Pacific Symposium on Biocomputing 12:343-354(2007)

BIOMEDIATOR DATA INTEGRATION AND INFERENCE FOR FUNCTIONAL ANNOTATION OF ANONYMOUS SEQUENCES

EITHON CADAG[†1,5], BRENT LOUIE[1], PETER J. MYLER[1,2,5], PETER TARCZY-HORNOCH[1,3,4]

Depts. of Medical Education and Biomedical Informatics[1], Pathobiology[2], Pediatrics[3], and Computer Science and Engineering[4], University of Washington, Seattle, WA USA
Seattle Biomedical Research Institute[5], Seattle, WA USA

Scientists working on genomics projects are often faced with the difficult task of sifting through large amounts of biological information dispersed across various online data sources that are relevant to their area or organism of research. Gene annotation, the process of identifying the functional role of a possible gene, in particular has become increasingly more time-consuming and laborious to conduct as more genomes are sequenced and the number of candidate genes continues to increase at near-exponential pace; genes are left un-annotated, or worse, incorrectly annotated. Many groups have attempted to address the annotation backlog through automated annotation systems that are geared toward specific organisms, and which may thus not possess the necessary flexibility and scalability to annotate other genomes. In this paper, we present a method and framework which attempts to address problems inherent in manual and automatic annotation by coupling a data integration system, BioMediator, to an inference engine with the aim of elucidating functional annotations. The framework and heuristics developed are not specific to any particular genome. We validated the method with a set of randomly-selected annotated sequences from a variety of organisms. Preliminary results show that the hybrid data integration and inference approach generates functional annotations that are as good as or better than "gold standard" annotations ~80% of the time.

1. Introduction

The increasing rate of genomic discovery has left biologists with an overwhelming amount of new and tentatively novel genes to examine. One of the first steps in scrutinizing a new genome is to annotate its genes with biochemical characteristics, cellular localization, and other functional properties to quickly identify targets of interest for further study. The re-visitation of "hypothetical" proteins using multiple updated molecular databases can reveal valuable biological information as well. It is estimated that between 25-66% of genes, depending on the organism, are annotated as "hypothetical" [1].

Annotation, however, is often a slow and laborious process, and the complete annotation of even a modestly-sized genome can take a small team of skilled annotators years to finish. Even with a large group of scientists the task remains non-trivial; collaborating scientists working on *Drosophilia*

[†] Corresponding author (ecadag@u.washington.edu)

melonogaster organized a two week "jamboree" to accomplish functional annotation [2]. Coupled with the necessity to maintain currency as sequence information is revised and molecular reference databases are updated, annotation becomes a Sisyphean effort.

Much of the challenge involved in annotating genes stem from scientists needing to consult various molecular databases to ensure complete and thorough annotations. Online data sources such as those furnished by the National Center for Biotechnology Information (NCBI)[a], the Wellcome Trust Sanger Institute[b] and many more made freely available by other researchers have become invaluable in helping annotators assign genes putative functions based on computational results. The nature of how biologic information is stored, *i.e.* in separate, heterogeneous data sources, dictates that data integration is the first step in gene annotation [3]. Information regarding functional properties of genes is fragmented in various online databases which were developed independently and do not inherently interoperate. To annotate genes biologists must manually query many individual data sources.

Considerable research has been done investigating automated methods of annotation, which in addition to alleviating manual efforts have the capability of querying and analyzing a far larger volume of information. While many of the automated annotation systems created thus far are very effective and successful at generating annotations, most are meant as one-off solutions to specific organisms or set of organisms, or utilize only a select number of databases and analyses on which the annotation process is tailored; data integration is frequently *ad hoc*. As the number of molecular databases increases, scalable automated annotation systems will be more necessary.

In this paper we present and evaluate a hybrid approach that addresses both the data integration and analytical needs of gene annotation. Recognizing that an effective annotation system must first be an effective data integration system and that biological expertise is indispensable in developing accurate annotations, we incorporated a robust inference engine on top of an already-existing data integration platform, BioMediator[c]. We identified several promising online biologic databases based on the processes used for model and non-model genome annotation projects and formulated a set of pilot heuristics for the inference engine which would reason over database query results and draw conclusions toward the annotations for submitted sequences.

---

[a] http://www.ncbi.nlm.nih.gov
[b] http://www.sanger.ac.uk
[c] http://www.biomediator.org

To evaluate our methodology, 116 annotated genes were selected randomly from GenBank [4] as a sample set. These genes were re-annotated using our BioMediator-based approach, and our computational annotations were compared to the actual annotations as listed in GenBank. Relying on manual inspection to resolve ambiguity, we found that our automated method yielded functional annotations as good as or better than the listed annotation for 78% of the sample.

2. Related Work

Automated gene annotation is a well-studied subfield of bioinformatics, and many projects have arisen out of the need for expedient gene annotation. Most automated annotation systems rely on a pipeline-based approach [5-7], whereby data is transformed or analyzed step-wise to reach a predicted function. Often the data sources used for the pipeline are replications of publicly available online databases and housed in local data warehouses. Kasukawa *et al*., for instance, relied on a custom annotation pipeline with a well-defined control structure to generate first-pass annotations for the mouse genome, and provided an interface for human curation and modification of automated annotations [5]; Potter *et al*., included a protein annotation pipeline in the ENSEMBL analysis pipeline [7] which assigns InterPro [8] domains to putative proteins after the gene identification stage, derived from species-specific curated data.

Marrying inference to gene annotation systems has also received research attention. Similar to MAGPIE [9], which uses PROLOG to reason over analytical results, FIGENIX uses Java-based JLog to enact intelligent reasoning over specific portions of its annotation pipelines [6]. Like most other automated annotation systems, FIGENIX uses a data warehouse approach to storing information.

In contrast to the automated annotation methods already mentioned, our approach uses a federated database system. The system does not store information locally; rather, queries are sent to the sources, normalized, cleansed and then analyzed in real-time, providing a small client-side footprint. This has the advantage of always providing up-to-date data, a limitation of the aforementioned warehousing approaches [10]. Additionally, data integration is accomplished with the use of a mediated schema, which provides the necessary semantic linkage between data sources as well as a common ontology for the development of general heuristics that are not specific to any single data source or genome. Because of the multi-tiered architecture used for the data integration process, new data sources can be readily added and incorporated into the mediated schema with minimal overhead cost and without a large increase in system complexity such a change might provoke in a pipeline-based system.

And, unlike other systems that rely on inference in annotation, the reasoning system is not restricted by an algorithmic pipeline, and is free to enact rules at arbitrary points in the data gathering process.

3. Methods

3.1. *Identifying Annotation Sources and Heuristics*

To test our combined method of data integration and inference, it was first necessary to select a set of data sources as well as initial logical inference rules to reason over returned information. We created a list of online databases and other resources for use in the process of functionally annotating genomes derived from methods used for the annotation of a set of organisms at the Seattle Biomedical Research Institute (SBRI).

Scientists from SBRI participate in an international effort to sequence and annotate the genomes of three disease-causing parasites, *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi* [11]. While the three genomes share considerable sequence similarity, most of the genes have little homology with genes in other species; approximately 66% of their genomes are annotated as "hypothetical". Additionally, we attempted to emulate the annotation of *Haemophilus influenzae*, the first non-viral genome to be completely sequenced. As such, *H. influenzae* is a far more studied genome than any of the Trypanosomatids. Our experience with these genomes provided the data sources and annotation processes on which we based our system.

Understanding annotation processes for a set of non-model genomes and a model genome gave us interesting results. Many of the data sources relied on by scientists for annotating the Trypanosomatids were based on computational analyses, and with the aid of Perl scripts, submission to multiple analytical services was done in parallel. Parsing through and drawing knowledge from the information, however, was a manual endeavor. Annotators for *H. influenzae*, while also employing some computational services, primarily NCBI's BLAST [12] and domain searches, relied more heavily on literature searches and some species-specific databases. From the sources used by scientists for the aforementioned genomes, a subset was selected to act as the data sources for the evaluation of our automated annotation system: the NCBI BLAST database [12], the NCBI Conserved Domain Database (CDD) [13], Wellcome Trust Sanger's Pfam database [14], PROSITE database [15], Fred Hutchinson Cancer Research Center's BLOCKS database [16] and the ProDom database [17].

Information on how to apply expert knowledge on returned data was also elicited from scientists, and provided the basis for initial logical inference rules. For example, heuristics provided by one scientist working on the

Trypanosomatid genomes noted that in examining BLAST scores, it was not necessarily preferable to use the top-scoring results because best BLAST hits are not always the closest relation to the sequence in question [18].

3.2. *Data Integration for Annotation with BioMediator*

The BioMediator data integration system is the querying, retrieval and normalization platform for our automated annotation method. Developed at the University of Washington, BioMediator is a general-purpose biologic data integration system whose adaptability to various biomedical domains has been demonstrated in the past by providing a data integration platform for linking expression array data with analytics software and uniting disparate neuroscience databases to identify locations in the cortex related to language processing [19-21].

A federated data warehouse that queries sources in real-time, BioMediator relies on a multi-tiered architecture whose core is a mediated schema that translates data from heterogeneous data sources into entity instances from the schema, thus collecting all query results under a single semantic framework (see Figure 1).

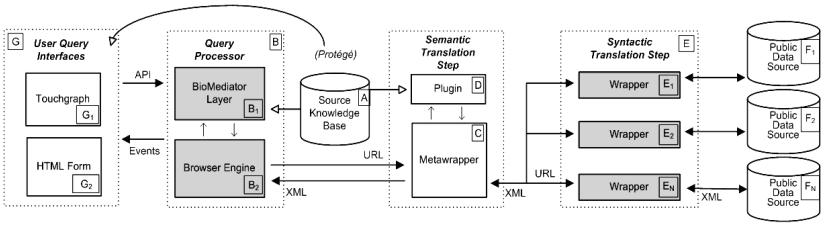

Figure 1. Diagram of BioMediator's architecture; data comes from sources (far right, F) via wrappers (E), which serialize data to schema-mapped XML (A) via the metawrapper (C,D) layer and sent to the BioMediator query processor (B) and interface (G). Original image adapted from [c].

We manually created a mediated schema for generalized, non-genome-specific functional annotation using the Protégé[d] ontology editor, wrappers to serialize data from the sources and source-to-schema mappings. During the evaluation of our annotation system, the schema contained 57 entities to represent data across genomic databases (*e.g.* 'Protein', 'ProteinDatabaseHit') and 55 binary relationships between those entities (such as 'ProteinHasProteinDatabaseHit' to describe a protein homology relationship).

[d] http://protege.stanford.edu

3.3. *Heuristics for Anonymous Sequence Annotation*

Utilizing BioMediator's plug-in architecture, we added the Java Expert System Shell (Jess) rule engine [22] to BioMediator, giving us the capability to formulate flexible sets of rules against mediated result sets. Unlike other previous annotation systems that employ rule engines to manage pipelines or make decisions based on analyses, our approach to integrating Jess into BioMediator does not compartmentalize the scope of the rule engine by limiting when or where the rules may fire; the Jess component is free to enact rules over any data as it enters the system piecemeal, after all data is loaded in aggregate or any combination thereof and treats all received data as part of the working memory. As a result of our approach, rules are applied in a consistent fashion for all annotations.

For our evaluation, three classes of rules were created to emulate as best as possible some of the annotation processes used by the genome annotators at SBRI. A total of 16 rules[e] were developed for the pilot evaluation of our system (see Figure 2 for rule example).

```
(defrule evalue-threshold-homologs
  (threshold (type evalue) (max ?M) (db ?D))
  ?F <- (homolog-pattern (evalue ?X&:(>= ?X ?M))
        (db ?B&:(eq ?B ?D)) (property ?P))
 =>
  (delete-reason ?P ?F "High expect value"))
```

Figure 2. Example rule that prunes homologs from the result set that do not pass a threshold. Specific thresholds for individual databases may be optionally set, and the final line above saves the reason for the removal of the record.

3.3.1 *Filtering Rules*

Filtering rules are heuristics that were limited to strictly ruling out possible annotations or other relevant data from further use by the inference engine. Rules that examined quantitative values for a minimum threshold, for instance, fell under this classification.

Also, based on techniques utilized by the FANTOM2 annotation pipeline [5], a filtering rule for the perceived quality of information was created. 12 regular expressions whose patterns indicate a possibly uninformative annotation were used. For example, homologous proteins that contained "unnamed" in their functional annotation were removed from further consideration. Data classified as removed does not leave the working memory; rather, they are restructured so that the reason for their removal is noted, and can be retrieved again if need be.

[e] Supplementary material on rules at: http://www.biomediator.org/publications

3.3.2 *Evidence-Building Rules*

The second class of rules uses information returned to increment evidence levels of tentative annotations. Homologous proteins enter the system as working memory with a low evidence level. As evidence is found to support that protein annotation (*e.g.* corroborating domains or large number of similar protein annotations returned), its evidence level is increased. This rule is analogous to the confidence classification system used by scientists annotating the *H. influenzae* genome at SBRI, with an ordinal scale to represent the level of evidence. Domains that recur in working memory multiple times, for example, may have their evidence level increased, as their likelihood to be associated with a target sequence is improved; likewise, functional annotations that are correlated with domain support will also reflect an increase in evidence.

Because our initial annotation system does not yet make use of formal biomedical vocabularies, such as the Gene Ontology [23] (GO), and there is no universally-accepted nomenclature in practice for all genomic databases, we establish correlations between the text of functional annotations provided by our data sources using a modified edit distance algorithm. Consider two strings, *k* and *l* with lengths *m* and *n* respectively; a matrix *G* of (*m* + 1) x (*n* + 1) is created, where row 0 is initialized to 0...n, and column 0 initialized to 0...m. The remaining positions in the matrix, G(*i*,*j*) are computed by[f]:

$$min(G(i-1,j)+1, G(i,j-1)+1, G(i-1,j-1)+c)$$
$$c = \begin{cases} 0, & char(k,i) = char(l,j) \\ 1, & otherwise \end{cases} \quad \text{(Eq. 1)}$$

The value given by *G*(*m*,*n*) / *q* is the phrase similarity measure we use between *k* and *l*, where *q* is the length of the longer of the two strings. In our annotation system, various evidence-building rules invoke this string-comparing algorithm, such as when protein homologies share similarly-phrased annotations.

3.3.3 *Annotation Selection Rules*

The third classification in our initial rule-base is those that select likely functional annotations from the working memory, based on evidence levels. All possible annotations are stratified by their level of evidence; related annotations are percolated to the top of the list if they appear repeatedly, and the highest-level annotation with the greatest amount of evidence is provided as the automated functional annotation, though the remaining possible annotations are

[f] Where *char*(*k*,*i*) represents the i^{th} character in the string *k*

available for viewing as well. If no annotation is available at any evidence level, the default "hypothetical" result is presented.

3.4. *Evaluation*

To evaluate the efficacy of our BioMediator-based automated annotation system, we randomly selected 116 genes from a local copy of the GenBank database from April 2006 [4]. The GenBank annotations for the 116 genes served as our "gold standard". We parsed out species names so that results from the source organism could be excluded from query returns; protein sequences from 58 bacteria, 31 eukaryotes, three viruses and one archaea were represented.

Once the genes were annotated by our system, the automated annotation and actual annotation were compared and individual automated annotation results scored as incorrect, correct but inferior to actual, same as actual or superior to actual. This quaternary scoring rubric was adapted to adjust for the known danger of outdated or incorrect GenBank annotations [24]. We used two measures in scoring, specificity and utility. Specificity is in reference to the level of granularity and precision provided in the annotation, *e.g.* "peptidase" would be a less specific annotation in comparison to "lysosomal cysteine-type endopeptidase", provided both are correct. Utility was used as a measure to compare how informative annotations are based on the textual content. An annotation that is based on a GO term, for example, would be considered more informative than one that uses idiosyncratic nomenclature. In cases where the automated annotation did not match the actual annotation, we used manual annotation methods and referred our findings to a domain expert for final scoring.

4. Results of Automated Annotation Using Inference

Our evaluation showed the automated annotations had specificity at the same level, or better, than the GenBank annotations 78% of the time. Additionally, the automated annotation was equal to or more informative than the GenBank annotation in 85% of the sample genes. As putative genes from non-model organisms are generally less likely to register sequence similarity hits in databases versus well-studied model organisms, we also compared the systems performance along a model- and non-model organism stratification as determined by the NCBI Model Organisms Guide [25] (see Table 1). Of the 116 automated annotations generated, seven were deemed to be incorrect when compared to the GenBank annotations. Upon manual inspection, reasons for the system assigning incorrect annotations were attributable to either a) the genes having short sequences, and were subsequently expunged by expect-value rules,

or b) pertinent information returned originated from the organism which the sequence was taken and were thus pruned out.

Table 1. Results of automated annotation in comparison to GenBank annotations.

	Model Organisms (n=56)				Non-model Organisms (n=60)			
	Wrong	Worse	Same	Better	Wrong	Worse	Same	Better
Spec.	2 (3.6%)	8 (14.3%)	30 (53.6%)	16 (28.6%)	5 (8.3%)	10 (16.7%)	37 (61.7%)	8 (13.3%)
Util.	2 (3.6%)	6 (10.7%)	41 (73.2%)	7 (12.5%)	5 (8.3%)	4 (6.7%)	42 (70.0%)	9 (15.0%)

Individual results varied in quality and nomenclature. The databases we relied on as sources did not share a common terminology so semantically equivalent, though syntactically different, annotations were commonplace. In some cases, lower evidence levels provided superior annotations than those at higher evidence levels, though we used the highest evidence level presented in scoring. In seven cases, the automated annotation system presented a function for a gene for which GenBank records show either none or list "hypothetical". Manual annotation indicated that there was evidence in four of the seven to suggest that the automated annotation was correct; for the remaining three annotations some evidence suggested their correctness, though their true annotation remained relatively ambiguous (see Table 2 for example results).

Table 2. Selected automated annotation results juxtaposed with actual annotations from GenBank, with notes.

Automated Annotation	Actual Annotation	Notes
Hypothetical	Ribosomal protein L34	Sequence was small; relevant entries removed by expect-value rules
Anion exchange transporter	SLC26A5 protein	Automated is less specific but more informative
Nicotinic acetylcholine receptor alpha4 subunit	Unnamed protein product	Evidence for automated is very convincing; affirmed with manual inspection
COG4619: ABC-type uncharacterized transport system, ATPase component	ABC-type uncharacterized transport system, ATPase component	Automated and actual match, controlled vocabulary used
GTP-binding protein RAB4	PREDICTED: similar to ras-related GTP-binding protein 4b	Annotations are essentially the same, but varying naming conventions used

5. Discussion

The framework and methodology on which we base our approach to gene annotation is unique from previous automated gene annotation solutions. By using BioMediator as a data integration platform to handle sequence queries and

retrieve results, we avoid the overhead involved with maintaining large repositories that replicate already-available data sources; responsibility for updating the data sources we use falls on the originators of the source data itself, and generally remove users of our system from most maintenance tasks. Because of the system's relatively small memory and processor footprint, it can be used on the desktop computers of annotating scientists. BioMediator's tiered architecture also allows us to add and remove sources with relative ease, and without the effort often necessary in warehouse systems, where database schemata and workflows may need to be altered considerably as data sources and tasks change over time. Scientists researching a novel genome, for example, could map any local in-house databases to the databases linked to BioMediator, thereby rapidly integrating their species-specific data with any sources already supported by BioMediator.

Also, building the inference system around the schema rather than individual sources afforded us a method of quickly developing annotation rules without having to necessarily address each data source individually. Inference rules are also a natural, transparent way of capturing annotator knowledge. Once the rules were conceived, the development time in Jess was rapid. It is important to note, though, that our results were obtained using a set of rules that were not tuned or optimized, and thus we expect results will be better as rules are improved based on feedback from annotators.

The scalability and flexibility of our approach, however, did come at cost, and online data sources do experience downtime. While testing the system, one of our sources was unavailable for several hours. Theoretically, we hope that by utilizing many more sources in the future that have partial redundancy, the loss of any single source may be somewhat offset. Still, as a federated data system, our ability to retrieve data is subject to the real-time availability of the data sources.

An important handicap was that we did not rely on a structured ontology such as GO for our initial evaluation. While the schema we utilized was ontology-based, none of the sources we relied on used any controlled vocabulary on a consistent basis. Phylogenic information was not represented in our evaluation, and could have provided valuable data in relating evolutionary linkage to target sequences. Despite these shortcomings, the initial evaluation of our annotation system and methodology gives encouraging results; the efficacy of our approach is comparable to that of a previously evaluated species-specific and pipeline-based automated annotation system, 75.1-78.6% estimated accuracy for FANTOM2 [5], with the additions of being non-specific to any genome and having an architecture oriented toward scalability.

6. Conclusion

The growing size, disparity and heterogeneity of biologic data and the necessity for expert curation in determining the protein functions for the myriad of newly sequenced genomes means that an automated annotation system that can address future gene annotation requirements must to be both a robust data integration platform and a powerful expertise-based system. In this paper, we have presented a technique and framework that couples the two important tasks in gene annotation into a cohesive platform, and evaluated its performance.

Future iterations of the system will annotate genes using a controlled vocabulary with the addition of data sources such as InterPro which regularly and consistently include GO terms in their records. While our initial system relies on online databases, incorporating analytical services like transmembrane-locating or phylogeny-inferring software into the schema and developing rules to take advantage of such information would be a valuable addition. Alteration of current rules will also improve our annotation capabilities, such as a dynamically-determined threshold to account for sequences of variable length.

Additionally, in the future, we hope to evaluate our system against more ongoing genome annotation projects, to compare automated annotation results with further manually-created annotations. The true test of our system would be to annotate a novel genome in parallel with expert scientists.

Acknowledgements

This work is supported by NHGRI grant R01HG02288 and the National Library of Medicine training grant T15LM07442. The authors of this paper would like to acknowledge Elizabeth Worthey and Alice Erwin for lending their knowledge of annotation to our research, as well as Ron Shaker, Janos Barberos and Dhileep Sivam for their technical assistance.

References

1. Worthey, E., Myler, P., *Protozoan genomes: gene identification and annotation.* International Journal for Parasitology, 2005(35): p. 495-512.
2. Adams, M., Celniker, S., *et al*., *The Genome Sequence of Drosophilia melanogaster.* Science, 2000. **287**(5461): p. 2185-2195.
3. Garrels, J.I., *Yeast genomic databases and the challenge of the post-genomic era.* Functional & Integrative Genomics, 2002. **2**(4-5): p. 212-237.
4. *GenBank*. 2006 [cited April 2006]; Available from: http://www.ncbi.nlm.nih.gov/Genbank/
5. Kasukawa, T., Furuno, M., *et al*., *Development and Evaluation of an Automated Annotation Pipeline and cDNA Annotation System.* Genome Research, 2003. **13**.
6. Gouret, P., Vitiello, V., *et al*., *FIGENIX: Intelligent automation of genomic annotation: expertise integration in a new software platform.* BMC Bioinformatics, 2005. **6**.

7. Potter, S., Clarke, L., *et al*., *The Ensembl Analysis Pipeline.* Genome Research, 2004. **14**.
8. Apweiler, R., Attwood, T., *et al*, *The InterPro database, an integrated documentation resource for protein families, domains and functional sites.* Nucleic Acids Research, 2001. **29**(1): p. 37-40.
9. Gaasterland, T., Sensen, C., *MAGPIE: automated genome interpretation.* Trends in Genetics, 1996. **12**(2): p. 76-78.
10. Louie, B., Mork, P., *et al*., *Data Integration and Genomic Medicine.* Journal of Biomedical Informatics, 2006.
11. El-Sayed, N., Myler, P., *et al*., *Comparative Genomics of Trypanosomatid Parasitic Protozoa.* Science, 2005. **309**(5733): p. 404-409.
12. Altschul, S., Gish, W., *et al*., *Basic Local Alignment Search Tool.* Journal of Molecular Biology, 1990. **215**.
13. Marchler-Bauer, A., Anderson, J., *et al*., *CDD: a Conserved Domain Database for protein classification.* Nucleic Acids Research, 2005. **33**(D): p. 192-196.
14. Bateman, A., Coin, L., *et al*., *The Pfam protein families database.* Nucleic Acids Research, 2004. **32**(D).
15. Hulo, N., Bairoch, A., *et al*., *The PROSITE database.* Nucleic Acids Research, 2006. **34**(D): p. 227-230.
16. Henikoff, S., Henikoff, J., *Protein family classification based on searching a database of blocks.* Genomics, 1994. **19**(1): p. 97-107.
17. Corpet, F., Gouzy, J., *et al*., *The ProDom database of protein domain families.* Nucleic Acids Research, 1998. **26**(1): p. 323-326.
18. Koski, L., Golding, B., *The Closest BLAST Hit Is Often Not the Nearest Neighbor.* Journal of Molecular Evolution, 2001. **52**: p. 540-542.
19. Donelson, L., Tarczy-Hornoch, *et al*., *The BioMediator System as a Data Integration Tool to Answer Diverse Biologic Queries.* Proceedings of MedInfo, IMIA, 2004.
20. Wang, K., Tarczy-Hornoch, P., *et al*., *BioMediator Data Integration: Beyond Genomics to Neuroscience Data*. in *American Medical Informatics Association 2005 Symposium Proceedings*. 2005.
21. Mei, H., Tarczy-Hornoch, P., *et al*., *Expression Array Annotation Using the BioMediator Biological Data Integration System and the BioConductor Analytic Platform*. in *American Medical Informatics Association 2003 Symposium*. 2003.
22. *Jess, the Rule Engine for the Java Platform*. 2006 [cited 2006]; Available from: http://herzberg.ca.sandia.gov/jess/
23. Ashburner, M., Ball, C., *et al*., *Gene ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.
24. Harris, J., *Can you bank on GenBank?* Trends in Ecology and Evolution, 2003. **18**(7): p. 317-319.
25. *National Center for Biotechnology Information, Model Organisms Guide*. 2006 June 2006 [cited; Available from: http://www.ncbi.nih.gov/About/model/index.html