (SAGE) enable researchers to study genetic and metabolic networks and whole genomes in a parallel manner (Shalon et al., 1996, Spellman et al., 1998, Weinstein et al., 1997, Ross et al., 2000). These technologies can generate data for thousands of genes per experiment, creating a need for data mining strategies to interpret and understand experimental results. DNA microarrays contain probe DNA of known sequence attached to a slide, which is exposed to target samples that have been differentially labeled (Schena et al., 1995). The expression of genes in the target samples can be detected and quantified by their level of competitive hybridization to the probe DNA. Affymetrix, Inc. developed oligonucleotide chips, which use a single probe followed by exposure to target samples. DDD is a method for comparing sequence-based cDNA pools, using UniGene clusters to narrow sequences to genes expressed in humans. Serial analysis of gene expression (SAGE) is a methodology using sequence tags representing specific transcripts assembled into long molecules which are cloned and sequenced, allowing for the measurement of each transcript by the detection of its sequence tags (Velculescu et al., 1995). Microarrays and SAGE can be used complementarily, for example, microarrays can be used to identify cell-specific transcripts and SAGE can be used to determine the percentage of these that are mitochondrial (Gnatenko et al., 2003). SAGEmap at NCBI provides a mapping between SAGE tags and UniGene clusters (Lash et al., 2000).

Microarrays and Affymetrix chips have the advantage of being fast and comprehensive; however, they are expensive and are subject to hybridization and image analysis artifacts. DDD and SAGE have a cost advantage, but there are fewer data analysis tools for them and the data are not comprehensive (SAGE data are available for a limited number of organs).

Gene expression data can be combined with protein data to find key patterns involving gene expression and protein function (Nishizuka et al., 2003). Data mining for relationships between gene expression and protein function is vital, because protein function can be uncorrelated with gene expression and proteins, not genes, are usually the targets for therapeutic intervention. Since proteins are often valuable drug targets, gene and protein expression data are crucial components of what has been termed the "genomification of medicine" (CMAJ, 2003).

2.4.1 Cancer Gene Discovery

The Cancer Genome Anatomy Project (CGAP) at the National Cancer Institute was established in 1996 to uncover the "molecular anatomy" of cancer cells (Strausberg et al., 1997). Digital differential display (DDD) can be used to mine CGAP's EST databases for cancer genes. For example, Scheurle et al. used DDD to identify genes of interest in breast, colon, lung, ovary, pancreas, and prostate solid tumor tissues (Scheurle et al., 2000). DDD allows for normal and tumor cDNA libraries to be compared by generating transcript fingerprints using statistical analysis. Combined with hits from the UniGene database, DDD predicted 12 genes up- or down-regulated in colon tumor tissue, and 3 genes were verified by laboratory experimentation to fit the DDD prediction, making them potential diagnostic and therapeutic targets for colon cancer.

2.4.2 Prognosis Prediction

Gene expression data help cancer researchers depict the states of several genes at a time under varying experimental conditions. The resulting molecular profiles are useful for predicting the prognosis of some types of cancer. For example, van de Vijver et al. established a 70-gene expression profile for breast cancer metastases (van de Vijver et al., 2002). At 10 years, the probability of remaining free of distant metastases was 50.6±4.5 percent in the group with a poor-prognosis signature and 85.2±4.3 percent in the group with a good-prognosis signature. The authors concluded that the gene expression profile was a more effective predictor of disease outcome than clinical/histological criteria. A similar study by Shipp et al. analyzed the expression of 6,817 genes in tumor specimens from Diffuse large B-cell lymphoma (DLBCL) patients who received chemotherapy and applied a weighted-voting classification algorithm to identify cured versus fatal or refractory disease (Shipp et al., 2002). The weighted-voting algorithm is a supervised machine learning algorithm that distinguishes two classes of inputs. An idealized expression profile is created for each class where the expression is high in one class and low in the other class. The voting scheme involves one vote per gene for class A or class B, depending on the similarity of its expression profile to each class's idealized profile. The winner of the votes, either class A or class B, is the predicted class. The weighted-voting algorithm classified two categories of patients with very different five-year overall survival rates (70% versus 12%), indicating that supervised learning classification techniques can predict outcome in DLBCL and identify rational targets for intervention. These results suggest that the integration of a patient's particular constellation of gene expression patterns

with other diagnostic tools and resources would pave the way towards personalized medicine.

2.4.3 Tumor Classification

Different classes of tumors are associated with variations in therapeutic response to anti-cancer agents. Golub et al. determined that gene expression profiles can be used to assign tumors to established classes, using human acute leukemias as a test case (Golub et al., 1999). Distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) is imperative for effective treatment of the disease. Using an expression profile of 6,817 genes, Golub et al. applied a self-organizing map (SOM) to cluster 38 leukemia samples into AML (24/25 samples correctly classified) and ALL (10/13 samples correctly classified). An SOM is learned from an unsupervised algorithm applied to a network of nodes that tunes the inputs to pattern classes (Kohonen, 1981). These results suggest that individualized treatment would be possible using a tumor classification based on gene expression profiles.

3. CASE STUDY: THE GENE EXPRESSION OMNIBUS

The Gene Expression Omnibus (GEO) was developed to address a need in the genomics community for a standardized data repository for gene expression data, including microarray, oligonucleotide chip, hybridization filter, SAGE, and protein data (Wheeler et al., 2004). GEO data are assembled into comparable sets (GDS) which can be searched using Entrez GDS or Entrez GEO. GEO retrieves pre-computed graphical representations of experimental data, as well as gene name, GenBank accession, clone ID, ORF, mapping information, the dataset title, and additional flags regarding outliers and detection calls. Retrievals are listed in order of most-interesting-first, based on a scoring scheme which considers flagged effects, expression level, outliers, and variability.

The following examples from GEO's online tutorial illustrate the types of queries that can be posed to GEO.

To identify all dual channel nucleotide microarray experimental datasets exploring metastasis in humans, enter:

```
"dual channel"[Experiment Type] AND metastasis AND
human[Organism]
```

To view profiles of kallikrein family genes across all datasets, enter:
`kallikrein`

To limit these kallikrein retrievals to datasets investigating progesterone, enter:
`kallikrein AND progesterone`

To view profiles that fall into the top 1% abundance rank bracket in at least one sample in dataset GDS186, enter:
`GDS186 AND 100[Max Value Rank]`

A range can also be specified. To view the top 5%, enter:
`GDS186 AND 96:100[Max Value Rank]`

To view profiles that fall into the top 1% variable molecular abundance profiles in dataset GDS186, enter:
`GDS186 AND 100[Ranked Standard Deviation]`

GEO BLAST can be used to query Entrez GEO for expression profiles based on sequence similarity. GEO datasets can be browsed and mined using online tools including hierarchical clustering. Hierarchical clustering is an unsupervised method for detecting similar sets of data based on shared features, which can be visualized as a dendrogram. The connectivity of the dendrogram depends upon the clustering algorithm and similarity measure used. Details on hierarchical clustering can be found in a separate chapter in this book.

The following example begins with selecting the "Mammary epithelial cells and breast cancer" dataset GDS90, which returns a GEO record showing a summary of the data, including description of experiment, organism, type of experiment, number of probes, and date. Twenty six unordered samples are indicated, which can be selected or deselected using checkboxes. A cluster analysis can be performed using user-defined distance metrics and hierarchical clustering methods (see Figure 19-1).

4. CONCLUSIONS AND DISCUSSION

The genomics data mine includes text data, map data, sequence data, and expression data. Public repositories house much of the genomic data mine, along with computational tools to organize, integrate, and understand the data. Text data represent the genomics knowledge base and can be mined

for relationships, literature networks, and new discoveries by literature synthesis and relational chaining. Map data are crucial for the discovery of new disease genes. Sequence data can be mined to gain insight about protein function and evolution. Microarray experiments generate expression data on thousands of genes at a time, requiring data mining tools that help to visualize, analyze, and interpret the data. Gene and protein expression data are crucial components of what has been termed the "genomification of medicine" (CMAJ, 2003), and careful and effective mining of genomic data has huge potential for future clinical applications. The integration of text, map, sequence, and expression data will contribute towards an understanding of the systems biology of living organisms, contributing to improved diagnoses and individualized medicine.

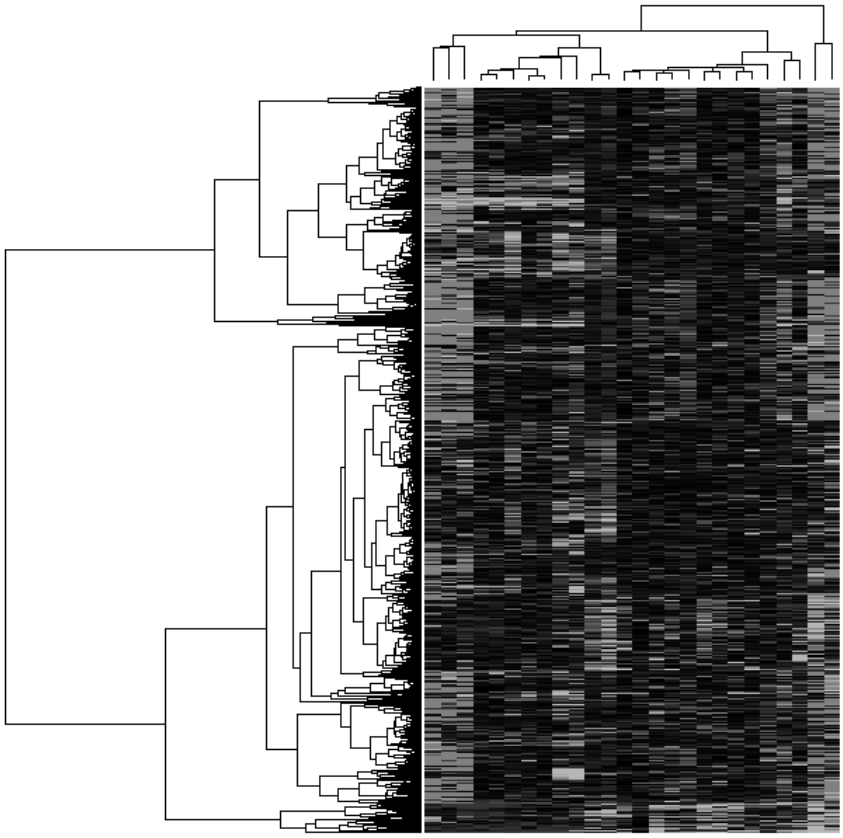

Figure 19-1. GDS90 dataset. Dark = high expression levels, light = low expression levels. Rows = genes, columns = cell/tissue type. Genes are clustered according to cell/tissue type (left hand side) and the 26 samples according to gene expression (upper area). Users can zoom in on areas of interest and download the selected data.

REFERENCES

Abascal, F. and Valencia, A. (2002). "Clustering of Proximal Sequence Space for the Identification of Protein Families," *Bioinformatics*, 18:908-21.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990). "Basic Local Alignment Search Tool," *J Mol Biol.*, 215:403-10.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., and Yeh, L.-S.L. (2004). "UniProt: The Universal Protein Knowledgebase," *Nucl. Acids. Res.,* 32:D115-D119.

Aronson, A.R. (1996). "The Effect of Textual Variation on Concept Based Information Retrieval," *Proc AMIA Annu Fall Symp.*:373-7.

Baldi, P. and Brunak, S. (1998). *Bioinformatics: The Machine Learning Approach (Adaptive Computation & Machine Learning)*, MIT Press.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2004). "GenBank: Update," *Nucl. Acids. Res.*, 32:D23-D26.

Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A. (1999). "Automatic Extraction of Biological Information from Scientific Text: Protein-protein Interactions," in *Proc Int Conf Intell Syst Mol Biol.*, 60-7.

Blaschke, C. and Valencia, A. (2001). "The Potential Use of SUISEKI as a Protein Interaction Discovery Tool," *Genome Inform Ser Workshop Genome Inform.*, 12:123-34.

Blaschke, C. and Valencia, A. (2002). "The Frame-based Module of the SUISEKI Information Extraction System," *IEEE Intelligent Systems*, 17(2):14-20.

Boguski, M.S., Lowe, T.M., Tolstoshev, C.M. (1993). "DbEST--Database for 'Expressed Sequence Tags'," *Nat Genet.*, 4:332-3.

Bult, C.J., Blake, J.A., Richardson, J.E., Kadin, J.A., Eppig, J.T. and the Mouse Genome Database Group (2004). "The Mouse Genome Database (MGD): Integrating Biology with the Genome," *Nucl. Acids. Res.*, 32:D476-D481.

Chang, J.T., Schutze, H. and Altman, R.B. (2002). "Creating an Online Dictionary of Abbreviations from MEDLINE," *J Am Med Inform Assoc.*, 9(6):612-20.

Cheng, J., Hatzis, C., Hayashi, H., Krogel, M.A., Morishita, S., Page, D. and Sese, J. (2002). "KDD Cup 2001 Report," *SIGKDD Explorations*, 3(2):47--64.

Chiang, J.H. and Yu, H.C. (2003). "MeKE: Discovering the Functions of Gene Products from Biomedical Literature via Sentence Alignment," *Bioinformatics*, 19(11):1417-22.

Chiang, J.H., Yu, H.C., and Hsu, H.J. (2004). "GIS: A Biomedical Text-mining System for Gene Information Discovery," *Bioinformatics*, 20(1):120-1.

CMAJ. (2003). "The Genomification of Medicine," *CMAJ*, 168(8):949-951.

Craven, M. and Kumlien, J. (1999). "Constructing Biological Knowledge Bases by Extracting Information from Text Sources," *Proc Int Conf Intell Syst Mol Biol.*:77-86.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). "Extracting Human Protein Interactions from MEDLINE Using a Full-sentence Parser," *Bioinformatics*, 20(5):604-11.

Donaldson, I., Martin, J., De Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G.D., Michalickova, K., Pawson, T., and Hogue, C.W. (2003). "PreBIND and Textomy--Mining the Biomedical Literature for Protein-protein Interactions Using a Support Vector Machine," *BMC Bioinformatics*, 4(1):11.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository," *Nucleic Acids Res.*, 30(1):207-210.

Eom, J.-H. and Zhang, B.-T. (2004). "PubMiner: Machine Learning-based Text Mining Systems for Biomedical Information Mining," *Lecture Notes in Artificial Intelligence*, 3192:216-225.

Faria-Campos, A.C., Cerqueira, G.C., Anacleto, C., De Carvalho, C.M., Ortega, J.M. (2003). "Mining Microorganism EST Databases in the Quest for New Proteins," *Genet Mol Res*., 2:169-77.

Friedman, C., Kra, P., Yu, H., Krauthammer, M. and Rzhetsky, A. (2001). "GENIES: A Natural-language Processing System for the Extraction of Molecular Pathways from Journal Articles," *Bioinformatics*, 17 Suppl 1:S74-82.

Gaizauskas, R., Demetriou, G., Artymiuk, P.J., and Willett, P. (2003). "Protein Structures and Information Extraction from Biological Texts: The PASTA System," *Bioinformatics*, 19(1):135-43.

Gene Ontology Consortium. (2004). "The Gene Ontology (GO) Database and Informatics Resource," *Nucl. Acids. Res.*, 32:D258-D261.

Gnatenko, D.V., Dunn, J.J., McCorkle, S.R., Weissmann, D., Perrotta, P.L., Bahou, W.F. (2003). "Transcript Profiling of Human Platelets Using Microarray and Serial Analysis of Gene Expression," *Blood*, 101:2285-93.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286:531537.

Hahn, U., Romacker, M., and Schulz, S. (2002). "Creating Knowledge Repositories from Biomedical Reports: The MEDSYNDIKATE Text Mining System," in *Pacific Symposium on Biocomputing:*338-49.

Harris, T.W., Chen, N., Cunningham, F., Tello-Ruiz, M., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Chan, J., Chen, C.-K., Chen, W.J., Davis, P., Kenny, E., Kishore, R., Lawson, D., Lee, R., Muller, H.-M., Nakamura, C., Ozersky, P., Petcherski, A., Rogers, A., Sabo, A., Schwarz, E.M., Van Auken, K., Wang, Q., Durbin, R., Spieth, J., Sternberg, P.W. and Stein, L.D. (2004). "WormBase: A Multi-species Resource for Nematode Biology and Genomics," *Nucl. Acids. Res.* 32:D411-D417.

Hu, Y., Hines, L.M., Weng, H., Zuo, D., Rivera, M., Richardson, A., and LaBaer, J. (2003). "Analysis of Genomic and Proteomic Data Using Advanced Literature Mining," *J Proteome Res.*, 2(4):405-12.

Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A.A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J. H., Hunter, P.J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novere, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M. R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., and Wang, J. (2003). "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models," *Bioinformatics*, 19(4):524-531.

Jenssen, T.K., Laegreid, A., Komorowski, J., and Hovig, E. (2001). "A Literature Network of Human Genes for High-throughput Analysis of Gene Expression," *Nat Genet.*, 28(1):21-8.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004). "The KEGG Resource for Deciphering the Genome," *Nucl. Acids. Res.*, 32:D277-D280.

King, R.D., Karwath, A., Clare, A., Dehaspe, L. (2000). Accurate Prediction of Protein Functional Class from Sequence in the Mycobacterium Tuberculosis and Escherichia Coli Genomes Using Data Mining," *Yeast*., 17:283-93.

Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S.Y. and Karp, P.D., 2004). "MetaCyc: A Multiorganism Database of Metabolic Pathways and Enzymes," *Nucl. Acids. Res.*, 32:D438-D442.

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., Van Den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Garcia-Pastor, M., Harte, N., Kanz, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Stoehr, P., Stoesser, G., Tuli, M. A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2004). "The EMBL Nucleotide Sequence Database," *Nucl. Acids. Res.*, 32:D27-D30.

Lash, A.E., Tolstoschev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., and Altschul, S.F. (2000). "SAGEmap: A Public Gene Expression Resource," *Genome Res.*, 10(7):1051-60.

Kohonen, T. (1981a). "Automatic Formation of Topological Maps of Patterns in a Self-organizing System," in *Proceedings of 2SCIA, Scand. Conference on Image Analysis*, Helsinki, Finland:214-220.

Mitelman, F., Mertens, F., Johansson, B. (1997). "A Breakpoint Map of Recurrent Chromosomal Rearrangements in Human Neoplasia," *Nature Genet.*, 15:417-474.

Muggleton, S. (1991). "Inductive Logic Programming," *New Generation Computing*, 8:295-318.

Ng, S.K. and Wong, M. (1999). "Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts," *Genome Inform Ser Workshop Genome Inform.*, 10:104-112.

Nishizuka, S., Charboneau, L., Young, L., Major, S., Reinhold, W.C., Waltham, M., Kouros-Mehr, H., Bussey, K.J., Lee, J.K., Espina, V., Munson, P.J., Petricoin, E. 3rd, Liotta, L.A., Weinstein, J.N. (2003). "Proteomic Profiling of the NCI-60 Cancer Cell Lines Using New High-density Reverse-phase Lysate Microarrays," *Proc Natl Acad Sci U S A*, 100:14229-34.

NLM. (2000). Unified Medical Language System Knowledge Sources.

Novichkova, S., Egorov, S. and Daraselia, N. (2003). "MedScan, A Natural Language Processing Engine for MEDLINE Abstracts," *Bioinformatics*, 19(13):1699-706.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi T. (2001). "Automated Extraction of Information on Protein-protein Interactions from the Biological Literature," *Bioinformatics*, 17(2):155-61.

Palakal, M., Stephens, M., Mukhopadhyay, S., Raje, R., and Rhodes, S.J. (2002). "A Multi-level Text Mining Method to Extract Biological Relationships," *IEEE CSB:*97-108.

Perez-Iratxeta, C., Bork, P. and Andrade, M. A. (2001). "XplorMed: A Tool for Exploring MEDLINE Abstracts," *Trends Biochem Sci.*, 26(9):573-5.

Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2002). "Association of Genes to Genetically Inherited Diseases Using Data Mining," *Nat Genet.*, 31(3):316-9.

Pruitt, K.D. and Maglott, D.R. (2001). "RefSeq and LocusLink: NCBI Gene-centered Resources," *Nucleic Acids Res.*, 29(1):137-140.

Pustejovsky, J., Castano, J., Zhang, J., Kotecki, M., and Cochran, B. (2002). "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," *Pac Symp Biocomput.:*362-73.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Rindflesch, T.C., Hunter, L., and Aronson, A.R. (1999). "Mining Molecular Binding Terminology from Biomedical Text," *Proc AMIA Symp.:*127-31.

Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van De Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon,

D., Myers, T.G., Weinstein, J.N., Botstein, D., Brown, P.O. (2000). "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nat Genet.*, 24(3):227-35.

Salwinski, L, Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). "The Database of Interacting Proteins: 2004 Update," *Nucl. Acids. Res.*, 32:D449-D451.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995). "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science*, 270:467-70.

Scheurle, D., DeYoung, M.P., Binninger, D.M., Page, H., Jahanzeb, M., Narayanan, R. (2000). "Cancer Gene Discovery Using Digital Differential Display," *Cancer Res.*, 60:4037-43.

Schuler, G.D., Epstein, J.A., Ohkawa, H., and Kans, J.A. (1996). "Entrez: Molecular Biology Database and Retrieval System," *Methods Enzymol.*, 266:141-62.

Schuler, G.D. (1997). "Pieces of the Puzzle: Expressed Sequence Tags and the Catalog of Human Genes," *J. Mol Med.*, 75:694-698.

Sekimizu, T., Park, H.S., Tsujii, J. (1998). "Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts," in *Proc. 9th Workshop Genome Informatics*, Universal Academy Press, Tokyo:62-71.

Shalon, D., Smith, S.J., and Brown, P.O. (1996). "A DNA Microarray System for Analyzing Complex DNA Samples Using Two-Color Fluorescent Probe Hybridization," *Genome Res.*, 6:639-645.

Shatkay, H., Edwards, S., Wilbur, W.J., and Boguski, M. (2000). "Genes, Themes and Microarrays: Using Information Retrieval for Large-scale Gene Analysis," in *Proc Int Conf Intell Syst Mol Biol.* 8:317-28.

Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., and Golub, T.R. (2002). "Diffuse Large B-cell Lymphoma Outcome Prediction by Gene-expression Profiling and Supervised Machine Learning," *Nat Med.*, 8(1):68-74.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B. (1998). "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces Cerevisiae By Microarray Hybridization," *Mol Biol Cell.*, 9:3273-3297.

Srinivasan, P. (2001). "MeSHmap: A Text Mining Tool for MEDLINE," in *Proc AMIA Symp.:*642-6.

Stapley, B.J. and Benoit, G. (2000). "Biobibliometrics: Information Retrieval and Visualization from Co-occurrences of Gene Names in Medline Abstracts," in *Pac Symp Biocomput.:*529-40.

Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R., Mostafa, J. (2001). "Detecting Gene Relations from Medline Abstracts," *Pac Symp Biocomput.:*483-95.

Strausberg, R.L., Dahl, C.A., and Klausner, R.D. (1997). "New Opportunities for Uncovering the Molecular Basis of Cancer," *Nat. Genet.*, Spec No 17:415–416.

Swanson, D.R. (1990). "Medical Literature as a Potential Source of New Knowledge," *Bull Med Libr Assoc.*, 78:29-37.

Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L., and Weinstein, J.N. (1999). "MedMiner: An Internet Text-mining Tool for Biomedical Information, with Application to Gene Expression Profiling," *Biotechniques*, 27(6):1210-4, 1216-7.

Tanabe, L. (2003). "Text Mining the Biomedical Literature for Genetic Knowledge [dissertation]," George Mason University, *AAT 3079362*.

Tatusov, R.L., Koonin, E.V., Lipman, D.J. (1997). "A Genomic Perspective on Protein Families," *Science*, 278:631-7. Genet Mol Res. 2003 Mar 31;2(1):169-77.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V. (2000). "The COG Database: A Tool for Genome-scale Analysis of Protein Functions and Evolution," *Nucleic Acids Res.*, 28:33-6.

Temkin, J.M. and Gilder, M.R. (2003). "Extraction of Protein Interaction Information from Unstructured Text Using a Context-free Grammar," *Bioinformatics*, 19(16):2046-53.

Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. (2000). "Automatic Extraction of Protein Interactions from Scientific Abstracts," in *Pac Symp Biocomput.* 5:538-549.

Tu, Q., Tang, H. and Ding, D. (2004). "MedBlast: Searching Articles Related to a Biological Sequence," *Bioinformatics*, 20(1):75-7.

van De Vijver, M.J., He, Y.D., Van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., Van Der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., and Bernards, R. (2002). "A Gene-expression Signature as a Predictor of Survival in Breast Cancer," *N Engl J Med*., 347(25):1999-2009.

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W. (1995). "Serial Analysis of Gene Expression," *Science*, 270:484-7.

Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., De Jong-van Den Berg, L.T., and Vos, R., (2000). "Text-based Discovery in Biomedicine: The Architecture of the DAD-system," in *Proc AMIA Symp.*:903-7.

Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J., Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., Buolamwini, J.K., Van Osdol, W.W., Monks, A.P., Scudiero, D.A., Sausville, E.A., Zaharevitz, D.W., Bunow, B., Viswanadhan, V.N., Johnson, G.S., Wittes, R.E., and Paull, K.D. (1997). "An Information-intensive Approach to the Molecular Pharmacology of Cancer," *Science*, 275:343-349.

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., and Wagner, L. (2003). "Database Resources of the National Center for Biotechnology Information," *Nucleic Acids Res.*, 31(1):28-33.

Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Suzek, T.O., Tatusova, T.A., and Wagner, L. (2004). "Database Resources of the National Center for Biotechnology Information: Update," *Nucleic Acids Res.*, 32:D35-40.

Wilbur, W.J. (2002). "A Thematic Analysis of the AIDS Literature," in *Pac Symp Biocomput.*:386-97.

Wong, L. (2001). "PIES, a Protein Interaction Extraction System," in *Pac Symp Biocomput.*:520-31.

Wu, Z. and Leahy, R. (1993). "An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Application to Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:1101-13.

Yakushiji, A., Tateisi, Y., Miyao, Y., and Tsujii, J. (2001). "Event Extraction from Biomedical Papers Using a Full Parser," in *Pac Symp Biocomput.*:408-19.

Yeh, A.S., Hirschman, L., Morgan, A.A. (2003). "Evaluation of Text Data Mining for Database Curation: Lessons Learned from the KDD Challenge Cup," *Bioinformatics*, 19 Suppl 1:i331-9.

SUGGESTED READINGS

Baldi, P. and Brunak, S., 1998, Bioinformatics: The Machine Learning Approach (Adaptive Computation & Machine Learning), MIT Press.
Machine learning Methods are covered (including neural networks, hidden Markov Models, and belief networks), aimed at biologists and biochemists. It also allows physicists/mathematicians/computer scientists to explore applications of Machine learning in Molecular biology.

Causton, H., Quackenbush, J. and Brazma, A.. (2003, *Microarray Gene Expression Data Analysis: A Beginner's Guide*, Blackwell Publishers.
Microarray experimental design and analysis, geared towards graduate students and researchers in bioinformatics, with emphasis on underlying concepts.

Jurafsky, D. and Martin, J. H.. (2000, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice-Hall.
Large reference on foundations of natural language and speech processing.

Korf, I., Yandell, M., and Bedell, J., Eds.. (2003, *BLAST. The Definitive Guide. Basic Local Alignment Search Tool*, O'Reilly & Associates, 1st ed., Sebastopol, CA.
A detailed look at the BLAST suite of tools. It enables users to experiment with parameters and analyze their results, and includes tutorial and reference sections.

Shatkay H., and Feldman R.. (2003, Mining the biomedical literature in the genomic era: an overview, *J Comput Biol.* **10**:821-55.
A comprehensive review of the state of the art in biomedical text Mining.

ONLINE RESOURCES

Abascal and Valencia's Protein Family Annotation
http://www.pdg.cnb.uam.es/funcut.html

Broad Institute's datasets
http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
http://www.broad.mit.edu/mpr/lymphoma/

Cancer Genome Anatomy Project, NCI
http://cgap.nci.nih.gov/

Jeffrey Chang's BioNLP Server, Stanford University
http://bionlp.stanford.edu/

Critical Assessment of Information Extraction Systems in Biology, CNB Protein Design Group and MITRE Corp.
http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

European Molecular Biology Laboratory (EMBL), Links to XplorMed and other genomic data Mining tools http://www-db.embl-heidelberg.de/jss/SearchEMBL?services=x

GENIA corpus for biomedical NLP, University of Tokyo
http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/

Genomics and Bioinformatics Group, Molecular Pharmacology, NCI
http://discover.nci.nih.gov/index.jsp

GEO tutorial, NCBI
http://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html

Kyoto Encyclopedia of Genes and Genomes (KEGG)
http://www.genome.ad.jp/kegg/kegg.html

Lymphoma/Leukemia Molecular Profiling Project data
http://llmpp.nih.gov/lymphoma/

Mitelman Database of Chromosome Aberrations in Cancer
http://cgap.nci.nih.gov/Chromosomes/Mitelman

NCBI Data Mining Tools including BLAST, COGs, Map Viewer, LocusLink, and UniGene
http://www.ncbi.nlm.nih.gov/Tools/

Oxford University, ILP applications and datasets
http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/applications.html

Predicting Protein Function from Sequence using Machine Learning
http://www.aber.ac.uk/~dcswww/Research/bio/ProteinFunction/

PubGene literature and sequence networks
http://www.pubgene.org/tools/Network/Browser.cgi

SAGEmap, NCBI
http://www.ncbi.nlm.nih.gov/SAGE/

Leming Shi's site on DNA Microarrays
http://www.gene-chips.com/The Institute for Genomic Research (TIGR)
http://www.tigr.org/

Stanford University, Yeast cell cycle data
http://genome-www.stanford.edu/cellcycle/data/rawdata/

UMLS Knowledge Source Server, NLM
http://umlsks4.nlm.nih.gov/

QUESTIONS FOR DISCUSSION

1. How will genomics change the practice of medicine? How long will it take until personalized medicine is possible? What are some of the obstacles that will need to be overcome?

2. What role does the genomic data mine have in systems biology? How can data from different perspectives be integrated into a complete picture of a biological system?

3. Describe the genomics knowledge base – what is it? What are its themes? What is it missing? How can the missing knowledge be discovered?

4. What are the potential benefits of text mining the biomedical literature? How can this potential be realized?

5. Perez-Iratxeta et al. used cytogenetic data and functional annotation to score candidate disease genes using fuzzy set theory. Describe an alternative approach to finding candidate disease genes.

6. NCBI's TaxPlot is a tool for 3-way comparisons of genomes on the basis of the protein sequences they encode. What are some applications of TaxPlot? What are some of the questions that can be explored using this tool?

7. Use GEO to explore a dataset of interest to you. Interpret the hierarchical cluster tree. What does PCA tell you about this dataset?

8. How do data-driven and hypothesis-driven genomics methodologies complement each other? What are the advantages/disadvantages of each?

9. How much genomics should medical schools require? What would be the best way to teach medical students genomics? Do bioinformaticians and computational biologists need to know biology and/or medicine? If so, how much?

10. Describe an ideal genomics data mining tool that would help researchers understand human diseases. What would the program need to do? What data would it take as input? What output would it return? How accurate would it need to be? How fast would it need to run? Would the system require human reasoning or be fully automatic? How could its performance be tested?