A HYBRID APPROACH FOR TRANSLATIONAL RESEARCH

Yue Wang Webster

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics,
Indiana University

April 2010

Accepted by the Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

____

Josette F. Jones, Ph.D., Chair

____

Mathew J. Palakal, Ph.D.

Doctoral Committee

____

Gunther Schadow, M.D., Ph.D.

March 2, 2010

____

Ernst R. Dow, Ph.D.

ACKNOWLEDGEMENTS

A sincere gratitude for all who have supported me through these years of study, especially:

Dr. Josette F. Jones

Dr. Mathew J. Palakal

Dr. Ernst R. Dow

Dr. Gunther Schadow

Dr. Anna McDaniel

Dr. Jacob Koehler

Dr. Ranga C. Gudivada

Dr. Jian Wang

Dr. John Calley

Dr. Lee O’Brien

Mr. Gabor Varga

ABSTRACT

Yue Wang Webster

A HYBRID APPROACH FOR TRANSLATIONAL RESEARCH

Translational research has proven to be a powerful process that bridges the gap between basic science and medical practice. The complexity of translational research is two-fold: integration of vast amount of information in disparate silos, and dissemination of discoveries to stakeholders with different interests. We designed and implemented a hybrid knowledge discovery framework. We developed strategies to leverage both traditional biomedical databases and Health Social Network Communities content in the discovery process. Heuristic and quantitative evaluations were carried out in Colorectal Cancer and Amyotrophic Lateral Sclerosis disease areas. The results demonstrate the potential of our approach to bridge silos and to identify hidden links among clinical observations, drugs, genes and diseases, which may eventually lead to the discovery of novel disease targets, biomarkers and therapies.

Josette F. Jones, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

Opportunities of Translational Research

As basic science and medical practice continue their exponential growth in complexity and scope, the need of bridging research with practice becomes more urgent. Molecular or cellular level discoveries made at lab “bench” need to progress to the patient’s “bedside” as therapies [1]. On the other hand, knowledge gained at “bedside” is important for the researchers at the “bench” to better understand the molecular mechanisms of diseases [2]. At present, discoveries in one discipline are not efficiently transformed into executable knowledge that can be used in the other [3]. Translational research is a branch of research that attempts to develop insight into such cross-disciplinary knowledge transformation and collaboration [4]. In the following sections, we show examples of the important roles that translational research plays in life science and health care practice.

Opportunities in Drug Reposition

Identifying new indications for existing drugs is an important strategy of drug discovery. For example, Atorvastatin is an FDA approved drug used to lower cholesterol [5]. A drug hunter may want to find other possible diseases that might be treated by Atorvastatin. Using a non-translational approach, the drug hunter would query for pathways related to Atorvastatin in KEGG and would retrieve no result from KEGG.

Using a translational approach, the drug hunter would first search for the indications of Atorvastatin. The drug hunter would then search for genes that are associated with those indications. Next the drug hunter would look for pathways associated with those genes in KEGG. The drug hunter would find one of the pathways to be the Alzheimer’s Disease (AD) pathway and may consider AD as a possible new indication for Atorvastatin. This hypothesis is supported by other studies that show clinical benefit of Atorvastatin in AD patients [6, 7]. The drug hunter can rapidly identify such opportunities because clinical and genomic disciplines are effectively connected in a translational approach.

Opportunities in Target Identification

Identifying the disease targets is often the first step towards discovering the cure. For example, to develop AD treatment, a scientist may first search for all drug-able genes associated with AD. A non-translational approach is to search against various biological databases for drug-able genes associated with AD. Since the disease has few known pathway steps, this approach yields limited success and no drug-able GPCR target can be identified.

Qu et al. described a translational approach in [8]. They first retrieved all genes participating in the AD pathway. Next they searched for all pathways associated with those genes. Then they repeated the first two steps by searching for genes participated in all pathways found in the previous iteration. Using this approach, the authors reported that they could identify more novel targets. For instance, they found eight GPCR targets that are implicitly associated with AD at the second iteration. Those associations are supported by evidence found in multiple studies [9, 10]. One of the genes, SLIT2, is found to be involved in neuron recombination and axonogenesis [11-13]. AD-related deficits have been observed for the functionally related RNA messages encoding the SLIT2 axon guidance receptor factor and the neuroglycan C precursor [14].

Opportunities in Cross-disease Research

Today's scientists and investigators are faced with a deluge of data from various disciplines. Translational research provides a systematic way to identify implicit associations and insights "hidden" in large and heterogeneous datasets. Ruttenberg proposed in [4] that knowledge should be shared among specialists in different disease areas. One success story of this approach is the discovery of NST-729 as a cross-disease biomarker for neurodegeneration and as the first molecular probe for Amyotrophic Lateral Sclerosis (ALS) [15]. Based on the knowledge that Parkinson's Disease (PD), AD, Huntington's Disease (HD), and ALS share common features at the clinical [16], neural [17, 18], cellular [19, 20], and molecular levels [21], Shirvan et al. studied and compared the performance of NST-729 cross the transgenic models of two neurodegenerative disorders, AD and ALS. Without translational knowledge and tools such comparison will not be possible.

Besides the use cases we described above. There are other opportunities for translational research, such as clinical trial management, longitudinal patient health record, epidemiology studies, and public health surveillance [22].

Challenges in Translational Research

As show in Figure 1, the knowledge of life science and medical practice evolves in a spiral fashion: basic science discoveries lead to clinical studies, and then to medical practice tools, which influence the health care policy makers and the public. Knowledge gained from clinical studies, medical practice and public health can in turn inform chemists and biologists in the laboratory. However, the knowledge transformation is not always effective. New discoveries are often stored in their discipline sources that form information-silos; and experts tend to interact within their own circles which are social-silos. One major task of translational research is to overcome these two types of silos.

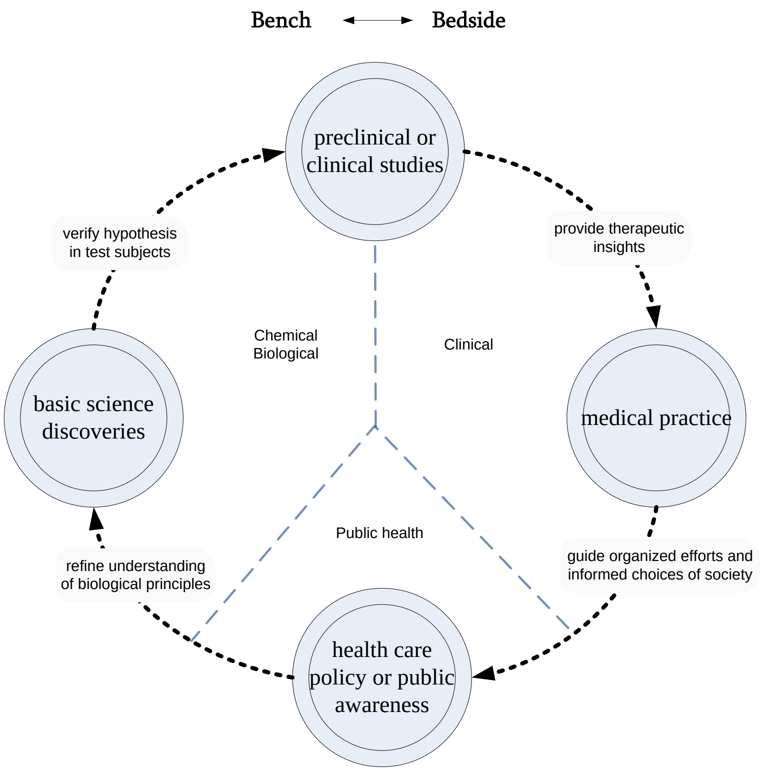

Bench
Bedside
preclinical or clinical studies
verify hypothesis in test subjects
provide therapeutic insights
Chemical Biological
Clinical
basic science discoveries
medical practice
Public health
refine understanding of biological principles
guide organized efforts and informed choices of society
health care policy or public awareness

Figure 1. Knowledge evolves in a spiral fashion across bench and bedside

Heterogeneous Data Sources

A wide variety of data types and artifacts from different discipline sources are involved in translational process. We grouped them into the following major categories: chemical, biological, clinical, and public health data.

Chemical data is relative straightforward comparing with biological and clinical data. The main subject of chemical data is molecules and their interactions both with each other and with the environment. Such information (physical properties, chemical properties, and reactions) can often be captured in basic data types (numbers and stings). 2D or 3D structure information can be expressed in special text format. Graphical representation is used mostly in visualizing structures, studying molecular dynamics, and protein-ligand docking.

Biological data are more heterogeneous because it encompasses many domains of knowledge (molecular and cell biology, genetics, structural biology, pharmacology, physiology, etc.). According to Topaloglon [23], most of the information that biologists are interested in is available in public reference databases, specialized private data sources, and scientific literature. It is estimated that 80% of the biological data are in text format, and the rest resides in databases that range from indexed files, to formal databases.

Clinical data is concerned with or based on the actual observation and treatment of diseases in patients rather than experiment or theory [24]. It is created, rendered and consumed during the health care process. Different from chemical and biological data, temporal information is an intrinsic component of clinical data. It is often probabilistic or fuzzy in nature. For instance diagnoses are always specified with a degree of certainty. Because of its complexity and diversity, data standards play a key role in handling clinical data in translational research. There are four major types of clinical data standards [25]. Terminology standards define accepted vocabulary and how they should be used. In ICD-9 coding, for example, “chest pain” is a valid term and has a specific code associated with it, whereas “pain in the chest” is not. Conceptual standards define how certain concepts are conveyed. Document standards define information required in a certain document and the location of the information. Messaging standards define how information is packaged and communicated between parties.

Most clinical data are concerned with a single patient (individual level data), such as laboratory test results, patient demographics, discharge summaries, and progress notes. On the other hand, public health data is often concerned with a group of people (high-level aggregated data) who have common characteristics, such as data used to study disease outbreaks and epidemics. Since the focus of public health is to promote healthy behaviors and to prevent diseases through surveillance of cases. Social context is an important part of public health data.

Multiple Data Levels

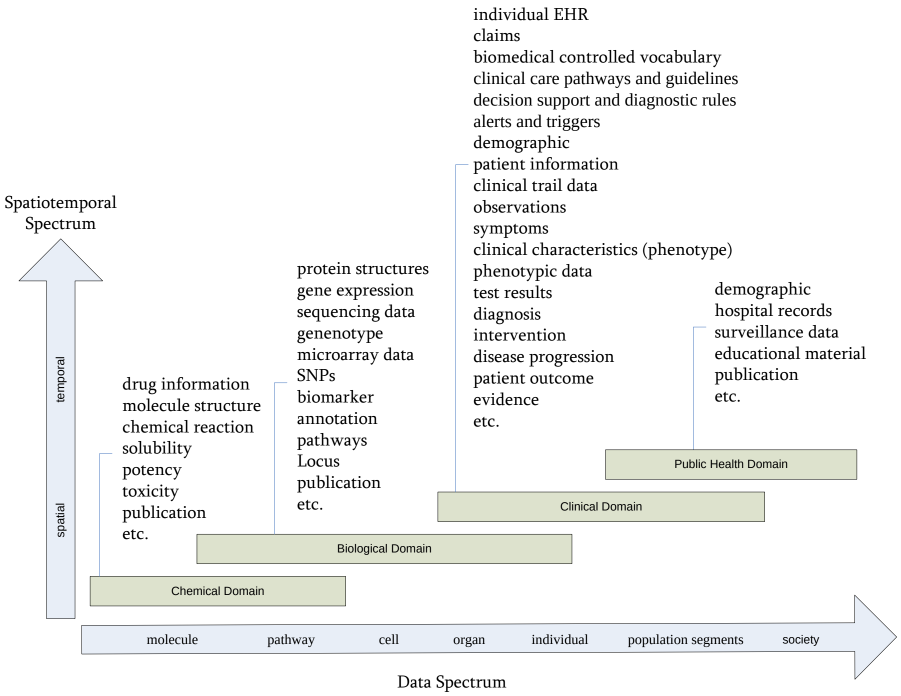


Figure 2. Multi-level data involved in translational research

Translational research involves capturing and mining a wide spectrum of data types across multiple levels: molecule →pathway →cell → organ → individual → population segments, subgroups → society. As shown in Figure 2 the data ranges from the descriptions of molecular events, to the descriptions of complex biological systems, and to the nature-language descriptions of an individual or a group. In addition, both spatial and temporal data types are important for translational research. Spatial data is necessary for describing compound or protein structures because the same set of atoms could have multiple orientations in 3D space. Temporal data types are important for clinical and public health discipline, where it is essential to track the condition of a patient or the spread of a disease.

The challenge is in connecting data from different levels. For example, how to link molecular data such as gene sequence with data specific to an individual such as a patient's Electronic Health Record, and with data collected at population level such as genome-wide association study of patients from a certain ethnical group.

Unique Opportunities and Challenges Presented by HSNC

Web2.0 refers to web applications that facilitate interactive information sharing, interoperability, user-centered design, and collaboration on the World Wide Web, such as *YouTube*, *Facebook*, and *Twitter*. Health social network communities (HSNC) are online communities where users search, self-track, share and discuss health-related information using Web2.0 technologies. Examples of popular HSNC include PatientsLikeMe.com (PLM), DailyStrength.org and MedHelp.org. Their primary users are patients with similar medical conditions. With the aid of HSNC, the role of those patients is changing. Instead of being passive test subjects, they are becoming active participants, information owners, or peer leaders.

For instance, when a fifteen-month clinical trial with 44 patients reported that lithium, a drug used to treat bipolar disorder, may delay progression of amyotrophic lateral sclerosis (ALS) [26]. Within a few months, an ALS patient gathered 250 ALS patients inside PLM to self-experiment with lithium tracking their conditions using social networking tools. This patient-driven trial included more test cases than any published study of lithium to date [27]. The conclusion of the 250-sample trial was different from the previous 44-sample trial. The preliminary analysis of PLM members’ data showed no correlation between lithium and reduced ALS disease progression. This example highlighted the potential of HSNC research model, especially in fighting orphan diseases that do not fit into the current business model of pharmaceutical industry.

However, there are many challenges in harvesting the consumer-generated information and translating it from “bedside” to “bench”. One obvious issue is the consumer-professional vocabulary gap. Vocabularies used in patient-oriented online communities are consumer English. On the other hand, most biomedical databases are developed for professionals and use discipline-specific vocabularies. For example, PLM allows patients to describe their conditions using folksonomy, a user-generated taxonomy. Smith and Wicks pointed out that less than half of the symptom terms contributed by PLM patients can be mapped to UMLS concepts or synonyms [28].

Another challenge comes from the fact that the information organization of HSNC and biomedical databases are established differently. Information in HSNC is organized and stratified by consumers through collaborative filtering, tagging, voting and other Web 2.0 techniques, which is a “bottom-up” approach. On the other hand, professionals often define the data schema of research-oriented databases before loading the data, which is a “top-down” approach. Therefore, it is not surprising that 62% of the “symptoms” used by PLM patients were not categorized as “Signs or Symptoms” in UMLS.

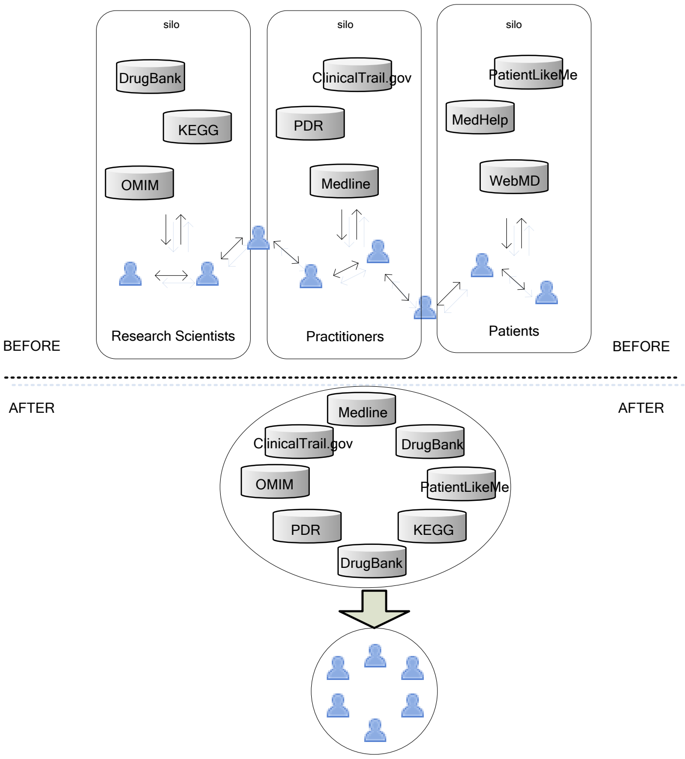


Figure 3. Bridging the information-silos and the social-silos

In summary, to address the challenges in translational research as represented by Figure 3, a successful knowledge discovery system has to accomplish at least two tasks: 1) to bridge the silos and discover novel associations; and 2) to deliver the results to the users based on their specific needs. Here we propose a hybrid approach that combines several technologies to achieve these two aims. The background chapter briefly reviews the technologies that form the basis of this approach, as well as the two diseases relevant to the case studies. The methodology chapter discusses the uniqueness of the design and describes the implementation details. The case study results and observations are reported in the results chapter. Key strategies and findings are highlighted in the conclusion chapter. Limitations of the approach and future research directions are disused in the discussion chapter.

CHAPTER TWO: BACKGROUND

Semantic Web

One of the fast-growing research areas of informatics, Semantic Web (SW), shows great promise in navigating and drawing sophisticated inference from diverse digital resources [4, 29, 30]. SW has four building blocks: a mechanism to uniquely identify the web resources; a framework to describe web resources; a query language; and an ontology language. Each building block can be implemented differently. Here, we only discuss the standards recommended by the World Wide Web Consortium (www.w3.org).

Uniform Resource Identifier (URI) is “a compact sequence of characters that identifies an abstract or physical resource”. It distinguishes one resource from all others. It is the foundation of SW.

Resource Description Framework (RDF) is a format that describes web resources by triples (subject-predicate-object). The subject and object are the resources, and the predicate is the relationship between the subject and the object. A subject or object can be an organic compound, a gene, a pathway or a patient. Predicate (a.k.a. property) can be any relationship between the subject and the object (i.e. causes, regulates, transcribes, etc.) For instance, “drug A causes disorder B” can be represented as <A treats B>, where A is the subject, B is the object, and “treat” is the property.

Simple Protocol and RDF Query Language (SPARQL) is a SQL-like query language for query SW. It essentially consists of a standard query language, a data access protocol and a data model. Using SPARQL, users can form semantic queries that otherwise require lengthy and complex SQL statements in a relational database.

RDF Schema (RDFS) and *Web Ontology Language* (OWL) have both been used in SW applications to encode ontologies. OWL is more expressive than RDFS. The data described by OWL ontology is interpreted as a set of “individuals” or “classes” and a set of “property assertions” which relate these individuals to each other. The axioms in OWL ontology define constraints on the individuals and the types of relationships permitted between them. A SW system, therefore, can infer additional information based on the axioms.

In this work, SW technologies are used to integrate information-silos. We derived all URIs from NIH authoritative identifiers, such as EntrezGene ID or UMLS concept ID. Therefore entities sharing the same URI are merged into one concept regardless of their sources. Two concepts from isolated date sources are connected if they are both associated with the same concept, which is the key for semantically bridging the information-silos.

Graph Analysis

The history of graph analysis in mathematics can be tracked as far back as the eighteenth century [31]. Some common terminologies are defined in . A *graph* G = (V,E) is a collection of nodes (V) and edges (E) connecting those nodes. An *edge* e = (u, v) ∈ E connects two *nodes* u and v. The nodes u and v are said to be *incident* with the edge e and *adjacent* to each other. The *degree* d(v) of a node (a.k.a. vertex) v is the number of its incident edges. Let (e_1, … , e_k) be a sequence of edges in a graph G = (V,E). This sequence is called a *path* if there are nodes v_0, … , v_k such that $e_i = (v_{i-1}, v_i)$ for i = 1, … , k and the edges e_i are pair-wise distinct and the nodes v_i are pair-wise distinct. The *length* of a path is given by its number of edges, $k = |(e_1, \dots, e_k)|$. A *shortest path* between two nodes u, v is a path with minimal length. The *diameter* is the maximum *shortest path* length amongst all pairs of nodes in a graph.

Average degree, diameter, and average shortest path are topology measurements often used in graph analysis. Average degree reflects the “connectivity” of a graph. Diameter and average shortest path reflect the “compactness” of a graph. A small diameter or a low average shortest path length indicates that all the nodes are in proximity to each other. A highly connected, compact graph often represents a dense knowledge space where concepts are closely related to each other. In such a graph, the changes of one node have greater impact on other nodes than it would be in a loosely connected graph.

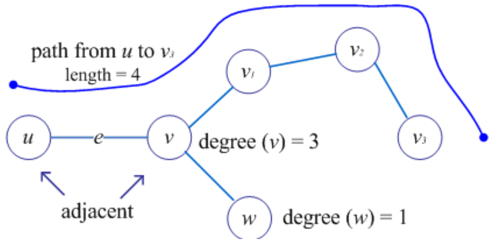


Figure 4. Common terminologies used in graph analysis

Graph analysis has drawn much interest among bioinformatics researchers due to the rapid growth of publicly available high throughput data [32-39]. Such data have provided linkages among chemical, biological, and clinical entities. Chen and colleagues surveyed multiple applications that encoded knowledge using graph-based data structures. In those applications, biomedical entities are modeled as nodes and the relationships as edges (links). The graphs can then be analyzed using conventional graph analysis technique or extension of it [38-41].

In this case, we are especially interested in applying graph algorithms to rank search results. Common approaches [42-45] include concept structure analysis, PageRank, and Hyperlink-Induced Topic Search (HITS). PageRank with Priors proposed by White and Smyth [46] simulates the steps of a Web surfer, who starts from any of the root nodes on the Internet and follows a random link at each step with β as the probability of returning to the root nodes. A score is computed for each node on the Internet to reflect its probability of being reached by the surfer. This score is used to measure the relative “closeness” of a node to the root nodes. K-Step Markov method simulates a similar Web surfing scenario as in PageRank, except that the surfer returns to the root nodes after K steps and restarts the process. K-Step Markov algorithm estimates the relative probability that a surfer will spend time at a node given that the surfer starts in a set of root nodes and stops after K steps. HITS with Priors proposed by Kleinberg measures two properties of a node: 1) authority score estimates the importance of the node itself; and 2) hub score measures the importance of other nodes linked to the current node [47]. Therefore, HITS with Priors not only considers the number of links to and from a node but also its neighbors’.

Gudivada et al. have proposed a modified algorithm to rank genes [48]. In traditional WWW ranking analyses, all links are considered equally significant. But in the context of biological networks, the importance of a link also depends on the nodes connected with it. Using gene and pathway association as an example, Gudivada explained that a gene participates in multiple pathways is more important than a pathway that has multiple genes since most pathways will include multiple genes. To model this nature of biological networks, each link is assigned a subjectivity weight and an objectivity weight. Link such as ‘*Gene-HasAssociated-Pathway*’ is assigned a higher subjectivity weight (for gene) and lower objectivity weight (for pathway). The only constraint is that for each link the sum of subjectivity and objectivity weights must be equal to 1.

User Profiling

Translational research has many stakeholders with different perspectives; therefore, we use profiling technologies to capture users' interests and to control how the results are presented. Information that seems to be trivial or irrelevant to one user may be important to the other. For example, a drug hunter is interested in discovering novel molecules that interact with a certain enzyme. Concepts such as active sites of the enzyme, electronic density map of the enzyme, etc. would be most relevant, whereas pathway regulated by the enzyme would not. On the other hand, a biologist who is interested in the mechanism of a disorder may be interested in the pathway regulated by the enzyme.

There are three stages of user profiling at the information level: information representation, information classification, and user profile learning. The primary challenge comes from dealing with heterogeneous data encountered in the three stages. Most user profiling systems deal with domain knowledge represented by a thesaurus or a linear list of terms or concepts which are assumed to be independent of each other [49]. To represent a nonlinear structure with inter-related concepts, an ontology can be used to form the basis for user personalized searching and browsing [50].

CHAPTER THREE: METHODOLOGY

Design

Based on the background research, we propose a hybrid approach, combining SW, graph algorithm and user profiling. In terms of implementation, this approach has two main steps: 1) to construct a full semantic graph using associations extracted from multiple discipline sources; 2) to subtract a sub graph for each user based on the profile using pseudo- relevance feedback strategy. We hypothesize that this approach is advantageous in discovering hidden associations across disciplines and tailoring the results for individual user, which are essential to translational research. The reasons are as following. An association can be modeled as two nodes linked by an edge in a graph. Therefore, we may model discipline silos as a large, connected graph. Conventional graph algorithms can be extended to mine the graph based on user profiles, which allows us to deliver the mining results in a user-centric manner to the respective stakeholder. It is expensive to conduct graph analyses on a large-scale graph, but we can build virtual sub graphs to reduce the cost of the analyses. The sub graphs also capture the personalized views of the full knowledge space. Such views are specific to each user and are part of the strategy to tailor search results for different stakeholders of translational research.

To demonstrate this approach and to evaluate its feasibility, we designed and implemented a knowledge discovery framework called HyGen. Figure 5 shows the major components (layers) of HyGen. Layer A consists of data mining protocols that extract associations from various silos. The associations form the full graph in Layer B. Layer C manages the user profiles; D creates sub graphs; and E produces the personalized views. In the following sections, we describe the design and implementation details of each layer.

Implementation

Layer A: Extract Associations

Disease titles, clinical synopses and text fields are downloaded from OMIM (ftp.ncbi.nih.gov/repository/OMIM). To insure that a clinical term and its synonyms are merged to the same node in the semantic graph, we normalize the terms against UMLS. We use MetaMap (http://mmtx.nlm.nih.gov) to extract clinical-relevant terms from short phrases in titles and clinical synopses sections and map them to UMLS concepts. For the long free-text fields of OMIM records, we implemented text-mining protocols using Pipeline Pilot's Text Analytics Collection (accelrys.com) to extract clinical-relevant terms, and to map them to UMLS concepts. Our experiments showed that the protocols have a higher recall rate (75% to 80%) than MetaMap (less than 25%) when mining OMIM text fields. Each gene in OMIM Morbid Map (downloaded from ftp.ncbi.nih.gov/repository) is linked with the UMLS concepts extracted from its corresponding OMIM record, including the title, the clinical synopsis, and the text field. We use similar approach to normalize the drug indications of DrugBank records, as well as disease terms of PharmGKB and GAD.

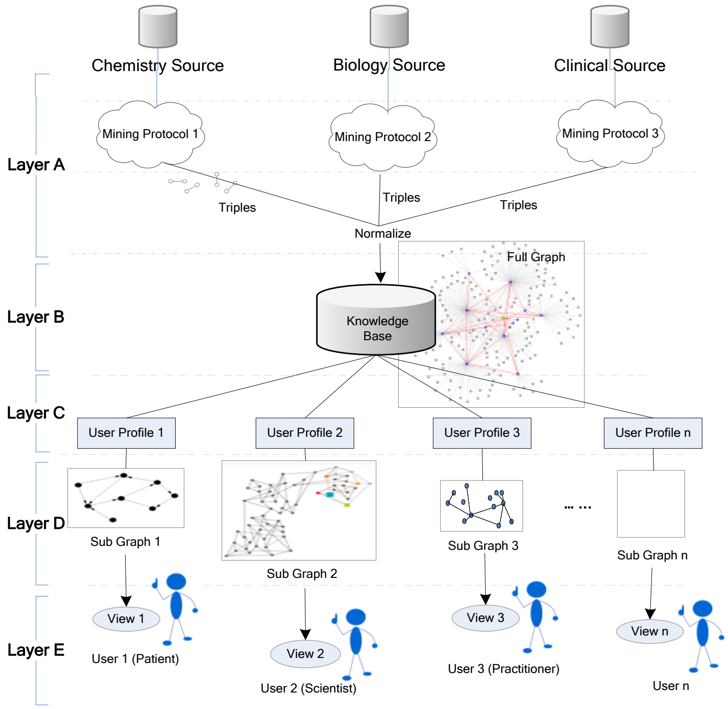
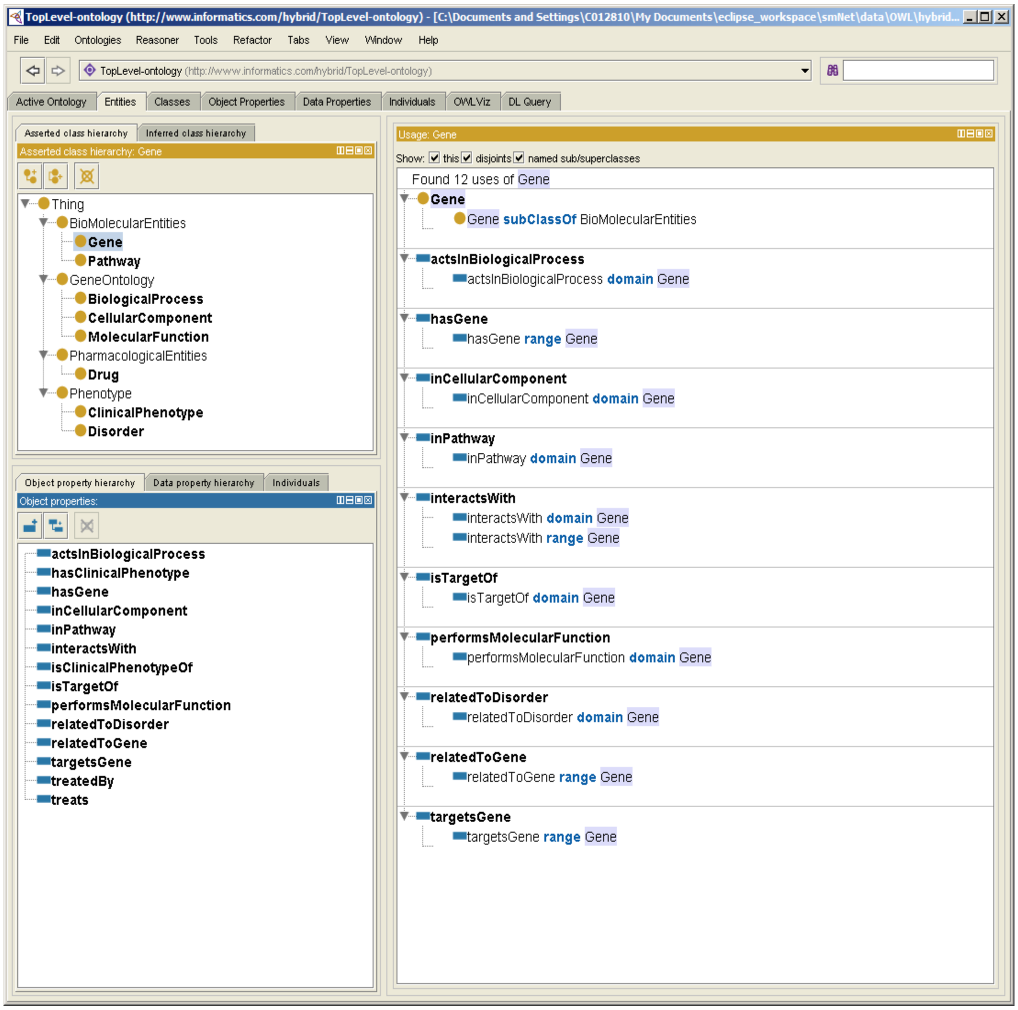


Figure 5. Major components of HyGen

The chemical compounds in OMIM record, DrugBank, PharmGKB, and KEGG are normalized against CHEMLIST, a dictionary for identifying chemical information in the literature [51]. From the normalized compound list, marketed drugs or drug candidates in clinical trials phase-II or later are extracted. Each drug or drug candidate is then connected with its target genes based on PharmGKB and DrugBank records. Other types of associations from genomic, pharmacological, and proteomic sources are also incorporated in the full graph. Appendix A lists all associations in the current version of HyGen.

Layer A was coded in Java. It interacts with MetaMap and Pipeline Pilot's Text Analytics Collection to recognize and map source-dependent terms to standard dictionaries. It uses Jena API (www.jena.sourceforge.net) to create and manipulate the associations as RDF triples. The associations harvested by A were loaded into AllegroGraph RDFStore (http://www.franz.com/agraph/allegrograph) based on a custom-built top-level ontology developed in Protégé (shown in Figure 6).



Show: ☑ this ☑ disjoints ☑ named sub/superclasses

Found 12 uses of Gene

Gene
- Gene subClassOf BioMolecularEntities

actsInBiologicalProcess
- actsInBiologicalProcess domain Gene

hasGene
- hasGene range Gene

inCellularComponent
- inCellularComponent domain Gene

inPathway
- inPathway domain Gene

interactsWith
- interactsWith domain Gene
- interactsWith range Gene

isTargetOf
- isTargetOf domain Gene

performsMolecularFunction
- performsMolecularFunction domain Gene

relatedToDisorder
- relatedToDisorder domain Gene

relatedToGene
- relatedToGene range Gene

targetsGene
- targetsGene range Gene

Figure 6. HyGen's top-level ontology

Layer B: Construct Full Graph

In Layer B, the associations are converted to nodes and edges of the full graph, where nodes represent biomedical entities, such as genes, diseases, or compounds; and edges represent the relationships between entities. The final full graph is an un-weighted directed acyclic graph.

HyGen's full graph is different from most existing life science networks (graphs) in that it allows any types of nodes. For example Chen's protein-protein network analysis [43] was conducted on proteins linked by their interactions. Yildirim's drug-target network [35] has two types of node: drugs and proteins linked by drug-target binary associations. Goh's human disease network [52] has two types of nodes: disorders and disease genes linked by the fact that mutations in a gene lead to a specific disorder. Campillos' drug-drug network has one type of node: drugs linked by their side-effect similarity [53]. Li's human disease-disease network [54] has one type of node: diseases linked by their shared pathways. Different from them, Layer B can integrate many types of nodes and enable HyGen to discover associations among different types of biomedical entities.

It is worth mentioning that even though SW technologies are used in this implementation, the full graph can be constructed by other tools as long as they allow HyGen to merge and connect entities from diverse sources.

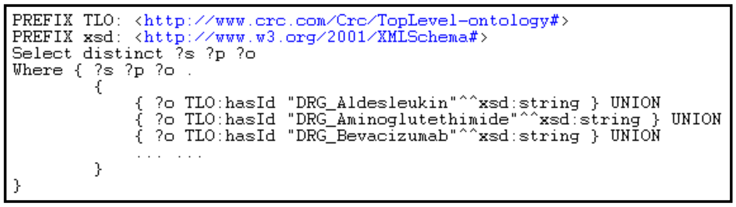
Layer C: Define User Profile

A user profile (P_i) defines the "seeds", which are special nodes relevant to a certain user. They can be any type, for example "fatigue" (a symptom), "SOD1" (a gene), or "riluzole" (a drug). Those nodes are the starting points of the knowledge discovery, hence the name "seeds". In addition, the user can specify in P_i whether the user prefers 1) long associations; 2) rare data types; 3) highly connected concepts (hotspots); 4) new information; and 5) how to measure the relative importance of a biomedical entity.

Based on literature research [55] and our experience, we constructed seventeen user profile templates for different user groups involved in translational research (see Appendix B). Preference and default parameters were empirically defined in the templates, based on which users can build their own profiles. Appendix C shows a sample profile stored in HyGen. One user can have multiple profiles. Through Layer C, the profiles can be updated, deleted, published and copied.

Layer D: Build Sub Graph

Based on the full graph, different virtual sub graphs are constructed for different user profiles, following an iterative process inspired by the pseudo-relevance feedback used in some document retrieval systems. Given a user profile, HyGen begins by traversing the full graph to find all the neighbors of the seeds and marking them as discovered. Figure 7 is a sample SPARQL query used to retrieve neighbors of a set of nodes. Numerical weights (from 0 to 1) are assigned to the edges based on the data sources where the associations are originated. Users can adjust the data source weight based on their own experience and needs in the user profiles. For example, if a user is very familiar with database A and would like to search for novel information beyond the scope of A, then the user can assign a lower score to A.



```
PREFIX TLO: <http://www.crc.com/Crc/TopLevel-ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
Select distinct ?s ?p ?o
Where { ?s ?p ?o .
        {
            { ?o TLO:hasId "DRG_Aldesleukin"^^xsd:string } UNION
            { ?o TLO:hasId "DRG_Aminoglutethimide"^^xsd:string } UNION
            { ?o TLO:hasId "DRG_Bevacizumab"^^xsd:string } UNION
            ... ...
        }
}
```

Figure 7. Sample SPARQL query to expand sub graph iteratively

Next, HyGen performs graph analysis to rank all the discovered nodes based on the criteria defined in the user profile. Any low ranking node is turned back to undiscovered. This counts as one step.

In the subsequent iterations, HyGen searches for *undiscovered* nodes that are neighbors of the *discovered* nodes; ranks all the *discovered* nodes; and re-labels them *discovered* or *undiscovered* according to the ranks. If the user specified the maximum number of steps *X* in the profile, then HyGen stops searching after *X* steps. Otherwise, HyGen stops after it has exhausted all nodes in the full graph. At the end of the final iteration, all the *discovered* nodes and their edges form the virtual sub graph specific to the given user profile.

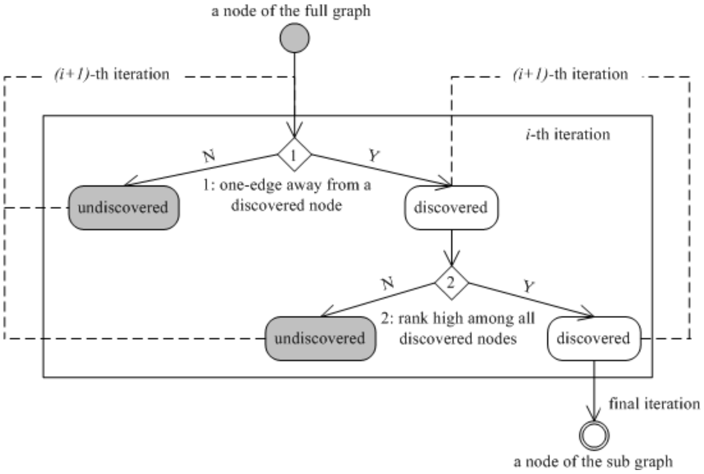
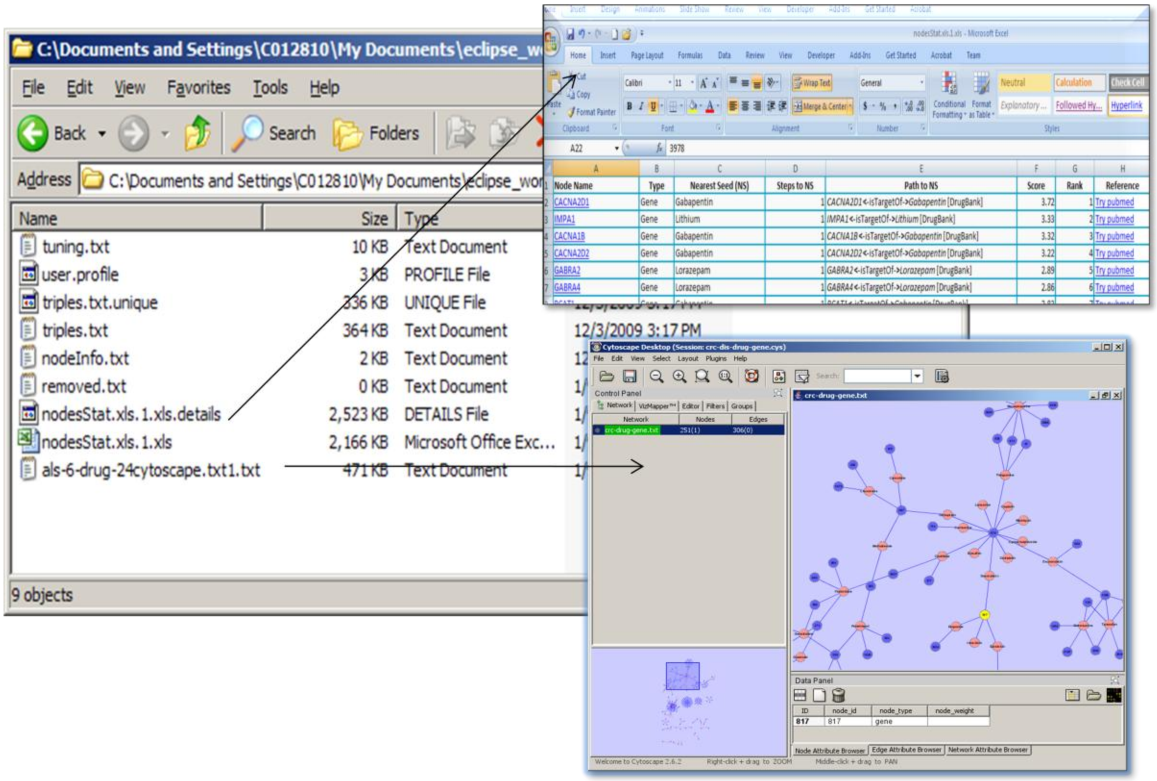


Figure 8. State diagram of a node

The state diagram, Figure 8, illustrates how one node's state is influenced by the node's ranking and by its neighbors in each iteration. A node is in one of the two states during this process: *discovered* or *undiscovered*, with the initial state being *undiscovered* except for the seeds defined by the user profile. An *undiscovered* node becomes *discovered* if any of its neighbors has been *discovered*; a node's state changes from *discovered* to *undiscovered* if its ranking is too low.

In other words, HyGen's exploration radiates out slowly from the original seeds, acquiring or disowning nodes at each iteration. We named the process of re-ranking *discovered* nodes pseudo- relevance feedback. The term was borrowed from document retrieval systems. However, instead of using pseudo- relevance feedback for query expansion, HyGen applies this strategy to re-rank and re-arrange the nodes that have already been discovered. Adopting pseudo-relevance feedback in this novel way, HyGen can quickly construct a user-specific view of the full knowledge space with high sensitivity and selectivity.

Layer E: Output Personalized Views



C:\Documents and Settings\C012810\My Documents\eclipse_w

Address C:\Documents and Settings\C012810\My Documents\eclipse_wor

Name	Size	Type	
tuning.txt	10 KB	Text Document	
user.profile	3 KB	PROFILE File	
triples.txt.unique	336 KB	UNIQUE File	
triples.txt	364 KB	Text Document	12/3/2009 3:17 PM
nodeInfo.txt	2 KB	Text Document	12
removed.txt	0 KB	Text Document	1/
nodesStat.xls.1.xls.details	2,523 KB	DETAILS File	1/
nodesStat.xls.1.xls	2,166 KB	Microsoft Office Exc...	1/
als-6-drug-24cytoscape.txt1.txt	471 KB	Text Document	1/

9 objects

nodesStat.xls.1.xls - Microsoft Excel

A22 3978

Node Name	Type	Nearest Seed (NS)	Steps to NS	Path to NS	Score	Rank	Reference
CACNA2D1	Gene	Gabapentin	1	*CACNA2D1*<-isTargetOf->*Gabapentin* [DrugBank]	3.72	1	Try pubmed
IMPA1	Gene	Lithium	1	*IMPA1*<-isTargetOf->*Lithium* [DrugBank]	3.33	2	Try pubmed
CACNA1B	Gene	Gabapentin	1	*CACNA1B*<-isTargetOf->*Gabapentin* [DrugBank]	3.32	3	Try pubmed
CACNA2D2	Gene	Gabapentin	1	*CACNA2D2*<-isTargetOf->*Gabapentin* [DrugBank]	3.22	4	Try pubmed
GABRA2	Gene	Lorazepam	1	*GABRA2*<-isTargetOf->*Lorazepam* [DrugBank]	2.89	5	Try pubmed
GABRA4	Gene	Lorazepam	1	*GABRA4*<-isTargetOf->*Lorazepam* [DrugBank]	2.86	6	Try pubmed

Cytoscape Desktop (Session: crc-dis-drug-gene.cys)

Network	Nodes	Edges
crc-drug-gene.txt	251(1)	306(0)

crc-drug-gene.txt

ID	node_id	node_type	node_weight
817	817	gene	

Welcome to Cytoscape 2.6.2 Right-click + drag to ZOOM Middle-click + drag to PAN

Figure 9. Output files of HyGen

One of the key steps of pseudo- relevance feedback algorithm is to rank and re-rank nodes based on graph analysis. The rank of a node v is computed as the weighted mean of seven factors, each of which has a value between 0 and 1. The definition and calculation of each factor can be found in Appendix D. Users can further control the ranking by adjusting weights k_1 to k_7 in their profiles. The final score is computed by equation (3-1). The node with the highest score achieves the top ranking.

$$\overline{X_{SA}} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, X = \{C_{SA}, S_{SA}, R_{SA}, P_{SA}, L_{SA}, T_{SA}, F_{SA}\} \; and \; W = \{k_1, \dots, k_7\} \qquad (3\text{-}1)$$

Once the sub graph is constructed, HyGen can issue semantic queries against it. The results are ranked based on the same criteria and weights described above. At the end, HyGen generates two main artifacts for the user: an annotated sub graph file viewable in CytoScape (www.cytoscape.org); and a set of spreadsheets of sorted associations, with their graph attributes, scores, shortest path to the seeds and other computed properties. Figure 9 shows a screen capture of the various output files.

CHAPTER FOUR: RESULTS

Case Studies in Colorectal Cancer

Introduction

Colorectal Cancer (CRC) also called colon cancer or large bowel cancer, is the second leading cause of cancer-related death [56]. CRC disease progression is believed to be a step-wise process, where cells change from normal epithelium through polyp form to carcinoma. Mutations in two classes of genes: tumor-suppressor genes and proto-oncogenes, are found to increase the risk of developing CRC. Studies have also shown that CRC is related to at least one of three different pathways: termed chromosomal instability, microsatellite instability, and the CpG island methylator phenotype [57]. Different pathways tend to affect different sets of tumor suppressor genes and are characterized by different biological behaviors.

Early and accurate detection of different types of colorectal cancer may greatly improve the chances of survival. Medical interventions should be tailored for individual patient. However, this is not the case. Even though great progress was made in understanding the molecular basis of CRC, it has translated into few genetic biomarkers that are currently used in clinical practice [58]. We believe a translational approach is needed to help transform the biological discoveries into clinically diagnostic tools and personalized CRC medicine.

To test HyGen, we carried out a set of case studies tailored for multiple users involved in colorectal cancer (CRC). The first user is a health care consumer with little medical knowledge; the second user is an experienced practitioner having rich medical knowledge; the third user is a pharmacologist with deep understanding of chemistry and biology. Three sample profiles were developed for the users (Table 1). Based on the profiles, three sub graphs (views) were created and analyzed.

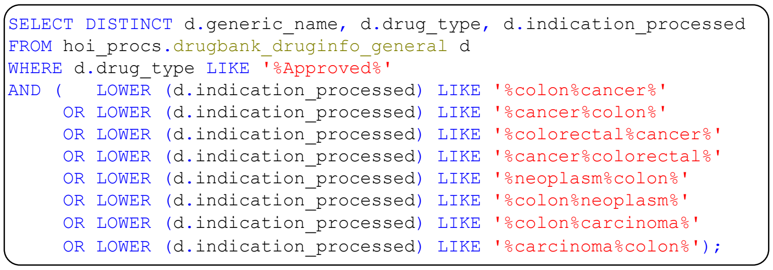
Table 1. Three sample user profiles in CRC

	User Description	**Seeds**	**Preference**
#1	**A health care consumer** with little knowledge of biology or medicine, wanting to know possible treatments for CRC.	a single term of interest: colorectal cancer	well-established information, short associations
#2	**A health care practitioner** who read a review paper on CRC physiology	32 CRC-related genes cited in the	fresh information, medium-length

	and wondered if such knowledge in the biology discipline can help him in clinical practice.	review paper	associations, hotspots
#3	**A pharmacologist** interested in disease target and biomarkers, wanting to know if knowledge from the clinical field can help in drug discovery.	52 drugs often prescribed by doctors to CRC patients	fresh information, long associations, hotspots

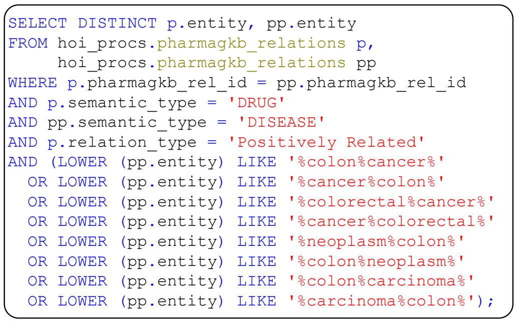
View for a Health Care Consumer

We simulated the user profile of a CRC patient interested in new therapeutic options. The profile P_{pat}, contained a single seed “colorectal cancer”, and was set up to award well-established information and short associations. The single seed expanded to a sub graph with 831 edges and 602 nodes. We compared this results with direct SQL queries against pharmacological sources that provide drug and disease information (DrugBank, PharmGKB, and CTD). Searching for drugs with inductions being CRC or its synonyms, we retrieved eight hits using direct SQL query in DrugBank (Figure 10) and no hits from PharmGKB (Figure 11). Since the chemical-disease relationship in CTD has a broad definition, the search in CTD returned 1030 hits but most of them are not compounds that can be taken by human as drugs.



```
SELECT DISTINCT d.generic_name, d.drug_type, d.indication_processed
FROM hoi_procs.drugbank_druginfo_general d
WHERE d.drug_type LIKE '%Approved%'
AND (   LOWER (d.indication_processed) LIKE '%colon%cancer%'
     OR LOWER (d.indication_processed) LIKE '%cancer%colon%'
     OR LOWER (d.indication_processed) LIKE '%colorectal%cancer%'
     OR LOWER (d.indication_processed) LIKE '%cancer%colorectal%'
     OR LOWER (d.indication_processed) LIKE '%neoplasm%colon%'
     OR LOWER (d.indication_processed) LIKE '%colon%neoplasm%'
     OR LOWER (d.indication_processed) LIKE '%colon%carcinoma%'
     OR LOWER (d.indication_processed) LIKE '%carcinoma%colon%');
```

Figure 10. SQL used to search for CRC drugs in DrugBank



```
SELECT DISTINCT p.entity, pp.entity
FROM hoi_procs.pharmagkb_relations p,
     hoi_procs.pharmagkb_relations pp
WHERE p.pharmagkb_rel_id = pp.pharmagkb_rel_id
AND p.semantic_type = 'DRUG'
AND pp.semantic_type = 'DISEASE'
AND p.relation_type = 'Positively Related'
AND (LOWER (pp.entity) LIKE '%colon%cancer%'
  OR LOWER (pp.entity) LIKE '%cancer%colon%'
  OR LOWER (pp.entity) LIKE '%colorectal%cancer%'
  OR LOWER (pp.entity) LIKE '%cancer%colorectal%'
  OR LOWER (pp.entity) LIKE '%neoplasm%colon%'
  OR LOWER (pp.entity) LIKE '%colon%neoplasm%'
  OR LOWER (pp.entity) LIKE '%colon%carcinoma%'
  OR LOWER (pp.entity) LIKE '%carcinoma%colon%');
```

Figure 11. SQL used to search for CRC drugs in PharmGKB

On the other hand, the top 5% associations[1] identified by HyGen contained twenty five drugs, including the eight hits from DrugBank search. Those drugs are listed in Table 2. Their ranking scores are the weighted sums of the seven factors discussed in Chapter Three.

Literature research confirmed that it is possible to use those drugs in treating CRC patients [59-67]. HyGen's full graph contains no more data than what is in the databases we queried directly. HyGen helped the patient to identify additional treatment options because it has connected the clinical features with the genomic information, and pharmacology information. This demonstrates that HyGen is greater than the sum of its parts, an important benefit of bridging silos.

Table 2. Drugs ranked at top 5% of HyGen's results

Drug	**Indications**
cisplatin	For the treatment of metastatic testicular tumors, metastatic ovarian tumors and advanced bladder cancer.
vincristine	For treatment of acute leukemia, malignant lymphoma, Hodgkin's disease, acute erythraemia, acute panmyelosis.
topotecan	For the treatment of metastatic carcinoma of the ovary and small cell lung cancer following the failure of first-line chemotherapy.
idarubicin	For the treatment of acute myeloid leukemia (AML) in adults. This includes French-American-British (FAB) classifications M1 through M7.
daunorubicin	For remission induction in acute nonlymphocytic leukemia (myelogenous, monocytic, erythroid).
cytarabine	For the treatment of acute non-lymphocytic leukemia, acute lymphocytic leukemia and blast phase of chronic myelocytic leukemia.
methotrexate	For the treatment of gestational choriocarcinoma, chorioadenoma destruens and hydatidiform mole. Also for the treatment of severe psoriasis and severe, active, classical or definite rheumatoid arthritis.
mercaptopurine	For remission induction and maintenance therapy of acute lymphatic leukemia.
tamoxifen	For the treatment of breast cancer.
capecitabine	For the treatment of patients with metastatic breast cancer resistant to both paclitaxel and an anthracycline-containing chemotherapy regimen.
epirubicin	For use as a component of adjuvant therapy in patients with evidence of axillary node tumor involvement following resection of primary breast cancer.
etoposide	For use in combination with other chemotherapeutic agents in the treatment of refractory testicular tumors and as first line treatment in patients with small cell lung cancer.

[1] 5% is used as the standard cutoff for all results returned by HyGen

trimetrexate	For use, with concurrent leucovorin administration (leucovorin protection), as an alternative therapy for the treatment of moderate-to-severe Pneumocystis carinii pneumonia (PCP) Also used to treat several types of cancer including colon cancer.
raltitrexed	For the treatment of malignant neoplasm of colon and rectum.
pamidronate	For the treatment of moderate, severe hypocalcaemia associated with malignancy.
paclitaxel	Used in the treatment of Kaposi's sarcoma and cancer of the lung, ovarian, and breast.
oxaliplatin	Used in combination with infusional 5-FU/LV, is indicated for the treatment of advanced carcinoma of the colon.
mitomycin	For treatment of malignant neoplasm of lip, oral cavity, pharynx, digestive organs, peritoneum, female breast, and urinary bladder.
levamisole	For adjuvant treatment in combination with fluorouracil after surgical resection in patients with Dukes' stage C colon cancer.
leucovorin	For the treatment of osteosarcoma (after high dose methotrexate therapy). Also used in combination with 5-fluorouracil to prolong survival in the palliative treatment of patients with advanced colorectal cancer.
irinotecan	For the treatment of metastatic colorectal cancer (first-line therapy when administered with 5-fluorouracil and leucovorin).
gemcitabine	For the first-line treatment of patients with metastatic breast cancer, locally advanced metastatic non-small cell lung cancer and as first-line treatment for patients with adenocarcinoma of the pancreas.
fluorouracil	For the topical treatment of multiple actinic, solar keratoses. also useful in the treatment of superficial basal cell carcinomas when conventional methods are impractical, Fluorouracil injection is indicated in the palliative management of some types of cancer, including colon, rectum, breast, and stomach.
docetaxel	For the treatment of patients with locally advanced metastatic breast cancer after failure of prior chemotherapy. In combination with prednisone, in the treatment of patients with androgen independent (hormone refractory) metastatic prostate cancer.
bevacizumab	For treatment of metastatic colorectal cancer.

View for a Health Care Practitioner

Scientific discoveries can help practitioners to provide better care to patients. In this case study, we simulated the user profile of a medical doctor who is a CRC specialist and is interested in other disorders and complications related to CRC. The profile, P_{doc}, contained 32 known genes related to CRC. Since the user is an expert looking for knowledge applicable to medical practices, we set up P_{doc} to award fresh but medium-length associations and nodes with higher degree of connectivity, because highly connected nodes are more likely to be organizing functional modules and critical for survival. The maximum iteration X was set to 2 and the sub graph finished with 4591 edges and 2392 nodes. The top disorders associated with CRC are listed in Table 3.

To judge the novelty of the results, we searched PubMed for original papers where the new disorder name and CRC occurred together in the same titles or abstracts. The number of co-occurrences is listed in Table 3 as well. The low numbers seem to indicate that some suggestions are quite novel, assuming novel information will be less known and appear in fewer papers.

Table 3. Disorders suggested to a practitioner (5% cutoff)

Disorder	**Score**	**# Paper**
neural tube defects	1.00	5
rippling muscle disease	0.97	0
polydactyly, preaxial IV	0.88	0
vitamin d-dependent rickets, type I	0.85	0
mismatch repair cancer syndrome	0.82	11
dyssegmental dysplasia, silverman-handmaker type	0.82	0
osteoporosis	0.81	29
spondylometaphyseal dysplasia, kozlowski type	0.78	0
premature chromatid separation trait	0.77	0
neurofibromatosis, type I	0.76	3
meningioma, familial	0.73	1
otospondylomegaepiphyseal dysplasia	0.68	0

The high-ranking disorders may shine lights on the common disease mechanisms of CRC and other disorders. For example, we have found that the link between CRC and Neural Tube Defects (NTD) is scientifically possible. Folate supplements have been used to prevent NTD [68, 69]. There are studies claiming that folate may also lower CRC risk [70, 71]. The initial full graph has no direct links between CRC and NTD. Zooming into the sub graph as shown in Figure 12, we noticed that mutation in TP53 is linked to increase risk of CRC; TP53 is a gene encoding tumor protein p53, which is involved in DNA repair and changes in metabolism; TP53 connects to MTHFR via a drug that is a pyrimidine analogue and inhibits the cell's ability to synthesize DNA; and MTHFR polymorphism is linked to an increased risk for NTD. Those links suggest that CRC and NTD pathways may share some common components. By comparing of their pathways and biological process, an expert may arrive at new hypnoses about the disease mechanisms of CRC and NTD.

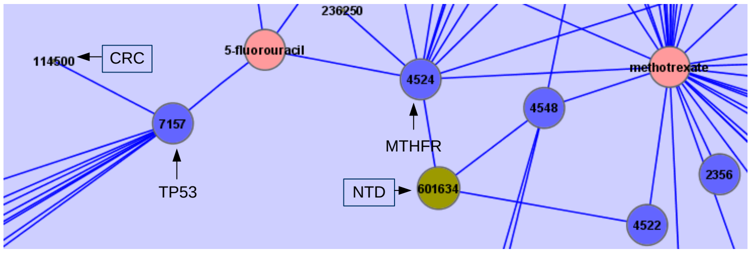


Figure 12. Zoom-in view of the sub graph

The next high-ranking association identified by HyGen, Rippling muscle disease (RMD), is a rare autosomal dominant disorder that may occur sporadically [72]. It was reported that sporadic RMD could be treated by thymectomy or immunosuppression [73, 74]. Some RMD patients' symptoms were reduced after treated with anti-cancer drugs [75]. Experts suggest that sporadic RMD may be a new paraneoplastic or autoimmune disease, characterized by certain antibodies response against self [73]. Similar antibodies can also sometimes be found in patients with CRC [76, 77].

The third high-ranking association in Table 3 is preaxial polydactyly. It is a congenital anomaly characterized by the presence of more than the normal number of fingers. Many believe it is part of a complex genetic syndrome [78, 79]. The gene or set of genes responsible for preaxial polydactyly have been localized to chromosome 7q36 [80, 81]. A homeobox gene HB9 is within the critical region of 7q36 and is also expressed in pancreas, small intestine, and colon [82]. The association seems to suggest that the two phenotypes, preaxial polydactyly and CRC, are linked through HB9.

The last example we discuss here is the possible link between CRC and type I vitamin D-resistant rickets, VDDRI (ranked 4th in Table 3). Even though no PubMed paper contains those two disorders in the same abstract, the practitioner, being an experienced physician, may realize that vitamin D is recommended to lower the risks of both diseases. Further research may show him that VDDRI is associated with mutations in the gene of the vitamin D receptor (VDR) [83]. Meanwhile, colorectal cells contain vitamin D Receptors and are able to convert 25(OH) vitamin D into 1,25(OH)2 vitamin D, which may prevent tumor progression in colon. The practitioner may follow up on this interesting connection by comparing the lab results of VDDRI and CRC patients, especially their 1,25(OH)2 VD level tests. This finding may eventually lead to novel diagnosis tools or therapies.

View for a Biomedical Researcher

Knowledge discovered during clinical practice provides novel insights of disease pathology to researchers working in the basic science discipline. In this case study, we simulated the profile of a pharmacologist who is interested in novel disease targets and biomarkers related to CRC. This user's profile, P_{sci}, consisted of 52 drugs that are often prescribed by doctors to CRC patients. Since the user is an expert looking for novel associations, we set the maximum iteration X=3. The sub graph expanded to 550 edges and 492 nodes including drugs, disorders, genes, pathways, and interactive partners.

P_{sci} was set to award long associations, recent information, and hotspots because highly connected nodes in biological networks are generally found to be essential for viability and the delineation of those nodes often leads to new insights and hypotheses [37, 39]. The top ranking genes are listed in Table 4. We searched PubMed for original papers where the suggested gene and CRC occurred together in the titles or abstracts. The number of co-occurrences is listed in Table 4. Assuming novel information will be reported by few papers, genes such as CAV3, LYST, TYROBP and DRD1 seem to be more "interesting" and may be candidates for future CRC genetic studies.

Table 4. Genes suggested to a pharmacologist (5% cutoff)

Gene	**Score**	**#Paper**
KRAS	1.00	289
FGFR1	0.88	3
CAV3	0.76	0
LYST	0.69	0
ATM	0.68	32
TYROBP	0.63	0
YWHAE	0.59	1
DRD1	0.58	0
IFNG	0.58	2

TAP2	0.56	1
CCND1	0.56	9
CPS1	0.55	0
NR0B1	0.52	0
AGT	0.50	1
EPOR	0.49	0
RAG1	0.49	0
SCARB2	0.49	0
AIRE	0.48	0

Compare the Three Views

We compared the topology properties of the three sub graphs discussed above. Sub graph No.1 had 831 edges and 602 nodes; No.2 had 4591 edges and 2392 nodes; No.3 had 550 edges and 492 nodes. Their node types included drugs, genes, pathways, clinical features and disorders. Basic network topology analyses were carried out using Network Analyzer (http://med.bioinf.mpi-inf.mpg.de/netanalyzer). Figure 13 shows how the user preferences have affected the shortest path distribution of the three sub graphs. User profile No.1 was set to reward short associations, thus the Average Shortest Path Length (ASPL) of sub graph No.1 was 3.6 and its diameter was 7, the shortest of all three sub graphs. Profile No.3 preferred long associations, thus sub graph No.3 had the largest ASPL and diameter: 5.91 and 16. Sub graph No.2's ASPL and diameter were 4.4 and 11 respectively, because its user preferred medium-length associations.

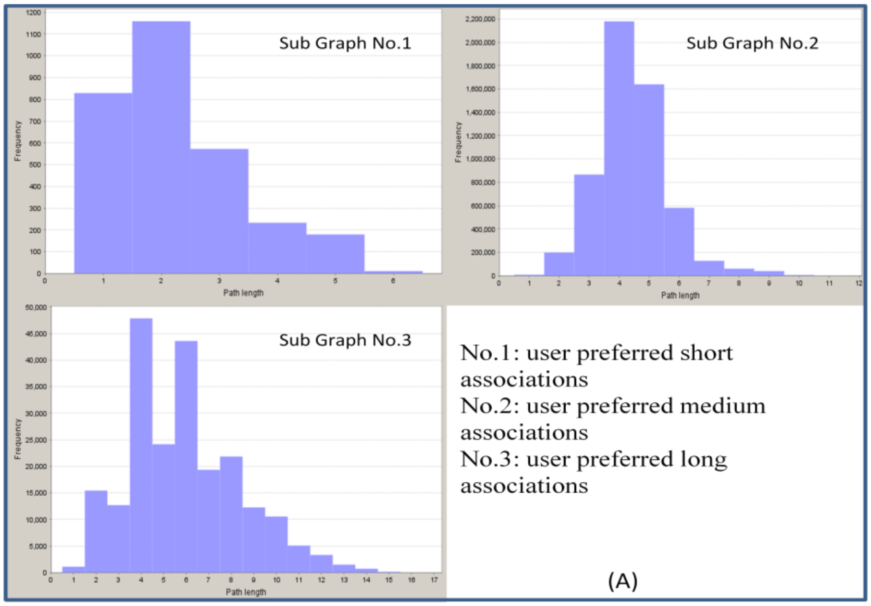


Figure 13. Compare the shortest path distributions of the sub graphs

Figure 14 shows the degree distributions of the sub graphs with a fitted power law. Each point on the graph indicates the number of nodes with a particular degree k for k = 0,…,n. A power-law degree distribution is often seen in a scale-free network such as many biological networks [84]. In spite of the differences in the size and shortest path, the three graphs exhibit similar trend in the degree distributions, suggesting that the scale-free feature of the full graph has been preserved in the sub graphs independent of the user profiles. Such observation indicates that sub-graphing based on user profiles may be used to lower the cost of large-scale graph analyses without distorting the nature of the original graph. However, it is necessary to point out that we must be cautious when extrapolating from sub graph to the properties of the full graph as Stumpf and colleagues have pointed out [85].

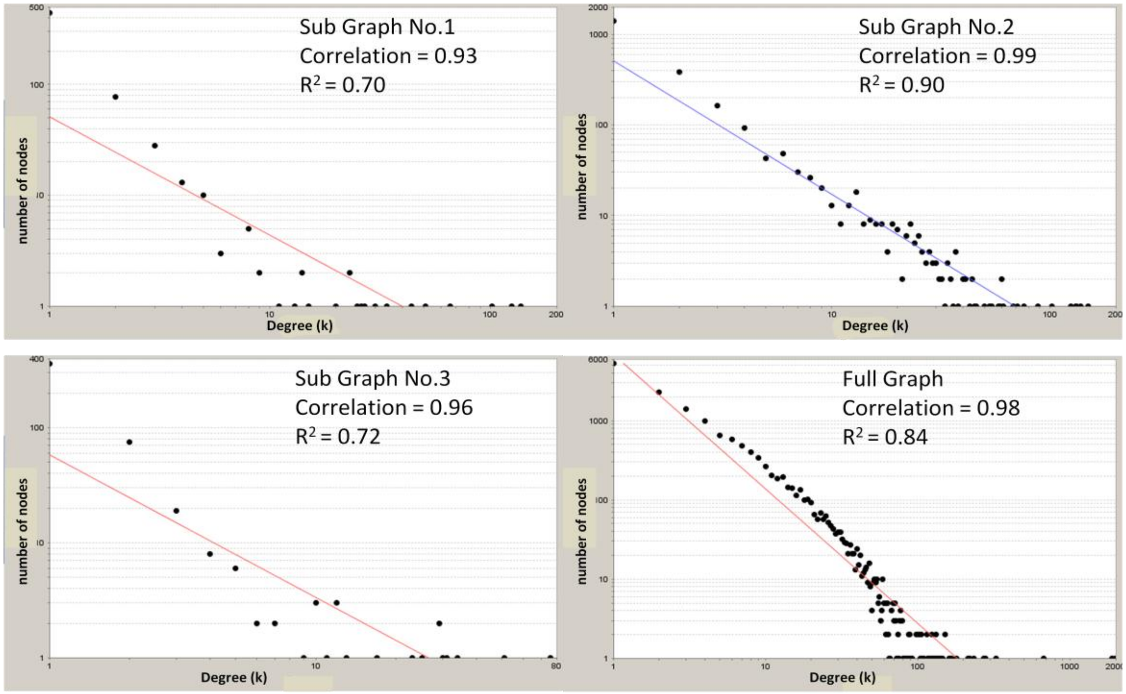


Figure 14. Compare the degree distributions of the sub graphs

Case Studies in Amyotrophic Lateral Sclerosis

Introduction

Amyotrophic Lateral Sclerosis (ALS), also known as Lou Gehrig's disease, is a progressive, fatal, neurodegenerative disease which usually leads to paralysis and death within five years of symptom onset [86]. Since its first report in the mid 1800s, extensive research efforts have been spent in battling ALS. However, no cure has been found yet; the only FDA-approved drug, riluzole, can prolong life by 3 to 6 months but cannot change the course of the disease. Understanding ALS disease mechanism not only can lead to early diagnosis tools and effective treatments, but can also improve the knowledge of other neurodegenerative diseases.

To date, the molecular underpinnings of ALS remains elusive [87]. 10% of ALS cases are familial ALS (fALS). Various genes have been identified in fALS patients, the most important of which is SOD1, whose mutation accounts for up to 20% of all familial cases [88]. Other genes implicated in fALS include ALS2, SETX, VAPB, ANG, TARDBP, MAPT and DCTN1 [88, 89]. Little is known about the other 90% of the ALS cases, namely sporadic ALS (sALS) [90, 91]. Two main strategies have been used in identifying causative genes of ALS [92-98]. Candidate gene studies search for genes based on priori hypotheses about the disease mechanisms. Genome-wide association studies (GWAS) do not make assumptions about the nature or location of the genes but need large sample size for association analysis. Although it is believed that genetic factors play a central role, very few genes have been found unequivocally implicated in ALS [93]. Experts believe that multiple genetic and environmental factors are implicated in ALS [99]; to understand its biological heterogeneity requires cross-disciplinary collaborations and translational efforts, such as meta-analysis of genetic, toxicology, pharmacological, health outcome and environmental data.

ALS is a complex disease, affected by many factors [92, 99], such as the multiple effects of single genes, the interactions of multiple genes, and the interactions of genes with environment. HyGen's graph approach, therefore, seems ideal to study the intricate links among those factors. On the other hand, ALS is characterized by late onset and short survival. Association and analysis of data from unrelated individuals are necessary because it is difficult to obtain sufficient number of cases required for classical family-based studies. It is beneficial to leverage HSNC content in ALS research, because information in HSNC is generated by a large number of patients with diverse ethnic and environmental backgrounds. For instance, 5% of all ALS patients in the U.S. are registered members of PLM. The quantified self-tracked information generated by those patients is arguably the largest data set for ALS translational research. We decide to expand the scope of HyGen by using traditional research-oriented databases in combination with the aggregated content from PLM's ALS community. The following sections discuss the quantities evaluations and the hypotheses highlighted by HyGen.

Convert HSNC Content to Graph Nodes

ALS patients can enter both structured and unstructured data in their PLM profiles. Individual-level data are aggregated and reflected in PLM's community reports. A community symptom report contains the prevalence and severity of the symptoms. We extracted the most frequently reported (MFR) symptom terms from ALS community's symptom report (www.patientslikeme.com/als/symptoms). We then used MetaMap to map those symptom terms to UMLS concepts, followed by manual inspection. The top ten MFR symptoms and the matching UMLS concepts are displayed in Appendix E.

A community treatment report contains information such as dosage distribution, side-effects reported by patients, patients' time on the treatment, and reasons patients have started or stopped the treatment. We extracted MFR prescription drug names from community treatment report (www.patientslikeme.com/als/treatments) and normalized them against CHEMLIST compounds or their synonyms. Since a UMLS concept Id or a CHEMLIST Id defines the URI of a node, each MFR symptom term or drug name reported in PLM is uniquely mapped to the node that represents the same clinical concept or chemical substance in the full graph. To establish mappings of instance-level terms is the key to connect patient-generated content with research-oriented biomedical data sources. Biomedical ontologies and thesaurus such as UMLS and their companion linguistic tools have made it possible to automate a large part of the mapping process.

The other challenge mentioned previously is that more than half of the symptoms submitted by PLM patients were not "Signs or Symptoms" in UMLS [28]. We circumvented this problem by defining one general type called "clinical-feature" for concepts belonging to multiple UMLS types listed in Table 5. The relationships between clinical-features and other types of nodes were loosely defined (e.g. “related_to_gene” and “related_to_drug”). Obviously, the penalty of this approach is a higher false positive rate. Therefore the pseudo-relevance feedback strategy (described in the methodology chapter) is critical to reduce the number of irrelevant connections.

Table 5. Merged UMLS semantic types

semantic group	semantic type id	semantic type name
Disorders	T019	Congenital Abnormality
Disorders	T020	Acquired Abnormality
Disorders	T033	Finding
Disorders	T037	Injury or Poisoning
Disorders	T046	Pathologic Function
Disorders	T047	Disease or Syndrome
Disorders	T048	Mental or Behavioral Dysfunction
Disorders	T049	Cell or Molecular Dysfunction
Disorders	T050	Experimental Model of Disease
Disorders	T184	Sign or Symptom
Disorders	T190	Anatomical Abnormality
Disorders	T191	Neoplastic Process
Phenomena	T034	Laboratory or Test Result
Physiology	T039	Physiologic Function

Capture Treatment-Symptom Correlation in HSNC

We carried out heuristic evaluations in CRC case studies. Heuristic approach has limitations in comparison across use cases, due to its relative and qualitative nature. In addition, heuristic metrics are hard to automate and to apply in larger tests. However, due to the lack of gold standards and quantitative test methods for biomedical hypothesis generation systems, heuristic tests are still the most common approaches in evaluating systems like HyGen.

Here we present a quantitative approach using HSNC. The initial full graph of HyGen has been compiled from traditional research-oriented data sources, whose content is the result of systematic research and analysis. On the other hand, content in HSNC is the by-product of health care, which is a Complex System. The associations embedded in HSNC reflect the emergent and self-organizing properties of Complex Systems. We desire to test whether HyGen can identify the associations in “real-world” health care practice (the “bedside”) based on data extracted from research-oriented data sources (the “bench”). In other words, starting from PLM’s MFR drugs, we expect HyGen to retrieve

PLM’s MFR symptoms from the full graph and to rank them high. Similarly, we expect HyGen to highlight the proper MFR drugs for PLM’s MFR symptoms. We expect both tests to achieve sufficient statistical significance with p-value less than 0.01.

The p-value of HyGen’s results was obtained by permutation testing. Permutation test is also called randomization test, where one computes statistic for all possible permutations of the data to calculate the exact p-value. In practice, however, an approximate p-value is computed by sampling a sufficient large number of possible permutations. To establish independent permutations, we constructed N random graphs (permutations) by reassigning the edges between nodes in the real full graph. We ran HyGen against the real graph and saved the result as the observed statistic. We then ran HyGen against each random graph (permutation) and labeled the permutation as “success” if it achieved similar or better results than the real graph.

Therefore, each permutation is a Bernoulli trial with p-value being the proportion of all “success” runs represented by equation (4-1) where x is the number of successes; N is the total number of permutations; p is the p-value.

$$Pr(p = \frac{x}{N}) = \binom{N}{x} p^{x}(1-p)^{N-x} \tag{4-1}$$

In the first experiment, we invoked HyGen using the top ten MFR drugs as the seeds. The top ten MFR symptoms were identified and ranked at upper 5% by HyGen with p-value less than 0.01 based on 10,000 permutations. In the second experiment, we used the top ten MFR symptoms as the seeds. Based on the same condition, HyGen identified the top ten MFR drugs with p-values less than 0.01.

We also computed the enrichment factor (EF) of HyGen’s results. EF is the ratio of the abundance of a particular entity in an enriched environment to its abundance in the original environment. The EF of the two experiments can be computed by equation (4-2) and equation (4-3) respectively. The top ten MFR symptoms were identified with 36 fold enrichment. The top ten MFR drugs were identified with 8 fold enrichment.

$$\frac{\text{abundance of FMR symptoms in HyGen result}}{\text{abundance of FMR symptoms in the full graph}} = \frac{{}^{10}\!/_{\text{\# symptoms in top 5\% of the sub graph}}}{{}^{10}\!/_{\text{\# symptoms in the full graph}}} = 36 \tag{4-2}$$

$$\frac{\text{abundance of FMR drugs in HyGen result}}{\text{abundance of FMR drugs in the full graph}} = \frac{{}^{10}\!/_{\text{\# drugs in top 5\% of the sub graph}}}{{}^{10}\!/_{\text{\# drugs in the full graph}}} = 8 \tag{4-3}$$

Those two simple tests boost the belief that HyGen can use knowledge from the scientific discipline to identify associations relevant to health care practices. They also illustrate the potential of HSNC as supplementary, empirical data sources for translational research.

Effects of Seeds on Hit Rate

Having converted PLM's MFR terms into graph nodes in previous evaluations, we could use all of them as seeds for HyGen to identify potential ALS genes. However we decided to find a systematic process for selecting the optimal set of seeds. Based on detailed literature research, we defined the gold standard to be twenty ALS genes reported in some of the most salient studies to date [54, 90, 91, 94-98, 100-105]. In Figure 15, we summarized them according to their dates of publication.

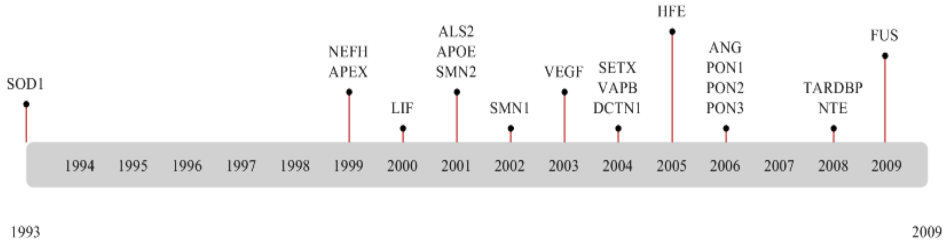


Figure 15. ALS genes used as the gold standard

Next, we conducted sets of experiments to study how the selection of seeds affects HyGen's ability to identify candidate genes. A pool of twenty seeds was derived from the top MFR terms in PLM community reports. In each set of experiments, HyGen built twenty sub graphs (G_1 to G_{20}) with increasing number of seeds. The first sub graph started from one seed $x_1$randomly picked from the pool; the second sub graph started from x_1 plus another random seed x_2; …; the last sub graph G_{20} was constructed from all twenty seeds.

$$\{ x_1\} \rightarrow G_1$$

$$\{x_1, x_2\} \rightarrow G_2$$

…

$$\{x_1, x_2, \ldots, x_{20}\} \rightarrow G_{20}$$

For each sub graph, HyGen produced a sorted list of genes associated with ALS. We compared the top 5% genes in HyGen's list with the gold standard, and called the overlapping genes "hits". Four properties: number of hit, number of nodes, number of edges, and average degree were calculated for G_1 to G_{20} and plotted in one chart. Each property *p* is represented by a line in the chart, and each point (*x*, *y*) on the line corresponds to a sub graph G_i. The value of *x* is the number of seeds used to build G_i, and *y* is a normalized value: $(p_x-p_{min})/(p_{max}-p_{min})$ where *p* is one of the four properties of G_i.

A sample chart (Figure 16) represents a set of twenty sub graphs. Assuming the number of nodes and edges reflect the size of a graph, and the average degree roughly corresponds to the connectivity, such chart can be used to study how the seeds affect the sub graph and its hit rate. With twenty seeds in the pool, there are more than 2×1018 (20!) possible charts. We manually analyzed twenty samples and observed two general trends:

Num. Of seeds	Num. of nodes	Num. of edges	Average degree	Num. of hits
1	2973	4464	3.0	0
2	3996	6417	3.2	0
3	5016	7936	3.2	3
4	5224	8331	3.2	4
5	5438	8626	3.2	4
6	7925	14859	3.7	5
7	8180	15385	3.8	5
8	11392	27568	4.8	6
9	11637	28240	4.8	5
10	11893	28592	4.8	4
11	12119	29699	4.9	3
12	12351	30038	4.9	3
13	12394	30754	5.0	3
14	12907	32126	5.0	3
15	16475	45386	5.5	6
16	16976	46674	5.5	6
17	16979	46688	5.5	6
18	17615	48679	5.5	6
19	17648	48859	5.5	6
20	17663	48899	5.5	5

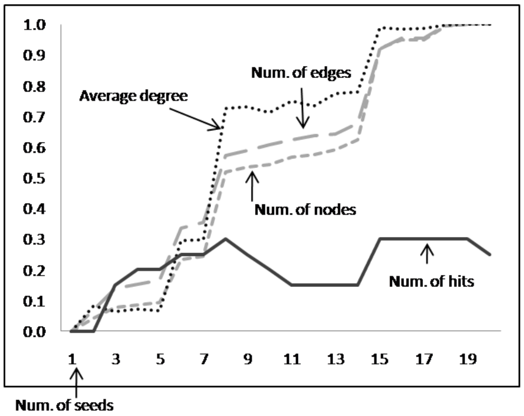


Figure 16. Visualize properties of twenty sub graphs in one chart

Observation 1: given more seeds, HyGen built larger and more connected networks. The growth of edges and nodes followed roughly the same trend as the growth of average degree. In Figure 17, the three dashed lines are of similar shape.

Observation 2: given more seeds, HyGen did not necessarily identify more hits. When connectivity grew rapidly, hit rate increased. In the chart, sharp climbs of the dark dashed lines co-occur with climbs of the solid line. In a sufficiently large network, addition of low value seeds could negatively impact the hit rate. Plateaus of the dark dashed lines often co-occur with the dips of the solid line, especially in the right-hand side of the chart.

Considering the origin of the seeds and HyGen's work flow, those trends are not difficult to explain. All twenty seeds have been derived from MFR terms reported by ALS patients. The fact that they have emerged from thousands of other terms, suggests that they are interrelated, central concepts in ALS knowledge space. Therefore, addition of each seed increases both the size and the connectivity of the knowledge space (Observation 1). Sharp climbs of the connectivity indicate that the seeds added have filled important knowledge gaps in the previous sub graphs. Such high value seeds are more likely to increase hit rate. Plateaus of connectivity indicate that the seeds added have not remarkably enriched the knowledge in the sub graph. When the graph is large, however, adding a low value seed can still increase the total complexity and negatively impact the hit rate (Observation 2). As shown by the charts, HyGen becomes more susceptible to low value seeds as the graph grows bigger. Such trend has also been observed when applying HyGen to CRC. Both ALS and CRC are complex diseases. More experiments in other disease areas are needed to study whether this observation can be generalized to other multi-system, complex diseases.

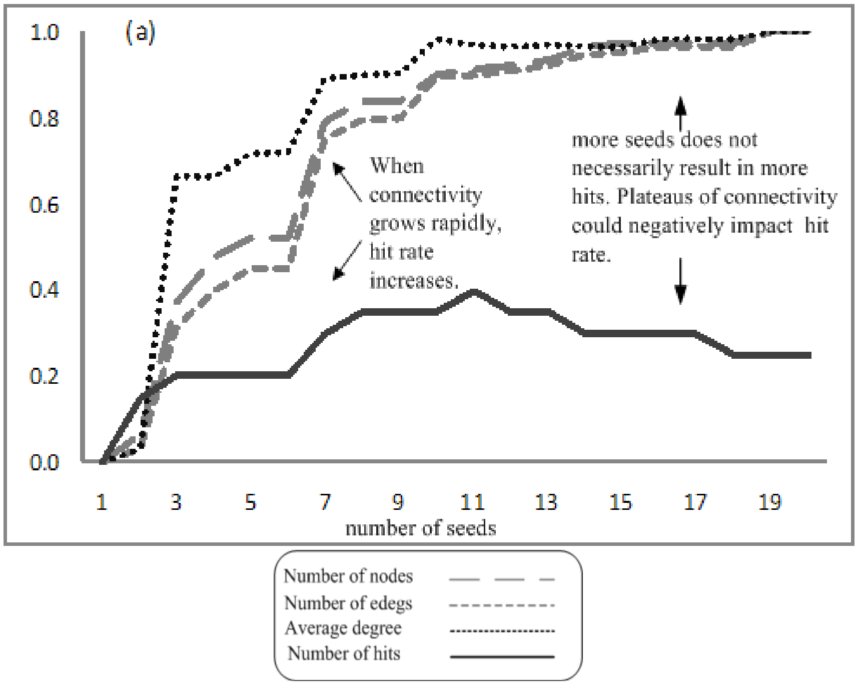


Figure 17. A typical chart generated by one of the random trials

Since adding more seeds does not necessarily improve hit rate, we desire to study whether the order in which the seeds are added will make a difference. We compared the charts where the seeds were added in random order, with a special case where the seeds were added in a specific order. We extracted the number of patients who reported each MFR term from the community reports and used it as an indication of the seed's prevalence. When building G1-G20 in the special case, we added the most prevalent seed first. The chart of the special case is shown in Figure 18. We noticed that the average degree increases more rapidly and a high hit rate is achieved with smaller sub graphs.

The hypothesis is that prevalent terms in patient communities are more relevant to their disease and consequently more valuable seeds. Based on the observations, we designed a systematic approach to optimize the seeds selection. A sub graph should begin with the most relevant seeds. Assuming the connectivity of a sub graph G is a function of its seeds $y = f(x)$, any additional seed should be selected such that G could achieve higher $f'(x)$ with less nodes and edges. When $f'(x)$ is approaching 0 after adding x_{n+1}. It is possible that $\{x_1 \ldots x_n\}$ may be the optimal set of seeds and adding more seeds could reduce the hit rate. When we do not know which nodes are more relevant, multiple optimization experiment sets are needed. Each set follows above process with the starting seed x1 being randomly selected. After completing all experiments, we may select the set that produces the best $f'(x)$ with the smallest sub graph and the fewest seeds.

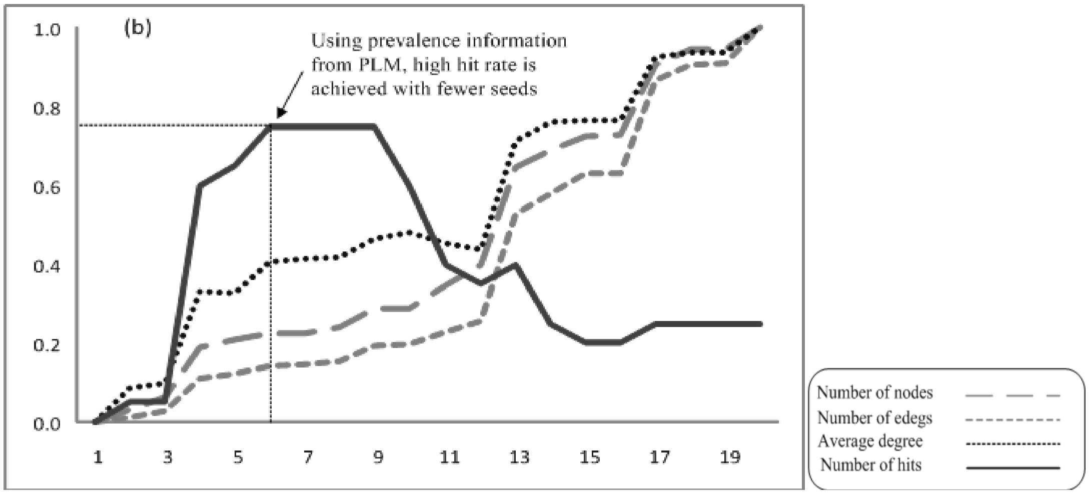


Figure 18. The chart generated by the special trial

However, if we can infer the relevancy of the seeds beforehand, the number of required experiments is greatly reduced. Data in social networking web sites is usually annotated by frequency, customer votes, and other information that can be used to infer the relevancy of a term. HSNC, therefore, may be a promising data source for selecting and optimizing seeds.

Identify ALS Candidate Genes

We followed the optimization process described above and added the seeds in descending order of their prevalence. HyGen has identified fifteen of the twenty gold standard genes (75% hit rate) with the six drugs that are used by the most ALS community members. The enhancement factor computed by equation (4-4) is 94 fold.

$$\frac{\text{abundance of hits in HyGen result}}{\text{abundance of hits in the full graph}} = \frac{15/_{\text{\# genes in top 5\% of the sub graph}}}{20/_{\text{\# genes in the full graph}}} = 94 \qquad (4\text{-}4)$$

Besides identifying the genes in the gold standard, HyGen has also suggested ALS genes that have not been published. Those genes are listed in Appendix F. Each gene has been suggested because it has a hidden connection to at least one of the seeds. For example, MTHFR has been highlighted based on the following reasoning: genetic variation in MTHFR influences a person’s susceptibility to acute leukemia according to Genetic Association Database (GAD); and acute leukemia may be treated by riluzole according to DrugBank. Since riluzole is also the drug most commonly taken by ALS patients, one possible hypothesis may be that genetic variation in MTHFR could also influence one’s susceptibility to ALS. Based on this suggestion and the informatics evidence, a biologist may further study MTHFR’s potential as a candidate for mutation screening in ALS. On the other hand, a specialist may decide to investigate the usage of MTHFR genetic test as an early diagnostic tool. Furthermore, in a cross-disease research effort, public GWAS data of both ALS and acute leukemia may be integrated to identify the connections and interactions between MTHFR and known ALS genes.

Link Riluzole with Alcoholism

Starting from the optimal seeds, HyGen has identified other interesting associations. For example, HyGen has suggested that riluzole may be relevant to several conditions including anxiety, impulsive disorders and alcohol abuse. It is believed that the pharmacological properties of riluzole include an inhibitory effect on glutamate release. Glutamate system is an important contributor to the pathophysiology of mood and anxiety disorders. Thus it is not surprising that HyGen has connected riluzole with anxiety and impulsive disorders. Such associations are not novel; using riluzole to treat severe mood, anxiety and impulsive disorders was proposed in papers such as [106].

However, the association between alcohol abuse and riluzole is more “interesting” and may inspire novel hypotheses. HyGen has linked riluzole to alcohol abuse by two steps: <riluzole-inhibits-NMDA> and <NMDA-related_to_disorder-alcoholism>. Based on the literature research, HyGen’s suggestion appears to be biologically relevant. Chronic alcohol ingestion increases the binding of glutamate to NMDA receptors. During withdrawal, a rebound activation of these receptors occurs and causes alcohol withdrawal syndrome, such as seizures and delirium tremens [107]. One may hypothesize that by inhibiting NMDA, riluzole could relieve alcohol withdrawal syndrome and reduce alcohol abuse problems. In fact, a recent animal study found that riluzole can selectively reduce alcohol self-administration and reduce the severity of alcohol withdrawal seizures in mice [108]. This study was published after we have already compiled all the associations in the full graph. Therefore, we believe HyGen has identified the connection between riluzole and alcohol abuse independently. It suggests that starting from knowledge embedded in HSNC, HyGen can discover hidden connections and suggest relevant hypotheses.

CHAPTER FIVE: CONCLUSION

Summary of Findings

To address the challenges in translational research, we proposed a hybrid approach and implemented a knowledge discovery framework called HyGen [109, 110]. Modeling life science and health care knowledge as semantic graph allows HyGen to aggregate and connect loosely associated disease and molecular level information into a formal structure. Graph-based pseudo- relevance feedback strategy has been developed to discover and prioritize associations relevant to users' interests. Heuristic and quantitative approaches involving two complex diseases have been used to evaluate the framework. The results demonstrate that starting from knowledge gained in the "bench", HyGen can discover novel connections and suggest possible hypotheses to users at the "bedside", and vice versa.

In Colorectal Cancer (CRC) case studies, we simulated three types of stakeholders: a health care consumer, a health care provider, and a biomedical researcher. A profile was defined for each user; a sub graph (view) was constructed; and novel associations were suggested. Literature research has confirmed that the associations suggested by HyGen are scientifically possible. Comparison with direct search in the original data sources has shown that HyGen can identify additional connections. Topological analysis of the sub graphs has suggested that HyGen can deliver views reflecting the users' preference without distort the nature of the original full graph.

Health Social Network Communities (HSNC) are emerging resources for translational research. In Amyotrophic Lateral Sclerosis (ALS) case studies, we converted the most frequently reported terms in the community report of PatientsLikeMe.com (PLM) into graph nodes using linguistic tools and large biomedical ontologies. By combining content from HSNC ("bedside") and traditional research-oriented databases ("bench"), HyGen has identified fifteen of the twenty gold standard ALS genes. In addition, HyGen has suggested new candidate genes for future investigations, as well as a novel association between riluzole and alcohol abuse.

In the various case studies, we designed and implemented various strategies to overcome the challenges of handling HSNC content. We explored two ways of using HSNC content: 1) use it as empirical data for quantitative evaluation; and 2) use it as the starting point (seed) for hypothesis generate. We observed that adding more seeds does not necessary improve the hit rate. A rapid increase of the graph connectivity often leads to increased hit rate; whereas a plateau may lead to declined hit rate especially in a large sub graph. Based on the observations, we proposed a systematic approach to evaluate the value of seeds and to optimize the selection of seeds.

In conclusion, case studies in both disease areas (CRC and ALS) illustrate that the knowledge model and approach of HyGen can be applied to other diseases using data from both traditional biomedical resources and HSNC. We believe this approach can help to bridge information-silos and to accelerate the communication between the "bench" and the "bedside".

Uniqueness of this Approach

Although concepts and technologies supporting user profiling have been studied by many researchers in the context of retrieving information from the World Wide Web [111, 112], fewer reports have been published on applying user profiling technologies to rank multi-level and cross-disciplinary biomedical data based on graph attributes.

With a few exceptions [48], most existing life science networks (graphs) have few types of nodes. This approach integrates many types of nodes and discovers associations among different types of biomedical entities.

Currently no studies have been published on the potentials and challenges in using HSNC data in translational research. With over twenty large HSNC being launched in the last few years [113], there is an increasing need for novel approaches and methods to leverage HSNC content. We explored various ways of using HSNC content in HyGen. To overcome the challenges in handling HSNC content, we developed a process for converting patient-generated terms into graph nodes and reducing false negatives by a graph-based pseudo- relevance feedback strategy.

In summary, we believe this approach is unique in three main aspects:

a. Personalized ranking is produced based on profiles using graph algorithms.
b. Graph-based pseudo-relevance feedback strategy is used to refine intermediate results.
c. Online patient community is used as a supplementary data source to traditional research-oriented databases.

Translational research aims to connect basic research at molecular, cell and organism level with clinical practice at individual and population level. By sharing the preliminary results here, we hope to elicit a greater interest within the informatics community in the development of novel methods and systems for translational research.

CHAPTER SIX: DISCUSSION

Implications of HSNC for Translational Research

From 2005 to 2009, participation in social networking sites more than quadrupled [114]. Many people anticipate the rapid changes in the communication landscape will have direct impact on translational research in both knowledge production and dissemination [113, 115]. First, the folksonomy constructed by patients of HSNC can elicit new health care concepts, coding sets, and classifications for translational research. In the future, we may see adoption of consumer-professional vocabulary mapping technologies by translational research systems. Certain translational approaches may need to completely replace the research-driven taxonomy with consumer-drive folksonomy.

Second, there are large cohorts of patients who share similar health conditions and come from diverse background in HSNC. Access to such individual-level data is critical for identifying disease causing factors and for translating biological knowledge into personalized medicine. Openness may become an emerging theme in translational research community. From technical perspectives, transparency, interoperability, open source, and open programming interfaces will become important design philosophies for translational research systems. From social perspectives, individuals are more open in sharing information when seeking solutions for their health problems [27]. Therefore, translational research community needs to increase the effort in engaging and empowering health care consumers. Historically translational research tools were designed for professionals. In the future, we may see growing demands for consumer-centric translational research tools and services, especially from HSNC.

Thirdly, the collective wisdom of patients and professionals in HSNC will contribute to the knowledge of treating diseases, as well as preventing them. Historically, translational studies concentrated on developing therapies; the focus of future translational research may shift toward earlier stages of health care cycle. For example, translational researcher may be able to identify risk factors and predict health outcome specific for an individual by analyzing the member profiles and the corresponding health-related information in HSNC.

Safety Concerns of Patient-driven Research Model

One concern of the patient-driven research model is that some patients, desperately wanting to cure their illness, may do harm to themselves by overly aggressive self-testing. The associations discovered by HyGen or other knowledge discovery tools are intended to inspiring novel hypotheses. Those hypotheses should be examined and tested under the supervision of experts. Without the proper guidance from medical professionals, health care consumers may possibly do damage to self and others. Unguided self-testing is especially dangerous when the drugs have a high potential for toxicity. Here we caution health care consumers in taking drugs for off-label usage. In the case of self-testing drugs and treatments, evidence and rigorous data analysis are necessary to ensure the safety and effectiveness of the therapies.

The legal and ethical concerns related with patient-driven research also need to be addressed at social level. We suggest that those issues should be examined in a separate and more comprehensive study.

Limitations and Future Research

Optimize Ranking Criteria

One of the limitations of this study is the selection of the ranking criteria. The criteria currently used by HyGen were selected based on experience and literature review [116, 117]. Assuming an association discovered by HyGen contains a certain piece of information. We measure seven aspects of each piece of information: how relevant is the information; how specific is the information; how important is the source of the information; how fresh is the information; whether the information is directly or indirectly associated with the search terms; how rare is the concepts involved in this piece of information; and whether this piece of information occupies a central position in the knowledge space. Some of those criteria may not be necessary. Some criteria may even prevent HyGen from obtaining the optimal ranking. On the other hand, criteria that can provide better rankings may have been missed from the list of seven criteria. Therefore, the current ranking produced by HyGen may not be optimal.

To obtain the optimal set of criteria, we need to develop sets of ranked associations based on different user profiles and use them as the training sets or gold standards. We can then try different combinations of the criteria, compare the results with each gold standard, and identify the optimal criteria set for each type of profile. When starting a search, the user's profile will be populated with an optimal criteria set pre-tailored for the given user type. A guideline should also be provided indicating which set of criteria is suitable for the types of search questions the user wants to address.

To establish optimal criteria for different stakeholders and to reduce personal bias, a user study of large number of pilot users, preferably consumers and professionals interested in multiple disease areas, is needed. We may pursue this study in the future; however it is outside the scope of the current dissertation.

Edge Properties

HyGen's graphs are matrix-based graphs. The edges have only one property: a numerical weight assigned according to the data sources. In the future, to reflect the complex relationships exist in life science and health care knowledge space, we should enable property-based graphs, where properties can be attached to the edges. In property-based graphs, edges can have any number of key-value pairs as their properties. Today, we analyze the graphs based on their topological features. Having edge properties will give us the ability to analyze the graph using property-based algorithm.

Another advantage of a property-based graph is in delivering more user-specific views. In a property-based graph, two neighboring nodes may be connected by multiple edges. In other words, any types of relationships (edges) can exist between the same pair of concept. Some relationships may be present in one view but not in the other based on the different user profiles.

Individual-level Data

Another limitation of the approach is that HyGen used only aggregated community-level data provided by PLM. The major challenge in using individual-level data is to keep sensitive patient data private while preserving the connections between data points that are necessary for association mining. As the open research model of HSNC continues to develop, it may be possible to probe the backend databases for associations at the individual-level using social science methods.

In the future, we need to address the quality shortcomings of patent-generated data, such as potential bias, input errors, intentional or unintentional false information. In this document, we discussed HyGen's technical strategies to alleviate some of the concerns, such as using biomedical ontologies to insure data consistence, using community-level frequency count to minimize individual bias and data error, and using pseudo-relevance feedback to reduce false associations. However, not all above concerns can be overcome by technology. Legal and ethical concerns, intentional false information, and other professionalism issues largely have to be addressed by the combined efforts from patients, researchers and health care providers.

Agent-based Information Extraction

HyGen's current data extraction layer is not fully automated. We propose the use of intelligent mining agents in the future. A mining agent extracts data from a data source and converts the information into triples $\{c_1, r_{12}, c_2\}$. New triples that have been validated will be posted on a "blackboard" to share with other mining agents. Each agent decides whether the new triples are relevant and takes actions accordingly, for example one agent may decide to retrieve additional data, and another agent may decide to re-evaluate some conflicting data. There are two key requirements for the agents: a) they shall be able to adjust the discovery process as more information becomes available; and b) they shall be able to influence the discovery process of each other.

The major challenge of designing an agent-based information extraction layer for HyGen is caused by the complex and transient interrelationships among biomedical data sources. Agents need to know the rules for processing the information and rules for interacting with each other. Some data sources publish the metadata that can be used to derive processing rules. There are also techniques for deriving metadata based on the data in the data sources [118, 119]. However, automatic generation of interaction rules for agents attached to diverse data sources is still a difficult problem [120]. In the case of HyGen, the interaction rules are especially complex due to the heterogeneous nature of translational data sources.

APPENDICES

Appendix A. Associations in the Full Graph

Associations	Count	Source Database
gene and CF[2]	150292	OMIM (www.ncbi.nlm.nih.gov) GAD (http://geneticassociationdb.nih.gov) PHARMAGKB (www.pharmgkb.org)
gene and gene	310842	BioGrid (thebiogrid.org) BIND (bond.unleashedinformatics.com) MINT (mint.bio.uniroma2.it) IntAct (www.ebi.ac.uk/intact) Reactome (reactome.org)
gene and pathway	91771	KEGG (www.genome.jp/kegg) Reactome (www.reactome.org) WikiPathways (www.wikipathways.org) Panther (www.pantherdb.org) PID (pid.nci.nih.gov)
drug and gene	6552	DrugBank (www.drugbank.ca) PharmGKB (www.pharmgkb.org)
drug and CF	6742	DrugBank (www.drugbank.ca) PharmGKB (www.pharmgkb.org)

[2] CF stands for Clinical Features

Appendix B. Major Stakeholders and Profiles

Stakeholder	Sample Query	Preference

Strategic or Portfolio Manager	• Therapeutic focuses of other companies
	• Similar treatment developed by other companies
	• What subset of the population is most likely to have a success outcome from this treatment?
	• What subset of the population is most likely to have an adverse event from this treatment?
	• Who are the opinion leaders for a therapeutic area/drug/pathway?
		X[3] = 1
		Long[4] = 0
		Rare[5] = 0
		Hotspot[6] = 1
		New[7] = 1
		R[8] = Company, Drugs, Diseases, Regulators, Payers, Patents, Market

Immunologist	• Which immunization regime delivers the most antibodies?
	• What process delivers the greatest diversity of antibodies?
		X = 2
		Long = 1
		Rare = 1
		Hotspot = 1
		New = 1
		R = Antibodies, immune response, molecule (large)

Cheminforma tic-ians	• What molecules are active against this target?
	• Related targets
	• Compounds that are related to an active compounds
	• Liabilities associated with a compound
		X = 3
		Long = 1
		Rare = 1
		Hotspot = 1
		New = 0
		R = Molecular Structure, Activity, Target, Pathway, Assay

Systems Physiologist	• Function of a target (gene/protein/phenotype)
	• Publications on a target
	• Who are the experts on this topic?
	• Do patients have variations in the drug target?
	• What hypotheses have been made about this gene?
	• What are the changes in the sequence of a miRNA and variations in the miRNA target region of a transcript?
		X = 2
		Long = 0
		Rare = 1
		Hotspot = 1
		New = 1
		R = Receptors, Proteins, Genomic Sequences, Pathway,
		Biological System

Cellular and Molecular Biologist	• Interactions for X enzymes that are involved in Y disease. For all these genes, get the expression and aCGH values for all disease samples
	• How do variations in the sequences of genes in the pathway X correlate with the extent of disease severity, vulnerability and familial predisposition to Disease Q?
	• What targets are associated with a disease?
	• Active compounds that affect a target
	• Epigenic regulators for a target
	• Demographics associated with a disease
	• How does the gene variant affect patient survival for this disease?
		X = 2
		Long = 1
		Rare = 1
		Hotspot = 1
		New = 1
		R = Disease, Pathway, Proteins and Genes Associated with Disease

[3] X is a number, indicating the maximum number of iteration
[4] 1 indicates the user prefers long associations, 0 otherwise
[5] 1 indicates the user prefers rare associations, 0 otherwise
[6] 1 indicates the user prefers highly connected entities, 0 otherwise
[7] 1 indicates the user prefers new information, 0 otherwise
[8] R is a list of concepts that the user is interested in.

	• What is the underlying aspect of the profile that distinguishes high and low risk patients (pathway)? • What variations are there in disease tissue versus germ line genes?	

Medicinal Chemist	• What classes of compounds appear to have activity?
	• What is the patent landscape?
		X = 1
		Long = 0
		Rare = 0
		Hotspot = 1
		New = 1
		R = Assay, Activity, Fragment, Synthesis, chemical properties, polypharmacology

In-Vitro Biologist	• Cell lines in which RNAi data has been generated using X reagents
	• Experiments conducted on a target
	• Do proteomic assays of tumor material and serum samples identify patterns reflecting outcomes?
		X = 1
		Long = 0
		Rare = 0
		Hotspot = 1
		New = 1
		R = Bioassay, pathway

In-Vivo Biologist	• What variations in metabolites correlate with the efficacy of compounds against Targets
	• Which animal species has the closest genome for the pathway/target of this disease?
	• Safety concerns for a compound
	• Markers for clinical assays
		X = 1
		Long = 0
		Rare = 1
		Hotspot = 1
		New = 1
		R = Toxicity, Efficacy, animal study

Clinical Trial Formulator or Lead Physician	• Which patients are most likely to respond in a clinical trial?
	• Other companies running clinical trials in the same therapeutic area
	• Known issues in this type of formulation in a drug class
	• Known side effects in therapies for this target?
	• Are there risk factors associated with the drug that should be taken into consideration (genetic, environmental)?
	• What is the treatment regime for this drug?
	• What are the clinical care guidelines for this drug?
	• Are there longitudinal studies available?
	• Do patients have variations in the drug target itself?
		X = 2
		Long = 0
		Rare = 1
		Hotspot = 1
		New = 1
		R = Hypothesis, disease, biomarker,
		Clinical Trials, Interventions

Sales and Marketing	• Who are the opinion leaders I need to influence?
	• What physicians should I visit?
	• What media should we use for advertising?
	• Should we target internet based advertising?
		X = 3
		Long = 1
		Rare = 0
		Hotspot = 1
		New = 1
		R = Physician, disease, market

Primary Care Clinician	• What are the alternative names for the disease/condition/ symptom?
	• What are the diagnostic criteria for the disease?
	• Treatment history
	• Difference between treatments
	• What issues have been seen in this type of
		X = 2
		Long = 0
		Rare = 0
		Hotspot = 1
		New = 1
		R = Patient, Patient

	formulation in this drug class? • What subpopulation demographics are associated with this disease? • Which population is most likely to respond to this therapy? • Are there risk factors associated with the drug that should be taken into consideration (genetic, environmental)? • What are the current hot research topic areas for disease x? • Are there tests I need to perform before prescribing this drug? • What is the standard disease progression? • What is the likely disease progression for this patient? • What is the predicted outcome for the patient? • How quickly will this patient metabolize the drug? • Is it likely that the patient will experience a recurrence? • Where is a patient on a complex, multi-dimensional risk spectrum based on detailed, individual molecular characteristics at genomics scale? • Are there natural alternatives to this drug? • What lifestyle changes should I recommend? • Is there a combination of drugs that would work best?	Diagnosis, Disease Symptoms, Referral, Treatment/Management Plan

Provider	• What symptoms or test would show that the patient does not have the disease/condition?
	• How is treatment A different from B?
	• Is this course of treatment economical?
		X = 1
		Long = 0
		Rare = 0
		Hotspot = 1
		New = 1
		R = Cost/benefit of Therapy, Differential Diagnosis, Prognosis for Patient

Patient	• Treatment options for disease X
	• What symptoms or tests are related to disease X?
	• Am I at high risk because of my family history, lifestyle, and condition?
	• What lifestyle changes should I make
		X = 2
		Long = 1
		Rare = 0
		Hotspot = 1
		New = 1
		R = Cost/benefit of Therapy, Diagnosis, symptoms, disease, risk factors, treatment

Appendix C. A Sample User Profile

```
# start the search with the following entities/terms
seeds.DRG.drugName=Riluzole||Baclofen||Lithium||Amitriptyline||Gabapent
in||Lorazepam||Zolpidem||Sertraline||Salbutamol||Citalopram
# what type of 'things' are you interested in
dataTypeOfInterest=Gene
# what centrality algorithm(s) do you prefer?
scoreTypes=PageRank,Degree,avgDistanceToSeeds
# how to normalize the graph attributes
PageRank.max=0.25
PageRank.min=0.0001
PageRank.shape=sigmUp
PageRank.convert=none
PageRank.weight=1
Degree.max=100
Degree.min=4
Degree.shape=sigmUp
Degree.convert=none
Degree.weight=2
DistanceCentrality.max=10
DistanceCentrality.min=0
DistanceCentrality.shape=sigmUp
DistanceCentrality.convert=none
DistanceCentrality.weight=1
avgDistanceToSeeds.max=5
avgDistanceToSeeds.min=3
avgDistanceToSeeds.shape=bell
avgDistanceToSeeds.convert=none
avgDistanceToSeeds.weight=2
# do you prefer specific associations
Specific=No
# discards any nodes (entities) that are connected to more than
MaxDegree number of other nodes?
MaxDegree=100
MinDegreeScore=0.3
# do you prefer fresh information
Fresh=1
# how much do you trust the sources (0 least trust, 1 most trust)
OMIM.confidence=1
KEGG.confidence=.9
DrugBank.confidence=.4
GKB.confidence=.4
Reactome.confidence=.8
GAD.confidence=.5
NCI_Nature.confidence=1
IPA.confidence=.7
DrugMatrix.confidence=.7
BioGRID.confidence=.2
GeneGo.confidence=.7
Wiki.confidence=.8
Panther.confidence=.8
Biobase.confidence=.7
BIND.confidence=.7
```

Appendix D. Ranking Criteria

	Meaning	**Value**
C_{SA}	how relevant is *v*	if *v is_a* R, $C_{SA} = 1$; else $C_{SA} = 0$
S_{SA}	how specific is *v*, e.g. "Epithelial Neoplasm" is more specific than "Neoplasm"	$S_{SA} = H_v / H$, where H_v is the position of *v* in the domain ontology and H is the total height of the branch where *v* is found
T_{SA}	how important are the sources	average source weights of all edges.
F_{SA}	how fresh is the information	the time difference between now and when *v* first appeared in its source
L_{SA}	how close is *v* to the seeds	average shortest path from *v* to the seeds
R_{SA}	how rare is *v*	$R_{SA} = \|N\| / \|M\|$ N = number of nodes that are of the same semantic type as *v*. M = number of nodes in the full graph
P_{SA}	*v*'s probability of being reached from the seeds	PageRank with Priors was computed by the following iterative equation [121]: $\pi(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(v \mid u)\pi^{(i)}(u)\right) + \beta P_v$ p_v was set to 1/|R| for all seeds and 0 for the rest. 0< β <1, $d_{in}(v)$ is the in-degree of a *v*. p(*v*|*u*) is the probability of reaching *v* from another node *u*. We assigned $\pi(v)^0$ to 1 for all nodes in the first iteration. HITS with Priors was computed by the following iterative equation [121]: $a^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} \frac{h^{(t)}(u)}{H^{(i)}}\right) + \beta P_v, \; and \; h^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{out}(v)} \frac{a^{(t)}(u)}{A^{(i)}}\right) + \beta P_v$ $H^{(i)} = \sum_{v=1}^{\|V\|} \sum_{u=1}^{d_{in}(v)} h^{(i)}(u), \; and \; A^{(i)} = \sum_{v=1}^{\|V\|} \sum_{u=1}^{d_{out}(v)} a^{(i)}(u)$ $d_{in}(v)$ and $d_{out}(v)$ are the in-degree and out-degree of *v*. We assigned a^0 and h^0 to 1 in the first iteration.

Appendix E. Symptoms Reported by ALS Patients

MFR[9] symptom	UMLS concept	Concept Id	UMLS type
Fatigue	Actual Fatigue	C2364051	Finding
	Fatigue	C0015672	SS[10]
	Fatigue	C2024893	Finding
Fasciculations	Muscular fasciculation	C0015644	SS
Stiffness/Spasticity	Stiffness	C0427008	SS
	Muscle Spasticity	C0026838	SS
Anxiety	Anxiety symptoms	C0860603	Finding
Emotional lability	Mood swings	C0085633	MBD[11]
Excess saliva	Sialorrhea	C0037036	DS[12]
Depression	Actual Depression	C2364072	Finding
	Depressed symptom	C1579931	SS
	Depressed mood	C0344315	Finding
	Depressive disorder	C0011581	MBD
	Depressive episode, unspecified	C0349217	MBD
	Mental Depression	C0011570	MBD
Pain	Actual Pain	C2364139	Finding
	Pain	C0030193	SS
Insomnia	Sleeplessness	C0917801	SS
Constipation	Constipation	C0009806	SS

[9] MFR = most frequently reported

[10] SS = Sign or Symptom

[11] MBD = Mental or Behavioral Dysfunction

[12] DS = Disease or Syndrome

Appendix F. Candidate Genes Suggested by HyGen

Gene	Reason for the suggestion[13]

MTHFR	MTHFR <-relatedToDisease-> Acute leukemia {GAD}
	Acute leukemia <-treatedBy-> **Riluzole** {DrugBank}

ALB	ALB<-isTargetOf->Nortriptyline {DrugBank}
	Nortriptyline<-treats->**Depressive disorder** {DrugBank}

ADRA1A	ADRA1A<-isTargetOf->Maprotiline {DrugBank}
	Maprotiline<-treats->**Depressive disorder** {DrugBank}

TNF	TNF<-relatedToDisease->AMYLOIDOSIS {GAD}
	AMYLOIDOSIS<-treatedBy->**Riluzole** {DrugBank}

TP53	TP53<-relatedToDisease->Acute leukemia {OMIM}
	Acute leukemia <-treatedBy->**Riluzole** {DrugBank}

ABCB1	ABCB1<-relatedToDisease-> Disorder, Bipolar {GAD}
	Disorder, Bipolar <-treatedBy-> **Amitriptyline** {DrugBank}

GRIN1	GRIN1<-interactsWith->DRD1{BKL_Proteome}
	DRD1<-relatedToDisorder->**Depressive disorder** {GAD}

CDKN2A	CDKN2A <-relatedToDisease-> Acute leukemia {OMIM}
	Acute leukemia <-treatedBy-> **Riluzole** {DrugBank}

HRAS	HRAS <-relatedToDisease-> Acute leukemia {OMIM}
	Acute leukemia <-treatedBy-> **Riluzole** {DrugBank}

BAALC	BAALC <-relatedToDisease-> Acute leukemia {OMIM}
	Acute leukemia <-treatedBy-> **Riluzole** {DrugBank}

ZFYVE26	ZFYVE26<-relatedToDisease-> clonus {OMIM}
	clonus <-treatedBy-> **Baclofen** {DrugBank}

[13] Each row corresponds to an association. The seeds are in bold font. Data sources are inside curly brackets and relationships are inside angle brackets.

REFERENCES

1. Philip Pizzo, Letter from the dean. Standford Medicine Magazine, 2002. 19(3): p. 1-2.
2. Dauphinee, D. and Martin, J.B., Breaking down the walls: Thoughts on the scholarship of integration. Journal of Medical Education, 2000. 75(9): p. 881-886.
3. Gaughan, A., Bridging the divide: The need for translational informatics. Pharmacogenomics, 2006 7(1): p. 117-122.
4. Ruttenberg, A., et al., Advancing translational research with the semantic web. BMC Bioinformatics, 2007 8(3): p. 1-2.
5. Refolo, L., et al., A cholesterol-lowering drug reduces beta-amyloid pathology in a transgenic mouse model of alzheimer disease. Neurobiol Dis, 2001. 8(5): p. 890-899.
6. Blain, J.-F. and Poirier, J., Cholesterol homeostasis and the pathophysiology of alzheimer's disease. Expert Review of Neurotherapeutics, 2004. 4(5): p. 823-829.
7. Sparks, D.L., et al., Atorvastatin for the treatment of mild to moderate alzheimer disease. Arch Neurol, 2005. 62(1): p. 753-757.
8. Qu, X., et al. Semantic web-based data representation and reasoning applied to disease mechanism and pharmacology. in IEEE International Conference on Bioinformatics & Biomedicine. 2007. Freemont, USA.
9. Beek, J., et al., Activation of complement in the central nervous system. Ann of the New York Academy of Sciences, 2003. 992(1): p. 56-71.
10. Xiong, Z.-Q., et al., Formation of complement membrane attack complex in mammalian cerebral cortex evokes seizures and neurodegeneration. The Journal of Neuroscience, 2003. 23(3): p. 955-956.
11. Ozdinler, P.H. and Erzurumlu, R.S., Slit2, a branching-arborization factor for sensory axons in the mammalian cns. The Journal of Neuroscience, 2002. 22(11): p. 4540-4549.
12. Yeo, S.-Y., et al., Involvement of islet-2 in the slit signaling for axonal branching and defasciculation of the sensory neurons in embryonic zebrafish. Mechanisms of Development, 2004. 121(4): p. 315-324.
13. Xu, H., et al., Calcium signaling in chemorepellant slit2-dependent regulation of neuronal migration Proceedings of the National Academy of Sciences, 2004. 101(12): p. 4296-4301
14. Vittorio, C., et al., Gene expression profiling of 12633 genes in alzheimer hippocampal ca1: Transcription and neurotrophic factor down-regulation and up-regulation of apoptotic and pro-inflammatory signaling. Journal of Neuroscience Research, 2002. 70(3): p. 462-473.
15. Shirvan, A., et al., Molecular imaging of neurodegeneration by a novel cross-disease biomarker. Experimental Neurology, 2009. 219(1): p. 274-283.
16. Royall, D.R., et al., Executive control function: A review of its promise and challenges for clinical research. A report from the committee on research of the american neuropsychiatric association. J Neuropsychiatry Clin Neurosci, 2002. 14(4): p. 377-405.
17. Levy, Y., et al., Therapeutic potential of neurotrophic factors in neurodegenerative diseases. BioDrugs, 2005. 19(2): p. 97-127.

18. Planells-Cases, R., et al., Pharmacological intervention at ionotropic glutamate receptor complexes. Curr Pharm Des, 2006. 12(28): p. 3583-3596.
19. Sauer, S.W., et al., Bioenergetics in glutaryl-coenzyme a dehydrogenase deficiency: A role for glutaryl-coenzyme a. J Biol Chem, 2005. 280(23): p. 21830-21836.
20. Dawson, T.M. and Dawson, V.L., Molecular pathways of neurodegeneration in parkinson disease. Science, 2003. 302(5646): p. 819-822.
21. Bertram, L. and Tanzi, R.E., The genetic epidemiology of neurodegenerative disease. J Clin Invest, 2005. 115(6): p. 1449-1457.
22. Novacek, V., et al. Dynamic integration of medical ontologies in large scale. in Health Care and Life Sciences Data Integration for the Semantic Web. 2007. Banff, Canada.
23. Topaloglon, T. Biological data management: Research, practice and opportunities. in 30th VLDB Conference. 2004. Toronto, Canada.
24. Dictionary.Com, Random house unabridged dictionary. (accessed in Feb. 2009).
25. California Heathcare Foundation, Clinical data standards explained. 2004.
26. Fornai, F., et al., Lithium delays progression of amyotrophic lateral sclerosis. PNAS, 2008. 105(6): p. 2052-2057.
27. Allison, M., Can web 2.0 reboot clinical trials? Nat Biotech, 2009. 27(10): p. 895-902.
28. Smith, C.A. and Wicks, P.J. Patientslikeme: Consumer health vocabulary as a folksonomy. in AMIA Annual Symposium. 2008. Washington, D.C., USA.
29. Mukherjea, S., Information retrieval and knowledge discovery utilising a biomedical semantic web. Briefings in Bioinformatics, 2005 6(3): p. 252-262.
30. Robu, I., et al., An introduction to the semantic web for health sciences librarians. Journal of the Medical Library Association, 2006. 94(2): p. 198-205.
31. Gibbons, A., Algorithmic graph theory. 1985: Cambridge University Press.
32. Braun, P., et al., Networking metabolites and diseases. Proceedings of the National Academy of Sciences, 2008. 105(29): p. 9849-9850.
33. Hopkins, A.L., Network pharmacology. Nat Chem Biol, 2007. 4(11): p. 682-690.
34. Feldman, I., et al., Network properties of genes harbouring inherited disease mutations. Proceedings of the National Academy of Sciences, 2008. 105(11): p. 4323-4328.
35. Yildirim, M.A., et al., Drug-target network. Nat Biotech, 2007. 25(10): p. 1119-1126.
36. Maayan, A., et al., Network analysis of fda approved drugs and their targets. Mt Sinai J Med, 2007. 74(1): p. 27-32.
37. Lage, K., et al., A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotech, 2007. 25(3): p. 309.
38. Xu, J. and Li, Y., Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics, 2006. 22(22): p. 2800-2805.
39. Lee, D.S., et al., The implications of human metabolic network topology for disease comorbidity. Proceedings of the National Academy of Sciences, 2008. 105(29): p. 9880-9885.
40. Assenov, Y., et al., Computing topological parameters of biological networks. Bioinformatics, 2008. 24(2): p. 282-284.

41. Chen, H., et al., Semantic web for integrated network analysis in biomedicine. Briefings in Bioinformatics, 2009. 10(2): p. 177-192.
42. Ding, L., et al. Finding and ranking knowledge on the semantic web. in 4th International Semantic Web Conference. 2005. Galway, Ireland.
43. Chen, J., et al., Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics, 2009. 10(73): p. 1-2.
44. Harith, A. and Christopher, B., Ontology ranking based on the analysis of concept structures, in Proceedings of the 3rd international conference on Knowledge capture. 2005: Banff, Canada.
45. Scardoni, G. and Laudanna, C., Analyzing biological network parameters with centiscape. Bioinformatics, 2009. 25(21): p. 2857-2859.
46. White, S. and Smyth, P. Algorithms for estimating relative importance in networks. in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003. Washington, D.C., USA.
47. Kleinberg, J.M. Authoritative sources in a hyperlinked environment. in Proceedings of the ninth ACM-SIAM symposium on Discrete algorithms. 1998. San Francisco, USA.
48. Gudivada, R.C., et al., Identifying disease-causal genes using semantic web-based representation of integrated genomic and phenomic knowledge. Journal of Biomedical Informatics, 2008. 41(5): p. 717-729.
49. Lahlou, A. and Urien, P. Sim-filter: User profile based smart information filtering and personalization in smartcard. in 15th Conference on Advanced Information Systems Engineering. 2003. Velden, Austria.
50. Susan Gauch, et al., Ontology-based user profiles for search and browsing Web Intelligence and Agent Systems, 2003. 1(3): p. 219-234.
51. Hettne, K.M., et al., A dictionary to identify small molecules and drugs in free text. Bioinformatics, 2009. 25(22): p. 2983-2991.
52. Goh, K.-I., et al., The human disease network. Proceedings of the National Academy of Sciences, 2007. 104(21): p. 8685-8690.
53. Campillos, M., et al., Drug target identification using side-effect similarity. Science, 2008. 321(5886): p. 263-266.
54. Kwiatkowski, T.J., Jr., et al., Mutations in the fus/tls gene on chromosome 16 cause familial amyotrophic lateral sclerosis. Science, 2009. 323(5918): p. 1205-1208.
55. Denney, C., et al. Creating a translational medicine ontology. in Proceedings of the First International Conference on Biomedical Ontology. 2009. Buffalo, USA.
56. Markowitz, S.D. and Bertagnolli, M.M., Molecular basis of colorectal cancer. N Engl J Med, 2009. 361(25): p. 2449-2460.
57. Boland, C., Molecular basis for stool-based DNA tests for colorectal cancer: A primer for clinicians. Rev Gastroenterol Disord, 2002. 1(1): p. 12-19.
58. De Roock, W., et al., Clinical biomarkers in oncology: Focus on colorectal cancer. Mol Diagn Ther, 2009. 13(2): p. 103-114.
59. Cassidy, J., et al., First-line oral capecitabine therapy in metastatic colorectal cancer: A favorable safety profile compared with intravenous 5-fluorouracil/leucovorin. Ann Oncol, 2002. 13(4): p. 566-575.

60. Javle, M. and Hsueh, C.-T., Updates in gastrointestinal oncology - insights from the 2008 44th annual meeting of the american society of clinical oncology. Journal of Hematology & Oncology, 2009. 2(1): p. 9-10.
61. Galmarini, C.M., et al., Polymeric nanogels containing the triphosphate form of cytotoxic nucleoside analogues show antitumor activity against breast and colorectal cancer cell lines. Mol Cancer Ther, 2008. 7(10): p. 3373-3380.
62. Balendiran, G., Fibrates in the chemical action of daunorubicin. Curr Cancer Drug Targets, 2009 9(3): p. 366-369.
63. Kurteva, G., et al., Different chemotherapy schedules in metastatic colorectal cancer: A response and survival analysis. J BUON, 2002 7(2): p. 121-125.
64. Coss, A., et al., Increased topoisomerase iia expression in colorectal cancer is associated with advanced disease and chemotherapeutic resistance via inhibition of apoptosis. Cancer Letters, 2009. 276(2): p. 228-238.
65. Drevs, J., et al., Antiangiogenic potency of various chemotherapeutic drugs for metronomic chemotherapy. Anticancer Res, 2004. 24(3): p. 1759-1764.
66. Andrews, J.M., et al., Systematic review: Does concurrent therapy with 5-asa and immunomodulators in inflammatory bowel disease improve outcomes? Aliment Pharmacol Ther, 2009. 29(5): p. 459-469.
67. Hubner, R.A. and Houlston, R.S., Folate and colorectal cancer prevention. Br J Cancer, 2008. 100(2): p. 233-239.
68. Pitkin, R.M., Folate and neural tube defects. Am J Clin Nutr, 2007. 85(1): p. 285-288.
69. Ray, J.G., et al., Vitamin b12 and the risk of neural tube defects in a folic-acid-fortified population. Epidemiology, 2007. 18(3): p. 362-366.
70. Guerreiro, C.S., et al., Risk of colorectal cancer associated with the c677t polymorphism in 5,10-methylenetetrahydrofolate reductase in portuguese patients depends on the intake of methyl-donor nutrients. Am J Clin Nutr, 2008. 88(5): p. 1413-1418.
71. Duthie, S.J., Folic acid deficiency and cancer: Mechanisms of DNA instability. Br Med Bull, 1999. 55(3): p. 578-592.
72. Arimura, K., Rippling muscle syndrome. Internal Medicine, 2002. 41(5): p. 325-326.
73. Steven, A.G., Acquired rippling muscle disease with myasthenia gravis. Muscle & Nerve, 2004. 29(1): p. 143-146.
74. Muller-Felber, W., et al., Immunosuppressive treatment of rippling muscles in patients with myasthenia gravis. Neuromuscular Disorders, 1999. 9(8): p. 604-607.
75. Takagi, A., et al., Rippling muscle syndrome preceding malignant lymphoma. Internal Medicine, 2002. 41(2): p. 147-150.
76. Ditzel, H., Human antibodies in cancer and autoimmune disease. Immunologic Research, 2000. 21(2): p. 185-193.
77. Akira, B.M., et al., Autoimmune hemolytic anemia associated with colon cancer. Cancer, 1974. 33(1): p. 111-114.
78. Lettice, L.A. and Hill, R.E., Preaxial polydactyly: A model for defective long-range regulation in congenital abnormalities. Current Opinion in Genetics & Development, 2005. 15(3): p. 294-300.

79. Temtamy, S.A., et al., Expanding the phenotypic spectrum of the baller-gerold syndrome. Genet Couns, 2003. 14(3): p. 299-312.
80. Radhakrishna, U., et al., An autosomal dominant triphalangeal thumb: Polysyndactyly syndrome with variable expression in a large indian family maps to 7q36. Am J Med Genet, 1996. 66(2): p. 209-215.
81. Tsukurov, O., et al., A complex bilateral polysyndactyly disease locus maps to chromosome 7q36. Nature Genetics, 1994. 6(3): p. 282-286.
82. Zguricas, J., et al., Clinical and genetic studies on 12 preaxial polydactyly families and refinement of the localisation of the gene responsible to a 1.9 cm region on chromosome 7q36. J Med Genet, 1999. 36(1): p. 33-40.
83. Nicolaidoua, P., et al., Vitamin d receptor polymorphisms in hypocalcemic vitamin d-resistant rickets carriers. Horm Res, 2007. 67(4): p. 179-183.
84. Barabasi, A.-L. and Oltvai, Z.N., Network biology: Understanding the cells functional organization. Nature, 2004. 5(2): p. 101-113.
85. Stumpf, M.P.H., et al., Subnets of scale-free networks are not scale-free: Sampling properties of networks. Proceedings of the National Academy of Sciences, 2005. 102(12): p. 4221-4224.
86. Brooks, B.R., et al., El escorial revisited: Revised criteria for the diagnosis of amyotrophic lateral sclerosis. Amyotrophic Lateral Sclerosis, 2000. 1(5): p. 293-299.
87. Garber, K., The elusive als genes. Science, 2008. 319(5859): p. 20.
88. Pasinelli, P. and Brown, R.H., Molecular biology of amyotrophic lateral sclerosis: Insights from genetics. Nat Rev Neurosci, 2006. 7(9): p. 710-723.
89. Beleza-Meireles, A. and Al-Chalabi, A., Genetic studies of amyotrophic lateral sclerosis: Controversies and perspectives Amyotrophic Lateral Sclerosis, 2009 10(1): p. 1-14.
90. Goodall, E.F., et al., Association of the h63d polymorphism in the hemochromatosis gene with sporadic als. Neurology, 2005. 65(6): p. 934-937.
91. Veldink, J.H., et al., Homozygous deletion of the survival motor neuron 2 gene is a prognostic factor in sporadic als. Neurology, 2001. 56(6): p. 749-752.
92. Simpson, C.L. and Al-Chalabi, A., Amyotrophic lateral sclerosis as a complex genetic disease. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2006. 1762(11): p. 973-985.
93. Dupré, N. and Valdmanis, P., Amyotrophic lateral sclerosis: Genome-wide association studies in amyotrophic lateral sclerosis. European Journal of Human Genetics, 2009. 17(1): p. 137-138.
94. Drory, V.E., et al., Association of apoe 4 allele with survival in amyotrophic lateral sclerosis. Journal of the Neurological Sciences, 2001. 190(1-2): p. 17-20.
95. Hadano, S., et al., A gene encoding a putative gtpase regulator is mutated in familial amyotrophic lateral sclerosis 2. Nat Genet, 2001. 29(2): p. 166-173.
96. Schymick, J.C., et al., Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: First stage analysis and public release of data. The Lancet Neurology, 2007. 6(4): p. 322-328.
97. Rosen, D., et al., Mutations in cu/zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature, 1993 (362): p. 59-62.

98. Lambrechts, D., et al., Vegf is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. Nat Genet, 2003 34(4): p. 383-394.

99. Strong, M.J., The evidence for altered rna metabolism in amyotrophic lateral sclerosis (als). Journal of the Neurological Sciences, 2010. 288(1-2): p. 11-12.

100. Corcia, P., et al., Abnormal smn1 gene copy number is a susceptibility factor for amyotrophic lateral sclerosis. Ann Neurol, 2002 51(2): p. 243-246.

101. Van Es, M.A., et al., Genetic variation in dpp6 is associated with susceptibility to amyotrophic lateral sclerosis. Nat Genet, 2008. 40(1): p. 29-31.

102. Cronin, S., et al., A genome-wide association study of sporadic als in a homogenous irish population. Hum Mol Genet, 2008. 17(5): p. 768-774.

103. Van Es, M.A., et al., Itpr2 as a susceptibility gene in sporadic amyotrophic lateral sclerosis: A genome-wide association study. The Lancet Neurology, 2007. 6(10): p. 869-877.

104. Munch, C., et al., Point mutations of the p150 subunit of dynactin (dctn1) gene in als. Neurology, 2004. 63(4): p. 724-726.

105. Crawford, T.O. and Skolasky, R.L., Jr., The relationship of smn to amyotrophic lateral sclerosis. Ann Neurol, 2002. 52(6): p. 857-858.

106. Zarate, C.A. and Manji, H.K., Riluzole in psychiatry: A systematic review of the literature. Expert Opinion on Drug Metabolism & Toxicology, 2008. 4(9): p. 1223-1234.

107. Hendricson, A.W., et al., Aberrant synaptic activation of n-methyl-d-aspartate receptors underlies ethanol withdrawal hyperexcitability. Journal of Pharmacology and Experimental Therapeutics, 2007. 321(1): p. 60-72.

108. Besheer, J., et al., Preclinical evaluation of riluzole: Assessments of ethanol self-administration and ethanol withdrawal symptoms. Alcoholism: Clinical and Experimental Research, 2009. 33(8): p. 1460-1468.

109. Webster, Y.W., et al. A hypbrid method to discover and rank cross-disciplinary associations. in IEEE International Conference on Bioinformatics & Biomedicine. 2009. Washigton D.C., USA.

110. Webster, Y., et al. A framework for cross-disciplinary hypothesis generation. in ACM 25th Symposium on Applied Computing. 2010. Sierre, Switzerlan.

111. Gils, B.V. and Schabell, E.D. User-profiles for information retrieval. in 15th Belgian-Dutch Conference on Artificial Intelligence. 2003. Nijmegen, Netherlands.

112. Arezki, R., et al., Web information retrieval based on user profile in Adaptive hypermedia and adaptive web-based systems. 2004, Springer Berlin. p. 275-278.

113. Swan, M., Emerging patient-driven health care models: An examination of health social networks, consumer personalized medicine and quantified self-tracking. Int J Environ Res Public Health, 2009. 6(2): p. 492-525.

114. Chou, W.-Y.S., et al., Social media use in use in the united states: Implications for health communication. J Med Internet Res, 2009. 11(4): p. 48-50.

115. Eysenbach, G., Medicine 2.0: Social networking, collaboration, participation, apomediation, and openness. J Med Internet Res, 2008. 10(3): p. 5-6.

116. Moskovitch, R., A comparative evaluation of full-text, concept-based, and context-sensitive search. J Am Med Inform Assoc, 2007. 14(2): p. 164-174.

117. Bamba, B. and Mukherjea, S., Utilizing resource importance for ranking semantic web query results in Semantic web and databases. 2005, Springer Berlin. p. 185-198.

118. Bayardo, R.J., Jr., et al., Infosleuth: Agent-based semantic integration of information in open and dynamic environments, in Proceedings of the 1997 ACM SIGMOD international conference on Management of data. 1997: Tucson, USA.

119. Patel, S., et al., Development of agent-based knowledge discovery framework to access data resource grid. Advances in Computational Sciences and Technology, 2010. 3(1): p. 23-31.

120. Walczak, S., Managing personal medical knowledge: Agent-based knowledge acquisition. International Journal of Technology Management 2009. 47(1): p. 22-36.

121. Chen, J., et al., Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics, 2009. 10(1): p. 73.

CURRICULUM VITAE

Yue Wang Webster

Education

Ph.D. in Health Informatics (2005-2010)
Indiana University, Indianapolis, Indiana, USA
M.S. in Electrical & Computer Engineering (2000-2002)
Purdue University, Indianapolis, Indiana, USA
B.S. in Chemistry and Molecule Engineering (1995-1999)
Peking University, Beijing, China

Experience

Ph.D. student, Indiana University-Purdue University Indianapolis (2005-2010)

- Lead the design and implementation of a framework for discovering and prioritizing novel associations
- Conducted case studies in Colorectal Cancer and Amyotrophic Lateral Sclerosis

Associate Consultant, Eli Lilly & Company (2009-Present)
Sr. System Analyst, Eli Lilly & Company (2007-2009)
System Analyst, Eli Lilly & Company (2002-2006)

- Collaborating with multi-disciplinary teams in large-scale discovery and translational research projects
- Actively participating and sometimes leading international collaborations.
- Developing algorithms and implementing small- to mid-sized prototypes to fill immediate informatics need
- Integrating chemical, biological and health outcome data from both external and internal sources to improve decision-making at the early phases of drug discovery
- Working closely with computational chemists and toxicologist to build in-silico models for predicating efficacy and liability of drug candidates

M.S. student and Research Assistant, Indiana University-Purdue University Indianapolis (2000-2002)

- Contributed to the design and development of an biological and chemical data integration system; this system was awarded National Science Foundation funding from 2001 to 2003
- Developed an ontology for integrating biological and chemical data

Skills

- Familiar with drug discovery process
- Service-oriented architecture
- Relational and object oriented databases
- Semantic Web technologies
- 2D/3D modeling and SAR methodology

- Perl, C#, Java, Ruby, JavaScript, SAS
- Pipeline Pilot, Spotfire
- Strong problem-solving skills and strategic thinking.
- Organized, good at self-management and prioritization
- Detail oriented

Publications

- Webster, Y., Gudivada, R., Dow, E., Koehler, J. and Palakal, M. (2009) A Framework for Cross-Disciplinary Hypothesis Generation. ACM 25th Symposium on Applied Computing
- Webster, Y., Gudivada, R., Dow, E., Koehler, J. and Palakal, M. (2009) A Hybrid Method to Discover and Rank Cross-domain Associations. IEEE BIBM conference processing
- Liao, Q., Wang, J., Webster, Y. and Watson, I. (2009) GPU Accelerated Support Vector Machines for Mining High-Throughput Screening Data. J. Chem. Info. Modeling
- Vieth, M., Erickson, J., Wang, J., Webster, Y. Mader, M., Higgs, R., Watson, I. (2009) Kinase Inhibitor Data Modeling and de Novo Inhibitor Design with Fragment Approaches. J. Med. Chem.
- Erickson, J., Mader, M., Watson, I., Webster, Y., Higgs, R., Bell, M., and Vieth, M. Structure-Guided Expansion of Kinase Fragment Libraries Driven by Support-Vector Machine Models. BBA Proteins and Proteomics Journal
- Erickson, J., Watson, I., Higgs, R., Bell, M., Sutherland, J., Mader, M., Webster, Y., and Vieth, M. (2009) Applications of the Fragment Concept in Drug Design. ICSM conference processing
- Zhang, H., Wang, J., Webster, Y., Mahoui, A., and Robertson, D. (2009) SPrime: Computational Chemistry Framework for Drug Discovery. ChemAxon US User Group Meeting
- Ben Miled, Z., Webster, Y., Li, N., and Liu, Y. (2003) An Ontology for Semantic Integration of Life Science Web Databases. Int. J. Cooperative Inf. Syst.
- Ben Miled, Z., Wang, Y., Li, N., Bukhres, O., Martin, J., Nayar, A. and Oppelt, R. (2002) BAO, A Biological and Chemical Ontology for Information Integration. Online J. of Bioinformatics
- Ben Miled, Z., Bukhres, O., Wang, Y., Li, N., Baumgartner, M., and Sipes, B. (2001) Biological and Chemical Information Integration System. Network Tools and Applications in Biology, Genoa, Italy

Honors

Eli Lilly Solution Achievement Award 2008
Honored speaker at Women in Engineering and Computing Conference 2004
Eli Lilly Solution Achievement Award 2003
1st Place IUPUI ECE Outstanding Graduate Student 2002
IUPUI ECE University Fellowship 2000-2002