Bioinformatics Advance Access published May 30, 2007

Data and text mining

Clustering Microarray-Derived Gene Lists through Implicit Literature Relationships

Mark F. Burkart*[1], Jonathan D. Wren[2], Jason I. Herschkowitz[3,4], Charles M. Perou[3,4,5] and Harold R. Garner[1]

[1]Departments of Internal Medicine and Biochemistry, The McDermott Center for Human Growth and Development, Division of Translational Research, The University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, Texas 75390, USA, [2]Arthritis & Immunology Program, Oklahoma Medical Research Foundation, 825 N.E. 13th Street, Oklahoma City, Oklahoma 73104,USA, [3]Lineberger Comprehensive Cancer Center, [4]Department of Genetics, [5]Department of Pathology & Laboratory Medicine, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Associate Editor: Dr. Limsoon Wong

ABSTRACT

Motivation: Microarrays rapidly generate large quantities of gene expression information, but interpreting such data within a biological context is still relatively complex and laborious. New methods that can identify functionally related genes via shared literature concepts will be useful in addressing these needs.

Results: We have developed a novel method that uses implicit literature relationships (concepts related via shared, intermediate concepts) to cluster related genes. Genes are evaluated for implicit connections within a network of biomedical objects (other genes, ontological concepts, and diseases) that are connected via their co-occurrences in Medline titles and/or abstracts. On the basis of these implicit relationships, individual gene pairs are scored using a probability-based algorithm. Scores are generated for all pairwise combinations of genes, which are then clustered based on the scores. We applied this method to a test set composed of nine functional groups with known relationships. The method scored highly for all nine groups and significantly better than a benchmark co-occurrence-based method for six groups. We then applied this method to gene sets specific to two previously defined breast tumor subtypes. Analysis of the results recapitulated known biological relationships and identified novel pathway relationships unique to each tumor subtype. We demonstrate that this method provides a valuable new means of identifying and visualizing significantly related genes within gene lists via their implicit relationships in the literature.

Contact: mark.burkart@utsouthwestern.edu

Supplementary information: Supplemental Figures 1 - 5. Supplemental Tables 1 and 2.

1 INTRODUCTION

DNA microarray experiments can be used to study the expression levels of thousands of genes for the analysis of cell signaling pathways, disease marker discovery, and research and development of therapeutics (Cooper, 2001). However, interpretation of the subsequent results within biological context is often daunting, due one's limited knowledge of specific genes and the typically complex nature of biological relationships.

Analysis of microarray datasets generally begins with unsupervised clustering of genes based on expression patterns, or supervised analysis to identify gene sets, followed by retrieval of gene list annotations with ontological descriptions (Hosack, *et al*., 2003). However, while ontology definitions, such as those from the Gene Ontology Consortium (Ashburner, *et al*., 2000), can help provide insights into the properties and functions of individual genes, they are often incomplete, lacking information related to specific molecular interactions, disease states, or associated phenotypes (Khatri and Draghici, 2005). Electronically available Medline abstracts, on the other hand, offer a more comprehensive source of information that can be mined for more diverse and biologically relevant relationships.

Several methods that use Medline-derived relationships to group functionally related genes have been reported (Shatkay and Feldman, 2003). Jenssen, *et al*. (2001) grouped genes by identification of co-occurring gene pairs in Medline abstracts to construct gene relationship networks. Chaussabel and Sher (2002) and Alako, *et al*. (2005) identified shared terms or concepts that co-occurred with gene names in Medline abstracts and then used those relationships to cluster both the genes (based on their concept score profiles) and the concepts (based on their gene score profiles). Jelier, *et al*. (2005) used both gene co-occurrences and co-occurrences of genes with shared concepts to map genes in a Euclidean space in which distance represented semantic relatedness. Relationships involving indirect linkage of concepts via co-occurrence in the literature have been termed implicit (Swanson, 1986) and are useful for literature-driven discovery (Weeber, *et al*., 2003; Srinivasan and Libbus, 2004; Hristovski, *et al*., 2003, Wren, *et al*., 2004).

Here we describe a method that uses implicit literature relationships to score pairs of genes for relatedness and subsequently cluster the full gene set based on these scores. In our method, biomedical concepts (objects) are connected to each other in a network by their mutual co-occurrences in Medline titles and abstracts. Within

*To whom correspondence should be addressed.

this object network genes are evaluated for implicit connections through other network objects and scored for relatedness to other genes by the ratio of their observed/expected implicit connections in the network using probabilistic methods. The relatedness scores for all gene pairs in the set are then used to cluster the gene set.

Our method used a thesaurus of primary names and synonyms derived from electronically available bioscience-oriented databases to efficiently map terms with spelling variations, synonyms, or aliases to a single corresponding database object. Thesauri have been previously used to increase the sensitivity of the analysis (Alako*, et al.*, 2005; Jelier, *et al*., 2005). We tested our implicit analysis method by grouping control sets of genes having known functional relationships and then comparing the results to a gene co-occurrence-based method as a benchmark. The implicit analysis compared favorably against the gene co-occurrence based method in all control sets, indicating the general efficacy of the method.

We next applied this method to microarray-derived gene sets characteristically expressed by Basal-like and Luminal breast tumor subtypes (Sorlie, *et al*., 2003; Hu, *et al*., 2006). The analysis identified gene clusters with functional relationships unique to each tumor subtype. Several of these relationships corresponded to previously described, tumor-specific phenotypes.

2 METHODS

2.1 Construction of the literature-derived network

Both the literature derived network of biomedical objects used in this study and the methodology used in filtering and scoring connections have been previously described (Wren, *et al*., 2004). The network consisted of a collection of concepts (biomedical objects) extracted from electronically available, curated biomedical databases, including Locus Link, HGNC, GDB, OMIM, and GO. Objects were classified as genes, diseases, or ontologies based on the source: genes - Locus Link, HGNC, GDB; diseases – OMIM; ontologies - GO. Classifications allowed for refined filtering of network associations when context-specific associations were desired. Only human genes were used in this study. An object's definition consisted of both a primary name and any synonyms derived from the source database (thesaurus), allowing concepts identified within texts to be matched to corresponding primary database objects.

Titles and abstracts from over 15 million Medline records dating from 1967 to 2005 were processed to catalogue all co-occurrences of objects. Co-occurrences for each pair of objects were totaled and classified as either sentence or abstract co-occurrences. Records were stored in a Microsoft Access 2003 database, and queries were executed by SQL statements with partial automation by VBA macros.

2.2 Implicit analysis and clustering Process

For a given list of genes, the literature derived database was queried to identify gene objects matching identifiers in the gene list. The matching identifiers produced a query list of genes that were found to be present in the database. The database was then searched to identify all other objects found to co-occur in Medline records with any of the gene objects in the query list. This produced a list of co-occurrences between the query genes and other objects in the database (Fig. 1A). Filters were applied to remove co-occurrences from the list that did not achieve preset cut-off values (Filtering is discussed in depth in Section 2.3).

An implicit connection between two genes resulted when two genes in the co-occurrence list shared a co-occurrence with a common object (Fig. 1B, C). The common object thus served as the intermediate in the implicit connection between the two genes. This differs from co-occurrence based methods of identifying related genes, since relationships are made though intermediate, shared objects, rather than through a literature co-occurrence

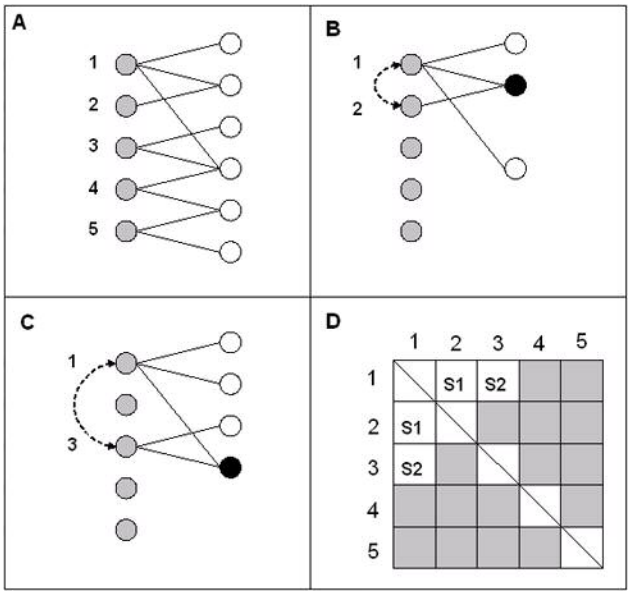

Fig. 1. Measurement of pairwise observed/expected scores for a set of genes. (**A**) A gene set (gray circles) was examined for co-occurring biomedical objects (connecting lines to white circles). (**B**) individual gene pairs were compared to identify both unshared (white circles) and shared objects (black circles). Shared objects formed implicit relationships. An observed/expected score was calculated based on shared and unshared objects for each gene pair (see scoring function). (**C**) subsequent pairs were compared and scored. This process was repeated until all gene pairings were scored (**D**). A pairwise matrix was populated with the scores (Gray cells must also be filled, values determined in (B) and (C) are shown filled).

of the two genes. For all possible pairings of genes in the query list, the number of observed implicit connections was counted. The number of expected implicit connections between each gene pair was then calculated as the number expected in an equally sized, randomly connected network (scoring function is discussed in Section 2.4). From the two values an observed/expected ratio was determined for each gene pair. Pairs with no observed implicit connections received a score of zero.

When the observed/expected scores were rank-sorted and graphed, the shape followed a power law distribution in which the top-ranking scores increased exponentially with respect to the rest of the distribution. A threshold equal to the observed/expected ratio at the 95th percentile was therefore applied to those rank-sorted scores above that percentile to prevent very high values from masking those in the rest of the distribution. Each gene's pairwise self-identity score was set equal to the threshold score to represent the highest attainable score, and a symmetric matrix of observed/expected scores was created from this list (Fig.1D). Hierarchical clustering (average linkage using standard Pearson correlation) of the pairwise score matrix was performed using Cluster (Eisen, *et al*., 1998) and visualized using Java Treeview (Saldanha, 2004).

2.3 Network connection filters

A set of filters was applied to object connections within the network to specify the types of relationships used in the analysis. Previous studies had shown that a pair of objects that co-occurred more frequently in an abstract were more likely to be meaningfully related and that objects that co-occurred rarely were more likely to be relationships of a trivial nature (Jenssen *et al*., 2001; Wren *et al*. 2004; Alako, *et al*. 2005). Wren *et al*. (2004) experimentally determined false-positive error rates of 42% for abstract co-occurrences and 17% for sentence co-occurrences, similar to rates reported by Ding, *et al*. (2002). Following these studies, object con-

nections resulting from fewer than three abstract or two sentence co-occurrences were omitted from the analysis.

During testing it was observed that objects with relatively high connection frequencies forming implicit connections between genes were less likely to represent real or significant relationships between the genes. For example, the object "chromosome" has 31,951 connections to other objects in the database, whereas "Y-chromosome" has only 2,848. Many genes would be expected to co-occur with the object "chromosome", and any resulting implicit relationship would probably be deemed less interesting than an implicit connection formed via the more specific object, "Y-chromosome". While object frequencies are used to weight connections in the scoring function, very common shared objects can still occasionally result in misleading or uninteresting associations between genes. Therefore, as a tradeoff between specificity and sensitivity, a filter was applied to remove implicit connections formed via objects having more than 5,000 connections in the network.

Implementation of these filters is similar to TF*IDF (Term Frequency)*(Inverse Document Frequency) weightings used in natural language processing and information retrieval. The low sentence/abstract co-occurrence filter is similar to TF because object frequencies in abstracts can be used to assign importance to objects, and the high connection frequency filter is similar to IDF in that common words are assigned reduced values. However, the filters are applied as absolute cut-offs whereas TF*IDF weightings are relative. Relative weighting is employed within the scoring function for object connections passing the filter settings.

2.4 Scoring function

For a gene list, every possible gene pairing was scored for relatedness based on their shared implicit relationships. Scores were determined using an observed/expected ratio, which was the observed number of implicit connections between the genes in the actual network divided by the number expected in an equally sized, randomly connected network. As previously described by Wren, *et al*. (2004), the probability, *P*, of a direct connection between a gene (*A*), and any other object in a random network (*B*), can be estimated using the formula:

$$P(A \leftrightarrow B) = 1 - \left(1 - \frac{K_A}{N_t}\right) * \left(1 - \frac{K_B}{N_t}\right) \quad (1)$$

where K_A and K_B are the number of network connections for *A* and *B*, respectively, and N_t is the total number of objects in the network. An implicit connection between a pair of genes is the combination of two such direct connections. Therefore, the probability that two genes (*A* and *C*), will be implicitly connected via another object, *B*, can be estimated as:

$$P(A \leftrightarrow B \leftrightarrow C) = P(A \leftrightarrow B) * P(B \leftrightarrow C) \quad (2)$$

The expected number of connections, *E*, for a given gene pair in a random network can be estimated by summing the individual probabilities for each possible implicit connection:

$$E(A \leftrightarrow B \leftrightarrow C) = \sum_{i=1}^{n} P(A \leftrightarrow B_i) * P(C \leftrightarrow B_i) \quad (3)$$

The number of implicit connections that are possible, B_n is taken to be the union of all unique B_i connected to either *A* or *C* in the real network, where B_a is the set of all *B* connected to *A*, and B_c is the set of all *B* connected to *C*. For every gene pairing in the list, both the expected number of random connections and the number of real implicit connections are determined to obtain observed/expected scores.

Table 1. Pathways, Gene Ontologies, and diseases used as control groups

Database	Category	Genes
Biocarta	caspase cascade in apoptosis	21
Biocarta	sonic hedgehog pathway	10
Biocarta	adhesion and diapedesis of lymphocytes	12
Gene Ontology	biological process: telomere maintenance	12
Gene Ontology	cellular constituent: cornified cell envelope	9
Gene Ontology	molecular function: DNA helicase	21
MeSH	disease: retinitis pigmentosa	10
MeSH	disease: chronic pancreatitis	10
MeSH	disease: nephroblastoma (Wilm's Tumor)	10

2.5 Selection of functionally related gene sets

To test the efficiency of the implicit analysis method, nine control groups of genes with known functional relationships (Table 1) were selected from Biocarta (http://www.biocarta.com), GO (http://www.geneontology.org), and MeSH (http://www.nlm.nih.gov/mesh/meshhome.html). The three lists of genes from each database were non-overlapping and represented those related by function within particular cell signaling pathways (Biocarta), common ontologies (GO), and implication in a particular disease (MeSH). Genes for the MeSH disease categories were derived by selecting the top ten genes found to co-occur with each disease in Medline records. The control groups were selected based on their diversity of functional relationship types and because they were relatively unambiguous and distinct.

2.6 Evaluation

The methodology used for performance measurements was similar to that used by Jelier, *et al*. (2005) for assessing their associative concept space (ACS) method's performance for grouping genes. The nine control groups were combined into a single group of 115 genes to measure the efficiency of the implicit analysis method in identifying the original control group from which each gene was originally derived. The implicit analysis was performed as described above to obtain pairwise observed/expected scores. For each gene, the pairwise Pearson correlations generated by the clustering program were used to rank-sort all other genes by the correlation coefficient. For a method to be considered a good measure of relatedness between genes, genes from a gene's particular functional group were expected to be top-ranked for that gene.

The implicit analysis method was compared to gene co-occurrence rankings as a performance benchmark. For the gene co-occurrence method, the number of Medline co-occurrences between all gene pairs within the set is determined. For each gene, all other genes are rank-sorted from greatest to least co-occurrences with the gene. For the purpose of the comparison, genes from a particular gene's control group were termed 'positives' and genes not in the same functional group were termed 'negatives'. A receiver operating characteristic (ROC), which plots the true-positive rate (correctly identified positives divided by all positives) against the false-positive rate (incorrectly classified negatives divided by all negatives), was calculated for each gene from the rank-sorted list. The area under the ROC curve (AUC) was used as a measure of the method's performance for each gene (Hanley and McNeil, 1982). AUC values range from 0 to 1, with 0.5 representing random sorting, 1 representing perfect sorting (all of a gene's group members are rank-sorted in the top ranked slots), and 0 representing the worst possible sorting. To test for significance between the two methods, the AUCs for each group were compared using the Wilcoxon signed-ranks test, a non-parametric alternative to the paired Student's t-test. Because the AUC distributions were weakly dependent and skewed, the p values from the Wilcoxon test are only approximations. More accurate p-values would normally be obtained through bootstrapping to simulate data independency, however this was not affordable due to the large number of repeated analy-

ses that would need to be individually and manually analyzed using substituted gene sets.

2.7 Selection of biologically relevant gene sets

Gene lists used in this study were selected from a recent microarray study of 105 different human breast tumors (Hu, *et al*., 2006). Hierarchical clustering analysis of the expression data resulted in successful identification of four previously defined tumor subtypes: Basal-like, Luminal, HER2+/ER- and Normal Breast-like (see Suppl. Fig. 1 for the complete cluster diagram). We performed two-class unpaired comparisons using Significance Analysis of Microarrays (SAM) (Tusher, *et al*., 2001) for each subtype versus all other tumors individually at a $\leq 5\%$ false discovery rate to identify genes specific for the subtypes. Subtype-specific genes were selected from the Basal-like and Luminal sets for this study. There were no shared genes between the sets. The data discussed in this publication are available in NCBI's Gene Expression Omnibus (Barrett, *et al*., 2005) (http://www.ncbi.nlm.nih.gov/geo/, GEO Series accession number GSE1992).

3 RESULTS

3.1 Gene set identification by functional relationships

To test the effectiveness of the implicit analysis method for grouping genes, we developed a test set in which nine control gene lists representing different biological pathways, ontologies, or disease states were combined (Table 1). The goal of the test was to determine if the genes in the test set could be grouped with functionally related genes from their original control groups. To be assured that a 'correct' assignment of a gene would be to the original control group, it was necessary to determine that the original groups were sufficiently distinct that for most genes a correct gene assignment would be to a single group. The rate of gene co-occurrences between control groups was analyzed and found to be 3.81%. This was deemed sufficiently low to assume that positive assignment of genes to their respective categories would be relatively unambiguous. The set was analyzed using both the implicit analysis method and the gene co-occurrence method, in which genes' relatedness is assessed by the number of co-occurrences observed between gene pairs in abstracts. The co-occurrence method was used as a benchmark because it has been previously used for assessing methods of literature analysis for gene lists (Jelier, *et al*., 2005), and is the most direct approach to finding gene-gene relationships in the literature.

For the combined set of 115 genes, 111 had implicit relationships to other genes in the set, and 102 genes had co-occurrences with other genes in the set. Four genes were not detected by either method, and it was found that these were not discussed in any Medline abstract. Nine genes had only implicit relationships to other genes in the set. These included, for example, DYRK1B, which is part of the sonic hedgehog pathway (Mao, *et al.* 2002), BRIP1, which is a DNA helicase (Cantor, *et al.* 2001), and CNFN, which is part of the cornified cell envelope (Michibata, *et al.*, 2004). Thus, the implicit method was able to capture more genes for subsequent analysis, ~9% ,because some of the genes could be related only implicitly.

The 102 genes that were identified using both methods were used to compare the ability of the methods for grouping genes in the combined set with those in their original control groups. For each gene, the other 101 genes were ranked by either pairwise correlations determined using the implicit clustering method or by the number of observed pairwise co-occurrences. ROC curves were generated for each gene. In ROC graphs, the true positive (within control group) rate is plotted against the false-positive (not within control group) rate for each rank to generate a curve. Area under the curve (AUC) can be measured to assess the quality of the rankings for each gene (Hanley and McNeil, 1982). AUC values ranged from 0 to 1, with 1 representing perfect ranking of control genes, 0.5 representing random ranking, and 0 the worst possible ranking.

The average of the median AUC values for the implicit method exceeded 0.94 for all nine functional groups, with none falling below 0.85 (Fig. 2).

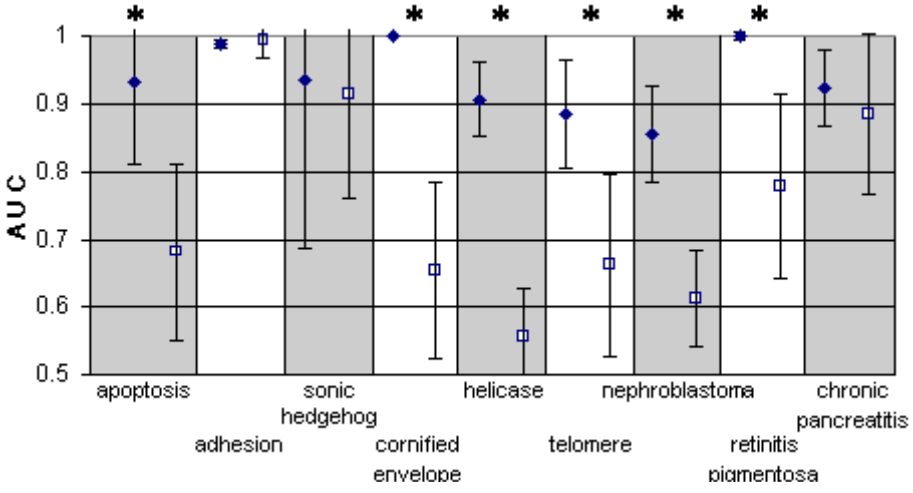

Fig. 2. Performance of the implicit analysis vs. co-occurrence methods and random grouping. For each of the nine functional groups, the Median AUC score and standard deviation are shown for the implicit (diamonds) and co-occurrence (boxes) methods. Asterisks above columns denote statistical significance for a difference between the two methods. P-values were: apoptosis 0.0005, adhesion 0.3652, sonic hedgehog 0.9375, cornified cell envelope 0.0312, DNA helicase < 0.0001, telomere maintenance 0.0078, nephroblastoma 0.002, retinitis pigmentosa 0.002, chronic pancreatitis 0.1934.

Table 2. Breakdown of relationships for gene pairs within control groups.

Control Group	AUC, co-occur.	Percent co-occur.	AUC, implicit	Percent implicit
Adhesion	0.99	100%	0.99	100%
Sonic hedgehog	0.91	71%	0.93	71%
Chronic pancreatitis	0.89	78%	0.92	100%
Retinitis pigmentosa*	0.78	47%	1.00	91%
Apoptosis*	0.68	39%	0.93	76%
Telomere maintenance*	0.66	39%	0.89	93%
Cornified cell envelope*	0.65	33%	1.00	83%
Nephroblastoma*	0.61	31%	0.86	100%
DNA helicase*	0.56	20%	0.91	82%
Mean	0.75	51%	0.94	89%
Std. dev.	±0.15	±26%	±0.05	±11%

Median AUC and percent of related gene pairs are reported for each method and were sorted by median AUC of co-occurrence method. Asterisks indicate statistically significant differences.

Median AUC was 0.75 for co-occurrence with one group scoring just above random, 0.56. The median AUC for co-occurrence was marginally better than the implicit method for only one group (Biocarta adhesion and diapedesis of lymphocytes). This difference, however, was not statistically significant. The implicit method out-performed co-occurrence for six groups at or above the 0.05 significance level, as determined by the Wilcoxon signed-ranks test. A notable example of improved performance was the GO cornified cell envelope category, for which perfect sorting was achieved for genes in the group using the implicit method whereas co-occurrence achieved 0.65.

Gene pairs in each control group were analyzed to determine rates of implicit relationships and co-occurrences. Implicit- or co-occurrence-related gene pairs were summed for each group and expressed as the percentage of all possible pairwise connections. It was found that a higher percentage of genes in each group were implicitly related than were related by co-occurrence (Table 2). All intra-group pairs related by co-occurrence were simultaneously implicitly related with the exception of a single gene pair in the DNA helicase category. The average number of gene pairs related implicitly or by co-occurrence was ~89% and ~51%, thus an average of ~38% of gene pairs were related only implicitly for all sets. This demonstrates that related genes can frequently be identified through implicit relationships where co-occurrences do not exist. Significantly, it also shows that a large percentage (at least ~38%) of the implicit relationships could not have been made only via objects related implicitly within the same abstract, i.e., where two genes are mentioned together with some other intermediate object in an abstract.

The median AUC by the co-occurrence method was found to be highly correlated with the percentage of co-occurring gene pairs (r^2=.96), thus the performance of the co-occurrence method was generally dependent on percentage of co-occurring gene pairs within the control group. For those groups in which greater than ~71% of gene pairs co-occurred, the implicit method did not significantly out-perform the co-occurrence method. However, for all other groups the implicit method significantly out-performed the co-occurrence method (Table 2).

Table 3. Method performance using different object type intermediates.

Control Group	Combined	Gene	Ontology	Disease
Apoptosis - Biocarta	0.93*	0.91	0.85	0.67
Adhesion - Biocarta	0.99*	0.99*	0.98	0.95
Sonic hedgehog - Biocarta	0.93*	0.93*	0.81	0.82
Cornified cell envelope - GO	1.00*	1.00*	1.00*	0.30
DNA helicase - GO	0.91*	0.87	0.89	0.90
Telomere maintenance - GO	0.89	0.80	0.90*	0.74
Nephroblastoma - MeSH	0.86	0.88*	0.81	0.88*
Retinitis pigmentosa - MeSH	1.00*	1.00*	0.90	1.00*
Chronic pancreatitis - MeSH	0.92	0.92	0.93	0.96*

Measured by median AUC. Asterisks indicate high scores for the group.

Thus, the implicit method appears to be able to infer relationships between genes where less literature directly mentions genes together in abstracts. Since the scientific literature is a perpetual work in progress and many related genes will not co-occur in abstracts, implicit methods may find related genes in their absence.

We next examined the performance of the implicit method while restricting object types used as intermediates in the implicit connections. Not surprisingly, genes sharing functional relationships within a specific context (gene, ontology, or disease-related) were most efficiently identified using implicit relationships of the same type (Table 3). For example, gene-type implicit connections worked best for cell signaling pathways and ontology-type implicit connections worked best for GO categories. Interestingly, sensitivity of the method when using all three object types was generally equal to or better (6/9 groups) than the most effective single object type. Thus, using relationships of different contexts to analyze gene sets should in some cases improve performance, particularly when the contexts most relevant to a given set are unknown.

3.2 Implicit analysis of breast tumor-specific gene sets

In order to test the implicit analysis using biologically relevant gene sets, we analyzed gene lists specific to two breast tumor subtypes, Basal-like and Luminal, previously identified in gene expression studies (Hu, *et al.*, 2006). The lists shared no common genes. Cancer gene sets were chosen because of the central role altered gene expression has in the development and pathology of the disease and because literature relationships between the genes could provide insight into disease mechanisms. Subtype-specific gene sets were compared to determine if insights unique to the specific subtypes could be provided.

The two gene lists were clustered based on pairwise observed/expected scores as described in Methods. Clustering on both axes produced symmetric, gene-x-gene arrays in which the magnitude of the observed/expected score was represented by the color intensity of the cells (Fig. 3). Groups of genes having high correlations between their respective pairwise scores were generally situated together along the array diagonal and formed identifiable clusters.

Implicit analysis and clustering was performed separately using each of the three object types (genes, ontologies, and diseases) present in the database in order to break-down implicit relationships by category, e.g., all gene-type implicit relationships were analyzed simultaneously to identify only pathway relationships. Arrays clustered based upon gene-type implicit relationships for

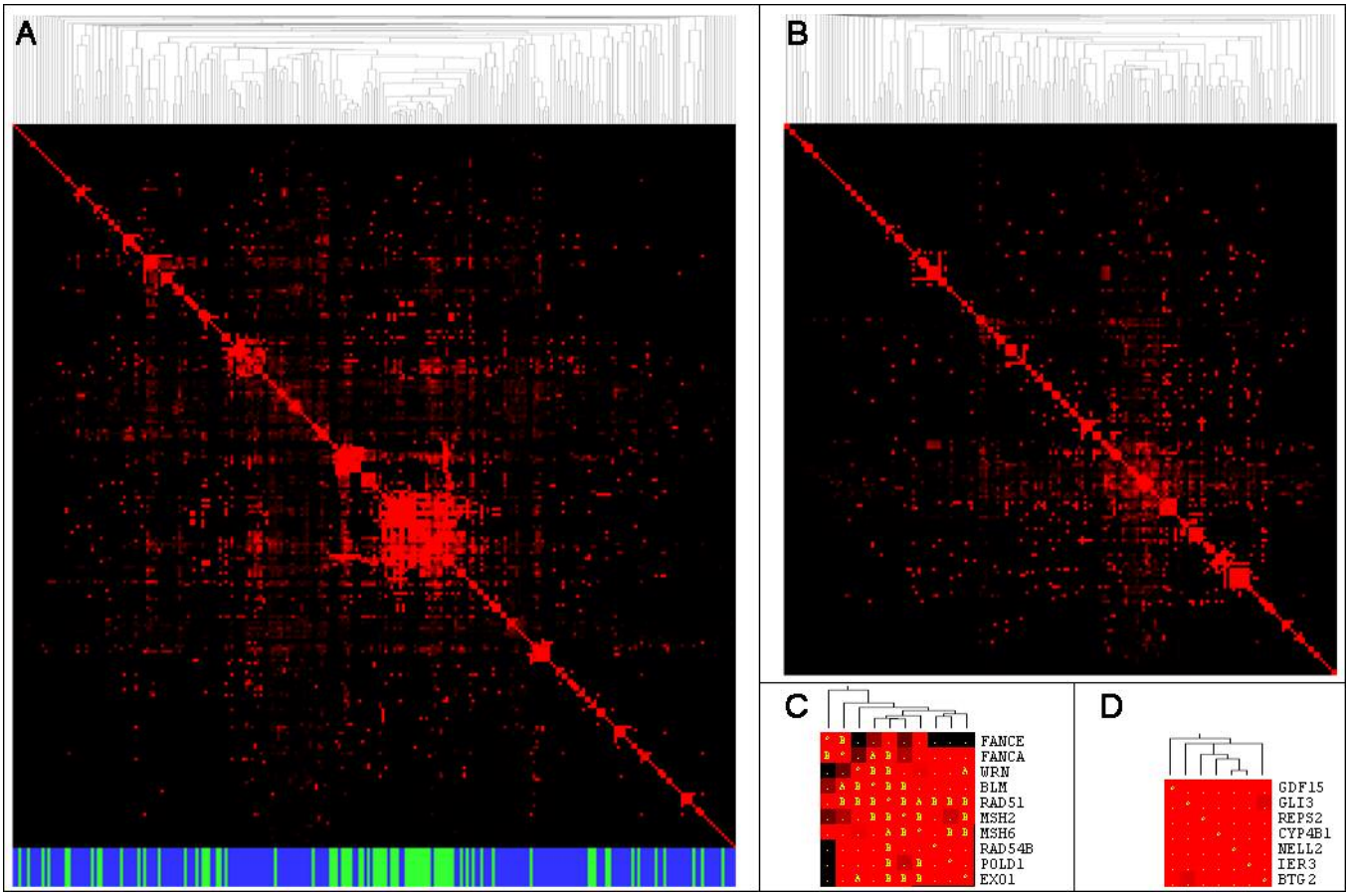

Fig. 3. **(A)** Basal-like tumor gene set clustered by implicit analysis method with gene-type intermediate connections. Clusters discussed in text were: (a) NF-kB signaling pathway; (b) Wnt signaling; (c) DNA repair; (d) DNA synthesis; (e) Cell cycle control; (f) Keratins; (g) Kallikreins. Proliferation expression cluster genes are represented below the array by green bars; blue bars represent other genes from the Basal-like expression cluster. **(B)** Luminal tumor gene set clustered with gene-type intermediate connections. Clusters discussed in text were: (a) tight-junction formation; (b) secretion and protein trafficking; (c) androgen receptor related; (d) ERBB2 and AF-6 related; (e) JAK/STAT pathway; (f) EP300/NCOR1 related; (g) GRB2 and PI3K related; (h) Wnt signaling; (i) EGR1 related. **(C)** Magnified Basal-like DNA repair cluster overlaid with gene co-occurrences: A = fewer than 2 co-occurrences; B = 2 or more co-occurrences. **(D)** Magnified Luminal EGR1 related cluster shows no co-occurrences of genes.

the Basal-like and Luminal sets are shown in Figure 3A and 3B. Arrays based upon ontology and disease-type implicit relationships for the sets are shown in Supplemental Figures 2 and 3 for Basal-like and Supplemental Figures 4 and 5 for Luminal. Highly correlated clusters (Pearson corr. >= 0.4) broken down by genes and the top five implicit objects shared by them are detailed in Supplemental Tables 1 and 2 for Basal-like and Luminal sets respectively.

3.3 Functionally related clusters in the Basal-like set

Clusters having genes with previously known functional relationships within the Basal-like set were identified, e.g., keratin (Fig. 3A-[f]) and kallikrein (Fig 3A-[g]) families. Several functional clusters not previously described for the set are detailed below.

3.3.1 NF-κB signaling Genes involved in NF-κB signaling were found separately via all three implicit relationship types (Table 4-[1]) (Fig. 3A-[a], Suppl. Figs. 2-[a], 3-[c]). Constitutive activation of the NF-κB pathway can squelch apoptotic induction signals (Shishodia and Aggarwal, 2002) and might confer apoptotic resistance to Basal-like tumors. Notably, TNFRSF21, (also known by its alias, DR6) is induced through NF-κB activation (Kasof, *et al*., 2001) and did not co-occur with any of these genes and could therefore be identified only implicitly.

3.3.2 Wnt signaling / developmental proteins Genes involved in cell fate and development decisions, including several members of the Wnt signaling pathway, were identified by gene-type relationships such as MYOD, BMP4, and other WNT genes (Table 4-[2]) (Fig 3A-[b]). Wnt signaling is known to be involved in breast cancer and other human cancers (Howe and Brown, 2004). Several of these genes are suppressors of myogenic development (EZH2, MDFI, ID4, NOTCH1) and could inhibit differentiation of Basal-like tumor cells into normal breast myoepithelial cells. This hypothesis is supported by the fact that Basal-like tumors express some markers (Keratins 5, 4, and 17) of differentiated myoepithelial cells but do not express several other classic markers (smooth muscle actin, p63, or membrane metallo-endopeptidase - CD10/CALLA) (Livasy, *et al*., 2006). Notably, EZH2, MDFI, and ID4 did not co-occur and could have been related only implicitly.

3.3.3 Proliferation signature genes Three clusters found via gene-type relationships (Fig. 3A-[c],[d],[e]), as well as clusters

Table 4. Examples of clusters found in Basal-like and Luminal sets

	Set	Obj. type	Genes in cluster	Top five implicit objects in cluster and pct. shared	% co-occ.
1	Basal	disease	BIRC3, BIRC2, MYBL2, BCL2A1, **PBK**, MYBL1, RELB, RIPK2, **TNFRSF21**	Burkitt's lymphoma (78), inhibitor of apoptosis (56), tumor necrosis (22), spontaneous apoptosis (22), sarcoidosis (22)	16.6
2	Basal	gene	**MDFI**, **EZH2**, **ID4**, SFRP1, FZD9, WNT6, WNT10A, FZD6, SOX9, FGF9, **CRABP1**, **ETV6**, NOTCH1, **KLF5**, FOXC1	WNT1 (40), MYOD (40), BMP4 (40), WNT3A (33), CCND1 (33)	7.6
3	Basal	gene	FANCE, FANCA, WRN, BLM, RAD51, MSH6, RAD54B, POLD1, EXO1, MSH2	RAD52(70), BRCA2(60), ERCC3(50), HPRT1(50), MRE11(50)	37.8
4	Basal	ontology	NEK2, CENPA, PTTG1, CDC20, BUB1, KIF11, **ORC6L**, KIF23, KIF2C, KNTC2, **MID1**	chromosome segregation (100), kinetochore (90), centrosome (70), cytokinesis (70), meiosis (60)	25.5
5	Luminal	gene	**SPDEF**, **TRPS1**, **IDS**, **KRT37**, **SH3BGRL**, TBK1, MAP3K1	AR(71), MAP3K14(29), CHUK(29), TXN(29), KLK3(29)	9.5
6	Luminal	gene	**CCND1**, IRS1, CISH, MAP2K4, **GATA3**, **ELF1**, (SYT1)	IL3(100), STAT5A(100), STAT3(86), STAT1(86), NF-kappa B(86)	9.5
7	Luminal	gene	**SIN3A**, **MYB**, **NRIP1**, **EID1**, **TAF9**	EP300(100), NCOR1(80), GAL4(60), CREBBP(40), GCN5(40)	0

Obj. type: intermediate object type (gene, ontology, or disease) forming implicit connection; Genes in cluster: bold-face genes had no co-occurrences with other genes in cluster, underlined genes had a single co-occurrence with another gene, a false-positive identification is shown in parentheses; Top five implicit relationships (shows percent of genes shared by in cluster); % co-occ.: percentage of co-occurring gene pairs in the cluster.

found by ontology- and disease-type relationships (Suppl. Figs. 2, 3), consisted of 'proliferation signature' genes originally identified by gene expression-based clustering and typically observed in rapidly proliferating tumors (Whitfield, et al., 2006). A large percentage of these genes are highly expressed during the cell cycle (Whitfield, et al., 2002) and are predictive of poor prognosis in breast cancer patients (Dai, et al., 2005). It is of interest that these genes, originally clustered by expression, are also clustered by implicit literature relationships, because it is suggestive of shared or mutual functionalities. Implicit relationships in these clusters indicated involvement in cell cycle control, DNA synthesis, DNA damage repair, and chromosome segregation, detailed below.

3.3.4 DNA damage repair Most of the genes found in this cluster function in DNA recombination and/or repair. Several have also been shown to repair stalled replication forks and maintain telomere length via homologous recombination (Tarsounas and West, 2005). These clusters may be phenotypically correlated with the high rate of gross chromosomal changes observed in these tumors (Richardson, *et al*., 2006). This is supported by the fact that almost all human BRCA1 mutation carriers develop Basal-like tumors (Sorlie, *et al*., 2003). Interestingly, disease-type relationships included diseases involving chromosomal instability and a predisposition to cancer. These genes were identified with all three object types (Table 4-[3]) (Fig. 3A-[c], Suppl. Figs. 2-[d], 3-[b]).

3.3.5 Chromosome segregation Genes required for chromosome segregation were identified via ontology-type implicit relationships (Table 4-[4]) (Suppl. Fig. 2-[c]). ORC6L, associated with the origin recognition complex, and MID1, which stabilizes microtubules, had no co-occurrences with other genes in the cluster and could be identified only implicitly.

3.3.6 DNA synthesis, cell cycle control A large cluster of 34 genes was identified via gene-type implicit relationships that included many proliferation signature genes. Two highly correlated subgroups within this cluster correlated at 0.8 consisted of DNA synthesis (Fig. 3A-[d]) and cell cycle control (Fig. 3A-[e]) genes. DNA synthesis genes, especially those involved in the origin recognition complex, were also found via ontology-type implicit relationships (Suppl. Fig. 2-[e]). Several of these are regulated by transcription factor E2F-1 during G1/S phase of the cell cycle and form a complex required for initiation of DNA replication (Fang and Han, 2006).

3.4 Functionally related clusters in the Luminal set

Luminal and Basal cells have different developmental fates and are likely to express genes differentially even in the wild-type state. Luminal cells, for example, have an apical surface facing the lumen and show greater expression of proteins required for secretory functions. Indeed, numerous secretion and trafficking-related genes were found in one cluster (Fig. 3B-[b]) and another contained genes involved in the formation of tight junctions found in luminal epithelia (Fig. 3B-[a]). Several functional clusters having not previously been described in this tumor subtype are detailed below.

3.4.1 Androgen receptor-related One cluster was found containing genes related by the androgen receptor (AR) (Fig 3B-[c]) (Table 4-[5]). Some (SPDEF, MAP3K1, TRPS1) regulate transcriptional activity of AR and others (KRT37, TRPS1) are regulated by AR at the transcriptional level. Interestingly, while AR is generally linked to prostate cancer, it may be of special significance in the Luminal subtype, because HER2/NEU, which is expressed in some tumors of the Luminal subtype, has been shown to modulate the activity of AR in prostate and breast cancer cell lines (Mellinghoff *et al.*, 2004). SPDEF, TRPS1, KRT37 had no co-occurrences with other genes in the cluster and could be identified only implicitly.

3.4.2 JAK/STAT pathway Signaling through the JAK/STAT pathway (Fig. 3B-[e])(Table 4-[6]) could be responsible for some of the anti-apoptotic and proliferative features of this subtype. Cyclin D1 (CCND1) and CISH transcription are increased through STAT5 signaling (Matsumura *et al*., 1999) (Mitchell *et al*., 2003). A false-positive identification, SYT1, resulted because the SYT1 alias, p65, is also an alias for the RELA gene.

3.4.3 EP300 / NCOR1 Genes which form transcriptional co-activator and co-repressor complexes with nuclear factors including EP300 and NCOR1 (Fig3B-[f]) (Table 4-[7]). None of these genes co-occurred. Most of these factors interact with EP300 or NCOR1, which interact with hormone receptors and other factors.

Other luminal clusters identified included genes related by HER2/NEU and AF6 (Fig. 3B-[d]), PI-3-kinase and GRB2 (Fig. 3B-[g]), WNT signaling (Fig. 3B-[h]), and EGR1 (Fig. 3B-[i]).

3.5 Implicit relationships show subtype specificity

Breast tumor subtypes show considerable biological and clinical diversity and may represent distinct disease processes (Hu, *et al*., 2006). It was therefore of interest to determine if groups of functionally interacting genes could be identified that were specific to the individual subtypes. To investigate this, we compared the degree of overlap between implicit relationships in clusters of the Basal-like and Luminal sets. Clusters obtained from gene-type implicit relationships were selected from each set having at least four genes and correlated at 0.4 or higher, resulting in 19 and 11 clusters for the Basal-like and Luminal sets, respectively. All pairwise combinations of the 19 Basal-like and 11 Luminal clusters, 209 unique cluster pairs, were compared for overlap of the top 10 implicit relationships (determined by counting the number of genes in a cluster sharing a given implicit relationship). It was found that 13 of the paired Luminal and Basal clusters had matching implicit relationships within the top 10. However, the overlap of these clusters was low, with three clusters exhibiting 20% overlap (two shared relationships) and 10 clusters exhibiting 10% overlap (one shared). The low degree of overlap indicated that the majority of genes of the respective subtypes were clustered through unique implicit relationships.

The relationships in the cluster pairs with two overlapping relationships were NF-kappaB and p65 (RelA, NF-kappaB subunit), STAT1 and STAT5, and RHOA and p85 (PI-3-kinase, regulatory subunit). These represent gene products often functioning as central mediators for several signaling pathways, possibly indicating that differing gene products of the respective tumor subtypes could function through some of the same pathway intermediates.

3.6 Most clusters are identifiable only implicitly

For the Basal-like and Luminal sets, only ~ 9.4 and ~ 9.6 percent of those genes that were implicitly connected also co-occurred. To more closely examine specific clusters for co-occurring gene pairs, co-occurrences were mapped onto the arrays of implicitly clustered genes, and it was observed that some of the clusters had gene pairs related by direct co-occurrences (Fig. 3C) and some did not (Fig. 3D). All clusters of four or more genes correlated at 0.4 or greater were manually examined in both sets, and it was found that ~ 12.2 percent of gene pairs within these clusters co-occurred in both sets.

However, even given the low percentage of direct relationships both within the gene lists and within the clusters themselves, there remained a possibility that significant implicit associations (i.e., those in highly correlated clusters) resulted as a consequence of co-occurring gene pairs occurring simultaneously in the same abstracts with intermediate objects connecting them implicitly.

To explore this possibility, implicit relationship scores between genes that also co-occurred were omitted from the list of pairwise scores to produce a truncated, implicit-only set of scores. Scores from the implicit-only set were used to cluster genes as before, producing clusters via purely implicit, and not simultaneously direct, connections. Unique clusters found correlated at 0.4 or greater were compared to those obtained previously at the same correlation. Original clusters were matched to the most similar clusters (>= 50 percent of genes shared) obtained using the implicit-only results. No clusters had more than one match. Of the original clusters, 39/43 (91%) of the Basal-like and 35/42 (83%) of the Luminal clusters were matched to similar clusters from the truncated set. Many of the lost clusters included those made up of genes having well known relationships, such as a kallikrein cluster (KLK5, 6, 8, and 10) and a keratin cluster (KRT6, 13, 16, and 17). Of the original clusters that matched clusters from the truncated set, 80.3% of Basal-like and 81.3% of Luminal genes were conserved in the matched clusters. Thus most gene relationships in clusters were formed via purely implicit connections.

4 DISCUSSION

In this study, we developed a method for implicit analysis and clustering of gene sets using a literature-derived biomedical object network. We began by showing that implicit analysis can identify functionally-related genes with improved performance over the gene co-occurrence method using a control set of genes with known functional relationships. This was particularly true for groups in which literature co-occurrence rates were lower and literature relationships could only be identified implicitly. This should be considered an advantage of the implicit method given that the biomedical literature is a perpetual work in progress. We also showed that genes sharing functional relationships of a specific semantic type (gene, ontology, or disease) were efficiently identified using the three object types simultaneously, a potential advantage when relevant relationship contexts for a given gene set are not known ahead of time.

We then employed implicit analysis and clustering against a real biological data set consisting of genes significantly expressed in Basal-like and Luminal breast tumor subtypes. Gene clusters were identified that shared functional relationships that reflected previously observed phenotypes (keratin cluster in Basal-like; secretion, trafficking and tight-junction formation clusters in Luminal). Previously unobserved functionalities were also observed (NF-κB-related anti-apoptotic mechanisms, DNA recombination and repair enzymes in the Basal-like subtype, hormone-activated receptors and JAK-STAT pathways in the Luminal). These represent potentially novel and relevant cancer-related pathway relationships for each subtype. Another interesting finding was that a subset of the Basal-set genes grouped by shared functional relationships consisted of proliferation signature genes previously identified by expression-based clustering. Comparison of implicit relationships shared by genes in clusters of both subtypes showed minimal overlap, and thus identified relatively subtype-specific clusters.

This method differs from previous literature-based gene clustering or grouping methods. In several previous methods (Chaussabel and Sher, 2002; Alako, *et al*., 2005; Jelier, *et al*., 2005) similarity measures between both genes and other concepts were used for subsequent clustering or arranging of genes. In this method, genes are not scored for literature similarity, but are instead compared using an observed/expected ratio of network connections to obtain

statistical connection strengths between genes. One possible advantage of this approach is that genes could be considered related without having significantly overlapping literature profiles if they have a sufficiently high ratio of observed/expected connections. This may be useful for genes not conceptually related in the literature but that may still share significant functional relationships.

Since genes are scored prior to the clustering step, each pairwise observed/expected score is represented within a single cell of the clustered matrix. Because the method produces a gene-x-gene matrix, it is possible to overlay the implicit matrix with co-occurrences (Fig 3C, D) to show which clusters have co-occurring genes and which do not. This could have discovery utilities, such as identifying potentially interacting genes with no literature co-occurrences. While the gene-x-gene matrix presented here simplifies visual comparison of gene pairs, it does somewhat reduce ease of use in identifying shared relationships, which must be identified via an additional query. In some methods (Chaussabel and Sher, 2002; Alako, *et al*., 2005) shared relationships are visible in the clustered matrix since genes are clustered against shared concepts.

One problem that was observed was that false positive relationships occasionally occurred due to misidentifications of terms with unrelated database objects during text processing. For example, the concept thesaurus would mistake one gene for another if they shared a common synonym or alias. Within the Basal-like and Luminal sets (Suppl. Table 1, 2), several genes were incorrectly grouped due to false positives of this type. Time-consuming manual analysis of the related literature abstracts was necessary to identify these errors. Advances in artificially-intelligent text processing may be required to address this type of error.

ACKNOWLEDGEMENTS

For helpful discussions regarding this paper we would like to thank John W. Fondon III, Wayne Fisher, My-Hanh T. Nguyen, Kristin Lennox, Cristi L. Galindo, Mounir Errami, and Ryan Weil. This research was supported by the P. O'B. Montgomery Distinguished Chair in Human Growth and Development, the Evelyn Hudson Foundation, and grants from NIH/NIAID Western Regional Centers of Excellence for Biodefense and Emerging Infectious Diseases (U54AI057156), NIH/NCI (CA096901), NIH/NCI SPORE (50CA70907), and UNC SPORE Breast Cancer (P50-CA58223).

REFERENCES

Alako, B.T.F., *et al*. (2005) Copub mapper: Mining medline based on search term co-publication. BMC Bioinformatics, 6(51), 1-15.

Ashburner, *et al*. (2000) Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat. Genet., 25(1), 25-29.

Barrett, T., *et al*. (2005) Ncbi geo: Mining millions of expression profiles—database and tools. Nucleic Acids Res., 33(Database Issue), D562-566.

Cantor SB, *et al*. (2001) BACH1, a novel helicase-like protein, interacts directly with BRCA1 and contributes to its DNA repair function. Cell, 105(1), 149-60.

Chaussabel, D. and Sher, A. (2002) Mining microarray expression data by literature profiling. Genome Biol., 3(10), 0055.0051-0055.0016.

Cooper, C.S. (2001) Applications of microarray technology in breast cancer research. Breast Cancer Res., 3(3), 158-175.

Dai, H., *et al*. (2005) A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. Cancer Res., 65(10), 4059-4066.

Ding, J., *et al*. (2002) Mining Medline: abstracts, sentences or phrases? Pac. Symp. Biocomput., Kauau, Hawaii, 7, 326-337.

Eisen, M.B., *et al*. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U S A, 95(25), 14863-14868.

Fang, Z.H. and Han, Z.C. (2006) The transcription factor e2f: A crucial switch in the control of homeostasis and tumorigenesis. Histol Histopathol., 21(4), 403-413.

Hanley, J.A. and McNeil, B.J. (1982) A simple generalization of the area under the ROC curve to multiple class classification problems. Radiology, 143, 29-36.

Howe, L.R. and Brown, A.M. (2004) Wnt signaling and breast cancer. Cancer Biol. Ther., 3(1), 36-41.

Hristovski, D., *et al*. (2005) Using literature-based discovery to identify disease candidate genes. Int. J. Med. Inform. 74, 289-298.

Hu, Z., *et al*. (2006) The molecular portraits of breast tumors are conserved across microarray platforms. B.M.C. Genomics, 7(1), 96.

Jelier, R., *et al*. (2005) Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics, 21(9), 2049-2058.

Kasof, G.M., *et al*. (2001) Tumor necrosis factor-alpha induces the expression of DR6, a member of the TNF receptor family, through activation of NF-kappaB. 20(55), 7965-75.

Khatri, P., *et al*. (2002) Profiling gene expression using onto-express. Genomics, 79(2), 266-270.

Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: Current tools, limitations, and open problems. Bioinformatics, 21(18), 3587-3595.

Livasy, C.A., *et al*. (2006) Egfr expression and her2/neu overexpression/amplification in endometrial carcinosarcoma. Gynecol. Oncol., 100(1), 101-106.

Lowe, H.J. and Barnett, G.O. (1994) Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. J.A.M.A., 271(14), 1103-1108.

Mao, J., *et al*. (2002) Regulation of Gli1 transcriptional activity in the nucleus by Dyrk1. J Biol. Chem., 277(38), 35156-61.

Matsumura I., *et al*. (1999) Transcriptional regulation of the cyclin D1 promoter by STAT5: its involvement in cytokine-dependent growth of hematopoietic cells. EMBO J., 18(5),1367-77.

Mitchell T.J., *et al*. (2003) Dysregulated expression of COOH-terminally truncated Stat5 and loss of IL2-inducible Stat5-dependent gene expression in Sezary Syndrome. Cancer Res. 63(24), 9048-54.

Mellinghoff, I.K., *et al*. (2004) HER2/neu kinase-dependent modulation of androgen receptor function through effects on DNA binding and stability. Cancer Cell, 6(5), 517-27.

Michibata, H., *et al*. (2004) Identification and characterization of a novel component of the cornified envelope, cornifelin. Biochem. Biophys. Res. Commun., 318(4), 803-13.

Richardson, A.L., *et al*. (2006) X chromosomal abnormalities in basal-like human breast cancer. Cancer Cell, 9(2), 121-132.

Rouzier, R., *et al*. (2005) Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin. Cancer Res., 11(16), 5678-5685.

Saldanha, A.J. (2004) Java treeview--extensible visualization of microarray data. Bioinformatics, 20(17), 3246-3248.

Shatkay, H. and Feldman, R. (2003) Mining the biomedical literature in the genomic era: an overview. *J. Comput. Biol.*, 10, 821-855.

Shishodia, S. and Aggarwal, B.B. (2002) Nuclear factor-kappab activation: A question of life or death. J. Biochem. Mol. Biol., 35(1), 28-40.

Sorlie, T., *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc. Natl. Acad. Sci. U S A, 100(14), 8418-8423.

Srinivasan, P. and Libbus, B. (2004) Mining medline for implicit links between dietary substances and diseases. Bioinformatics, 20(Suppl. 1), i290-i296.

Swanson, D.R. (1986) Fish oil, raynaud's syndrome, and undiscovered public knowledge. Perspect. Biol. Med., 30(1), 7-18.

Tarsounas, M. and West, S.C. (2005) Recombination at mammalian telomeres: An alternative mechanism for telomere protection and elongation. Cell Cycle, 4(5), 672-674.

Tusher, V.G., *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. U S A, 98(9), 5116-5121.

Weeber, M., *et al*. (2003) Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide. J. Am. Med. Inform. Assoc., 10(3), 252-259.

Whitfield, M.L., *et al*. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell, 13(6), 1977-2000.

Whitfield, M.L., *et al*. (2006) Common markers of proliferation. Nat. Rev. Cancer, 6(2), 99-106.

Wren, J.D., *et al*. (2004) Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics, 20(3), 389-398.