Pacific Symposium on Biocomputing 14:87-98 (2009)

QUERYING PARSE TREE DATABASE OF MEDLINE TEXT TO SYNTHESIZE USER-SPECIFIC BIOMOLECULAR NETWORKS

LUIS TARI[1], JÖRG HAKENBERG[1], GRACIELA GONZALEZ[2], CHITTA BARAL[1]

[1]Department of Computer Science and Engineering,
Arizona State University, Tempe, AZ 85287, USA

[2]Department of Biomedical Informatics,
Arizona State University, Phoenix, AZ 85004, USA

{luis.tari, jhakenbe,graciela.gonzalez,chitta}@asu.edu

Curated biological knowledge of interactions and pathways is largely available from various databases, and network synthesis is a popular method to gain insight into the data. However, such data from curated databases presents a single view of the knowledge to the biologists, and it may not be suitable to researchers' specific needs. On the other hand, Medline abstracts are publicly accessible and encode the necessary information to synthesize different kinds of biological networks. In this paper, we propose a new paradigm in synthesizing biomolecular networks by allowing biologists to create their own networks through queries to a specialized database of Medline abstracts. With this approach, users can specify precisely what kind of information they want in the resulting networks. We demonstrate the feasibility of our approach in the synthesis of gene-drug, gene-disease and protein-protein interaction networks. We show that our approach is capable of synthesizing these networks with high precision and even finds relations that have yet to be curated in public databases. In addition, we demonstrate a scenario of recovering a drug-related pathway using our approach.

1. Introduction

Modeling large-scale biological knowledge in the form of networks is a common approach to advance our understanding of the mechanisms that govern the behavior of a cell. The steadily increasing amount of 'omics' data facilitates the synthesis of a wide variety of biological networks, from modeling physical protein-protein interactions [1], synthesizing networks of genes that share certain properties such as coexpression based on gene expression data [2], to proteins that share biological processes [3]. The synthesis of biological networks, such as gene-disease associations [4] and gene-drug interactions [5], provides insights to our understanding of the role of genetics in diseases. Descriptions of other kinds of biological networks can be found in [6].

As a way to share this wealth of biological knowledge, the data is made available in various databases, such as IntAct [7] and MIPS [8] for protein-protein interactions. Such interaction databases are typically curated manually by a team of scientists, aided by automated extractors or provided by external contributors in some cases. While interaction data from these databases are highly useful as a concise resource for biologists, the level of detail about the interactions is a priori defined by the databases. The interactions are often restricted to specific kinds of information so that information one might be interested, such as the structure or strength of the interactions, might not be encoded in the databases [9]. Biologists who use these interactions have to be aware of the limitations of the data, which can be unclear if the biologists are not familiar with the curation protocol for the particular database. In other words, biologists can only use the interaction data in a passive manner as they are not engaged in the curation process of the interactions. Biologists can perform filtering or visualization on the interactions provided by the databases as users, but not how the interactions are collected. Such passive use of interactions limits the applicability of the interaction data into research. On the other hand, Medline abstracts are publicly accessible and encode the necessary information to synthesize different kinds of biological networks. For instance, it was estimated that 270,000 Medline abstracts are classified as abstracts with mentions of human, mouse and yeast protein-protein interactions [10]. A more recent work found 150,000 protein-protein interactions in 1 million Medline abstracts [11].

Our goal is to provide a mechanism that allows users to synthesize biomolecular networks specific to their needs through queries against Medline abstracts. Unlike the traditional approaches in querying biomolecular networks that are synthesized from existing curated data, the networks generated from querying Medline abstracts can be more suitable to the users' needs. By using simple-to-use queries to our specialized database of Medline abstracts, these networks convey the information needed by the users, such as strength of the interactions, and such information might be missing in the networks that are synthesized from curated data. Users can specify precisely what kind of information they need in the networks through queries, such as preconditions of the interactions. In addition, users do not have to depend on the time-consuming curation process and synthesize biological networks from curated data that do not include the latest findings.

To implement this mechanism, Medline abstracts are parsed by a natural language parser to represent the syntactic structures of the sentences called parse trees. Entities such as genes, diseases and drugs are automatically identified with the use of entity recognizers as semantic information. The parse trees and the semantic information are then stored in a specialized relational database. Having the parse trees in the database enables the users to extract sentences using simplified linguistic queries. The resulting sentences allow the users to synthesize biomolecular networks specific to their needs. This new paradigm enables biologists to utilize information in Medline abstracts effectively and synthesize their own biomolecular networks.

Our proposed mechanism is different from automated extraction tools that perform biological relationship extraction. Typical relationship extractors such as iHop [12] rely on their own dictionaries of entities to identify their associations based on coccurrences. Other systems such as AliBaba [13] utilize linguistic patterns to extract relationships. From the user's point of view, such extraction systems are treated as black boxes and users cannot specify how the interactions should be obtained. Our method allows users to issue their own search criteria through queries to generate biomolecular networks of interest. A text analysis engine called TLM [14] is based on the idea of a retrieval system that retrieves sentences using textual patterns as queries. Unlike our approach, TLM does not utilize grammatical structures of the sentences to retrieve sentences. Interaction databases such as IntAct, MINT, MIPS for protein-protein interactions and PharmGKB [15] for gene-drug interactions rely on their curators to identify interactions, and users have no influence in the curation process. Proprietary pathway analysis tools such as Ingenuity IPA*, ActiveMotif† utilize their manually curated database to analyze experimental data, but the curation protocol is not accessible to the users. The use of query languages can be found in querying pathways, such as PQL [16], QPath [17] and PATIKA [18], and finding information that require multiple data sources, such as semCDI [19], GenoQuery [20], Cytoscape [21]. However, these query languages depend on curated data in order to return answers.

In this paper, we describe a new paradigm in how users can synthesize biomolecular networks. We illustrate how this approach can lead to the synthesis of gene-drug relationship networks, gene-disease association networks and protein-protein interaction networks.

2. Methods

Our proposed method is to place the user in control of synthesizing their own biomolecular networks through queries to our specialized database of Medline abstracts, and the results returned by the database are utilized to generate the resulting biomolecular networks. Suppose a user is interested in constructing a network of gene-drug relations, in which the drugs are metabolized by enzymes. The following query can be used:

```
<DRUG> _ metabolized by <GENE>
```

* Ingenuity Pathway Analysis Tool: http://www.ingenuity.com

† Active Motif: http://www.activemotif.com

The symbols `<DRUG>` and `<GENE>` infer that the sequences of words have to be a drug name and a gene/protein name in the matching sentences. The order of the tokens in the query matters, so that the above query specifies that the grammatical structures of the matching sentences include a syntactic dependency between the words "*metabolized*" and "*by*". Similarly, "*by*" has to be syntactically dependent on `<GENE>`. The operator _ is a wildcard operator that `<DRUG>` and "*metabolized*" may not have any syntactic dependency between them in the matching sentences. This query can retrieve support evidences such as "Diclofenac is widely used in the treatment of rheumatic diseases and is mainly *metabolized* in the liver *by* CYP2C9."(PMID: 8793607). The grammatical structure of the sentence reveals that there are syntactic dependencies between "*metabolized*" and "*by*", as well as "*by*" and "*CYP2C9*". By allowing users to perform their own queries, users can specify their own criteria in their target interactions. One way of specifying the strength of the interaction is to include the word "*extensively*" in the query as follows:

```
<DRUG> _ extensively metabolized by <GENE>
```

Here we are interested in drug-enzyme metabolic relations in which the strength of the interactions is described as "extensive". The support evidence "Tacrine is extensively *metabolized by* CYP1A2." (PMID:9209244) is an example retrieved by the query. There are cases when negative relations are reported in the literature. Our current system simply disregards sentences with words that indicate negation, such as "not", "no", so that sentences such as "Hesperetin was not metabolized by human CYP1A2" (PMID:10781868) are not retrieved as support evidences.

The essential component of our method is *parse trees* of Medline abstracts; parse trees are syntactic structures that represent the grammatical structures of sentences. Parse trees include *constituent trees* and *linkages*, in which constituent trees are hierarchical syntactic structures of sentences and linkages are composed of *links* that represent syntactic dependencies between pairs of words. These parse trees are generated automatically by the Link Grammar parser [22]. Such parse trees are ideal to be used for expressing linguistic patterns, which are commonly utilized in automated extraction systems. To store the parse trees, a database is needed to capture the hierarchical representation of abstracts, which include the sections of the abstracts such as title or body of the abstracts, parse trees and the semantic information of words. Semantic information includes the entity type of a sequence of words, such as whether it is a gene/protein name[‡], a drug name or a disease name. To cope with the high variation of gene names, an entity recognition system based on a statistical machine learning technique named BANNER [23] is utilized to identify gene names in text. Lists of drug and disease names from MeSH[§], DrugBank[**] and PharmGKB are employed to recognize drug and disease names. We called the database as the *parse tree database*, and the database is implemented using a relational SQL database. Since standard SQL queries are not ideal for expressing queries that involve linguistic patterns, we develop a query language called *parse tree query language* (*PTQL*) that are used to express linguistic patterns and query parse trees. The details of the PTQL query language and its implementation can be found in [24]. Similar to standard database query languages such as SQL, PTQL is designed to be used by developers and people who are familiar with linguistics. To facilitate the synthesis of biomolecular networks through querying of parse trees of sentences in Medline abstracts by biologists, we offer a simpler query language called $PTQL^{LITE}$ that is not as expressive as PTQL but the syntax is close to keyword-based queries used in search engines. The sample queries shown in the beginning of this section are $\text{PTQL}^{\text{LITE}}$ queries.

Figure 1 shows an overview of our approach in using $\text{PTQL}^{\text{LITE}}$ queries to synthesize biomolecular networks. The *processor* utilizes the named entity recognizers and parses the MedLine abstracts, and stores the processed information in the parse tree database and the inverted index. The *middleware* handles the communication between the web interface and the parse tree database. The middleware takes $\text{PTQL}^{\text{LITE}}$ queries as input and generates PTQL queries. The PTQL queries are translated into standard SQL queries before querying the parse tree database. Due to the complexity of the translated SQL queries, retrieving results from a large parse tree database can be slow. We increase the efficiency of our system by utilizing an off-the-shelve information retrieval (IR) system so that $\text{PTQL}^{\text{LITE}}$ queries are first translated into IR queries to retrieve the matching sentences. The PTQL queries are applied to only the parse trees of the sentences retrieved by the IR system rather than the entire database of parse trees so that the process can be performed efficiently. We summarize the process of translating $\text{PTQL}^{\text{LITE}}$ queries into SQL queries to retrieve answers by the middleware as follows:

[‡] Gene, protein and enzyme names are indistinguishable by current automated entity recognizers, and sometimes even by human readers. From here on, we use "gene" to refer to genes/proteins/enzymes.

[§] Medical Subject Headings (MeSH): http://www.nlm.nih.gov/mesh/

[**] DrugBank: http://www.drugbank.ca/

1. The *IR query generator* generates an IR query based on an PTQL^{LITE} query provided by the user.
2. The IR query is used to retrieve relevant documents *D* and sentences *S* from the *inverted index*.
3. The *PTQL generator* translates the PTQL query into an SQL query and instantiate the query with document id $d \in D$ and sentence id $s \in S$.
4. The SQL query generated in Step 3 is applied to the *parse tree database*.
5. The *bionetwork generator* delivers the resulting network.

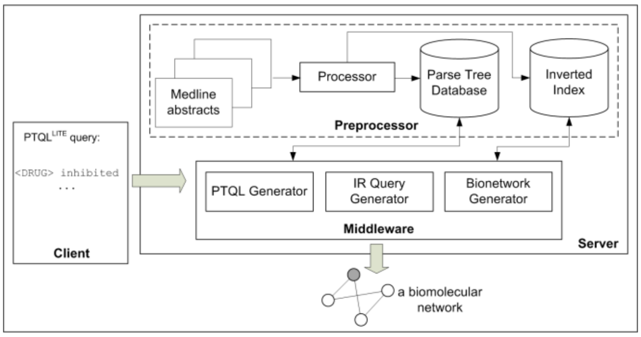

Figure 1 – A client-server system architecture for synthesizing bionetworks through querying parse trees of Medline abstracts. The middleware generates PTQL and IR queries based on the user's input, and retrieves information from the database and the index.

3. Synthesis of various biomolecular networks

In this section, we illustrate our approach with the synthesis of gene-drug relation, gene-disease association and protein-protein interaction networks. We evaluate our approach by using PharmGKB to verify the relations in the generated networks. As PharmGKB only covers part of the knowledge in the literature, we manually evaluated the correctness of the relations based on their corresponding supported evidences that were extracted by our approach.

3.1. *Gene-drug relationship networks*

Drug metabolism influences the effects of drug chemicals, and genetic variations can affect the effectiveness of drug metabolism. It is therefore essential to study the metabolic relations between enzymes and drugs. Here we illustrate the synthesis of a network of gene-drug relationships using our approach, specifically capturing the relations of drugs that are metabolized by enzymes. We use a collection of 13015 Medline abstracts from [25] that focus on topics about gene-drug relations to demonstrate the feasibility of our approach.

Sentences such as "*Triazolam is metabolized by CYP3A4*" (PMID:8612379) are typical examples of how gene-drug metabolic relations are described in biomedical articles. Biologists who are interested in such relations would use the following PTQL^{LITE} query to synthesize their networks:

```
<DRUG> _ metabolized by <GENE>
```

By default, the system filters out any sentences that infer negative relations. This query results a network of 141 genes and drugs with 138 relations generated from 178 supporting sentences, and each relation is supported by at least 1 sentence. The gene-drug network took about 10 seconds to be generated on a 2-GHz Intel DualCore CPU with 2 GB of RAM. To verify the correctness of the network, we used the gene-drug relations from PharmGKB [15], which is one of the largest curated databases of relations among genes, drugs and diseases that are publicly available. Among the 138 relations in the network, 43 of them can be found in PharmGKB. We further manually evaluated the correctness of the relations in the network based on their evidences that were extracted by our method. We observed that 122 out of 138 (i.e. precision of 88.41%) are indeed correct. We analyzed the incorrect relations and categorize the errors into two sources: (i) errors in extraction due to the sentence structure; (ii) errors due to recognition of entities. We list out some of these errors in Table 1. Example 1 in Table 1 is incorrect due to the fact that the clause describing the drug-enzyme metabolic relation for CYP2C9 is not in the same clause as the drugs lovastatin, simvastatin and atorvastatin. On the other hand, incorrect identification of drug names, such as recognizing "*important drugs*" and "*widely used drugs*" as drug names, leads to incorrect support evidences as shown in example 2 of Table 1. A careful, manual revision of the lexicon used for drugs would eradicate many of the type (ii) errors.

Table 1 – Support evidences that are extracted incorrectly by the query `<DRUG> _ metabolized by <GENE>`

	Gene/Drug	Incorrectly extracted evidence

1	CYP2C9/Lovastatin;
	CYP2C9/Simvastatin;
	CYP2C9/Atorvastatin
		Lovastatin, simvastatin, and atorvastatin are substrates of CYP3A4, whereas *fluvastatin is* metabolized by CYP2C9. (PMID:11029845)

2	CYP2E1/important drugs	Among important drugs *metabolized by* CYP2E1 … (PMID:2134674)

We synthesize another network using the same PTQL^{LITE} query but specifying that the relations in the network have to be supported by at least 2 different publications. A smaller network of 33 vertices (10 genes and 23 drugs) with 27 edges is generated with this criterion, as shown in Figure 2. Such network allows the discovery of potential relations. For instance, the drugs omeprazole are metabolized by CYP3A4 and CYP2C19, and users might want to study a potential relation between CYP3A4 and CYP2C19. Table 2 lists some of the relations encoded by this network as well as the corresponding supported sentences. Among the 27 relations, only 2 of them are considered as an error in the extraction (i.e. precision of 92.59%). This experiment serves as a proof-of-concept that a biologist can easily synthesize a user-specific network with PTQL^{LITE} queries. This also shows that our approach can overcome the time-consuming process of expert curation, which generally results in a low coverage of the knowledge that has already been published in the literature.

Table 2 – A partial list of gene-drug relations generated by our approach using the pattern `<DRUG> _ metabolized by <GENE>`. Each gene-drug relation is supported by at least 2 different publications, and the relations are yet to appear in PharmGKB.

Gene/Drug	Support evidence
CYP3A4/Triazolam	Triazolam is *metabolized by* CYP3A4. (PMID:8612379)
CYP3A4/Terfenadine	… therapeutics (terfenadine and cyclosporine) known to be *metabolized by* CYP3A4 (PMID:10752642)
Alcohol dehydrogenase / Ethanol	Ethanol is *metabolized by* alcohol dehydrogenase in the human stomach. (PMID:9693201)
CYP1A2/Propafenone	Propafenone is mainly *metabolized by* CYP2D6 (PMID:10917404)
CYP2E1/Chlorzoxazone	Chlorzoxazone is mainly *metabolized* to 6-OHchlorzoxazone *by* CYP2E1. (PMID:7910460)

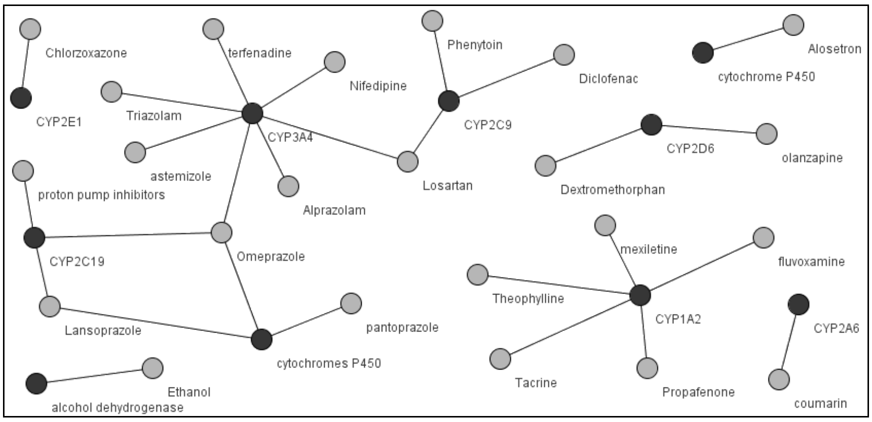

Figure 2 – A gene-drug network in which each edge represents a drug metabolized by an enzyme. Each edge is supported by at least 2 support evidences.

We illustrate our network synthesis approach with another example of drug-enzyme inhibition. The following PTQL^{LITE} query can be used:

```
<DRUG> _ inhibit <GENE> | <DRUG> _ inhibits <GENE> |
inhibition of <GENE> by <DRUG> |
<GENE> _ inhibited by <DRUG>
```

This query is composed of 4 subqueries, which are separated by the operator |, to capture the drug-enzyme inhibition relations. Our system essentially synthesizes the drug-enzyme inhibition network as the union of the relations resulted from the 4 subqueries. Using the criteria that each relation has to be supported by at least 2 publications, the resulting network is composed of 14 enzymes and 13 drugs with 19 relations. Among these 19 relations, 7 of the relations can be verified with the PharmGKB database. We further look into the other 12 relations, and realize that 9 relations are well supported by our extracted evidences, as shown in Table 3.

Table 3 – A list of correct drug-enzyme inhibitions and the corresponding support evidences. These correct relations are currently not contained in PharmGKB.

Gene/Drug	Support evidence

CYP3A4/Indinavir;
CYP3A4/Nelfinavir;
CYP3A4/Amprenavir
	The HIV protease inhibitors amprenavir, indinavir, nelfinavir, ritonavir and saquinavir *inhibit* CYP3A4. (PMID:10926350)

CYP2D6/Terbinafine	Terbinafine *inhibits* CYP2D6. (PMID:11475469)
alcohol dehydrogenase/ 4-methylpyrazole	… 4-methylpyrazole to *inhibit* alcohol dehydrogenase. (PMID:2994256)
CYP2D6/Quinidine	*Inhibition of* CYP2D6 *by* quinidine … (PMID:10510150)
catechol-O-methyltransferase/ tolcapone	*Inhibition of* catechol-O-methyltransferase *by* tolcapone has been shown …. (PMID:9343116)
Thiorphan/Bradykinin	… formation of the major metabolite bradykinin 1-7 was *inhibited by* thiorphan. (PMID:1629199)
benzoyl-Gly-His-Leu/ captopril	The metabolism of benzoyl-Gly-His-Leu was completely *inhibited by* captopril (PMID:7588745)

3.2. *Gene-disease relationship networks*

We illustrate how we can synthesize gene-disease association network using our approach. We use the following query to construct such gene-disease network:

```
<GENE> _ associated with <DISE> |
      <GENE> _ risk of <DISE>
```

Using the same 13015 Medline abstracts that were used in constructing gene-drug networks as described in the previous subsection, a network with 88 genes and diseases with 76 gene-disease relations is generated with the above query. Each of the relations in this network is supported by at least 1 publication. The evaluation of the network using PharmGKB shows that 7 of the relations can be confirmed as correct. Our manual evaluation shows that 54 of the 76 relations (i.e. precision of 71.05%) are correct by analyzing the extracted support evidences. We conclude that 11 of the incorrect relations are due to errors from the entity recognizers. For instance, *ACE inhibition* was incorrectly recognized as a gene name when in fact it is considered as a drug/treatment (even though *ACE* itself is a gene name). The rest of the incorrect relations are caused by incorrect extraction. We also synthesize another network using the criteria of at least 2 publications as support for the relations in the network. This results a small network with 11 genes and diseases with 6 relations. One of the reasons for such a small network is that our current system does not utilize normalization techniques to realize that terms such as "*vitamin D receptor*" and "*VDR*" refer to the same entity. Table 4 lists out some of these associations.

Table 4 – A list of gene-disease associations and the corresponding support evidences.

Gene/Disease	Support evidence
VDR/Osteoporosis	To determine whether a polymorphism of the VDR gene, already *associated with* osteoporosis …. (PMID:9259424)
UGT1A1/Gilbert's syndrome	The presence of an additional TA repeat in the TATA sequence of UGT1A1 has been *associated with* Gilbert's syndrome. (PMID:10340924)
VDR/ Hyperparathyroidism	Polymorphism of the VDR gene has recently been shown to be related to bone mineral density, and also *associated with* hyperparathyroidism …. (PMID:10508794)

3.3. *Protein-protein interaction networks*

We constructed a network of protein-protein interactions using the BioCreative 2 dataset [26]. The task in the BioCreative 2 IPS benchmark is to find protein-protein interactions for which a text provides evidence for a physical interaction between the proteins. A sample query is as follows:

```
<GENE> _ binds with <GENE>
```

We generated 11208 PTQL queries from the BioCreative 2 training dataset, and achieved a precision of 83.6% and recall of 58.6%.

4. Scenario

We illustrate a scenario on how users who are not familiar with linguistic structures can utilize our system to synthesize networks. Suppose the user is interested in synthesizing a pathway about the drug tamoxifen, the following query can be issued to first find all sentences that describe associations between genes and tamoxifen.

```
<GENE> _ <DRUG=”tamoxifen”>|<DRUG=”tamoxifen”> _ <GENE>
```

The retrieved sentences contain cooccurrences of genes and tamoxifen. The user can examine some of these sentences and refine the relations with this query:

```
<DRUG=”tamoxifen”> _ involvement of <GENE> |
  Binding _ <DRUG=”tamoxifen”> _  <GENE>
```

A pathway that involves the genes CYP3A4, CYP2B6 and CYP2D6 with tamoxifen can then be synthesized. This pathway is supported by the evidences as shown in Table 5, and can be verified with PharmGKB. This example illustrates the feasibility of pathway synthesis using our approach.

5. Conclusion

We demonstrate that our approach is capable of synthesizing biological networks with high precision without the use of curated data. This new paradigm of network synthesis tailors the specific needs of the users. Future work includes normalization of entities to handle name variations, and parse all Medline abstracts so that networks can be synthesized with respect to the latest findings. Inclusion of query templates that are typically used in describing gene-drug, gene-disease and protein-protein relations will be provided through the interface, so that synthesizing biological networks with our approach can even be simpler. A prototype based on 13015 Medline abstracts from [25], mainly focusing on gene-drug relations, is available at http://cbioc2.eas.asu.edu/netsynthesis.

Table 5 – A list of support evidences for the pathway that involves the drug tamoxifen

PMID	Support evidence
7748182	*Binding of* tamoxifen correlated with CYP3A4 and CYP2B6 content.
9037249	The proportion of activity inhibited by quinidine correlated positively with total microsomal tamoxifen 4-hydroxylation activity, indicating a major *involvement of* CYP2D6 in this reaction.

Acknowledgement

The authors would like to acknowledge the support of NSF through grant number 0412000 and a grant from SFAz. We also appreciate the helpful comments provided by the reviewers of this paper.

References

1. P. Uetz, L. Giot, G. Cagney, et al. A comprehensive analysis of protein-protein interactions in S. cerevisiae. *Nature,* **403**, 6770, 623-627 (2000).
2. V. van Noort, B. Snel and M. A. Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.,* **5**, 3, 280-284 (2004).
3. L. Tari, C. Baral and P. Dasgupta. Understanding the global properties of functionally related gene networks using GO. *PSB,* 209-220 (2005).
4. J. Lim, T. Hao, et al. A PPI network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell,* **125**, 4, 801-814 (2006).
5. G. Giaever, P. Flaherty, et al. Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *PNAS,* **101**, 3 (2004).
6. X. Zhu, M. Gerstein and M. Snyder. Getting connected: analysis and principles of biological networks. *Genes Dev.,* **21**, 9, 1010-1024 (2007).
7. H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, et al. IntAct: an open source molecular interaction database. *NAR,* **32**, suppl 1, D452-455 (2004).
8. H. W. Mewes, D. Frishman, U. Guldener, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res,* **30**, 31-34 (2002).

9. M. J. Betts and R. B. Russell. The hard cell: from proteomics to a whole cell model. *FEBS Lett.*, **581**, 15, 2870-2876 (2007).
10. I. Donaldson, J. D. Martin, B. d. Bruijn, et al. PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11 (2003).
11. K. Fundel, R. Kuffner and R. Zimmer. RelEx--Relation extraction using dependency parse trees. *Bioinformatics*, **23**, 3, 365-371 (2007).
12. R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nat. Genet.*, **36**, 7, 664 (2004).
13. C. Plake, T. Schiemann, M. Pankalla, et al. AliBaba: PubMed as a graph. *Bioinformatics,* **22**, 19, 2444-2445 (2006).
14. J. D. Martin. Fast and Furious Text Mining. *IEEE Data Eng. Bull.,* **28**, 4, 11-20 (2005).
15. M. Hewett, D. E. Oliver, D. L. Rubin, et al. PharmGKB: the Pharmacogenetics Knowledge Base. *NAR,* **30**, 1, 163-165 (2002).
16. U. Leser. A query language for biological networks. *Bioinformatics*, **21**, Supplement 2, ii33-ii39 (2005).
17. T. Shlomi, D. Segal, E. Ruppin, et al. QPath: a method for querying pathways in a PPI network. *BMC Bioinformatics*, **7**, 199-207 (2006).
18. E. Demir, O. Babur, U. Dogrusoz, et al. PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics,* **18**, 996-1003 (2002).
19. E. P. Shironoshita, Y. R. Jean-Mary, R. M. Bradley, et al. semCDI: A Query Formulation for Semantic Data Integration in caBIG. *J. Am. Med. Inform. Assoc.*, **15**, 4, 559-568 (2008).
20. F. Lemoine, B. Labedan and C. Froidevaux. GenoQuery: a new querying module for functional annotation in a genomic warehouse. *Bioinformatics*, **24**, 13, i322-9 (2008).
21. P. Shannon, A. Markiel, O. Ozier, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 11, 2498-2504 (2003).
22. D. Grinberg, J. Lafferty and D. Sleator. A Robust Parsing Algorithm For Link Grammars. **CMU-CS-TR-95-125** (1995).
23. R. Leaman and G. Gonzalez. BANNER: An executable survery of advances in biomedical named entity recognition. *Pacific Symposium of Biocomputing (PSB)*, 652-663 (2008).
24. P. H. Tu, C. Baral, et al. Generalized text extraction from molecular biology text using parse tree database querying. **TR-08-004** (2008).
25. J. T. Chang and R. B. Altman. Extracting and characterizing gene-drug relationships from the literature. *Pharmacogenetics*, **14**, 9, 577-586 (2004).
26. M. Krallinger, F. Leitner and A. Valencia. Assessment of the Second BioCreative PPI task: Automatic Extraction of Protein-Protein Interactions. *Proceedings of the Second BioCreative Challenge Workshop*, 41-54 (2007).