COMMON EVIDENCE NETWORK: AN INTEGRATED APPROACH TO INVESTIGATING GENE RELATIONSHIPS

by

Alison Meynert
B.Sc. Honours, University of Victoria, 2002

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

In the
School
of
Computing Science

SIMON FRASER UNIVERSITY

Summer 2005

APPROVAL

Name: **Alison Meynert**

Degree: **Master of Science**

Title of Thesis: **Common Evidence Network: An Integrated Approach to Investigating Gene Relationships**

Examining Committee:

Chair: **Valentine Kabanets**
Assistant Professor of Computing Science

Arvind Gupta
Senior Supervisor
Professor of Computing Science

BF Francis Ouellette
Senior supervisor
Adjunct Professor of Molecular Biology and Biochemistry,
Simon Fraser University;
Associate Professor of Medical Genetics,
University of British Columbia

Anne Condon
Supervisor
Professor of Computer Science,
University of British Columbia

Jenny Bryan
Supervisor
Assistant Professor of Statistics,
University of British Columbia

Frederic Pio
External Examiner
Assistant Professor of Molecular Biology and Biochemistry

Date Defended: August 3, 2005

SIMON FRASER UNIVERSITY

PARTIAL COPYRIGHT LICENCE

ABSTRACT

A common evidence network is a data structure that integrates evidence for relationships between genes from disparate data sources and across data types. It is an undirected weighted graph where nodes represent genes and edge weights are quantitative measures of confidence in the evidence linking two genes. We describe methods for producing edge weights for two evidence types: literature co-citation and similarity of Gene Ontology annotations. A tool was developed for identifying genes across multiple databases and consolidating selected annotations. Using gene synonym lists obtained from this tool, we extracted co-citations of genes from annotated biomedical abstracts as evidence. We developed a novel approach to interpreting the similarity of Gene Ontology terms annotated to genes. The method produces a score that quantitatively describes the similarity of Gene Ontology term annotations between two genes. We tested both methods on a set of genes sharing a common sequence feature.

ACKNOWLEDGEMENTS

Thanks are due to all members of my supervisory committee: Dr. Arvind Gupta, Francis Ouellette, Dr. Anne Condon, and Dr. Jenny Bryan. I thank my senior supervisor and training program mentor Dr. Arvind Gupta for the advice and support he has given me over the past two years. Francis Ouellette and Dr. Anne Condon supervised my thesis research. Their insightful questions kept my research focused, and their confidence in my abilities was invaluable.

I thank Stefanie Butland at the UBC Bioinformatics Centre for her guidance and support throughout this project. Sharon Ruschowski of the Bioinformatics Training Program and the graduate secretaries at SFU Computing Science provided significant assistance with navigating the forms and requirements of two universities.

I especially thank my parents for their support and for providing a restful space during stressful times. I am deeply grateful to my partner Colin for his patience, encouragement, and understanding.

My studies were supported financially by the School of Computing Science at Simon Fraser University, the Natural Sciences and Engineering Research Council of Canada, the British Columbia Advanced Systems Institute, and the Michael Smith Foundation for Health Research/Canadian Institutes for Health Research Bioinformatics Training Program for Health Research, and I express my gratitude to these institutions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Investigating relationships between genes

Large sets of molecular biology data have become publicly available over the last 20 years. These include genome and protein sequences, gene expression, biomolecular interactions, protein structure, and biomedical literature. The integration and analysis of different data types from multiple data sources is required to bring cohesion to the vast amounts of information. At one time, biologists exclusively studied one gene or gene family at a time, due to the time and expense of laboratory techniques. The explosion of high-throughput experiments has resulted in a flood of data that cannot be directly absorbed or comprehended. To deal with this, one strategy is to develop computational tools that synthesize the data.

The research field of bioinformatics is "the science of managing and analyzing biological data using advanced computing techniques.[1]" Bioinformatics began with the problem of identifying genes in genomic sequence data, and moved on to elucidating the function of gene products[2]. The focus is now shifting to the problem of fitting individual pieces of information together to discover how systems work as a whole[3]. Of particular interest is data, or evidence, that supports hypothesized relationships between genes[4].

A relationship between a pair of genes could be one of many kinds. The proteins produced by the two genes can interact in some way. If two genes

share the same patterns of expression, it is likely that they are regulated in a similar fashion. The protein product of one gene may regulate the expression of the other. The two genes may be related through evolution. The protein products may have similar functions, be involved in similar biological processes, or localize to the same part of the cell. The two genes may be susceptible to the same type of mutation, or cause similar diseases when mutated.

Similarly, the evidence used to infer a relationship between two genes could be of several types. One strong piece of evidence or the combined effect of weaker multiple correlations of evidence may provide enough confidence of a hypothesized relationship so that a researcher would be motivated to validate it experimentally. We designed a data structure to incorporate multiple types of evidence for inferring and investigating relationships between genes. We developed tools and methods for collecting and quantitatively measuring confidence of two types of evidence. We tested our methods on a set of genes sharing a similar sequence characteristic.

We call our data structure a "common evidence network." It consists of a graph in which nodes represent genes and each edge represents a single piece of evidence connecting two genes. An edge is weighted by a quantitative measurement of the quality of the evidence it represents. We investigated two types of evidence: co-citation of genes in biomedical abstracts, and similarity of annotated Gene Ontology[5] terms.

As a graph, a common evidence network can be visualized[6, 7] and analyzed[4] using standard methods. It could be built for preliminary exploration of

a set of genes and their relationships, or as a hypothesis assessment tool. The latter purpose is of particular interest to the Genomic Mutational Signature Sequences (GeMS) project discussed in section 1.3 below. This project is examining a specific set of genes that share a common structural feature, and are considered candidates for causing neurodegenerative diseases when mutated. A common evidence network built around evidence connecting these genes could be used to suggest laboratory experiments targeted towards verifying relationships between genes that could be related to the disease progression or its prevention.

1.2 Molecular biology background

The molecule deoxyribonucleic acid (DNA) is the encoding material of genomes. DNA is a polymer of nucleotides that forms strands. Two spiralling strands of millions of nucleotides can be held together by weak bonds between nucleotides on opposite strands form the DNA that constitutes our genomes. Nucleotides all have the same structure: a sugar (deoxyribose), a phosphate, and a nitrogenous base. There are four different types of nucleotides depending on the nitrogenous base: adenosine, cytosine, guanine, and thymine (A, C, G, and T). In a double helix of DNA, A pairs with T and C pairs with G. Humans have 46 long DNA molecules, called chromosomes. Twenty-two pairs are referred to as autosomes and in humans, are numbered in order of decreasing length. The remaining two chromosomes, amongst other functions, determine the sex of an individual. Females have two X chromosomes and males have one X and one Y chromosome.

The Mendelian definition of a gene is a segment of DNA that, if mutated, causes the organism to display a change in phenotype, or physical characteristic of an organism. This change in phenotype must additionally be heritable; that is, it can be passed on to the descendents of the organism in which the mutation occurred. This definition encompasses segments of DNA that are not generally considered genes by modern molecular biologists. In part, this is because the Mendelian definition does not work well with sequence databases, which require DNA coordinates, or locations. Mostly, however, it is because many changes in phenotypes cannot be observed.

Most of the genes that are researched and discussed are protein-coding genes. Protein-coding genes are segments of DNA that can be "read" and translated into proteins. Proteins are both the building blocks of an organism's structure, and the machinery that makes it run. The DNA reading process is called transcription, and results in a copy of the gene being made in RNA (ribonucleic acid).

DNA and RNA molecules differ only in the sugar component of each nucleotide, and in the substitution of uracil (U) for thymine (T) in RNA. When a gene is transcribed from DNA to RNA, a protein complex called RNA polymerase causes the two strands of the DNA molecule to "unzip", allowing RNA nucleotides to pair with nucleotides on one of the DNA strands. As the RNA nucleotides bind to the DNA, they also bind to each other to form a single-stranded molecule. As the RNA strand forms, it detaches from the DNA, and the two strands of DNA are "zipped" back together by the RNA polymerase complex.

The single-stranded RNA molecule is called a transcript of the gene, or messenger RNA.

Unlike the DNA in chromosomes, messenger RNA can travel outside the nucleus of the cell. RNA transcripts are "read" by a protein-RNA complex called a ribosome, which sequentially translates the messenger RNA into proteins. A protein is a chain of amino acid molecules. There are 20 different amino acids, and every three nucleotides in the RNA molecule translate to a single amino acid that the ribosome will add to the protein at that location. Because there are four different nucleotide bases, there are 64 unique combinations of three nucleotides. We call each of these combinations a codon. There are more codons than amino acids, therefore multiple codons can encode for the same amino acid.

1.2.1 Types of genetic diseases

Most genes have two versions, or alleles: one from the mother and one from the father. These alleles may be the same or they may be different. This leads to the two major types of mono-genetic diseases. A given mutation (change in the DNA of a gene) can be recessive or dominant. When two copies of the disease-associated allele are required for the disease to occur, we call the disease recessive. In this case, individuals with only one disease-associated allele will not develop the disease. Recessive genetic diseases are associated with a loss of the normal gene function. An example of a recessive genetic disease is sickle cell disease. Recessive diseases caused by a disease-

associated allele on the X chromosome, such as haemophilia, require two such alleles in females, but only one in males.

A dominant genetic disease requires only one disease-associated allele, and can be associated with a toxic effect due to this mutated allele. The mutated allele produces a protein that has gained some function due to the mutation. The new function is toxic to the cell or its environment. This is called a gain of function mutation. The other type of mutation is called loss of function, where the mutated allele produces non-functioning proteins. In this case, only one copy (allele) of the gene is producing a protein that is performing the normal function. This means the individual will only have half of the normal amount of that protein in their cells, which is sometimes insufficient for the cell to remain in a healthy condition.

In a dominant genetic disease, an individual with one normally functioning allele and one disease-associated allele will have the disease. Two examples of dominant genetic diseases are the polyglutamine expansion diseases discussed below and listed in Table 1[8], and amyotrophic lateral sclerosis (ALS)[9]. Two questions immediately arise about dominant genetic diseases, once the gene in question has been identified. First, what is the normal function of the gene? Second, is the mutation cause a gain or loss of function, or possibly both? Both of these biological questions can be explored using a common evidence network.

1.3 Motivation: Genomic Mutational Signatures (GeMS) Project

The GeMS project[10] is a collaborative initiative of the University of British Columbia's Bioinformatics Centre and Centre for Molecular Medicine and Therapeutics. This project focused on genetic diseases caused by a type of mutation that may occur individually in nine different human genes, causing nine separate associated diseases. The differences between the healthy alleles and the disease alleles of these genes follow very similar patterns, hence the "mutational signatures" project name. That is, at the mutation site, the healthy alleles all share a sequence characteristic that is changed in a similar fashion in the disease alleles.

Butland *et al* (in preparation) scanned the human genome sequence for genes that had similar starting sequence characteristics to the healthy alleles of the nine known disease genes. This set of genes was considered a set of candidate disease genes. We examined patients with symptoms similar to those caused by one of the known diseases (Huntington disease), but who tested negative for mutations in the associated gene. These patients were screened for the mutational signature in the candidate disease genes. Bioinformatics support analyzed the candidate disease genes for possible functional linkages. A common evidence network was developed as part of the bioinformatics analyses for the known and candidate disease genes from the GeMS project. However, these tools may also be applied to any reasonably small set of genes.

1.3.1 Polyglutamine expansion diseases

Polyglutamine domain expansion diseases, such as Huntington disease, were the first disease category under consideration by the GeMS project researchers. Glutamine, abbreviated Gln or Q, is one of the 20 amino acids that make up proteins. It is encoded by the triplet codons CAG and CAA. When one or both of these codons are repeated sequentially, the resulting protein will contain a repeated stretch of glutamine amino acids. We refer to these stretches as polyglutamine (polyQ) domains or repeats.

Polyglutamine domains are polymorphic in some genes. This means that within the normal population, there is variance in the length of a polyglutamine domain for a given gene. For example, in the gene that can cause Huntington disease, healthy individuals will have fewer than 26 CAG repeats. Individuals with 27-35 CAG repeats will be healthy themselves, but they may be at risk of having children with the disease[11]. Individuals with between 36 and 40 CAG repeats are at risk, but may not develop the disease[12, 13]. Those with 41 or more repeats will get Huntington disease: the more repeats an individual has, the earlier in life the onset of the disease will be[14, 15].

There are nine human diseases known to be associated with a specific type of mutation in a gene containing a polyglutamine domain. All of these diseases occur when a polyglutamine domain mutates to be longer than the normal range. Thus, they are referred to as polyglutamine domain expansion diseases. The mutations that give rise to the known polyglutamine domain expansion diseases all involve expansions in a series of repeated CAG codons,

sometimes interspersed with CAA codons. There is no known reason for the preference of CAG codons over CAA codons in the repeat expansion process.

The location of polyQ domains in the sequences of the nine known disease genes does not follow any pattern. The position of the first glutamine amino acid in the domain ranges from 0.6% to 92% of the sequence length, with a considerable spread of points in between. There are no determined three-dimensional structures of the proteins produced by polyglutamine expansion disease genes. Domain structures of ATXN1[16] and HD[17] have been determined, although these structures do not include the polyglutamine domains of the respective proteins. A partial structure of ATXN3 has been determined through nuclear magnetic resonance (NMR)[18]. The protein appears to be primarily globular, with a flexible tail segment from the C-terminus sequence, which contains the polyQ domain. Similarly, the polyQ domain of CREB-binding protein, a gene that is not known to be associated with disease, is in a flexible loop structure that binds with multiple ligands[19, 20].

All of the polyglutamine expansion diseases are neurodegenerative in nature. That is, they cause the death of neurons, specialized cells that make up nervous system tissues such as the brain and spinal cord. For example, an expansion in the polyglutamine domain of the huntingtin gene causes Huntington disease. Table 1 shows the list of known polyglutamine domain expansion diseases and the associated genes.

Table 1 Polyglutamine domain expansion diseases

Disease name	Gene name	Gene symbol
Dentatorubral pallidoluysian atrophy[21]	Atrophin 1	ATN1
Huntington disease[22]	Huntingtin	HD
Spinal bulbar muscular atrophy (Kennedy disease)[23]	Androgen receptor	AR
Spinocerebellar ataxia type 1[24]	Ataxin 1	ATXN1
Spinocerebellar ataxia type 2[25]	Ataxin 2	ATXN2
Spinocerebellar ataxia type 3 (Machado-Joseph disease)[26, 27]	Ataxin 3	ATXN3
Spinocerebellar ataxia type 6[28]	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit	CACNA1A
Spinocerebellar ataxia type 7[29]	Ataxin 7	ATXN7
Spinocerebellar ataxia type 17[30, 31]	TATA-box binding protein	TBP

1.3.2 Genes with polyglutamine domains

The single genetic link between the known polyglutamine domain expansion disease genes is the type of mutation that causes the diseases. An immediate question is whether other genes with polyglutamine domains cause diseases when those domains expand by mutation. The GeMS project considered patients with Huntington disease symptoms, but who tested negative for the mutation in the associated gene. The project screened these patients for expansions in the polyglutamine domains of a set of candidate disease genes.

The GeMS project restricted the screening process to only those genes with five or more sequential CAG codons, coding for five or more glutamine amino acids. Appendix A, Table 20 contains the complete list of 56 candidate disease genes and the 9 genes known to cause disease, for a total of 65 genes of interest.

CHAPTER 2: COMMON EVIDENCE NETWORK

2.1 Gene networks

Building a network is a popular approach to visualizing and analyzing large numbers of genes and their relationships to each other[3]. There are many types of gene networks, generally based on the type of evidence for the inferred relationships. The standard example is a protein-protein interaction network that documents the experimentally verified interactions between gene products[32]. Gene regulation relationships[33, 34], gene co-expression data[35-37], and evolutionary patterns[38] can also be viewed as networks.

2.1.1 Top-down vs. bottom-up

Most gene network projects attempt to display the relationships between all genes in a given genome or data set. We refer to this approach as top-down. These projects set a high priority on having a low number of false positive relationships. The trade-off is that there is often a low priority on maximizing the number of true positives[39]. Top-down projects provide valuable, high-confidence baseline data for researchers. However, because of the accuracy priority, true evidence is sometimes missed. This problem is especially acute when a network is based on identifying genes in the biomedical literature. For example, the STRING project at the European Molecular Biology Laboratory includes a text-mining component that does not include gene names that map to more than one gene[40]. Because of this, many references are missed.

An alternate or complement to the top-down whole genome approach is to focus on a smaller set of genes in what we refer to as a bottom-up approach. Researchers choose a set of genes based on a shared feature or a biological question. A smaller network is constructed, based on this set of genes. The inclusion of as much true evidence as possible becomes a priority, at the expense of including false evidence. The results from a top-down network for the genome under study can be used as validation data.

2.2 Common evidence network

Thus far, we have discussed networks based on a single type of evidence. Information linking genes is available from multiple sources, but these are difficult to synthesize. By combining information from disparate sources, we gain two major benefits: a clearer overview of the linkages between genes of interest, and a sense of the information coverage for each gene. For example, certain types of genes are difficult to study using some types of experimental approaches, which may lead to systemic biases in data sources that deal with those experiment types. Presenting multiple types of evidence can lend credibility to a given relationship.

A gene network is a graph in which we represent each gene as a node. Edges in the graph represent evidence common to two genes. Different types of evidence are represented as sets of edges that can be added and removed from the network for analysis or visualization. This type of graph is called a multi-graph, because a pair of nodes may be joined by multiple edges. Confidence in

a given piece of evidence is quantified as a weight on the resulting edge. We call this data structure a common evidence network.

2.2.1 Naturally paired and un-paired data

For the purposes of a common evidence network, we divide gene annotations (evidence) into two major classifications: naturally paired, and not naturally paired. Naturally paired data is existing evidence linking genes or gene products, such as protein-protein interactions or co-expression profiles. This data is relatively easy to co-opt for usage in a common evidence network, because the relationships between genes or their products are already quantified.

Most data falls into the not naturally paired category. Using data that is not naturally in pairs of genes requires analysis to first identify and then quantify relationships. This research concentrates on two sources of naturally unpaired data: citation of genes in medical literature, and Gene Ontology[5] annotations.

Table 2 Examples of evidence for linkages between genes

Evidence	Type	Linkage hypothesis
Co-citation in medical literature	Unpaired	Genes mentioned in the same abstract are more likely to be functionally linked
Similarity of Gene Ontology terms	Unpaired	Genes that are annotated with similar Gene Ontology terms are more likely to be functionally linked
Documented protein-protein interactions	Paired	Genes with products that have experimentally verified protein-protein interactions are functionally linked; those with predicted interactions are more likely to be functionally linked

Evidence	Type	Linkage hypothesis
Gene co-expression	Paired	Genes with similar microarray expression profiles across multiple experiments are more likely to be functionally linked

2.3 Network details

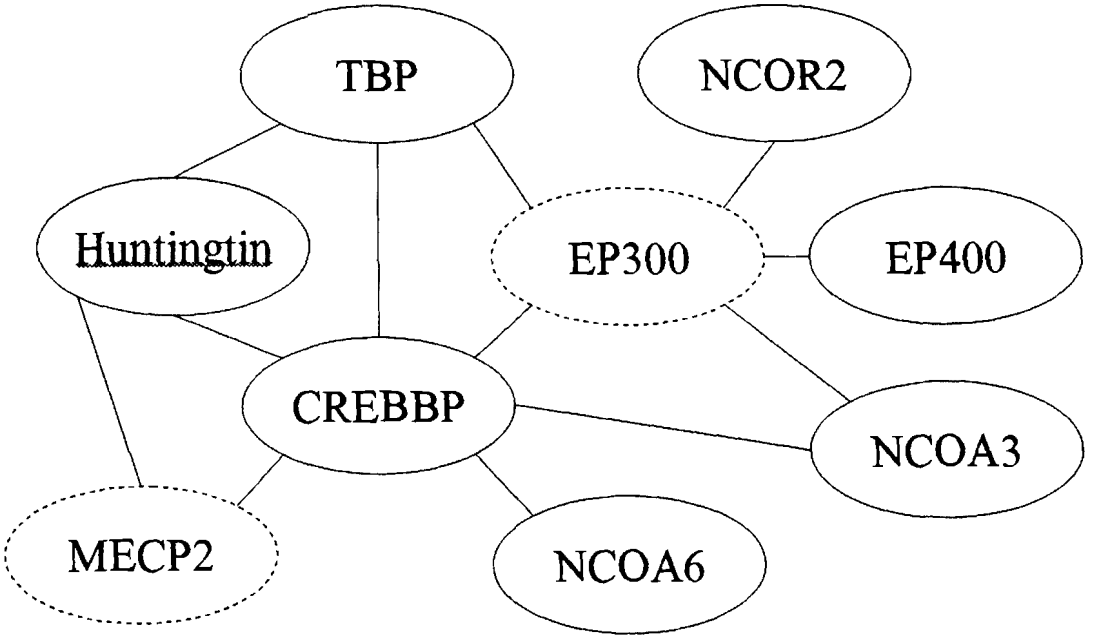
We can apply a scoring function to the evidence linking two genes, to indicate confidence in that piece of information. The resulting value becomes a weighted edge in the network. We can use different scoring functions to calculate edge weights for different evidence types. However, we want to be able to compare edges resulting from different types of evidence. We refer to the edges from a single evidence type as a layer in the network. The scoring functions for all layers should result in a consistent range of edge weights. For this project, we chose to use the range of 0 to 1.

2.3.1 Indirect relationships

A researcher would begin to examine his or her genes of interest by building a common evidence network that contained only direct links between genes in that set. However, this network may not give the researcher a sufficiently complete picture. In the example shown in Figure 1, we see that including only direct links would leave genes NCOR2 and EP400 disconnected from the rest of the graph. By including the gene EP300, which is connected to many of our genes of interest, we have expanded our gene list and potentially gained new information.

Figure 1 **An example of a gene network**

Nodes with solid outlines represent genes in the set of immediate interest. Nodes with dashed outlines represent genes outside that set.



We do not attempt to solve the problem of identifying indirect relationships for the two methods presented here, since a solution would require a top-down approach to identify relationships between all human genes. For the literature co-citation evidence, we would need to obtain a comprehensive, unambiguous list of all gene names to search the literature effectively for mentions of each gene. The Gene Ontology[5] similarity scoring method relies on randomly drawn sets of genes of the same size as the set of interest. We would need to calculate scores between every pair of human genes.

CHAPTER 3: GENE NAME AMBIGUITY

3.1 Sources of gene names

We want to compare pairs of genes. To do so, we must be able to uniquely identify our genes of interest. Most genes have more than one identifier, and many identifiers are not unique to one gene. The solution to this gene name ambiguity problem is twofold: a set of unique primary identifiers for each gene in the set of interest, and an unambiguous mapping of every other gene identifier to a single gene.

Researchers have been identifying human genes for many years. The sequencing of the genome identified many more[41]. DNA and protein sequences for genes and their products are stored in a number of databases[42-44], where alphanumeric strings called accession numbers serve as sequence identifiers. In addition, biologists working on a specific gene assign a user-friendly symbol. Gene symbols historically have been assigned based on one of the following:

- A function of the gene's protein product, e.g. TBP = TATA-box binding protein
- A disease caused by the mutation of the gene or a malfunctioning of its protein product, e.g. HD = Huntington Disease
- A base name used for a related family of genes, e.g. FOXP2 = Forkhead box domain family P, type 2

These human gene symbols are now sanctioned and approved by an international body: the Human Genome Organization (HUGO).

3.1.1 Human Genome Organization Nomenclature Committee

In 1979, the Human Genome Organization (HUGO) created a Nomenclature Committee (HGNC) to manage a standard set of gene symbols and descriptions for human genes[45]. As of May 1, 2005, the HGNC had approved 23,522 human gene symbols and descriptions. Each gene symbol is unique within the HGNC database, Genew[46]. Entries can be changed if required, or removed entirely if the record is a duplicate or if a hypothesized gene is shown to not exist. In general, once curators assign a symbol to a gene, the symbol can be expected to be relatively stable. The primary purpose of the HUGO nomenclature is to define a unique gene symbol to facilitate database searches and utilization. The existence of a unique, stable gene identifier list also has the side benefit of greatly assisting publication searches and the annotation of biomedical literature abstracts.

3.2 Ambiguities

As researchers discover more about a gene, the gene's primary or official symbol can change to take into account the new information. This results in several symbols that refer to the same gene. Additionally, we have situations where the same symbol can refer to multiple genes within the same species. For eukaryotic (organisms that have cells with a nucleus) genomes, researchers have found up to 5% intra-species symbol ambiguity in gene names identified in

publications[47]. Nomenclatures obtained from a gene-centric database (Entrez Gene[48], downloaded June 1, 2005) confirm this. We found 1.3% ambiguity internal to the nomenclatures of *Saccharomyces cerevisiae* (yeast), 0.4% for *Caenorhabditis elegans* (worm), 7.5% for *Drosophila melanogaster* (fruit fly), 3.3% for *Mus musculus* (mouse), and 4.1% for *Homo sapiens* (human). This translated to 2.8%, 0.9%, 16.1%, 5.7%, and 8.9% of gene records (respectively by species) containing an ambiguous name.

Many species have genes with similar functions and sequences; therefore researchers frequently use the same symbol for a given gene across multiple species. The use of annotated genomes to aid in the identification of genes in related, newly sequenced genomes exacerbates this problem from a text-mining perspective. Annotaters often give the gene in the new genome the same or similar symbol to its putative ortholog in the annotated genome. While this aids biological interpretation of the gene's function, text-mining applications parsing an article mentioning that gene cannot automatically identify to which species it belongs. Unlike sequence database references, taxonomic information is generally not formally or explicitly identified in biomedical abstracts.

The human and mouse genomics communities have attempted to solve this problem by the use of standardized character cases. Alphabetic characters in officially sanctioned human gene symbols are always upper case; for mouse genes, only the first character is capitalized. These differences are only useful when attempting to distinguish between these two species, and only if authors adhere consistently to the standards. Even when considering case, researchers

have found up to 14% inter-species gene symbol ambiguity across 21 eukaryotic species[47]. When the same researchers removed character case from consideration, the ambiguity rose to approximately 25%. Our results across the yeast, nematode, fruit fly, mouse, and human nomenclatures from Entrez Gene show 26.1% of gene records contain an ambiguous name using a case-insensitive comparison, and that 13.0% of gene names are ambiguous.

3.2.1 Locating database records

This mix of gene symbols results in two problems with gene databases. First, the ambiguity problem makes it difficult for researchers to identify specific gene records automatically, because multiple records can be returned for a given search symbol. For example, authors have used the gene symbol "AR" to refer to both the human androgen receptor gene and the human amphiregulin gene. Therefore, the symbol is included in database records for both of these genes, although it is officially only the symbol of the human androgen receptor gene. Second, different data sources use different symbols as primary record labels for genes, and contain different lists of alternate symbols. Generally, databases contain fields indicating the species for each gene record, restricting the problem to intra-species ambiguity.

We investigated the ambiguity of human gene names in the Entrez Gene database (downloaded June 1, 2005). A total of 74,824 names were identified from the official symbol, primary symbol, and synonym fields for 32,801 gene records. The official symbol field contains a symbol that has been assigned by the nomenclature authority for the species in question. 64.1% of human genes in

Entrez Gene have official symbols assigned by the HGNC. The primary symbol is the same as the official symbol for those records having an official symbol; otherwise it is generally the most commonly used symbol for the gene. Synonyms are any other names by which that gene is referred to by scientists.

There was no ambiguity found within the official symbol field (21,026 distinct names). This was unsurprising as all of these official symbols were designated by the HGNC, which has a mandate to provide unique symbols to every human gene. Within the primary symbol field, 24 of the 32,775 names occurred more than once (<0.1% ambiguity). The synonym field contained higher internal ambiguity, with 2,068 of the 43,165 names (4.8%) occurring more than once, some over 10 times.

Table 3 **A summary of ambiguity of human gene names in Entrez Gene**

The first three columns describe an occurrence pattern for a gene name. The fourth column contains the number of gene names that have this occurrence pattern.

Occurrences in official symbol field	Occurrences in primary symbol field	Occurrences in synonym field	Names with this occurrence pattern
0	0	1	40,129
0	0	>1	1,926
0	1	0	11,558
0	1	>0	187
0	>1	0	3
0	2	1	2
1	*1	0	20,094
1	*1	>0	912
1	*2	0	10
1	*2	>0	10

* Where an official symbol exists in a gene record, the primary symbol is identical to the official symbol.

We examined the overall ambiguity (both within and between fields) and summarized the results in Table 3. For each of the 73,696 gene names, we counted the number of records in which the name appeared in the official, primary, and synonym fields. We then counted the number of gene names having the same occurrence pattern across those fields. For example, the first row of the table says that 40,129 human gene names occurred only in the synonym field of one gene record each.

When we take into account the fact that if an official symbol has been assigned to a gene, the primary symbol is the same as the official symbol, we find that 3,050 names are assigned to more than one gene (4.1% of names, 8.9% of gene records). From Table 3, this is the sum of the number of names in all rows except the first, third, and seventh. The majority of the ambiguity between fields comes from names that are synonyms for multiple genes (second row of Table 3). In only 20 cases do we have an official symbol for one gene that is a primary symbol for another gene.

3.2.2 Searching publications

Another ambiguity problem becomes apparent when we consider text-mining applications. The identification of gene references in biomedical literature is a quickly growing subfield of text-mining applications. There are two major divisions within this subfield, based on the text source: full text articles and abstracts only. Abstracts are more widely available and contain a condensed version of information; however, they frequently do not have references to all of the genes mentioned in the full text of the article[49]. After considering the problem

of differentiating genes and the species from which they arise, researchers face the challenge of distinguishing gene symbols from English words and abbreviations of phrases.

Roughly 1% of gene symbols are ambiguous with general English words[47]. However, some of these gene symbols are not only general, but also common English words. Including these gene symbols in a simple string-matching search of publications will result in a large number of false positive matches. A more subtle and potentially larger source of false positive symbol matches is from abbreviations of phrases. For example, "HD" is the official gene symbol for the huntingtin gene. It is also an abbreviation for "Hodgkin's Disease" and "hemodialysis", two very common terms in the biomedical literature.

CHAPTER 4: INFORMATION GATHERING UTILITY

4.1 Problem definition

Molecular biologists investigating a problem often generate a list or lists of genes in which they are interested and wish to gather currently known information about these, including any commonalities. Shared information can shed light on the function or process under study. This information may include alternate symbols (synonyms), accession numbers across multiple databases, descriptions, publications, and functional annotations. They may want to have this information indexed by their chosen gene symbols.

One of the early challenges the biologist frequently faces is to decide which data sources to use. Many relatively comprehensive data sources exist, yet none can claim to contain all of the relevant information. This could be due to the specialized purpose of the source, the level of data curation, the limitations of the data source schema, or licensing issues related to using data from other sources. A researcher will frequently combine data from multiple sources to create his or her own comprehensive set of information for a particular set of genes.

This gives rise to the second and third problems. Locating the correct record for a gene in a given data source can be time consuming if multiple records match the biologist's search symbol. After locating all of the records for

the genes of interest, the biologist must then combine the records obtained from each data source.

Constant database updates and releases mean that more information becomes available on a regular basis. Researchers need to re-check data sources from time to time to ensure they have the most up-to-date information prior to publication. The true magnitude of the problem becomes fully apparent when we consider that the biologist is most likely doing all of the work manually, using web-based interfaces to access the various data sources.

An application to assist researchers by automatically finding the information they require could become an invaluable tool. This application must have a simple graphical user interface. It must allow researchers to use their preferred data sources and their preferred versions (releases) of those data sources. In general, a researcher will simply use the most up-to-date release; however, if the project stretches over a long period of time that may span multiple releases from the data source, standardizing on a single release will ensure consistent and reproducible results. Users must be able to search for genes using multiple symbols to increase the likelihood of complete coverage. The application must additionally allow for user-mediated resolution of multiple matches for a gene within a data source, and for multiple symbol-to-gene mappings between data sources.

4.2 User input

The user must specify at least one gene symbol for each gene in which he is interested. These symbols will be used as search terms against the various information sources. Additionally, the user may include multiple symbols to improve the likelihood of locating relevant records. The user has the option of guaranteeing the inclusion of these symbols in the output list of gene synonyms, which is especially applicable in situations where a researcher wants a list of gene symbols and their synonyms for text mining. An expert in that research domain may know of gene symbols that appear primarily in older publications, but are not necessarily included in current information sources.

4.3 Information sources

The information gathering application includes parsing and searching modules for plain text (flat) files from three main sources: Genew, the official HUGO human gene name database[46], UniProt, a protein-centric database[50], and Entrez Gene, a gene-centric database[48]. The information available from these sources is summarized in Table 4.

Table 4 Sources of information for human genes

A tick indicates that at least some of the indicated information is available for the given source. For instance, a gene may have multiple UniProt accessions, but only the primary accession will be included in Entrez Gene.

	Genew (HUGO)	UniProt	Entrez Gene
Gene symbols			
Official HGNC symbol	X	X	X
Synonyms	X	X	X
Protein names		X	X
Accession numbers			
NCBI GenBank	X	X	X
EMBL Nucleotide Sequence Database	X	X	X
DNA Databank of Japan	X	X	X
NCBI RefSeq	X		X
UniProt (SwissProt/TrEMBL)	X	X	X
Protein Information Resource (PIR)		X	
EnsEMBL		X	
GDB Human Genome Database	X		
Gene descriptions (long names)	X	X	X
Gene Ontology[5] terms		X	X
Publications		X	X

4.3.1 Genew – The Human Genome Nomenclature Database

Genew is the database of generally accepted human gene names[46]. Researchers studying the human genome have agreed to standardize human gene symbols and names. The Human Genome Organization (HUGO) is a consortium of human genomics researchers that includes the HUGO Nomenclature Committee (HGNC). Researchers submit proposals for gene symbols and longer, more descriptive names to the HGNC. The committee approves proposals and resolves conflicts. Approved gene symbols and names are included in the Genew database. Entries in the database become publicly

available upon publication of the submitter's paper. The motto of the HGNC is "Giving unique and meaningful names to every human gene"[51].

Sometimes, the approved gene symbol and name change when more information about the gene's function becomes available. Genew records maintain a list of previously approved symbols and names, as well as a separate list of gene symbol aliases, or synonyms. Each record may also contain links to other databases in the form of accession numbers.

4.3.2 Universal Protein Resource (UniProt)

UniProt[50] is a protein-centric database hosted by a consortium of European genome research centres, including the European Bioinformatics Institute, the Swiss Institute of Bioinformatics, and the Protein Information Resource (PIR), based in the United States of America. UniProt is divided into two main sections, based on the curation status. SwissProt contains records that have been manually curated from the literature or have had computational curation evaluated by a human. TrEMBL (translated EMBL) records have been computationally analyzed but have not yet been manually curated. Together, these records are referred to as the UniProt Knowledgebase[52].

4.3.3 Entrez Gene

Entrez Gene is a curated database at the National Center for Biotechnology Information (NCBI) that integrates information about genes into tracked records with the intention of representing all genes from all organisms. It makes extensive use of RefSeq, a database which contains one genomic, one

RNA, and one protein record of all such molecules known to exist[53]. Entrez Gene currently has entries from a limited number of organisms, but the intent and goal is to be comprehensive. Every gene record has a unique database identifier. The primary symbol for a human gene is its HGNC-assigned official symbol, if one exists. If the HGNC has not approved a symbol for a gene, its Entrez Gene record will have a generally accepted symbol as the primary symbol. The record will also include other known symbols for the gene, although the list will not necessarily be complete. A longer description is assigned in the same way: the HGNC-assigned name is given priority.

The database covers accession numbers for sequences obtained from the three major sequence databases: NCBI's GenBank[42], the EMBL Nucleotide Sequence Database[43], and the DNA Data Bank of Japan[44]. Records are shared across all three of these databases, and accession numbers are unique. Each gene record also includes its primary accession number from the UniProt database, if it exists.

Entrez Gene records also include two separate lists of publications associated with a gene. The first is a manually curated list called Gene References Into Function, or GeneRIFs. For a publication to be associated with a gene in the GeneRIFs list, it must mention something about the function of that gene. The second, Gene2Pubmed, is a more general list of publications that simply cite the gene. Gene records also contain Gene Ontology[5] annotations obtained from the Gene Ontology Annotation Database[54].

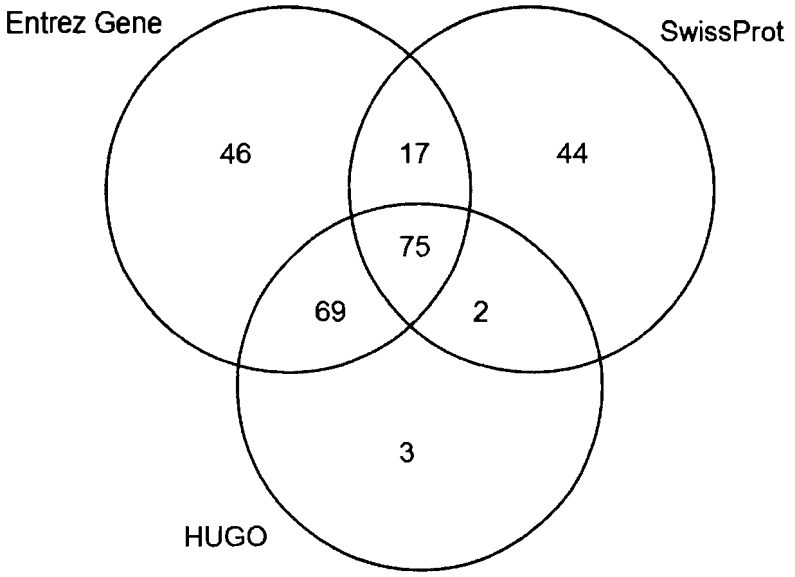
4.4 Comparison of information sources

While the three information sources defined above contain much duplicated information, each also contains some unique. We examined the information overlap for two different types of information: gene/protein names and associated publications. We did not examine accession numbers as these are explicitly focused on the database in question. Gene Ontology annotations will be further explored in Chapter 6.

The examination was limited to information associated with our 65 genes of interest (genes containing polyglutamine repeats). For the UniProt information source, we used only the human-curated (SwissProt) division. The computationally generated TrEMBL division frequently contained multiple records for the same gene. Only 52 genes were referenced in all three information sources. Genes that were not referenced in all information sources were excluded. Figure 2 and Figure 3 display the results.

Figure 2 Information overlap for gene/protein names

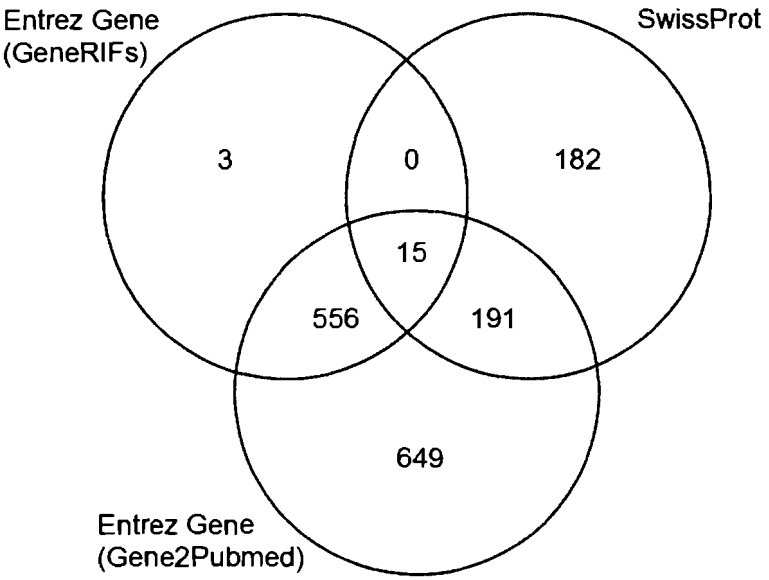
The SwissProt locus names (i.e. ANDR_HUMAN) were not included in the results shown here. These locus names are specific to the SwissProt database and are rarely if ever seen elsewhere.



For gene and protein names, Entrez Gene appears to be somewhat more comprehensive than the other two sources. It almost fully contains the names from HUGO. However, Entrez Gene is certainly not complete, as evidenced by the 19% of names only found in SwissProt or HUGO. It is of interest to note that only 29% of names are found in all three information sources. One possible explanation for this (and the distribution of names between Entrez Gene and HUGO) is that SwissProt is a protein-centric database and the other two sources are both DNA-centric.

Figure 3 Information overlap for associated publications

HUGO's Genew database does not record associated publications. The two sources of publication associations from Entrez Gene were compared. Note that most (124) of the 182 publications found exclusively outside Entrez Gene are annotated to a single gene (AR, or androgen receptor).



Here we see that the set of GeneRIFs publications is almost completely contained within the Gene2Pubmed set. This is expected, as both sets are from the same organization. The Gene2Pubmed set is a list of general associations, while the GeneRIFs set is a manually curated set of associations to papers that specifically discuss the function of the gene in question. In general, it appears that the GeneRIFs set is a subset of the Gene2Pubmed gene-to-publication associations. Only three gene-to-publication associations are in the GeneRIFs set that are not also in the Gene2Pubmed set. We see that although Entrez Gene appears comprehensive, 11% of the gene-to-publication associations occur exclusively in the SwissProt information source, demonstrating the increased data coverage that can be gained from combining data from multiple sources.

4.5 Applications for text-mining

The gene names and accessions output by the information gathering application can be used as input for text-mining applications. We provide two options in the application to assist in producing a more comprehensive and unambiguously mapped list of gene names. First, the user can choose to check for gene names that map to more than one gene. The application allows the user to choose a single name-to-gene mapping for ambiguous names. Second, the user can opt to have the gene name list "padded" with variants on the output names.

This "padding" involves the application of two rules. If a gene name contains a forward slash (e.g. ELD/OSA1), the output will include the original name and both components to either side of the slash (e.g. ELD/OSA1, ELD, OSA1). The second rule applies to names that begin with a series of alphabetical characters and end in a one or two digit number (e.g. SCA2). The application will output three variants on this type of name: with no space between the alphabetical characters and the number, with a single space, and with a dash (e.g. SCA2, SCA 2, SCA-2). These two rules cover some of the common variants of gene names that appear in biomedical journal articles.

CHAPTER 5: CO-CITATION EVIDENCE

5.1 Co-citations

When a biomedical journal article (a citation) mentions two or more genes, we say that these genes are co-cited, or that a co-citation exists for these genes. We examined co-citations of genes in the abstracts of articles. A co-citation can be broken down into pairs of genes. We hypothesized that when two genes are co-cited, they are more likely to be functionally linked. We identified instances of gene names, symbols and accessions in a body of biomedical literature abstracts through basic string matching. The results were used to identify co-citations.

The gene name ambiguity problem discussed above led to many false identifications of genes in biomedical literature. We identified the most commonly matched gene names and symbols, and estimated the precision of each of these. We evaluated the sensitivity of the application using gold standard lists of gene to publication associations obtained from Entrez Gene[48] and the SwissProt section of UniProt[50]. A scoring function based on the number of times two genes were cited together was used to produce an edge weight for evidence linking the two genes in the common evidence network.

5.2 Input

A comprehensive list of gene names, symbols, and accession numbers was assembled using the information gathering utility application described

above, for the genes listed in Table 20 (Appendix A). A biologist familiar with the genes of interest augmented the names and symbols in the list. Names and symbols were further padded by the rules described in 4.5 above.

Table 5 Input gene names, symbols and accession numbers

Pre-padding numbers include only names and symbols that were identified directly from the databases. Post-padding indicates that names and symbols were added to by the rules described in 4.5 above.

Total genes	65
Total names and symbols (pre-padding)	628
Total names and symbols (post-padding)	956
Total accession numbers	1663
Average names and symbols per gene (pre-padding)	9.7
Average names and symbols per gene (post-padding)	14.7
Average accession numbers per gene	25.6

5.2.1 MEDLINE XML

MEDLINE is a curated bibliographic database of abstracts from biomedical journals[55]. The United States of America's National Library of Medicine maintains, updates, and publishes the database. It is the primary source of literature citations in the biomedical research field. It contains only abstracts, not full articles. As of May 1, 2005, it contained over 15 million unique citations dating back to 1950.

Users can license and download the MEDLINE database as a series of XML format files. XML is an acronym for "Extensible Markup Language", and is a data format standard published by the World Wide Web Consortium (W3C)[56].

MEDLINE XML is an implementation of this data format standard. Every November, the MEDLINE XML implementation is updated and a new baseline set of files is created. The following analysis uses the 2005 baseline files and daily updates dating to May 1, 2005.

Curators annotate citations in MEDLINE. In addition to the standard bibliographic information such as author names, journal, publication date, title and abstract, MEDLINE citations have standard fields for summarizing the content of the article. Fields of particular interest for text-mining and searching applications are listed in Table 6. An example of a MEDLINE XML citation is shown in Figure 13 (Appendix B).

Table 6 MEDLINE citation annotation fields

Annotations for a citation are either submitted by the authors via the journal, or added by human curators at MEDLINE.

Annotation field	Description
Databank accession numbers	Accession numbers of sequences mentioned in the article.
Gene symbols	Gene symbols mentioned in the article.
Chemicals	Names of chemicals mentioned in the article, along with their registry numbers from various sources
Keywords	General terms from a controlled vocabulary
Medical Subject Headings (MeSH)[57]	Biomedical terms from a controlled, structured vocabulary

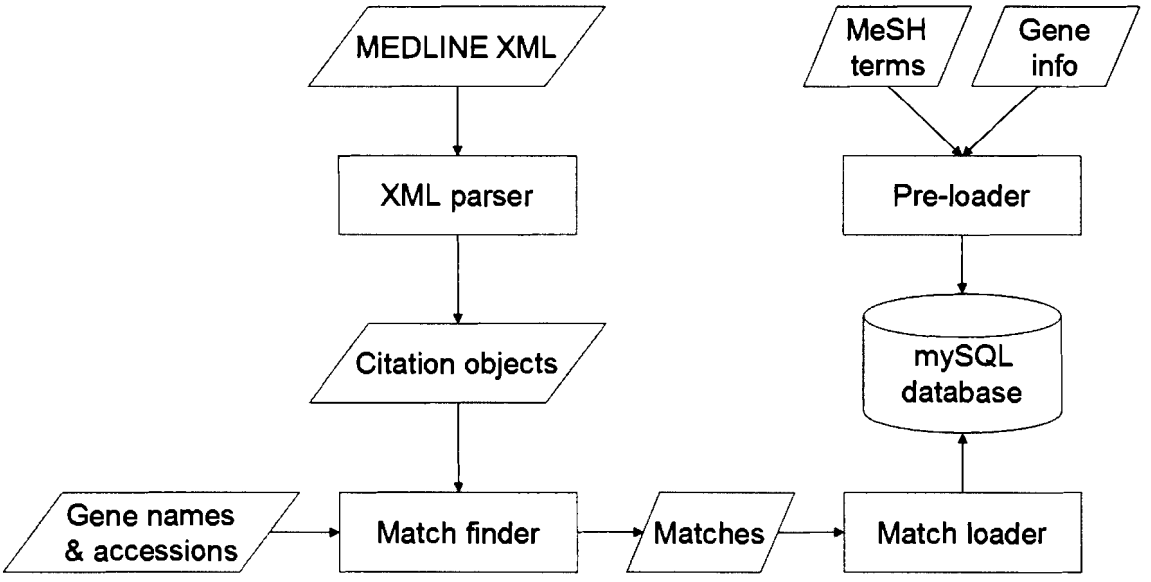
5.3 Gene finding application design

The application takes a set of gene names, symbols and accession numbers keyed by a primary gene symbol chosen by the user. We refer to the set of names, symbols, and accession numbers for a gene as its set of

synonyms. The synonym sets for each gene are not necessarily mutually exclusive. All of the synonyms are used to search for matches in MEDLINE XML citations.

The output of this application is a MySQL database (see Appendix C for the database schema). Figure 4 shows the application flow. The database is pre-loaded with the gene synonym sets and the MeSH vocabulary. The database allows a gene synonym to map to more than one gene. An optional set of gene descriptions can also be loaded.

Figure 4 Flowchart of a simple text-mining application for locating references to genes in MEDLINE XML



Once the initial data is loaded, the user can run the application in match finding mode. An XML parsing module reads the MEDLINE XML file. As it finishes reading the description of a single citation, it populates a data structure (a Citation object in Figure 4) with the information from that citation and informs the match-finding module. The match-finding module searches fields in the data

structure for matches to gene synonyms. If it finds any, it passes the citation data structure and the information about those matches to a module that loads that information into the database.

5.3.1 Match finder

When the match-finding module receives a citation from the parsing module, it examines the title and abstract for matches to gene synonyms. In addition, it searches all of the five fields listed in Table 6. The text in each field is divided into individual words. Each word is checked against the gene synonym list. To handle multiple-word gene synonyms, the match-finding module maintains a mapping of the first word to the complete synonym. Words are checked against this first word list, and if a match is found, the module looks for a match to the remainder of the synonym.

Table 7 Examples of gene synonym matches found in MEDLINE XML citations

Synonym	Gene	Field
TATA-box binding protein	TBP	Abstract
1E3G	AR	Accession
CG-1	CXorf6	Chemical
BRN-2	POU3F2	Gene symbol
Huntington	HD	MeSH Heading
Ataxin-1	ATXN1	Title

Matches are recorded as a combination of the matched synonym and the field in which the match was found. Table 7 shows some examples of actual matches. When the match finder completes its search of a citation, it checks to

see if it found any matches. If so, it passes the citation and the match information to the database-loading module.

5.4 Results

The match-finding module ran on 15,936,854 XML citations. These were obtained from the MEDLINE 2004 baseline, plus daily updates up to and including May 1, 2005. The citations represented 15,284,276 unique articles from the biomedical literature. Table 8 shows a breakdown of the non-unique citations by their MEDLINE status.

Table 8 MEDLINE status of scanned citations

Count	MEDLINE status	Description of status
13,455,005	MEDLINE	Fully curated citation, 1965 - present
1,770,731	OLDMEDLINE	Added from the NLM's* print indexes, 1950-1965
392,920	In Process	Basic citation information has been checked, additional data elements have not been added
222,091	In Data Review	Publisher has sent the basic information to the NLM*
96,107	PubMed, not MEDLINE	Non-MEDLINE records found in PubMed

*NLM = National Library of Medicine

We identified 192,681 unique articles containing at least one match to one of our gene names or accessions for a total of 268,707 matches. Most of the matches were found in the free text (title or abstract) of the citations (Table 9). Of the 834 gene names and 1,728 accessions used as search input, 401 names and 234 accessions were matched at least once.

Table 9 Matches to gene names and accessions, by matched field

Multiple matches to the same gene name/accession may occur in the same citation. Multiple names/accessions may be matched in the same citation.

Number of matches	Number of articles	Field in XML citation where match occurred
188,990	170,357	Abstract
31,578	30,231	Title
24,637	22,856	Chemical annotation
22,309	21,506	Medical Subject Heading (MeSH) term
594	550	Gene symbol annotation
475	287	Accession annotation
124	124	Keyword annotation

5.4.1 Sensitivity analysis

The first analysis for any text-mining application is to examine the sensitivity of the method. Sensitivity, or recall, is the proportion of all true matches that are correctly identified to the total number of matches (Equation 1). In general, the complete set of true matches is not known. To estimate sensitivity, therefore, we use a known subset of true matches, referred to as a gold standard. The gold standard is often a manually curated list. We estimated sensitivity using three overlapping datasets: Entrez Gene References Into Function (GeneRIFs), Entrez Gene PubMed links, and Uniprot's SwissProt Pubmed links (Figure 3 above). Recall that these datasets consist of gene-to-publication associations, where an association exists if the publication is relevant to that gene.

Equation 1 Sensitivity (recall)

TP = # of true positives, FN = # of false negatives

$$Sensitivity = \frac{TP}{TP + FN}$$

Table 10 Match finder sensitivity analysis results

The number of unique articles for each column is in brackets following the number of matches. In some cases, the number of articles with missed matches plus the number with correct matches exceeds the total number of articles. This is because some articles referring to more than one gene had some matches that were missed and some that were correctly identified, resulting in the article being counted twice. The merged dataset is the union of the three overlapping sets from different sources.

	Genes*	Publication associations	False negatives	True positives	Sensitivity
GeneRIFs	43	588 (576)	6 (6)	582 (570)	98.98%
Gene2Pubmed	64	1,471 (1,331)	166 (86)	1,305 (1,251)	88.72%
SwissProt	52	388 (340)	61 (27)	327 (315)	84.28%
Merged	64	1,656 (1,506)	186 (100)	1,470 (1,413)	88.77%

*The number of genes with at least one gene-to-publication assocation in the dataset.

We examined all 100 of the articles listed as containing matches that were not correctly identified by the match finder module. Of these, 26 were large-scale genomics papers that did not provide any specific information about the genes in question. These were responsible for 106 of the missed matches. A further 80 matches in 74 papers were missed because the genes in question were not specifically mentioned in the title, abstract, or other annotated fields.

The majority of these papers described protein families, protein complexes and protein-protein interactions. It is likely that the specific genes were

mentioned only in the full text, which is outside the scope of the match-finding module. In several cases, it appeared that the true match might have been assigned incorrectly to the gene in question, as the abstract referred to a closely related gene.

5.4.2 Specificity analysis

We assessed the specificity of each gene name/accession, to determine how much confidence we could have in a match. Specificity, or precision, is the number of true matches against the total number of matches (Equation 2). We can think of specificity as a measure of confidence. Since we had 635 gene names and accessions that were matched at least once, and 192,681 recorded gene-to-publication links, it was infeasible to check every gene name/accession and every match.

Equation 2 Specificity (precision)

TP = # of true positives, FP = # of false positives

$$Specificity = \frac{TP}{TP + FP}$$

Instead, we estimated the specificity of each gene name by manually examining a random selection of citations in which matches have been identified. Before we checked every one of the 635 names and accession numbers, we reduced the number to examine. We assumed that matches to accession numbers are unambiguous, since accession numbers are unique by design. Genomic accession numbers, however, may indicate large chunks of DNA

sequence containing multiple genes. These are therefore potentially ambiguous. This left us with 401 gene names. We further broke down the problem by examining the fields that were matched (Table 11).

Table 11 Number of gene names found each matched field in XML

Unique names	Field in XML citation where match occurred
390	Abstract
289	Title
138	Chemical annotation
72	Gene symbol annotation
12	Medical Subject Heading (MeSH) term
6	Keyword annotation
4	Accession annotation

The four names that matched in the Accession field of the XML citations translated into only three articles. Three gene names (CACNA1A, EA2, SCA6), all synonyms of a single gene, are contained in the Accession field of a single citation. In this case, the database that is being referred to is Online Mendelian Inheritance in Man (OMIM)[58], which uses gene names or symbols as its accessions. The other gene name (NP), incorrectly matched to two separate accession numbers in the RefSeq curated protein format ("NP_" followed by 9 digits).

For gene names matching in the Keyword and MeSH term fields, we examined the terms that might have given rise to the matches (Table 12). It appeared that longer, more specific gene names were more likely to be accurately matched. The exception in this table was the name "TFIID", which

refers to a well-known, long-studied gene[59]. We therefore assumed that gene names consisting of multiple words, or those that are fully spelled out, are unambiguous. There were 88 such names, reducing the total under examination to 311.

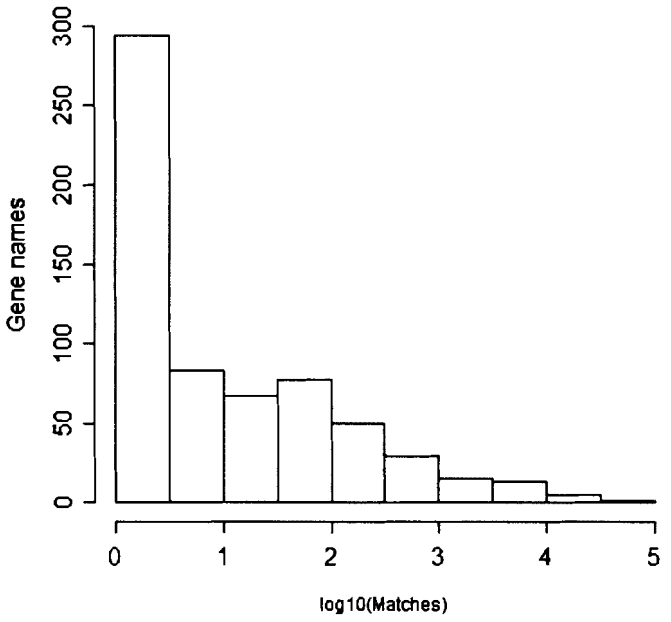
Table 12 Gene name matches in Keyword and MeSH term fields

Field	Gene name	Matched string(s)	Correct?
Keyword	AR	NASA Experiment Number AR-002	No
	BI	bi GUANIDE DERIVATIVES	No
	CAP	Cervical Cap*	No
	Huntington('s) Disease[§]	Huntington Disease*	Yes
	MINK	MINK[#]	No
MeSH	Androgen receptors	Androgen Receptors	Yes
	CAP	Cervical Cap*	No
		RNA Cap*	No
		Plant Root Cap*	No
		RNA Cap-Binding Proteins*	No
		Nuclear Cap-Binding Protein Complex*	No
	F18	Fluorodeoxyglucose F18*	No
	Huntington Disease	Huntington Disease*	Yes
	Machado-Joseph Disease	Machado-Joseph Disease*	Yes
	MINK	Mink*[#]	No
	NOD	Inbred NOD Mice*	No
	Rubenstein-Taybi Syndrome	Rubinstein-Taybi Syndrome	Yes
	TATA binding protein	TATA binding protein*	Yes
	TATA box binding protein	TATA box binding protein*	Yes
	TFIID	Transcription Factor TFIID*	Yes

*Variants represented with a single string
[§]Both "Huntington Disease" and "Huntington's Disease" matched in the Keyword field
[#]The animal, of the *Mustela* genus

We continued to reduce the problem by looking at the number of matches to each gene name, and assuming that only names with a very large number of matches are likely to be ambiguous. We considered only the 311 single-word or abbreviated gene names, and did not consider the matches we have manually checked for the Accession, Keyword, and MeSH term fields. This resulted in a distribution of match counts as in Figure 5.

Figure 5 **Distribution of match counts after removing manually identified false positive matches**



The distribution above shows clearly that a very few gene names have extremely high match counts, while most names have very few, and are therefore less likely to be problematically ambiguous. We identified the gene names with over 1,000 matches (26 names, or the top 8.3%). For each of these names, we drew a random sample of 20 citations from those that were matched, and

manually verified the citations for evidence that the matched string was in fact a gene name.

Table 13 Specificity estimates for gene names with over 1,000 matches

The first column is the number of matches that were found in MEDLINE citations for the gene name in that row. The second column is the number of unique citations represented by these matches. The third column is the number of matches divided by the unique citations. The fourth column is the gene name in question, and the fifth is the primary gene symbol to which this gene name is mapped. The sixth column is the number of correct matches for that gene name out of a random sample of 20, and the last is the resulting specificity estimate.

Matches	Unique citations	Matches per citation	Gene name	Primary symbol	Correct	Specificity estimate
49,417	45,011	1.10	KD	AR	0	<5%
18,899	15,557	1.21	CAP	BRD4	0	<5%
14,848	12,604	1.18	AR	AR	3	15%
13,733	12,196	1.13	HD	HD	8	40%
13,084	10,628	1.23	BI	CACNA1A	0	<5%
9,279	8,063	1.15	NP	ZNF384	0	<5%
6,737	2,544	2.65	CBP	CREBBP	14	70%
6,730	6,320	1.06	PH-1	PHC1	0	<5%
6,450	6,240	1.03	F18	CXORF6	0	<5%
5,795	3,852	1.50	NOD	ATN1	0	<5%
4,112	2,833	1.45	MINK	MINK1	0	<5%
3,611	3,284	1.10	DRIP	PCQAP	0	<5%
3,286	2,756	1.19	CCD	RUNX2	0	<5%
2,550	1,603	1.59	IRS-1	IRS-1	20	>95%
2,447	1,266	1.93	TFIID	TBP	20	>95%
2,382	1,775	1.34	TBP	TBP	19	95%
1,982	1,553	1.28	SMS	RAI1	0	<5%
1,879	1,183	1.59	CIP	NCOA3	0	<5%
1,835	1,476	1.24	PH1	PHC1	0	<5%
1,725	1,577	1.09	RTS	CREBBP	0	<5%
1,689	705	2.40	RUNX2	RUNX2	20	>95%

Matches	Unique citations	Matches per citation	Gene name	Primary symbol	Correct	Specificity estimate
1,664	1,421	1.17	K3	KCNN3	0	<5%
1,609	1,394	1.15	AIS	AR	4	20%
1,107	966	1.15	NRC	NCOA6	0	<5%
1,085	471	2.30	CBFA1	RUNX2	20	>95%
1,056	801	1.32	PEO	POLG	1	5%

The results of the specificity estimates are shown in Table 13. The higher the average number of matches per citation, the higher the likelihood that the match is to something important in the paper. That is, the string is frequently matched in multiple fields for a given individual citation. The five gene names with the highest average number of matches per citation are in the top six with the highest specificity estimates (CBP, RUNX2, CBFA1, TFIID, and IRS-1).

Of note, there are only a few gene names that actually showed evidence of ambiguity. Of the 26 gene names investigated, four had no false positives and 16 had all false positives in their respective samples. The large number of gene names with all false positives in their samples combined with the very small sample size could simply mean that the names are synonymous with common terms in biomedical literature.

5.5 Scoring co-citations

To score the linkage between two genes from their co-citations, we used the mutual information measure (Equation 3), a standard scoring function for measuring the degree of similarity between lists of items associated with a pair of objects[60]. The mutual information measure takes into account some of the

ambiguity and low specificity of some of the gene names as discussed above, without actually measuring the specificity of each name individually. It can also be translated directly into a score for the common evidence network, as it ranges from 0 to 1.

Equation 3 Mutual information measure

P_A, P_B = Proportion of citations associated with genes A and B, respectively
P_{AB} = Proportion of citations associated with both genes A and B
Proportions are the number of citations of interest divided by the total number of citations scanned by MEDLINE.

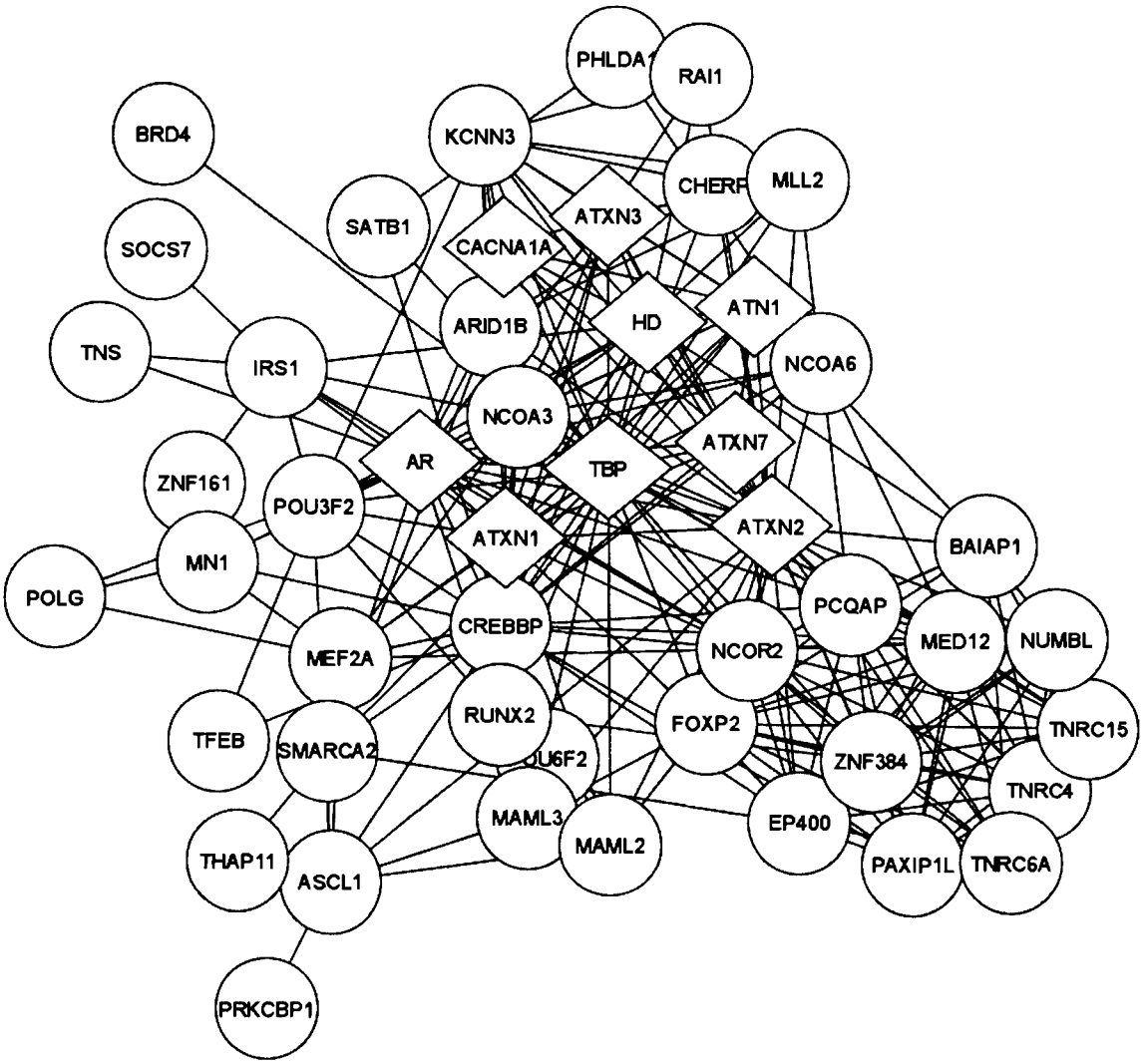
$$S(A,B) = \frac{P_{AB}}{P_A \times P_B}$$

Adapted from Alako *et al* 2005[60]

We consider only citations that have matches for more than one gene. This helps to reduce the number of false positive hits that might otherwise skew the data. It also has the effect of reducing the scores for genes that are co-cited with many other genes, and thus are likely better studied. This is of interest for biologists who are interested in discovering speculative or poorly studied linkages between genes.

Figure 6 Co-citation network

Nodes in diamonds are known disease genes. An edge indicates the genes are mentioned together in at least one citation.



A total of 3,220 citations had matches for more than one gene (for a total of 5,245 pairwise co-citations). This number was small enough that a combination of automated searching and manual examination sufficed to identify false positives. Only 1,042 citations (2,949 pairwise co-citations) remained after this process was complete, with matches for 228 unique gene names and accessions covering 49 of the 65 genes of interest. Figure 6 shows the resulting network.

The majority of the 2,949 pairwise co-citations are between pairs of known disease genes (2,183). A further 375 pairs have one member as a known disease gene. The remaining 391 are between pairs of genes that are not known to cause disease.

5.6 Comparison to other biomedical literature-mining tools

Text mining of biomedical literature is currently a topic of great interest in the bioinformatics community[61]. The results of the first BioCreAtIvE (Critical Assessment of Information Extraction in Biology) challenge were presented at a conference in March 2005[62]. Two tasks were set to competitors: identifying gene names in abstracts, and functionally annotating genes cited in an abstract. The first task was further divided into two sub-tasks: to identify mentions of genes in sentences from MEDLINE abstracts, and to map these names to unique genes. Since we are primarily interested in the co-occurrence of genes in the literature, we reviewed the results of the first task only, and concentrated on the second sub-task.

The first sub-task was a pure text-mining information extraction problem. Pre-tokenized text was analyzed to identify references to genes and entities related to genes (e.g. domains or binding sites), with no reference to a synonym list. Post-competition analysis showed that the tokenization boundaries were an important factor in correct identification of gene-related names. F-scores (Equation 4) of over 80% were reported for this sub-task[63].

Equation 4 Harmonized mean (F-score) of sensitivity and specificity
An F-score is a combined measure of accuracy.

$$F - score = \frac{2 \times Sensitivity \times Specificity}{Sensitivity + Specificity}$$

Of more interest to finding co-citations of genes, the second sub-task provided a gene synonym list and did not include the identification of gene-related entities, only genes and their products[64]. The task inputs were three sets of abstracts from MEDLINE, one for each of three model organisms (yeast, mouse, and fruit fly)[65]. A list of gene names was provided for each organism. The inputs were obtained from publication-to-gene association lists available from the model organism databases[66-68]. This meant that the problem of disambiguating gene names synonymous across different organisms could be mostly ignored for these tools.

Three of the eight teams that submitted systems for this task have published their tools. All three commented on the influence of a comprehensive gene synonym list and the problems with ambiguous gene names[69-71]. One team implemented a string matching method for comparison[69]. It achieved an estimated sensitivity of 86.1% for identifying fruit fly genes and 58.3% for mouse. However, the specificity for each of these organisms was only 3.3% and 15.1%, respectively. This reflects our low specificity estimates for some of the most frequently matched human gene names.

F-scores of over 90% were reported for some of the submitted tools[64]. For comparison, the F-score for our results, prior to manual pruning, was 68.8%. We based our specificity estimate on the manually pruned results (56.2%), and our

sensitivity estimate on the combined gene-publication associations from all three of the databases (88.8%). This result shows that string matching, which is the simplest of gene identification methods, can provide valuable information to researchers, given a comprehensive synonym list and a secondary source of evidence (i.e. requiring more than one gene occurring in an article).

5.6.1 STRING

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a top-down protein network project that includes a text-mining component[40]. It uses gene names obtained from SwissProt and organism-specific resources such as HUGO to identify instances of genes in biomedical literature. The search method is not described; however, as a component of an evidence network project, it provides a useful comparison for our co-citation evidence.

We were able to identify 63 of our 65 genes of interest in the STRING database, release 6.0. The two missing genes did not have any matches in citations with any other genes. STRING contained 846 gene-to-publication associations for our genes of interest, with the condition that each publication was associated with two or more of those genes. This represented 307 citations. STRING contained co-citations for only 21 of the genes from the list. Of interest, one of the known disease genes (ATXN2) is found in the STRING database, but it is never matched in the list of co-citations, although names of this gene do appear in some of the citations that contain references to more than one of the other genes in the list. This is because the most commonly used name for this gene, SCA2, is also a name for another gene, and the STRING literature co-

citation tool only included names that could be unambiguously mapped to a single gene (L.J. Jensen, personal communication, June 9, 2005).

Thirty-three of the STRING co-citations were not included in the manually pruned list. One citation had been removed during manual pruning because it was referring to a different pair of genes, although using the same names as a pair of our genes of interest. In a second case, the STRING application had incorrectly identified the gene name hSKCa3 as referring to the CACNA1A gene. Ours had correctly mapped hSKCa3 to the KCNN3 gene. Both CACNA1A and KCNN3 were in our list of genes of interest. A further three cases involved the mis-mapping (by STRING) of the BiP gene name to the CACNA1A gene. The remaining 28 cases all involved the mis-identification of the PCQAP gene by STRING. It was not clear why this occurred.

There were no cases where STRING had identified valid co-citations and our text-mining application had not. However, our method involved manual intervention to remove false positive matches to gene names in text, and STRING's method did not. If we use our manually pruned list as a standard, and we compare the STRING results to the results of our text-mining application prior to manual examination, it is clear that STRING has much higher specificity (96.4% vs. our 56.2%).

Higher specificity comes at a cost of lower sensitivity: only 39.5% of our manually verified matches were automatically identified by STRING. This only includes matches in citations entered before January 1, 2005, which was the cut-off date for MEDLINE citations in STRING release 6.0. If we use this sensitivity

estimate, STRING's text-mining results had an F-score of 56.0%. This is lower than our result of 68.8%. However, we suspect the STRING sensitivity result would be much higher if the algorithm had consistently identified the ATXN2 disease gene, which was mentioned frequently together with the other disease genes and accounts for 44.1% of our co-citation pairs.

CHAPTER 6: GENE ONTOLOGY EVIDENCE

6.1 Gene Ontology

The Gene Ontology (GO) project addresses the problem of biologists using multiple terms to describe the same concept[5]. This problem is another manifestation of the same root situation that gives rise to ambiguity in gene names: the evolution of vocabulary with increased information. The Gene Ontology is a controlled vocabulary of terms linked by defined relationships. It covers three categories of molecular biology terms: biological process, cellular component, and molecular function.

Each GO term consists of an identifier, the term itself, and a descriptive definition. Collaborating databases create mappings of gene products to GO terms. Annotators assign the most specific GO terms possible to a gene product. Specific in this case means the most refined description of the concept the annotator can assign to the gene. For example, if an annotator knew the gene's product is localized to the nucleolus, he would assign the GO term "nucleolus" for the gene's cellular component description, rather than "nucleus", because the nucleolus is a more specific location description. The relationships of those GO terms to more general terms imply the assignment of the general terms to that gene product. The evidence used for annotation varies from strictly computational to papers read by human curators. Table 21 in Appendix D contains the list of evidence types for annotations.

Note that GO terms are annotated to the products of genes, not to the genes themselves. We consider the GO terms annotated to the product(s) of a gene as annotated to that gene.

6.1.1 Relationships between GO terms

Defined relationships connect all GO terms in a graph. There are two types of relationships. The most common relationship is "is a". When term A "is a" child of term B, term A is a specialization of term B. The other major relationship is "part of". This relationship is especially common in the cellular component category, and means exactly what it says. For example, the term "membrane" (GO:0016020) is "part of" the term "cell" (GO:0005623). According to the Gene Ontology website, every term in the ontology will eventually be linked to the root node of its category via "is a" relationships[72].

The relationships between GO terms create a graph, with each GO term represented as a node, and each relationship represented as a directed edge from child to parent. (This research does not distinguish between "is a" and "part of" relationships.) Many biologists refer to this graph as a hierarchy. Strictly speaking, this is incorrect, because GO terms may have multiple parents. For example, in the biological process category, the term "potassium ion transport" (GO:0006813) is a child of both "monovalent inorganic cation transport" (GO:0015672) and "metal ion transport" (GO:0030001). This means that the GO term relationships define a directed acyclic graph (DAG), not a hierarchy, which is a more specialized form of a directed acyclic graph in which nodes may have only a single parent.

6.2 Methods of analyzing GO term annotations

6.2.1 Overrepresentation analysis

The most popular method of GO term annotation analysis is overrepresentation. Ten different published tools and two unpublished web-based tools can conduct this type of analysis[73-84]. To begin, the researcher chooses a set of GO terms. Given a gene set of interest, the researcher counts the number of genes that are annotated to each of the chosen GO terms, or to a descendent of one of those terms. For a baseline comparison, the researcher also counts the number of annotations from a larger set of genes (generally the entire set of known genes for the species under consideration).

The null hypothesis is that the genes of interest will be distributed among the chosen GO terms in the same proportions as the baseline set. With correction for multiple testing, the researcher can calculate a p-value for this null hypothesis. GO terms with p-values below a threshold are considered to be overrepresented for the genes of interest.

Multiple testing corrections impose limits on the number of terms that can be used, so the researcher may only select a subset of the entire ontology. This means that the choice of GO terms for the analysis is subjective and potentially biased. Across all three GO categories, there are 28 terms one level down from the roots and 536 terms two levels down. At one level down, there are already terms without children: if analysis is done for terms two levels down, it should include childless terms at the first level down.

The GO annotations themselves are biased towards better-studied genes and pathways. Statistically significant overrepresentations can occur simply because the set of all current annotations is not fully representative. It must be noted that this problem will likely affect all types of analyses. Another general problem of note is the possibility of annotation error. Most importantly, for the purposes of building a common evidence network, overrepresentation analysis does not result in information that connects pairs of genes.

A recent paper by Cheng *et al*[85] claims to describe a Gene Ontology graph-based metric for measuring similarity between pairs of GO terms. However, the scoring system they describe is dependent only on the graph location of the lowest common ancestor GO term for the pair of terms in question (see 6.4.2 for lowest common ancestor definition), and not on the relationship of those two terms to their lowest common ancestor. This means that we can assign a score to every GO term independently, based on its graph location. Terms above an arbitrary threshold score, which additionally do not have any ancestor terms with scores above that threshold, are chosen as labels for the subgraph below them. Genes are clustered by assignment to the groups defined by their annotated GO terms (genes may belong to multiple clusters), and the resulting clusters are analyzed for overrepresentation. Thus, we can view this method as an extension of overrepresentation analysis that incorporates a novel selection system for which terms to use as categories.

6.2.2 Partially ordered sets

We can view the GO DAG and the genes annotated to its terms as a group of partially ordered sets (posets). Each set is composed of the genes annotated to a single term. The partial ordering arises from the directed edge relationships between terms in the DAG. Josyln *et al*[86] described an algorithm based on poset theory that takes a list of genes and attempts to produce the best description of that set using GO terms.

Like in overrepresentation analysis, this method results in labelling groups of genes rather than pairs of genes. Unlike the aforementioned, it allows the use of all terms in the ontology. Therefore, the results of this method could be compared with the clustered results of a pairwise method. However, the authors have not provided a publicly available implementation of their method.

6.3 Data source

In order to compare the similarity of GO annotations for our gene set to a background distribution, we chose to use a single data source. We obtained GO annotations from the Entrez Gene database[48]. UniProt's manually curated SwissProt database also contains GO annotations[50]. However, for our 65 genes of interest, the annotations from Entrez Gene almost completely contain the annotations from SwissProt.

6.4 Comparing GO annotations of genes

Our goal is a quantitative measure of evidence linking a pair of genes. We can treat the similarity of GO term annotations for two genes as evidence for

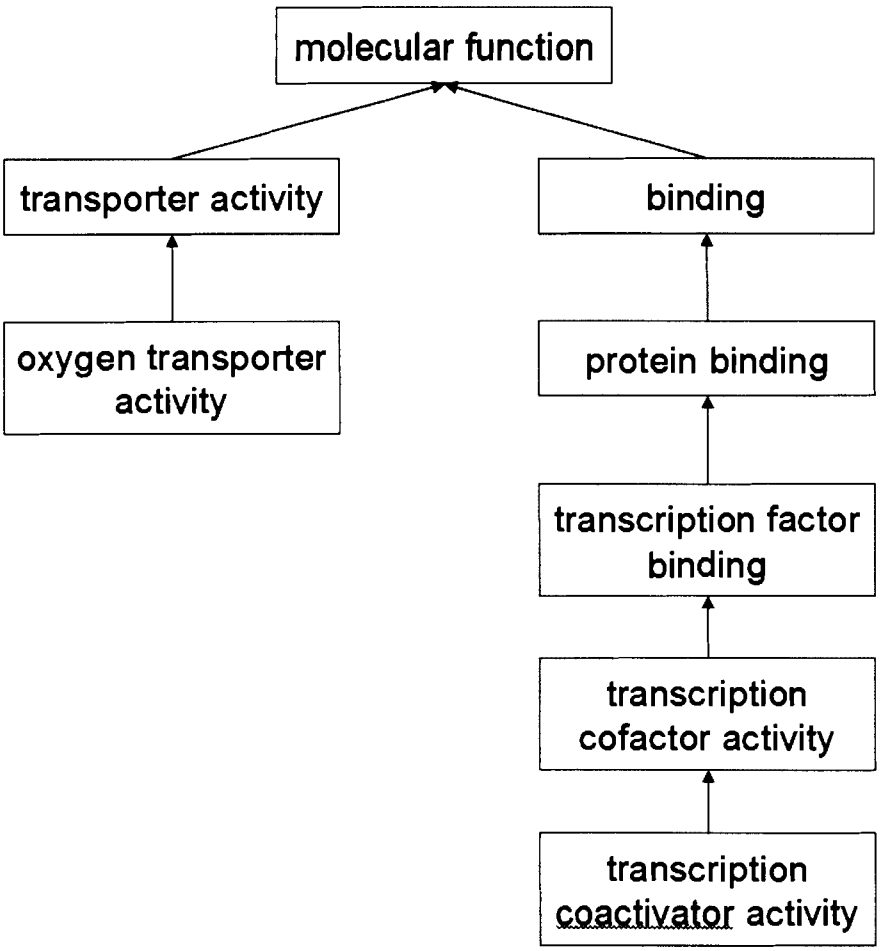
some type of relationship between them. Assuming a set of genes has been annotated with GO terms, we therefore want to determine a quantitative measure of how similar those annotations are between pairs of genes. It becomes apparent that the relationships between GO terms, while useful, do not imply any quantitative assignment of specificity to each level in the graph. To illustrate, examine the subgraph of the molecular function category shown in Figure 7.

The terms "oxygen transporter activity" and "protein binding" are the same number of edges down from the root of the graph, but the first term is the most specific term in that branch, while the second has more levels below it. This problem of differing specificity between terms at the same depth is most apparent at deeper levels of the graph, or between very divergent branches. This implies that one cannot compare different branches using an analysis method based solely on the depth of terms, because concepts on different branches become more refined at different rates.

Certain subgraphs of GO are very well defined and have many levels of refinement because research groups interested in that area have participated in the ontology development process. Other areas have received less attention. However, the problem of varying levels of concept specificity cannot be solved only by filling in the more poorly defined areas of the GO. This is because some concepts will always require more levels of refinement than others. Consistent analysis will require a measure of GO term specificity that incorporates information about the local structure of the graph.

Figure 7 **A subgraph of the Gene Ontology molecular function category**

The "transporter activity" branch becomes more specific more quickly than the "binding" branch.



6.4.1 Specificity of a GO term

Nodes in the GO graph at the same depth can have very different degrees of specificity. We cannot measure specificity exactly; however, we can make an approximation, as long as the limitations of that approximation are understood. Table 14 summarizes some potential specificity measures and their limitations. All of the measures that include the descendents of a term share the general limitation that terms with poorly defined (thus smaller and shallower) descendent subgraphs will have the same score as terms with well-defined truly small,

shallow subgraphs. The last measure appears to have the most reasonable limitations, given that we know the Gene Ontology is incomplete.

Table 14 Measures of GO term specificity

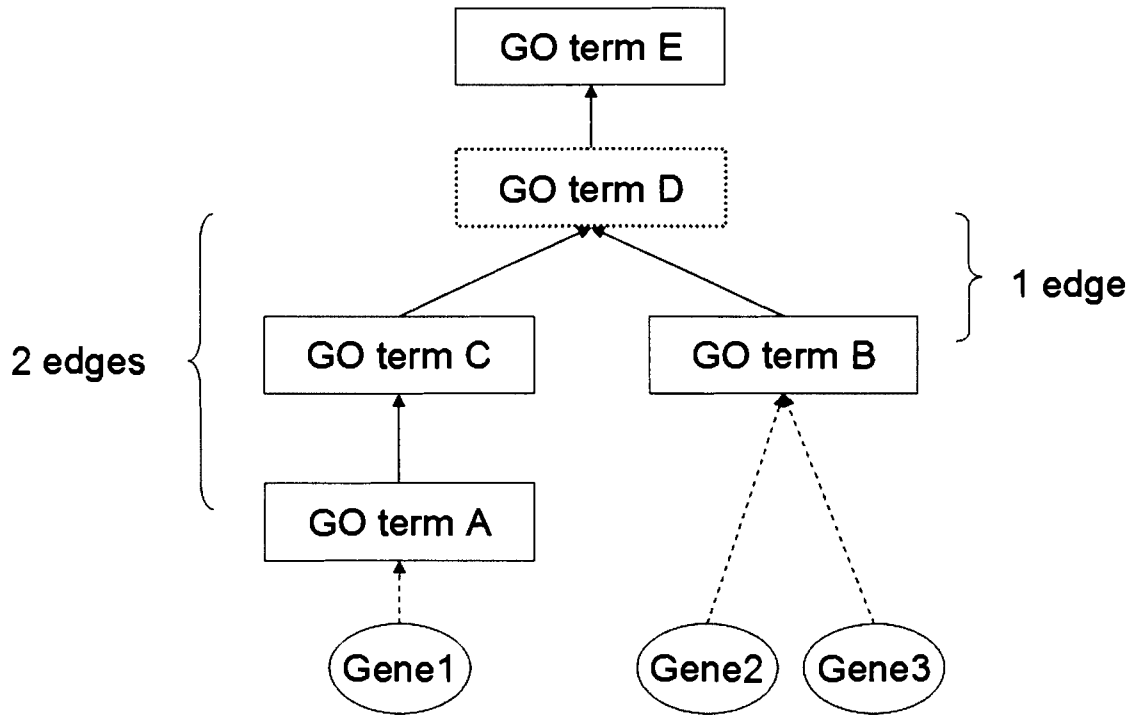
Measure	Limitations
Maximum depth of term	Does not say anything about how much more specific this term could be
Maximum path length to a leaf node descendent of this term	A single very long path will have too much influence No inclusion of information about the path(s) from the root to this term
Number of descendents below this term	Broad and shallow subgraphs will have the same score as narrow and deep subgraphs No inclusion of information about the path(s) from the root to this term
Average length of paths to leaf node descendents of this term	No inclusion of information about the path(s) from the root to this term

6.4.2 Distances between two GO terms

For a common evidence network, we need a quantification of the relationships in the GO graph that connect terms annotated to a pair of genes. We make the simplifying assumption that distances between GO terms in the graph are comparable. We define a lowest common ancestor of two nodes A and B as an ancestral node that has the minimum number of edges to each of A and B. Recall that multiple paths with different numbers of edges are possible because of multiple parents. Therefore, it is also possible that there exist multiple such lowest common ancestors.

Figure 8 Measuring distances between annotated GO terms

Gene 1 has been annotated with GO term A, and Genes 2 and 3 have been annotated with GO term B. GO term B is the lowest common ancestor annotation for Genes 2 and 3. GO term D is the lowest common ancestor annotation for both Genes 1 and 2, and for Genes 1 and 3.



In Figure 8, we illustrate the lowest common ancestor concept and its application to gene annotations. We can now measure the distance between the annotations for these genes based on the distances to the lowest common ancestor of each pair. Note that because of the possibility of multiple parents for a node in the GO graph, we will frequently see multiple lowest common ancestors for a given pair of GO terms. It is also possible that the distances from each descendent term to each lowest common ancestor could be different, resulting in the need to choose a single lowest common ancestor or combine the results from multiple lowest common ancestors.

Table 15 Distances between GO terms for example in Figure 8

In the fourth column, we have the distance from the first gene to the lowest common ancestor and the distance from the second gene to the lowest common ancestor. The distances are separated by a colon for clarity.

Gene pair	Annotated GO terms	Lowest common ancestor	Distances to lowest common ancestor
Genes 1 & 2	GO terms A & B	GO term D	2 : 1
Genes 1 & 3	GO terms A & B	GO term D	2 : 1
Genes 2 & 3	GO terms B & B	GO term B	0 : 0

6.4.3 Similarity of GO terms

Once we have calculated the distances between annotated GO terms for a given pair of genes, we need to describe a function that will quantify the similarity between each pair of GO terms. This score will help us to compare the levels of similarity between GO annotations for different pairs of genes. A naïve approach is to look at the reciprocal of the sum of the two distances to the lowest common ancestor, as in Equation 5.

Equation 5 A simple scoring function for two GO terms

***A,B* = GO terms**
C* = A lowest common ancestor of *A* and *B
|X,Y| = The number of edges between GO terms X and Y in the GO graph, plus one to avoid division by zero

$$Score = \frac{1}{1 + |A,C| + |B,C|}$$

This function does not differentiate between situations where the two GO terms in question are equally distant from the lowest common ancestor and those where one term is much closer. However, a new problem presents itself when we separate the effects of the two branch lengths to the lowest common

ancestor. If we have multiple lowest common ancestors, there is the potential that they will have different branch lengths. We must either choose one lowest common ancestor or somehow combine scores from multiple ancestors into a single result. Choosing one lowest common ancestor is the simpler solution. A simple criterion is to choose the most specific lowest common ancestor.

A more appropriate function will include factors to handle the different branch lengths and the total distance between the two terms, such as the function described in Equation 6. For a given total distance between two annotated GO terms through a lowest common ancestor, this function will put heavier weight onto paths with more evenly balanced distances. It will also put lower weight onto paths where terms are farther apart in the tree.

Equation 6 A scoring function for measuring the similarity of two GO terms, using the relative distances from annotated GO terms to a lowest common ancestor

See Equation 5 for descriptions of A, B, and C.
The first factor is the reciprocal of the sum of the two distances, as in Equation 5. The second factor is the minimum ratio of the two distances.

$$Score = \frac{1}{1+|A,C|+|B,C|} \times \frac{1+\min\{|A,C|,|B,C|\}}{1+\max\{|A,C|,|B,C|\}}$$

Both of these functions lack a factor for the specificity of the lowest common ancestor. As discussed in 6.4.1, specificity cannot truly be measured, but only approximated. We consider the inclusion of a factor for the average path length to a descendent term from the lowest common ancestor in Equation 7. We use the reciprocal of the average path length because we want to increase the score as the average path length decreases.

Equation 7 A scoring function for measuring the similarity of two GO terms, including a factor for the specificity of the lowest common ancestor

See Equation 6 for a description of the first two factors.
L_i = leaf node descendent of the lowest common ancestor C
The denominator of the third factor is the number of leaf node descendents of the lowest common ancestor C, and the numerator is the sum of path lengths from each leaf node descendent to C. Thus, the third factor is the reciprocal of the average path length from a leaf node descendent to the lowest common ancestor.

$$Score = \frac{1}{1+|A,C|+|B,C|} \times \frac{1+\min\{|A,C|,|B,C|\}}{1+\max\{|A,C|,|B,C|\}} \times \frac{1+|\{L_i\}|}{1+\sum_i |L_i,C|}$$

6.4.4 Similarity of GO term annotations for genes

A single gene can have multiple GO term annotations in the same category (cellular component, molecular function, or biological process), and thus share multiple distinct lowest common ancestors with other genes. Therefore, our function needs to take into account multiple sets of distances arising from all possible pairs of GO terms annotated to each gene in a pair. We could sum the scores for each pair of GO terms; we could take the average; we could take the maximum score; or we could allow multiple scores for each gene pair.

Combining the scores will favour genes with multiple links with more general lowest common ancestor terms over those with fewer links with more specific lowest common ancestor terms. For example, if two genes share only one identical GO term annotation, that pair could score less than two genes that share multiple distant annotations. Outputting all of the scores is also more useful for analysis, as individual scores can be combined later if the researcher wishes, whereas combined scores cannot be broken down into their component parts.

6.4.5 Score distributions

Given a pair of genes, their GO term annotations, and a comparison scoring function, we can calculate the scores for each pair of GO terms. We now want to know how significant those scores are. To find out, we want to measure the scores against a background distribution. We can then normalize a given score by mapping it to its percentile in the distribution. There are two possible sources for such a distribution: the set of all pairs of GO terms, or the set of all pairs of genes from some background set. The remainder of this section assumes we are discussing a single category of terms from the ontology.

The scores resulting from the set of all pairs of GO terms provide a background distribution that quantifies the graph structure. If we measure the score for a pair of GO terms annotated to a pair of genes against this distribution, we are measuring distances regardless of how gene annotations are actually distributed within that structure. The total number of genes assigned to a given node in the graph does not influence its weight in the distribution. Additionally, we are potentially including data from the GO graph that is not relevant to the background set of genes. For example, the GO graph contains vocabulary for plants that will never be applied to genes from animal organisms.

Because we chose to obtain GO annotations solely from Entrez Gene, the background set for our 65 polyglutamine domain genes is the set of all human protein-coding genes from that database. The list of these genes and their annotated GO terms was obtained from Entrez Gene on April 29, 2005. This

snapshot of Entrez Gene contains records for 25,370 human protein-coding genes. Of these, 15,277 have GO annotations.

We examined the similarity of the level of annotation detail for our set of genes compared to the entire data set. We took a random sample of 65 from the set of human protein-coding genes, and did this 100 times.

Table 16 compares the annotations for the random samples to those for our genes of interest and to the complete set of annotations from Entrez Gene. It is immediately obvious that our genes of interest are much better annotated than the average random sample. Across all categories, there are fewer genes without any annotations.

When we look at only those genes that actually have GO term annotations, we see that our set of interest has approximately the same (within 2 standard deviations) average number of annotations per term as the random samples. As a control, we also compared the average GO terms per gene for the random samples to the averages over the entire Entrez Gene data set. The statistics for the random samples are virtually identical to those for the whole set, which gives confidence that the random samples are representative of the set.

Table 16 **Comparing random samples of GO annotations for protein-coding genes to our genes of interest.**

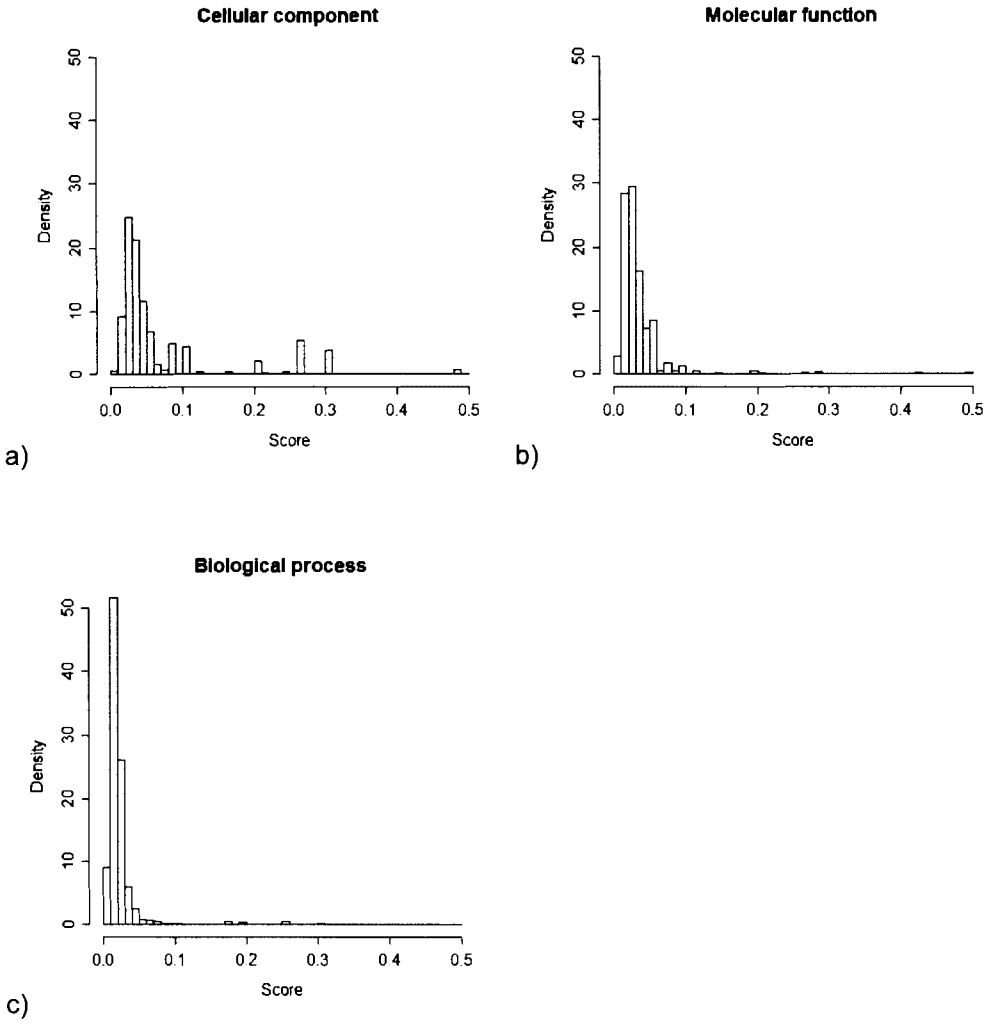
We used 100 random samples of 65 genes obtained from the Entrez Gene database downloaded April 29, 2005. The standard deviation of the values over the random samples is given in brackets for those cells.

		Number of genes with no GO annotations	Average GO terms per gene	Average GO terms per gene with at least one annotation
Cellular component	All	n/a	0.76	1.64
	Random	34.7 (3.91)	0.77 (0.12)	1.66 (0.16)
	PolyQ	10	1.03	1.37
Molecular function	All	n/a	1.21	2.42
	Random	32.1 (3.85)	1.24 (0.20)	2.46 (0.28)
	PolyQ	7	1.74	2.51
Biological process	All	n/a	1.25	2.30
	Random	29.4 (3.69)	1.26 (0.17)	2.30 (0.23)
	PolyQ	11	2.05	2.66

It was infeasible to calculate the scores for all pairs of genes in the data set; therefore, we used the random samples and combined the results. For each pair of genes within each random sample, we considered the score for each pair of GO terms annotated to those two genes. This takes into account both a measurement against the GO graph structure and the relative distribution of genes among nodes in that graph. The score distributions are shown in Figure 9.

Figure 9 **Estimated score distributions for pairs of Gene Ontology annotations**

Scores were calculated from 100 randomly selected sets of 65 genes and combined. We used Equation 7 to calculate the score distributions.



For each of the three categories of the Gene Ontology, we estimated the 99th percentile of the score distribution for a group of 65 randomly selected genes using the median of that value across the 100 random samples (Table 17). The median was preferred to the mean as an estimator due to the high variance of estimates across the random samples. In particular, the distribution of the estimator for the molecular function category had a long tail towards the higher

values. We chose to use a higher percentile to result in a stricter evaluation of the utility of the score function.

Table 17 Estimated 99th percentile of score distributions
We used Equation 7 to calculate the score distributions.

Category	Minimum	Median	Maximum
Cellular component	0.262195	0.303665	0.483871
Molecular function	0.112442	0.280657	1.000000
Biological process	0.068735	0.175160	0.281553

6.5 Results

We applied Equation 7 to score the relationships between GO term annotations for pairs of genes from our list of interest (genes encoding polyglutamine-domains with CAG repeats). Pairs of annotations scoring above the estimated 99th percentile were retained. Only one pair was above the score cut-off for the cellular component category, so we did not consider this category any further.

There were 468 gene pairs with scores above the cut-off in the biological process category, representing 42 genes. Likewise, there were 355 pairs for 46 genes in the molecular function category. Table 18 and Table 19 summarize the lowest common ancestor terms that gave rise to these links. The tables include all links above the cut-off, not only the highest-scoring link for each pair of genes. This ensures that we can evaluate all of the potentially interesting linkages.

Table 18 Lowest common ancestor terms linking gene pairs – biological process

Term ID	Term definition	Number of genes
GO:0006355	regulation of transcription, DNA-dependent	24
GO:0006350	transcription	16
GO:0007399	neurogenesis	9
GO:0007165	signal transduction	5
GO:0006366	transcription from RNA polymerase II promoter	4
GO:0006357	regulation of transcription from RNA polymerase II promoter	3
GO:0006367	transcription initiation from RNA polymerase II promoter	3
GO:0007268	synaptic transmission	3
GO:0000398	nuclear mRNA splicing, via spliceosome	2
GO:0006281	DNA repair	2
GO:0006461	protein complex assembly	2
GO:0006468	protein amino acid phosphorylation	2
GO:0006810	transport	2
GO:0007242	intracellular signaling cascade	2
GO:0007417	central nervous system development	2
GO:0007601	visual perception	2
GO:0030518	steroid hormone receptor signaling pathway	2
GO:0045893	positive regulation of transcription, DNA-dependent	2

Table 19 Lowest common ancestor terms linking gene pairs – molecular function

Term ID	Term definition	Number of genes
GO:0003677	DNA binding	17
GO:0003700	transcription factor activity	14
GO:0005515	protein binding	11
GO:0008270	zinc ion binding	8
GO:0005524	ATP binding	7
GO:0003713	transcription coactivator activity	5
GO:0003723	RNA binding	3
GO:0046966	thyroid hormone receptor binding	3
GO:0003702	RNA polymerase II transcription factor activity	2

Term ID	Term definition	Number of genes
GO:0003714	transcription corepressor activity	2
GO:0004386	helicase activity	2
GO:0004402	histone acetyltransferase activity	2
GO:0004674	protein serine/threonine kinase activity	2
GO:0005509	calcium ion binding	2
GO:0016563	transcriptional activator activity	2
GO:0030374	ligand-dependent nuclear receptor transcription coactivator activity	2

We created graphs for the molecular function and biological process categories where each node represented a single gene, and weighted edges represented pairs of GO term annotations and their scores. We applied a simple visual clustering algorithm that assigns shorter lengths to edges with higher weights. The resulting graphs were labelled with the GO terms that best described each cluster, based on the groupings in Table 18 and Table 19. Note that not all of the terms from the tables are used in the graphs, as this is visually too complex.

Figure 10 Gene clusters based on scored relationships between annotated GO terms – biological process

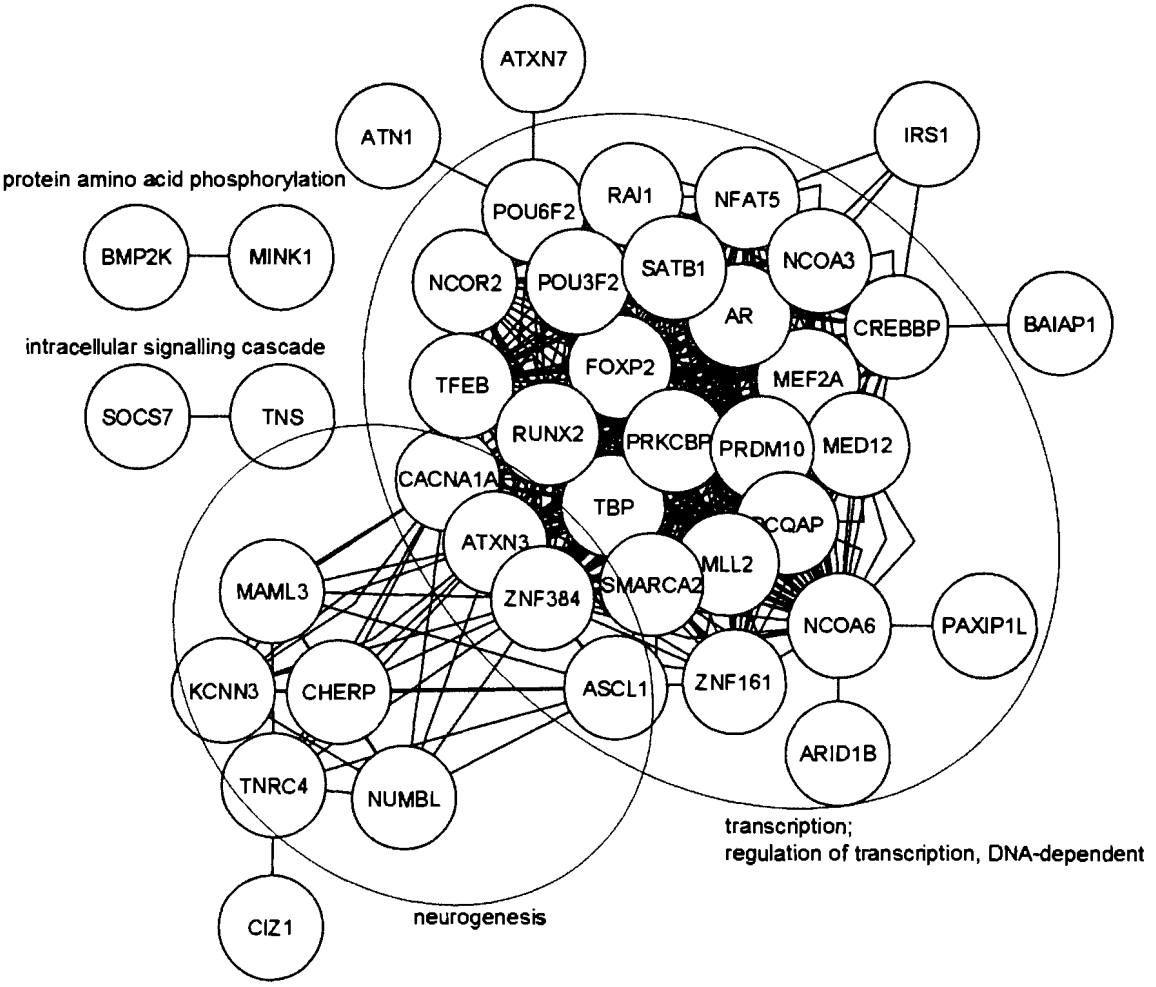
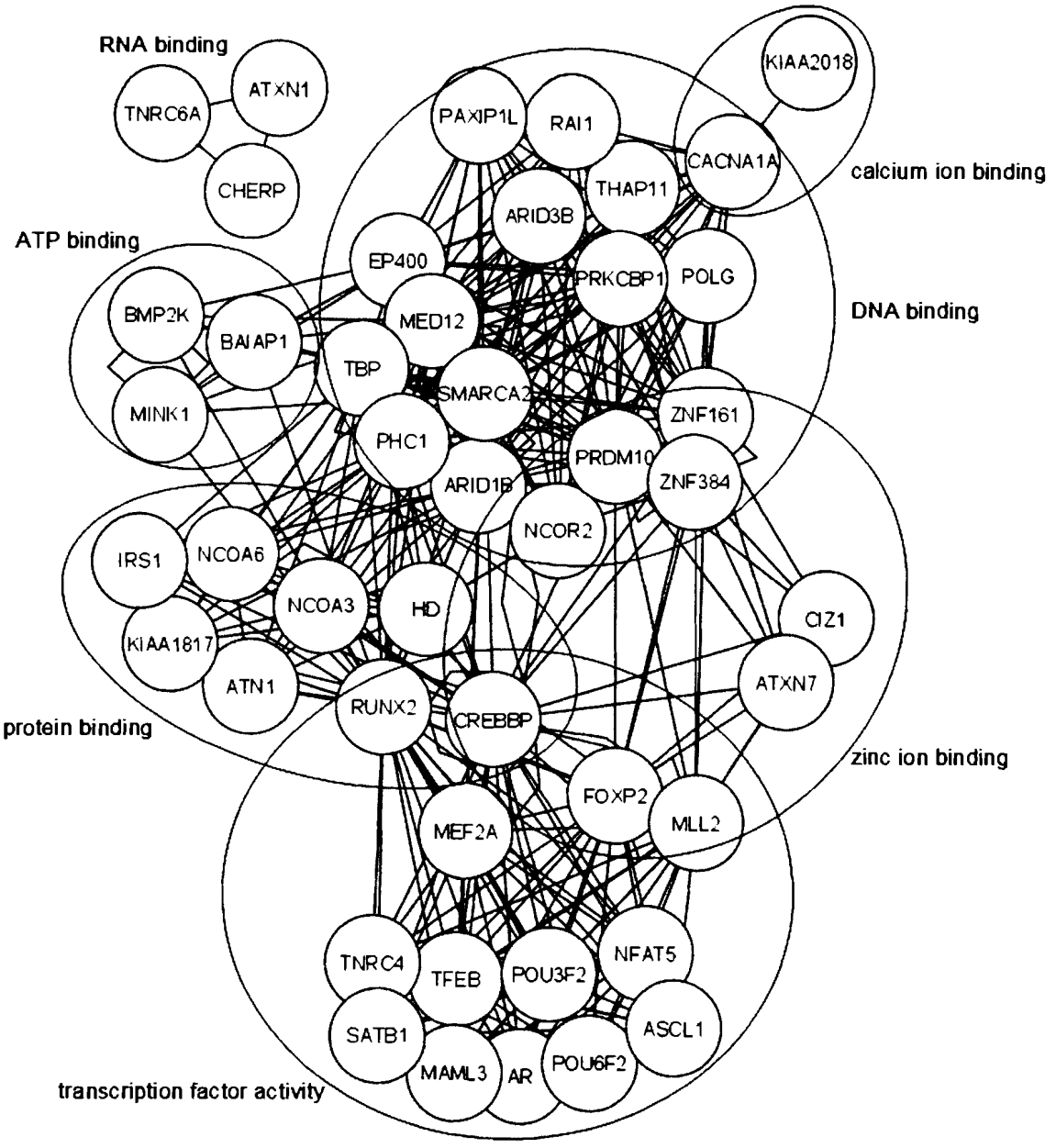


Figure 11 Gene clusters based on scored relationships between annotated GO terms – molecular function



CHAPTER 7: DISCUSSION

7.1 Summary of results

The original goal of this research was to produce a network of evidence supporting relationships between genes, based on co-citation of genes in biomedical literature abstracts. The network was to link genes containing polyglutamine repeats as defined by the Genomic Mutational Signatures project. Indirect, or off-by-one, relationships were to be included if possible. The co-citation evidence was to form the first layer of edges in a network that would integrate multiple sources of evidence, called a common evidence network (Chapter 2).

7.1.1 Gene name ambiguity

Exploration of biomedical text-mining literature and our first attempts at building a co-citation network brought an underlying problem to light. We found ambiguity of gene names within species, between species, with disease names, abbreviations of biomedical terms, and with the English language (Chapter 3). These ambiguities caused two problems. First, we needed a comprehensive list of gene names and accessions, all mapped to their respective genes. However, no individual gene- or protein-centric database had a complete list for all of the genes on our list of interest.

We built an application to identify gene records in three selected databases: Entrez Gene[48], SwissProt[50], and HUGO Genew[46]. It can be extended by adding modules to read and extract information from new databases. The application searches these databases using one or more names for each gene, and allows the user to resolve instances where more than one gene record matches a given name (Chapter 4). It extracts gene names, descriptions, accessions, associated publications, and annotated Gene Ontology terms from each gene record and outputs text files in a simple format for each information type. For each gene and a given information type, the data is combined into a single output line.

7.1.2 Co-citation evidence

The second problem was to resolve ambiguities detected by our text-mining application. The purpose of this application was to uniquely identify gene mentions in biomedical citations (Chapter 5). Given the comprehensive synonym list from our information gathering utility application, we exactly matched gene names and accessions in the abstract, title, and various curated annotations for each citation. An examination of the synonyms with the most matches showed generally very low specificity (precision) estimates when we considered all 192,681 citations with at least one match. In contrast, we estimated between 84.28% and 98.98% sensitivity (recall) based on three separate gene-to-publication association lists.

We found only 3,022 citations (5,245 pairwise co-citations) with matches to at least two distinct genes. Of these, 1,042 (2,949 pairwise co-citations) were

manually validated as containing correct matches. A citation was removed from this list if it was found to contain incorrect matches that resulted in less than two distinct genes being associated with that citation. From these results, we estimated a specificity of 56.22% for the automated system when co-citation is a requirement. This gave an F-score (combined sensitivity and specificity) of 68.8%, which compared favourably with the co-citation results using the STRING application[40]. STRING had specificity of 96.4% and sensitivity of 39.5% for an F-score of 56.0% on the same set of genes.

We noted that the whole genome, top-down approach used by the STRING project placed greater emphasis on specificity, and did not include gene names that mapped to more than one gene. One gene name that was not included by STRING accounted for 44.1% of the co-citation pairs found in our manually validated set. The STRING co-citation pairs were a proper subset of our results; we did not miss any of the matches they found. We conclude that a sensitivity-focused, bottom-up approach centred on a relatively small set of genes has the potential to fill in gaps in the results of specificity-focused, top-down whole genome approaches.

The manually validated co-citation pairs were scored using the mutual information measure. By using only the co-citation results, this measure assigns higher scores to pairs of genes that occur rarely in pairs, but tend to occur together. It assigns lower scores to genes that commonly occur in pairs, such as the known ataxia disease genes. Scores range between 0 and 1, and are thus directly usable as edge weights in a common evidence network.

7.1.3 Gene Ontology evidence

During the course of this project, the GeMS team began examining its set of polyglutamine repeat containing genes for similarity of Gene Ontology[87] terms. They approached the problem using overrepresentation analysis provided by the GoMiner tool[76]. The depth limitations of this type of analysis led us to explore methods for pairwise comparison of the GO terms annotated to a set of genes (Chapter 6). We designed a novel scoring system based on a measure of the specificity of the path linking two GO terms in the GO DAG. Significant scores were considered to be those above the 99th percentile of a bootstrapped score distribution based on samples of the same size as that of the gene set of interest.

All three GO categories (biological process, molecular function, and cellular component) were considered as separate layers of evidence. This was partly because some genes have annotations in one category but not others, but mostly because the three categories are not comparable. We found only one gene pair from our test set of genes that had a score over the 99th percentile in the cellular component category, therefore we did not continue analysis for that category. We normalized all scores from the biological process and molecular function categories to the range of 0-1 for use as edge weights in a common evidence network.

7.2 Future work

A common evidence network is only as useful as the data on which it is based. The more layers of accurate data that can be incorporated, the more complete the network becomes, and the more potential use it has as an

interpretive tool of use to scientists. A relationship between two genes connected by multiple low-weight edges may be worth further investigation. We propose the inclusion of more evidence types and the inclusion of more genes through indirect linkages (as discussed in 2.3.1 above) as a means to increase confidence in more weakly supported putative relationships.

The integration of multiple evidence types into a single network is the end goal of this research. It is not enough to simply layer the results of each evidence type together. We will need tools and methods for visualizing and analyzing a common evidence network. Open source graph visualization tools such as Cytoscape[6] and Osprey[7] can be adapted for this purpose. It is possible to write additional software modules for clustering and other analysis methods. These modules can be directly incorporated into the aforementioned visualization tools.

7.2.1 Additional data

Two naturally paired evidence types are obvious candidates for layers in a common evidence network: protein-protein interactions and gene co-expression. If two proteins interact, the research hypothesis is that they perform some biological function together. Data from protein-protein interaction databases is readily available for public use[88-93]. Some groups have begun to integrate data from across multiple databases[40, 94]. Of particular interest for networks focusing on human genes, interactions of human proteins can be predicted based on interactions of the corresponding proteins found in model organisms such as mouse, fruit fly, or worm[95]. A quantitative measure for interaction confidence

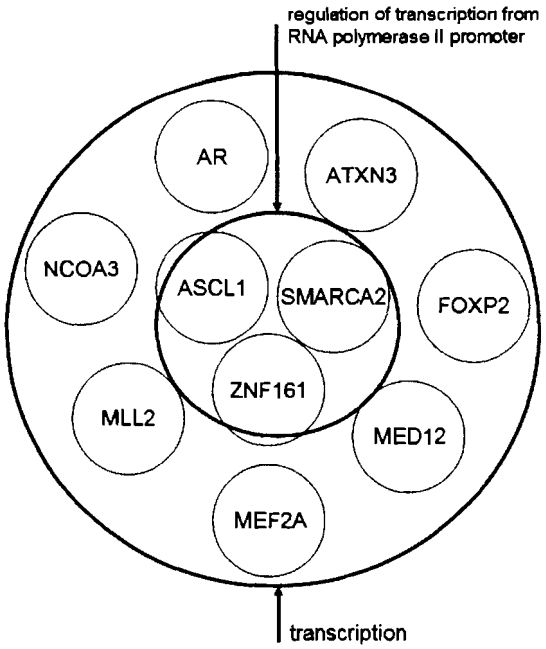
could be related to the experiment type, including its classification as high- or low-throughput.

Gene expression experiments measure the relative amounts of mRNA produced by genes in a cell at a given moment in time. Genes with similar changes in expression levels across multiple cell conditions are said to be co-expressed. The hypothesis is that genes with similar changes in expression are similarly regulated[33, 34]. Gene expression data is frequently noisy due to the limits of the experimental techniques involved[96]. Confidence in a linkage grows with the number of sets of experiments in which a pair of genes is considered co-expressed. A quantitative measure of the co-expression linkage between two genes could be based simply on the number of experiment sets in which the co-expression is seen[97].

7.2.2 Gene Ontology clustering and graph visualization tool

A particularly useful visualization module would specifically target the results of the Gene Ontology analysis to automatically produce graphs such as Figure 10 and Figure 11. The clustering algorithm used for those graphs is based on spring tension. Higher edge weights equate to higher spring tension between two genes, drawing them together. The graph reaches equilibrium when the spring tensions are balanced. We propose a modification to this method that considers the relationships between the GO terms corresponding to edge labels.

Figure 12 Example output from proposed Gene Ontology cluster visualization tool (Edges not shown)



Recall that the lowest common ancestor between two GO terms is the primary factor in the equation scoring the relationship between those terms. We can label the resulting edge not only with the weight that comes from the score, but also with the lowest common ancestor GO term. Consider a cluster of genes in which a small core group is linked by GO term A, and a peripheral group is linked by GO term B, where term A is a descendent of term B. Additionally, genes in the core group are linked to genes in the peripheral group by term B. A visually appealing cluster algorithm would place the core group in the centre of the cluster, labelling it by term A, place the genes in the peripheral group around those in the core group, and label the entire group by term B (Figure 12).

The tool would need to be aware of the relationships between the subset of Gene Ontology terms that are in use as edge labels. It would start with genes

linked by GO terms that do not have any descendents as edge labels in the graph, and work upwards in the GO DAG. Some form of conflict resolution would be required for cases where genes belong to multiple groups.

7.3 Conclusion

The goal of this research was to describe a data structure to represent evidence for relationships between genes, and to design methods for obtaining some types of evidence and measuring confidence in that evidence. A network is an appropriate data structure. Existing visualization and analytic tools can be applied, and new types of evidence can be added. This type of bottom-up approach is targeted towards small lists of genes focused around a specific biological question. Methods for adding evidence types can therefore favour sensitivity over specificity, as opposed to top-down, whole genome approaches.

To add evidence, we first needed to identify gene records in databases. We found that the ambiguity of gene names was an underlying issue, requiring a tool for extracting and collating information from gene records from multiple databases. We developed a method to gather evidence based on co-citations in biomedical abstracts. The co-citation method compares favourably to other published tools. We presented a novel approach to quantifying the similarity of Gene Ontology terms annotated to a pair of genes. We tested both methods on the list of human polyglutamine domain containing genes.

APPENDIX A: CANDIDATE DISEASE GENES

Table 20 Genes of interest (candidate and known disease genes)

The gene symbol and name shown are the official approved symbols and names from the Human Genome Organization Nomenclature Committee, where available.

Gene symbol	Gene name
AR*	androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease)
ARID1B	AT rich interactive domain 1B (SWI1-like)
ARID3B	AT rich interactive domain 3B (BRIGHT- like)
ASCL1	achaete-scute complex-like 1 (Drosophila)
ATN1*	dentatorubral-pallidoluysian atrophy (atrophin-1)
ATXN1*	ataxin 1
ATXN2*	ataxin 2
ATXN3*	ataxin 3
ATXN7*	ataxin 7
BAIAP1	BAI1-associated protein 1
BMP2K	BMP2 inducible kinase
BRD4	bromodomain containing 4
C14ORF4	chromosome 14 open reading frame 4
C9ORF43	chromosome 9 open reading frame 43
CACNA1A*	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit
CHERP	calcium homeostasis endoplasmic reticulum protein
CIZ1	CDKN1A interacting zinc finger protein 1
CREBBP	CREB binding protein (Rubinstein-Taybi syndrome)
CXORF6	chromosome X open reading frame 6
DCP1B	DCP1 decapping enzyme homolog B (S. cerevisiae)
EP400	E1A binding protein p400
FOXP2	forkhead box P2
HD*	huntingtin (Huntington disease)
IRS1	insulin receptor substrate 1
KCNN3	potassium intermediate/small conductance calcium-activated channel,

	subfamily N, member 3
KIAA0476	KIAA0476
KIAA1817**	KIAA1817 protein
KIAA2018	KIAA2018
MAML2	mastermind-like 2 (Drosophila)
MAML3	mastermind-like 3 (Drosophila)
MED12	mediator of RNA polymerase II transcription, subunit 12 homolog (yeast)
MEF2A	MADS box transcription enhancer factor 2, polypeptide A (myocyte enhancer factor 2A)
MINK1	misshapen-like kinase 1 (zebrafish)
MLL2	myeloid/lymphoid or mixed-lineage leukemia 2
MN1	meningioma (disrupted in balanced translocation) 1
NCOA3	nuclear receptor coactivator 3
NCOA6	nuclear receptor coactivator 6
NCOR2	nuclear receptor co-repressor 2
NFAT5	nuclear factor of activated T-cells 5, tonicity-responsive
NUMBL	numb homolog (Drosophila)-like
PAXIP1L	PAX transcription activation domain interacting protein 1 like
PCQAP	PC2 (positive cofactor 2, multiprotein complex) glutamine/Q-rich-associated protein
PHC1	polyhomeotic-like 1 (Drosophila)
PHLDA1	pleckstrin homology-like domain, family A, member 1
POLG	polymerase (DNA directed), gamma
POU3F2	POU domain, class 3, transcription factor 2
POU6F2	POU domain, class 6, transcription factor 2
PRDM10	PR domain containing 10
PRKCBP1	protein kinase C binding protein 1
RAI1	retinoic acid induced 1
RUNX2	runt-related transcription factor 2
SATB1	special AT-rich sequence binding protein 1 (binds to nuclear matrix/scaffold-associating DNA's)
SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2
SOCS7	suppressor of cytokine signaling 7
ST6GALNAC5	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 5
TBP*	TATA box binding protein

TFEB	transcription factor EB
THAP11	THAP domain containing 11
TNRC15	trinucleotide repeat containing 15
TNRC4	trinucleotide repeat containing 4
TNRC6A	trinucleotide repeat containing 6A
TNRC6B	trinucleotide repeat containing 6B
TNS	tensin
ZNF161	zinc finger protein 161
ZNF384	zinc finger protein 384

* Known polyglutamine expansion disease gene
** No HGNC-approved symbol/name available

APPENDIX B: MEDLINE XML EXAMPLE

Figure 13 An example of a MEDLINE XML citation

Displayed fields are those contained in the citation data structure produced by the XML parsing module. Some fields have been condensed for the purposes of display. The original citation does not contain the GeneSymbol or Accession fields. These were manually added.

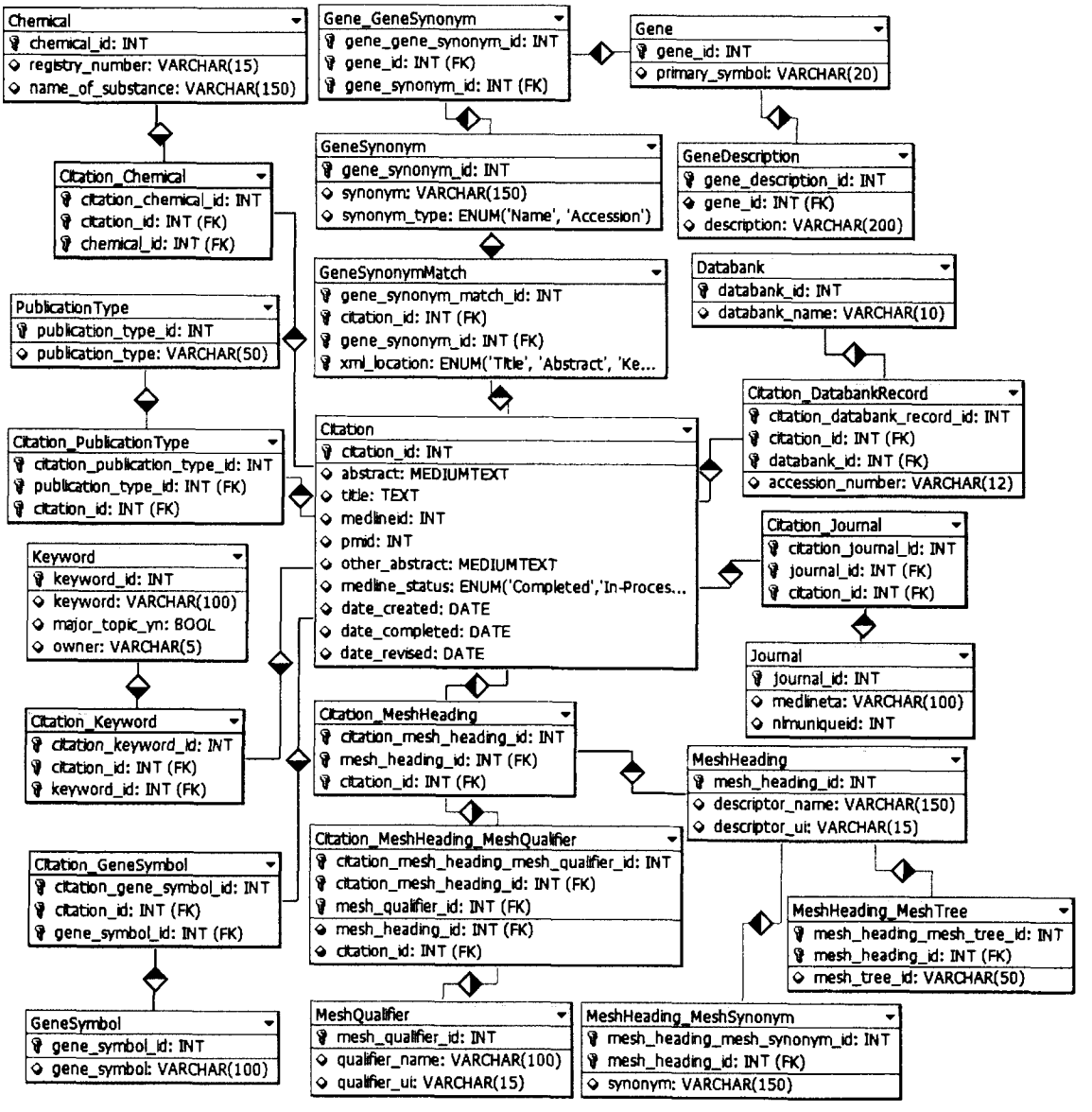
```
<MedlineCitation Status="MEDLINE">
    <PMID>12881722</PMID>
    <DateCreated>2003-08-29</DateCreated>
    <DateCompleted>2003-09-25</DateCompleted>
    <DateRevised>2004-11-17</DateRevised>
    <Article>
        <ArticleTitle>Huntingtin interacts with REST/NRSF to modulate the transcription of
        NRSE-controlled neuronal genes.</ArticleTitle>
        <Abstract>
            <AbstractText>Huntingtin protein is mutated in Huntington disease. We
            previously reported that wild-type but not mutant huntingtin stimulates
            transcription of the gene encoding brain-derived neurotrophic factor (BDNF;
            ref. 2). Here we show that the neuron restrictive silencer element (NRSE) is
            the target of wild-type huntingtin activity on BDNF promoter II. Wild-type
            huntingtin inhibits the silencing activity of NRSE, increasing transcription
            of BDNF. We show that this effect occurs through cytoplasmic sequestering of
            repressor element-1 transcription factor/neuron restrictive silencer factor
            (REST/NRSF), the transcription factor that binds to NRSE. In contrast,
            aberrant accumulation of REST/NRSF in the nucleus is present in Huntington
            disease. We show that wild-type huntingtin coimmunoprecipitates with REST/NRSF
            and that less immunoprecipitated material is found in brain tissue with
            Huntington disease. We also report that wild-type huntingtin acts as a
            positive transcriptional regulator for other NRSE-containing genes involved in
            the maintenance of the neuronal phenotype. Consistently, loss of expression of
            NRSE-controlled neuronal genes is shown in cells, mice and human brain with
            Huntington disease. We conclude that wild-type huntingtin acts in the
            cytoplasm of neurons to regulate the availability of REST/NRSF to its nuclear
            NRSE-binding site and that this control is lost in the pathology of Huntington
            disease. These data identify a new mechanism by which mutation of huntingtin
            causes loss of transcription of neuronal genes.</AbstractText>
            </Abstract>
        <DatabankList>
            <Databank>
                <DatabankName>GENBANK</DatabankName>
                <AccessionNumberList>
                    <AccessionNumber>NP_002102</AccessionNumber>
                    <AccessionNumber>NP_001700</AccessionNumber>
                    <AccessionNumber>NP_005603</AccessionNumber>
                </AccessionNumberList>
            </Databank>
        </DatabankList>
        <PublicationTypeList>
            <PublicationType>Journal Article</PublicationType>
        </PublicationTypeList>
    </Article>
    <MedlineJournalInfo>
        <MedlineTA>Nat Genet</MedlineTA>
        <NlmUniqueID>9216904</NlmUniqueID>
    </MedlineJournalInfo>
    <ChemicalList>
        <Chemical>
            <RegistryNumber>0</RegistryNumber>
            <NameOfSubstance>Brain-Derived Neurotrophic Factor</NameOfSubstance>
        </Chemical>
        <Chemical>
            <RegistryNumber>0</RegistryNumber>
            <NameOfSubstance>Huntingtin protein, human</NameOfSubstance>
        </Chemical>
```

```
    <Chemical>
        <RegistryNumber>0</RegistryNumber>
        <NameOfSubstance>Nerve Tissue Proteins</NameOfSubstance>
    </Chemical>
    <Chemical>
        <RegistryNumber>0</RegistryNumber>
        <NameOfSubstance>Nuclear Proteins</NameOfSubstance>
    </Chemical>
    <Chemical>
        <RegistryNumber>0</RegistryNumber>
        <NameOfSubstance>Repressor Proteins</NameOfSubstance>
    </Chemical>
    <Chemical>
        <RegistryNumber>0</RegistryNumber>
        <NameOfSubstance>Transcription Factors</NameOfSubstance>
    </Chemical>
    <Chemical>
        <RegistryNumber>0</RegistryNumber>
        <NameOfSubstance>transcription factor REST</NameOfSubstance>
    </Chemical>
</ChemicalList>
<GeneSymbolList>
    <GeneSymbol>HD</GeneSymbol>
    <GeneSymbol>NRSE</GeneSymbol>
    <GeneSymbol>REST</GeneSymbol>
    <GeneSymbol>BDNF</GeneSymbol>
</GeneSymbolList>
<MeshHeadingList>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Animals</DescriptorName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Brain-Derived Neurotrophic
        Factor</DescriptorName>
        <QualifierName MajorTopicYN="Y">genetics</QualifierName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Cell Line</DescriptorName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="Y">Gene Expression Regulation</DescriptorName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Humans</DescriptorName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Huntington Disease</DescriptorName>
        <QualifierName MajorTopicYN="N">genetics</QualifierName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Mice</DescriptorName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Mice, Knockout</DescriptorName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Nerve Tissue Proteins</DescriptorName>
        <QualifierName MajorTopicYN="N">genetics</QualifierName>
        <QualifierName MajorTopicYN="Y">physiology</QualifierName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Neurons</DescriptorName>
        <QualifierName MajorTopicYN="Y">physiology</QualifierName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Nuclear Proteins</DescriptorName>
        <QualifierName MajorTopicYN="N">genetics</QualifierName>
        <QualifierName MajorTopicYN="Y">physiology</QualifierName>
    </MeshHeading>
    <MeshHeading>
        <DescriptorName MajorTopicYN="N">Promoter Regions (Genetics)</DescriptorName>
```

```
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Rats</DescriptorName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Rats, Sprague-Dawley</DescriptorName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Repressor Proteins</DescriptorName>
            <QualifierName MajorTopicYN="Y">genetics</QualifierName>
            <QualifierName MajorTopicYN="N">physiology</QualifierName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Research Support, Non-U.S.
            Gov't</DescriptorName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Silencer Elements,
            Transcriptional</DescriptorName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Transcription Factors</DescriptorName>
            <QualifierName MajorTopicYN="Y">genetics</QualifierName>
            <QualifierName MajorTopicYN="N">physiology</QualifierName>
        </MeshHeading>
        <MeshHeading>
            <DescriptorName MajorTopicYN="N">Transcription, Genetic</DescriptorName>
        </MeshHeading>
    </MeshHeadingList>
</MedlineCitation>
```

APPENDIX C: DATABASE SCHEMA

Figure 14 Database schema for text-mining application



APPENDIX D: GENE ONTOLOGY EVIDENCE TYPES

Table 21 Types of evidence for GO term annotations

Evidence type	Description
Inferred by curator (IC)	• To be used for those cases where an annotation is not supported by any evidence, but can be reasonably inferred by a curator from other GO annotations, for which evidence is available.
Inferred from direct assay (IDA)	• Enzyme assays • In vitro reconstitution (e.g. transcription) • Immunofluorescence (for cellular component) • Cell fractionation (for cellular component) • Physical interaction/binding assay (sometimes appropriate for cellular component or molecular function)
Inferred from electronic annotation (IEA)	• Annotations based on "hits" in sequence similarity searches, if they have not been reviewed by curators (curator-reviewed hits would get ISS) • Annotations transferred from database records, if not reviewed by curators (curator-reviewed items may use NAS, or the reviewing process may lead to print references for the annotation)
Inferred from expression pattern (IEP)	• Transcript levels (e.g. Northerns, microarray data) • Protein levels (e.g. Western blots)
Inferred from genetic interaction (IGI)	• "Traditional" genetic interactions such as suppressors, synthetic lethals, etc. • Functional complementation • Rescue experiments • Inference about one gene drawn from the phenotype of a mutation in a different gene
Inferred from mutant phenotype (IMP)	• Any gene mutation/knockout • Overexpression/ectopic expression of wild-type or mutant genes • Anti-sense experiments • RNAi experiments • Specific protein inhibitors

Evidence type	Description
Inferred from physical interaction (IPI)	• 2-hybrid interactions • Co-purification • Co-immunoprecipitation • Ion/protein binding experiments
Inferred from sequence or structural similarity (ISS)	• Sequence similarity (homologue of/most closely related to) • Recognized domains • Structural similarity • Southern blotting • Protein features, predicted or observed (e.g. hydrophobicity, sequence composition)
Non-traceable author statement (NAS)	• Database entries that don't cite a paper (e.g. UniProt Knowledgebase records, YPD protein reports) • Statements in papers (abstract, introduction, or discussion) that a curator cannot trace to another publication
No biological data available (ND)	• Used for annotations to "unknown" molecular function, biological process, or cellular component.
Inferred from reviewed computational analysis (RCA)	• Predictions based on large-scale protein interaction experiments • Predictions based on integration of large-scale datasets of several types • Text-based computation (e.g. text mining)
Traceable author statement (TAS)	• Anything in a review article where the original experiments are traceable through that article (material from introductions to non-review papers will sometimes meet this standard; discussion sections should usually be regarded with greater skepticism) • Anything found in a text book or dictionary; usually text book material has become common knowledge (e.g. "everybody" knows that enolase is a glycolytic enzyme).
Not recorded (NR)	• Used for annotations done before curators began tracking evidence types (appears in SGD and FlyBase annotations). It should not be used for new annotations: use TAS or NAS.

Source: GO Evidence Code Guide[98]

REFERENCE LIST

1. Genome Glossary. Available at: http://www.ornl.gov/sci/techresources/Human_Genome/glossary/glossary.shtml.

2. Yu U, Lee SH, Kim YJ, Kim S. Bioinformatics in the post-genome era. *J Biochem Mol Biol*. 2004;37:75-82.

3. Bader GD, Enright AJ. Intermolecular interactions and biological pathways. In: Baxevanis AD, Ouellette BFF, eds. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Third ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2005:254.

4. Grindrod P, Kibble M. Review of uses of network and graph theory concepts within proteomics. *Expert Rev Proteomics*. 2004;1:229-238.

5. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. the gene ontology consortium. *Nat Genet*. 2000;25:25-29.

6. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498-2504.

7. Breitkreutz BJ, Stark C, Tyers M. Osprey: A network visualization system. *Genome Biol*. 2003;4:R22.

8. Gunawardena S, Goldstein LS. Polyglutamine diseases and transport problems: Deadly traffic jams on neuronal highways. *Arch Neurol.* 2005;62:46-51.

9. van Vught PW, Veldink JH, Baas F, van Muiswinkel FL, van den Berg LH. From gene to disease: Amyotrophic lateral sclerosis. *Ned Tijdschr Geneeskd.* 2004;148:2125-2127.

10. GeMS - Genomic Mutational Signature Sequences. Available at: http://bioinformatics.ubc.ca/gems.

11. Goldberg YP, Kremer B, Andrew SE, et al. Molecular analysis of new mutations for huntington's disease: Intermediate alleles and sex of origin effects. *Nat Genet.* 1993;5:174-179.

12. Rubinsztein DC, Leggo J, Coles R, et al. Phenotypic characterization of individuals with 30-40 CAG repeats in the huntington disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats. *Am J Hum Genet.* 1996;59:16-22.

13. McNeil SM, Novelletto A, Srinidhi J, et al. Reduced penetrance of the huntington's disease mutation. *Hum Mol Genet.* 1997;6:775-779.

14. Brinkman RR, Mezei MM, Theilmann J, Almqvist E, Hayden MR. The likelihood of being affected with huntington disease by a particular age, for a specific CAG size. *Am J Hum Genet.* 1997;60:1202-1210.

15. Langbehn DR, Brinkman RR, Falush D, Paulsen JS, Hayden MR, International Huntington's Disease Collaborative Group. A new model for prediction of the age of onset and penetrance for huntington's disease based on CAG length. *Clin Genet.* 2004;65:267-277.

16. Chen YW, Allen MD, Veprintsev DB, Lowe J, Bycroft M. The structure of the AXH domain of spinocerebellar ataxin-1. *J Biol Chem.* 2004;279:3758-3765.

17. Takano H, Gusella JF. The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-kB/Rel/dorsal family transcription factor. *BMC Neurosci.* 2002;3:15.

18. Masino L, Musi V, Menon RP, et al. Domain architecture of the polyglutamine protein ataxin-3: A globular domain followed by a flexible tail. *FEBS Lett.* 2003;549:21-25.

19. Lin CH, Hare BJ, Wagner G, Harrison SC, Maniatis T, Fraenkel E. A small domain of CBP/p300 binds diverse proteins: Solution structure and functional studies. *Mol Cell.* 2001;8:581-590.

20. de Chiara C, Menon RP, Adinolfi S, et al. The AXH domain adopts alternative folds the solution structure of HBP1 AXH. *Structure (Camb).* 2005;13:743-753.

21. Koide R, Ikeuchi T, Onodera O, et al. Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nat Genet.* 1994;6:9-13.

22. The Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell.* 1993;72:971-983.

23. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature.* 1991;352:77-79.

24. Banfi S, Servadio A, Chung MY, et al. Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat Genet.* 1994;7:513-520.

25. Pulst SM, Nechiporuk A, Nechiporuk T, et al. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat Genet.* 1996;14:269-276.

26. Kawaguchi Y, Okamoto T, Taniwaki M, et al. CAG expansions in a novel gene for machado-joseph disease at chromosome 14q32.1. *Nat Genet.* 1994;8:221-228.

27. Schols L, Vieira-Saecker AM, Schols S, Przuntek H, Epplen JT, Riess O. Trinucleotide expansion within the MJD1 gene presents clinically as spinocerebellar ataxia and occurs most frequently in german SCA patients. *Hum Mol Genet.* 1995;4:1001-1005.

28. Zhuchenko O, Bailey J, Bonnen P, et al. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat Genet.* 1997;15:62-69.

29. David G, Abbas N, Stevanin G, et al. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. *Nat Genet.* 1997;17:65-70.

30. Koide R, Kobayashi S, Shimohata T, et al. A neurological disease caused by an expanded CAG trinucleotide repeat in the TATA-binding protein gene: A new polyglutamine disease? *Hum Mol Genet.* 1999;8:2047-2053.

31. Zuhlke C, Hellenbroich Y, Dalski A, et al. Different types of repeat expansion in the TATA-binding protein gene are associated with a new form of inherited ataxia. *Eur J Hum Genet.* 2001;9:160-164.

32. Gilbert D. Biomolecular interaction network database. *Brief Bioinform.* 2005;6:194-198.

33. Schlitt T, Brazma A. Modelling gene networks at different organisational levels. *FEBS Lett.* 2005;579:1859-1866.

34. Deng X, Geng H, Ali H. EXAMINE: A computational approach to reconstructing gene regulatory networks. *BioSystems.* 2005.

35. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. Conservation and coevolution in the scale-free human gene coexpression network. *Mol Biol Evol.* 2004;21:2058-2070.

36. van Noort V, Snel B, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* 2004;5:280-284.

37. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. *Science*. 2003;302:249-255.

38. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. Prolinks: A database of protein functional linkages derived from coevolution. *Genome Biol*. 2004;5:R35.

39. Jansen R, Gerstein M. Analyzing protein function on a genomic scale: The importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol*. 2004;7:535-545.

40. von Mering C, Jensen LJ, Snel B, et al. STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*. 2005;33:D433-7.

41. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001;409:860-921.

42. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res*. 2005;33:D34-8.

43. Kanz C, Aldebert P, Althorpe N, et al. The EMBL nucleotide sequence database. *Nucleic Acids Res*. 2005;33:D29-33.

44. Tateno Y, Saitou N, Okubo K, Sugawara H, Gojobori T. DDBJ in collaboration with mass-sequencing teams on annotation. *Nucleic Acids Res*. 2005;33:D25-8.

45. Shows TB, Alper CA, Bootsma D, et al. International system for human gene nomenclature (1979) ISGN (1979). *Cytogenet Cell Genet.* 1979;25:96-116.

46. Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S. Genew: The human gene nomenclature database, 2004 updates. *Nucleic Acids Res.* 2004;32:D255-7.

47. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics.* 2005;21:248-256.

48. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Res.* 2005;33:D54-8.

49. Schuemie MJ, Weeber M, Schijvenaars BJ, et al. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics.* 2004;20:2597-2604.

50. Bairoch A, Apweiler R, Wu CH, et al. The universal protein resource (UniProt). *Nucleic Acids Res.* 2005;33:D154-9.

51. HUGO Gene Nomenclature Committee Homepage. Available at: http://www.gene.ucl.ac.uk/nomenclature.

52. Apweiler R. Sequence databases. In: Baxevanis AD, Ouellette BFF, eds. *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* Third ed. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2005:4.

53. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2005;33:D501-4.

54. Camon E, Magrane M, Barrell D, et al. The gene ontology annotation (GOA) database: Sharing knowledge in uniprot with gene ontology. *Nucleic Acids Res.* 2004;32:D262-6.

55. MEDLINE Fact Sheet. Available at: http://www.nlm.nih.gov.proxy.lib.sfu.ca/pubs/factsheets/medline.html.

56. Extensible Markup Language (XML). Available at: http://www.w3.org/XML.

57. Medical Subject Headings. Available at: http://www.nlm.nih.gov.proxy.lib.sfu.ca/mesh.

58. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 2005;33:D514-7.

59. Davidson I, Martianov I, Viville S. TBP, a universal transcription factor? *Med Sci (Paris).* 2004;20:575-579.

60. Alako BT, Veldhoven A, van Baal S, et al. CoPub mapper: Mining MEDLINE based on search term co-publication. *BMC Bioinformatics.* 2005;6:51.

61. Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: An overview. *J Comput Biol.* 2003;10:821-855.

62. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: Critical assessment of information extraction for biology. *BMC Bioinformatics*. 2005;6:S1.

63. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: Gene mention finding evaluation. *BMC Bioinformatics*. 2005;6:S2.

64. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: Normalized gene lists. *BMC Bioinformatics*. 2005;6:S11.

65. Colosimo M, Morgan A, Yeh A, Colombe J, Hirschman L. Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics*. 2005;6:S12.

66. String: functional protein association networks. Available at: http://string.embl.de.

67. The Mouse Genome Database. Available at: http://www.informatics.jax.org.

68. The FlyBase Database. Available at: http://flybase.org.

69. Crim J, McDonald R, Pereira F. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*. 2005;6:S13.

70. Hanisch D, Fundel K, Mevissen H, Zimmer R, Fluck J. ProMiner: Rule-based protein and gene entity recognition. *BMC Bioinformatics*. 2005;6:S14.

71. Fundel K, Guttler D, Zimmer R, Apostolakis J. A simple approach for protein name identification: Prospects and limits. *BMC Bioinformatics*. 2005;6:S15.

72. The Gene Ontology Website. Available at: http://www.geneontology.org.

73. Robinson MD, Grigull J, Mohammad N, Hughes TR. FunSpec: A web-based cluster interpreter for yeast. *BMC Bioinformatics*. 2002;3:35.

74. Berriz GF, King OD, Bryant B, Sander C, Roth FP. Characterizing gene sets with FuncAssociate. *Bioinformatics*. 2003;19:2502-2504.

75. Castillo-Davis CI, Hartl DL. GeneMerge--post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*. 2003;19:891-892.

76. Zeeberg BR, Feng W, Wang G, et al. GoMiner: A resource for biological interpretation of genomic and proteomic data. *Genome Biol*. 2003;4:R28.

77. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: A web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*. 2004;20:578-580.

78. Beissbarth T, Speed TP. GOstat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*. 2004;20:1464-1465.

79. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: Functional analysis of gene datasets based on gene ontology. *Genome Biol*. 2004;5:R101.

80. Masseroli M, Martucci D, Pinciroli F. GFINDer: Genome function INtegrated discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res*. 2004;32:W293-300.

81. Shah NH, Fedoroff NV. CLENCH: A program for calculating cluster ENriCHment using the gene ontology. *Bioinformatics*. 2004;20:1196-1197.

82. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree machine (GOTM): A web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. *BMC Bioinformatics*. 2004;5:16.

83. SGD Gene Ontology Term Finder. Available at: http://db.yeastgenome.org/cgi-bin/GO/goTermFinder.

84. eGON - explore Gene Ontology. Available at: http://nova2.idi.ntnu.no/egon.

85. Cheng J, Cline M, Martin J, et al. A knowledge-based clustering algorithm driven by gene ontology. *J Biopharm Stat*. 2004;14:687-700.

86. Joslyn CA, Mniszewski SM, Fulmer A, Heaton G. The gene ontology categorizer. *Bioinformatics*. 2004;20 Suppl 1:I169-I177.

87. Harris MA, Clark J, Ireland A, et al. The gene ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2004;32:D258-61.

88. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*. 2002;30:303-305.

89. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: A molecular INTeraction database. *FEBS Lett*. 2002;513:135-140.

90. Bader GD, Betel D, Hogue CW. BIND: The biomolecular interaction network database. *Nucleic Acids Res.* 2003;31:248-250.

91. Breitkreutz BJ, Stark C, Tyers M. The GRID: The general repository for interaction datasets. *Genome Biol.* 2003;4:R23.

92. Han K, Park B, Kim H, Hong J, Park J. HPID: The human protein interaction database. *Bioinformatics*. 2004;20:2466-2470.

93. Hermjakob H, Montecchi-Palazzi L, Lewington C, et al. IntAct: An open source molecular interaction database. *Nucleic Acids Res*. 2004;32:D452-5.

94. Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF. Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics*. 2005;6:34.

95. Huang TW, Tien AC, Huang WS, et al. POINT: A database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics.* 2004;20:3273-3276.

96. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet*. 2005;365:488-492.

97. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. *Genome Res*. 2004;14:1085-1094.

98. GO Evidence Code Guide. Available at: http://www.geneontology.org/GO.evidence.shtml.