THE IRIDESCENT SYSTEM: AN AUTOMATED DATA-MINING METHOD TO IDENTIFY, EVALUATE, AND ANALYZE SETS OF RELATIONSHIPS WITHIN TEXTUAL DATABASES

APPROVED BY SUPERVISORY COMMITTEE

_____
Dr. Harold R. Garner

_____
Dr. John Minna

_____
Dr. Ronald Butow

_____
Dr. Roger Schultz

Dedication

To Daniel, my father, who gave me love and understanding in my times of need. To Karen, my mother, who gave me my life and her love, and through her untimely death taught me that while we can plan for tomorrow, we must live as if there is not one. To Thanya, my wife, for her support and love. And finally, to Karen, my daughter, it is by your existence that the song of my life will be sung long after I am gone.

THE IRIDESCENT SYSTEM: AN AUTOMATED DATA-MINING METHOD TO IDENTIFY, EVALUATE, AND ANALYZE SETS OF RELATIONSHIPS WITHIN TEXTUAL DATABASES

By

Jonathan Daniel Wren, B.B.A, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas
Dallas, Texas
January, 2003

THE IRIDESCENT SYSTEM: AN AUTOMATED DATA-MINING METHOD TO IDENTIFY, EVALUATE, AND ANALYZE SETS OF RELATIONSHIPS WITHIN TEXTUAL DATABASES

Publication No. _____

Jonathan Daniel Wren, B.B.A, B.S., Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2003

Supervising Professor: Harold R. Garner, Ph.D.

Individuals are limited in their ability to read, remember and compare relationships within the vast amount of literature available in science and, indeed, most other fields. This is not only because the amount of literature is increasing exponentially, but the number of things being researched within is as well. Adding to the scale of analysis are new technologies that increase the rate by which data is being gathered from scientific experiments. For most areas of research interest, the scale of analysis exceeds an individual's ability to be aware of all the relationships contained within. Thus, an informatics approach is necessary to identify large-scale trends, shared relationships and novel relationships that are not contained within the literature, but are the logical consequence of relationships that are. A system has been designed to establish a network of relationships between "objects" of research interest (e.g. genes, chemical compounds, drugs, diseases and clinical phenotypes) by extracting information from scientific text in an automated manner. This system, called IRIDESCENT (Implicit Relationship IDEntification by in-Silico Construction of an Entity-based Network from Text), enables the discovery of novel relationships by identifying and scoring objects sharing large sets of relationships with an object of interest. IRIDESCENT also allows sets of objects to be analyzed for shared relationships, such as responding genes from a microarray experiment. Herein is described the development and workings of IRIDESCENT as well as several well-developed applications of the system.

The IRIDESCENT System: An Automated Data-mining Method to Identify, Evaluate and Analyze Sets of Relationships within Textual Databases

By

Jonathan Daniel Wren, B.B.A, B.S.

A dissertation in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Abstract

Individuals are limited in their ability to read, remember and compare relationships within the vast amount of scientific literature available. This is not only because the amount of literature is increasing exponentially, but the number of things being researched within is as well. Adding to the scale of analysis are new technologies that increase the rate by which data is being gathered from scientific experiments. For most areas of research interest, the scale of analysis exceeds an individual's ability to be aware of all the relationships contained within. Thus, an informatics approach is necessary to identify large-scale trends, shared relationships and novel relationships that are not contained within the literature, but are the logical consequence of the relationships that are. A system has been designed to establish a network of relationships between "objects" of research interest (e.g. genes, chemical compounds, drugs, diseases and clinical phenotypes) by extracting information from scientific text in an automated manner. This system, called IRIDESCENT (Implicit Relationship IDEntification by in-Silico Construction of an Entity-based Network from Text), enables the discovery of novel relationships by identifying and scoring objects sharing large sets of relationships with an object of interest. IRIDESCENT also allows sets of objects to be analyzed for shared relationships, such as responding genes from a microarray experiment. Herein is described the development and workings of IRIDESCENT as well as several well-developed applications of the system.

Table of Contents

Table of Contents (continued)

Prior Publications unrelated to this dissertation work

O'Brien KM, Wren JD, Dave VK, Bai D, Anderson RD, Rayner S, Evans G, Dabiri AE, Garner HR, "ASTRAL, a hyperspectral imaging DNA sequencer", *Review of Scientific Instruments*, p. 2141-6, Vol 69(5), May 1998

Fondon JW 3rd, Mele GM, Cummings D, Pande A, Wren JD, O'Brien KM, Kupfer KC, Lerman M, Minna JD and Garner HR, "Computationally Assisted Polymorphic Marker Identification: Identification and Verification of Multiple New 3p21.3 Polymorphic Markers", *Proceedings of the National Academy of Science,* 95:7514-7519, June 23, 1998

Garner HR, Wren JD, Minna JD, Fondon JW 3rd. "Polymorphic Repeats in Human Genes". US Pat No. 6,472,154. Filed Dec 31, 1999, Issued Oct 29, 2002.

Wren JD, Forgacs E, Fondon JW 3rd, Pertsemlidis A, Cheng S, Gallardo T, Williams RS, Shohet RV, Minna JD, and Garner HR "Repeat Polymorphisms Within Gene Regions: Phenotypic and Evolutionary Implications". *American Journal of Human Genetics*, August 2000, Vol. 67, p. 345-56

Forgacs E, Wren JD, Kamibayashi C, Kondo M, Xu XL, Markowitz S, Tomlinson GE, Muller CY, Gazdar AF, Garner HR and Minna JD "Searching for microsatellite mutations in coding regions in lung, breast, ovarian and colorectal cancers" *Oncogene*, 22 February 2001, Vol. 20, No. 8 pp.1005-1009

Wren JD, Kulkarni A, Joslin J, Butow R and Garner HR "Cross-hybridization on PCR spotted microarrays" *IEEE Engineering In Medicine and Biology* 2002 Mar-Apr; 21(2):71-5.

Wren JD, Mittleman D, Garner HR "SIGNAL – Sequence Information and GeNomic AnaLysis" *Computer Methods and Programs in Biomedicine* 2002 May; 68(2): 177-181

Kulkarni AV, Williams NS, Lian Y, Wren JD, Mittleman D, Pertsemlidis A, Garner HR "ARROGANT: An application to manipulate large gene collections". *Bioinformatics* 2002 Nov. 18(11): 1410-7

Chapter 1

Introduction

1.1 The amount of biological information is increasing exponentially

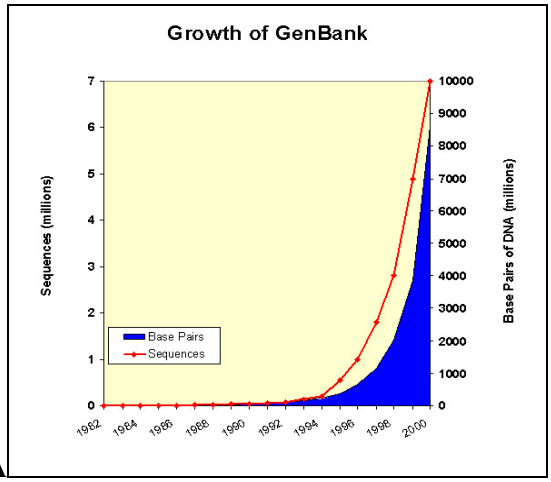
"We are drowning in information and starved for knowledge."
-John Naisbitt, *Megatrends*

There has been an explosive growth in the amount of biomedical information in the past decade (Figure 1), driven by increased technological capacity and fueled by resources devoted to research. The Human Genome Project has been declared completed[1], along with a number of model organisms and pathogens, and at the beginning of 2002 there were DNA sequences deposited for a total of 117,764 species[2]. There were 117,481 molecular structures established for 352,924 known chemical compounds[3], 13,700 human genes with a defined function and location[4], and 13,034 human diseases or potential diseases[5]. The central repository for biomedical publications, MEDLINE, contained approximately 12.7 million records at the time of this writing, and was estimated to be increasing at a rate of 500,000 records annually. While the coefficient of growth varies among these databases, they are all growing exponentially.

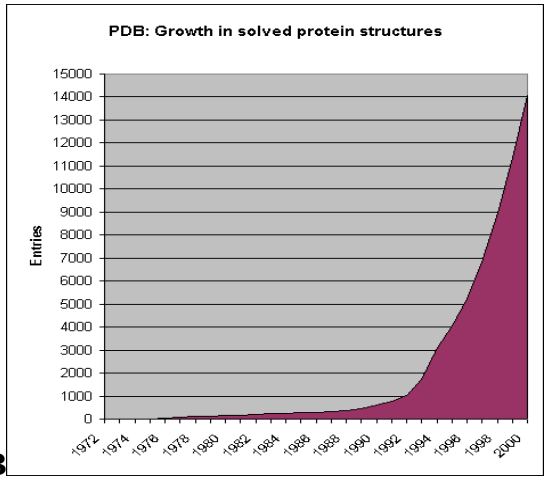
In most data-gathering efforts, there comes a point of diminishing returns – a point where gathering more data yields less and less understanding per unit gathered. And regardless of how much data are available, there is a difference between having data and understanding it. Data are usually gathered within a limited context, often to address a specific question about the relationship between two entities, and sometimes reported just as

observations. Despite any expertise a researcher may have, as individuals we are aware of only a very small portion of any of this. Since biological science is reductionist in nature, this does not limit our ability to discover new relationships between a relatively small set of entities, but we are limited in the scale by which we can analyze relationships.

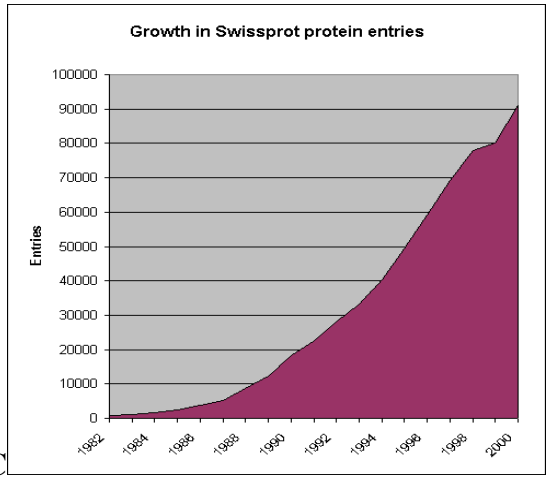
A



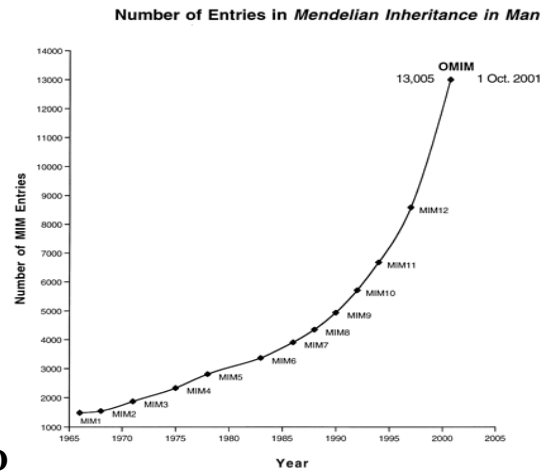
B

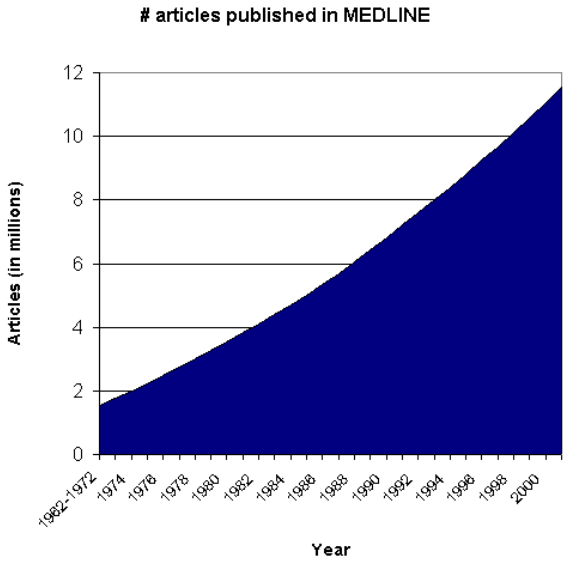


C



D





E

Figure 1: The number of entries in biomedical databases is increasing exponentially. Shown are the number of entries for: **a**) DNA sequences (Genbank[6]), **b**) Known proteins (Swissprot[7,8]), **c**) 3-D structural databases (PDB[9,10]), **d**) Inherited human diseases (Online Mendelian Inheritance in Man[5,11]), and **e**) Scientific publications (MEDLINE).

It is this exponentially increasing amount of data that creates the need (or opportunity) for greater information management tools – a need emphasized by the curators of these databases[12]. Despite the increases in technological capacity, we know that whether we measure the expression levels of one gene or all of them, the implications of expression level changes are not as straightforward as their measurements. Microarrays and other such developments in technology are just tools that give us more power in our quest for understanding, they do not in themselves bring understanding. Understanding comes from making connections between empirical observations and the implications of such observations on other previous observations and knowledge. The scientific value of a sequence, for example, does not lie as much in the sequence itself as it does in the meaning and implications of specific sequence patterns (e.g. genes, promoters, transposons). We

assume a new sequence inherits a limited amount of information (e.g. molecular function, biological role) from another, previously studied, sequence that it is highly similar to. Even the functions and actions of the ultimate protein products of the sequence data are rarely of importance in isolation, but rather by how they affect a biological system by their relationships to each other and how that system relates to other systems to impact the organism as a whole. We have increased our capacity to gather data, but can we also say that we have increased our capacity to use it?

1.2 Data, information and knowledge are distinct entities

Since the words "data" and "information" are frequently used interchangeably, as are "information" and "knowledge", it is useful to draw a distinction between these terms. Data are the most fundamental unit of the three terms, consisting of an empirical measurements or set of measurements. Datum is compiled to contribute to information, but it is fundamentally independent of it. Information, by contrast, is derived from *interests*. For example, data may be gathered on height, weight, race and diet for the purpose of finding variables correlated with risk of heart disease. But the same data could be used to develop a formula to create information about height/weight or race/diet correlations. Knowledge can be loosely defined as a set of information that gives sufficient understanding of a system to model cause and effect. To extend the previous example, information on race and diet could be used to develop a regional marketing strategy for food sales while information on height/weight ratios could be used as guidelines for physicians to recommend alterations in diet. There are no strict boundaries between the three terms, as datum can potentially be equivalent to

knowledge. For example, a high Geiger-counter reading coming from a ham sandwich is raw data, but it also gives information about the composition of the sandwich and the knowledge that eating it is unwise. In short: Data comes from examining, information comes from correlating, and knowledge comes from modeling.

This distinction between data, information and knowledge needs to be emphasized because science is primarily about increasing our collective knowledge, but the bulk of what is being generated with our high-throughput technologies (e.g. sequencers, microarrays, proteomic 2-D gels) is data. Data are gathered to gain information/knowledge about an item of interest, but may also contain useful information about other items not originally intended for study. Similarly, the bulk of publications within MEDLINE contain information regarding the analysis of data, but they are usually intended to address a specific question and cannot foresee all the implications of each relationship discovered within. MEDLINE contains data, information and knowledge, all of which are expressed as interactions (co-mentions) of various "objects". There are a number of anecdotes in science about discoveries inspired by accidents or sudden insights that arise from research in unrelated fields – providing a critical relationship necessary to unify a set of relationships. Information is derived from interests, and while most data are gathered in pursuit of a single interest, there is the possibility that it could contribute more information to other interests and enable the creation of more knowledge. It is this distinction that creates a significant proportion of the value in this proposal. Each datum gathered is understood in the context of a limited set of informational interests, and the potential applications of each bit of information gathered are understood in a limited context. Herein exists the potential for further discovery.

1.3 Bridging the gap between the growth of data and knowledge requires the ability to relate data

There is a need to understand the relevance and implications of new data in the light of as many previous observations as possible. Basic science, after all, is an endeavor undertaken to advance knowledge even though its immediate applications are not usually intended to address a direct human need (e.g. cure a disease). The real worth of data is their contribution to knowledge, which in turn contributes to the power to choose action(s) from almost infinite choices. The ability to understand observations in multiple contexts can lead to new insights and discovery, and thus, automated methods of discovery and understanding become more valuable.

Scientists are excellent at finding patterns and elucidating relationships within data, but are limited in the amount and rate by which they can assimilate it. Computers, conversely, are limited in their ability to find patterns or understand relationships but are faster and comprehensive in assimilating data. We are not yet able to endow humans with the ability to assimilate more data, so if we are to attempt to comprehensively search existing data for patterns, it will be necessary to use a computer. The problems addressed by this project are the following: First, how can a domain of knowledge be obtained in electronically readable format? Second, how can we enable software recognition of data contained within this domain? Third, how can we identify valuable informational relationships between datum contained therein? And finally, how can we use these relationships to identify broad trends in relationships and present them in such a way to enable the discovery of new knowledge?

This report details the development and testing of a software package entitled IRIDESCENT (Implicit Relationship IDEntification by in-Silico Construction of an Entity-based Network from Text) in an attempt to address these problems. IRIDESCENT uses MEDLINE to represent a domain of knowledge, database entries (referred to as "objects") to recognize data within text, and co-citation of database entries within the same MEDLINE record to exhaustively identify potential informational relationships between these objects. IRIDESCENT is a general text object correlation tool, and here we focus on its application, validation and demonstration within the biomedical domain by using a biomedical object set in concert with MEDLINE. Identified relationships are stored within a database and used to create a comprehensive network of relationships for analysis. Groupings of relationships identified within this network are then ranked against a random network model to ascertain a quantitative measure of how exceptional any particular grouping is. The system is developed to address several different biologically relevant questions, and tested on each. Advantages, limitations and implementation of each of these will be discussed in detail.

Chapter 2

Background and Previous Work

2.1 Extracting Informational Relationships by Text-Mining

There has recently been an increased focus on the development of tools for processing free-form textual information, such as is found in literature databases such as MEDLINE[13-16]. The need to apply greater organizational and filtering power to the abundance of recorded information is becoming more salient as not only the amount of information increases, but also as the number of sources for it increase (e.g. journals, web pages) along with their heterogeneous formats (e.g. text, HTML, PDF, etc.). This ability to greater utilize the biomedical literature in aiding scientific discovery is a topic of increasing interest[17].

Programs have been developed to identify biological entities such as genes[18-20], proteins[21], chemicals[22], drugs and cell lines[19] within free-form textual input, and even the associative terminology used with them[23]. Doing so enables the automated construction of a reference work or topical database, otherwise known as a knowledge base[24]. Importantly, it enables relationships between these objects to be identified as well as the nature of the relationship. For example, programs have been designed to identify and catalog binding interactions between proteins[25-31] and their associated molecular compounds[32] as well as construct a network of potentially interacting genes[33-35].

These programs are intended to identify associations between a relatively homogeneous set of entities, whereas IRIDESCENT attempts to assimilate a more heterogeneous set of terms from established databases rather than attempting to recognize them de-novo. This heterogeneous set of terms, insofar as they represent relationships deemed to be of biological interest, will be useful in identifying shared and implicit relationships for the purpose of knowledge discovery.

2.2 Arrowsmith: New information from old data

In 1986, when MEDLINE had less than half the number of entries it does today in 2002, a researcher named Don Swanson first demonstrated that two biological phenomena without a known link could be related through an intermediate link in an semi-automated way[36]. The concept is relatively straightforward, as illustrated in Figure 2. Here we see that while relationships between A and B have been studied as well as a relationships between B and C, for whatever reason, no relationship has been identified between A and C. Swanson termed these relationships "Non-interactive literatures". Swanson thus developed a method of pairing keywords from the titles of MEDLINE records to identify commonalities between two sets of literature. Using this method, he identified a relationship between Raynaud's Disease (A) and fish oil (C) by the associated blood and vascular changes related to both phenomena (B)[36]. Raynaud's Disease is a circulatory disorder in which blood flow is reduced in the extremities. Fish oil, conversely, increases a number of circulatory variables

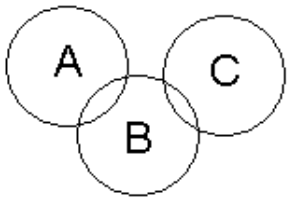


Figure 2: Sets A and C have something in common which is not obvious from examining either one independently.

decreased in Raynauds, enabling Swanson to hypothesize that fish oil might have a positive effect on Raynaud's patients. Swanson was eventually shown to be correct[37]. This method was then applied to identify other relationships, such as between magnesium levels and migraine headaches[38] as well as connections between arginine intake and blood levels of somatomedins[39].

Swanson had some initial success with his method and eventually published his program, Arrowsmith[40,41], to accomplish the search for "non-interactive" literatures, which is available today on the Internet[42]. A conceptual diagram of how Arrowsmith works is shown in Figure 3. Figure 3a shows the process of a directed search between two concepts, A and C. In this diagram, A (represented by a circle) is a general concept of interest in the form of keywords or phrases to be used in a topical search of MEDLINE. The titles obtained from the search are then parsed into a set of individual words. From this set, words that are presumably uninformative are filtered out, leaving a set of keywords (represented by the rectangular boxes underneath). C consists of a different topical search and is, presumably, one not known to overlap with A. That is, if one searched MEDLINE for the combined set "A and C", one should find nothing (or at least no entries that suggest a relationship). Arrowsmith identifies a set of keywords found in **both** A and C, represented by B. It is in this set that undocumented connections may be found. It is left to the judgment of the researcher whether or not the connections in B are relevant.

Figure 3a represents a directed search, the type of search one would be interested in if one were hypothesizing a connection between A and C, yet could find no available literature. Figure 3b represents an undirected search, the approach one might take if one was interested

in simply finding any new or interesting connections related to A. From an initial set of keywords derived from a topical search, A, one would conduct another independent search on this entire set of keywords. The results could be combined into another set of keywords, B and again, from each of these keywords, another search is conducted. This third list of references, obtained from a search on all of the keywords in B, can be processed to **exclude** references already found in the initial set, A. We are then left with a final set, C.

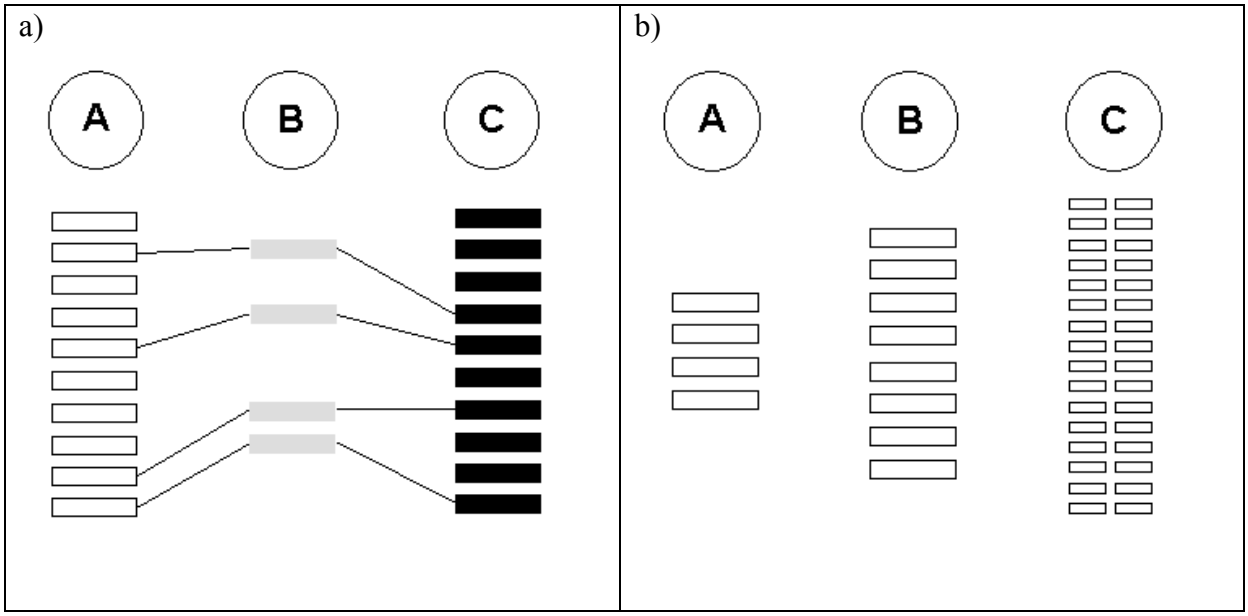


Figure 3: Swanson's approach to searching for related but non-interactive literatures. **a)** When two concepts (A and C) are hypothesized to be related, but there is little or no supporting literature, they may be related through an intermediate, B. **b)** When attempting to discover new connections for a concept, A, one can expand the search to all related items, B, and then conduct another search for a set of items, C, which are not found when searching A.

There are a number of reasons why Swanson's method is highly inefficient. First, Arrowsmith only uses the titles of articles. While it would be a trivial matter to extend the same analysis to abstracts as well, this restriction serves a practical purpose in reducing the number of keywords a user has to analyze into a more concentrated set. However, the titles

do not always describe the discovery in specific terms, nor do they include much of the relevant information found in the abstract. Second, only key words (rather than phrases) are used, leaving no distinction, for example, between articles about "cardiac arrest" and "cardiac development". It has been proposed that this shortcoming could be overcome by using the Unified Medical Language System's Metathesaurus concepts in place of keywords[43], but no demonstration of efficacy was given except to show Swanson's discovery could be replicated. Third, while the method is termed "automated" it is actually semi-automated because it requires a manual compilation of records as input, and manual expert evaluation of each matching keyword for relevance. While this expert evaluation is to some degree unavoidable, it is because of the data explosion that this requirement makes the method less and less efficient. One group, however, has used a normalized statistical frequency of keyword and keyphrase occurrences in an attempt to buoy the most relevant words and phrases to the top of the list[44]. The disadvantage of a keyword-based approach, aside of allowing only a limited context, is in the size of the domain analyzed. Even after stop words are screened out, the number of unique keywords grows rapidly, as illustrated conceptually in Figure 3b. This helps to illustrate that, for all practical purposes, one must begin by hypothesizing a relationship between A and C, because the search space grows very quickly when dealing with an undirected search. Finally, no method of scoring the results is provided, leaving the user without any way of estimating how relevant a shared word might be.

The scale of analysis is the limiting factor for any method using word-pairing or co-occurrence of terms. At the end of 2002, MEDLINE contained 12,725,686 records,

6,980,030 of which had abstracts. When these 12 million records were parsed, they contained over 4,500,000 unique words. To show how quickly unique words from a set of abstracts related to a common topic would grow, titles and abstracts from 973 MEDLINE records were obtained from a topical search on the keyword "wnt" and processed into individual words using the word parsing routine of IRIDESCENT. A total of 11,226 unique words were found within a total of 191,165 words. Merging only the simple root variants of these words (e.g. counting "bind", "binds" and "binding" as one word) trimmed the list down to 9,479 words. A filter was then applied to exclude 220 uninformative words (e.g. "hence", "where", "did", "at") and probable adverbs (words ending in "ly"). The final list contained 8,495 keywords. A number of these were words with more complex word root variants (e.g. bind/bound, cell/cellular), proper nouns (e.g. "Beckman", "Smith"), numbers or percentages, a few uninformative words that weren't screened (e.g. "hundred", "liter"), a large number of words whose usefulness in conducting another search was probably low (e.g. "agarose", "filter") and a large number of words whose usefulness was uncertain because they represent extremely broad concepts (e.g. "cell", "development", "Drosophila"). By querying MEDLINE abstracts cumulatively using the most frequent keywords on this list with PubMed (i.e. 1 word, then 2, then 3, up to 50), and calculating the asymptote, we estimate that a total of 6,100,000 MEDLINE articles contained one or more of the keywords from the Wnt list in its abstract, which represents almost 97% of the total MEDLINE records that contain an abstract. Therefore, examining this domain of implicitly related articles for potential relationships would be tantamount to reading most of MEDLINE anyway.

Since papers describing the wnt signaling pathway use many of the same terms, we also plotted the growth rate of keywords in general, as records are examined randomly. The total growth in unique keywords from the wnt abstracts is plotted against the same number of effectively random abstracts (obtained from MEDLINE using the keyword “result”). All the words in the abstracts were recorded into a database, adding to the cumulative total every time a new word was found. As Figure 4 shows, a relatively small set of 100 abstracts quickly balloons to 4,000 unique words found.

What is evident from the wnt keyword growth analysis is that an undirected search on anything but a small starting domain quickly becomes impractical from the standpoint of human time and effort. This suggests that some method must be found to reduce all these irrelevant keywords from analysis. This is especially true if the utility of the end results is to be judged by a human.

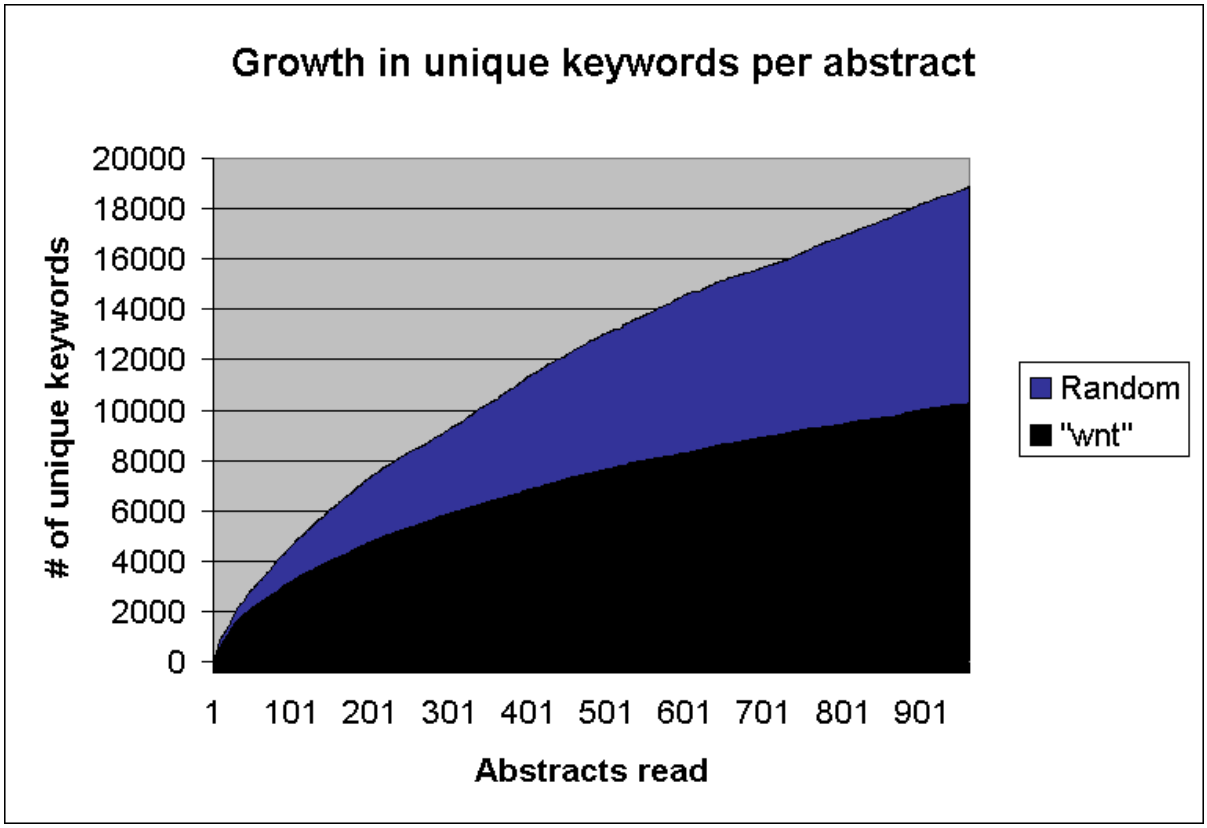


Figure 4: As the number of abstracts analyzed increases, so does the total number of unique keywords. Within a "topical" domain, the growth is not as rapid as it is within a domain where the abstracts are chosen randomly.

While the approach may be inefficient, Swanson made a very valuable point: Relating new discoveries to old ones in the literature is not perfectly efficient, even in a well-studied system. Given that Swanson demonstrated there are undiscovered connections within existing literature, at least three questions naturally arise from this: How many more undiscovered relationships are there, how valuable are they and how can they best be found? The first two questions we will probably not be able to address here, but the last question represents the focus of IRIDESCENT.

2.3 Data-mining and literature-based knowledge discovery

A very practical way to evaluate any literature-based analysis is based upon three factors: How comprehensive is it, what is the rate of error (false-positives and false-negatives), and how much work must a human be required to do to identify interesting relationships? Given the very real limitations of time, attention and concern that human experts face in evaluating not only the validity of a relationship, but its potential utility, it is practical to restrict analysis to entities of direct research interest. IRIDESCENT in its current implementation, restricts analysis to things we know to be of concern to the biomedical community: genes, diseases, clinical phenotypes and small molecules such as drugs and chemical compounds. Analysis is restricted to titles and abstracts for two reasons: First, and most practical, these represent the vast majority of electronically available information. Second, these two portions of MEDLINE records are considered to represent, in summary form, the discoveries of the article.

There are a number of difficulties inherent in processing text from scientific abstracts. Often, there are manipulation experiments contained within the article where reference to a protein and its interaction is solely in the context of artificial conditions set up by the experimenters. For example, when a gene knockout animal strain is constructed and the effects of a drug on the strain are discussed to help elucidate the interaction between drug and gene – extracting information from a sentence like “Drug ABC was shown to be lethal” might be misleading. While such contextual information is usually apparent to humans, it will be difficult for a computer to accommodate for this because it will require keeping an ongoing record of the conditional circumstances that apply to each sentence. If an object

happens to fall in this category of special circumstances, the documented relationship should have a proportionately small counter when compared to the sum of the occurrences of the object.

Another problem involves the use of non-standard notation to describe artificial constructs and it will be difficult to get a computer to recognize the meaning inherent in some descriptions. For example, take the statement "The ABCΔ130-140 protein was unable to bind DEF". Biomedical researchers easily understand two things from this statement: ABC normally binds DEF (implied) and without amino acids 130-140 it is unable to. Such notation could easily be accommodated if it was standard, but the researcher might have defined the 130-140 deletion as ABCΔ1d (for 1st domain), ΔABC-2 (for 2nd deletion construct), ABC-DEFBR (ABC without DEF Binding Region) or any number of ways related to what is being studied. Even if the primary object could be resolved, it becomes slightly problematic to assign relationships at that point since it is not the object itself (e.g. ABC) that is being referred to but a modified version constructed for a specific experimental purpose. Nonetheless, these notations are relatively uncommon within abstracts, and IRIDESCENT will only attempt to identify objects cataloged in its object database.

In addition to the problems mentioned above, there are at least two more types of false-positive errors possible: The system incorrectly identifies an object/relationship or the conclusions/results of the research are in error. The latter case should be relatively rare and since we are not yet equipped with the capacity for computationally-based rational thought, such instances will have to be written off as possible systematic false-positive errors. The former case of computational errors is much more likely to occur, and the rate by which it

occurs will have to be identified by manual evaluation of accuracy. This can be done by taking subsets of the entries in the Object-Relationship Database (ORD), going back to the original reference and evaluating how many are accurate. Once an estimate of error is obtained, we can assign "fuzzy" relationships – that is, the system will assign a confidence score ranging from 0 to 1 that represents a numerical estimation of the probability the relationship is non-trivial. It is more difficult to ascertain whether or not a relationship is "real" as it is to estimate if it is non-trivial in nature. This evaluation of accuracy will be critical to providing scores to rank potentially undocumented relationships. One of the early goals in refining IRIDESCENT was to reduce the systematic errors in building the ORD. The other type of error that might occur from rare or poor semantic phrasing will be more difficult to deal with. While processing abstracts, IRIDESCENT will emphasize accuracy over thoroughness (precision over recall), which is to say that we are willing to miss identifying relationships mentioned infrequently within the literature in favor of being confident that the relationships identified are correct.

Many of these problems could be resolved by providing consistent and standard classification to objects of study. In fact, some of these issues of classification and standardization have been recently tackled by the Genome Ontology (GO) project at Stanford[45,46], whose goal is "to produce a dynamic controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing". Projects such as this help underline the importance of defining/elucidating the fundamental units of cells and their processes for informational purposes, and we will discuss the GO effort in more detail later. While an individual

researcher may not find it important to have a standardized description of the protein he is studying because he has a thesaurus already in his head, someone attempting to build a database to help elucidate relationships will find it very valuable. As it is, the National Library of Medicine uses a tool for their Metathesaurus called MetaMap, which is a way of attempting to match phrases and word variants with concepts contained within the Metathesaurus[47-49]. This helps users select a variety of topical areas once they input their general interests in a "freehanded" manner.

Chapter 3

Experimental Approach

In brief, the goal of IRIDESCENT is to use the scientific literature as a source by which relationships can be identified and evaluated as a *set*. The ability to make statements about sets of relationships enables us to justify the existence of a new relationship by virtue of shared relationships, and to identify commonalities. To do this, informational relationships need to be identified in a manner comprehensive enough such that we can begin to make statements about what is known and not known, within an established margin of error. Furthermore, when a set of relationships can be identified, there must be a means of evaluating its relevance as a set. The experimental approach is thus divided into several parts, each necessary to accomplish the overall goal. Each of these parts is discussed and evaluated in turn.

First, a knowledge domain must be defined. That is, we are attempting to identify informational relationships to be able to make a statement about the current state of knowledge, which necessarily includes historical archives. In this case, MEDLINE is assumed to represent a source of knowledge about the biomedical domain since it is the central repository for titles and abstracts from thousands of biomedical journals. Second, to engage in the discovery of new relationships, informational relationships from within a domain of knowledge must be assimilated. Recognition of meaningful relationships within MEDLINE is based upon the assumption that the primary subjects of biomedical research are categorized in a general manner (e.g. genes, diseases, phenotypes, chemical compounds are

used to study, related and understand specific phenomena) and these subjects are of sufficient importance to be contained within specific databases. Third, IRIDESCENT attempts to comprehensively identify informational relationships within MEDLINE through the co-occurrence of objects within these titles and abstracts. Fourth, a comprehensive network of relationships is stored in a database and then used to create queries that involve shared relationships and those that are only known implicitly. Fifth, these shared and implicit relationships are evaluated statistically using bounded network models. And finally, the system is tested by application to existing problems in biomedical research.

3.1 General Computational Approach and Design

Development was initially conducted on a Desktop 800 MHz Pentium III (named "GESTALT") with 256 MB RDRAM and 36 gigabyte (GB) SCSI Hard Drive. In early 2002, development was switched to a Pentium-4 Personal Computer named "IRIDESCENT" with 1 GB RDRAM, an ultra-fast 36 GB SCSI drive and backup 72 GB SCSI drive. MEDLINE was stored locally on the 72 GB drive, taking up a total of 42.7 GB of drive space. IRIDESCENT is written in Visual Basic 6.0 (VB6) supplemented with Service Pack 4. It uses Open Database Connectivity (ODBC) extensions to enable database access from Microsoft Access 2000. This Access database is used to store the Object Recognition Database (ORD - see appendix). VB6 also accommodates Structured Query Language (SQL) server extensions via ODBC, which allows for an upgrade, which will become necessary as the database grows in size. Access can handle databases up to 1 GB in size. The latest version of the ORD database contains 302,549 recognizable entries (107,451 unique objects)

and stores 7,589,042 total relationships recognized within MEDLINE, all requiring a total of 390 MB of space. Thus, upgrading will not be required in the short term, but will become a future concern.

Although the system has been implemented on a desktop computer, processing time is relatively rapid. Given the size of the object database and the MEDLINE flat files above, all of MEDLINE was processed in 22 days by the system. A year's worth of updates (~500,000 records) could be processed within less than 2 days at this speed. While a greater speed is always beneficial, this amount of time required for a database build does not significantly limit the system at this point in time. As the total number of objects recognized increases, the total time it takes to process MEDLINE increases at a much greater rate than simply increasing the number of records processed. Fortunately, IRIDESCENT consists of a number of processes that could, in theory, run in parallel. Scanning each abstract, for example, can be considered a separate recording event. On a 64 CPU machine, 64 abstracts could theoretically be sent at one time for processing. Separating the database and source (input) files on different drives allows faster processing time since the read-head does not need to constantly reposition itself while first reading from the source and then updating the database.

Figure 5 illustrates a flowchart of the general system logic, which corresponds to each of the major tasks undertaken by IRIDESCENT. Each of the steps in this overall chart are expanded in later sections by number. In Step 1 (expanded in Figure 6), all object classes to be recognized are assimilated into one central database. This entails reformatting and error-checking for each of the entries as read from each database. In Step 2 (expanded in Figure

10), objects from the different databases are checked against one another for entries that belong together (e.g. a disease named after a gene) and those that should be separate. Acronyms are flagged if they refer to more than one gene as being potentially ambiguous, and lexical variants are identified for each of the object entries. In Step 3 (expanded in Figure 7), MEDLINE records are input sequentially and searched for any objects that are mentioned within. As objects are co-mentioned, they are evaluated and put into the relationship database (Step 4). At this point, processing of MEDLINE is finished and the object-relationship database (ORD) has been constructed for analysis and contains a network of biomedical relationships. A this point, the database can be analyzed to find sets of shared relationships between two or more objects. This enables common relationships to be identified such that a researcher, sufficiently versed in biomedicine, can examine the shared relationships identified by the system to ascertain the nature of the overall relationship between two or more objects. This database is then used in the analysis steps (A.1, A.2 and A.3).

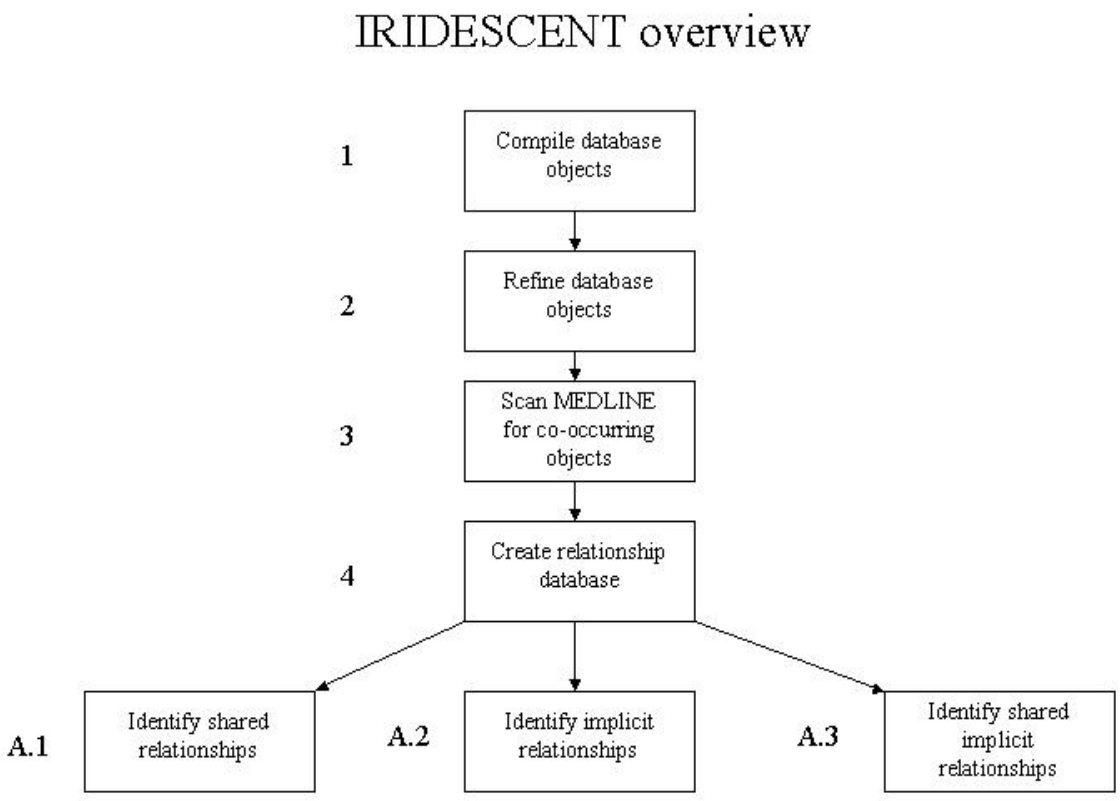


Figure 5: Objects of interest are assimilated into a central database (1) and processed such that they can be recognized within text (2). MEDLINE records are then sequentially input and searched for all objects found within their titles and abstracts (3). A database is created from these co-occurring objects (4) and is used to create a relationship network that enables a user to identify relationships shared among a set of objects (A.1), objects that are implicitly related to a central query object by virtue of shared relationships (A.2) and objects implicitly related to a set of objects by virtue of the important groupings within the set (A.3).

3.2 Object-based analysis: Defining what is interesting within the literature

Most biomedical knowledge is summarized in MEDLINE, which is freely available as electronic text in XML (eXtended Markup Language) format from the National Library of Medicine (NLM). Databases are considered repositories for raw data, even if various

informational facets can be found within individual fields. The scientific literature is considered the central repository of information and knowledge, allowing for extreme diversity in descriptions and methods. The drawback is that while databases are highly amenable to computerized analysis, information in the scientific literature is not. It is diverse in format, complex in structure and has no well-defined standards. Searching databases for a gene names or keywords is relatively straightforward task, but searching the literature for specific items of interest can be an arduous task.

Databases are rich and concentrated sources of data and information, and because items of interest such as gene names are more easily and accurately extracted from databases, they provide an excellent source for term recognition. Routines in IRIDESCENT have been written to process a number of diverse textual formats in order to populate the ORD with biological objects. Gene entries were obtained from the Genome Data Base (GDB) and the Human Genome Nomenclature Committee (HGNC), which has developed the accepted standard that the National Center for Biotechnology Information (NCBI) uses for gene nomenclature, and LocusLink, which is curated by the NCBI. The first database version of IRIDESCENT had a total of 35,579 gene name entries (13,104 unique gene names) from combining these three lists. Online Mendelian Inheritance in Man (OMIM) entries on inherited disorders (and potential disorders) numbered a total of 28,733 entries that included 11,464 unique disease/phenotypes. A total of 7,713 subheadings from the Medical Subject Heading (MeSH) index were incorporated and categorized as Small Molecules (drugs, metabolites, chemicals, elements) if they were in the “D” main category. If the entry was under the MeSH “C” category, the entry was categorized as a disease/phenotype. In the

second IRIDESCENT database build, 148,281 chemical compound entries were added from ChemID (37,855 unique names) and 10,138 drug entries (2,032 unique names) were obtained from the Food and Drug Administration (FDA). The first database build was constructed in January 2001 and the second build in July 2002. The Internet locations of the downloaded files are given in Table 1.

Entries in these databases require formatting since they are to be used for text matching rather than categorization. Entries such as “Cassette, ATP-Binding” are more likely to be written as “ATP-Binding Cassette” in abstracts. Similarly, parenthetical comments that are useful to people perusing the database such as “Color Blindness (x-linked) Syndrome” are not likely to be matched against textual input. It has been necessary to address a number of such formatting issues, but only a few of the more broadly applicable ones will be mentioned. A process flowchart of database creation and formatting is given in Figure 6.

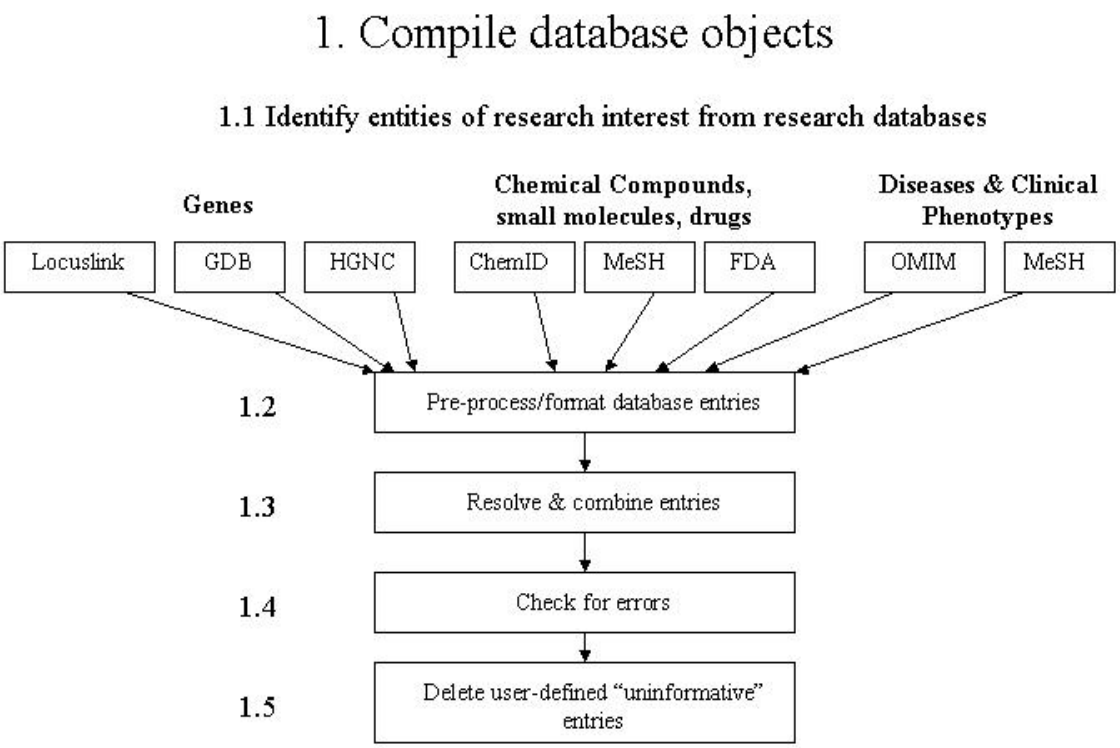


Figure 6: Entries from various databases considered to represent areas of research interest are combined into one central database. Each entry is processed such that it is likely to be recognized within the text it occurs in. Errors and uninformative entries are deleted.

As shown earlier with reference to Swanson's work, a keyword-based approach is currently impossible to implement (at least on any system available here at UTSW) since there are over 4.2 million unique words within MEDLINE. But mainly because this approach will lead to excessively high false positive and false negative correlations, greatly diminishing the desired utility and efficacy of a knowledge discovery system. It would be necessary to identify phrases as well (e.g. "cardiac" as a keyword is far less informative than "cardiac development" or "cardiac arrest"), and it is not apparent how such phrases would be

identified *de novo*, although approaches such as grammar induction[50] may be fruitful in the future (grammar induction is an attempt at discovering common structures within text, structures which arise as a consequence in the construction of meaning). The number of stored relationships would scale exponentially with the number of words, which would require an enormous database. It would furthermore be exceedingly slow in the storing and analysis of relationships, with the bulk of computational power being devoted to uninteresting terms such as “the” and “what”. Even with a stopword list, other uninformative words would still predominate such as “survey” and “liter”. Therefore, centering the analysis on pre-defined objects enables a focus on relationships with a high probability of being informative to any researcher involved in the study of the general class it was derived from (i.e. genes, diseases, phenotypes, chemicals). Other object types such as tissue types, protein motifs or species names can be added as desired depending upon research interest. The only drawback to the object-centered approach is that certain informative relationships (e.g. a common promoter sequence) might be missed.

Besides the sources used to construct the ORD, Table 1 contains a compendium of additional online text-based sources that can be used to provide supplemental data such as additional synonyms or additional object types. It is not necessary for the success of this project to assimilate as many objects as possible, but rather it is important to have a reliable set of objects representing very broad and popular areas of research.

Name	Location	Data
Human Gene Nomenclature Committee (HGNC)	http://www.gene.ucl.ac.uk/nomenclature/	Official (HUGO) gene names
Genome Database (GDB)[51]	http://gdbwww.gdb.org/gdb/advancedSearch.html	Gene names & synonyms; diseases; cytolocs;
Online Mendelian Inheritance in Man (OMIM)[52]	ftp://ncbi.nlm.nih.gov/repository/OMIM/	Human diseases & phenotypes
Medical Subject Headings(MeSH)	http://www.nlm.nih.gov/mesh/filelist.html	Diseases, phenotypes, chemicals, drugs, tissues, pathogens
Center for Disease Control (CDC)	ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD9-CM/2002/	Pathogenic diseases & drugs
Kyoto Encyclopedia of Genes and Genomes (KEGG)[53]	http://www.genome.ad.jp/kegg/	Pathways, genes, orthologs, functions, enzymes and ligands
MEDLINE Plus	http://www.nlm.nih.gov/medlineplus/druginformation.html	Drug names & synonyms, phenotypes (side effects)
Locuslink[54,55]	http://www.ncbi.nlm.nih.gov/LocusLink/	Gene names, aliases, OMIM links, cytolocs, homology
Enzyme and co-factor database[56]	ftp://ftp.expasy.ch/databases/enzyme	Enzymes, co-factors, diseases, metabolite associations
The U. Minn. Biocatalysis /Biodegradation Database[57]	http://www.labmed.umn.edu/umbbd/index.html	Pathways, enzymes (&EC), metabolic compounds
Swiss-Prot[7]	ftp://expasy.cbr.nrc.ca/databases/swiss-prot/	Gene names, protein families & members; DB xrefs
FlyBase[58]	http://flybase.bio.indiana.edu/ (for inputting data on Drosophila genes homologous to Human ones)	Drosophila homologs: their cellular locations & functions
Mouse Genome Database[59]	http://www.informatics.jax.org/	Mouse homologs & human gene names, GO classifications
Genome Ontology Project[46]	http://www.geneontology.org/	Biological processes, Molecular functions & cellular components.

Unified Medical Language System (UMLS)[60]	http://www.nlm.nih.gov/research/umls/	Acronyms, drug names, medical vocabulary, biological objects
Structural Classification of Proteins (SCOP)[61]	http://scop.mrc-lmb.cam.ac.uk/scop/	Protein Structural Classifications: Folds, families, superfamilies
Alliance For Cellular Signalling (AFCS)	http://afcs.swmed.edu/	G-protein coupled receptor database
Food and Drug Administration (FDA)	http://www.fda.gov/cder/ndc/listings.txt http://www.fda.gov/cder/ndc/formulat.txt	Approved drugs (brand names + chemical names)

Table 1: Some online databases and data sources amenable to either direct or query-based text mining. Among these will be a varying degree of entry overlap (e.g. Locuslink, GDB, and HGNC all have human gene names), but can be useful for identifying synonyms and lexical variants.

3.3 Using co-occurring terms to exhaustively identify potential relationships

Natural Language Processing (NLP) engines have the potential to identify relationships between objects much more accurately than co-mentions because they attempt to structure the relatedness of words in a sentence[62]. However, they also have a higher false-negative (FN) rate due to their limitations on resolving distant references. For example, it is quite common for biomedical abstracts to begin the first sentence defining the objects being studied, describe the experimental setup and then summarize results. Within the result summary, references are sometimes made in terms of acronyms defined earlier or simply "control" versus "experimental" groups. Current NLP methods are not without false-negatives, and have difficulty understanding distant references within a body of text. Under certain circumstances, they can also fail to understand that while no explicit relationship for two objects may be stated within an abstract, it is implicitly understood from reading the abstract that the first is mentioned only because it is relevant to the study of the second. NLP

methods are additionally problematic to implement within the context of another software system, since they are frequently stand-alone executable systems. Thus it would be difficult to obtain a software package capable of NLP that could be readily integrated into the system we are designing. Finally, NLP methods devote a significant amount of CPU time to parsing relationships between each word in a sentence, when all we are really interested in is the relationships between objects. For example, NLP methods are able to more accurately identify relationships in the sense that the subject and object of a sentence can be recognized by their relative positions within the sentence and intervening words that provide context, allowing for a greater confidence in the existence of a relationship. They will even potentially recognize the existence of a negative or speculative relationship, and lower the false-positive rate. But to do so, all words within the sentence must be mapped to their parts of speech, resolving ambiguity in word usage, and diagrammed to determine the nature of relationships between all words within the sentence. These operations are compute-intensive and while such methods could be potentially useful, for the task we are attempting here they will provide little additional information at a much higher cost in processing time and potential recognition of object relationships.

We attempt to identify as many relationships as possible by postulating that a potential relationship exists between two objects when they are observed to co-occur within the same MEDLINE record, an approach also taken by others[33,34,63]. Co-occurrences are calculated both within abstracts as well as sentences, with the hypothesis that two objects mentioned in the same sentence are more likely to represent a non-trivial relationship. This hypothesis will be tested in a later section and although this approach increases the recall

(defined as the number of true positives identified divided by the total number of true positives), it will reduce the precision (defined as the number of true positives divided by the total number of predicted positives). Because of this reduced precision, we will rely upon an approach based upon fuzzy logic. Fuzzy logic is named as such because the existence of a relationship is not represented in a binary manner (e.g. on or off), but rather as a measurement within a spectrum of values[64] (Figure 7).

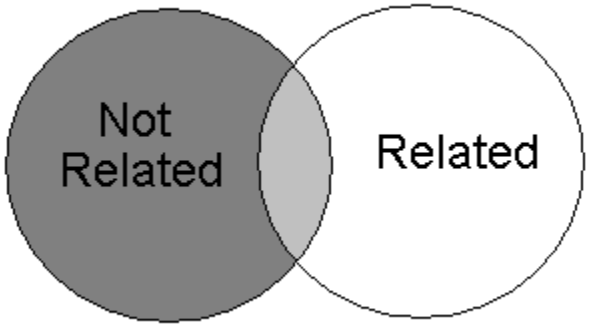


Figure 7: Fuzzy set theory represents membership in a category as a spectrum of values. Membership between the two sets "Related" and "Not related" is looked at from the standpoint of how far into one set a relationship is. An object could, for example, be 73% related to another and thus 27% not related. Representing a relationship in this manner enables us to assign "confidence" scores.

A random set of 25 MEDLINE records (titles and abstracts) was chosen and objects co-occurring within each of them were manually evaluated to establish whether they shared a non-trivial relationship with one another. It was determined that two objects co-mentioned within the same sentence were more likely (83%) to be related to one another in a non-trivial manner than objects co-mentioned in the same abstract (58%). Sentence co-mentions, however, have a relatively high rate of false-negatives, missing 43% of the non-trivial

relationships within an abstract. This proportion of correct relationships among abstract co-mentions is similar to the estimates others have obtained[34,65].

Two types of false positive (FP) errors were observed, one of which appeared to be relatively random in nature while the other was more systematic. Random FP errors would occur, for example, when an object within an abstract was part of an assay and not the study (e.g. sodium, EDTA), when a relationship was declared not to exist (e.g. "We found *no* relationship between A and B"), or when speculative/extraneous information was included in the abstract (e.g. "We hypothesize a possible role in…"). The more co-mentions observed between two objects, the less important this random source of error is, since even if the number of relationships was inaccurate, the existence of a relationship was true. For statements declaring no relationship such as "We found no relationship between A and B", we do not anticipate this will be more than a minor problem for two reasons. First, for better or worse, reporting of negative results is infrequently given in journals. Second, when negative results are reported, it is usually only noteworthy once. We would not anticipate the continued reporting of negative results regarding a specific relationship in future reports.

Figure 8 provides an overview of the process of populating the ORD with specific relationships identified within MEDLINE. The process begins with the inputting of text from MEDLINE and parsing it into individual abstracts (Step 3.1), where fields of interest such as title, abstract, date and PubMed ID are identified and extracted (Step 3.2). The records are pre-processed to remove double spaces, unusual characters and carriage returns (Step 3.3). The abstract is then parsed into individual sentences for analysis (Step 3.4), during which sets of words are analyzed in different set sizes by parsing the sentence into individual words

(Step 3.5) that are then put into an array (Step 3.6). Starting with the longest string of words recognized by IRIDESCENT (5 in its current version), these strings of words are matched against the objects in the ORD (Step 3.7). When a set of words is matched, it is marked as being an object. If no match is found, smaller and smaller word sets are searched down to individual words. By starting with the longest and moving to the shortest, we match phrases such as “polycystic kidney disease” before the phrase “kidney disease”. If a term has been flagged by ARGH as being ambiguous within text (i.e. it has multiple definitions), then IRIDESCENT checks to see if that acronym-definition pair was resolved earlier in the abstract (Step 3.8). Similarly, if there is a particular capitalization requirement for the term, a special routine is called to check if the capitalization patterns are consistent with the database representation (Step 3.9). For example, the term “KD” when capitalized can stand for Kidney Disease, but when it is mixed case “kD” or lowercase “kd” it will refer to kilodaltons. As objects are identified, they are added to an array that is processed once the abstract ends. If the relationship is new, a new database entry is created. If not, then the counter within the existing relationship is incremented. After the relationship database is populated, the user can then use IRIDESCENT to analyze relationship sets.

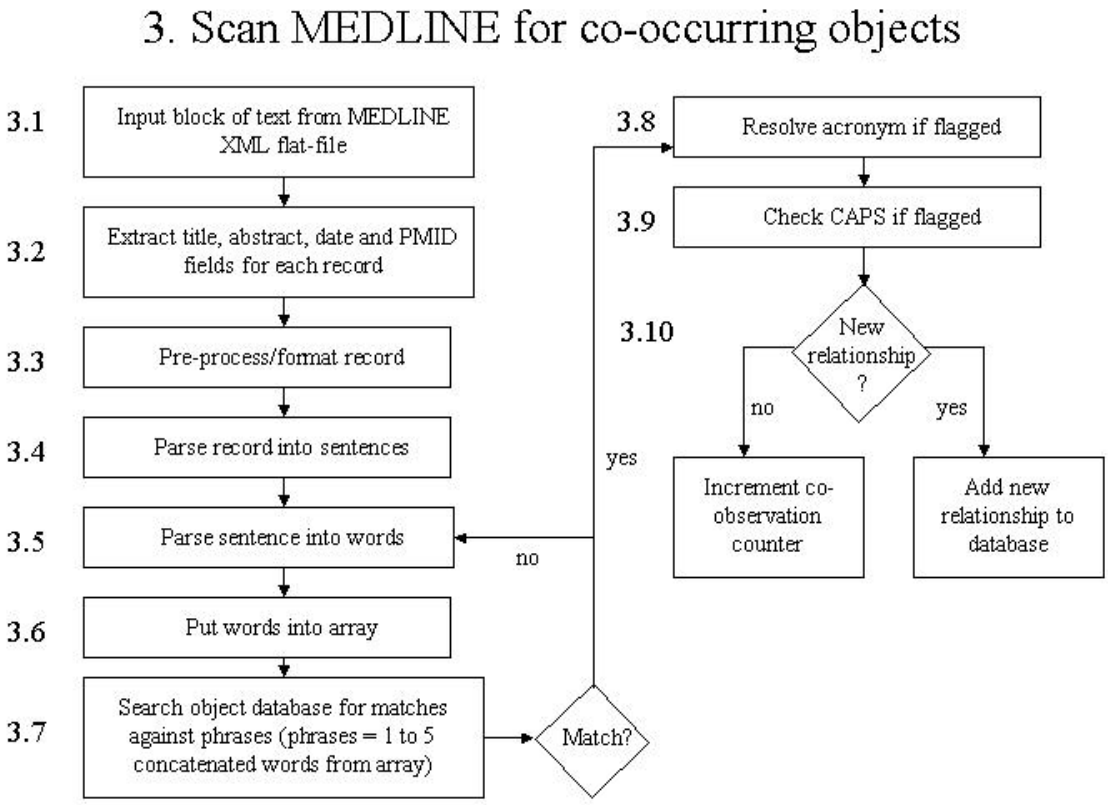


Figure 8: MEDLINE records are sequentially processed from large (>100MB) XML files. Fields of interest for each record (title, abstract, date and PubMed ID) are determined from the tags within these files and used to extract it. Each record is then processed to identify objects contained within. Their relative location to other objects (same sentence or same abstract) is ascertained and included in the co-occurrence database. CAPS = capitalization patterns.

3.4 Acronym resolution: A critical step in increasing both precision and recall

Systematic false-positive errors proved to be highly problematic when attempting to catalog a relationship between objects, invalidating from 1% to 100% of the observed co-mentions used to establish a relationship. The primary contributors to these systematic errors were homonym-like and polynym-like terms. Homonyms are words spelled identically but

with different meanings, and we use the term homonym-like to denote that the matching terms are not necessarily words but can encompass acronyms and abbreviations as well. Polynyms are acronyms spelled identically but with multiple definitions, and we use the term polynym-like to more broadly encompass terms such as gene symbols (e.g. p40) that may not necessarily be acronyms, per se, but are used to refer to different genes. At this point, it became necessary to resolve acronyms and provide a greater quality control for term recognition within abstracts, so separate modules were created for IRIDESCENT.

Acronyms, abbreviations and other forms of word or phrase shortening (hereafter just collectively referred to as "acronyms") aid in the efficiency of communication, but are confusing for text-mining software like IRIDESCENT when the acronym has multiple definitions (i.e. it is a polynym). Table 2, for example, shows several examples of ambiguous acronyms within MEDLINE.

Gene Name	**Definition**	**Most popular alternative meaning(s)**	**DPA score**
GAS	Gastrin	Group A Streptococci, Global Assessment Scale	3%
NM	Neutrophil Migration gene	Nuclear Matrix, Nodular Melanoma	1%
SD	Segregation Distortion gene	Standard Deviation, Sprague-Dawley	<1%
CT	Cytidylyltransferase 1	Computed Tomography, Calcitonin	<1%
ACT	Activator of CREM in Testis	Activated Clotting Time, Antichymotrypsin	<1%

Table 2: Acronyms frequently have different meanings within the literature. How frequently they have an alternative definition within a body of literature (e.g. MEDLINE) can be estimated by the Definition Percentage of unique Acronym (DPA) score. DPA is calculated by dividing the # of times one specific definition is used for a unique acronym by the total # of definitions used for the acronym.

To allow disambiguation of such terms, an automated, accurate and scalable method was developed by which acronym-definition pairs could be identified within text. The program written for this was entitled Acronym Resolving General Heuristic (ARGH), the implementation of which was published in Methods of Information in Medicine[66]. ARGH will be recapitulated here in part because resolving acronyms is a critical step in being able to recognize objects within text. The ARGH manuscript was also written for a more general audience, and it will be informative to discuss it specifically in the context of IRIDESCENT.

ARGH enables IRIDESCENT to resolve author-defined acronyms within text, or at least assign a probability score between an acronym and its set of potential definitions. It was used to create, in an automated manner, a reference work on acronym definitions. This compilation of reference works using information retrieval and extraction is also known as knowledge base construction[24]. This reference work has several advantages over manual or semi-automated methods, besides time and effort saved, such as enabling identification of relative frequencies for alternate acronyms and definitions as well as spelling, phrasing and hyphenation variants for a unique acronym-definition pair. It also aids in identifying acronym/definition variants present in the literature that may not necessarily be in biomedical databases. To resolve and identify acronyms, a set of heuristics to accurately locate and identify the boundaries of acronym-definition pairs was developed and refined in terms of precision and recall on subsets of MEDLINE records. These training sets were gradually increased in size and heuristics re-evaluated to ensure scalability.

However, as evidenced by several collections of acronym definitions in both printed and electronic formats, acronyms are not always defined before being used. Recent printed

biomedical acronym reference books are manually compiled from a subset of available literature and contain from 4,000 to 32,000 acronyms[67-69]. As we later show by calculating the annual growth rate of biomedical acronyms, any printed reference will be quickly outdated by the time it is published. Many online sources for acronym and abbreviation definitions are relatively narrow in their scope such as the Human Genome Acronym listing (http://www.ornl.gov/hgmis/acronym.html), the WorldWide Web Acronym and Abbreviation Server (WWWAAS) http://www.ucc.ie/acronyms/acro.html with 18,000+ acronyms, the Pharmaceutical Lexicon (http://www.pharma-lexicon.com) with 27,000+ acronyms, Acronym Search (http://www.acronymsearch.com) with 40,000+ acronyms, and Bioabacus, which is a conglomeration of many such sites[70], containing 6,000 entries as of its 5th release version. Perhaps the most comprehensive site on the web is Acronym Finder, claiming over 88,000 acronyms/abbreviations and their 202,000+ definitions (as of 8/8/01, http://www.acronymfinder.com), which consists primarily of terms from highly acronym prone fields such as "computers, technology, telecommunications, and military". So far, the utility of such databases has stemmed from their general value to a community of users as a source of reference. Use of manually compiled and curated databases in natural language processing and information extraction efforts is of limited value for several reasons: First, the sources they are derived from are not always literature-based, some are submitted by users or assimilated from other published lists. Hence, the user cannot be certain the given spelling, hyphenation and/or word phrasing can be considered either standard or even common, and while this is perfectly convenient for conveying the meaning of an acronym to a human reader, it can prove problematic for computational analysis. In addition, IRIDESCENT is

concerned with retrieval and/or analysis of terms within MEDLINE only, so inclusion of terms from other domains contributes nothing towards recognition and will even slow down processing. Third, while a list of possible meanings is given in such dictionaries, it is not clear what their relative abundance is. While users may discover from these databases that an acronym has multiple definitions, they are given no information on how common or rare each definition is. This will be critical for IRIDESCENT to decide which acronyms will require resolution within MEDLINE. Any acronym whose primary meaning consists of less than 90% of recognized definitions is flagged in the ORD to require acronym resolution whenever it occurs within text before a relationship is established.

Other automated methods have been developed and applied to the same or highly similar problems. However, most of these methods usually pre-define what an acronym is supposed to look like and then write rules for its recognition. I believe it is somewhat disingenuous to pre-define what an acronym is "supposed" to look like (e.g. must begin with an alphabetical character, must be between 3-6 characters long, etc.) and then measure the precision and recall of one's rule set afterward in terms of the rules you have just laid down. ARGH was approached from the standpoint of identifying as many acronyms as possible, as determined by human assessment, and then gradually adding heuristics that cover the way in which acronyms are defined within MEDLINE to reduce the amount of false positives. With any rule not based upon an absolute truth or applied to an imperfect dataset will come false-negatives, and so the goal with ARGH was to gradually refine these rules, keeping track of the FP and FN rates. Additionally, most other approaches examined are not scalable, having

been tested on only hundreds or thousands of abstracts. Building a database using such heuristics would have ultimately failed when applied to 12 million MEDLINE abstracts.

ARGH is superior to all previous programs in almost every aspect, with the possible exception of AcroPhile[71]. AcroPhile was designed to search web pages for potential acronyms, is the most sophisticated of the group and the only one to be applied to a relatively large body of text (936,550 government and military web pages)[72]. AcroPhile consists of four different algorithms that vary in their performance in obtaining acronym definitions from such web pages, with precisions ranging from 87% to 94% and recalls from 59% to 88%. AcroPhile has an advantage over ARGH in that it is able to identify acronym-definition patterns outside of parentheses. ARGH is most similar to their contextual approach, but differs from AcroPhile primarily in that it does not pre-define patterns for acronym-definition pairs. Instead, ARGH first attempts to move right-to-left, matching consecutive letters found within the acronym to letters within the definition and then uses a heuristic set to distinguish between valid and invalid pattern matches. ARGH also imposes very loose length restrictions on the length of definitions and acronyms (255 characters) and instead of using a list of "noise words" to be skipped in matching patterns, ARGH simply allows a finite number of non-matching intermediate words (e.g. "rats" may be a skipped word in "Sprague-Dawley rats (SD)").

Shown below in Table 3 are some examples of how acronyms are constructed within MEDLINE. These categories proved useful in deciding which heuristics would be likely to discard the least number of relevant terms.

Type	Freq.	Term	Definition	Comments
I	38%	AD	Alzheimer Disease	Sequential matching of capital acronym letters 1st letter to each word in definition
I	1%	bpm	beats per minute	Acronym letters correspond to 1st letters in definition words, capitalization unimportant
I	5%	OTG7	Orchid Transitional Growth related gene 7	More words in definition than letters in acronym
I	2%	scFv	single-chain variable fragments	Acronym letters are not in the same order as major letters in the definition
Ib	2%	TBK	Total Body Potassium	Consecutive 1st letter matches, except a symbol is substituted for a definition word
Ic	4%	EPNP	1,2-epoxy-3-(p-nitrophenoxy)-propane	First acronym letter is not first word letter in definition
II	9%	TGFbeta	Transforming Growth Factor beta	Acronym is a mixture of 1st letter capitals and spelled-out symbol/word
II	14%	GGA	Geranylgeranylacetone	Definition is concatenation of multiple words, acronym letters correspond to each word
II	22%	MVA	Mevalonic acid	Some acronym letters match 1st letters in definition words, others are intermediate
II	1%	Dsh	Dishevelled	Abbreviation consists of letters within nearest word
II	<1%	Botox	botulinum toxin	Abbreviation is concatenation of first letters from adjacent words
II	1%	EcoRec	ecotropic retrovirus receptor	Abbreviation is concatenation of first letters from separated words
IIa	1%	EP	Phosphoenzyme	Acronym letters rearranged within the same word

Table 3: Variation in acronym construction. Shown are examples of the variation in the ways acronyms and abbreviations are formed within a set of 100 abstracts examined, making comprehensive identification a non-trivial matter. Such constructs can be categorized into two basic types: Acronym-like (Type I) and abbreviation-like (Type II) and within each type are variations. Type Ib matches first letters, but one of the letters is symbolic in nature. Type Ic matches first letters, but such letters may come after other punctuation besides spaces. Type IIa deviates from the standard method of constructing abbreviations by using definition letters in non-sequential order. Also shown are the relative frequencies of each type of abbreviated construct.

ARGH defines acronyms as *any* abbreviatory shortening of words or phrases, not purely symbolic in nature, from a corresponding definition. Potassium (K) and Silver (Ag) are examples of purely symbolic representations, since the symbols used to represent the words are not derived from the word itself, but rather its parent language (latin abbreviations for Kalium and Argentum, respectively). However, since some acronyms are derived from a combination of their representative words and a symbolic reference, we do count those as valid acronyms (e.g. triiodothyronine (T3)). Definitions and acronyms are restricted to be no more than 255 characters long. Since none of the observed definitions were over 200 characters long and abstracts are usually limited in length, it is not likely that a significant number of acronyms or definitions will be excluded by this restriction.

We additionally distinguish between rates of systematic precision (defined as true positives/(true positives + false positives)) and systematic recall (defined as true positives/(true positives + false negatives)) and per-identification-event rates of precision and recall. Systematic rates refer to database entries and are reflective of how accurate and inclusive ARGH is in compiling acronym-definition patterns from a set of literature. Per-identification-event rates refer to the ability of the system to recognize instances of acronym-definition patterns within text. We distinguish between the two because a system can have an impressive rate of 98% accuracy per-identification-event on relatively small sets of literature, which may be adequate for automated recognition of terms in text-processing, but it is insufficient for automated methods of database construction because as more literature is processed, errors accumulate.

Database entries were considered false positives when they contained words unrelated to the definition of the acronym. For example, a definition of "In interleukin-2" for the acronym "IL-2" would be considered a false positive error. If a heuristic was added that excluded this entry and it was the only one containing "interleukin-2" as a definition for IL-2, the exclusion would affect the systematic recall. However, if the heuristic excluded this entry but no other entries containing valid definitions for IL-2, it would only lower the per-identification-event recall. A definition such as "Interleukin-2 gene" for IL-2 would not be considered an error because, even though the word "gene" is not represented by any symbols within the acronym, it is directly relevant to the description of what IL-2 is and can be considered a definition variant. Finally, only entries that were a result of a software identification error were counted as FPs. For example, the definition "Interleukine-2" for IL-2 is most likely a spelling error, but could also be a valid variation (e.g. American "Armor" versus the British "Armour"). It is beyond the scope of ARGH to attempt to discern the two. The set of heuristics used are summarized in Tables 4a and 4b.

Basic heuristics for locating acronyms & definitions (n=100)	**Total Pos.**	**True Pos.**	**False Negs.**	**Systematic Precision**	**Recall Per ID event**	**Systematic Recall**
Term encased within parentheses	520	165	4	32%	97.6%	100%
Term consists of one word only	311	165	4	53%	97.6%	100%
Term must contain at least one alphabetic character	211	165	4	78%	97.6%	100%
All acronym letters also in definition, in consecutive order	162	159	10	97.9%	94.1%	93.8%
Allow non-sequential 1st letter matches in definition words	163	160	9	97.9%	94.7%	93.9%
Additional heuristics for boundary definition (n=1,000)					(est.)	(est.)
None	1054	825	--	78.3%	94.7%	93.9%
Require 1st letter match on abbreviation-type acronyms	1054	869	+0	82.4%	94.7%	93.9%
Limit # of definition words to # of letters in acronym+2	876	867	+2	99.0%	94.6%	93.7%

Table 4a: Heuristics to locate acronym-definition pairs and their boundaries. A set of heuristics was cumulatively applied to batches of MEDLINE records (titles and abstracts) to identify acronym-definition patterns. As the size of the dataset increased, more variation was observed in the way acronym-definition patterns were constructed, requiring the addition of new heuristics to increase the overall precision. False negatives for the additional rules are reported in terms of how many additional valid entries are excluded from the database.

Large scale heuristics for validating acronym/definition patterns (n=1,000,000)	**Dataset total entries**	**Dataset valid entries**	**Total # entries matching criteria***	**# valid entries discarded (est.)**	**Systematic Precision**	**Syst. Recall (est.)**
None	500	433	--	--	86.6%	93.7%
Certain words in definition restrict which acronym types are valid	468	433	7,950	809	92.5%	93.1%
Allow only certain punctuation within acronyms & defs.	465	433	1,485	119	93.1%	93.1%
Restrict types of valid parentheticals within def.	458	433	3,616	217	94.5%	92.9%
Restrict occurrence of acronym as contiguous substring of def.	450	433	7,999	80	96.2%	92.8%
Acronym/definition ratio restrictions	448	433	2,294	138	96.6%	92.8%
Restrict automatic extension for units	445	433	164	0	97.3%	92.8%
Require 1st letter matches for "II", "III" and "OH"	443	433	2,312	0	97.7%	92.8%
All of MEDLINE processed (n=12,037,763)						
None	500	481			96.2%	92.8%

Table 4b: Heuristics developed to reduce error rates in large-scale (over 1 million records) datasets. Basic heuristics for identifying acronym-definition patterns work well on smaller datasets, but the variability in constructing these patterns eventually lowers the systematic precision (# of correct entries / total # of entries) as more text is analyzed. A total of 153,616 unique acronym-definition patterns were recognized within 1,000,000 MEDLINE records, an estimated 133,031 of which are valid entries. *Some entries match more than one criterion.

With this set of heuristics, we processed all available MEDLINE records obtained from the National Library of Medicine (NLM) in XML format, representing a total of 12,037,763 records (37.3 gigabytes in size) dating up to February 2002. From a total of

6,418,919 abstracts, ARGH recognized 4,562,567 acronym-definition patterns, of which 98.8% were found in the format *definition(acronym)* and the other 1.2% in the format *acronym(definition)*. From these patterns, a database of 737,330 records was created, containing 174,940 unique acronyms/abbreviations and 638,976 unique definitions. Of the unique acronyms, 63,440 (36%) had more than one definition associated with them and 62,974 definitions (10%) had more than one acronym associated with them.

To estimate overall precision per database entry, we chose 3 random subsets of 500 records (again by generating random record ID numbers from within the database) and found 19, 15 and 18 FP errors, giving an estimated overall systematic precision rate of 96.5 ± 0.4% per entry. From observing the number of unique acronym-definition patterns excluded, we estimated the systematic recall rate to be 92.8%. To verify the accuracy of this estimate, we obtained 3 sets of 100 (effectively) random abstracts different than our original set by searching PubMed on the non-topical keywords "determined", "below" and "set". We then manually counted the number of acronyms defined in any manner within the titles and abstracts, and checked our database for the existence of the corresponding acronym-definition pair. Ratios of identified/existing acronym-definition pairs were 139/152 (91.4%), 101/105 (96.1%) and 86/94 (91.5%) for the sets, respectively, giving an overall rate of 93.0 ± 2.7%.

ARGH can be accessed online at http://lethargy.swmed.edu/ARGH/argh.asp. Frequency statistics were compiled for each acronym-definition pattern found within MEDLINE, and used in the online interface to sort acronyms or definitions by their relative abundance within the literature. This enables users to quickly identify which

acronyms/definitions are more common and more likely to be implied in the absence of additional information. These frequency rankings also enable users to determine which spelling, hyphenation or phrasing variant could be considered the "standard" one, by popular use. In addition, for each acronym definition, the date of its earliest occurrence was included in the database, allowing a historical perspective of the approximate time a definition for an acronym was coined and also enabling us to construct a plot of the growth in the number of unique acronyms and abbreviations over time versus the number of their definitions. Figure 9 shows that the number of definitions (this number includes definition variants as well as completely different definitions) is increasing at a faster rate than the number of unique acronyms available to represent them.

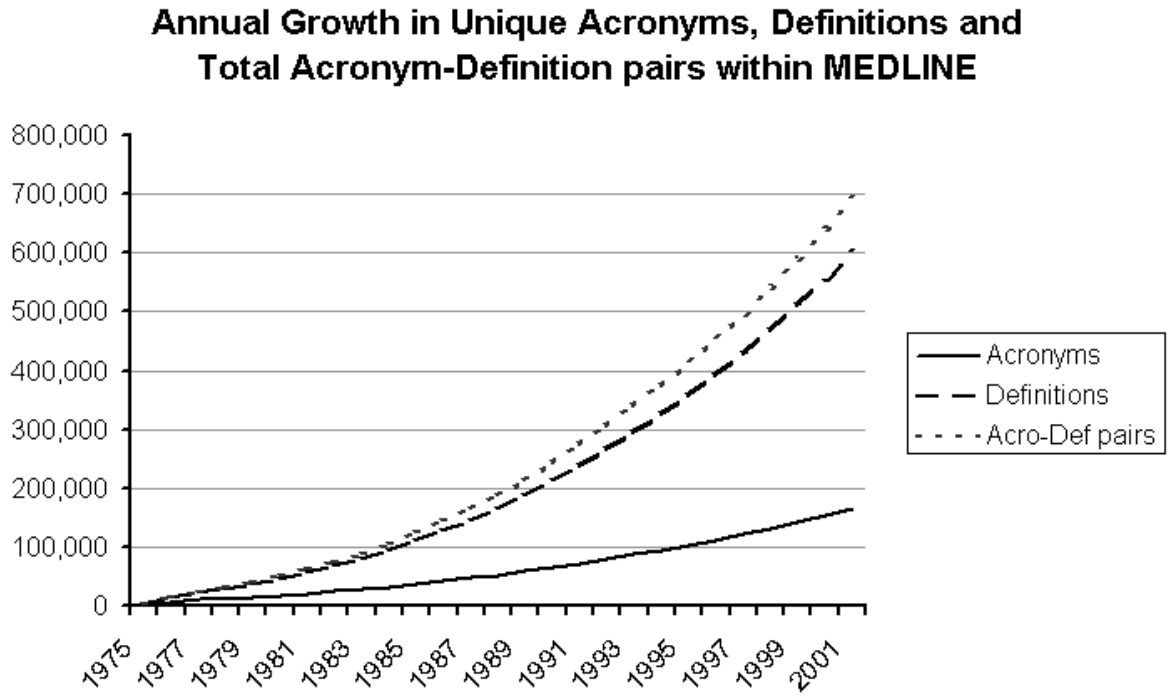


Figure 9: The growth in the number of unique acronyms, definitions and acronym-definition pairs is increasing exponentially. Acro-Def pairs = Acronym-Definition pairs.

In MEDLINE, therefore, it is becoming less likely that in the absence of a definition within the originating text, many acronyms will be unambiguously associated with the intended definition. Because of this ambiguity, it would be useful to know how likely a given acronym is associated with one particular definition and vice versa. Therefore, we can calculate the Definition Percentage of unique Acronym (DPA) and Acronym Percentage of unique Definition (APD) as a way of estimating the likelihood of a specific acronym being associated with a specific definition in the absence of an explicit definition. This is useful for IRIDESCENT to ascertain which acronyms require resolution when they occur in text. Currently, IRIDESCENT considers an acronym unambiguous if the database definition comprises at least 95% of all identified definitions. Table 5 shows an example of acronyms with a large number of alternative definitions, giving the two most popular definitions in the database and their DPA scores. Here we see that some acronyms such as CT are predominantly associated with one definition (or its variant), while others such as PA are not.

Acronym	# of unique defs.	total # of all defs.	Most popular definitions	# times this def. found	DPA
CA	1,206	6,857	Calcium	1,376	20%
			Carbonic Anhydrase	598	9%
PA	1,084	6,466	Plasminogen Activator	745	12%
			Phosphatidic Acid	703	11%
PC	1,068	7,548	Phosphatidylcholine	2,741	36%
			Phosphorylcholine	315	4%
CS	1,002	5,527	Conditioned Stimulus	566	10%
			Circumsporozoite	310	6%
PS	925	5,236	Phosphatidylserine	1,269	24%
			Paradoxical Sleep	409	8%
PI	921	9,419	Phosphatidylinositol	1,978	21%
			Inorganic Phosphate	1,010	11%
SC	887	4,810	Superior Colliculus	757	16%

			Subcutaneous	548	11%
AP	879	7,026	Alkaline Phosphatase	1,120	16%
			Action Potential	590	8%
CP	868	5,537	Cyclophosphamide	607	11%
			Cerebral Palsy	462	8%
CT	866	25,899	Computed Tomography	14,033	54%
			Computed Tomographic	3,414	13%

Table 5: Within MEDLINE, a number of acronyms have many different definitions (polynyms). Going by the total number of different definitions found within MEDLINE, the ten most ambiguous acronyms are shown. Not surprisingly, many of the most ambiguous acronyms were those with the least number of letter combinations to represent them. The Definition Percentage of unique Acronym (DPA) scores provide a quantitative estimate of how likely an acronym is to be specifically associated with a definition within the body of literature examined in the absence of a definition.

This ambiguity extends to the creation of acronyms from definitions as well (as shown in Table 6).

Definition	**# times definition found**	**# of different acronyms**	**Most popular acronyms**	**# times acronym used**	**APD**
alkaline phosphatase	3,227	38	ALP	1,624	50%
			AP	1,120	35%
beta-glucuronidase	848	36	GUS	654	77%
			BG	40	5%
glucose-6-phosphate dehydrogenase	1,585	35	G6PD	910	57%
			G-6-PD	262	17%
alpha-tocopherol	246	29	alpha-T	63	26%
			AT	38	15%
beta-endorphin-like immunoreactivity	113	27	beta-END-LI	28	25%
			bet-EI	14	12%
beta-Endorphin	822	25	beta-EP	349	42%
			beta-END	199	24%
5`-nucleotidase	194	25	5'-NT	37	19%
			5'-Nase	29	15%
peripheral blood mononuclear cells	6,953	25	PBMC	4,933	71%
			PBMCs	1,370	20%
glyceraldehyde-3-phosphate dehydrogenase	650	25	GAPDH	474	73%
			G3PDH	42	6%

2-chloroadenosine	172	24	2-CADO CADO	33 32	19% 19%

Table 6: Multiple acronyms can exist for a unique definition within MEDLINE. Acronyms can be created from definitions in a variety of ways, adding a different kind of ambiguity in uniquely associating acronyms with a definition. Shown are ten definitions with the most corresponding acronyms and/or abbreviations within our database, and their Acronym Percentage of unique Definition (APD) score, providing an estimate of how frequently a specific acronym is used to represent a unique definition. Note that the APD score does not take into account the ambiguity of an acronym in representing other definitions. For example, while BG was defined in this table as beta-glucuronidase 40 times, it was also defined as Blood-Glucose 199 times.

The DPA score can be useful in estimating how ambiguous an acronym is within a body of literature (in the absence of a definition), but we found it was limited in its utility when a definition varied widely in its spelling, hyphenation patterns or phrasing. For example, JNK had 77 different definitions in our database, but they were all variants on the definition “c-Jun N-terminal kinase”. The DPA score of 41.6% for the most common definition might give the impression that JNK has alternative definitions, when it does not. As a partial solution to this problem, we have created a “stemmed” version of the ARGH database where plural endings, spacing and punctuation have been removed. Stemming reduced the number of unique definitions to 540,821 (85% of the original size) and helped in some cases, but for entries like JNK where the second most common definition is “c-Jun NH2-terminal kinase”, it did not. We developed a routine to align the definitions and compare similarity scores, and found it worked well under most circumstances (Table 7a) but was unable to distinguish circumstances under which a minor variance was critical to the meaning of the definition (Table 7b). It provided a limited and imperfect solution to the problem of matching conceptually identical definitions from their semantic variants, and will require more work.

Acronym	Definitions	Similarity
DMH	Dimethylhydrazine 1,2-dimethylhydrazine ----++++++++++++++++	81%
IL-2	Interleukin-2 Interleukin-2 gene ++++++++++++++-----	76%
12-HETE	12-hydroxy eicosatetraenoic acid 12-hydroxy-5,8,10,14-eicosatetraenoic acid ++++++++++----------++++++++++++++++++++	73%

Table 7a: Some term variants contain additional descriptive words or symbols. Aligning their letters is one way of measuring similarity. Where the terms fail to align is useful in determining whether or not they should be considered identical. Plural endings or additional descriptors, such as in the IL-2 example above, do not represent the existence of a conceptually different entity. Prefixes such as in DMH are sometimes unimportant as well, but sometimes are quite critical to certain chemical aspects of activity (e.g. L-alanine versus R-alanine).

By establishing that the difference exists in one contiguous block of text and that the terms are otherwise identical over a given percentage of their length, we can estimate which terms are identical in meaning. While this works well for most of the variants found within MEDLINE, there are instances in which it does not, as in Table 7b.

Acronym	Definitions	Similarity
ABP	Androgen binding protein Auxin binding protein ------++++++++++++++++	71%
AD	Alzheimer's disease gene Aujeszky's disease gene ---------+++++++++++++++	63%
ACG	Acetylgalactosamine Acetylgluc osamine ++++++++----++++++	74%

Table 7b: Relatively minor differences between terms can be critical to the meaning. Simple percentage cutoffs (e.g. 66%) would erroneously recognize 2 of these terms as being identical when they are in fact entirely different entities.

3.5 Using the Merriam-Webster dictionary to determine capitalization requirements for objects and screen out uninformative words

When conducting direct textual comparisons, capitalization patterns matter. Not all gene names are capitalized (e.g. alpha-2 microglobulin) in the database, but if they begin a sentence then the capitalization is forced. Similarly, if capitalization patterns are inconsistent between the object as given by the database and the object as it appears within text, references within MEDLINE will be missed. Consequently, IRIDESCENT conducts all word comparisons in lower-case. This, however, is not without its drawbacks. Shown in Table 8 are gene names that match common words.

Gene symbol	Full Name	Term Frequency
LARGE	Like-acetylglycosyltransferase	346,940
MICE	MHC class I polypeptide-related E	252,904
END	Endoglin	194,157
LIGHT	Ligand invasive growth herpes transmembrane	177,995
SEX	Sex chromosome X (Plexin A3)	127,176

Table 8: The five genes with the most entries returned from a PubMed query. These 5 words, aside from being a generally true statement about the disruptive presence of oversized rodents upon casual procreation, also happen to share the same spelling with several common words. During text scanning, this type of error can sometimes be corrected by checking capitalization patterns.

In these cases, it is useful to know if the capitalization pattern within a word matters or not. To resolve this problem, the Merriam-Webster dictionary was assimilated from Project Gutenberg (http://www.gutenberg.org/), which is an effort to make classical printed works available electronically. While any sufficiently large non-scientific source would do (e.g. Cosmopolitan magazine), the electronic availability combined with what should be a comprehensive coverage of English words made Merriam-Webster a more attractive source.

Words in the ORD that match entries from the Merriam-Webster dictionary are flagged so that when they are identified within text, their capitalization patterns are checked with the capitalization pattern as given in the ORD. This works well for most entries, but for some abbreviations there is still a problem as shown in Table 9. Difference in capitalization patterns was recognized early as a potential problem, and was simply incorporated as a routine - no evaluation of impact on FP/FN rates was conducted.

Abbreviation	**Full name(s)**
For	Formate, Forssman antigen
As	Arsenic, anti-sense, Aspermia
And	Androstenedione
If	Fetal insulin, Free inhibitor
But	Butanol, Butirosin

Table 9: The abbreviations for some terms match common words as well, but capitalization patterns do not enable consistent discernment. Methods of context determination are necessary.

All 150,922 words found within the Merriam-Webster dictionary were assimilated into a database and compared with each of the single-word entries in the above popular databases. For some model organisms, such as *Drosophila melanogaster*, it is more difficult to apply information extraction methods to identify gene names within text whereas within the *Saccaromyces cerevisiae* literature, gene names are almost completely unambiguous (Table 10). By conducting this comparison we are able to flag which entries require capitalization checking to be considered valid and which have a high probability of being confused with common words regardless of capitalization. This way, when a lower-case match is made with IRIDESCENT's object database (e.g. LARGE is matched with "large"),

the software knows that the same capitalization pattern must be followed for the relationship to be recorded (i.e. the word must be in all capital letters to be recognized as the gene name).

Database	# of single-word entries	Entries matching common words
OMIM	15,859	580 (3.6%)
HGNC/GDB	24,736	604 (2.4%)
Locuslink Human	16,767	343 (2%)
Locuslink Mouse	16,102	563 (3.5%)
Locuslink Drosophila	6,249	1,163 (18.6%)
SGD	6,626	9 (0.1%)

Table 10: The number of terms identical to 'common' words (common being defined as found within the Merriam-Webster dictionary) varies by database. OMIM = Online Mendelian Inheritance in Man, HGNC = Human Genome Nomenclature Committee, GDB = Genome Database, SGD = *Saccharomyces cerevisiae* Genome Database.

In summary, there are a series of steps that need to be taken to ensure terms are accurately identified within free-form textual input (Figure 10). Ambiguous acronyms (polynyms) need to be identified so that the system can recognize which terms require the system to resolve which definition is intended for the abbreviated term in question (Step 2.1). Common words must be identified to recognize terms in which capitalization patterns are important, as well as when a database search is to be constructed. For example, when querying PubMed, one could not use batch query to retrieve records on the acronym "LARGE", as case is disregarded. This flag enables only the definition to be searched upon (Step 2.2). In the entries where capitalization patterns are important, IRIDESCENT then knows that when the term is encountered, it must be in the proper form to be counted as representing the object within the database it was intended to represent (Step 2.3). Lexical variants are critical to recognition, as many database entries will represent either a preferred

term, the most commonly used term, or simply the term the database curator is aware of. Within text, it is important to recognize terms that are sometimes hyphenated or are in plural form (Step 2.4). Similarly, acronyms for objects are not always given within the database, so it is necessary to recognize those acronyms when they occur within text (Step 2.5).

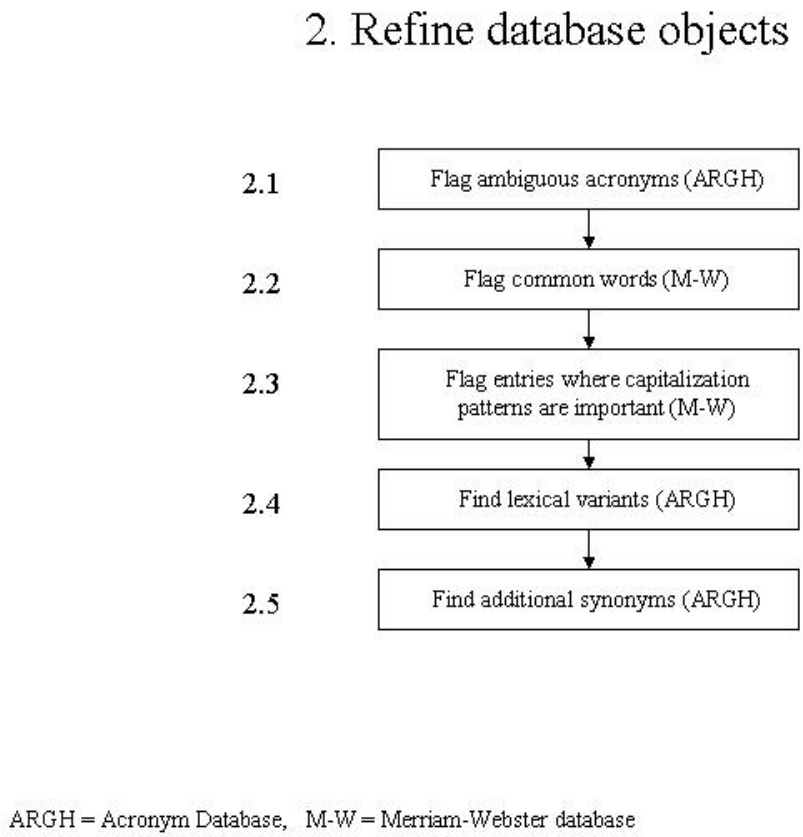


Figure 10: Quality checks required for the recognition of objects within free-form textual input. Acronyms must be either unambiguous or resolved within the text they are found (using ARGH), capitalization patterns distinguish some abbreviations and acronyms from common words (distinguished by comparing terms against the Merriam-Webster dictionary), and some terms as written in text vary in their lexical construction when compared to their corresponding database entry.

So far, our goal has been to enable heuristic-based refinement of the database, such that the user is not required to engage in large-scale specification of objects to be included or discarded from analysis. One of the goals in automating this effort is to decrease the amount of user-intervention that is required. However, there is still a need for fine-tuning and user flexibility. Thus, the final step in refining the ORD is to delete user-defined entries as well as incorporate other entries defined by the user. Currently, 742 entries are deleted. Reasons for deletion include terms that refer to vague entities or entities of a sufficiently broad class (e.g. antigen, arrest), entries that are considered to common to be informative (e.g. acid, age), entries that match with common words (e.g. for, the, next), and entries that are in error (e.g. "1,3").

3.6 Other Text-Mining Considerations: Term Variance and Identification

When the ambiguity in an acronym involves the difference between an object class of interest (e.g. gene, disease, phenotype, chemical compound) and a class that is not of interest (e.g. society name, journal name), the solution is simply to resolve it. But when the same acronym refers to two different objects within the same database (Table 11), it must be flagged as ambiguous. Genes are assigned official names by the Human Gene Nomenclature Committee (HGNC) to avoid duplication of symbols, but many of their synonyms already published in the literature conflict with the standard names. Recent literature will more likely contain the updated "correct" symbol, but determining what literature qualifies as "recent" varies between each term.

Gene Symbol	Gene Name
P40	Nucleolar protein p40 Laminin receptor 1 (alias) Proteasome 26S subunit (alias)
TPO	Thyroid Peroxidase Thrombopoietin (alias)
RSS	Russel-Silver Syndrome gene Rigid Spine Muscular Dystrophy (alias)
MCD	Malonyl CoA Decarboxylase Medullary Cystic Kidney Disease (alias)

Table 11: Synonyms for some genes are also primary names for others, necessitating an automatic flagging of ambiguity regardless of DPA score.

Lexical expansion in the recognition of terms is highly useful when the term varies a lot, examples of which are shown in Table 12. By aligning definitions whose acronyms match, a cutoff of 80% similarity within one contiguous block (i.e. only one mismatching gap is allowed) is used to determine whether or not two terms are similar. For most terms this works well, but some alignments fail (e.g. TNFR2 in Table 12 below) because of nested acronyms. This type of nested abbreviation is relatively rare, however.

Symbol	Definitions	# of times observed
JNK	c-Jun N-terminal kinase c-Jun NH2-terminal kinase c-Jun amino-terminal kinase	538 150 58
TNFR2	Tumor Necrosis Factor Receptor 2 TNF receptor 2 TNF-receptor type 2	13 7 1
TIF2	Transcriptional Intermediary Factor 2 Transcription Intermediary Factor 2 Transcriptional Intermediate Factor 2	7 6 2

Table 12: The way a biomedical term is spelled can vary wildly between authors and journals. This can be problematic in proper recognition of the terms. Shown here is the

number of times a specific symbol has been observed within MEDLINE to be associated with a specific definition.

For acronyms such as TNFR2, this can be dealt with in part by expanding nested acronyms (e.g. TNF) into their full definitions before comparisons are made to determine if two definitions are equal. If two terms are still not equal, as would be the case with the definition "TNF-receptor type 2", an imperfect solution is to "align" the different definitions as discussed earlier.

3.7 User Interface

VB6 offers a development environment that includes the easy implementation of a graphical user interface (GUI) coupled with the power of object-oriented programming and the flexibility of the Microsoft Visual Studio component add-in libraries. Shown in Figure 11 is the start-up screen for IRIDESCENT, providing access to both analysis and maintenance features within IRIDESCENT. First is the Object Analysis option, which allows the user to conduct an Implicit Relationship Analysis (IRA) or a Shared Relationship Analysis (SRA). The Array Analysis takes the user to a screen that enables an array of objects to be compared versus

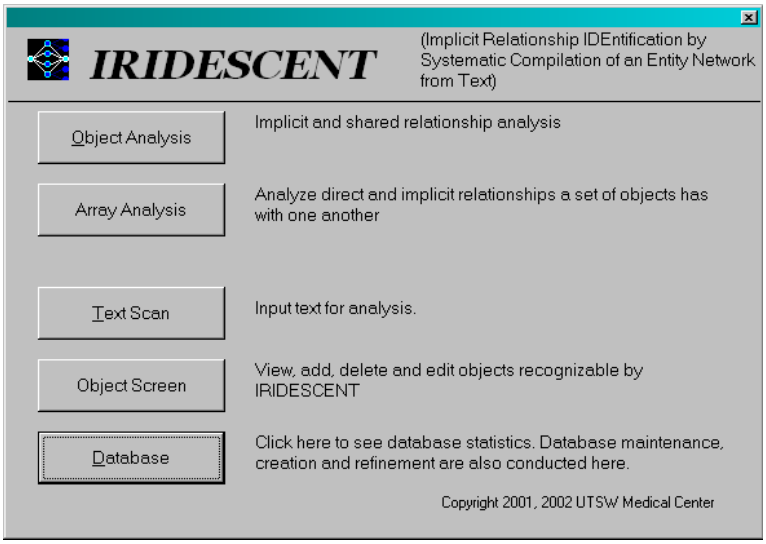


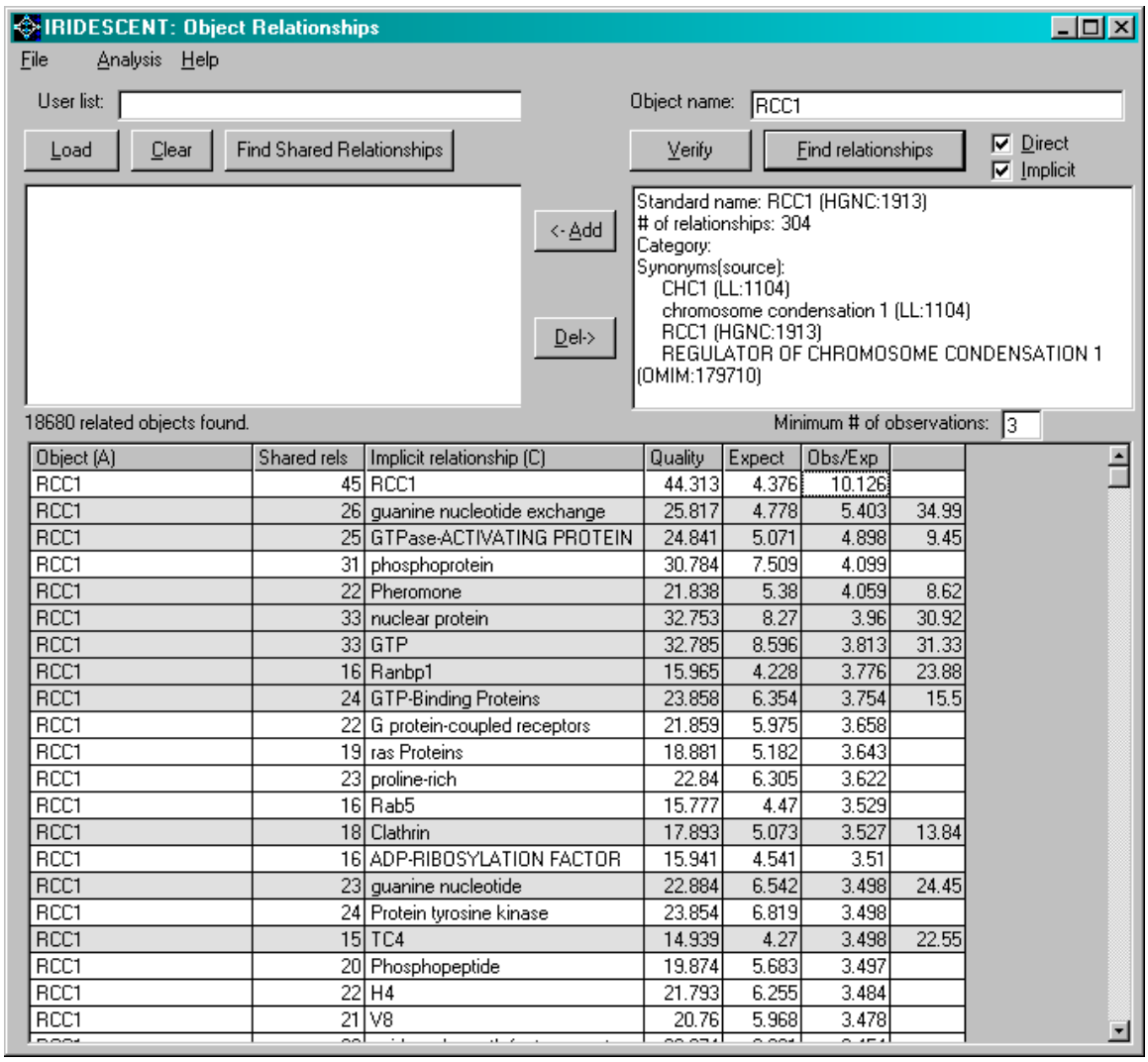
Figure 11: IRIDESCENT's entry screen displays what features and analyses are readily available to the user.

itself for direct and implicit relationships, which is useful for examining how a large set of objects might be interrelated such as in a microarray experiment. The Text Scan option enables the user to input and analyze text in terms of what the IRIDESCENT system is capable of seeing. That is, a user may input an abstract and see what objects and relationships IRIDESCENT is capable of analyzing. This option can also be used to scan text for additional object relationships to be added to the ORD. The Object Screen allows the user to view objects within the ORD, add new objects and/or synonyms as well as edit or delete existing objects. Finally, the Database option allows the user to perform tasks involving the ORD such as rebuilding the object recognition database, adding new records, and gathering statistics.

3.7.1 Implicit Relationship Analysis

Figure 12 shows the screen where the user can input lists of objects for analysis and analyze single objects for the known (direct) and implicit (indirect) relationships they have within the relationship network. The window in the upper right shows an example of how the existence of an object within the ORD is verified. The user inputs an object (e.g. RCC1) in the textbox entitled “Object Name” and presses the button “Verify”. If the object exists then information on the object is displayed in the window below. Here, we see that RCC1 stands for “Regulator of Chromosome Condensation 1”, along with other synonyms such as “chromosome condensation 1”. In parentheses next to these names and synonyms are the sources from which the name/synonym was obtained (LL=Locuslink, OMIM = Online Mendelian Inheritance in Man, HGNC = Human Genome Nomenclature Committee).

IRIDESCENT calculates how many times each of these synonyms is observed within MEDLINE. When all of MEDLINE has been processed, IRIDESCENT makes the most commonly used synonym the "standard" name. This helps users more easily recognize what object is being referred to. The number of relationships within the ORD (304 in this case) is also displayed within the window to give the user of how many potential relationships (co-mentions) of other objects with the analyzed object were identified within MEDLINE.



Object (A)	Shared rels	Implicit relationship (C)	Quality	Expect	Obs/Exp	
RCC1	45	RCC1	44.313	4.376	10.126	
RCC1	26	guanine nucleotide exchange	25.817	4.778	5.403	34.99
RCC1	25	GTPase-ACTIVATING PROTEIN	24.841	5.071	4.898	9.45
RCC1	31	phosphoprotein	30.784	7.509	4.099	
RCC1	22	Pheromone	21.838	5.38	4.059	8.62
RCC1	33	nuclear protein	32.753	8.27	3.96	30.92
RCC1	33	GTP	32.785	8.596	3.813	31.33
RCC1	16	Ranbp1	15.965	4.228	3.776	23.88
RCC1	24	GTP-Binding Proteins	23.858	6.354	3.754	15.5
RCC1	22	G protein-coupled receptors	21.859	5.975	3.658	
RCC1	19	ras Proteins	18.881	5.182	3.643	
RCC1	23	proline-rich	22.84	6.305	3.622	
RCC1	16	Rab5	15.777	4.47	3.529	
RCC1	18	Clathrin	17.893	5.073	3.527	13.84
RCC1	16	ADP-RIBOSYLATION FACTOR	15.941	4.541	3.51	
RCC1	23	guanine nucleotide	22.884	6.542	3.498	24.45
RCC1	24	Protein tyrosine kinase	23.854	6.819	3.498	
RCC1	15	TC4	14.939	4.27	3.498	22.55
RCC1	20	Phosphopeptide	19.874	5.683	3.497	
RCC1	22	H4	21.793	6.255	3.484	
RCC1	21	V8	20.76	5.968	3.478	

Figure 12: The Object Analysis screen within IRIDESCENT. From this screen, the user can input lists of objects for analysis or analyze an individual object in terms of its direct and indirect relationships. The highest Obs/Exp ratio should be the query object (at top), which serves as a positive control for the query and an upper boundary for all Obs/Exp scores.

Figure 12 also shows the result of an implicit relationship analysis on the query object, RCC1. Note that both checkboxes ("Direct" and "Implicit") are checked, indicating to the system that all objects that share relationships with the query object are to be displayed, whether the relationship is already known or not. The related objects are displayed in the grid below and when the relationship is known, the row is shaded. The strength by which objects are related to the query object, when known, is displayed in the rightmost column. When a relationship exists only through intermediates, it is unshaded. The column headers can be double-clicked to sort on any column that is displayed. In the figure above, it is sorted in descending order of the Obs/Exp ratio calculated for each set of shared relationships, bubbling the most statistically exceptional groupings to the top. A user might wish to also sort by most shared relationships (either the "Shared Rels" or the "Quality" column), to identify broad trends regardless of how statistically exceptional they are. This IRA is in summary form, displaying only how many relationships are shared. When the user clicks on the implicit relationship column, IRIDESCENT then expands the screen to display the related objects (B) shared by both (A) and (C). If the user were to click on the implicit entry "phosphoprotein", the information in Figure 13 would be displayed to the user.

A	AB_str	B	BC_Str	C	Implicit_Str
RCC1	49.07	Nucleus	340.22	phosphoprotein	49.07
RCC1	31.33	GTP	57.26	phosphoprotein	31.33
RCC1	30.92	Nuclear protein	91.08	phosphoprotein	30.92
RCC1	26.94	Chromatin	72.2	phosphoprotein	26.94
RCC1	26.28	replication	179.92	phosphoprotein	26.28

RCC1	22.21	mutations	136.97	phosphoprotein	22.21
RCC1	15.5	GTP-Binding Proteins	19.73	phosphoprotein	15.5
RCC1	12.43	membrane	865.32	phosphoprotein	12.43
RCC1	10.36	Vesicles	169.58	phosphoprotein	10.36
RCC1	9.45	GTPase-ACTIVATING PROTEIN	12.43	phosphoprotein	9.45
RCC1	13.84	Clathrin	7.79	phosphoprotein	7.79
RCC1	7.21	Guanosine	23.87	phosphoprotein	7.21
RCC1	24.45	Guanine nucleotide	6.63	phosphoprotein	6.63
RCC1	5.64	H2b	5.8	phosphoprotein	5.64
RCC1	22.55	TC4	5.39	phosphoprotein	5.39
RCC1	5.22	Kinase	731.48	phosphoprotein	5.22
RCC1	5.22	Tumor	429.92	phosphoprotein	5.22
RCC1	23.88	Ranbp1	4.81	phosphoprotein	4.81
RCC1	6.3	NTF2	4.56	phosphoprotein	4.56
RCC1	4.31	protein A	42.02	phosphoprotein	4.31
RCC1	3.98	UBIQUITIN	6.05	phosphoprotein	3.98
RCC1	3.73	Guanine	6.05	phosphoprotein	3.73
RCC1	3.73	repetitive sequence	19.98	phosphoprotein	3.73
RCC1	4.89	RANBP2	3.65	phosphoprotein	3.65
RCC1	3.73	X-linked	3.48	phosphoprotein	3.48
RCC1	34.99	guanine nucleotide exchange factor	3.48	phosphoprotein	3.48
RCC1	8.62	Pheromone	3.48	phosphoprotein	3.48
RCC1	3.4	Alternative splicing	45.5	phosphoprotein	3.4
RCC1	3.15	reticulum	149.53	phosphoprotein	3.15
RCC1	3.15	Nucleosome	15.17	phosphoprotein	3.15

Figure 13: The relationships (B) that RCC1 (A) shares with the object "phosphoprotein" (C). The strength of each A to B relationship is shown in the column "AB_Str", and each B to C relationship in the column "BC_Str". Since IRIDESCENT assumes an implicit relationship is only as strong as its weakest link, the output is sorted by the column "Implicit Str", which is the lesser of the two strength columns.

If the user were interested the MEDLINE abstracts that IRIDESCENT used to obtain the relationship, the user would then double-click on the strength column of the relationship of interest. For example, if the user were interested in the nature of the relationship that RCC1 and guanosine triphosphate (GTP) share, double-clicking on the strength column between them would bring up the abstracts within MEDLINE that contain the two terms (Figure 14). In this way, the user can begin examining the relationships of most interest for their biological relevance.

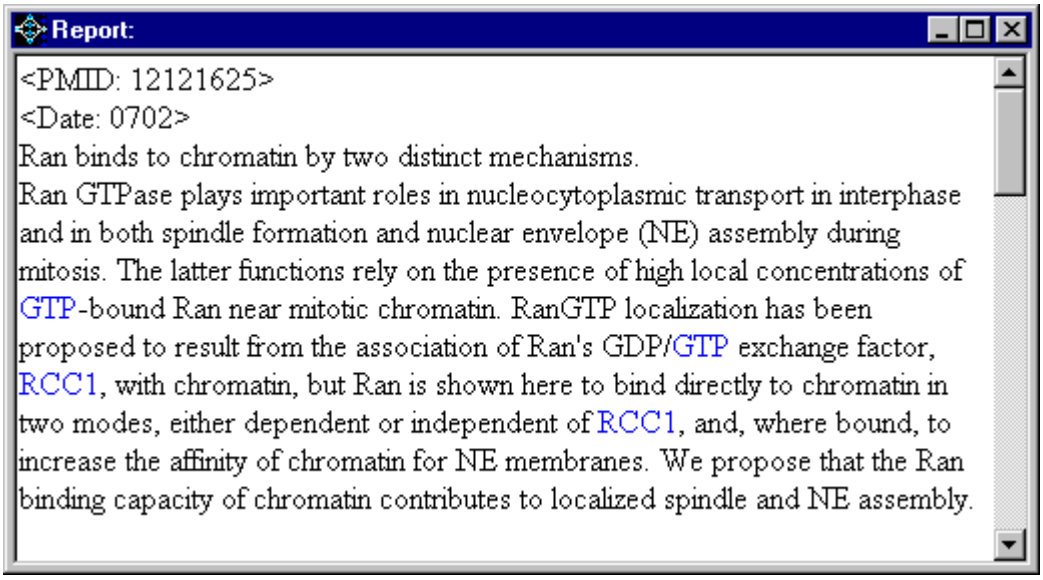


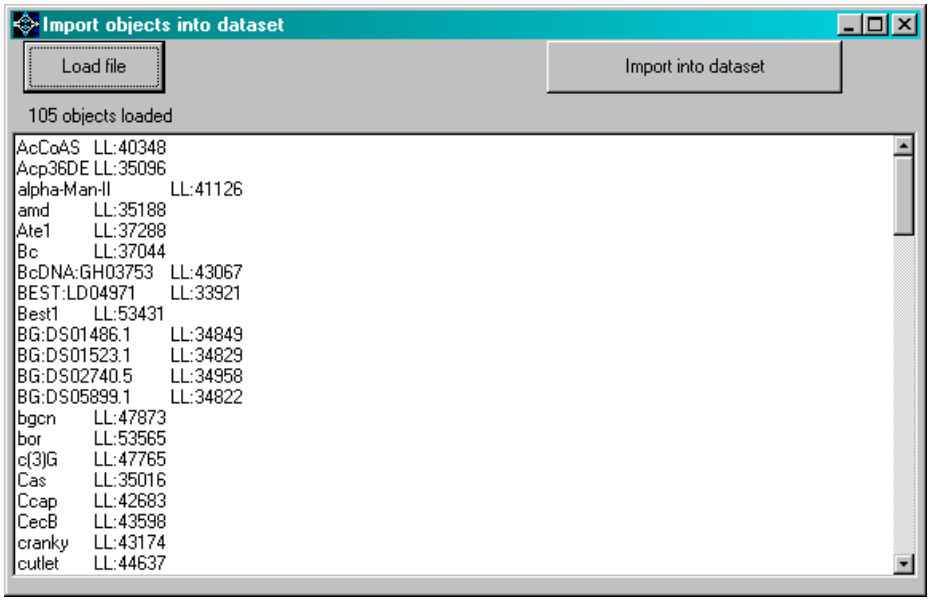
Figure 14: IRIDESCENT displays the MEDLINE abstracts that contain a name/synonym for both RCC1 and GTP. Recognized names/synonyms for these two terms are colored blue where they occur to draw the users attention to relevant areas of the text.

3.7.2 Shared Relationship Analysis

To conduct a shared relationship analysis, the user accesses the Object Analysis screen (Figure 12). Here, the user can create a dataset consisting of all the objects to be analyzed. The user would first name the dataset in the text box entitled “User list” and thereafter add or delete objects from this set. This is done by entering objects one by one in the textbox named “Object Name”, verifying it with the “verify” button, and then pressing the “Add” button. Because this procedure can be cumbersome for large datasets generated from external sources such as microarrays, the user also has the option to import a simple text-file consisting of the object name and an object ID (if any) separated by a tab. The ID field is only necessary in case the object name is ambiguous – IRIDESCENT will still import

the file if no ID is given, but in this event will not warn the user of any potential ambiguity but rather take the first entry with the ambiguous name. The user then accesses the “File” menu option in Figure 12 and chooses “Import Objects into dataset”. A file browser window will be opened to select the file. Upon opening the file, the objects will be imported into the window (Figure 15a).

a



b

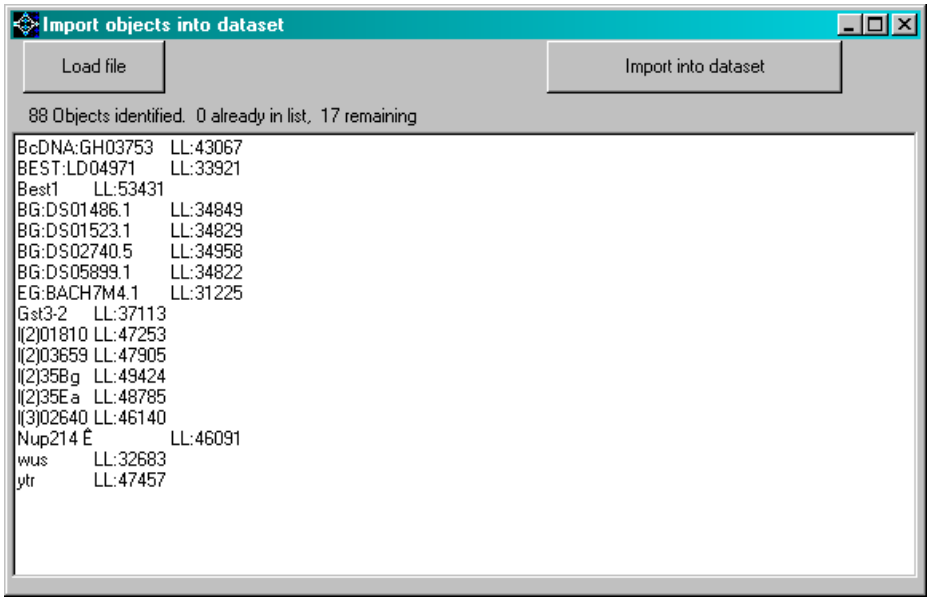
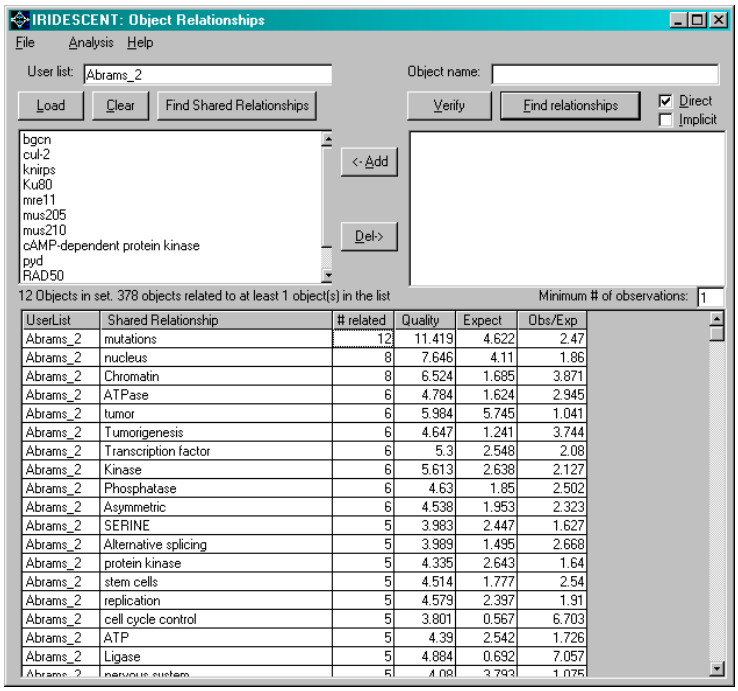


Figure 15: Semi-automated matching of entries from a text file to ORD entries. **a)** A user list is imported consisting of gene names and database identifiers. **b)** After "Import into dataset" is chosen, IRIDESCENT attempt to match each gene and returns a list of genes that did not match any database entry.

After opening the object file, when the user hits the "Import into dataset" button, IRIDESCENT will then attempt database matches for each of the objects and remove them from the import list if a match is found (Figure 15b). At this point, the user can manually

refine the list and hit “import” to try matching again, or can discard the genes from analysis. For example, the entries beginning in “BG:” represent genes without known functions and thus would not be very useful in a literature analysis. The user continues to iterate until all objects are imported or deleted. Then, the user has created a dataset named after the file opened (and should thus name input files by some convention the user is likely to remember). Members in this dataset can be added and deleted at the entry screen in Figure 12. When satisfied, the user clicks the button “Find Shared Relationships” and IRIDESCENT analyzes the objects for shared relationships and scores the sets of shared relationships. Figure 16 shows an example of an analyzed list.

a



b

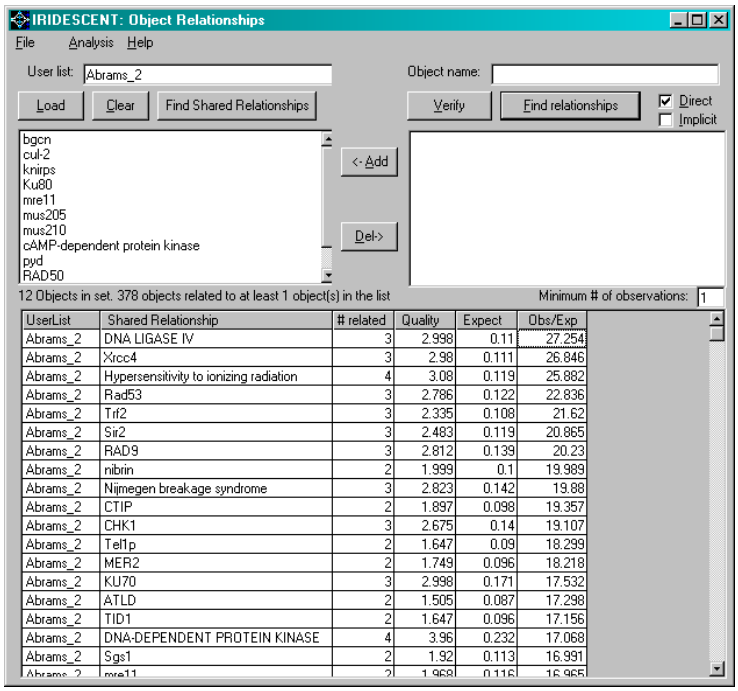


Figure 16: A microarray dataset related to apoptotic stimuli is analyzed using IRIDESCENT. **a)** The list is sorted by the number of objects in the user list that are related to the object in the "Shared Relationship" column, showing the most common relationships. **b)** The list is sorted by the Obs/Exp ratio, showing the most exceptional relationships.

In Figure 16, objects related to 2 or more members of the list are displayed in the column “Shared Relationship”. The number of objects in the list that are connected to them is shown along with the Obs/Exp score, ranking the statistical exceptionality of the match. When the list is sorted by the most matching relationships, this gives the user an idea of what general commonalities the objects share. This list, for example, was a study of the transcriptional response associated with the exposure of cells to ionizing radiation (courtesy of John Abrams). Going from top to bottom and examining the literature associated with the shared relationships in Figure 16a gives the user an idea of the nature of the associations (shared relationships underlined). Most of these genes are found in the nucleus, and are involved in repairing mutations to DNA. As part of the cell cycle control process, they are able to remodel chromatin structure and conduct DNA replication and repair. Defects in several of these products have been shown to lead to tumorigenesis.

To identify some of the more exceptional relationships, the user can sort by the Obs/Exp ratio (Figure 16b). Here, we see a key phenotype associated with 4 of the genes on this list “Hypersensitivity to ionizing radiation”, along with a number of key DNA repair enzymes associated with the response to ionizing radiation: XRCC4, Rad53 and DNA Ligase IV. To the user, this suggests that these genes, if not on the microarray, should be added or at least examined for their role in the response if not already known.

3.7.3 Array Analysis

Analyzing members of a dataset in relation to each of the other members is no different conceptually than the Direct/Implicit relationship analysis just presented. The layout, however, enables a user to rapidly identify which relationships between the genes in the dataset are known and which may share a large number of implicit relationships with other members of the set. A sparsely populated matrix indicates little is known about the relationships between the objects in the dataset, which suggests that either novel relationships are being identified or the objects simply do not have much in common. An example of an array analysis is shown in Figure 17.

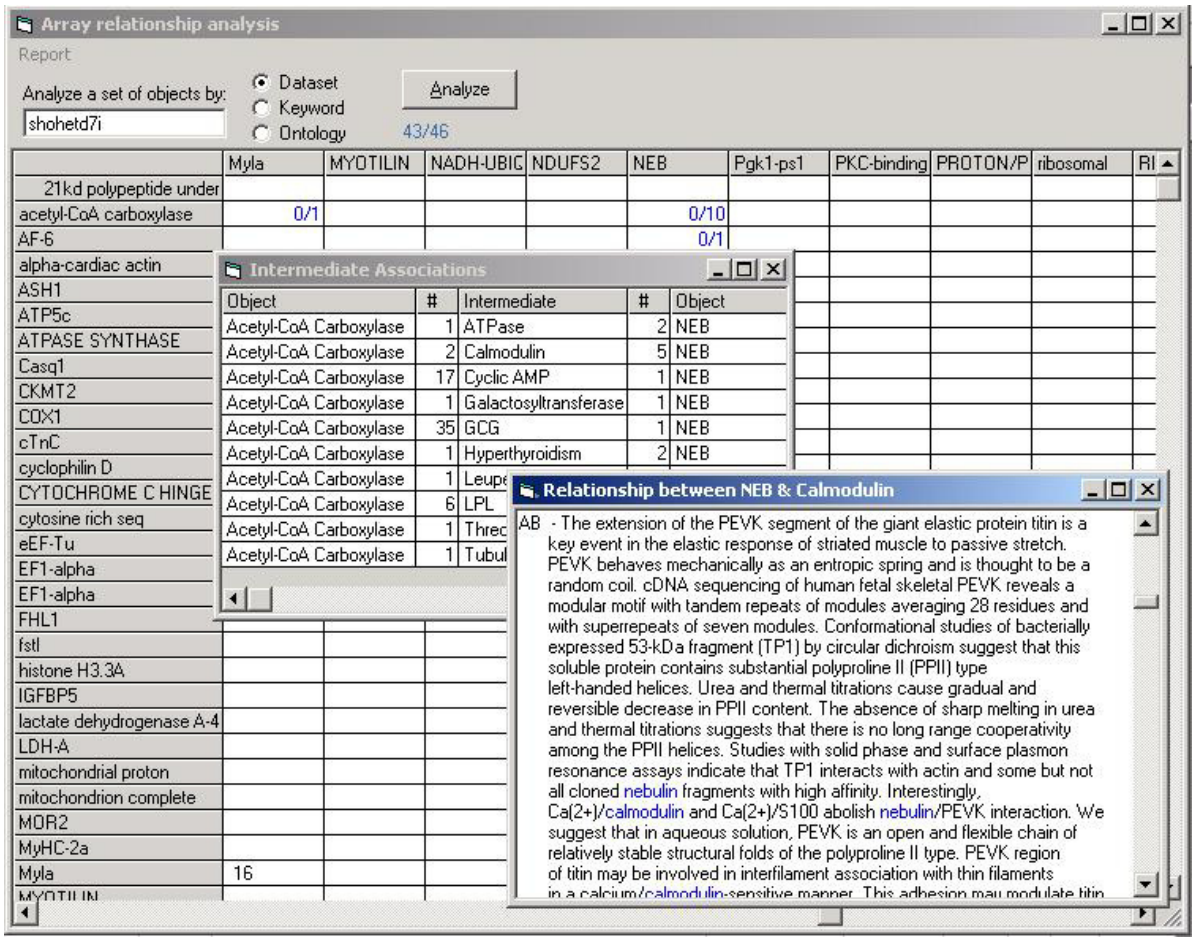


Figure 17: A dataset involving a study on the transcriptional effects of isoproterenol on cardiac myocytes after 7 days is shown. On top and on the far left are the names of the genes being analyzed. The first inset (left) shows the shared relationships in the implicit relationship between Acetyl-CoA Carboxylase (ACC) and Nebulin. Double-clicking on the entry between nebulin and calmodulin expands to show the original source of the relationship within MEDLINE and allows the user to ascertain the nature of the relationship. To the right of the dataset textbox is the option to compare objects within a dataset, sharing a common keyword (e.g. “kinase”), or within an ontology (e.g. “microtubule binding”).

3.7.4 Scanning Text

The Text Scan screen is where all tasks related to the input of textual data are conducted. Shown in Figure 18 is an example of a set of abstracts processed by IRIDESCENT. Words or phrases recognized by IRIDESCENT are colored.

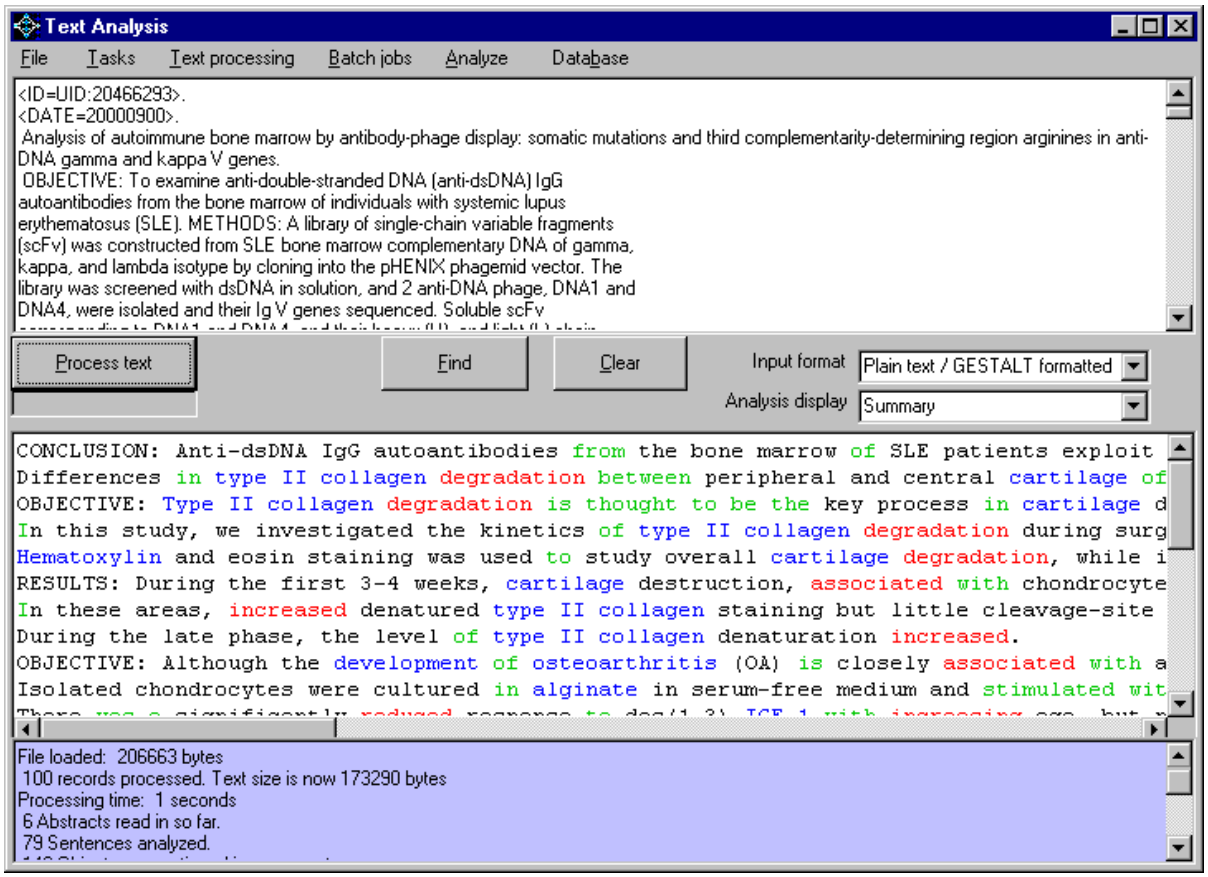
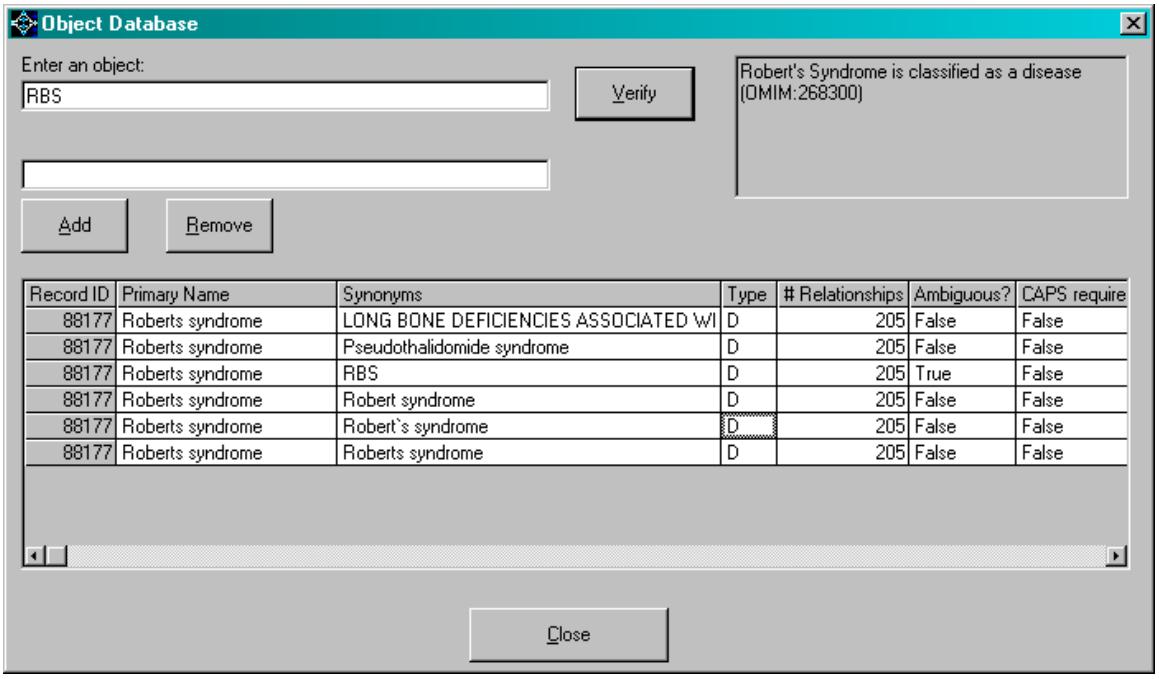


Figure 18: Recognition of named entities within input text using IRIDESCENT. In the top window, an abstract is formatted to display fields relevant to IRIDESCENT: Unique ID (UID), date of publication, title and abstract. The bottom window shows what IRIDESCENT recognizes within text. Blue words/phrases are objects recognized by the system. In red font are words that denote the nature of a relationship (e.g. “increased”), while green font denotes phrases that link objects together (e.g. “of”, “in”).

All of MEDLINE is processed under the “Batch Jobs” option, which turns off visual feedback so the system may process records faster. Also included under this menu are the options to identify acronyms within MEDLINE records (construct the ARGH database) and to identify words found within MEDLINE (construct the MEDLINE_Words database). Any text may be analyzed here as free-form input.

3.7.5 Object Screen

In the Object Screen (shown in Figure 19), users may manipulate the objects within IRIDESCENT's database by adding or deleting synonyms, as well as creating new objects to be recognized within text or permanently deleting an object from the recognition database. Here, users can also delete relationships identified by IRIDESCENT (e.g. ones deemed to be "uninteresting" or in error) as well as add new relationships.



Record ID	Primary Name	Synonyms	Type	# Relationships	Ambiguous?	CAPS require
88177	Roberts syndrome	LONG BONE DEFICIENCIES ASSOCIATED WI	D	205	False	False
88177	Roberts syndrome	Pseudothalidomide syndrome	D	205	False	False
88177	Roberts syndrome	RBS	D	205	True	False
88177	Roberts syndrome	Robert syndrome	D	205	False	False
88177	Roberts syndrome	Robert`s syndrome	D	205	False	False
88177	Roberts syndrome	Roberts syndrome	D	205	False	False

Figure 19: The Object Manipulation Screen. Here, objects can be examined within the ORD and synonyms created and deleted. The term "RBS" maps to the disease "Roberts Syndrome". When a field is highlighted by single-clicking, information regarding the field is displayed in the upper right text box. RBS has a number of synonyms associated with it, and the column "Ambiguous?" indicates whether or not the term is an acronym with multiple definitions, requiring resolution within MEDLINE. Some entries are in all upper-case letters such as the first entry on this list. These objects typically come from OMIM, which formats its entries in such a manner. IRIDESCENT is not case sensitive unless the rightmost box "CAPS required" is set to "true".

3.7.6 Database Screen

The Database Screen provides summary statistics for the ORD as well as a venue to perform all database-centric tasks. Figure 20 shows the statistics for the most recent build of ORD (12/1/02). A total of 124,120 unique objects are within the ORD and are associated with 323,500 total terms representing primary names and synonyms. There are 742 user-entered words/phrases that are deleted from this database, and 1,863 that are added (the real number is actually much smaller – the user added object database originally doubled as the dataset analysis database where object sets were entered. This was changed, but entries were not deleted in the interests of preserving recognition, and many of these entries were merely duplicates of objects already in the ORD). Also included are 1,025 chromosomal loci recognized by the system, 13,414 ontology entries provided by GO, 219 meta-relationships (discussed later and shown in red in Figure 18) and 86 linker entries (shown in green in Figure 18). Finally, the SORD database is functionally separated from the ORD database and represents a method of preserving the recognition build of each IRIDESCENT run. Within the SORD are all the recognized relationships as well as the objects known at the time of the run. As new objects are added, new IRIDESCENT runs are necessary. But there is a need to preserve what was done and revert to previous versions if necessary. Thus, the separation of the ORD database, which is constantly refined and the SORD database, which represents previous runs of IRIDESCENT. The version shown here is the first version run in 2001 with 3,444,326 relationships recognized. A total of 1,123 user dataset entries are also in the database.

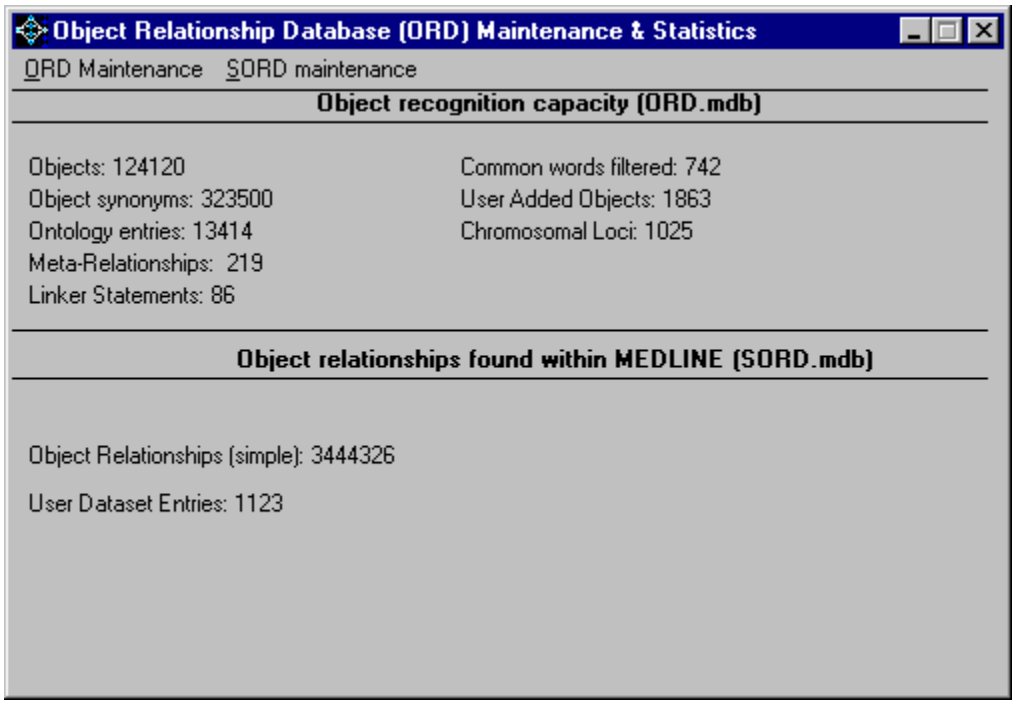


Figure 20: IRIDESCENT's database screen. Here, the user can perform database maintenance tasks as well as gather statistics on all databases associated with IRIDESCENT.

Chapter 4

IRIDESCENT: Completed Work, Analyses and Results

The first version of IRIDESCENT processed 12,037,763 MEDLINE records from 1967 to January 2002, creating a network of 3,482,204 unique relationships between objects. Approximately 2/3 of the objects in the database found exact literal matches within the literature, identifying at least one relationship for 22,482 of the 33,539 unique objects (85,234 total terms when including synonyms) within the database. Most of the experiments and analyses discussed hereafter will have been conducted using this version.

4.1 Evaluating MEDLINE records as a source of knowledge and database entries as a basis for object identification

Recall rates for IRIDESCENT were estimated from a set of review articles. Four objects were randomly chosen from the collective object database, representing one of each object type, with the stipulation that at least 2 review articles had been written about the object within the past 3 years. A set of 2-3 review articles was then selected, and a list of all other objects mentioned therein having any non-trivial relationship to the original query object was compiled. Only objects of the same type as those in the central database were counted (i.e. genes, diseases, phenotypes and small molecules). Review articles were selected for CTLA-4 (gene)[73-75], Fragile-X Syndrome (disease)[76-78], cachexia (clinical phenotype)[79-81], and dynorphin (small molecule)[82,83]. The list from each set of reviews

was then compared to the relationships identified by IRIDESCENT after processing all of MEDLINE.

As Table 13 shows, objects contained within the collective database represent an estimated 78% (141/181) of the total number of objects of their type found within review articles. Of the 40 objects mentioned in the literature but not found in the database, 2 were diseases, 9 phenotypes, 7 genes, and 22 small molecules. The 2 disease names (Graves' Opthalamopathy and Relapsing-remitting Experimental Autoimmune Encephalomyelitis) and 9 phenotypes were simply not mentioned in OMIM. Three of these phenotypes, however, were simply the result of a semantic difference between the OMIM entry and the article ("rocking" versus "body-rocking", "greater interocular distance" versus "increased interocular distance", "fetal akinesia" versus "akinesia"). The most problematic category was small molecules, for which many chemicals and drugs widely mentioned in the literature (e.g. DAMGO, DADLE, isoprenaline) were simply not found in the MeSH trees database.

Of the 141 database objects cited in the reviews as being related to one of the central query objects, 17 were not mentioned within any MEDLINE title or abstract related to the query object. Of these, 9 were not found because of spelling/phrasing differences between the database entry and the literature, 1 was missed because it was flagged by IRIDESCENT as an ambiguous acronym and not defined in the abstract (PKI), and 1 was given as a gene family name (NFAT) in the review while only the specific family members were listed in MEDLINE abstracts. The remaining 6 objects represented relationships not mentioned in the titles/abstracts of the articles. And of these six, 3 were discussed in the review in the context of a closely related (implicit) phenomenon. Of these 138 relevant relationships mentioned in

MEDLINE titles and abstracts IRIDESCENT identified 127 of them, giving it a recall rate of 92% in terms of identifying the conceptual occurrence of database objects within textual input. In terms of identifying informative relationships between object types within MEDLINE, the assimilated databases provide IRIDESCENT with the potential to recognize an estimated 78% (141/181) of relevant relationships. Overall, in terms of identifying relevant relationships within a domain, IRIDESCENT has an estimated recall rate of 70% (127/181).

Some of IRIDESCENT's FN failures to identify objects within text were systematic (e.g. the MeSH entry 5,8,11,14,17-Eicosapentaenoic Acid is almost always referred to in MEDLINE simply as eicosapentaenoic acid) while other failures varied in their rates (e.g. JNK was found to be spelled 81 different ways including "c-Jun N-terminal kinase" 605 times, "c-Jun NH2-terminal kinase" 154 times and "c-Jun amino-terminal kinase" 62 times[84]).

Name (# of reviews)	**Category**	**# of Literature references[a]**	**Total rels. in review**	**Total rels. found in DB[b]**	**Object in DB, but no rel. found**	**Object in review but not in DB**
CTLA-4 (3)	Gene	1,191	44	37	2	5
Dynorphin (2)	Molecule	2,647	40	23	4	13
Fragile-X (3)	Disease	2,141	35	22	6	7
Cachexia (3)	Phenotype	2,933	62	42	5	15
TOTAL			**181**	**124**	**17**	**40**

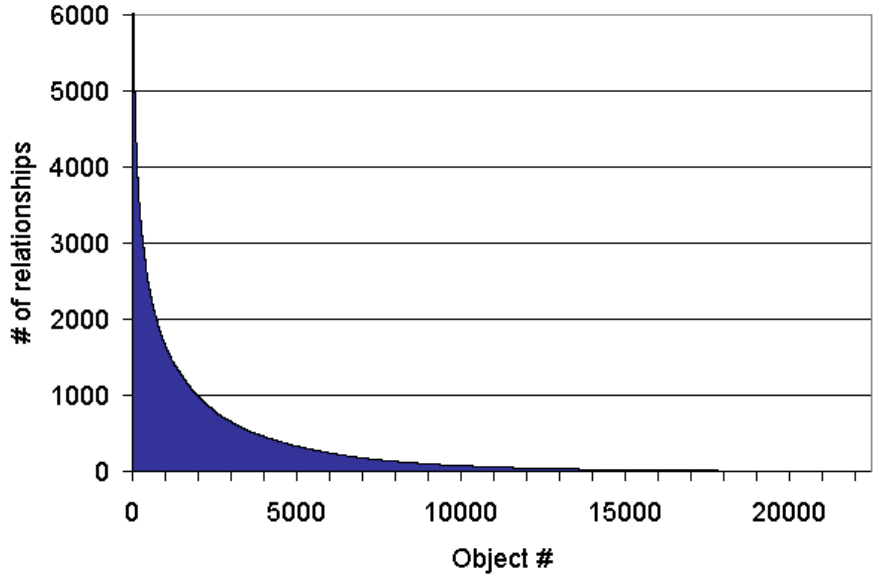
Table 13: Database objects used by IRIDESCENT to identify relevant relationships within MEDLINE records are compared to the relevant relationships between objects that are given within review articles. [a]As of 1/23/02. [b]DB=IRIDESCENT's identified relationship database. This analysis was conducted after all MEDLINE records were processed.

4.2 Developing a scoring mechanism based upon the statistical properties of relationships in a network

The number of relationships identified per object followed an exponentially decreasing distribution (Figure 21a), indicating a highly disproportionate distribution of object terms within the literature. Sodium was the most abundantly mentioned object, and found at least once in the same abstract with 8,868 other objects (~40% of all objects identified). Using this network of relationships, we plotted the number of direct connections for each object versus the number of purely indirect (implicit) connections associated with it (Figure 21b). Note that as the number of direct relationships increases, the number of implicit relationships rapidly approaches the theoretical maximum, which is the total number of nodes in the network. Even objects with relatively few direct relationships can still be implicitly related to the *vast majority* of objects in the network. While this high degree of implicit connectivity may in part be due to some objects being associated with extremely abundant terms, such as sodium, this demonstrates that the mere fact that two things are implicitly related is trivial. Two objects may share sodium as a relationship, but this would not likely be informative because many objects share sodium as a relationship and one could not conclude very much from this alone. However, if two objects shared a relationship with a gene related to only a few other objects, this would be more informative to a user since the relationship is likely to be within a much more specific context. The fundamental challenge in identifying novel relationships with the potential to be of value to scientific research is being able to assign some measure of potential relevancy to each of these numerous implicit

relationships, as well as ascertain how exceptional shared relationships are within the context of the network connectivity properties (Figure 21a).

a



b

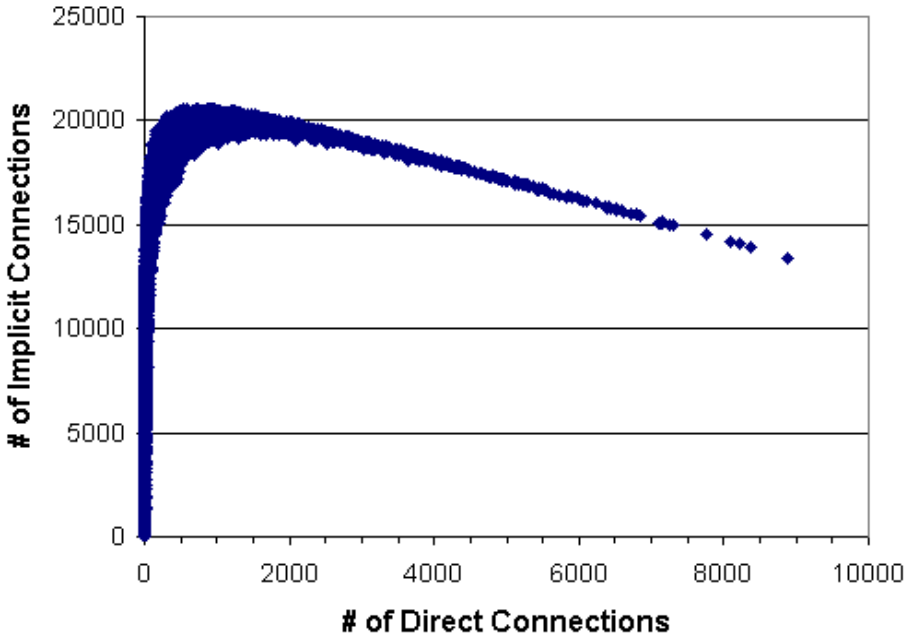


Figure 21: a) Distribution of the number of relationships each object in the database has. A relatively small fraction of the objects in the database are directly related to a large percentage of the total, contributing to a rapid explosion in the number of implicitly related objects **(b)**. Most objects are either directly or implicitly related to the majority of other objects in the database - highlighting the need for a method to score implicit connections for their potential relevance.

For direct relationships (e.g. a list of genes related to a disease), it is relatively straightforward to assign strength scores to each relationship based upon estimated error rates and frequency of co-occurrence. Since terms that co-occur more frequently are more likely to represent valid relationships[34], we assign object relationships a score based off of the number and type (i.e. abstract or sentence) of co-mentions observed and their corresponding error rates. The probability a relationship between A and B is an error is represented as a function of the number of times, *n*, the two objects are co-mentioned and the random error rate, *r*, associated with the co-mention metric used to establish the relationship, written as:

$$P(err) = r^n \quad (1)$$

Thus, the probability it is valid can be written as:

$$P(valid) = 1 - r^n \quad (2)$$

The strength of a relationship can be seen as a function of the number of times it has been observed and the collective probability of each observation being an error. Since we calculate two different relationship metrics, number of sentence co-mentions (C_s), and number of abstract co-mentions (C_a), we assign an overall strength of association score (S) based upon their individual error rates, r_s (17% FP) and r_a (42% FP) respectively, by the formula:

$$S = C_s*(1-r_s) + C_a*(1-r_a) \quad (3)$$

For example, if two objects are co-mentioned 5 times within a sentence and 10 times within an abstract, the strength score would be 5*0.83 + 10*0.42 = 8.35. For implicit relationships, it is not yet clear what statistical parameters correlate with the probability of it representing a valid relationship, but we can surmise that the probability of an implicit relationship (A-B-C) being valid would not be greater than the least probable of the two individual (A-B or B-C) relationships linking them. We can thus estimate that $P(A \leftrightarrow C) \leq P(A \leftrightarrow B)*P(B \leftrightarrow C)$, where the symbol ↔ is defined as the existence of a non-directional relationship between two objects.

It is important to provide a control for sets of relationships and implicit relationships, to ascertain whether or not such a grouping of objects is meaningful. It is somewhat difficult to prove that if a common object such as “cancer” consistently shows up as a strongly

implicit relationship (i.e. many shared intermediates) within most analyses, that it is not meaningful. We can, however, assign a measure of exceptionality to the relationship based upon the total number of relationships each object has within the network.

Assuming a number of objects were randomly connected in a network with the same connectivity shown in Figure 21a, we can calculate the odds that any two objects would be implicitly related and how many intermediate relationships we would expect them to share (Note: M. Huebschman helped provide some of the mathematical rigor that follows to better explain the formula I derived). To evaluate the relationships shared by a set of objects within a network of relationships, an expectation value for the number of connections was developed based upon the relative connectivity of each object involved. In this network, objects can be seen as nodes and adirectional relationships (represented by the symbol $\leftrightarrow$) as connections. Let $P(A(K_A))$ represent the probability that node A, having K_A connections in the network, is randomly connected to any one single node. The probability of A not being connected to that one node is then $1- P(A(K_A))$. Let all node connections be independent of all other nodes.

Consider two nodes A with K_A connections and B with K_B connections in a network of nodes, N_t+1. The probability that A connects to B and/or B connects to A, $P(A\leftrightarrow B)$, can be equated to picking marbles our of a jar. Let the colors of the marbles represent the different nodes in the network; there are N_t colors. Now assume there are two identical jars of marbles. The total probability is the probability of picking, say, a red colored marble out of the one jar on K_A tries and/or picking a red marble out of the other jar with K_B tries. That is, if we have at least one red marble out of the two jar we have a connection between A and

B, but we also know from our node relationships that if there is a relationship between A to B, it is adirectional and, thus, there is a relationship B to A. To meet our network design of only one connection between nodes, we add the restriction that when a marble is picked it is not replaced in the jar.

After picking marbles out of the jars there are four possible outcomes. 1.) There are two red marbles (A connected to B and B connected to A). 2) There is one red marble and it came from jar A (A connected to B but B not connected to A). 3.) The reverse of outcome 2), one red marble but it came from jar B (A not connected to B but B connected to A). 4.) There are no red marbles (A not connected to B and B not connected to A).

$$P(1) + P(2) + P(3) + P(4) = 1 \tag{4}$$

In our data network, as noted above, our probability is

$$P(A \leftrightarrow B) = P(1) + P(2) + P(3) \tag{5}$$

or simply

$$P(A \leftrightarrow B) = 1 - P(4) \tag{6}$$

$P(A(K_A))$ is equal to the probability of picking a red out of N_t colors on the first try; plus picking a red on the second try with one less color present times not picking red on the first try; plus picking a red on the third try with two less colors available times not picking a red on the first try times not picking a red on the second try; etc. on down to the K_A^{th} try. The probability of picking red on first try is $1/N_t$. The probability of picking red on the second try assuming it did not occur on the first is $1/(N_t - 1)$. The probability that it was not picked on the first try is $(1- 1/N_t)$, etc. Then the probability is

$$P[A(K_A)] = \frac{1}{N_t} + \frac{1}{N_t - 1}\left(1 - \frac{1}{N_t}\right) + \frac{1}{N_t - 2}\left(1 - \frac{1}{N_t - 1}\right)\left(1 - \frac{1}{N_t}\right) + \cdots + \frac{1}{N_t - K_A + 1}\left(1 - \frac{1}{N_t - K_A + 2}\right)\cdots\left(1 - \frac{1}{N_t}\right) = \frac{K_A}{N_t} \tag{7}$$

The probability of not picking a red on K_A tries is $1\text{-}P(A(K_A)) = 1\text{-}K_A/N_t$.

Likewise the probability for B to select A:

$$P[B(K_B)] = \frac{K_A}{N_t} \tag{8}$$

and the probability for B not to select A is $1\text{-}K_B/N_t$. The probability of not picking a red marble from either jar is:

$$P[4] = \left(1 - \frac{K_A}{N_t}\right)\left(1 - \frac{K_B}{N_t}\right) \tag{9}$$

Substituting the result of Equation (9) into Equation (6), our total probability $P(A \leftrightarrow B)$ is:

$$P[A \leftrightarrow B] = 1 - \left(1 - \frac{K_A}{N_t}\right)\left(1 - \frac{K_B}{N_t}\right) \tag{10}$$

A plot of this probability for K_A and K_B is shown in Figure 22, which is valid for all non-zero values of K_A and K_B. Intuitively, if either K_A or K_B were equal to N_t, we would expect that $P(A \leftrightarrow B) = 1$, since if one or the other node has a connection to all other nodes (N_t), then the connection between the nodes is certain.

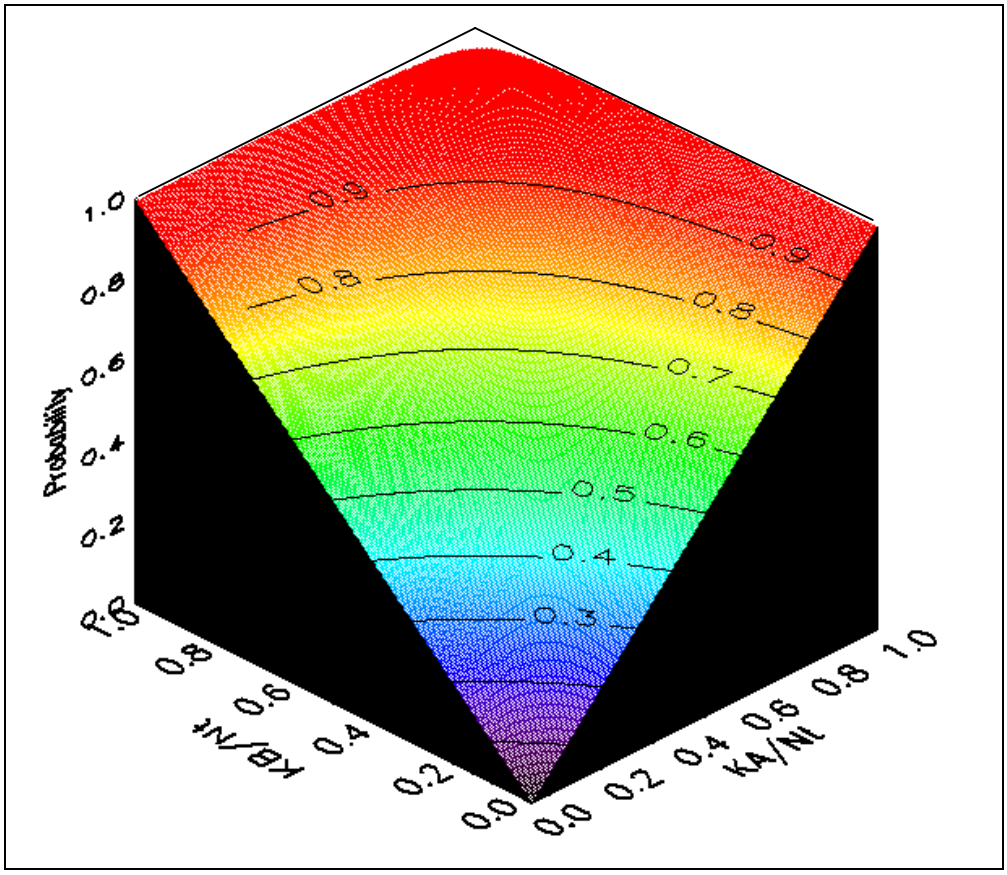


Figure 22: Probability of P(A↔B) for all K_A and K_B for a network. Contour lines of constant probability are shown. Color shading represents black = 0, rainbow color order = mid-levels and white = 1.

The ability of Equation 10 to predict the probability of two objects being associated, assuming a randomly connected network, was confirmed by assigning a random number of relationships (1 to 10,000) to two objects within a 10,000 node network and determining whether or not one of those relationships connected the two objects. When this was done for 10,000 iterations and compared with the expected number of relationships, the observed/expected ratio converged to 1.0 as the set size increased, demonstrating that in this limit the equation accurately predicted behavior in this type of network. This was then repeated for IRIDESCENT's literature-derived network, randomly picking two objects, each

having at least 1 relationship within the network, 10,000 times, and the ratio of observed to expected relationships was determined to be 0.40. A ratio less than 1 is consistent with a network whose connectivity is not random.

Additionally, Equation 10 can be looked at from the standpoint of probable membership in a set. The probability B will lie in the domain of A (written as $P(B \in A)$) is K_A/N_t and the probability A will be in the domain of B (written as $P(A \in B)$), is K_B/N_t, each of which is independent of the other. Because the formula $P(A \in B)$ OR $P(B \in A)$ cannot be as easily represented as the probability B is not related to A and vice versa, written as NOT ($P(A \notin B)$ AND $P(B \notin A)$). This formulation also converts mathematically into Equation 10.

The probability of making various connections from A, with K_A connections in the network, to different nodes in the network based solely on that nodes connections in the network, is the sum of each various connection probability. Consider a group of n nodes, $\{G\}=\{B_1, B_2, B_3,\ldots B_n\}$. The probability that A will connect to all the nodes in this group is

$$P[A \leftrightarrow \{G\}] = \sum_{i=1}^{n} P[A(K_A) \leftrightarrow B_i(K_i)] \tag{11}$$

$$P[A \leftrightarrow B_n] = \sum_{i=1}^{n} 1 - \left(1 - \frac{K_A}{N_t}\right)\left(1 - \frac{K_{Bi}}{N_t}\right) \tag{12}$$

Equation 12 provides the probability of random connections in the group. It can also be thought of as the number of occurrences per unit node. For example, suppose the probability is 0.0015 that a node, A, with 10 connections to the network of 1001 nodes will connect to a node B_1 with 50 connections and B_2 with 100 connections. We would expect to

see only 1.5 such configurations in our network (Probability or Connections/Node)*N_t (nodes) = 0.0015*1000 = 1.5 Connections).

Equation 12 is the exact random probability for the criteria described. Given an actual literature-derived network of connections we can determine the confidence that the relationships of the network are not random. There are many ways to make use of Equation (10) in the network. We wish to first address how any gene is related to objects in a Gene Ontology (GO) category. A group of {B} objects in a GO category was selected. A group {E}={Expected Connections by A-object (gene)} is determined by Equation (10) for all the A-objects that connect to two or more members of {B}. Group {E} is the probability of all the different A(K_A) satisfying this criteria. Let group {M} be the number of actual A's in the literature network connecting with this criteria. We used 99 different sized groups from 2 to 100, $\{B_n\}$ where n is the number in the group, $\{B_2\}$, $\{B_3\}$, …$\{B_{100}\}$. For each size-group $\{B_n\}$, the objects were varied 100 times to give 100 samples for each group size. The sample groups become

$$\{\{B_2^1\}, \{B_2^2\}\ldots.\{B_2^{100}\}\}, \{\{B_3^1\}, \{B_3^2\},\ldots\{B_3^{100}\}\{,\ldots\{\ldots\{B_{100}^{100}\}\}.$$

Likewise the associated probability groups:

$$\{E_2^1, E_2^2\ldots.E_2^{100}\}, \{E_3^1, E_2^2\ldots E_2^{100}\}, \ldots\{E_{100}^1\ldots E_{100}^{100}\} \text{ and}$$

associated number of connections groups:

$$\{M_2^1, M_2^2\ldots.M_2^{100}\}, \{M_3^1, M_2^2\ldots M_2^{100}\}, \ldots\{M_{100}^1\ldots M_{100}^{100}\}.$$

For each of the size-groups, $\{B_n^{j=1 \text{ to } 100}\}$, a mean and standard deviation was determined for the 100 samples for both the probability $\{E_n^{j=1 \text{ to } 100}\}$ and counted number of connections, $\{M_n^{j=1 \text{ to } 100}\}$. Let X_n^E and $\sigma'_n{}^E$ be the mean and sample standard deviations,

respectively, for the probability samples; and X_n^M and $\sigma'_n{}^M$ be the mean and sample standard deviations, respectively, for the counted connections. Estimating the population standard error of the mean for each distribution is $\sigma_n^E = \sigma'_n{}^E/\sqrt{N} = \sigma'_n{}^E/\sqrt{100}$ or $\sigma_n^E = 0.1\ \sigma'_n{}^E$ and $\sigma_n^M = 0.1\sigma'_n{}^M$.

We now have population means and their standard errors for the number of connections any object, A, in the network has with sets of two or more objects, B, in the network. These are for random connections. Likewise we have the population means and their standard errors for the counted objects, A, in the network which were connected to the sets of objects, B. These are for the observed connections. We can compare the two sets of means to make a confidence estimate on whether our observed connections in the network are random or not. We name the counted mean as the Observed Value, X_n^M, and the random mean as the Expected Value, X_n^E. The ratio of Observed to Expected is the Obs/Exp score, X_n^M / X_n^E. The closer the ratio is to one, the more probable the observed value is random. We can thus assign a statistical confidence that the measured mean, X_n^M, is not random if the difference is greater than the expected random mean plus a confidence interval ($X_n^E \geq 2\sigma_n^E$). That is

$$\left| X_n^E - X_n^M \right| \geq 2\sigma_n^E \tag{13}$$

Since the area under a normal distribution from +2σ to +infinity has only 2.5% of the area as does the area from -2σ to –infinity, we have a 97.5% confidence level that the Observed mean, X_n^M, is not random using Equation 13.

However, in our first set of calculation we have used only the sample and not he population standard error in the means. Thus we have been more restrictive in stating our

confidence than required since $\sigma_n^E < \sigma'^E_n$. We will incorporate this factor in the future. Additionally, since we have sampled two populations for every size-set, in the future we could further increase confidence in our assessment of the relationships in the network by evaluating the differences

$$\left|X_n^E - (X_n^M - 2\sigma_n^M)\right| \geq 2\sigma_n^E \tag{14a}$$

$$\left|X_n^E - (X_n^M + 2\sigma_n^M)\right| \geq 2\sigma_n^E \tag{14b}$$

For our Observed values greater than the Expected values, satisfying Equation 14a would mean that the Observed values have a further reduced chance of being random than would be indicated by Equation 13. Observed values satisfying Equation 14b would also be considered not random but with less confidence than Equation 13 indicates. We will be refining the confidence assessments with Equations 14a and 14b in future analysis.

To establish that Equation 12 aids in quantitatively evaluating relevant groupings, sets of objects created at random from our database were compared with sets of objects expected to share common elements, which were obtained by using genes within specifically defined ontological categories from the Genome Ontology database[46]. Using Equation 12 to calculate an average observed to expected ratio for the 10 most frequently shared relationships between objects, we find that this ratio is consistently higher for the topical set than for that of the random set (Figure 23).

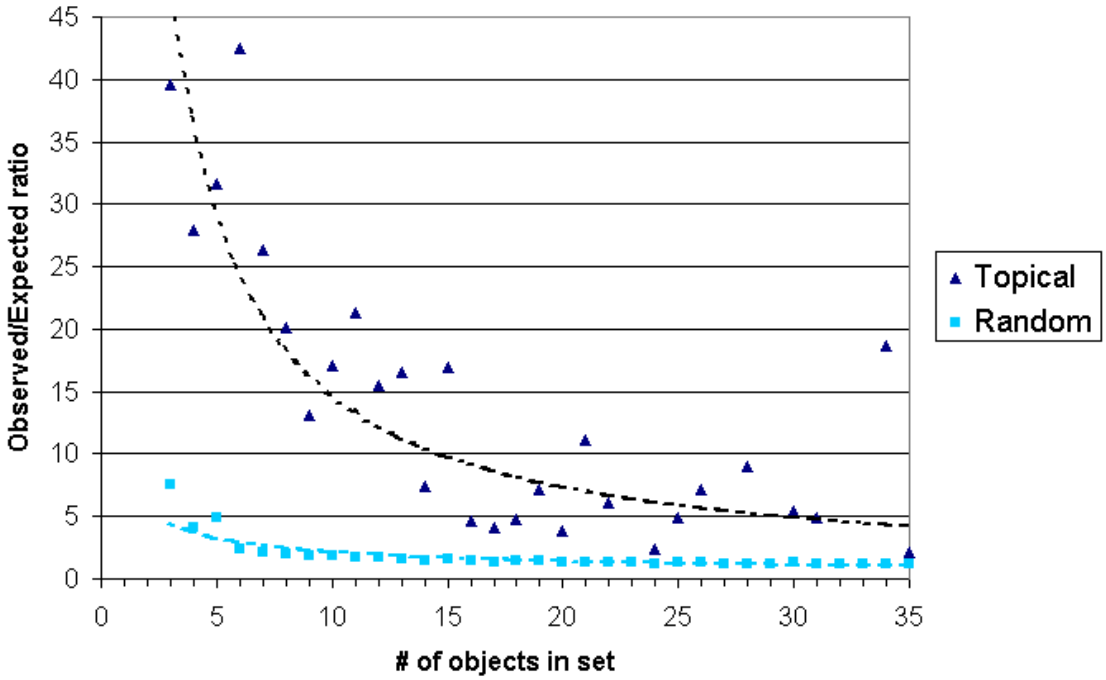


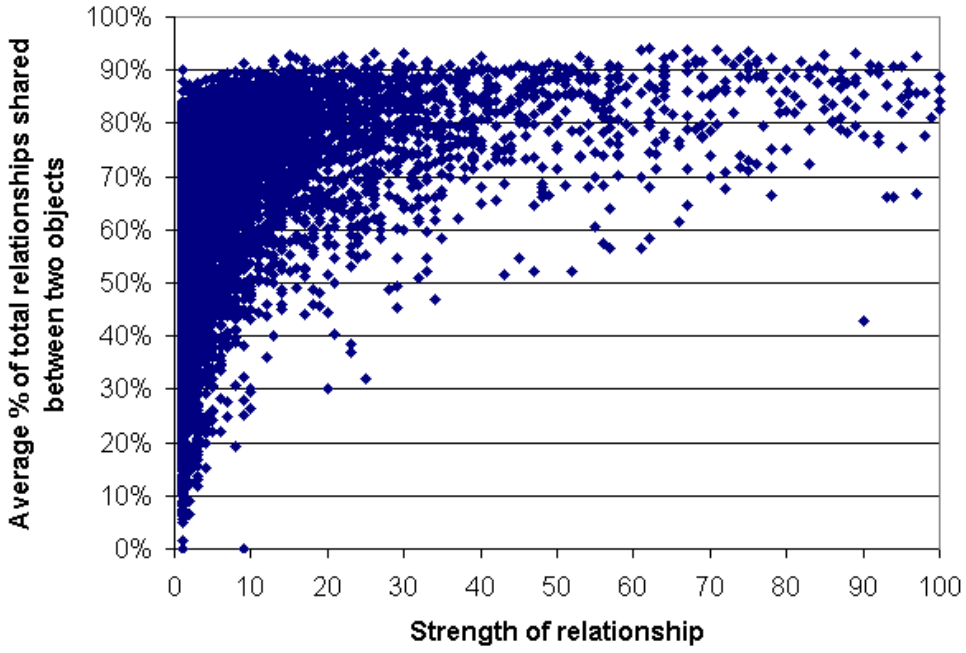
Figure 23: Comparison of the average observed to expected ratio for the 10 most highly related objects between random and topical sets. Sets of objects, ranging in size, were either generated randomly or obtained from classifications within the Genome Ontology database. The sets were analyzed to identify other objects related to members of the set and the observed to expected ratio was calculated and averaged for the 10 objects related to the most members of the set. These averages were again averaged by the size of the set and graphed above (n=10 for random sets, while n varies for the topical sets but is at least 5).

4.3 Estimating the relatedness of two objects by virtue of their shared relationships

We can use Equation 12 to estimate how exceptional an implicit relationship is, given the relative abundance of each of the two objects within the network. When evaluating implicit relationships, we are interested in how likely a specific relationship between A and C is to be of relevance. Generally speaking, relevance is highly subjective. That is, a relationship may indeed exist between A and C, but how important such a relationship is with regards to an area of research interest will be dependant upon the examiner. Thus, it is

difficult to assign an objective measure of relevance, but by evaluating the quantitative statistical properties of relationships known to be relevant, we can then compare them with the same properties of objects we suspect to be implicitly related. Among a number of properties, we find that the greater the strength of the relationship between two objects (as determined by Equation 3), the more relationships they tend to share in common (Figure 24a) and the stronger these shared relationships tend to be (Figure 24b). Therefore, the more relationships two objects share and the stronger those shared relationships are, the more likely the two objects are related. We can provide a quantitative estimate of how related two objects are by calculating what percentage of their total relationships overlap.

(a)



(b)

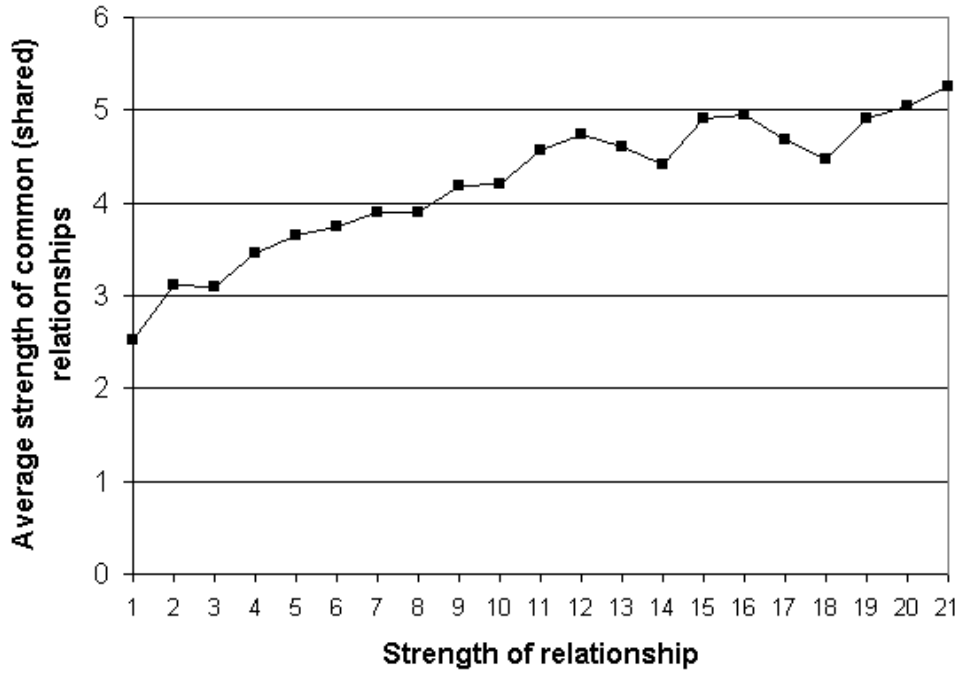


Figure 24: Statistical properties of related objects that are correlated with the strength of relationship. 20,000 related objects were randomly chosen from the relationship database and analyzed for the average percentage of the total known relationships they shared (**a**) and the average strength of their shared relationships (**b**).

Given these trends, we reasoned that two objects that share more relationships than would be expected by chance should have a greater probability of being related themselves. To establish this, we evaluated the observed to expected ratio for all objects related directly and indirectly to a central query object, plotting the strength of the relationship, if it was known, on the *y* axis. For the object "Cardiac Hypertrophy", we see that the higher the observed to expected ratio (Obs/Exp), the more likely the relationship is known (Figure 25). Furthermore, we note that the higher the Obs/Exp ratio, the stronger the relationship tends to be.

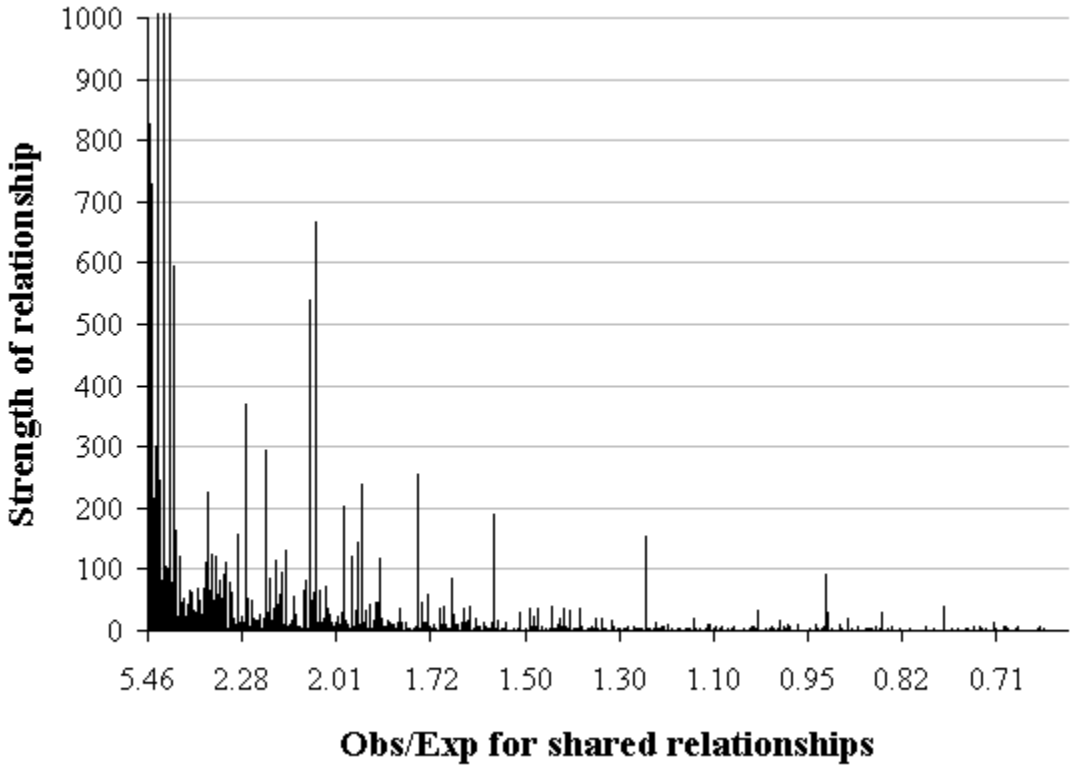


Figure 25: Objects were analyzed for their implicit "relatedness" to cardiac hypertrophy solely on the basis of the relationships they shared (Obs/Exp). If a relationship in MEDLINE has been established, its strength (based upon frequency of co-occurrence within MEDLINE) is plotted on the *y*-axis, otherwise it will appear as a gap (meaning no relationship has been established). Shown is a subset of 4,887 objects sharing at least 100 relationships with cardiac hypertrophy, sorted by their calculated observed to expected ratio. Due to *x*-axis compression, not all gaps will be visible on this graph.

To confirm that the trend observed in Figure 25 is not specific to the analysis of cardiac hypertrophy, but rather a general trend, we randomly picked 100 objects from the database that had between 500 and 1000 relationships within the network (this range was chosen simply to ensure that the approximate scale of analysis for each object was similar). Implicit relationships were identified for these objects and ranked by their Obs/Exp values. The top 1000 Obs/Exp scores were taken for each analysis and ranked from 1 (highest

Obs/Exp) to 1000 (lowest), and a normalized strength score calculated for each object analyzed, ranging from 1.0 (strongest direct relationship observed) to 0.0 (no relationship observed). Figure 26 shows this average strength plotted against the Obs/Exp rankings, showing that this is a general trend.

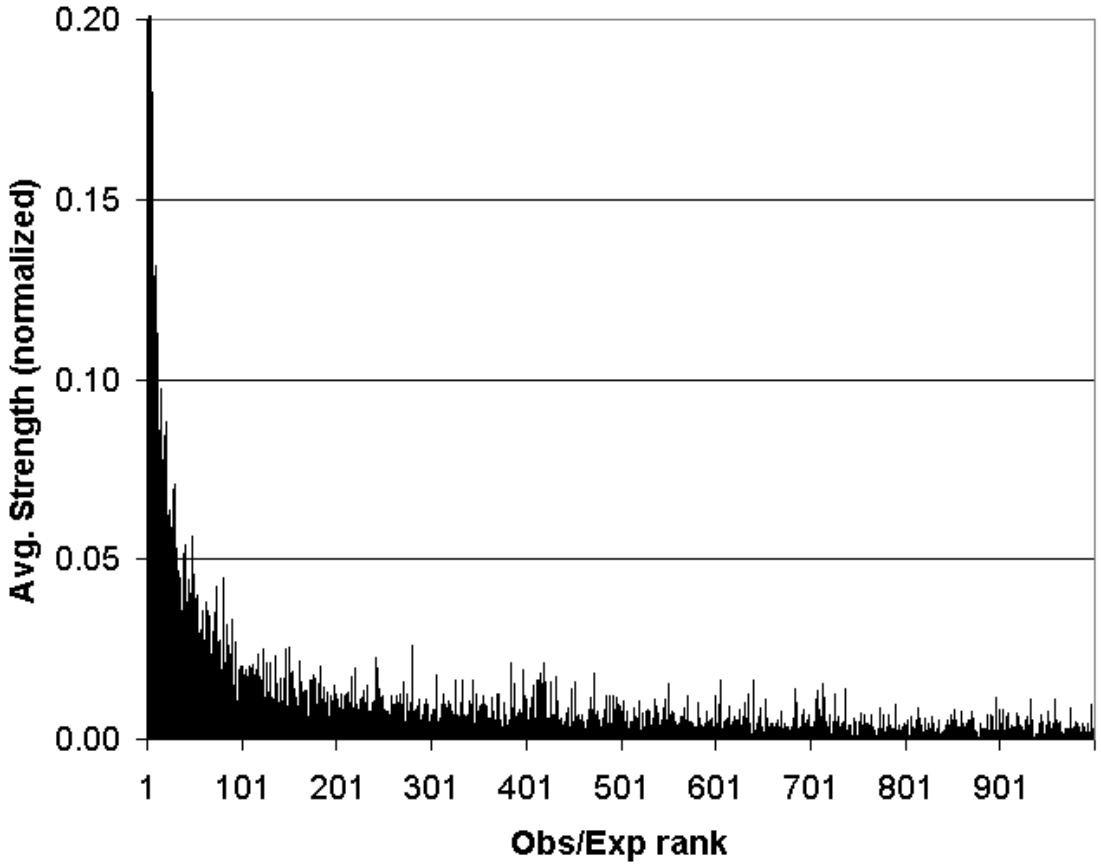


Figure 26: The observed to expected ratio obtained from identifying and analyzing shared relationships correlates with the existence and known strength of a relationship. This enables novel (implicit) relationships to be correlated with the probability that it is relevant (as judged by existing relationships) and important (as judged by the strength/frequency of historical reporting).

In some ways, the correlation of exceptional groupings with known relationships is not too surprising, as we would expect that two objects with very similar purposes, function, or involvement in a biological process should interact and be studied with many of the same objects. This does, however, establish that the relatedness of two objects can be correlated with the relationships they share and provides us with a means to quantitatively evaluate implicit relationships. This numeric evaluation should enable us to identify new relationships, not found within MEDLINE records, that are more likely to be logically plausible and relevant to the query object because of the relationships they share.

4.4 Using relationship strength in analysis

There are several different ways that shared relationships can be evaluated based upon their strength (i.e. frequency and type of co-occurrence). Typical object relationships within MEDLINE observed so far follow an exponentially decreasing distribution in strength, as illustrated in Figure 27. In part, the leftmost portion of the curve can be attributed to the continual reiteration and refinement of the earliest known relationships while other, newer, relationships will have been studied less and appear on the rightmost tail of the curve. Also within this rightmost portion, unfortunately, are the majority of random errors. This is because strength is a function of frequency and frequency is inversely related to the probability a co-mention does not reflect a non-trivial relationship. That is, the more times two objects are co-mentioned together and the closer together these co-mentions tend to occur, the more likely this co-mentioning reflects a non-trivial relationship. Confidence in the relative strength and importance of a relationship can be estimated from this frequency of co-

occurrence. It could be argued that this frequency reflects how well or how long a relationship has been established rather than how strong the relationship is. In part, this will be true because the longer a relationship has been known, the more opportunities researchers have had to study various aspects of it. However, given that the purpose of an abstract is to summarize the important findings of a research report and is limited in the number of words that one is allowed in writing it (usually 250 words or less), it does not seem reasonable to suppose that most authors would mention something in the abstract if it were not at all relevant to the work done.

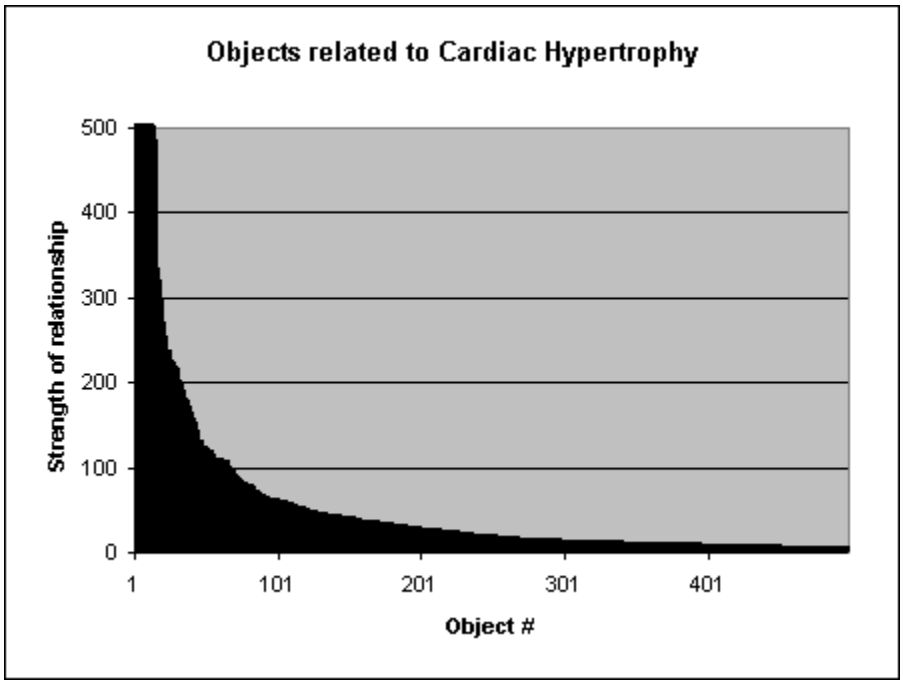
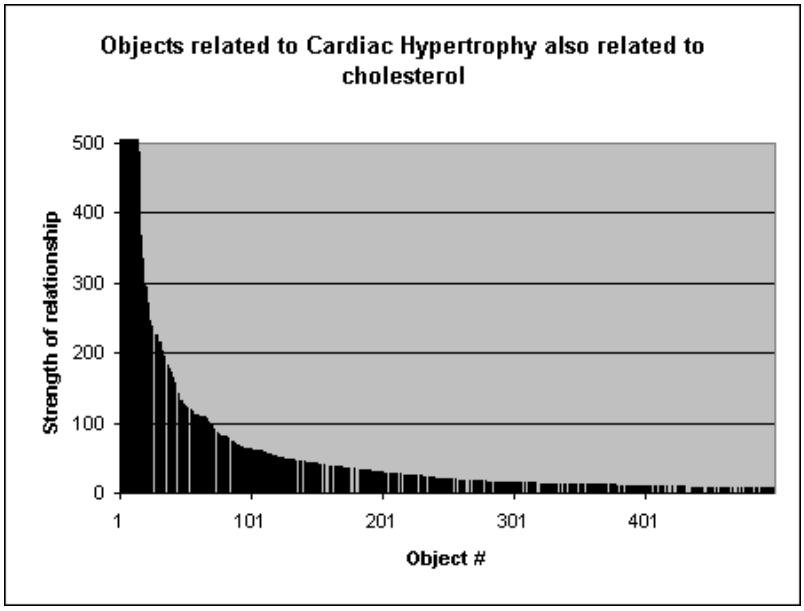


Figure 27: Distribution in the relationship strengths of objects related to cardiac hypertrophy. Only the 500 objects with the strongest relationships are plotted here, and the strength cutoff has been set at 500 to enable viewing of relevant features. This exponentially decreasing curve is typical for any relatively well-studied object. Objects are usually related strongly to a few items, and weakly to very many. In part, this is due to the time a relationship has been known, since most authors briefly summarize known relationships to introduce their experimental approach to discovering new ones. This distribution is also a

function of the relevancy of a relationship, as strongly relevant phenomena tend to be mentioned and studied together.

When an object, A, is implicitly related to another object, C, by a number of intermediates, B, we would anticipate that if A and C both shared a set of strong relationships that the probability of a relationship between A and C would be greater than if they shared a set of weak relationships. Dividing the total strength of the shared relationships by the total strength of all relationships, we can estimate what proportion of the important relationships that are shared (Figure 28). The area underneath the curve (AUC) can be calculated as the integral of the total strength of the relationship. This number can be calculated for the relationships shared by A or by C, reflecting in part the directionality of the relationship. For example, high cholesterol levels contribute to the development of cardiac hypertrophy through a number of different mechanisms including arterial hypertension, myocardial ischemia from blood clots, and membrane phospholipid composition. Cholesterol is the 27th strongest relationship on the list of objects related to cardiac hypertrophy with 244 abstract co-mentions. Consequently, a number of the strongest relationships with cardiac hypertrophy are relationships also shared with cholesterol (Figure 28a). However, cholesterol is related to a number of other biological processes, of which cardiac hypertrophy is only a small part. Consequently, a larger proportion of the strongest relationships cholesterol has are not ones shared with cardiac hypertrophy (Figure 28b).

a



b

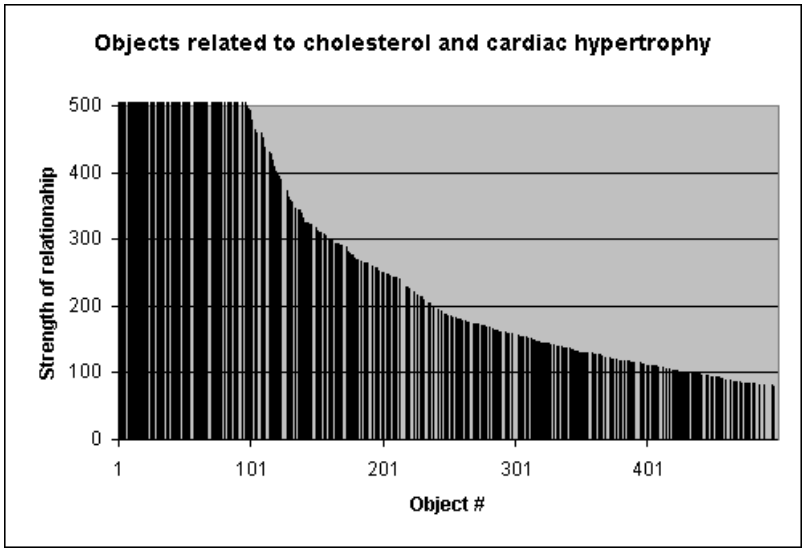
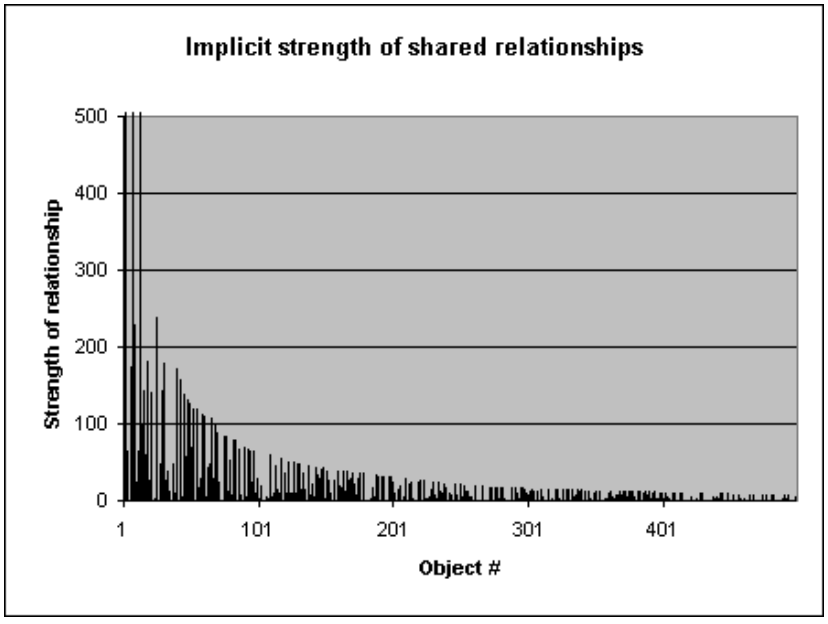


Figure 28: The relative importance of a related object as seen through the strength distribution of shared relationships. **a)** The distribution in the strength of relationships to cardiac hypertrophy. **b)** The distribution in the strength of relationships to essential hypertension. Dividing the integral strength of shared relationships by the total area underneath the curve allows an estimate of how important the shared relationships are. Cholesterol levels contribute in a number of ways to the development of cardiac hypertrophy,

but cardiac hypertrophy has little to do with cholesterol metabolism. The integral strength ratio for cardiac hypertrophy is 0.89, while for cholesterol it is 0.77.

This disparity is even more evident when considering the implicit strength of these relationships (Figures 29a and 29b). The areas underneath these curves can be used to estimate the relatedness of two objects by comparing the strength of the matching relationships to estimate their relative importance with respect to the query object. The relative strength of the matches can also be taken into account to assess the relative strength of the matching relationships with respect to both objects.

a



b

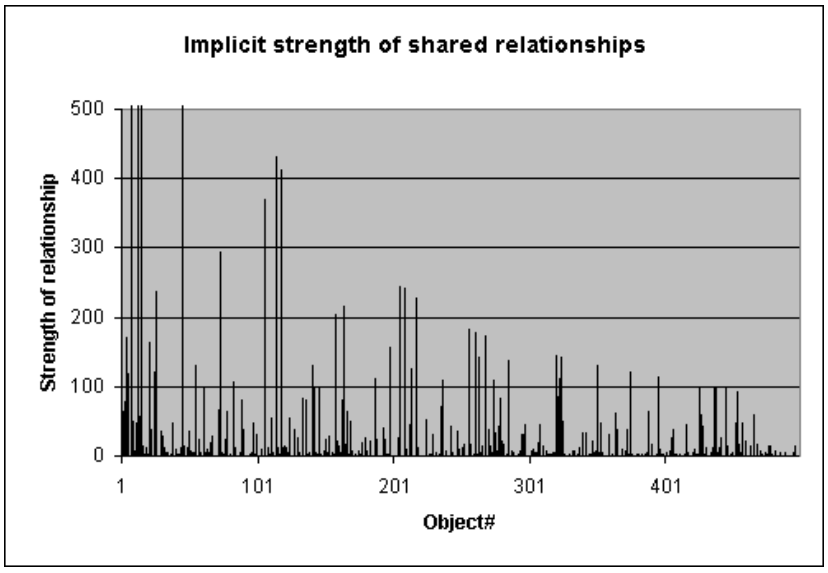
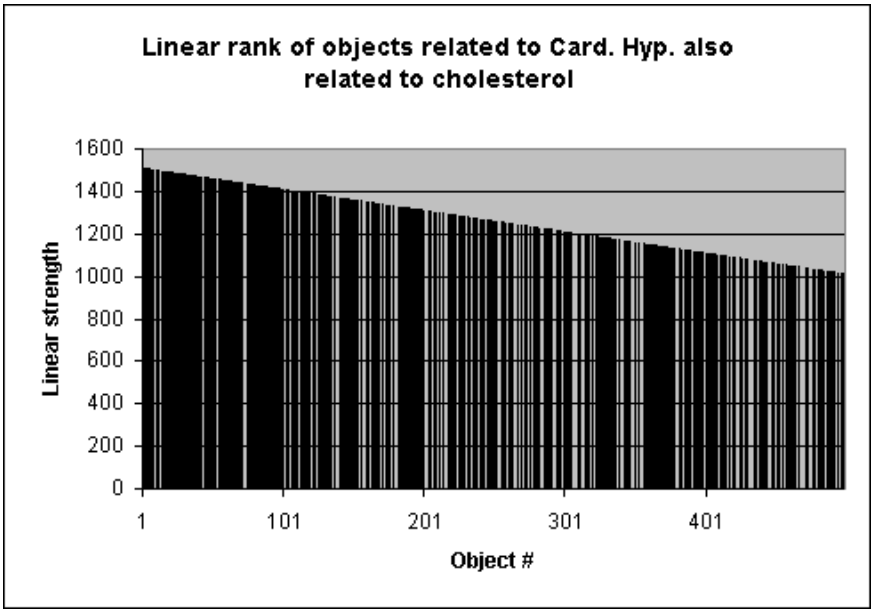


Figure 29: Implicit strength of the shared relationships between cardiac hypertrophy and cholesterol. **a)** Shaded areas show which relationships cardiac hypertrophy shares with cholesterol. **b)** Shaded areas show the relationships cholesterol shares with cardiac hypertrophy. Taking the integral strength ratios reveal a stronger disparity between the directional importance of the relationship. Cardiac hypertrophy has an implicit integral strength ratio (IISR) of 0.35 with respect to cholesterol, while cholesterol has an IISR of 0.07 with respect to cardiac hypertrophy.

The disadvantage of this exponential weighting scheme is that high priority is given to the few relationships that comprise the leftmost portion of the curve. While these relationships are better established than others, they may consequently be less interesting. As mentioned previously, part of this high frequency of co-occurrence will be a function of how long a relationship has been known. New, very important relationships will simply not have had sufficient time to accumulate such a high frequency of co-occurrence. We can thus flatten the curve into a linear ranking of relationships by their strength to reduce, but not eliminate, the relative importance of time as a factor (Figure 30).

a



b

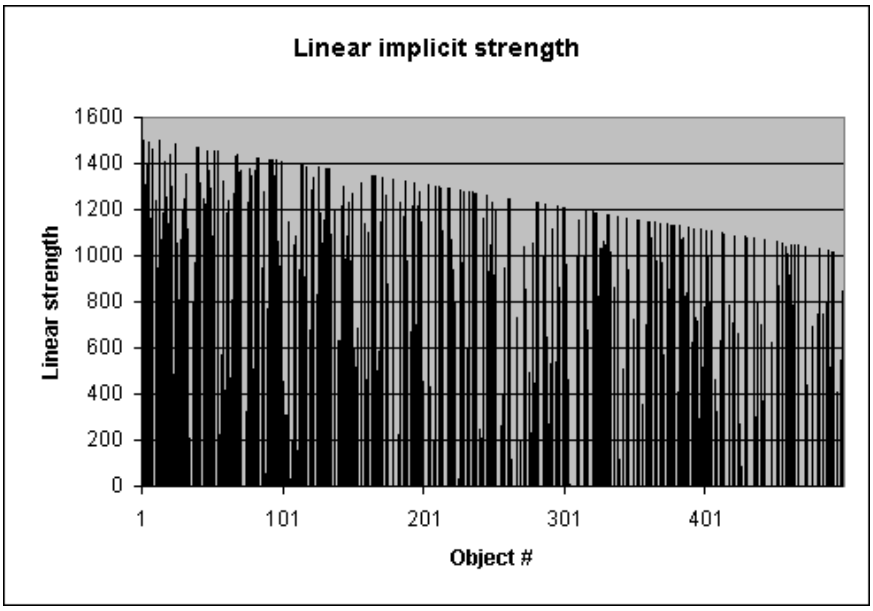


Figure 30: Ranking shared relationships by their relative position in a list results in a linear distribution in scores, thus reducing the relative importance of long-standing or well-known relationships. **a)** Objects related to cardiac hypertrophy that are also related to cholesterol by their relative position in a 1500 member list. Gaps indicate objects that are related to the query object only implicitly – that is, there are no documented relationships. **b)** The linear implicit strength reduces the effect of extreme differences in relative strengths when comparing AUC ratios.

For example, hypertension has been known for a long time to be highly correlated with the risk of developing cardiac hypertrophy as well as its severity. Its relative contribution to the area under the curve in the non-linear model (Figure 28) is the strength of the cardiac hypertrophy-hypertension relationship (3811) divided by the total of all the relationship strengths that any object has with cardiac hypertrophy (43510) = 3811/43510 = 8.8% of the total value, giving this one relationship a relatively high weight. In the linear model above, its contribution is only 1509/1139295 = 0.1% of the total value.

Unfortunately, a problem that arises with the linear ranking model is that the large number of low-quality relationships (i.e. objects co-occurring only once in the same abstract) will be weighted higher. For example, there were 600 total objects with only one co-occurrence with cardiac hypertrophy. In the linear ranking model, one of these objects will arbitrarily receive a 600-fold higher weight than another. This problem may be corrected by assigning a linear rank to each object based not upon its relative rank in the list of all objects, but the relative rank of its strength score within the list of all strength scores. This scoring scheme, the Linear Ranked Relationship Strength Score (LRRSS), will be used instead of the straight linear model.

Finally, we can envision a circumstance in which the strength of a relationship is not as important as the certainty of one. For example, if two objects shared a subset of relationships to objects collectively responsible for a specific biological process (e.g. acute-phase immune response, cell division, microtubule assembly, etc.), the relative strength of such relationships is not necessarily as important as the fact that the relationships are shared. Under this circumstance, we would prefer to evaluate how confident we are that the co-

mentions represent actual relationships. Using the veracity score (Equation 2), we assume that if the odds of one co-mention being a FP error is 50%, then the odds of two co-mentions both being errors would be 50%*50%=25%. The veracity score for any given relationship thus will range from the lowest possible FP rate measured for co-mentions to 1. It has not yet been established whether or not this assumption is true, and will be the subject of future study, but for now will remain in place as an operative assumption about the overall veracity of any given relationship. We can then plot shared relationships in terms of their integral veracity scores.

4.5 Implicit Relationship Analysis: Chlorpromazine and Cardiac Hypertrophy

We sought to validate the utility of IRIDESCENT to identify novel and useful implicit relationships by applying it towards a disease studied by researchers at UTSW, cardiac hypertrophy. Our objective was to use the system to identify compounds implicitly related to cardiac hypertrophy that could be of use in affecting the course of the disease.

Cardiac hypertrophy is a process by which the myocytes in the heart muscle expand in size, decreasing the capacity of the heart to pump blood. This phenomenon can occur as a response to environmental stimuli such as increased physical stress, chemical/toxic insults or genetic modification[85]. It is a relatively widely studied field as evidenced by the 3,654 articles in MEDLINE containing the key phrase “cardiac hypertrophy” as of 6/12/2002. IRIDESCENT identified a total of 2,102 objects mentioned within these articles and a total of 19,718 unique objects implicitly related to cardiac hypertrophy through 1,842,599 different paths. This example helps illustrate the highly interconnected nature of this

MEDLINE-derived relationship network and the need for a method of scoring potential relevancy, since this one object is either directly or implicitly related to 97% of all the identified objects in the literature. Using Equation 12 to calculate the expected number of times a set of objects would connect to the implicit object and compare this to the observed number of times, IRIDESCENT generated a ranked list of small molecules (e.g. drugs, metabolites, and chemical compounds) that were implicitly related to cardiac hypertrophy (Table 14). We chose chlorpromazine in part because of the molecular targets by which it is known to interact.

Rank	Implicit Relationship	Unique Paths	# of rels	Quality Estimate	Expect	Obs/ Exp	Score
1	Endotoxins	1301	3280	1025.2	307	4.24	1004.8
2	Progesterone	1448	4190	1131.8	392	3.70	966.6
3	Morphine	1217	3029	939.3	283	4.30	932.6
4	Bromide	1368	4079	1048.2	381	3.59	868.7
5	Concanavalin A	1317	3802	1002.3	355	3.70	857.9
6	Globulin	1130	2836	849.7	265	4.26	836.6
7	**Chlorpromazine**	**1089**	**2691**	**824.5**	**252**	**4.33**	**824.5**
8	Polyethylene Glycol	1153	2986	862.7	279	4.13	823.2
9	Cisplatin	1129	2932	862.0	274	4.12	820.2
10	Methotrexate	1190	3297	897.1	308	3.86	800.1
11	Esterase	1197	3394	907.6	317	3.77	791.0
12	Neomycin	1105	2908	841.5	272	4.06	790.1
13	Casein	1165	3289	894.9	308	3.79	783.3
14	Phytohemagglutinin	1099	2848	807.3	266	4.13	769.8
15	Isoleucine	1142	3134	852.2	293	3.90	767.3
16	Methanol	1221	3781	930.5	354	3.45	742.5
17	Galactose	1104	3040	826.3	284	3.88	741.5
18	Polysaccharide	1092	3160	829.4	295	3.70	708.2
19	Acetone	1075	3045	804.2	285	3.78	701.5
20	Tetracycline	1066	3022	799.9	283	3.77	697.2

Table 14: Objects of the class "Small Molecule/Drug" within the composite database that are implicitly related to cardiac hypertrophy, ranked by their score, which is a composite function of the probability each individual relationship is valid, the number of relationships each object is expected to have given its relative abundance in the network, and the implicit strength of each connecting relationship. The number of shared relationships between cardiac hypertrophy and the implicitly related objects is given in the "Unique Paths" column. A statistical estimate of how many of these

represent valid relationships is given in the "Quality Estimate" column. How frequent each implicit object is within the network is given under the column "# of rels", and the number of relationships we would expect by chance given the relative frequencies of each object is in the column labeled "Expect".

Chlorpromazine is an aliphatic phenothiazine compound (Figure 31) used principally as an anti-psychotic and anti-emetic drug[86]. It has a number of physiological effects and molecular targets that suggest it might provide an anti-hypertrophic effect in the heart, one of which is its alpha-adrenergic blocking activity[87]. Hypertrophy can be induced through overstimulation of alpha-adrenergic receptors by agonists and this effect can be blocked by alpha-adrenergic antagonists[88]. Chlorpromazine is also known to induce hypotension, and cardiac hypertrophy can result from hypertension. Within the cell, one factor known to induce cardiac hypertrophy is the overexpression of calcineurin[89], which is dependent upon calmodulin for its phosphatase activity[90]. Chlorpromazine is also known to interact with calmodulin[91] as an antagonist, suggesting a potential role beyond the alpha-adrenergic receptor. Despite the potential mechanistic connections between cardiac hypertrophy and chlorpromazine, there is no indication within MEDLINE that any relationship between the two has been established.

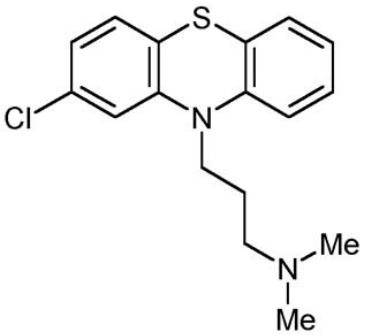


Figure 31: The molecular structure of chlorpromazine

While it would be useful to study chlorpromazine's effect upon alpha-adrenergic or hypertension induced hypertrophy, we sought to establish whether or not chlorpromazine's effects might extend to a more general model of cardiac hypertrophy. Using a beta-

adrenergic agonist known to induce cardiac hypertrophy, isoproterenol, 2 groups of 8 mice were fitted with osmotic micro-infusion pumps, one group given a steady dose of 20 mg/kg/day isoproterenol and the other 20 mg/kg/day isopreterenol + 10 mg/kg/day chlorpromazine. A smaller dose of chlorpromazine was chosen in preference to a larger one so that alterations in feeding behavior would be minimized. Additionally, we noted an adverse lethal reaction between mice given chlorpromazine while under anesthesia induced by avertin (tribromoethanol). Echocardiograms were taken before treatment and again 7 days later before the surviving mice were sacrificed to allow measurement of their heart weights. We find that the amount of cardiac hypertrophy was reduced in the chlorpromazine treated mice by a number of different measurements. Individual data points are shown graphically in Figure 32 and summarized in Table 15.

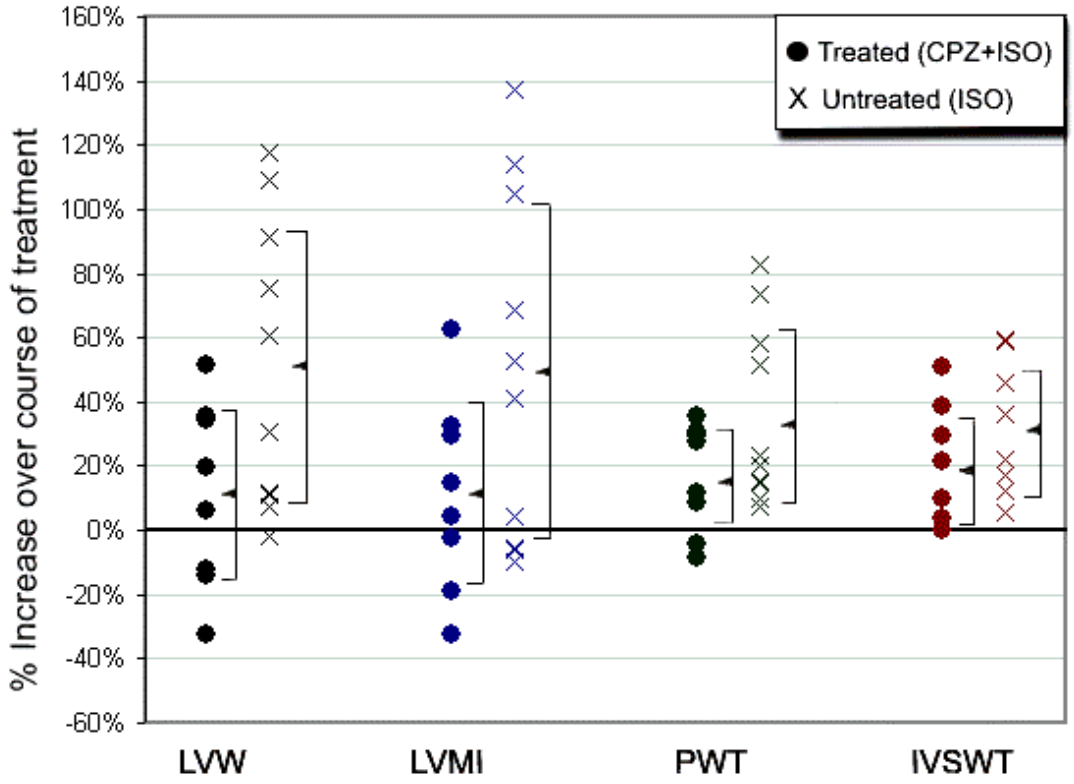


Figure 32: Chlorpromazine protects against the development of cardiac hypertrophy. Echocardiography was use to estimate the change in weight or thickness of several different cardiac structures over the course of treatment. One group of mice received isoproterenol only (ISO, n=10) and the other received both isoproterenol and chlorpromazine (CPZ+ISO, n=8). LVW = Left Ventricle Weight (CPZ+ISO 11±27%, ISO 51±43%, P<0.02), LVMI = Left Ventricular Mass Index (CPZ+ISO 11±28%, ISO 50±52%, P<0.04), PWT = Posterior Wall Thickness (CPZ+ISO 16±16%, ISO 36±27%, P<0.05), IVSWT = Intraventricular Septum Wall Thickness (CPZ+ISO 19±18%, ISO 31±20%, P<0.12).

Group	**ΔLVW**	**ΔLVMI**	**ΔPWT**	**ΔIVSWT**
CPZ+ISO	11%±29%	11%±30%	16%±17%	19%±19%
ISO	53%±45%	50%±55%	36%±28%	31%±21%
t-test	0.02	0.04	0.05	0.12

Table 15: Mice simultaneously treated with chlorpromazine and isoproterenol develop less cardiac hypertrophy than mice treated with isoproterenol alone. LVW = Left Ventricle Weight, LVMI = Left Ventricle Mass Index, PWT = Posterior Wall Thickness, IVSWT = Intraventricular Septum Wall Thickness. T-test = 1 tailed student's t-test

4.5.1 Materials and Methods used in the Chlorpromazine-Cardiac Hypertrophy Study

Male, 8-10 week old C57/BL6J mice were obtained from Jackson Labs (Bar Harbor, ME), and isoproterenol and chlorpromazine from Sigma Labs. Isoproterenol (ISO) and chlorpromazine (CPZ) were delivered by micro-osmotic pump (Alzet, model 1007D). One group received 20 mg/kg/day ISO and the other 20 mg/kg/day ISO + 10 mg/kg/day CPZ. Two separate experiments were conducted using this dosage regimen, one for 4 days (CPZ+ISO, n=7 and ISO, n=7) with a pre-treatment of 10 mg/kg i.p. one day before pumps were inserted, and the other for 7 days (CPZ+ISO, n=8 and ISO n=6) without pre-treatment. Initially, we observed a higher rate of mortality during surgery in the 4-day group (5 died in the CPZ+ISO group, 1 died in the ISO group) in which the mice were given chlorpromazine pre-treatment concurrently with Avertin (1.25% tribromethanol, 12µl/g, i.p.). No difference in mortality rates was observed in the 7-day group not given pre-treatment (2 died in the CPZ+ISO group, 2 in the ISO group).

Transthoracic echocardiography was performed before implantation of the pump and immediately before sacrificing at day 7 (Sonos 5500, Agilent; 12MHz transducer). Mice were sedated with intraperitoneal injection of low-dose Avertin. Echocardiographic examination was started 10 minutes after initiation of sedation to limit anesthesia-induced impairment of cardiac function[93]. A parasternal short-axis view was obtained for left ventricular M-mode imaging at the papillary muscle level by an operator unaware of

treatment status (R.B.). Three independent M-mode images were analyzed for measurements of left ventricular end-diastolic internal diameter (LVEDD), intraventicular septum (IVS) and posterior wall thickness (PWT) in two consecutive beats according to the American Society of Echocardiography leading edge method[94]. Estimation of left ventricular weight (LVW) was calculated as LVW(mg)= $((LVEDD+IVS+PWT)^3-LVEDD^3)x1.055mg/mm^3$[95]. Left Ventricular Mass Index (LVMI) was calculated from pretreatment (pre) and posttreatment (post) echocardiogram calculations of LVW and body weight (BW) as LVMI = $(LVW_{pre} / BW_{pre}) / (LVW_{post} / BW_{post})$. Fractional shortening (FS) was calculated as FS%=((LVEDD-LVESD)/LVEDD)x100.

4.6 Implicit Relationship Analysis: NIDDM and methylation

The etiological origin of non-insulin dependant diabetes mellitus (NIDDM) has long been controversial. Decades of research on NIDDM have created tens of thousands of articles scattered across many specialties, each containing their own set of observations. This makes a broad perspective difficult, if not impossible. Fortunately, IRIDESCENT enables a broad perspective to be elucidated. Following is an analysis of relationships NIDDM was discovered to share with epigenetic changes such as DNA methylation and chromatin structure. These relationships, when considered collectively and in the context of fundamental observations on the nature of the disease, are highly suggestive that NIDDM is the result of epigenetic changes within adipocytes, leading to a gradual dysregulation of cytokines or cytokine-like factors that are responsible for the NIDDM phenotype.

NIDDM is an increasingly prevalent disease in the world, especially the United States, where the number of new patients grew 49% between 1991 and 2000. The economic cost of NIDDM is staggering, estimated at $98 billion annually in 1997[96] and affecting as much as 6% of the population in the United States. Many factors that correlate with the risk of developing NIDDM have been identified, but causality has proven elusive. NIDDM has consequently been termed a "complex" disorder[97], thought to be a result of a complex interaction between environmental influence and genetic background. To date, no association has been reported between the etiology of NIDDM and epigenetic alterations such as changes in DNA methylation status or chromatin condensation.

DNA methylation is a fundamentally important phenomenon within eukaryotes, serving as a means to distinguish host DNA from foreign[98], to determine which strand of DNA is newly replicated[99] and to provide a signal for chromatin condensation such that transcriptional programs can be inactivated, a process especially important during normal development[100]. Loss of methylation in regulatory DNA regions has been an active research area in cancer, with a number of genes known to be dysregulated from a loss of methylation in certain tumors[101]. While loss of DNA methylation can be induced chemically (e.g. 5-aza-2'-deoxycytidine), it is not clear what factors may be present in the environment that would have a similar effect.

The first barrier scientists face in hypothesizing a novel relationship between objects is awareness of common relationships. Assuming one had a reason to hypothesize a novel relationship between epigenetic modification and NIDDM, it would still be necessary to read and organize 24,752 articles on NIDDM and 25,338 articles on methylation to identify

commonalities (statistics as of July 5, 2002 as determined by MEDLINE keyword query). A bioinformatics approach is necessary to collate data of such scale. By examining the entire body of MEDLINE literature associated with NIDDM, IRIDESCENT can identify all potential relationships that NIDDM has to other objects by their co-occurrence within the same journal abstract.

We used IRIDESCENT to identify and rank objects within MEDLINE implicitly related to NIDDM. We found that NIDDM shares many implicit relationships with two other objects in our database: “Methylation” and “Chromatin” (Table 16).

Rank	Paths	Implicit Relationship	Quality	Expect	Obs/Exp
---	2105	NIDDM	1421	329	4.32
1	1361	Endotoxin	1054	308	3.42
2	1312	Hydrocortisone	991	296	3.35
3	1301	Neuroblastoma	975	339	2.88
4	1287	Methylation	959	346	2.77
5	1256	Chromatin	938	339	2.77

Table 16: Top 5 objects (genes, diseases, phenotypes, and small molecules) implicitly related to NIDDM (shown at top as a positive control for the query). The nature of each implicit relationship will vary and must be determined by examination of the intermediate connections.

From the 33,534 unique objects IRIDESCENT is capable of recognizing within text, a total of 2,105 were found to be directly related to NIDDM. IRIDESCENT then analyzed MEDLINE for all objects directly related to these 2,105 objects, removing those already in the list of direct relationships. The resulting list contains relationships that are known only implicitly, which is to say that no relationship between the two objects should be found within the body of MEDLINE titles & abstracts. These implicit relationships are then

evaluated by IRIDESCENT based upon the number of shared relationships they have with each other, relative strength of each relationship, quality of the relationships (statistical probability that each relationship is valid), and the likelihood the two objects would share a set of relationships by chance, given the relative abundance of both objects and their shared intermediates within the network. Not all of the 1,287 relationships shared between "methylation" and "NIDDM" are necessarily causal, correlative or even meaningful, but many are. Collectively, they provide evidence that a relationship does exist between epigenetic control and NIDDM and enabled us to develop a more comprehensive theory regarding an epigenetic etiology and pathogenesis of NIDDM. We will limit discussion of shared relationships to those we believe are most pertinent (Figure 33)

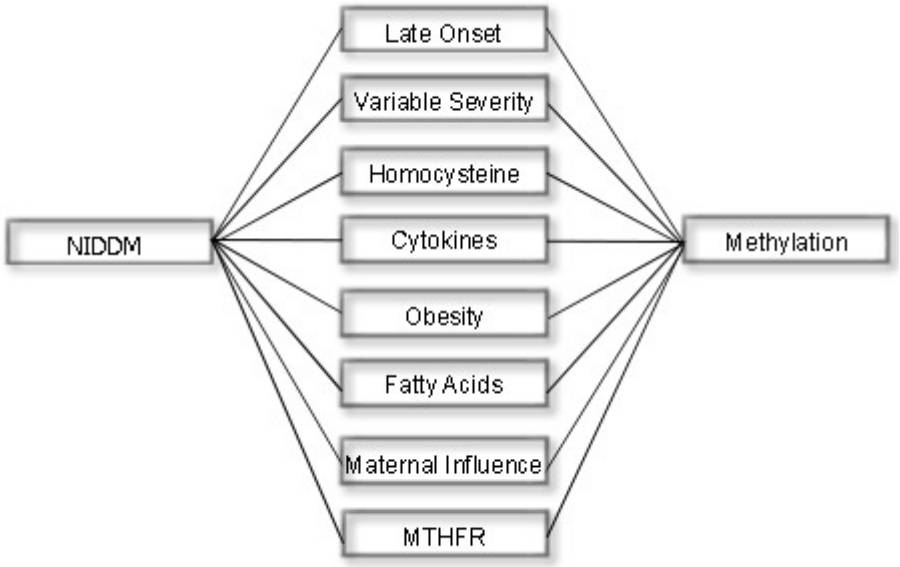


Figure 33: Important shared relationships between methylation and NIDDM. A total of 1,287 co-cited objects were identified between the two, of which an estimated 959 of these represent actual relationships of a non-trivial nature. Only relationships emphasized within this report are shown here. A full list is available online at http://innovation.swmed.edu/IRIDESCENT/NIDDM_theory.htm

4.6.1 Shared relationships linking alterations in DNA methylation to NIDDM

IRIDESCENT identified a number of common phenotypes in the onset and pathology of NIDDM that are also shared by diseases associated with a change in methylation state. These shared relationships offer a perspective on some of the puzzling properties of NIDDM not easily explained by environmental or genetic mutation models. For example, NIDDM is a disease with variable and late onset, a phenotype linked to some epigenetic disorders through DNA hypomethylation such as aberrant expression of X-linked genes[102] onset of Huntingtons Disease[103] and oncogenesis of tumors[104,105]. Not all late-onset illnesses are caused by epigenetic changes, but most others share phenotypic abnormalities that are unique to the disease, such as the accumulation of amyloid precursor proteins in Alzheimers[106] or Lewy bodies in Parkinsons[107]. NIDDM is highly correlated with the presence of obesity and Advanced Glycosylation End products (AGEs), but neither is a requirement for its development nor unique to it as a disease. NIDDM also varies in its severity, generally increasing over time. This is a phenotype shared with some tumors that have undergone methylation changes in promoter sequences, leading to higher gene expression and a more aggressive phenotype[105,108]. Another interesting observation about NIDDM is the "maternal effect" in which NIDDM patients report a higher frequency of maternal history of diabetes[109]. While this is not without controversy[110], such an effect could be explained if *de novo* methylation of DNA sequences during development was due to maternal influence. This type of phenomenon, in fact, has been observed in mice[111].

IRIDESCENT also identified a number of metabolic alterations in the body's ability to methylate DNA that correlate with the existence of or predisposition to NIDDM. For example, elevated levels of homocysteine have been found in NIDDM patients, correlating with increased severity of the disease as defined by mortality[112]. Homocysteine is a critical metabolic intermediate responsible for carrying out methylation reactions, and elevated serum levels of it are also correlated with DNA hypomethylation[113]. It has also been reported that sulfur-poor diets that force synthesis of cysteine from methionine predispose individuals to Type II Diabetes later in life [114,115]. Since methionine affects S-adenosyl methionine (SAM), which is the methyl donor for the methylation of newly-synthesised DNA, these individuals develop with an impaired ability to establish *de novo* DNA methylation patterns. Genetic factors that lead to deficiencies in the methylation pathway have also been shown to predispose individuals to develop NIDDM. There is a well-known polymorphism (C677T) in the methylenetetrahydrofolate reductase (MTHFR) gene that reduces its efficiency, leading to a global hypomethylation of DNA[116]. Individuals with this mutation are also predisposed to develop NIDDM and other complications of the metabolic syndrome[117].

Aberrant methylation patterns have been shown to induce diabetic symptoms in another form of diabetes, Transient Neonatal Diabetes Mellitus (TNDM), which is a result of genetic imprinting[118]. The same imprinted region responsible for TNDM, however, is not known to be responsible for NIDDM[119]. If epigenetic alterations are responsible for NIDDM, then three questions naturally arise: First, what secreted factors are responsible for

the NIDDM phenotype? Second, what tissue-type(s) is responsible for expressing the factors that induce the NIDDM phenotype? And third, what environmental factors could lead to a loss of methylation and consequent dysregulation of the secreted factors?

Insight into an answer for the first question comes from the highest scoring object on IRIDESCENT's list of implicitly related objects(Table 16): Endotoxins. While endotoxins are not known to be associated or causal in NIDDM, they have been shown to induce obesity and insulin resistance[120,121]. Most of the relationships shared between NIDDM and endotoxins are objects that either affect or are involved in the immune response, especially cytokines and inflammatory factors. Expression of acute-phase markers such as C-reactive protein, and pro-inflammatory cytokines such as IL-6 and TNF-alpha are highly correlated with the presence and severity of NIDDM symptoms[122-124]. These pro-inflammatory cytokines are also positively correlated with obesity[125]. Furthermore, TNF-alpha has been found to induce insulin resistance[126-128]. Indeed, there is a growing body of evidence that cytokines, more specifically the pro-inflammatory cytokines, are responsible for the NIDDM phenotype. It has been observed, for example, that a reversal of NIDDM symptoms can be induced by disruption of the inflammatory pathway with high doses of aspirin[129]. Troglitazone, a widely used medication to treat NIDDM, has also been found to have anti-inflammatory properties[130], and the lifestyle changes of exercise and dietary changes prescribed to NIDDM patients that have been successful in reversing NIDDM phenotypes have also been associated with reductions in inflammatory cytokines[131,132].

Since there is evidence that pro-inflammatory cytokines are the causal factor in NIDDM, it is of interest to identify their origin. Besides B-cells and T-cells, adipocytes and endothelial cells are the only other cell types known to normally produce cytokines. We see that within T-cells, cytokine expression is determined by DNA methylation patterns[133] and can be altered by demethylating agents[134]. Neither T-cells nor B-cells seem a likely candidate since they are not very metabolically active in their naïve or memory forms, and their more active differentiated forms are relatively short-lived. Adipocytes, however, are the primary repository for lipids and produce cytokines in proportion to factors such as their size and surrounding obesity[135]Interestingly, a study by Benjamin and Jost demonstrated that short-chain fatty acids (SCFAs) can promote the demethylation of actively transcribed regions[136]. SCFAs can also affect chromatin structure by inhibiting HDAC, causing hyperacetylation of histones[137] and making regions of DNA more accessible to transcription factors. SCFAs are not normally present in high concentrations within adipocytes, but are normal metabolic byproducts of the long-chain fatty acids stored within. Since the rate of lipolysis within adipocytes is increased in NIDDM[138], and can be induced by factors such as TNF-alpha already known to be elevated in NIDDM[139], this would have an effect upon the relative concentrations of SCFAs within adipocytes. Higher amounts of SCFA metabolites within adipocytes might provide an environment in which loss of DNA methylation could occur and, coupled with active transcriptional activity, could lead to the hypomethylation and consequent dysregulation of cytokines or cytokine-like factors that lead to NIDDM. We see suggestive evidence of this in a study by Laimer *et al* involving IL-6 and TNF-alpha levels in

20 women before and 1 year after gastric banding surgery. They found that the levels of other obesity markers such as C-Reactive Protein (CRP) declined, while IL-6 and TNF-alpha did not[140].

Within the proposed model, the etiology of NIDDM occurs within adipocytes, involving a gradual loss of DNA methylation around the promoters of cytokines and/or cytokine-like factors normally secreted by the adipocyte. This loss of methylation is favored under the conditions provided by obesity and is caused by transcriptional activity. The subsequent loss of methylation leads to a dysregulation of these factors, resulting in a constitutive increase in the production of cytokines from adipocytes. Negative regulatory factors reduce the expression of these factors, enabling a management of the NIDDM phenotype, but only as long as they are present.

4.6.2 Etiological models of NIDDM

We examine this new proposed model in the context of the three existing models for the etiology and pathogenesis of NIDDM: Genetic, environmental, and a complex interaction of both factors. Genetic studies have shown that inheritance plays a role in determining an individual's risk of developing NIDDM[141]. Linkage studies, while delineating a number of potential susceptibility regions, have yet to be successful in identifying a specific gene or set of genes responsible for the most popular form of NIDDM, despite the large cohorts involved. The well-established correlation between obesity and NIDDM also indicates that environmental variables affect the pathogenesis of NIDDM. The prevailing theory is that the onset of NIDDM is caused by one or more environmental variables acting upon a genetic

background, of which there may be many contributing genes[97]. This theory explains how susceptibility to NIDDM correlates with genetic background, such as race, as well as with environmental variables such as diet and exercise. There are other observations about the nature of NIDDM that the complex model does not explain but the epigenetic model does: Time-dependency and systemic memory.

Even when environmental variables are present on a susceptible genetic background, the onset of NIDDM is still time-dependent. That is to say, the risk of developing NIDDM is positively correlated with age. This is not easily explained by the complex disease model except to postulate an as-yet-unknown "trigger" event, such as an infection. Even if this were true, it would not explain the persistence of NIDDM after onset. NIDDM is diagnosed by the levels of insulin resistance and glucose intolerance experienced by a patient, levels which can be altered to pre-diabetic levels by sufficient changes in lifestyle. NIDDM, however, cannot be reversed[142]. None of the existing models account for a mechanism by which the body can "remember" its state. The methylation status of genes, however, is considered to be a relatively persistent phenomenon, responsible for committing cells into their differentiated states[143]. Given that loss of DNA methylation is correlated with age[144], that the number of methylated sites in a genome is determined by inheritance, and that loss of methylation can be affected by environmental variables, it would seem that the proposed epigenetic model merits serious consideration.

Contrary to the mutation-centric model, which assumes alterations in function or activity based upon either somatic or inherited mutations in DNA, an epigenetic model implies a dysregulation of a gene or set of genes. Thus, phenotypes resulting from the

expression of such genes would make biological sense under other physiological conditions. Preventing energy influx into cells by inducing insulin-resistance makes sense when considered within the context of the role of the immune system. Acquired immunity in the form of B-cell maturation and antibody production takes time during which pathogens are able to replicate. Part of the early immune response consists of an increase in the presence of pro-inflammatory cytokines within the circulating bloodstream[145,146]. It would make sense that one role of these early-responders would be to stem the influx of resources like glucose into cells to prevent their utilization by invading pathogens. Since adipocytes contain a large reservoir of energy, this makes them ideal targets for invading pathogens and could necessitate their taking a more active role in fighting infection beyond that of other somatic cells.

A candidate list for genes that have undergone epigenetic dysregulation can be obtained by identifying expression changes via microarray analysis and subsequently examining the methylation status of their promoters. If this theory is ultimately shown to be correct, it will allow us the ability to diagnose the current level of epigenetic progression towards NIDDM in patients and offer hope for a NIDDM cure that could not be easily provided in a mutation-centric model. It is not apparent how region-specific methylation could be reintroduced to affected regions, but since *de novo* methylation is a normal process during development, it stands to reason that the mechanism to do so is already in place.

4.6.3 Experimental approach

To test the hypothesis that loss of methylation in the promoter regions of pro-inflammatory cytokines or within genes responsible for the release of pro-inflammatory cytokines (e.g. an increase in pro-inflammatory cytokine receptor number could make adipocytes hypersensitive to normal levels of cytokines) is responsible for the NIDDM phenotype, we take the following approach:

1) Identify candidate genes
    a. Pare IRIDESCENT output of genes related to NIDDM through the user interface
    b. Identify promoter or CpG islands within each candidate gene
    c. Design primers that can amplify bisulfite-treated DNA within these regions
    d. Test primers on bisulfite-treated DNA to ensure they amplify the appropriate product
2) Obtain adipocyte samples from NIDDM patients and non-NIDDM patients
3) Use primers to amplify bisulfite-treated DNA from patient samples from each group
4) Clone amplicons into bacterial vectors to isolate individual clones
5) Sequence 10 clones for each gene to obtain a sampling of the methylation status for each gene being studied
6) Compare average number of sites methylated for each gene between the NIDDM and non-NIDDM patient samples

Table 17 shows an example of one of the genes analyzed for promoter methylation from a human adipocyte sample. The NIDDM status of this patient is unknown, the data were gathered to test the experimental pipeline from primer design to DNA isolation from adipocytes to the sequencing of bisulfite-treated DNA. Each CpG site is queried within the promoter region for its methylation status and summed for an overall estimate of percent methylation. The table shows the status of each CpG position (going 5’ to 3’ within the CpG island) within each of the 8 clones analyzed. There are 33 positions total that could be

methylated within this region and we see that only an average of 5.88 CpG sites (17.8% of the total) are methylated.

CpG position 5' -> 3'	Leptin-1	Leptin-2	Leptin-3	Leptin-4	Leptin-5	Leptin-6	Leptin-7	Leptin-8	Total
1	1					1		1	3
2									0
3					1	1		1	3
4				1					1
5	1		1	1		1	1		5
6								1	1
7							1		1
8						1	1		2
9								1	1
10		1						1	2
11					1	1	1		3
12						1		1	2
13							1		1
14							1		1
15								1	1
16		1				1			2
17	1				1				2
18			1						1
19									0
20									0
21		1	1		1				3
22								1	1
23	1								1
24	1								1
25	1								1
26	1			1					2
27	1							1	2
28			1						1
29	1			1					2
30									0
31									0
32			1						1
33									0
Total	9	3	5	4	4	7	6	9	
Avg. # methylated:	5.88								
Std. dev.:	2.30								
Avg. pct. Methylation:	17.8%								

Table 17: Eight clones containing bisulfite-treated DNA from the putative promoter region of the leptin gene are sequenced to ascertain which CpG positions are methylated. A "1" within a column indicates that position is methylated within that clone.

If no difference is identified between samples, then the candidate gene list must be refined and the approach reassessed. If a positive result can be achieved, then it is relatively straightforward to continue testing of the hypothesis, but what is more difficult is concluding with confidence that the hypothesis can be rejected. If, at this point, no significant difference in percent methylation between patient samples is found, then the following questions must be addressed:

1) With what confidence are we certain the adipocyte samples are from NIDDM and non-NIDDM patients? Our current source of adipocyte samples comes from surgical waste where diagnoses of diabetic status are ascertained by the prescription medication the patient was on at time of surgery. Some of the "non-diabetics" may actually be undiagnosed diabetics. At this point, the most informative experiment would be to obtain a set of control samples from patients whose levels of insulin resistance have been established.
2) Is the gene list exhaustive? Genes directly implicated in the pro-inflammatory response will have been tested, but are there other genes that could indirectly be implicated in this process? A reassessment of the candidate gene list will also be necessary at this point.
3) Can the hypothesis be more directly tested via an induction system? That is, using a factor that is known to induce transcription of a normally silenced gene, can we show loss of methylation within an adipocyte sample when subjected to this factor? Furthermore, can we show that this loss of methylation is adipocyte-specific or that the rate of loss is increased in adipocytes?

If a positive result is obtained and a statistically significant difference is observed between NIDDM and non-NIDDM patients, then the following remains to prove the hypothesis, either by its existence in the literature or experimentation. Depending on the gene(s) that are differentially methylated between patient samples, the answer may already be known or not.

1) Is the gene product secreted from adipocytes?

2) Can the secreted factor induce insulin resistance by itself? If not, then in concert with other circulating factors present in NIDDM?
3) Can the concentrations of the secreted factor account for the presence of insulin resistance? That is, would enough of the factor be produced by adipocytes to induce the phenotype?
4) Does removal of the factor result in the restoration of normal phenotype? This would have to be an experiment conducted using a conditional induction system in a model organism such as rats.

In conclusion, the cardiac hypertrophy-chlorpromazine relationship and the NIDDM-Methylation relationship are two examples of how IRIDESCENT can be used to elucidate novel relationships based upon identifying and ranking sets of shared relationships between two objects, screening for those relationships which have not already been documented. Analysis of shared relationships, however, extends beyond just those that exist between two objects to those that exist between 2 or more objects. IRIDESCENT is also able to identify and rank other objects related to a set of query objects. This enables a user to address the question "What do these things have in common?" or even "Do these things have anything in common beyond what could be observed among a set of randomly assembled objects?". We now turn to discuss IRIDESCENT's application to identifying commonalities within a set of objects.

4.7 Shared Relationship Application: Gene Ontology construction

There are a variety of scientific applications that are able to benefit from an analysis that answers the question: What does this group of objects have in common? One of the primary advantages of modern large-scale data-gathering technologies such as microarrays is that they can detect novel or unexpected changes in transcriptional response across an

extremely broad spectrum of genes. Conducting a clustering analysis on the responders aids researchers in identifying similar changes in expression levels between samples, experiments, conditions, cell lines or other variables, but further interpretation beyond this is left to the experimenter. Approaches to doing this will vary because the specific interests of an individual researcher can be very diverse, such as identifying genetic pathways affected by a change in experimental conditions, new genes that may be a member of a pathway, or drugs that affect a similar set of genes. In many cases, there is an open-ended aspect to each microarray experiment where no specific answer is sought at all but rather the data are expected to speak for themselves – that is, new patterns that reflect a biologically meaningful event will be uncovered. This is one of the aspects of microarray technology that makes it so appealing to the individual researcher, the idea that completely unanticipated discoveries can be revealed by conducting as broad a survey as possible.

One of the challenges in interpretation of microarray data is identifying the biological significance in the change of a *set* of genes. Beyond the similarities apparent from the gene names themselves or any annotation associated with them, MEDLINE is the primary source of information researchers consult to identify the common relationships that define a cluster of responding genes. Searching MEDLINE for information on genes can be a daunting task, as the number of articles published in MEDLINE containing the names of known genes ranges from zero (unpublished, yet transcript identified) to over 135,000 (Insulin). MEDLINE contained an estimated 12 million records at the beginning of 2002 and is growing at an annual rate of approximately 500,000 records/year, making manual evaluation of large sets problematic at best. Awareness of commonalities within large-scale data-

gathering experiments is central to the process of insight and discovery, and efforts to link literature information to experimental data provided by microarrays have recently been the focus of much effort[147,148].

Useful methods of linking genes to informational descriptors have been used in programs like MedMiner[149] and ARROGANT[150], but more sophisticated methods of analysis are needed to begin to make statements about the responding genes as a whole. Towards this end, methods have been developed such as the mapping of responding genes to a core set of relevant literature based upon a single best "kernel" document[151], giving the user the ability to identify keywords relevant to the retrieved documents. And Masys *et al*. developed a method to map keywords within the published literature on gene sets to a MeSH keyword hierarchy using the UMLS Metathesaurus[152]. These efforts to identify literature-based shared relationships have centered upon microarrays, but there is a more global need to find relationships among sets of objects assembled by any means. For example, clinicians could analyze a set of phenotypes to identify associated diseases or chemicals in the hopes they might provide insight into disease etiology or pharmacology.

We sought to determine whether IRIDESCENT's relationship network derived from co-occurring objects within MEDLINE could be used to evaluate the relatedness of a set of objects based upon the relationships they share (illustrated in Figure 34). This type of analysis has several benefits, such as allowing experimenters at least one way of verifying that their experimental grouping is purposeful (assuming the grouped objects are adequately represented within the literature). It enables "themes" (e.g. cancer, apoptosis, diabetes) to be identified via objects related to the most set members, as well as a method of scoring

exceptional groupings within the list. Finally, it also represents a potential method of identifying "missing members" in a set, by their relatedness to the group as a whole. We thus chose to apply this system to the analysis of Gene Ontology (GO) categories.

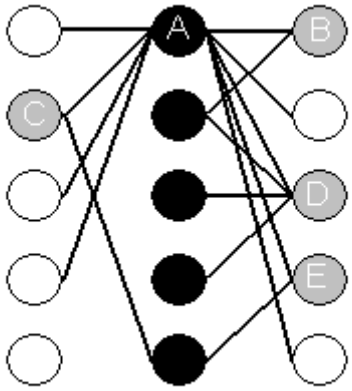


Figure 34: Identifying shared relationships within a network. A set of objects (represented by the black nodes) is queried to identify connections shared by at least 2 members of the set (gray nodes). Problematic is determining how exceptional such relationships are. Node A, for example, is connected to almost all other nodes outside the set, thus connections shared with A (nodes C,B,E) do not seem particularly exceptional. Node D, however, is connected to most of the members of the set, including nodes with very few connections, suggesting such a grouping is more exceptional.

The need for a dynamic controlled vocabulary in biology has been established[46] and efforts are underway to categorize known genes in terms of their ontology (http://www.geneontology.org/). Gene ontology construction is accomplished primarily via a manual, volunteer effort. Ultimately, the task could span the curation of tens of thousands of genes for potentially hundreds of thousands of species. Of course, the total number of assigned ontology categories would be far less, as homologous genes across species will

share all or most of an ontological classification. Yet a large number of ontological classifications have yet to be made, and there is a need to reassign function in response to changing knowledge as well as ascertain which classifications have yet to be assigned based upon current knowledge. As such, the value of automation in this process has been recognized. For example, Raychaudhuri *et al.* recently developed a document classifier to associate documents with GO categories[153], providing a supervised machine learning method of predicting gene annotation.

We know from our earlier studies (see Figure 23) that topical sets have a higher average Obs/Exp score than random sets. Therefore, we reasoned that genes IRIDESCENT identifies as being related to members of an ontological category would represent an enriched set of potential additions to that category. If so, then this method could represent a powerful and automated manner of assisting in the ontology development effort. These suggested new members could, after evaluation, be added to the original set. Genes related to an ontology category for reasons other than membership in it (e.g. associated with a closely related process) could be removed from the list and the number of remaining members would provide a dynamic estimate of how much curation remains, assuming the relationship network was kept up-to-date with the most current literature. Such an estimate would, of course, be limited by the method itself as well as the availability of literature documenting potential relationships.

First, gene ontology records were downloaded (on 11/11/2002) and found to contain a total of 13,414 unique ontological identification numbers and 13,106 unique descriptions. 115,303 Locuslink records were downloaded on the same date and processed by

IRIDESCENT so that only entries that represented actual known genes were included in the database (e.g. no tentative assignments based upon weak homology, ORFs or predictive methods), leaving 42,345 entries. A total of 21,452 of these Locuslink entries had at least one existing ontology category.

Sets of objects (2 to 100 members) were chosen at random from the literature-derived network, with the only stipulation being that they must have at least one connection in the network. All objects connected to at least 2 members of the set were evaluated according to Equation 12, while objects connected to fewer than 2 of the set were discarded. An average was calculated for all of these connecting objects. This was repeated 100 times for each size set to obtain a set size average and standard deviation for the set size average. The same was then done for sets of genes within each ontology category, except the sets were displayed as individual data points rather than averaged so that the distribution may be evaluated. We hypothesized that the average Obs/Exp score should be higher due to a larger number of shared relationships within the set and, in fact, this is what we saw earlier in Figure 23 when averages were plotted. Figure 35, below, shows the individual data points (ontology categories) as compared to the random set (shown with bars indicating one standard deviation). This allows a method of scoring a set for its overall "cohesiveness". For smaller sets, it is apparent that there is much more overlap with the random sets than is observed for the larger sets.

We examined some of the topical entries that scored within 2σ of the random average to see if their low scores might perhaps be in error. We find that a number of ontological categories can have genes that serve a common purpose, yet are sufficiently separate in terms

of their genetic associations that they are not frequently mentioned together in the literature (e.g. sensory perception genes, anion transporters). This represents a potential limitation of the method.

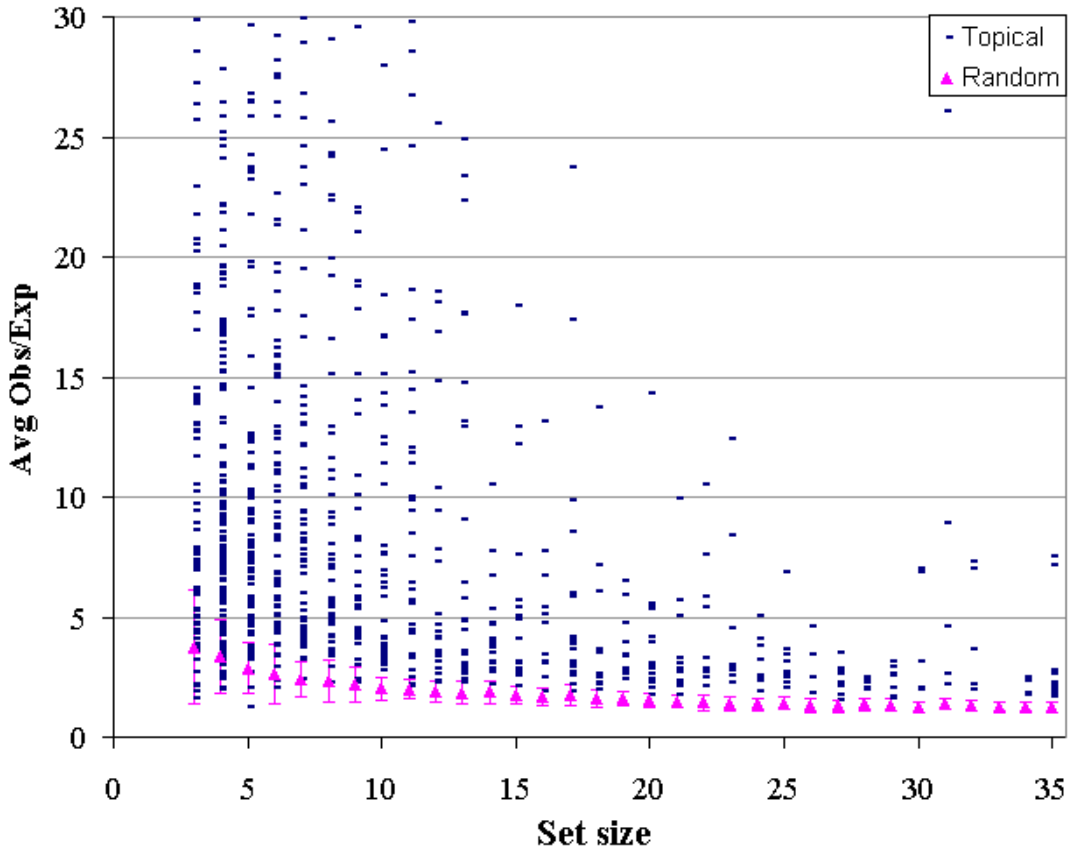


Figure 35: Comparison of the average observed to expected ratio between topical and random sets. Sets of objects, ranging in size, were either generated randomly or obtained from classifications within the Gene Ontology database. For the random sets, only the average Obs/Exp value is shown, along with bars indicating one standard deviation. The average Obs/Exp is shown for each topical set. There are 127 data points with an Obs/Exp above 30 that are not shown in this graph.

Genes associated with GO categories were output if their observed to expected ratio (Obs/Exp) was calculated to be at least 2 standard deviations (2σ) above the average Obs/Exp value for the same number of relationships given the size set as determined by

random network simulations, and if the object was related to at least a minimal (5%) portion of the set. Only GO categories with at least 3 members were processed. Entries were manually deleted from the output when the gene name was ambiguous or identical to a common word or phrase. Deleted gene names were: Autoantigen, cell surface protein, G protein-coupled receptor, membrane protein, unknown function, unassigned, transcription factor, transcriptional co-activator, nuclear protein, inflated, copper.

Given that a higher Obs/Exp ratio is indicative of a greater cohesiveness within a set, we reasoned that genes sharing many relationships and having a high Obs/Exp ratio with respect to the genes in a given ontological category, but are not themselves included in the category, might represent an enriched set of candidate genes for possible inclusion in the ontological category analyzed. Table 18 shows an example of a set of genes within an ontological category (brain development) along with the number of relationships each gene has within the network.

Gene Name	# relationships	LocusLink ID
NT-3	1111	LL:18205
TTF-1	425	LL:21869
BF1	406	LL:2290
DLX2	152	LL:1746
PGDH	152	LL:26227
SIX3	140	LL:6496
Hesx1	128	LL:15209
FMR2	121	LL:2334
ZIC	101	LL:7545
ZIC2	88	LL:7546
BMI1	87	LL:648
Cart-1	69	LL:8092
BF2	68	LL:2291
RB18A	21	LL:5469
Lhx6	20	LL:26468

hyh	14	LL:15591
Vax1	12	LL:22326
PITPNM	11	LL:9600
UNC5C	6	LL:8633
NKX2B	0	LL:4821

Table 18: Genes in the ontological category of "Brain development" (Gene Ontology ID# 7420) sorted by the number of related objects identified within MEDLINE by IRIDESCENT.

Table 19 shows the output produced by analysis of this set using IRIDESCENT. Within this table are a number of object names related to the genes in Table 18, and illustrates in part the nature of the problem we are attempting to address. Some of these relationships, while perhaps true, are not particularly exceptional such as the objects "tumor", "nucleus" and "apoptosis", which are all very highly related (common) objects within IRIDESCENT's literature-based network, and their relative abundance is reflected by a low Obs/Exp ratio. Examining the gene names within this list, however, reveals a number of genes also implicated in brain development but not annotated as such within Locuslink's Gene Ontology, such as engrailed (human homologs EN1 and EN2)[154], SHH [155], as well as BMP-4 and FGF8[156]. These genes have a much higher Obs/Exp ratio, suggesting a strong association with this ontological category. Another gene name, caudal, is in the list but scores low because "caudal" is also a word used to describe structures towards the tail end of the body.

Object Name	# shared	Expect	Obs/Exp	Locuslink ID
Nervous system	14	6.28	2.23	
Transcription factor	14	4.20	3.34	
Neurons	13	6.16	2.11	

Tumor	13	9.55	1.36	
Fibroblasts	10	5.51	1.81	
Lymphoma	9	3.81	2.36	
Nucleus	9	6.81	1.32	
SHH	**9**	**0.46**	**19.48**	**LL:6469**
Alternative splicing	8	2.43	3.29	
Secreted	8	4.40	1.82	
Apoptosis	7	4.65	1.50	
DNA-binding protein	7	1.54	4.54	
Hypoplasia	7	2.32	3.02	
Oncogene	7	2.03	3.45	
Zinc	7	4.02	1.74	
BMP-4	**6**	**0.39**	**15.22**	**LL:652**
Caudal	**6**	**2.52**	**2.38**	**LL:1044**
Cysteine	6	3.95	1.52	
Ectodermal	6	1.05	5.70	
engrailed	**6**	**0.27**	**22.01**	**LL:2019**
FGF8	**6**	**0.28**	**21.68**	**LL:2253**

Table 19: Objects related in the literature to one or more of the genes in Table 18 (only first 21 relationships shown), sorted by the total number of shared relationships identified by IRIDESCENT. Four out of the five genes on the list have high obs/exp ratios, suggesting their presence on the list is due to strong relationships with the specific members of the set.

Genes associated with the set of genes in each GO category, but not within the category, were output for further analysis. A total of 163,791 new annotations were predicted. Associations by co-mention are gene-specific, since no attempt to discern species is made when scanning MEDLINE, but the list compares the known ontology annotation of each species within Locuslink to the identified literature association. Thus, a number of the predicted associations on the list will be for genes in species in which the ontology association has not been annotated, but may be annotated in a homolog. For example, the gene GRM3 (metabotropic glutamate receptor 3) is currently annotated with the GO term "synaptic transmission" in humans (Locuslink ID# 2913) but not in rats (Locuslink ID# 24416).

We sampled a subset of these suggested ontology additions to estimate how many genes potentially belonged to the ontological category yet were not documented as such and how many suggestions could be considered false-positives. We did this given the caveat that official inclusion in some ontological categories is not always obvious. For example, if a gene regulates microtubule binding, but does not itself bind to microtubules, does it belong in that category? Could upstream genes critical to a biological process, but not specific to it, be considered part of the same category as other genes in the process? Nonetheless, we evaluated each suggested addition as to whether or not its inclusion in the ontological category it was associated with could be considered reasonable by asking whether or not the literature associated with the gene suggested that it play a direct role (biological process/molecular function) or be localized in the appropriate cellular compartment (cellular component). We randomly chose 50 of the entries and conducted a literature search using the gene name(s) in concert with ontology keywords/phrases, trying various search combinations. Of the sample surveyed, 26 (52%) played a role in or were a part of the ontological category, 12 were related to the category in some way but did not belong in it, 9 genes were not related in any direct or obvious manner, and 3 genes represented erroneous associations due to ambiguous gene symbols (*drosophila's* "urogenital" gene, CCT2 which stands for "chaperonin subunit 2" in mammals but "phosphocholine cytidylyltransferase 2" in *drosophila*, and MT2 which stands for "metallothionein 2" in mammals but "methyltransferase 2" in *drosophila*).

To aid in user evaluation, each gene in the list was hyperlinked to its Locuslink ID in an Excel spreadsheet and posted on the web at:

http://innovation.swmed.edu/IRIDESCENT/GO_relationships.htm

Given the volume of genetic information in the literature and the limited amount of time available to curate and develop ontologies, this type of approach can be highly useful in aiding the process. What might potentially be of use, although yet to be determined, are the relationships ontological categories have with other, non-gene objects such as diseases, phenotypes, chemicals or drugs. These types of relationships could suggest the creation of new ontological categories.

Since addition of noise (i.e. unrelated or random entries) to any set of related objects will reduce their Obs/Exp ratio and obscure existing commonalities, this will be problematic in experiments where a number of interrelated subsystems are present within a much larger whole. The quality of the output and reliability of the calculated observed to expected ratio will depend upon the ability of the experimenter to accurately define a set of interest.

IRIDESCENT has also been applied to a number of microarray experiments to identify commonalities within sets of transcriptional responders, elucidating general "themes" within the analyzed sets as well as suggesting additional genes that should be studied due to their relatedness, yet were not on the array experiment analyzed. It has also proven useful as a positive control for the existence of a cohesive transcriptional response. That is, if an array is analyzed and the average Obs/Exp ratio falls within 2 standard deviations of the random mean, it can be concluded that the genes that responded have little to do with each other – at least in terms of how they have been studied within the literature.

One study, for example, focused on the dendritic cell response to external stimuli. IRIDESCENT was used to determine if the responding genes were cohesive as a set, confirm that the genes were involved in the immune response, and identify additional genes related to the responders that were not on the microarray[157].

4.8 Historical Analysis of Indirect Connections

The validation studies directly demonstrate the utility of the system as applied to real-world problems. Another, more theoretical approach could also prove to be useful in quantitatively ascertaining the utility of IRIDESCENT – the analysis of historical predictions. If the date a relationship is first observed is recorded, then this information can be used in a query to identify implicit relationships only known before that date. Thus, we can compile a list of direct relationships existing as of the date specified and in turn generate a list of indirect connections that were known at the time. We can then compare this list of historical implicit connections to the modern-day direct connections to determine how many of IRIDESCENT's predicted relationships would come true.

One can envision two basic ways by which scientific research enables the progress of knowledge: By a completely *de novo* discovery, or based upon existing knowledge. A *de novo* discovery might be completely accidental or could come from systematic testing of random approaches that culminates in a connection that could not have been anticipated otherwise. Similarly, existing knowledge can lead to explicit hypotheses, whether specific (e.g. A and C interact) or broad (e.g. a target protein of interest could be isolated from a yeast two-hybrid screen). Historical discovery will be composed of a mixture of these two basic

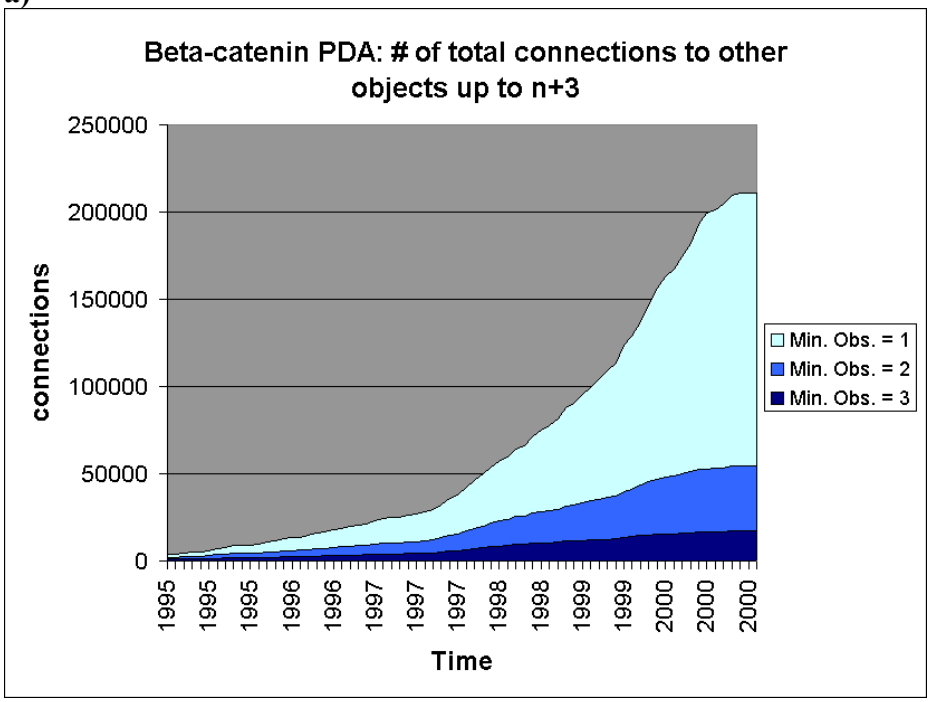
types of discoveries. We can attempt to measure the discoveries that were achievable by knowledge-based reasoning, whether they actually were achieved in such a manner or not, by cataloging the relationships an object has with other objects. At any given point in time, an object should have a number of direct relationships with other objects as well as a number of indirect relationships with the objects that these objects are related to. One would suspect that some number of these indirect relationships might someday be discovered to be direct relationships, much as the migraine-magnesium connection that Swanson made[38].

To provide an example of such connections, let's assume that in 1995, A (a gene) is discovered to be related to B (a disease). At this time it was also known that B was related to C (a phenotype). One could reasonably surmise there might be a connection between A and C, depending on the nature of the relationships. Perhaps the phenotype is seen in other diseases and A may be responsible, in part or in whole. For some, a connection may be obvious and the A-C connection would be quickly closed with documented research. But in other cases, how any given relationship is relevant may not be obvious until further information becomes available. The human bottleneck to discovery, as Swanson illustrated, is the limitation of human expertise in knowing exactly what is relevant. Most of the time, it is difficult to objectively say what sorts of key words, phrases or relationships are directly relevant in making an informational connection. Such relevant secondary connections may not have been made because the connection is largely tangential to the primary focus of the research.

A set of 1,000 abstracts was downloaded from PubMed by searching on the keyword "wnt", which is a developmental gene studied primarily in *Drosophila*. Another group of

1,270 abstracts was downloaded from MEDLINE using the keyword “beta-catenin”. Beta-catenin is a protein involved in the formation of adherens junctions in mammalian epithelia[158] and located on human chromosome 3p21, a region important because of its implication in tumor development[159]. We will refer to objects as *n* and the objects they are directly associated with as *n+1*. Objects directly associated with the *n+1* objects but not *n* are implicitly related and are referred to as *n+2*. Figure 36a shows how the number of total connections increases exponentially over time while Figure 36b shows how many objects known today to be direct connections were only known then to be connected indirectly, through intermediates (# of different intermediates not shown). Because some connections may be spurious, the minimum number of observations required to establish a downstream connection were varied between 1 and 3. The minimum number of connections between n and n+1 were kept at 1 because, presumably, we want the system to be sensitive to new discoveries related to our object of interest and at the same time give us downstream connections that are established.

a)



b)

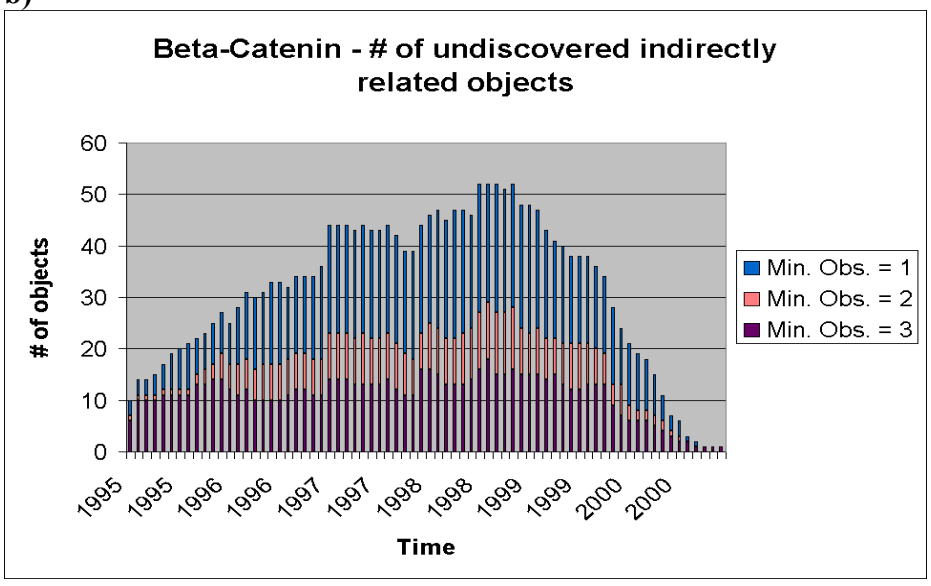


Figure 36: Objects related to the gene Beta-catenin and the effects of varying the minimum number of observations for a connection to be considered valid. **a)** The growth in the total number of connections is exponential with time. **b)** A retrospective look at how many objects were known to be related to Beta-catenin indirectly at any given point in time. As minimum observation requirements are relaxed, the total number of objects goes up. Since we are using present-day direct connections to evaluate how many undiscovered indirect connections

existed at the time, the graph necessarily falls to zero as it approaches the present-day (February 2001 with this data set).

Analysis of this specific test set centered around one object (wnt) and all indirect associations are derived solely from within that set of literature, which we will call a Primary Domain Analysis (PDA). We can assume that when an object such as "beta-catenin" appears in an abstract, it falls into one of three general categories: It is either the primary focus of the article, a direct but secondary consideration (e.g. parameter varied to study effects on another object), or of tangential concern (e.g. mentioned as a member of a gene family, part of the background on why another object is being used, etc.). We can anticipate that the behaviors of the graphs would change depending upon how many connections were already known at the time an object was discovered. We examine in Figure 37 how indirect connections expand as we move beyond the PDA to incorporate more prior knowledge. As all of the graphs show, the percentage of indirect connections of modern-day relevance declines over time. This is expected, because we are attempting to peer into the past using significantly more knowledge about relationships as the amount of time increases. That is to say, what was known yesterday is not that different than what is known today, and thus not many discoveries could be predicted on such a short time frame. However, what was known 10 years ago is significantly less than what is known today, and thus we can identify more areas in which discoveries have yet to be made.

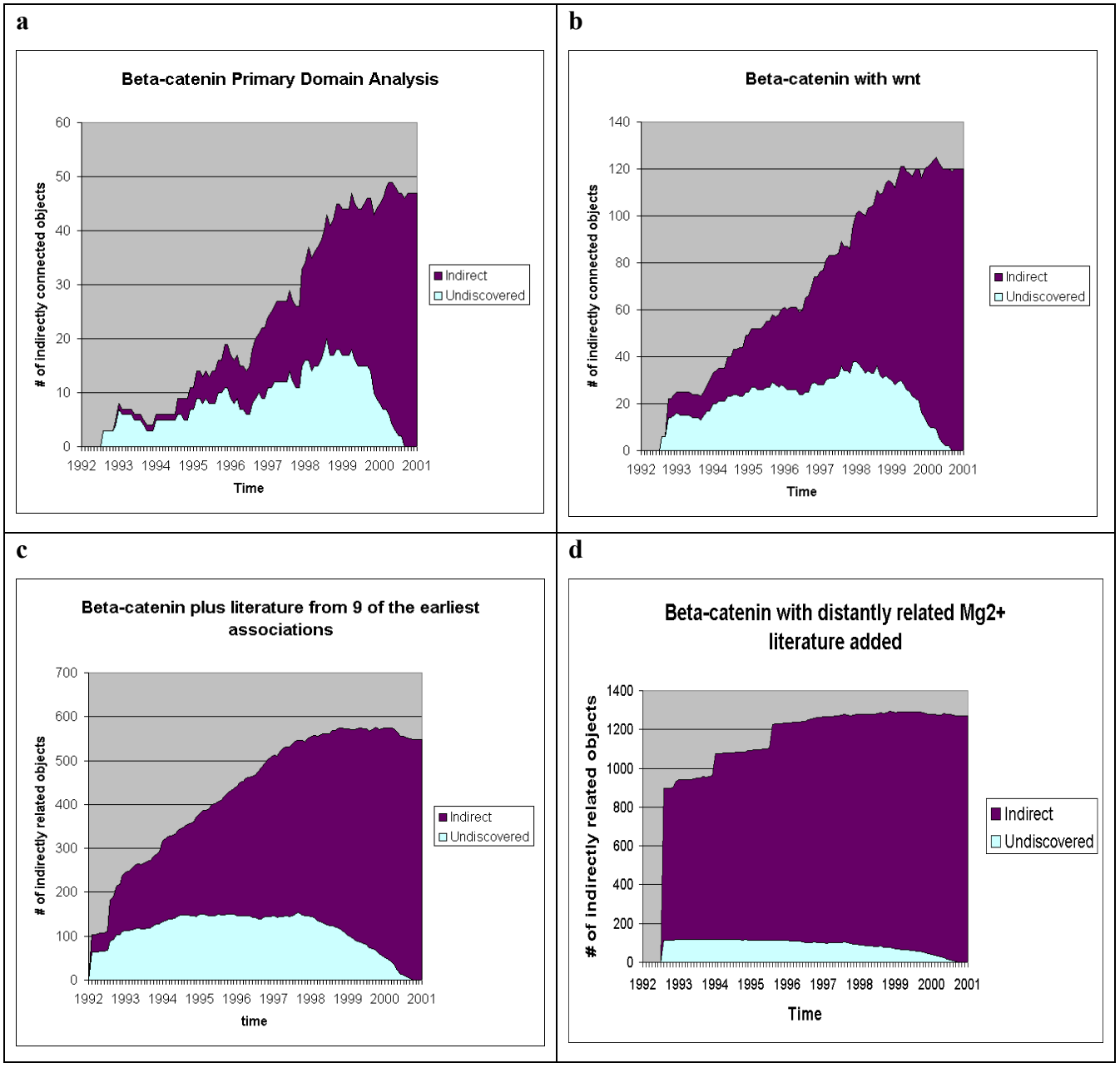


Figure 37: Graphs of the total number of objects indirectly associated with beta-catenin over time. **a)** Primary Domain Analysis(PDA) using only the 1,270 abstracts obtained by searching MEDLINE with the keyword "beta-catenin". This set only goes as far back as February 1992. **b)** Addition of literature involving wnt, an object closely related to beta-catenin, to the PDA. The wnt set dates back to March 1989 and brings the total amount of abstracts analyzed to 1,970. **c)** Further addition of prior literature on the earliest direct associations with beta-catenin that IRIDESCENT was aware of (before 1993): Wingless, alpha-catenin, armadillo, N-cadherin, E-cadherin, plakoglobin, uvomorulin and p120. There were 4,028 total abstracts in this extended set. **d)** Addition of 9,490 abstracts obtained from

searching on the MeSH domain “magnesium” and keyword “increase*”. While the literature was expected to be only remotely related, this graph shows how the addition of only a few indirect connections can greatly expand the number of total connections. Note the sudden jumps in 1994 and August 1995. In 1994 the addition of EGF (Epidermal Growth Factor) and connected the two literatures, as it shares a number of relationships with magnesium. In 1995, the direct connection of “calcium” to beta-catenin further expanded the number of indirect connections.

4.9 Future directions

The primary purpose of developing IRIDESCENT on a Windows-based platform in Visual Basic 6 (VB6) was to provide a means by which a method could be easily implemented and tested within a reasonable time frame, but other implementations would offer performance improvements. Moving development to a Unix-based platform would more readily enable multiple processors to be included in the processing of text and database searches. VB6 is not currently implementable on Unix-based platforms, so IRIDESCENT would have to be converted to another language such as C or even Java. The process of reading in and processing abstracts is highly amenable to distributed processing. One drawback of the current implementation is that if a user wants to analyze an object not within the ORD, it is easy to add but then all of MEDLINE must be re-processed. A routine could be written, however, to process only those records relevant to the new object and disclude all other relationships that were already processed from being reprocessed.

On a basic level, there is always room for refinement of object recognition. ARGH proved to be very useful in identifying spelling variants, but some variations are harder to recognize. Many genes are the subunit of a larger protein and, as such, will be referred to in many ways that a simple text alignment as described in Tables 7a and 7b would not easily

pick up. It might be referred to as “TNF-alpha receptor subunit 7-beta” or “beta subunit of TNFAR7” or “TNFAR7 beta heterodimer”. The upside of the problem is that the effort of writing long phrases is almost always reduced to an acronym (e.g. TNFAR7b), but the downside is that the acronym can be written in different ways (e.g. TNFR7-beta). Ambiguity within database entries is also a problem as noted earlier for the gene MT2 in Locuslink, standing for “Methyltransferase 2” and “Metallothionein 2”. IRIDESCENT easily handles the ambiguity, assuming the two definitions are different entries. If the two are erroneously recognized as synonyms, then that’s where the error arises.

More object classes could be assimilated into the ORD, but one should bear in mind the limitations of this type of analysis. When a tentative relationship is found between a gene and a disease or a phenotype and a chemical, the biologist or medical expert can foresee the potential for an interesting relationship (e.g. the gene is involved in the disease, the chemical causes a phenotype). When another object class such as a cell line name is added to the ORD, the nature of the relationship is not as obvious. Several genes co-mentioned with a cell line in the literature might simply imply that they were studied using the cell line, the rationale simply being availability rather than biological meaningfulness. A number of other objects might have been co-mentioned for reasons not really cohesive as well. The classes of objects assimilated were chosen for just this reason – potential relationships between them had readily apparent implications.

An approach that might be useful in the absence of a well-developed object class database such as MeSH would be to identify objects *de novo* within text based upon semantic

patterns. In Figure 38, for example, IRIDESCENT is used to identify certain keywords that suggest that the name of a disease is nearby within a sentence.

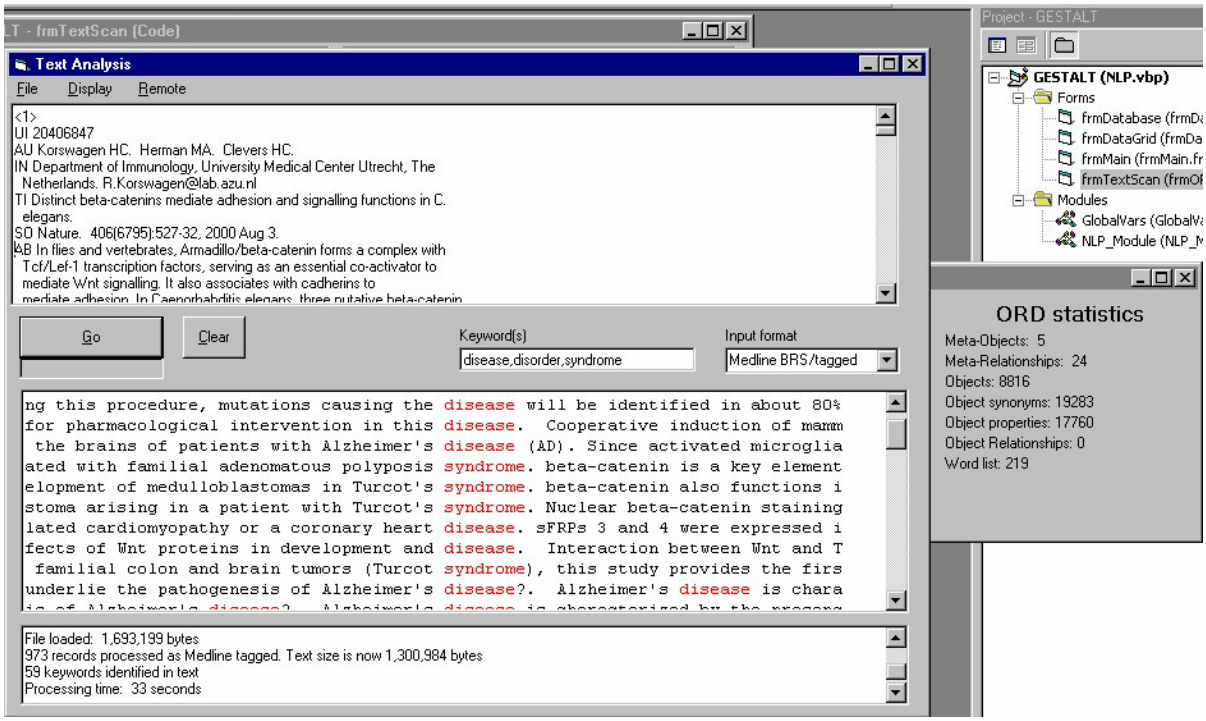


Figure 38: MEDLINE abstracts are scanned in IRIDESCENT for the presence of contextual keywords that would suggest the presence of a nearby object class. Words such as “disease”, “disorder” and “syndrome” suggest that a name might frequently precede these words. Such a name, when certain pattern-matching rules are applied, could be elucidated to form a knowledge base encompassing an object class. Approaches discussed early in this manuscript have been used to identify protein names by such contextual clues, but has yet to be applied to disease-based names.

Similarly, besides expanding the object classes recognized, one could expand the literature processed by IRIDESCENT. It would be relatively trivial to extend it to other abstracts, but more problematic would be to extend analysis to full-text. Full text access to scientific reports within MEDLINE has been an ongoing desire of the biomedical community for some time. Having more text available would enable IRIDESCENT to recognize more

relationships. We would anticipate, however, that it would also provide a much higher rate of false-positives since many different topics may be discussed within an article. Background may include tangential information, materials and methods sections may include information on genetic constructs or chemical compounds used in assays that have nothing to do with the primary objects of study, and conclusions may be overly speculative in nature. In short, much more analysis will have to be conducted on how to handle full-text articles, including each section within a report and the variability in naming of these sections that many journals have (e.g. The journal "Bioinformatics" has a "Systems and Methods" and an "Algorithms" section – names rarely observed in other journals). There is indeed a large potential for greater recognition capacity by using full-text, but abstracts intentionally constrain the user to present a summary of the most pertinent aspects of the study with limited speculation.

Adding contextual information is a future aim, as the adirectional relationships modeled here in this project are limiting. If the nature of a connection is known such as A upregulates B (A↑B) or A affects the activity of B (A→B), then we can engage in a bit more sophistication when searching for implicit relationships. We would expect to see, for example, in our chlorpromazine-cardiac hypertrophy analysis that the number of A↑B connections would be similar to the number of C↓B connections (note the directionality in this notation, as B↓C is not equivalent to C↓B), since chlorpromazine antagonizes the effects of cardiac hypertrophy. This is where NLP or more sophisticated IE techniques will be of use. However, it will still be problematic to represent context. Take the insulin-glucose relationship as an example. Insulin signaling (A) is supposed to increase the import of glucose (B) into target tissues. So with reference to intracellular concentrations, A↑B, but

with reference to blood concentrations of glucose, A↓B. Even then, this is still not completely accurate. It would better be written as A(extracellular)↑B(intracellular) or A(extracellular)↓B(extracellular). These problems fall in the general domain of knowledge representation – somewhat akin to translating between English and symbolic/algorithmic.

There is also the issue of subnetworks within the whole relationship network built by IRIDESCENT. In the Obs/Exp formulas, we score using all objects in the set. It is quite possible that when a smaller subset is examined, the paired relationships are far more exceptional than when the large subset is. If the set B consisted of objects with 1000, 100 and 10 connections within a network, the expected # of connections with another object A might be 1.0. One would expect that the object with 1,000 connections would be the most likely partner. But what if it was the node with 10 connections? Under the formula we have developed here, it would be unexceptional. It can be somewhat problematic, however, to evaluate the connections based only upon the shared ones and not the total number of chances to connect offered by the set.

Recognizing *drosophila* names is a recurring problem within IRIDESCENT. The gene names are synonymous with common words, so no method of strict term-based recognition can be applied. Contextual approaches will have to be developed. For example, if the word “frizzled” is used in an abstract, it could be referring to appearance or the *drosophila* gene name. We might be able to presume that if the abstract specifically states that *drosophila* is being used for the study, the word then refers to the gene name.

As mentioned earlier, with further development of the Knowledge Representation ability of the system, we could then use other connections and perhaps some

inductive/deductive logic to hypothesize what sort of properties or behaviors an object should have given similar sets of relationships among other similar objects. Some work towards this end has been accomplished and a discussion can be found in the Appendix.

All things considered, perhaps the best future direction to take with IRIDESCENT is simply to continue to use it in an attempt to discover new and relevant relationships.

Appendix

IRIDESCENT's Databases

a. Object-Relationship Database (ORD):

Somewhat of a misnomer, this database now just holds objects. It was designed to hold relationships as well, but is not normally used to do so because of the high storage overhead associated with storing all literature references associated with each relationship. The streamlined ORD (SORD) was created to store relationships in a more compact form, and is the central relationship database used for queries. The ORD is the central repository for object names and synonyms to be recognized within text, as well as user-defined entries and excluded entries. ORD also contains information from the Genome Ontology project so that genes can be analyzed in terms of their ontology as well as literature references.

Table	**Field**	**Description**
TblMetaRelationship	**ID**	**Key**
	Type	General type of relationship (e.g. association, increase, subset, etc.)
	Subtype	Relationship subclass (e.g. association.physical, location.cellular, ID.Genbank, etc.)
	WordForm	Grammatical form of the verb
	Keyword	Keyword indicating a relationship
TblObjRel	**ID**	**Key**
	Object1	Object #1
	Relationship	Has this relationship to
	Object2	Object #2
	Source	Source of this information
	SentenceNum	Sentence # that this relationship was found in
	Date	Most recent date this relationship was seen (yyyymmdd)

	Observed	# of times this relationship was observed
TblObjectSynonyms	RecordID	Unique number assigned to an object and all of its synonyms
	ObjectName	Standard name encompassing an object and all of its synonyms. Initially determined from input databases, it is changed after processing all of MEDLINE to the most frequently used synonym name.
	ObjectSynonym	**Synonyms for the standard object name**
	ObjectType	(G)ene, (D)isease, clinical phenotype (CP) or small molecule/drug (SM)
	SourceID	Origin of this object.
	Net_freq	# of relationships in the network an object has
	Int_str	Integral strength for an object – used in calculations
	Ver_str	Integral sum of the veracity scores
	Occurances	# of times this term was found in MEDLINE
	Chromosome	Chromosomal location (genes only)
	CAPS_flag	Flag to ensure the object is only recognized if textual capitalization patterns match the database entry
	AA_flag	Ambiguous Acronym flag. Object will not be recognized unless the acronym is defined within the textual input.
	CW_flag	Common Word flag. When retrieving records from PubMed, these synonyms will not be used in the search.
	Asyn_flag	Acronym can be used to refer to two different objects in the database. Not used.
TblCommonWords	**ID**	**Key**
	Word	Word to be deleted from the database during construction
	Category	Why this word is deleted (vague, common, error, etc.)
TblUserAddedObjects	**ID**	**Key**
	ObjectName	Object Name
	ObjectType	Object Type
	ObjectSynonym	Object Synonym
	SourceID	User name entering this new object
	CAPS_flag	CAPS flag
	AA_flag	Ambiguous Acronym flag
	CW_flag	Common Word flag
TblLinker	Word	(not necessary)
	KeyPhrase	Phrase to be recognized within text
	Type	How this phrase should be treated by the IE engine (e.g. link two concepts, negate a concept, etc.)

TblLLOntology	**ID**	**Key**
	Onto_ID	Genome Ontology ID number
	LocuslinkID	Locuslink ID number of gene
TblOntology	**Record_Key**	**Key**
	ID	Ontology ID
	Description	Ontology description
	Master_node	ID number for next highest branch in tree
TblTemp	**ObjectName**	**Name of object**
	Obj1Type	Object 1 type
	Obj2Type	Object 2 type
	Date	Used for historical analysis
	AB_min_obs	Minimum A-B observations to count as a relationship
	BZ_min_obs	Minimum B-C observations to count as a relationship
	Min_abs_obs	Minimum # of abstract observations to count as a relationship
	Min_SAR	Minimum Sentence to Abstract Ratio to count as a relationship

Table A1: Tables and their fields within the Object-Relationship Database. Record Key fields are indicated by **bold** type.

b. Streamlined Object-Relationship Database (SORD):

The SORD was created to hold all identified relationships between objects identified by the processing of all MEDLINE records. Relationships between objects are stored in the table tblCo_mention in terms of the unique RecordID number associated with each object. SORD contains a copy of the table tblObjectSynonyms from the ORD version used to create it so that all relationships (stored by RecordID) can be traced back to the name and synonyms used to recognize them. It is critical that the tblObjectSynonyms from the ORD version used to create each SORD be copied to the SORD either before or after processing all of MEDLINE records. This is necessary to ensure that each version of the database used to scan MEDLINE is preserved in a stand-alone version while development and refinement of the

ORD can continue. For example, as routines are changed to load, process and refine database entries, entries in the ORD will change. This will not be reflected in terms of identified relationships until the next time MEDLINE is processed. Thus, it is important to preserve the exact synonyms and database flags used to create each SORD so that all relationships can be traced.

TblTemp is a very important table in this database. Values to be input into queries must be stored here before the queries can be run. For example, when querying a list of objects for what they have in common, the user will input the RecordID for each object of interest into the table tblUserList. Then, the name of the userlist to be analyzed will be put in this table. All related queries will then use this value. Similarly, when conducting an implicit analysis, the RecordID for the object of interest will be input here and then all related queries can be run.

Table	**Field**	**Description**
TblCo_Mention	**ID**	**Key**
	Object1	Record ID for related object (object 1 will be the lowest numerical value between objects 1 and 2)
	Object2	Object 1 is related to this object (given as Record ID)
	Sent_Obs	# of times a co-mention has been observed between the two within a sentence
	Abstr_Obs	# of times a co-mention has been observed between the two within an abstract
TblTemp	**ID**	**Key**
	ObjectID	RecordID for object of interest (implicit analyses)
	UserList	Name of userlist (for shared relationship analyses)
	Min_Strength	Minimum strength to considered a relationship
TblUserList	**ID**	**Key**
	Name	Name of user-defined list of objects (e.g. Jonathan1)
	RecordID	RecordID for each object being analyzed within a list

Table A2: IRIDESCENT's co-mention database format

c. MW_Dictionary:

The Merriam-Webster dictionary was obtained from Project Gutenberg in electronic format. It is used to identify "common" words – that is, words used in non-scientific speech. This database is the summary of the electronic processing of all entries.

Table	Field	Description
TblMW_words	**ID**	**Key**
	Word	English word
	Frequency	# of times this word was seen in the Merriam-Webster Dictionary

Table A3: The MW dictionary database format

d. ARGH and Stemmed ARGH databases:

The stemmed version of ARGH is used to compare ORD entries. It differs only in that entries are "stemmed" (word endings and special symbols removed). The table tblObjectSynonyms should be copied from the most recent version of the ORD to be able to run queries. This is not necessary for IRIDESCENT to use the database entries, but rather for independent queries.

Table	Field	Description
TblAcronym	**RecordKey**	**Key**
	Acronym	Acronym
	Definition	Definition for the acronym
	DefStemmed	Stemmed definition
	Observed	# of times this acronym-definition pair was seen within MEDLINE
	First_Observed	Date this pair was first observed
	Context_example	PMID of the first observation
	Flag	Processing flag

TblObjectSynonyms	RecordID	Unique number assigned to an object and all of its synonyms
	ObjectName	Standard name encompassing an object and all of its synonyms. Initially determined from input databases, it is changed after processing all of MEDLINE to the most frequently used synonym name.
	ObjectSynonym	**Synonyms for the standard object name**
	ObjectType	(G)ene, (D)isease, clinical phenotype (CP) or small molecule/drug (SM)
	SourceID	Origin of this object.
	Net_freq	# of relationships in the network an object has
	Int_str	Integral strength for an object – used in calculations
	Ver_str	Integral sum of the veracity scores
	Occurances	# of times this term was found in MEDLINE
	Chromosome	Chromosomal location (genes only)
	CAPS_flag	Flag to ensure the object is only recognized if textual capitalization patterns match the database entry
	AA_flag	Ambiguous Acronym flag. Object will not be recognized unless the acronym is defined within the textual input.
	CW_flag	Common Word flag. When retrieving records from PubMed, these synonyms will not be used in the search.
	Asyn_flag	Acronym can be used to refer to two different objects in the database. Not used.

Table A4: The Stemmed acronym database generated by ARGH

Information Extraction efforts

IRIDESCENT currently relies upon co-citations to establish relationships, which are adirectional in nature. Different types of analyses can be conducted if the nature of the relationship is known, such as searching for antagonistic and complementary phenomenon. This can be accomplished by information extraction (IE) methods. IRIDESCENT contains the ability to engage in pattern identification through IE methods, enabling the nature of relationships to be identified. Unfortunately, the IE methods developed have a relatively high false-negative rate associated with them, making them of less utility than co-citations. However, the IE does enable the nature of some relationships to be identified and further development has potential to enable more sophisticated analyses. The work done in developing IRIDESCENT's IE engine will be discussed. Use of an IE engine holds utility beyond that of determining the nature of a relationship, as it can also help to identify certain types of phenomena that would be difficult to locate using traditional information retrieval (IR) methods.

We are all familiar with the standard query-based approach (e.g. PubMed) to searching for items of research interest and, no doubt, familiar with the frequency by which results not relevant to our queries are returned. This interface is simple and intuitive, yet it has its limitations. Counter-intuitively, the more information that becomes available the harder it becomes to find items of interest. For example, suppose a researcher is interested in identifying phenomena known to cause an increase in intracellular magnesium levels. A Boolean keyword query might consist of the words "magnesium" and "increase", or some variants thereof. "Increase" is a common word, and therefore most results returned will not

be those in which the word “increase”is used to modify “magnesium. In the entries where the two words occur closer together, one may still not be certain whether the returned results are about the effects of an increase in magnesium on something or about something that causes an increase in magnesium. A phrase-based search like “increases magnesium levels” is more likely to give more precise results, but there are numerous ways a writer could phrase such a concept. For example, it could be written as “found to increase magnesium concentration” or “observed elevated intracellular levels of magnesium”. If you construct a Boolean query to ensure both “increase” and “magnesium” are found, you are then faced with a number of false-positive results containing phrases matching your search words such as “…can cause intracellular **magnesium** depletion and an **increase** in intracellular calcium”. Because one would also want to ensure that word root variants like “increasing” and “increased” are not left out, one could employ the use of wild cards like “increas*”. Wildcards will help make the search more comprehensive, but also quickly increase the number of false positives. Worse, synonyms that describe the same phenomena, such as “Mg^{2+}” or “elevation”, “rise” and “higher levels of” are not included in the search. Table A5 illustrates the keyword variance in returned results from MEDLINE searches. If nothing else, it is evident that there is a need for more efficient way of searching the literature for phenomena of interest because there are too many false positive results.

Query	**Results**[a]
Magnesium	58,011
Mg^{2+}	22,141
Magnesium (MeSH: all subheadings)	46,151
Increase*	1,396,427
magnesium and increase	5,773

magnesium and increases	2,171
magnesium and increased	7,936
magnesium and increasing	2,241
magnesium and (increase or increases or increased or increasing)	13,291
"increases magnesium"	13
"elevates magnesium"	0
"higher magnesium concentration"	5
(MeSH: Magnesium) and increas*	9,490

Table A5: The results obtained from a query will vary depending on how it is constructed. [a]Results in table are from all MEDLINE records as of 11/21/2000, obtained using the Ovid search engine.

MEDLINE has attempted to deal with the problem of synonymous names for phenomena by providing a method of mapping words to a controlled vocabulary for informational categorization, called MeSH (Medical Subject Headings)[160]. MeSH allows the mapping of a word or phrase onto topical (Subject Headings) searches, which helps include synonyms in a search and enables the ability to find documents where commonly used keywords relevant to the study may not be included in the title or abstract. Even though not all biomedically relevant synonyms have been mapped, MeSH usually works very well when searching for information on individual topics, and even allows for selection of subtopics. However, MeSH is primarily limited to nouns and will not enable you to focus your search on the types of interactions such nouns might have. Neither does it provide context or an efficient way of elucidating relationships between one item of interest and others.

It is this incredible amount of data and information that is available to us that, ironically, makes it harder to find relevant information. Scientists use a variety of shortcuts to

aid in this task, such as narrowing the range of journals they read to ones they consider focused and high-quality in the hope that relevant information will be published there as well as attending national meetings to keep in touch with colleagues and current research in their field. While this is effective to an extent, they both rely upon other people who are just as limited as they are to provide coverage and screening of information. And unfortunately, while these strategies help keep people informed, it does not put them at the forefront of knowledge.

Meta-Relationships: Many ways of describing the same thing

Objects have their synonyms, whether a word or a phrase, that can enable a many-to-one mapping. Similarly, descriptions of actions, reactions, changes, variance or any other type of relationship an object might have with another object can be described in many different ways. Determining synonyms for relationships is not sufficient because we are interested primarily in the general type of relationship. Such a general type of relationship, or categorical clustering, could be said to encompass a large variety of interactions we will term a Meta-relationship. For example, observations could be made on the interactions of two proteins and described using terms such as “associate”, “dissociate”, “adhere” or “bind”. Whereas “associate” may have a subtly different meaning than “bind”, it is not entirely incorrect to catalog the interaction under a general terms such as “physical association” than under each individual heading. An example of such categorical clusterings can be seen in NCI’s MedMiner, which attempts to group together sentences containing search keywords into a general category[161,162], but a more accurate comparison would be what the NIH’s

UMLS system calls a "semantic relationship" and similarly encompasses a broad number of terms[163]. There are four basic types of Meta-relationships included in this project: Positive effect (increase), negative effect (decrease), physical association and logical association.

I have compiled a list of root forms of the keywords denoting such relationships (as shown in Table A6), along with how common their usage is in MEDLINE. Word spelling variants (e.g. releaser vs. releasor, disassociate vs. dissociate) have been checked for each one and will not be included because they comprise a small portion (typically < 2%) of their usage. Terms and phrases included in Meta-relationships can be added and modified to fit future project goals if necessary, an example of some is given in Appendix Table along with how they might be used.

These specific Meta-relationships were chosen for the purposes of end-utility. Ultimately, we must define not only the types of things we are interested in studying, but what we are interested to know about them. General associations and categorizations can be useful for a variety of purposes, and obtaining quantitative, rather than qualitative, changes enable the system to search for complementary and antagonistic phenomena. Knowing the phenotypes of a disease and which other phenomena are responsible for generating similar phenotypes and opposite phenotypes can aid in determining the origins of the disease and searching for potential cures. For example, a medical condition may cause a decrease in alcohol dehydrogenase (ADH). This quantitative phenotype would be of interest to the system because a way of treating this symptom would involve increasing ADH levels. The same condition may have another phenotype of liver toxicity, but the opposite of toxicity is hard to define even though we could envision possible antagonistic words like "restoration",

"regeneration" or "growth". Toxicity is a relatively generic term, qualitative in describing a phenomenon and difficult to define what its antagonist or complement might be. However, it might be useful as a link to understanding if one is working with patients suffering from liver toxicity due to unknown causes.

ROOT Meta-relationship keywords in MEDLINE

As of 12/18/2000

Increase		**Decrease**	
		Degrad*	(86,234)
		Ubiquitinat*	(1,244)
Activat*	(415,310)	Inactivat*	(77,008)
Enabl*	(53,244)	Deactivat*	(3,877)
Induc*	(905,161)	Block*	(271,393)†
		Repress*	(28,562)
Increas*	(1,396,427)	Suppress*	(172,959)
Upregulat*	(13,369)	Decreas*	(686,727)
Up-regulat*	(379,907)	Downregulat*	(8,636)
Rais*	(98,364)	Down-regulat*	(24,282)
Elevat*	(209,038)†	Depress*	(182,205)
Enhanc*	(296,430)†	Reduc*	(769,287)
Releas*	(275,316)	Inhibit*	(743,450)
Stabiliz*	(54,136)	Sequest*	(12,092)
Higher	(518,292)	Destabiliz*	(5,965)
Agonist*	(103,108)	Lower	(410,993)
		Antagonist*	(167,073)
Association, Physical		**Association, Logical**	
Bind*	(519,336)	Modif*	(245,349)
Cleav*	(63,683)	Regulat*	(382,435)
Cataly*	(98,809)	Acetylat*	(12,142)
Interact*	(321,075)	Phosphorylat*	(78,924)
Dissociat*	(62,378)	Mediat*	(323,761)†
Heterodimer*	(10,190)	Control*	(935,431)†
Complex*	(356,990)†	Affect/s	(187,119)
Associat*	(879,398)†	Effect*	(1,872,664)
		Correlat*	(475,991)
Symptom*	(267,651)		
Abnormal*	(283,924)		
Deficien*	(153,465)		

Table A6: Words used to signify a type of relationship (Meta-relationship) between objects and how many MEDLINE abstracts contain a root form variant of the word. Asterisks (*) are used to denote wildcards. †Noun form of verb or alternative use of words throws off accurate estimate of total (e.g. "blocks of time", "elevator accidents", "enhancer element", "complex

behavior", "association of physicians", "experimental control", "mediated discussion groups")

Quantitative relationships are those in which verbs and verb phrases such as "increases", "upregulates", or "elevates the levels of" are used to describe them. Qualitative relationships are those that can be quantifiably measured, but are put in broader terms of "more" or "less" of a characteristic. They are denoted by the use of adjectives or nouns such as "hypertrophic", "hypoplasia", or "megalencephaly".

Relationships: Linking A to B

Relationships between objects are stored in terms of their Meta-relationship, but the same type of relationship can be worded in the literature with a variety of different grammatical constructs, as shown in Table A7. Being able to extract these relationships ("inhibit": Meta-relationship=decrease) as well as their objects ("wnt", "the quaternary complex") is a critical part of the project. The original sentence used in this example reads "wnt signaling somehow inhibits the kinase activity of the quaternary complex"[164].

Phrase	**Form of the verb "to inhibit"**
Wnt signaling acts to inhibit the kinase activity…	Verb (root form)
Wnt signaling somehow inhibits the kinase activity…	Verb (3rd sing. pres.)
QC kinase activity is somehow inhibited by wnt…	Verb (past)
Wnt signaling somehow inhibiting kinase activity…	Verb (pres. particip.)
Wnt signaling somehow leads to the inhibition of kinase activity…	Noun form (gerund)
Wnt signaling somehow acts as an inhibitor of kinase activity…	Noun form (sing.)
Wnt signaling is one of the inhibitors of kinase activity…	Noun form (pl.)
…study the QC inhibition. **It** is somehow due to wnt signaling…	Pronoun reference
Wnt signaling somehow has inhibitory effects upon the QC…	Adjective

Wnt signaling somehow becomes inhibitive towards the kinase…	Adjective

Table A7: The many grammatical ways to describe the effect of the gene wnt upon the kinase activity of the quaternary complex.

Semantic Parsing and Information Extraction

Abstracts are input and parsed sentence by sentence, checking for Meta-objects and relationships. A flowchart of the IE portion of IRIDESCENT is shown in Figure A1.

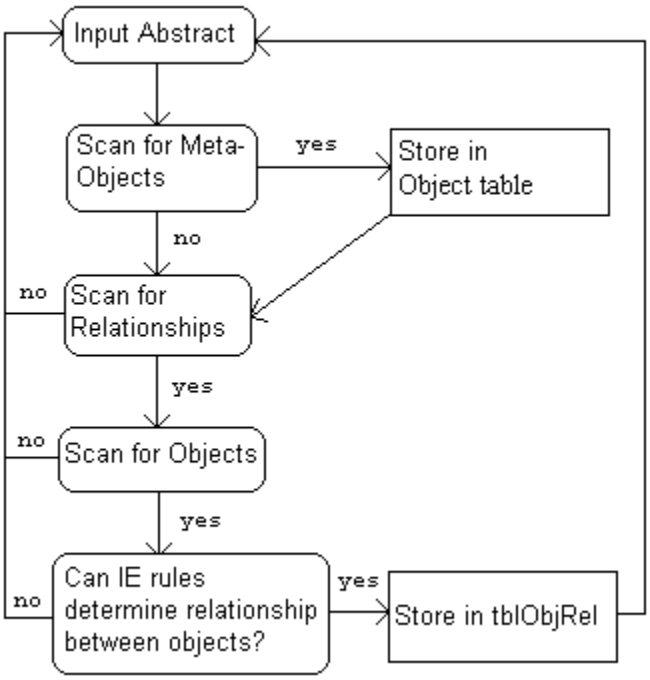


Figure A1: The Information Extraction (IE) step of IRIDESCENT involves scanning sentences from abstracts (from MEDLINE or other sources) for Meta-objects to be cataloged in the Object table (tblObjectSynonyms). Then the text is scanned for the Meta-relationship keywords that indicate a possible relationship. If a relationship is found, IRIDESCENT then scans for objects, if less than two objects are found it goes to the next sentence. If a relationship and two objects are found, IRIDESCENT sends the sentence to the grammar parser and then to the IE rule determination set in an attempt to properly catalog the relationship. If a good match is found, it is stored.

Other potential Meta-relationships and their uses		
Meta-relationship	**Keywords/patterns**	**Usage**
Subset.family	The * family;	Members of the same family can be assumed to have similar properties.
Similarity.sequence	Homologous; orthologous; paralogous	Homologs will be assumed to have the same roles and associations as their counterparts in other species
Similarity.structure	Domain is similar to; has a conserved fold	Structural similarities could mean functional similarities. If a domain is associated with a function and a protein has that domain, it will be assumed to have that function.
Location.cellular	localiz*; found in; located in; membrane-spanning; transmembrane	Association/exclusion studies
Location.systemic	Expressed in; found in * tissues, found in *cytes	When all else fails, it can be useful to go over a list of all known ESTs expressed only in the specific tissue of interest and suggest one of them based upon functional domain similarity.
Logic gate	and; along with; in addition to; or; but not; without; in the absence of;	Logic gates are the core of complex behavior
Subset	part of the; belongs to the; is within the; is a;	Logical consistency checking of relationships.
Variation	varies/vary in/with *x*;	Correlation can be used for prediction, association or diagnostics as well as a potential window into causation

Table A8: Other potential Meta-relationships that could be used as other Meta-objects are added (e.g. tissue types, gene families, protein domains).

Glossary

ARGH (Acronym Resolving General Heuristic) – A software module responsible for identifying acronym definitions when given.

Artificial Intelligence (AI) – Methods by which computers perform tasks that Humans would deem "intelligent" such as identifying pictures, understanding spoken words or written text, and solving problems.

Data(pl.) **or Datum** (sing.) – A measurement or statistic. Fundamental unit of **information**.

Fuzzy Relationship – A confidence score ranging from 0 to 1 that connects two objects, corresponding to the estimated certainty of relationship.

Gene Ontology Consortium – A group of researchers dedicated to constructing a dynamic, controlled vocabulary for gene function.

IRIDESCENT – **I**mplicit **R**elationship **IDE**ntification by in-**S**ilico **C**onstruction of an **E**ntity-based **N**etwork from **T**ext. A system designed to extract information from databases and textual sources for the purpose of cataloging and understanding the **relationships** of the **objects** contained within.

Information – Factual relationships resulting or derived from a set of data.

Information Extraction (IE) – The process of identifying informational elements of specific interest within textual sources

Knowledge – Sufficient information about a set of objects to make predictions, deductions and/or inductions.

MEDLINE – A bibliographic database curated by the National Library of Medicine, covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. Currently, MEDLINE contains bibliographic information from over 4,600 biomedical journals published in the United States as well as 70 other countries.

Meta-object – A general area of interest defined in terms of a keyword or key phrase. Used to conduct searches for related **objects** of interest.

Meta-relationship – A general type of **relationship** between two **objects** that can be described using one or more terms.

Natural Language Processing (NLP) – Understanding language written in a natural context (i.e. as one would speak).

NCBI – National Center for Biotechnology Information, a subsidiary organization of the National Institutes of Health, located in Bethesda, MD. Its primary function is to provide biological information and tools to analyze it to the scientific community.

Object – A noun or noun phrase corresponding to a biological entity of interest (e.g. genes, proteins, metabolites, drugs, phenotypes, diseases, protein families, protein domains).

Object Recognition Database – Database which contains object names, synonyms, lexical variants as well as the relationships identified between all of them.

Relationship – A non-directional connection between two data entities.

UniGene – A database curated by **NCBI** consisting of complete or partial sequence reads from the transcribed regions of genes.

References

1. CONSORTIUM, T.G.I.S. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=howmany
3. http://chem.sis.nlm.nih.gov/chemidplus/
4. http://www.ncbi.nlm.nih.gov:80/LocusLink/statistics.html
5. http://www.ncbi.nlm.nih.gov/Omim/Stats/mimstats.html
6. http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html
7. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45-8 (2000).
8. http://www.expasy.ch/sprot/sprot-top.html
9. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42 (2000).
10. http://www.rcsb.org/pdb/
11. Hamosh, A., Scott, A.F., Amberger, J., Valle, D. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* **15**, 57-61 (2000).
12. http://www.ncbi.nlm.nih.gov/Web/Newsltr/Summer99/decade.html
13. Andrade, M.A. & Bork, P. Automated extraction of information in molecular biology. *FEBS Lett* **476**, 12-7 (2000).
14. Srinivasan, P. & Rindflesch, T. Exploring Text Mining from MEDLINE. *Proc AMIA Symp* , 722-6. (2002).
15. Srinivasan, P. MeSHmap: a text mining tool for MEDLINE. *Proc AMIA Symp* , 642-6. (2001).
16. de Bruijn, B. & Martin, J. Getting to the (c)ore of knowledge: mining biomedical literature. *Int J Med Inf* **67**, 7-18. (2002).
17. Yandell, M.D. & Majoros, W.H. Genomics and natural language processing. *Nat Rev Genet* **3**, 601-10. (2002).
18. Proux, D., Rechenmann, F., Julliard, L., Pillet, V. & Jacq, B. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Inform Ser Workshop Genome Inform* **9**, 72-80 (1998).
19. Rindflesch, T.C., Tanabe, L., Weinstein, J.N. & Hunter, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput* , 517-28 (2000).
20. Collier, N., Nobata, C. & Tsujii, J. Extracting the Names of Genes and Gene Products with a Hidden markov Model. in *Conference on Computational Linguistics (COLING) 2000* (, 2000).
21. Fukuda, K., Tamura, A., Tsunoda, T. & Takagi, T. Toward information extraction: identifying protein names from biological papers. *Pac Symp Biocomput* , 707-18 (1998).
22. Wilbur, W.J. *et al.* Analysis of biomedical text for chemical names: a comparison of three methods. *Proc AMIA Symp* , 176-80 (1999).
23. Rindflesch, T.C., Hunter, L. & Aronson, A.R. Mining molecular binding terminology from biomedical text. *Proc AMIA Symp* , 127-31 (1999).

24. Ohta, Y., Yamamoto, Y., Okazaki, T., Uchiyama, I. & Takagi, T. Automatic construction of knowledge base from biological papers. *Proc Int Conf Intell Syst Mol Biol* **5**, 218-25 (1997).
25. Thomas, J., Milward, D., Ouzounis, C., Pulman, S. & Carroll, M. Automatic extraction of protein interactions from scientific abstracts. *Pac Symp Biocomput* , 541-52 (2000).
26. Blaschke, C., Andrade, M.A., Ouzounis, C. & Valencia, A. Automatic extraction of biological information from scientific text: protein-protein interactions. *Ismb* , 60-7 (1999).
27. Wong, L. A Protein Interactions Extraction System. in *Pacific Symposium in Biocomputing* 520-531 (, 2001).
28. Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res* **28**, 289-91. (2000).
29. Xenarios, I. *et al.* DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res* **29**, 239-41. (2001).
30. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**, 303-5. (2002).
31. Marcotte, E.M., Xenarios, I. & Eisenberg, D. Mining literature for protein-protein interactions. *Bioinformatics* **17**, 359-63. (2001).
32. Bader, G.D. *et al.* BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res* **29**, 242-5. (2001).
33. Stapley, B.J. & Benoit, G. Biobibliometrics: information retrieval and visualization from co- occurrences of gene names in Medline abstracts. *Pac Symp Biocomput* , 529-40. (2000).
34. Jenssen, T.K., Laegreid, A., Komorowski, J. & Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* **28**, 21-8. (2001).
35. Sekimizu, T., Park, H. & Tsujii, J. Identifying the Interaction between Genes and Gene Products Based on Frequencly Seen Verbs in Medline Abstracts. in *Eights Workshop on Genome Informatics (GIW98)* (, 1998).
36. Swanson, D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med* **30**, 7-18 (1986).
37. DiGiacomo, R.A., Kremer, J.M. & Shah, D.M. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am J Med* **86**, 158-64 (1989).
38. Swanson, D.R. Migraine and magnesium: eleven neglected connections. *Perspect Biol Med* **31**, 526-57 (1988).
39. Swanson, D.R. Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspect Biol Med* **33**, 157-86 (1990).
40. Swanson, D.R. & Smalheiser, N.R. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artifical Intelligence* **91**, 183-203 (1997).

41. Smalheiser, N.R. & Swanson, D.R. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed* **57**, 149-53 (1998).
42. http://kiwi.uchicago.edu/
43. Weeber, M. *et al.* Text-based discovery in biomedicine: the architecture of the DAD-system. *Proc AMIA Symp* , 903-7 (2000).
44. Lindsay, R.K. & Gordon, M.D. Literature-Based Discovery by Lexical Statistics. *Journal of the American Society for Information Science* **50**, 574-587 (1999).
45. http://genome-www.stanford.edu/GO/
46. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
47. Rindflesch, T.C. & Aronson, A.R. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proc Annu Symp Comput Appl Med Care* , 240-4 (1994).
48. Aronson, A.R. The effect of textual variation on concept based information retrieval. *Proc AMIA Annu Fall Symp* , 373-7 (1996).
49. Aronson, A.R. & Rindflesch, T.C. Query expansion using the UMLS Metathesaurus. *Proc AMIA Annu Fall Symp* , 485-9 (1997).
50. Adriaans, P., Fernau, H. & Van Zaanen, M. *Grammatical Inference: Algorithms and Applications: 6th International Colloquium, ICGI 2002*, (Springer Verlag, Amsterdam, 2002).
51. Talbot, C.J. & Cuticchia, A. Human Mapping Databases. in *Current Protocols in Human Genetics* 1.13.1-1.13.12 (John Wiley & Sons, Inc., 1999).
52. McKusick, V.A. *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders.*, (Johns Hopkins University Press, Baltimore, 1998).
53. Kanehisa, M. *Post-genome Informatics*, (Oxford University Press, 2000).
54. Pruitt, K.D., Katz, K.S., Sicotte, H. & Maglott, D.R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* **16**, 44-7 (2000).
55. Maglott, D.R., Katz, K.S., Sicotte, H. & Pruitt, K.D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res* **28**, 126-8 (2000).
56. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res* **28**, 304-5 (2000).
57. Ellis, L.B., Hershberger, C.D. & Wackett, L.P. The University of Minnesota Biocatalysis/Biodegradation database: microorganisms, genomics and prediction. *Nucleic Acids Res* **28**, 377-9 (2000).
58. The FlyBase database of the Drosophila Genome Projects and community literature. The FlyBase Consortium. *Nucleic Acids Res* **27**, 85-8 (1999).
59. Blake, J.A., Eppig, J.T., Richardson, J.E. & Davisson, M.T. The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Res* **28**, 108-11 (2000).
60. Lindberg, D.A., Humphreys, B.L. & McCray, A.T. The Unified Medical Language System. *Methods Inf Med* **32**, 281-91 (1993).
61. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 536-40 (1995).

62. Manning, C. & Schutze, H. *Foundations of Statistical Natural Language Processing*, (MIT Press, 1999).
63. Craven, M. & Kumlien, J. Constructing biological knowledge bases by extracting information from text sources. *Proc Int Conf Intell Syst Mol Biol* , 77-86. (1999).
64. Kosko, B. *Fuzzy Thinking: The New Science of Fuzzy Logic*, (Hyperion, 1994).
65. Ding, J., Berleant, D., Nettleton, D. & Wurtele, E. Mining Medline: Abstracts, Sentences or Phrases? in *Pacific Symposium in Biocomputing* (, Kauau, Hawaii, 2002).
66. Wren, J.D. & Garner, H.R. Heuristics for Identification of Acronym-Definition Patterns Within Text: Towards an Automated Construction of Comprehensive Acronym-Definition Dictionaries. *Methods of Information in Medicine* **41**, 426-34 (2002).
67. Jablonski, S. *Dictionary of Medical Acronyms & Abbreviations.*, (Hanley & Belfus, 2001).
68. Delong, M. *Medical Acronyms, Eponyms & Abbreviations*, (, 1997).
69. Dupayrat, J. *Dictionary of Biomedical Acronyms and Abbreviations.*, (Wiley, New York, 1990).
70. Rimer, M. & O'Connell, M. BioABACUS: a database of abbreviations and acronyms in biotechnology and computer science. *Bioinformatics* **14**, 888-9 (1998).
71. Larkey, L., Ogilvie, P., Price, A. & Tamilio, B. Acrophile: An Automated Acronym Extractor and Server. in *ACM Digital Libraries conference* 205-214 (, 2000).
72. Larkey, L., Ogilvie, P., Price, A. & Tamilio, B. Acrophile: An Automated Acronym Extractor and Server. in *Proceedings of the ACM Digital Libraries conference* 205-214 (, 2000).
73. Tomer, Y. Unraveling the genetic susceptibility to autoimmune thyroid diseases: CTLA-4 takes the stage. *Thyroid* **11**, 167-9. (2001).
74. Green, J.M. The B7/CD28/CTLA4 T-cell activation pathway. Implications for inflammatory lung disease. *Am J Respir Cell Mol Biol* **22**, 261-4. (2000).
75. McCoy, K.D. & Le Gros, G. The role of CTLA-4 in the regulation of T cell immune responses. *Immunol Cell Biol* **77**, 1-10. (1999).
76. Jin, P. & Warren, S.T. Understanding the molecular basis of fragile X syndrome. *Hum Mol Genet* **9**, 901-8. (2000).
77. Bardoni, B., Mandel, J.L. & Fisch, G.S. FMR1 gene and fragile X syndrome. *Am J Med Genet* **97**, 153-63. (2000).
78. Kooy, R.F., Willemsen, R. & Oostra, B.A. Fragile X syndrome at the turn of the century. *Mol Med Today* **6**, 193-8. (2000).
79. Tisdale, M.J. Cancer anorexia and cachexia. *Nutrition* **17**, 438-42. (2001).
80. Hasselgren, P.O. & Fischer, J.E. Muscle cachexia: current concepts of intracellular mechanisms and molecular regulation. *Ann Surg* **233**, 9-17. (2001).
81. Barber, M.D. Cancer cachexia and its treatment with fish-oil-enriched nutritional supplementation. *Nutrition* **17**, 751-5. (2001).
82. Caudle, R.M. & Mannes, A.J. Dynorphin: friend or foe? *Pain* **87**, 235-9. (2000).
83. Steiner, H. & Gerfen, C.R. Role of dynorphin and enkephalin in the regulation of striatal output pathways and behavior. *Exp Brain Res* **123**, 60-76. (1998).

84. http://lethargy.swmed.edu/ARGH/argh.asp
85. Aronow, B.J. *et al.* Divergent transcriptional responses to independent genetic causes of cardiac hypertrophy. *Physiol Genomics* **6**, 19-28. (2001).
86. Shen, W.W. A history of antipsychotic drug development. *Compr Psychiatry* **40**, 407-14. (1999).
87. Morgan, J.P. & Van Maanen, E.F. the role of differential blockade of alpha-adrenergic agonists in chlorpromazine-induced hypotension. *Arch Int Pharmacodyn Ther* **247**, 135-44. (1980).
88. Colucci, W.S. Alpha-adrenergic receptor blockade with prazosin. Consideration of hypertension, heart failure, and potential new applications. *Ann Intern Med* **97**, 67-77. (1982).
89. Molkentin, J.D. *et al.* A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell* **93**, 215-28. (1998).
90. Feng, B. & Stemmer, P.M. Ca2+ binding site 2 in calcineurin-B modulates calmodulin-dependent calcineurin phosphatase activity. *Biochemistry* **40**, 8808-14. (2001).
91. Marshak, D.R., Watterson, D.M. & Van Eldik, L.J. Calcium-dependent interaction of S100b, troponin C, and calmodulin with an immobilized phenothiazine. *Proc Natl Acad Sci U S A* **78**, 6793-7. (1981).
92. Zhang, W. Old and new tools to dissect calcineurin's role in pressure-overload cardiac hypertrophy. *Cardiovasc Res* **53**, 294-303. (2002).
93. Roth, D.M., Swaney, J.S., Dalton, N.D., Gilpin, E.A. & Ross, J., Jr. Impact of anesthesia on cardiac function during echocardiography in mice. *Am J Physiol Heart Circ Physiol* **282**, H2134-40. (2002).
94. Sahn, D.J., DeMaria, A., Kisslo, J. & Weyman, A. Recommendations regarding quantitation in M-mode echocardiography: results of a survey of echocardiographic measurements. *Circulation* **58**, 1072-83. (1978).
95. Gardin, J.M., Siri, F.M., Kitsis, R.N., Edwards, J.G. & Leinwand, L.A. Echocardiographic assessment of left ventricular mass and systolic function in mice. *Circ Res* **76**, 907-14. (1995).
96. http://www.cdc.gov/diabetes/pubs/estimates.htm
97. Marx, J. Unraveling the causes of diabetes. *Science* **296**, 686-9. (2002).
98. Doerfler, W. *et al.* On the insertion of foreign DNA into mammalian genomes: mechanism and consequences. *Gene* **157**, 241-5. (1995).
99. Woodcock, D.M. *et al.* Delayed DNA methylation is an integral feature of DNA replication in mammalian cells. *Exp Cell Res* **166**, 103-12. (1986).
100. Attwood, J.T., Yung, R.L. & Richardson, B.C. DNA methylation and the regulation of gene transcription. *Cell Mol Life Sci* **59**, 241-57. (2002).
101. Esteller, M. & Herman, J.G. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *J Pathol* **196**, 1-7. (2002).
102. Anderson, C.L. & Brown, C.J. Variability of X chromosome inactivation: effect on levels of TIMP1 RNA and role of DNA methylation. *Hum Genet* **110**, 271-8. (2002).
103. Reik, W., Maher, E.R., Morrison, P.J., Harding, A.E. & Simpson, S.A. Age at onset in Huntington's disease and methylation at D4S95. *J Med Genet* **30**, 185-8. (1993).

104. Qu, G., Dubeau, L., Narayan, A., Yu, M.C. & Ehrlich, M. Satellite DNA hypomethylation vs. overall genomic hypomethylation in ovarian epithelial tumors of different malignant potential. *Mutat Res* **423**, 91-101. (1999).
105. Kim, Y.I. *et al.* Global DNA hypomethylation increases progressively in cervical dysplasia and carcinoma. *Cancer* **74**, 893-9. (1994).
106. Hardy, J. The Alzheimer family of diseases: many etiologies, one pathogenesis? *Proc Natl Acad Sci U S A* **94**, 2095-7. (1997).
107. Nussbaum, R.L. & Polymeropoulos, M.H. Genetics of Parkinson's disease. *Hum Mol Genet* **6**, 1687-91 (1997).
108. Sharrard, R.M., Royds, J.A., Rogers, S. & Shorthouse, A.J. Patterns of methylation of the c-myc gene in human colorectal cancer progression. *Br J Cancer* **65**, 667-72. (1992).
109. Alcolado, J.C. & Alcolado, R. Importance of maternal history of non-insulin dependent diabetic patients. *Bmj* **302**, 1178-80. (1991).
110. Thorand, B. *et al.* Can inaccuracy of reported parental history of diabetes explain the maternal transmission hypothesis for diabetes? *Int J Epidemiol* **30**, 1084-9. (2001).
111. Pickard, B. *et al.* Epigenetic targeting in the mouse zygote marks DNA for later methylation: a mechanism for maternal effects in development. *Mech Dev* **103**, 35-47. (2001).
112. Stehouwer, C.D., Gall, M.A., Hougaard, P., Jakobs, C. & Parving, H.H. Plasma homocysteine concentration predicts mortality in non-insulin- dependent diabetic patients with and without albuminuria. *Kidney Int* **55**, 308-14. (1999).
113. Yi, P. *et al.* Increase in plasma homocysteine associated with parallel increases in plasma S-adenosylhomocysteine and lymphocyte DNA hypomethylation. *J Biol Chem* **275**, 29318-23. (2000).
114. Rees, W.D. Manipulating the sulfur amino acid content of the early diet and its implications for long-term health. *Proc Nutr Soc* **61**, 71-7. (2002).
115. Barker, D.J. & Osmond, C. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *Lancet* **1**, 1077-81. (1986).
116. Stern, L.L., Mason, J.B., Selhub, J. & Choi, S.W. Genomic DNA hypomethylation, a characteristic of most cancers, is present in peripheral leukocytes of individuals who are homozygous for the C677T polymorphism in the methylenetetrahydrofolate reductase gene. *Cancer Epidemiol Biomarkers Prev* **9**, 849-53. (2000).
117. Benes, P. *et al.* Methylenetetrahydrofolate reductase polymorphism, type II diabetes mellitus, coronary artery disease, and essential hypertension in the Czech population. *Mol Genet Metab* **73**, 188-95. (2001).
118. Temple, I.K. *et al.* Further evidence for an imprinted gene for neonatal diabetes localised to chromosome 6q22-q23. *Hum Mol Genet* **5**, 1117-21. (1996).
119. Shield, J. *et al.* Maturity onset diabetes of the young (MODY) and early onset Type II diabetes are not caused by loss of imprinting at the transient neonatal diabetes (TNDM) locus. *Diabetologia* **44**, 924. (2001).
120. Sugita, H. *et al.* Inducible nitric oxide synthase plays a role in LPS-induced hyperglycemia and insulin resistance. *Am J Physiol Endocrinol Metab* **282**, E386-94. (2002).

121. Nilsson, C. *et al.* Maternal endotoxemia results in obesity and insulin resistance in adult male offspring. *Endocrinology* **142**, 2622-30. (2001).
122. Pickup, J.C., Chusney, G.D. & Mattock, M.B. The innate immune response and type 2 diabetes: evidence that leptin is associated with a stress-related (acute-phase) reaction. *Clin Endocrinol (Oxf)* **52**, 107-12. (2000).
123. Pradhan, A.D., Manson, J.E., Rifai, N., Buring, J.E. & Ridker, P.M. C-reactive protein, interleukin 6, and risk of developing type 2 diabetes mellitus. *Jama* **286**, 327-34. (2001).
124. Kern, P.A., Ranganathan, S., Li, C., Wood, L. & Ranganathan, G. Adipose tissue tumor necrosis factor and interleukin-6 expression in human obesity and insulin resistance. *Am J Physiol Endocrinol Metab* **280**, E745-51. (2001).
125. Das, U.N. Is obesity an inflammatory condition? *Nutrition* **17**, 953-66. (2001).
126. Liu, L.S., Spelleken, M., Rohrig, K., Hauner, H. & Eckel, J. Tumor necrosis factor-alpha acutely inhibits insulin signaling in human adipocytes: implication of the p80 tumor necrosis factor receptor. *Diabetes* **47**, 515-22. (1998).
127. Moller, D.E. Potential role of TNF-alpha in the pathogenesis of insulin resistance and type 2 diabetes. *Trends Endocrinol Metab* **11**, 212-7. _00000272 _00000272 (2000).
128. Ruan, H., Hacohen, N., Golub, T.R., Van Parijs, L. & Lodish, H.F. Tumor necrosis factor-alpha suppresses adipocyte-specific genes and activates expression of preadipocyte genes in 3T3-L1 adipocytes: nuclear factor-kappaB activation by TNF-alpha is obligatory. *Diabetes* **51**, 1319-36. (2002).
129. Yuan, M. *et al.* Reversal of obesity- and diet-induced insulin resistance with salicylates or targeted disruption of Ikkbeta. *Science* **293**, 1673-7. (2001).
130. Aljada, A. *et al.* Nuclear factor-kappaB suppressive and inhibitor-kappaB stimulatory effects of troglitazone in obese patients with type 2 diabetes: evidence of an antiinflammatory action? *J Clin Endocrinol Metab* **86**, 3250-6. (2001).
131. Ziccardi, P. *et al.* Reduction of inflammatory cytokine concentrations and improvement of endothelial functions in obese women after weight loss over one year. *Circulation* **105**, 804-9. (2002).
132. Pedersen, B.K. *et al.* Exercise and cytokines with particular focus on muscle-derived IL-6. *Exerc Immunol Rev* **7**, 18-31 (2001).
133. Wilson, C.B., Makar, K.W. & Perez-Melgosa, M. Epigenetic regulation of T cell fate and function. *J Infect Dis* **185 Suppl 1**, S37-45. (2002).
134. Rothenburg, S., Koch-Nolte, F., Thiele, H.G. & Haag, F. DNA methylation contributes to tissue- and allele-specific expression of the T-cell differentiation marker RT6. *Immunogenetics* **52**, 231-41 (2001).
135. Coppack, S.W. Pro-inflammatory cytokines and adipose tissue. *Proc Nutr Soc* **60**, 349-56. (2001).
136. Benjamin, D. & Jost, J.P. Reversal of methylation-mediated repression with short-chain fatty acids: evidence for an additional mechanism to histone deacetylation. *Nucleic Acids Res* **29**, 3603-10. (2001).
137. Sealy, L. & Chalkley, R. The effect of sodium butyrate on histone modification. *Cell* **14**, 115-21. (1978).

138. Foley, J.E., Kashiwagi, A., Verso, M.A., Reaven, G. & Andrews, J. Improvement in in vitro insulin action after one month of insulin therapy in obese noninsulin-dependent diabetics. Measurements of glucose transport and metabolism, insulin binding, and lipolysis in isolated adipocytes. *J Clin Invest* **72**, 1901-9. (1983).
139. Green, A., Dobias, S.B., Walters, D.J. & Brasier, A.R. Tumor necrosis factor increases the rate of lipolysis in primary cultures of adipocytes without altering levels of hormone-sensitive lipase. *Endocrinology* **134**, 2581-8. (1994).
140. Laimer, M. *et al.* Markers of chronic inflammation and obesity: a prospective study on the reversibility of this association in middle-aged women undergoing weight loss by surgical intervention. *Int J Obes Relat Metab Disord* **26**, 659-62. (2002).
141. Haffner, S.M., Stern, M.P., Hazuda, H.P., Pugh, J.A. & Patterson, J.K. Hyperinsulinemia in a population at high risk for non-insulin-dependent diabetes mellitus. *N Engl J Med* **315**, 220-4. (1986).
142. Nathan, D.M. Prevention of long-term complications of non-insulin-dependent diabetes mellitus. *Clin Invest Med* **18**, 332-9. (1995).
143. Michalowsky, L.A. & Jones, P.A. DNA methylation and differentiation. *Environ Health Perspect* **80**, 189-97. (1989).
144. Catania, J. & Fairweather, D.S. DNA methylation and cellular ageing. *Mutat Res* **256**, 283-93. (1991).
145. Cooper, M.A. *et al.* Interleukin-1beta costimulates interferon-gamma production by human natural killer cells. *Eur J Immunol* **31**, 792-801. (2001).
146. Toossi, Z. The inflammatory response in Mycobacterium tuberculosis infection. *Arch Immunol Ther Exp* **48**, 513-9 (2000).
147. Noordewier, M.O. & Warren, P.V. Gene expression microarrays and the integration of biological knowledge. *Trends Biotechnol* **19**, 412-5. (2001).
148. Masys, D.R. Linking microarray data to the literature. *Nat Genet* **28**, 9-10. (2001).
149. Tanabe, L. *et al.* MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* **27**, 1210-4, 1216-7. (1999).
150. Kulkarni, A. *et al.* ARROGANT: An application to manipulate large gene collections. *Bioinformatics* **11**, 1410-7 (2002).
151. Shatkay, H., Edwards, S., Wilbur, W.J. & Boguski, M. Genes, themes and microarrays: using information retrieval for large- scale gene analysis. *Proc Int Conf Intell Syst Mol Biol* **8**, 317-28 (2000).
152. Masys, D.R. *et al.* Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics* **17**, 319-26. (2001).
153. Raychaudhuri, S., Chang, J.T., Sutphin, P.D. & Altman, R.B. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res* **12**, 203-14. (2002).
154. Sarnat, H.B., Benjamin, D.R., Siebert, J.R., Kletter, G.B. & Cheyette, S.R. Agenesis of the mesencephalon and metencephalon with cerebellar hypoplasia: putative mutation in the EN2 gene--report of 2 cases in early infancy. *Pediatr Dev Pathol* **5**, 54-68. (2002).

155. Marti, E. & Bovolenta, P. Sonic hedgehog in CNS development: one signal, multiple outputs. *Trends Neurosci* **25**, 89-96. (2002).
156. Crossley, P.H., Martinez, S., Ohkubo, Y. & Rubenstein, J.L. Coordinate expression of Fgf8, Otx2, Bmp4, and Shh in the rostral prosencephalon during development of the telencephalic and optic vesicles. *Neuroscience* **108**, 183-206 (2001).
157. Mizumoto, N. *et al.* Classification of Immunoregulatory Stimuli by Dendritic Cell-based Biosensor. *(submitted)* (2003).
158. Hulsken, J., Birchmeier, W. & Behrens, J. E-cadherin and APC compete for the interaction with beta-catenin and the cytoskeleton. *J Cell Biol* **127**, 2061-9. (1994).
159. Kraus, C. *et al.* Localization of the human beta-catenin gene (CTNNB1) to 3p21: a region implicated in tumor development. *Genomics* **23**, 272-4. (1994).
160. http://www.nlm.nih.gov/mesh/meshhome.html
161. Tanabe, L. *et al.* MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques* **27**, 1210-4, 1216-7 (1999).
162. http://discover.nci.nih.gov/textmining/relevant.html
163. Baclawski, K., Cigna, J., Kokar, M.M., Mager, P. & Indurkhya, B. Knowledge representation and indexing using the unified medical language system. *Pac Symp Biocomput* , 493-504 (2000).
164. Waltzer, L. & Bienz, M. The control of beta-catenin and TCF during embryonic development and cancer. *Cancer Metastasis Rev* **18**, 231-46 (1999).

VITA

Jonathan Daniel Wren was born in Tallahassee, Florida, on October 12th, 1968, the firstborn son of Dr. Karen Tower Wren and Dr. Daniel Alan Wren. His family moved to Norman, Oklahoma in 1973, where he grew up and graduated from Norman High School in May 1986. He then entered The University of Oklahoma in August 1986 and in May 1991 was awarded the degree of Bachelor of Business Administration (B.B.A.) in Management of Information Systems. From there, he went on to work as a database programmer for Rapp Collins Worldwide until the summer of 1993. Having gained an interest in science through independent study during this time, he then returned to The University of Oklahoma to obtain a degree in science with the intention of pursuing an advanced degree thereafter. He was awarded the degree of Bachelor of Science (B.S.) in Biochemistry in May, 1996 and was then accepted into graduate school at the University of Texas Southwestern Medical Center at Dallas, Texas. In March 1998, he married Thanya Del Valle Santos and in January 1999 their daughter, Karen Nicole Wren was born.

Permanent Address: 4017 Oxford Way
Norman, OK 73072