Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Functional gene clustering via gene annotation sentences, MeSH and GO keywords from biomedical literature

Jeyakumar Natarajan[1], * and Jawahar Ganapathy[1]

[1]Centre of Excellence in Bioinformatics, School of Biotechnology, Madurai Kamaraj University, Madurai 625021, India; Jeyakumar Natarajan* - E-mail: jkumar@mrna.tn.nic.in; *Corresponding author

received November 18, 2007; accepted December 30, 2007; published online December 30, 2007

Abstract:
Gene function annotation remains a key challenge in modern biology. This is especially true for high-throughput techniques such as gene expression experiments. Vital information about genes is available electronically from biomedical literature in the form of full texts and abstracts. In addition, various publicly available databases (such as GenBank, Gene Ontology and Entrez) provide access to gene-related information at different levels of biological organization, granularity and data format. This information is being used to assess and interpret the results from high-throughput experiments. To improve keyword extraction for annotational clustering and other types of analyses, we have developed a novel text mining approach, which is based on keywords identified at the level of gene annotation sentences (in particular sentences characterizing biological function) instead of entire abstracts. Further, to improve the expressiveness and usefulness of gene annotation terms, we investigated the combination of sentence-level keywords with terms from the Medical Subject Headings (MeSH) and Gene Ontology (GO) resources. We find that sentence-level keywords combined with MeSH terms outperforms the typical 'baseline' set-up (term frequencies at the level of abstracts) by a significant margin, whereas the addition of GO terms improves matters only marginally. We validated our approach on the basis of a manually annotated corpus of 200 abstracts generated on the basis of 2 cancer categories and 10 genes per category. We applied the method in the context of three sets of differentially expressed genes obtained from pediatric brain tumor samples. This analysis suggests novel interpretations of discovered gene expression patterns.

Keywords: text mining; functional clustering; microarray data analysis

Background:

In recent years, increasing amounts of biological data have become available through techniques such as DNA microarrays and other high-throughput gene and protein assays. **[1, 2]** As large numbers of genes can be included in such studies, the task of assigning meaningful biological function to gene patterns or gene clusters is a considerable challenge. Typical analyses using supervised (classification) or unsupervised (clustering) methods require the user to incorporate the necessary background knowledge. **[3]** This ability to incorporate background knowledge is fundamental to effective and efficient scientific discovery. A substantial amount of biomedical knowledge is captured in free-text form in abstracts and full-text articles and also in specialized biological information systems such as Gene Ontology (GO) **[4]**, Medical Subject Headings (MeSH) **[5]**, Database of Interacting Proteins (DIP) **[6]** etc. Until only a few years ago, human reasoning was the primary method for the extracting, synthesizing and interpreting the information contained in the biomedical literature and supporting biological information systems.

However, in recent years the number of online documents (and other biological information repositories) has grown tremendously. This is both an opportunity and a challenge. On one hand, such resources facilitate automated processing of the knowledge and information contained in these documents. On the other hand, such processing poses considerable algorithmic and computational challenges **[7]**. For example, the biomedical abstract database MEDLINE **[8]** currently contains about 15 million citations and about 40 000 citations are added monthly.

Text mining is the application of techniques from machine learning, natural language processing (NLP), information extraction and statistical/mathematical approaches to automated extraction of useful knowledge from text **[9]**. Text mining of biomedical literature has been applied successfully to various biological problems. Many studies focus on protein-protein **[10-13]** and gene-protein interactions. **[14]** Other specific relationships between biological entities such as sub-cellular localization of proteins **[15, 16]**, molecular binding relationships **[17]** and interaction between genes and drugs **[18]** are also explored. Text analysis of biomedical literature has also been applied successfully to incorporate functional information of gene expression data **[19-23]**. For example, MedMOLE **[19]** identifies the functions among a group of genes by simple text clustering of entire MEDLINE documents associated with the genes. Blaschke *et al.* **[20]** extracted information about the common biological characteristics of gene clusters from MEDLINE using a statistical term weighting approach. This method returns an ordered set of keywords with a high probability of occurrence in abstracts. Liu *et al*. **[24]** extended this approach by clustering such keywords to find gene-to-gene relationships. Clustering genes by functional keyword association can provide direct information about the nature of genes and their functional association **[25]**. However, the quality of the keyword lists extracted from the biomedical literature for each gene significantly affects the clustering results. Commonly, these approaches represent genes by extracting keywords from entire abstracts **[25]**. These keywords may undergo transformations such as *weighting* or *dimension reduction* with the goal of improving clustering quality and efficiency. However, gene clustering using entire abstracts has the following main drawbacks. (a) Abstracts normally contain a large number of irrelevant sentences. These sentences may influence the clustering process and are likely to obscure information useful for gene annotation. (b) The number of unique terms in abstracts is typically very large. This requires the ability to deal with sparse data spaces or methods for dimensionality reduction. (c) Dimension reduction methods such as principal component analysis or signal-to-noise methods increase the computational complexity, may lead to the loss of important keywords and do not guarantee that the reduced dimensionality will yield better clustering/annotation information. Also, the composite features may be hard to interpret.

To avoid the above drawbacks and improve the clustering process, we decided to use gene annotation sentences from abstracts instead of using full abstracts to extract the keywords. Current NLP techniques allow such sentence extraction from documents. Using clustering on the basis of sentence-extraction techniques has the advantage of avoiding complex dimensionality reduction and term weighting techniques. Further, this approach is likely to yield more specific terms which are easier to interpret. We first extracted the potential sentences describing gene annotation information from abstracts using a NLP method utilizing gene/protein name dictionaries and pattern-matching-based rules. In addition to this sentence-keywords approach, we carried out two further experiments involving MeSH terms and GO terms as supplementary keywords. Hence, in our method, each gene is represented by set of keywords extracted from *sentences*, MeSH terms and GO terms. To demonstrate the usefulness of the proposed text mining methods, we performed hierarchical clustering of a *gene* × *keyword* matrix to find functionally discrete sub-groups of genes. The overall experimental design and its components are illustrated in the Figure 1.

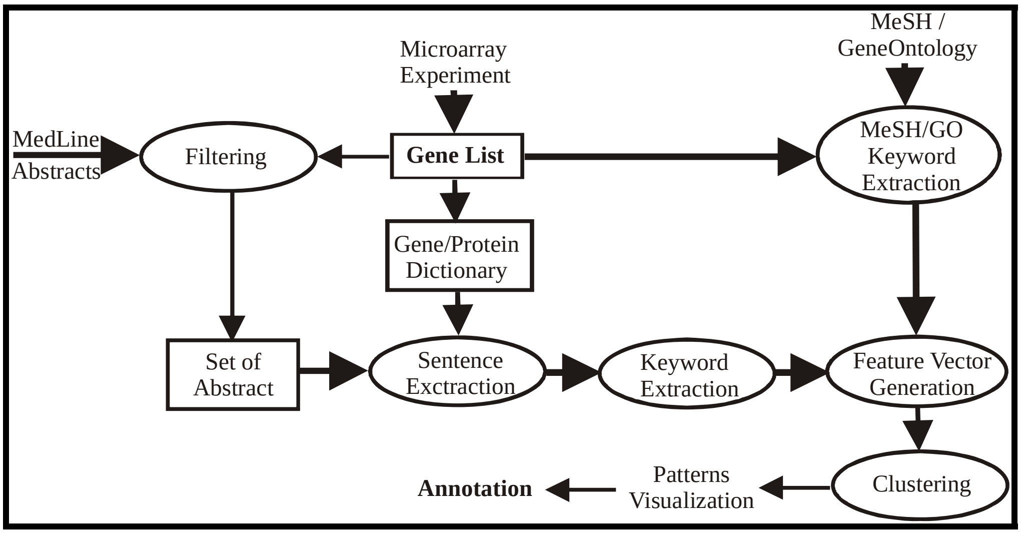

Figure 1: Experimental design of gene clustering with sentences-level, MeSH and GO keywords

We validated the performance of keywords extracted by our method using a manually annotated corpus of 200 abstracts. We also evaluated the usefulness of our method by sorting differentially expressed genes from a microarray experiment into functional sub-groups. The objective of our gene clustering process using functional keywords is to identify and summarize potential functional gene groups and to complement the conventional gene expression data clustering tasks.

Methodology:

Gene/Protein name and synonym dictionary creation

One of the major obstacles in biomedical literature processing is the variety of names each gene or protein is known by. To address this problem in the present study, we developed a gene/protein name dictionary. Essentially, each entry of this dictionary consists of a preferred (or canonical) name for a gene/protein and a list of synonyms used for this gene/protein. The dictionary was created on the basis of the Entrez Gene **[26]** (previously LocusLink) database, one the most stable and complete sources of information on genes. Since our study is part of a wider investigation in the context of human brain tumor research, we focused specifically on human genes/proteins. We developed PERL scripts to extract and select from the Entrez Gene entries the official symbols as preferred name of the gene and other aliases as known synonyms. In addition, we augmented the dictionary with relevant synonyms from other publicly available databases including GeneCards

[27], SwissProt **[28]**, GoldenPath **[29]** and HUGO **[30]**. The final dictionary contains 26 731 unique human gene/protein names and 274 845 synonym names.

Keywords extraction from biomedical literature

In our study each gene is represented by a list of keywords extracted from MEDLINE abstract sentences, MeSH terms and GO terms. The procedure for extracting keywords from each data source is discussed below.

MEDLINE abstracts keywords extraction

To extract the keywords associated with each abstract, we decided to use gene annotation *sentences* from the abstracts instead of constructing a large keyword vector based on the entire abstract. The assumption is that the information given on sentence-level is much more specific and therefore useful to characterize the function of the genes. Only sentences that contain one or more genes reference from our gene lists will be considered as gene annotation sentences, all other sentences are discarded from the analysis. We applied the following three steps to extract sentence-level keywords (1) gene-name normalization, (2) sentence filtering, and (3) keyword extraction.

Gene-name normalization

This process replaces all the gene names in the abstract with its unique canonical identifier (Entrez gene ID) using the gene-synonym dictionary specially constructed for this study.

Sentence filtering

This process extracts all the gene annotation sentences from abstracts that contain one or more gene names from our gene lists using regular-expression pattern matching rules. We used different regular expressions (which rely on matching of pre-defined patterns or rules such as arrangement of gene/protein names with articles, prepositions and other keywords) to filter sentences containing one to three genes. We defined our regular expressions as nouns describing agents, passive verbs, active verbs and nouns describing actions. Table 1 (in supplementary material) depicts an example for each type of expression. For example, the regular expression

($gene @{0,6} $action (of|with) @{0,2} $gene)

extracts sentences that match the structure shown below the expression. The notational construct ‘A → B → ...’ is interpreted as ‘A followed by B followed by ...’.

gene name → 0-6 *words* → *action verb* → ‘of’ or ‘with’ → 0-2 *words* → *gene name*

Sentence keyword extraction

Sentences containing one or more gene names were parsed using the Brill part-of-speech tagger. **[31]** This program labels each word in a sentence with its part-of-speech information such as word category like noun, verb, adjective, preposition, etc. This information plays a critical role in identifying corresponding noun and verb phrases. Then, with a simple PERL program, noun phrases containing gene names were filtered out and the remaining noun phrases and verb phrases were extracted as keywords. Initial tests showed that certain keywords were common for most of the genes in the list (e.g., activates, associates, stimulates etc.). We manually removed these common keyword words from the list. The following example illustrates this process:

(1) Sentence
BRCA1 physically associates with p53 and stimulates its transcriptional activity.

(2) Brill-POS-tagged sentence
BRCA1/NNP physically/RB associates/VBZ with/IN p53/NN and/CC stimulates/VBZ its/PRP$ transcriptional/JJ activity/NN. /.

(3) Sentence keywords
associates, stimulates, transcription activity

(4) Sentence keywords after manual curation
transcription activity

MeSH keywords extraction

To extract MeSH keywords, we searched for the gene names in our gene lists in the title and abstract of MEDLINE citations related to each gene and extracted the associated MeSH terms for each gene. The extracted gene-MeSH term list was represented by scores indicating the frequency of gene-MeSH term co-occurrence. Initial tests showed that certain MeSH keywords in the list were common biological terms and less useful from the point of view of gene annotation (e.g., human, DNA, animal, Support U.S Govt etc.). A collection of MeSH stop words was created manually and these terms were removed from the gene-MeSH term lists. Finally, from the thus filtered gene-MeSH lists, the 20 highest-frequency MeSH terms associated with each gene were taken as MeSH keywords associated with each gene. For example the MeSH keywords associated with a gene “FOS” in our gene list are oncogene, felypressin, transcription-factor, thermoreceptors, DNA-binding, antibiosis, inflammatory-response, zinc-fingers, gene-regulation, and neuronal-plasticity.

GO keyword extraction

We used the GO keywords information incorporated in Gene Ontology **[**Error! Bookmark not defined.**]** to extract GO keywords associated with each gene. Out of the three GO annotation categories we included only molecular function and biological process as we believe that cellular component (e.g. nucleus, cell membrane etc.) is less important for characterizing genes in the context of this study. Further, due to the hierarchical nature of GO and multiple inheritance in the GO structure, we consider with every ancestor up to level 2 in the GO tree in assigning GO keywords. This enables us to use more generalized GO terms. For example the GO keywords associated with the gene “FOS” in our gene list are protein-dimerization, DNA binding, RNA polymerase, transcription factor, DNA methylation, inflammatory-response, and nucleus.

Keyword representation and calculation of numeric vectors

After the keyword extraction phase, each gene was described by a list of keywords extracted from MEDLINE abstract sentences, MeSH terms and GO terms. These keyword vectors then served as a basis for clustering (i.e., unsupervised class discovery). To do this, each term vector needed to be represented by a numeric vector representing the relative importance of keywords for each gene. This process is concerned with computing the numeric weight, w_{ij}, for each gene-term pair (g_i, t_j) (i = 1, 2,…n and j = 1, 2, … k) to represent the gene's characteristics in terms of the associated keywords. Common techniques for such numeric encoding includes (1) Binary, the presence or absence of a keyword relative to a gene, (2) Term frequency, he frequency of occurrence of a keyword with a gene (3) Term frequency × inverse document frequency (TF*IDF), the relative frequency of occurrence of a keyword with a gene compared to other genes.

As we derived the keywords from gene annotation sentences but not from full abstracts, we found the number of keywords associated with each gene is small. We noticed also that absolute frequency of most keywords tended be one. Therefore, we adopted the binary encoding scheme as illustrated in Table 2 in supplementary material, in which each gene is represented by a vector of 'normalized' absolute keywords frequencies. The 'normalized' absolute frequency of each vector element (keyword) is either zero or one.

Gene clustering

Clustering is a data mining technique that groups or clusters data components (typically represented as numeric vectors) according to their similarity or dissimilarity. The goal is to maximize intra-cluster and minimize inter-cluster similarity among the components. **[32, 33]** Clustering is typically used to identify sample groups in data. Unlike supervised learning methods that require explicit class label information, clustering is unsupervised and no information about target groups (classes) is used. Two basic approaches to clustering can be distinguished, hierarchical clustering (e.g., agglomerative and divisive) and non-hierarchical clustering (e.g., k-means/c-means clustering). Agglomerative hierarchical clustering starts with each object representing a cluster and then merges the clusters in sequence. Divisive hierarchical clustering starts with all samples in one cluster and successively split clusters. In hierarchical clustering the distance (similarity) between clusters is measured using different techniques such as single linkage, average linkage or complete linkage **[32]** and basic distance and similarity metrics (e.g., Euclidean, Minkowski, Hamming distance). K-means clustering requires a priori specification of the desired number of clusters, k. This method clusters data into groups by iteratively optimizing the positions of cluster centers (means) so that the sum of within-cluster similarities (the similarity between data points and their cluster centers) is maximized.

Essentially, the sentence-level binary coding scheme adopted in this study consists of numeric row vectors representing genes (via the associated biological function/process terms), and numeric column vectors representing annotation terms (via the associated genes). These two sets of vectors can be independently clustered using available clustering algorithms and tools. This approach can produce useful and specific information about the biological characteristics of sets of genes. In this study, we have used average linkage hierarchical clustering algorithm. **[33]** Using this algorithm has two advantages for this study. First, clustograms, a visualization of the substructures contained in a gene collection are produced, and second, individual clusters of genes are identified by clustogram splits at different levels. Clustering was performed using Cluto **[34]** and Cluster/Treeview **[35]** facilitates visualization of the clustograms.

Results and Discussion:

Evaluation

To obtain a quantitative measure on the performance of the various keyword encoding schemes, we developed a text corpus of 200 manually annotated abstracts based on two cancer categories brain tumor and breast cancer of our interest (see Table 4 under supplementary material). We used the following procedure to establish the corpus: (1)Determine randomly two cancer categories (brain tumor and breast cancer), (2) For each cancer category, select randomly 10 genes from Entrez such that species = human and number of associated abstracts ≥ 50, (3)For each gene identified in this way, select randomly 10 abstracts, resulting in a total of 200 abstracts; 10 abstracts for each of the 10 genes associated with each of the two cancer categories, (4) For each of the 200 abstracts, identify manually the keywords characterizing biological function and processes from abstracts, MeSH terms and GO terms.

With this text corpus we were able to construct a matrix containing all 20 genes and their associated keywords and keyword frequencies from abstracts, MeSH terms and Go terms. The manually annotated corpus of 200 abstracts and the matrix of 20 annotated genes served as gold standard for our evaluation experiments. We carried our four evaluation experiments: (1) Abstract keywords (baseline). Extracts gene annotation terms based on term frequencies * inverse document frequencies (TF*IDF) within the entire abstract without regard to sentence structure, (2) Sentence keywords. Extracts gene annotation terms based sentence-level keywords, (3) Sentence + MeSH keywords. As in (2) above plus MeSH terms (see Section MeSH keywords extraction), (4) Sentence + MeSH + GO keywords. As in (2) above plus MeSH terms (see Section MeSH keywords extraction) and GO terms (see Section GO keyword extraction).

Essentially, in each evaluation experiment the input is the text corpus of 200 abstracts and the output is a list of genes with its predicted annotation terms. Informally, the closer the predicted annotation terms match the manually established annotation terms, the better is the method. Performance is measured via commonly used criteria such a recall (analogous to sensitivity), precision (analogous to positive predictive value) and the F-measure (a score that combines recall and precision). The results we obtained are shown in Table 5 (below in supplementary material).

We notice that the baseline method comprising TF*IDF keywords fares worst among all four approaches. We interpret this as evidence for the validity of the methods involving sentence-level processing as this information is likely to carry most specific characterizing terms. The 'brute-force' abstract-level processing will have difficulty in extracting these terms correctly and consistently. We further notice that the substantial improvements of precision and recall when we include MeSH terms and GO terms. This may be because these two categories are more specific and MeSH and GO annotations were done using full-papers and these biological functions and process are not described in all abstracts.

Clustering of genes resulting from microarray experiment

To demonstrate the usefulness of the presented keyword-extraction techniques to microarray data analysis, this method was applied to annotate and cluster gene lists that were found differentially expressed in a microarray experiment investigating the impact of two mitogenic proteins, Epidermal growth factor (EGF) and Sphingosine 1-phosphate (S1P), on glioblastoma cell lines **[36]**. The microarray data set reveals three sets of differentially expressed genes ($p<0.05$), namely, genes differentially expressed with response to EGF, G(EGF), genes differentially expressed with respect to S1P, G(S1P) and genes differently expressed in response to both, G(COM).

Genes were considered differentially expressed if their p-value is smaller than 0.05. We found that, when compared to the resting state, 19 genes were significantly differentially expressed as a response to EGF, 35 genes as a response to S1P and 30 genes as a response to COM, i.e., combined stimuli of S1P and EGF. The three gene lists are referred to as G(EGF), G(S1P) and G(COM), respectively (see Table 6 in supplementary material).

Using these the three gene lists obtained from the microarray experiment (Table 6 shown in supplementary material) as query in MEDLINE returned the three corresponding sets of abstracts A(EGF), A(S1P) and A(COM), respectively. The abstracts were processed with the keyword extraction method involving sentence-level, MeSH and GO terms and the resulting representations were clustered using average linkage hierarchical clustering algorithm. Our gene clustering strategy and clustering algorithms are explained in the Methodology section. The resulting clustograms are presented in Figure 2, Figure 3, and Figure 4, respectively.

The clustograms depict associations between genes and biological function/process terms derived from the abstracts obtained with the various gene lists. For the investigating scientist, the clustograms fulfill the following main functions: (1) Squares highlighted in a horizontal line link a gene to one or more biological functions or processes. This is useful to see which genes are associated with which functions/processes and which genes have few or many associations. The interpretation of many and few is very much dependent on the associated biological function/process categories, the particular scientific question under investigation, and also on how extensively a particular gene has been researched and reported in the literature. (2) Users may visually delineate clusters, i.e., rectangular areas with many highlighted squares in them and few highlighted squares around them. Any cluster, small or large, is potentially very useful to have discovered. Each cluster identified in this way relates a set of genes to a group of biological functions and processes. In a sense, each gene in the clustered is characterized by the same set of biological function and process concepts, a kind of 'guilt by association'. This information is extremely useful as it provides clues as to the roles genes may play collectively in pathways and functions, processes, and possible phenotypes, that are associated with these pathways.

Summary of analysis of EGF cluster, G(EGF)

The clustograms in Figure 2 show the results obtained from extracting the sentence-level function/process keywords (plus MeSH and GO terms) from 28,913 abstracts (for the 19 genes detected in response to EGF stimulus) and the subsequent clustering. In Figure 2a several individual genes with very many (e.g., CALD1, CLU, FOS) and very few (e.g., HRY, DUSP6) associations stand out. Another interesting feature is the large cluster at the lower left corner of Figure 2a (reproduced in more detail in Figure 2b) containing the genes DUSP, ID1, KLF2, CALD1, ABCA, CLU, FOS, JUN and SLC5A3. Many genes in this cluster are associated with the same set of keywords (transcription factor, cell death and secretion).

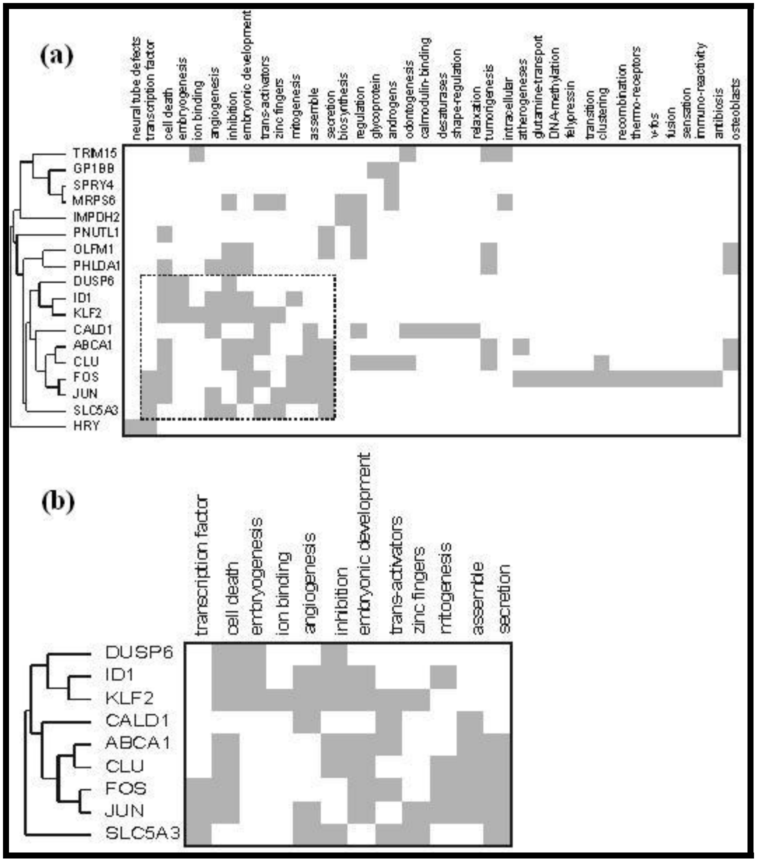

Figure 2: Characterization 19 genes differentially expressed genes in response to EGF. (a) All 19 genes against the discovered biological function/process terms. (b) Detailed view of group of manually selected cluster sharing common features (9 genes and 12 function/process terms)

Summary of analysis of S1P cluster, G(S1P)

The clustograms in Figure 3 show the results obtained from extracting the sentence-level function/process keywords (plus MeSH and GO terms) from 19,705 abstracts (for the 30 genes detected in response to S1P stimulus) and the subsequent clustering. In Figure 3a several individual genes with very many (e.g., CCL3, IL6, IL8, F3) and very few (e.g., HERB2, DOC1) associations stand out. Another interesting feature is the large cluster at the upper left corner of Figure 3a (reproduced in more detail in Figure 3b containing the genes TNAIP, KLF5, BCL6, NAB1, BTG1, NFKBIA, NR4A1, SOCS5, CITED2, NRG1, JAG1, PLAU, CCL2, IL8, IL6, GLIPR1, F3, MAP2K3, and EHD1. Many genes in this cluster are associated with the same set of keywords (atherogenesis, mitogenesis, assemble, inflammation, focal-contact, …, and protein-binding).

Summary of analysis of the common gene cluster, G(COM)

The clustograms in Figure 4 show the results obtained from extracting the sentence-level function/process keywords (plus MeSH and GO terms) from 39,890 abstracts (for the 30 genes detected in response to EFG and S1P stimuli) and the subsequent clustering. In Figure 4a several individual genes with very many (e.g., MYC, MAFF, ATF3) and very few (e.g., DIPA, UGCG, SNARK) associations stand out. Another interesting feature is the large cluster at the upper left corner of Figure 4a (reproduced in more detail in Figure 4b containing the genes SPRY2, GEM, ZYX, NEDD9, MYC, LIF, SERPINE1, DTR, MUCL1, C8FW, MAFF, ATF3, RTP801, EGR1, JUNB, FOSL1, CEPED, TIEG, EGR2, EGR3, and ZFP36. Many genes in this cluster are associated with the same set of keywords (DNA binding, zinc fingers, repressor proteins, …, and mitosis).

An important aim in microarray data mining is to bind transcriptionally modulated genes to functional pathways or to understand how transcriptional modulation can be associated with specific biological events such as genetic disease phenotype, molecular mechanism of drug action, cell differentiation etc. However, the amount of functional annotation available with each transcriptionaly modulated genes is still a limiting factor because not all genes are well annotated. Our functional clustering/grouping will enable to select literally informative genes (Figure 2b, Figure 3b, and Figure 4b) for further investigations in the above data mining and knowledge discovery pipeline. Our evaluation suggests that this approach will provide more specific and useful information than typical approaches using abstract-level information. This is particularly the case when the sentence-level terms are augmented by MeSH and GO keywords.

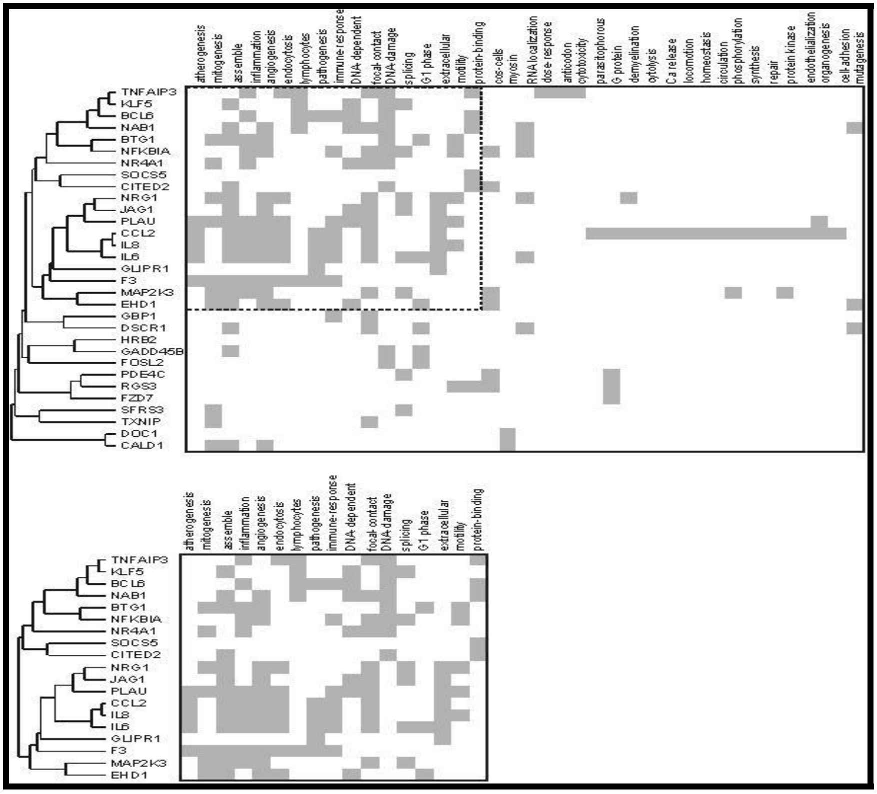

Figure 3: Characterization 30 genes differentially expressed genes in response to S1P. (a) All 30 genes against the discovered biological function/process terms. (b) Detailed view of group of manually selected cluster sharing common features (19 genes and 17 function/process terms)

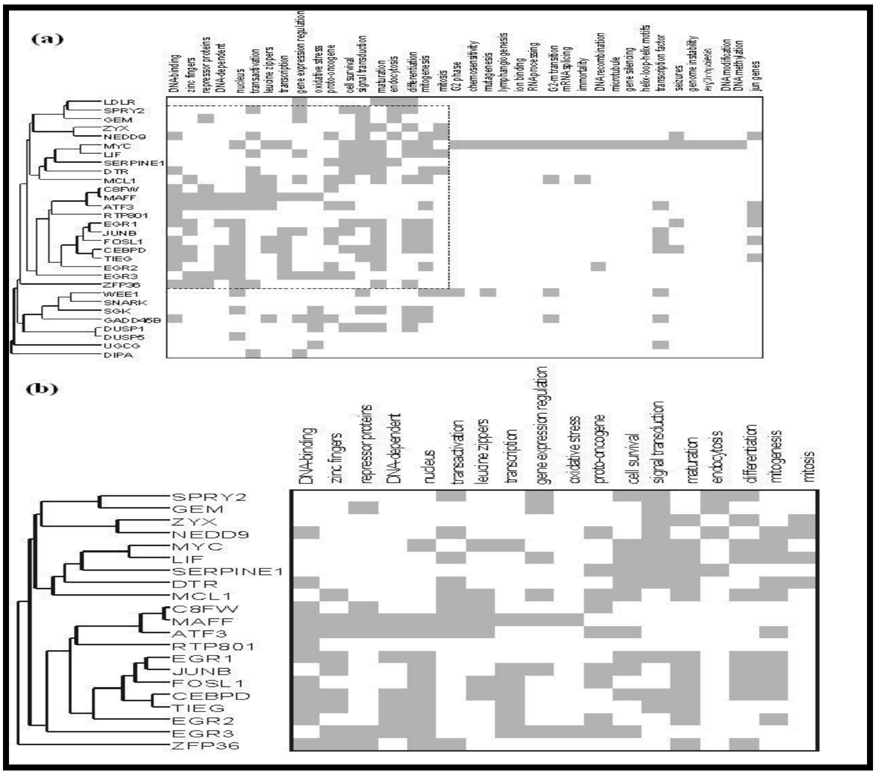

Figure 4: Characterization 30 genes differentially expressed genes in response to both EGF and S1P. (a) All 30 genes against the discovered biological function/process terms. (b) Detailed view of manually selected cluster sharing common features (21 genes and 18 function/process terms)

Conclusion:

The sequencing of whole genomes and the introduction high throughput analysis (e.g., oligonucleotide and cDNA chips, MALDI/SELDI-TOF MS) provides biomedical research with a global perspective, which necessitates the development of novel mining tools to explore and interpret data in timely manner. This paper presents a novel approach to combine sentence-level keywords with GO and MeSH terms. In our evaluation experiment, this approach has shown promising results. The present evaluation suggests that this approach will provide more specific information than typical approaches using abstract-level information. This is particularly the case when the sentence-level terms are complemented by MeSH and GO terms. Further, clustering of genes into different functional groups based on literature keywords has the potential to help biologists identify and characterize literally informative genes of interest for further investigations.

Future work:

Future enhancements of the system will include additional data resources (OMIM. DIP, KEGG) and the generation of association rules to identify correlations among genes in the same cluster. Association rules between the genes in the same cluster seem particularly interesting because it allows one to find the presence of regularities between gene groups. Finally, abstracts were used in this study as they are readily and easily available but they are limited in content. As full-text contains large number of irrelevent sentences compared to abstracts this approach may be useful for full-text analysis too, as it performs filtering of irrelevant sentences before clustering. The plan to perform the current study with full-text articles and compare the results with that of abstracts is on the way.

References:

[01] A. Schulze & J. Downward, *Nat. Cell. Biol.*, 8: E190 (2001) [PMID: 11483980]

[02] M. Schena, *et al., Science,* 270: 467 (1995) [PMID: 7569999]

[03] M. B. Eisen, *et al., Proc. Natl. Acad. Sci.,* 95: 14863 (1998) [PMID: 9843981]

[04] http://www.geneontology.org

[05] http://www.ncbi.nlm.nih.gov/mesh/meshhome.html

[06] http://dip.doe-mbi.ucla.edu/

[07] J. Natarajan, *et al., Crit. Rev. Biotechnol.,* 869: 139 (2005) [PMID: 15999851]

[08] http://www.ncbi.nlm.nih.gov

[09] M. Krallinger & A. Valencia, *Genome Biology*, 6: 224 (2005) [PMID: 15998455]

[10] L. Wong, *Pac. Sym. on Biocomp.*, 6: 520 (2001) [PMID: 11262970]

[11] J. C. Park, *et al., Pac. Sym. on Biocomp.,* 6: 396 (2001) [PMID: 11262958]

[12] A. Yakushiji, *et al., Pac. Sym. on Biocomp.,* 6: 408 (2001) [PMID: 11262959]

[13] C. Friedman, *et al., Bioinformatics Suppl.,* 1: 74 (2001) [PMID: 11472995]

[14] T. Sekimizu, *et al., Genome Inform Ser Workshop Genome Inform.,* 9: 62 (1998) [PMID: 11072322]

[15] M. Craven & J. Kumlien, *Proc Int Conf Intell Syst Mol Biol.*, 77 (1999) [PMID: 10786289]

[16] B. J. Stapley, *et al., Pac. Sym. on Biocomp.,* 7: 374 (2002) [PMID: 11928491]

[17] T. C. Rindflesch, *et al., Proc AMIA Symp.,* 127 (2000) [PMID: 10566334]

[18] T. C. Rindflesch, *et al., Pac. Sym. on Biocomp.,* 5: 517 (2000) [PMID: 10902199]

[19] http://www.bioinformatica.unito.it/bioinformatics/medmole/welcome.html.

[20] M. A. Andrade & A. Valencia, *Bioinformatics,* 14: 600 (1998) [PMID: 9730925]

[21] H. Shatkay, *et al., Proc Int Conf Intell Syst Mol Biol.*, 317 (2000) [PMID: 10977093]

[22] T. K. Jenssen, *et al., Nat. Genet.,* 28: 21 (2001) [PMID: 11326270]

[23] L. Tanabe, *et al., Biotechniques,* 27: 1210 (1999) [PMID: 10631500]

[24] Y. Liu, *et al., IEEE/ACM Trans. on Comput. Biol. and Bioinf.,* 2: 62 (2005) [PMID: 17044165]

[25] D. Chaussabel & A. Sher, *Genome Biology,* 3: 10 (2002) [PMID: 12372143]

[26] http://www.ncbi.nih.gov/entrez/

[27] http://bioinformatics.weizmann.ac.il/cards/

[28] http://ca.expasy.org/sprot/

[29] http://www.cse.ucsc.edu/centers/cbe/Genome/

[30] http://www.gene.ucl.ac.uk/hugo/

[31] http://www.cs.jhu.edu/~brill/

[32] P. Willet, *Information Processing and Management,* 24: 577 (1988)

[33] A. K. Jain, *et al., ACM Computing surveys,* 31: 264 (1998)

[34] http://www-users.cs.umn.edu/~karypis/cluto/

[35] http://bonsai.imas.u-tokyo.ac.jp/~mdehoon/software/cluster

[36] J. Natarajan, *et al., BMC Bioinformatics,* 7: 373 (2006) [PMID: 16901352]

Edited by T. W. Tan & S. Ranganathan

Citation: Natarajan & Ganapathy, Bioinformation 2(5): 185-193 (2007)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Supplementary material

Name of Expression	Expression Pattern	Sentence Output
Nouns describing agents	($gene (is)? (the\|an\|a) @{0,2}$action of @{0,2} $gene)	IL6, a known mediator of STAT3 response
Nouns describing actions	($gene @{0,6} $action (of|with) @{0,1} $gene)	abi5 domains required for interaction with abi3
Passive verbs	($gene @{0.6} (is|was|be|are|were) @{0,1} $action $(by|via|through) @{0,3} $gene)	Protein kinase c (PKC) has been shown to be activated by parathyroid hormone
Active verbs	($gene $sub-action @{0,1} $action @{0,2} $gene)	Insulin mediated inhibition of hormone sensitivity lipase activity

Table 1: An example set of regular expressions for nouns describing agents and agents, and passive and active verbs

Genes / Terms	t_1	t_2	...	t_k
g_1	$w_{11} = 0$	$w_{21} = 1$	...	$w_{k1} = 1$
g_2	$w_{12} = 1$	$w_{22} = 1$	...	$w_{k2} = 0$
...	...	...	...	...
g_n	$w_{1n} = 0$	$w_{2n} = 0$	...	$w_{kn} = 1$

Table 2: Binary representation of genes: w_{ij} represents the 'normalized' absolute keyword frequency of the keyword (or term) t_j for gene g_i (see also illustration in Table 3)

Genes / Terms	cell death	zinc fingers	…	DNA methylation
HRY	0	1	…	1
KLF2	0	0	…	1
ID1	1	1	…	0
JUN	1	0	…	0
DUSP6	0	0	…	0
…	…	…	…	…

Table 3: Rudimentary example of gene representation based on gene list *G*(*EFG*)

Genes	Category
ADAM23, DKK1, IGF2, LRRC4, L3MBTL, MMP9, MSH2, PTPNS1, SFMBT1, ZIC1	Brain Tumor
AMPH, ATM, BRCA1, BRCA2, CHEK2, CDH1, PHB, TFF1, TSG101, XRCC3	Breast Cancer

Table 4: Test set of 20 human genes manually grouped into two cancer classes

Keywords Extraction Method	Precision	Recall	F-measure (%)
Abstract keywords (baseline)	0.31	0.24	27.05
Sentence keywords only	0.57	0.38	45.60
Sentence + MeSH keywords	0.64	0.47	54.19
Sentence + MeSH + GO keywords	0.78	0.72	74.88

Table 5: Precision, recall and F-measure of extracted keywords

Gene List	Name of Genes
G(*EGF*) (19 genes)	HRY, KLF2, ID1, JUN, DUSP6, IMPDH2, GP1BB, PNUTL1, CGI-96, CALD1, TRIM15, FOS, SPRY4, CLU, SLC5A3, MRPS6, ABCA1, OLFM1, PHLDA1
G(*S1P*) (35 genes)	F3, NR4A1, KLF5, GADD45B, IL8, CITED2, CALD1, IL6, BCL6, LBH, HRB2, KIAA0992, NFKBIA, TNFAIP3, CCL2, DSCR1, TXNIP, NAB1, EHD1, GBP1, GLIPR1, MAP2K3, FZD7, RGS3, SOCS5, FOSL2, JAG1, DOC1, NRG1, BTG1, PDE4C, KIAA1718, KIAA0346, SFRS3, PLAU
G(*COM*) (30 genes)	MAFF, DUSP5, EGR3, SERPINE1, ZFP36, DUSP1, LIF, DTR, MYC, GADD45B, RTP801, ATF3, JUNB, SNARK, WEE1, EGR2, TIEG, SPRY2, CEBPD, SGK, GEM, NEDD9, LDLR, EGR1, C8FW, UGCG, MCL1, ZYX, FOSL1, DIPA

Table 6: Gene lists for differentially expressed genes