Exploring Genomic Medicine Using Integrative Biology

by

Atul Janardhan Butte

B.A. Computer Science
Brown University, 1991

M.D.
Brown University Medical School, 1995

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN
PARTIAL FULFILLMENT OF THE REQURIEMENTS FOR THE
DEGREE OF

DOCTOR OF PHILOSOPHY IN HEALTH SCIENCES AND TECHNOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2004

Signature of Author: ______________________________
Harvard-MIT Division of Health Sciences and Technology
May 12, 2004

Certified by: ______________________________
Isaac Kohane, M.D., Ph.D.
Associate Professor of Pediatrics, Harvard Medical School
Henderson Professor of Health Sciences and Technology
Harvard-MIT Division of Health Sciences and Technology
Thesis Supervisor

Accepted by: ______________________________
Martha L. Gray, Ph.D.
Edward Hood Taplin Professor of Medical and Electrical Engineering
Co-Director, Harvard-MIT Division of Health Sciences and Technology

Exploring Genomic Medicine Using Integrative Biology

by

Atul Janardhan Butte

SUBMITTED TO THE HARVARD-MIT DIVISION OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQURIEMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN HEALTH SCIENCES AND TECHNOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ABSTRACT

Instead of focusing on the cell, or the genotype, or on any single measurement modality, using integrative biology allows us to think holistically and horizontally. A disease like diabetes can lead to myocardial infarction, nephropathy, and neuropathy; to study diabetes in genomic medicine would require reasoning from a disease to all its various complications to the genome and back.

I am studying the process of intersecting nearly-comprehensive data sets in molecular biology, across three representative modalities (microarrays, RNAi and quantitative trait loci) out of the more than 30 available today. This is difficult because the semantics and context of each experiment performed becomes more important, necessitating a detailed knowledge about the biological domain. I addressed this problem by using all public microarray data from NIH, unifying 50 million expression measurements with standard gene identifiers and representing the experimental context of each using the Unified Medical Language System, a vocabulary of over 1 million concepts. I created an automated system to join data sets related by experimental context. I evaluated this system by finding genes significantly involved in multiple experiments directly and indirectly related to diabetes and adipogenesis and found genes known to be involved in these diseases and processes. As a model first step into integrative biology, I then took known quantitative trait loci in the rat involved in glucose metabolism and build an expert system to explain possible biological mechanisms for these genetic data using the modeled genomic data.

The system I have created can link diseases from the ICD-9 billing code level down to the genetic, genomic, and molecular level. In a sense, this is the first automated system built to study the new field of genomic medicine.

Biographical Note

Atul Butte is currently on staff in the Children's Hospital Informatics Program, is a practicing pediatric endocrinologist at Children's Hospital, Boston, and is an Instructor at Harvard Medical School. Dr. Butte received his undergraduate degree in Computer Science from Brown University in 1991, and worked in several stints as a software engineer at Apple Computer (on the System 7 team) and Microsoft Corporation (on the Excel team). He graduated from the Brown University School of Medicine in 1995, during which he worked as a research fellow at NIDDK through the Howard Hughes/NIH Research Scholars Program. He completed his residency in Pediatrics and Fellowship in Pediatric Endocrinology in 2001, both at Children's Hospital, Boston.

Dr. Butte has authored 25 publications in bioinformatics, medical informatics, and molecular diabetes and has delivered more than 30 presentations world-wide on bioinformatics, including four at the National Institutes of Health. During his research work under Dr. Isaac Kohane, he developed a novel methodology for analyzing large data sets of RNA expression, called Relevance Networks. This technique was published in the Proceedings of the National Academy of Science (2000, 97:12182). Dr. Butte's recent awards include the 2003 Emory University School of Medicine, Pathology Residents' Choice Award, 2002 American Association for Clinical Chemistry Outstanding Speaker Award, 2002 Endocrine Society Travel Award based on presentation merit, 2001 American Association for Cancer Research Scholar-In-Training Award and the 2001 Lawson Wilkins Pediatric Endocrine Society Clinical Scholar Award. Dr. Butte's research is supported by grants from NCI, NIDDK, NHLBI, NINDS, NIAID, NLM, the Endocrine Fellows Foundation, the Genentech Center for Clinical Research and Education, the Lawson Wilkins Pediatric Endocrinology Society, and Merck. Along with Isaac Kohane and Alvin Kho, Dr. Butte has co-authored one of the first books on microarray analysis titled "Microarrays for an Integrative Genomics" published by MIT Press.

ॐ

Acknowledgements

More than 10 years ago, I first heard of Dr. Isaac Kohane while I interviewed at Children's Hospital for a residency position. I knew immediately that I had to work with him. Six years ago, he introduced me to the new field of functional genomics as he handed me one of the first microarray data sets. My life has forever changed. Dr. Isaac Kohane, my mentor and friend, I cannot thank you enough.

Six years ago, I met Dr. Peter Szolovits as I enrolled as a student at MIT. Dr. Szolovits introduced me to courses at MIT that first inspired me to apply data modeling techniques to genomic data, which led to my Masters Degree in Medical Informatics. He has always encouraged discipline in my writing. Dr. Szolovits will always be a prominent role-model for me, as my work increasingly spans both clinical and bioinformatics. My work has forever changed. Dr. Szolovits, I cannot thank you enough.

More than 10 years ago, I heard of the incredible work in the field of insulin receptor signal transduction being accomplished by Dr. C. Ronald Kahn at the Joslin Diabetes Center. I am still amazed that I continue to have the opportunity to benefit from his wisdom, knowledge, guidance, and mentorship. Because of Dr. Kahn, I now have the ability to focus my bioinformatics skills and ideas on the growing problem of diabetes mellitus. My career has forever changed. Dr. Kahn, I cannot thank you enough.

My wife, Dr. Tarangini Deshpande, entered my life 4½ years ago. I cannot imagine life without her. Beyond absorbing my role and responsibilities in the home while this dissertation was being written, she continues to inspire my life and my work, from the most theoretical to the deepest technical levels. In all ways, I could not have focused on this thesis without her. I cannot thank her enough.

A little girl entered my life 16 months ago. Her name, Kimayani, literally means "by a miracle." I apologize to her for the time away that this work required, and I promise to make it up to her starting now.

My parents, Janardhan and Mangala Butte, have encouraged my academic development since birth. For heading out at night to purchase new toys or puzzles immediately after I mastered the ones I had as a toddler, to teaching me biology in elementary school, to introducing me to computer programming as a teenager, I cannot thank them enough.

My brother, Dr. Manish Butte, who inspires me by pushing himself constantly to learn and develop as a physician scientist. I cannot thank him enough.

My inlaws, Sudhir and Neela Deshpande, have been enormously helpful in caring for Kimayani, allowing me to intensely pursue this academic work. Their love and support for me have been immeasurable. I cannot thank them enough.

My fellow colleagues, faculty, fellows, students and staff of the Children's Hospital Informatics Program, who have provided an inspirational sounding-board and test-bed for many of my ideas over the past six years.

Dr. Joseph Majzoub, for supporting my pursuit of this Ph.D. while I was still a fellow and attending physician in the Division of Pediatric Endocrinology at Children's Hospital.

My grandparents, aunts and uncles, and ancestors for their continued blessings.

Without financial support from the following institutions, I could not have completed this work: the National Library of Medicine, the National Heart Lung and Blood Institute, the National Institute of Diabetes and Digestive and Kidney Diseases, the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, the Endocrine Fellows Foundation, the Lawson Wilkins Pediatric Endocrinology Society, the Harvard Center for Neurodegeneration and Repair, the Genentech Center for Clinical Research and Education, the Merck / Massachusetts Institute of Technology fellowship, Children's Hospital Division of Endocrinology, and the Harvard-MIT Division of Health Sciences and Technology.

1. Introduction

Genomic medicine has been defined by Alan Guttmacher and Francis S. Collins as the application of our rapidly expanding knowledge of the human genome to medical practice. [4] We commonly define genomic medicine by the individual experimental modalities available to study the genome. As an example of this, as of this writing, there are over 3,600 publications in MEDLINE referencing the use of single nucleotide polymorphisms in humans, and over 7,600 publications referencing the use of microarrays or gene expression profiling in humans. However, there are currently fewer than 150 publications in which both technologies are cited together, and most of these are review articles.

The current movement in molecular biology has led to a revolution in the tools available to study and deconstruct processes and systems at a new resolution and level of comprehensiveness. This movement has already started to improve our understanding of human health. Technologies such as DNA sequencing and parallel expression measurements by microarray have allowed for novel diagnostic tests to assist in diagnosing diseases that are sometimes difficult to distinguish [5], to differentiate subgroups of disease that differ in prognosis [6], and to determine populations of patients that may respond to novel therapeutics. [7,8]

However, genomic medicine has the potential to be more than just the sum of our measurement modalities. A complex disease may have multiple manifestations, complications and models. Physiology and diseases (pathophysiology) can not only be described by component profiles in time and space (e.g. in what part of a body does a disease occur, the timing of disease progression), by also by the coordinate values of all molecular biological measurements and how they change across time and space. The more dimensions we consider coordinately in an analysis and the more components we consider in each dimension, the most accurate the picture we get of the disease process. I hypothesize that there may be synergy in applying all measurements and modalities in a coordinated manner to address a complex disease. To date, a systematic study of integration across measurements and experimental modalities in molecular biology has not been done. This may be because of the enormous scope of the problem. Sequencing and microarrays are just two of at least 30 large-scale measurement or experimental modalities available to investigators in molecular biology (listed in Table 1).

Leroy Hood has noted "what distinguishes systems biology from the more classical biology of the past 35 years or so, which looked at genes and proteins one at a time, is the attempt to look at all, or at least most, of the elements and their interrelationships." [9] As distinguished from "genome-scale", which carries an implication of genes being measured, and "large-scale," which does not carry an implication as to how large, I am using the term "nearly-comprehensive" for these measurement and experimental modalities, in that they represent a catalog of conditions and quantities that are close to being fully inclusive. I am not restricting to modalities with finite catalogs; I will also use the term "nearly-comprehensive" for large, uncountable catalogs where some strategy has been applied to add order to these catalogs.

These nearly-comprehensive modalities can be divided into three broad categories:

1. Nearly-comprehensive measured data specific to an experimental context: Results from measurements systems that output quantitative or qualitative data for a large set of related properties (such as nucleotide sequence, gene expression, or protein identification), yet are specific for a given context (such as cell type, species, or physiological context).

2. Nearly-comprehensive computed data specific to an experimental context: Calculated data derived from data in the first category.

3. Nearly-comprehensive applied contexts: A large set of related conditions (such as a comprehensive sequential gene deletion or suppression) under which a measurement or set of measurements are made. This category refers to multiple environments into which a cell or organism may be placed, as opposed to the first category, which refers to multiple measurements from a cell or organism.

Intersection of data sets *within a single measurement modality* is commonly done for a variety of reasons. This is made casually apparent by the scores of papers in which a Venn diagram is sketched showing the lists of genes significantly up and down regulated in one condition that intersect with the corresponding lists under other conditions. [10-13] Free and commercial software packages, such as Bullfrog, GeneSpring, and others, offer functions to create Venn diagrams from gene expression data analysis. [14,15]

Drawing conclusions from the integration of microarray data sets is an important inferential process that requires an understanding of the implications and semantics behind set operations such as *union, intersection,* and *difference,* when applied to expression data—a single measurement modality. This dissertation will address similar inferential processes involved in the combination of large-scale measurements *across modalities*. Robust understanding of the many aspects of disease requires that researchers integrate across multiple classes of data. The number of (and prior evidence for) generated hypotheses could increase on the order of n^2, where n is the number of near-comprehensive modalities available. This is the equivalent of Metcalf's law, stating that a network becomes more valuable as it reaches more users. [16] It also generates a lot more "red-herrings" or false positives, a phenomenon which is well known to researchers just working with one large-scale modality like expression microarrays.

While there are large classes of important inferential processes that can be brought to bear to research in integrative biology, the science behind these inferences has not been fully worked out. How does an expression experiment in mouse inform one about co-expression in humans? How is a human association study informed by quantitative trait loci in the rat? How does a mouse expression study relate to a human proteomic study? With the large amounts of data already collected and available online, these are questions that need to be asked if we are to realize the promise of "integrative biology." [17-27]

John Weinstein has used the term "integromic" to describe the application of genomic, proteomic, and bioinformatics methods to yield a validated answer. [28] Here, I will generalize, and use the term "*integrome*" to mean the space of inferences from all possible combinations of large-scale modalities in biology and medicine. I see as an important challenge the development of robust and specific automated inferential processes to help map out the integrome, similar to how development of automated sequencers was crucial in mapping out the genome.

An analogy to the integrome problem may be found in development of web-services

Technically, most of the modalities listed in Table 1 can be enumerated in terms of their constituent nucleotides, genes, proteins, or chemicals, and identifiers and vocabularies to describe these atoms have already been under development. [29-31] Increasing numbers of standards have been proposed for the representation or storage of these data. [32-37] Increasing numbers of international databases have been set up to store these data. [38-40]

However, even though analyzing the results of each individual modality may be routine, understanding how to consider results across different modalities has not been comprehensively described scientifically. A close analogy to this may be found in the domain of web-services in the information technology industry. XML, a standard text-based method for information storage and representation, has been increasingly accepted for data exchange. Standards for web-services (called Simple Object Access Protocol, SOAP, and Web Services Description Language, WSDL) have been developed allowing software to automatically store, retrieve, and exchange data across the Internet. Lincoln Stein and others have promoted the use of these standards for life-sciences data; for instance, Ensembl allows web-services based queries of its databases. [41,42]

Even though SOAP specifies the format of exchanged messages, the actual content of the messages needed to achieve complex operations is unspecified. Thus, even if there is agreement as to how to send messages, without higher level specifications, methods to achieve particular tasks will evolve differently, leading to eventual incompatibility. With this realization, the World Wide Web Consortium has been sponsoring the Web Services Choreography Working Group, which is describing higher level requirements around these services. The focus for this group is on how to accomplish the task, versus how to structure the message.

Analogous to this, I am proposing a model for understanding how nearly-comprehensive data can be joined, what questions can be asked from this joining, and how to draw inferences from these operations. Our focus in connecting nearly-comprehensive data sets is on how to accomplish and understand these experiments, versus how to structure the file formats for the individual modalities. Similar to the Web Services Choreography Working Group, our goal is to describe these processes in a way such that automated systems can be designed and implemented to operate these functions.

Challenges in understanding the integrome

Understanding how to draw conclusions from the integrome is challenging for several reasons. First, the context of the samples used becomes more important, necessitating a detailed knowledge about the biological domain. Yet this type of information is often represented by free-text descriptions, without the use of structured vocabularies, making automated inference impossible.

In addition, there is often a lack of intuition in interpreting even simple operators, such as union, intersection, and difference, which do not have obvious meanings when applied to, for instance, integration between genes differentially expressed in an experiment and genes in linkage disequilibrium with a phenotype related to the experiment.

Third, the specifics behind each experiment are crucial. For instance, there is a subtle but important distinction between gene expression measurements where measurements represent relative expression levels versus absolute expression levels. Yet as shown above, investigators are increasingly asking how these types of data can be intersected. The implications for a gene in the intersection between two of these data sets is non-trivial, and depends on the reference sample being used. [43-45] Blind intersections of data sets may lead to over-interpretation or misinterpretation.

Modeling experiments with nearly-comprehensive modalities and contexts

The model I propose is to consider every nearly-comprehensive experiment or data set as three orthogonal components: *context*, *catalog*, and *content*. The *context* of an experiment represents the steps of molecular biology experimentation performed (e.g. the decrease of gene expression at the RNA level via RNAi) and the source of the samples. This information might include the species used (e.g. *Caenorhabditis elegans*), the organ in which samples were taken (e.g. the intestine), the cell type, disease, and other characteristics of the source of the measurements.

The *catalog* represents the biological operations performed to gather data (e.g. labeling RNA transcripts, hybridizing, then scanning to measure absolute gene expression), and the list of elements for which a measurement may be obtained. The *content* represents the set of actual data associated with the given measured elements and given context. A graphical representation of this model with two example data sets modeled this way is shown in Figure 1.

Nearly-comprehensive data sets rarely contain enough information on their potential relations to other data sets. Instead, joining two data sets often requires external information. Sometimes this information is available in a structured form, such as tables relating protein identifiers with gene identifiers. Occasionally, this information is not available without in depth review and interpretation of the biological literature, such as relating two cell types with similar properties.

Drawing inferences from two or more nearly-comprehensive data sets is dependent on the types of data. One of the simplest inferential operations is intersection, and I will consider how to implement intersection as a model for working out the other operations.

Phenotypic, gene, and data intersection

Given the context/catalog/content framework outlined above, I find there are three separate aspects to the intersection of nearly-comprehensive data sets. The first is intersection of experimental context. Intuitively, though it depends on the question being asked, two nearly-comprehensive experiments can be intersected if they share “biological relatedness.” The second aspect of intersection is that of the measured elements (for example, two genes across species may be held as equivalent through homology), and the third aspect is intersection of the data elements themselves (for example, relating a ratio of relative expression measurement to an absolute expression measurement). I will consider the last two of these aspects first, since they have been worked out the most.

To perform an intersection, the catalog of measured items, such as gene transcripts, proteins, enzymes, and metabolites, must also be identified and unified across modalities and contexts. An example of unification across modalities is when genes measured by expression level need to be unified with genes positioned near genetic markers. An example of unification by context is when gene expression measurements need to be unified between *C. elegans* and *M. musculus*. Though by no means solved, unification across some of these contexts has been addressed, especially the use of gene homolog and ortholog tables. [29,46] Unification across modalities is possible through the use of translation tables that map a set of measured gene identifiers (e.g. those provided by a microarray manufacturer) with more global gene identifiers (e.g. LocusLink or RefSeq identifiers). As will be illustrated in chapter 5 of this dissertation, incorrectly determined mappings can lead to false positives, including genes mapped within and across species that are not appropriate.

After mappings for the measurement items have been identified, the actual content, or measurements themselves must be unified. Whereas expression measurements may appear to be directly comparable, reasoning between a LOD score and expression measurements to infer properties of genes or phenotypes is more challenging, yet increasing values for both measurements typically signify increasing significance.

The analysis of most large-scale modalities typically reports not only a series of measurements, but also which of those measurements are significant. In these modalities of molecular biology, significance itself can represent a statistical finding (e.g. a finding probably caused by something other than random chance), a numerical finding (e.g. a large difference), or a biological finding (e.g. a finding only indirectly implied by the data). Resolving significant findings between measurement modalities requires careful consideration of the semantics.

The hardest of the three aspects of intersection to implement is contextual. Specifically, unifying between the phenotypes studied in two nearly-comprehensive experiments is an automation challenge. Though any nearly-comprehensive data set has a number of contextual properties that can be represented using structured vocabularies (such as the species used, or tissue type), many are not so easily represented (such as the phenotype, the experiment performed, or the abnormality seen in the patient or organism). Due to the efforts of the Microarray Gene Expression Data Society, a data model has been established to represent the protocols and methods used for microarray hybridization and scanning, and similar models have been proposed for proteomics and protein-protein interaction data. [35,37,47] However, this is not the case for representing every possible aspect of the sample, nor is it the case for many of the other nearly-comprehensive modalities.

Even today, a researcher studying any particular biological process with a nearly-comprehensive modality would ideally want to be able to gather as many relevant data sets as possible. Unfortunately, the degree of relevance of a data set is currently more likely to be assessed using unstructured contextual properties or free-text descriptions. It is precisely this aspect of intersection that will be addressed in this dissertation.

Different inferential operations will have different sets of requirements in terms of how the context, catalog and content have to be unified. As shown in Figure 2, the experimental contexts, the catalogs of measured elements and the content of data elements will each need to be individually related to fully intersect two data sets. Relating these may require the use of external knowledge. Even if the experimental context and the measured elements are unified, one still has to reconcile the data elements. The example shown in Figure 3 illustrates that directly interchanging relative and absolute expression measurements is conceivable, but not trivial and might involve modeling of the reference sample or considering the ratio of two absolute expression measurements. [43]

However, even if only one of the three aspects of integration can be unified, one can still use tools such as two-dimensional hierarchical clustering [48] or relevance networks [49] to visualize relations between the measurements and generate novel hypotheses. This single axis relation is demonstrated in Figure 4.

Example of manual integration

I recently used a manual approach using the integrome to undertake a successful study of the process of adipogenesis, or how fat cells develop the ability to respond to insulin stimulation and store fatty acids. The process of fat cell development is crucial in the development of diabetes in many individuals. I am listing this use-case here (1) to demonstrate the value of studying the integrome, (2) to serve as an example of the steps required to manually integrate across different measurements modalities, (3) and to give an example of the overall biological domain for the problem addressed in this dissertation.

Type 2 diabetes affects approximately 15 million people in the United States. More importantly, a U.S. child born in 2000 now has a lifetime risk of 33% for development of type 2 diabetes. This is most likely due to the current epidemic of obesity and inactivity in the children in the United States and worldwide.

Diabetes as a result of obesity involves increased hepatic glucose output and decreased glucose uptake into peripheral tissues; though muscle is the most important site for uptake quantitatively, adipocytes are also crucially involved, suggested by the correlation between adiposity and insulin resistance. At a cellular level, obesity results from both increased size of white adipose tissue and adipocytes, from increased lipid accumulation and differentiation of preadipocytes into adipocytes.[50] Preadipocytes do continue to undergo differentiation throughout life under the appropriate stimulation.[51] In obesity, adipocytes secrete increased levels of hormones, such as leptin, TNFα, resistin, and Interleukin 6, and decreased levels of adiponectin and adipsin. These hormones are involved in energy balance, metabolism, and feedback response to the brain, and may be involved in insulin resistance.[52] Thus, determining the molecular process of adipogenesis is crucial for the development of diagnostics and therapeutics for both obesity and type 2 diabetes mellitus.

The process of adipogenesis is defined by the gain of several abilities by the preadipocyte: insulin responsive glucose uptake, insulin responsive lipoprotein lipase, and hormonal secretion. Several transcription factors involved in differentiation from preadipocyte to adipocyte have been described including PPARγ [53,54], C/EBPβ, C/EBPδ [55], SREBP1 [50], and E2F/DP. [56] All of these transcription factors play a predefined role *during* adipogenesis. Extracellular promoters and inhibitors of adipogenesis are also known, such as insulin, IGF1, fatty acids, and many others. Negative regulators include *wnt* ligands, TGFβ, and others.[57] However, few intracellular factors are known that impact adipogenesis.

I have been studying the genes that alter the process of fat storage in three unique models of disturbed adipogenesis: (1) altered adipogenesis in Hutchinson-Gilford Progeria Syndrome (HGPS), (2) defective adipogenesis in mouse models missing key insulin signaling components, and (3) altered fat storage from comprehensively knocking out genes in the worm *C. elegans*. Adipogenesis in humans may actually be the sum of many parallel developmental processes, including the acquisition of ability to store fatty acids as well as the ability for insulin stimulated glucose uptake. The process of fat storage in each of these models is not identical. My hypothesis is that there is a core set of genes that impact the ability to store fatty acids, and the only way to find these genes in an efficient manner is by integrating genome-scale lists of differentially regulated genes from the three models.

Insulin receptor signaling dependence in mouse models of adipogenesis

In collaboration with the laboratory of C. Ronald Kahn at the Joslin Diabetes Center, we have already obtained microarray expression measurements from brown preadipocyte cell lines derived from mice lacking IRS-1, -2, -3 and -4, all immediate downstream targets of the insulin receptor. Affymetrix U74v2A microarrays were used to measure approximately 12,000 expressed transcripts. Quadruplicate measurements were made on 4 cell lines from each of the 4 knockouts and 3 separate control cell lines for biological reproducibility, totaling 28 microarrays. When considering the phenotype of adipogenic potential as a continuum, from wildtype, IRS-4, IRS-2, IRS-3, to IRS-1, we found 20 genes upregulated and 61 genes down regulated across the knockouts corresponding to the degree of difficulty of adipogenesis (an example is shown in Figure 5).

Differential gene expression in human fibroblasts from patients with HGPS

Another unexpected model of adipogenesis may be found in Hutchinson-Gilford Progeria syndrome (HGPS), a rare condition affecting 1 in 8 million births now known to be associated with a mutation within codon 608 of the Lamin A gene on chromosome 1.[58,59] In addition to a number of clinical manifestations, children with HGPS progressively develop failure to thrive and loss of subcutaneous fat. At autopsy, the subcutaneous adipose tissue is atrophic. In addition, previous reports have indicated insulin resistance in these patients,[60] including decreased insulin receptors in HGPS lymphoblasts [61] as well as impaired insulin binding and insulin-insensitive hexose transport.[62] Finally, other defects in Lamin A are associated with familial lipodystrophies.

To further explore the molecular pathogenesis of HGPS, in collaboration with the Progeria Research Foundation, we obtained and analyzed the gene expression patterns of three HGPS fibroblast cell lines heterozygous for the codon 608 mutation to three normal control lines. Out of 33,000 measured genes, 366 (1.1%) showed a 2-fold significant difference, with 198 up- and 168 down-regulated genes. The products of the differentially regulated genes participate in a large number of different biological processes, and many are known to function in tissues that are severely affected in HGPS.

Differential gene expression in human fibroblasts from patients with HGPS

The laboratory of Gary Ruvkun has published a list of genes involved in *C. elegans* adipogenesis, determined by a global knock-out strategy using RNAi. [63] Out of over 16 thousand genes knocked out, they found 112 genes that when knocked out increase the fat content of the worm, and 305 genes when knocked out decrease fat content.

Manual approach to integration

After considering the caveats, we manually applied an intersection to the three lists of genes. I am describing the manual approach here to illustrate that while such an intersection is not hard to conceptualize, it is difficult to operationalize, and remains beyond the reach of many biomedical researchers. Thus, this difficulty makes a good case for automation.

1. The mouse data contains gene expression measurements made in brown preadipocytes in four knockout mice and their littermate controls, in basal conditions as well as in time-series response to insulin and IGF-1. I determined that for the intersection (1) brown preadipocytes may be an acceptable alternative to white preadipocytes, and (2) the data regarding insulin and IGF-1 stimulation was unnecessary.

2. The human data contains gene expression measurements made in fibroblasts from three patients with Progeria validated to have the characterized mutation, one patient with Progeria not validated to have the characterized mutation, and

three age-matched patients. I determined that for the intersection, I would only use data from patients verified to have the mutation.

3. The worm data contains lists of genes that cause increased or decreased fat storage when transiently knocked out in wildtype worms or in worms where either *daf-2*, *tph-1*, or *tub-1* was already deleted. The increase and decrease was characterized on an eight point scale. I determined that for the intersection, I would flatten the five point scale into three points (increase, decrease, no change) and I would use the list of genes causing changes in wildtype.

4. For mouse data: Affymetrix accession numbers from the U74Av2 array were translated into GenBank accessions for their target sequences. GenBank accessions were translated into current UniGene clusters. UniGene identifiers were translated into LocusLink identifiers.

5. For human data: Affymetrix accession numbers from the U133A and U133B arrays were translated into GenBank accessions for their target sequences. GenBank accessions were translated into current UniGene clusters. UniGene identifiers were translated into LocusLink identifiers.

6. For worm data: GenePair accession numbers references in Ashrafi, et al., were translated into WormBase gene accession numbers. WormBase accession numbers were translated into LocusLink symbols. LocusLink symbols were translated into LocusLink identifiers. LocusLink identifiers were translated into UniGene clusters.

7. UniGene cluster identifiers from two and three datasets were joined using Homologene. This resulted in few genes in the intersection.

8. Gene families were determined arbitrarily, based on gene symbol, for each significant gene in the mouse, human and worm data. Homologous gene families were identified in each of the three lists.

After these eight complex steps, 19 genes or gene families were found in the intersection. One gene family found in the intersection is the *wnt* family of ligands, specifically *wnt6* and *wnt10a*. *Wnt5a* and *wnt10b* are already known to inhibit adipogenesis. [64]

Importantly, several tripartite-motif (TRIM) containing proteins were found in the intersection, which represent a potentially novel set of genes in fat storage. Mouse *Trim30* is 3.4 fold down-regulated in IRS-1 knockout preadipocytes as compared to wildtype. The human homolog, *Trim5*, is 2.3 fold up-regulated in Progeria. In biological validation in a separate mouse model of adipogenesis, we found *Trim30* decreases 11-fold when measured by RT-PCR during 3T3-L1 adipogenesis (shown in Figure 6). Though this certainly does not prove that *Trim30* is necessary and sufficient for adipogenesis, it does suggest that *Trim30* is involved in the process of adipogenesis.

This case of manual intersection between three biologically relevant data sets illustrates the value of studying the integrome, and serves as an example of the onerous steps required to manually integrate across different measurement modalities.

Skills needed for integrative biology

Intersection between modalities itself is useful, but the efficient use of this method requires a practitioner trained in a specific set of skills. Automated assistance with reasoning across nearly-comprehensive modalities is needed, and I am addressing this need with this dissertation. At the current time, however, intersection of data across modalities is mostly performed manually. Successful manual integration requires biomedical knowledge, programming infrastructure skills and computer science skills:

Biomedical knowledge

1. To find samples related to a disease in order to integrate their measurements, one needs to understand the known or hypothesized causes of diseases as well as known implications and complications of diseases. This is context-specific knowledge. For example, unless a researcher studying type 2 diabetes understands that onset of that disease may be preceded by insulin resistance, he or she may miss genomic data collected on relevant samples.

2. A researcher needs to know of the biological implications of a true-positive finding in any measurement modality. For example, a significant difference in the measurement of a gene's expression level by microarray could be an indication of increased transcription, but it could also indicate a decrease in transcript degradation, a change in the number and type of cells in the sample being studied, or even a change in alternative splicing in the transcript, depending on the microarray probes. If no change is seen in the level of the protein coded by this gene, some of these other explanations may become more likely.

3. Inferences between modalities involving genes will be more fruitful if a causal chain of reasoning can be established from the gene-level finding and the context of the samples. This is not just knowledge of biological pathways; this is knowledge as to what might happen if a pathway or its components are altered or disrupted. For example, though it may be significant if a gene appears in the intersection of a microarray data set and a genetic association study, that gene will be more significant if it can "explain" the disease studied in the microarray and the genetic trait studied. Of course, novel explanations will need new biological validation. Though limited in scope and detail, tools to automatically link genes and proteins to pathways can suggest biological processes that may be involved. [65,66]

4. To find relevant samples, a researcher would need to know how samples and data sets may be indexed or stored. This may involve knowledge of clinical vocabularies, such as identifiers from the International Classification of Diseases. [67]

Programming infrastructure skills

5. The translation of modality-specific identifiers to common identifiers is important in integrating data sets. Web-based tools are available that translate some modality-specific identifiers to global identifiers, such as UNCHIP (www.unchip.org), NetAffx, [68] and Resourcerer. [69]

6. Finding homologs of genes is also important, especially for integrating data between species. Data sets describing similar processes across species can be

a rich source for the application of integrative biology, and specifically comparative genomics. HomoloGene, developed and maintained by the NCBI, is a database relating gene orthologs and homologs across 23 eukaryotic species. [29] HomoloGene relations are designated as either curated or calculated by nucleotide sequence homology. The TIGR Eukaryotic Gene Orthologs database contains ortholog and paralog relations for genes (represented either by tentative consensus sequences from gene or EST sequencing) for 61 eukaryotic species. [46] Familiarity with these services can allow data from additional species to be found and integrated with existing data.

7. Fewer tools are available for determining gene families or functional families. The results of an intersection may not be obvious. Two transcription factors with similar protein domain structure may be involved in similar processes in two species, yet may have different names and symbols. Paralogs, or genes from the same super-family formed by gene duplication, can be found using HomoloGene. [29] Linking genes to their parent families or known functional groups (such as GeneOntology [70]) before performing intersections may yield a richer result.

Computer science skills

8. Integration of data sets requires some software application or platform in which to perform operations. Unfortunately, there is currently no commonly available bioinformatics tool to assist with intersections across species, never mind modalities. However, sets of genes and operations, such as intersection, union, and difference, can be modeled using relational databases. Lists of significant

findings can be uploaded into a relational database system (such as Microsoft Access, MySQL, Oracle, and others), along with tables that map between identifiers. Queries can then be written or graphically created that join the lists of findings across the mappings. Facility with the use of a relational database system is crucial for integrative biology, and knowledge of query languages, such as SQL is optimal.

9. The formal representation of biological knowledge requires an understanding of the methods used for several decades by the artificial intelligence community. Several formal knowledge-bases have been created for narrowly-scoped domains in molecular biology, such as EcoCyc and MetaCyc for metabolic and genomic data in *E. coli* and other bacterial species. [71,72] Though many biologists are familiar with the graphical representation of pathways, and even how these can be drawn using programs like GenMAPP, representing new pathways in a computational form, like EcoCyc, requires familiarity with predicate logic and programming languages for these predicates, such as Prolog. [73]

How the rest of this dissertation is organized

Integrative biology is more than a technical issue. Proper intersection is more than just lining up the correct columns in two data sets; it requires an understanding of the biological context of both data sets, so that the intersection operation itself is justifiable. Currently, that understanding is not an automated process. The biology cannot be ignored.

The integration of large-scale data sets across measurement modalities has already demonstrated itself to be a synergistic process to create new knowledge and testable hypotheses. Operations such as union, intersection, and difference can quickly expand and focus findings around the most significant. In addition, with the large number of publicly available data sets, from modalities including gene expression, protein identification, and phenotypic and clinical measurements, one can get these benefits without the increased cost of additional measurements.

As others increase the number and resolution of measurements, as the discipline of systems biology suggests, exploration of the integrome represents approaching a problem from multiple vantage points and focusing on the common or core question. To do so successfully will require even more multidisciplinary expertise that is grounded in deep understanding of the biology while embracing comprehensive quantitative methods. Fundamental biological questions, such as aging and development, can be asked at multiple levels, from the molecular to the cellular to the organism, so being able to capture how those level integrate is an important step towards an operable genomic medicine.

This dissertation will explore how integrative biology can be used to explore diseases in genomic medicine. Specifically, I will use this dissertation to model the context, catalog and content of genomic data within the largest publicly available database, and will use this model to address the specific question of how large-scale genetic and genomic data studying the same biological process can be integrated in an automated manner to validate each other. More importantly, the dissertation will show how together they can lead to causal mechanistic explanations of disease.

Chapter 2 starts with a review of previous work related to this dissertation, including studies representative of and demonstrating the benefits of integrative biology, previous work in integrating clinical, genomic and pharmacological data, and other work in the representation of biological pathways and taxonomies. The previous work also includes previous work in resolving genes from identifiers, and integrating genomic and genetic findings.

Chapter 3 will directly address the modeling of the experimental context of genomic samples and experiments with a structured vocabulary using automated methods. First, I will show how I have successfully created an automated system to model the context of genomic samples and experiments from annotations in the largest publicly available gene expression repository. Moreover, I will show how the largest biomedical vocabulary can be used to represent the majority of these contextual annotations. In chapter 4, I will then show how to determine the cell type and disease studied by experimenters as an application of this contextual modeling.

Chapter 5 will address how I can model catalogs of measured genes from the largest public repository of expression measurement data. With context and catalog measured, I will then show how to model the content in chapter 6. Specifically, I will show how I have created an automated system to extract and determine the significant genes from every possible comparison of groups of microarrays. I will then show how modeling the content itself provides a valuable tool in studying the effect of experimental variables on gene expression across dozens of experiments.

To test the validity of the model build to represent the context, catalog, and content from the largest public repository of expression measurement data, I will then show in chapter 7 how I have used an automated method to find microarray data sets related to a particular disease and perform an intersection of these data sets to successfully find genes relevant in the disease process.

Finally in chapter 8, I address how the context / catalog / content model can address the problem of integrating genetic and genomic data that study the same biological process. I will demonstrate a system that contains a model knowledge base of how changes in expression can lead to a particular disease, and a model knowledge base of how quantitative trait loci and gene expression measurements can be related to each other. I will then show how the system can be given input genetic data and queried for genes that match the genetic data, have an ortholog in expression data related to the trait, and can explain the trait through known biology pathways and pathophysiology.

The dissertation concludes with the known limitations of intersecting large-scale molecular biological data in chapter 9, a summary and future directions in chapter 10, and references.

2. Previous work

This dissertation will explore how integrative biology can be used to explore diseases in genomic medicine. Specifically, I will use this dissertation to model the context, catalog and content of genomic data within the largest publicly available database, and will use this model to address the specific question of how large-scale genetic and genomic data studying the same biological process can be integrated in an automated manner to validate each other. More importantly, the dissertation will show how together they can lead to causal mechanistic explanations of disease, with the appropriate application of prior biological knowledge.

Thus, the steps required to accomplish this include (1) integration across experimental modalities, (2) representing and using biological pathway knowledge, (3) mapping identifiers to genes. This chapter provides a review of previous work in each of these steps.

Specifically, I will start with previous publications that demonstrate the advantages of integrating data sets within a single measurement modality. I will review several published examples of integration across two or more large-scale measurement or experimental modalities. I will cover previous attempts to model to process of integration itself.

I will then review previous attempts at qualitative and quantitative representations of metabolic pathways, ending with tools that help in visualizing these pathways. The chapter continues with coverage of biomedical taxonomies and ontologies, though the review of this subject continues in chapters 3 and 5. After covering taxonomies and standardized formats for representing genomic data, I will describe my and others previous work in mapping identifiers to genes.

Using these methods, we have previously built a system integrating gene expression data across 11 multi-center programs in genomics, linking expression data to universal gene identifiers regardless of platform and file format. I will describe this work in this chapter.

I will end this chapter with a short discussion how the integration of modalities still cannot fully explain a simple step of a metabolic pathway without biological knowledge.

Demonstrated benefits of integrating data sets

There have been several demonstrated benefits resulting from integrating data sets of a single measurement modality. First, a second set of microarray data measured under similar experimental conditions as the first can serve as validation of important findings. For example, Michael Primig, et al., compared their results on yeast meiosis to earlier published findings using Venn diagrams that showed their gene list was a superset of previous lists. [74,75]

Alternatively, a second data set may be directly joined with the first to increase the number of samples available and improve the power and the ability to draw statistical conclusions. Zambon, et al., compared their list of genes in human skeletal muscle regulated in a diurnal manner with lists of circadian-regulated genes in mouse heart, liver and suprachiasmatic nuclei made by others. [76,77] This resulted in a set of candidate genes that were validated as circadian in mouse skeletal muscle. [14]

A second data set may be used to probe another aspect of a common process, or to study aspects of a process that are unique to a single experimental context. For example, David Fruman, et al., measured the gene expression response in B cells with partial loss of function in phosphoinositide 3-kinase (PI3K) and Bruton's tyrosine kinase (Btk), and used the significant overlap in gene lists as evidence that PI3K acts through Btk. [78]

A second data set may also be used to filter out biological noise, such as eliminating genes known to be differentially expressed across the circadian cycle from genes differentially expressed in samples acquired from organisms. Whitney, et al., showed that many genes involved in proteins synthesis were regulated in a diurnal manner. [79] Future microarray studies on other processes reflected in whole blood could take advantage of this result by filtering out this circadian cluster, thereby improving the specificity of their results.

Examples of integrating experiments across multiple modalities

There are several published examples of nearly-comprehensive measurements under many or nearly-comprehensive conditions. In a now classic study, Michael Eisen, et al., demonstrated hierarchical clustering of over 2,000 yeast gene transcripts measured under eight time-series conditions. [80] In later work, Timothy Hughes and others measured near-comprehensive gene expression differences in yeast under a variety of conditions, including 200 gene deletion strains. [81] Even though both experiments measured roughly the same number of transcripts per microarray, the later work differs from the former work in an important way. Hughes's work measured gene expression in a set of systematically constructed contexts: that of serial gene deletion of over 4% of the yeast genome. In contrast, Eisen's work measured gene expression response to an arbitrary set of contexts.

There are fewer examples of the intersection of two nearly-comprehensive measurements. Uwe Scherf, Douglas Ross, and others joined a data set of baseline RNA expression levels from the NCI60, a set of 60 human cancer cell lines used by the National Cancer Institute Developmental Therapeutics Program to screen anti-cancer agents since 1989, [82] to a data set of drug susceptibility in the same cell lines. They showed how clustering cell lines by genes can differ from clustering based on drug susceptibility, and proposed mechanisms how expression differences of specific genes can impact susceptibility to specific agents or classes of drugs. [48,83]

We developed a method termed *relevance networks* and applied it to a similar pharmacogenomic data set, generating hypotheses of putative functional relationships between pairs of genes and pharmaceuticals. [49] We joined baseline RNA expression levels of 6,701 genes measured from the NCI60 to a database of measures of cancer susceptibility to 4,991 anti-cancer agents. We studied this data to understand how the baseline RNA expression levels in the cell lines correlated with the inhibition of growth of these same cell lines to thousands of anti-cancer agents.

The relevance networks formed from associations with correlation coefficient beyond ±0.80 are shown in Figure 7. At this threshold, only one network contains an association between a gene expression and a measure of anti-cancer agent susceptibility. The association suggests that increased expression of lymphocyte cytosolic protein-1

(*LCP1*) is associated with increased susceptibility to the anti-cancer agent NSC 624044, a thiazolidine carboxylic acid derivative. Though a specific role for *LCP1* in tumorogenicity had been postulated [84] and though other thiazolidine carboxylic acid derivatives are known to inhibit tumor cell growth, [85] there is no known relationship between this specific anti-cancer agent and gene in the biomedical literature. This relation remained significant after 100 permutations of the data. [49]

Vamsi Mootha, et al., integrated four publicly available expression data sets with linkage data and proteins identified from mitochondria to ascertain the gene and mutation responsible for Leigh syndrome, French-Canadian type. [86] Monica Stoll, et al., integrated 125 phenotypes with linkage data from rats to determine candidate genes potentially involved in cardiovascular function. [87] Boris Rolinski, et al., are using mass spectrometry to determine differential levels of several amino acids and acylcarnitines in randomly mutagenized mice (using ENU). [88] Petra Ross-Macdonald, et al., disrupted nearly 2,000 genes in the yeast genome randomly inserting transposons and studied their effects across a panel of twenty phenotypic tests. [89]

In an integration study incorporating genetic and expression data, Eric Schadt and others started with gene expression differences in liver between two inbred strains of mice, then used the most significantly different genes as traits which were then mapped as quantitative trait loci. They first highlighted genes with known polymorphisms that affected their transcript levels, then showed loci associated with fat pad mass that would otherwise have been insignificant had the expression data not been considered. [90,91]

Albertha Walhout, et al., combined three modalities of data measured in the *C. elegans* germ-line: protein-protein interactions, phenotypes measured after RNAi against each gene in the genome, and RNA expression measurements. In their study, they determined that interacting proteins tend to be co-expressed and knocking out many of the interacting partners often led to similar phenotypes. [25]

Others have written about the potential of integrating the results of cross-modality experiments. Marc Vidal has noted that integration of multiple functional maps can lead to novel informatics algorithms, and he comprehensively covers many of the genome-scale modalities available. [24,92] Vidal uses the analogy of binding maps into an atlas. However, he does not consider near-comprehensive measurements and contexts outside of the genome. To continue the analogy, how does one connect a subway map with a street map? One can only perform this integration with the prior knowledge of the location of each train station. I believe the potential is much greater than just integrating gene-centric maps; there is a greater potential in linking modalities by phenotype, but this requires a level of sophistication and biological knowledge beyond simple database joins.

Again, these examples are given to illustrate that exploration into integration of multiple near-comprehensive modalities has started, but has not yet been studied in any formal way. None of the above referenced publications have comprehensively studied the large class of important inferential processes that can be used in integrative biology, or given any generalized systematic method for interpreting the results of integration. Yet as shown above, integrome exploration appears to be the implicit goal of several leading researchers. In this dissertation, I am proposing a framework for modeling nearly-comprehensive experiments and a first step towards the rational intersection of these data sets.

Quantitative, qualitative and visual representations of biological pathways

Quantitative representations

Many simulation systems are available for quantitatively representing particular series of biological reactions, including BioSpice [93], Cellerator [94], and Virtual Cell [95]. E-Cell is a platform allowing the creation of models consisting of variables, processes, and systems. [96] E-cell can model discrete or continuous processes, and though it can use ordinary differential equations to model changes in variables, any arbitrary process can be written in the language Python and incorporated. E-Cell is being used to model cellular components such as mitochondria, cells such as neurons, and diseases including diabetes.

Schoeber, et al., used custom software in Matlab to model pathways from the epidermal growth factor (EGF) receptor to gene expression. Using a system of ordinary differential equations with 94 variables, they showed how *c-fos* expression matched predicted levels as EGF concentrations were varied.

Qualitative representations

The BioCyc knowledge library contains the two literature-derived component libraries of EcoCyc and MetaCyc, and currently contains 13 computationally-derived pathway libraries, all in the domain of bacterial organisms and their metabolic, biosynthetic, degradation, and energy metabolism pathways. EcoCyc was one of the first efforts to integrate metabolic and genomic data and covers *E. coli.* [71] MetaCyc is a metabolic-pathway database that describes the action of over one thousand enzymes, with links to various species implementing the pathways but without links to genes. [72] The underlying representation for the BioCyc libraries is a Pathway/Genome Database (PDGB) in which is described a bacterial genome, gene structure and sequence on its chromosome(s) including transcription factors and binding sites, the protein coded for each gene and the reaction in which the protein participates. Each reaction has substrates, and collections of reactions are called pathways. Transcription factors are given special object status; aside from this, reactions involving transcription or translation are not covered. Though SRI is working on a HumanCyc, it is not yet clear what biological scope will be covered by this new database, or whether it will expand beyond those metabolic pathways common to all the currently maintained organisms.

Karp and Riley noted that the *knowledge acquisition problem* was one of the most serious challenges they faced in the construction of EcoCyc. [97] The problem in constructing a knowledge assembly for molecular biology is particularly difficult because of the rich set of objects and classes needed to qualitatively describe even basic cellular operations, and rapid access and resolution of these objects is difficult while curators attempt to create factual assertions. Karp and others have attempted to address the knowledge acquisition problem in three ways: (1) creating domain-specific graphical tools for entry of biological information, (2) creating a domain-specific browsing tool for editing of assertions, and (3) training curators and those who perform knowledge entry. [97]

KEGG, a database within GenomeNet, is similar in that it also contains information on enzyme-substrate reactions, [98] initially derived from the Boehringer Mannheim and

Roche Applied Science Biochemical Pathways wall chart. [99] Though KEGG contains manually created figures representing pathways of reactions, its main representation for genes, proteins, pathways, and transcripts is as nodes in a graph. Edges contain relations between these objects of the same class (there are few, if any, edges between different classes of objects).

Hofestädt and Thelen described the use of Petrinets to represent metabolites and enzymes, [100] building on the work of Reddy, et al. [101] They encountered difficulties in representing gene regulatory processes, due to the lack of temporal representation.

The National Center for Genome Resources PathDB contains information about protein complexes, metabolic, signaling, regulatory and other cellular pathways, phenotype categories, kinetic parameters, protein-molecule and protein-protein interactions, and genetic interactions, for the domain of Arabidopsis and yeast. The repository holds data from 1,250 articles entered over 3 years in a graph-theoretic data structure. It is not clear whether the system supports reasoning beyond user driven queries.

WIT is a hierarchical view of similar metabolic pathways connected to the sequenced genomes of approximately 40 organisms. [102] The University of Minnesota Biocatalysis/Biodegradation Database contains metabolic reactions in microbes especially focused around the conversion and breakdown of environmental pollutants, including metals, metalloids, and metal chelators. [103]

All of these databases are biased towards pathways involving the processing of classical biochemical substrates present across single-cellular and complex organisms.

As an example, none of these databases models even the most superficial understanding of the action of insulin on a target tissue.

There have been several applications developed that take advantage of established taxonomies and metabolic knowledge bases. Badea demonstrated how inductive logic programming could be applied to a microarray data set and prior information from GeneOntology to learn functional differences between subtypes of adenocarcinoma. [104] Hanisch, et al., demonstrated an application showing correlation of gene measurement corresponds to distance of the two genes in a metabolic knowledge base. [105] Zien, et al., suggested interpreting gene expression data in the context of graph-theoretic pathway scores derived from a metabolic knowledge base. [106]

Visual representations

A number of graphical visualizations of biochemical networks have been made, and a number of software packages are currently available. Kohn created a diagrammatic representation for the mammalian cell cycle and DNA repair pathways [107], serving as a visual indicator of the molecular interactions in this domain, similar to the Boehringer Mannheim Biochemical Pathways wall chart.

GenMAPP allows the graphical visualization of gene and proteins involved in a biological process, and can color-code the genes based on experimentally derived measurement data. [66] However, the underlying data structure stores only the visualization information and does not contain a computable network. KnowledgeEditor similarly uses biological pathway information and maps genomic measurements onto these pathways. [108]

Krishnamurthy, et al., divide biological pathways into three categories: (1) metabolic and biochemical, (2) transcription, regulation and protein synthesis, and (3) signal transduction, and use a physical representation to divide these pathways into four layers: (1) structures of molecules, (2) functional use of molecules in processes, (3) pathways of processes, and (4) complex networks of related pathways. They provide a query mechanism and graphical viewer for the networks, with content similar to above mentioned metabolic pathways. [109]

BioCarta provides additional variety of pathways with graphics available through the Internet; this representation is also not computable. BioJake allows for the entering of biological reaction information.[110] The Alliance for Cellular Signaling has provided a web-site that provides a protein-specific integration of sequence, domain, and molecular data, and has provided graphical representations for a handful of signaling pathways. [111] The Expasy server similarly provides a web-accessible index for the Roche Applied Science Biochemical Pathways wall chart, but like these others, the only representation is graphical. [112] GSCope uses a hyperbolic projection to view arbitrarily defined networks. [113]

It is important to note the limitations in these graphical methods. The graphical representation of a biological process is not computable. Typically, the appropriate graphical view is chosen for the results of a gene expression analysis based on the number of genes in common between the graphics and the list of differentially expressed genes (even if the graphics are representing proteins, not genes). A chain of causal reasoning explaining how a pathway might explain a set of differentially expressed genes is never produced.

Summary of representation in biomedical systems

A range of abstraction has been used in modeling biomedical pathways and systems. [114] At one end of this range are pure visualization methods that can depict changes in component state, but cannot be used to computationally make predictions. Moving from this extreme are the statistical systems, such as relevance networks and Bayesian networks, that can model the co-occurrence or correlation of measurements from a genome-scale catalog of components, but with little commitment, context or top-down experimental design. At the other end of the range are systems of differential equations that provide a high resolution predictive ability, but often at the expense of comprehensiveness or generalizability.

For this dissertation, I am creating a knowledge base that is qualitative in nature, with the ability to relate to every known gene and biomedical condition. I will not be able to make detailed predictions as to the exact measurement levels of components, but because of this, I will be able to incorporate physiological states simply by reference.

Biological taxonomies, ontologies, and formats

GeneOntology is a hierarchical taxonomy and vocabulary for the molecular functions, biological processes, and cellular components in which proteins participate, and includes associations to commonly used external identifiers for genes and proteins. [70] Other vocabularies used for gene and protein annotation include InterDom [115], InterPro [116], and PRINTS [117] for protein domains, and LocusLink [29] and SOURCE [30] for other types of annotations. Other gene identifiers include the stable NCBI LocusLink identifier [31] and the unstable NCBI UniGene identifier [118].These and other taxonomies are described in greater detail in chapters 3 and 5.

The Systems Biology Markup Language (SBML) is a free XML-based format allowing the exchange of computational biological models. SBML can be used to store quantitative mathematical models in an open format, but is too specified for qualitative models. There are many other existing file and storage formats for genome-scale data. These include Genome Annotation Markup Elements XML, Minimum Information about a Microarray Experiment (MIAME) [32], the NCBI Haplotype Set XML, the PharmGKB schema for genotypes [34], the SNPPR XML format for genotypes , the NCBI Seqset XML format for GenBank data, the Distributed Sequence Annotation System XML scheme, Tagged Image File Format for CCD camera images, DAT and PRE format files for GENEHUNTER, netCDF and ANDI formats for mass spectrometry, and other CORBA Life Science Research formats.

Mapping from identifiers to genes

A major problem in interpreting results from microarray analyses is one of nomenclature. Each microarray output file lists a probe-set accession number, an expression quantity, and degree of confidence of that measurement. A single microarray may reference genes described in a variety of data-bases, including expressed sequence tags, GenBank identifiers, and UniGene cluster identifiers. This is currently a problem because different microarray model years from the same manufacturer may use a different set of accession numbers.

Before microarray manufacturers were pressured by the research community to make translation tables available, we solved this problem by constructing a publicly-available web-based integration tool called UNCHIP (www.unchip.org). UNCHIP finds the latest information on all stored accession numbers through periodic downloads of and integration with LocusLink, OMIM, UniGene, Golden Path, PROSITE, and GeneOntology. The architecture of UNCHIP is shown in Figure 8. From any Affymetrix probe-set accession, we can gather the latest information from these databases.

More importantly, we can start with an arbitrary hypothesis and work backwards to query the microarray database. This differs from the web-based translation tables (www.netaffx.com) provided by Affymetrix. [68] Several examples of queries are shown here. First, we learned that a significant number of Affymetrix probe sets were linked to more than one LocusLink gene. This was because the GenBank accession provided for the Affymetrix identifiers indicated the probe set was likely designed against a region of a chromosome, instead of an expressed product. Thus, without sequence level information, it would be impossible to distinguish which gene in that region of DNA was actually being measured. These types of probe sets can be found using UNCHIP by the query shown in Figure 9.

If prior data implicating a particular chromosomal region is known, for example from a genetic study, those probes measuring genes in that region can be specifically found using a series of queries shown in Figure 10. Queries can also be written making use of additional prior knowledge, such as GeneOntology classification or protein domain structure.

In summary, UNCHIP was our earliest example demonstrating that we could dynamically map from microarray platform identifiers to global gene identifiers, and that we could query these gene catalogs based on prior biological knowledge, including chromosomal localization and base-pair position.

An integrated gene-centered store of genomic data

The Programs for Genomic Applications (PGA) is a research consortium of 11 multi-center projects funded by the National Heart Lung and Blood Institute with the goal of discovering genes and proteins associated with heart, lung, blood, and sleep health disorders.

The PGA programs have already generated and publicly released vast amounts of pre-publication microarray and sequencing data, including over 1,200 microarrays (measured using Affymetrix oligonucleotide and spotted cDNA microarrays), over 500 genes sequenced for single nucleotide polymorphisms (SNPs), and 18 genes sequenced for mutations. Though the raw data files are publicly available through the Internet, this data has unfortunately remained inaccessible to the majority of researchers unaccustomed to retrieval and advanced analysis of genomic data. Poor secondary use of this data was reflected in low web-site “hit rates”, few requests for data, and few publications citing the primary data sources.

Our hypothesis was that making this data usable by the research community requires more than making raw data files available on the Internet. To address this problem, we created PGAGENE, a web-based gene-specific genomic data search engine. PGAGENE consists of four components: (1) a set of cross-referencing tables between

Affymetrix and LocusLink identifiers, (2) a gene information database holding expression, mutation and polymorphism data, (3) the indexing agent which traverses a list of web-sites and gathers information about the known genes, and (4) the web-site allowing retrieval of all data using gene identifiers, symbols, names, or disease name.

PGAGENE uses LocusLink as its master list of genes. As the indexing agent traverses through PGA web-sites, it encounters tab-delimited files that resemble gene expression data and maps these using its cross-referencing tables. After traversal is complete, PGAGENE rank normalizes all gene expression data by sorting expression values for each array and representing each value by its position scaled to a number between 0 and 1, assuming a uniform distribution. [119]

Using the data contained with PGAGENE, we demonstrated a cluster of diabetes-related genes maintained across species which could not have been found without this data integration. Specifically, we took the expression measurements of a subset of genes measured at least once in the PGA in human, mouse and rat and hierarchically clustered their measurements, as shown in Figure 11. We found that four genes, insulin autoantigen 1 (ICA1), fatty acid binding protein 1 (FABP1), leptin receptor (LEPR), and peroxisome proliferative activated receptor, gamma, coactivator 1 (PPARGC1) were placed in the same cluster, indicating similarity in gene expression measurement. The roles of all four of these genes continue to be studied as important players in type 1 and type 2 diabetes mellitus, but the majority of samples in the PGA were measured to study heart, lung, blood, and sleep disorders.

In summary, this was our earliest example demonstrating that in the right context, gene expression could be compared and integrated across species. More importantly, this work showed that integration of gene expression data across experiments could yield important findings for the study of a particular disease, even when the original experiments did not address that disease.

Pathway analysis and network determination

The current movement in molecular biology has led to a revolution in the tools available to deconstruct processes and study systems at a high resolution and level of comprehensiveness. In *systems biology*, one hopes to be able to model all the components in biology, ascertain networks and pathways linking these components, perturb the systems *in vitro* and *in silico*, and update the networks with new information. [120] Despite all the progress made in high throughput measurements in the past five years, there are still major areas of knowledge integration across these measurement modalities that are undeveloped. If our goal in functional genomics is to be able to ascertain biological regulatory pathways from genome-scale data sets, then it is crucial that our piecewise *a priori* knowledge be put together. To date, there have been several efforts to try to reconstruct pathways from gene expression measurements. [121,122] Given the success of recapitulating the first few steps of the glycolytic pathway from substrate measurements, it would appear plausible to do this. [123] However, without careful consideration of the pathways that one is trying to reconstruct, there is significant risk for a methodological and metaphorical flaw in this analysis.

A simple example is given here, where pathways are nested within several molecular and physiological levels, and the genes that are regulated in one pathway play a role in another pathway. Consider the genes coding for lactic acid dehydrogenase. The gene LDHA is expressed mostly in muscle and is located on chromosome 11p15.4, while LDHB is located on chromosome 12p12.2 and is expressed mostly in heart. LDHA is known to have binding sites for HIF-1, and at least six other transcription factors. LDHB is thought to have a binding site for SP1. There is likely to be post-transcriptional level regulation of these two genes. Because of the differences in promoter regions, it is safe to assume that both LDHA and LDHB participate in their own gene expression regulatory networks.

LDHA and LDHB code for protein subunits. The final protein product contains four subunits. Combinations of the two types of subunits as assembled into the five lactic acid dehydrogenase isozymes, LDH-1 (four LDHB subunits) to LDH-5 (four LDHA subunits). These five isozymes are found in a binomial distribution in mammals. [124] Additionally, the final protein itself is an enzyme assisting in the conversion of lactic acid to pyruvate, called lactic acid dehydrogenase and defined by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology as Enzyme Commission (EC) 1.1.1.27. Lactic acid and pyruvate are substrates in the glycolytic pathway of anaerobic energy production.

Thus, in this example, one finds *at least four layered pathways,* as shown in Figure 12. The operators formed in the enzyme level pathway participates as one enzyme in a major biochemical substrate pathway, that of glycolysis. That operator, however, is assembled stochastically in the assembly pathway from subunit components. Finally, those subunits are produced within their own genetic regulatory pathways involving a multitude of transcription factors.

Even with the improving measurements techniques from systems biology, the complexity of this one step may still be missed. Measurements of the genome give us nucleotide sequence and variants of this sequence. Measurements using microarrays give us comprehensive RNA expression. In the future, measurement of the proteome may give us comprehensive quantitative protein levels. Measurements using mass spectrometry and yeast two-hybrid methods might give us the three protein-protein interactions shown here. Measurements of metabolite levels might give us the relative amounts of lactic acid and pyruvate.

Integration across these measurement modalities not only requires the ability to identify and unify components spanning multiple levels, it also requires the ability to create models that span these levels. For example, it is known that exogenously added pyruvate in INS-1 cells (which model pancreatic beta cells) potentiates glucose-stimulated insulin secretion, an effect that is abolished if LDH-A over-expressed. [125] A model of this single effect necessarily spans at least six measurement modalities, including expression and protein levels, metabolic substrates, signal transduction, protein modifications, endocrinological and physiological systems, and must also take into account the cellular, tissue and disease context, and the experimental design itself of testing a system with glucose and gene over-expression.

Thus, in considering these layers of pathways, it becomes clear that it is not sufficient to simply represent or visualize the glycolytic pathway in terms of measurements from individual genome-scale modalities. Instead, when interpreting genome-scale measurements simultaneously, a knowledge-base that integrates across these pathways is a necessity.

Summary of previous work

In this chapter, I reviewed previous publications that have demonstrated the advantages of integrating data sets within and between measurement and experimental modalities. I covered quantitative, qualitative, and visual representations of pathways and systems. I reviewed several biological taxonomies and standardized formats and addressed the problem of identifying genes from identifiers. I covered a previous example of integrating genomic data, regardless of format and platform, into a single database and web-site. I ended with a discussion of how integration of experimental modalities still cannot fully explain a simple step of a metabolic pathway without additional biological knowledge that spans these modalities.

The system I will be describing in this dissertation attempts to draw upon the advantages of many previous successful systems. In chapter 5, I will describe how I have built from the UNCHIP system to expand the ability to resolve genes from identifiers. In chapter 6, the work on extracting and unifying gene expression data builds upon the PGAGENE system. In chapter 8, I will explain how genetic and genomic data can be used to causally explain a disease, using a model qualitative knowledge-base, similar to those presented in this chapter.

3. Modeling the context of genomic samples and experiments with a structured vocabulary using automated methods

Motivation

A researcher studying any particular biological process with a genome-scale modality would ideally want to be able to gather as many relevant data sets as possible. Unfortunately, determining whether a data set is relevant is particularly challenging. Even for data sets that are stored in standardized formats, such as microarray data, the useful annotations of this data, such as the experimental context of an experiment, are likely to be represented only by unstructured narrative text (hereafter designated free-text). Furthermore, deciding whether a data set is relevant to ones study requires biological knowledge and expertise.

Manually reading descriptions of data sets is not a scalable approach. At the time of this writing, microarray data for nearly a thousand experiments are stored in international repositories. Reanalysis and further discovery from collections of data is going to be dependent on extracting the annotations from these data sets; a manual approach is not scalable for the amount of data already present.

Automated techniques to ascertain biomedical assertions from free-text have been in development. Krauthammer, et al., have been developing GeneWays, a system that uses the GENIES natural language processing system to automatically collect molecular interaction information from literature. [126] They studied the properties of this information and found that waiting time of the adoption of unique findings follows an exponential distribution. Others have described methods in which pathways and biological relations can be ascertained from the published literature or from publication abstracts. [127-129] However, none of these have addressed the specific problem of ascertaining information from the annotations of genome-scale experiments.

I have addressed this problem by creating an automated system called GENOTEXT (GENOmics conTEXT) to extract the contextual identifiers and annotations of samples from the largest publicly available gene expression repository. Moreover, I have been able to model the majority of these contextual annotations using the largest available biomedical vocabulary.

Introduction to the Gene Expression Omnibus

The Gene Expression Omnibus (GEO) is an international repository for gene expression data, developed and maintained by the National Library of Medicine. [130] GEO consists of a database-backed web-site (http://www.ncbi.nlm.nih.gov/geo) and a publicly-available File Transfer Protocol (FTP) site where data can be downloaded.

The GEO data model consists of four data types. GEO platforms (abbreviated GPL) represent a mapping between local gene identifiers and external identifiers, gene names, symbols, and other descriptors. Each GPL also describes the manufacturer of the method and the species for which the platform is used. Platforms can be defined as providing absolute measurements from a single sample, or relative measurements between two samples.

GEO samples (abbreviated GSM) relate expression measurements of multiple RNA transcripts with local identifiers, and are themselves related to a single GPL. Each sample corresponds to one or two biological sources, depending on whether absolute or relative expression measurements are represented.

Multiple samples may have been measured in a single experiment; each GEO series (abbreviated GSE) relates to multiple GSM. A GSM may participate in more than one GSE. GSM annotations include additional contextual information, such as author, organism, submitter, and submitter contact information.

Finally, a subset of the GSE have previously been manually validated as containing internally comparable data; these are represented as GEO data sets (abbreviated GDS). Each GDS relates to a single GSE. Most GDS further define experimental variables that were delineated in the original measurement design, such as “age” and “time”, and qualitatively define these variables for samples, such as “old” and “young”. In this way, a GDS may relate to GSM that are outside the designated GSE.

The relations between GSM, GSE, GDS and GPL are shown in Figure 13. At the time of this writing, the GEO web-site contains 16,448 samples (GSM) contained in 900 series (GSE) measured using 721 platforms (GPL). The GEO FTP site holds a subset of this: 8,519 GSM contained in 524 GSE measured using 195 GPL. The FTP site also defines 448 GDS, which relate to a total of 6,612 GSM.

Introduction to the Unified Medical Language System

The problem of finding biologically relevant data sets would be aided if the contexts and phenotypes behind genome-scale experiments were labeled with terms from a structured vocabulary, the same way that the genes, proteins, and other elements in these experiments are labeled using global identifiers, such as LocusLink and RefSeq. The vocabulary that might best serve this role is the Unified Medical Language System.

The Unified Medical Language System (UMLS) is the largest unification of over 60 biomedical vocabularies containing approximately one million inter-related concepts. [131] UMLS was developed and is maintained by the National Library of Medicine.

UMLS would be ideal for representing contextual and experimental terms for five reasons. First, UMLS concepts already include both genes and phenotypes. Second, it is currently the only large system freely available to academic researchers (needing only a signed license agreement). Third, UMLS covers concepts in both human and model organisms, and scales from the molecular level, to the physiological, to the pathological. The UMLS already includes GeneOntology and the NCBI taxonomy. [70,130] Fourth, UMLS already contains over 20 million relations between concepts, and these could be readily taken advantage of for relating phenotype concepts. Fifth, and most important, UMLS is meant to serve as a standard format for distributing terminologies. Thus, providing UMLS labels or methods to unify existing nearly-comprehensive data sets with UMLS will allow investigators to gain synergy with other projects involving UMLS (for example, joining with hospital databases of patients).

The UMLS has three components. The Metathesaurus contains a catalog of unified biomedical concepts, relations between concepts, and text strings mapped to each concept. The Semantic Network contains a catalog of 135 higher level categories for all concepts in the Metathesaurus, as well as relations between these categories. The

SPECIALIST lexicon and other resources provide data and tools for processing the text strings associated with UMLS concepts.

Though UMLS has largely been used for research and applications in medical informatics, component vocabularies within UMLS have been extensively used in molecular biology and genomics. The NCBI Taxonomy provides over 150,000 concepts representing known species and their proper taxonomy. [130] GeneOntology provides nearly 13,000 concepts representing the normal molecular function, biological process, and cellular components for proteins in eukaryotic cells. [70] Incorporation of the SNOMED International vocabulary version 3.5 provides over 113,000 terms, especially covering pathological terms and concepts. The International Classification of Diseases 9th revision adds nearly 19,000 concepts related to human diseases. Addition of the Online Mendelian Inheritance in Man and Digital Anatomist provides support for concepts for genes identified to be related to human disease and human anatomical structures, respectively. It is important to note that the latest versions of these vocabularies are not necessarily in the most recent releases of UMLS.

An example of the basic relations in the Metathesaurus is shown in Figure 14. The UMLS Metathesaurus relates the component vocabularies by creating concepts that span the vocabularies. Each unique concept thus relates to one or more source vocabularies. A concept may have multiple listed synonyms and terms; each term is uniquely specified in the Metathesaurus.

Asserted structural or hierarchical relations between concepts are primarily stored in the UMLS related concepts table (MRREL), while statistical relations between concepts appear in the UMLS co-occurrence table (MRCOC). These statistical relations represents frequencies of both concepts appearing together in a source database; pragmatically, the majority of concepts participating in these relations are MeSH headings, and the relations represent the number of instances when both headings appeared together in a MEDLINE record.

Introduction to MetaMap

MetaMap is a software program written and maintained by Alan Aronson at the National Library of Medicine. [132] MetaMap takes formatted or free text and generates a list of potentially matching concepts from the UMLS Metathesaurus. This is done in a chain of five steps, including first parsing the text sentences and word phrases, generating variants for words, finding candidate UMLS strings matching words and variants, evaluating these candidates, then finally mapping the final candidates to the word phrase.

Methods

Using the METAMORPHOSYS tool, I created a subset of the 2003AC release of UMLS, dropping vocabularies I initially considered as less relevant. These are listed in Table 2. Though we excluded the 1993 Online Mendelian Inheritance in Man (OMIM) vocabulary, I did keep the 1998 OMIM vocabulary. All subsequent analysis was performed using the remaining subset, which contains 878,496 concepts described by 1,724,070 text strings. These concepts are related in 22,524,248 defined relations and 13,864,516 statistical relations.

MetaMap exists as a standalone program as well as a set of programming libraries. I wrote software in PERL that extracted seven text items from data from the Gene Expression Omnibus, listed in Table 3, and stored these within a relational database implemented in MySQL. Using the MetaMap toolkit, I created software in Java that processed each of these seven text items and stored in a relational database the UMLS string unique identifier (SUI), the score of the match, and the phrase of original text in which the string was found. As concept unique identifiers (CUI) were needed in subsequent analyses, they were automatically determined using the SUI to CUI relations in the UMLS concepts table (MRCON).

I defined gross mapping errors as those caused by the incorrect interpretation of abbreviations leading to multiple strings, such that it was highly unlikely that there was any proper reference for these mapped strings. I chose to eliminate strings instead of concepts, since these could be globally eliminated. I wrote a program that took specified SUI and text fragments (described as regular expressions) and eliminated these mappings.

Results

Success of MetaMap in extracting UMLS concepts from GEO text strings

I manually evaluated whether the final-candidate concepts designated by MetaMap were sufficient to represent the text items, or whether all candidate concepts would be needed. I took the titles of the first six GEO data sets, shown in Table 4, and generated the list of candidate and final-candidate concepts. There was no difference in the mapped candidate and final-candidate concepts for one of the six titles; the differences for the other five titles are shown in Table 5. MetaMap disregarded as final-candidate concepts important terms such as “Gene Expression Profiling”, “malignant”, and “development”. Based on this manual review and opting towards dealing with false positives rather than missing true positives, I decided to store and consider all candidate concepts as being mapped.

Taxonomy of mistakes in mapping

The errors made by MetaMap in determining concepts from the GEO annotations fall into a taxonomy of at least eleven categories. There is a common theme behind all of these errors. As more authors and journals call for microarray data to be made publicly available, there is also an increasing level of detail being placed in the descriptions of the followed experimental protocols with the goal of aiding others in reproducing the findings. In general, as more text is entered having little to do with the experimental design, or as more words are abbreviated, the potential for errors greatly rises during automated text processing.

MetaMap mapping to unknown concepts

MetaMap created 21,867 mappings using 465 string unique identifiers that were not present in our subset of UMLS. It is possible these identifiers are present in the other vocabularies in UMLS that I excluded.

Missing concepts in UMLS

A significant number of important concepts related to investigations in molecular biology are missing in UMLS. For example, phosphate-buffered saline (PBS) is commonly used during protein isolation. As an investigational agent, PBS may be used to treat a control group. The abbreviation "PBS" maps to the concepts Lead (C0023175, the element), Lead (a homeopathic remedy), and Peripheral Blood (C0229664), but there is no available concept for phosphate-buffered saline.

Similarly, the abbreviation "TG" and term "transgenic" do not map to a UMLS concept, and instead mistakenly map to 12 concepts (such as thyroglobulin and tumor growth).

The commonly used technique of Serial Analysis of Gene Expression, [133] abbreviated SAGE, maps to a food item and to a homeopathic preparation. The technique of SAGE does not exist as a concept in UMLS.

Samples from the Cancer Genome Anatomy Project (CGAP) are annotated as being funded by the National Cancer Institute. However, the National Cancer Institute and other institutes of National Institute of Health are not concepts in UMLS, though the NIH itself is (C0027468).

Abbreviation errors

A number of abbreviations are used by investigators in writing the GEO annotations. Currently, MetaMap has no way to avoid parsing and mapping these abbreviations to UMLS concepts. Correcting many of these errors required a thorough reading of the descriptions, and in some cases, finding and reading any associated MEDLINE abstracts related to the data set in question.

A few examples are noted here. A previously manufactured microarray by Affymetrix to study gene expression in the rat was called the Rat Genome U34 Set, and often abbreviated “RG U34”. “RG” is mistakenly mapped to both retinal ganglion and radical gastrectomy.

Many of the data sets in GEO are referenced in publications. A few of these manuscripts were published in the journal BMC Bioinformatics, and this was noted in the description annotations of the associated GEO data sets. Unfortunately, “BMC” was mapped to bone marrow cells.

Other abbreviation errors were easier to find. State symbols and abbreviations for the United States are not present in UMLS, but are frequently used in descriptions. The symbol CA, for California, maps to six concepts, such as calcium, cancer, carcinoma, cardiac arrest, and coronary artery. MO, for Missouri, maps to the element Molybdenum.

The term “CGS,” meaning the compound CGS-21680 hydrochloride, a selective agonist for an adenosine receptor, maps to colloid goiter, while “DC”, abbreviating dendritic cells, maps to Dupuytren's contracture.

Missing synonym variants in UMLS

The majority of the source text used for concept matching by MetaMap comes from the string (STR) field in the UMLS Concept Names table (MRCON). In the subset of UMLS used for this study, there are 1,724,070 text strings for 878,496 concepts, suggesting that many of the concepts are mapped to multiple text strings as synonyms. Practically, however, the list of synonyms and variants is not complete. In our subset, only 386,297 concepts (44%) included more than one string per concept, and for these concepts, the average number of synonyms was still only 3.2. Of note, I eliminated additional supported languages from our UMLS subset besides English, with the hope of preventing mapping errors; this may have inadvertently contributed to errors, as shown below.

Several of the samples and data sets included the term “Type 2 diabetes.” Unfortunately, this phrase maps to the concepts “type 2” (C0441730) and “diabetes” (C0011847), because the closest synonym in my subset of UMLS is written as “type 2 diabetes mellitus.” However, the term “type 2 diabetes” is listed in the International Classification of Primary Care, Version 2-Plus, Australian Modification, and is an acceptable term for this concept.

One missing synonym with significant implications is the term “wild type,” used in over 400 GEO samples, series and data sets. The closest concept in UMLS is wild-type genetics (C0678926), which enters UMLS through the Alcohol and Other Drug Thesaurus and has no synonyms. Unfortunately, wild-type is not currently a MeSH heading. Most of the annotations with the term “wild” were mapped to the concept Wild (C0445392), which is a subtype of serotype typing.

The description for GEO sample 578 includes the text “granulocyte colony stimulating factor mobilized peripheral blood CD34 cells”. Though “peripheral blood” and “CD34” mapped properly, the combination, which should have mapped to monocytes, did not.

Poor text formatting

The GEO file format is simply described as using the ASCII character set. Presumably because the majority of investigators view these descriptions at the GEO web-site, some investigators have formatted and submitted their descriptions using the Hypertext

Markup Language (HTML). These tags are not properly filtered by MetaMap and create a source of errors. For example, the HTML tag "<li>" is used to introduce a list-item in HTML, but MetaMap mistakenly assigns the concept Lithium. Similarly, the HTML tag to introduce bold text, "<b>", maps to the concepts Bath, Brothers, Bacillus and Behavior.

Mistaken identity from experimental description

A number of the GEO description annotations contain lengthy coverage of the experimental design and protocols used in created the data. Unfortunately, MetaMap finds acceptable concepts for biomedical terms used in protocols and company names, yet these were not contributory to the accurate description of the samples. In some cases, these were misleading when no additional accurate descriptions of the same semantic type were present.

For example, GEO sample 12011 has in its description "...GenePix software analysis... Axon Instruments..." which maps to Axon. This sample was obtained from a murine microglial cell line called BV2, which is not represented in UMLS. Unfortunately, there is no other mapped concept that would permit an understanding of the cellular and tissue source for this sample; thus, the mapping to Axon is misleading. Similarly, "HP" is a commonly used abbreviation for Hewlett-Packard, a manufacturer of microarray scanners. "HP" maps to the concept Health Promotion (C0018738).

Occasionally, a product name led to an incorrect mapping. The Cyclone phosphorimaging system, by PerkinElmer, maps to the natural phenomenon Cyclone (C0337000).

Sometimes, the location of the company making products used for the experiment leads to mapping errors. Axon Instruments is located in Foster City, California. Ambion is located in Austin, Texas. Both of these companies made products used in the creation of GEO sample 12011, and thus the sample is incorrectly assigned the concepts Fostering (C0242298), related to foster homes, and austin (C0605411), an organic chemical. The worst example of this was for the city Saint Louis, where the abbreviation “St” maps to 116 incorrect concepts, such as shock therapy, skin test, stroke, and sinus tachycardia.

Salmon sperm DNA may be used to improve the signal in microarray studies by reducing the background. Samples with “sperm” in their description were mapped to Spermatozoa. Similarly, growth media is commonly used to maintain cell lines and yeast and bacterial cultures. The term “medium” maps to the anatomic concept of Tunica Media (C0162867). The abbreviation “TET” for tetracycline-regulatable alleles incorrectly maps to Tetanus (C0039614). [134]

Processing of RNA for hybridization with microarrays is commonly performed using material provided in pre-packaged units commonly called “kits.” Unfortunately, the term “kit” itself maps to the KIT Oncogene (C0812225).

Occasionally, publications are cited in GEO descriptions, commonly by listing the last name and first initials of an author. The description of GEO series 7 indicates the study is described further in a publication by “Khodursky AB et al.(2000)”, but “AB” was mapped to Spontaneous abortion (C0000786).

Unfortunate choice of experimental identifiers

I searched for concepts in the titles of GEO data sets, series, and samples. Occasionally, investigators put complete words in these titles, which mapped correctly, but abbreviations and identifiers referenced in associated publications were also often used. For example, the title of GEO sample 6751 is "Ova A/J 3" which maps to Ovum. The investigators used "Ova" to abbreviate "ovalbumen," mentioned in the description of the parent GEO series. However, "ovalbumen" is not an acceptable spelling in our subset of UMLS, compared to "ovalbumin", which does map to a concept that was not applied to this sample.

Similarly, the title of GEO sample 6768 title is "DM(LL)" which maps to Diabetes Mellitus. This sample is from a skin lesion from a patient with leprosy. GEO sample 4833 holds the title "KT1008c_DT/IN", where "DT" maps to Alcohol Withdrawal Delirium or delirium tremens. In an extreme example, the abbreviation "SF", such as the title of GEO sample 2138 "SF-295_CL12015_CNS", maps to many tissues and diseases, including spinal fluid, spontaneous fracture, swine fever, scarlet fever, seminal fluid, synovial fluid, seizure frequency, and more. The symbols "AB" and "MGM", used in GEO sample titles, mapped to Abortion and Meningioma.

Mismatch of semantic-type

Occasionally, the same term can map to multiple concepts. This is handled in UMLS through separate identifiers for each term for each term-concept relationship. Resolving between concepts is non-trivial, but one is aided by the UMLS Semantic Network, where each concept is mapped to 189 semantic types. The classic example of resolving the word "cold" is aided by the knowledge that cold temperature (C0009264) has a semantic type of Natural Phenomenon of Process (T070), different than the common cold (C0009443) and chronic obstructive airway disease (C0024117, occasionally abbreviated COLD), which have a semantic type of Disease or Syndrome (T047).

One extreme example of a semantic mismatch concerned a phrase used in over five-hundred sample descriptions. The phrase “spot quality assessment” mapped to the disease Exanthema (“spots”).

Unfortunately, proper word sense disambiguation for certain terms in molecular biology may require a species-level understanding during parsing. For example, the title of GEO sample 1740 is “Non-embryogenic callus of Medicago truncatula.” This mapped to the concept Bone Callus, yet *Medicago truncatula* is a species of green plant, which obviously has no bones. Here “callus” was referring to a cluster of undifferentiated plant cells. The proper concept does not appear to exist in UMLS.

Incorrect variant processing

MetaMap has the ability to take terms and create lexical variants, in order to more accurately match to UMLS terms. Unfortunately, this can lead to incorrectly assigned terms as well. One of the most frequently mapped concepts was Saw, meaning the surgical instrument. MetaMap takes the word “see”, used in over a thousand sample descriptions, creates the lexical variant “saw”, then maps that term to the surgical instrument.

MetaMap also has the ability to map to and from abbreviations. The description for GEO sample 51 contains the text “...Gene expression profile in developing mouse cerebellum at postnatal day 7." The term "postnatal day" internally maps to "PND", which then maps to the UMLS concept of Paroxysmal Dyspnea.

Irrelevant text in GEO

Over seventy GEO samples included the entire MIAME check list in their descriptions, beginning with "The MIAME Checklist Experiment Design: A. Type of experiment: for example, is it a comparison of normal vs. diseased tissue..." The MIAME check list is a suggested outline of critical points to cover in a description of a microarray experiment, including experimental design, experimental factors, hybridization design, labeling protocols, and more. [32] Because these investigators included the entire text of the checklist itself, these sample descriptions are nearly 1,500 words long, and cannot be parsed by MetaMap. This is an example of how including more text in a description can lead to less specific information about the experiment actually being transferred to a reader.

Spelling errors in GEO text

Certain terms in GEO annotations contained spelling errors. For example, the source of GEO sample 4100 is listed as "murine subcontaneous adipose" instead of "subcutaneous." The description for GEO sample 3258 noted specific "growth condtions" instead of "conditions." Descriptions for five additional samples contained the misspelled word "postive" instead of "positive."

Correcting mapping mistakes

I defined gross errors as those caused by the interpretation of abbreviations leading to multiple strings that were likely to be incorrect in all references and could be globally eliminated. Though by no means complete, the majority of these were found using three strategies. First, I manually studied the most commonly mapped concepts to determine why they were mapped. For example, I discovered the concept Kidney to be mapped to many samples; further analysis showed that MetaMap maps the single letter “K” to concepts like kidney and potassium. This was eliminated, and the list was studied again.

One of the worst mapping errors was found this way. I discovered a large number of annotations mapping to the UMLS concepts of ETS transcription factor and Aluminum. This led to the discovery that MetaMap was mapping to these concepts from the term “et al.” This term is used in the descriptions of over one thousand samples.

Second, I manually studied concepts used that are under semantic types not typically associated with gene expression analysis. For example, I found multiple concepts with the semantic type “Professional or Occupational Group” were assigned to GEO annotations. Further analysis revealed MetaMap was incorrectly mapping occupational terms such as “pilot”, “principal”, and “messenger” to terms such as “pilot studies”, “principal hypothesis” and “messenger RNA”.

Third, GEO sample, series and data set titles were studied to ensure abbreviations were not inadvertently leading to improperly chosen concepts. For example, I noted a large number of references to folic acid. The mappings were made through the synonym pteroylglutamic acid, which mapped to titles with the abbreviation “PGA”. The National Heart, Lung, and Blood Institute funded Programs in Genomic Applications, abbreviated PGA, have contributed over 600 samples into GEO.

I wrote a program that took these specified string unique identifiers (SUI) and text fragments (described as regular expressions) and eliminated these mappings. This approach was limited, however. Strings that were even rarely chosen correctly could not be eliminated in this way; otherwise a crucially correct mapping for a description would be lost. This program removed a total of 47,089 mappings involving 1,010 unique strings (SUI).

Success of Mapping Concepts to Gene Expression Omnibus Annotations

After applying the automated concept assignment and after manual elimination of incorrectly assigned concepts, both described above, a total of 286,398 string assignments remained from the 7 types of annotations associated with the GEO samples, series, and data sets. These strings map to 4,190 unique concepts.

Table 6 indicates the success of extracting any concepts from GEO text strings. The GEO series description annotation was the most information-rich, in that it provided the highest number of unique concepts. This was due to relative uniqueness of each series description, compared to sample descriptions which were often repeated across all samples within a series.

It is important to note that 8,454 out of 8,519 GEO samples (99.2%) were successfully directly matched to at least one UMLS concept, through some combination of title, description, source or keyword annotations. The 65 GEO samples with no directly matching UMLS concept had no descriptions and no discernable words in their titles; all were a variant of "S*nn*_EC_JYK" where *nn* indicates a number. These 65 samples belonged to a single human series which had titles and descriptions that were successfully matched to GEO concepts. Thus, I interpret these results as meaning that every GEO sample can be mapped either directly to UMLS concepts, or indirectly through its parent GEO series.

In general, the number of unique concepts elicited from a text item correlates with the length of the text item. This is shown in Table 7. In addition, the number of concepts assigned to a sample based on its keywords correlates with number of keywords provided.

The correlation between the number of unique concepts that map to an annotation and the length of the annotation is shown from Figure 15 to Figure 24 for each of the seven types of GEO annotations.

In general, the annotations on which MetaMap failed in parsing fell into two categories. Many GSM titles were short and contained laboratory identifiers with few recognizable words. In contrast to this, many GSM descriptions were so long that MetaMap could not even determine an initial set of concepts. Both of these extremes prevented MetaMap from mapping text to concepts.

Semantic Types of Mapped Concepts

Though there are 189 UMLS semantic types, our subset of the Metathesaurus includes concepts from only 135 semantic types. The 4,190 unique concepts assigned to the GEO annotations were drawn from 123 (91%) of these 135 semantic types, shown in Table 8.

I gave special consideration of those semantic types contributing relative few concepts, such as Vertebrate (T010) and Reptile (T014). The majority of these semantic types governed concepts that were assigned correctly. However, a few of these semantic types represented concepts that were entirely incorrectly mapped. For example, every concept in the semantic type Professional or Occupational Group (T097) was incorrectly assigned.

The simplest, most direct method to correct this problem would be to eliminate all concepts from rarely used semantic types. However, this method would occasionally yield invalid results. For example, the semantic type Human (T016) contained only two concepts: Human (C0020114) and *Homo sapiens* (C0086418), but these two concepts were used 2,340 times in the GEO annotations. The semantic type Experimental Model of Disease (T050) contributed only one concept, Disease Model (C0684309), which was assigned once in only one GEO annotation, the title from GDS 22 "Parkinson's Disease model". Despite the rare usage of this concept and semantic type, this assignment was correct. Thus, one cannot use paucity of assigned concepts as a way to potentially eliminate incorrectly assigned concepts (and types).

There was correlation between the number of concepts mapped and the number of unique concepts used within each semantic type (correlation coefficient 0.77). In other words, in general it was not the case that only a few concepts from each semantic type were used repeatedly. Two exceptions are noted here. Only two concepts were mapped from the semantic type Human, but these two were mapped in 1,780 GEO annotations. At the other extreme, only five concepts were mapped from the semantic type Nucleotide Sequence, but these five concepts were mapped a total of only five times.

Of note, the semantic types Cell (T025) and Tissue (T024) together contributed less than 3% of the unique concepts assigned from GEO annotations, while contributing 5% of the total number of concept assignments. This could be interpreted as very few of the words were spent in describing the source of the samples, and that there was relative similarity of samples.

There were 13 semantic types that contributed no terms during the automated assignment, shown in Table 9. The root semantic type, Entity (T071) is only associated with 5 concepts, none of which were used during assignment. Subjectively, most of the semantic types that are associated with significant numbers of concepts in UMLS, yet were not used during the automated concept assignment, govern concepts that are not currently or typically studied using gene expression microarrays (like language and drug delivery devices).

Table 10 indicates the top 50 concepts mapped to GEO samples, series, and data sets. I manually studied these concepts to ascertain exactly what concepts were mapped and why.

Since this list was generated after gross error correction (described above), and since manual error correction focused on the most frequently made errors, only a few concepts on this list are unexpected. The concept Utilization (C0042153) was frequently selected due to the significant use of the word “using”. The concept Sampling - Surgical action (C0441621) and Sampling (C0441621) were also frequently selected, due to frequent use of the word “sample”.

The concept Expression (C0185117) maps repeatedly, but the mapped concept is a subtype of surgical manipulation. Instead, the term “expression” as commonly used in molecular biology should have mapped to Gene Expression (C0017262), but the single word “expression” is not an accepted synonym for this important concept in UMLS. A similar error was made in the selection of control (C0243148), meaning an attribute, such as image control or volume control, instead of control groups (C0009932). Again, no additional synonyms are present for C0243148 that would allow for this match with just the word “control”. Labeling (C1167624) above refers to the process of stigmatizing an individual. The closest accurate concepts in UMLS would be Staining and Labeling (C0886517) or Stable Isotope Labeling (C1257948); use of the concept biotinylation (C0525026) would be appropriate for some microarray protocols.

The concept Robinson (C0443050) refers to a named strain of organism; a significant number of sample descriptions referred to a publication by Wen-Tao Peng, Mark D. Robinson, and others. [134]. A number of samples used the word “sex”, resulting in Coitus (C0009253). The concepts Seen (C0205397) and Vision (C0042789) were chosen for the frequently used word “saw”.

The concept Strain typing (C0449945) indicates the use of “strain”, in the context of a particular strain of organism studied. Unfortunately, the concept Muscle Strain (C0080194) was also selected because of its allowed synonym “strain”.

The concept Wild (C0445392) was incorrectly selected for the multiple instances of “wild-type” and “wild type”. Though the concept Wild-type genetics (C0678926) does exist in the Metathesaurus, there are no additional synonyms that would allow for matches to this concept; in other words, only exact matches with the string “wild-type genetics” are likely to map to this concept.

The concepts day (C0439228) and hour (C0439227) were often chosen correctly, but these concepts were also incorrectly chosen in response to many biomedical abbreviations, such as “DS domain”, “oligo d(T)”, “Protein H”, and “H. pylori”. Similarly, the concept Muscle (C0026845) was mapped properly in most instances, but was also incorrectly assigned through its synonym “musculus” as used in the common term “Mus musculus.”

Discussion

In this chapter, I have demonstrated an automated system called GENOTEXT that can successfully generate mappings to the Unified Medical Language System (UMLS) for every microarray sample stored in the Gene Expression Omnibus. Every sample can be directly or indirectly modeled using UMLS concepts. The results of this project suggest that UMLS, even in its current state, is sufficient to represent a number of the concepts held in the text-based annotations of genome-scale data. I showed that, with some exceptions, a longer annotation results in the mapping of more unique concepts.

However, I found eleven types of errors in creating these mappings. Most of these errors came from deficiencies in the mapping software, vocabularies, and the written text. Though I was successfully able to create a program to eliminate errors caused by incorrect mapping of abbreviations, proper modeling of concepts from the most common abbreviations used in molecular biology needs to be done in the future.

This work immediately suggests ways that text-based annotations can be written to facilitate automated extraction of concepts. Excess and superfluous text in annotations can lead to a large number of falsely mapped concepts. Even simple spell checking can help.

The highest-use semantic types in terms of concepts used are Amino Acid, Peptide, or Protein, Pharmacologic Substance, Organic Chemical, Functional Concept, Body Part, Organ, or Organ Component. This suggests that automated determination of specific annotations of these types might be possible.

The failure to properly map particular terms suggests a number of vocabularies that, if added to UMLS, could improve representation of these annotations. A listing of all authors of articles indexed in MEDLINE, the largest cities in the world and the United States, companies making products used in molecular biology research and their product names, and the institutes of NIH could be easily generated from semi-automated sources, such as almanacs, company catalogs, or web-sites. A vocabulary of other terms used in experimentation in molecular biology, such as buffer abbreviations, experimentation on animals, genetics and genomics, could be developed manually.

Future work in mapping concepts from annotations is suggested. The mapping could benefit from taking advantage of the proximity of terms within a sentence, which is currently ignored. Another strategy to limit improper assignment of concepts would be to eliminate mapped concepts that have never appeared in publication with relevant anchor concepts, such as RNA or Gene Expression Profiling. However, though abstract co-occurrence data is provided in UMLS tables, these are restricted to concepts that are also MeSH headings.

Additional mapping strategies could include counting the number of instances of a concept within a GEO annotation and restricting to those mentioned the most, or using concepts mentioned in as many different annotations as possible.

GEO series can be associated with a publication through its PUBMED identifier. The abstract and MeSH headings for this publication might be a better source of annotations for the GEO series, and this needs to be evaluated. However, only 279 GEO series have a PUBMED identifier listed, out of 524. Most of these GEO series listing an identifier are earlier sets. This might have to do with the submission of data sets before a publication has been approved; authors might not be going back to update their submissions once a PUBMED identifier is known.

Finally, success in mapping genomic samples to UMLS immediately suggests that the genes implicated in these samples and experiments could be mapped to cells, diseases, procedures, and patients whose data are increasingly being represented by UMLS.

4. Automated determination of disease and cell type studied in microarray samples from a large public repository by parsing text based annotations

Motivation

Technologies such as DNA sequencing and parallel expression measurements by microarray have allowed for novel diagnostic tests to assist in diagnosing diseases that are sometimes difficult to distinguish [5], to differentiate subgroups of disease that differ in prognosis [6], and to determine populations of patients that may respond to novel therapeutics. [7,8] With over 17,000 samples stored, and nearly 10,000 microarray samples available for downloading from international repositories, such as the Gene Expression Omnibus, [130] a researcher studying any disease or biological process using a genome-scale modality would ideally want to be able to gather as many of these relevant data sets as possible, to compare and contrast with her own data. Two important ways in which molecular biologists deem a sample as relevant are by the cell type and disease studied.

However, as the data in repositories continues to grow at an exponential rate, manual reading of the descriptions of data sets will no longer be feasible. Reanalysis and further discovery from collections of data is going to be dependent on extracting the annotations from these data sets; a manual approach is not scalable for the amount of data already present.

Automated determination of the cell type and disease studied by a gene expression experiment is particularly challenging. Even for data sets that are stored in standardized formats, such as microarray data, annotations of this data, including the experimental context of an experiment, are likely to be represented only by free-text. Furthermore, deciding whether a data set is relevant to ones study currently requires biological knowledge and expertise, especially when an exact match is not found.

There are several potential applications for the automated determination of the cell type and disease studied by a gene expression experiment. One could arrange and compare samples in known hierarchies of cell types, including those that reflect cellular differentiation processes, like those manually found that explain the differentiation of hematopoetic cells. [135] One could also find relevant samples across a variety of tissues reflective of diseases being studied.

I have previously described an automated system called GENOTEXT that extracts contextual identifiers and annotations from samples in the largest publicly available gene expression repository, the Gene Expression Omnibus. [130] Using GENOTEXT, I have been able to model the majority of these annotations using the largest biomedical vocabulary, the Unified Medical Language System (UMLS). [131]

An important example application of GENOTEXT is the identification of cell type and disease. Here, I will show how the mapping between GEO samples, series, and data sets and UMLS concepts can be used to automate the determination of cellular and tissue type of GEO samples.

Methods

As described previously, I have successfully loaded data from the Gene Expression Omnibus, including samples, series, and data sets. Through the GENOTEXT system, annotations of GEO data were successfully mapped to 4,190 unique concepts from the Unified Medical Language System (UMLS).

The Unified Medical Language System is a unification of multiple biomedical vocabularies. UMLS defines both inter-vocabulary and intra-vocabulary relations between biomedical concepts. The inter-vocabulary relations primarily represent synonymous concepts in multiple vocabularies, and are implicitly contained in the concept structure, where a single concept may have multiple terms from source vocabularies.

The intra-vocabulary concepts primarily represent relations between different concepts. Structural or hierarchical relations are primarily stored in the UMLS related concepts table (MRREL). These relations connect two UMLS concepts in a number of ways, including broader-to-narrower, parent-to-child, allowed qualification, and sibling relations. Statistical relations appear in the UMLS co-occurrence table (MRCOC). These statistical relations represents frequencies of both concepts appearing together in a source database; pragmatically, the majority of concepts participating in these relations are MeSH headings, and the relations represent the number of instances when both headings appeared together in a MEDLINE record.

Here, I created a software program in Java called CONTRAVERSE that implements a minimum-spanning tree (breadth-first search) across both types of UMLS relations. As applied here, to determine cell type and disease from annotations, I used the UMLS related concepts table (MRREL) because of the increased clarity of semantics in its relations.

Traversal started with the concepts Cells (C0007634) and Cultured Cell Line (C0682516) and was limited to relations leading to child concepts of semantic type Cell (T025). Without the semantic type restriction, traversal otherwise quickly descends into subcomponents of the cell, including DNA and its associated concepts, which were not desired for this analysis. Traversal continued from concept to concept to GEO data set to GEO series to GEO sample. A traversal path was ended when a GEO sample was reached.

In addition, direct traversal from the ten specific concepts listed in Table 11 to GEO samples was not permitted; these concepts participated in the hierarchy of concepts under Cells and Cultured Cell Line, but were not specific enough to assign a cell type to the sample. Without this restriction, due to the minimal-spanning tree approach used, premature assignment of a GEO sample to one of these ten concepts prevented further assignment to a more specific concept.

In a similar manner, I repeated this approach using a starting concept of Disease (C0012634) and restricted traversal to concepts with semantic type Disease (T024).

The breadth-first search resulted in trees starting from the initial concept(s) as the root and reaching the GEO samples as the leaves. Trees were formatted and printed using the freely available yEd Java Graph Editor (yWorks GmbH, Tübingen, Germany).

Results

Using CONTRAVERSE, I was able to predict a cell type in 3,381 out of 8,519 GEO samples (40%). Only 27 were related to the concept Cultured Cell Line (C0682516); the rest, shown in Figure 25, were related to the concept Cells (C0007634).

Of the unclassifiable 5,138 GEO samples, 1,981 (39%) were samples from plants or single cellular organisms (including *Arabidopsis thaliana* and *Saccharomyces cerevisiae*) and 273 (5%) were from non-mammalian multi-cellular organisms (such as *Drosophila melanogaster*). Of the remaining 2,884 unclassified mammalian samples, an additional 2,627 could be directly or indirectly mapped to a UMLS concept in the semantic types Body System (T022), Body Part, Organ, or Organ Component (T023), Tissue (T024), Body Substance (T031), Embryonic Structure (T018), or Neoplastic Process (T191). Though these might specify information about where the sample was obtained, these concepts are not at a cellular resolution, because UMLS is missing crucial relations between tissue belonging to an organism and the cell types contained within that tissue. For example, in UMLS, there is no asserted relationship between the concepts vastus lateralis muscle (C0224444) and muscle cells (C0596981).

Finally, this left only 257 unclassified mammalian samples that were also not mapped to any concepts in the six UMLS semantic types above. After manual review, I determined the reasons why each of these samples was unclassifiable. Eighty-two samples, including GEO samples 1130, 1678, 4424, 6684, and 6603, had no text in their description annotation and nothing contributory in any of the other six types of annotations.

I found 71 of these 257 unclassified samples belong to GEO series that contain mappings to two MEDLINE abstracts. The first is a publication referring to the concordance of expression levels between twins, for which 70 samples were placed in GEO. [136] MetaMap parsing of this abstract did not find any usable term in the semantic type Cells, or in any of the other six semantic types listed above. The term "lymphoblastoid cells" is present in this abstract, but MetaMap does not match this variant to the concept lymphoblastoid cell line (C0682526). The MeSH heading "Lymphocytes/*metabolism" was assigned to this MEDLINE record, however. The second abstract covers a single unmapped sample and does contain terms that map to the concept Breast Cancer of type Neoplastic Process. [137]

GEO sample 11963 and others reference a cell line "BV2" but does not otherwise contain text stating what type of cell or cell line this is. Similarly, GEO sample 1718 has the text "aCGH cell line", and GEO sample 1725 mentions "HMEC cell line" but UMLS does not have a matching concept for any of these. GEO samples 69, 3340, and others, only state "Individual LCL vs Pooled Control LCLs" in their descriptions. Similarly, GEO sample 8635 and its series repeated refer to "DC's" without once mentioning what it stands for.

GEO sample 1728, and others, refer to an experiment involving polio virus infection, but do not state the cells type used. GEO sample 2144 has a description stating a similar infection on HeLa cells, but this was missed by MetaMap and not mapped to the proper concept. GEO sample 823 used "NIH3T3", which is not a listed synonym for "NIH 3T3," and GEO sample 4100 used "Murine subcontaneous adipose," while "adipose" is not a sufficient synonym for adipose cell (C0206131).

Errors in cell type determination

Obviously, mistaken mappings from text annotations to UMLS concepts will lead to incorrect determination of cell type. In addition to this trivial source of errors, I found four additional types of classification errors seen in several of the samples that were considered classified. These were errors that persisted even after the removal of gross MetaMap mapping errors that had resulted from the misinterpretation of abbreviations. Though some of these errors occurred in GEO descriptions that are unnecessarily complicated, or contain text requiring sophisticated understanding of pathophysiology, there were several assumptions I made about GEO samples that were not always true, and important to relay here.

GEO description states interpretation

The description of GEO series 343 notes “DNA microarray profiling identifies molecular heterogeneity and suggests a role of B-cells in acute renal allograft rejection.” Because the conclusion of the experiment was stated in the methods, the samples in this series were mistakenly mapped to the concept of B-Lymphocytes (C0004561).

Single cell type assumption was incorrect

GEO series 272 is described as “T and B lymphocyte development profiles.” The search strategy I used stopped after any appropriate concept was reached for a sample, and did not continue to try to find additional related concepts. Thus, this series was incorrectly marked only with B-Lymphocytes (C0004561).

Experiment involved cells besides those used in the sample

The experimental protocol behind some samples involved cell types that were different than the ones studied using microarrays. For example, GEO data set 434 is titled “Embryonic stem cell cholesterol metabolism mutants”, while one particular sample within the data set is titled “mouse control, liver, RNA B6477A.” This particular sample maps to embryonic stem cell (C0596508); a more correct result would have been to additionally map to liver cell (C0227525).

Reduced specificity in cell type determination

Another type of error was found in specificity. GEO sample 3977 has a description listed as “CD4 lymphocytes…” yet UMLS does not have that phrase as a synonym for Helper-Inducer T-Lymphocytes (C0018894). Instead, MetaMap maps the more general term Lymphocytes (C0024264). Similarly, the description for GEO series 640 contains both “mammalian spermatozoan” and “18 day old testis” and its description contains the term “germ cells”. This series was mapped to the more general concept Germ Cells (C0017471) instead of the more specific concept Spermatozoa (C0037868). Because of the breadth-first search algorithm used, the closest concept to the starting concepts will preferentially be mapped and these are typically more general concepts.

Correct cell type classifications

Despite the errors noted above, many samples were correctly classified; a few representative examples are noted in Table 12.

Classifying GEO samples by disease

I repeated the traversal strategy in CONTRAVERSE starting with the concept Disease (C0012634) restricting to child concepts of semantic type Disease or Syndrome. Using this approach, I was able to predict a disease in 1,505 out of 8,519 GEO samples (18%). I found five types of classification errors in determining the diseases related to GEO samples.

Terms from experimental protocols that inadvertently map to diseases

Samples obtained from *C. elegans* that mentioned the term "worms," such as GSM 468, are classified as Helminthiasis (C0018889). Similarly, samples with the term "cold" in the description, such as "cells were scraped into 1ml ice cold PBS," map to Common Cold (C0009443). Samples describing a protocol involving the adenoviral transfection of genes map to Adenovirus Infections (C0001486).

Additional abbreviations leading to mapping errors

While reviewing the disease classifications of GEO samples, I found additional abbreviations that inadvertently mapped to terms suggestive of disease. The inclusion of the abbreviation "SSPE," a buffer containing sodium chloride, sodium phosphate, and anhydrous EDTA disodium dihydrate commonly used in molecular biology, causes one sample to map to Subacute Sclerosing Panencephalitis (C0038522). The acronym ORF, commonly used in genetics and genomics to mean open reading frame, is a synonym for Contagious Ecthyma (C0013570).

Similarly, GSM 2116 has the title "CCRF-CEM_CL7003_LEUKEMIA," where "CEM" maps to Contagious equine metritis (C0276037). GSM 12520 involved a cell line called

"JP 253"; "JP" maps to Juvenile Periodontitis (C0031106). Fortunately, this type of error occurred in very few samples, particularly due to the large number of mapping errors previously eliminated.

This problem was not restricted to abbreviations and acronyms; I saw a similar problem with mismatches of small words. Unexpectedly, portions of location names were considered as abbreviations and inadvertently mapped to diseases. The "Le" from the phrase "Le Genest-Saint-Isle France" in the description of GSM 2559 maps to Lupus Erythematosus (C0409974). Similarly, the "St" from "St Quentin-Fallavier" and "St Louis" maps to Esotropia (C0014877).

Finding a disease when none was studied

A number of samples in GEO represent data sets are provided as reference samples, where a particular disease was not studied. For example, the title of GSE 513 is "Cynomolgus monkey testicular cDNAs for discovery of novel human genes." Based on the word "testicular", this sample maps to Testicular dysfunction (C0405581). While it may be true that the list of genes found in this experiment may assist investigators in studying many testicular diseases, including testicular dysfunction, this experiment was not specifically studying testicular dysfunction. I discovered that our method has no way to accurately report the lack of a studied disease.

Ambiguity in mapping terms to diseases

There are several terms used to describe diseases that are not unique for a single disease. For example, the GEO data set GDS 485 has the title "Hypertension induced by angiotensin." The term "hypertension" maps to two concepts in UMLS: Hypertension or Hypertensive vascular disease (C0020538), and Hypertension induced by pregnancy (C0340274). MetaMap mapped all samples with the word “hypertension” to both concepts; I cannot resolve this ambiguity without additional context and samples representative of both concepts.

Reduced specificity in disease determination

Similar to the difficulties seen in cell type determination, I found errors in reduced specificity for disease determination. The title of GDS 274 is “Hepatocellular carcinoma metastasis”, which maps to both Neoplasm Metastasis (C0027627) in addition to Hepatocellular Carcinoma (C0019204). Due to the breadth-first search algorithm, this data set was assigned Neoplasm Metastasis, even though a more specific concept was available.

Correct disease classifications

Many samples were correctly classified by disease; a few representative examples are noted in Table 13.

More importantly, since these diseases exist in a hierarchy, samples can be searched in the context of that hierarchy. Using CONTRAVERSE to search for samples related to the Peripheral Nervous System Diseases (C0031117) and its child concepts finds samples from GEO series 465 (Expression profiling in the muscular dystrophies), GEO data set 412 (Amyotrophic lateral sclerosis), and GEO data set 198 (Inflammatory myopathy), as well as false positives. Performing a free-text search on the GEO website for “peripheral nervous system diseases” results in no samples or data sets found.

Discussion

I was successfully able to create an automated system, called CONTRAVERSE, that can determine a relevant cell type for a large number of samples in GEO, and a relevant disease for a smaller number of samples.

Use of these cell type assignments can take advantage of the hierarchy of cell types within UMLS. For example, though there are 60 samples that map directly to the cell type Lymphocytes (C0024264), there are an additional 36 samples that map to child concepts of Lymphocytes, including T-Lymphocytes (C0039194), Null Lymphocytes (C0024265), and Helper-Inducer T-Lymphocytes (C0018894), making a total of 96 samples. This is illustrated in Figure 26. An investigator searching for samples related to lymphocytes can now take advantage of samples directly matching that concept and its descendant concepts.

However, the network architecture for the concepts and relations within UMLS is not truly hierarchical, but instead has a network topology. This is true even for subsets of the network under the concept Cells. Figure 27 illustrates this problem. The concept Neutrophils has two parents: Phagocytes and Granulocytes. Phagocytes has Cells has a direct parent, and thus samples involving Neutrophils will be assigned using this shortest-path, instead of the alternate path which would have provided greater hierarchical detail.

A number of lessons were learned during this process. The shortest-path algorithm, as used in CONTRAVERSE, attempts to assign a single concept to each sample. This is obviously incorrect when a sample consists of multiple-cell types. In addition, the shortest-path to a sample may assign a relevant superficial concept, at the expense of a more specific concept. Alternative search strategies should be tried in the future.

Some of the GEO series provide an identifier to a MEDLINE record. It is possible that using the MeSH headings from these records, or applying MetaMap to the abstract text may yield more accurate mappings to cell types and diseases, especially since the MeSH headings are assigned by trained personnel at the National Library of Medicine. However, these identifiers are only provided for 279 of the 524 GEO series (53%), and will only provide data on the entire experiment, not on individual samples. In addition, only 2,148 of the 238,072 concepts (0.9%) that are MeSH headings have a semantic type of Cells or one of the other six types studied in this chapter. If one of these 2,148 concepts is not applied to the abstract, it will not help in cell-type determination. However, MEDLINE abstracts may preferentially map to diseases. This could be used to improve accuracy and could even be used to positively state when a GEO sample or data set has nothing to do with a disease. The advantages and disadvantages of using MEDLINE data versus GEO annotations clearly need to be studied.

This work immediately suggests ways that text-based annotations can be written to facilitate automated extraction of cell type and disease. GEO titles should be written with whole words and numbers, and should be carefully evaluated for inadvertent abbreviations. GEO descriptions should be written without excess verbiage; the MIAME checklist should be followed, but the text of the checklist should obviously not be included. Descriptions should be written in plain text without the use of HTML. Descriptions should ideally be written without using a specialist jargon, and without requiring users to read the original publication. The original tissue or cell type of cell lines should obviously be included.

A number of terms were repeatedly missed during MetaMap processing suggesting vocabularies that could be added to UMLS to improve processing. A listing of journal titles and cell lines used in research could be easily generated from semi-automated sources, such as almanacs, company catalogs, and web-sites including those maintained by the American Type Catalog Collection and the Coriell Cell Repository.

In addition to improving the vocabularies in UMLS, additional links between anatomical concepts and histological concepts are also suggested, such as providing relations between all organs and anatomy structures and their component cell types. Incorporating relations from vocabularies such as SNOMED may provide this in an automated manner.

There are additional areas of analysis that are immediately suggested by this work. Groups of samples for which cell type and disease annotations have been mapped can be compared in an unsupervised manner to discover connections; two cell types or diseases may resemble each other in terms of their gene expression profile. Samples falling under a cell type branch may not exactly resemble each other; these differences may even suggest previously unknown sub-types of cells.

5. Extracting catalogs of measured genes from the largest public repository of expression measurement data

Motivation

At the time of this writing, over 8,000 RNA expression measurement sets are publicly available from the Gene Expression Omnibus (GEO), an international repository for gene expression data developed and maintained by the National Library of Medicine. [130] GEO consists of a database-backed web-site (http://www.ncbi.nlm.nih.gov/geo) and a publicly-available File Transfer Protocol (FTP) site where data can be downloaded. GEO platforms (abbreviated GPL) represent a mapping between local gene identifiers and external identifiers, gene names, symbols, and other descriptors. Each GPL also describes the manufacturer of the method and the species for which the platform is used. Platforms can be defined as providing absolute measurements from a single sample or relative measurements between two samples.

GEO samples (abbreviated GSM) relate expression measurements of multiple RNA transcripts with local gene identifiers, and are themselves related to a single GPL. Each sample corresponds to one or two biological sources, depending on whether absolute or relative expression measurements are represented.

Currently, GEO expression measurements are not mapped to fixed, universal identifiers, such as LocusLink. [31] Our goal here was to create a single relational database with as many GEO expression measurements mapped to LocusLink identifiers as possible. There are numerous difficulties that make this a non-trivial process. First, there is no specified standard as to which universal identifier must be listed in a GPL; these include GenBank, LocusLink, or UniGene identifiers, or references to clones from over seventy catalogs. Second, the column structure for a GPL is not specified, and the universal identifier could appear in any column. Third, contrary information is often provided in a GPL; for example, if both a GenBank accession and a UniGene accession are provided as the mapping for a local gene identifier, the GenBank identifier may more closely represent the actual sequence used on that platform, compared to the UniGene identifier, which may have already been retired.

The value in providing a mapping from GEO expression measurements is manifold. First, as of this writing, there are over 100 million expression measurements in GEO. The properties of expression for each gene across this many measurement contexts have yet to be studied. Similarly, characteristics of a nearly-comprehensive set of expressed products can be studied globally. Second, mapping to LocusLink allows data to be compared across paralogs and orthologs, using Homologene. Thus, gene expression measured in similar contexts can be compared across species, and similarities and differences in the contexts themselves can be studied.

Previous attempts at mapping local gene identifiers to global identifiers

Many database-backed web sites are now available to translate gene identifiers from one type to another. Jennifer Tsai, and others, created a database-backed web site called RESOURCERER that serves as a cross-reference database between 21 microarray platforms. [69] Pinglang Wang, and others, created ProbeMatchDB to translate five types of clone and microarray identifiers. [138] Kimberly Bussey, and others, created MatchMiner which provides a similar function. [139] We have constructed a publicly-available web-based integration tool called UNCHIP (www.unchip.org) that relates Affymetrix identifiers to LocusLink. Though there are many other tools that provide similar functionality to these four, none specifically address the translation of all the various identifiers and platforms in the Gene Expression Omnibus to LocusLink.

Methods

Mapping GEO gene expression measurements to LocusLink was accomplished in three steps, depicted in Figure 28. First, I mapped many commonly used identifiers to LocusLink. Second, I loaded all GEO samples, containing expression data referenced by a local gene identifier, GEO platforms, containing the mapping between the local gene and external identifiers, and created relations between the two. Third, I created the mapping between each GEO platform's external identifier and a commonly used identifier. Each of these steps is described below.

Mapping from commonly used identifiers to LocusLink

Building from the previous UNCHIP program, I wrote a program in PERL to parse the LocusLink, UniGene, HomoloGene and Affymetrix data files. The LocusLink data file provided a catalog of and information for genes in a number of species, importantly excluding *Escherichia coli*, and other bacteria, *Arabidopsis thaliana*, and other plants, and *Saccharomyces cerevisiae* and other yeast. The LocusLink record for a gene provided a relation to a UniGene cluster as well as a few GenBank identifiers. The UniGene data files provided many mappings from GenBank and many clone identifiers to UniGene clusters. The HomoloGene file represented orthologs and paralogs by relating multiple LocusLink identifiers across species through a HomoloGene identifier.

Finally, the Affymetrix data files provided mappings from Affymetrix identifiers and GenBank.

From these files, I created a list of potential identifiers for clones from the UniGene files for each species. I manually implemented an iterative strategy that clustered the clone identifiers by the first 4 to 6 characters, and I determined the characteristics of the largest groups. Representative examples of these were manually queried using the Entrez web-site which resulted in the name of the grouping containing those identifiers. I will use the term *identifier-spaces* to refer these groupings of identifiers. A regular expression was created to describe the identifier-space, a name was given, and these were removed from the list to be processed.

Loading and linking GEO samples and platforms

I loaded all publicly available GEO samples and platform data into a relational database. The GEO platform data describes how GEO samples using that platform are organized. Each GEO platform is stored as tab-delimited text with a single header line explaining each of the columns. One of the column headers (not necessarily the first) is marked "ID" and marks that column as holding local identifiers for that platform. Each row subsequent to the header row represents a single gene being measured.

As shown in Figure 29, the GEO platform data was transformed into three tables: (1) GPL Header holding the GEO platform identifier, a column number, and the text of one column from the header row, (2) GPL ID holding the GEO platform identifier, the text of one row in the "ID" column, and a newly generated number unique for that platform and local gene identifier, and (3) GPL Data holding all data from each row and column.

Expression data from GEO samples are stored in a tab-delimited text format, with the column structure described by the associated GEO platform. Each row represents a single gene expression measurement, and a single column in each row contains the local gene identifier. I created a single table with holding all gene expression measurements, along with the local gene identifier, GEO sample identifier, and platform identifier.

Mapping from GEO platform external identifiers to commonly used identifiers

I manually inspected sample records measured under 195 GEO platforms to determine which column in the platform might be a valid identifier from an established identifier-space. I preferentially chose columns with identifiers that were fixed, stable, with little commitment.

Finally, I performed a database join across six tables as shown in Figure 30 and assessed the success in mapping gene expression data to LocusLink identifiers. For the expression measurements I could map to LocusLink, I also mapped them to orthologs and paralogs within and across species using HomoloGene relations.

Results

Mapping from commonly used identifiers to LocusLink

Using the manual iterative approach described above, I was able to map identifiers from 70 identifier-spaces with LocusLink, some of which are listed in Table 14.

I created over 28 million mappings between an identifier in one of these identifier-spaces to a LocusLink gene. As an extreme example of this mapping, the gene gamma actin (LocusID 71) is the most identified gene with 93,745 identifiers. However, the genes human glyceraldehyde-3-phosphate dehydrogenase (LocusID 2597) and human eukaryotic translation elongation factor 1 alpha 1 (LocusID 1915) are referenced in the most identifier-spaces, having 31 types of identifiers. Other frequently referenced genes are listed in Table 15.

A subset of the mappings for the gene N-acetyltransferase 2 (LocusID 10) are shown in Table 16. Some of these mappings are direct and trivial: the numeral "10" directly maps to the gene with LocusID 10. Several identifiers, such as "X14672" and "AAA59905" are directly referenced in the LocusLink record for the gene, and thus map without translation. Other identifiers require one or two levels of indirect reference. For example, the GenBank GI identifier "10286060" refers to GenBank "AV684197", which itself refers to UniGene cluster Hs.2, which refers to this gene. Mappings from clone identifiers are also provided; IMAGE 1870937 maps to both GenBank AI262683 as well as BX095770, but both of these are in the same UniGene cluster Hs. 2. Note that mappings for official and commonly used symbols, as well as various common permutations of identifiers are also provided.

Loading and linking GEO samples and platforms

The publicly available GEO data as of this writing included 8,519 samples measured using 195 platforms. I was able to load these data into a relational database, resulting in 104,171,741 gene expression measurements related to 2,452,203 platform local gene identifiers.

Twenty-one GEO data sets had some formatting problem with GEO platform data. Two examples are given here. GEO platform 561 has external identifiers provided for only nine of 28,800 measurements. The submitters of these arrays have claimed that "clone identification information was proprietary" and that they "only included the identifiers that were going to be made public in an upcoming publication." GEO platform 248 has header items transposed; column 1, currently labeled as "ID", was switched with column 3, currently labeled as "Column".

Mapping from GEO platform external identifiers to commonly used identifiers

I was able to manually create a mapping from the external identifiers specified in 163 GEO platforms to commonly used identifiers. For 99 of these GEO platforms, this external identifier was found to be GenBank identifier; an additional 11 platforms specified GenBank GI numbers. Three platforms contained identifiers that matched UniGene and two platforms provided LocusLink identifiers.

Using these tables, I was able to create 985,670 relations between GEO platform local gene identifiers and LocusLink identifiers. Specifically, this mapping related 922,382 GEO platform identifiers (37% of the total 2,452,203) and 61,648 (29% of the total 211,433). Of the 922,382 GEO platform identifiers, 878,349 (95.2%) refer uniquely to one LocusLink identifier, while 44,033 (4.8%) refer to more than one gene. For example, local gene identifier 33,377 from GEO platform 371 refers externally to GenBank identifier M94081, but that identifier relates to 62 different LocusLink identifiers (all different T-cell receptor genes). One could consider removing from analysis those platform identifiers that relate to more than gene, unless data increasing the specificity is available; otherwise, follow up and validation of results involving these identifiers is problematic.

Interestingly, of the 61,648 genes measured in LocusLink, 56,421 (92%) are measured by more than one GEO platform, but 5,227 (8%) are measured by only a single GEO platform. In other words, a finding based on one of these uniquely measured genes may currently be impossible to validate on another platform. Of note, human gamma actin has more probes in more platforms than any other gene (1,184). Other genes represented with many probes on many platforms include human glyceraldehyde-3-phosphate dehydrogenase (691), human beta actin (567), human eukaryotic translation elongation factor 1 alpha 1 (466), and human ribosomal protein L3 (458).

Quantitatively, it was the case that the more identifiers that were available for a gene in LocusLink, the greater the number of probes on GEO platforms; the correlation between identifier count and probe count was 0.69, and between log identifier count and log probe count was 0.84 (shown in Figure 31).

Discussion

With this work, I can now unify over 50 million expression measurements from the Gene Expression Omnibus with LocusLink gene identifiers, and can now study the properties inherent in gene selection for microarray catalogs.

Microarrays are one of the first nearly-comprehensive measurement systems commonly available to molecular biologists. We commonly treat measurements from microarrays as behaving in characteristic distributions, such as a normal or gamma distribution.

However, this places critical assumptions on the underlying set of genes in each measurement platform.

With this work, I am showing that genes are not measured equally within or across expression measurement platforms. Contrary to this, the number of probes available to measure a gene is a function of the number of identifiers available for it.

Most identifiers are either manually assigned during targeted sequencing efforts, or arbitrarily assigned during high-throughput cDNA sequencing efforts. This may introduce two independent biases. At one point, a gene may have been deemed as interesting, leading to multiple sequencing efforts and having multiple identifiers assigned. These genes may now be overrepresented on platforms. Alternatively, genes that are highly expressed, and thus repeatedly found and re-identified during high-throughput cDNA sequencing, may also be overrepresented on platforms.

Regardless of the cause, if the most familiar genes are the ones measured the most often, this will introduce a subtle bias if microarrays are used in a “hypothesis free” manner. If multiple independent probes are created for these genes, some of these probes may work better than others, leaving the user to speculate as to the exact level of expression of the gene. Additional effort may have gone into designing better probes for these probes, resulting in more accurate measurements than for other genes. In addition, multiple measurements for a single gene can change the overall distribution of expression measurements on the array, and can alter the prior probabilities of particular functional categories being implemented in an experiment.

Future directions

I used a manual approach to create bridges between GEO expression measurements, GEO platforms, commonly used gene identifiers, and LocusLink. In the future, this type of mapping could be automated for new GEO platforms. Software could be written to study and score each column in the GEO platform description based on the number of matches to terms within a single identifier-space and species.

Though seven species are included in the mapping from GEO expression data to LocusLink (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, *Drosophila melanogaster*, and *Gallus gallus*), other important species are notably absent, including *Escherichia coli*, and other bacteria, *Arabidopsis thaliana*, and other plants, and *Saccharomyces cerevisiae*. Though these species have fixed global identifiers, they are not presently incorporated into LocusLink. The National Center for Biotechnology Information has indicated that LocusLink identifiers will be subsumed by NCBI Gene identifiers in the near future. If these identifiers include these missing species, it will open the data already available for these species for unification in a single identifier-space.

6. Automating the extraction and determination of significant genes from GEO data sets

Introduction

The number of expression measurements already available studying particular biological processes and diseases is surprising. At the time of this writing, 448 defined data sets containing 6,612 parallel expression measurements are publicly available via the Gene Expression Omnibus (GEO), an international repository for gene expression data developed and maintained by the National Library of Medicine. [130]

In chapter 3, I showed how we can model the annotations of GEO samples using the Unified Medical Language System (UMLS). In chapter 5, I demonstrated how I can unify over 50 million expression measurements from GEO with LocusLink gene identifiers.

My hypothesis here is that if significant genes can be found from all possible comparisons of biologically interesting groups of samples, I could find a more significant list of genes involved with experimental variables more successfully than possible from any one data set.

Out of the 448 GEO data sets available, 409 (91%) define experimental variables that were delineated in the original measurement design, and provide values for these variables for each sample. As an example, GDS 200 is an experiment in which the investigators measured gene expression as a function of two tissue types ("Ca1 hippocampus" or "dentate gyrus hippocampus"), three treatments ("context and shock", "control", or "shock only"), four time points (1, 2, 4, or 6 hours), and three strains ("C57BL/6J", "50% C57BL/6, 50% SJL Alzheimers", or "MK-801-injected C57BL/6J").

There are 72 possible combinations of these variables, yet only 64 microarray measurements are provided in this data set, indicating that not every possible combination of experimental parameters was tested.

There have been many published methods on determining genes most significantly different between two groups of expression measurements. Our goal here was to comprehensively apply a single commonly used method of significance determination across all possible two group comparisons in all GEO data sets, in order to study the characteristics of genes most often found to be different across the variety of experimental contexts found in GEO.

Methods

GEO consists of a database-backed web-site (http://www.ncbi.nlm.nih.gov/geo) and a publicly-available File Transfer Protocol (FTP) site where data can be downloaded. GEO data sets (abbreviated GDS) represent collections of data that have been manually validated as containing internally comparable data. Using methods previously described in chapter 5, I placed all GEO gene expression measurements from GEO samples assigned to GEO data sets into a single database, and mapped the GEO gene identifiers for these measurements to LocusLink.

I created a software program in Java, called COMPRARE, that iterates over every GEO data set and over each experimental variable. The iterated experimental variable is considered the *test-variable*; the other available variables in the data set, if any, are considered *background-variables*. COMPRARE iterates through all possible values for all the *background-variables*, and within this iteration, it creates all possible pairings of values within the *test-variable*. Thus, in putting these values together, I had a pair of states that differ in value at the *test-variable*, but are exactly similar for all the *background-variables*. An example of a two group comparison is shown in Figure 32.

Expression samples are then found that match these states, and each gene is compared across the two groups, using one or two tests. If more than two measurements are available for a gene in both groups, an f-test is applied; if significant, the gene is evaluated using a t-test with unequal variance, otherwise a t-test with equal variance is used. If the *p* value for the comparison is under 0.01, the gene is recorded.

If one or more measurement is available for a gene, a mean expression value is calculated for each group and a fold difference is calculated. If a gene shows a difference greater than 2 fold, it is recorded.

Data measured using a platform indicated as dual channel (measuring relative expression levels) are excluded from the fold difference calculation, due to the presence of positive and negative values. Any gene with any negative measurements measured on any platform is similarly excluded.

For experimental variables of interest, I looked at the genes implicated in all two-group comparisons involving that variable, and sorted them by the number of positive comparisons. I then joined multiple genes into ortholog/homolog families using the HomoloGene relations, assigning each family the sum of the individual gene counts. [130]

Results

Using the nested iterative approach implemented in COMPRARE, as described above, I performed 11,121 two group comparisons across all GEO data sets, evaluating a total of 133,288,138 genes by t-test and/or fold difference. Roughly 13%, or 17,104,588, of these tests resulted in a positive gene; 3,663,177 (2.7%) were positive by t-test and 13,441,411 (10.1%) were positive by fold-difference. These findings implicated 44,838 genes out of 52,189 (86%) across 7 species: *Mus musculus*, *Rattus novegicus*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Canis familiaris*, and *Bos taurus*. Out of the 17,104,588 genes implicated in a comparison, 1,989,685 of the genes were implicated in both the t-test and fold-difference test. If the t-test and fold-difference test were independent, I would have expected only 369,412 genes implicated in both.

The GEO data sets contain 21 named variables, of which two are catchall variables called "unclassified" and "error". These are listed in Table 17, along with the number of data sets using the variable and the number of values used. Variables and values in GEO are specified as free text, and are apparently drawn from no established vocabulary.

To take one of these variables as an example, the variable "gender" was used in four GEO data sets. Four values have been specified for "gender": "female", "male", "male and female", and "mix."

I studied the variable "age" and the genes implicated across the 401 two-group comparisons in 24 GEO data sets. Each gene could be implicated in as many as 401 positive comparisons; however, because data sets involved multiple species, operationally, this maximum was much lower. I then joined orthologous and homologous genes into homology families.

A total of 19,725 genes were implicated in one or more comparisons involving "age," and these belonged to 11,488 homology families. The distribution of the sum of the count of positive comparisons for these families is shown in Figure 33.

The insulin and IGF-1 receptors have been previously implicated in aging in both mice and worms, suggesting that caloric restriction may lead to longer life. [140] Matthias Blüher showed in that mice genetically engineered to be missing the insulin receptor in fat tissue had an average 18% longer life-span than controls. [141] Koutarou Kimura showed that decreases in daf-2 signaling, the worm ortholog of the insulin/IGF-1 receptor, increases life-span. [142]

I found the insulin receptor family was highly implicated in two-group comparisons involving "age." The human insulin receptor was positive in 3 comparisons, while the mouse insulin and IGF-1 receptors were positive 133 and 102 comparisons, respectively, The mouse insulin receptor-related receptor is also a member of this homology family and was implicated in 39 comparisons. The mouse insulin and IGF-1 receptors were at the top 5.5%ile and 1.9%ile of the 19,725 gene list. In total, however, this family was implicated in 277 comparisons, placing it at the 0.3%ile of the 11,488 homology families. Only 36 additional homology families were more highly implicated than the insulin receptor family.

Similarly, other known genes involved in lifespan, including the sirtuin family of genes, were likely implicated in repeated experiments. [143] Human *sirt1* was implicated in 99 comparisons, placing it in the top 6.2%ile of the 19,725 gene list.

Discussion

Building on our previous work of loading all data sets from the Gene Expression Omnibus, including gene expression measurements, experimental variables, and values, I created an automated system called COMPRARE that can comprehensively make every possible two-group comparison in each data set, and can record the genes whose expression is significantly different in each comparison. I can then study the experimental variables themselves, such as “age”, and have found that important genes are repeatedly implicated across multiple comparisons.

More genes were positive than I expected. Though our p-value threshold for the t-test was set to 0.01, almost 3% of the genes tested were positive. The current implementation of COMPRARE does not adjust for multiple-comparison testing, and this may account for a greater number of positive genes than expected. [144] Over 10% of genes tested had increased or decreased over 2 fold. In total, 17,104,588, or 13% of the gene tests performed, were positive. Restricting to genes that pass both the fold-difference test and t-test cuts the number of positives to 1,989,685, a 10 fold reduction.

The thresholds for the t-test and fold difference testing were set arbitrarily. Future version of COMPRARE could recomputed significance in the context of permuted data, and could then determine appropriate thresholds dynamically to meet an expected false-discovery rate. [145] Additional methods to determine whether a gene is significantly different can be implemented as well, including those that consider each gene in the context of other measurements in the same samples, like Bayesian approaches. [146] Analytic methods that take advantage of time series measurements could also be used. [147,148]

Future directions for this work include an in depth pursuit of the specific genes involved in aging, as well as making a similar analysis of the other experimental variables, especially time and developmental stage.

It is interesting that 86% of genes undergoing testing were positive. The remaining 14% represent a set of genes that do not appear to change in expression values between states. In some ways, this could be a novel definition for a "housekeeping" gene, or a gene that is thought to be relatively constant in expression.

However, the qualitative nature of the variables and values will quickly become a limitation. Future work needs to be done on unifying these variables and values with concepts in the Unified Medical Language System.

7. Automated search and integration of multiple genome-scale measurements related to a disease

Introduction

Instead of focusing on the cell, or the genotype, or on any single measurement modality, using integrative biology allows us to think holistically and horizontally. A disease like diabetes mellitus can lead to myocardial infarction, nephropathy, and neuropathy; to study diabetes mellitus in genomic medicine would require reasoning from a disease to all its various complications to the genome and back. A researcher studying any particular biological process with a genome-scale modality would ideally want to be able to gather and reason over as many relevant data sets as possible.

In chapter 3, I showed how we can model the annotations of GEO samples using the Unified Medical Language System (UMLS). In chapter 5, I demonstrated how I can unify over 50 million expression measurements from GEO with LocusLink gene identifiers. In chapter 6, I demonstrated how I can generate lists of significant different genes from all possible comparisons in every GEO data set.

My hypothesis here is that by starting with a disease concept, I can now find biologically relevant genomic data sets and determine genes that are significantly different across several of these data sets. The hypothesis is that these genes are highly likely to be involved in the normal physiology, and possibly the pathophysiology, of the disease process.

Methods

As described in chapter 5, I have loaded data from the Gene Expression Omnibus, including samples, series, and data sets. Annotations of GEO data were successfully mapped to 4,190 unique concepts from the Unified Medical Language System (UMLS) using the automated system called GENOTEXT. I have also shown previously that a cell type can be predicted for 39% of GEO samples.

Asserted structural or hierarchical relations between concepts are primarily stored in the UMLS related concepts table (MRREL), while statistical relations between concepts appear in the UMLS co-occurrence table (MRCOC). These statistical relations represents frequencies of both concepts appearing together in a source database; pragmatically, the majority of concepts participating in these relations are MeSH headings, and the relations represent the number of instances when both headings appeared together in a MEDLINE record. For example, whereas MRREL contains a relation between adipogenesis and adipose cell, MRCOC contains a relation between adipose cell and obesity.

As described in chapter 4, I created a software program called CONTRAVERSE that implements a minimum-spanning tree (breadth-first search) using both types of relations in UMLS. As applied here, to find samples whose annotations are related to a starting concept, I used the UMLS co-occurrence table (MRCOC) because of the greater number of relations available that span source vocabularies. However, the majority of these relations connect two UMLS concepts that are also MeSH headings. Because of this, concepts that are not also MeSH headings essentially have no relations specified in this table. I added an optional process to “boost” the number of starting concepts, such that if the program noted that the starting concept(s) had no co-occurrence relations available, it added additional related starting concepts. These related concepts were determined using the UMLS related concepts table (MRREL) by traversing from starting concepts through relations designated as source-asserted relatedness and possibly synonymy. The “boosting” continued until at least one of the starting concepts had a co-occurrence relation.

Breadth-first traversal was initiated with the starting set of concepts after the optional “boosting” process and proceeded through the UMLS co-occurrence relations. Traversals continued from concept to concept to GEO data set to GEO series to GEO sample. A traversal path was ended when a GEO sample was reached. I considered that if traversal reached a GEO sample, it had already effectively reached that sample’s parent GEO series and GEO data set. Traversal to concepts was not restricted by semantic type. The only co-occurrence relations used were those marked as being from the citations to the published literature and with a source from the recent MEDLINE database.

To improve sensitivity and specificity, I designed and implement four optional filters that could be applied during the traversal step from one concept to another. Instead of traversing all UMLS co-occurrence relations, our first filter eliminated relations that were relatively insignificant given the two concepts being related. This was determined by counting the total number of MEDLINE records involving both concepts separately, then calculated the fraction of these totals covered by the number of abstracts cited in the relation between the two concepts. The second optional strategy was to test this fraction against a threshold for the source concept, the destination concept, both, or neither.

The third optional strategy was to ensure traversal always moved from a concept with more references in MEDLINE records to a concept with fewer references; I eliminated relations in which this was not the case. Our fourth strategy was to eliminate relations between concepts and GEO samples, series and data sets that had a score below a threshold. Scores for concepts mapped to GEO annotations are assigned by MetaMap, and are designed to reflect the degree of certainty of the match.

I manually determined all available GEO data sets to determine which may have an experimental design that was directly related to the study of diabetes mellitus. This list was compared to a list of data sets returned after searching the GEO web-site for the term “diabetes.” The concatenation of both lists is shown in Table 18.

The GEO web site returns GDS365 from a search query of “diabetes” because some of the samples are from diabetics after transplant. The web-site indicated GDS541 and GDS217 as relevant because both reference GSM761, a sample of normal cerebellum from a female who had diabetes mellitus. GDS167 was indicated as possible because “diabetes” is mentioned in the submitter’s institution annotation. These four data sets were felt to be false-positives.

I evaluated both the co-occurrence filter and the score filter by changing the thresholds and creating receiver-operating characteristic (ROC) curves for each filter and threshold. Sensitivity and specificity was measured based on the ability of each strategy to find the true-positive list of data sets after starting from the concept Diabetes Mellitus (C0011849).

The breadth-first search resulted in a tree starting from the initial concepts(s) as the root and reaching the GEO samples as the leaves. Trees were output in GraphML, then formatted and printed using the freely available yEd Java Graph Editor (yWorks GmbH, Tübingen, Germany).

As described in chapter 6, I created the COMPRARE system to perform every possible comparison between two named groups of samples across all GEO data sets, evaluating 11,121 two group comparisons involving 133,288,138 gene expression comparisons using t-test and/or fold difference. Using the results of the ROC curves, I chose the optimal parameters for sensitivity and specificity, and then traversed from the starting concept of Diabetes Mellitus (C0011849). At the depth determined to be the optimal for sensitivity and specificity, I determined the list of GEO samples (and from these, GEO data sets). I then kept only those GEO data sets containing at least one cell type reference to Muscle (C0596981). Using the results of COMPRARE, I created a list of those genes significantly different in any one of these GEO data sets, then folded this list taking into account the multiple species that were studied using the NCBI Homologene tables.

Results

In previous work, I demonstrated how microarray samples and data sets from the Gene Expression Omnibus (GEO) could be stored in a single database, with local gene identifiers mapped to LocusLink identifiers, annotations mapped to UMLS concepts, and all genes significant in any possible comparison of groups of samples listed.

Here, I have put all of this together. I first evaluated the sensitivity and specificity of using a breadth-first search traversal across the UMLS co-occurrence relations, starting from the UMLS concept of Diabetes Mellitus (C0011849) and reaching a true-positive set of GEO data sets and samples studying Type 1 and Type 2 diabetes mellitus. I determined initially that due to the large number of co-occurrence relations, I was able to reach every GEO data set and sample from the concept Diabetes Mellitus in eight or fewer steps. This is primarily due to the particular nature of the UMLS co-occurrence relations. Though UMLS is the concatenation and unification of multiple vocabularies, the network of concepts and co-occurrence relations follows that of a scale-free network. A few UMLS concepts participate in a large number of relations, as shown in Figure 34, and the involvement in relations drops exponentially.

As shown in Figure 35, the log of the count of UMLS co-occurrence relations for a concept is inversely linearly proportional to the number of concepts with that count, characteristic for a scale-free network. This implies the existence of a small fraction of highly-connected hub concepts.

Other relevant scale-free networks include airline routes, social networks, power-grids, the world wide web, the collaboration network of scientists, and some peer-to-peer file sharing networks. Search algorithms can take advantage of scale-free networks. Requests for music files in peer-to-peer networks typically move one computer to its most highly connected neighbor. These requests then travel randomly in the central core of highly connected computers until the appropriate file is found. [149] Typically, given the nature of the data, identical copies of the file are found in many locations in the network.

However, our goal with CONTRAVERSE is not to find samples, but to order GEO samples, series and data sets by equating distance from a concept as relevance to that concept. In addition, since I am traversing the UMLS co-occurrence network looking for related concepts without a specific correct answer in mind, our problem is not in sensitivity (i.e. finding the concept I am looking for) but in specificity (i.e. eliminating unrelated concepts). Knowing UMLS is scale free may allow us to take advantage of highly connected or cited concepts, by quickly eliminating those concepts and samples that are least relevant.

I implemented four strategies to improve the specificity in finding GEO samples and data sets related to starting concepts, while maintaining sensitivity in finding a true-positive list of related samples. The first strategy took advantage of the scores provided for each relation in the UMLS co-occurrence network: each relation between two concepts also contains the number of MEDLINE records in which both concepts are covered. Our first strategy was to eliminate UMLS co-occurrence relations that were relatively insignificant given the source and destination concepts being related. I determined this by counting the total number of MEDLINE records involving both concepts separately, then determined the fraction of each involving the two concepts together. I tested whether this fraction was greater than a threshold fraction, in the source, destination, or both concepts; these were varied for the second strategy. As shown in Figure 36, as I decreased the threshold fraction determining relevance eliminating more relations, our ability to find the true-positive list of diabetes mellitus-related samples and data sets improved. In addition, as shown in Figure 37, sensitivity improved when only the source concept was tested.

Our third optional strategy was to compensate for the scale-free nature of the co-occurrence network and ensure that traversals always moved towards a concept with fewer references in MEDLINE records. As shown in Figure 38, sensitivity and specificity worsened with this strategy.

I implemented a fourth strategy taking advantage of the assignment scores between UMLS concepts and GEO samples. As MetaMap was used to determine UMLS concepts from seven unstructured descriptive GEO annotations (sample title, sample description, sample keywords, sample source, series title, series description, and data set title), each mapped concept was assigned a score reflecting the degree of certainty of the mapping. Our fourth strategy was to eliminate mappings with scores falling below a threshold. As shown in Figure 39, as I increased the threshold score eliminating more mappings, our sensitivity and specificity worsened.

Based on this, I determined that the optimal search strategy involved traversing the co-occurrence relations while eliminating relations where the ratio of MEDLINE records referenced in the relation fell below 1% of the total MEDLINE records referenced by the destination concept, without otherwise taking into account the relative numbers of referenced records by the source and destination concepts, and using all MetaMap mappings between GEO and UMLS. I traversed the co-occurrence relations starting from the concept Diabetes Mellitus (C0011849) using this optimized strategy, and gathered all GEO samples at a depth of 3 or fewer relations away from Diabetes Mellitus, resulting in 1,273 samples. I took the 117 parent GEO data sets for all of these samples and kept only those that were mapped to a cell type of Muscle (C0596981), leaving the 16 data sets listed in Table 19. I then gathered the list of genes already determined to be differentially expressed in these data sets.

Several of these represent false positives. GDS11 contains a number of fibroblast samples from a patient with multiple congenital malformations, including a heart murmur. The concept Heart Murmurs (C0018808) was related to Dental Care for Chronically Ill (C0206196), which was related to Diabetes Mellitus (C0011849). GDS182 was implicated because it contains expression measurements from mouse duodenum (C0013303), which is related to the pancreas (C0030274), which in turn is related to insulin (C0021641). GDS276 contains samples of rat leg muscles, related to leg (C1140621), which is related to Amputation (C0002688).

A total of 12,876 genes were found to be differentially expressed in at least one of these data sets, of which 996 genes were different in 4 or more data sets. These 12,876 genes came from experiments and studies from three species, with the majority of implicated genes coming from studies involving mouse samples.

Implicated genes were then grouped by homology, where multiple ortholog and paralogs for a gene were combined into a single homolog reference. A total of 7,348 homolog families were found, with only 277 families (3.8%) containing genes implicated across 7 or more independent data sets. For example, SLC2A4 (glucose transporter 4, the insulin-responsive glucose transporter) was implicated in 4 rat (as LocusID 25139) and 2 mouse (as LocusID 20528) independent diabetes-related data set.

Within this list of 277 gene families are multiple gene families known to play a role in insulin signaling. Several of genes are known to play an important role in insulin receptor signal transduction. Grb2 is a protein that acts as an adaptor protein, that binds phosphorylated insulin-receptor substrate-1 (IRS-1) and SOS. [150] SLC2A4 is glucose transporter 4, the insulin-stimulated glucose transporter in fat and muscle cells. [151,152] MAPK3 is the mitogen activated protein kinase 3, downstream of Ras and well known to be phosphorylated in response to insulin-stimulation. [153,154]

There are complex interactions between the leptin and insulin receptor signaling pathways, and polymorphisms in the leptin receptor in humans have been shown to be associated with insulin resistance. [155,156] A number of adipocyte related genes appeared on our list, including UCP1, UCP3, adipsin and *DLK1*, also known as *Pref-1*. Adipsin is secreted by fat cells and expression is impaired in genetic and acquired obesity. [157-159] Pref-1 is a known inhibitor of adipogenesis. [160]

Finally, other genes known previously implicated with insulin resistance were also in this list of 277 gene families, including RRAD, [161,162] 11-beta hydroxysteroid dehydrogenase 1, [163,164] growth hormone receptor, [165] glycogen synthase 1 and 2 (muscle), glycerol-3-phosphate dehydrogenase 2 (mitochondrial), the epidermal growth factor receptor family, the STAT family of transcription factors, and the PPARA family of transcription factors.

Discussion

In this section, I describe how I have taken the largest set of publicly available gene expression data, mapped the experimental context of the data to concepts from the Unified Medical Language System, mapped the measured genes to NCBI LocusLink and Homologene identifiers, and calculated the significantly differentially expressed genes across all experiments. After testing and determining an optimal search strategy to traverse the scale-free network of the UMLS concept co-occurrence relations, I created a list of data sets computationally determined to be related to Diabetes Mellitus, intersected the significantly different genes from all of these and determined the gene families most involved in these data sets. The top 3.7% of these gene families appear to be enriched with genes known to be involved in insulin signaling, adipogenesis, glucose metabolism, and energy production.

A significant limitation of this work is the use of the HomoloGene table to specify ortholog relations. Though these relations are thought to be accurate and specific, since these homolog families include paralogs (homologous species within the same species), they essentially become many-to-many relations of genes across species, which may add non-specificity to matches and knowledge bases built using homology families.

As with any data reduction technique, there is a potential danger in eliminating the rarely cited relations between concepts, as this may eliminate potentially novel ways to consider data sets related to concepts. Future work is called for in developing additional strategies with improved sensitivity and specificity, and to draw from algorithms being developed for searching alternate scale-free networks.

This work demonstrates the importance of annotating genome-scale databases using a structured vocabulary, and here I show that UMLS is sufficient for this purpose. The system created here can link from a conceptual understanding of diseases, even from the ICD-9 billing code level, down to the genetic, genomic, and molecular level. In a sense, this is the first automated system built to study the new field of genomic medicine.

8. Automated reasoning to explain genetic data using genomic data and integrative biology

Introduction

Drawing conclusions from the integration of microarray data sets is an important inferential process that requires an understanding of the implications and semantics behind set operations such as *union, intersection,* and *difference,* when applied to expression data—a single measurement modality. In chapter 1, I addressed similar inferential processes involved in the combination of large-scale measurements *across experimental modalities*. The model I proposed for intersecting two nearly-comprehensive experiments or data sets is to address three independent intersections: that of *context*, *catalog*, and *content*. In chapter 3, I modeled the contextual annotations from the Gene Expression Omnibus (GEO), the largest database of publicly available gene expression data using the Unified Medical Language System. In chapter 5, I unified the catalogs of genes measured in GEO with the global identifiers in LocusLink. In chapter 6, I comprehensively compared all possible groups of samples from all GEO data sets to determine a list of genes significantly different in any comparison, or the content of this database. In chapter 7, I validated the context, catalog and content of this model by searching for experiments related to diabetes mellitus. I found that genes significantly different across more of the experiments found appeared to be more involved in insulin signal transduction and diabetes mellitus.

In this chapter, I will specifically apply this model to the problem of unifying quantitative trait loci (QTL) with gene expression data. A QTL is a region of a chromosome that is statistically significantly associated with a particular trait. The region is typically delimited by genetic markers, such as a single nucleotide or variable length polymorphism. A QTL signifies a difference in DNA around that area, such as a change in the regulation of a gene's expression or function, but this difference might only appear in a certain tissue and at a particular time or stage of development. In contrast to this, expression data involves the amount of expression of a number of genes in a set of samples. A graphical representation the two example types of data modeled is illustrated in Figure 40.

Our goal here was to create an automated system that could integrate any available and relevant genomic data and findings with modeled QTL so that the QTL could be explained. In other words, I desired a causal chain of biologically plausible events leading from the potential implications of having a polymorphism in a region of DNA to a reason for the trait explained by the QTL.

Methods

As described in chapters 3, 5, and 6, I have successfully modeled the context, catalog and content of a large amount of data from the Gene Expression Omnibus. Specifically, using the GENOTEXT automated system, I took annotations describing GEO data and successfully mapped these to 4,190 unique concepts from the Unified Medical Language System (UMLS). In addition, I have successfully mapped 37% of the local gene identifiers in GEO to LocusLink identifiers. Using the COMPRARE automated system, I performed every possible comparison between two named groups of samples across all GEO data sets, evaluating 11,121 two group comparisons involving 133,288,138 gene expression comparisons using t-test and/or fold difference. In this way, I have modeled the context, catalog, and content for many GEO samples and data sets.

Here, I modeled quantitative trait loci in a similar manner. First, I took the 140 quantitative trait loci described in the NCBI rat genome project. To model the context, I manually mapped the text description of each trait to the closest UMLS concept. The exact base pair coordinates and chromosome for each QTL were provided in the rat genome "seq_pheno.md" file, providing a model of the catalog of measurements. The content was trivially modeled such that any gene with a transcriptional start site within the QTL coordinates was considered positive.

At this point, I had modeled the context, catalog, and context of a large amount of genomic data in the form of gene expression profiles, as well as example genetic data in the form of quantitative trait loci. As stated above, my goal was to create an automated system that could generate a causal chain of biologically plausible events leading from the potential implications of having a polymorphism in a region of DNA to a reason for the trait explained by the QTL. Thus, intersecting the genetic and genomic data was going to require more than just mapping genes to a universal identifier, such as LocusLink.

Two separate types of causal chains needed to be developed. The first causal chain is *cross-modality*, and relates a QTL to significant microarray findings in general. A QTL indicates some difference in that region of DNA is statistically associated with a trait. There are many downstream biological implications for genes within a QTL, including nothing, decreased expression levels, altered transcriptional regulation, or altered protein amount, structure or function. I modeled the first of these implications, in that a gene physically residing within a QTL might be significantly differentially expressed in a microarray study related to the trait.

A second type of causal chain needed to be developed; this was a *gene-dependent* chain. Even if a gene was found to be within a QTL and significantly differentially expressed in a microarray study related to the trait, the changes seen in the gene across the two modalities still has to causally explain the trait.

Thus, I needed to implement knowledge for both the cross-modality and gene-specific causal chains. Many approaches have been used in capturing biological knowledge, ranging from first-order predicate models to probability based models. In building the two limited knowledge bases here, I recognize that the relations I am modeling are probabilistic in reality, and cannot be fully represented by first-order predicate languages. The aim here was to see how far we can proceed with simplistic assumptions and still learn from the process of integration. I recognize that this may be a limiting oversimplification.

I needed to implement this knowledge base in a system allowing for the depth-first traversal of assertions towards a goal. While this could have been implemented using graph traversals across a semantic network, it was optimal to choose a system allowing maximal flexibility in making assertions, minimal requirements in specifying the exact goal-based traversal strategy, and allowing for backward "reasoning". Thus, I implemented a biological reasoning system with the following five rules in PROLOG: [73]

1. A gene is in the location of a QTL if the transcriptional start site for the gene is within the coordinates of the QTL, on the same chromosome, and refers to the same species.

2. A gene is differentially expressed in a gene expression experiment if it is implicated in any two-group comparison, involving any experimental variable and background conditions.

3. A gene expression experiment is related to a trait if there is prior knowledge that a concept annotating the context of the gene expression experiment is related to a concept annotating the trait, within a set concept traversal distance.

4. Gene A can causally influence gene B if there is prior knowledge stating this.

5. Gene A is a homolog of gene B if they are in the same homologous cluster (data derived from NCBI Homologene).

6. Finally, a quantitative trait locus could explain its trait if there was a gene A in the location of the QTL and a gene B differentially expressed in a related gene expression experiment, and if a homolog of gene A could causally influence a homolog of gene B, and if there was a causal chain of reasoning back to the trait.

I also created a knowledge base with five assertions:

1. A decrease in the genes in homology group 2517 (which includes *Ucp3*) causes a decrease in electron transport (UMLS concept C0013846).

2. A decrease in electron transport causes an increase in fatty acids (UMLS concept C0015684).

3. An increase in fatty acids causes insulin resistance (UMLS concept C0021655).

4. Insulin resistance causes diabetes mellitus (UMLS concept C0011849).

5. An abnormal glucose tolerance test (UMLS concept C1260441) is diagnostic of diabetes mellitus (UMLS concept C0011849).

These rules and assertions were implemented in the Prolog language, given in Appendix A. I tested these rules by using the backward-chaining rules to "explain" the QTL Niddm40. [166]

Additional knowledge was also required to perform this intersection. To be able to unify between genome coordinates and genes, I mapped all genes referenced in LocusLink to their current precise genome locations for human, mouse and rat, using the identical base pair coordinate system.

Pragmatically, the easiest way to intersect the data across any set of modalities is to convert each measurement into a binary value, namely significant or insignificant, then perform the desired set operation. Instead of using raw expression measurements or log-odds scores, this is what I did here.

Results

Starting with the quantitative trait locus between base pairs 139105187 and 167448551 on chromosome 1 in the rat, the system found 406 genes contained in that region. The trait for this locus was manually mapped to the UMLS concept abnormal glucose tolerance test (C1260441). The system successfully related that concept to Diabetes Mellitus (C0011849) and found 81 GEO data sets significantly related to Diabetes Mellitus.

The system then found a causal chain leading from the decrease of a gene in the region of the locus to diabetes mellitus, specifically UCP1 and UCP3. It recognized that a decrease in UCP3 could lead to a decrease in electron transport, which could lead to a build up of fatty acids and insulin resistance, causing diabetes mellitus.

Finally, UCP3 was chosen as a final candidate since expression of UCP3 was found to be differentially expressed in several diabetes related GEO data sets.

Discussion

I have previously modeled the context, catalog and content of a large amount of data from the Gene Expression Omnibus, modeling the context using concepts from the Unified Medical Language System (UMLS), modeling the catalog by relating local gene identifiers to LocusLink, and modeling the content by performing every possible comparison between two named groups of samples across all GEO data sets. I have previously shown how I can successfully traverse the UMLS concept network to find data sets related to a biomedical concept.

Here, I developed a model framework that puts all this modeled expression data to biologically explain how a quantitative trait locus can lead to its trait. To explain a quantitative trait locus, the system was programmed to find genes based on location, chromosome and species. To find relevant gene expression data sets, the system was given two knowledge bases. The first was high level assertions about the downstream implications of traits, modeled using relations between UMLS concepts. The second was a gene-based pathway, with relations between the states of genes to UMLS concepts and diseases. The use of knowledge bases in molecular biology is not novel, but our ability to create a knowledge base grounded with concepts of genes and other biomedical concepts from two vocabularies (LocusLink and UMLS) may be. [71,72,98,102]

Gene-specific knowledge bases are becoming available from more sources. Several companies and research groups are generating assertions by automated parsing of MEDLINE abstracts. [126,129] Others are using human readers to build knowledge bases from manual reading of publications and curation.

However, cross-modality knowledge bases are not generally available. It is not clear how the size of the ideal cross-modality knowledge base compares to the size of the ideal gene-specific knowledge base. We can gain some insights using the UMLS as an analogous system. There are over 20 million relations between the approximately 880,000 concepts in UMLS, but only 612 relations between the 189 semantic types of concepts. Though this analogy is full of inaccuracies, the 250,000 genes in LocusLink might participate in 5.6 million gene-specific relations, but the thirty experimental and contextual modalities used in molecular biology may be related with as few as 100 assertions.

Additional work on expanding the cross-modality knowledge base is crucial.

I represented the cross-modality and gene-specific assertions using the first-order predicate language Prolog. In reality, the assertions I modeled are truly probabilistic, and would be better represented in a probabilistic framework, like Bayesian Networks. Future work should include re-implementing this system in such a probabilistic framework.

A significant limitation of this work is the use of the HomoloGene table to specify ortholog relations. Though these relations are though to be accurate and specific, since these homolog families include paralogs (homologous species within the same species), they essentially become many-to-many relations of genes across species, which may add non-specificity to matches and knowledge bases built using homology families.

LocusLink has long been acknowledged as an ideal way to identify genes, in that unique fixed identifiers represent genes grounded by actual positions in a genome, and because external relations to other sources are available, such as protein domains, official symbols and names, and diseases. To date, no similar vocabulary has been accepted to universally represent the experimental context of genome-scale data. This work suggests that gene expression data, sequencing and quantitative trait data, and other large-scale measurements in molecular biology can and should be annotated with identifiers from a vocabulary such as UMLS. By linking with both LocusLink and UMLS, this system demonstrates that high level computable pathways can be represented and not only used to join two genome-scale modalities in molecular biology, but also to link with the complexity of human diseases and our understanding of them.

9. Limitations to intersection

In this dissertation, I have been using a structured exploration of integrative biology as a biomedical derivative of Exploratory Data Analysis (EDA) proposed more than 25 years ago by Tukey. [167] More than a set of techniques, the EDA approach suggested techniques on how to find structure in and gain insight from data sets, and how to develop simple models from that data. With this dissertation, I have developed the first automated tools to assist in determining and exploring relationships between experimental modalities in molecular biology. But just as improper use of EDA can lead to false discoveries and missed relations between variables, so too are there caveats to our use of integrative biology, and specifically the use of intersection.

There are significant potential limitations in using intersection across data sets. First, as shown in Figure 41, one experiment may have examined a small portion of a large biological process and another experiment may have examined a separate, non-overlapping portion of the same process. Though several genes may be involved in the same large process, the intersection will not retrieve them, possibly leading to false negatives.

Second, if genes are implicated in a process in one data set but are not even measured in another, these genes will not be present in the intersection and therefore some may lead to false negatives. This is illustrated in Figure 42. This can happen commonly in comparisons across species, where one species has fewer genes than another. Related to this, intersection is crucially dependent on the definition of identity. For example, two genes could be defined as identical across species based on sequence homology or pattern and timing of expression. Differing identity relationships will result in different results after intersection.

Third, it is often the case that when intersecting two near-comprehensive data sets that the list of genes implicated in either experiment is a function of input parameters (i.e. thresholds of significance, or algorithm used to determine significance). As the input parameters change, the lists available for intersection change, as does the result of the intersection. This is illustrated in Figure 43. One may need to weaken thresholds of significance to increase the number of overlapping elements.

Fourth, it is easiest to reason across genes and proteins when all of these elements across the intersection can be represented using the same identification schema. An ideal common identifier for measured elements is comprehensive and open. For example, as shown in Figure 44, translating microarray identifiers into LocusLink identifiers can be done for a number of species. However, if one wishes to compare findings from *S. cerevisiae* with *M. musculus*, one needs to use an alternative identifier, since genes from *S. cerevisiae* are not covered by LocusLink identifiers. Whereas relating two open identifiers is still practicable, relating proprietary identifiers is obviously not.

Fifth, and perhaps hardest to solve, the intersection of some near-comprehensive data sets may not be possible without the application of *a priori* knowledge to transduce one modality to another. An example is shown in Figure 45. On first pass, the intersection of a near-comprehensive gene expression data set and a near-comprehensive measurement of metabolites on the same samples would seem to be impossible.

However, some of the genes code for proteins that act on the metabolites. With the right knowledge, the intersection becomes possible; however, with differing knowledge, the intersection will change.

10. Summary and future directions

Summary

In chapter 3, I demonstrated an automated system called GENOTEXT, which has successfully generated mappings to the Unified Medical Language System (UMLS) for every microarray sample stored in the Gene Expression Omnibus. Every sample can be directly or indirectly modeled using UMLS concepts. The results of this project suggest that UMLS, even in its current state, is sufficient to represent a number of the concepts held in the text-based annotations of genome-scale data. As an application of this, in chapter 5, I described an automated system, called CONTRAVERSE, that can determine a relevant cell type for a large number of samples in GEO, and a relevant disease for a smaller number of samples.

In chapter 6, I showed how all data sets from the Gene Expression Omnibus, including gene expression measurements, experimental variables, and values, can be loaded into a single database. I created an automated system called COMPRARE that comprehensively makes every possible two-group comparison in each data set, and records the genes whose expression is significantly different in each comparison. I showed how the experimental variables themselves, such as “age”, can be studied, and found important genes are repeatedly implicated across multiple comparisons.

With the experimental contexts, the catalogs of measured elements and the content of data elements modeled from the Gene Expression Omnibus, I then showed in chapter 7 how data sets can be found and sorted based on relevance given a query concept, and how genes known to be related to the concept can be found in the intersection. Finally, in chapter 8, I showed how a model knowledge base could be used to integrate genomic data, in the form of gene expression profiles, with genetic data, in the form of quantitative trait loci (QTL). The integrated data could then be used to explain the trait of a quantitative trait locus.

What I accomplished that could not be done before

Several aspects of this work are novel. First, the creation of an automated system that can represent the experimental context of genomic experiments using the Unified Medical Language System is new. Though I have shown how errors in mapping to concepts can occur, the results suggest that UMLS, even in its current state, is sufficient to represent a number of the concepts held in the text-based annotations of genome-scale data. Related to this, there was no previous automated method to determine cell type and disease from the annotations of genome-scale experimental data before this work. A user of this system can now find genomic data of a particular tissue type, and can cluster samples based on established taxonomies of cell types and diseases.

The mapping of identifiers used in the Gene Expression Omnibus to global gene identifiers, such as LocusLink, is novel. Currently, even the National Center for Biotechnology Information does not offer this translation, and instead provides only a free-text search for genes through the GEO web-site. This mapping allowed me to unify over 50 million gene expression measurements with LocusLink. A user of this mapping can now discover the expression pattern of individual genes over thousands of samples.

The creation of an automated system to comprehensively make every possible two-group comparison in each experimental data set is novel. A user of this system can now comprehensively determine under which experimental conditions a particular gene is differentially expressed, resulting in hypotheses between the gene and these conditions.

The traversal of UMLS to find related concepts is not novel, but to use traversal to sort genomic data sets based on relevance to a particular concept is new. A user of this system can now start with a disease concept, even identified using an ICD-9 disease code, and find genes differentially expressed in samples related to that disease.

Finally, the use of a knowledge-base to represent biomedical assertions is not new. I implemented assertions allowing the joining of two measurement modalities (gene expression profiling and quantitative trait loci) and a model set of assertions linking genes to a disease. The building of a knowledge-base with the ability to reason on both genes (specified by LocusLink and Homologene) and biomedical concepts (specified by the Unified Medical Language System) is novel, and its use for the automated explanation of a genetic finding is novel. A user can now query this system with additional QTLs from another species and can find a causal chain of plausible biomedical events leading to the trait being measured. With additional knowledge, this system can expand beyond its model domain of diabetes mellitus.

Near-term questions that can now be addressed

There are several important questions that this system can be used to answer with slight or no additions. First, what transcription factors are most often implicated in experiments annotated with each UMLS concept? What receptors? Genes from these two categories that are implicated repeatedly in experiments annotated with a UMLS concept may be more amenable to biological testing and validation. Computationally, a

Bayesian networks could be constructed to relate transcription factors and receptors with genes that are consistently co-implicated with them. This kind of probabilistic co-implication network linking genes, ortholog families, and concepts from experimental context, could be mined to find strong, yet currently undiscovered, relations.

What are the characteristics of those genes that are consistently unimplicated across multiple comparisons and experiments? Are these genes what are typically considered as "housekeeping genes?" Do these genes have fewer predicted transcription factor binding sites in their promoter regions, or fewer single nucleotide polymorphisms in their promoter regions? The property of being invariant can and should be studied with this system.

Over 140 GEO experiments have used time as an experimental variable. Currently, the values for time used in these experiments are qualitative and represented by free-text. If we manually translated these into standard quantitative units (e.g. seconds), we might be able to take advantage of these multiple time-series experiments using signal processing methods that need fixed-time intervals, or modeling these processes using Markov chains or other probabilistic methods. [122,148,168,169] We might be able to find novel and testable transcription factor-gene relations from this modeling.

Can we pull together diseases that are not otherwise thought of as being related, based on genomics? Related to this, can we comprehensively assess the similarity of animal models for human diseases, based on genomics? Similarity can be defined in at least two ways: based on similar patterns of gene expression, or based on similar genes being implicated in an experiment. Relating diseases for which little is currently known about etiology to animal or cellular models could be fruitful in the determination of causal mechanisms, and could be a useful outcome of genomic medicine.

We are increasingly using the results of genomic data to help predict genes that may have single nucleotide polymorphisms or mutations that are significantly associated with a disease trait. Is this a valid assumption? How often is it the case that genes that are differentially expressed are the ones that are actually involved in disease? Moreoever, how often is it the case that genes are differentially expressed in the specific diseases they are associated with through mutations or polymorphisms? By joining data in this system to genes known to be associated with diseases, found in sources such as the Online Mendelian Inheritance in Man, we can now answer these questions. [130]

Trisomy 21 is caused by an extra chromosome, often as a result of nondisjunction in parent gametes. What UMLS concepts (including cell and tissue types) are most associated with experiments in which genes on chromosome 21 are differentially expressed. Are these concepts related to the known clinical problems seen in trisomy 21? Can we make predictions about other genetic diseases in this way? Moreover, can we operate in reverse: given the known clinical problems seen in a genetic disease, could we find experiments with annotations related to these clinical problems, then find genes differentially expressed across these experiments? Novel hypotheses regarding gene mutations causative of genetic diseases could be established and tested.

Similarly, we could start with a gene and determine which tissue types are most likely to be associated with experiments in which that gene is differentially expressed. We could test whether we can predict the tissues with clinical problems for complex genetic diseases, such as the Hutchinson-Gilford Progeria Syndrome thought to be caused by a mutation in Lamin A.

A tissue sample with a mixture of cell types is clearly not the same as a more homogeneous type of tissue. This is especially true in the study of pathophysiological and degenerative processes. For example, even a small sample from a plaque from a brain with multiple sclerosis is considered very differently in the bioinformatics analysis, because the entire sample depends on how many invading white blood cells are present within the sample. Without knowing that many cell types are present within the sample, this mix of cell types is otherwise thought of as increasing the noise in this sample, and that noise needs to be measured and compensated for. We could now use this system to find genes that uniquely define a tissue, as a potential marker. We could then develop an application that uses these genes as components in a mixture model, to computationally determine the cellular makeup of unknown samples.

Finally, we could study how gene expression relates to metabolic pathways. How are programs, such as glycolysis or gluconeogenesis, activated? Is it the case that multiple enzymes are upregulated when a pathway is activated, or just a few enzymes acting as control points? The data in this system could be related to KEGG or other established repositories of metabolic pathways to determine patterns of activation.

In addition to questions that can now be answered, several applications can now be written to take advantage of this work. First, automated assignment of GeneOntology categories to genes is sequence dependent; rules can be written to assign categories to genes based on the presence of particular protein domains. [170] Using my system, we can propose a novel method for automated assignment of GeneOntology categories. If a gene is repeatedly found to be differentially expressed in comparisons involving experiments annotated with a concept mapping to a GeneOntology category, then we could propose a gene to category relation.

The "genomic profile" of a hospital could now be studied. An application could be written that takes the ICD-9 discharge diagnosis codes from a health care center and finds genes significantly different in comparisons involving data sets related to those diagnosis codes. Such a list of genes might be useful some day, if a health care center wishes to develop local gene-expression tests to address diagnostic questions.

Currently, most statistical relations between concepts in UMLS are determined by co-occurrence in MEDLINE records. However, this system could now be used to provide a novel source of co-occurrence: if concepts share a significant number of genes (i.e. concepts annotate experiments in which comparisons implicate a similar set of genes), those concepts could now be related to each other.

There are several existing software applications that attempt to find categories of genes that are statistically overrepresented given an input list of genes. [65,171,172] However, the typical output from these programs is an ordered list of GeneOntology categories that might best explain the gene list, not a testable hypothesis. We can use the set of implicated genes from all possible comparisons of GEO data sets to perform a similar function. An application could be designed to take an input list of genes and find the comparison that best matches the input list. The advantage of such an application, compared to previous work, is that the output is an actual description of an experiment and the specific experimental variables that led to the matching list of genes. These experimental details could more directly lead to a testable hypothesis.

Interface needed for biologists

Raw database queries to implement and answer these questions will not be sufficient for the majority of biologists and physician scientists. At least two database-backed web-sites could be developed allowing these scientists to immediately use some of the results of the work presented in this dissertation.

First, a web-site needs to be created that takes as input a text string, which gets mapped to UMLS concepts. After the user has indicated which of these is appropriate, the web-site would use the CONTRAVERSE system to find related GEO data sets. These would be displayed as hyperlinks back to GEO. The user could then select one or more of these data sets to then perform an intersection, resulting in genes significantly different in all of these data sets. Subsequent web-pages would allow the specific gene expression profile within an experiment to be automatically drawn as a bar-graph.

A second web-site could be developed that takes a list of genes as input. This list would represent those genes implicated in a user's experiment, determined using existing bioinformatics techniques. After receiving the list, the web-site would search the comprehensive group comparisons within the GEO data set generated by the COMPRARE system, finding the comparison whose list of significant genes matches the users list most closely (using a nearest-neighbor algorithm), taking into account cross-species orthologs if needed. The output from the web-site would be a list of the matching genes as well as the specific experimental variables and values that resulted in that list.

Integrating existing knowledge-bases and scalability to newer genomic modalities

In chapter 8, I showed how a model knowledge base could be used to integrate genomic data, in the form of gene expression profiles, with genetic data, in the form of QTLs. Though a gene-specific knowledge-base had to be created allowing genes to be related to diseases, the actual integration was only possible through the creation of a cross-modality knowledge-base allowing automated reasoning over how a gene finding in a genetic experiment can be resolved with a gene finding in a genomic experiment.

Several groups and companies are addressing the creation of gene-specific knowledge bases which could be incorporated into this system. Proteome, now part of Incyte Corporation, created a set of BioKnowledge Library Databases through manual curation. The annotations of genes in the human, mouse and rat knowledge bases include patterns of expression in tissue, cell, and tumors, known associated diseases, and mutant phenotypes. For worm and yeast, even more annotations are available, including known protein-protein interactions, complex formation, genetic interactions, known inducers and repressors, protein modifications, as well as free-text statements on function and activity. All of this data is otherwise not available through LocusLink. Proteome data can be incorporated in two ways. First, it can be used as a gold-standard for testing algorithms trying to predict gene-tissue relations. Second, relations on tissue expression and disease associations can be represented in first-order predicate calculus and directly translated Prolog assertions, for use in integromic queries. With improved text parsing technology, the free-text statements could be represented as well. Proteome currently provides information over 734,000 annotations on over 70,000 genes in human, mouse, rat, yeast and worm.

Ingenuity Systems has created a similar knowledge base, but with many more annotations for human, mouse and rat genes. Assertions for a gene include its protein interactions, other genes that regulate and are regulated by this gene, its role in cellular processes, organismal processes, and disease, and its functional roles. For example, the gene coding for the human insulin receptor is annotated as phosphorylating 44 specific named proteins, activating 18 proteins, altering the expression level of 17 genes, and much more. Though Ingenuity provides much information that is not otherwise available, they currently provide no method for automated reasoning across this knowledge base. Ingenuity uses its own private ontology of over 280,000 concepts. Globally relating all of these concepts to UMLS will be difficult; instead, unifying small portions of the ontology relevant to the study of a disease (such as cellular processes used to describe genes known to be involved in diabetes mellitus) might be quickly fruitful.

Ariadne Genomics sells a knowledge base of biological pathways called ResNet. Assertions in this knowledge base were derived from natural language processing techniques applied to abstracts in MEDLINE. Their knowledge base contains over 200,000 assertions of gene regulation, interaction and modification, including over 16,000 assertions on gene expression. Modeled components include proteins, enzymes, functional classes, complexes, cellular processes, small molecules, and treatments. Assertions of one component's action on another are drawn from a taxonomy of at least 10 types of activities, from binding and regulation to enzymatic activity. With a smaller taxonomy than that used by Ingenuity, it might be possible to globally unify the Ariadne taxonomy with UMLS.

However, integrating across experimental and contextual modalities requires more than just a gene-specific knowledge base. Additional work will need to be spent in modeling the integration of additional modalities. The current cross-modality knowledge base has just one way a gene can be in a quantitative trait locus and be differentially expressed. If we now want to integrate proteomic measurements, we will have to model exactly how protein levels relate to (impact or be impacted by) QTLs or gene expression levels.

Connecting n modalities could require as little as $n-1$ relations or as many as n^2 relations. Ideally, additional modalities could initially be incorporated by directly linking assertions about measurements in that modality to one or two existing modalities. For example, changes in proteomic measurements can be immediately incorporated as being causally downstream of expression measurements, as a rough approximation. With additional effort, more relations can then be added; to continue the example, changes in the proteomic measurements of transcription factors can be causally downstream of expression measurements.

Though it may not change as frequently as gene-specific knowledge, cross-modality knowledge is not static, as occasionally additional ways are found in how genes, proteins, enzymes, and substrates can interact. Any such finding immediately implicates how measurements of these elements can relate to each other.

Immediate next steps

The list of genes differentially expressed in comparisons involving age needs to be pursued. This list will be compared to other data sets derived from models of aging, such as replicative senescence and Hutchinson-Gilford Progeria Syndrome, that are not yet present in GEO. Computationally, we will test to determine whether genes in this list are preferentially on the outer portions of chromosomes, near telomeres, which are known to shorten during aging.

Existing pathways related to insulin signaling and glucose metabolism from public sources will be incorporated so that additional QTLs can be explained. Additional cross-modalitity knowledge will be added allowing incorporation of proteomic and polymorphism data sets, especially in the domain of diabetes mellitus. By definition, however, those QTLs that defy explanation might be the most interesting, as they potentially reflect a novel causal pathway from a change in DNA to a trait.

Even if the integrome becomes fully explored, with models allowing integration between all experimental modalities, the human intuition that goes into finding a novel causal pathway will be difficult to model. However, just as automated DNA sequencers provided access to information that now enables questions resulting in the blooming of the field of genomics, so too do I believe that the integromic framework provides a new platform on which future researchers can ask questions that we cannot even conceive of today.

Table 1: Over 30 nearly-comprehensive measurement or experimental modalities are available for experiments in molecular biology. While analyzing data from a single modality is becoming routine, analyzing data across multiple modalities or contexts remains a challenge.

Nearly-comprehensive measured data specific to an experimental context

a) DNA sequencing: includes map of sequence tagged sites; over 600 bacterial, 600 eukaryotic, and 1500 virus species underway or completed.[173]

b) DNA genetic distance: Recombination maps have been published for human, mouse, and other species. [174-176]

c) DNA polymorphisms: microsatellite repeats [177], single nucleotide polymorphisms [178,179], and haplotypes [180]

d) RNA sequencing: Expressed sequence tags [181,182], cDNA sequencing [183,184], alternative splicing [185,186]

e) RNA absolute expression: absolute expression measurements theoretically comparable across genes [133,187]; over 690 experiments with over 11,500 microarrays publicly available [38,130]

f) RNA relative expression: relative expression measurements theoretically comparable across genes [188]

g) Protein identification: two-dimensional gel electrophoresis [189], tandem mass spectrometry [190]

h) Protein absolute quantitation: tagging each gene, then Western blotting [191], or microarray [192,193]

i) Protein relative quantitation: using isotope-coded affinity tags [194]

j) Protein activity: change in state [195], or across activities to determine gene function [196]

k) Protein relative activity: interaction with small molecules [197]

l) Protein interactions with DNA: chromatin immunoprecipitation of labeled proteins binding to promoter regions [198-202], binding to CpG microarrays [201,203], oligonucleotide microarrays [204], or tiling path microarrays [205]

m) Protein interactions with proteins: Yeast two-hybrid [206-210] or mass spectrometry of complexes [211,212]

n) Protein interactions with small molecules: effect on protein interaction with small molecules [213]

o) Protein interactions with carbohydrates [214,215]

p) Protein localization: genome wide HA-tagging [89] or GFP-tagged fusion proteins [216]

q) Protein modification: tyrosine phosphorylation using Western blots [217] or mass spectrometry [218]

r) Metabolites: HPLC [219] or NMR [220]

s) Multi-phenotype characterization: after effects of RNAi [221]

Nearly-comprehensive computed data specific to an experimental context

a) DNA quantitative trait loci: genome regions statistically associated with quantitative traits [222]

b) DNA predicted transcription factor binding sites [223]

c) RNA co-expression: determined using correlation coefficients [49,80]

Nearly-comprehensive applied contexts

a) Gene disruption: targeted [224] or random insertional [89] or chemical [225]
b) Gene double disruption: mutations with deletions [226]
c) Gene activation: random [227]
d) RNA transient knockdown using RNAi or siRNA [63,228,229]
e) Multiple species [230]
f) Tissue microarray [231-233]
g) Presentations of a diagnosis [6,234-236]
h) Sampling across microscopic field of view, or spatial position [237]
i) Extracellular environment [238]: in the presence of an array of carbohydrates [239], in the presence of pharmacological agents [82,240], in the presence of multiple ligands [111]
j) Sampling across time: lifespan [241], or circadian [242-245]

Table 2: Vocabularies excluded from our Metathesaurus subset.

Source vocabularies
Alternative Billing Concepts
Beth Israel Vocabulary, 1.0
Descritores em Ciencias da Saude (Portuguese translation of the Medical Subject Headings), 2003
Descritores en Ciencias de la Salud (Spanish translation of the Medical Subject Headings), 2003
Canonical Clinical Problem Statement System, 1999
Current Dental Terminology (CDT), 4
Medical Entities Dictionary, 2003
Physicians' Current Procedural Terminology, 2003
Physicians' Current Procedural Terminology, Spanish Translation, 2001
German translation of the Medical Subject Headings, 2003
German translation of ICD10, 1995
German translation of UMDNS, 1996
DSM-III-R, 1987
Nederlandse vertaling van Mesh (Dutch translation of MeSH), 2003
Finnish translations of the Medical Subject Headings, 2003
HCPCS Version of Current Dental Terminology (CDT), 4
HCPCS Version of Current Procedural Terminology (CPT), 2003
Home Health Care Classification, 2003
ICPC2E-ICD10 relationships from Dr. Henk Lamberts, 1998
ICD10, American English Equivalents, 1998
International Statistical Classification of Diseases and Related Health Problems, Australian Modification, Americanized English Equivalents, 2000
International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Australian Modification, January 2000 Release
International Classification of Primary Care 2nd Edition, Electronic, 2E, 1998
International Classification of Primary Care, Version 2-Plus, 2000
ICPC, Basque Translation, 1993
ICPC, Danish Translation, 1993
ICPC, Dutch Translation, 1993
ICPC, Finnish Translation, 1993
ICPC, French Translation, 1993
ICPC, German Translation, 1993
ICPC, Hebrew Translation, 1993
ICPC, Hungarian Translation, 1993
ICPC, Italian Translation, 1993
ICPC, Norwegian Translation, 1993
International Classification of Primary Care, Version 2-Plus, Americanized English Equivalents, 2000
ICPC, Portuguese Translation, 1993
ICPC, Spanish Translation, 1993

Source vocabularies
ICPC, Swedish Translation, 1993
Thesaurus Biomedical Francais/Anglais [French translation of MeSH, 2003
Italian translation of Medical Subject Headings, 2003
Online Congenital Multiple Anomaly/Mental Retardation Syndromes, 1999
Master Drug Data Base, 2003_03
Medical Dictionary for Regulatory Activities Terminology (MedDRA), 6.0
Medical Dictionary for Regulatory Activities Terminology (MedDRA), American English Equivalents, 6.0
Medical Dictionary for Regulatory Activities Terminology (MedDRA), American English Equivalents with expanded abbreviations, 6.0
Medical Dictionary for Regulatory Activities Terminology (MedDRA), with expanded abbreviations, 6.0
Medical Dictionary for Regulatory Activities Terminology (MedDRA), 6.0, Portuguese Edition
Medical Dictionary for Regulatory Activities Terminology (MedDRA), 6.0, Spanish Edition revised
Online Mendelian Inheritance in Man, 1993
Multum MediSource Lexicon, 2003_03
Micromedex DRUGDEX, 2001-08
Metathesaurus CPT Hierarchical Terms, 2003
Classification of Nursing Diagnoses, 1999
First DataBank National Drug Data File, 2001-07
Nursing Interventions Classification, 1999
Nursing Outcomes Classification, 1997
Omaha System, 1994
Patient Care Data Set, 1997
Pharmacy Practice Activity Classification, 1998
Thesaurus of Psychological Index Terms, 2001
Russian Translation of MeSH, 2003
UltraSTAR, 1993
UMDNS: product category thesaurus, 2003
WHO Adverse Reaction Terminology, 1997
WHOART, French Translation, 1997
WHOART, German Translation, 1997
WHOART, Portuguese Translation, 1997
WHOART, Spanish Translation, 1997

Table 3: Seven free-text annotations extracted from the Gene Expression Omnibus containing contextual or experimental information regarding a sample

Annotation
GEO sample title
GEO sample description
GEO sample source
GEO sample keyword
GEO sample series
GEO series title
GEO series description
GEO data set title

Table 4: Titles of six data sets used to determine whether final-candidate mappings would be sufficient to represent the necessary concepts, or whether all mapped concepts were needed.

GEO data set	Title
1	Testis gene expression profile
2	Melanoma, cutaneous malignant, classification
3	Cerebellar development time course
4	Tissue-specific and development-regulated genes in maize
5	Diurnal and circadian-regulated genes (I)
6	Germline development and function

Table 5: UMLS strings mapped to the titles listed in Table 4. Those strings that were additionally deemed as final-candidates are listed in the fourth column. There was no difference between candidates and final-candidates for GEO data set 5.

GEO data set	String	Candidate SUI	Final Candidate SUI
1	Testis	S1044598	S1044598
1	TESTIS	S1790542	S1790542
1	Gene Expression Profiling	S1684714	
1	Gene Expression	S0044005	
1	Gene	S0043986	
1	Expression, NOS	S0298422	
2	Melanoma	S0060982	S0060982
2	Cutaneous	S0365139	S0365139
2	malignant	S1556770	
2	Malignant	S1466658	
2	Classification	S0007165	
2	classification	S0007166	
2	classification	S1465738	
3	Cerebellar	S0362852	S0362852
3	Development	S1802517	
3	development	S1802518	
3	Time course	S1046322	
3	Time	S0093786	
3	Course	S0849879	
4	tissue	S0290001	S0290001
4	Specific	S0324791	
4	Development	S1802517	S1802517
4	development	S1802518	S1802518
4	Regular	S1467339	S1467339
4	regulatory	S1423373	S1423373
4	Regulation	S0081187	S0081187
4	Genes	S0044269	S0044269
4	maize	S2759961	S2759961
4	maize	S2759962	S2759962
6	Germ Line	S0044605	S0044605
6	Development	S1802517	
6	development	S1802518	
6	function	S1086266	S1086266
6	Function	S0889213	S0889213
6	FUNCTION	S1466234	S1466234

Table 6: Number of concepts mapped to each of the seven GEO free-text annotations.

Source of text item	**Number available**	**Number in which any concept was successfully found**	**Percent successful**	**Count of unique concepts assigned**	**Concepts per successful text item**
GSM Title	8,519	5,401	63%	968	0.18
GSM Description	7,123	6,292	88%	2,630	0.42
GSM Source	8,518	8,123	95%	1,282	0.16
GSM Keyword	2,337	2,203	94%	492	0.22
GSE Title	524	450	86%	917	2.04
GSE Description	247	243	98%	2,376	9.78
GDS Title	449	443	99%	832	1.88

Table 7: Correlation between the number of unique concepts elicited and the length of the annotation. The second column indicates the average and standard deviation of the length of each annotation, with the description annotations being the longest. The third column indicates the correlation between the length of each text item and the number of unique concepts mapped to it.

Source of text item	**Average length (characters)**	**Correlation between length of successfully parsing text item and number of unique concepts**
GSM Title	22.5 ± 16.3	0.552
GSM Description	790.3 ± 1745.0	0.969
GSM Source	25.7 ± 18.8	0.627
GSM Keyword	21.6 ± 24.7	0.793
GSE Title	35.0 ± 20.5	0.633
GSE Description	587.0 ± 773.4	0.971
GDS Title	37.9 ± 11.7	0.488

Table 8: The 4,190 unique concepts assigned to the GEO annotations were drawn from these 123 UMLS semantic types. Column 3 indicates the number of concepts mapped from each semantic type to any GEO annotation. Column 4 indicates the average MetaMap score of the mapping. Column 5 indicates the number of GEO annotations using a concept of each semantic type.

Type	Name of semantic type	Count of unique concepts used	Average score of association	Number of GEO annotations using a concept of this semantic type
T116	Amino Acid, Peptide, or Protein	320	623.11	5147
T121	Pharmacologic Substance	272	614.60	5968
T109	Organic Chemical	260	615.89	7234
T169	Functional Concept	209	632.53	15317
T023	Body Part, Organ, or Organ Component	201	632.64	8145
T123	Biologically Active Substance	184	625.25	7764
T080	Qualitative Concept	172	590.31	13320
T081	Quantitative Concept	168	600.33	13596
T082	Spatial Concept	157	619.10	6997
T047	Disease or Syndrome	148	653.53	2796
T061	Therapeutic or Preventive Procedure	134	630.34	4104
T170	Intellectual Product	132	625.84	11906
T191	Neoplastic Process	116	644.37	3567
T079	Temporal Concept	112	618.54	11872
T033	Finding	105	632.09	4205
T025	Cell	91	668.43	7763
T126	Enzyme	77	627.61	1538
T059	Laboratory Procedure	77	640.08	3625
T073	Manufactured Object	75	614.25	4701
T129	Immunologic Factor	74	618.31	1760
T114	Nucleic Acid, Nucleoside, or Nucleotide	67	642.98	5899
T074	Medical Device	64	636.75	3335
T028	Gene or Genome	56	622.77	1684
T130	Indicator, Reagent, or Diagnostic Aid	53	641.95	2460
T032	Organism Attribute	50	615.40	4098
T046	Pathologic Function	48	650.34	2357
T015	Mammal	46	642.71	6221
T118	Carbohydrate	46	620.63	828
T029	Body Location or Region	44	602.87	480
T196	Element, Ion, or Isotope	44	610.97	2239
T070	Natural Phenomenon or Process	42	618.39	1572
T002	Plant	40	639.39	1078

Type	Name of semantic type	Count of unique concepts used	Average score of association	Number of GEO annotations using a concept of this semantic type
T040	Organism Function	40	627.07	2566
T042	Organ or Tissue Function	40	635.43	1588
T060	Diagnostic Procedure	39	631.85	4226
T041	Mental Process	38	632.36	2673
T168	Food	36	628.13	1400
T024	Tissue	36	652.24	2423
T184	Sign or Symptom	34	663.86	2144
T192	Receptor	34	607.50	578
T083	Geographic Area	34	640.42	2351
T034	Laboratory or Test Result	34	684.80	517
T125	Hormone	34	630.86	447
T119	Lipid	33	646.51	619
T019	Congenital Abnormality	33	654.93	360
T122	Biomedical or Dental Material	33	623.18	1277
T197	Inorganic Chemical	32	624.60	1184
T044	Molecular Function	32	616.59	544
T009	Invertebrate	32	653.09	406
T062	Research Activity	32	637.86	2932
T131	Hazardous or Poisonous Substance	30	596.85	931
T026	Cell Component	28	558.24	439
T007	Bacterium	28	623.53	524
T037	Injury or Poisoning	26	661.66	1426
T045	Genetic Function	25	609.58	2147
T043	Cell Function	25	609.90	752
T058	Health Care Activity	24	642.46	921
T039	Physiologic Function	24	652.13	970
T091	Biomedical Occupation or Discipline	24	657.02	667
T078	Idea or Concept	24	632.83	3463
T054	Social Behavior	23	630.77	475
T031	Body Substance	22	644.48	738
T185	Classification	21	601.26	697
T110	Steroid	20	633.29	288
T057	Occupational Activity	20	593.60	528
T098	Population Group	19	585.50	1919
T055	Individual Behavior	19	663.25	1217
T018	Embryonic Structure	16	637.54	324
T063	Molecular Biology Research Technique	16	628.38	344
T067	Phenomenon or Process	16	642.12	853
T020	Acquired Abnormality	16	636.44	230
T092	Organization	15	622.34	835
T048	Mental or Behavioral	15	672.16	171

Type	Name of semantic type	Count of unique concepts used	Average score of association	Number of GEO annotations using a concept of this semantic type
	Dysfunction			
T005	Virus	15	604.78	201
T167	Substance	14	596.22	926
T056	Daily or Recreational Activity	13	597.84	448
T093	Health Care Related Organization	13	603.26	315
T004	Fungus	12	590.47	489
T090	Occupation or Discipline	12	664.07	541
T124	Neuroreactive Substance or Biogenic Amine	12	613.56	289
T099	Family Group	12	585.97	343
T100	Age Group	11	621.21	443
T104	Chemical Viewed Structurally	10	640.39	398
T022	Body System	9	611.33	811
T111	Eicosanoid	9	601.78	216
T195	Antibiotic	9	657.43	271
T049	Cell or Molecular Dysfunction	9	638.79	887
T075	Research Device	8	654.49	492
T038	Biologic Function	8	673.48	326
T013	Fish	8	598.91	685
T120	Chemical Viewed Functionally	7	635.83	326
T068	Human-caused Phenomenon or Process	7	654.04	120
T115	Organophosphorus Compound	6	627.62	59
T097	Professional or Occupational Group	6	562.78	31
T030	Body Space or Junction	6	654.36	24
T127	Vitamin	6	640.64	142
T017	Anatomical Structure	6	617.21	239
T052	Activity	6	650.23	444
T201	Clinical Attribute	6	577.82	275
T012	Bird	5	666.14	38
T001	Organism	5	597.11	528
T086	Nucleotide Sequence	5	682.20	5
T103	Chemical	4	621.43	160
T064	Governmental or Regulatory Activity	4	655.69	49
T200	Clinical Drug	4	550.66	88
T089	Regulation or Law	3	557.82	71
T008	Animal	3	632.68	287
T190	Anatomical Abnormality	3	651.46	114

Type	Name of semantic type	Count of unique concepts used	Average score of association	Number of GEO annotations using a concept of this semantic type
T065	Educational Activity	3	640.01	119
T102	Group Attribute	3	687.48	23
T021	Fully Formed Anatomical Structure	2	646.58	128
T096	Group	2	633.27	296
T069	Environmental Effect of Humans	2	645.27	64
T066	Machine Activity	2	606.31	51
T016	Human	2	552.14	1780
T101	Patient or Disabled Group	2	565.83	315
T053	Behavior	1	645.69	39
T050	Experimental Model of Disease	1	691.00	1
T051	Event	1	678.50	2
T077	Conceptual Entity	1	517.00	17
T095	Self-help or Relief Organization	1	616.43	24
T014	Reptile	1	687.41	34
T010	Vertebrate	1	573.33	3

Table 9: List of UMLS semantic types from which no concepts were mapped to GEO annotations. Column 3 indicates the number of concepts within each semantic type that exist in UMLS.

Type	Name of semantic type	Count of concepts in UMLS
T003	Alga	3301
T011	Amphibian	1550
T203	Drug Delivery Device	1455
T194	Archaeon	824
T171	Language	714
T006	Rickettsia or Chlamydia	502
T095	Self-help or Relief Organization	39
T072	Physical Object	37
T087	Amino Acid Sequence	28
T094	Professional Society	17
T085	Molecular Sequence	7
T071	Entity	5
T088	Carbohydrate Sequence	4

Table 10: The top 50 UMLS concepts mapped to GEO annotations. Column 3 indicates the number of unique GEO annotations for which the concept was mapped.

Concept	Concept Name	Count
C0007634	Cells	3275
C0035668	RNA	2591
C0042153	Utilization	2383
C0025914	House mice	1738
C0025929	Laboratory mice	1738
C0441621	Sampling - Surgical action	1657
C0870078	Sampling	1657
C0332307	With type	1559
C0439810	Total	1403
C0445392	Wild	1388
C0037585	Computer software	1307
C1167624	Labeling	1231
C0439227	Hour	1190
C0439228	Day	1171
C0205397	Seen	1126
C0042789	Vision	1124
C0205173	Duplicate	1002
C0333052	Version, NOS	986
C0439232	Minute of time	953
C0441633	Scanning	946
C0243148	control	918
C0086418	Homo sapiens	903
C0020114	Human	877
C0439242	mL	868
C0337051	Pool, NOS	849
C0080194	Muscle strain	848
C0040300	Tissues	825
C0020202	Hybridization, Genetic	802
C0449945	Strain typing	793
C0681814	experiment	793
C0017337	Genes	790
C0026809	Mus	786
C0596988	mutant	786
C0205307	Normal	743
C0205409	Isolated	733
C0001779	Age, NOS	731

Concept	Concept Name	Count
C0026845	Muscle	728
C0596981	Muscle Cells	727
C0185117	Expression, NOS	722
C0024554	Male gender	717
C0683312	categories	678
C0040223	Time	629
C0002778	Analysis of substances	610
C0936012	Analysis	610
C0004561	B-Lymphocytes	601
C0443050	Robinson	598
C0010453	Anthropological Culture	593
C0220814	Cultural	593
C0430400	Laboratory culture	593
C0009253	Coitus	587

Table 11: Mappings between these ten concepts and GEO samples, series, and data sets were ignored because of the lack of specificity in these concepts.

CUI	Concept
C0007600	Cell Line
C0007634	Cells
C0007635	Cells, Cultured
C0009013	Clone Cells
C0012634	Disease
C0021311	Infection
C0027651	Neoplasms
C0040300	Tissues
C0449475	Type of cell
C0682516	Cultured Cell Line

Table 12: Examples of data in GEO with cell type successfully represented.

GEO Object	**Annotation**	**Mapped to**
GSM 4843	Source “K562 erythroleukemia cells”	K562 Cells (C0600432)
GSM 3509	Source “Primary Acute Lymphoblastic Leukemia Cells”	Lymphoblast (C0229613)
GSM 8724	Source “HeLa CD4+ cells”	Hela Cells (C0018873)
GSM 8893	Title “Cx43 KO cortical astrocytes 1”	Astrocytes (C0004112)
GSM 8530	Description “Uninfected Vero cell culture control”	Vero Cells (C0042542)
GSM 8509	Keyword “neutrophil”	Neutrophils (C0027950)
GSM 820	Source “NIH3T3 fibroblasts treated with E2F1 expressing adenovirus”	Fibroblasts (C0016030)
GSE 13	Title “Murine bone marrow B cell precursors”	Bone Marrow Cells (C0005955)
GSE 609	Title “SCID vs Normal Thymocyte Comparisons”	Thymocyte (C0814999)
GDS 45	Title “Cochlear hair cell line differentiation time course (Mu11K-A)”	Hair Cells (C0018496)

Table 13: Examples of data in GEO with disease successfully represented.

GEO Object	Annotation	Mapped to
GDS 157	Title "Type 2 diabetes and insulin resistance (HuGeneFL)"	Insulin Resistance (C0021655)
GDS 167	Title "Autoimmune disease mechanisms"	Autoimmune Diseases (C0004364)
GDS 22	Title "Parkinson's Disease model"	Parkinson Disease (C0030567)
GDS 238	Title "Skin tumors and vitamin A supplements"	Skin Neoplasms (C0037286)
GDS 252	Title "Lung hypertension recovery (U74Av2)"	Pulmonary Hypertension (C0020542)
GDS 26	Title "Copper regulon in S. Cerevisiae"	Hypocupremia, NOS (C0268070)
GDS 274	Title "Hepatocellular carcinoma metastasis"	Neoplasm Metastasis (C0027627)
GDS 351	Title "Pulmonary fibrosis model (129/SV, bleomycin sensitive)"	Pulmonary Fibrosis (C0034069)
GDS 386	Title "Arthritis synoviocyte response to TNF alpha"	Arthritis (C0003864)
GDS 76	Title "Macrophages infected with Salmonella (SHZ)"	Salmonella infections (C0036117)
GSE 415	Title "Mouse models of cardiac remodeling"	Myocardial Infarction (C0027051)
GSE 443	Title "Leprosy lesion gene expression"	Leprosy (C0023343)
GSE 445	Title "Alpha Thalassaemia Myelodysplasia Syndrome (ATMDS)"	Thalassemia (C0039730)
GSE 465	Title "Expression profiling in the muscular dystrophies"	Muscular Dystrophies (C0026850)
GSE 480	Title "Sleep apnea and glucose metabolism"	Sleep apnea and glucose metabolism (GSE480)
GSE 485	Title "Genetic basis of sensitivity to pulmonary fibrosis"	Pulmonary Fibrosis (C0034069)
GSE 493	Title "Gene expression profiling in DQA1*0501+ children with untreated dermatomyositis"	Dermatomyositis (C0011633)
GSE 495	Title "Hyperoxic lung injury"	Hyperoxia (C0242706)
GSE 505	Title "Serial analysis of gene expression in the corneal endothelium of Fuchs' dystrophy"	Fuchs' Endothelial Dystrophy (C0016781)
GSE 513	Title "Cynomolgus monkey testicular cDNAs for discovery of novel human genes"	Testicular dysfunction (C0405581)
GSE 609	Title "SCID vs Normal Thymocyte"	Severe Combined Immunodeficiency (C0085110)
GSE 620	Description "Most individuals with cystic fibrosis (CF) carry..."	Cystic Fibrosis (C0010674)
GSE 621	Title "Interstitial cystitis and antiproliferative factor treatment"	Cystitis (C0010692)
GSE 675	Title "Time course analysis of response to HCMV infection"	Cytomegalovirus Infections (C0010823)
GSE 768	Title "Neural stem and neuroblastoma cells"	Neuroblastoma (C0027819)
GSE 77	Title "Exercised Induced Hypertrophy"	Cardiomegaly (C0018800)
GSE 828	Title "Genes/pathways underlying lipoprotein homeostasis"	Hyperlipidemia (C0020473)
GSE 89	Title "Bladder tumour stage classification"	Bladder Neoplasms (C0005695)
GSM 2135	Title "GSM2135: RPMI-8226_CL7010__LEUKEMIA"	Leukemia (C0023418)
GSM 2206	Description "A catalytic antioxidant ... attenuates expression of inflammatory genes in stroke"	Cerebrovascular accident (C0038454)
GSM 941	Title "Heat Shock 000 minutes"	Shock (C0036974)

Table 14: Representative examples of 70 identifier-spaces of gene identifiers and regular expressions that accurately match these identifier-spaces.

Identifier-space	Regular expression matching identifiers

RIKEN clone	^[0-9GIKLEFfBCAD][0-9CE][0-9BCD][0-9][0-9][0-9][0-9][A-Z][0-9][0-9]
	^[0-9]{2}B[0-9]{6}[A-Z][0-9]{2}
	^[0-9]{2}B[0-9]{5}[A-Z]{2}[0-9]{2}

University of Iowa clone	^UI-.*
Max Planck Institut fuer Molekulare Genetik clone	^[AB]9[A-Z][0-9]{2}[A-Z][0-9]{2}
Columbia University clone	^Hy18-.* ^b4HB3M.* ^N3H.*
Centre d'Immunologie INSERM/CNRS	^MTA.[A-Z][0-9]{2}\.[0-9]{3}

Table 15: The most frequently identified genes in LocusLink. Column 3 indicates the number of unique identifiers for each gene, across all identifier-spaces.

LocusID	Gene name	Count of unique identifiers
71	actin, gamma 1	93745
1915	eukaryotic translation elongation factor 1 alpha 1	93366
2597	glyceraldehyde-3-phosphate dehydrogenase	58313
13627	eukaryotic translation elongation factor 1 alpha 1	49901
60	actin, beta	40729
6187	ribosomal protein S2	37499
6122	ribosomal protein L3	34714
15135	hemoglobin Y, beta-like embryonic chain	32650
15129	hemoglobin, beta adult major chain	32650
15130	hemoglobin, beta adult minor chain	32630
18367	olfactory receptor 66	32620
2512	ferritin, light polypeptide	31911
6175	ribosomal protein, large, P0	31285
15481	heat shock protein 8	29751
1937	eukaryotic translation elongation factor 1 gamma	28936
11576	alpha fetoprotein	28711
11657	albumin 1	28701
280662	afamin	28682
2023	enolase 1, (alpha)	28473
2495	ferritin, heavy polypeptide 1	28359

Table 16: Some of the identifiers for the gene N-acetyltransferase 2. The first column indicates the actual text string serving as the identifier, while the second column indicates the identifier-space in which each identifier was classified. To simplify retrieval, this table was stored in a denormalized manner. Strings in the third column, when present, indicate how the particular identifier was mapped to the particular LocusLink gene. For example, the text "Hs.2<-BX095770" for the identifier "IMAGE:1870937" indicates information was present allowing mapping of that identifier to "BX095770", which maps to "Hs.2", which maps to this gene.

Identifier	Identifier-Space	Translate
AI262683	GenBank	Hs.2
NM_000015	GenBank	Hs.2
X14672	GenBank	
Hs.2	UniGene	
AAA59905	Protein	
NP_000006	Protein	
P11245	Protein	
D90042_at	Affymetrix hu6000_merged	Hs.2<-D90042
38912_at	Affymetrix u95_a	Hs.2<-D90042
38912_at	Affymetrix u95v2_a	Hs.2<-D90042
206797_at	Affymetrix u133_a	Hs.2<-NM_000015
D90042_at	Affymetrix hu6000_merged	D90042
10	LocusLink	
NAT2	LocusLink official symbols	
NAT2	LocusLink all symbols	
AAC2	LocusLink all symbols	
IMAGE:1870937	IMAGE clone	Hs.2<-AI262683
1870937	IMAGE clone	Hs.2<-AI262683
IMAGp998I184581	IMAGE clone	Hs.2<-BX095770
IMAGE:1870937	IMAGE clone	Hs.2<-BX095770
UI-H-FG1-bgl-g-02-0-UI	University of Iowa clone	Hs.2<-BU624903
IMAGp998I184581_, _IMAGE:1870937	Institute of Molecular Biology and Genetics Ukraine clone	Hs.2<-BX095770
10286060	GenBank GI	Hs.2<-AV684197

Table 17: Variables in use in GEO data sets. The second column indicates the number of data sets in GEO using each variable. The third column indicates the number of unique qualitative values given for each variable.

Variable	Data sets using this variable	Unique values
time	146	171
strain	73	157
treatment	66	122
disease	50	105
tissue	45	158
agent	42	79
age	24	37
cell type	23	50
cell line	14	72
development stage	14	47
infection	14	20
growth medium	10	22
dose	8	29
specimen	8	35
error	6	46
metabolism	5	2
gender	4	4
stress	3	8
temperature	3	8
unclassified	3	10
shock	2	3

Table 18: GEO data sets manually determined to be directly related to the study of diabetes mellitus, or result in a query on the GEO web-site with the term “diabetes.”

GEO Data Set	Description	Results when queried in GEO	Manually determined true positive
GDS10	Analysis of the NOD model of type 1 diabetes	X	X
GDS157	Gene expression involved in susceptibility for type 2 diabetes	X	X
GDS158	Gene expression involved in susceptibility for type 2 diabetes	X	X
GDS160	Gene expression involved in susceptibility for type 2 diabetes	X	X
GDS161	Gene expression involved in susceptibility for type 2 diabetes	X	X
GDS162	Gene expression involved in susceptibility for type 2 diabetes	X	X
GDS167	Autoimmune disease mechanisms	X	
GDS217	Cancer Genome Anatomy Project SAGE library collection	X	
GDS256	Temporal analysis of skeletal muscle response to corticosteroid methylprednisolone		X
GDS268	Identification of proteins involved in fatty acid oxidation in skeletal muscle of obese individuals		X
GDS272	Hypoxia and glucose metabolism		X
GDS279	Comparison of effects of low fat and high fat diets on liver gene expression in LDL receptor deficient mice		X
GDS365	B-cells and acute renal allograft rejection	X	
GDS402	Type 2 diabetes and renal function	X	X
GDS541	CGAP libraries: brain	X	

Table 19: Sixteen data sets that result after an optimized traversal starting at concept Diabetes Mellitus (C0011849).

GEO data set	Title
GDS11	DNA copy-number aberrations
GDS157	Type 2 diabetes and insulin resistance (HuGeneFL)
GDS158	Type 2 diabetes and insulin resistance (Hu35k-A)
GDS160	Type 2 diabetes and insulin resistance (Hu35k-B)
GDS161	Type 2 diabetes and insulin resistance (Hu35k-C)
GDS162	Type 2 diabetes and insulin resistance (Hu35k-D)
GDS182	Large-scale analysis of the mouse transcriptome
GDS233	Muscle regeneration (U74Av1)
GDS254	Muscle, normal extraocular, profile
GDS256	Pharmacogenomic effect of corticosteroid in skeletal muscle
GDS268	Obesity and fatty acid oxidation
GDS272	Hypoxia and glucose metabolism
GDS276	Muscle profiles
GDS278	Muscle response to acute resistance exercise
GDS2	Melanoma, cutaneous malignant, classification
GDS461	Aortic stiffness

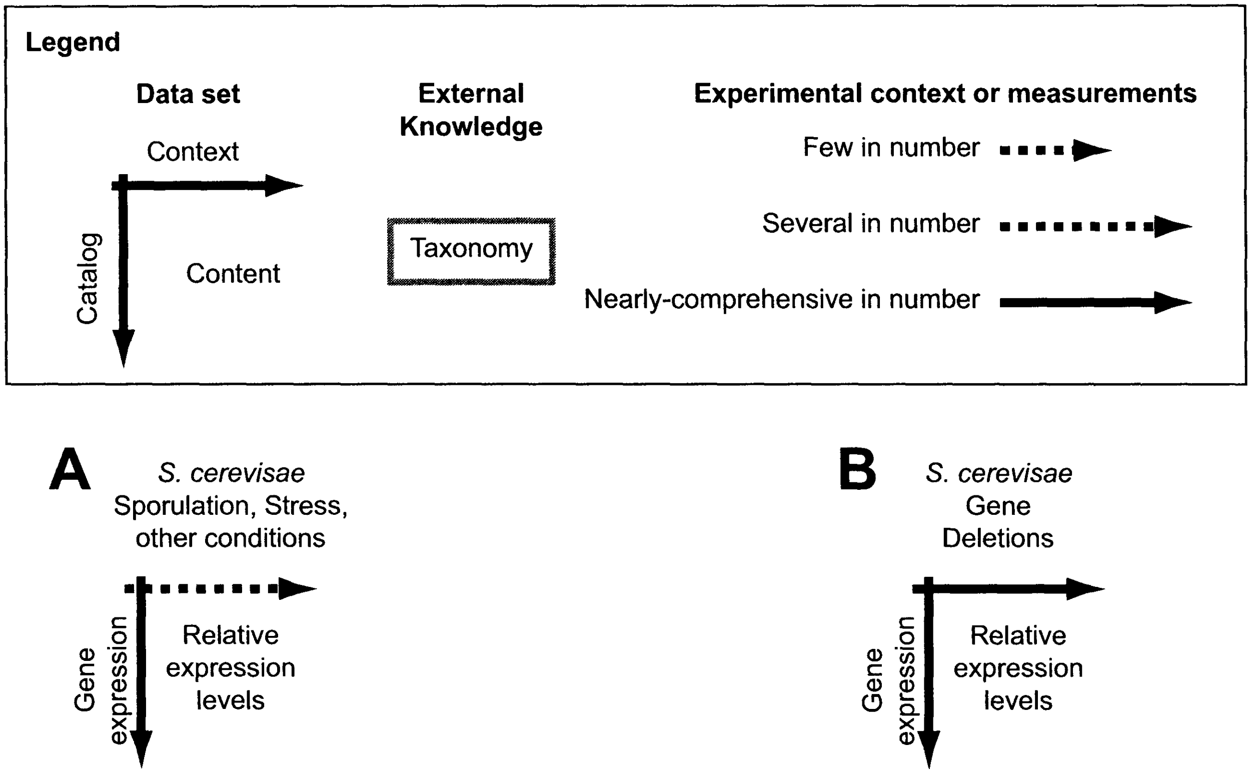


Figure 1: Nearly-comprehensive data sets and experiments can be modeled by considering the experiment context, catalog of measurements, and measured content separately. In the graphical representation of this model, a short broken arrow represents a few contexts or measurements, while a solid arrow represents a nearly-comprehensive set. (A) Eisen, et al., reported on multiple discrete biological processes in yeast, including cell-cycle by α-pherome, sporulation, and stress. [80] (B) Hughes, et al., measured gene expression by microarray in yeast under a variety of conditions, including 200 gene deletion strains. [81] Though the measurement axis is virtually the same, the context axis is significantly different, in that the Hughes data set attempts a systematic approach to the experimental conditions, which can become nearly-comprehensive.

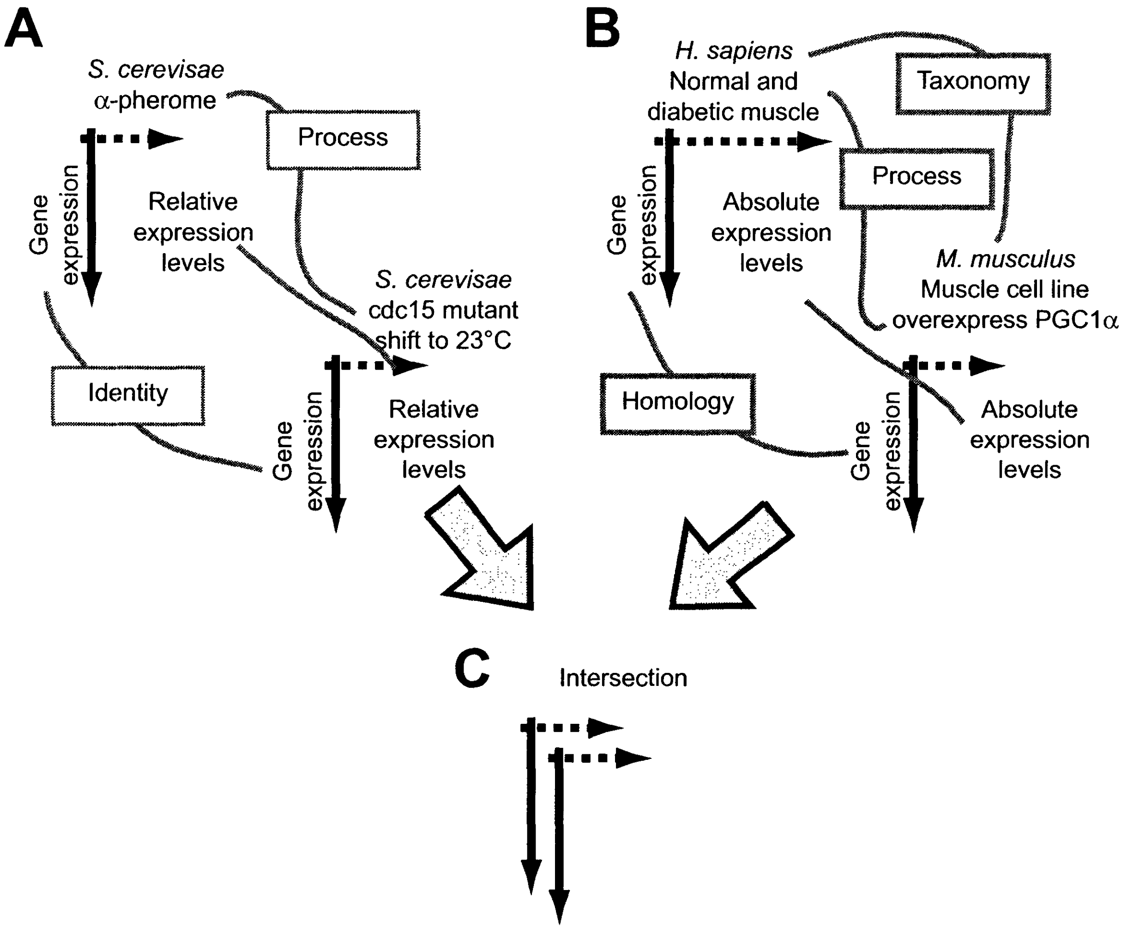


Figure 2: Performing inferential operations on combinations of nearly-comprehensive data sets requires unifying the context, the measurements, and the data elements. (A) Spellman, et al., reported on genes involved in the cell cycle in yeast after α-pherome and elutriation. [246] An automated system attempting to intersect the expression patterns in response to these factors would need to (1) recognize that α-pherome addition and temperature shift in a cdc15 mutant both synchronize cells into the cell cycle, (2) trivially know how to match symbols for the genes measured, and (3) trivially know that the relative expression measurements are directly comparable. (B) Mootha, et al., reported on expression differences in genes involved in oxidative phosphorylation between human diabetic and non-diabetic samples. [247] Hypothesizing that many genes found were downstream of PGC1α (PPARGC1A), they related the expression patterns of these same genes in a publicly available panel of human tissues [248] with mouse skeletal muscle cell lines over-expressing PGC1α. Here, proper intersection could only occur with the knowledge that (1) mouse and human muscle are equivalent for this experiment, (2) mouse genes downstream of PGC1α might be expected to be highly expressed in PGC1α expressing human tissues if the hypothesis is true, and (3) many mouse and human homologies can be compared, and (4) trivially that absolute expression levels are comparable. (C) For both intersections, addition of external prior knowledge is crucial.

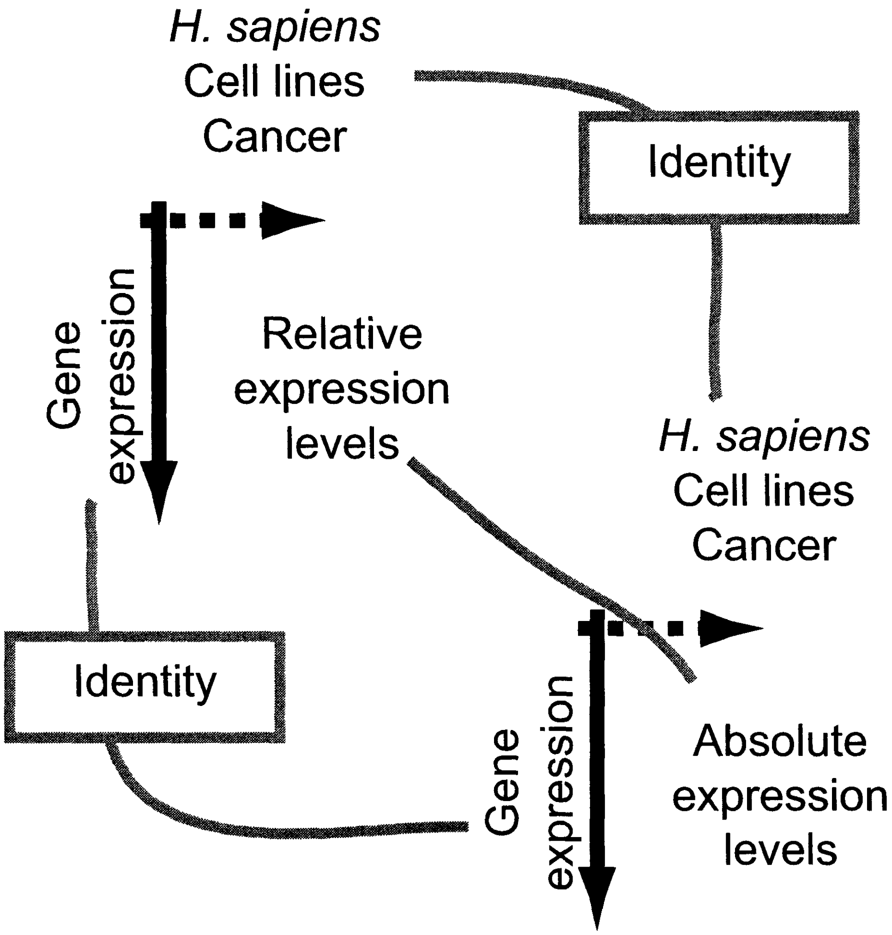


Figure 3: At least two data sets are available for basal gene expression measurements from the NCI60 cancer cell lines. [49,83] Relating the two not only requires the trivial matching of gene and cell line identifiers, but also dealing with the differences between relative and absolute expression levels.

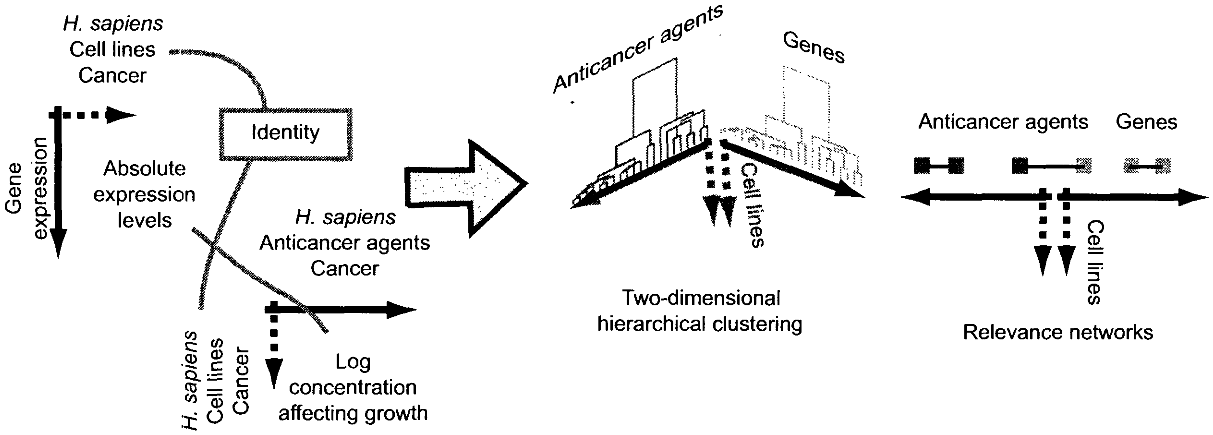


Figure 4: Non-intersection relations between data sets can also occur without matching both context and measurement. Nearly-comprehensive gene expression measurements were made from the NCI60, and these were related to susceptibility measurements in the NCI60 across a nearly-comprehensive set of anti-cancer agents. In this case, the cell lines used as the context of the expression data are identical to the cell lines on which measurements were made in the susceptibility data. The measurements of the expression data set do not match the context of the susceptibility data set; thus it does not make sense to directly intersect these two data sets. Instead, they can be joined across the single common axis, and appropriate analytic tools, such as two-dimensional hierarchical clustering or relevance networks may be used.

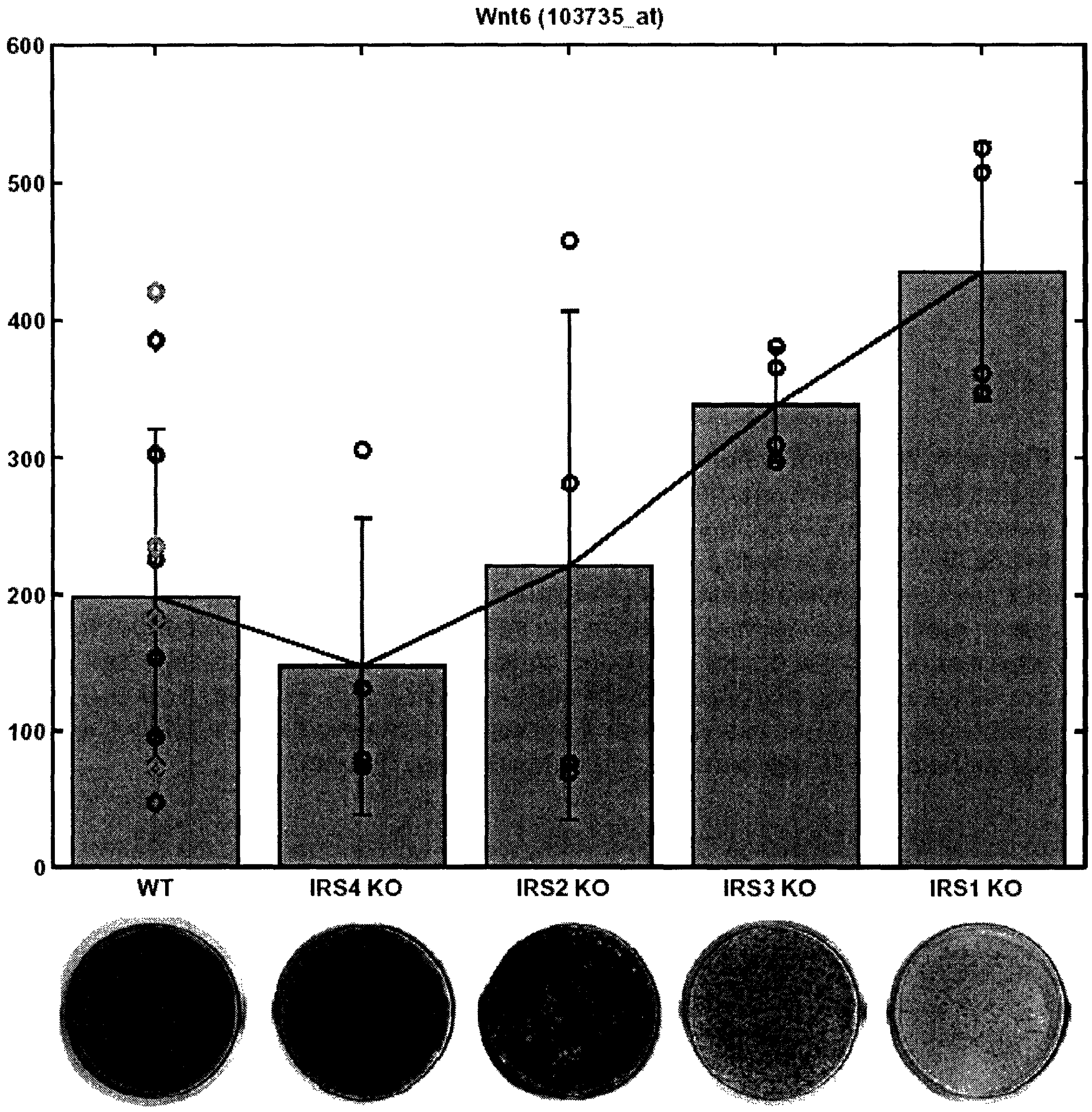


Figure 5: Wnt6 is one of only 20 genes where expression levels correlate with genotype, when ordered from best-differentiating (wildtype), to worst-differentiating (IRS-1 knockout).

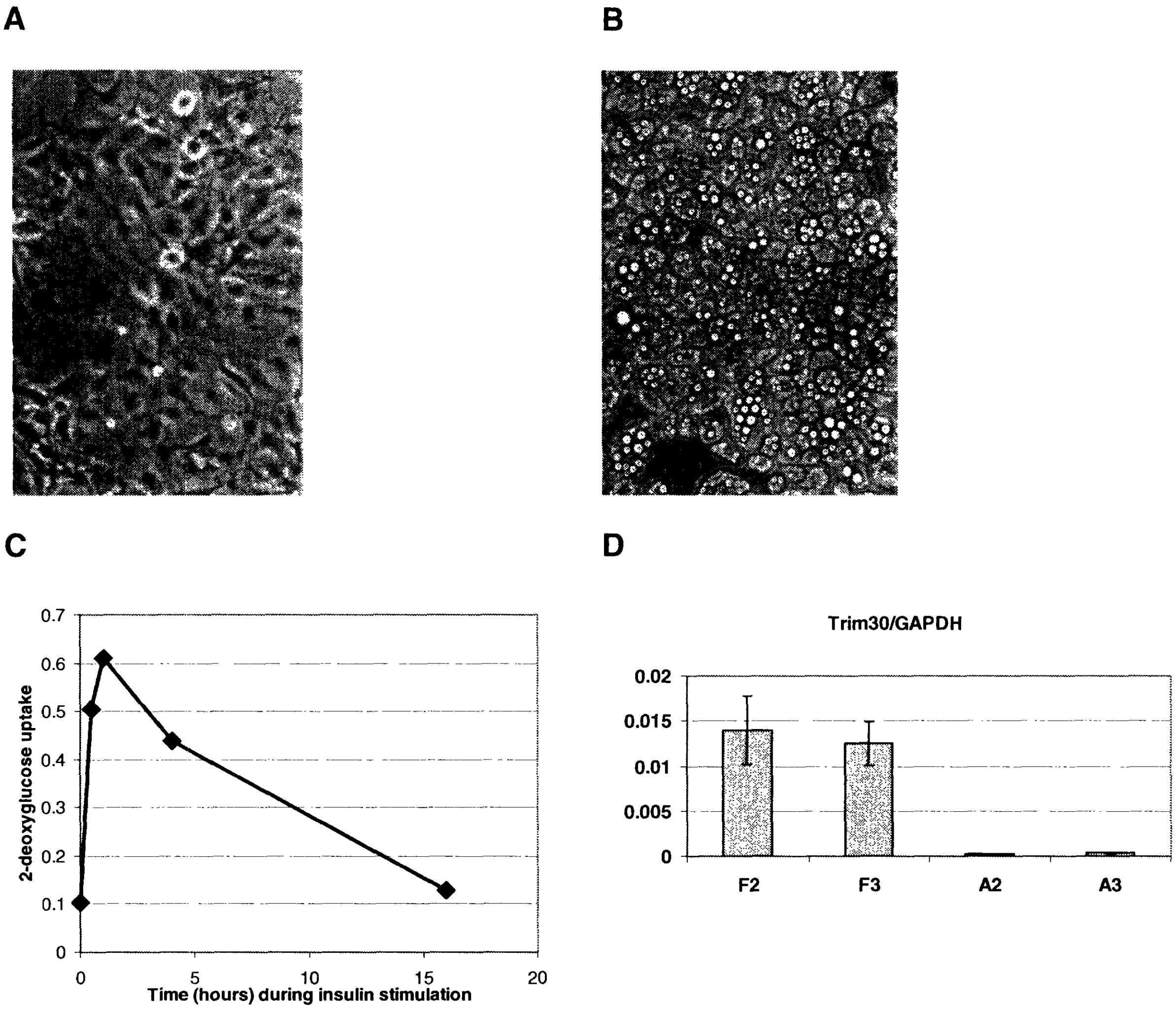


Figure 6: We have differentiated model fibroblasts (panel A) into adipocytes (panel B). Panel C shows insulin-stimulated 2-deoxyglucose uptake in these differentiated adipocytes. Panel D shows *Trim30* expression before and after adipogenesis, with duplicate biological and triplicate technical replication.

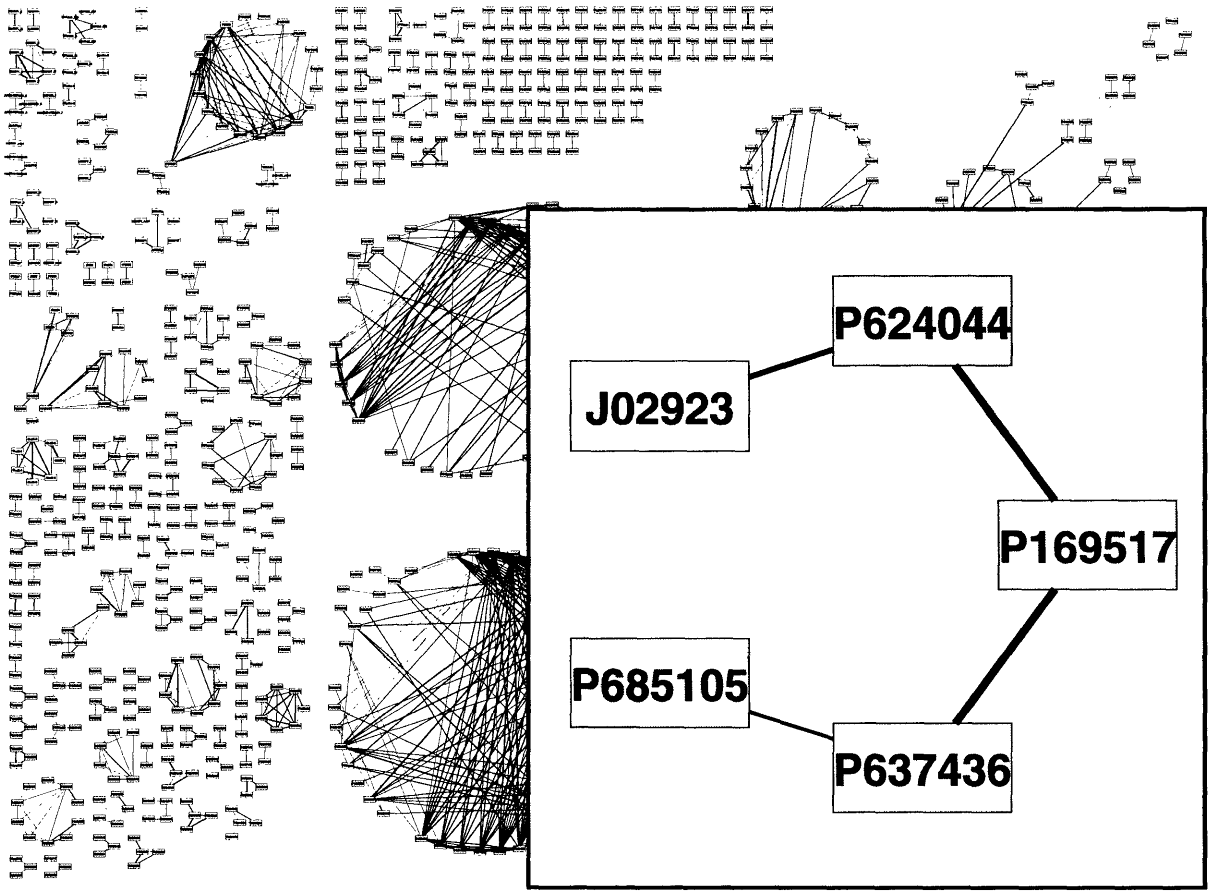


Figure 7: Relevance networks constructed from baseline gene expression in sixty cancer cell lines joined with susceptibility of the same cell lines to anti-cancer agents. The pairs of features (anti-cancer agents in shaded boxes, genes in white boxes) with correlation coefficient beyond ±0.80 were drawn with line thickness proportional to correlation coefficient. The inset shows the association between LCP1 expression (J02923) and susceptibility to a thiazolidine carboxylic acid (P624044).

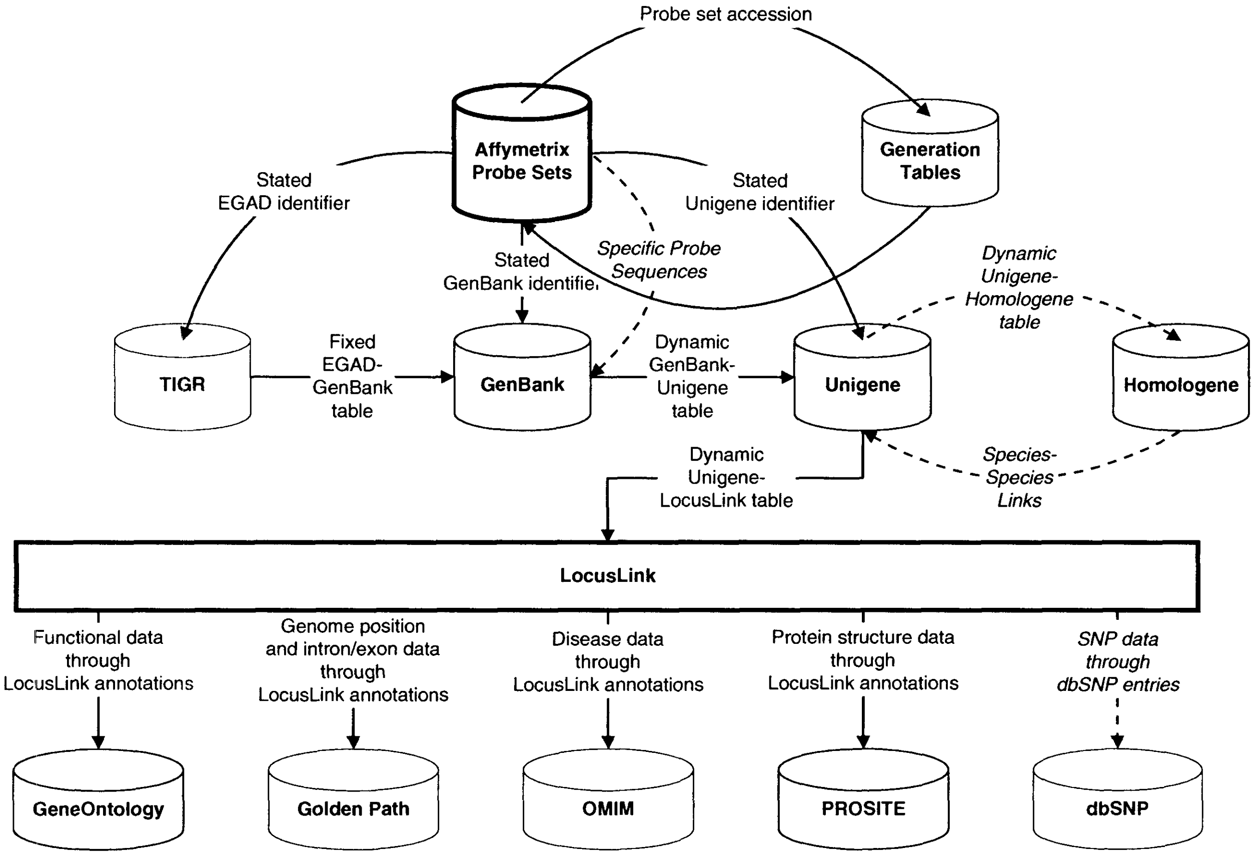

Probe set accession
Affymetrix Probe Sets
Generation Tables
Stated EGAD identifier
Stated Unigene identifier
Stated GenBank identifier
Specific Probe Sequences
Dynamic Unigene-Homologene table
TIGR
Fixed EGAD-GenBank table
GenBank
Dynamic GenBank-Unigene table
Unigene
Homologene
Dynamic Unigene-LocusLink table
Species-Species Links
LocusLink
Functional data through LocusLink annotations
Genome position and intron/exon data through LocusLink annotations
Disease data through LocusLink annotations
Protein structure data through LocusLink annotations
SNP data through dbSNP entries
GeneOntology
Golden Path
OMIM
PROSITE
dbSNP

Figure 8: UNCHIP architecture for cross-microarray, cross-revision knowledge management of microarray probe set annotations. Dashed lines are proposed cross-species connections.

```
Select b.link_symbol from chip_accession a, chip_accession_link b where b.parent_accession = a.id and
    b.link_type = 1 and a.accession = '737_at' and a.array_type = 'u95_a' and b.version = 2;
+-------------+
| link_symbol |
+-------------+
| D87002      |
+-------------+

Select b.link_symbol from chip_accession a, chip_accession_link b where b.parent_accession = a.id and
    b.link_type = 26 and a.accession = '737_at' and a.array_type = 'u95_a' and b.version = 2;
+-------------+
| link_symbol |
+-------------+
| Hs.284380   |
| Hs.296429   |
+-------------+

Select distinct b.link_int, c.link_type, c.link_text, d.link_type, d.link_text from chip_accession a,
    chip_accession_link b, chip_accession_link c, chip_accession_link d where b.parent_accession = a.id
    and c.source_id = b.id and d.source_id = b.id and b.link_type = 27 and c.link_type in (18,31,38) and
    d.link_type in (17,37) and a.accession = '737_at' and a.array_type = 'u95_a' and b.version = 2;
+----------+-----------+-----------+-----------+-----------------------------------------------------+
| link_int | link_type | link_text | link_type | link_text                                           |
+----------+-----------+-----------+-----------+-----------------------------------------------------+
|    25812 |        38 | POM121L1  |        37 | similar to rat integral membrane glycoprotein POM121 |
|     2678 |        31 | D22S732   |        17 | gamma-glutamyltransferase 1                         |
|     2678 |        31 | D22S672   |        17 | gamma-glutamyltransferase 1                         |
|     2678 |        31 | GTG       |        17 | gamma-glutamyltransferase 1                         |
|     2678 |        31 | GGT       |        17 | gamma-glutamyltransferase 1                         |
|     2678 |        18 | GGT1      |        17 | gamma-glutamyltransferase 1                         |
+----------+-----------+-----------+-----------+-----------------------------------------------------+
6 rows in set (0.01 sec)
```

Figure 9: A query can be written in SQL to explain why probe set 737_at on the Affymetrix U95A microarray refers to more than one LocusLink gene. The queries indicated that GenBank D87002 cluster in both UniGene clusters Hs.284380 and Hs.296429, which link to LocusLink genes 25812 (POM121L1) and 2678 (GTG).

```
Select distinct a.link_text, b.link_int, c.link_text, d.accession, d.array_type
From chip_accession_link a, chip_accession_link b, chip_accession_link c,
    chip_accession d
Where
a.link_text = ‘Homo sapiens’ and a.link_type = 35 and
b.link_type in (15,27,23) and
c.link_text = ‘4’ and c.link_type = 28 and
a.parent_accession = b.parent_accession and
a.parent_accession = c.parent_accession and
a.parent_accession = d.id;
```

```
+--------------+----------+-----------+------------------+---------------+
| link_text    | link_int | link_text | accession        | array_type    |
+--------------+----------+-----------+------------------+---------------+
| Homo sapiens |    85462 | 4         | 44096_at         | u95_b         |
| Homo sapiens |    85438 | 4         | 76249_at         | u95_e         |
| Homo sapiens |    85013 | 4         | 65166_at         | u95_c         |
| Homo sapiens |    84992 | 4         | 45734_at         | u95_b         |
| Homo sapiens |    84869 | 4         | 48697_at         | u95_b         |
| Homo sapiens |    84803 | 4         | 52883_at         | u95_b         |
| Homo sapiens |    84740 | 4         | 66234_at         | u95_c         |
```

Figure 10: SQL query listing genes on human chromosome 4 that are measured across microarray types.

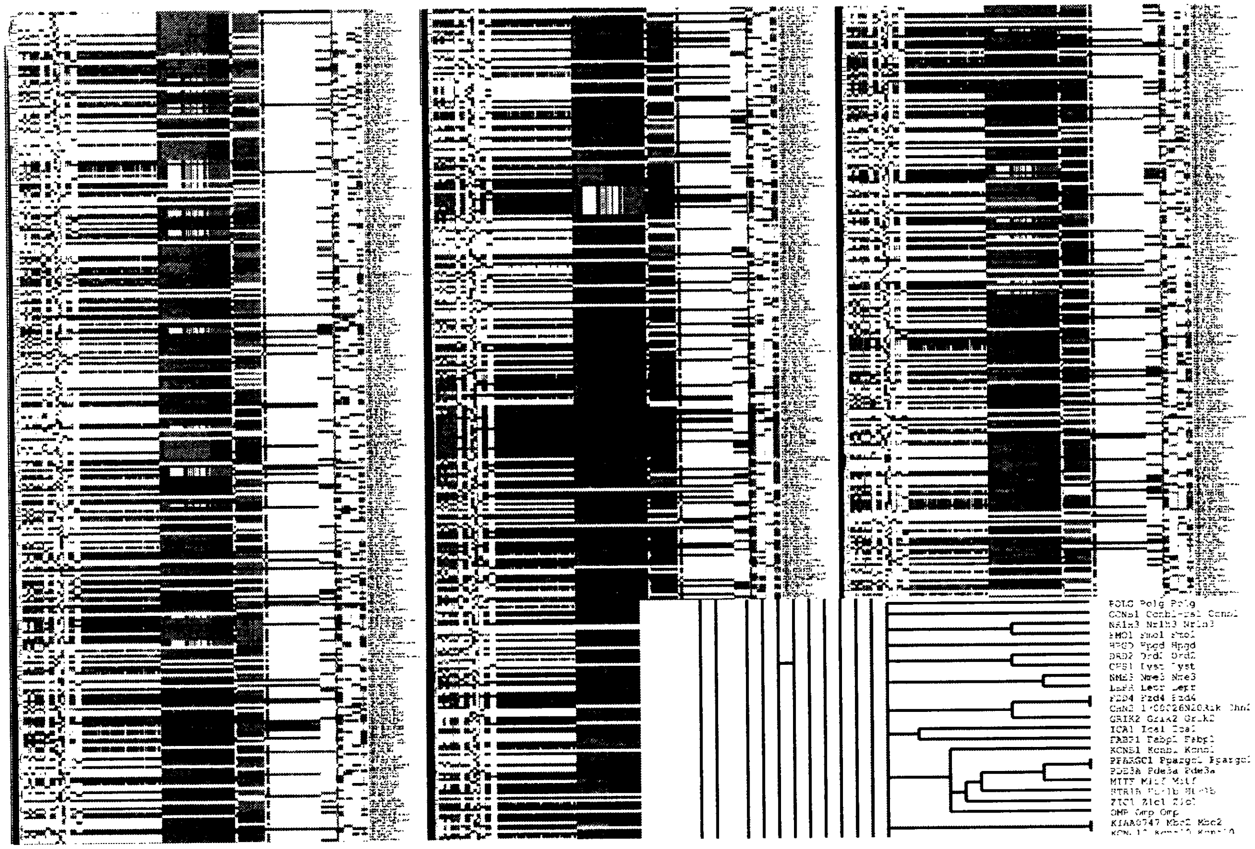


Figure 11: To demonstrate the cross-species, cross-platform, and cross-institution abilities of PGAGENE, we constructed a dendrogram from the 893 genes having HomoloGene information linking across *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*, with at least one PGA-measured expression value from each species. The columns represent 997 microarrays in one of two platforms (spotted cDNA and Affymetrix GeneChip) from three PGAs (CardioGenomics, HopGenes, and TREX). Expression values were normalized to rank ordered percentile and color-coded for a measurement, or white for unmeasured data. The inset shows an interesting subset of this dendrogram, where the four diabetes-related genes insulin autoantigen 1 (ICA1), fatty acid binding protein 1 (FABP1), leptin receptor (LEPR), and peroxisome proliferative activated receptor, gamma, coactivator 1 (PPARGC1) all appear in the same sub-branch, indicating shared expression patterns across the three species.

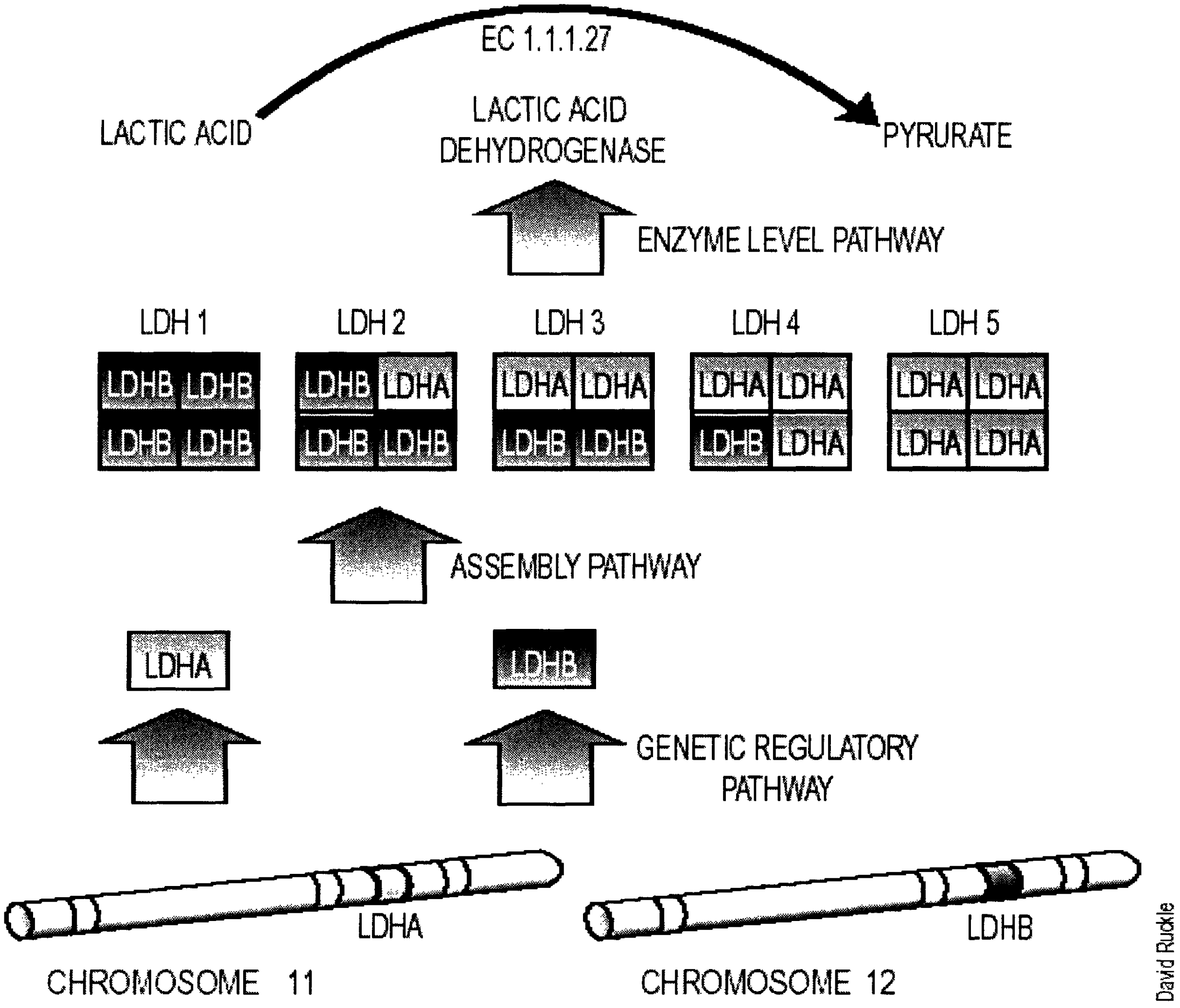


Figure 12: From Kohane, et al. [249] An example of why pathway discovery solely using gene expression measurements is difficult, using the nested pathways involved in lactic acid conversion to pyruvate. This is also an example of why it is not sufficient to simply represent or visualize the glycolytic pathway in terms of measurements from individual genome-scale modalities. Instead, when interpreting genome-scale measurements simultaneously, a knowledge-base that integrates across these pathways will be necessary.

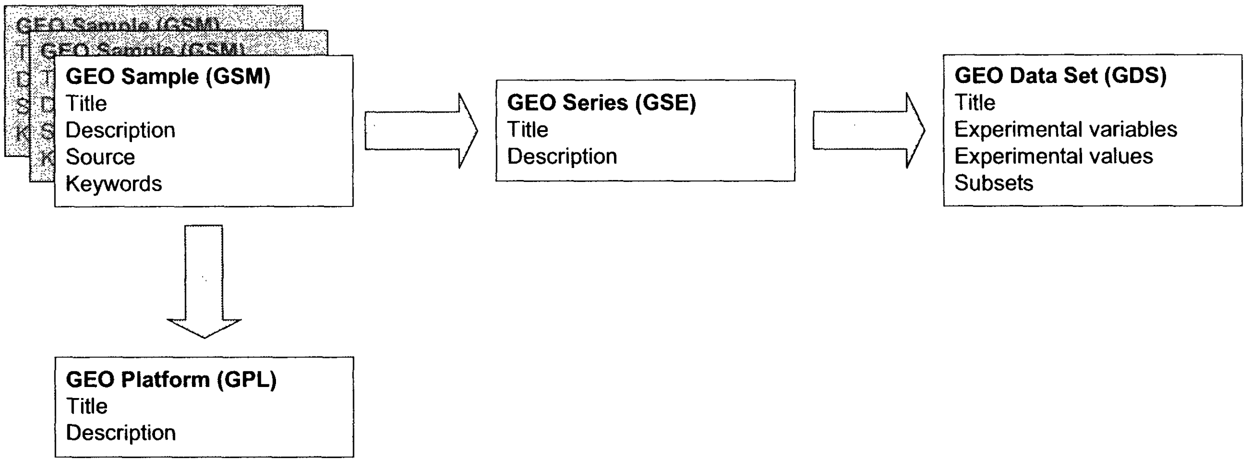


Figure 13: Relation between GEO samples, series, data sets and platforms.

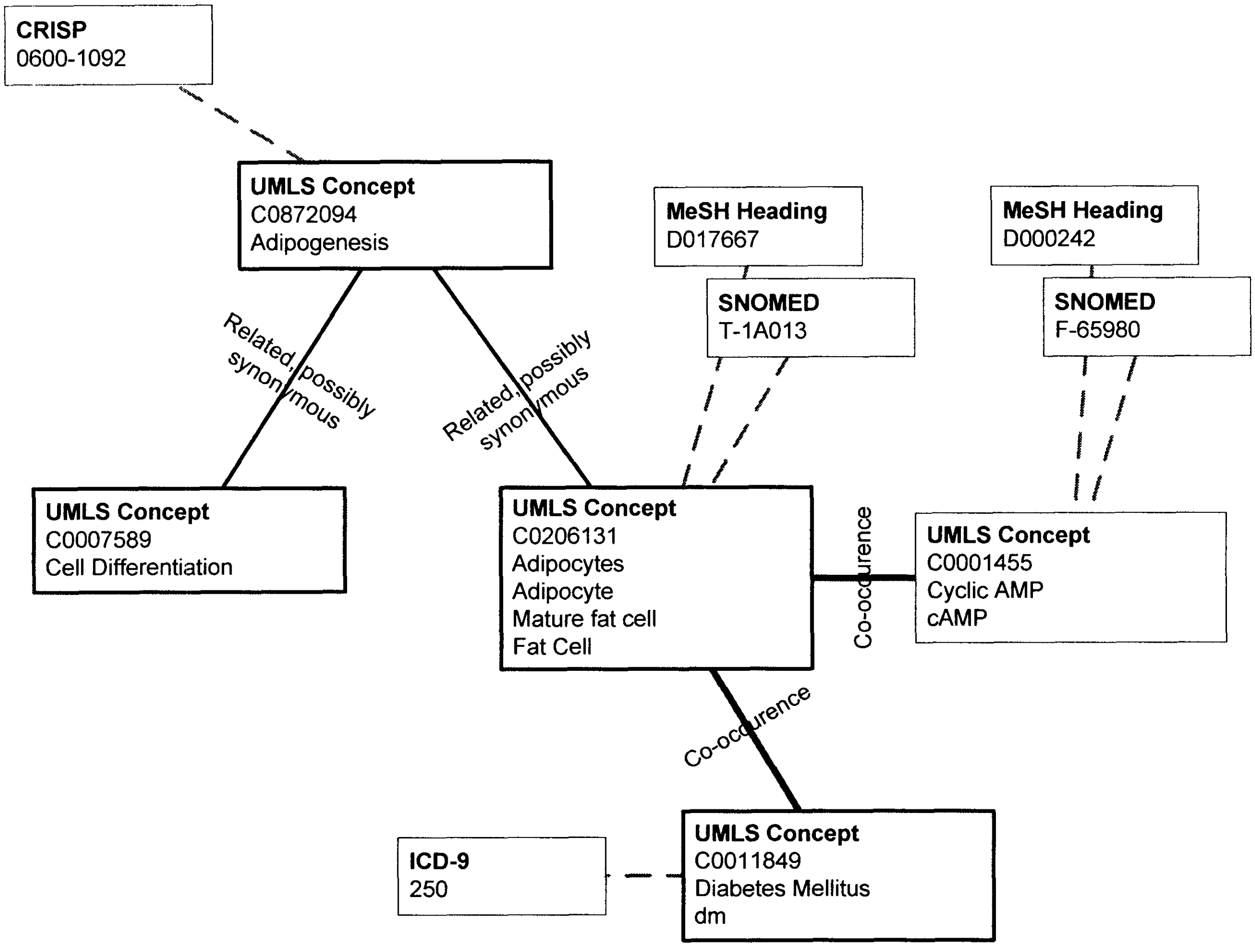


Figure 14: An example illustrating concepts, terms and synonyms, relations to source vocabularies, and relations between concepts including asserted relationships and statistical relationships.

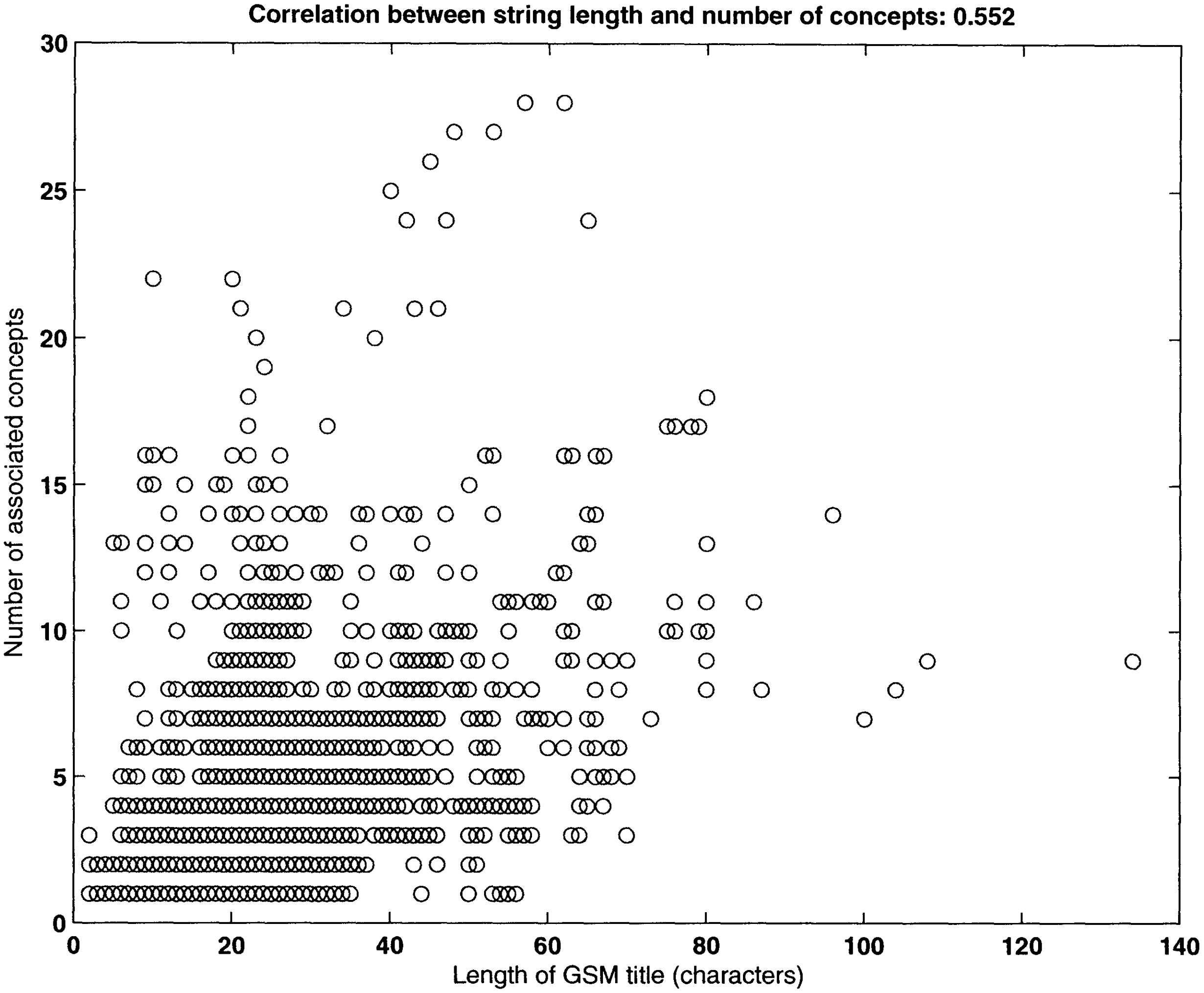


Figure 15: Plot between the length of a GEO sample title and the number of unique concepts mapped.

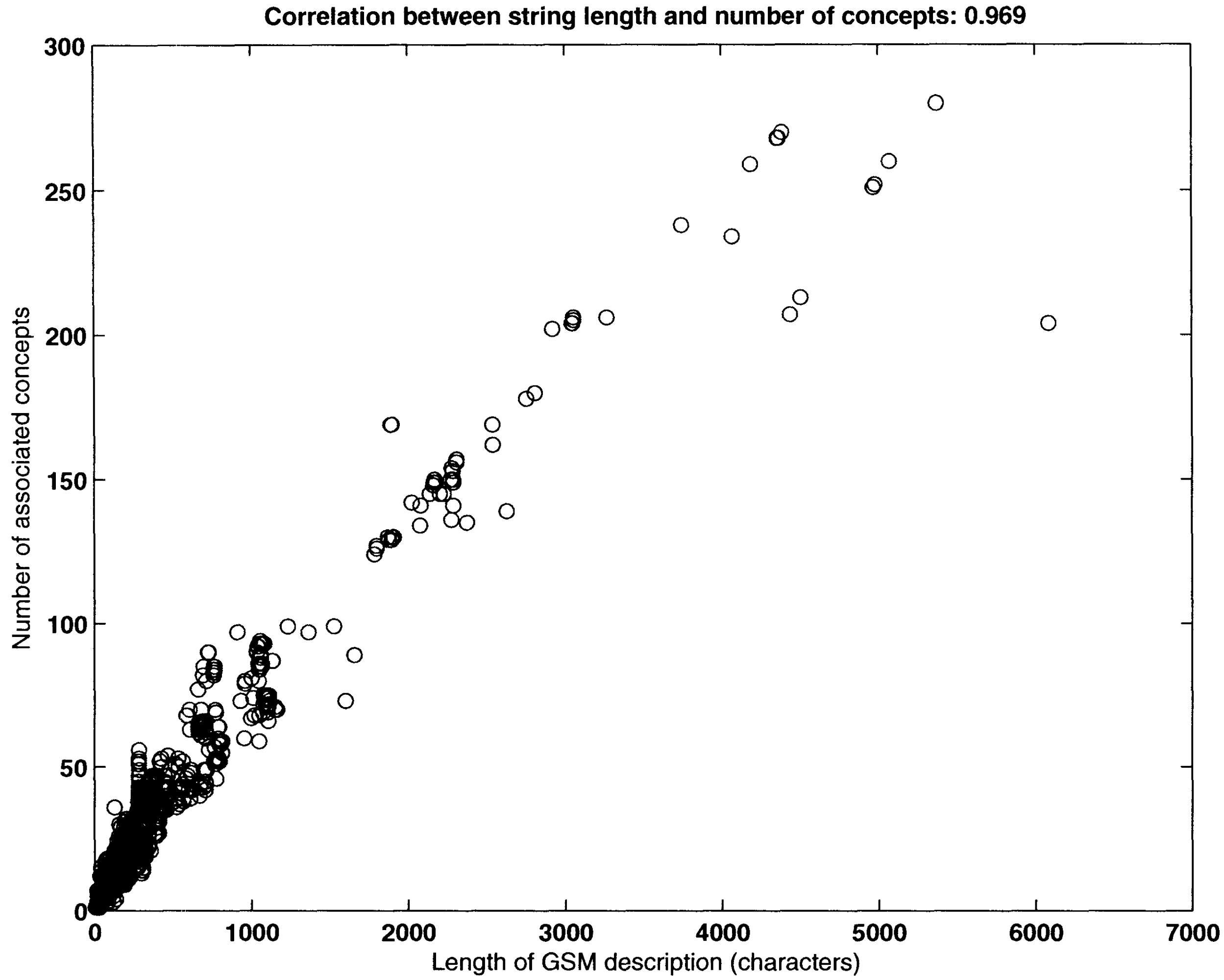


Figure 16: Plot between the length of a GEO sample description and the number of unique concepts mapped.

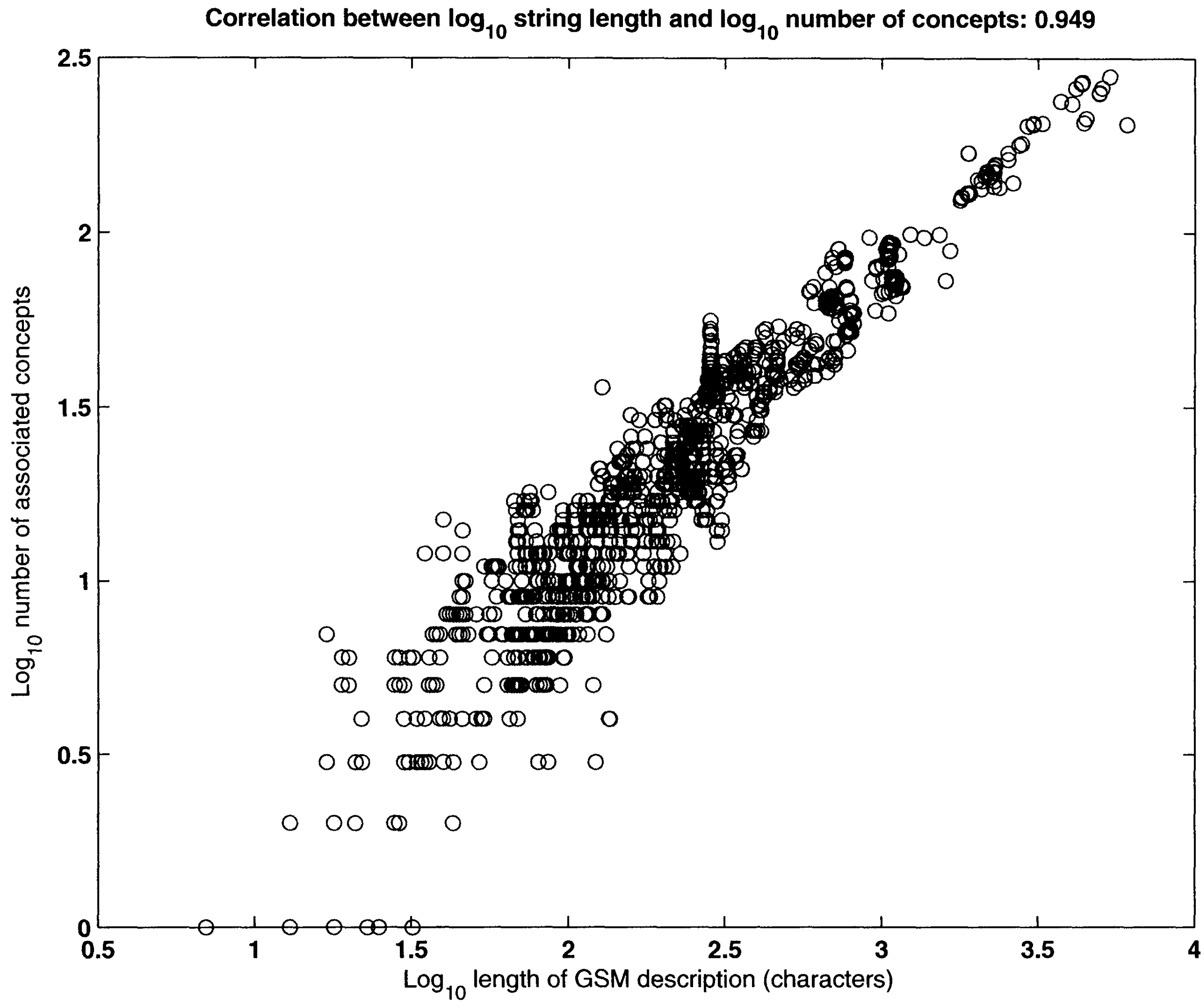


Figure 17: Plot between the log of the length of a GEO sample description and the number of unique concepts mapped.

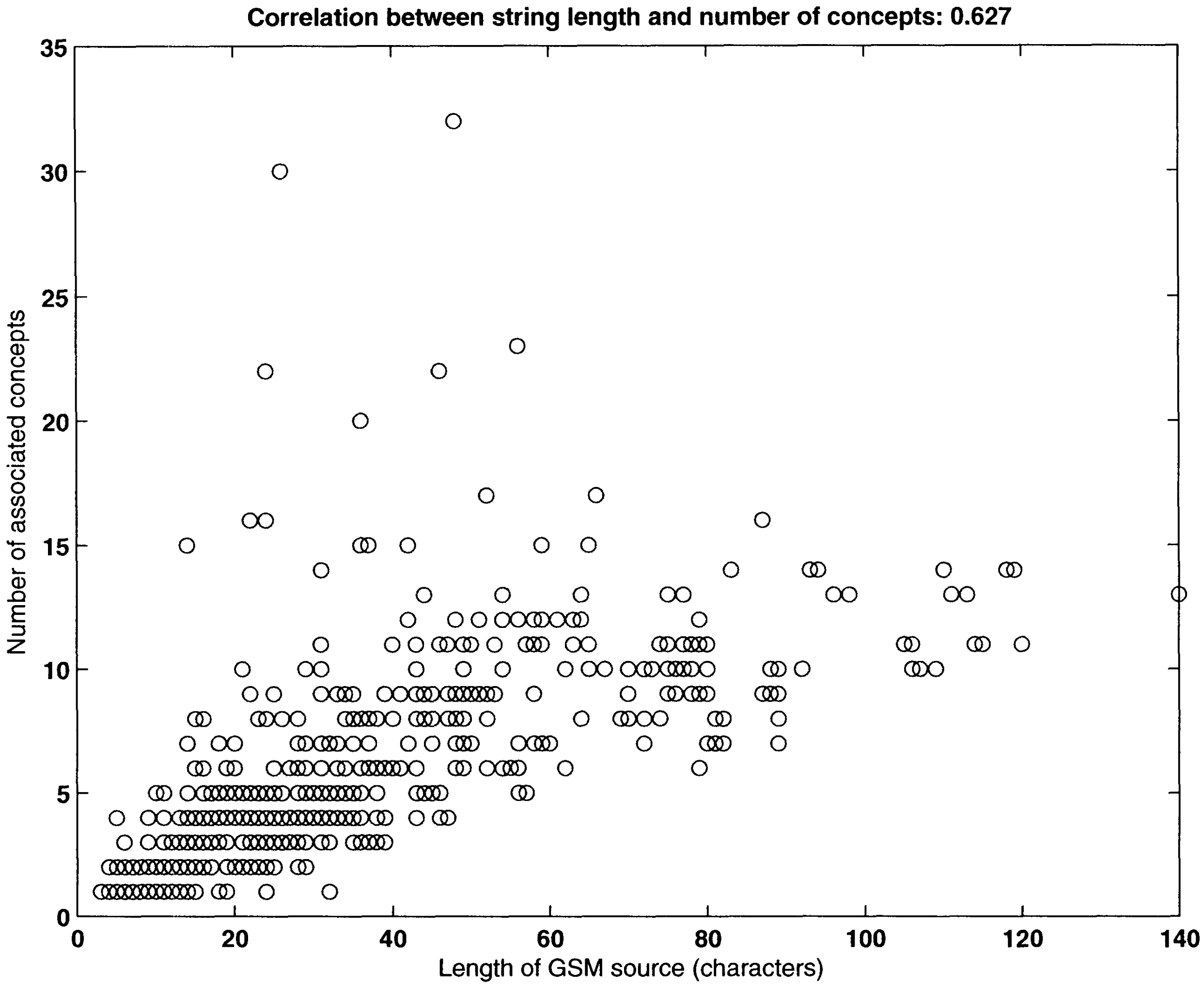


Figure 18: Plot between the length of a GEO sample source and the number of unique concepts mapped.

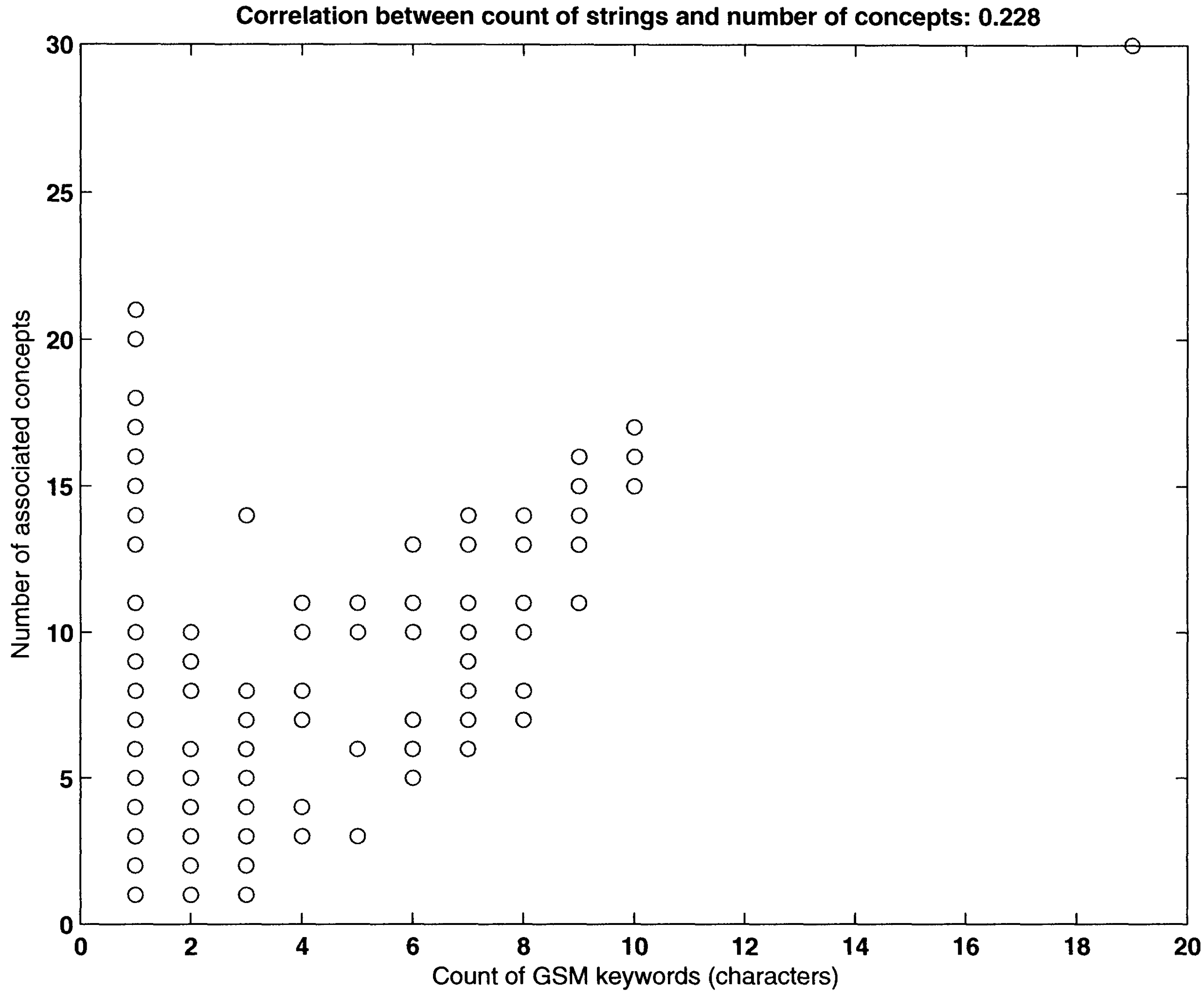


Figure 19: The keyword annotation is the only one in a GEO sample that may be repeated. Plot between the number of GEO keyword annotations for a sample and the number of unique concepts mapped.

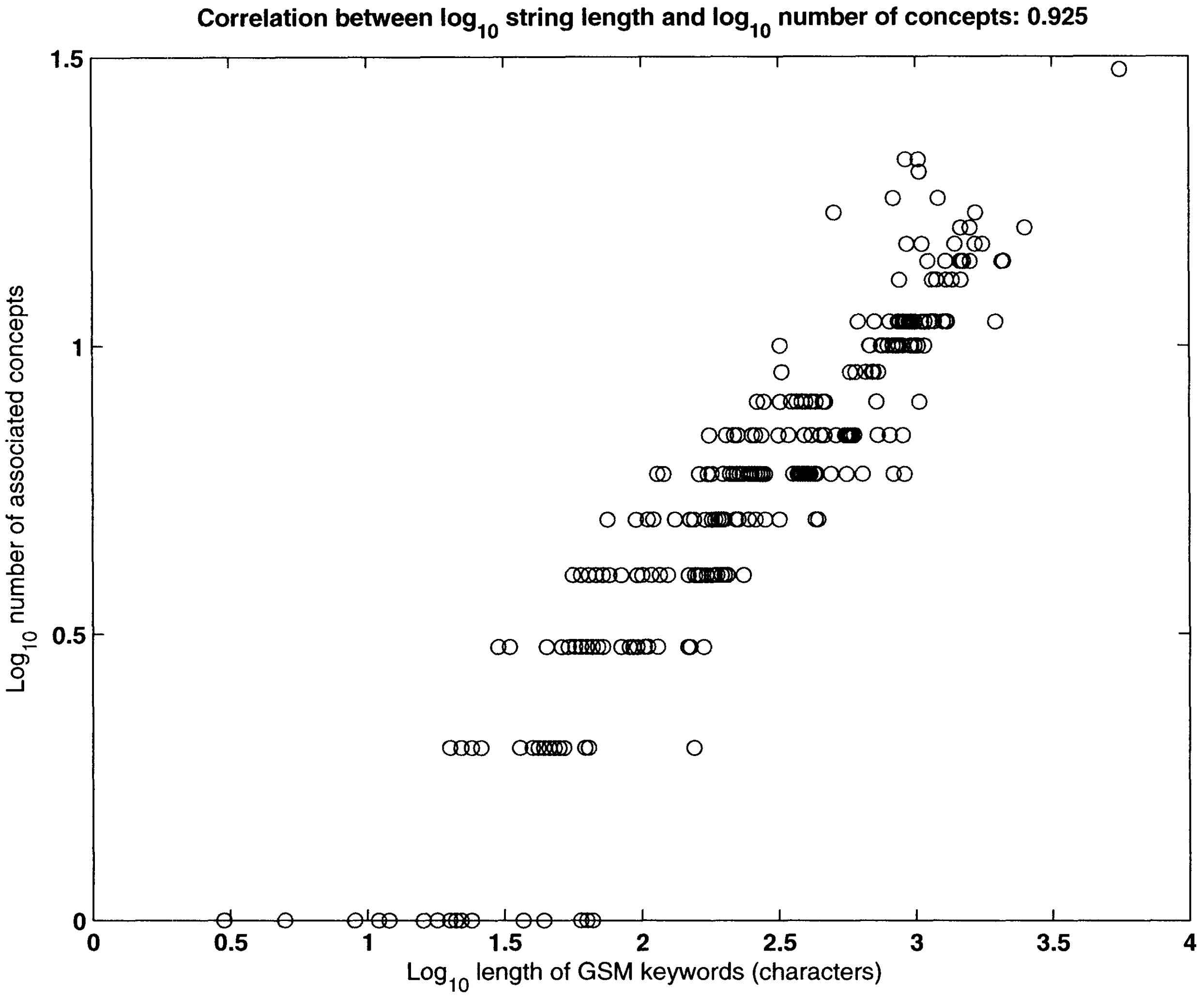


Figure 20: Plot between the log length of all the keyword annotations for a GEO sample and the number of unique concepts mapped.

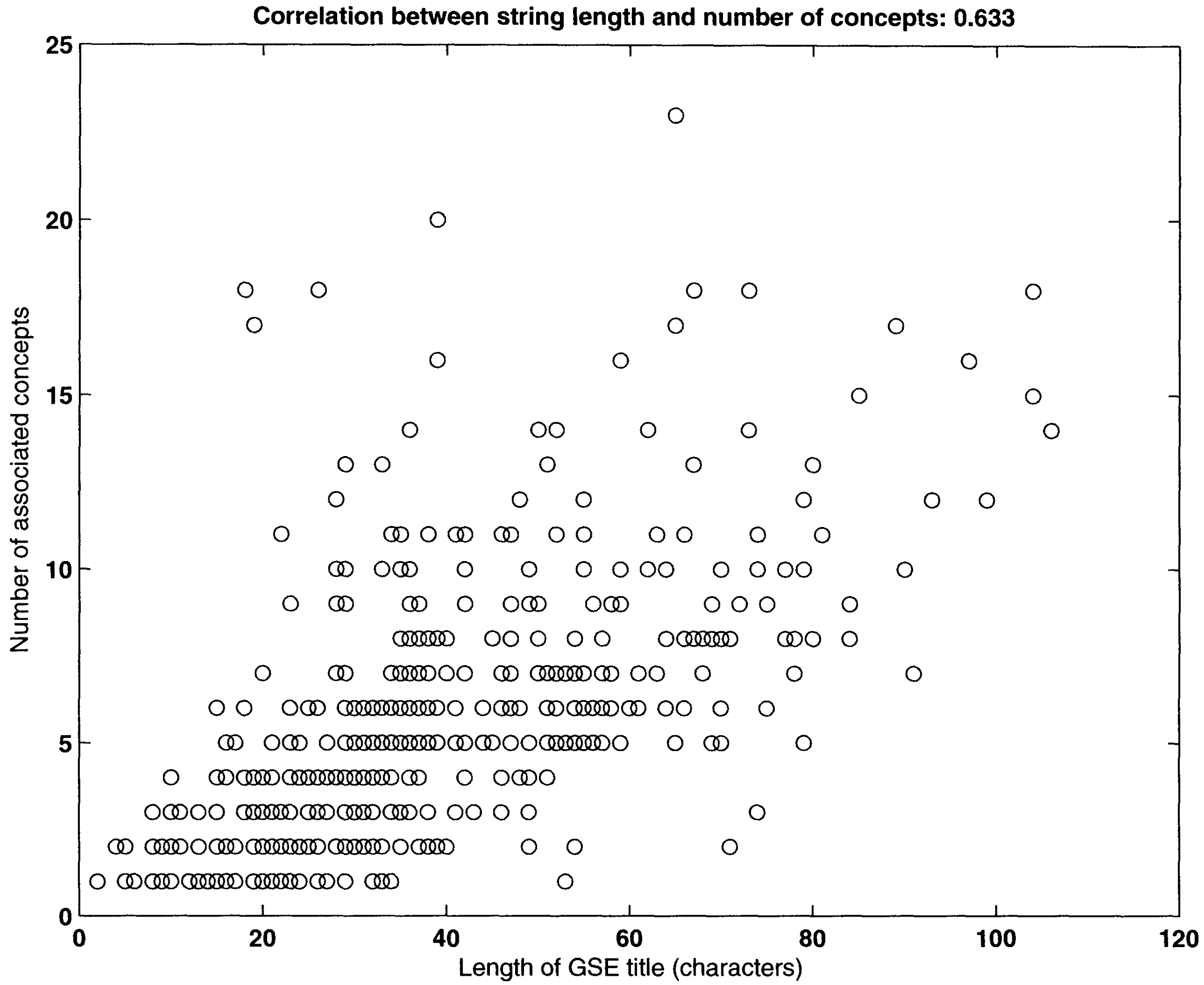


Figure 21: Plot between the length of a GEO series title and the number of unique concepts mapped.

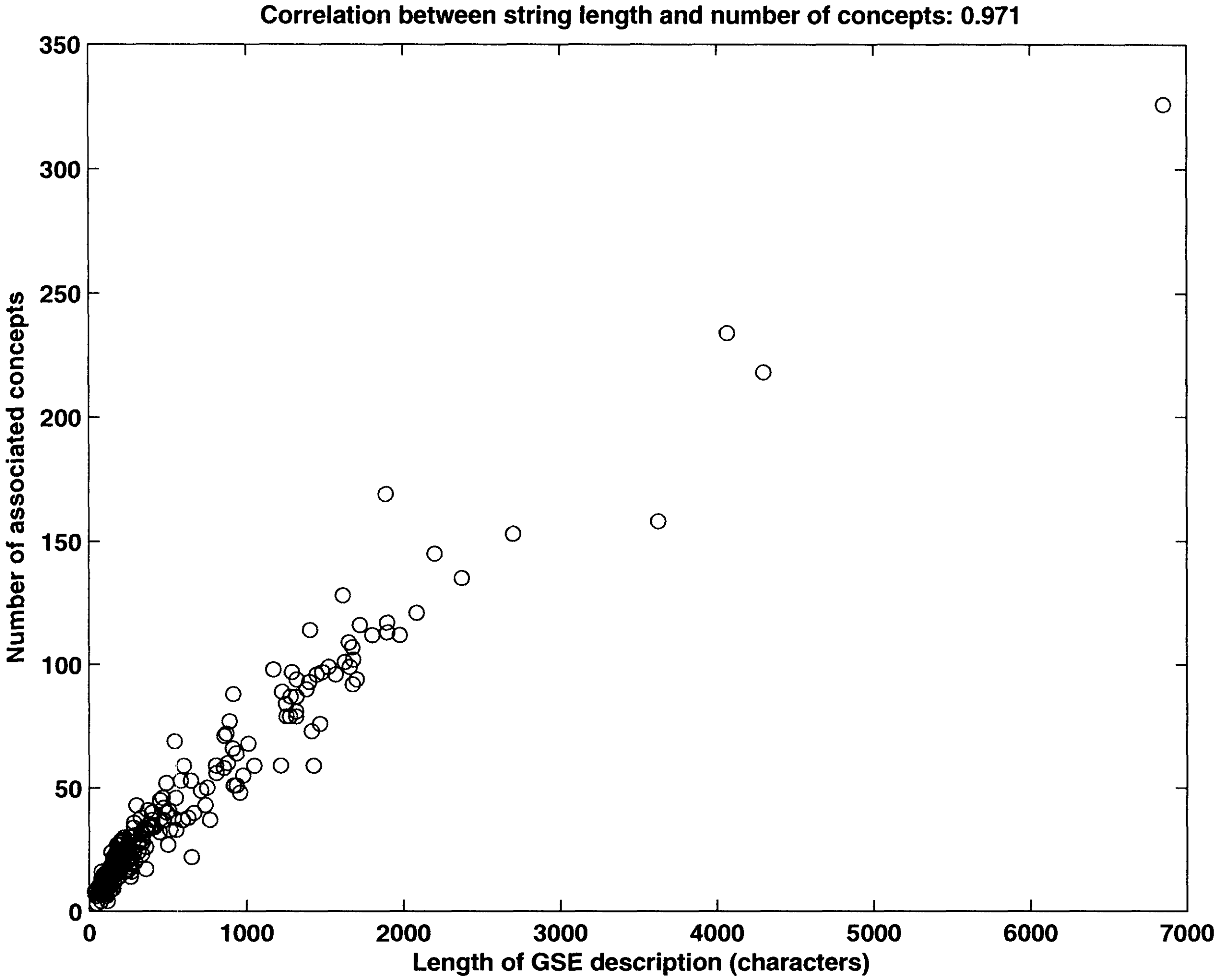


Figure 22: Plot between the length of a GEO series description and the number of unique concepts mapped.

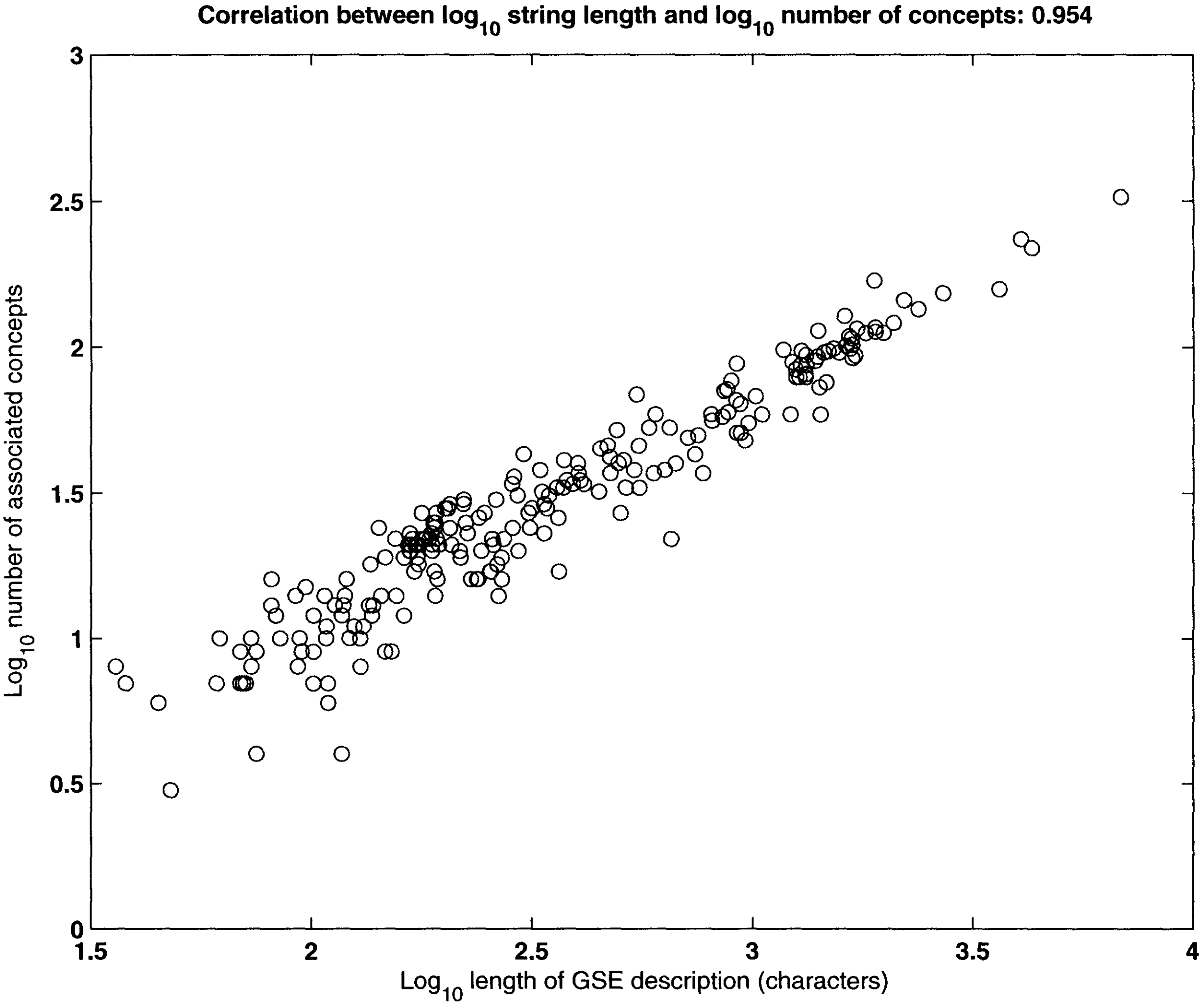


Figure 23: Plot between the log length of a GEO series description and the number of unique concepts mapped.

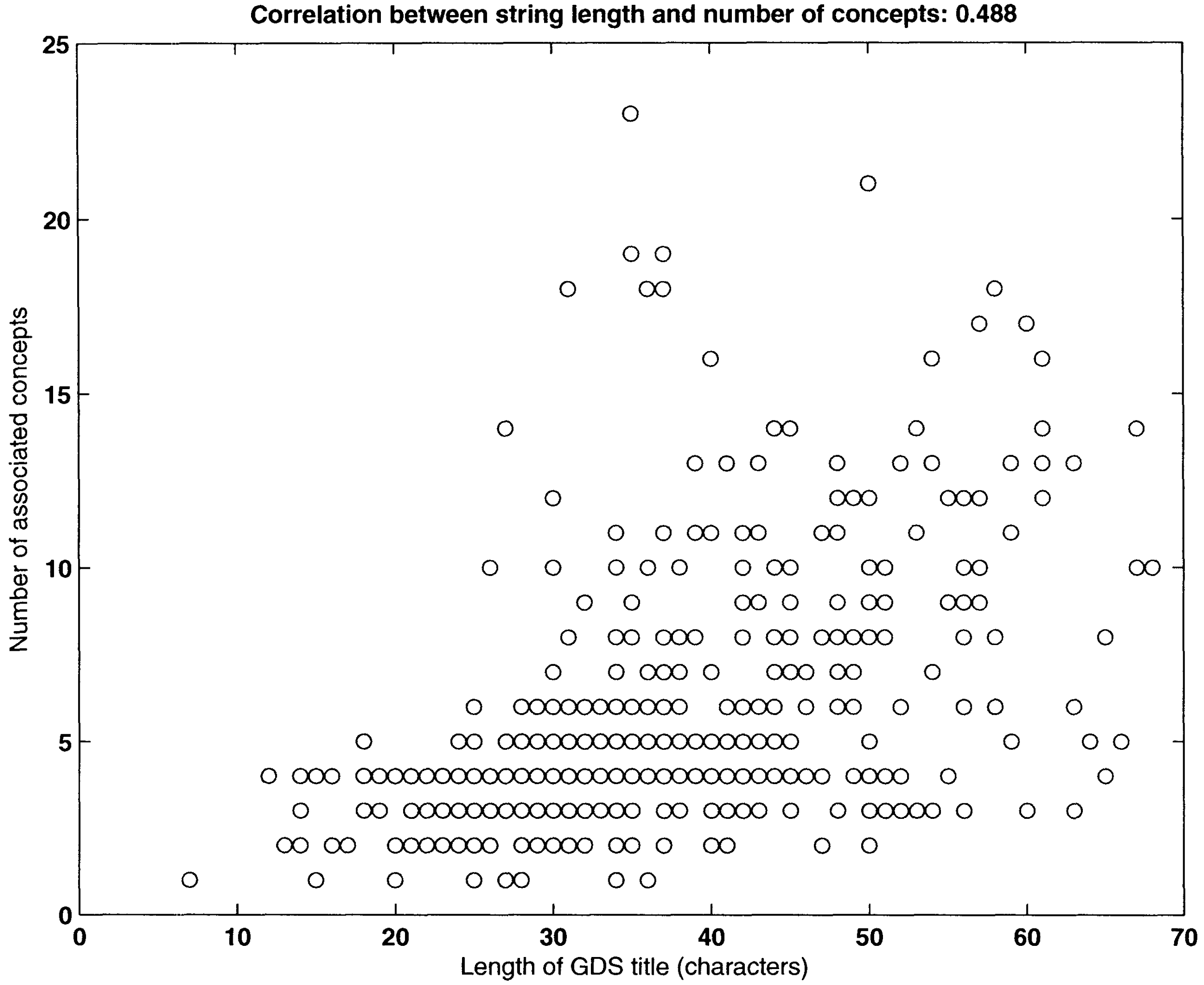


Figure 24: Plot between the length of a GEO data set title and the number of unique concepts mapped.

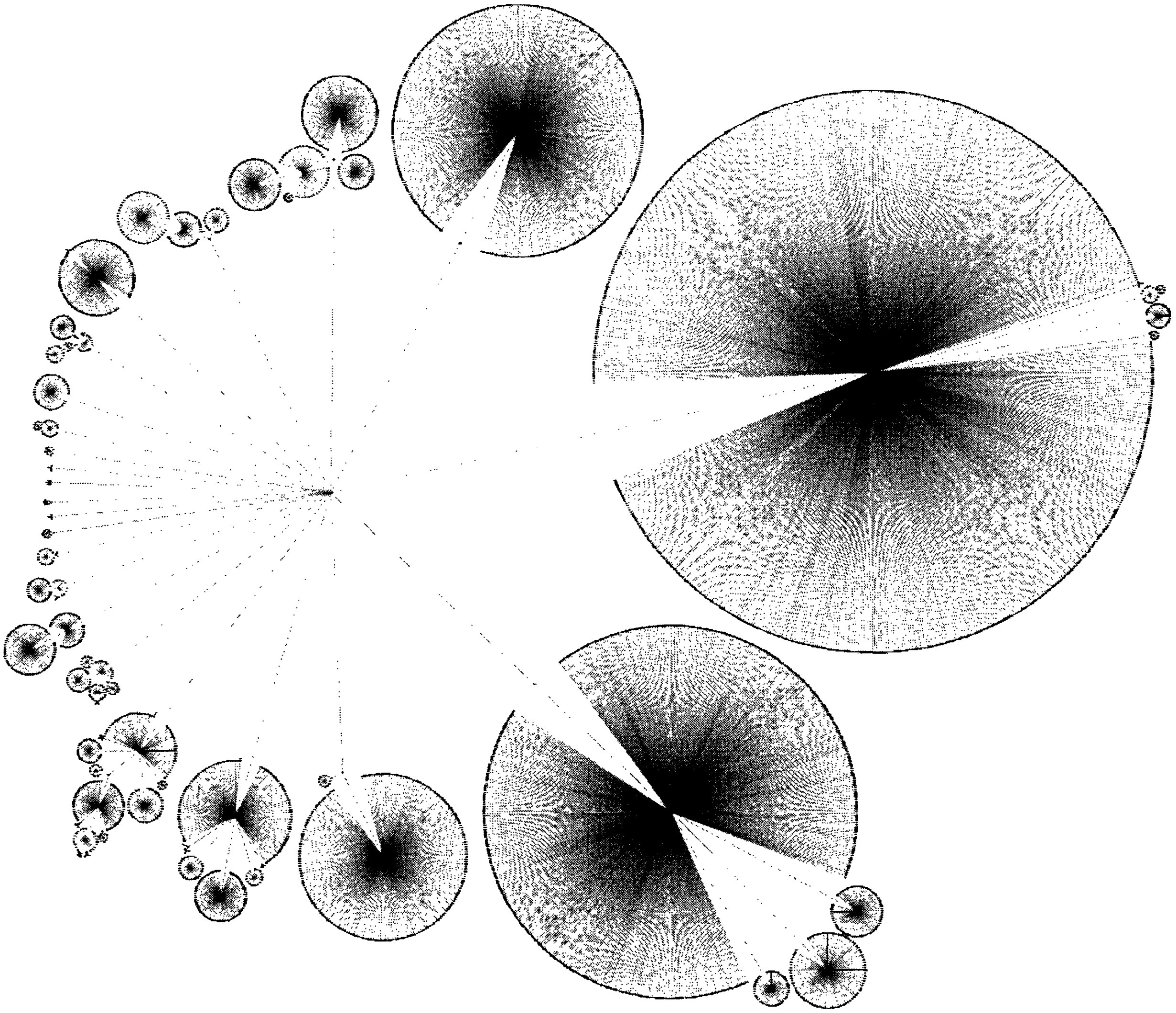


Figure 25: Graphical representation of the breadth-first traversal starting at concept Cells (C0007634) (the center point) and continuing until GEO samples are reached.

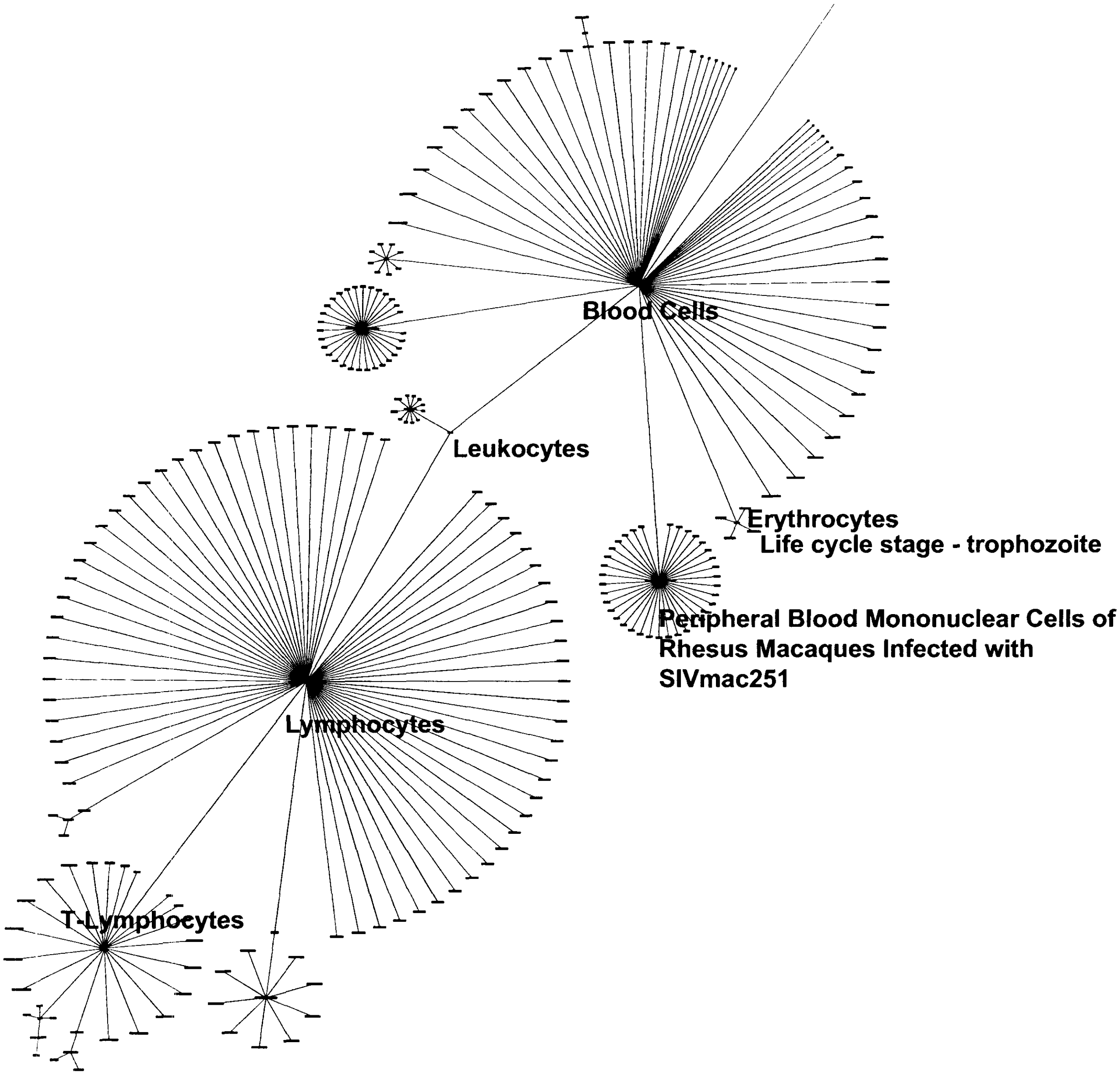

Blood Cells
Leukocytes
Erythrocytes
Life cycle stage - trophozoite
Peripheral Blood Mononuclear Cells of Rhesus Macaques Infected with SIVmac251
Lymphocytes
T-Lymphocytes

Figure 26: Hierarchy of cell types under Blood Cells and their associated GEO samples.

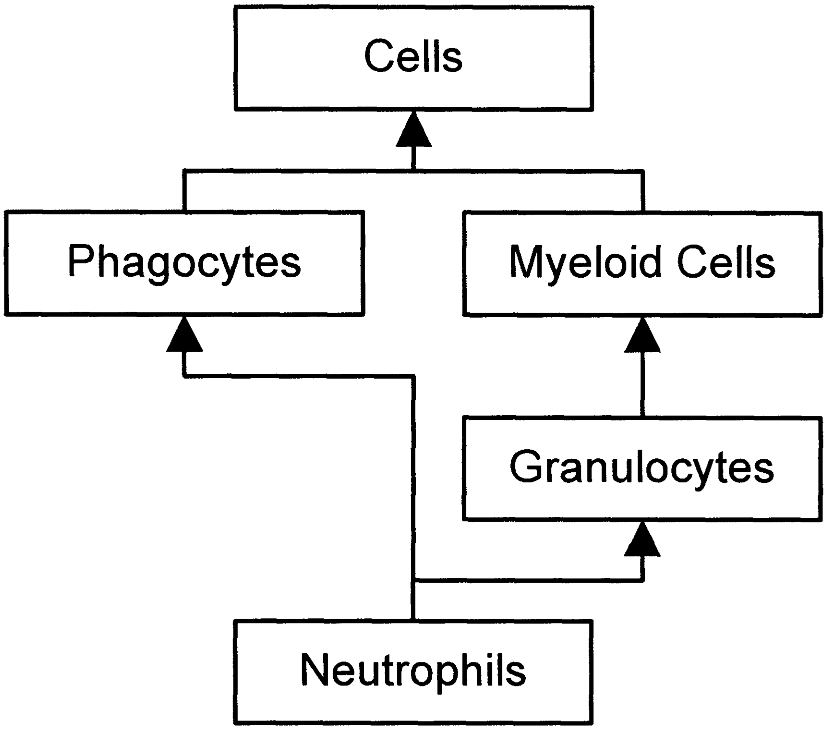


Figure 27: Two paths exist from concept Neutrophils to concept Cells. Though the right path adds additional specificity, the left path will be chosen during a breadth-first search starting at Cells because it is shorter.

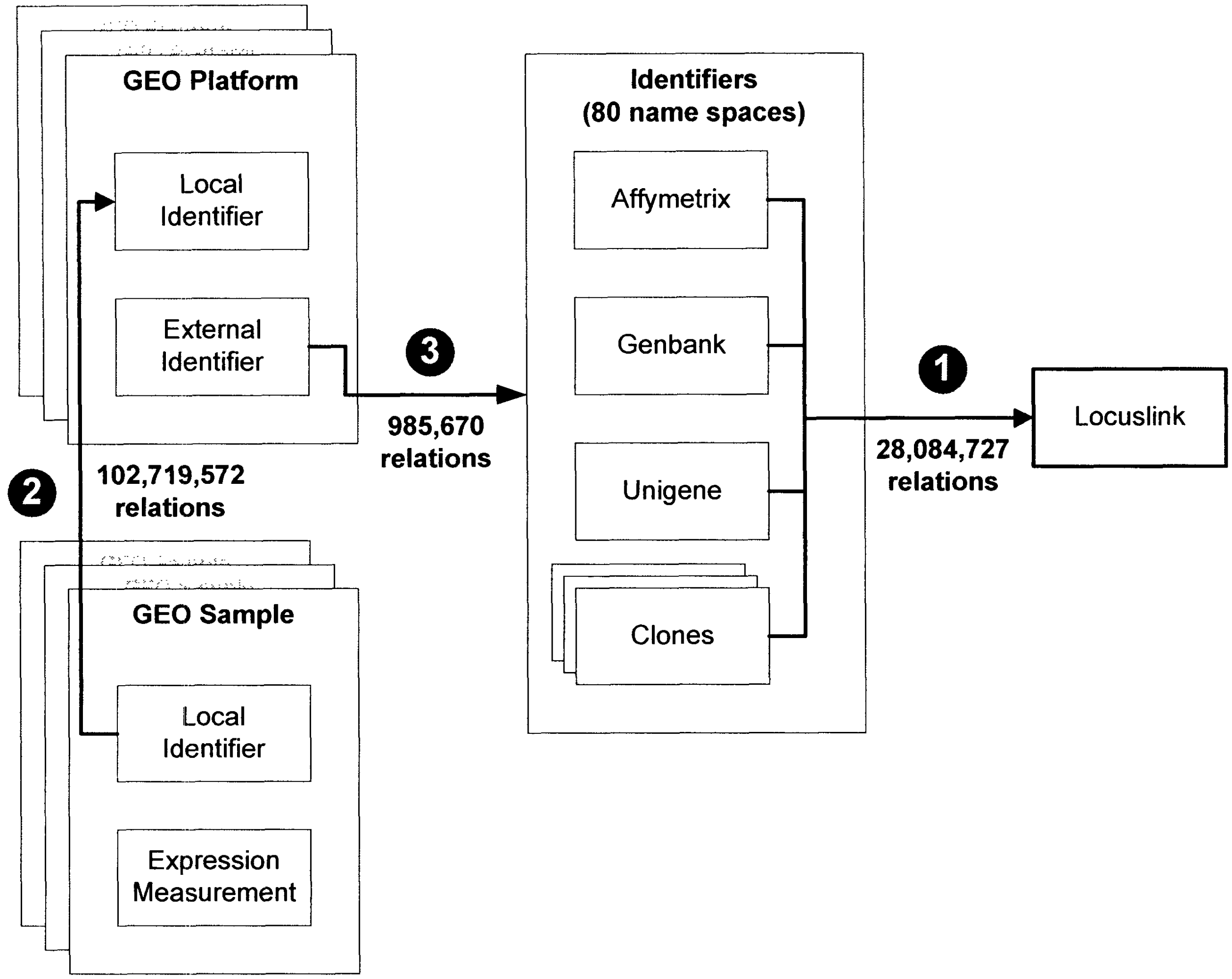
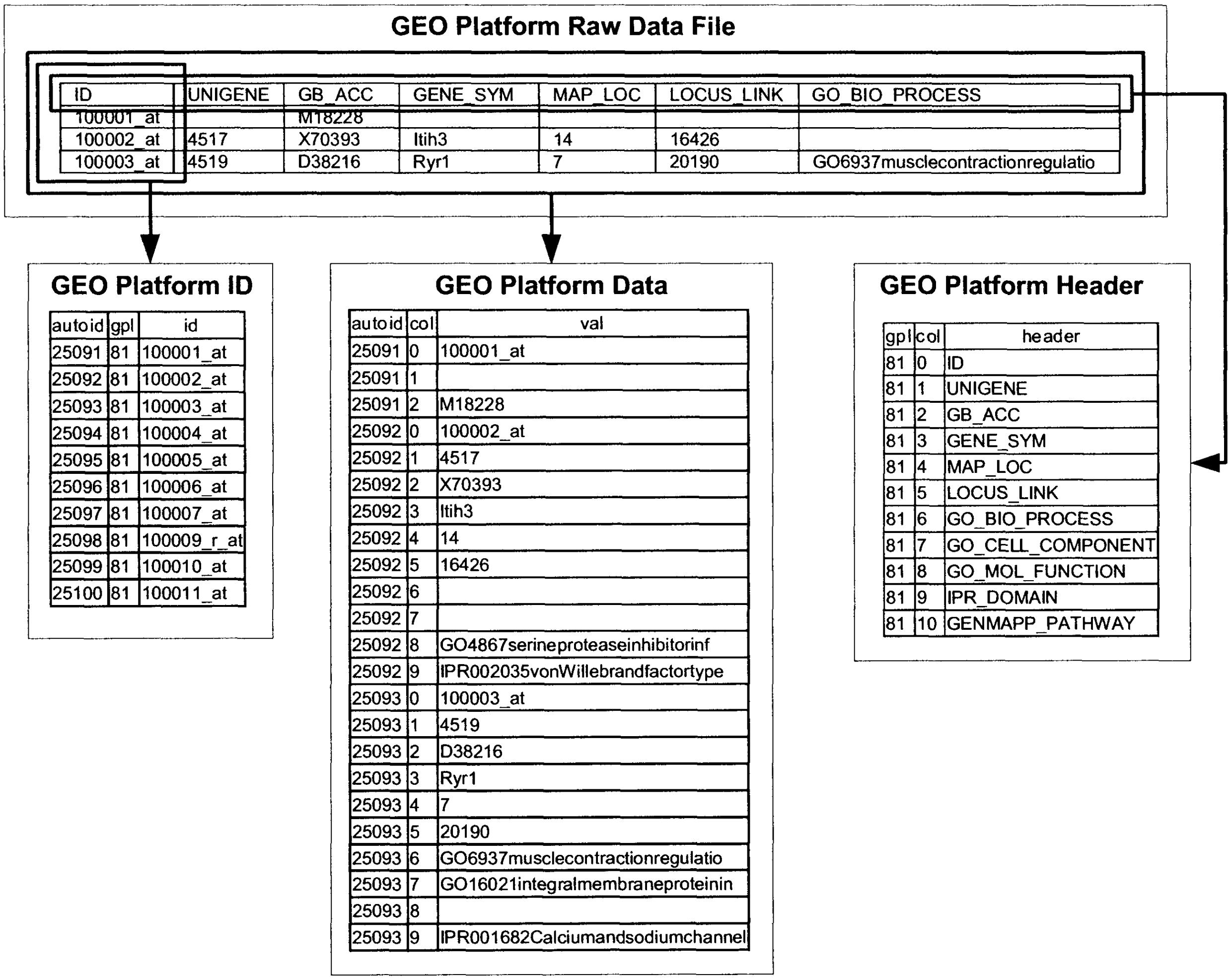


Figure 28: Mapping GEO expression measurements to LocusLink gene identifiers required creating three sets of relations.



GEO Platform Raw Data File

ID	UNIGENE	GB_ACC	GENE_SYM	MAP_LOC	LOCUS_LINK	GO_BIO_PROCESS
100001_at		M18228				
100002_at	4517	X70393	Itih3	14	16426	
100003_at	4519	D38216	Ryr1	7	20190	GO6937musclecontractionregulatio

GEO Platform ID

autoid	gpl	id
25091	81	100001_at
25092	81	100002_at
25093	81	100003_at
25094	81	100004_at
25095	81	100005_at
25096	81	100006_at
25097	81	100007_at
25098	81	100009_r_at
25099	81	100010_at
25100	81	100011_at

GEO Platform Data

autoid	col	val
25091	0	100001_at
25091	1	
25091	2	M18228
25092	0	100002_at
25092	1	4517
25092	2	X70393
25092	3	Itih3
25092	4	14
25092	5	16426
25092	6	
25092	7	
25092	8	GO4867serineproteaseinhibitorinf
25092	9	IPR002035vonWillebrandfactortype
25093	0	100003_at
25093	1	4519
25093	2	D38216
25093	3	Ryr1
25093	4	7
25093	5	20190
25093	6	GO6937musclecontractionregulatio
25093	7	GO16021integralmembraneproteinin
25093	8	
25093	9	IPR001682Calciumandsodiumchannel

GEO Platform Header

gpl	col	header
81	0	ID
81	1	UNIGENE
81	2	GB_ACC
81	3	GENE_SYM
81	4	MAP_LOC
81	5	LOCUS_LINK
81	6	GO_BIO_PROCESS
81	7	GO_CELL_COMPONENT
81	8	GO_MOL_FUNCTION
81	9	IPR_DOMAIN
81	10	GENMAPP_PATHWAY

Figure 29: Data from GEO platforms was extracted and separated into three tables: identifiers (ID), header information (Header), and all data (Data).

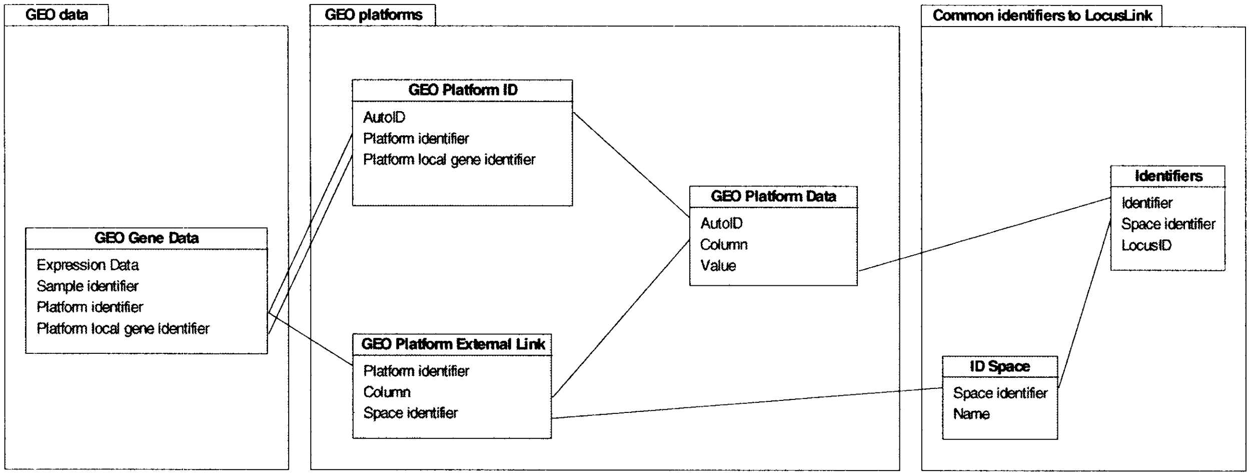


Figure 30: Database join across six tables relating gene expression data from GEO to LocusLink identifiers.

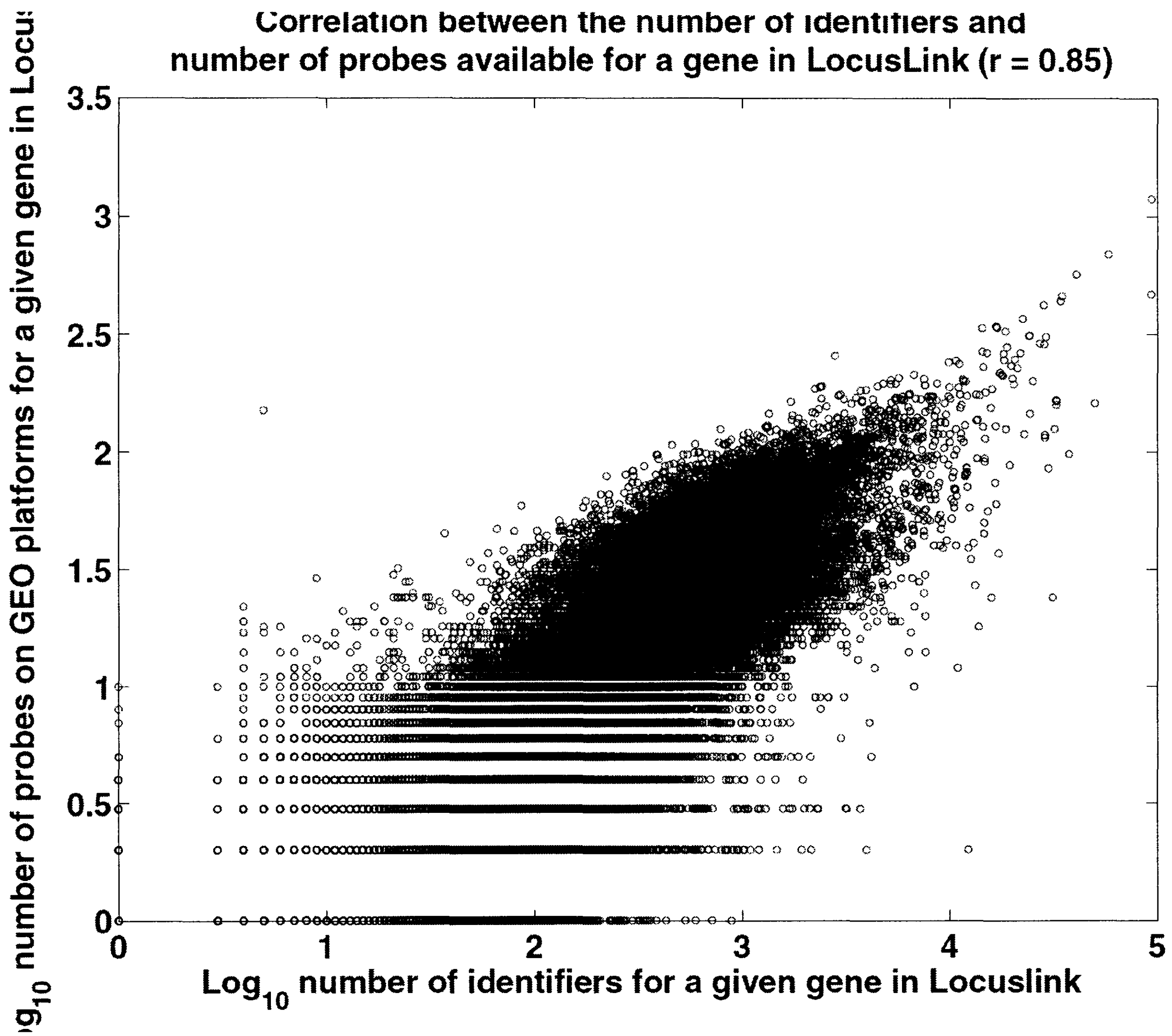


Figure 31: Plot between the log of the number of identifiers available for a gene in LocusLink and the log of the number of probes on GEO platforms for that gene, showing strong correlation.

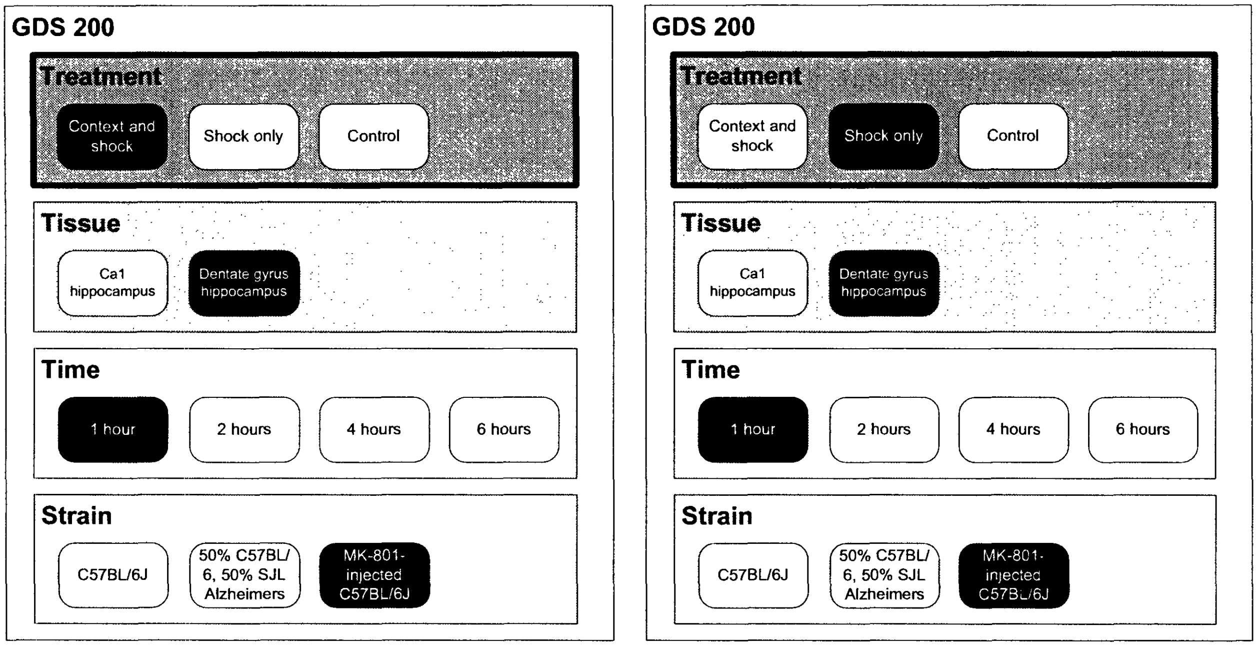


Figure 32: Example of two groups of samples being compared. The test-variable is Treatment, where one group is assigned “Context and shock” while the other is assigned “Shock only.” Values for the three background variables, Tissue, Time, and Strain, are set to equal values in the two groups.

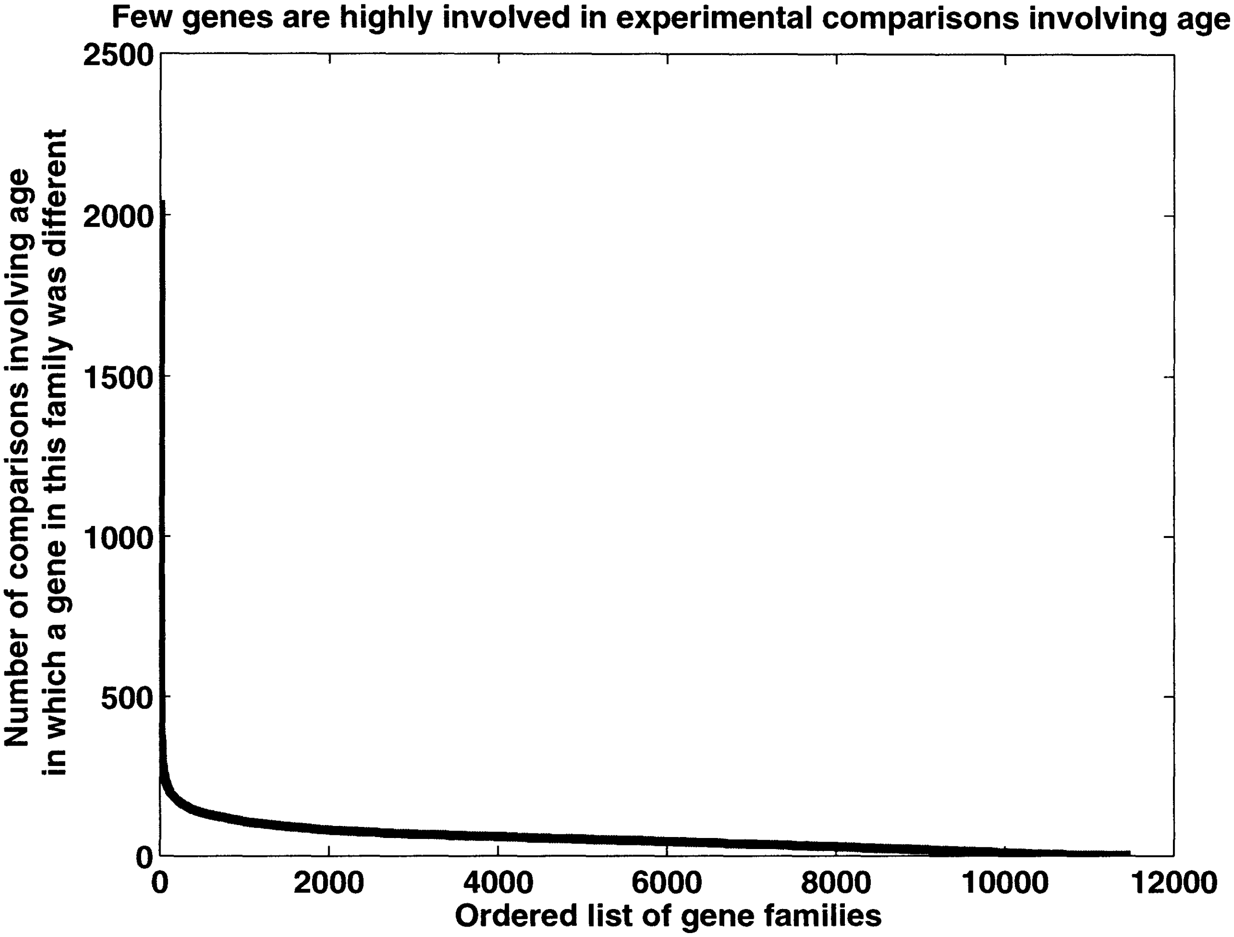


Figure 33: Over 11,000 homology families were implicated in one or more two-group comparisons involving the variable "age." Very few families were implicated in over 100 comparisons.

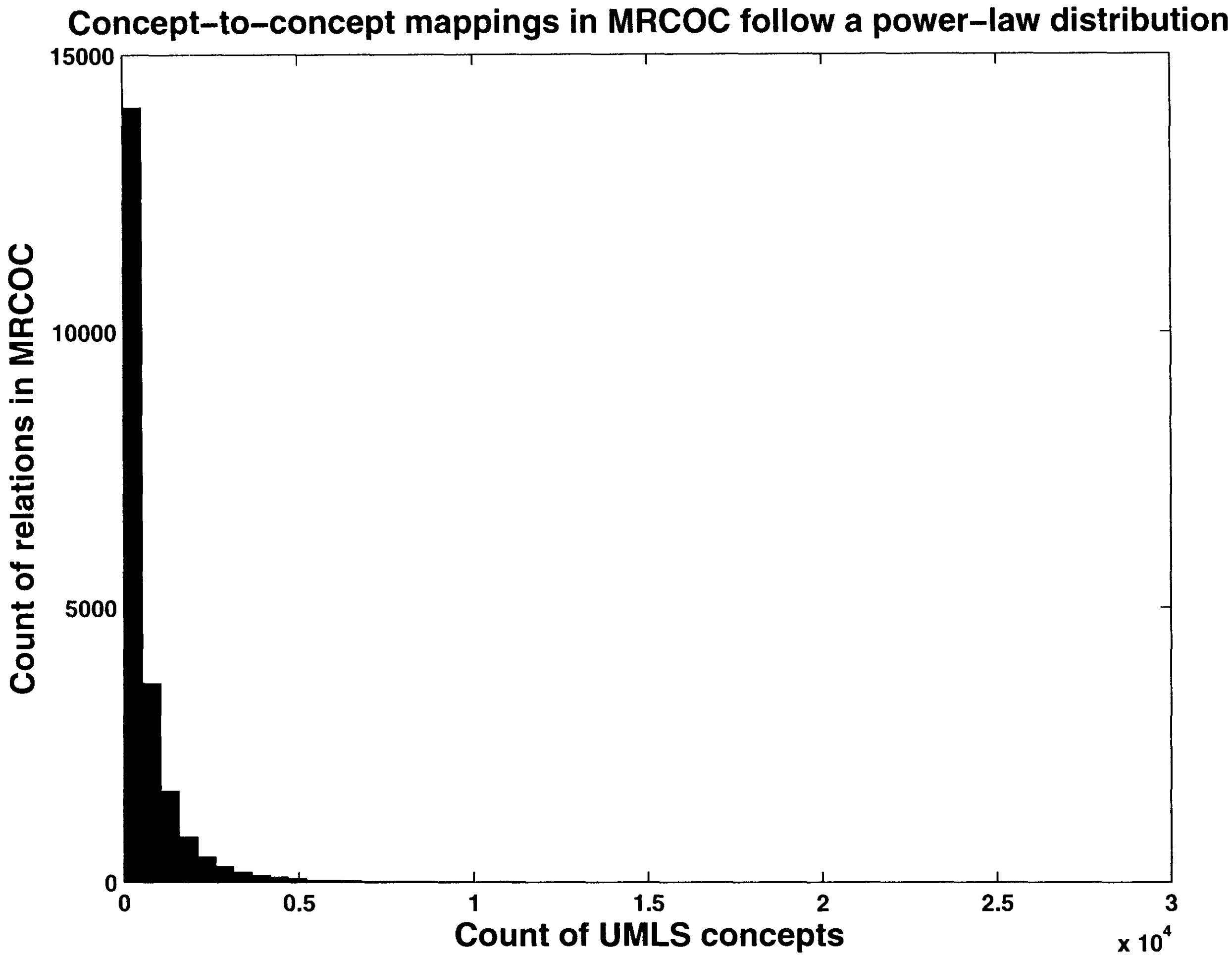


Figure 34: Plot of the count of relations in the UMLS co-occurrence table (MRCOC) across the number of concepts with each count. Relatively few concepts have over ten-thousand relations.

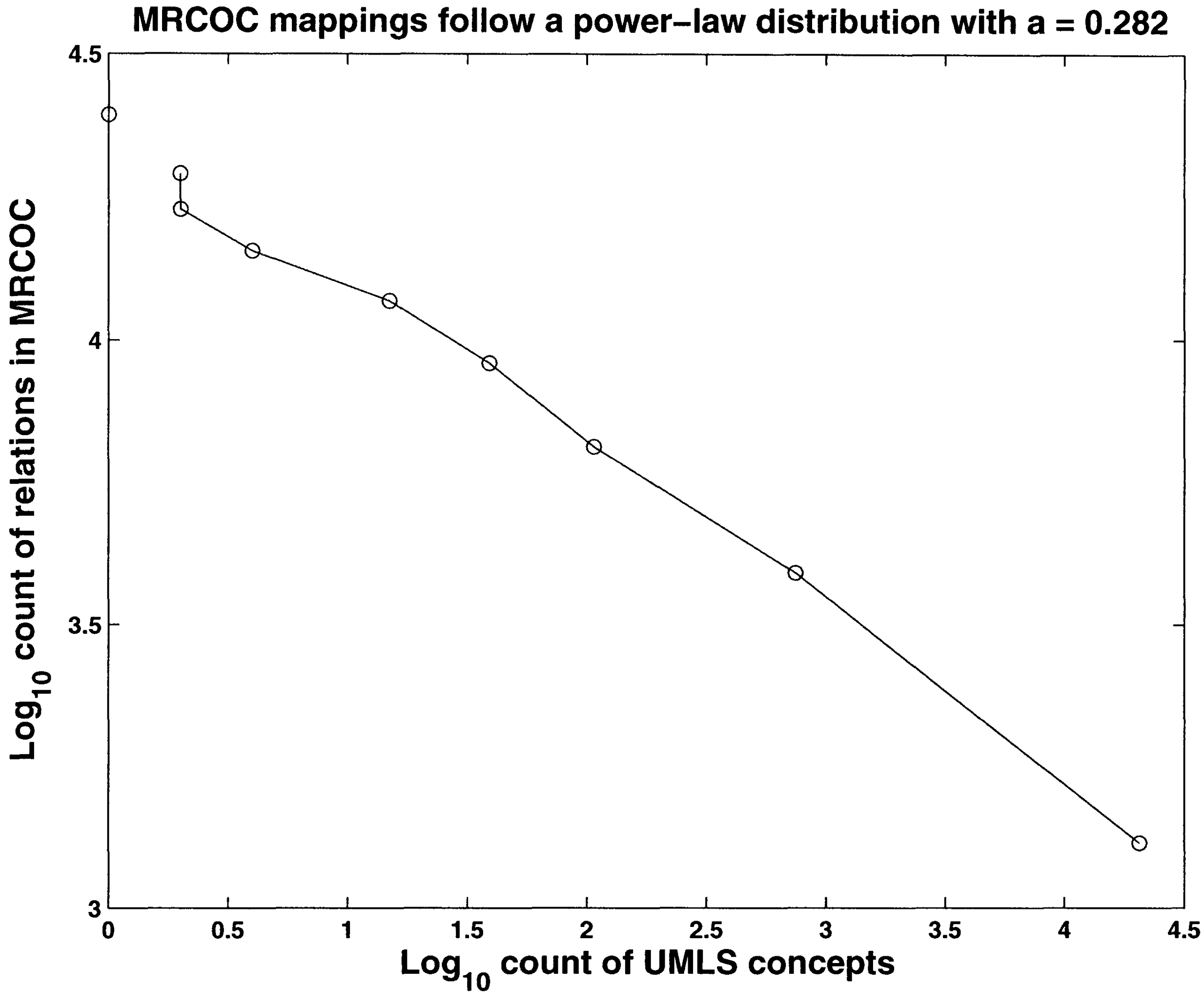


Figure 35: Plot of the log of count of relations in the UMLS co-occurrence table (MRCOC) across the log of the number of concepts with each count.

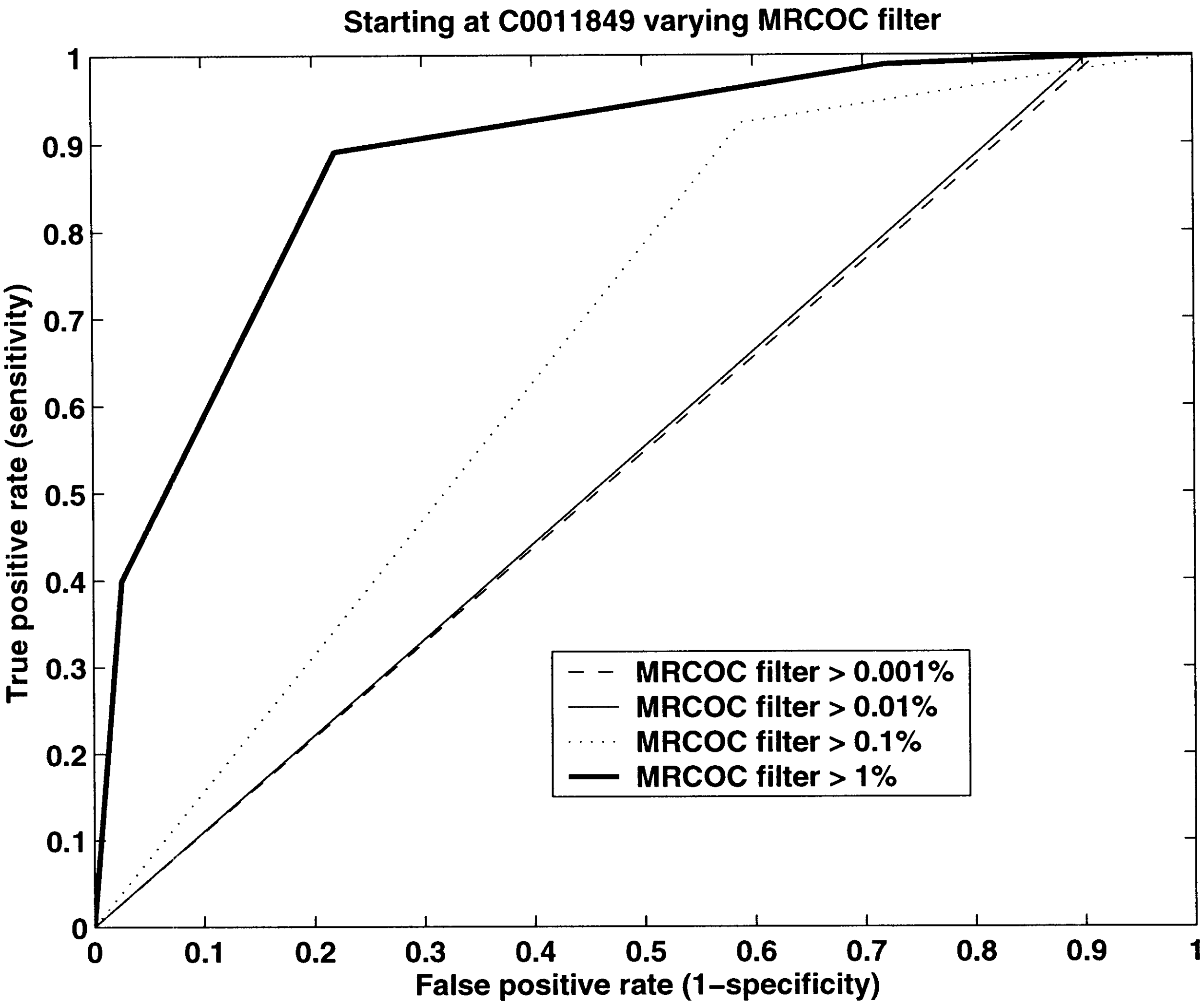


Figure 36: Receiver-operating characteristic curves indicating the effect on sensitivity and specificity as more relations are excluded from the UMLS MRCOC co-occurrence table. The most optimal curve was in the strategy excluding the least significant 1% of the relations.

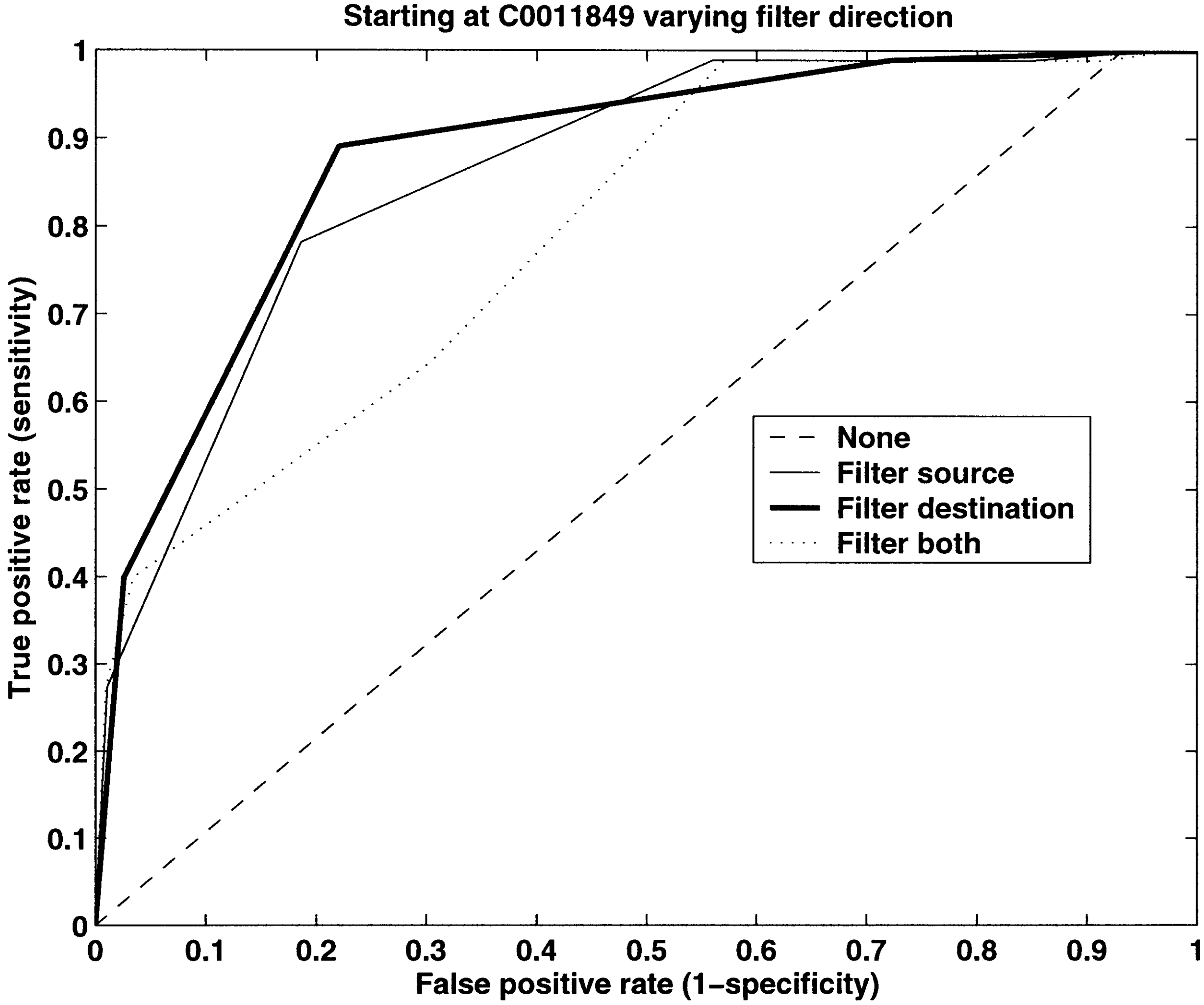


Figure 37: Receiver-operating characteristic curves indicating the effect on sensitivity and specificity as co-occurrence significance is evaluated in the source, destination, or both concepts during traversal. The most optimal curve was in the strategy evaluating co-occurrence significant in the destination concept.

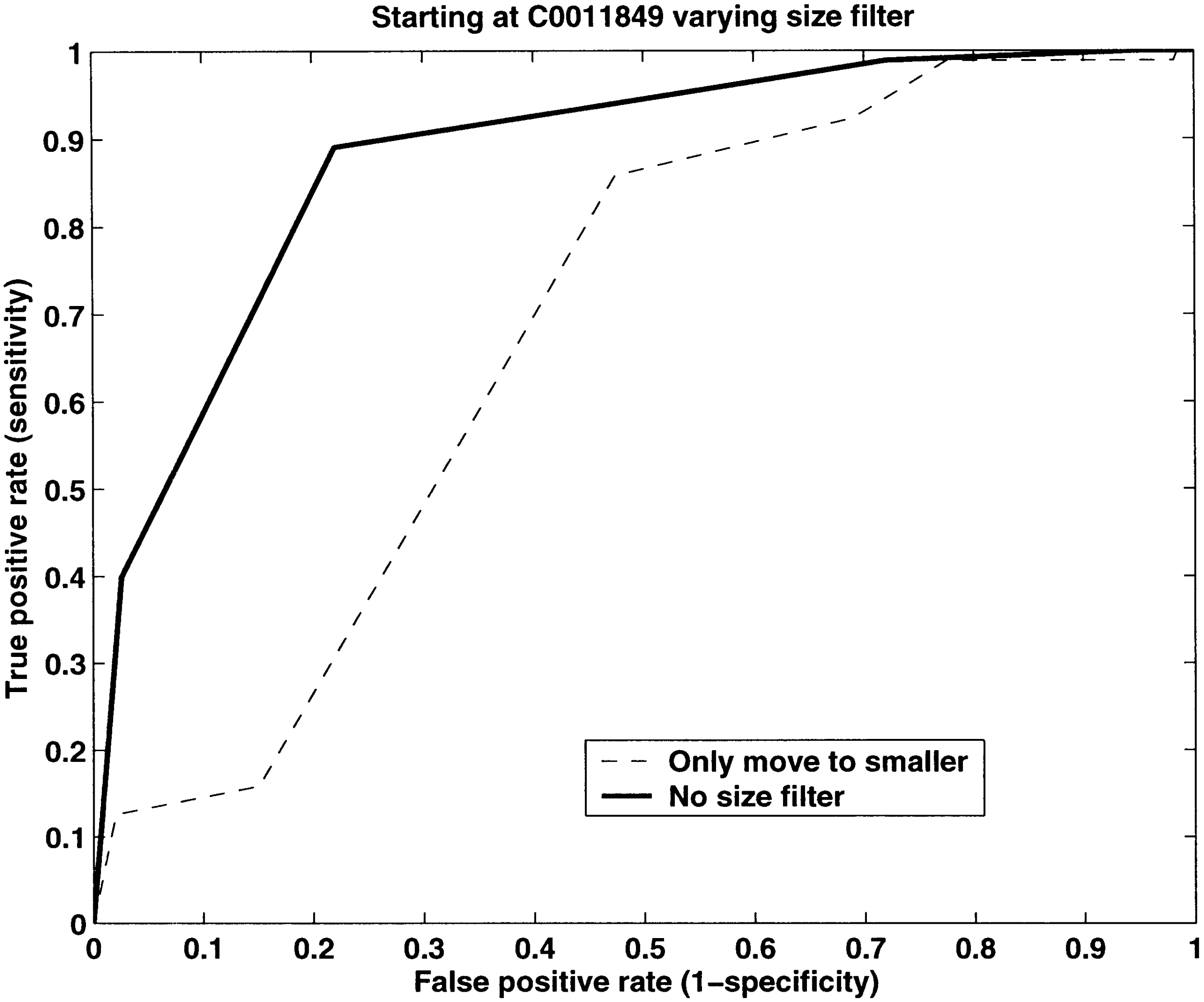


Figure 38: Receiver-operating characteristic curves indicating the effect on sensitivity and specificity when traversal is restricted to moving to concepts with fewer citations. The most optimal curve was the strategy without this restriction.

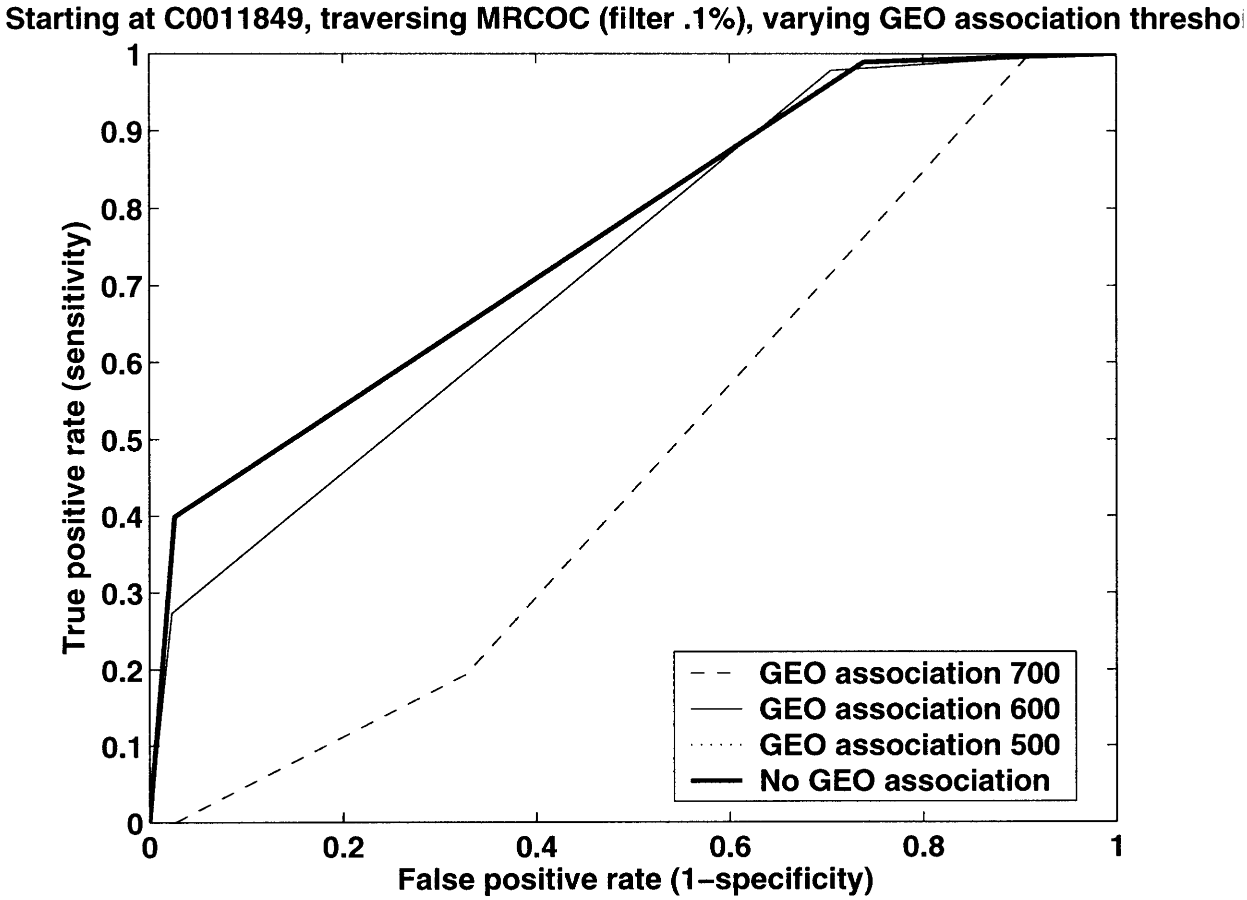


Figure 39: Receiver-operating characteristic curves indicating the effect on sensitivity and specificity when traversal to GEO samples, series and data sets is restricted to relations with a MetaMap score meeting a threshold. The most optimal curve was the strategy without this restriction.

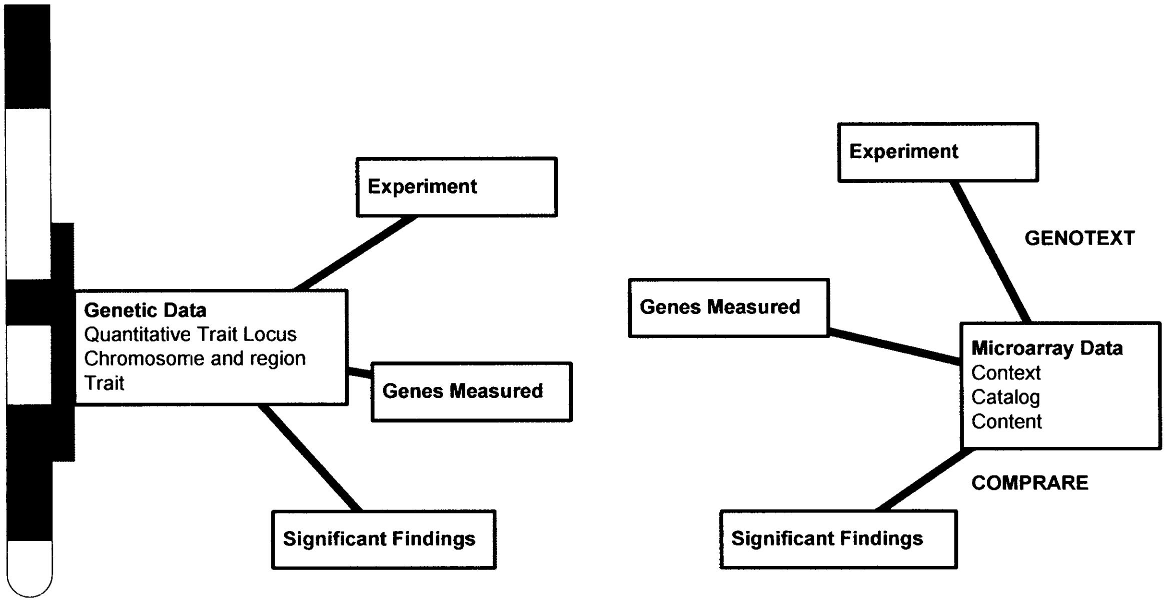


Figure 40: Genetic data and microarray data can be modeled in terms of their experimental context, gene measurement catalog, and content of significant findings. Without additional knowledge, automated intersection of these two modalities is not possible.

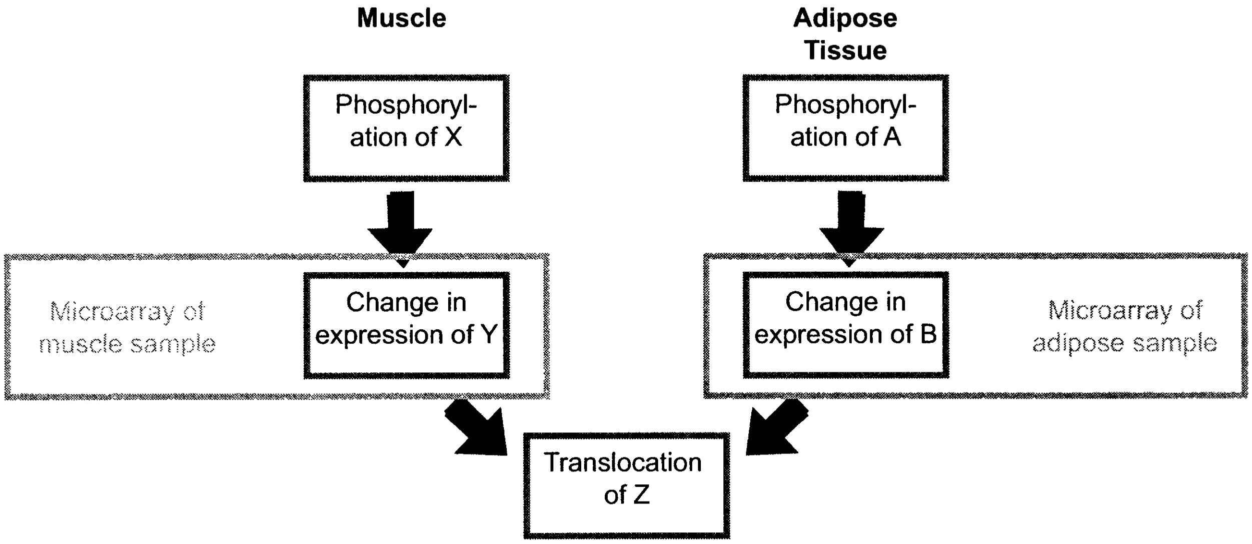

Muscle
Adipose Tissue
Phosphorylation of X
Phosphorylation of A
Microarray of muscle sample
Change in expression of Y
Change in expression of B
Microarray of adipose sample
Translocation of Z

Figure 41: Two experiments may have examined small non-overlapping portions of a large biological process. Though several genes and proteins may be involved in the same large process, the intersection will not retrieve them.

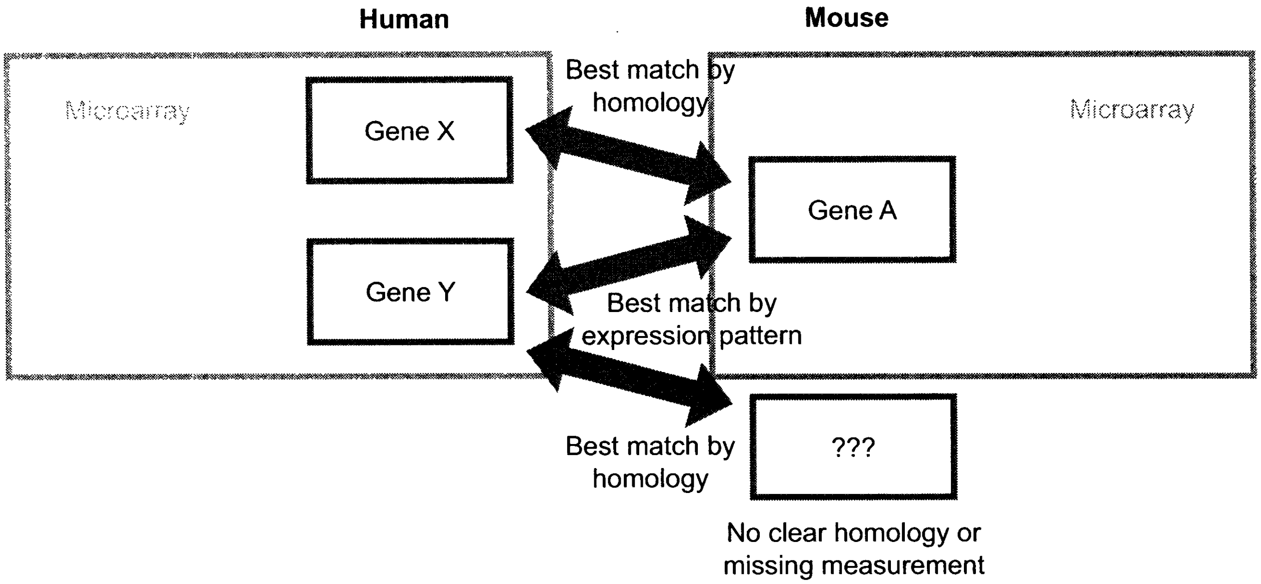


Figure 42: If genes are implicated in a process in one data set but are not even measured in another, these genes will not be present in the intersection. Differing equivalence relations between species will also alter intersection results.

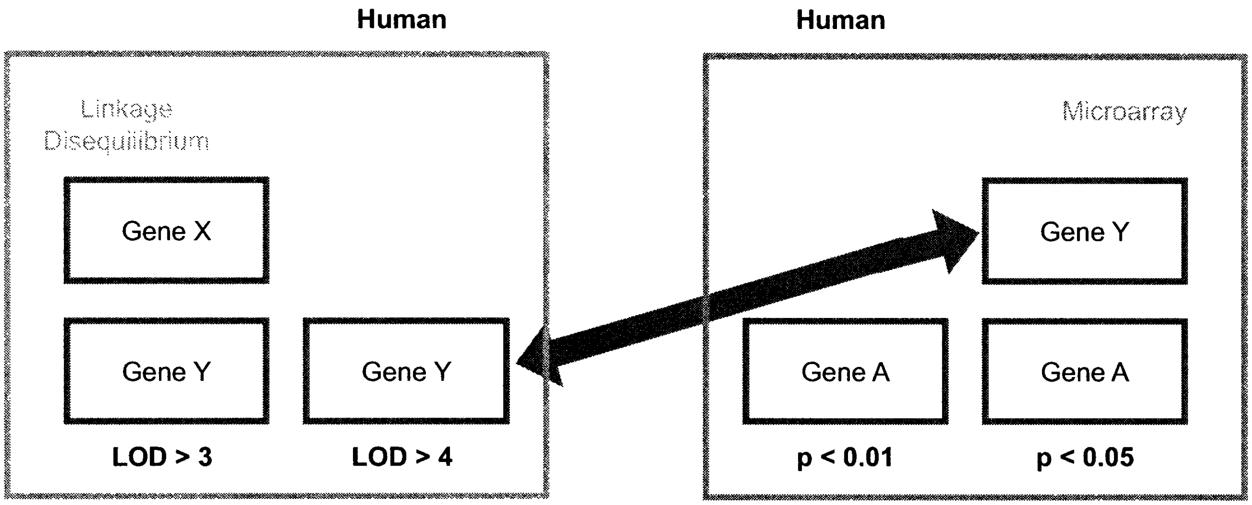


Figure 43: Intersection relations between two experiments depend on the parameters and analytic method used in each experiment.

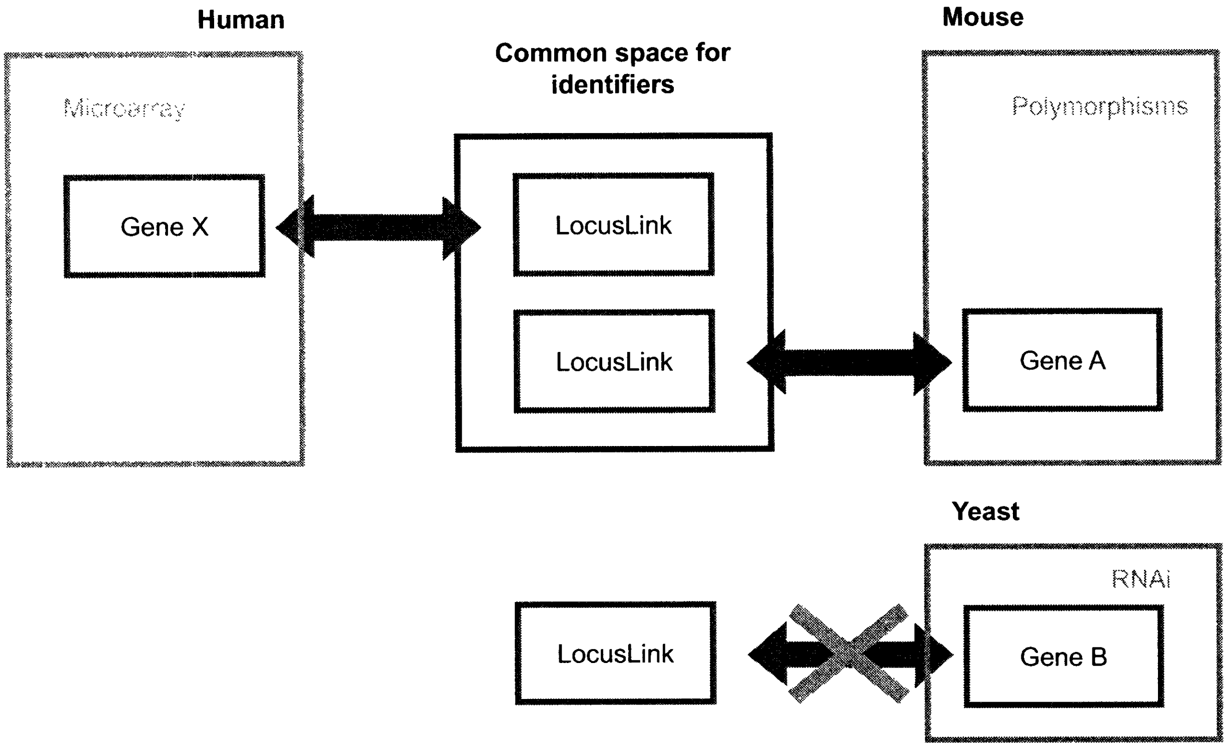


Figure 44: Finding intersection relations between species is difficult when a common identification scheme is not available.

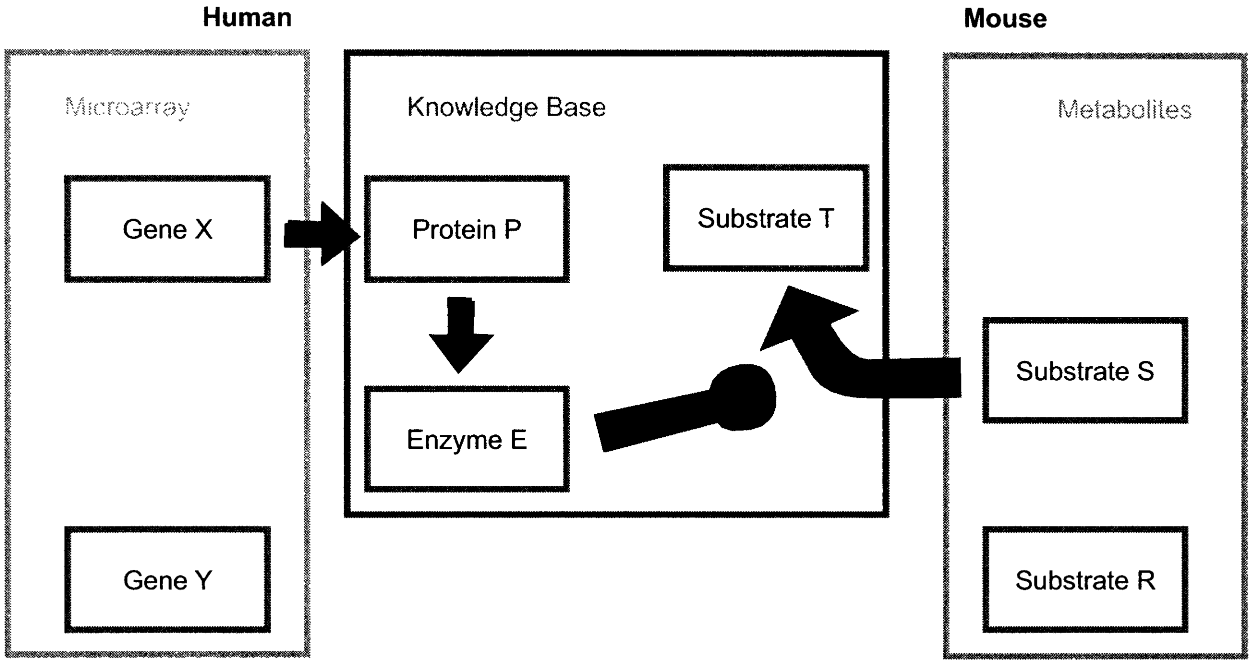


Figure 45: Intersection between some types of nearly-comprehensive data sets may not be possible without the application of *a priori* knowledge to transduce one modality to another.

Appendix

Prolog program that implements the reasoning that provides a biological explanation of the trait of a quantitative trait locus.

```
/* relating traits to diseases */

relation( concept(c1260441), rel("diagnoses"), concept(c0011849)).


/* relating genes to causative pathways to diseases */

pathway( decrease(hid(2517)), rel("causes"), decrease(concept(c0013846)) ).
pathway( decrease(concept(c0013846)), rel("causes"),
increase(concept(c0015684)) ).
pathway( increase(concept(c0015684)), rel("causes"), concept(c0021655) ).
pathway( concept(c0021655), rel("causes"), concept(c0011849) ).

qtl( qtl_symb("Niddm40"), species(c0034693), concept(c1260441), chrom(1),
range(139105187,167448551) ).
qtl( qtl_symb("Niddm23"), species(c0034693), concept(c1260441), chrom(1),
range(203569474,203569596) ).
qtl( qtl_symb("Niddm2"), species(c0034693), concept(c1260441), chrom(2),
range(197218978,197219067) ).
qtl( qtl_symb("Niddm3"), species(c0034693), concept(c1260441), chrom(10),
range(86331202,86331424) ).
qtl( qtl_symb("Niddm22"), species(c0034693), concept(c1260441), chrom(11),
range(84600690,84600838) ).
qtl( qtl_symb("Niddm18"), species(c0034693), concept(c1260441), chrom(14),
range(32584631,32584754) ).
qtl( qtl_symb("Niddm28"), species(c0034693), concept(c1260441), chrom(14),
range(81225172,81225329) ).
qtl( qtl_symb("Niddm29"), species(c0034693), concept(c1260441), chrom(16),
range(19510771,80047391) ).
qtl( qtl_symb("Niddm32"), species(c0034693), concept(c1260441), chrom(17),
range(59107689,59107785) ).

related_concepts( C1, C1, TRAIL ).
related_concepts( C1, C3, TRAIL ) :-
      ( relation( C1, _, C2 ) ; relation( C2, _, C1 ) ),
      legal( C2, TRAIL ),
      related_concepts( C2, C3, [C2|TRAIL] ).


biological_causal( H1, H1, TRAIL ).
biological_causal( H1, H3, TRAIL ) :-
      pathway( H1, _, H2 ),
      legal(H2, TRAIL),
      biological_causal( H2, H3, [H2|TRAIL] ), !.

legal( C, [] ) :- !.
legal( C, [H|T] ) :- C \== H, legal(C,T).
```

```
get_gene_symbol( LOCID, SYMBOL ) :-
      gene(LOCID,SYMBOL,_,_,_).

range_within_range(range(SMLOW,SMHIGH),range(LGLOW,LGHIGH)) :- SMLOW >=
LGLOW,
      SMLOW =< LGHIGH,
      !.

gene_in_qtl(LOCID, QSYMB) :- qtl(qtl_symb(QSYMB), QSPECIES, _, QCHROM,
QRANGE),
      gene(LOCID,_,QSPECIES,QCHROM,GRANGE),
      range_within_range(GRANGE,QRANGE).

explain_qtl_symbol(QSYMB,QTL_GENE,SIG_GENE) :- gene_in_qtl(QTL_GENE, QSYMB),
      qtl(qtl_symb(QSYMB), _, QTRAIT, _, _),
      gene_homology( H1, QTL_GENE ),
      related_concepts( QTRAIT, CONCEPT2, TRAIL1 ),
      expr_context( GDS, CONCEPT2, DEPTH ),
      DEPTH =< 3,
      biological_causal( decrease(H1), CONCEPT2, TRAIL2 ),
      expr_sig( GDS, SIG_GENE ),
      gene_homology( H1, SIG_GENE ).
```

References

1. Moir, R. D. & Spann, T. P. The structure and function of nuclear lamins: implications for disease. *Cell Mol Life Sci* **58**, 1748-57 (2001).
2. Ellis, L. B., Speedie, S. M. & McLeish, R. Representing metabolic pathway information: an object-oriented approach. *Bioinformatics* **14**, 803-6 (1998).
3. Henrich, T. et al. MEPD: a Medaka gene expression pattern database. *Nucleic Acids Res* **31**, 72-4 (2003).
4. Guttmacher, A. E. & Collins, F. S. Genomic medicine--a primer. *N Engl J Med* **347**, 1512-20 (2002).
5. Golub, T. R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-7 (1999).
6. Alizadeh, A. A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-11 (2000).
7. Lynch, T. J. et al. Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib. *N Engl J Med* (2004).
8. Paez, J. G. et al. EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy. *Science* (2004).
9. Hood, L. Leroy Hood expounds the principles, practice and future of systems biology. *Drug Discov Today* **8**, 436-8 (2003).
10. Zeitlinger, J. et al. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395-404 (2003).
11. Saban, R., Saban, M. R., Nguyen, N. B., Hammond, T. G. & Wershil, B. K. Mast cell regulation of inflammation and gene expression during antigen-induced bladder inflammation in mice. *Physiol Genomics* **7**, 35-43 (2001).
12. Cadet, J. L., Jayanthi, S., McCoy, M. T., Vawter, M. & Ladenheim, B. Temporal profiling of methamphetamine-induced changes in gene expression in the mouse brain: evidence from cDNA array. *Synapse* **41**, 40-8 (2001).
13. Amundson, S. A., Bittner, M., Meltzer, P., Trent, J. & Fornace, A. J., Jr. Induction of gene expression as a monitor of exposure to ionizing radiation. *Radiat Res* **156**, 657-61 (2001).
14. Zapala, M. A., Lockhart, D. J., Pankratz, D. G., Garcia, A. J. & Barlow, C. Software and methods for oligonucleotide and cDNA array data analysis. *Genome Biol* **3**, SOFTWARE0001 (2002).
15. Conway, A. R. *GeneSpring User Manual version 4.2* (Silicon Genetics, Redwood City, CA, 2002).
16. Metcalf, B. From the Ether Metcalfe's Law: A network becomes more valuable as it reaches more users. *InfoWorld* **17**, 53 (1995).
17. Chien, K. R. Genomic circuits and the integrative biology of cardiac diseases. *Nature* **407**, 227-32 (2000).
18. Palsson, B. The challenges of in silico biology. *Nat Biotechnol* **18**, 1147-50 (2000).

19. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2**, 343-72 (2001).
20. Delneri, D., Brancia, F. L. & Oliver, S. G. Towards a truly integrative biology through the functional genomics of yeast. *Curr Opin Biotechnol* **12**, 87-91 (2001).
21. Oltvai, Z. N. & Barabasi, A. L. Systems biology. Life's complexity pyramid. *Science* **298**, 763-4 (2002).
22. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662-4 (2002).
23. Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295**, 1664-9 (2002).
24. Ge, H., Walhout, A. J. & Vidal, M. Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet* **19**, 551-60 (2003).
25. Walhout, A. J. et al. Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline. *Curr Biol* **12**, 1952-8 (2002).
26. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193-7 (2003).
27. Kafatos, F. C. & Eisner, T. Unification in the century of biology. *Science* **303**, 1257 (2004).
28. Center for Cancer Research Office of Communication. (National Cancer Institute, 2003).
29. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31**, 28-33 (2003).
30. Diehn, M. et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* **31**, 219-23 (2003).
31. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* **29**, 137-40. (2001).
32. Brazma, A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* **29**, 365-71. (2001).
33. Spellman, P. T. et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* **3**, RESEARCH0046 (2002).
34. Hewett, M. et al. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* **30**, 163-5. (2002).
35. Taylor, C. F. et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat Biotechnol* **21**, 247-54 (2003).
36. Berman, J. J., Edgerton, M. E. & Friedman, B. A. The tissue microarray data exchange specification: a community-based, open source tool for sharing tissue microarray data. *BMC Med Inform Decis Mak* **3**, 5 (2003).
37. Hermjakob, H. et al. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol* **22**, 177-83 (2004).
38. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-10. (2002).
39. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res* **30**, 13-6. (2002).
40. Wu, C. H. et al. The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res* **30**, 35-7. (2002).
41. Stein, L. Creating a bioinformatics nation. *Nature* **417**, 119-20 (2002).

42. Birney, E. et al. Ensembl 2004. *Nucleic Acids Res* **32 Database issue**, D468-70 (2004).
43. Kuo, W. P., Jenssen, T. K., Butte, A. J., Ohno-Machado, L. & Kohane, I. S. Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18**, 405-12. (2002).
44. Nimgaonkar, A. et al. Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics* **4**, 27 (2003).
45. Tan, P. K. et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-84 (2003).
46. Lee, Y. et al. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res* **12**, 493-502 (2002).
47. Ball, C. A. et al. Standards for microarray data. *Science* **298**, 539 (2002).
48. Ross, D. T. et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat Genet* **24**, 227-35 (2000).
49. Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R. & Kohane, I. S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A* **97**, 12182-6 (2000).
50. Rosen, E. D., Walkey, C. J., Puigserver, P. & Spiegelman, B. M. Transcriptional regulation of adipogenesis. *Genes Dev* **14**, 1293-307 (2000).
51. Ahima, R. S. & Flier, J. S. Adipose tissue as an endocrine organ. *Trends Endocrinol Metab* **11**, 327-32 (2000).
52. Steppan, C. M. & Lazar, M. A. Resistin and obesity-associated insulin resistance. *Trends Endocrinol Metab* **13**, 18-23 (2002).
53. Chawla, A. & Lazar, M. A. Peroxisome proliferator and retinoid signaling pathways co-regulate preadipocyte phenotype and survival. *Proc Natl Acad Sci U S A* **91**, 1786-90 (1994).
54. Tontonoz, P., Hu, E. & Spiegelman, B. M. Stimulation of adipogenesis in fibroblasts by PPAR gamma 2, a lipid-activated transcription factor. *Cell* **79**, 1147-56 (1994).
55. Rosen, E. D. The molecular control of adipogenesis, with special reference to lymphatic pathology. *Ann N Y Acad Sci* **979**, 143-58; discussion 188-96 (2002).
56. Altiok, S., Xu, M. & Spiegelman, B. M. PPARgamma induces cell cycle withdrawal: inhibition of E2F/DP DNA-binding activity via down-regulation of PP2A. *Genes Dev* **11**, 1987-98 (1997).
57. MacDougald, O. A. & Mandrup, S. Adipogenesis: forces that tip the scales. *Trends Endocrinol Metab* **13**, 5-11 (2002).
58. De Sandre-Giovannoli, A. et al. Lamin A Truncation in Hutchinson-Gilford Progeria. *Science* (2003).
59. Eriksson, M. et al. Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293-8 (2003).
60. Rosenbloom, A. L. et al. Progeria: insulin resistance and hyperglycemia. *J Pediatr* **102**, 400-2 (1983).
61. Briata, P., Bellini, C., Vignolo, M. & Gherzi, R. Insulin receptor gene expression is reduced in cells from a progeric patient. *Mol Cell Endocrinol* **75**, 9-14 (1991).

62. Elsas, L. J. & Longo, N. Impaired insulin binding and excess glucose transport in fibroblasts from a patient with leprechaunism. *Enzyme* **38**, 184-93 (1987).
63. Ashrafi, K. et al. Genome-wide RNAi analysis of Caenorhabditis elegans fat regulatory genes. *Nature* **421**, 268-72 (2003).
64. Ross, S. E. et al. Inhibition of adipogenesis by Wnt signaling. *Science* **289**, 950-3 (2000).
65. Doniger, S. W. et al. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* **4**, R7 (2003).
66. Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. & Conklin, B. R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* **31**, 19-20 (2002).
67. *International Classification of Diseases: 9th revision, Clinical Modification (ICD-9-CM)* (Centers for Medicare & Medicaid Services, Washington DC, 2003).
68. Liu, G. et al. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* **31**, 82-6 (2003).
69. Tsai, J. et al. RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biology* **2**, 1-4 (2001).
70. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9. (2000).
71. Karp, P. D., Riley, M., Paley, S. M. & Pelligrini-Toole, A. EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Res* **24**, 32-9 (1996).
72. Karp, P. D., Riley, M., Paley, S. M. & Pellegrini-Toole, A. The MetaCyc Database. *Nucleic Acids Res* **30**, 59-61 (2002).
73. Clocksin, W. F. & Mellish, C. S. *Programming in Prolog* (Springer-Verlag, Berlin ; New York, 1987).
74. Primig, M. et al. The core meiotic transcriptome in budding yeasts. *Nat Genet* **26**, 415-23. (2000).
75. Chu, S. et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).
76. Storch, K. F. et al. Extensive and divergent circadian gene expression in liver and heart. *Nature* **417**, 78-83 (2002).
77. Panda, S. et al. Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell* **109**, 307-20 (2002).
78. Fruman, D. A. et al. Phosphoinositide 3-kinase and Bruton's tyrosine kinase regulate overlapping sets of genes in B lymphocytes. *Proc Natl Acad Sci U S A* **99**, 359-64 (2002).
79. Whitney, A. R. et al. Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci U S A* **100**, 1896-901 (2003).
80. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868 (1998).
81. Hughes, T. R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26 (2000).

82. Weinstein, J. N. et al. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343-349 (1997).
83. Scherf, U. et al. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* **24**, 236-44 (2000).
84. Park, T., Chen, Z. P. & Leavitt, J. Activation of the leukocyte plastin gene occurs in most human cancer cells. *Cancer Res* **54**, 1775-1781 (1994).
85. Prevost, G. P. et al. Inhibition of human tumor cell growth In vitro and In vivo by a specific inhibitor of human farnesyltransferase: BIM-46068. *Int J Cancer* **83**, 283-287 (1999).
86. Mootha, V. K. et al. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. *Proc Natl Acad Sci U S A* **100**, 605-10 (2003).
87. Stoll, M. et al. A genomic-systems biology map for cardiovascular function. *Science* **294**, 1723-6 (2001).
88. Rolinski, B. et al. The biochemical metabolite screen in the Munich ENU Mouse Mutagenesis Project: determination of amino acids and acylcarnitines by tandem mass spectrometry. *Mamm Genome* **11**, 547-51 (2000).
89. Ross-Macdonald, P. et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413-8 (1999).
90. Schadt, E. E., Monks, S. A. & Friend, S. H. A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochem Soc Trans* **31**, 437-43 (2003).
91. Schadt, E. E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).
92. Vidal, M. A biological atlas of functional maps. *Cell* **104**, 333-9 (2001).
93. Arkin, A. P. Synthetic cell biology. *Curr Opin Biotechnol* **12**, 638-44 (2001).
94. Shapiro, B. E., Levchenko, A., Meyerowitz, E. M., Wold, B. J. & Mjolsness, E. D. Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations. *Bioinformatics* **19**, 677-8 (2003).
95. Schaff, J. & Loew, L. M. The virtual cell. *Pac Symp Biocomput*, 228-39 (1999).
96. Tomita, M. et al. E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72-84 (1999).
97. Karp, P. D. & Riley, M. in *Bioinformatics : databases and systems* (ed. Letovsky, S.) (Kluwer Academic Publishers, Boston, 1999).
98. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**, 42-6 (2002).
99. Michal, G. *Biochemical Pathways Wall Chart* (Boehringer Mannheim GmbH Biochemica and Spektrum Akademischer Verlag GmbH, Heidelberg, Germany, 1982).
100. Hofestadt, R. & Thelen, S. Quantitative modeling of biochemical networks. *In Silico Biol* **1**, 39-53 (1998).
101. Reddy, V. N., Mavrovouniotis, M. L. & Liebman, M. N. Petri net representations in metabolic pathways. *Proc Int Conf Intell Syst Mol Biol* **1**, 328-36 (1993).
102. Overbeek, R. et al. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* **28**, 123-5 (2000).

103. Ellis, L. B., Hou, B. K., Kang, W. & Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining. *Nucleic Acids Res* **31**, 262-5 (2003).

104. Badea, L. Functional discrimination of gene expression patterns in terms of the gene ontology. *Pac Symp Biocomput*, 565-76 (2003).

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

106. Zien, A., Kuffner, R., Zimmer, R. & Lengauer, T. Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol* **8**, 407-17 (2000).

107. Kohn, K. W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol Biol Cell* **10**, 2703-34 (1999).

108. Toyoda, T. & Konagaya, A. KnowledgeEditor: a new tool for interactive modeling and analyzing biological pathways based on microarray data. *Bioinformatics* **19**, 433-4 (2003).

109. Krishnamurthy, L. et al. Pathways database system: an integrated system for biological pathways. *Bioinformatics* **19**, 930-7 (2003).

110. Salamonsen, W., Mok, K. Y., Kolatkar, P. & Subbiah, S. BioJAKE: a tool for the creation, visualization and manipulation of metabolic pathways. *Pac Symp Biocomput*, 392-400 (1999).

111. Gilman, A. G. et al. Overview of the Alliance for Cellular Signaling. *Nature* **420**, 703-6 (2002).

112. Boeckmann, B. et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365-70 (2003).

113. Toyoda, T., Mochizuki, Y. & Konagaya, A. GSCope: a clipped fisheye viewer effective for highly complicated biomolecular network graphs. *Bioinformatics* **19**, 437-8 (2003).

114. Ideker, T. & Lauffenburger, D. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol* **21**, 255-62 (2003).

115. Ng, S. K., Zhang, Z., Tan, S. H. & Lin, K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res* **31**, 251-4 (2003).

116. Mulder, N. J. et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res* **31**, 315-8 (2003).

117. Attwood, T. K. et al. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* **31**, 400-2 (2003).

118. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **28**, 10-4. (2000).

119. Lee, K., Kohane, I. S. & Butte, A. J. PGAGENE: integrating quantitative gene-specific results from the NHLBI Programs for Genomic Applications. *Bioinformatics* **19**, 778-9 (2003).

120. Milburn, J. Beyond the genome: turning data into knowledge. *Drug Discov Today* **6**, 881-883. (2001).

121. DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-6 (1997).

122. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. *J Comput Biol* **7**, 601-20 (2000).
123. Arkin, A., Shen, P. & Ross, J. A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements. *Science* **277**, 1275-9 (1997).
124. Jungmann, R. A., Huang, D. & Tian, D. Regulation of LDH-A gene expression by transcriptional and posttranscriptional signal transduction mechanisms. *J Exp Zool* **282**, 188-95 (1998).
125. Alcazar, O., Tiedge, M. & Lenzen, S. Importance of lactate dehydrogenase for the regulation of glycolytic flux and insulin secretion in insulin-producing cells. *Biochem J* **352 Pt 2**, 373-80 (2000).
126. Krauthammer, M. et al. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics* **18 Suppl 1**, S249-S257 (2002).
127. Shatkay, H., Edwards, S., Wilbur, W. J. & Boguski, M. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc Int Conf Intell Syst Mol Biol* **8**, 317-28 (2000).
128. Stevens, R., Goble, C. A. & Bechhofer, S. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform* **1**, 398-414 (2000).
129. Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* **28**, 21-8. (2001).
130. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* **32 Database issue**, D35-40 (2004).
131. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32 Database issue**, D267-70 (2004).
132. Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, 17-21 (2001).
133. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484-7. (1995).
134. Peng, W. T. et al. A panoramic view of yeast noncoding RNA processing. *Cell* **113**, 919-33 (2003).
135. Tenen, D. G. Disruption of differentiation in human cancer: AML shows the way. *Nat Rev Cancer* **3**, 89-101 (2003).
136. Cheung, V. G. et al. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet* **33**, 422-5 (2003).
137. Seth, P., Krop, I., Porter, D. & Polyak, K. Novel estrogen and tamoxifen induced genes identified by SAGE (Serial Analysis of Gene Expression). *Oncogene* **21**, 836-43 (2002).
138. Wang, P. et al. ProbeMatchDB--a web database for finding equivalent probes across microarray platforms and species. *Bioinformatics* **18**, 488-9 (2002).
139. Bussey, K. J. et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol* **4**, R27 (2003).
140. Holzenberger, M. et al. IGF-1 receptor regulates lifespan and resistance to oxidative stress in mice. *Nature* **421**, 182-7 (2003).

141. Bluher, M. et al. Adipose tissue selective insulin receptor knockout protects against obesity and obesity-related glucose intolerance. *Dev Cell* **3**, 25-38 (2002).
142. Kimura, K. D., Tissenbaum, H. A., Liu, Y. & Ruvkun, G. daf-2, an insulin receptor-like gene that regulates longevity and diapause in Caenorhabditis elegans. *Science* **277**, 942-6 (1997).
143. Anderson, R. M. et al. Manipulation of a nuclear NAD+ salvage pathway delays aging without altering steady-state NAD+ levels. *J Biol Chem* **277**, 18881-90 (2002).
144. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368-75 (2003).
145. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-21. (2001).
146. Baldi, P. & Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* **17**, 509-19 (2001).
147. Reis, B. Y., Butte, A. J. & Kohane, I. S. Extracting knowledge from dynamics in gene expression. *J Biomed Inform* **34**, 15-27. (2001).
148. Butte, A. J., Bao, L., Reis, B. Y., Watkins, T. W. & Kohane, I. S. Comparing the similarity of time-series gene expression using signal processing metrics. *Journal of Biomedical Informatics* **34**, 396-405 (2002).
149. Adamic, L. A., Lukose, R. M., Puniyani, A. R. & Huberman, B. A. Search in power-law networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **64**, 046135 (2001).
150. Skolnik, E. Y. et al. The function of GRB2 in linking the insulin receptor to Ras signaling pathways. *Science* **260**, 1953-5 (1993).
151. Kanai, F. et al. Direct demonstration of insulin-induced GLUT4 translocation to the surface of intact cells by insertion of a c-myc epitope into an exofacial GLUT4 domain. *J Biol Chem* **268**, 14523-6 (1993).
152. Kanai, F. et al. Insulin-stimulated GLUT4 translocation is relevant to the phosphorylation of IRS-1 and the activity of PI3-kinase. *Biochem Biophys Res Commun* **195**, 762-8 (1993).
153. Boulton, T. G. et al. ERKs: a family of protein-serine/threonine kinases that are activated and tyrosine phosphorylated in response to insulin and NGF. *Cell* **65**, 663-75 (1991).
154. Boulton, T. G. et al. An insulin-stimulated protein kinase similar to yeast kinases involved in cell cycle control. *Science* **249**, 64-7 (1990).
155. Szanto, I. & Kahn, C. R. Selective interaction between leptin and insulin signaling pathways in a hepatic cell line. *Proc Natl Acad Sci U S A* **97**, 2355-60 (2000).
156. Chiu, K. C., Chu, A., Chuang, L. M. & Saad, M. F. Association of leptin receptor polymorphism with insulin resistance. *Eur J Endocrinol* **150**, 725-9 (2004).
157. White, R. T. et al. Human adipsin is identical to complement factor D and is expressed at high levels in adipose tissue. *J Biol Chem* **267**, 9210-3 (1992).

158. Lowell, B. B. et al. Reduced adipsin expression in murine obesity: effect of age and treatment with the sympathomimetic-thermogenic drug mixture ephedrine and caffeine. *Endocrinology* **126**, 1514-20 (1990).

159. Flier, J. S., Cook, K. S., Usher, P. & Spiegelman, B. M. Severely impaired adipsin expression in genetic and acquired obesity. *Science* **237**, 405-8 (1987).

160. Boney, C. M., Fiedorek, F. T., Jr., Paul, S. R. & Gruppuso, P. A. Regulation of preadipocyte factor-1 gene expression during 3T3-L1 cell differentiation. *Endocrinology* **137**, 2923-8 (1996).

161. Reynet, C. & Kahn, C. R. Rad: a member of the Ras family overexpressed in muscle of type II diabetic humans. *Science* **262**, 1441-4 (1993).

162. Moyers, J. S., Bilan, P. J., Reynet, C. & Kahn, C. R. Overexpression of Rad inhibits glucose uptake in cultured muscle and fat cells. *J Biol Chem* **271**, 23111-6 (1996).

163. Paterson, J. M. et al. Metabolic syndrome without obesity: Hepatic overexpression of 11{beta}-hydroxysteroid dehydrogenase type 1 in transgenic mice. *Proc Natl Acad Sci U S A* **101**, 7088-7093 (2004).

164. Morton, N. M. et al. Novel adipose tissue-mediated resistance to diet-induced visceral obesity in 11 beta-hydroxysteroid dehydrogenase type 1-deficient mice. *Diabetes* **53**, 931-8 (2004).

165. Dominici, F. P. & Turyn, D. Growth hormone-induced alterations in the insulin-signaling system. *Exp Biol Med (Maywood)* **227**, 149-57 (2002).

166. Nakaya, A., Hishigaki, H. & Morishita, S. Mining the quantitative trait loci associated with oral glucose tolerance in the OLETF rat. *Pac Symp Biocomput*, 367-79 (2000).

167. Tukey, J. W. *Exploratory data analysis* (Addison-Wesley Pub. Co., Reading, Mass., 1977).

168. Moloshok, T. D. et al. Application of Bayesian decomposition for analysing microarray data. *Bioinformatics* **18**, 566-75 (2002).

169. Shmulevich, I., Dougherty, E. R., Kim, S. & Zhang, W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **18**, 261-74 (2002).

170. Schug, J., Diskin, S., Mazzarelli, J., Brunk, B. P. & Stoeckert, C. J., Jr. Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res* **12**, 648-55 (2002).

171. Draghici, S. et al. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res* **31**, 3775-81 (2003).

172. Toronen, P. Selection of informative clusters from hierarchical cluster tree with gene classes. *BMC Bioinformatics* **5**, 32 (2004).

173. Collins, F. S. et al. New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**, 682-9 (1998).

174. Kong, A. et al. A high-resolution recombination map of the human genome. *Nat Genet* **31**, 241-7 (2002).

175. Hudson, T. J. et al. A radiation hybrid map of mouse genes. *Nat Genet* **29**, 201-5 (2001).

176. Van Etten, W. J. et al. Radiation hybrid map of the mouse genome. *Nat Genet* **22**, 384-7 (1999).
177. Rhodes, M. et al. A high-resolution microsatellite map of the mouse genome. *Genome Res* **8**, 531-42 (1998).
178. Irizarry, K. et al. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat Genet* **26**, 233-6 (2000).
179. Wiltshire, T. et al. Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc Natl Acad Sci U S A* **100**, 3380-5 (2003).
180. Johnson, G. C. et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233-7 (2001).
181. Marra, M. et al. An encyclopedia of mouse genes. *Nat Genet* **21**, 191-4 (1999).
182. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25**, 232-4 (2000).
183. Strausberg, R. L. et al. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci U S A* **99**, 16899-903 (2002).
184. Ota, T. et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat Genet* **36**, 40-5 (2004).
185. Xu, Q., Modrek, B. & Lee, C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* **30**, 3754-66 (2002).
186. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nat Genet* **30**, 13-9 (2002).
187. Lockhart, D. J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-80. (1996).
188. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-70 (1995).
189. Rabilloud, T. Detecting proteins separated by 2-D gel electrophoresis. *Anal Chem* **72**, 48A-55A (2000).
190. Link, A. J. et al. Direct analysis of protein complexes using mass spectrometry. *Nat Biotechnol* **17**, 676-82 (1999).
191. Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature* **425**, 737-41 (2003).
192. Espina, V. et al. Protein microarrays: molecular profiling technologies for clinical specimens. *Proteomics* **3**, 2091-100 (2003).
193. Liotta, L. A. et al. Protein microarrays: meeting analytical challenges for clinical applications. *Cancer Cell* **3**, 317-25 (2003).
194. Gygi, S. P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994-9 (1999).
195. Hestvik, A. L., Hmama, Z. & Av-Gay, Y. Kinome analysis of host response to mycobacterial infection: a novel technique in proteomics. *Infect Immun* **71**, 5514-22 (2003).
196. Martzen, M. R. et al. A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286**, 1153-5 (1999).

197. Jessani, N., Liu, Y., Humphrey, M. & Cravatt, B. F. Enzyme activity profiles of the secreted and membrane proteome that depict cancer cell invasiveness. *Proc Natl Acad Sci U S A* **99**, 10335-40 (2002).
198. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-9 (2000).
199. Lee, T. I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).
200. Wang, L. et al. Analyses of p53 target genes in the human genome by bioinformatic and microarray approaches. *J Biol Chem* **276**, 43604-10 (2001).
201. Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H. & Farnham, P. J. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* **16**, 235-44 (2002).
202. Odom, D. T. et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378-81 (2004).
203. Adorjan, P. et al. Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res* **30**, e21 (2002).
204. Cawley, S. et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499-509 (2004).
205. Sun, L. V. et al. Protein-DNA interaction mapping using genomic tiling path microarrays in Drosophila. *Proc Natl Acad Sci U S A* **100**, 9428-33 (2003).
206. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7 (2000).
207. Ito, T. et al. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* **97**, 1143-7 (2000).
208. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74 (2001).
209. Walhout, A. J. & Vidal, M. Protein interaction maps for model organisms. *Nat Rev Mol Cell Biol* **2**, 55-62 (2001).
210. Davy, A. et al. A protein-protein interaction map of the Caenorhabditis elegans 26S proteasome. *EMBO Rep* **2**, 821-8 (2001).
211. Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-3 (2002).
212. Gavin, A. C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7 (2002).
213. Giaever, G. et al. Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc Natl Acad Sci U S A* **101**, 793-8 (2004).
214. Fukui, S., Feizi, T., Galustian, C., Lawson, A. M. & Chai, W. Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. *Nat Biotechnol* **20**, 1011-7 (2002).
215. Wang, D., Liu, S., Trummer, B. J., Deng, C. & Wang, A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat Biotechnol* **20**, 275-81 (2002).

216. Huh, W. K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-91 (2003).
217. Nollau, P. & Mayer, B. J. Profiling the global tyrosine phosphorylation state by Src homology 2 domain binding. *Proc Natl Acad Sci U S A* **98**, 13531-6 (2001).
218. Zhou, H., Watts, J. D. & Aebersold, R. A systematic approach to the analysis of protein phosphorylation. *Nat Biotechnol* **19**, 375-8 (2001).
219. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol* **19**, 125-30 (2001).
220. Nicholson, J. K., Connelly, J., Lindon, J. C. & Holmes, E. Metabonomics: a platform for studying drug toxicity and gene function. *Nat Rev Drug Discov* **1**, 153-61 (2002).
221. Kamath, R. S. et al. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**, 231-7 (2003).
222. Biola, O. et al. The nature and identification of quantitative trait loci: a community's view. *Nat Rev Genet* **4**, 911-6 (2003).
223. Conkright, M. D. et al. Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol Cell* **11**, 1101-8 (2003).
224. Winzeler, E. A. et al. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901-6 (1999).
225. Justice, M. J. Capitalizing on large-scale mouse mutagenesis screens. *Nat Rev Genet* **1**, 109-15 (2000).
226. Tong, A. H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-8 (2001).
227. Harrington, J. J. et al. Creation of genome-wide protein expression libraries using random activation of gene expression. *Nat Biotechnol* **19**, 440-5 (2001).
228. Clemens, J. C. et al. Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways. *Proc Natl Acad Sci U S A* **97**, 6499-503 (2000).
229. Kamath, R. S. & Ahringer, J. Genome-wide RNAi screening in Caenorhabditis elegans. *Methods* **30**, 313-21 (2003).
230. Behr, M. A. et al. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520-3 (1999).
231. Camp, R. L., Chung, G. G. & Rimm, D. L. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nat Med* **8**, 1323-7 (2002).
232. Bubendorf, L. et al. Survey of gene amplifications during prostate cancer progression by high-throughout fluorescence in situ hybridization on tissue microarrays. *Cancer Res* **59**, 803-6 (1999).
233. Perrone, E. E. et al. Tissue microarray assessment of prostate cancer tumor proliferation in African- American and white men. *J Natl Cancer Inst* **92**, 937-9 (2000).
234. Bhattacharjee, A. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98**, 13790-5 (2001).

235. Prakash, K. et al. Symptomatic and asymptomatic benign prostatic hyperplasia: molecular differentiation by using microarrays. *Proc Natl Acad Sci U S A* **99**, 7598-603 (2002).

236. Sarwal, M. et al. Molecular heterogeneity in acute renal allograft rejection identified by DNA microarray profiling. *N Engl J Med* **349**, 125-38 (2003).

237. Brown, V. M. et al. High-throughput imaging of brain gene expression. *Genome Res* **12**, 244-54 (2002).

238. Causton, H. C. et al. Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell* **12**, 323-37 (2001).

239. Nimrichter, L. et al. Intact cell adhesion to glycan microarrays. *Glycobiology* **14**, 197-203 (2004).

240. Dolma, S., Lessnick, S. L., Hahn, W. C. & Stockwell, B. R. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. *Cancer Cell* **3**, 285-96 (2003).

241. Tomancak, P. et al. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol* **3**, RESEARCH0088-8 (2002).

242. Schaffer, R. et al. Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis. *Plant Cell* **13**, 113-23 (2001).

243. Akhtar, R. A. et al. Circadian cycling of the mouse liver transcriptome, as revealed by cDNA microarray, is driven by the suprachiasmatic nucleus. *Curr Biol* **12**, 540-50 (2002).

244. Duffield, G. E. et al. Circadian programs of transcriptional activation, signaling, and protein turnover revealed by microarray analysis of mammalian cells. *Curr Biol* **12**, 551-7 (2002).

245. McDonald, M. J. & Rosbash, M. Microarray analysis and organization of circadian gene expression in Drosophila. *Cell* **107**, 567-78. (2001).

246. Spellman, P. T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-97 (1998).

247. Mootha, V. K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-73 (2003).

248. Su, A. I. et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99**, 4465-70 (2002).

249. Kohane, I. S., Kho, A. T. & Butte, A. J. *Microarrays for an Integrative Genomics* (MIT Press, Cambridge, Massachusetts, 2002).