BIOINFORMATICS

Vol. 17 no. 4 2001
Pages 319–326

Use of keyword hierarchies to interpret gene expression patterns

Daniel R. Masys[1,4,7], John B. Welsh[2,8], J. Lynn Fink[3], Michael Gribskov[5], Igor Klacansky[4] and Jacques Corbeil[1,4,6]

[1]Departments of Medicine, [2]Pathology, [3]Biology, [4]UCSD Cancer Center, University of California, San Diego, San Diego, CA 92093, USA, [5]San Diego Supercomputer Center and [6]Veterans Medical Research Foundation, 3350 La Jolla Village Drive, San Diego, CA 92161, USA

Received on August 11, 2000; revised on December 8, 2000; accepted on December 13, 2000

ABSTRACT

Motivation: High-density microarray technology permits the quantitative and simultaneous monitoring of thousands of genes. The interpretation challenge is to extract relevant information from this large amount of data. A growing variety of statistical analysis approaches are available to identify clusters of genes that share common expression characteristics, but provide no information regarding the biological similarities of genes within clusters. The published literature provides a potential source of information to assist in interpretation of clustering results.

Results: We describe a data mining method that uses indexing terms ('keywords') from the published literature linked to specific genes to present a view of the conceptual similarity of genes within a cluster or group of interest. The method takes advantage of the hierarchical nature of Medical Subject Headings used to index citations in the MEDLINE database, and the registry numbers applied to enzymes.

Availability: We have created a publicly accessible website that provides this form of gene expression interpretation at http://www.array.ucsd.edu.

Contact: dmasys@ucsd.edu

INTRODUCTION

The analytical challenges of interpreting gene expression data obtained from high-density cDNA and oligonucleotide probe microarrays are formidable. A variety of statistical approaches to expression analysis for gene microarray data have been reported, based on the correlation of numerical values of expression intensities (Carlisle *et al.*, 2000; Cho *et al.*, 1998; Eisen *et al.*, 1998; Ermolaeva *et al.*, 1998; Tamayo *et al.*, 1999; Golub *et al.*, 1999; Ross *et al.*, 2000). A common characteristic of purely numerical techniques is that they identify groups of genes of potential interest, but leave to the user the task of discovering and interpreting the biological similarities that may underlie the expression pattern. Since gene groups of interest may include dozens, hundreds or potentially even thousands of different genes, it is beyond the limits of unaided human cognition to detect and organize these data along multiple lines of conceptual similarity by inspection alone.

A plausible approach to categorizing the characteristics of known genes within a group of interest is to use the information content of published literature linked to those genes. A variety of approaches to this type of data mining have been described. Commercially available microarray interpretation software generally allows searching for results associated with specific genes by words included in a gene definition or description field, and may also provide HTML hyperlinks to specific citations in MEDLINE. Tanabe *et al.* (1999) created a web-based question answering utility for gene expression that exploits data linkages contained in GeneCard and PubMed database retrievals. Shatkay *et al.* (2000) have used information retrieval algorithms to find the literature most closely related to all of the genes contained in a microarray, and to predict relationships among genes independent of experimental values. This method depends upon identification (currently by a human expert) of a single best 'kernel' paper describing a gene. Marcotte *et al.* (1999) used SwissProt keywords to functionally characterize groups of genes identified by a variety of experimental and computational predictive methods.

A common approach to information retrieval from the biomedical literature is the use of 'keywords' representing the essential concepts contained within a text. The biomedical literature in the National Library of

[7]To whom correspondence should be addressed at: University of California, San Diego School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093-0602, USA.

[8]Present address: Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, CA 92121, USA.

Medicine's MEDLINE database is indexed by keywords drawn from a controlled terminology called Medical Subject Headings (MeSH) that was originally developed to categorize the citations contained in *Index Medicus*. The MeSH vocabulary contains 19 000 terms ('main headings'), approximately 300 000 synonyms for those terms, and 103 500 chemical names. An average of 10 MeSH indexing terms are applied to each MEDLINE citation by professional indexers, who choose these keywords after reading the full text of the article. Thus MeSH keywords serve as a telegraphic surrogate of the concepts contained in a journal article. Since 1987, GenBank accession numbers and other molecular database identifiers have been added as searchable descriptors for all articles written about specific gene sequences where the author provides the GenBank accession number in the text or as part of a footnote associated with the article. Further detail on the content and structure of MeSH is available at http://www.nlm.nih.gov/mesh/meshhome.html.

The MeSH indexing terms are organized into concept hierarchies (in formal information science terms, directed acyclic graphs) that represent 'is-a' and 'part-whole' relationships. For example, in the MeSH Disease terminology hierarchy, Multiple Sclerosis is an example of an Autoimmune Demyelinating disease, which is in turn an example of a Nervous System Disease. Similarly, in the Anatomy hierarchy the Lens is a part of the Eye, which is itself an instance of the Sense Organs. A common display convention for concept hierarchies is to display more specific terms indented under the more general term of which they are an example. Table 1 shows the top level nodes of the concept hierarchies in MeSH that determine the thematic areas by which the biomedical literature is indexed. A single term may exist in multiple hierarchies. In addition, for articles that discuss specific enzymes there are keywords and numerical codes drawn from the hierarchy of Enzyme Commission (EC) codes assigned by the Commission on Biological Nomenclature of the International Union of Pure and Applied Chemistry (IUPAC). See http://www.chem.qmw.ac.uk/iubmb/enzyme/ for a complete discussion of the EC code hierarchy. Use of hierarchical groupings of keywords defined for a specific database has been described by Tavazoie *et al.* (1999) in the functional characterization clusters of yeast genes derived from expression data.

The importance and utility of terminology hierarchies for assisting in characterizing the literature associated with specific groups of genes derive from several sources. The first is that variability in the terms applied by human indexers to the literature means that no single term can reliably retrieve all of the articles related to a particular term or concept. Secondly, the scope of similarity varies with the interests of the researcher: for example, for a specific hypothesis it may be of interest to find all genes indexed by a single molecule (e.g. glycosylphosphatidylinositol) while another hypothesis might relate to all molecules of the class 'phospholipids'; selecting the proper granularity depends upon knowledge of the hierarchy and of the number of matches that exist with any particular group of genes. In addition, participation in multiple concept hierarchies enables a gene to be viewed from multiple perspectives (e.g. arthritis is both an inflammatory disorder and a disease of bones and joints).

Table 1. Concept hierarchies of medical subject heading terms

• Anatomy
• Organisms
• Diseases
• Chemicals and drugs
• Analytical, diagnostic and therapeutic Techniques and equipment
• Psychiatry and psychology
• Biological sciences
• Physical sciences
• Anthropology, education, sociology and social phenomena
• Technology and food and beverages
• Information science
• Humanities
• Persons
• Health care
• Geographic locations

SYSTEM AND METHODS

We constructed a database of gene microarray identifiers (i.e. the names and codes used by the manufacturers for the individual detection units on their chip, filter, or array) and their associated GenBank accession numbers. We used a variety of array sources, including GeneChip® HuGeneFL, Cancer G100, U95a and Mu11K arrays (Affymetrix, Santa Clara, CA), Human UniGEM™ V 2.0 Clone Lists (Incyte Genomics, Palo Alto, CA), and also cluster identifiers from NCBI UniGene Build 108. Each of the GenBank accession numbers and UniGene cluster identifiers was used in an automated database search of MEDLINE via the PubMed interface, and up to 20 matching citations were downloaded. (20 citations was taken as the convenience sample for this proof-of-concept application, as it represented the upper limit of citations contained on a results page generated by PubMed.) The program that executed the search script for each gene identifier used the syntax published by the National Center for Biotechnology Information (NCBI) for automated searching of PubMed MEDLINE via the Internet (available at http://www.ncbi.nlm.nih.gov:80/entrez/utils/pmqty_help.html). A parsing program (i.e. a program designed to analyze the text contained in the

Table 2. Representative data linkages among commercially available microarray identifiers and the published literature. Where multiple alternative forms exist for the identifier of an individual locus in array, all forms are included, thus the number of identifiers exceeds the number of loci for most arrays

Array name	Array identifiers	GenBank accession nos	Citations	Unique citations	Loci with no match	MeSH terms	Registry number terms	EC nos	Total index terms	Fraction of array with 1 or more matching citations
Affy-HuFL	8 693	6 941	8 771	6866	1551	54 455	26 498	5 190	80 953	77.6
Affy-U95a	17 768	13 488	9 061	7679	2934	107 493	52 377	10 287	159 870	78.8
Affy-Cancer	2 643	2 223	3 179	2553	452	20 801	10 197	2 275	30 998	79.6
Incyte Unigem v2	8 820	8 717	3 586	2654	6357	23 534	11 676	2 241	35 210	27.0
Unique totals	37 051	14 197	10 378	8106	7612	66 054	32 079	6 174	98 133	46.3

downloaded results) was used to extract MeSH (MH) terms and chemical Registry Number (RN) keywords from the saved search results of the individual gene identifier searches.

A linking file was constructed that included the microarray locus identifier and the unique MEDLINE identifiers of the citations retrieved using that identifier. To compute and display the keyword hierarchies we used Unified Medical Language System (UMLS^{TM}) Metathesaurus files provided via research license from the National Library of Medicine. The UMLS Metathesaurus is a vocabulary resource that contains linkages between commonly used biomedical coding and naming systems. Each term in the Metathesaurus is assigned a unique Concept Identifier contained in the UMLS Metathesaurus Relational CONcept file (MRCON), and relationships between concepts are represented in a Metathesaurus Relationships file (MRREL). Further detail on these publicly available terminology resources of the UMLS is available at http://www.nlm.nih.gov/research/umls/.

Each MeSH term and RN entry was translated via automated term lookup to its Concept Identifier, and the corresponding MeSH and Enzyme Commission hierarchical code numbers were retrieved.

A data mining method was developed that involves retrieving all keywords associated with the literature linked to any group of submitted gene identifiers. In a second step, the MeSH term and EC hierarchy numbers are used to retrieve all 'parent' terms (i.e. more general terms in the branching tree of terminology), up to and including the root of the term hierarchy. The number of matching gene and citation records is summed and displayed as a set of term hierarchy trees with more specific terms indented under their predecessors. At each node in the term hierarchy a cumulative total of matching records is displayed in HTML format as a hypertext link that retrieves a summary-by-gene of the matching keywords. This summary includes links to the PubMed citation that caused the match to occur, the GenBank sequence record, GeneCard summary, and the submitted experimental expression values for each gene identifier. The highest frequency matches are displayed in a color that makes them easily discernible among the set of all keywords.

To provide a measure of whether any particular keyword would be expected by chance to be associated with a group of genes, we applied the method of retrieving literature-associated keywords to 500 groups of 100 randomly selected genes drawn from the list of 37 000 possible unique gene identifiers in our database. Individual keyword frequencies were expressed as the count of the number of times a keyword appeared in association with a randomly chosen gene group, divided by the total number of keywords associated with that group. Mean frequencies and standard deviations were calculated, and frequency distributions were evaluated and shown to follow Gaussian distributions. These keyword frequencies from randomly generated gene 'clusters' were then used to compare the observed versus expected frequency of each keyword retrieved in association with a newly-defined set of genes, and a P-value generated to represent an estimate of the likelihood that the keyword would appear at or above the observed frequency by chance, in a group of keywords of the same size as that observed. These P-value estimates are displayed for each term in the hierarchy.

IMPLEMENTATION

A representative set of linkages between microarray identifiers of commercially available arrays and the controlled vocabulary terms describing literature associated with those identifiers is shown in Table 2. Overall to date, we have found one or more published citations associated with 46% of genes on commercially available arrays, with a range from 79% of the human genes found on chips made by photolithography-based oligonucleotide synthesis to 27% of genes based on UniGene loci that include ESTs. The number of literature citation links and associated keyword to microarray identifiers are grow-

Table 3. Genes whose relative overexpression is predictive of ALL and AML, as described by Golub *et al.* (1999)

Genes predictive of Acute Lymphocytic Leukemia (ALL)	Genes predictive of Acute Myelogenous Leukemia (AML)
U22376 c-myb	M55150 Fumarylacetoacetate hydrolase
X59417 Proteasome iota PROS-27	X95735 Zyxin 2
U05259 MB-1	U50136 LTC4S Leukotriene C4 synthase
M92287 cyclin D3	M16038 LYN tyrosine kinase
M31211 Myosin light chain	U82759 HoxA9 Homeodomain protein
X74262 RbAp48 retinoblastoma binding protein	M23197 CD33 Human differentiation antigen
D26156 Transcriptional activator hSNF2b	M84526 Adipsin/complement factor D
S50223 HKR-T1 = Kruppel-like zinc finger protein	Y12670 Leptin receptor
M31523 E2A transcription factor	M27891 CST3 cystatin C
L47738 Inducible protein	X17042 Hematopoetic proteoglycan core protein
U32944 Dynein light chain 1	Y00787 MDNCF monocyte-derived neutrophil
Z15115 TOP2 DNA topoisomerase II)	chemotactic factor
X15949 IRF2 Interferon regulatory factor	M96326 Azurocidin
X63469 TFIIE beta transcription factor	U46751 p62 for the Lck SH2 domain
M91432 MCAD medium-chain acyl-CoA	M80254 hCyP3 Cyclophilin isoform
dehydrogenase	L08246 MCL1 Myeloid cell differentiation protein
U29175 BRG1 Transcriptional activator	M62762 Vacuolar H+ ATPase proton channel subunit
Z69881 Ca^{2+} ATPase	M28130 Interleukin 8 (IL8)
U20998 SRP 9 Signal recognition particle subunit 9	M63138 Cathepsin D (catD) gene
D38073 MCM3 hRlf beta subunit (p102 protein)	M57710 Epsilon-BP IgE-binding protein
U26266 Deoxyhypusine synthase	M69043 MAD-3 mRNA encoding IkB-like activity
M31303 Op18 Oncoprotein 18	M81695 Leukocyte adhesion glycoprotein p150,95
Y08612 Rabaptin Nup88 protein	X85116 Epb72
U35451 Heterochromatin protein	M19045 Lysozyme mRNA
M29696 IL-7 Interleukin-7 receptor	M83652 Properdin
M13792 ADA Adenosine deaminase (ADA)	X04085 Catalase

ing incrementally as we add additional array identifier links and as additional literature indexed by GenBank accession numbers and UniGene identifiers is published and indexed in MEDLINE. The 8106 unique literature citations identified to date are characterized by just over 98 000 controlled vocabulary terms, about one third of which are chemical RNs.

A sample use of this approach to data mining is shown by applying the keyword analysis methods described here to the publication by Golub *et al.* (1999) describing statistical methods for classifying cancers. This study used commercially available GeneChips to derive a set of genes whose expression value differences were able to predict, without prior biological knowledge, whether a leukemia sample was derived from a myeloid or lymphoblastic cell line. Their statistical method selected a group of 50 genes, 25 of which were highly expressed in Acute Lymphoblastic Leukemia (ALL) relative to the overall expression mean, and 25 of which were highly expressed in Acute Myelogenous Leukemia (AML). Table 3 shows a listing of these genes. The GenBank identifiers of the two groups of 25 genes identified in this study were analyzed to determine the common themes evident in the published literature linked to these identifiers, and compare the keyword characterizations of genes predictive of AML versus the characterizations of genes predictive of ALL. A total of 70 citations describing 44 of 50 genes were found (22 of 25 in each group), and these citations were indexed by 814 controlled terminology descriptors. A sample display of the organization of total matches by term hierarchy is shown in Figure 1. This summary

Keyword distribution of literature associated with this set of genes

Subject Keyword Areas	Term Matches
Enzyme Registry Numbers	40
Anatomy	101
Organisms	30
Diseases	31
Chemicals and Drugs	249
Analytical Techniques	36
Psychiatry and Psychology	1
Biological Sciences	110
Physical Sciences	14
Information Science	1
Health Care	1

Fig. 1. Summary of terminology matches by concept hierarchy for 50 genes described by Golub *et al.* (1999).

Enzyme Commission/ Registry Entries for ALL-predictive genes

```
Oxidoreductases (3) {>.6}
   Amine Oxidoreductases (3)  {<.001}
      Pyrroline Carboxylate Reductases (1)  {<.005}
Transferases (3) {>.13}
   Phosphotransferases (3)  {>.3}
      RNA Polymerase III (3)  {<.01}
Complement Activating Enzymes (30) {<.001}
   Endonucleases (2)  {>.3}
      DNA Restriction Enzymes (2)  {>.3}
   Carboxypeptidases (7)  {<.05}
      Cysteine Endopeptidases (4)  {<.001}
      Aspartic Endopeptidases (2)  {<.001}
         Renin (2)  {<.001}
      prorenin (1)  {<.005}
         multicatalytic endopeptidase complex (1)  {<.001}
   Amidohydrolases (13)  {<.001}
      Nucleoside Deaminases (13)  {<.001}
         Adenosine Deaminase (8)  {<.001}
   Acid Anhydride Hydrolases (8)  {<.005}
      rac GTP-Binding Proteins (8)  {<.005}
         Dynein ATPase (5)  {<.001}
         Ca(2+-Transporting ATPase) (3)  {<.001}
Lyases (2) {~.07}
      Hydro-Lyases (2)  {<.001}
         Carbonate Dehydratase (2)  {<.001}
Isomerases (7) {<.001}
   DNA Helicases (7)  {<.001}
      DNA topoisomerase II alpha (5)  {<.001}
         DNA Topoisomerase (ATP-Hydrolysing) (5)  {<.001}
```

Enzyme Commission/ Registry Entries for AML-predictive genes

```
Oxidoreductases (4) {>.3}
          phospholipid-hydroperoxide glutathione
              peroxidase(4){<.001}
          Catalase (2)  {<.001}
          Peroxidase (2)  {<.001}
  Transferases (16) {~.03}
     Acyltransferases (2)  {~.07}
        dihydrolipoamide acyltransferase (2)  {~.03}
           Chloramphenicol O-Acetyltransferase (2)  {<.001}
     Alkyl and Aryl Transferases (5)  {<.001}
        p21(ras farnesyl-protein transferase) (5)  {<.001}
           Spermidine Synthase (2)  {<.001}
           Glutathione Transferase (2)  {<.005}
           leukotriene-C4 synthase (1)  {<.001}
     Phosphotransferases (9)  {>.13}
        c-CrkII protein (7)  {~.07}
        Receptor Protein-Tyrosine Kinases (2)  {>.13}
           Protein-Tyrosine Kinase (5)  {<.005}
  Complement Activating Enzymes (44) {<.001}
     Endonucleases (9)  {~.03}
        Phospholipases A (2)  {<.01}
        DNA Restriction Enzymes (5)  {<.01}
           Deoxyribonucleases, Type II Site-Specific (2)  {<.001}
           Aspergillus Nuclease S1 (1)  {<.001}
        Glucosidases (4)  {<.005}
           Muramidase (2)  {<.001}
           beta-Galactosidase (2)  {<.001}
     Carboxypeptidases (16)  {<.001}
        Kallikreins (14)  {<.001}
           Pancreatic Elastase (2)  {<.001}
           Complement Factor D (2)  {<.001}
           Complement Factor B (2)  {<.001}
           myeloblastin (1)  {<.001}
        Aspartic Endopeptidases (2)  {<.001}
           Cathepsin D (2)  {<.001}
     Acid Anhydride Hydrolases (7)  {<.01}
        rac GTP-Binding Proteins (7)  {<.01}
           Adenosinetriphosphatase (2)  {<.005}
           H(+-Transporting ATP Synthase) (4)  {<.001}
           fumarylacetoacetase (3)  {<.001}
  Isomerases (8) {<.001}
     Racemases and Epimerases (5)  {<.001}
        Amino Acid Isomerases (5)  {<.001}
        cyclophilin C (3)  {<.001}
           Peptidylprolyl Isomerase (3)  {<.001}
```

Fig. 2. Summary of concept hierarchy matches for Enzyme Commission terms (EC numbers) for genes described by Golub *et al.* (1999). Left panel shows hierarchy of genes predictive for ALL; right panel shows analogous term hierarchy for genes predictive of AML. Matching term numbers are in parentheses, and are hyperlinks to detail pages that provide Entrez links to gene sequence, GeneCard record, and the specific MEDLINE citation that caused the match. Values in curly braces {} are *P*-value estimates of the probability that a keyword would appear with the observed frequency by chance.

serves as a 'table of contents' to more detailed concept hierarchies.

Figure 2 shows a comparison of the hierarchical display of Enzyme Commission descriptors for ALL versus AML predictive genes. In both sets of genes, the most common enzymatic descriptor class is that of Complement-Activating Enzymes. In the ALL-predictive set of genes, these enzyme descriptors including endonucleases, endopeptidases, amidohydrolases, and acid anhydride hydrolases. In the AML-predictive set, several plasminogen activators (e.g. kallikreins) occur as keywords, a finding that potentially correlates with defibrination syndromes and other hemostatic abnormalities that are associated with AML but not with ALL. Overall, complement activation is a common and potentially clinically significant phenomena in acute leukemias, and the high frequency of this descriptor in the set of highly expressed genes is consistent with the authors' observations that informative genes were not merely markers of hematopoeitic lineage, but encoded proteins important

Diseases Associated with ALL-predictive genes

```
Virus Diseases (1) {>.3}
Neoplasms (5) {>.3}
   Neoplasms by Histologic Type (4) {>.6}
      Leukemia (4) {<.001}
         Leukemia, Lymphocytic (3) {<.001}
            Leukemia, B-Cell (1) {<.001}
               Leukemia, B-Cell, Acute (1) {<.001}
            Leukemia, Lymphocytic, Acute (1) {<.001}
               Leukemia, B-Cell, Acute (1) {<.001}
            Leukemia, T-Cell (1) {<.001}
   Precancerous Conditions (1) {<.001}
      Preleukemia (1) {<.001}
Nervous System Diseases (2) {>.3}
   Autoimmune Diseases of the Nervous System (1) {<.001}
      Demyelinating Autoimmune Diseases, CNS (1) {<.001}
         Multiple Sclerosis (1) {<.001}
   Demyelinating Diseases (1) {<.001}
      Demyelinating Autoimmune Diseases, CNS (1) {<.001}
         Multiple Sclerosis (1) {<.001}
Female Genital Diseases and Pregnancy Complications
       (1){>.6}
   Genital Diseases, Female (1) {>.6}
      Infertility (1) {<.001}
         Infertility, Female (1) {<.001}
Hemic and Lymphatic Diseases (1) {>.6}
   Hematologic Diseases (1) {>.6}
      Preleukemia (1) {<.001}
Neonatal Diseases and Abnormalities (3) {>.3}
   Hereditary Diseases (2) {>.13}
      Werner Syndrome (1) {<.001}
   Infant, Newborn, Diseases (1) {<.001}
      Severe Combined Immunodeficiency (1) {<.001}
Immunologic Diseases (4) {<.01}
   Autoimmune Diseases (1) {>.13}
      Autoimmune Diseases of the Nervous System (1){<.001}
         Demyelinating Autoimmune Diseases, CNS (1){<.001}
            Multiple Sclerosis (1) {<.001}
   Immunologic Deficiency Syndromes (3) {<.001}
      Common Variable Immunodeficiency (1) {<.001}
      Severe Combined Immunodeficiency (1) {<.001}
Pathological Conditions, Signs and Symptoms (1) {>.3}
   Pathologic Processes (1) {>.13}
      Disease Attributes (1) {<.001}
         Acute Disease (1) {<.001}
```

Diseases Associated with AML-predictive genes

```
Neoplasms (5) {>.13}
   Cysts (1) {<.001}
      Kidney, Cystic (1) {<.001}
         Kidney, Polycystic (1) {<.001}
   Neoplasms by Histologic Type (4) {>.6}
      Leukemia (4) {<.001}
         Leukemia, Hairy Cell (1) {<.001}
         Leukemia, Myeloid (3) {<.001}
            Leukemia, Myelomonocytic, Acute (1) {<.001}
            Leukemia, Nonlymphocytic, Acute (1) {<.005}
               Leukemia, Myelocytic, Acute (1) {<.001}
Urologic and Male Genital Diseases (2) {>.3}
   Urogenital Diseases (1) {>.13}
      Urogenital Abnormalities (1) {<.005}
         Kidney, Polycystic (1) {<.001}
   Urologic Diseases (1) {>.13}
      Kidney Diseases (1) {~.07}
         Kidney, Cystic (1) {<.001}
            Kidney, Polycystic (1) {<.001}
Female Genital Diseases and Pregnancy Complications (1) {>.6}
   Genital Diseases, Female (1) {>.6}
      Urogenital Diseases (1) {>.13}
         Urogenital Abnormalities (1) {<.001}
            Kidney, Polycystic (1) {<.001}
Hemic and Lymphatic Diseases (4) {>.3}
   Hematologic Diseases (1) {>.6}
      Bone Marrow Diseases (1) {~.03}
         Myelodysplastic Syndromes (1) {~.03}
            Leukemia, Myeloid (1) {<.005}
   Lymphatic Diseases (3) {<.005}
      Lymphoproliferative Disorders (3) {<.005}
         Leukemia, Hairy Cell (1) {<.001}
         Leukemia, Myeloid (2) {<.001}
            Leukemia, Nonlymphocytic, Acute (1) {<.001}
               Leukemia, Myelocytic, Acute (1) {<.001}
Neonatal Diseases and Abnormalities (1) {>.3}
   Abnormalities (1) {>.6}
      Urogenital Abnormalities (1) {<.005}
         Kidney, Polycystic (1) {<.001}
Immunologic Diseases (3) {~.07}
   Immunoproliferative Disorders (3) {<.005}
      Lymphoproliferative Disorders (3) {<.001}
         Leukemia, Hairy Cell (1) {<.001}
         Leukemia, Myeloid (2) {<.001}
            Leukemia, Nonlymphocytic, Acute (1) {<.001}
               Leukemia, Myelocytic, Acute (1) {<.001}
Pathological Conditions, Signs and Symptoms (1) {>.6}
   Pathologic Processes (1) {>.13}
      Inflammation (1) {<.001}
```

Fig. 3. Summary of concept hierarchy matches for Disease-related MeSH terms for genes described by Golub *et al.* (1999). Left panel shows hierarchy of genes predictive for ALL; right panel shows analogous term hierarchy for genes predictive of AML. Summary of concept hierarchy matches for Disease-related MeSH terms.

in cancer pathogenesis. These conceptual similarities, revealed by the automated summing and organization of literature keywords associated with these 50 genes, is a new finding that complements the interpretations of the original paper's authors.

Figure 3 shows a comparison of the hierarchical display of Disease descriptors for ALL versus AML predictive genes. The disease hierarchy display for this gene set shown in Figure 3 not surprisingly contains a majority of links to literature citations about hematologic and neoplastic diseases. Appropriately, the keywords associated with the ALL predictive set are associated with literature describing B and T cell lymphocytic leukemias, and the analogous hierarchy of the AML predictive set is linked to citations describing myelogeneous leukemia. However, genes in the ALL predictive set are also implicated in inherited combined immunodeficiency, and multiple sclerosis. Both ALL and AML-predictive sets of genes contain accession numbers that link to literature on polycystic kidney.

The numbers in parentheses to the right of each term are a hyperlink to detailed displays such as that shown in Figure 4, that link to the GenBank sequence, corresponding GeneCard record, and the specific citation (PubMed link) that caused the match to occur.

Leukemia C4.557.337

GeneCards Link	Accession # (Entrez Link)	Citation (PubMed link)	UNIQID	NAME
GeneCard	M31303	92011487	M31303	Human oncoprotein 18 (Op18) gene, complete cds
GeneCard	M31523	90150282	M31523	Human transcription factor (E2A) mRNA, complete cds
GeneCard	S50223	93043304	S50223	HKR-T1=Kruppel-like zinc finger protein [human, MOLT 4 T-cells, mRNA, 798 nt]

Return to Diseases index

Leukemia, Lymphocytic C4.557.337.428

GeneCards Link	Accession # (Entrez Link)	Citation (PubMed link)	UNIQID	NAME
GeneCard	M31523	90150282	M31523	Human transcription factor (E2A) mRNA, complete cds
GeneCard	S50223	93043304	S50223	HKR-T1=Kruppel-like zinc finger protein [human, MOLT 4 T-cells, mRNA, 798 nt]

Return to Diseases index

Leukemia, B-Cell C4.557.337.428.500

GeneCards Link	Accession # (Entrez Link)	Citation (PubMed link)	UNIQID	NAME
GeneCard	M31523	90150282	M31523	Human transcription factor (E2A) mRNA, complete cds

Leukemia, Lymphocytic, Acute C4.557.337.428.511

GeneCards Link	Accession # (Entrez Link)	Citation (PubMed link)	UNIQID	NAME
GeneCard	M31523	90150282	M31523	Human transcription factor (E2A) mRNA, complete cds

Return to Diseases index

Fig. 4. Detail of terms associated with gene loci, concept hierarchy numbers, and links to online sources of sequence and citations.

We have implemented this 'cluster mining' approach as a component of a publicly available set of microarray analysis tools named 'HAPI'—the High-density Array Pattern Interpreter system at http://www.array.ucsd.edu. The prototype website enables users to upload tab-delimited gene expression array data representing clusters or groups of interest containing up to 750 gene identifiers. There is no firm upper bound on the number of genes that can be analyzed in this manner, however, and future implementations on more powerful servers will enable larger numbers of genes to be analyzed simultaneously.

DISCUSSION

The measurement of the simultaneous expression values of thousands of genes creates a large amount of data whose interpretation by inspection may be likened to 'attempting to drink from a fire hose' (Waldrop, 1990). To date, most gene expression analysis tools described in the literature attempt to classify patterns of experimental results by statistical approaches. We describe an approach to microarray data interpretation based on the subject index terms that have been associated with the genes of the microarray as extracted from bibliographic and molecular sequence databases, combined with an estimate of the likelihood that such keyword associations would appear by chance. The method's principal strength is the use of concept hierarchies that are more robust than individual keyword comparisons for representing multiple potential contexts of similarity within groups of genes. The method is capable of simultaneously depicting similarities that may exist at the level of biological structure, molecular function, physiology and pathophysiology, and clinically manifest diseases, just as a single published article about a gene of interest may report findings in several of these dimensions. This characterization of groups of genes by biological concept complements purely mathematical approaches to gene microarray data analysis.

The method has several limitations. The analysis cannot help characterize anonymous ESTs and genes for which there is no computable link to the published literature, and overall at present less than half of the genes available on commercially available arrays have one or more literature citations linked to them. This situation should improve over time as additional literature is published that

assigns functions to genes that are currently anonymous. Commonly occurring keywords such as 'Amino Acid Sequence', 'Base Sequence', and keywords describing experimental methodologies such as 'Blotting, Northern' add little or no insight into the common properties of expressed genes. Better known and therefore better categorized genes, will yield more literature links and may bias the analysis by overrepresentation. Because ours is a data mining and exploration tool, we have not attempted to correct for such biases.

Over time, as the biomedical literature grows with new publications correlating primary nucleotide sequence with biological function, this method will become increasingly useful. We are developing methods to link keyword data mining to a variety of statistical clustering approaches, and to automatically update these linkages as new articles relevant to specific genes are published. We are also currently pursuing more speculative approaches to establishing links between gene groups and the literature, such as retrieving literature associated with homologous genes in other species based on sequence similarity, and expanding retrieval based on all named genes contained within UniGene clusters when any single gene or EST within that cluster is included in an analysis. However, even with its current limitations, it is evident that using the controlled terminology keywords of the published literature associated with groups of genes, and the organization of those keywords in biological concept hierarchies, is a useful 'cluster mining' approach that complements purely mathematical approaches to gene microarray data analysis.

ACKNOWLEDGEMENTS

We thank the UCSD Center for AIDS Research Genomics Core, UCSD Cancer Center, UCSD General Clinical Research Center, the University-wide AIDS Research Program, San Diego Veterans Medical Research Foundation (J.C.) and the School of Medicine, Office of the Dean (D.R.M.). Reena Deutsch and Deborah Hilton of the UCSD General Clinical Research Center provided consultation on the statistical methods for keyword frequency interpretation. We also thank T.Gingeras, D.Looney, D.Richman and A.Leigh Brown for thoughtful comments on the manuscript. Supported in part by National Institutes of Health Grants 5P30 AI3614-06 (author J.C.), 5U01CA84998-02 (authors D.R.M. and I.K.) and 5M01RR00827-24 (biostatisticians Deutsch and Hilton).

REFERENCES

Carlisle,A.J. *et al.* (2000) Development of a prostate cDNA microarray and statistical gene expression analysis package. *Mol. Carcinog.*, **28**, 12–22.

Cho,R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Ermolaeva,O. *et al.* (1998) Data management and analysis for gene expression arrays. *Nature Genet.*, **20**, 19–23.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Marcotte,E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Ross,D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.*, **24**, 227–235.

Shatkay,H. *et al.* (2000) Genes, themes and microarrays. Using information retrieval for large-scale gene analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 317–328, August 2000.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Tanabe,L. *et al.* (1999) MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, **27**, 1210–1214, 1216–1217.

Tavazoie,S. *et al.* (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.

Waldrop,M.M. (1990) Learning to drink from a fire hose. *Science*, **248**, 674–675.