A Framework for Discovering Associations from the Annotated Biological Web

Louiqa Raschid and Padmini Srinivasan and Woei-Jyh Lee

1 Introduction

During the last decade, biomedical researchers gained access to the entire human genome, reliable high-throughput biotechnologies, and affordable computational resources and network access. In combination, these new tools created a new model for biomedical research that no longer uses computational tools merely to monitor research, but instead exploits these tools to acquire knowledge and make discoveries. Consider a simplified Web of three publicly accessible resources Entrez Gene, OMIM and PubMed, in Figure 1. Data entries in each resource are annotated with terms from multiple controlled vocabularies (CVs). The hyperlinks between data entries in any two resources form a relationship between the two resources and is represented by a (virtual) link. Thus, an entry in Entrez Gene, annotated with GO terms, can have hyperlinks to multiple entries in PubMed that are annotated with MeSH terms.Similarly, OMIM entries, annotated with terms from SNOMED CT may have hyperlinks to entries in Entrez Gene and PubMed. This forms a rich Web of annotated data entries. Our objective in this research is to develop tools to discover meaningful patterns across resources and ontologies. As a first stage in teasing out patterns, we execute a protocol to follow hyperlinks, extract annotations, and generate *LSLink* datasets. We then mine the term-links of the *LSLink* datasets to find *potentially meaningful associations*. Biologically meaningful associations of pairs of CV terms may yield actionable *nuggets* of previously unknown knowledge. Moreover, the *bridge* of associations across CV terms will reflect the *practice* of how scientists annotate data across hyperlinked repositories.

2 Methodology

2.1 *LSLink* Datasets

We identify a *background* dataset associated with a specific experiment protocol. It represents a broad and representative sample of data entries, hyperlinks and annotations. An *LSLink* dataset is a collection of *term-links*. Figure 2 illustrates 3 sample hyperlinks between 2 Entrez Gene and 2 PubMed entries. The hyperlinks are between entries e_1 and p_1, e_2 and p_1, and e_2 and p_2. The terms g_a, g_b, g_c and m_a, m_b, m_c, m_d annotate these entries. Each entry is associated with two terms. If we consider the hyperlink between e_1 and p_1, the two CV terms g_a and g_b annotating

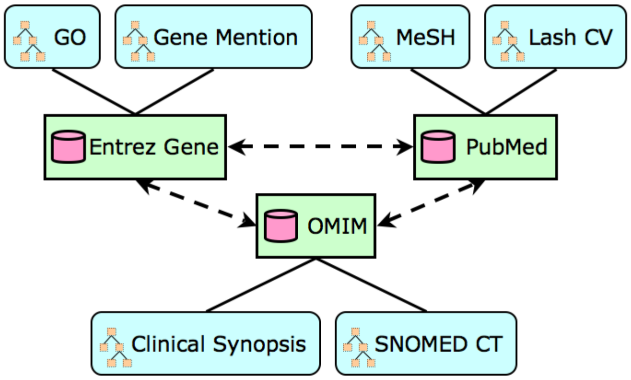

Figure 1: Web of Entrez Gene, OMIM and PubMed Resources

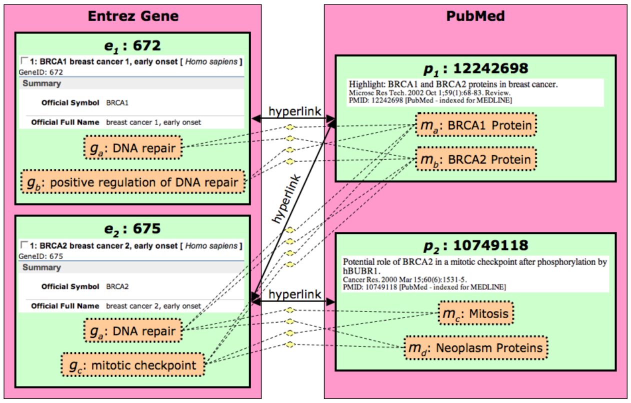
Entrez Gene
PubMed
e_1 : 672
1: BRCA1 breast cancer 1, early onset [*Homo sapiens*]
GeneID: 672
Summary
Official Symbol BRCA1
Official Full Name breast cancer 1, early onset
g_a: DNA repair
g_b: positive regulation of DNA repair
e_2 : 675
1: BRCA2 breast cancer 2, early onset [*Homo sapiens*]
GeneID: 675
Summary
Official Symbol BRCA2
Official Full Name breast cancer 2, early onset
g_a: DNA repair
g_c: mitotic checkpoint
hyperlink
hyperlink
hyperlink
p_1 : 12242698
Highlight: BRCA1 and BRCA2 proteins in breast cancer.
Microsc Res Tech. 2002 Oct 1;59(1):68-83. Review.
PMID: 12242698 [PubMed - indexed for MEDLINE]
m_a: BRCA1 Protein
m_b: BRCA2 Protein
p_2 : 10749118
Potential role of BRCA2 in a mitotic checkpoint after phosphorylation by hBUBR1.
Cancer Res. 2000 Mar 15;60(6):1531-5.
PMID: 10749118 [PubMed - indexed for MEDLINE]
m_c: Mitosis
m_d: Neoplasm Proteins

Figure 2: Sample hyperlinks between Entrez Gene and PubMed

Number of active human gene records in Entrez Gene	38,529
Number of GO annotations extracted	116,513
Number of distinct GO terms extracted	6,177
Number of distinct PubMed records which are reached via four link types	143,450
Number of distinct MeSH descriptors that are major topics	11,419
Number of term-links generated	13,770,651
Number of distinct associations between pairs of CV terms (GO and MeSH pairs)	1,855,992

Table 1: Background *LSLink* dataset of human genes and publications

e_1, and the two CV terms m_a and m_b annotating p_1, then we can generate four *term-links*. An example term-link is the following: $(g_a, m_c, e_2, p_2) = (DNA\ repair, Mitosis, 675, 10749118)$. These 3 hyperlinks from Figure 2 generate 12 term-links. Note that both hyperlinked data entries must be annotated in order to generate a term-link.

Consider a background *LSLink* dataset that includes all term-links generated from the human gene records in Entrez Gene with GO annotations that have hyperlinks to publications in PubMed with MeSH annotations. We limit our protocol to only generate term-links for the MeSH terms identified as major topic headings in the PubMed publications. The statistics for this background dataset as of May 31st, 2007 is reported in Table 1. Details of the specific experiment protocol to generate this background *LSLink* dataset is in [12].

2.2 Relevant Metrics

We propose two classes of metrics to identify significant associations of pairs of CV terms. The first class is based on the logarithm of the odds (LOD) ratio [4, 11, 20] and the second class on the hypergeometric distribution [21, 22].

Notation

- (G, M, E, P) is the background dataset of term-links between entries in Entrez Gene E annotated with terms G from GO that have hyperlinks to entries in PubMed P annotated with terms M from MeSH. $\#(G, M, E, P)$ is the cardinality of the term-links in (G, M, E, P). (G, M, E', P') and $\#(G, M, E', P')$ correspond to the user dataset, a subset of the background dataset that is of interest to a scientist.

- $\#(g_u \wedge m_w, E, P)$ is the cardinality of term-links containing the pair of terms g_u and m_w in the background dataset.

- $\#(g_u \vee m_w, E, P)$ is the cardinality of term-links containing either term g_u or term m_w in the background dataset.

LOD Confidence and LOD Support

The metrics based on the LOD ratio are a measure of the extent to which a specific association of CV terms deviates from one resulting from chance alone (a random association). A random association is one where each data entry in the background dataset is equally likely to be annotated with a particular CV term, and any pair of entries is equally likely to have a hyperlink occurring between them. Using the well known association rule approach [1, 2, 9], we define LOD based confidence and support. The LOD confidence and LOD support metrics that we use include a *term-freq* correction based on the logarithm of the odds (LOD) measure to account for the term frequencies of the associated terms in the background dataset. This *term-freq* correction is novel to our work. We note that given the universe of CV terms and annotations, data entries and hyperlinks between data entries, and term-links, there are many possible approaches to obtain expressions for LOD support and confidence. We have used our judgement to pick some reasonable choices.

- Term probability reflects how commonly a CV term is used to annotate a data entry. Term probability may be estimated using annotation level term frequencies, i.e., by counting up the total number of annotations in some background dataset (annotation level). Alternately, it can be estimated using the cardinality of data entries that are annotated in the background dataset (data entry level). We chose to calculate term probability at the annotation level.
  - $Pr\_term(g_u,E) = \frac{number\ of\ annotations\ that\ are\ g_u\ in\ E}{total\ number\ of\ annotations\ in\ E}$
  - $Pr\_term(m_w,P) = \frac{number\ of\ annotations\ that\ are\ m_w\ in\ P}{total\ number\ of\ annotations\ in\ P}$
- Link annotation probability for the pair (g_u, m_w) estimated from the user query dataset:
  - $Pr\_link(g_u,m_w,E',P') = \frac{\#(g_u \wedge m_w,E',P')}{\#(G,M,E',P')}$
- Conditional link annotation probability for the pair (g_u, m_w) in the user query dataset:
  - $Pr\_cond(g_u,m_w,E',P') = \frac{\#(g_u \wedge m_w,E',P')}{\#(g_u \vee m_w,E',P')}$
- LOD support equals to the logarithm of the link annotation probability divided by the corresponding term probabilities:
  - $LODSupport(g_u,m_w,E',P') = log(\frac{Pr\_link(g_u,m_w,E',P')}{Pr\_term(g_u,E)Pr\_term(m_w,P)})$
- LOD confidence equals to the logarithm of the conditional link annotation probability, given the appearance of either CV term, divided by the corresponding term probabilities:
  - $LODConf(g_u,m_w,E',P') = log(\frac{Pr\_cond(g_u,m_w,E',P')}{Pr\_term(g_u,E)Pr\_term(m_w,P)})$

Hypergeometric Distribution and the $P-$value

The hypergeometric distribution (HG) describes the discrete probability of selecting particular associations of CV terms (g_u, m_w) from a background dataset when sampling items without replacement. The HG distribution gives a quantification of the level of one's *surprise* at finding *over-representation* for a particular item in a given sample of size k drawn from a larger population of size n [6]. The P-value of the HG distribution, when applied to our problem, will provide the expectation of picking at least r term-links annotated with the CV term pair (g_u, m_w), when picking exactly k term-links to create a user query dataset.

Consider a background dataset of $n = \#(G,M,E,P)$ term-links generated from the hyperlinks between data resource E and P annotated with G and M. There are $s = \#(g_u \wedge m_w,E,P)$ term-links containing the specific pair of CV terms (g_u, m_w) in the background dataset. We then consider a user query dataset of $k = \#(G,M,E',P')$ term-links which is a subset of the background dataset. An observation of a term-link with this particular pair of CV terms (g_u,m_w) in the user query dataset is defined to be a success.

- The HG distribution probability and P-value to observe r occurrences of term-links containing the pair (g_u, m_w), given n, s and k, are as follows:
  - $Pr(r|n,s,k) = \frac{\binom{s}{r}\binom{n-s}{k-r}}{\binom{n}{k}}$
  - P-value $= \sum_{q=r}^{min(s,k)} Pr(q|n,s,k)$

The smaller the P-value for the observed r, in the user query dataset, the greater the over-representation of term-links representing an association between the pair of CV terms (g_u,m_w).

2.3 User Dataset and Significant and Meaningful Associations

From the background *LSLink* dataset, we consider a simple *user query* dataset. A user query dataset is a subset of the background *LSLink* dataset that is of interest to a scientist. Here our user query subset are all the term-links associated with an Entrez Gene entry for some gene of interest, e.g., `BRCA1/BRCA2` or `CFTR`. (Complex queries such as gene families will be evaluated in our project.) There can be a potentially large number of possible associations even for a single gene. For example, for `BRCA1/BRCA2`, there were 81,428 term-links and they represented 12,296 distinct associations between pairs of CV terms! Using the LOD confidence and LOD support metrics, one can rank these pairs of associations of CV terms and identify the Top 25 potentially significant pairs for each gene. Experts rated the associations of pairs of CV terms along two independent dimensions, as follows: `(Meaningful, Maybe Meaningful, Not Meaningful)`, and `(Widely Known, Somewhat Known, Unknown/Surprising)`. A majority of the Top 25 pairs of associations for each gene were identified as meaningful or possibly meaningful and widely known or somewhat known (a true positive). Several of the pairs were unknown and might lead to new knowledge. For example, for `BRCA1/BRCA2`, the association of the GO term `negative regulation of centriole replication` with the MeSH term `Fallopian Tube Neoplasms (Neoplastic Process)` might be interesting, because it indicates that the tumor and the negative regulation might have a causal relationship [7]. For `CFTR`, the association of the GO term `ATP-binding and phosphorylation-dependent chloride channel activity` and the MeSH term `Fimbriae Proteins` was also found to be interesting since this is a previously unknown activity of these proteins [18]. The background dataset of term-links from this preliminary study and the potential associations among pairs of GO and MeSH terms are available at the following site: `http://www.cbcb.umd.edu/research/lslink/lodgui/`

3 Research Directions

3.1 Learning Associations from Indirect Links or Paths

Figure 2 illustrates how a hyperlink between two data entries generates a set of term-links. Such hyperlinks are direct links in that there are no intermediate data entries between the pair of annotated data entries of interest. A generalization is to consider *indirect* links between objects. Indirect links are interesting to consider for data entry pairs that are *not* directly linked. As an example, there is no direct hyperlink from human gene *BRCA1 (GeneID:672)* in Entrez Gene to document *PMID:17081976* in PubMed. However the two are indirectly related via an OMIM entry. *BRCA1* points to OMIM record *MIM:113705* which in turn points to Pubmed record *PMID:17081976*. In this example the length of the path, defined as the number of hyperlinks that are traversed to connect the two data entries of interest, is 2.

While we believe that such indirect paths may also yield interesting term-links, we are likely to be less confident in these term-links as compared to the term-links generated from a direct (often curated) hyperlink between data entries, where the path length is 1. We may also consider confidence in the coverage provided by a particular path between two resources. For example, suppose we observe that 60% of the genes in Entrez Gene that have a path to a publication in PubMed via OMIM also have a direct link from Entrez Gene to Pubmed. Then, this 60% coverage can represent a confidence score in any of these term-links generated from this path Entrez Gene to OMIM to PubMed, compared to the term-links from a direct hyperlink from Entrez Gene to Pubmed. We note that determining and combining such *path length* and *confidence* corrections may be non-trivial.

3.2 Learning Associations from Retrieval Links

Another dimension for research is the use of *retrieval links* in augmenting the term-links from hyperlinked data entries. While a given gene object in Entrez Gene may have several hyperlinks to PubMed documents, these hyperlinks may be incomplete in their capability to identify a comprehensive pool of relevant documents. Consider that hyperlinks between Entrez Gene and PubMed are often created by manual, labor intensive protocols that involve intellectual effort. Thus, there may be a significant time lag before these hyperlinks are created. In this interval, many documents that are relevant to a gene may be missed. This could significantly limit our set of interesting term-links and hence interesting association between CV terms.

We will explore methods to augment the pool of hyperlinks with *retrieval links*. We will explore retrieval links specifically in the context of term-links between gene entries and PubMed documents. Retrieval links are supported by our prior research on GeneDocs [8]. GeneDocs is a MEDLINE document ranking system system that targets retrieval for gene queries. GeneDocs takes as input an Entrez Gene id, or a gene symbol. It ranks the retrieved results using its

MeSH descriptor w/ major topic	Number of associated GO terms
Candidiasis, Cutaneous (Disease or Syndrome)	5
Central Nervous System Infections (Disease or Syndrome)	5
Hyperlipoproteinemia Type V (Disease or Syndrome)	5
Tinea Versicolor (Disease or Syndrome)	5
Akathisia, Drug-Induced (Disease or Syndrome)	3
Hyperlipoproteinemia Type III (Disease or Syndrome)	3
Dyslipidemias (Disease or Syndrome)	2
Hyperlipoproteinemia Type IV (Disease or Syndrome)	2
Hyperlipoproteinemias (Disease or Syndrome)	2
Optic Neuritis (Disease or Syndrome)	2
Vitamin K Deficiency (Disease or Syndrome)	2

MeSH descriptor w/ major topic	GO term	Number of term-links
Hyperlipoproteinemia Type V (Disease or Syndrome)	apolipoprotein E receptor binding	1
Hyperlipoproteinemia Type V (Disease or Syndrome)	regulation of axon extension	1
Hyperlipoproteinemia Type V (Disease or Syndrome)	response to reactive oxygen species	1
Hyperlipoproteinemia Type V (Disease or Syndrome)	tau protein binding	1
Hyperlipoproteinemia Type V (Disease or Syndrome)	vasodilation	1

Table 2: Frequency Analysis of MeSH to GO associations for gene APOE.

ranking logic; the user is also given the option to use relevance feedback. The web based system is available at the following site:
http://sulu.info-science.uiowa.edu/genedocs/.

GeneDocs considers gene name synonyms that have been harvested from several sources including Entrez Gene and SwissProt. It also handles the three major varieties of name ambiguity quite successfully. The effectiveness of GeneDocs, especially in dealing with name ambiguity, is shown in [15]. GeneDocs is also based on our prior research on ranking strategies [16, 17].

GeneDocs ranks retrieved documents with a relevance rating in the range (0,1]. We will incorporate the GeneDocs score as a *retrieval rating confidence* score and use this as an additional correction in determining LOD support and LOD confidence. We note that if there is a direct hyperlink to a retrieved document, then the term-links that are obtained will automatically be given the highest rank of 1.

3.3 Tools to Determine Patterns of Annotation

Controlled vocabularies and ontologies are designed to annotate specific classes of objects, e.g., genes, diseases, drugs, etc. They typically have *distinct and independent* orientations, governing bodies, histories and application strategies. Our research will identify significant associations between pairs of CV terms, or an *association bridge*, and may offer users a unique opportunity to identify and perhaps explain patterns in annotation *practice*. We note that this research complements a growing body of research exploring annotation practices within a single ontology [10, 5, 26, 19], as well as research in ontology alignment, matching and integration [3, 14, 23, 24]. Due to space limitations we do not review this literature.

We motivate finding patterns in the cross ontology *bridge* of associations using a simple example. Suppose we consider a user query dataset for the APOE gene and the most significant associations among MeSH and GO terms. We can perform a frequency analysis on the MeSH terms and identify how many GO terms were associated with each MeSH term. Next, for each of the (MeSH, GO) associations, we can report on the number of term-links from the user query dataset. (We note that we can also report on the LOD support or LOD confidence or the $P-$value of the HG metric.) The frequency analysis for a subset of the APOE dataset is in Table 2.

We note that such a frequency analysis, while interesting, only provides partial insight into any patterns in the bridge of associations across ontologies. An extension is to consider a bi-partite graph where the nodes are the set of GO terms and the set of MeSH terms. There will be an edge in the graph if there is a (significant) association between the corresponding pair of CV terms - this forms the *bridge of associations*. Users can customize the bridge by choosing a user query dataset and further selecting a threshold on any of the statistics defined for the association between the corresponding pair of CV terms, e.g., the number of term-links, LOD support or LOD confidence or the $P-$value of the HG metric. Users may also be interested in specific patterns, e.g, a fully connected component between some N GO terms and M MeSH terms.

Our objective is to abstract useful properties and features for inclusion in our tool to better understand the bridge of associations. A good starting point would be simple properties such as clustering coefficients and node degree distributions. We could also provide features to identify components and other patterns as a basis for identifying regions of strong connections in the association bridge across the CVs. Node characteristics such as betweenness centrality [13] might offer an interesting approach to identify important nodes (important CV terms) in the bi-partite graph.

By focusing on the association bridge at the level of independent associations between pairs of CV terms, we are able to simplify the problem of finding patterns. However, we are ignoring key information. The two sets of CV terms participating in the association bridge typically come from ontologies or structured vocabularies. By ignoring their structural properties, we may be loosing valuable insight. Thus, we plan to explore a second level of more advanced functions that reflect these properties. For example, we can use the ontology structure to create CV superterms and thereby aggregate associations in the bridge. CV superterms are associated with the roots of subtrees in the ontology. Subtrees may be selected in different ways. We can also use the class sub-class relationship in the ontology structure to augment associations in the bridge.

4 Conclusion

Our research focus is to identify biologically meaningful and statistically significant associations in the biological Web. We develop multiple metrics to identify potentially significant associations between pairs of CV terms. We will develop a suite of tools for the exploration and evaluation of the *LSLink* datasets, to identify patterns of annotation practice, and to target potentially meaningful associations.

References

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD Record*, 22(2):207–216, June 1993.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceeding of the 20th International Conference on Very Large Data Bases (VLDB 1994)*, pages 487–499, San Francisco, CA, USA, 12-15 September 1994.

[3] D. Aumueller, H. H. Do, S. Massmann, and E. Rahm. Schema and ontology matching with COMA++. In *Proceedings of the 2005 ACM SIGMOD international Conference on Management of Data (SIGMOD 2005)*, Baltimore, Maryland, USA, 13-16 June 2005.

[4] George A. Barnard. Sequential tests in industrial statistics. *Supplement to the Journal of the Royal Statistical Society*, 8(1):1–26, 1946.

[5] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(Database issue):D262–D266, 1 January 2004.

[6] Cristian I. Castillo-Davis and Daniel L. Hartl. GeneMerge post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7), 1 May 2003.

[7] Chi-Ping Day. Personal communiction, 2007.

[8] Gene Ranked Document List (Retrieval for Gene Queries - Sehgal and Srinivasan). `http://sulu.info-science.uiowa.edu/genedocs/`.

[9] Jiawei Han and Micheline Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco, California, USA, 3 November 2005.

[10] Craig E Jones, Alfred L Brown, and Ute Baumann. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, 8(170), 2007.

[11] Jan O. Korbel, Tobias Doerks, Lars J. Jensen, Carolina Perez-Iratxeta, Szymon Kaczanowski, Sean D. Hooper, Miguel A. Andrade, and Peer Bork. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biology*, 3(5), 5 April 2005.

[12] Woei-Jyh Lee, Louiqa Raschid, Padmini Srinivasan, Nigam Shah, Daniel Rubin, and Natasha Noy. Using annotations from controlled vocabularies to find meaningful associations. In *Fourth International Workshop on Data Integration in the Life Sciences (DILS 2007)*, Philadelphia, Pennsylvania, USA, 27-29 June 2007.

[13] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 101(Suppl. 1):5200–5205, 6 April 2004.

[14] M. A. Noy, N. F.; Musen. Anchor-prompt: Using non-local context for semantic matching. *International Journal of Human-Computer Studies*, 59(6):983–1024, 2003.

[15] Aditya K Sehgal and Padmini Srinivasan. Retrieval with gene queries. *BMC Bioinformatics*, 7(220), 21 April 2006.

[16] Aditya Kumar Sehgal and Padmini Srinivasan. Manjal: a text mining system for medline. In *SIGIR*, page 680, 2005.

[17] Aditya Kumar Sehgal, Padmini Srinivasan, and Olivier Bodenreider. Gene terms and english words: An ambiguous mix. In *SIGIR*, 2004.

[18] Nigam Shah. Personal communiction, 2006.

[19] Mary Shultz. apping of medical acronyms and initialisms to Medical Subject Headings (MeSH) across selected systems. *J Med Libr Assoc*, 94(4):410–414, October 2006.

[20] Mir S. Siadaty and William A. Knausg. Locating previously unknown patterns in data-mining results: a dual data- and knowledge- mining method. *BMC Medical Informatics and Decision Making*, 6(13), 7 March 2006.

[21] Robert R. Sokal and F. James Rohlf. *Biometry: the principles and practice of statistics in biological research*. W. H. Freeman, New York, New York, USA, August 1969.

[22] Robert R. Sokal and F. James Rohlf. *Biometry*. W. H. Freeman, New York, New York, USA, 15 September 1994.

[23] G. Stumme and A. Maedche. Fca-merge: A bottom-up approach for merging ontologies. In *JCAI '01 - Proceedings of the 17th International Joint Conference on Artificial Intelligence*, San Francisco, California, USA, 2001. Morgan Kaufmann.

[24] Octavian Udrea, Lise Getoor, and Renée J. Miller. Leveraging data and structure in ontology integration. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 449–460, New York, NY, USA, 2007. ACM Press.

[25] R. Varadarajan, V. Hristidis, and L. Raschid. Explaining and reformulating authority flow queries. In *In preparation*, 2007.

[26] W. John Wilbur. The Dimensions of Indexing. In *AMIA 2003 Annual Symposium*, pages 714–718, Washington, DC, USA, 8-12 November 2003.

[27] Li X, Chen H, Huang Z, Su H, and Martinez JD. Global mapping of gene/protein interactions in PubMed abstracts: A framework and an experiment with P53 interactions. *J Biomed Inform*, 17 January 2007.