AUTOMATED NATURAL-LANGUAGE PROCESSING FOR INTEGRATION AND FUNCTIONAL ANNOTATION OF COMPLEX BIOLOGICAL SYSTEMS

by

Carlos F. Santos

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2008

Doctoral Committee:

Professor David J. States, Co-Chair
Associate Professor Brian D. Athey, Co-Chair
Professor Daniel M. Burns Jr.
Professor Gilbert S. Omenn
Associate Professor Barbara E. Mirel

To my parents, Byron and Susy

ACKNOWLEDGEMENTS

I would like to thank my advisor and friend, Dr. David States, for his guidance, support and friendship all these years. An incredibly thoughtful and brilliant scientist, Dr. States welcomed me to his lab in Washington University and Michigan Bionformatics from its early stages. He has truly established my career in this field.

I owe equal thanks to Dr. Brian Athey for his guidance and friendship throughout my graduate career and for the extensive feedback he gave me as I developed this thesis.

Many thanks Dr. Gilbert Omenn, Dr. Daniel Burns, and Dr. Barb Mirel for participating on my thesis committee and providing thoughtful feedback and encouragement on this project. Dr. Mirel provided much assistance to us in the design of the BioSearch-2D search engine, and her suggestions on the application's user interface were critical to readying the site for launch. The search engine would also not have been possible without Vasu Mahavishnu's expert application coding as well as Alex Ade's database curation work. Drs. Dragomir Radev, Angel Lee, and H.V. Jagadish have also provided very helpful and critically important feedback for this project.

Many thanks to Yuri Ikeda for managing my graduation process and my years in Bioinformatics, she is a real friend indeed.

I would last, but certainly not least, like to especially thank Dr. Frank O'Donnell of the Hopkins Capital Group, for his mentorship, guidance, and support while I attended graduate school. The incredible opportunity he offered me to contribute and participate in the development of the HCG's biotechnology portfolio complemented my graduate school experience in more ways than I can count.

This project was supported in part by a grant for the NIH/National Library of Medicine R01 LM008106 and the National Center for Integrative Biomedical Informatics (NCIBI), NIH Grant # U54-DA021519.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

AUTOMATED NATURAL-LANGUAGE PROCESSING FOR INTEGRATION AND FUNCTIONAL ANNOTATION OF COMPLEX BIOLOGICAL SYSTEMS

by

Carlos F. Santos

Co-Chairs: David J. States and Brian D. Athey

This dissertation discusses the use of automated natural language processing (NLP) for characterization of biomolecular events in signal transduction pathway databases. I also discuss the use of a dynamic map engine for efficiently navigating large biomedical document collections and functionally annotating high-throughput genomic data. An application is presented where NLP software, beginning with genomic expression data, automatically identifies and joins disparate experimental observations supporting biochemical interaction relationships between candidate genes in the Wnt signaling pathway. I discuss the need for accurate named entity resolution to the biological sequence databases and how sequence-based approaches can unambiguously link automatically-extracted assertions to their respective biomolecules in a high-speed manner. I then demonstrate a search engine, BioSearch-2D, which renders the contents of large biomedical document collections into a single, dynamic map. With this engine, the prostate cancer epigenetics literature is analyzed and I demonstrate that the summarization map closely matches that provided by expert human review articles. Examples include displays which prominently feature genes such as the androgen receptor and glutathione S-transferase P1 together with the National Library of Medicine's Medical Subject Heading (MeSH) descriptions which match the roles described for those genes in the human review articles. In a second application of BioSearch-2D, I demonstrate the engine's application as a context-specific functional annotation system for cancer-related gene signatures. Our engine matches the annotation produced by a Gene Ontology-based annotation engine for 6 cancer-related gene

signatures. Additionally, it assigns highly-significant MeSH terms as annotation for the gene list which are not produced by the GO-based engine. I find that the BioSearch-2D display facilitates both the exploration of large document collections in the biomedical literature as well as provides users with an accurate annotation engine for ad-hoc gene sets. In the future, the use of both large-scale biomedical literature summarization engines and automated protein-protein interaction discovery software could greatly assist manual and expensive data curation efforts involving describing complex biological processes or disease states.

CHAPTER I

Introduction

1.1 Overview

This dissertation discusses the use of computational natural language processing to cull data from the research literature and place extracted observations in biomedical context. Natural language understanding is difficult to automate, but an increasing number of successful implementations of automated biomedical knowledge extraction from free text are being reported[1-3]. We discuss the need for accurate named entity resolution to biological sequence databases and how sequence-based approaches can unambiguously link automatically extracted assertions to their respective biomolecules in a high-speed manner.

We also describe a large-corpus summarization engine which clusters and maps articles, named entities, and biological topics from standardized ontologies into a single user-browseable window in real time. The system enables efficient partitioning of large document clusters into easily-browsed clusters of biologically-related topics.

Recent, successful applications of natural language parsing in molecular biology include recognizing molecular interactions[3, 4], inhibition relationships[5], and pathways[1, 6]. To date, much work has focused on extracting specific classes of relationships from article text ("binds", "inhibits", etc.), but relatively little attention has been given to the problem of defining when, where and under what circumstances these relationship apply. In a related problem, biomedical text search results are still primarily returned in a text manner not easily amenable to large-scale review. Mapping the distribution of annotations relevant to biomolecules in a literature corpus (in effect the contextual role of individual genes within the corpus) remains a daunting challenge even in cases where review articles exist covering those documents.

Representing and capturing biological knowledge and context from free-form biomedical text are major goals of this project. To that end, both the automated parsing system we apply to the Wnt pathway as well as the context-specific search engine, BioSearch-2D we have developed provide novel ways of extracting or mapping biological facts (protein interactions and functional annotation) in a high speed, contextual manner from the literature.

The major focus of this work is recent peer reviewed literature indexed in the National Library of Medicine's Pubmed database. This includes the vast majority of the academic biomedical literature.

1.2 Signal Pathway Annotation

Detailed signal pathway annotation and model construction is by nature an arduous task for human readers to accomplish. The task is complicated for heavily-investigated pathways like the Wnt signal transduction cascade or other major cellular pathways due to the large volume of papers published for biological interactions involving members of the pathway. For the Wnt signal transduction literature, for instance, there were 239 MeSH-annotated "Signal Transduction" Wnt pathway MEDLINE articles in 2003 and 889 articles for the period from 2000 to 2004. Expanding the search to include other co-factors or major proteins in the pathway expands the results to many thousands of articles.

For a pathway like Wnt/Frizzled, up-to-date models are essential for investigators in the field; without an accurate model, experimental evidence may be annotated out of biological context or inconsistent with experimental evidence. Comprehensively annotated models of complex pathways like Wnt are also essential for hypothesis-generation and experiment validation, yet with the exception of periodic reviews on the subject, there are few sources of Wnt-signaling information that are kept up to date with the latest published literature.

Previous authors[7-14] have used NLP-based systems to extract biological molecule annotation information[7], to detect protein-protein interaction information[8, 15, 16], or to improve indexing and recall into searches from MEDLINE abstracts[12, 17]. Methods employed include a mixture of text mining and indexing for terms which can be classified by Bayesian statistics[10], structured grammar matches[18], or word filtering of known entities, as well as the use of partial and full parsers. Full parsers have been employed to discover

protein-protein interactions with promising results, highlighting the utility of this approach. In contrast to full-sentence parse systems reported previously, our application is fully open-source and structured in an XML format that can be easily translated into other representations, including diagramming applications or ontologies[14]. The named entity module we present employs a word-statistic chi-squared test, but begins with a partial parser to derive the necessary named entities; the full parser module provides deeper phrase attachment, syntax information, and grammatical relations, but requires as a pre-filter, a hand-selected list of verbs for protein-protein interaction and most importantly, the named entity list derived from the partial parse.

A continuing challenge in protein-protein interaction detecting remains the detection of specific biologically-relevant molecule names from source literature, and domain-specific usage of names that requires extensive ontological development behind an NLP pipeline before the results can be usefully represented.

In our annotation system, we avoid the need to generate and maintain a large-scale ontology by taking advantage of both the Link full parser[19]'s phrase attachment facilities, as well as fast partial-parser [20] (Cass) noun-phrase annotation to generate a list of words specific to Wnt signal transduction and the general MeSH-annotated signal transduction literature. The fast partial parser's ability to detect and annotate multiple-word noun phrases within the text, coupled with a simple statistical test allows the system to automatically build a corpus-specific named entity list without requiring maintenance of an extensive set of background annotation or dictionaries. While this approach is only a first-pass disambiguation of the named entities found within the corpus (e.g. it does not link to actual sequence or cross-reference data), for the queries likely to be of interest to a human domain expert, we find this automated named entity annotation to be at least as specific as the human-constructed signaling pathway entities available in the public domain, and in some cases, the entities we detect are actually more specific instances of proteins in the human model.

Following the named entity generation, we detect the actual interaction and protein-associations, with a full parser, the CMU Link parser[19] to reduce grammatically complicated sentences into simplified "tuples", which roughly correspond to specific

biological assertions made in any particular sentence. This representation allows us to query the corpus for named entity interactions, where the assertion "tuple" syntax provides a direct linking verb between two named entities (rather than a simple search for co-occurrence, the parse logic behind each "tuple" reproduces with a high degree of accuracy and flexibility the core assertions made by each sentence in a paper). Coupling our specific over-represented Wnt signaling terms, with the parser output yields various relevant possible additions to the canonical Wnt pathway, as well as provides provenance and annotation for a majority of the interactions present in the pathway where source material was not annotated.

1.3 A Visual Map to the Literature

The biomedical literature continues to grow at an accelerated rate, yet the search engines most commonly used to access it remain the keyword-based retrieval engines like NCBI Entrez-Pubmed (http://www.ncbi.nlm.nih.gov/PubMed/) and Google Scholar (http://scholar.google.com). In active fields like cell signaling or oncology, the size of these engines' query results quickly overwhelms human reading ability. Making matters worse, due to the context-dependent nature of scientific research, the first or most recent article(s) returned are typically only a small fraction of those required to comprehensively describe the full body of knowledge contained in the literature on the queried disease condition or biomolecular process.

In order to interpret the results of any given returned result, then, users must not only select a few articles of interest from their search, but then also undertake the additional task of browsing at least in passing the co-referencing papers and related publications returned by the search engine. Quite often, review articles exist which assist by offering expert opinion and summarization of bodies of literature, but these typically focus on specific sub-disciplines within the literature and once published do not update themselves to reflect new findings.

Even considering the publication of review articles, however, the overall growth in the literature is now such that even relatively limited searches often return overwhelming volumes of results. As of early 2008, a query of MEDLINE for the phrase "cancer AND epigenetics" retrieves 5,348 articles; limiting the same query to "epigenetics AND prostate cancer" reduces this number to a still-substantial 285 articles. Similarly, a query for

"prostate cancer AND apoptosis" results in well over 3,000 articles and 472 reviews, an intractable number of papers for all but the most determined reader. Overall, the process of discovering the context and function of gene or disease processes within a result remains a formidable and time-consuming task for a human reader. The problem becomes even worse when discussing complex systems in biology within variable contexts, such as multi-factorial disease or signal transduction pathways with variable roles. A literature search of "Wnt AND signal transduction" for example (returning papers relating to the Wnt family of secreted signaling proteins) yields 3,525 articles, of which roughly 1,500 discuss Wnt-related genes in a developmental biology context and 845 discuss Wnt's in the context of cancer biology. Currently, approximately 50 genes are believed to comprise the core of this pathway[21], yet extracting the oft-varying role of these genes from the hundreds of experimental publications describing them remains a task which challenges even expert human readers.

A number of biomedical search alternatives to the Pubmed search engine have been developed which attempt to better organize the result sets returned by queries. These include text displays of ontology-based clustered results [22], graphical [23] and textual [23] displays of clusters of documents. Also, some search engines include documents not indexed by MEDLINE (Google Scholar) but still present results in a series of text-pages like the Pubmed search engine. All of these primarily return abstracts or titles in lists or as node-edge graphs. Search results from these engines often do not directly display the precise distribution of named entities within those results in a single comprehensive view. Furthermore, in some cases the engines are often limited in retrieval size [22] on the underlying corpus, leading to undercoverage when analyzing the relationships between many hundreds of entries actually present in the result (for example, MeSH headings corresponding to documents and their genes within a given corpus).

Gene- and MeSH-based topic clustering applications in the biomedical literature have been reported in prior work, for instance PubGene [24], a system for automated extraction of explicit and implicit biomedical knowledge from publicly available gene and text databases to create a gene-to-gene co-citation network. The system described does not function as a search interface to article subsets; rather, it explores relationships and similarities within genes in MEDLINE abstracts. Other approaches describe clustering strategies using MeSH

topics, such as the gene-to-phenotype clusters reported by Jennsen Korbel [25] but these are largely one-time analyses rather than search engines in their own right. In yet others, such as the heatmap queries in Lydia [26], the analyses or engines are not focused specifically on biomedical content.

The RefViz literature analysis tool [27] may perhaps be the closest available overview heatmap utility available to that which we discuss in this project, as it displays a literature clustering and retrieval heatmap for documents. Unlike Biosearch-2D, however, RefViz does not cluster results based on organism-specific gene lists or controlled external ontologies. RefViz instead renders the distribution of topics into more of a word-based map rather than a gene-concept-centered map.

Previous work on information extraction in biomedicine includes a number of reports which attempt to extract information about genes from scientific texts using the co-occurrence of terms in a sentence or abstract[17, 28-31]. These approaches, like ours, extract genes within an actual biological context [24], [17], but unlike our current implementation, they do not attempt to summarize a corpus specifically using this approach nor allow for reclustering specific subsets of documents according to user-selectable criteria. Both do report, however, that co-occurrence of gene names in an abstract frequently reflects an actual biological relationship between co-occurring genes.

Masys,et al.[32] describe a system of keyword profiles for genes based Medical Subject Headings (MeSH), but the system is not presented as a user-navigable search engine.
A close comparison to our utility could be CoPubMapper by Alako, et al. [33] but like the other approaches, the analysis presented does not form a direct interface for a search into the literature (so the actual keyword clusters are hidden) and it is not implemented as a web-based utility, but rather was performed as a one-time analysis task. Alako, et al. [33] also report differences in the name tagging algorithm and normalization to our name matching algorithm.

Our search engine in contrast is primarily gene-versus-concept centered, and is a true web-based application, motivated by a need to analyze and explore the role of genes and their roles as described in a literature subset chosen at query time. Our first application for the system explores the prostate cancer genomic literature for those papers describing

methylation and epigenetic changes in tumor progression. Rendering a heatmap of the genes versus MeSH topics relating to articles discussing the genes, the application scales to cover the many hundreds of genes observed in the corpus and the correspondingly large collection of MeSH topics corresponding to articles in which those genes are found. The map itself is rendered and presented via a Flash-based website, allowing rich, interactive, corpus-wide exploration and document retrieval guided by the image features themselves.

To demonstrate the coverage of these maps, we analyze the results from a focused major disease query, "prostate cancer AND epigenetics", as well as the literature discussing a major signaling pathway, the Wnt pathway. We select topics within these collections and analyze the map coverage against human-authored reviews in both cases and a curated web resource in the case of Wnt. Our results suggest that an automated mapping of even a complex corpus in a heatmap corresponds closely to the gene-concept discussion provided by the human reviews and reference websites.

1.4 Contextual Functional Annotation

The annotation of gene list results produced by high-throughput genomics and proteomics experiments has resulted in a vast number of gene expression signatures and canonical reference lists corresponding to important disease and clinical states. Typically, the functional annotation of these gene lists into biological context relies on annotation utilities which calculate the relative enrichment of ontology terms for genes found in the input list compared to the term frequency assigned to genes in a genome-wide context. The majority of these annotation utilities employ the Gene Ontology[34] as their primary annotation ontology. Additionally, some provide additional annotations such as protein-protein interaction lists, protein functional domains, disease associations, pathways, sequence features, homologies, and selected curated literature references [35-38] [39-41]. These utilities are varied, and include both executable software as well as websites like GoMiner [42], EASEonline [35], GeneMerge [43], eGOn [44], FuncAssociate [45], GOTree Machine (GOTM) [46], GOSurfer [47, 48], Ontology Traverser, CLENCH [49], GOToolBox [50], FatiGO [39, 40, 51], and DAVID [35-38]. A complete review of these utilities is described by Khatri, et al. [52].

Additionally, annotation tools like the Molecular Concept Maps described by Rhodes, et al. [53-55] are available which link microarray studies to a number of oncology-related ontologies in order to better allow annotation of clinically distinct cancer gene profiles. In one published report, Tomlins, et al. describe common shared genes between cancer signatures annotated between different cancer types and specific gene repression signatures in both breast and prostate cancers, demonstrating the power of incorporating non-GO ontologies in a highly-focused biological context. [53-55].

To date, Gene Ontology-based annotation engines rely on an intermediate curation step to assign genes to ontology terms based on literature or experimental observation. As Khatri, et al. note, these mapping efforts have historically been fairly accurate [56] and extensive yet mostly assigned in an automated fashion (as of February 2008, there exist 182573 GO annotations for 35113 human genes, of which only 52,246 were not derived electronically) (http://www.geneontology.org). By contrast, MeSH annotation is performed manually by human curators on individual MEDLINE articles. Linking article-derived MeSH terms to genes, therefore, could provide a more tightly-coupled gene annotation than annotations obtained through secondary-source ontologies.

Khatri, et al. further highlight a key limitation to the current batch of annotation engines, in that annotations "related to those genes [which] are involved in several biological processes" are limited to single contexts. Due to the nature of the GO hierarchy, most current tools weight biological processes equally. In effect, these tools make "restricting the query to specific clinical areas…a challenge since the basic annotation itself is largely restricted to basic biological processes". They describe a specific example in the case of BRCA, which has a distinct biological roles as both tumor suppressor as well as in carbohydrate metabolism [52]. Depending on the gene signature in which it is found, the annotations may differ for the gene, which in turns impacts the accuracy of any biological inferences made on that annotation.

In terms of user-interface, the vast majority of existing utilities remain largely text-based, with results returned being large term lists with statistical significance values assigned to each term. These text lists are often produced in batch manner and returned as series of dense text annotations which seldom reflect internal categories between the genes analyzed.

A few graphical interfaces have been developed to address the usability limitations of these text results, including two-color plots rendered by DAVID, where they are described as "… the most powerful graphic presentations in DAVID applications" by the authors. [35, 57]

We have developed an integrated MeSH annotation system in conjunction with a literature concept mapping utility, BioSearch-2D. From a user-submitted gene list, the system renders hierarchically-clustered, dynamic two-dimensional maps representing the distribution of a large set of human gene identifications in biomedical text versus selected MeSH terms. Coloring on the map corresponds to statistically-significant annotations assigned to MeSH terms. These maps directly represent the distribution of MeSH terms corresponding to submitted gene lists as well as the statistical significance in a single unified display, instead of in a series of text lists. We find that the maps match key functional annotation assignments produced by GO-based engines, as well as use a two-dimensional map to render context-specific annotations clustering and intuitive distribution plots which identify functional subgroups in submitted gene lists.

CHAPTER II

Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction

2.1 Background

Wnt signaling is a very active area of research with highly relevant publications appearing at a rate of more than one per day. Building and maintaining databases describing signal transduction networks is a time consuming and demanding task that requires careful literature analysis and extensive domain specific knowledge. For instance, more than 50 factors involved in Wnt signal transduction have been identified as of late 2003. In this chapter we describe a natural language processing (NLP) system that is able to identify references to biological interaction networks in free text and automatically assembles a protein association and interaction map.

A "gold standard" set of names and assertions was derived by manual scanning of the Wnt genes website [58] (http://www.stanford.edu/~rnusse/wntwindow.html) including 53 interactions involved in Wnt signaling. This system was used to analyze a corpus of peer reviewed articles related to Wnt signaling including 3,369 Pubmed and 1,230 full text papers. Names for key Wnt-pathway associated proteins and biological entities are identified using a chi-squared analysis of noun-phrases over-represented in the Wnt literature as compared to the general signal transduction literature. Interestingly, we identified several instances where generic terms were used on the website when more specific terms occur in the literature, and one typographic error on the Wnt canonical pathway. Using the named entity list and performing an exhaustive assertion extraction of the corpus, 34 of the 53 interactions in the "gold standard" Wnt signaling set were successfully identified (64% recall). In addition, the automated extraction found several interactions involving key Wnt-related molecules which

were missing or different from those in the canonical diagram, and these were confirmed by manual review of the text. These results suggest that a combination of NLP techniques for information extraction can form a useful first-pass tool for assisting human annotation and maintenance of signal-pathway databases.

2.2 Introduction

Detailed signal pathway annotation and model construction can be an arduous task for human readers to accomplish. The task is complicated for heavily-investigated pathways like the Wnt signal transduction cascade or other major cellular pathways due to the large volume of papers published for biological interactions involving members of those pathways. In the Wnt signal transduction literature, for example, there were 239 MeSH-annotated "Signal Transduction" AND Wnt pathway articles in 2003, and 889 articles for the period from 2000 to 2004. Expanding the search to include other co-factors or major proteins in the pathway expands the results to many thousands of articles.

For a pathway like the Wnt pathway, up-to-date models are essential for investigators in the field; without accurate models, experimental results may be placed outside of the proper biological context or key insights may be missed altogether if the model structure is incorrect. Comprehensively-annotated models of complex pathways like Wnt are also essential for hypothesis-generation and experiment validation, yet with the exception of periodic reviews on the subject, there are few sources of Wnt-signaling information that are kept consistent with the latest published literature.

In the past, various groups [7-14] have used NLP-based systems to extract biological molecule annotation information [7], to detect protein-protein interaction information [8, 15, 16], or to improve indexing and recall into searches from MEDLINE abstracts [12, 17]. Methods included a mixture of text mining and indexing, with some groups using classification by Bayesian statistics [10], structured grammar matches [18], or word filtering of known entities, as well as the use of partial and full parsers. Full parsers have been employed to discover protein-protein interactions with promising results, highlighting the utility of this approach[14], however they are not available as open-source.

We have developed an automated NLP-based system to assist in the generation of up-to-date pathway models from the literature, that can automatically detect and rank key interacting proteins in an article corpus like that of Wnt signaling.

The named entity module we present employs a word-statistic chi-squared test, but begins with a partial parser to derive the necessary named entities. Then, the full parser module provides deep phrase attachment, syntax annotation, and grammatical relations, and extracts interaction statements by filtering results with a list of verbs and the named entity list derived from the partial parse.

We avoid the need to generate and maintain a large-scale named entity list by taking advantage of both the Link parser's [19] phrase attachment facilities, as well as fast partial-parser's [20] noun-phrase annotation to generate a list of words specific to Wnt signal transduction. Our system uses the fast partial parser coupled with a simple statistical test to automatically build a corpus-specific named entity list without requiring an extensive pre-computed synonym list. While this approach is only a first-pass disambiguation of the named entities found within the corpus, for the queries likely to be of interest to a human domain expert, we find this automated named entity annotation to be at least as specific as the human-constructed signaling pathway entities available in the public domain.

Following named entity extraction, we detect the actual interaction and protein-associations with the Link parser[19]. The parser allows us to reduce grammatically complicated sentences into simplified "tuples" which roughly correspond to specific biological assertions made in any particular sentence. The 3-tuple representation permits fast searches for a direct linking verb between two named entities. The search we perform yields various relevant possible additions to the canonical Wnt pathway, as well as provides provenance and annotation for a majority of the interactions present in the pathway where source material was not annotated.

2.3 Methods: Article XML Processing and Full Parse

2.3.1 HTML Retrieval and XML Conversion

Full-text and MEDLINE articles are retrieved using NCBI's LinkOut e-retrieval utility [59]. For an initial query, an XML file of retrieved UI (Pubmed ID) entries serves as a corpus index, from which local Perl script retrieves where possible the full-text article (via LinkOut URL) and MEDLINE entry. The latter entry serves as a backup entry for cases where full-text may not be present, or where the NCBI LinkOut URL yields only a PDF file.

For the Wnt signaling pathway, we queried Pubmed with:

("Signal Transduction"[MeSH] OR Wnt[All fields] OR Akt[All Fields] OR catenin[All Fields] OR frizzled[All Fields])

The query yielded 3523 articles (full analysis in supplementary data), of which 3369 could be retrieved in XML. Of these 3369 documents, the majority (2914) had a parseable abstract field (either from HTML or MEDLINE record), and of the 455 that did not, the papers were often review papers, with the XML tag marked as "TOP". The full corpus composition is available as supplementary data. The query was restricted to the past five years (1999/03/03 to 2004/03/01).

2.3.2 XML Document Structure Parsing

To normalize successfully-retrieved HTML papers, we developed a document-structure parsing script in Perl (v. 5.6.0) that extracts into XML-format the Titles, PMID, Abstract, Methods/Materials, Conclusions, Figures, Tables, and References sections of full-text articles: We parse sentences within all sections by default, only explicitly excluding sections parsed as "References". It is important to note that of the 3369 retrieved papers, over 10% had no explicitly-labeled "abstract" section (even if one was provided in the MEDLINE).]

2.3.3 Pre-Processing and Parse

For parsing, we process and exclude non-parseable sections like references and tables in each paper. Articles are then processed through a Link grammar parser [19] (version 4.1a; http://www.link.cs.cmu.edu/link/ftp.html) on a 16-node Linux cluster.

2.3.4 Link Parser Output

For each sentence, the parser yields word associations as a flat list with left-hand terms "attached" by a grammar relation to terms on the right. The "subject-verb-object" relations provided by the parser form the core assertions we wish to capture from the parse. The parser captures the main verb of each clause or sentence, links it with the proper subject noun, and object if present, yielding a subject-verb-object assertion which we extract as a 3-tuple.

2.4 Methods: Assertion Representation via Link Parsing: Subject-Verb-Object Tuples

2.4.1 Tuple Format

The structures we call tuples are Link-grammar-parser derived structured, hierarchical representations of grammatical relations between phrases and words within sentences. Generally, each tuple takes the form of a three-component structure:

In our tuple format:

```
<int pmid="12952940">
      <protA>Wnt</protA>
      <protB>Frizzled</protB>
      <assert>
            <src_sent>...</src_sent>
            <tuple>
                  <subj>...</subj>
                  <verb> ... </verb>
                  <obj> ... </obj>
            </tuple>
      </assert>
</int>
```

Each interaction (*int),*contains two named entities *protA* and *protB* , with *assert* element which contains a sentence (*src_sent)*, and a tuple element *(tup).* The *tup* contains a subject (*subj)*, verb (*verb),* and an object (*object).* The subject and object terms can be either single or multi-word nouns, attached to modifying prepositional phrases, adjectives, and articles. Verbs are single words, and are marked as *verb.* Objects follow the specific verb marked.

Some authors [9] employ sophisticated template-matching with partial parse-based algorithms when detecting interactions. These systems are faster than our parse, but often require substantial manual template generation for the partial parser.

Our interaction detection searched for phrases with two named entities flanking any of a select group of stemmed verbs. The verb list itself was manually compiled from a listing of verbs found in the corpus and from verbs in general usage likely to be found describing protein-interactions. These "direct" and "indirect" physical interaction verbs are split into:

```
Direct interaction verbs:
bind (bound), interact(-s,-ed), stabilize(-s,-d), phosphorylate(-s,-d),
ubiquinate(-s,-d), sumoylate(-s,-d), degrade(-s,-d), block(s)

Indirect interaction verbs:
induc(-es,-ed), trigger(-s,-ed), block(s), enhance(s), synergize(s),
cooperate(s), localizes, regul(-ates)(-ion), activate(s), inhibit(s),
control(s), translocate(s), antagonize(s), amplif(-y)(-ies), transduce(s),
degrade(s), trigger(s)
```

2.4.2 Tuple Examples

The system outputs tuple assertions from sentences in XML:

```
<assert>
      <src_sent>
            Wnt8 binds to LRP6 and Frizzled8.
      </src_sent>
      <tup>
            <subj>Wnt8</subj>
            <verb mod="v">binds.v</verb>
            <obj><p pp="to">LRP6</p></obj>
      </tup>
</assert>
```

In the sentence above, "Wnt8 binds to LRP6 and Frizzled8." yields two assertion tuples: the binding of "Wnt8" to "LRP6" and a matching tuple (not shown) for the binding of "Wnt8" to "Frizzled8".

In addition to direct interactions, sentences where a verb suggesting an interaction is found within the object, we make the assertion as being the closest preceding matching verb or gerund matching within the phrase for the named entity in the object.

2.5 Methods: Automatic Name Extraction from a Partial Parser

The Cass parser [60] is a fast (10000 sentences/hour) deterministic partial parser that we use to construct a named entity set specific to the current domain. The parser has several key advantages over a parser like Link that make it a worthwhile choice for a named entity recognizer, primarily its good specificity for detecting selected "phrase chunks" of sentences at speeds which are many orders of magnitude greater than those achieved with a full parser like Link. This markup allows us to statistically compile named-entity candidates (noun phrases) from the small topic-specific corpus against a massive background corpus (all "signal transduction"), while reserving the use of a computationally-expensive full parser only for determining tuples in the small corpus.

We used the Cass parser to select named entities (noun phrases) for the Wnt pathway by comparing the occurrence of named entities in the Wnt-specific article corpus against their occurrence in a "background" signal-transduction literature corpus (10000 records, yielding 8873 parsed articles corresponding to the PubMed query "Signal Transduction"[MeSH] from the previous two years).

By comparing the frequency of "Wnt" to "signal transduction" noun phrases, we calculated one-degree of freedom chi-squared values for Wnt Cass noun-phrases relative to the Signal Transduction corpus and ranked them according to that chi-squared value. Significance was set as $p<0.001$. Examples of over-represented Wnt terms included both single phrases, as well as compound phrases.

For every NX term, X^2 was calculated as:
w_i *:* the number of occurrences of NX term *i* in the Wnt-specific corpus
W: the total number of NX terms in the Wnt-specific corpus
s_i *:* the number of occurrences of term *i* in the signal-transduction corpus
S_i *:* the number of occurrences of term *i* in the signal-transduction corpus

$$(1)\quad X^2 = \sum_{i=1}^{k} \frac{\left(\frac{w_i}{W} - \frac{s_i}{S}\right)^2}{\frac{s_i}{S}}$$

Note that not all terms were proteins, since the terms are noun-phrases in general. In the application, proteins of interest were filtered at search time manually where found. Noun

phrases we detected included both single (“Wnt”) and multiple-word forms that would otherwise be missed by a dictionary-based search (e.g. “casein kinase i epsilon”).

2.6 Methods: Automatic Name Extraction Using a Full Parser

2.6.1 Full-Parse Phrase-Derived Named Entity Extraction from the Link Parser

The second named entity-extracting module in the pipeline scans the tuples generated (Wnt-specific tuples) from the Link parse for tuples derived from sentences such as "*X* is … a protein" and "the *Y* protein". For every tuple formatted with "is" as the verb, we find the subject, and if it is a single word or phrase, capture the predicate phrase for that tuple, and append the subject into an index entry one word at a time, recursively. For example:

```
Sentence: E-cadherin is a transmembrane glycoprotein ..

E-cadherin >> is >> glycoprotein
E-cadherin >> is >> transmembrane glycoprotein

E-cadherin >> Append to "glycoprotein" file
E-cadherin >> Append to "transmembrane glycoprotein" file
```

After categories are formed and the first set of names is input, the system re-scans the entire corpus for phrases of the form *"article X Y"*, where *article* is either "a", "an", or "the", *Y* is a term category (e.g. "protein"), and *X* is a non-whitespace term. This second pass allows us to capture a small additional fraction of terms of the form "the Wnt protein", where the last word in the phrase is a solid term category like "protein".

The end result of both passes is a series of categories or category files, comprising a shallow ontology. This auto-categorization system yielded 7066 distinct categories for the 3306-article Wnt-signaling specific corpus, and 24474 terms within those categories, of which 24323 were unique terms. The largest categories are not surprisingly commonly discussed terms, including "protein", "gene", "proteins", etc.

We find the terms extracted are very specific as they are directly extracted from direct declarative statements in the corpus.

2.6.2 Manual Annotation Results

Our precision and recall are measured as to the correct fraction of overall interactions returned and the percentage of the interactions captured in the gold standard[58], respectively. Results are given in Table 1.

2.6.3 Calculation of Precision

We define precision as the fraction of correct tuples returned by the parser. These tuples are tuples where the sentence actually supported evidence for a direct physical binding interaction or mentioned an indirect but biological relationship between the two protein entities in the tuple.

From the corpus, we derived a set of 6787 Tuples/Interactions, of which 1210 were unique pair-wise. We tested 5% (randomly selected) of the data set (340 sentences), representing individual unique sentences with their tuples and the two interacting proteins, and hand-scored assertions for the accuracy of the tuple and named-entity search to determine if the sentences support the interactions noted. This tests the performance of the parse/extraction software without explicitly biasing the sampling towards a subset of the corpus (e.g. interactions which only contain a few papers in the entire corpus). For the parser evaluation, we tally but ignore from the final count all name-detection errors as these are a function of the named-entity module or of the human input.

"Direct" verb tuples are more useful for actual diagramming of physical pathways, but the "indirect" interactions are still indicative of relationships between distant pathway components. Tuples may be useful as a validation of models built with the system. We are not measuring interaction directionality at present in the system.

2.6.4 Calculation of Recall

The exact recall metric for a system like ours is difficult to calculate manually, as it would require determining the total number of "facts" made about binding proteins in the articles scanned. We therefore calculate recall as the fraction of the "gold standard" interaction set we are able to reproduce compared to the Wnt genes homepage, rather than as the fraction of interactions detected against the absolute "assertion or interaction" count in the corpus.

2.6.5 Domain Specificity

By default, all returned interactions that are "correct" are within the domain. The corpus itself is the domain we examine, and we expect a "Wnt" corpus to therefore contain only within-domain interactions.

2.6.6 Discussion: Use of a Partial Parser for Named Entity Extraction

The Cass parser lacks certain phrase attachment and coordination capabilities of Link, but we found that its relatively good accuracy and very high speed allowed us to use Cass as a named entity extractor. Cass' finite-state grammar rules allow us to extract multiple-word noun phrases without requiring the use of an external dictionary or coordination and integration with existing synonym lists.

In actual usage, we found that compiling extensive named-entity lists from other databases provided little benefit, as in the end, interactions adding to the "gold standard" will be manually verified before being submitted as authoritative. Extracting the named entities from the text itself yields word phrases that are guaranteed to match (even if they are spelling variants), and allows extraction of useful assertions that can later be verified for accuracy. As expected, this process is extremely fast, but can occasionally introduce spurious "interactions" between terms and common phrases.

2.7 Results: Comparison of Automatic Wnt Pathway Annotation and the Existing Gold Standard

The system discovered various high chi-squared terms with additional or different annotations than those present in the gold standard:

2.7.1 The Phosphorylation Interaction between CKI-epsilon (CK1e) and APC

In the diagrammed gold-standard Wnt-signaling pathway, no specific mention of CK1-epsilon (CKIe, CKI epsilon) interaction with APC is made, and on closer inspection, Kishida et al.[61] do make a statement of the direct phosphorylation between the two molecules.

2.7.2 The Phosphorylation of beta-catenin by CKII (CK2)

The Wnt genes gold standard mentions CK2 as CKII in the context of binding to Dishevelled, but does not specifically show direct interaction of CK2 with beta-catenin in the protein interaction figures although links to a paper describing phosphorylation of beta-catenin by CK2 are provided. Our search independently found two articles, including the cited articles[62] and a morphological study[63] which describe the direct interaction of CK2 with beta-catenin directly. The chi-squared values for CK2 and beta-catenin are 1179.50 and 40537.69, respectively, suggesting these terms are significantly over-represented in the Wnt literature as a whole, and suggesting this interaction should be a directly-featured pair in the gold standard map.

2.7.3 Six3 And Wnt Regulation

The Wnt genes website lists Six3 (Sine oculis homeobox (Drosophila) homologue 3) as a Wnt target gene[64]. Six3 also feedbacks to repress Wnt expression, an interaction note mentioned on the website and specifically not mentioned in the table of Wnt feedback target genes. A paper cited by the website describes this interaction[65].

2.7.4 Pathway Expansion: Wnt Downstream Targets

Chen, et al. report that Wnt-1 signaling inhibits apoptosis and caspase activation induced by cancer chemotherapy [66]. Such distant pathway cross-talk events of activation and regulation between Wnt and other pathways are difficult to curate manually and by definition are often not fully referenced in “canonical” diagrams. In particular, remote downstream

activation or cross-talk between proteins downstream of the canonical pathway are areas where statements in the literature could be mined by automatic annotation software.

2.7.5 Wnt-7a and LMX-1b

Lmx1b is induced in the mouse dorsal mesenchyme by wnt-7a and it is both necessary and sufficient to specify dorsal limb pattern[1]. The activation pattern was not noted in the Wnt genes website, but was found amongst the interactions by the machine parse (in article PMID 12588849) [1].

2.7.6 Typographical corrections: Pygopus and Pygopos

Human typists are not infallible, and the name recognizer component of the pathway automatically discovered the Pygopus name but missed the interaction with Pygopos. The latter term resulted in the term list after human entry, and manual review showed the spelling error arose from a spelling error on the annotation itself from the Wnt signaling canonical pathway. The example serves not as any particular criticism of the pathway map, but rather highlights the risk of relying on human typed input into pathway annotations; automated systems do not fatigue or commit unintentional typographical mistakes whereas human input can lead to a certain degree of error even in highly-curated databases.

2.8 Conclusion

Our results with automatic component identification and interaction detection in the Wnt signaling pathway suggest that natural language techniques are able to substantially improve the coverage of canonical reference literature and signaling models. The high precision and processing speed of this automated signaling interaction pipeline demonstrates the value of full-parsers and statistical techniques. Using this approach as a "first-pass" filter into the literature can usefully assist scientist maintaining databases and information resources in complex and rapidly evolving fields such as signaling pathways. As with any fully-automated system, however, the recall rates with respect to the known canonical models do not yet match those of an expert human reviewer.

In the future, we expect to capture directionality and type of interaction in a more robust way for our assertions. This will require more template development, and may require the

use of an ontology for an outside reference source for error-detection of incorrect assertions. The role we most expect this system to serve is a real-time scanning facility for new articles, searching for newly-discovered interactions. Automated computational methods are capable of analyzing a much broader coverage of literature than would be feasible for a human reviewer to perform. In this role, there is a premium on specificity to avoid overloading the manual reviewer with erroneous matches, and our results suggest that deep-parsing, automated natural language processing technology is now capable of achieving this requirement.

We found that our auto-categorization module, using statistical and natural-language parsing techniques allowed us to build a named entity list at run-time, rather than requiring a cumbersome fixed named entity assembler before the processing. This approach was perhaps our main advantage in this pipeline, because unlike general English-language texts, the biomedical literature enjoys a substantial human-hierarchical index via the MeSH tags provided by MEDLINE.

MeSH indexing provides a powerful tool for building reference and background article sets that can be used to search a specific article corpus for biologically-relevant named entities which are typically over-represented with high statistical significance. The fast partial parser CASS serves a useful role in assigning multiple-word entities. CASS is uniquely powerful in its ability to efficiently process very large collections of text. This speed is a result of algorithmic efficiencies which are unlikely to be matched by more complete full-parsers. The combination of fast partial-parse, exploiting MeSH indexing and statistical analysis of multiple word phrases significantly simplifies our task of assembling a comprehensive term list.

At a deeper level of text interpretation, the Link parser provides us with grammatical relations, which allows us to move beyond simple association statistics to access the information encoded in the grammatical structure of sentences. While some sentences in biomedical text are too complex to be accurately parsed using current technology, we find that parsers such as Link are able to accurately and efficiently parse the majority of sentences in the molecular biology literature. Using the integrated approach described above, we are beginning to be able to analyze the knowledge encoded in biomedical text.

Table 1: Performance of the Automated Pathway Analysis and Examples

	Count (%)	Interaction as Detected		Example Sentences of this Error or Type	Short-format Link Tuple (not in XML)
Total manually sample counted	370				
Total incorrect names (*ignored from both tallies below)*	22 (5.9)	Akt <-> Tir	12896980	Although **Akt** activity was also induced by **Tir**on and DPI, the other two free-radical scavengers examined , only selenite supported cell growth.	LIN: [**Akt** activity.n] v:<was.v> [m:<induced.v> only [pp by **Tiron**]]
Total indirectly/categorically correct interactions (A pathway...B pathway...ignoring name errors)	129 (37.1%)	Akt <-> PI3-kinase	14557259	**Akt** is activated by many growth factors and cytokines in a **PI3-kinase**-dependent manner.	**Akt** v:<is.v> [m:<activated.v> [pp in [a **PI3-kinase**-dependent manner.n]] [pp by [many cytokines. n]]]
Total directly/physical interaction correct (A->binds->B ignore name errors)	215 (61.7%)	Dvl <-> Axin	11113207	Consistent with these results, **Dvl** *interacts* with **Axin** and inhibits GSK-3 beta-dependent phosphorylation of beta-catenin, APC, and Axin in the Axin complex .	**Dvl** v:<*interacts.v*> [pp with **Axin**]
Total correct names, but error in the parse (ignoring name errors):	4 (1.1%)	Dvl <-> Axin	11113207	Consistent with these results, **Dvl** interacts with Axin and *inhibits* GSK-3 beta-dependent phosphorylation of beta-catenin, APC, and Axin in the **Axin** complex	**Dvl** v:<inhibits.v> [pp in [the **Axin** complex.n]]
Total Gold Standard Associations Detected				**31 of 53 (58.4)**	
Parse/Extract Precision Total correct (direct+indirect, ignoring name errors):				**344 of 370 (92.3)**	
Parse/Extract Recall with respect to Gold Standard Review Derived Set				**31/53 (58.4)**	
Separate Unique Interactions (overall)				**1176**	
Separate Unique With Correct Name Recognition				**1043**	

CHAPTER III

Grounding of Free Text to Biomolecular Sequence Databases

3.1 Background

Accurate mapping of free-text named entities to precisely defined biological entities remains a critical and necessary step for rapid integration of high-volume, automated information extraction methods into systems biology models, pathway or biomolecular interaction graphs. Here we describe a full-text pipeline focused on the Wnt signaling pathway which exploits short DNA primer sequences in full text to establish statistically-validated sequence alignments as the basis for mappings from free named entities to standardized Genbank sequence entries. Using the published literature as an intermediary database, we are able to map from the core Wnt signaling pathway to a more extensive set of precisely identified Wnt related molecules. We find that primers are ideally suited for unambiguous genomic localization, but are found with relatively low frequency in full text and abstract papers.

Modern natural language processing and information extraction systems are able to leverage massive computational power against the human-authored biomedical text databases in order to process heterogeneous text into machine-readable assertions that can form the basis for improved systems biology models, pathway or biomolecular interaction graphs, or biomolecular annotations. In many named entity discovery utilities published to date, however, the basis for assigning named entity to curated names arises largely from a variety of matching algorithms which scan free text and match the output entities to annotation lines or standard names in biological databases on the basis of name-to-name string match. [7, 8, 14, 67-70]

. We find that PCR primers, when present in biomedical articles, are well-suited as readily-alignable, unambiguous anchors into genomic sequence databases. These primers

can serve as high precision markers for data integration tasks, allowing precise anchoring of free text named entities to curated definition entries and standardized gene names in biomolecular databases. In many cases, primer sequences allow a greater degree of precision in entity definition than that which was used by the original author (due to colloquialism or 'canonical' entity naming). Of additional interest computationally, the search space reduction achieved when comparing individual article's named entities (e.g. noun chunked phrases) against the relatively small set of aligned definition lines allows even low-stringency, low-performance searches to efficiently match entries while still maintaining high accuracy. The results from our pipeline demonstrate the utility of exploiting these unambiguous PCR primer sequences to anchor free-text named entities to genomic coordinates and existing gene models and show how these experimental entries can perhaps yield higher precision matches to sequence than simple string matching alone.

3.2 Methods

We have developed a full-text and abstract-based automated text processing pipeline described previously[71] in order to mine the biomedical literature databases from HighWire Press (http://highwire.stanford.edu/), Pubmed Central (http://www.pubmedcentral.nih.gov/), and the NCBI's Pubmed/MEDLINE (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed). In the pipeline, full-text and MEDLINE articles are retrieved using NCBI's Linkout e-retrieval utility[59]. For a given MEDLINE/Pubmed query, an XML file of retrieved UI (Pubmed ID) entries is processed by a series of Perl scripts which retrieve when possible the full-text article (via LinkOut URL) and MEDLINE entries corresponding to individual articles. While full-text is the desired output, the pipeline in the vast majority of cases also maintains the latter as a backup entry for cases where full-text may not be present, or where the NCBI LinkOut URL yields only a PDF file.

As previously described, the pipeline focuses primarily on the corpus of Wnt signal transduction literature retrievable via the MEDLINE query:

("Signal Transduction"[MeSH] OR Wnt[All fields] OR Akt[All Fields] OR catenin[All Fields] OR frizzled[All Fields])

As of the time of this manuscript (mid-2005), the pipeline contains 3334 articles (MEDLINE, html, and/or PDF). Of these articles, 1269 are available as full-text articles which we process into XML format, and the remainder are available as abstract-only or in HTML which we fail to parse (but retain as HTML or MEDLINE source) (1967 papers); the remainder are only present as placeholders in the case of errors or missing abstract data .

3.2.1 Article Processing

Retrieved articles are processed into XML, then split into sentences and parsed by the Cass [60] partial parser for noun-phrase extraction as previously described[71]. Briefly, files from the original HTML are converted into one-sentence-per-line format, parsed by the Cass parser, and noun (NX) phrase entries are extracted and stored into a Microsoft SQL Server relational database. This database also maintains the MEDLINE records for individual articles, which allows the system to query the NLM Medial Subject Heading[1] entries provided for each paper in the database. These MeSH entries allow querying of standard species names ('Human', 'Humans', 'Rat', 'Rats', 'Mice') for each paper.

Table 2: Corpus Composition

Wnt Signal Pathway	Documents
HTML or abstract only	1967
Full Text (XML parseable)	1269
Error or Missing	108

Retrieved papers in either full-text or abstract format were indexed by species. DNA primers were extracted by scanning the original HTML and XML source for the regular expression /([ACGTRYSW]{8,})/. Species-specific sequence alignment to genomic locations were performed on primers by an NCBI BLAST [72] search (databases for species-specific DNA primer searches were the human genome NCBI release 35 [66], NCBI Mouse Genome Assembly 33[73], and NCBI Rat Genome v3.1 [74]; BLAST parameters included an e-value of <0.1 with MegaBLAST and output in tabular format with gi-lines displayed). BLAST-aligned DNA primer sequences were then mapped to their respective genomic locations by querying the Ensembl genome database for each species' respective set of primers in the corpus. Primers with matches against Ensembl were stored in the database

together with the Ensembl identifiers, definition names, and gene names as well as with their specific genomic locations.

Table 3: DNA Primers by Paper Format

Format	Distinct Primers Found
HTML-only (full text) or abstract	435
Full Text, HTML-to-XML Parseable	1981

3.2.2 Named Entity Matching (Ensembl to Cass Noun Chunks)

Named entities extracted from the articles in the corpus by NX noun-phrase chunking after parsing with the CASS partial parser were stored in a relational database. The resulting named entities included noun-phrases with protein names but these were not scanned directly by a dictionary matching these names against a definition line. Instead, from the BLAST results and the subsequently matched Ensembl database search, we matched the sequence gene name and description lines in Ensembl for each primer against the Cass named entities (NX chunks) as follows. Note that for each match operation, we also maintain a record of the method used to match ("criteria") in the database.

1) An exact match was performed if possible (case-insensitive stringwise comparison) on the definition line and the gene name itself against the NX phrase. (criteria label: "exact")
2) If the full-match failed, a match was performed on any parenthetical content in the Ensembl description line against the NX phrase (criteria label: "paren")
3) Also, the two longest words (special non-word characters excluded) in the Ensembl description line were matched against the NX phrase. (criteria label: "2-word")
4) As a last resort, a "stemming" operation was performed: the base match of terms like "Wnt7" were stripped of the trailing numerals, and the base name ("Wnt") was matched against the NX phrase. (criteria label: "basematch")
5) All matched names and NX's of length 2 characters or shorter were excluded for the scan.

Table 4: Results of Primer Scanning (Sorted by Species) in the Corpus

Species (MeSH)	Distinct Documents Containing Primers	Distinct Primers Found in Corpus	Distinct Documents with BLAST-alignable hits	Distinct Ensembl ID matches to primers (by BLAST)	Distinct name (Ensembl gene name) of primers (by BLAST)	Distinct Matched Named Entities (NX phrases in free text)*	Distinct Matched Ensembl Gene Names (vs. Named Entites)
Human(s)	209	1277	140	847	436	2999	160
Mice	151	1006	110	834	356	2181	122
Rats	56	335	39	460	162	554	46

***Named entities are labeled distinct tuple-wise, as they are contained within tuple assertions extracted by the pipeline. See [71] for a discussion of tuples.**

3.3 Results

Within the corpus, we resolved a large number of names from the Ensembl database's description and gene name fields in each species to noun phrases matched by the parser (see Table 4). Table 5 shows example matches, with the Ensembl Gene Name entry labeled as it matches to a Cass-derived NX phrase match. The phrase matches are often unable to match exactly, in which case the two-word (two longest definition line words) or the stemmed (base) were used to determine a match. For instance, Frizzled-1 was stemmed in the record for paper "PMID:11287180" (Pubmed id 11287180) to the term "Frizzled", which then matched a noun phrase entry "Frizzled" detected by Cass. The exact primer in this instance was the nucleotide sequence 'GTACTGAGCGGAGTGTGTTTTCT', mapping to the mouse gene Frizzled-1. It is interesting to note the generic "Frizzled" usage in this instance: the stemmed terminology used by the authors is not as informative as a free-text entry in its own right, but becomes readily-resolvable when anchored to a sequence by the DNA alignment of "GTACTGAGCGGAGTGTGTTTTCT" to the Ensembl Entry "ENSMUSG00000044674".

Table 5: Example Matches: DNA Primers Aligned to Ensembl and Matched Against CASS Partial Parse NX Phrases

Ensembl Gene Name	Source Paper	Ensembl Description Match (words/characters)	Cass NX Phrase Match (Noun Phrase Named Entity)	EnsemblID	Criteria
Apc	PMID:11854293	Adenomatous+polyposis	axin/adenomatous polyposis coli -directed targeting	ENSMUSG00000005871	2-word
AXN2_MOUSE	PMID:11809808	Axin+2	Axin2 cDNA	ENSMUSG00000000142	2-word
Catnb	PMID:10884377	catenin+Beta	B beta-catenin mRNA levels	ENSMUSG00000006932	2-word
PGR	PMID:12554765	Progesterone+receptor	progesterone receptor	ENSG00000082175	2-word
NM_199472	PMID:11809808	glyceraldehyde-3-phosphate+dehydrogenase	glyceraldehyde-3-phosphate dehydrogenase	ENSMUSG00000055676	2-word
AXIN2	PMID:11940574	Axil	Axil	ENSG00000168646	paren
Fzd1	PMID:11441081	Frizzled-1(Frizzled)	Frizzled	ENSMUSG00000044674	basematch
Sfrp2	PMID:11287180	frizzled-related+frizzled-related	secreted Frizzled-related proteins	ENSMUSG00000027996	2-word
CDKN1A	PMID:11463845	p21(p21)	Akt-mediated p21 phosphorylation	ENSG00000124762	basematch
Cdk5rap2	*PMID:12177059*	*Fragment(Fragment)*	*fragment*	*ENSMUSG00000039298*	*basematch*
EDG2	PMID:11485975	LPA-1(LPA)	Three LPA receptors	ENSG00000198121	basematch

Exact match is a highly-stringent criteria for matching names. Not surprisingly, the performance of the algorithm exceeds 99% precision when sequences are directly aligned and matched string-wise to names. When compared with the total group of NX phrases returned over the articles with primers in the corpus, the recall remains relatively low, however. These results are not surprising, however, as the named entities with primers occur rarely compared to general noun phrases. Nonetheless, as an anchor point for exact match for curation, the exceedingly high precision obtained with this method is a desirable outcome.

Table 6: Algorithm Performance

Precision	>99% (due to exact match stringency)
Recall (average per article where primers are present)	3.8%

Improving the recall measure for this algorithm remains a challenge, as primers are rarely included except for mention when authors discuss experimental methods. Errors observed with the algorithm include occasional mismatches (e.g. “fragment” matches to “fragment” in a noun phrase when both are present in the definition and NX phrase). The method offers an improvement for phrase expansion of ‘stemmed’ or ‘canonical’ phrases (like LPA or

Frizzled) which remain a challenge in traditional string-based match algorithms, as the necessary information in many cases is lacking from the terms and therefore must be inferred from the surrounding context. In contrast, exploiting secondary sources of information, like primer sequence-based matches, can help guide the string match with additional information and assist in accurate resolution of the ambiguous noun phrase to sequence. The run-time performance of the algorithm is an additional benefit to sequence-based resolution. Unlike string matches to dictionary, the relatively few entries and resulting miniscule search space of the aligned sequence description entries allows application in this case of otherwise intractable or very low-stringency methods (like combinatorial term matching, or word fragment matching). We used a two-longest-word match as a demonstration heuristic, but the individual term match methods can be readily altered to more complex variants if so desired.

An important aspect of this work is the precision of linking. This allows us to assign higher biological significance to rare matches. In the case of Wnt, a number of gene names were identified that are not part of the canonical Wnt signaling pathway. These include AMHR2, BRAF (mutated Raf), and BRCA1. With the thousand of named entities occurring in the corpus we scanned, these would not be significant if the mapping had even a 1% false positive error rate. By using PCR primer matching for confirmation, we can identify these named entity resolutions as significant.

3.4 Discussion

A central problem in named entity resolution is the frequent use of imprecise language in biomedical text. For knowledge extraction and database linking, we need to link named entities in text to precisely defined molecular entities, but this is frequently impossible based on sentence level text analysis. For example, authors typical specify the species used as the basis for a body of work only once or a few times in a manuscript, and rarely qualify individual gene names with the species of origin. Working at a sentence level, it is therefore impossible to know which species a gene name is referring to. We have identified PCR primers as a class of easily recognized named entity in text that encode precise molecular information and allow precise named entity resolution to be performed automatically and reliably.

PCR primers are surprisingly prevalent in the molecular biology literature with 2618 distinct primers associated with 328 distinct genes in 3334 papers. However, the distribution of PCR primer data across papers is not at all uniform, and many of the primers refer to controls (GAPDH) of little value in knowledge extraction.

Mapping the PCR primer to the genome is, of course, only a part of the problem. We also need to associate the identified gene with text. Effectively, we use the primer match to dramatically restrict the search space for named entity resolution to just the text appearing in the gene description field. In this way, even partial and incomplete matching can be made with high reliability.

In this work, we have not attempted to map primer positions within genes, but this represents a potentially fertile approach for future work. There is no standard way to refer to exons, particularly when a gene is subject to alternative splicing. For example, is the first exon associated with an alternative transcription start site "exon 1b" or "exon 2"? Authors are also inconsistent in referring to positions within exons. For example, is "codon 1 of exon 2" the first codon entirely contain in exon 2, or the first codon partially overlapping exon 2? When molecular sequence tags (both nucleic acid and peptide) are provided, it should be possible to resolve many of these ambiguities.

CHAPTER IV

Heatmap Concept Mapping and Search of Biomedical Document Collections

4.1 Background

Text lists such as those returned by the NCBI Entrez and Google search engines are to date the most widely adopted method of performing biomedical literature searches. With the rapid growth of the biomedical literature in recent years, however, even relatively focused queries yield large result sets which are difficult, if not impossible, for humans to read comprehensively. Making matters worse, in biomedical literature search, the first or most recent article(s) returned are typically only a fraction of those needed to fully describe a disease condition or biomolecular process. We describe a system which automatically renders real-time browseable heatmaps of large document collections by integrating an automated gene-tagging algorithm with NCBI MeSH tags found in those collections. These maps then serve as the interface into the query result document collection and provide users a visual concept map for query results. We demonstrate that this automatically-generated, web-based and user-searchable heatmap can accurately represent the contents of the query in a manner comparable to a human review article. The system scales to hundreds of genes and major topics in near real time. To evaluate the system's performance, we demonstrate the mapping of gene-concept clusters within a document collection from the prostate cancer literature against human review articles from that same literature. In a second example, we demonstrate the system achieves a high-level of agreement with expert reviews covering the literature of a major developmental pathway, the Wnt signal transduction pathway, both in the context of developmental biology as well as in the context of cancer progression.

4.2 Introduction

The biomedical literature continues to grow at an accelerated rate, yet the search engines most commonly used to access it remain the keyword-based retrieval engines like NCBI Entrez-Pubmed (http://www.ncbi.nlm.nih.gov/PubMed/) and Google Scholar (http://scholar.google.com). In active fields like cell signaling or oncology, the size of these engines' query results quickly overwhelms human reading ability. Making matters worse, due to the context-dependent nature of scientific research, the first or most recent article(s) returned are typically only a small fraction of those required to comprehensively describe the full body of knowledge contained in the literature on the queried disease condition or biomolecular process.

In order to interpret the results of any given returned result, then, users must not only select a few articles of interest from their search, but then also undertake the additional task of browsing at least in passing the co-referencing papers and related publications returned by the search engine. Quite often, review articles exist which assist by offering expert opinion and summarization of bodies of literature, but these typically focus on specific sub-disciplines within the literature. Once published, these articles do not update themselves to reflect new findings.

Even considering the publication of review articles, however, the overall growth in the literature is now such that even relatively limited searches often return overwhelming volumes of results. As of early 2008, a query of MEDLINE for the phrase "cancer AND epigenetics" retrieves 5,348 articles; limiting the same query to "epigenetics AND prostate cancer" reduces this number to a still-substantial 285 articles. Similarly, a query for "prostate cancer AND apoptosis" results in well over 3,000 articles and 472 reviews, an intractable number of papers for all but the most determined reader. Overall, the process of discovering the context and function of gene or disease processes within a result remains a formidable and time-consuming task for a human reader. The problem becomes even worse when discussing complex systems in biology within variable contexts, such as multi-factorial disease or signal transduction pathways with variable roles. A literature search of "Wnt AND signal transduction" for example (returning papers relating to the Wnt family of secreted signaling proteins) yields 3,525 articles, of which roughly 1,500 discuss Wnt-related

genes in a developmental biology context and 845 discuss Wnt's in the context of cancer biology. Currently, approximately 50 genes are believed to comprise the core of this pathway[21], yet extracting the oft-varying role of these genes from the hundreds of experimental publications describing them remains a task which challenges even expert human readers.

A number of biomedical search alternatives to the Pubmed search engine have been developed which attempt to better organize the result sets returned by queries. These include text displays of ontology-based clustered results [22], graphical [23] and textual [23] displays of clusters of documents. Also, some search engines include documents not indexed by MEDLINE (Google Scholar) but still present results in a series of text-pages like the Pubmed search engine. All of these primarily return abstracts or titles in lists or as node-edge graphs. Search results from these engines often do not directly display the precise distribution of named entities within those results in a single comprehensive view. Furthermore, in some cases the engines are often limited in retrieval size [22] on the underlying corpus, leading to undercoverage when analyzing the relationships between many hundreds of entries actually present in the result (for example, MeSH headings corresponding to documents and their genes within a given corpus).

Gene- and MeSH-based topic clustering applications in the biomedical literature have been reported in prior work, for instance PubGene [24], a system for automated extraction of explicit and implicit biomedical knowledge from publicly available gene and text databases to create a gene-to-gene co-citation network. The system described does not function as a search interface to article subsets; rather, it explores relationships and similarities within genes in MEDLINE abstracts. Other approaches describe clustering strategies using MeSH topics, such as the gene-to-phenotype clusters reported by Jennsen Korbel [25] but these are largely one-time analyses rather than search engines in their own right. In yet others, such as the heatmap queries in Lydia [26], the analyses or engines are not focused specifically on biomedical content.

The RefViz literature analysis tool [27] may perhaps be the closest available overview heatmap utility available to that which we discuss in this project, as it displays a literature clustering and retrieval heatmap for documents. Unlike Biosearch-2D, however, RefViz

does not cluster results based on organism-specific gene lists or controlled external ontologies. RefViz instead renders the distribution of topics into more of a word-based map rather than a gene-concept-centered map.

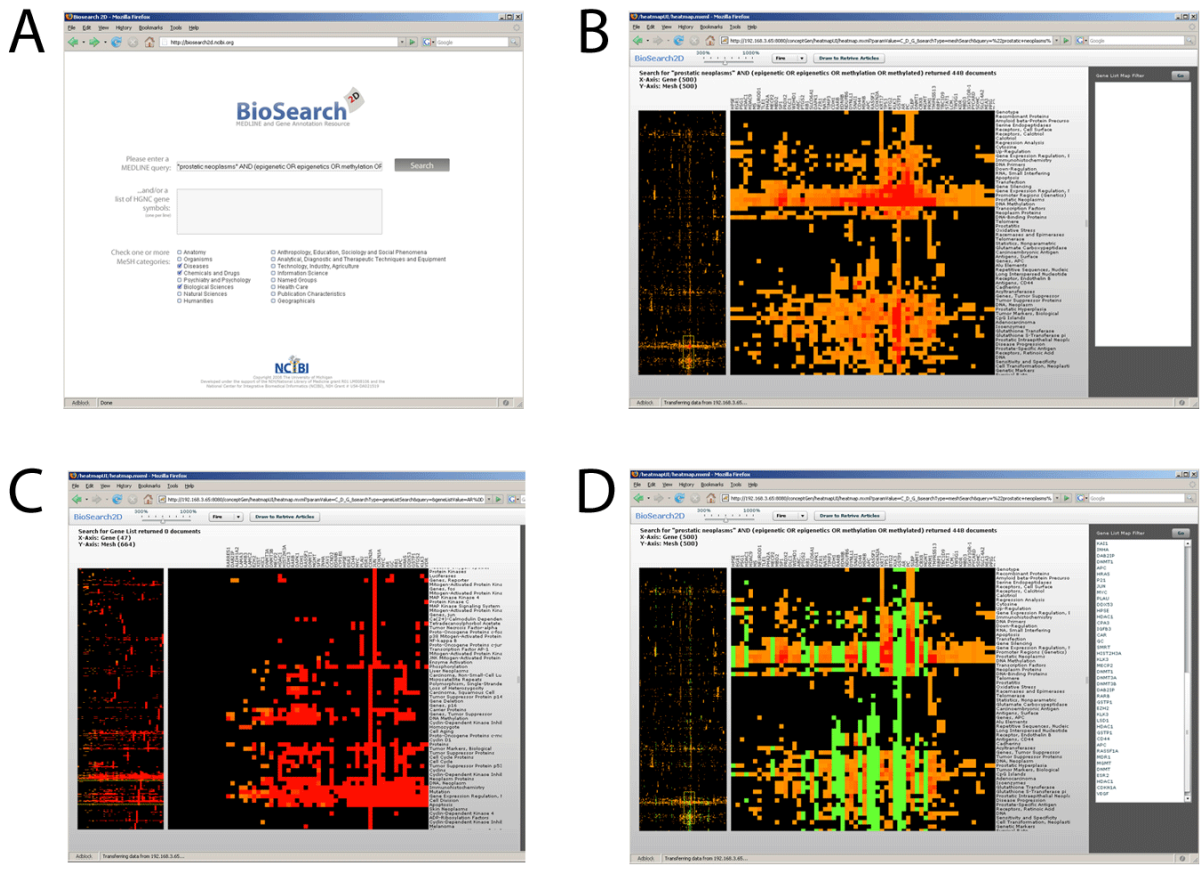
Previous work on information extraction in biomedicine includes a number of reports which attempt to extract information about genes from scientific texts using the co-occurrence of terms in a sentence or abstract[17, 28-31]. These approaches, like ours, extract genes within an actual biological context [24], [17], but unlike our current implementation, they do not attempt to summarize a corpus specifically using this approach nor allow for re-clustering specific subsets of documents according to user-selectable criteria. Both do report, however, that co-occurrence of gene names in an abstract frequently reflects an actual biological relationship between co-occurring genes.

Masys,et al.[32] describe a system of keyword profiles for genes based Medical Subject Headings (MeSH), but the system is not presented as a user-navigable search engine. A close comparison to our utility could be CoPubMapper by Alako, et al. [33] but like the other approaches, the analysis presented does not form a direct interface for a search into the literature (so the actual keyword clusters are hidden) and it is not implemented as a web-based utility, but rather was performed as a one-time analysis task. Alako, et al. [33] also report differences in the name tagging algorithm and normalization to our name matching algorithm.

Our search engine in contrast is primarily gene-versus-concept centered, and is a true web-based application, motivated by a need to analyze and explore the role of genes and their roles as described in a literature subset chosen at query time. Our first application for the system explores the prostate cancer genomic literature for those papers describing methylation and epigenetic changes in tumor progression. Rendering a heatmap of the genes versus MeSH topics relating to articles discussing the genes, the application scales to cover the many hundreds of genes observed in the corpus and the correspondingly large collection of MeSH topics corresponding to articles in which those genes are found. The map itself is rendered and presented via a Flash-based website, allowing rich, interactive, corpus-wide exploration and document retrieval guided by the image features themselves.

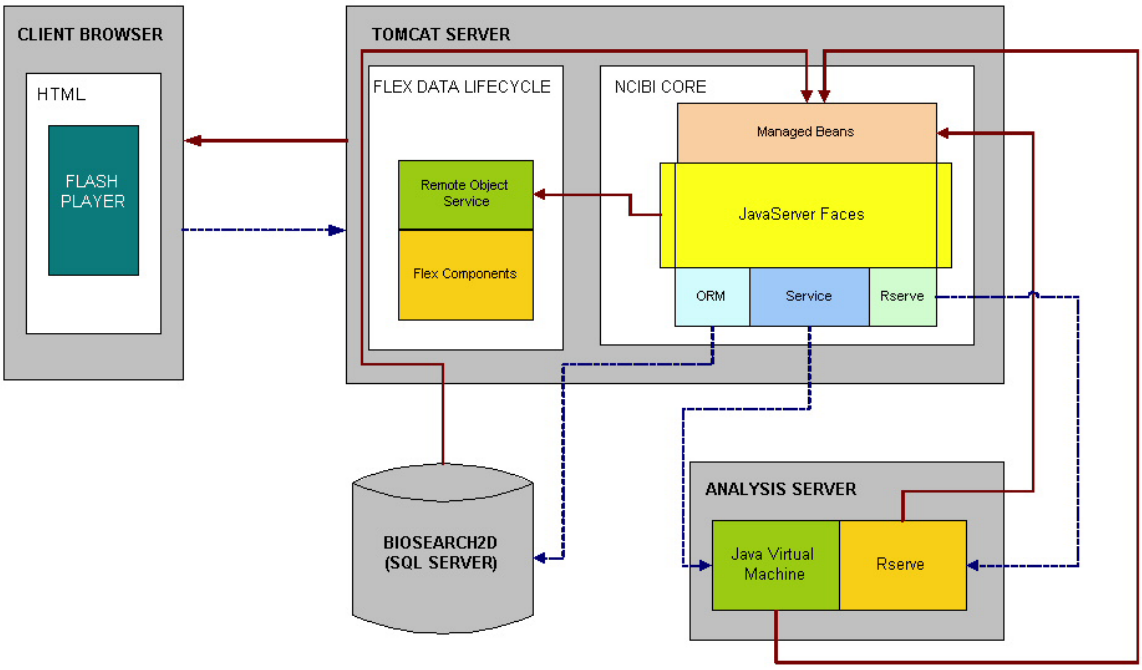
To demonstrate the coverage of these maps, we analyze the results from a focused major disease query, “prostate cancer AND epigenetics”, as well as the literature discussing a major signaling pathway, the Wnt pathway. We select topics within these collections and analyze the map coverage against human-authored reviews in both cases and a curated web resource in the case of Wnt. Our results suggest that an automated mapping of even a complex corpus in a heatmap corresponds closely to the gene-concept discussion provided by the human reviews and reference websites.

Figure 1: Screen displays from the BioSearch-2D website



Screen displays from the BioSearch-2D website, demonstrating the map results of a MEDLINE query "prostatic neoplasms AND (epigenetic OR epigenetics OR methylation OR methylated)". (A) Initial search screen for input of a search phrase (MEDLINE query) and/or an HGNC gene symbol list. (B) Result map for a MEDLINE query alone, showing a heatmap display of a portion of a genes-by-MeSH matrix in the document collection. (C) The result map for the MeSH topics corresponding to a user-submitted gene list (D) the gene-vs-MeSH heatmap from panel (A), showing genes from gene list from (C) labeled in green.

Figure 2: BioSearch-2D Architecture Overview



BioSearch-2D application architecture and data flow. Server-side components include the Apache Tomcat 6.0 web server in combination with Adobe Flex 3.0 (www.adobe.com), the Java Server Faces application framework (http://java.sun.com), the R statistical computing package [75] with R-serve (http://www.rosuda.org/Rserve/), and Microsoft SQL Server 2005. The client-side web-facing interface requires a modern web browser such as Mozilla Firefox or Microsoft Internet Explorer with Adobe Flash Player 9.0 browser plug-in installed (www.adobe.com). The Flash/Adobe Flex interface offers cross-browser compatibility and flexible user-interface features including the ability to instantly resize and recolor the image maps.

4.3 Methods

The biomedical literature enjoys substantial use of standardized nomenclature. A large proportion of published work has high quality metadata tags associated with individual papers (the standardized topic ontology known as MeSH). Our corpus visualization system detects named entities (genes, proteins, etc) in collections of biomedical articles and clusters the collection by gene versus function. The website then presents a two-dimensional searchable heatmap of tagged terms and their topics as an interface into the underlying collection.

4.3.1 Retrieval of Relevant Documents

First, document abstracts are retrieved from current (2008) NCBI MEDLINE XML data files (http://www.nlm.nih.gov/databases/leased.html) into a Microsoft SQL Server relational database. Tagging and gene-vs-MeSH tables are pre-computed with a Java-based gene name tagger (manuscript in preparation) and stored to minimize processing time. For individual user queries, the system only queries the NCBI e-utilities for Pubmed ID's (pmid) which are used as a basis for selecting the document collection to be displayed.

4.3.2 Gene Name Tagging

Gene name tags in our system are pre-computed following loading of the MEDLINE XML data into the relational database. The system employs a two stage approach: in the first phase a dictionary of names and synonyms is assembled. In the case of gene names, we use the NCBI Gene database as a source of both names and synonyms supplemented with synonyms from cross referenced databases including HGNC, the Jackson Laboratory Mouse Genome Informatics and Ensembl/EBI. Note that every dictionary entry is associated with an entry in the gene database. Thus, when a noun phrase is tagged by a dictionary match, we have an explicit link to a well defined information resource that can be used for further data integration.

Where MeSH index terms are available and a species index term is applies to the article, we use this species information to scope the dictionary of relevant gene names and synonyms.

To assess the performance of our gene name tagging, we use the NCBI GeneRIF sentences as a "gold standard". The GeneRifs are sentences deposited by users as examples of text referring to a gene[76]. On the task of gene name tagging, we used the NCBI GeneRif collection as a gold standard. Of 152,517 GeneRifs referring to human or mouse genes, we correctly tag 102,284 for a recall of 67%. Of the 132,582 GeneRifs that were tagged, the correct gene was identified 77% of the time. This is a lower bound on the precision of tagging because the relevant noun phrase was often simply not tagged in GeneRifs where our tagger failed to identify the correct gene. Our results compare favorably to other recent results reported in this field, but this is not a definitive comparison[77]. Interestingly, although each GeneRif is associated with only a single gene in the NCBI database, we identified 402,083 distinct references to genes in this data set. When biologists refer to genes, they typically refer to several genes. This reinforces the value of our document concept map paradigm.

4.3.3 MeSH Entries

MeSH terms (http://www.nlm.nih.gov/mesh/) are extracted from the MEDLINE records for individual articles in the analysis demonstrated.

4.3.4 Hierarchical Clustering and Generation of Heatmap Summary Display

Following name tagging, a matrix is generated of genes versus topics, and hierarchical clustering along both genes and MeSH axes is performed in the R statistical computing package *hclust* function, using the complete linkage method. Different clustering schemes are available and can be set via parameters into *hclust().*

After clustering, all headers and document heatmap data are rendered dynamically in a Flash-based application, which allow the use of a selector marquee box for users to select areas within the heatmap for retrieval of articles corresponding to selected areas on the heatmap.

As described in Figure 2, the system is deployed on the Apache Tomcat 6.0 web server in combination with Adobe Flex 3.0 (www.adobe.com), the Java Server Faces application framework (http://java.sun.com), the R statistical computing package [75] with R-serve (http://www.rosuda.org/Rserve/), and Microsoft SQL Server 2005. The client-side interface

requires a modern web browser such as Mozilla Firefox or Microsoft Internet Explorer with the Adobe Flash Player 9.0 browser plug-in installed (www.adobe.com).

The R processing server is a 2 GHz Intel Xeon CPU running Red Hat Linux with 8Gb of RAM and the web and database servers are 2Ghz Intel Xeon CPUs running Microsoft Windows Server 2003.

On our current hardware implementation, document retrieval limits are currently set to 5000 documents per query to avoid remote network timeouts. The application's overall load and response time is largely limited by the initial remote query step (NCBI Pubmed ID retrieval). Time from submission until map display currently ranges from an average of 10-15 seconds for submitted queries, regardless of document count. MeSH and gene symbol counts returned are currently limited to the top 500 MeSH terms returned for the overall gene list and the top 500 genes for the overall MeSH term count principally to preserve readability of the heatmap on a standard display.

In the initial search window, the system allows the user to select MeSH tree sub-categories in order to filter the search to a subset of MeSH. In the queries discussed in the following sections, queries were restricted primarily to the MeSH categories "Diseases", "Chemicals and Drugs", "Biological Sciences" and "Anatomy".

4.4 Review Article and Website Selection

The review articles selected for the prostate cancer evaluation, Li, et al. Epigenetics of Prostate Cancer *Front. Biosci.* 12, 3377-3397 and Nelson, W.G., Yegnasubramanian, S., Agoston, A.T., Bastian, P.J., Lee, B.H., Nakayama, M. and De Marzo, A.M. (2007) Abnormal DNA methylation, epigenetics, and prostate cancer, Front Biosci, 12, 4254-4266. were selected on the basis of their deep coverage of a rapidly-evolving subject (the role of epigenetic mechanisms and modifications in prostate cancer) as well as for their relevance to the study of a human disease with major clinical significance. Both reviews are written by highly-cited and published authors in the respective areas, and provide deep coverage into the field of investigation they cover, with the Li article citing 253 references and the Nelson article citing 143 references respectively.

4.5 Results

We find that browseable dynamic heatmaps can be a powerful aid in summarizing the function of genes in literature collections. In one case, we analyzed in detail a corpus of prostate cancer epigenetics articles as returned by NCBI Pubmed in January 2008 and compared the map coverage to the three most recent human review articles published on this topic. We chose this subject area as it is a rapidly-evolving field within a subject of substantial clinical importance. Accordingly, while the results discussed in this paper relate primarily to this focused corpus, the search engine in use accepts arbitrary NCBI/MEDLINE user queries for processing and is not limited to the oncology literature space.

In the MEDLINE query “"prostatic neoplasms" AND (epigenetic OR epigenetics OR methylation OR methylated)”, 448 documents are retrieved and processed by our system. A selection from the final clustered gene vs. MeSH image map is shown in Figure 1. On the website, the viewer sees a combined view, including a small “birds-eye” compressed map for browsing the results together with an exploration window for focusing on individual clusters. From this map, general trends specific to this MEDLINE query result set are easily observed, including a large vertical line corresponding to the androgen-receptor gene, PSA, DNA methyltransferases, as well as a number of MeSH terms including “prostatic neoplasms”, “DNA methylation”, and “Adenocarcinoma”. In addition, a number of smaller clusters of other groups of genes include the apoptosis regulators Akt, Bcl-2 and apoptosis-related caspases, EZH2, histone methyltransferase genes, GSTP1, and DNA methylases.

Transmembrane mucins MUC1 and MUC4 are described in a report by Singh, et al. as being regulated by epigenetic mechanisms in a cell line model[78]. The cluster includes genes associated with DNA hypermethylation in the context of prostate cancer including E-cadherin, pi-class glutathione S-transferase, and the tumor suppressor CDK2N.[79]

We find that the automated gene-by-MeSH clustering itself yields genes which often physically interact and are clearly related to the major disease process observed in the query. Examples include EZH2, which is known to associate with other PcG proteins, EED and SUZ12, within the context of PRC2/3 complexes[80]. Also co-clustering are androgen receptor, FAS, and the androgen-stimulated gene PSA.

In Figure 3, a cluster within the image generated for the MEDLINE query "EZH2" references MeSH topics detailing the function of proteins associated with the Polycomb-group protein EZH2[81] and its binding or co-regulated partners, including EED, HDAC (histone deacetylases), SUZ12 [82], DAB2IP [83]. The view also applies to other stages of disease: in a query of the term "TMPRSS2", the androgen-responsive TMPRSS2 fusion gene[84, 85] associated with histone genomic epigenetic reprogramming in prostate tumors is shown in conjunction with ERG, ETV1, ETV4 and the MeSH term "Prostatic Intraepithelial Neoplasia", "Recombinant Fusion Proteins", "Gene Rearrangement". These reflect the role of these fusion genes as described in early prostate cancer development in the literature. [84-87].

Table 7: Coverage of Gene/MeSH Clusters by BioSearch-2D Compared with Human Review Articles from the Prostate Cancer Epigenetics Literature.

.

Source	Sections (within chapter)	Sections Covered	Gene Name Accuracy
Prostate Cancer Epigenetics Reviews [88] [89]	6	5/6*	77-90%**
Wnt Signaling Review [90]	5	5/5	--
Wnt Signaling Website [21]	1	--	77-90%**

* website covers all topics in this review article except therapeutic areas.

** the accuracy of the gene name depends on the accuracy of the underlying tagging algorithm.

4.6 Evaluation of Coverage Against Human Review Articles

We evaluated the gene-concept map coverage of the heatmap against human reviews describing epigenetic modifications in prostate cancer.

In the first of these reviews, by Nelson, et al. [88] we find that the reviewer comprehensively describes major disease processes, including "DNA Hypermethylation", "Heterochromatin and Epigenetic Gene Silencing", "DNA Hypomethylation", "Demethylation and Loss of Imprinting", "DNA Methyltransferases and Cancer Development", "DNA Methylation Changes in Prostate Cancer", "Sensitive Detection of Hypermethylated CpG Islands as Prostate Cancer Biomarkers", and "Epigenetic Gene Silencing as a Therapeutic Target for Prostate Cancer Prevention and Treatment." Our search clusters capture the majority of the genes in the major topic areas, including EZH2, MeCP2, the histone deacetylase HDAC1, Mi-2, DNA methyltransferase, DAB2IP and INK4a. In the loss of imprinting/hypomethylation section, another cluster captures MDB2, SP1, DNA methylation, but does not capture IGF-2. In discussing the DNA methyltransferases, the cluster captured DMNT1 which formed the focus of the discussion. The author discusses the role of GSTP1 at length, including a discussion of TMPRSS2 and the ETS family genes involved in gene fusions. Of the topics mapped by our clustering algorithm, we find the author does not discuss the apoptosis genes (aside from TNF-associated apoptosis), the sirtuins, the carboxypeptidases, nor the cell cycle checkpoint genes in detail as described in our overall image map.

A second review by Li, et al. [89] divides the epigenetics of prostate cancers into similar sections. In our cluster maps, topics not covered include the specific details on age or dietary factors discussed by Li, et al.

We exceed the coverage seen for EZH2 and DAB2IP in the article (including genes such as EED and SUZ12 which are not discussed in the section on histone modifications in prostate cancer).

For the Wnt genes review, we focused on a recent review discussing the role of Wnt proteins in cancer authored by Nusse, et al. [90] and found substantial overlap with the genes mentioned in the review (APC, Axin, beta-catenin, LRP5, Dsh, and Dkk). Fig. 3 illustrates

the central result of this query. As with the prostate cancer example, we find additional topics relating to Wnt and signal transduction not discussed in the review. These include papers covering the induction of fibroblast growth factors in tumors, papers discussing the role of various frizzled family members in tumor progression, and a specific group of articles discussing the role of Wnts in regulating apoptosis different cancer types like hepatomas, renal cell carcinomas and hepatoblastoma.

In the Wnt genes website, we find substantial agreement between our genes mapped and the targeted annotation. Using as a reference the most current (2008) Wnt target gene list (http://www.stanford.edu/%7ernusse/pathways/targets.html), we generated for comparison a map of the MEDLINE query "Wnt AND signal transduction AND TCF AND target". As with the previous review, we find substantial overlap in the list, but with genes annotated with additional MeSH terms according to the literature, including those expected such as "colonic neoplasms", "Wnt proteins", "beta Catenin", "Promoter Regions (genetic)", "Phosphoproteins", and "Intercellular Signaling Peptides and Proteins".

Overall, we find that for genes vs. MeSH topics, coverage of the rendered searchable map matches that found in reviews excepting certain non-gene-rich topic areas such as novel therapeutics.

4.7 Conclusions

We have developed a system to quickly render the gene mentions within of large document collections into a single heatmap. Genes clustered according to human-curated document annotations can assist in the analysis of larger document collections by reducing the many hundreds of abstracts in a collection into a series of easily-identified pixel clusters on a heatmap. A side effect of the clustering is that the relative size and position of genes and topic clusters roughly corresponds to the importance of these topics as presented in the underlying corpus.

A current limitation of the system remains its limitation to abstract texts only. We anticipate that, as the availability of full-text document and open-access biomedical article collections improves, so will the coverage of the displayed heatmaps. Additionally,

improving the gene name tagging accuracy could reduce gene mislabeling and identification errors.

The overall heatmaps generated by BioSearch-2D are similar to "concept maps" in their rendering of gene sets into sub-groups described by common MeSH ontology terms. The map displays the distribution of terms amongst the genes of interest and can render genes together according to common disease processes. In one example, the metalloproteinases are clustered together in a group also containing the term "neoplasm invasiveness", an association which is widely established in the biomedical literature.

The use of a heatmap as the primary representation for the literature permits very fine-grained representation of the contents of the corpus while allowing a human viewer to very quickly observe the gene groupings in the collection (the androgen-receptor, Polycomb group proteins, histone deacetylases, anti-apoptotic proteins, and GSTP1) along with their function as described by individual documents. We are currently adapting the search engine to map full-text document collections and additional named-entity classes (cell lines, substance names, etc.) as they become more easily available.

4.8 Funding

This project was supported in part by a grant for the NIH/National Library of Medicine R01 LM008106 and is released as part of the National Center for Integrative Biomedical Informatics (NCIBI), NIH Grant # U54-DA021519.

BioSearch-2D can be accessed online at http://biosearch2d.ncibi.org

CHAPTER V

BioSearch-2D: Literature-Based Context-Specific Functional Annotation for Genomic Data

5.1 Abstract

In recent years, a large quantity of functional annotation software has been developed to interpret the biological function of signature gene lists from high-throughput genomic experiments. By primarily adopting the standardized Gene Ontology (GO), these systems annotate gene lists with statistically-significant terms describing major biological processes, cellular components and molecular function. In most cases, the output produced by these tools consists of static term lists of statistically significant matches ranked according to the relative enrichment of tagged terms present within the submitted list. A number of other ontologies remain largely underrepresented in these efforts, however, including the NCBI's MeSH vocabulary, which comprehensively annotates the biomedical literature and describes a broad range of topics in biomedicine, from clinical terminology to terminology about scientific research methodology. We have developed a dynamic web-based utility, BioSearch-2D, which automatically matches gene names to MeSH annotations and then automatically renders a browseable gene-vs-MeSH "topic map" of statistically significant terms from user-submitted gene lists. Unlike standard annotation engines, BioSearch-2D renders dynamic maps of literature-based topics for gene lists which cover the many clinical and physiological terms present in the MeSH ontology. In addition, our engine offers allows specific filtering of the annotation via MEDLINE queries in order to prioritize specific biomedical contexts. To demonstrate the performance of this engine, we analyze a set of six human-annotated reference gene sets and demonstrate that our coverage matches and in many sets augments the results from traditional Gene Ontology-based annotation engines. Our map annotation of these lists yields clinical and physiological relationships in data sets

from cancer signature lists to canonical pathways which are not easily identifiable by standard annotation software. Signature gene lists annotated include those involved in cellular adhesion processes, genes involved in the cell cycle, DNA repair, and genes relevant to cellular adhesion in metastatic lung disease.

5.2 Introduction

The annotation of gene list results produced by high-throughput genomics and proteomics experiments has resulted in a vast number of gene expression signatures and canonical reference lists corresponding to important disease and clinical states. Typically, the functional annotation of these gene lists into biological context relies on annotation utilities which calculate the relative enrichment of ontology terms for genes found in the input list compared to the term frequency assigned to genes in a genome-wide context. The majority of these annotation utilities employ the Gene Ontology[34] as their primary annotation ontology. Additionally, some provide additional annotations such as protein-protein interaction lists, protein functional domains, disease associations, pathways, sequence features, homologies, and selected curated literature references [35-38] [39-41]. These utilities are varied, and include both executable software as well as websites like GoMiner [42], EASEonline [35], GeneMerge [43], eGOn [44], FuncAssociate [45], GOTree Machine (GOTM) [46], GOSurfer [47, 48], Ontology Traverser, CLENCH [49], GOToolBox [50], FatiGO [39, 40, 51], and DAVID [35-38]. A complete review of these utilities is described by Khatri, et al. [52].

Additionally, annotation tools like the Molecular Concept Maps described by Rhodes, et al. [53-55] are available which link microarray studies to a number of oncology-related ontologies in order to better allow annotation of clinically distinct cancer gene profiles. In one published report, Tomlins, et al. describe common shared genes between cancer signatures annotated between different cancer types and specific gene repression signatures in both breast and prostate cancers, demonstrating the power of incorporating non-GO ontologies in a highly-focused biological context. [53-55].

To date, Gene Ontology-based annotation engines rely on an intermediate curation step to assign genes to ontology terms based on literature or experimental observation. As Khatri, et al. note, these mapping efforts have historically been fairly accurate [56] and extensive yet

mostly assigned in an automated fashion (as of February 2008, there exist 182573 GO annotations for 35113 human genes, of which only 52,246 were not derived electronically) (http://www.geneontology.org). By contrast, MeSH annotation is performed manually by human curators on individual MEDLINE articles. Linking article-derived MeSH terms to genes, therefore, could provide a more tightly-coupled gene annotation than annotations obtained through secondary-source ontologies.

Khatri, et al. further highlight a key limitation to the current batch of annotation engines, in that annotations "related to those genes [which] are involved in several biological processes" are limited to single contexts. Due to the nature of the GO hierarchy, most current tools weight biological processes equally. In effect, these tools make "restricting the query to specific clinical areas…a challenge since the basic annotation itself is largely restricted to basic biological processes". They describe a specific example in the case of BRCA, which has a distinct biological roles as both tumor suppressor as well as in carbohydrate metabolism [52]. Depending on the gene signature in which it is found, the annotations may differ for the gene, which in turns impacts the accuracy of any biological inferences made on that annotation.

In terms of user-interface, the vast majority of existing utilities remain largely text-based, with results returned being large term lists with statistical significance values assigned to each term. These text lists are often produced in batch manner and returned as series of dense text annotations which seldom reflect internal categories between the genes analyzed. A few graphical interfaces have been developed to address the usability limitations of these text results, including two-color plots rendered by DAVID, where they are described as "… the most powerful graphic presentations in DAVID applications" by the authors. [35, 57]

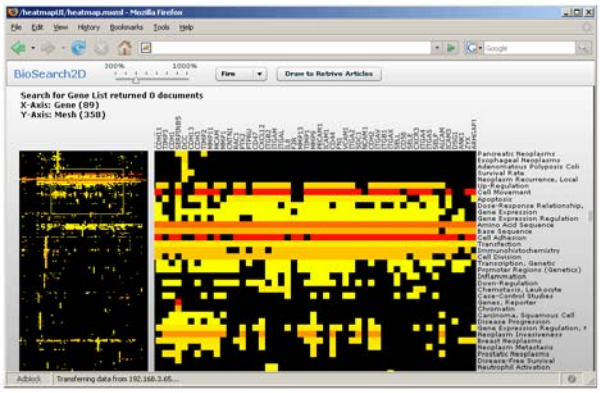
We have developed an integrated MeSH annotation system in conjunction with a literature concept mapping utility, BioSearch-2D. From a user-submitted gene list, the system renders hierarchically-clustered, dynamic two-dimensional maps representing the distribution of a large set of human gene identifications in biomedical text versus selected MeSH terms. Coloring on the map corresponds to statistically-significant annotations assigned to MeSH terms. These maps directly represent the distribution of MeSH terms corresponding to submitted gene lists as well as the statistical significance in a single unified

display, instead of in a series of text lists. We find that the maps match key functional annotation assignments produced by GO-based engines, as well as use a two-dimensional map to render context-specific annotations clustering and intuitive distribution plots which identify functional subgroups in submitted gene lists.

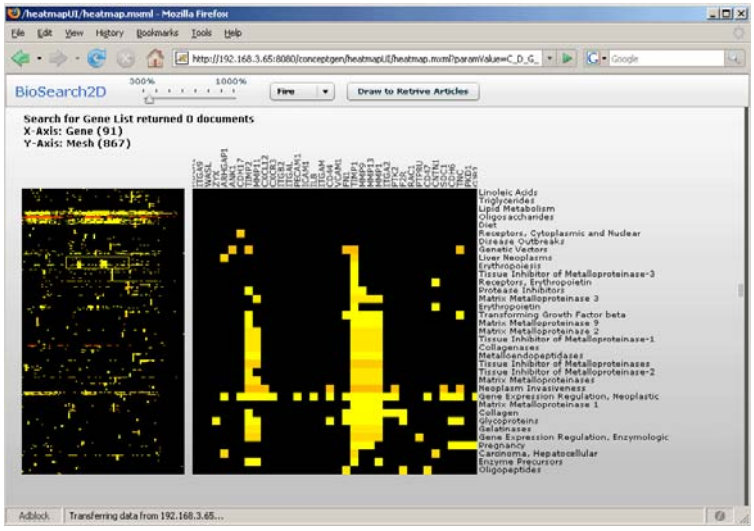
Figure 3: Screenshots of the BioSearch-2D Gene Annotation Website

A

B



C



Screenshots of the BioSearch-2D Gene Annotation Website. (A) Input of literature query or a gene list together with several MeSH sub-categories. (B) Results window showing the annotation of a 95-gene set of cancer-related genes involved in cell adhesion and metalloproteinases [91]. 89 of the 95 genes had highly-significant annotations matching the Brentani annotation ("cell adhesion", "cell movement", "neoplasm invasiveness", "cadherins", "metalloproteinases"). In the window, two scrollable maps (overview at left and detail at right) contain non-black pixels represent MeSH terms (rows) from documents in MEDLINE where the corresponding gene (columns) was detected. (C) Detailed view of a cluster of metalloproteinases (MMP9, MMP13, etc.) from the original list, and the corresponding MeSH terms ("Neoplasm Invasiveness", "Matrix Metalloproteinases", etc.).

5.3 Methods

The biomedical literature enjoys substantial use of standardized nomenclature, and a large proportion of published work has high quality metadata tags associated with individual papers (MeSH terms). Our corpus visualization system detects named entities (genes, proteins, etc) in collections of biomedical articles and clusters the collection by gene versus function. The website then presents a two-dimensional searchable heatmap of tagged terms and their topics as an interface into the underlying collection. For gene annotations, we expanded the map concept to include MeSH terms from MEDLINE which match genes in user-submitted gene lists.

5.3.1 Gene Name Tagging

As described by Santos, et al (manuscript in submission), gene name tags in our system are pre-computed following loading of the MEDLINE XML data into the relational database. The system employs a two stage approach: in the first phase a dictionary of names and synonyms is assembled. In the case of gene names, we use the NCBI Gene database as a source of both names and synonyms supplemented with synonyms from cross referenced databases including HGNC, the Jackson Laboratory Mouse Genome Informatics and Ensembl/EBI. Note that every dictionary entry is associated with an entry in the gene database. Thus, when a noun phrase is tagged by a dictionary match, we have an explicit link to a well defined information resource that can be used for further data integration.

Where MeSH index terms are available and a species index term is applies to the article, we use this species information to scope the dictionary of relevant gene names and synonyms.

As described in a companion manuscript, we assess the performance of our gene name tagging, using the NCBI GeneRIF sentences as a "gold standard". The GeneRIF's are sentences deposited by users as examples of text referring to a gene[76]. On the task of gene name tagging, we used the NCBI GeneRif collection as a gold standard. Of 152,517 GeneRIF's referring to human or mouse genes, we correctly tag 102,284 for a recall of 67%. Of the 132,582 GeneRIF's that were tagged, the correct gene was identified 77% of the

time. This is a lower bound on the precision of tagging because the relevant noun phrase was often simply not tagged in GeneRIF's where our tagger failed to identify the correct gene.

For each gene in the annotation database, the top 500 MeSH terms are computed and stored in the database, ranked according to the number of distinct PMID's returned for the gene-by-MeSH combination. For example: for the human gene BCL2 (NCBI GeneID 596), the top genewise tagged MeSH terms sorted by decreasing document count are: "Humans" (9837 articles), "Apoptosis" (6220 articles), "Proto-Oncogene Proteins c-bcl-2" (6136 articles), "Proto-Oncogene Proteins" (2395 articles), etc. MeSH terms are also assigned the major MeSH category, to allow selecting a subset of the MeSH tree for annotation and clustering.

For every submitted gene list, statistical significance scores for MeSH term enrichment are calculated for every MeSH term assignable to the submitted list versus the null set, which are all genes for which MeSH terms could be assigned by the above gene name tagging algorithm (15513 genes overall in the human genome). P-values for enrichment of resulting terms are calculated identically to the DAVID EASE score calculation (http://david.abcc.ncifcrf.gov/content.jsp?file=functional_annotation.html#summary): for every MeSH term resulting from the gene list, a 2x2 contingency table and the enrichment of the gene list analyzed with a modified Fisher's exact test. P-values for each MeSH term are multiplied by the total number of MeSH scores (Bonferroni correction).

For the submitted gene lists, a two-dimensional matrix is calculated, with genes as the columns and MeSH terms as the rows. This matrix is then hierarchically clustered with a Java processing pipeline submitting the matrix into an R server for clustering via the *hclust* function as previously described by Santos, et al (manuscript in submission). Finally, color is assigned to each row by the gene list's degree of enrichment for that term (Fisher's exact test; red: p-values < 0.01, orange: $0.01 \leq$ p-value ≤ 0.05: yellow: p-value > 0.05).

As positive control for the accuracy of our annotation we compared the annotation of 6 curated gene lists from the Broad Institute (http://www.broad.mit.edu/gsea/msigdb/genesets.jsp?collection=CGP) [92],selecting from the C2 "curated gene sets" tree, "chemical and genetic perturbations", and "canonical pathways" to that rendered in the DAVID Gene Ontology annotation engine (Table 8) We

also analyzed 2 positional gene sets as negative control, selected from chromosome 1. GO term annotations were assigned according to the DAVID assignments, using the default settings

In order to identify related genes from the search relating to a disease process, we searched the Brentani cell adhesion dataset (95 genes) against the query "cell adhesion AND neoplasms AND metastasis" in the BioSearch-2D main literature query window. An example of this search method is described in an accompanying manuscript by Santos, et al. Briefly, the map produced is produced from genes and MeSH terms from our literature database corresponding to documents returned by NCBI's Pubmed as matching the query. 1,698 documents were returned in the query, of which the maximum 500 genes and 500 MeSH terms by unique article count were clustered and mapped from query result corpus.

5.4 Results

Gene sets input into our initial screen (see Figure 3) are rendered into a navigable heatmap clustered and colored according to the most significant annotations (Fisher's exact test with Bonferroni correction applied to correct for multiple testing in each result map).

In Figure 3, the initial query window allows for input of a literature query or a gene list and selection of MeSH sub-category for annotation. These MeSH sub-categories include clinical annotations (disease types, biomolecules, drug molecules) which are not present in the standard Gene Ontology annotation. In the illustrated example, the results window shows the annotation of a 95-gene set of cancer-related genes involved in cell adhesion and metalloproteinases [91]. 89 of the 95 genes had highly-significant annotations matching the Brentani annotation ("cell adhesion", "cell movement", "neoplasm invasiveness", "cadherins", "metalloproteinases"). These terms are clinically meaningful and include terms like "Neoplasm Invasiveness" and metalloproteinases which are widely described as being related processes in the biomedical literature (255 articles published on these combined topics in 2006 alone, 1100+ articles altogether in the past five years). In the bottom view, two scrollable maps (overview at left and detail at right) contain non-black pixels represent MeSH terms (rows) from documents in MEDLINE where the corresponding gene (columns) was detected. The bottom row shows a cluster of metalloproteinases (MMP9, MMP13, etc.)

from the original list, and the corresponding MeSH terms ("Neoplasm Invasiveness", "Matrix Metalloproteinases", etc.).

To assess in more detail the completeness of the annotation, we selected 6 gene sets from the Broad Signature Gene List Database [92] published by Brentani, et al. [91], together with two positional gene sets selected at random as negative controls.

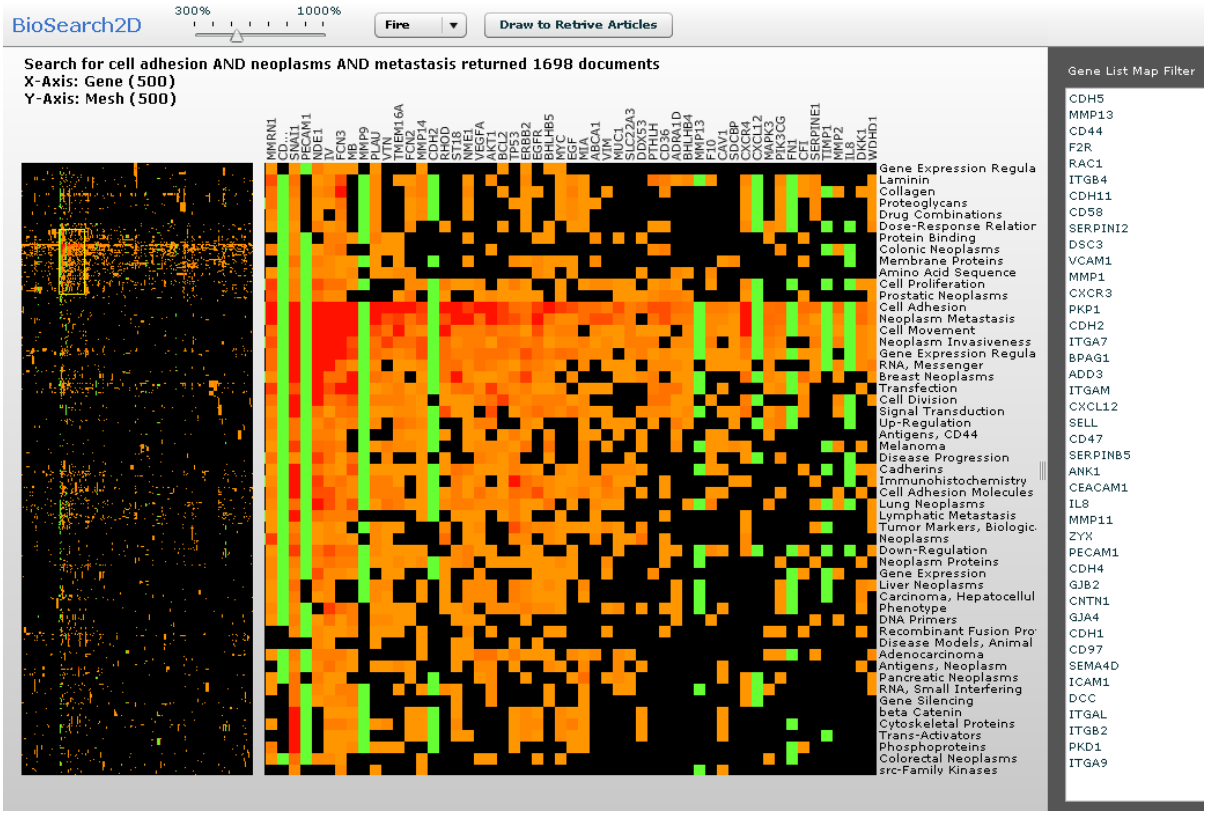
For both engines, the positional gene sets do not return meaningful MeSH or GO terms, and match a sharply reduced number of genes (<20-30% of either list) in the gene list. In contrast to the results returned by two-dimensional displays such as the DAVID functional map, our results can also be easily integrated into a literature search to find genes which may be related to the original list but are not included in the submitted gene list.

We find the agreement between gene tagging based on MeSH terms to be high on curated gene lists. Table 8 details the results from 6 curated sets from Brentani, et al. along with a negative control of two gene sets selected at random from the positional gene sets available. In the cell adhesion set, 89/95 genes found were tagged by our engine, and significant MeSH terms included cell adhesion, cell adhesion molecules, cell movement, cadherins, among others relating to the topic of cell adhesion (a complete list can be obtained from the map software). These terms agree well with the GO terms returned by the DAVID GO annotation engine, including cell adhesion, cell-cell adhesion, integrin, cell-matrix adhesion. In the CELL CYCLE dataset, 89/95 genes are tagged, with clusters returned that agree well with the most significant GO term annotations.

Table 8: Functional Annotation Comparison: Terms from BioSearch-2D MeSH Annotation Compared to Gene Ontology Annotation Terms from DAVID (Dennis, et al., 2003)

Set Name (genes tagged in BioSearch 2D/genes in gene list)	Description	BioSearch-2D Map Top Mesh Terms (MeSH categories C,D,G; ($p<0.01$ unless noted)	Top GO Clusters by Significance (DAVID default settings)	Clusters Observed in Maps where MeSH annotation ($p<0.01$)
Brentani CELL ADHESION (89/95 genes)	Cancer related genes involved in cell adhesion and metalloproteinases	cell adhesion, cell adhesion molecules, cell movement, cadherins, integrins.	cell adhesion, cell-cell adhesion, integrin, cell-matrix adhesion	Cell Adhesio, Matrix Metalloproteinases, Lymphocyte Activation, beta-catenin, cytoskeletal proteins, cadherins
Brentani CELL CYCLE (82/86 genes)	Cancer related genes involved in the cell cycle	Cell cycle proteins, cyclins, protein-serine-threonine kinases, cyclin-dependent kinases, mitosis,	cell cycle, regulation of progression through cell cycle, regulation of cell cycle	Cell cycle proteins, cyclins, protein-serine-threonine kinases, cyclin-dependent kinases, mitosis, G2 phase, phosphorylation.
Brentani CELL DEATH (70/76 genes)	Cancer-related genes involved in cell death	Apoptosis, DNA Damage, Cell Death ($p<0.05$), Cell Survival ($p<0.05$), DNA Sequence ($p<0.05$)	cell death, death, regulation of apoptosis, regulation of programmed cell death	Apoptosis, DNA Damage, Cell Death ($p<0.05$), Cell Survival ($p<0.05$), DNA Sequence ($p<0.05$)
Brentani DNA METHYLATION (22/24 genes)	Cancer-related genes involved in DNA methylation and modification	DNA damage, mutation ($p<0.05$)	DNA metabolism, nuclear protein, DNA binding	Binding sites ($p<0.05$), Apoptosis ($p<0.05$)
Brentani IMMUNE FUNCTION (51/54 genes)	Cancer-related genes involved in immune function	Lymphocyte activation, Sequence Homology ($p<0.05$)	Glycoprotein, response to biotic stimulus, defense response, humoral immune response, signal transduction	Lymphocyte Activation, Haplotypes
Brentani DNA REPAIR (36/41 genes)	Cancer-related genes involved in DNA repair	DNA Damage, DNA replication, Base Pair Mismatch, DNA ligases, Endonucleases, DNA repair enzymes	response to DNA damage stimulus, nuclear protein, DNA replication,, nuclease activity	DNA repair, DNA Damage, DNA replication, Endonucleases
Genes in cytogenetic band chr10p11 (33/142 genes)	Negative control (Genes in cytogenetic band chr10p11)	None: base sequence ($p<0.05$), carrier proteins ($p<0.05$)	Unavailable (<80% of the list not mapped to GO terms)	No significant clustering
Genes in cytogenetic band chr12q23 (33/106 genes)	Negative control (Genes in cytogenetic band chr12q23)	None: Amino Acid Sequence ($p<0.05$), Base Sequence ($p<0.05$)	No meaningful annotations: binding, calcium ($p=3.3e-2$)	No significant clusters except non-specific "Amino Acide Sequence"

Figure 4: Map generated of gene-vs-MeSH mappings from 1698 documents resulting from the query "cell adhesion AND neoplasms AND metastasis"



Map generated of gene-vs-MeSH mappings from 1698 documents resulting from the query "cell adhesion AND neoplasms AND metastasis" with highlighted (green) columns denoting matching genes from the Brentani cancer-related cell adhesion gene list. Relevant genes not present in the original list but highlighted from the literature as involved in cell adhesion and invasiveness include VEGFA, CD36, EGFR, AKT1, CXCR4, among others.

In Figure 4, we explore the literature heatmaps produced by the search engine independently of the cell adhesion gene list, and overlay the gene list to identify functionally related genes from the literature which are not present in the original list. A gene-vs-MeSH map generated from 1698 documents resulting from the query "cell adhesion AND neoplasms AND metastasis", with highlighted (green) columns denotes matching genes from the Brentani cancer-related cell adhesion gene list. Relevant genes not present in the original list but highlighted from the literature as involved in cell adhesion and invasiveness include VEGF (see [93] for a discussion), PTLHL (implicated in tumor migration into the bone microenvironment [94]), EGFR (implicated in lymphatic metastasis [95]), AKT1 (suppresses metastasis [96]), CXCR4 (implicated in cancer stem cell dissemination and metastasis [97] and a potential therapeutic target). It is important to note that several genes were mis-tagged by the automatic tagger, including "IV". These gene tagging errors are a consequence of the automated tagging algorithm and can be pruned by user feedback functions in development and by additional dictionary curation.

5.5 Conclusions

We have developed a MeSH annotation engine which graphically displays the most significant MeSH terms for a user gene list and which matches well the output from a reference standard GO annotation utility. Incorporating MeSH terms, with their associated clinical subheadings may assist in the functional annotation of gene lists relating to major disease processes like cancer metastasis. Furthermore, the ability to compare gene annotations with a whole-literature gene-vs-MeSH map can help identify related genes as described in the literature which are functionally related to the gene list submitted but which are not present in the original submission.

5.6 Online Access

BioSearch-2D can be accessed online at http://biosearch2d.ncibi.org

CHAPTER VI

Conclusion

Our results with automatic component identification and interaction detection in the Wnt signaling pathway suggest that natural language processing techniques are able to improve the coverage of canonical reference literature and signaling models. The high precision and processing speed of this automated signaling interaction pipeline demonstrates the value of full-parsers and statistical techniques. Using this approach as a "first-pass" filter into the literature offers a useful method for curation of databases and information resources in complex and rapidly evolving fields such as signaling pathways. We find that even though the recall rates with respect to the known canonical models do not yet match those of an expert human reviewer, the system could nonetheless succeed in detecting a large percentage of the protein-protein interactions reported in the literature.

In the future, we expect to capture directionality and type of interaction in a more robust way for our assertions; this will require additional template development, and may require the use of an external ontology for an outside reference source for error-detection of incorrect assertions. The role we most expect this system to serve is a real-time scanning facility for new articles, searching for newly-reported interactions. Automated computational methods are capable of analyzing a much broader coverage of literature than would be feasible for a human reviewer to perform. In this role, there is a premium on specificity to avoid overloading the manual reviewer with erroneous matches, and our results suggest that deep-parsing, automated natural language processing technology is now capable of achieving this requirement.

We found that our auto-categorization module, using statistical and natural-language parsing techniques, allowed us to build a named entity list at run-time, rather than requiring a cumbersome fixed named entity assembler before the processing. This approach was

perhaps our main advantage in this pipeline, because unlike general English-language texts, the biomedical literature enjoys a substantial human-curated hierarchical index via the MeSH tags provided by MEDLINE.

MeSH indexing provides a powerful tool for building reference and background article sets that can be used to search a specific article corpus for biologically-relevant named entities which are typically over-represented with high statistical significance. In our pipeline, the fast partial parser CASS served a useful role in assigning multiple-word entities. Moreover, its ability to efficiently process very large collections of text allowed us to extract these entities in a fairly comprehensive manner. The combination of fast partial-parse, exploiting MeSH indexing, and statistical analysis of multiple word phrases significantly simplified our task of assembling a comprehensive term list.

While some sentences in biomedical text are too complex to be accurately parsed using current technology, we find that parsers such as the Link parser [19] are able to accurately and efficiently parse the majority of sentences in the molecular biology literature. Using the integrated approach described above, we are beginning to be able to analyze the knowledge encoded in biomedical text.

Furthermore, our application of a heatmap for document search and genomic functional annotation demonstrates that context-specific data summarization can be successfully achieved in a very complete, near real-time manner over a large corpus.

In a second application, we explored the use of a dynamic map as a means of summarizing the biomedical literature. The system, BioSearch-2D, renders very fine-grained representations of large document collections while allowing human readers to very quickly observe the gene groupings in the collection. Examples shown include the androgen-receptor, Polycomb group proteins, histone deacetylases, anti-apoptotic proteins, and GSTP1, along with MeSH terms detailing their function. The overall coverage of this map closely matched the coverage of the same literature provided by expert human reviews in both written articles as well as curated web repositories. We are currently adapting the search engine to map full-text document collections with additional named-entity classes (cell lines, substance names, etc.) as they become more easily available.

The dynamic maps generated by BioSearch-2D seek to convey a similar meaning as a "concept maps" in their rendering of gene entities with the MeSH ontology terms. In contrast to standard node-edge graphs, however, the precise distribution of biological facts in the map can be assessed immediately by the viewer. By allowing an interactive search and rendering the content of hundreds of documents into a single map, the map intuitively displays related functions for genes within those documents and places them in a tightly-defined contextual role. Complex genomic pathways such as the Wnt pathway play differing roles depending on their context, and a heatmap representation coupled with a clustering algorithm allows for their improved annotation as the biomedical literature evolves. Further development of the BioSearch-2D engine could incorporate image feature-detection algorithms to assist users in selecting clusters of interest. As an initial screening component, we intend to color the side axes for the display according to major MeSH tree function. For many of the clusters displayed, such as that for EZH2 in the prostate cancer cluster, splits performed by the automated clustering algorithm occasionally partition important topics in the image into clusters too small for the user to immediately identify. We anticipate that users navigating the image may benefit from both additional map coloring options as well as different clustering algorithm parameters from the initial query.

We believe that both the automated assertion extraction software and the large-scale annotation and summarization abilities of the BioSearch-2D engine could greatly assist curators in reviewing and integrating data from the literature on complex signal pathways like the Wnt pathway. Both of these systems accomplish in hours (Wnt protein interaction software) or seconds (BioSearch-2D) tasks that would demand orders of magnitude more time from a human reviewer. As the accuracy of named-entity recognition improves and additional databases become available, the performance of these systems will improve. In the future, it may be possible to use these tools as the core of community and individual genomic data curation and integration efforts. These advances could have a very high impact on the ability to organize bioinformatics data in a cost-effective and scientifically-appropriate manner.

Bibliography

1. Chen, L., H. Liu, and C. Friedman, *Gene name ambiguity of eukaryotic nomenclatures.* Bioinformati cs, 2005. **21**(2): p. 248-56.
2. Yandell, M.D. and W.H. Majoros, *Genomics and natural language processing.* Nat Rev Genet, 2002. **3**(8): p. 601-10.
3. Rindflesch, T.C., et al., *EDGAR: extraction of drugs, genes and relations from the biomedical literature.* Pac Symp Biocomput, 2000: p. 517-28.
4. Ono, T., et al., *Automated extraction of information on protein-protein interactions from the biological literature.* Bioinformatics, 2001. **17**(2): p. 155-61.
5. Pustejovsky, J., et al., *Robust relational parsing over biomedical literature: extracting inhibit relations.* Pac Symp Biocomput, 2002: p. 362-73.
6. Park, J.C., H.S. Kim, and J.J. Kim, *Bidirectional incremental parsing for automatic pathway identification with combinatory categorial grammar.* Pac Symp Biocomput, 2001: p. 396-407.
7. Andrade, M.A. and A. Valencia, *Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts. Development of a prototype system.* Proc Int Conf Intell Syst Mol Biol, 1997. **5**: p. 25-32.
8. Blaschke, C., et al, *Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions.* Proc. of the AAAI Conf. on Intelligent Systems for Molecular Biology (ISMB), 1999: p. 60-67.
9. Koike, A., Y. Kobayashi, and T. Takagi, *Kinase pathway database: an integrated protein-kinase and NLP-based protein-interaction resource.* Genome Res, 2003. **13**(6A): p. 1231-43.
10. Wilbur, W.J. and Y. Yang, *An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts.* Comput Biol Med, 1996. **26**(3): p. 209-22.
11. Stephens, M., et al., *Detecting gene relations from Medline abstracts.* Pac Symp Biocomput, 2001: p. 483-95.
12. Iliopoulos, I., A.J. Enright, and C.A. Ouzounis, *Textquest: document clustering of Medline abstracts for concept discovery in molecular biology.* Pac Symp Biocomput, 2001: p. 384-95.
13. Raychaudhuri, S., H. Schutze, and R.B. Altman, *Using text analysis to identify functionally coherent gene groups.* Genome Res, 2002. **12**(10): p. 1582-90.
14. Daraselia, N., et al., *Extracting human protein interactions from MEDLINE using a full-sentence parser.* Bioinformatics, 2004. **20**(5): p. 604-11.
15. Marcotte, E.M., I. Xenarios, and D. Eisenberg, *Mining literature for protein-protein interactions.* Bioinformatics, 2001. **17**(4): p. 359-63.
16. Bader, G.D. and C.W. Hogue, *Analyzing yeast protein-protein interaction data obtained from different sources.* Nat Biotechnol, 2002. **20**(10): p. 991-7.
17. Stapley, B.J. and G. Benoit, *Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts.* Pac Symp Biocomput, 2000: p. 529-40.
18. Temkin, J.M. and M.R. Gilder, *Extraction of protein interaction information from unstructured text using a context-free grammar.* Bioinformatics, 2003. **19**(16): p. 2046-53.

19. Sleator, D., Temperly D., *Parsing English with a Link Grammar.* Carnegie Mellon University Computer Science Technical Report CMU-CS-91-916, 1991.
20. Abney, S., *Partial Parsing via Finite-State Cascades.* Journal of Natural Language Engineering, 1996. **2**(4): p. 337-344.
21. Nusse, R. *The Wnt gene Homepage*. [web] 2008 26 April 2004 [cited; Available from: http://www.stanford.edu/~rnusse/wntwindow.html.
22. Doms, A. and M. Schroeder, *GoPubMed: exploring PubMed with the Gene Ontology.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W783-6.
23. Eaton, A.D., *HubMed: a web-based biomedical literature search interface.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W745-7.
24. Jenssen, T.K., et al., *A literature network of human genes for high-throughput analysis of gene expression.* Nat Genet, 2001. **28**(1): p. 21-8.
25. Korbel, J.O., et al., *Systematic association of genes to phenotypes by genome and literature mining.* PLoS Biol, 2005. **3**(5): p. e134.
26. Lloyd, L., D. Kechagias, and S. Skiena. *Lydia: A System for Large-Scale News Analysis*. in *String Processing and Information Retrieval: 12th International Conference,*
SPIRE 2005,. Buenos Aires, Argentina,: Springer Verlag Lecture Notes in Computer Science.
27. ThompsonResearchSoft, *RefViz*. 2007, Thompson ResearchSoft.
28. Chaussabel, D. and A. Sher, *Mining microarray expression data by literature profiling.* Genome Biol, 2002. **3**(10): p. RESEARCH0055.
29. Tanabe, L., et al., *MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling.* Biotechniques, 1999. **27**(6): p. 1210-4, 1216-7.
30. Becker, K.G., et al., *PubMatrix: a tool for multiplex literature mining.* BMC Bioinformatics, 2003. **4**: p. 61.
31. Wren, J.D., et al., *Knowledge discovery by automated identification and ranking of implicit relationships.* Bioinformatics, 2004. **20**(3): p. 389-98.
32. Masys, D.R., et al., *Use of keyword hierarchies to interpret gene expression patterns.* Bioinformatics, 2001. **17**(4): p. 319-26.
33. Alako, B.T., et al., *CoPub Mapper: mining MEDLINE based on search term co-publication.* BMC Bioinformatics, 2005. **6**: p. 51.
34. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
35. Dennis, G., Jr., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery.* Genome Biol, 2003. **4**(5): p. P3.
36. Huang da, W., et al., *The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.* Genome Biol, 2007. **8**(9): p. R183.
37. Huang da, W., et al., *DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W169-75.
38. Sherman, B.T., et al., *DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis.* BMC Bioinformatics, 2007. **8**: p. 426.

39. Al-Shahrour, F., R. Diaz-Uriarte, and J. Dopazo, *FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.* Bioinformatics, 2004. **20**(4): p. 578-80.
40. Al-Shahrour, F., et al., *BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W460-4.
41. Beissbarth, T. and T.P. Speed, *GOstat: find statistically overrepresented Gene Ontologies within a group of genes.* Bioinformatics, 2004. **20**(9): p. 1464-5.
42. Feng, W., et al., *Development of gene ontology tool for biological interpretation of genomic and proteomic data.* AMIA Annu Symp Proc, 2003: p. 839.
43. Castillo-Davis, C.I. and D.L. Hartl, *GeneMerge--post-genomic analysis, data mining, and hypothesis testing.* Bioinformatics, 2003. **19**(7): p. 891-2.
44. Beisvag, V., et al., *GeneTools--application for functional annotation and statistical hypothesis testing.* BMC Bioinformatics, 2006. **7**: p. 470.
45. Berriz, G.F., et al., *Characterizing gene sets with FuncAssociate.* Bioinformatics, 2003. **19**(18): p. 2502-4.
46. Zhang, B., et al., *GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.* BMC Bioinformatics, 2004. **5**: p. 16.
47. Zhong, S., et al., *GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space.* Appl Bioinformatics, 2004. **3**(4): p. 261-4.
48. Zhong, S. and D. Xie, *Gene Ontology analysis in multiple gene clusters under multiple hypothesis testing framework.* Artif Intell Med, 2007. **41**(2): p. 105-15.
49. Shah, N.H. and N.V. Fedoroff, *CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology.* Bioinformatics, 2004. **20**(7): p. 1196-7.
50. Martin, D., et al., *GOToolBox: functional analysis of gene datasets based on Gene Ontology.* Genome Biol, 2004. **5**(12): p. R101.
51. Al-Shahrour, F., et al., *FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W91-6.
52. Khatri, P. and S. Draghici, *Ontological analysis of gene expression data: current tools, limitations, and open problems.* Bioinformatics, 2005. **21**(18): p. 3587-95.
53. Tomlins, S.A., et al., *Integrative molecular concept modeling of prostate cancer progression.* Nat Genet, 2007. **39**(1): p. 41-51.
54. Rhodes, D.R., et al., *Molecular concepts analysis links tumors, pathways, mechanisms, and drugs.* Neoplasia, 2007. **9**(5): p. 443-54.
55. Rhodes, D.R., et al., *Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles.* Neoplasia, 2007. **9**(2): p. 166-80.
56. Camon, E.B., et al., *An evaluation of GO annotation retrieval for BioCreAtIvE and GOA.* BMC Bioinformatics, 2005. **6 Suppl 1**: p. S17.
57. Hosack, D.A., et al., *Identifying biological themes within lists of genes with EASE.* Genome Biol, 2003. **4**(10): p. R70.
58. Nusse, R. *The Wnt gene Homepage*. [web] 2004 26 April 2004 [cited; Available from: http://www.stanford.edu/~rnusse/wntwindow.html.
59. Information, N.C.f.B., *NCBI - Entrez Programming Utilities*. 2004, NCBI.

60. Abney, S., *Statistical Methods and Linguistics*. The Balancing Act: Combining Symbolic and Statistical Approaches to Language, ed. J.K.a.P. Resnik. 1996, Cambridge, MA.: The MIT Press.
61. Kishida, M., et al., *Synergistic activation of the Wnt signaling pathway by Dvl and casein kinase Iepsilon.* J Biol Chem, 2001. **276**(35): p. 33147-55.
62. Song, D.H., et al., *CK2 phosphorylation of the armadillo repeat region of beta-catenin potentiates Wnt signaling.* J Biol Chem, 2003. **278**(26): p. 24018-25.
63. Rosner, A., et al., *Pathway pathology: histological differences between ErbB/Ras and Wnt pathway transgenic mammary tumors.* Am J Pathol, 2002. **161**(3): p. 1087-97.
64. Lagutin, O.V., et al., *Six3 repression of Wnt signaling in the anterior neuroectoderm is essential for vertebrate forebrain development.* Genes Dev, 2003. **17**(3): p. 368-79.
65. 2003. **130**(23).
66. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
67. Settles, B., *Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets.* Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA). 2004.
68. Chang, J., H. Schutze, and R.B. Altman. *GAPSCORE: Finding Gene and Protein Names One Word at a Time*. 2005 [cited 2005; Available from: http://bionlp.stanford.edu/gapscore/.
69. Alias-I, I., *Alias-i LingPipe.* 2005.
70. Novichkova, S., S. Egorov, and N. Daraselia, *MedScan, a natural language processing engine for MEDLINE abstracts.* Bioinformatics, 2003. **19**(13): p. 1699-706.
71. Santos, C., D. Eggle, and D.J. States, *Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction.* Bioinformatics, 2005. **21**(8): p. 1653-8.
72. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
73. NCBI. *NCBI Mouse Build 33*. 2005 [cited 2005 2005]; Available from: http://www.ncbi.nlm.nih.gov/genome/seq/NCBIContigInfo.html.
74. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.
75. Bates, D., R. chambers, and P. Dalgaard. *The R Project for Statistical Computing (Web Site) http://www.r-project.org*. 1996 [cited; Available from: http://www.r-project.org.
76. Mitchell, J.A., et al., *Gene indexing: characterization and analysis of NLM's GeneRIFs.* AMIA Annu Symp Proc, 2003: p. 460-4.
77. Narayanaswamy, M., K.E. Ravikumar, and K. Vijay-Shanker, *A biological named entity recognizer.* Pac Symp Biocomput, 2003: p. 427-38.
78. Singh, A.P., et al., *Aberrant expression of transmembrane mucins, MUC1 and MUC4, in human prostate carcinomas.* Prostate, 2006. **66**(4): p. 421-9.
79. Rennie, P.S. and C.C. Nelson, *Epigenetic mechanisms for progression of prostate cancer.* Cancer Metastasis Rev, 1998. **17**(4): p. 401-9.

80. Vire, E., et al., *The Polycomb group protein EZH2 directly controls DNA methylation.* Nature, 2006. **439**(7078): p. 871-4.
81. Varambally, S., et al., *The polycomb group protein EZH2 is involved in progression of prostate cancer.* Nature, 2002. **419**(6907): p. 624-9.
82. Cao, R. and Y. Zhang, *SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex.* Mol Cell, 2004. **15**(1): p. 57-67.
83. Chen, H., S.W. Tu, and J.T. Hsieh, *Down-regulation of human DAB2IP gene expression mediated by polycomb Ezh2 complex and histone deacetylase in prostate cancer.* J Biol Chem, 2005. **280**(23): p. 22437-44.
84. Iljin, K., et al., *TMPRSS2 fusions with oncogenic ETS factors in prostate cancer involve unbalanced genomic rearrangements and are associated with HDAC1 and epigenetic reprogramming.* Cancer Res, 2006. **66**(21): p. 10242-6.
85. Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**(5748): p. 644-8.
86. Liu, W., et al., *Comprehensive assessment of DNA copy number alterations in human prostate cancers using Affymetrix 100K SNP mapping array.* Genes Chromosomes Cancer, 2006. **45**(11): p. 1018-32.
87. Cerveira, N., et al., *TMPRSS2-ERG gene fusion causing ERG overexpression precedes chromosome copy number changes in prostate carcinomas and paired HGPIN lesions.* Neoplasia, 2006. **8**(10): p. 826-32.
88. Nelson, W.G., et al., *Abnormal DNA methylation, epigenetics, and prostate cancer.* Front Biosci, 2007. **12**: p. 4254-66.
89. Li, L.C., *Epigenetics of prostate cancer.* Front Biosci, 2007. **12**: p. 3377-97.
90. Nusse, R., *Wnt signaling in disease and in development.* Cell Res, 2005. **15**(1): p. 28-32.
91. Brentani, H., et al., *The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags.* Proc Natl Acad Sci U S A, 2003. **100**(23): p. 13418-23.
92. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
93. Napione, L., et al., *Integrins: a flexible platform for endothelial vascular tyrosine kinase receptors.* Autoimmun Rev, 2007. **7**(1): p. 18-22.
94. Mundy, G.R., *Metastasis to bone: causes, consequences and therapeutic opportunities.* Nat Rev Cancer, 2002. **2**(8): p. 584-93.
95. Eccles, S.A., *Cell biology of lymphatic metastasis. The potential role of c-erbB oncogene signalling.* Recent Results Cancer Res, 2000. **157**: p. 41-54.
96. Toker, A. and M. Yoeli-Lerner, *Akt signaling and cancer: surviving but not moving on.* Cancer Res, 2006. **66**(8): p. 3963-6.
97. Hermann, P.C., S.L. Huber, and C. Heeschen, *Metastatic cancer stem cells: A new target for anti-cancer therapy?* Cell Cycle, 2007. **7**(2).