U.S. DEPARTMENT OF ENERGY

PNNL-16211

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Aligning Ontologies and Integrating Textual Evidence for Pathway Analysis of Microarray Data

B Gopalan
C Posse
AP Sanfilippo
M Stenzel-Poore
SL Stevens
J Castano
N Beagley
RM Riensche
B Baddeley
RP Simon
J Pustejovsky

October 2006

Pacific Northwest
NATIONAL LABORATORY

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

This document was printed on recycled paper.
(9/2003)

PNNL-16211

Aligning Ontologies and Integrating Textual Evidence for Pathway Analysis of Microarray Data

B Gopalan[1]
C Posse[1]
AP Sanfilippo[1]
M Stenzel-Poore[2]
SL Stevens[2]
J Castano[3]
N Beagley[1]
RM Riensche[1]
B Baddeley[1]
RP Simon[4]
J Pustejovsky[3]

[1] Pacific Northwest National Laboratory
[2] Oregon Health and Science University
[3] Brandeis University
[4] Neurobiology Research Legacy Clinical Research and Technology Center

October 2006

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

ABSTRACT

Expression arrays are introducing a paradigmatic change in biology by shifting experimental approaches from single gene studies to genome-level analysis, monitoring the expression levels of several thousands of genes in parallel. The massive amounts of data obtained from the microarray data needs to be integrated and interpreted to infer biological meaning within the context of information-rich pathways. In this paper, we present a methodology that integrates textual information with annotations from cross-referenced ontologies to map genes to pathways in a semi-automated way. We illustrate this approach and compare it favorably to other tools by analyzing the gene expression changes underlying the biological phenomena related to stroke. Stroke is the third leading cause of death and a major disabler in the United States. Through years of study, researchers have amassed a significant knowledge base about stroke, and this knowledge, coupled with new technologies, is providing a wealth of new scientific opportunities. The potential for neuroprotective stroke therapy is enormous. However, the roles of neurogenesis, angiogenesis, and other proliferative responses in the recovery process following ischemia and the molecular mechanisms that lead to these processes still need to be uncovered. Improved annotation of genomic and proteomic data, including annotation of pathways in which genes and proteins are involved, is required to facilitate their interpretation and clinical application. While our approach is not aimed at replacing existing curated pathway databases, it reveals multiple hidden relationships that are not evident with the way these databases analyze functional groupings of genes from the Gene Ontology.

INTRODUCTION

A detailed knowledge of the functional roles played by the expressed genes in the context of the biological pathways and networks is essential to identify the role of these genes in different biological processes. The number and diversity of genes exceed the ability of any single investigator to track the complex relationships established by the data sets. Therefore additional tools are required for data analysis. At the same time, the massive amount of data obtained from the microarray data needs to be integrated and interpreted correctly and further information is required to achieve this goal.

The clustering of gene expression data based on common experimental features is a widely used technique for automating the analysis of gene expression patterns. However, these clustering techniques often yield too many groups, thus generating a crowded information environment in which relevant gene clusters and their functionality are not easily identified. A further limitation of current gene expression analytic approaches such as clustering is that

they do not incorporate comprehensive background knowledge about the genes—i.e., functional genomic information—into the analysis. Functional genomics is concerned with the establishment of a verifiable link between gene expression and cell/organ/tissue function and dysfunction. Different tools have been proposed for pathway analysis (or functional enrichment), however they can be limited to specific organisms.

There are two main sources of functional genomic evidence that can be used to support the automated identification of gene expression patterns from microarray data: textual evidence from relevant biomedical literature and the Gene Ontology. The exploitation of these sources presents great opportunities for biological discovery.

The amount of information about biological entities and their interactions is huge and growing fast. The biomedical literature database MEDLINE indexes over 12 million articles, and it is growing by over 2000 daily. The traditional practice of manually retrieving gene information from biomedical document repositories (e.g., PubMed) and then filtering this information for significant and meaningful relationships is very time-consuming, usually requires biological experts, and often is not very successful. Automatic information extraction from online biomedical articles and abstracts can provide a more efficient and comprehensive way of retrieving functional information about gene groups. For example, Raychaudhuri et al. (2003a, 2003b) show that genes can be clustered into functionally related groups by using textual evidence automatically extracted from biomedical articles and abstracts. However, to maximize the utility of information extraction for the elicitation of biomedical

evidence, it is necessary to construct a reliable model for Entity Recognition and Relation Identification, as described in Pustejovsky et al. (2002), that is capable of establishing

- equivalences across identical biological entities that can appear under diverse names (e.g., SPP1 = OPN = BNSP = Secreted phosphoprotein 1 = osteopontin)
- relevant links among them (e.g.,OPN is expressed during embryogenesis).

Moreover, the extraction process must be tuned to the focus problem at hand, in order to increase relevance and precision.

The Gene Ontology (GO, http://www.geneontology.org) provides three orthogonal networks of about 17,000 classificatory terms (GO codes) structured through semantic relationships such as inheritance (IS-A) and meronymy (PART-OF). GO codes encode biological process (BP), molecular function (MF) and cellular component (CC) properties of gene and gene products. Several web-based GO tools (http://www.geneontology.org/GO.tools.shtml) are now available that allow users to classify genes from gene expression microarrays (or other experimental studies). The Gene Ontology has also been used as a resource for measuring semantic similarity between gene products (see Lord et al. 2002, 2003; Couto et al. 2003; Azuaje et al. 2005, Bodenreider et al. 2005, Posse et al. 2006).

The Gene Ontology clearly constitutes a very important functional genomic knowledge resource and provides a structure for organizing genes into biologically relevant groupings. However, in its current form, it presents several critical limitations. For example, there are no associative relations across the three ontologies capable of indicating that a cellular component is the location of a molecular function and that a molecular function is involved in a biological process. The lack of such associative relations weakens reasoning power across GO codes, as recognized by Bada et al. (2004) and Bodenreider et al. (2005). Also, while GO contains about 200 terms corresponding to "pathways", it does not address the relations between processes/functions and pathways. As Bodenreider et al. (2005) show, associative relations can be inferred through a variety of statistical techniques that estimate the similarity of two GO codes inter-ontologically in terms of the distribution of the gene product annotations associated with the two GO codes in the GO database. However, as Posse et al. (2006) point out (see also section C3), these methods yield associative similarity values that are not commensurate with values obtained through methods that estimate GO code similarity intra-ontologically (Lord et al. 2002, 2003; Couto et al. 2003; Azuaje et al. 2005). Therefore, intra- and inter-ontological GO code similarity cannot therefore be used in conjunction with ensuing reduced inference power.

Finally, while there is an upsurge in the number of tools that use the Gene Ontology and Information Extraction techniques to assign biological significance to gene expression data, relatively little work has been devoted to the development of analytic methods that integrate GO annotations and textual evidence. This integration can be very powerful, especially in cases where one source of evidence alone is not sufficient to assist in the acquisition of the relevant biomedical knowledge.

The goal of the high-throughput data analysis such as microarray data is to infer biological meaning within the context of information-rich pathways. While GO has been used to annotate microarray data both by hand and by some software packages (Doniger et al. 2003), there has been no automated way to use it for pathway-based analysis except MAPPFinder, a tool that dynamically links gene-expression data to the GO hierarchy (http://www.genmapp. org/MAPPFinder.html). There are several other pathway databases like Metacore™ (www.genego.com) that use the Gene Ontology to map the functional processes onto pathway maps and signaling networks, but have limited functionality, e.g. simple look-up of existing annotations and significance analysis of the categories found. Moreover, none of these tools take into account one major limitation with the current GO structure: the gene ontologies address only one knowledge domain at a time and are not cross-referenced, despite the obvious similarities between terms across the three gene ontologies. This could not only lead to a biased representation of functional processes, but also increase the chance of missing important relationships between genes in the presence of evidence supporting such relationships.

The system we propose addresses these inadequacies by mapping genes to pathways in a semi-automated way by harnessing the combined functionality of (1) XOA, a methodology for computing semantic similarities between the gene annotations from cross-referenced gene ontologies, and (2) Medstract, an information extraction approach tuned to the biomedical domain. This system complements existing curated commercial pathway databases by revealing relationships that remain inscrutable for the GO-based functional analysis of gene grouping used in current pathway databases.

Stroke: a specific problem

Stroke is the third leading cause of death and the leading cause of disability in the United States. Of approximately 750,000 people afflicted, 158,000 die annually. The incidence of stroke is predicted to increase to over one million per year by 2050. The potential for neuroprotective stroke therapy is enormous, even if it requires treatment prior to ischemic event. For instance, 50% of patients who undergo coronary artery bypass surgery suffer permanent cognitive decline from intraoperative emboli. Preoperative treatment of such patients (336,000 annually) could reduce stroke incidence and morbidity.

Biologists at Oregon Health and Science University have developed a mouse model of neuroprotection in stroke and have carried out gene expression profiling studies to identify potential neuroprotective genes and their associated pathways (under a Program Project Grant funded by NINDS, PO1-NS035965). The goal of this work is to determine the endogenous molecular mechanisms involved in LPS preconditioning by examining the genomic changes associated with LPS-induced neuroprotection against focal ischemia. Studies have shown that LPS preconditioning, via prior systemic administration of low doses of LPS, induces marked neuroprotection against subsequent stroke injury. The cellular mechanisms of neuroprotection induced by preconditioning offer attractive targets for the development of therapeutic approaches (Stenzel-Poore et al. 2003). However, in order to uncover such mechanisms, it is necessary to develop a holistic systems perspective capable of providing a better understanding of the roles played by both the trigger (LPS) and its outcome (the expressed genes). For example, it is important to know how and in what experimental conditions or biological systems (e.g., heart, liver, brain) endotoxin (i.e., LPS) serves as a neuroprotectant. Deciphering the relationship among genes with similar regulation pattern, such as genes upregulated in experimental cohorts pretreated with LPS, can lead to the discovery of genes that act as mediators of the protective phenotype. For example, by identifying the central mediators involved in LPS preconditioning, we may be able to define essential pathways responsible for this important cellular program and thereby provide novel therapeutic strategies in stroke. Our hypothesis, based on these preliminary studies is that knowledge about relationships among genes is instrumental in predicting how LPS may be regulating these genes and how these genes may be involved in neuroprotection, with consequent identification of which genes act as mediators of the protective phenotype. A detailed knowledge of the functional roles played by the expressed genes in the context of the biological pathways and networks that lead to cell death or cell survival is essential in order to identify the biomarkers of stroke and uncover the endogenous mechanisms of neuroprotection.

METHODOLOGIES

Information Extraction using Medstract

Medstract (Pustejovsky et al. 2002) comprises a set of robust natural language processing (NLP) tools for the automated extraction of unstructured information and the creation of bio-entities and bio-relations from Medline publications. We outline the general architecture of these tools for developing databases for domain-specific information servers. Currently, two databases have been built: a Bio-Acronym Server, Acromed, and an Inhibitory-Relation Server. Acromed results from the following processing of Medstract NLP modules:

1. Preprocessing: tokenization, tagging, stemming
2. Shallow parsing and entity recognition and relation identification
3. Semantic typing: semantic tag look-up and type composition
4. Acronym and abbreviation recognition
5. Anaphora and coherence resolution
6. Database normalization.

The relation identification module was developed independently of the specifics of how the particular relation (e.g., *inhibit*) and associated nominals behave in Medline. This module was defined and designed to work on the output of the shallow parsing module to identify argument and relational chunks, independently of any specific lexical item. The extraction of a particular relation (e.g., *inhibit or regulate*) is accomplished by identifying lexical items that denote the target relation. This task is subdivided in two parts:

1. Sentence-level parsing identifies the predicate and the subject and object chunks as well as subordinate clauses and coordination
2. Nominal level parsing identifies relations within a noun phrase.

Using this machinery, a bio-relational database was built which contains regulatory and inhibitory relations of bio-entities (proteins, cellular processes, etc.) extracted from Medline data (2001 distribution). This database contains over 6,969,000 relations. The system has been measured at a performance of 90% precision, 59% recall, and 22% partial recall (Pustejovsky et al. 2002). The user can select the type of relation to be searched (i.e., *inhibit*, *regulate*, etc.) as well as the Unified Medical Language System (UMLS) type (e.g., *gene*, *amino acid*) or name of the bio-entities, which are the arguments to the relation. As a result of the search, all biological relations relevant to the specified bio-entities are returned in either the form of a database table or a navigable hyperbolic graph. Both forms link directly to the citations from the abstracts. The server is in beta release and is being extended to other types of biological relations (e.g., *acylation*, *phosphorylation*).

Cross-Ontological Analytics (XOA)

Until recently two orthogonal GO-based similarity approaches have been used. One approach assesses GO code similarity in terms of shared hierarchical relations, within each gene ontology (BP, MF, or CC) (Lord et al. 2002, 2003; Couto et al. 2003; Azuaje et al. 2005). For example, the relative semantic closeness of two biological processes would be determined by the informational specificity of the most immediate parent that the two biological processes share in the BP ontology. The other approach establishes GO code similarity by leveraging associative relations across the three gene ontologies (Bodenreider et al. 2005). Such associative relations make predictions such as which cellular component is most likely to be the location of a given biological process and which molecular function is most likely to be involved in a given biological process.

While these two approaches are fully complementary, no effort has been made so far to combine them. One major difficulty in carrying out such a combination resides in the heterogeneity of the two measures, one based on a hierarchical assessment and the other on an associative one. This means that results relative to inter-ontological similarity relationships are not commensurable with intra-ontological ones. Posse et al. (2006) present a methodology, XOA for Cross-Ontological Analytics, where this difficulty can be solved so that the two approaches can be integrated, with ensuing benefits in coverage and accuracy.

XOA provides a GO-based similarity algorithm capable of combining intra- and inter-ontological relations by "translating" each associative relation across the gene ontologies into a hierarchical relation within a single ontology, so that all GO similarities can be computed as intra-ontological relationships and therefore yield commensurable scores. More precisely, let c1 (c2) denote a GO code in the gene ontology O1 (O2), the XOA similarity score between c1 and c2 is defined as follows:

$$XOA(c1, c2) = \max\{sim(c1, c3)*\cos(c2,c3), sim(c2,c4)*\cos(c1, c4)\}$$

where cos(ci,cj) denotes the cosine associative measure proposed by Bodenreider et al. (2005), sim(ci,cj) denotes any of the three intra-ontological semantic similarities used by Lord et al. (2002, 2003), and Azuaje et al. (2005), that is, the similarity measures proposed by Resnik (1995), Jiang and Conrath (1997) and Lin (1998), and the maximum is taken over all GO codes c3 in O1 and c4 in O2.

In the next section, we present a case study of how an integrated analysis of XOA and Medstract reveals biologically meaningful relationships that are not identified with other data analysis approaches, including microarray analysis.

Case Study: Using XOA and Medstract to relate genes to the TGF-beta signaling pathway

In an effort to determine neuroprotective pathways associated with LPS preconditioning, we performed bioinformatics analysis, manual literature and database searches on the genes regulated in our microarray analysis. Promoter analysis of the microarray data, using PAINT (Vadigepalli et al. 2003), together with our data mining and manual literature curation, revealed a distinct role for TGFβ in LPS preconditioned mice that tips the balance in favor of SMAD1 signaling and away from SMAD3 pathways which may be harmful. The TGFβ superfamily of proteins have been associated with improved outcome following brain ischemia, therefore we were interested in determining if TGFβ may be a candidate mediator of LPS induced neuroprotection.

To date, 35 genes in our gene expression dataset have been identified as having associations with the TGFβ family of signaling molecules and the corresponding signaling pathways. As a case study, we used these 35 genes to determine whether the commonly used data analysis tools were able to map these genes to TGFβ pathways or to find linkages to TGFβ.

We examined tools from the four common methods of data analysis: GoStat (Beissbarth and Speed 2004) for Gene Ontology, Metacore™ (www.genego.com) for pathway databases, FatiWise (Mulder et al. 2003) for pathway prediction and Medstract for information extraction. The criteria for selecting these tools as a representative of each methodology were based on their demonstrated evidence in similar areas of research. As shown in Table 1, no single method succeeded in relating all 35 genes to TGFβ. Medstract and Metacore™ performed the best by revealing 19 and 27 gene-to-TGFβ associations respectively. If all four techniques are combined, 34 of the 35 genes are identified as associated with TGFβ; Stk11ip is the gene for which no TGFβ association is found.

Using XOA, we examined the potential TGFβ link of our 35 genes by calculating the XOA score between every GO code of the gene and the GO code of TGFβ receptor signaling pathway (GO:0007179). For every gene, the GO code with the highest XOA score to the TGFβ receptor signaling pathway is shown in Table 1. In cases where several GO codes shared the highest XOA score (which reflects the actual biological nature of more than one processes or function being attributed to a pathway), these codes were ranked according to their informational specificity (see previous section) and the most specific GO code was retained. For illustrative purposes, the XOA score is given with reference to the Resnik-based similarity measure. The other variants of XOA based on Lin and Jiang & Conrath similarity measures (not shown) strongly corroborate the findings in Table 1 (correlations between XOA scores are close to 1). Table 1 also reports p-values associated with the XOA scores. The p-value is obtained by comparing the XOA score with the distribution of all possible scores obtained by computing the XOA score between all possible pairs of GO codes from the three gene ontologies.

As Table 1 shows, XOA succeeds in finding a relation between all the 35 genes and the TGFβ receptor signaling pathway. Moreover, all but two XOA scores are associated with p-values ≤ 0.07. Stk11ip, the gene not identified by any of the above four methods, produces an XOA score corresponding to a p-value of 0.14 while the largest p-value of 0.15 is observed with the gene Lefty1. However, the XOA score by itself is not sufficient to make a biological interpretation of a link unless it is supported by evidence from literature. This is where Medstract complements XOA. Table 1 reveals that when XOA scores are high enough (e.g. XOA$\geq$ 5.83), the GO code linking one of the 35 genes to the TGFβ receptor signaling pathway is specific enough to provide an adept pathway link. In these cases, Medstract can be profitably queried using as input the gene and GO code link and the TGFβ pathway to obtain the appropriate information that validates and further characterizes the new pathway link discovered.

When lower XOA scores are obtained (i.e., XOA<5.83), the GO code link tends to be less specific and therefore has a higher chance to refer to genes that may not be associated with the TGFβ receptor signaling pathway. In these cases, several strategies for reducing these false positives are available. First, we can enrich the cohort of GO codes associated with the gene using GoPubmed (Doms and Schroeder 2005). GoPubmed is a web server that allows users to explore PubMed search results using the Gene Ontology for categorization and navigation purposes (available at http://www.gopubmed.org). Querying GoPubmed with the gene returns a list of GO codes that can be filtered according to their informational content. The new list of specific GO codes is then used to find links to pathways using the XOA approach. If these results are not satisfactory and continue to produce only generic GO code links, we can look for gene similarities between the gene of interest and the members of the TGFβ pathway. As shown in Posse et al. (2006) the XOA methodology can be efficiently extended to gene product similarities as follows:

Let *GP1* and *GP2* be two genes/gene products. Let *c11*,*c12*,…, *c1n* denote the set of GO codes associated with *GP1* and *c21*, *c22*,…., *c2m* the set of GO codes associated with *GP2*. The XOA similarity between *GP1* and *GP2* is given by

$$XOA(GP1,G\,P2) = max\{XOA(c1i\ ,\ c2j)\} \qquad (1)$$

where $i=1,\ldots,n$ and $j=1,\ldots,m$.

The higher the semantic similarity between the gene of interest and one of the genes in the pathway, the more likely it is that the gene of interest belongs to the pathway. Medstract can be used to further substantiate this finding.

Discussion & Evaluation

Unlike other GO tools or pathway prediction tools, XOA identifies which one amongst the many processes or functions that a gene performs can be correlated to a pathway. Such correlation is further substantiated by evidence from literature using Medstract.

Table 2, shows the GO terms found in association with the 35 genes spread across the three ontologies. This shows that a gene product can have one or more molecular functions, be used in one or more biological processes and may be associated with one or more cellular components. While GO acts as a repository of the known functional biological information on each gene, the ability to determine which of the many biological processes or molecular functions or cellular components that a gene product has, can be correlated to a pathway has not been explored by the existing ontological analysis tools. Though, GO molecular function terms have been used to predict subcellular locations (Lu and Hunter 2005), the current approaches used for ontological analysis are limited to looking up existing annotations and performing a significance analysis for the categories found. These approaches may not be able to discover previously unknown functions for known genes even if there is data justifying such inference across the three ontologies.

TGFβ exerts its action by binding to its transmembrane serine/threonine kinase receptors, which in turn triggers activation of various intracellular signaling pathways. Though a biological process is not equivalent to a pathway, the GO term descriptions (like regulation of TGFβ receptor signaling pathway, or transmembrane protein serine/threonine kinase signaling pathway, etc.,), explicitly shows biological relevance to the TGFβ pathway. The Stat1 gene has 15 biological processes, 6 of which returned the same highest XOA score to the pathway (signal transduction, intracellular signaling cascade, JAK-STAT cascade, tyrosine phosphorylation of STAT protein, STAT protein nuclear translocation, I-kappaB kinase/NF-kappaB cascade). More than one biological process can be involved in a pathway, and in the case of Stat1, all these 6 process terms seemed relevant in the context of TGFβ pathway. XOA correctly identified the term, 'JAK-STAT cascade' as the more specific in terms of their information content to the pathway.

In addition to mapping GO terms to pathways, XOA identifies relationships across ontologies. For example, the gene "Fos" returned the biological process term, 'Regulation of transcription, DNA dependent' as the highest correlation to the TGFβ receptor signaling pathway. When queried for highest XOA score to the pathway, for GO terms across ontologies, we observed the molecular function term 'DNA binding' and cellular location term 'Nucleus' as the best correlations. In the data observed, when a biological process can be defined as a series of events carried out by one or more ordered assemblies of molecular functions, then XOA has been able to identify

that the biological process,"regulation of transcription, DNA dependent" constitutes the molecular function "DNA binding", and the cellular location where this process or function happens is "Nucleus".

The XOA methodology has the ability to build semantic bridges between the three hierarchies of molecular functions, cellular components, and biological processes that provide a more clear view of how processes, functions and locations are related in the context of pathways.

A biological process is a recognized series of events but not equivalent to a pathway although some GO codes do describe pathways. Since TGFβ receptor signaling pathway already had a GO code, we used the same for this case study. However, in cases where the GO code for a specific pathway is not available, there are alternate ways of calculating semantic similarities, using XOA. For example, as discussed in the previous section, we could use XOA to calculate the gene similarities between all the members of the pathway and the gene(s) of interest. The higher the XOA score for common processes/functions, the closer is the semantic similarity between the gene of interest and the pathway under study. We can use Medstract to understand what these processes or functions contribute to the pathway.

SUMMARY & FUTURE WORK

The main motivation for exploring the relationship across the three gene ontologies is to understand the nature of their contribution in unraveling the biological phenomena performed by the genes and gene products. The case study presented above shows that the combination of XOA and Medstract has the potential of identifying biomarkers of stroke and uncovering the processes through which the genes carry out the biological phenomena. This provides a clear indication that XOA and Medstract have a very promising potential as pathway discovery tools. Moving forward, the evaluation of our preliminary study suggests that the alternative strategies have to be developed to properly address cases where more precision is required in associating genes to pathways.

ACKNOWLEDGEMENTS

This work was supported by the Computational and Information Science Initiative at the US Department of Energy's Pacific Northwest National Laboratory, Richland WA 99352. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the US Department of Energy under Contract DE-AC05-76RL01830.

REFERENCES

Andrade, M.A. (1999) Position-specific annotation of protein function based on multiple homologs. *ISMB* 28-33.

Azuaje F., H. Wang and O. Bodenreider (2005) Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies* 2005, pages 9-10.

Beissbarth T. and T. Speed (2004) GOstat: find statistically overrepresented Gene Ontologies within group of genes. *Bioinformatics* 20:1464-1465.

Bodenreider, O., M. Aubry and A. Burgun (2005) Non-lexical approaches to identifying associative relations in the gene ontology. In *Proceedings of Pacific Symposium on Biocomputing*, pages 104-115.

Chang, J.T., S. Raychaudhuri, and R.B. Altman (2001) Including biological literature improves homology search. In *Proc. Pacific Symposium on Biocomputing*, pages 374—383.

Couto, F. M., M. J. Silva and P. Coutinho (2003) Implementation of a functional semantic similarity measure between gene-products. *Technical Report*, Department of Informatics, University of Lisbon, http://www.di.fc.ul.pt/tech-reports/03-29.pdf.

Doniger, S.W., N. Salomonis, K.D. Dahlquist, K. Vranizan, S. C. Lawlor and B. R Conklin (2003) MAPPFinder : using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, *Genome Biology 2003, Vol.4, Issue1, Article R7*

Lin, D. (1998) An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI, pages 296-304.

Lin, D. and P. Pantel (2001) Discovery of inference rules for question answering. *Natural Language Engineering.* 7(4):343–360.

Lord P.W., R.D. Stevens, A. Brass, and C.A.Goble (2002) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19(10):1275-1283.

Lord P.W., R.D. Stevens, A. Brass, and C.A.Goble (2003) Semantic similarity measures as tools for exploring the Gene Ontology. In *Proceedings of Pacific Symposium on Biocomputing*, pages 601-612.

Lu Z. and L. Hunter (2005), GO Molecular Function Terms are predictive of subcellular localization. In *Proceedings of Pacific Symposium on Biocomputing* 10:151-161

Mulder, N.J., R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork et al. (2003) The InterPro database 2003 brings increased coverage and new features. *Nucleic Acids Res.* 31:315–318.

Posse C., A. Sanfilippo, B. Gopalan, R. Riensche, N. Beagley, and B. Baddeley (2006) Cross-ontological analytics: combining associative and hierarchical relations in the Gene Ontologies to assess gene product similarity. In *Proceedings to International Workshop on Bioinformatics Research and Applications*, University of Reading, UK, May 28-31

Pustejovsky J., A. Rumshinsky and J. Castaño (2002) Rerendering semantic ontologies: automatic extensions to UMLS through corpus analytics. In *Proceedings of the OntoLex-2002 workshop on Ontologies and Lexical Knowldge Bases*, Las Palmas de Gran Canaria, Spain, Mai 2002, pages 60-67.

Sanfilippo A., C. Posse and B. Gopalan (2004) Aligning the Gene Ontologies. In *Proceedings of the Standards and Ontologies for Functional Genomics Conference 2*, Philadelphia, PA, http://www.sofg.org/ meetings/sofg2004/Sanfilippo.ppt

Stenzel-Poore M.P., S.L. Stevens, Z. Xiong, N.S. Lessov, C.A. Harrington, M. Mori, R. Meller, H.L. Rosenzweig, E. Tobar, T.E. Shaw, X. Chu and R.P. Simon (2003) Effect of ischaemic preconditioning on genomic response to cerebral ischaemia: similarity to neuroprotective strategies in hibernation and hypoxia-tolerant states. *Lancet* 362: 1028–1037.

Vadigepalli R, P. Chakravarthula, D.E. Zak, J.S. Schwaber, G.E. Gonye (2004) PAINT: a promoter analysis and interaction network generation tool for gene regulatory network identification. *OMICS,* Fall;7(3):235-52.

.

Table 1: Comparison of methods relating genes to pathways using a set of 35 genes from microarray analysis of LPS preconditioning that have been manually identified as having associations with the TGFβ receptor signaling pathway.

Gene Symbols	Manual Curation	Medstract	Meacore™	FatiWise	GoStat	Semantically similar GO code of gene to TGFβ receptor pathway (GO:0007179)	XOA Score	
							Resnik	P-value
Serpine1	√	√	√	√		GO:0005576 (Extracellular region) - CC	5.32	0.07
Smad1	√	√	√	√	√	GO:0007182 (SMAD protein heteromerization) - BP	11.76	0.00
Foxh1	√	√	√			GO:0003677 (DNA binding) - MF	5.22	0.07
Map3k7	√	√	√			GO:0007179 (TGFβ receptor signaling pathway) - BP	11.76	0.00
Strap	√	√	√		√	GO:0030512 (negative regulation of TGFβ receptor signaling pathway) - BP	11.76	0.00
Tgfb3	√	√	√			GO:0007179 (TGFβ receptor signaling pathway) - BP	11.76	0.00
Tieg1	√	√	√			GO:0003676 (nucleic acid binding) - MF	5.31	0.07
Nodal	√	√		√	√	GO:0007179 (TGFβ receptor signaling pathway) - BP	11.76	0.00
Catnb	√	√	√	√		GO:0016055 (Wnt receptor signaling pathway) - BP	7.06	0.02
Fos	√	√				GO:0003677 (DNA binding) - MF	5.22	0.07
Jun	√	√				GO:0003677 (DNA binding) - MF	5.22	0.07
Runx2	√	√	√			GO:0040036 (regulation of fibroblast growth factor signaling pathway) - BP	9.76	0.00
Ski	√	√	√			GO:0005737 (cytoplasm) - CC	5.22	0.07
Stat3	√	√				GO:0007259 (JAK-STAT cascade) – BP	5.83	0.05
Tgif2	√	√	√			GO:0003677 (DNA binding) – MF	5.22	0.07
Thbs1	√	√	√			GO:0007155 (cell adhesion) – BP	5.57	0.05
Acvr2	√		√			GO:0007178 (transmembrane protein serine/threonine kinase signaling pathway) – BP	11.07	0.00
Gdnf	√		√		√	GO:0007179 (TGFβ receptor signaling pathway) – BP	11.76	0.00
Lefty1	√		√			GO:0007275 (development) – BP	3.51	0.15
Prss11	√		√		√	GO:0030512 (negative regulation of TGFβ receptor signaling pathway) – BP	11.76	0.00
Sara1	√		√			GO:0007264 (small GTPase mediated signal transduction) - BP	5.83	0.05
Bmp8a	√			√	√	GO:0007179 (TGFβ receptor signaling pathway) - BP	11.76	0.00
Inhba	√			√		GO:0007166 (cell surface receptor linked signal transduction) - BP	7.06	0.02
D0H4S114	√				√	GO:0017015 (regulation of TGFβ receptor signaling pathway) – BP	11.76	0.00
Tob1	√		√		√	GO:0007184 (SMAD nuclear protein translocation) – BP	11.76	0.00
Fhl2	√		√			GO:0008270 (zinc ion binding) – MF	5.22	0.07
Lef1	√	√	√			GO:0016055 (Wnt receptor signaling pathway) – BP	7.06	0.02
Lmo4	√		√			GO:0008270 (zinc ion binding) – MF	5.22	0.07
Miz1	√	√	√			GO:0008270 (zinc ion binding) – MF	5.22	0.07
Pias1	√		√			GO:0007259 (JAK-STAT cascade) – BP	5.83	0.05
OPN	√		√			GO:0007160 (cell-matrix adhesion) – BP	5.57	0.05
Stat1	√	√	√			GO:0007262 (STAT protein nuclear translocation) – BP	5.83	0.05
Stk11ip	√					GO:0016301 (kinase activity) – MF	3.76	0.14
Tcf4	√		√			GO:0003677 (DNA binding) – MF	5.22	0.07
Tgif	√		√			GO:0003677 (DNA binding) - MF	5.22	0.07

Table 2: Distribution of GO codes for 35 genes across the three ontologies – [Biological Process (BP), Molecular Function (MF), and Cellular Component (CC)]

Gene Symbols (# of GO codes)	# of BP codes	# of MF codes	# of CC codes	Gene Symbols (# of GO codes)	# of BP codes	# of MF codes	# of CC codes
Serpine1 (10)	3	4	3	Lefty1 (7)	3	3	1
Smad1 (23)	16	3	4	Prss11 (14)	4	8	2
Foxh1 (14)	9	3	2	Sara1 (11)	6	2	3
Map3k7 (14)	3	11	-	Bmp8a (10)	7	2	1
Strap (16)	2	4	-	Inhba (29)	21	6	2
Tgfb3 (19)	15	2	2	D0H4S114 (2)	1	1	-
Tieg1 (10)	5	4	1	Tob1 (7)	3	3	1
Nodal (10)	7	2	1	Fhl2 (5)	1	3	1
Catnb (49)	25	9	15	Lef1 (15)	8	4	3
Fos (14)	7	2	5	Lmo4 (11)	5	4	2
Jun (15)	7	4	4	Miz1 (8)	4	3	1
Runx2 (17)	10	5	2	Pias1 (10)	4	4	2
Ski (5)	1	1	3	OPN (32)	18	6	8
Stat3 (27)	15	8	4	Stat1 (20)	12	6	2
Tgif2 (6)	3	2	1	Stk11ip (1)	-	1	-
Thbs1 (8)	2	4	2	Tcf4 (16)	6	8	2
Acvr2 (20)	2	13	5	Tgif (9)	4	3	2
Gdnf (19)	15	3	1				