Mining The Biomedical Literature; A Key Capability For Genomics Research

James P. Sluka, Ph.D.

INPHARMIX INCORPORATED

Contact:
James P. Sluka
Chief Scientist
InPharmix Incorporated
Email: JSluka@InPharmix.com
Phone: (317) 422-1464
Web: www.InPharmix.com

Abstract

As the use of genomic tools increases, there is a growing need for tools to effectively exploit the resulting data. Lists of genes that are related under an experimental paradigm are a common result of genomics methods such as subtracted libraries, differential display, and gene chip experiments. Currently, there are few tools for extracting useful information from these data sets.

Here we present an analysis of a typical genomics data set in which we efficiently extract the information required to further a research project. As our data set, we chose the work of Golub and coworkers (*Science*, **286**:531–537, 1999), in which cancer classification based on gene expression is applied to human acute leukemias. These workers examined acute leukemias arising from myeloid precursors (acute myeloid leukemia, **AML**) using RNA from patients' bone marrow mononuclear cells analyzed using oligonucleotide microarrays.

Analysis was conducted using PDQ_MED, a web based program created by InPharmix Inc. PDQ_MED is based on the assumption that if two genes are found to be related under an experimental paradigm, such as a gene chip experiment, then any literature which relates the two genes is of interest. PDQ_MED searches MEDLINE for abstracts that contain two or more of the terms in the user's query set. This pair-wise approach allows the researcher to effectively mine the more than five million abstracts in MEDLINE for information relevant to their research project. In addition, PDQ_MED can do sentence level proximity searching and properly handles abstract specific pseudonyms and acronyms.

We have used PDQ_MED to analyze the 24 genes in the **AML** dataset and added "acute myeloid leukemia" as an additional term. PDQ_MED executed 325 queries to MEDLINE in 4.4 minutes and identified 58,760 abstracts which refer to at least one of the 25 terms. PDQ_MED identified and analyzed a set of 17 terms which can be grouped together via the literature. In addition, there is literature directly linking seven of the terms with **AML**. PDQ_MED provides several graphical views of the interrelationships found in the literature as well as an extensively hyperlinked listing of the relevant sentences and abstracts.

Background

As the use of genomic tools increases, there is a growing need for tools to effectively exploit the resulting data. Lists of genes that are related under an experimental paradigm are a common result of genomics methods such as subtracted libraries, differential display, and gene chip experiments. Currently, there are few tools for extracting useful information from these data sets.

PDQ_MED Strategy

The tool we are introducing for this work is **PDQ_MED** (**P**air-wise **D**ata **Q**uery to **MED**LINE). PDQ_MED exhaustively searches MEDLINE for abstracts that contain two or more of the terms in the user's data set. This pair-wise approach allows the researcher to effectively mine the more than five million abstracts in MEDLINE for information relevant to their research project.

PDQ_MED is based on the assumption that if two genes are found to be related under some experimental paradigm, such as a gene chip experiment, then any literature which relates the two genes is of interest. A "co-occurrence" is defined as any abstract that contains two or more of the query terms. The simplest embodiment of this idea is to search MEDLNE (or other database) with all possible pairwise combination of the query terms. For N terms, approximately $N^2/2$ searches need to be conducted. For small values of N this can be done manually. For larger values, the number of searches required quickly becomes impractical.

The basic input to PDQ_MED is a list of query terms encompassing the genes, proteins, diseases or other concepts under investigation. An individual query term can consist of more than one version of a particular name. For example, a query can consist of a full name and an abbreviated name; `"Interleukin-1b, IL-1b"`, or alternative names; `"proteasome iota, macropain iota"`. Each term, delimited by white space or quote marks, is joined by an implied OR. In addition, the user may explicitly join phrases by any of the other Boolean operators, such as `BUTNOT`, or use any of the field or date operators supported by MEDLINE.

Proximity Search

A refinement to the basic search strategy is to require a higher degree of dependence, i.e., closer proximity within the document, between the two terms. In "Proximity" searching PDQ_MED examines all abstract containing two terms and determines if the terms co-occur in the same sentence. Sentence level proximity searching is not supported by MEDLINE.

One challenge to effectively employing proximity searching in the scientific literature is the highly variable nature of the names of genes, proteins and small molecules. As mentioned above, PDQ_MED allows the user to enter multiple names for the same entity, however, acronyms which are either common words, or used for more than one concept, are problematic. For example, a common acronym of "Acute Lymphoblastic Leukemia" is **ALL**. Since **ALL** is a common English word MEDLINE will not even search for abstracts containing it. In addition, it is common for more than one gene, protein or concept to use the same acronym. These problems with acronyms make proximity searching in the bio-medical literature difficult. Consider, for example, the abstrac t;

> In acute lymphoblastic leukemia (**ALL**), the cell surface ...

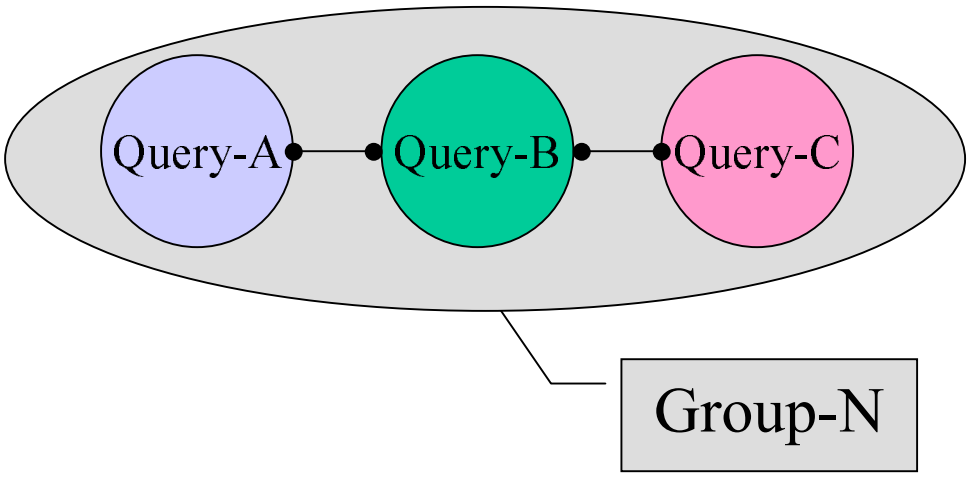
followed by several sentences and concluding with;

> ... GPRE also decreased the fraction of CD11-bearing **ALL** M2 and M5 cells.

In this case, the use of a "local acronym" (**ALL**) destroys simple proximity searching for the term "acute lymphoblastic leukemia". To circumvent this problem we use a proprietary algorithm to track local acronyms on a per abstract basis. This allows PDQ_MED to identify the GPRE + **ALL** sentence shown above as a proximity sentence without including the common word “ALL” in the MEDLINE query.

Term Grouping

After PDQ_MED has identified all of the abstracts containing two or more of the query phrases it uses a greedy clustering algorithm to organize the terms into groups. These groups represent sets of terms that co-occur in the literature. For example, if query-A and query-B co-occur in a set of abstracts and query-B and query-C co-occur in a different set of abstracts, then queries-A, B and C are clustered together in the same group. Groups suggest relationships between terms which are not explicitly present in MEDLINE. In this example, grouping suggests a possible relationship between query-A and query-C because of their common linkage to query-B, even though A and C do not explicitly co-occur.



Test Data Set

As an example data set, we have chosen the work of Golub *et al*., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring" (*Science*, 286 531–537,1999). In this work, cancer classification based on gene expression monitoring by DNA microarrays is applied to human acute leukemias. These workers compared the classification of acute leukemias into those arising from lymphoid precursors (acute lymphoblastic leukemia, **ALL**) or from myeloid precursors (acute myeloid leukemia, **AML**). The initial data set consisted of 38 bone marrow samples (27 **ALL**, 11 **AML**) obtained from acute leukemia patients. RNA prepared from the bone marrow mononuclear cells was hybridized to Affymetrix high-density oligonucleotide microarrays containing 6817 human genes. The key results for this study are shown in Figure 1.

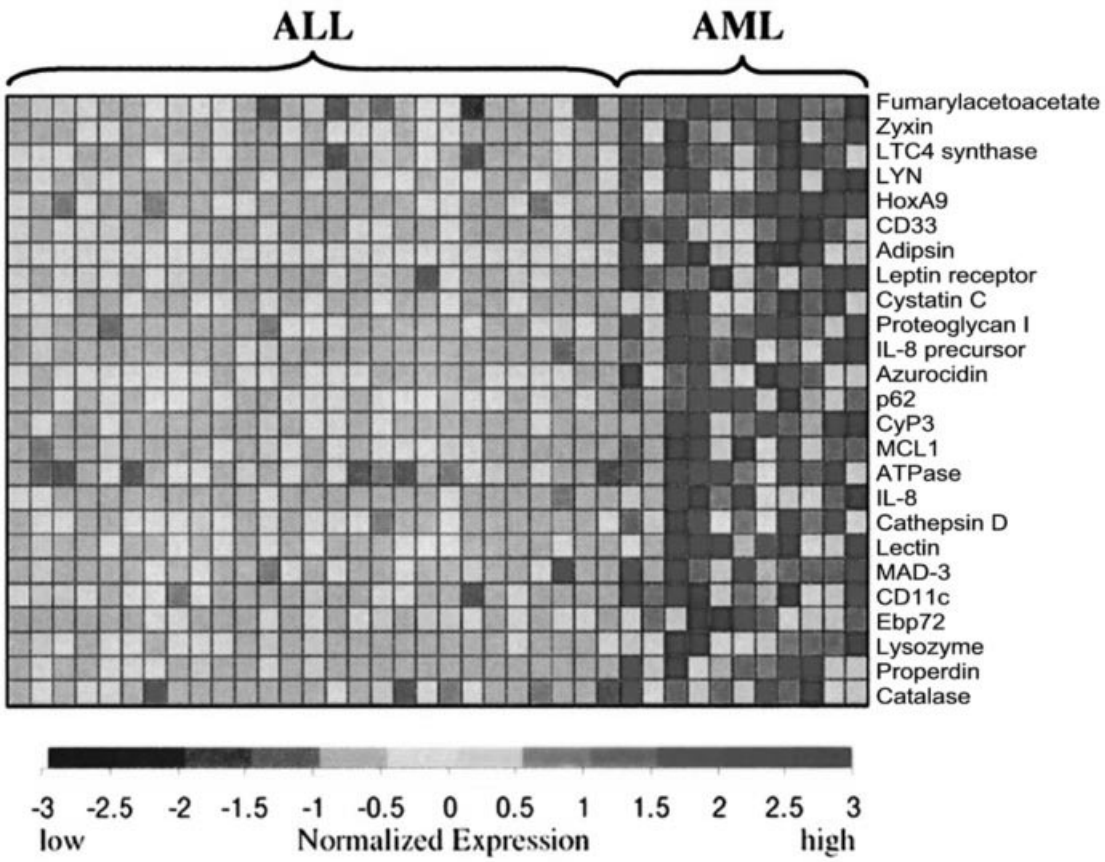


Figure 1: Genes distinguishing **ALL** from **AML**. The 25 genes most highly upregulated in **AML** vs. **ALL** are shown in descending order. Each row corresponds to a gene, with the columns corresponding to expression levels in samples from different patients. Expression levels greater than the mean are shaded in red, and those below the mean are shaded in blue.

Naming the Genes

The first step in the analysis is to name each gene in the data set using the names that are used in MEDLINE. In this example, the original names are those that appear in the FASTA formatted database for the microarray. Since these names tend to be brief, and not necessarily the currently recognized names for the genes, some work needs to be done to verify and/or correct the names. To assign the best possible name to each gene we used a combination of publicly available databases. These included GENBANK, OMIM and GeneCards. In addition, a preliminary run using PDQ_MED provides links to Entrez/PUBMED that simplifies this process.

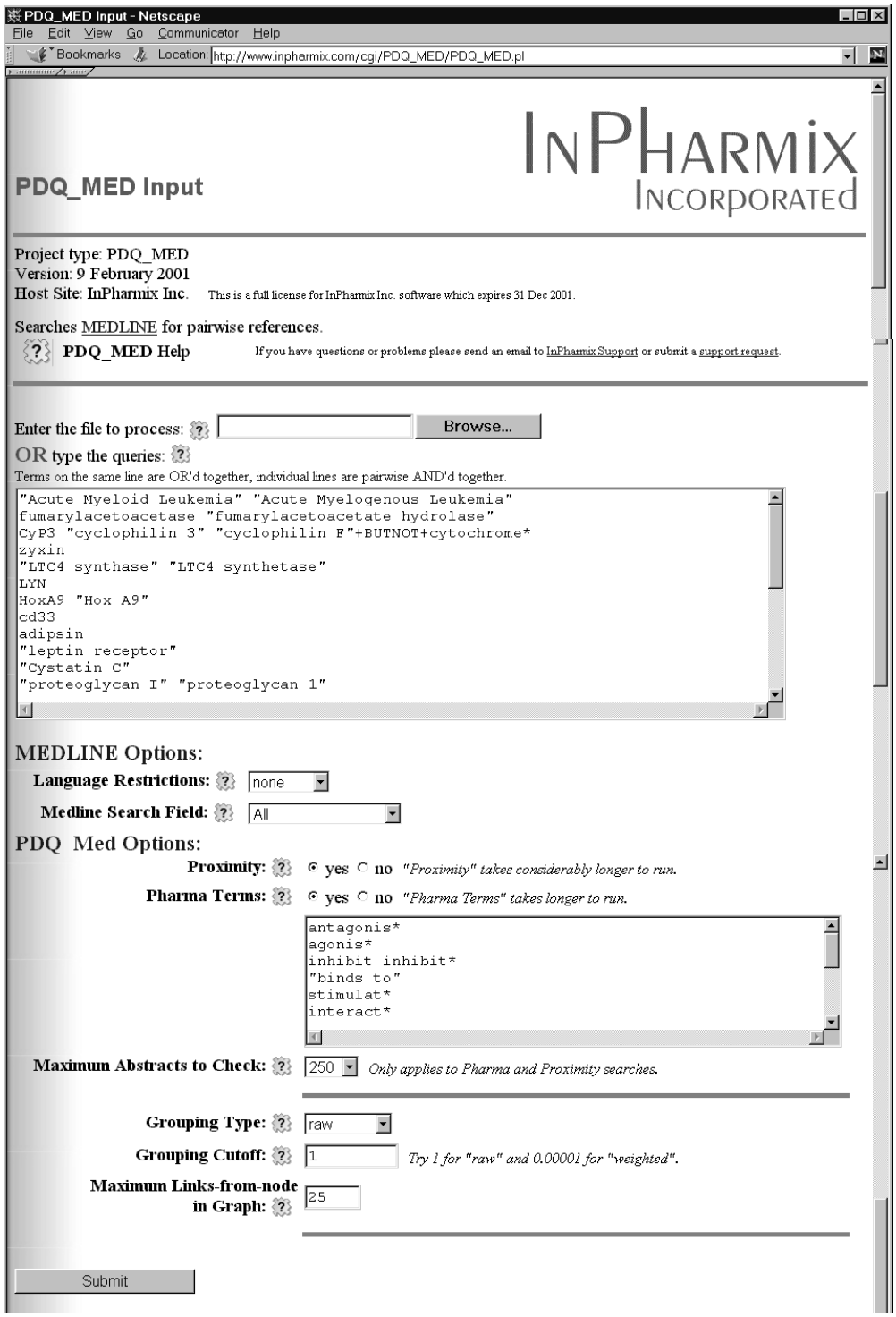
It is generally best to avoid acronyms since more than one gene or concept may use the same acronym. For example, in this data set one of the entries is labeled as CyP3. CYP is used as an acronym for both cytochrome P450s and cyclophilins. In this case, the acronym refers to a cyclophilin. In order to use the term CyP3, but avoid returning references to cytochromes, we will include "`BUTNOT cytochrom*`" in the query string.

Table 1: Listing of the genes highly expressed in AML vs. ALL.

	Original Name	UID	Expanded Name
Highly Expressed in AML			**"Acute Myeloid Leukemia" "Acute Myelogenous Leukemia"**
1	Fumarylacetoacetate	M55150	fumarylacetoacetase "fumarylacetoacetate hydrolase"
2	Zyxin	X95735	zyxin
3	LTC4 synthase	U50136	"LTC4 synthase" "LTC4 synthetase"
4	LYN	M16038	LYN
5	HoxA9	U82759	HoxA9 "Hox A9"
6	CD33	M23197	cd33
7	Adipsin	M84526	adipsin
8	Leptin receptor	Y12670	"leptin receptor"
9	Cystatin C	M27891	"Cystatin C"
10	Proteoglycan I	X17042	"proteoglycan I" "proteoglycan 1"
11	IL-8 precursor	Y00787	IL-8 Interleukin-8
12	Azurocidin	M96326	Azurocidin CAP37
13	p62	U46751	p62
14	CyP3 (1)	M80254	CyP3 "cyclophilin 3" "cyclophilin F"+BUTNOT+cytochrome*
15	MCL1	L08246	MCL1
16	ATPase	M62762	V-ATPase "vacuolar proton-ATPase"
17	IL-8 (2)	M28130	
18	Cathepsin D	M63138	"Cathepsin D"
19	Lectin	M57710	galectin-3 "galactose-specific lectin 3" MAC-2 LGALS2
20	MAD-3	M69043	MAD3 NFKBI
21	CD11c	M81695	CD11c
22	Ebp72 (3)	X85116	Epb72 Stomatin
23	Lysozyme	M19045	Lysozyme
24	Properdin	M83652	Properdin
25	Catalase	X04085	Catalase

PDQ_MED Web Interface

PDQ_MED is a web based application that will run on any web server capable of running Perl cgi's. Tested platforms include Windows, Linux and Unix servers. The input page for PDQ_MED is shown below.



Preliminary Search Results

We will use PDQ_MED to analyze the 24 genes for the AML dataset that are listed in Table 1. In addition to the 24 gene names we will include "acute myeloid leukemia" and "acute myelogenous leukemia" as additional terms. Because of the size of this dataset we will use PDQ_MED's proximity matching mode to identify pairs of terms which co-occur in the same sentence.

As a preliminary pass of the data, PDQ_MED was instructed to look for titles of papers which contain two or more of the query terms. This simplified proximity searching, the only type directly supported by MEDLINE, identifies the strongest linkages between terms in the query set. PDQ_MED identified 32 papers whose titles contain two of the query terms. These 32 titles contain 11 distinct term pairs.

PDQ_MED provides several views of the relationships between the members of a group. One such view is a "minimal spanning tree" (figure 2). Similar to a phylogenic tree, a minimal spanning tree shows the minimal set of linkages which allow the traversal of all members of the group. The minimal spanning tree presents a concise view of the interrelationships. However, this type of representation cannot show all of the interrelationships. In particular, since the tree is built with the minimal set of strongest links, most weak links are not shown. For example, AML is linked to eight of these terms but only the three strongest are shown.

An alternative to the minimal spanning tree is to use a geometric algorithm to attempt to solve for the complete set of linkages within the data (figure 3).

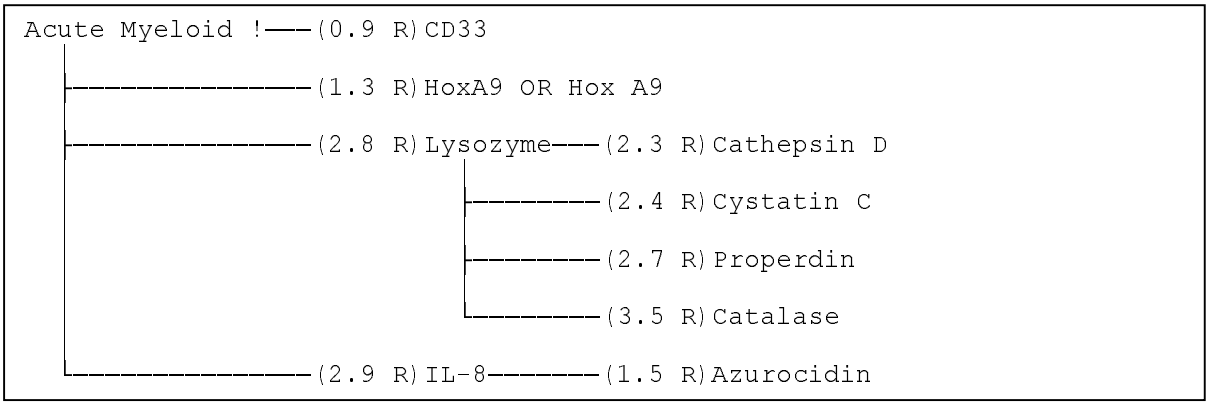


Figure 2: Minimal spanning tree for the 17 term group from the AML data set. Terms and term-pairs are hyperlinked to MEDLINE. Numbers in parenthesis are the $-\text{Log}_{10}$ of the normalized co-occurrence frequency for that pair of terms. This value is smaller for strong linkages and larger for weaker ones. The co-occurrence is hyperlinked to search MEDLINE for this pair of terms. The R is hyperlinked to MEDLINE for the pair of terms but further restricts the search to review articles. Terms are truncated to 14 characters to compress the display horizontally.

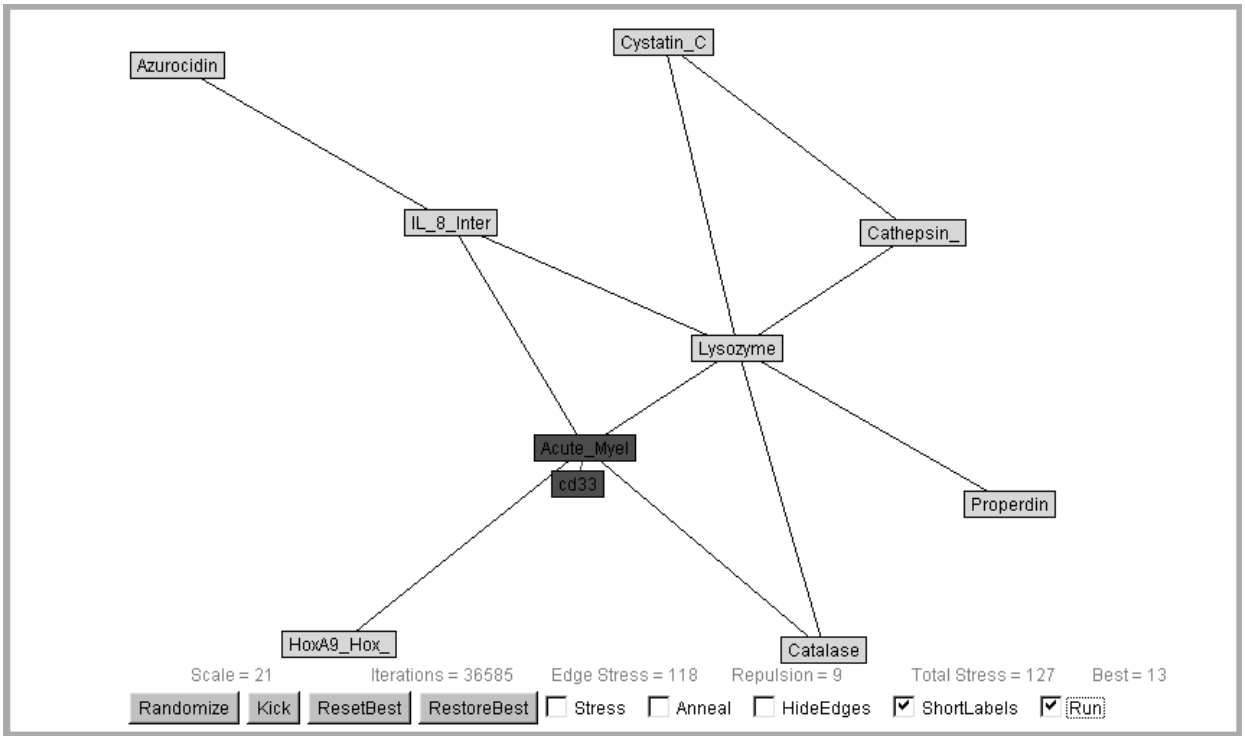


Figure 3: A distance geometry representation of the interrelationships found for the title search of terms from the AML dataset. Each box represents one query phrase. Connected boxes represent terms which co-occur in at least one abstract. The length of the interconnection is inversely proportional to the normalized frequency of co-occurrence, short lines represent frequent co-occurrence, longer lines less frequent co-occurrence. Black connecting lines represent relationships that the distance geometry algorithm considers correctly represented, red lines represent unresolvable interconnections.

Title Search Proximity Sentences

The complete set of article titles which PDQ_MED identified for the AML query set is shown below. Query terms are highlighted in red.

Acute Myeloid Leukemia* AND *Catalase

1. 6090009 Deficiency of erythrocyte superoxide dismutase and **catalase** activities in patients with malignant lymphoma and **acute myeloid leukemia**.

Acute Myeloid Leukemia* AND *CD33

1. 8441034 Modeling and dosimetry of monoclonal antibody M195 (anti-**CD33**) in **acute myelogenous leukemia**.
2. 1440849 The use of radiolabeled anti -**CD33** antibody to augment marrow irradiation prior to marrow transplantation for **acute myelogenous leukemia**.
3. 10537338 A phase I trial of humanized monoclonal antibody HuM195 (anti -**CD33**) with low -dose interleukin 2 in **acute myelogenous leukemia**.
4. 8175042 Prognostic significance of CD34 expression and **CD33**/CD13 ratio in **acute myeloid leukemia**.
5. 7908235 Acquisition of CD13 and **CD33** expression at relapse on **acute myeloid leukemia** cells with an unusual phenotype: MPO+CD13-**CD33**-.
6. 10884796 Characterization of CD13 and **CD33** surface antigen-negative **acute myeloid leukemia**.
7. 7517211 HLA-DR-, **CD33**+, CD56+, CD16- myeloid/natural killer cell acute leukemia: a previously unreco gnized form of acute leukemia potentially misdiagnosed as French-American-British **acute myeloid leukemia**-M3.
8. 10391105 We describe the morphological, cytochemical, immunologic, and c ytogenetic features of two patients with **AML** with maturation (FAB M2) and the phenotype MPO+, CD13 (-), **CD33**(-), CD56(+).

Acute Myeloid Leukemia* AND *HoxA9 OR Hox A9

1. 10221343 Low frequency of rearrangements of the homeobox gene **HOXA9**/t(7;11) in adult **acute myeloid leukemia**.

Acute Myeloid Leukemia* AND *IL-8 OR Interleukin-8

1. 8412317 **IL-8** mRNA expression and **IL-8** production by **acute myeloid leukemia** cells.
2. 7578521 Plasma levels of IL-1, TNF alpha, IL-6, **IL-8**, G-CSF, and IL1-RA during febrile neutropenia: results of a prospective study in patients undergoing chemotherapy for **acute myelogenous leukemia**.

Acute Myeloid Leukemia* AND *Lysozyme

1. 288968 The prognostic value of serum **lysozyme** activity in **acute myelogenous leukemia**.
2. 1546687 The significance of an elevated serum **lysozyme** value in **acute myelogenous leukemia** with eosinophilia.

Cathepsin D* AND *Lysozyme

1. 8093011 Distinctive inhibition of the lysosomal targeting of **lysozyme** and **cathepsin D** by drugs affecting pH gradients and protein kinase C.
2. 2590170 Calcitriol enhances transcriptional activity of **lysozyme** and **cathepsin D** genes in U937 promonocytes.
3. 10708885 Delta(9)-tetrahydrocannabinol selectively increases aspartyl **cathepsin D** proteolytic activity and impairs **lysozyme** processing by macrophages.

Cystatin C* AND *Cathepsin D

1. 2013314 Inactivation of human **cystatin C** and kininogen by human **cathepsin D**.

Interleukin-8* AND *Azurocidin OR CAP37

1. 8621683 Identification of defensin-1, defensin-2, and **CAP37**/azurocidin as T-cell chemoattractant proteins released from **interleukin-8**-stimulated neutrophils.

Interleukin-8* AND *Lysozyme

1. 10728932 Effects of TNF-alpha and IL-1 beta on mucin, **lysozyme**, IL-6 and **IL-8** in passage-2 normal human nasal epithelial cells.

Lysozyme* AND *Properdin

1. 948833 **Lysozyme**, complement and **properdin** dynamics in calves.

Full PDQ_MED Search with Proximity

For a complete search of the literature with proximity checking, PDQ_MED executed 325 queries to MEDLINE and identified 58,760 abstracts which refer to at least one of the 25 terms. Total run time for this analysis was less than 5 minutes. For AML, there are 6,884 abstracts and both catalase and lysozyme each occur in more than 16,000 abstracts. The complete literature co-occurrence matrix is shown in Table 2. Lysozyme, AML, IL-8, cathepsin D and catalase are the most extensively cross-referenced terms in this set with 10, 7, 7, 6, and 6 co-occurring terms respectively (table 2).

For the 25 terms in the AML data set, PDQ_MED identified a group of 17 terms which can be linked together via the literature. For these 17 terms there were a total of 245 sentences representing 31 term pairs. No co-occurrences were found for the remaining 8 terms.

Table 2: Literature co-occurrence matrix for the AML dataset.

	CyP3 OR cyclophilin 3	zyxin	V-ATPase	Azurocidin OR CAP37	adipsin	fumarylacetoacetase	LTC4 synthase	Lysozyme	CD11c	Hox A9	MCL1	proteoglycan I	CD33	Cathepsin D	Properdin	galectin-3	LYN	IL-8 OR Interleukin-8	p62	**AML**	leptin receptor	Cystatin C	Catalase	Epb72 OR Stomatin	MAD3 OR NFKBI
zyxin		65																							
V-ATPase			551																						
Azurocidin				74																					
adipsin					129																				
fumarylacetoacet						145																			
LTC4 synthase							92																		
Lysozyme								16477																	
CD11c								8	984																
Hox A9										48															
MCL1											126														
proteoglycan I												34													
CD33								7	39				1047												
Cathepsin D			1					21						3179											
Properdin								16							1418										
galectin-3			1					4	1					3		532									
LYN																	718								
IL-8				2				8	1					1	1			7245							
p62								1									4		513						
AML								13	5	9			53					3		6956					
leptin receptor																				1	574				
Cystatin C								2						2								429			
Catalase							1	33						1			1	6		4			17802		
Epb72, Stomatin																								64	
MAD3, NFKBI																									41

directly linked to AML

indirectly linked to AML

Figure 4 shows the relationships within this group as a "minimal spanning tree". The strongest relationships shown in Figure 4 are for AML and HoxA9, CD33 and leptin receptor. Figure 5 shows the "distance geometry" treatment of the 17 co-occurring terms in the AML dataset.

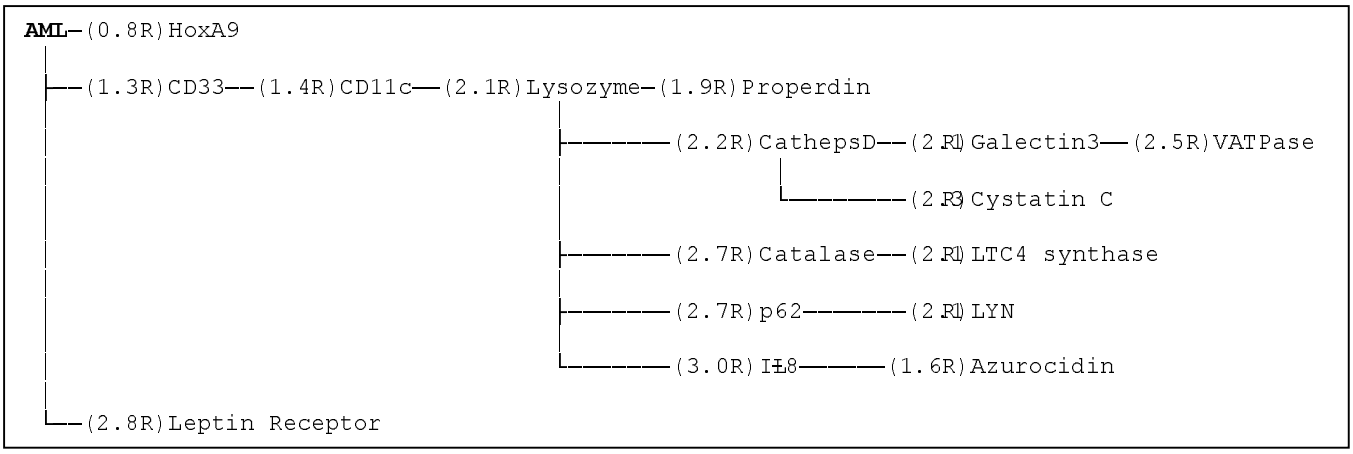


Figure 4: Minimal spanning tree for the 17 terms from the full AML data set. Terms are truncated to 14 characters to compress the display horizontally.

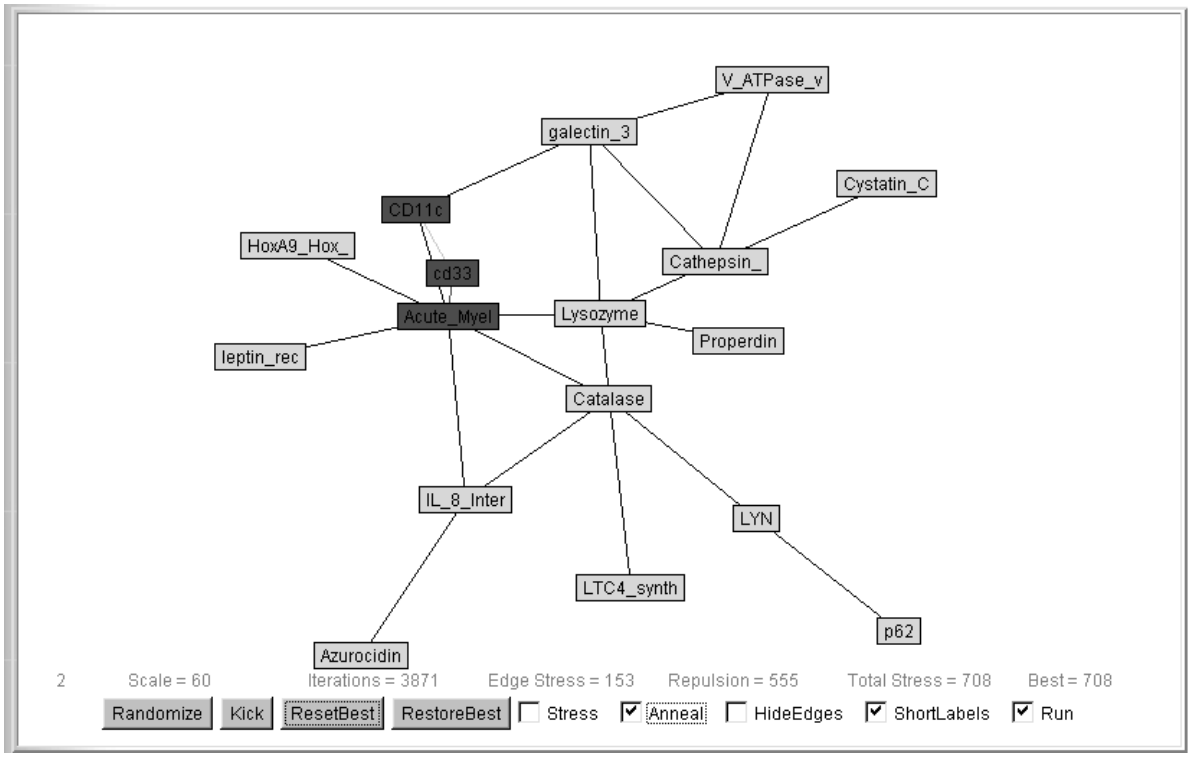


Figure 5: A distance geometry representation of the interrelationships found for the terms from the full AML dataset with proximity checking. In this view of the data each node was allowed to create no more than three connections.

Typical AML Proximity Sentences

A scree n shot of re pre se ntative proximity se nte nce s for AML is shown be low. The que ry te rms are highlighted in re d and the sentences for each term pair are ordered by relevance.

Acute Myeloid Leukemia OR Acu! **AND** *HoxA9 OR Hox A9*

11113197 (20.6) We have previously shown that **HOXA9** collaborates with MEIS1 in the induction of **acute myeloid leukemia** (AML).

11157742 (19.7) NUP98-**HOXA9** expression in hemopoietic stem cells induces chronic and **acute myeloid leukemias** in mice.

10602420 (18.7) We demonstrate that the expression of **HOXA9** and MEIS1 in leukemia cells is uniquely myeloid, and that these genes are commonly co-expressed in myeloid cell lines and in samples of **acute myelogenous leukemia** (AML) of all subtypes except in promyelocytic leukemia.

9695407 (16.6) In **acute myeloid leukemia** (AML) with t(7;11) translocation, the **HOX A9** gene is rearranged.

10757811 (16.3) The genes encoding **Hoxa9** and Meis1 are transcriptionally coactivated in a subset of **acute myeloid leukemia** (AML) in mice.

10221343 (15.9) Low frequency of rearrangements of the homeobox gene **HOXA9**/t(7;11) in adult **acute myeloid leukemia**.

9649441 (14.2) Primary bone marrow cells, retrovirally engineered to overexpress **Hoxa9** and Meis1a simultaneously, induced growth factor-dependent oligoclonal **acute myeloid leukemia** in <3 months when transplanted into syngenic mice.

10397741 (12.3) The nucleoporin gene NUP98 was found fused to the **HOXA9**, HOXD13, or DDX10 genes in human **acute myelogenous leukemia** (AML) with chromosome translocations t(7;11)(p15;p15), t(2;11)(q35;p15), or inv(11)(p15;q22), respectively.

10936866 (10.9) NUP98-**HOXA9** chimera mRNA, which is known to be involved in t(7;11)(p15;p15) translocation in **acute myeloid leukemia** (AML), was not detected by reverse transcriptase-polymerase chain reaction, and NUP98 rearrangement was not detected by Southern blot analysis of the blasts in the MDS phase.

Acute Myeloid Leukemia OR Acu! **AND** *IL-8 OR Interleukin-8*

8412317 (24.0) **IL-8** mRNA expression and **IL-8** production by **acute myeloid leukemia** cells.

8767523 (19.0) We investigated the profile of interleukin-8 (**IL-8**) expression and release by leukemic cells obtained at diagnosis from 42 untreated adult patients with **acute myeloid leukemia** of various FAB subtypes (2 M0, 7 M1, 6 M2, 6 M3, 10 M4 and 11 M5).

7578521 (16.5) Plasma levels of IL-1, TNF alpha, IL-6, **IL-8**, G-CSF, and IL1-RA during febrile neutropenia: results of a prospective study in patients undergoing chemotherapy for **acute myelogenous leukemia**.

Acute Myeloid Leukemia OR Acu! **AND** *leptin receptor*

10029596 (63.3) Results confirm the reported expression of **leptin receptor** in normal CD34(+) cells and demonstrate the frequent expression of **leptin receptors** in **AML** blasts.

Acute Myeloid Leukemia OR Acu! **AND** *Lysozyme*

4628693 (109.2) [Estimation of **lysozyme** in **acute myelogenous leukemia**].

288968 (69.7) The prognostic value of serum **lysozyme** activity in **acute myelogenous leukemia**.

Summary of All Linkages

A brief summary of the literature for each term pair for the AML dataset is shown below.

AML and Catalase (4 sentences)
Catalase activity is increased in AML cells.
AML and CD11c (5 sentences)
CD11c is a characteristic marker of AML cells.
AML and CD33 (52 sentences)
CD33 is a characteristic marker of AML cells and the target of anti-CD33 MoAb based therapies.
AML and HoxA9 (9 sentences)
HoxA9 (a homeobox gene) is upregulated in AML, is implicated in the induction of AML and involved in a chromosomal translocation event unique to AML.
AML and IL-8 (3 sentences)
IL-8 expression and secretion is up-regulated in isolated AML blasts and serum levels are higher in AML patients.
AML and Leptin Receptor (1 sentence)
Leptin receptor levels are higher in AML cells and leptin stimulates proliferation of cultured AML cells.
AML and Lysozyme (13 sentences)
Elevated serum and urinary lysozyme levels are clinical markers for classifying AML subtypes.
Cathepsin D and Catalase (1 sentence)
Co-occur eosinophil granules.
Cathepsin D and Galectin-3 (3 sentences)
Both are upregulated in epithelial injury.
Cathepsin D and Lysozyme (21 sentences)
Both are regulated by vitamin D in promonocytes.
CD11c and Lysozyme (8 sentences)
Both are markers of AML.
CD33 and CD11c (39 sentences)
Both are monocyte/histiocyte markers and are used to sub-classify AML types.
CD33 and Lysozyme (7 sentences)
Both markers of AML.
Cystatin C and Cathepsin D (2 sentences)
Cystatin C is an inhibitor of Cathepsin D.
IL-8 and Azurocidin (2 sentences)
Azurocidin, a T-cell chemoattractant, is released from interleukin-8-stimulated neutrophils.
IL-8 and Catalase (6 sentences)
The response to active oxygen species by inflammatory mediators such as IL-8 are reduced by catalase.
IL-8 and Lysozyme (8 sentences)
Both are immunological and inflammatory mediators and are frequently found together in biological specimens.
IL-8 and Properdin (1 reference)
Properdin secretion from neutrophils is stimulated by IL-8.
LTC-4 Synthase and Catalase (1 sentence)
Catalase inhibits LTC4 metabolism.
Lyn and Catalase (1 sentence)
Catalase inhibits the H2O2 induced activation of Lyn.
Lyn and p62 (4 references)
Both are src-like protein tyrosine kinases and p62 is a substrate for Lyn.
Lysozyme and Catalase (33 sentences)
Components of the cytoplasmic germicidal and lytic systems.
Lysozyme and Properdin (16 sentences)
Both are non-specific defense factors.
V-ATPase and Cathepsin D (1 sentence)
Both occur in acidic vacuoles of macrophages.
V-ATPase and Galectin-3 (1 sentence)
Osteoclastic markers (osteoclasts can be derived from blood monocytes).

Pharma Terms

Optionally, PDQ_MED will also search the query set against a list of "pharma terms". The "pharma term" list includes terms such as agonist, antagonist, regulates, inhibits etc. These search results can be used to generate a list of agonists, antagonist etc. for the genes in the query list as shown in Table 3.

Table 3: "Pharma term" linkages for the AML dataset.

Term	Agonist / Antagonist / Binder / Up & Down Regulators
fumarylaceto-acetase	4-(hydroxymethylphosphinoyl)-3-oxo-butanoic acid (HMPOBA) , X-ray crystal structure
zyxin	alpha-actin, vasodilator-stimulated phosphoprotein (VASP), Zyx16-30 (APAFYAPQKKFGPVV), h-warts/LATS1, CRP, NOC2
LTC4 synthase	magnolol, 5-LO inhibitors, thiopyrano[2,3,4-c,d]indoles, DEX, MK-886, CSA
LYN	SHPTP1 (binds), SyK (binds), betac (binds), homopoietic-specific poten HS1 (binds), GAL1 (regulates), PP1, AKT (regulates), Fc receptor gamma (modifies), Erk1 (binds)
HoxA9	MEIS1 (binds), PBX1 (binds), TALE (binds),
cd33	SHP-1(binds), SHP-2 (binds), immunoconjugates(bind),Fc receptor gamma(binds)
adipsin	Retinoic acid (regulated by), insulin (regulated by), ephedrine and caffeine (stimulates), RU38486 (regulates), glucocorticoids (regulates), antibodies
leptin receptor	c-fos (regulates), leptin (binds), SHP-2/ERK (acivates), SOCS3 (feedback), DAG - kinase z (interacts)
Cystatin C	cathepsin B (inhibitor of), methylprednisolone (up regulates), CSA (down regulates)
proteoglycan I	MSA (stimulates), CDF (stimulates)
Interleukin-8	antibodies, AP-1 (regulates), NFKB (regulates), TNF (upregulates), CSA (inhibits), DEX, IL-1ra (inhibits)
p62	Insulin rec.(substrate), TRAF6(binds), NUP93(binds), SP1 (binds), CD28(binds)
MCL1	PCNA (binds)
V-ATPase	bafilomycin (inhibits), VMA12p (binds), VMA22p (binds), concanamycin (inhibits), suramin (inhibits), NBD-C1(inhibits), (2Z,4E)-5-(5,6-dichloro-2-indolyl)-2-methoxy-N-[4-(2,2,6,6-tetramethyl)piperidinyl]-2,4-pentadienamide (inhibits), actin (binds)
Cathepsin D	Estrogen (up regulates), statine (inhibits), pepstatin A (inhibits), ceramide (inhibits), IFN-gamma (regulates), ZPAD (inhibits), CEL5-A (inhibits), CEL5-G (inhibits), EA-1 (inhibits)
galectin-3, MAC-2	SP-1 (regulates), leptomycin B (inhibits export), antibodies, betagalactosides (binds), p60 (binds), p90 (binds)
MAD3,NFKBI	cDc20 (binds)
CD11c	fibrin (bind), fibrinogen (binds), IAV (regulates), leptin (regulates), antibodies, collagen-I (binds), PyRo1(regulates), c-Myc (regulates), CD2 3(binds), CD18(binds)

Unlinked Terms

Seven out of 25 of the AML query terms were not linked to any other term by PDQ_MED using proximity searching;

Term	Abstract Count
Proteoglycan I †	34
Adipsin †	128
CyP3 OR "cyclophilin 3" †	29
Epb72 OR Stomatin	61
Fumarylacetoacetase	142
MAD3 OR NFKBI	41
MCL1 †	126
Zyxin	65

† This term has cross-references that were omitted because of proximity settings.

Of these seven terms, four (proteoglycan I, adipsin, Cyclophilin 3 and MCL1) co-occur in abstracts with a term from the 17 member group but the co-occurrence failed to pass the proximity test. Nonetheless, the four terms can be tentatively linked to the larger 17 member set. These seven genes may represent unique opportunities for study since their involvement in AML is without literature precedence. In addition, the "pharma term" list (see table 3) suggests potential starting points for addition research on these genes.

AML Summary

We have used a new text mining tool to examine the results of a typical genomics experiment. For the AML data set, 17 of the 24 gene names can be linked via the scientific literature and seven of the terms can be directly linked to the disease. Examination of the literature identified by PDQ_MED indicates that many of these genes are related to the non-specific host response cascade (catalase, IL-8, lysozyme, cathepsin D, cystatin C, azurocidin, properdin, LTC-4 synthase, galectin-3) characteristic of myeloblasts. In addition, a list of known agonists and antagonists for these genes was extracted from MEDLINE. The ana lysis provides a detailed explanation for why many of these genes appear in this dataset based on the scientific literature. The PDQ_MED analysis provides the researcher with a framework to help understand what is known about the interactions within a set of genes and highlights areas for future research.

Conclusion

We have demonstrated PDQ_MED, a novel tool for the search and analysis of the scientific literature. PDQ_MED allows researchers to effectively mine the more than 5 million abstracts in MEDLINE for information that will allow them to fully exploit the results of their genomics experiments. PDQ_MED quickly provides a framework, based on the biomedical literature, which helps to organize and explain why certain sets of genes are co-regulated. PDQ_MED also identifies pairs of genes or gene-disease relationships for which there is no literature precedence. In addition, the user may search for "pharma terms" such as "agonist", "antagonist" or "drug" and use these terms to highlight key concepts relevant to their research. Overall, PDQ_MED ensures that the researcher can effectively gather and analyze the relevant literature for large sets of genes, proteins and disease terms hence providing a key capability for a successful genomics research project.