Pacific Symposium on Biocomputing 12:28-39(2007)

MINING GENE-DISEASE RELATIONSHIPS FROM BIOMEDICAL LITERATURE: WEIGHTING PROTEIN-PROTEIN INTERACTIONS AND CONNECTIVITY MEASURES

GRACIELA GONZALEZ[∞], JUAN C. URIBE[*], LUIS TARI[*],
COLLEEN BROPHY[+], CHITTA BARAL[*],

[∞]*Department of Biomedical Informatics, Ira A. Fulton School of Engineering,*
[*]*Computer Science and Engineering Department, Ira A. Fulton School of Engineering,*
[+]*Center for Metabolic Biology, Department of Kinesiology*
Arizona Sate University
Tempe, Arizona 85281, USA

Motivation: The promises of the post-genome era disease-related discoveries and advances have yet to be fully realized, with many opportunities for discovery hiding in the millions of biomedical papers published since**.** Public databases give access to data extracted from the literature by teams of experts, but their coverage is often limited and lags behind recent discoveries. We present a computational method that combines data extracted from the literature with data from curated sources in order to uncover possible gene-disease relationships that are not directly stated or were missed by the initial mining.
Method: An initial set of genes and proteins is obtained from gene-disease relationships extracted from PubMed abstracts using natural language processing. Interactions involving the corresponding proteins are similarly extracted and integrated with interactions from curated databases (such as BIND and DIP), assigning a confidence measure to each interaction depending on its source. The augmented list of genes and gene products is then ranked combining two scores: one that reflects the strength of the relationship with the initial set of genes and incorporates user-defined weights and another that reflects the importance of the gene in maintaining the connectivity of the network. We applied the method to atherosclerosis to assess its effectiveness.
Results: Top-ranked proteins from the method are related to atherosclerosis with accuracy between 0.85 to 1.00 for the top 20 and 0.64 to 0.80 for the top 90 if duplicates are ignored, with 45% of the top 20 and 75% of the top 90 derived by the method, not extracted from text. Thus, though the initial gene set and interactions were automatically extracted from text (and subject to the impreciseness of automatic extraction), their use for further hypothesis generation is valuable given adequate computational analysis.

1. Introduction

Post-genome project data and techniques available to the research community have exponentially increased the capacity of researchers to conduct experiments and publish results. The resulting deluge of biomedical literature, however, has reached a point that exceeds the capacity of any researcher to process and assume, making it difficult to realize the full benefit of these findings. From 1994 to 2004, close to 3 million biomedical articles were published by US and European researchers [1]. This publication rate has resulted in approximately 16 million publications currently indexed in PubMed.

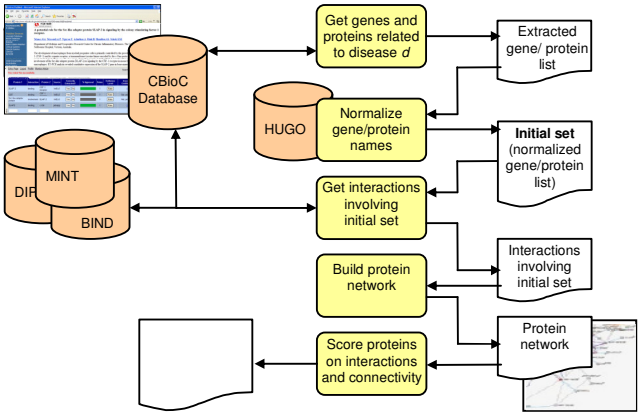


Figure 1. Overview and data flow of the computational method presented here to mine the biomedical literature for genes potentially related to a specific disease.

Efforts have been made to extract data from articles and abstracts. For example, Entrez's OMIM [2] has summaries of published work that relate genes to diseases. However, it covers only about 20% of the human genes in the Entrez Gene database. A similar initiative for gene function annotation, GeneRIF (Gene Reference Into Function), was started in 2002, but it covers only about 1.7% of all the genes in Entrez and 25% of human genes[3]. New findings usually take a long time to be reflected by curated sources such as these, and any computational method that relies solely on them will necessarily have its hands tied.

To fill this void, the Collaborative Bio Curation (CBioC) project [4, 5] was started to bring together nuggets of information automatically extracted from the published biomedical literature and the intellectual power of a social network of researchers, who can rate the accuracy of the extraction. Extracted facts include protein-protein interactions, gene-disease and gene-bioprocess relationships.

This paper describes a computational method that uses extracted facts from the CBioC database and integrates them with curated sources to find a set of proteins potentially related to a target disease, ranking them so that existing knowledge (known gene-disease relationships and curated protein-protein interactions) is balanced with the potential impact of new information (protein-protein interactions extracted from the literature) and the researcher's intuition. An assessment of the method through a study of atherosclerosis is also described and reported in the Results section.

This balance of different factors, notably a network connectivity impact measure for each gene, among others, marks the difference between our approach and others such as MedGene [6] and the method in [7]. The scope of the initial gene-disease data also differs, as well as the level of user interaction,

which is very limited in other approaches. A comparative view to these efforts is presented in the Related Work section and in Section 2.4.

The resulting ranked list of genes and gene products can provide the basis for further focused experiments to investigate the genetic determinants of any disease. On top of helping to find gene-disease relationships that were not discovered in the information extraction step (false negatives), this focused analysis could uncover yet unexplored genetic linkages and provide an insight into specific genetic and proteomic pathways related to any disease, as our study in atherosclerosis will show. The method is implemented in Java using SQL to access the CBioC database (which is stored as a MySQL database). On-demand runs can be requested by contacting the authors. A web-based interface to the software is in development.

Other sections cover the computational method, the results of applying the method to the study of atherosclerosis, and a comparison with related work.

2. The computational method

The computational method presented here takes a 4-step approach to the task of finding and ranking genes and gene products related to a given disease, relying not only on automatic computation, but allowing (not requiring) user input at different levels. The method can be summarized as follows:

1. Obtain a list of genes or gene products known to be involved with the target disease from the CBioC[5] database.
2. Apply heuristics to unify variants of extracted names, and use HUGO [8] to normalize both the set obtained in the previous step and the names stored in CBioC. This will be referred to as the ***initial set***.
3. Apply nearest-neighbor expansion to the initial set to build a protein interaction network using data from the CBioC database and curated databases. Analyze the connectivity of the network. The genes and proteins in this network (derived from the interactions) form the ***extended set***.
4. Apply a heuristic scoring formula to the extended set to predict the proteins most likely related to the disease.

One part of the formula measures the number of interactions of each gene in the extended set with proteins in the initial set, incorporating contextual information if indicated by the user. The second part measures the role of the protein in the connectivity of the protein network, since high degrees of local network interconnectivity can identify sets of functionally related proteins [9, 10]. Researchers can focus the analysis through different interventions. Figure 1 shows the data and process flow of the method. Each step is detailed next.

2.1. *Initial set of disease-related genes and gene products*

The initial set of genes and gene products of interest is obtained by querying the CBioC database using the disease of interest and any variants or synonyms of its name. CBioC uses a natural language processing extraction system, IntEx [11], which is based on identification of syntactic roles, such as subject, objects, verbs, and modifiers. English grammar dependencies reported by Link Grammar [12] are used to identify the roles and transform complex sentences of interest into triplets of the form (Entity1, interaction, Entity2). We extended IntEx to extract not only protein-protein interactions, but also gene-disease relationships, using MeSH [13] terms under the disease category to recognize them in the abstracts.

Even though the natural language processing approach allows for more precise extractions than co-occurrence[11], the gene-disease relationships and protein-protein interactions extracted directly from the literature are not perfect. In fact, IntEx reports a 65.7% precision in extracted interactions[11]. Thus, there will be genes and gene products in the initial set that are not related to the disease (false positives), just as there will be others that are not retrieved even though they are related (false negatives). The protein interaction network analysis and the incorporation of protein-protein interactions from curated sources helps assuage the impact of these problems. Also, users might filter the initial set to narrow the focus to a particular set of genes and gene products.

2.2. *Unifying extracted gene and protein names*

One of the challenges of using data extracted directly from biomedical texts is the great variety of names used for the same entity: one gene or gene product might appear under different synonyms and variants. For example, HNF4A might appear as hepatocyte nuclear factor 4 alpha or any of a number of aliases (such as HNF4, MODY, TCF, or TCF14), or variants of any of these, such as HNF4-alpha or HNF 4A. An additional problem is that the triplets in the CBioC database sometimes include modifiers that were in the same noun phrase or modifying phrase, such as “HNF4A protein” or “HNF4A mutation”. It was necessary to unify the names (*normalize* them) so when the protein network is built, all the interactions of the same protein are clustered into a single node. A naive normalization algorithm was applied to entries in the CBioC database to eliminate non-essential words (such as “protein” or “mutation” at the end of a name), in order to then find its official abbreviation in the HUGO[8] database.

2.3. *Build the protein network*

The CBioC database is queried for any and all interactions involving the genes and gene products in the initial set. On top of the extracted interactions, CBioC

integrates interaction data from BIND[14], MINT[15], DIP[16], IntAct [17], and BioGRID[18]. A nearest neighbor algorithm is run to build a protein interaction network, noting the *confidence level* for each interaction as follows:

1. If the interaction comes from any of the curated sources, its confidence level is noted as 1.
2. If the interaction comes from CBioC, and it has received "Yes" votes from the community of users, its confidence level is noted as .65 plus .07 for each "Yes" vote up to 1. CBioC counts only one vote per user per fact.
3. If the interaction comes form CBioC, and has not been rated by any user, its confidence level is given as .65 (the measured precision of IntEx [11]).

2.4. *Rank the genes and gene products in the expanded network*

To rank the genes in the resulting set, we score each gene or gene product based on the number and confidence levels of its interactions with proteins in the initial set, and combine this measure with another that reflects how relevant it is for maintaining the protein network connectivity. Both measures are important. The first helps discover the most active proteins with respect to the disease (high precision), preferring interactions with the highest confidence level (high fidelity), while the second finds those that could potentially play a crucial role in a pathway related to the disease or that are very likely related to the known (extracted) genes, as high degrees of local network interconnectivity can identify sets of functionally related proteins [9, 10].

The first score also incorporates user-defined weights. For example, given interactions as triplets (Entity1, interaction-term, Entity2), users might indicate that interactions that include "phosphorylates" as an interaction term should be given greater weight. Let us assume for now that no user weights are defined. We use a variation of the formula given in [7] for this level, removing a bias towards the initial set that the formula in [7] suffers from. Let

- A be the extended set of proteins (initial set plus interactions).
- $N(i)$ is the set of proteins in the initial set interacting with protein i.
- $p(i,j)$ be the confidence level of the interaction between proteins i and j.
- $N(i,j) = 1$ if protein $i \in A$ and $j \in N(i)$, and 0 otherwise.

Then a score t is assigned to each protein i by applying Eq.(1).

$$t_i = u_i^2 * |N(i)| \tag{1}$$

$$u_i = \frac{\sum_{j \in N(i)} p(i, j)}{|N(i)|} \tag{2}$$

Equation (2), u_i, is the average confidence level of the interactions involving i. Equation (1) results from expanding the formula used in [7], noting that in [7], $N(i)$ is the set of proteins interacting with protein i and $N(i,j) = 1$ if $j \in N(i) \cap A$.

$$t_i = \exp\left(k * \ln\left(\sum_{j \in N(i) \cap A} p(i,j) \right) - \ln\left(\sum_{j \in N(i) \cap A} N(i,j) \right) \right)$$

$$= \frac{\exp\left(k * \ln\left(\sum_{j \in N(i) \cap A} p(i,j) \right) \right)}{\exp\left(\ln\left(\sum_{j \in N(i) \cap A} N(i,j) \right) \right)} = \frac{\exp\left(\ln\left(\sum_{j \in N(i) \cap A} p(i,j) \right) \right)^k}{\sum_{j \in N(i) \cap A} N(i,j)} \quad (3)$$

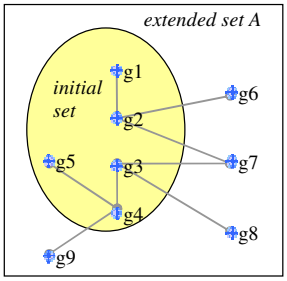
$$= \frac{\left(\sum_{j \in N(i) \cap A} p(i,j) \right)^k}{\sum_{j \in N(i) \cap A} N(i,j)} = \frac{\left(\sum_{j \in N(i) \cap A} p(i,j) \right)^k}{\left| N(i) \cap A \right|}$$

From the last expression, using k=2 as in [7], and noting that $|N(i) \cap A| = N(i)$ (since only interactions in A are included), we get Eq. (1). Since u_i<=1 remains relatively constant, the score t_i mainly depends on the number of interactions for *i.* By the definition of *N(i)* in the method presented here, only interactions with proteins in the initial set would be counted, whereas all of the interactions loaded for *i* are counted in [7].

Thus, since only interactions that have at least one member in the initial set are loaded, all of the interactions involving the proteins that belong to the initial set will be counted in [7], compared to only a small fraction of the interactions for the proteins not in the initial set. Naturally, this leads to a scoring bias in [7], with proteins in the initial set having a larger score than all other proteins. This explains why only one of the top ranked proteins in their final ranking was "novel" (not in the initial set –"derived" in this study–). Figure 2 illustrates the problem by comparing what would be the score given with each method given a confidence level of 1 for all interactions. The method in [7] (column (a)) would score g2, g3, g4 the highest, while the method presented here (column (b)) will rate g4 and g7 equally high.

Thus, *we count only interactions with proteins in the initial set*, which "evens out" the playing field for the proteins added later: if they interact with a good number of proteins in the initial set, their score will go up. Aside from the purely mathematical, this change makes biological sense: if a relationship has been reported between a gene and a disease, other proteins that are highly connected to the known facts might be important pieces in a pathway.

The fairness of the formula is obvious in our evaluation study (Section 3), where 45% of the top 20 ranked proteins in our resulting list were derived from interactions, compared to 1 in 20 in [7].



	(a)	(b)	(c)
g1	1	1	0
g2	3	1	2/8
g3	3	1	4/8
g4	3	2	2/8
g5	1	1	0
g6	1	1	0
g7	2	2	3/8
g8	1	1	0
g9	1	1	0

Figure 2. Simplified comparative scoring, assuming average confidence = 1, using scoring as in Eq. (1) but counting (a) all interactions involving protein *i*, as in [7], and (b) counting only interactions with proteins in the initial set, as done in the method presented here –before normalizing over 100-, with the corresponding (c) connectivity score (an innovative aspect of this method).

Further mathematical manipulations at this level include applying user-defined bias to certain proteins, as explained before, and normalization over 100 to reflect the relative ranking (thus, the interaction score t_i of protein i indicates the % of proteins with an interaction score less than or equal to that of i).

The second level of scoring involves evaluating the role of the gene on the overall connectivity of the protein network. It has been demonstrated that high degrees of local network interconnectivity can identify sets of functionally related proteins [9, 10]. The statistical validity of using network connectivity measures for sets of interacting proteins in this way has already been established [7, 9, 10, 19]. Here, this concept is applied to assessing the importance of a protein, measuring how much would the connectivity of the network be affected if it were removed. To formulate this precisely, let

- The path between two proteins p_1 and p_n be the set $\{p_1, p_2, \ldots p_{k-1}, p_k, \ldots p_n\}$ such that for $n > 2$, p_{k-1} interacts with p_k for every k in $2 \ldots n$.
- A set of interactions is called a network.
- The largest connected sub-network in a network is the largest subset of interactions from it that forms a path.
- The connectivity index (aka index of aggregation) of a network N, $C(N)$, is the ratio of the size of the largest sub-network of N to the size of N.

The connectivity score for a given gene or gene product i is given by Eq(4).

$$connectivity\_score_i = C(N) - C(N \setminus i) \quad (4)$$

In Eq.(4), $N \setminus i$ stands for all the proteins in N except i. This score is then combined with the interaction score t_i given by Eq. (1) using Eqs.(5) and (6).

$$combined\_score_i = t_i^{\,s} \quad (5)$$

$$s = 1 + (w * connectivity\_score_i) \quad (6)$$

The exponential combination of the scores was preferred over linear since the connectivity score is very small (less than 0.01 in most cases). The constant w is used to adjust the weight of the connectivity score in the overall ranking.

The study in atherosclerosis presented here uses w=4 to achieve an approximate even split in the extracted and derived proteins among the top 20.

The combined score allows distinctions among genes such as g4 and g7 in Figure 1, where the connectivity score (column (c)) will break the tie and favor g7, since removing g7 will disconnect the network.

3. Results

According to the American Heart Association, more than 71 million American adults have one or more types of cardiovascular disease. It is the underlying cause of death in 37% and a contributing cause in 58% of all deaths in the United States. It claims more lives than the next 4 leading causes of death combined [20]. Atherosclerosis is the deadliest of cardiovascular disease and accounts for nearly three-fourths of all deaths.

Atherosclerosis results from a complex process involving endothelial dysfunction, inflammation, and dyslipidemia (a process called atherogenesis). The process leads to the accumulation of lipid and extracellular matrix proteins in the intima of arteries. Even though the genetic basis of atherosclerosis is not completely understood [21], a number of genes have been associated with atherosclerosis. Gene expression profiling of atherosclerosis has been used to identify relevant genes and pathways [22]. Our tool will allow the incorporation of published data into these experiments, and could help form new hypothesis.

There were 98 genes and gene products in the initial set from the CBioC database, resulting in 9963 genes in the extended set. Coverage was calculated with respect to OMIM [2] at 0.70 (with 73 out of the 104 genes listed in OMIM included in the extended set), using edit distance <= 1 to match (*i.e.*, one or no characters were dropped or added to declare a match).

We researched the evidence supporting the relationship of the top-ranked genes as to atherosclerosis, annotating each gene according to the findings, as described in Table 1. The accuracy statistics for the top n proteins appear in Table 2, with definitions and formulas used for all measures in Table 1.

Table 3 presents the details for the top 20 unique proteins. Annotations for the top 90 unique proteins are available at http://www.cbioc.org. Those marked "found", like TNF alpha, Angiotensin II, IL1, Collagen, and PLAT, were verified by direct PubMed searches. Consider TNF alpha: among over 800 hits, PMID 16718633 reports the contribution of TNF-alpha, TGF-beta and IL-6 gene expression to systemic inflammation in atherosclerosis. It is also mapped to GO term 0008289, "lipid binding activity" [23]. However, they are not mentioned in OMIM and in some cases nor in Entrez Gene as being related to atherosclerosis, and would have been missed by relying only on the information in these sources.

Table 1. Definitions used in assessing the computational method (TP = true positives, FN = false negatives, FP= false positives).

Extracted	protein belongs to the initial set
Derived	protein belongs to the extended set
Known	protein is among those reported in OMIM as related to the disease
Found	protein found in the literature as related to the disease
Suspect	protein likely related to the disease based on its function or interactions
Some support	protein found to be related, small number of supporting articles
Not Found	protein not found to be related to the disease
Not a gene	extracted entity does not refer to a gene or protein
Duplicate	synonym or variant of a previously listed protein
strict_TP	known + found
relaxed_TP	known + found + some support + suspect
Coverage	TP / (TP + FN) = known / all OMIM genes related to disease
Accuracy_stp	TP / (TP + FP) = strict_TP / (extracted + derived – duplicates)
Accuracy_rtp	TP / (TP + FP) = relaxed_TP / (extracted + derived – duplicates)
Accuracy_stp w/ dups	TP / (TP + FP) = strict_TP / (extracted + derived)
Accuracy_rtp w/ dups	TP / (TP + FP) = relaxed_TP / (extracted + derived)

Table 2. Performance measures for the top *n* proteins. See definitions in Table 1.

	n = 27	*n* = 123
Unique proteins	20	90
Extracted	12	31
Derived	15	92
% derived	56%	75%
Known (in OMIM)	9	16
Found (in literature)	8	42
Some support	1	6
Suspect	2	8
Not Found	0	16
Not a gene	0	2
Duplicate	7	33
Coverage wrt OMIM	0.09	0.15
Accuracy_stp	0.85	0.64
Accuracy_rtp	1.00	0.80
Accuracy_stp-w/ dups	0.63	0.47
Accuracy_rtp-w/ dups	0.74	0.59

Proteins marked “Suspect” are those for which there are threads in the literature that suggest they might be involved in the disease due to their interactions or function, but no direct report linking the two was found. For example, for PRKCG, PMID 10617676 states: “the signaling pathway of protein kinase C is known to play a role in mediating the action of cytokines”. Other cytokines, such as IL1 and IL6 have strong evidence of linkage to atherosclerosis [24]. For ERVK2 (HERV), PMID 11672541 states that it “may cause type I diabetes by activating autoreactive T cells”, and that “endogenous retroviral (HERV) superantigens induced via IFN-alpha by viral infections is a novel mechanism through which environmental factors may cause disease in genetically susceptible individuals.” In turn, PMID 16973967 states that “Adaptive immunity, in particular T cells, is highly involved in atherogenesis”, relating T cells to the disease. Other articles support this idea.

Overall, the top genes identified fit into categories underlying pathogenetic mechanisms of atherosclerosis: insulin resistance (insulin, ALB, ERVK2), lipids (APOB, APOE, HDL and LDL), inflammation (IL6, TNFa, IL1 –cytokines-), hypercoagulability (Fibrinogen) and endothelial injury (NOS, and ICAM).

Table 3. Top genes and gene products, ranked by combined score, using w=4. Duplicates due to name variants are not shown.

Protein	Type	Interaction score	Connectivity score	Combined score	Evidence
INSULIN	extracted	100.0	0.2149	5239.1	Known
ALB	extracted	60.0	0.0375	110.9	Known
APOE	extracted	65.0	0.0314	109.8	Known
FIBRINOGEN	extracted	52.5	0.0334	89.1	Found
ICAM 1	extracted	42.5	0.0341	70.9	Known
IL6	extracted	40.0	0.0315	63.7	Known
HDL	extracted	52.5	0.0116	63.1	Found
TNF ALPHA	derived	62.5	0.0001	62.6	Found
LDL	extracted	50.0	0.0120	60.3	Known
NOS	extracted	37.5	0.0304	58.3	Found
APOB	extracted	45.0	0.0116	53.7	Known
ERVK2	derived	50.0	0.0001	50.1	Suspect
ANGIOTENSIN II	derived	50.0	0.0001	50.1	Found
IL 1	derived	50.0	0.0001	50.1	Found
PRKCG	derived	47.5	0.0001	47.6	Suspect
COLLAGEN	derived	45.0	0.0001	45.1	Found
TAT	derived	44.5	0.0001	44.6	Some support
VWF	extracted	32.5	0.0217	44.0	Known
PLAT	derived	42.5	0.0001	42.6	Found
LIPOPROTEIN L	derived	40.0	0.0002	40.1	Known

4. Related Work

The closest approaches to the one presented here are MedGene[6] and [7]. MedGene uses published literature to extract gene-disease passages, but then does not expand the initial list, ranking the passages using purely statistical methods related to co-occurrence in the text, not biological basis. The extraction tool uses co-occurrence rather than NLP, and user intervention is limited to choosing amongst different statistical ranking formulas. Aside from differences discussed in the previous section, the method in [7] uses an initial gene list from OMIM expanded with interactions from the Online Predicted Human Interaction Database (OPHID). Even though network connectivity is used for showing statistical validity of the method, the scoring formula does not account for it.

5. Conclusion and future work

The method presented here makes an innovative use of a combination of important measures to rank a list of proteins mined from biomedical literature as related to a disease, namely, number of interactions and connectivity impact. It can be a valuable tool in the analysis and exploration of proteins and pathways that relate to a disease. The resulting ranked list of genes and gene products can provide the basis for further focused experiments to investigate the genetic determinants of diseases, as the atherosclerosis study presented here showed.

Focused analysis helps uncover false negatives, and can potentially result in calling attention to yet unexplored genetic linkages and descriptive research on the disease, such as chromosomal aberrations, specific genetic mutations and amplifications that play a role in the disease that can then be investigated through wet lab experiments.

Future work includes improvement of the normalization module to reduce duplicates, since this pollutes the connectivity measures by generating unnecessary nodes, as well as tuning of the formula and methodology with other diseases to potentially incorporate other measures. A web-based interface to allow the public to use the tool is in development.

References

1. Soteriades, E.S. and M.E. Falagas, *Comparison of amount of biomedical research originating from the European Union and the United States.* BMJ: British Medical Journal. , 2005. 331 (7510): p. 192-194.
2. *Online Mendelian Inheritance in Man, OMIM (TM).* 2000, Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, NLM (Bethesda, MD).
3. Lu, Z., K.B. Cohen, and L. Hunter. *Finding GeneRIFs via Gene Ontology Annotations*, in *Pacific Symposium on Biocomputing*. 2006. Maui, Hawaii, USA: World Scientific Publishing Co. Pte. Ltd.
4. Baral, C., H. Davulcu, M. Nakamura, P. Singh, L. Tari, and L. Yu, *Collaborative Curation of Data from Bio-medical Texts and Abstracts and Its integration*. Lecture Notes in Computer Science. 2005. 309-312.
5. Baral, C., H. Davulcu, G. Gonzalez, G. Joshi-Topee, M. Nakamura, P. Singh, L. Tari, and L. Yu, *CBioC: Web-based Collaborative Curation of Molecular Interaction Data from Biomedical Literature*, in *Genetics Society of America 1st Biocurator Meeting*. 2005: Pacific Grove, CA.
6. Hu Y, H.L., Weng H, Zuo D, Rivera M, Richardson A, LaBaer J., *Analysis of genomic and proteomic data using advanced literature mining.* Journal of proteome research., 2003. 2(4): p. 405-412.
7. Chen, J.Y., C. Shen, and A.Y. Sivachenko. *Mining Alzheimer Disease Relevant Proteins from Integrated Protein Interactome Data*. in *Pacific Symposium on Biocomputing*. 2006. Maui, Hawaii, USA: World Scientific Publishing Co. Pte. Ltd.
8. *HUGO Gene Nomenclature Committee Database*. [cited; Available from: http://www.gene.ucl.ac.uk/nomenclature/.
9. Rives, A.W. and T. Galitski, *Modular organization of cellular networks.* PNAS, 2003. 100(3): p. 1128-1133.
10. LaCount, D.J., M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J.R. Hesselberth, L.W. Schoenfeld, I. Ota, S. Sahasrabudhe, C. Kurschner, S.

Fields, and R.E. Hughes, *A protein interaction network of the malaria parasite Plasmodium falciparum.* Nature, 2005. 438(7064): p. 103-107.
11. Ahmed, S.T., D. Chidambaram, H. Davulcu, and C. Baral, *IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for Bio-Medical Text.* , in *BioLINK SIG (Biolink 2005)*. 2005: Detroit, Michigan.
12. Sleator, D. and D. Temperley, *Parsing English with a Link Grammar.* Third International Workshop on Parsing Technologies, 1993.
13. Kostoff, R.N., J.A. Block, J.A. Stump, and K.M. Pfeil, *Information content in Medline record fields.* International Journal of Medical Informatics, 2004. 73(6): p. 515-527.
14. Bader, G., Betel, D., Hogue, C., *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res., 2003. 31: p. 248-250.
15. Zanzoni, A., L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, *MINT: a Molecular INTeraction database.* FEBS Letters, 2002. 513(1): p. 135-140.
16. Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S., Eisenberg, D., *DIP: The Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions.* NAR, 2002. 30: p. 303-5.
17. Hermjakob, H., L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, and R. Apweiler, *IntAct: an open source molecular interaction database.* Nucl. Acids Res., 2004. 32(suppl_1): p. D452-455.
18. Stark, C., B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, *BioGRID: a general repository for interaction datasets.* Nucl. Acids Res., 2006. 34(suppl_1): p. D535-539.
19. Ewens, W. and G. Grant, *Statistical Methods in Bioinformatics: An Introduction*: Springer; 1st edition (April 20, 2001).
20. Thom, T., N. Haase, W. Rosamond, V.J. Howard, J. Rumsfeld, T. Manolio, Z.-J. Zheng, K. Flegal, C. O'Donnell, S. Kittner, D. Lloyd-Jones, D.C. Goff, Jr., Y. Hong, S. Members of the Statistics Committee and Stroke Statistics, R. Adams, G. Friday, K. Furie, P. Gorelick, B. Kissela, J. Marler, J. Meigs, V. Roger, S. Sidney, P. Sorlie, J. Steinberger, S. Wasserthiel-Smoller, M. Wilson, and P. Wolf, *Heart Disease and Stroke Statistics--2006 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee.* Circulation, 2006. 113(6): p. e85-151.
21. Lusis, A.J., *Atherosclerosis.* Nature, 2000. 407(6801): p. 233-241.
22. Bijnens, A.P.J.J., E. Lutgens, T. Ayoubi, J. Kuiper, A.J. Horrevoets, and M.J.A.P. Daemen, *Genome-Wide Expression Studies of Atherosclerosis: Critical Issues in Methodology, Analysis, Interpretation of Transcriptomics Data.* Arterioscler Thromb Vasc Biol, 2006. 26(6): p. 1226-1235.
23. *PubGene*. [cited; Available from: http://www.pubgene.org.
24. Tedgui, A. and Z. Mallat, *Cytokines in Atherosclerosis: Pathogenic and Regulatory Pathways.* Physiol. Rev., 2006. 86(2): p. 515-581.