SANDIA REPORT

SAND2004-0161
Unlimited Release
Printed January 2004

Application of Multidisciplinary Analysis to Gene Expression

Mónica Mosquera-Caro, Shawn Martin, Erik Andries, Jeffrey Potter, Kerem Ar, Yuexian Xu, Huining Kang, Xuefei Wang, Maurice H. Murphy, Paul Helman, Robert Veroff, David M. Haaland, Susan Atlas, Jim Cowie, Chris Fields, Valeriy Sibirtsev, George Davidson and Cheryl Willman.

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Issued by Sandia National Laboratories, operated for the United States

Department of Energy by Sandia Corporation.

NOTICE: **This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.**

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2004-0161
Unlimited Release
Printed January 2004

Application of Multidisciplinary Analysis to Gene Expression

George S. Davidson and Shawn Martin
Computational Biology

David M. Haaland
Chemical and Biological Sensing, Imaging & Analysis

Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-0310

Mónica Mosquera-Caro, Jeffrey Potter, Karen Ar, Yuexian Xu, and Cheryl Willman
Cancer Research and Treatment Center
Department of Pathology
University of New Mexico
Albuquerque, New Mexico

Erik Andries, Huining Kang, Xuefei Wang, Maurice H. Murphy, Paul Helman,
Robert Veroff, and Susan Atlas
Computer Science, Physics and Astronomy, Mathematics and Statics
University of New Mexico
Albuquerque, New Mexico

Jim Cowie, Chris Fields, and Valeriy Sibirtsev
The Computing Research Laboratory
New Mexico State University
Las Cruces, New Mexico

Abstract

Molecular analysis of cancer, at the genomic level, could lead to individualized patient diagnostics and treatments. The development to follow will signal a significant paradigm shift in the clinical management of human cancers. Despite our initial hopes, however, it seems that simple analysis of microarray data cannot elucidate clinically significant gene functions and mechanisms. Extracting biological information from microarray data requires a complicated path involving multidisciplinary teams of biomedical researchers, computer scientists, mathematicians, statisticians, and computational linguists. The integration of the diverse outputs of each team is the limiting factor in the process to discover candidate genes and pathways associated with the molecular biology of cancer. Specifically, one must deal with sets of significant genes identified by each method and extract whatever useful information may be found by comparing these different gene lists. Here we present our experience with such comparisons, and share methods developed in analysis of an infant leukemia cohort studied on Affymetrix HG-U95A arrays. In particular, spatial gene clustering, hyper-dimensional projections, and computational linguistics were used to compare different gene lists. In spatial gene clustering, different gene lists are grouped together and visualized on a three-dimensional

expression map, where genes with similar expressions are co-located. In another approach, projections from gene expression space onto planes clarify how groups of genes can jointly have more predictive power than groups of individually selected genes. Finally, online literature is automatically rearranged to present information about genes common to multiple groups, or to contrast the differences between the lists. The combination of these methods has improved our understanding of infant leukemia. While the complicated reality of the biology dashed our initial, optimistic hopes for simple answers from microarrays, we have made progress by combining very different analytic approaches.

Analysis techniques for molecular classification in infant leukemia

Advances in the treatment and prognosis of childhood leukemia are considered remarkable successes in modern medicine (Greaves, 2002). However, even using the current risk classification systems (combining age, white blood cell count at presentation (WBC), morphology, cytogenetics, and other biologic parameters), infants with leukemia who will ultimately achieve complete clinical remission cannot be precisely identified (Biondi, 2000). Notably, those patients who will be primarily resistant, or more prone to relapse, are simply not completely predicted by cytogenetic parameters; they are distributed among all clinically defined risk groups. Refined recognition of patients who will respond to the less intensive therapies would be very desirable, particularly to increase survival and decrease therapy-related toxicity (Felix, 1999; Biondi, 2000). We are addressing the need for such discrimination diagnostics by developing gene expression-based classifications using Affymetrix U95Av2 oligonucleotide microarrays (with 12,625 probes). Here we discuss the methods used to analyze our infant cohort, which is a statistically designed group of 126 infant patients with acute leukemia. Of the 126 cases, 78 were Acute Lymphoid Leukemia (ALL, 62%), 48 were Acute Myeloid Leukemia (AML, 38%). In addition, 53 cases (42%) had translocations involving the MLL gene (chromosome segment 11q23), see Figure 1.

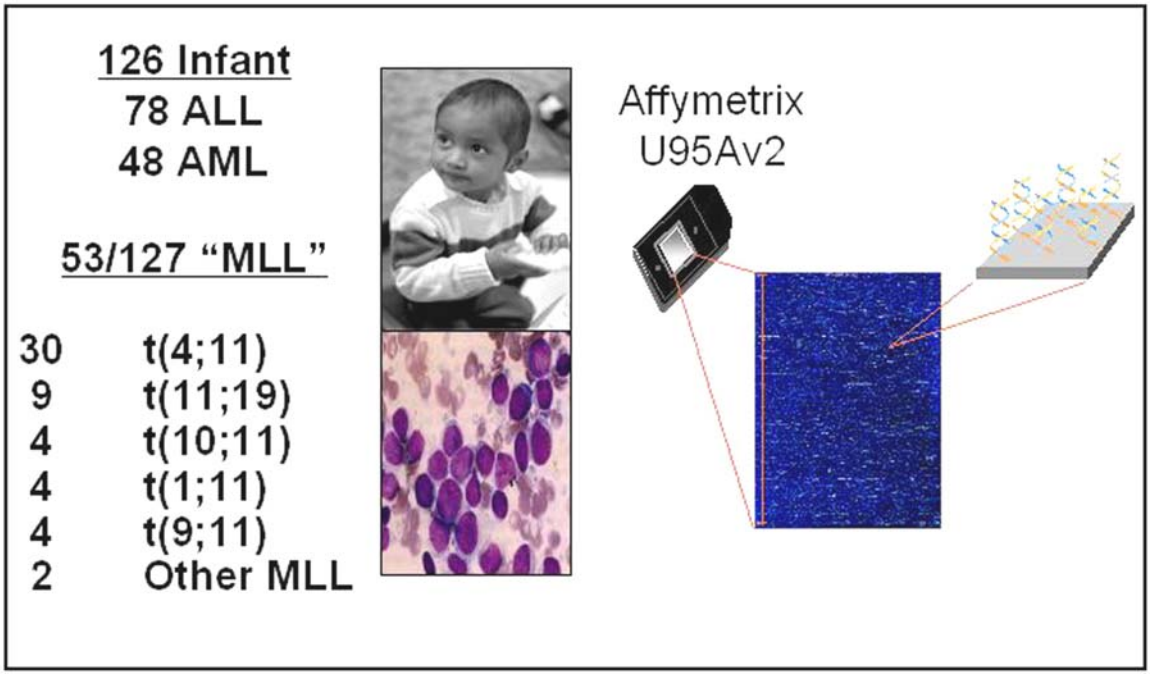

Figure 1. Design of the leukemia cohort. The statistically designed cohort contained 126 acute leukemia samples from infant patients. Of the 126 cases, 78 were Acute Lymphoid Leukemia (ALL, 62%), 48 were Acute Myeloid Leukemia (AML, 38%), and 56 (44%) cases had translocations involving the MLL gene (chromosome segment 11q23). Cases were studied using Affymetrix U95Av2 oligonucleotide microarrays (12,625 probes).

Traditionally, the analysis of microarray data has used both unsupervised methods, which group together genes or patients based on intrinsic similarities in expression, and supervised approaches, which exploit knowledge available in a training set to predict unknown groups of genes or patients. We began our analysis with an unsupervised search for two traits: 1) gene expression profiles related to leukemia type (AML vs. ALL, as defined by traditional morphology standards), and 2) initial outcome (remission vs. failure) in infant patients. Principal Component Analysis (PCA) (see Joliffe, 1986) was used to determine whether an apparent partition could be seen between the expression profiles of cases in each one of the classes (specifically, ALL versus AML and remission versus failure cases).

As shown in Figure 2, PCA uncovers a clear separation between the lymphoid cases (in blue) and the myeloid cases (in red). In fact, the first three principal components capture the infant ALL/AML lineage distinction. However, PCA failed to find a clear partition between the remission (shown in pink, Figure 3) and the failure cases (shown in green, Figure 3). In general, and despite an array of different methods, we found that predicting resistance and treatment failure was a much more complex problem than the “type” classification (ALL vs. AML).

Figure 2. Principal Components Analysis (PCA) ALL/AML separation. The figure shows the projections of the first and second (left panel) and third and fourth (right panel) principal components of the infant microarray data (using all genes). Each sphere represents an infant sample in the “gene expression” dimension. A separation of the gene expression profiles of lymphoid cases (ALL, shown in blue) versus the myeloid cases (AML, shown in red) can be seen in the third and fourth PCA projections.

Figure 3. Principal Components Analysis (PCA) remission vs. failure separation. The figure shows two-dimensional projections of the first and second (left panel) and second and third (right panel) principal components of the infant microarray data (all genes used). Each sphere represents an infant sample in the "gene expression" dimension, and is color-coded to indicate the specific outcome of the case: remission (pink) or failure (green). The first three principal components captured the infant ALL/AML lineage distinction (Figure 2), but failed in demonstrating a partition between remission and failure cases.

The next step in our analysis involved using supervised learning methods to predict patient outcome. Supervised learning methods are trained to recognize "known classes", creating classification algorithms that may be able to predict new cases. These algorithms are also capable of uncovering interesting and novel therapeutic targets by way of gene selection. A supervised method needs a training set (known examples) and a test set (for evaluating the effectiveness of the classifier). For these methods, the 126 infant samples were divided into representative training (82 cases) and test sets (44 cases), statistically balanced according to the clinical labels (leukemia lineage, cytogenetics and outcome).

Several supervised class prediction approaches were used including Bayesian networks (Helman et al. 2002), Recursive Feature Elimination in the context of Support Vector Machines (SVM-RFE) (Guyon et al. 2002), Neuro-Fuzzy Logic, and Discriminant Analysis. These classification algorithms were evaluated using fold-dependent, leave-one-out, cross validation (LOOCV) techniques. As shown in Table 1, outcome prediction (remission versus failure) was particularly poor for all of the methods employed.

Parallel comparison of discriminating genes

The process of defining the best set of discriminating genes identified by the different methods is difficult. Different methods of gene selection typically produce different lists of genes (note that

this is not necessarily bad because related genes can be essentially equally predictive). We developed two main methods for visualizing many gene lists at once and consequently comparing not only the lists but also the methods that produced them.

It is reasonable to imagine that two different gene lists carry nearly the same information if the genes in the lists generally cluster near each other. We tested this approach with an unsupervised clustering algorithm (Kim et al., 2001) that uses Pearson's correlation coefficient to estimate the similarity between any pair of genes. The 20 strongest positive correlations between each gene and its neighbors were used to assign that gene to an x-y coordinate in the two-dimensional plane using force-directed placement (Davidson et al., 2001). In this x-y ordination step, genes are positioned relative to each other under the influence of attractive and repulsive forces. Each gene is attracted to other genes with a force proportional to their similarity in gene expression, and is repelled by a constant force proportional to the local density of genes. A computer program called VxInsight was used to visualize the spatial distribution of the genes, resulting in a visualization wherein genes with a high correlation are placed near each other. As a further visual cue, the two-dimensional scatter plot is converted into a three-dimensional terrain map in which the z-axis denotes the density of genes within a given area. The genes identified by the various supervised methods were highlighted, and colored as shown in Figures 4 and 5. Not surprisingly, the gene lists that are successful in differentiating between ALL and AML do cluster near each other, as seen in Figure 4. However, as the various methods have difficulty with the remission/failure prediction, it is reasonable to assume that there is no readily identifiable set of differentiating genes for this prediction, and indeed, the various methods have no consensus and the suggested genes are scattered widely across the overall gene cluster (See Figure 5).

Table 1. Class Predictor Performance

Description	Bayesian Net		SVM		Fuzzy Inference		Discriminant Analysis	
	r	p-value[1]	r	p-value[1]	r	p-value[1]	r	p-value[1]
ALL vs. AML	.912	<.001**	.971	<.001**	.971	<.001**	.853	<.001**
Remission. vs. Fail	.568	.256	.622	.094	.405	.906	.568	.256
Remission. vs. Fail in ALL	.542	.419	.625	.153	.375	.924	.500	.580
Remission. vs. Fail in AML	.461	.709	.769	.046*	.461	.709	.461	.709

r = Success rate.
p-value[1] = Computed estimating the probability of successful prediction.
* means that the predictor is significant at level D=0.05
** means that the predictor is significant at level D=0.01.

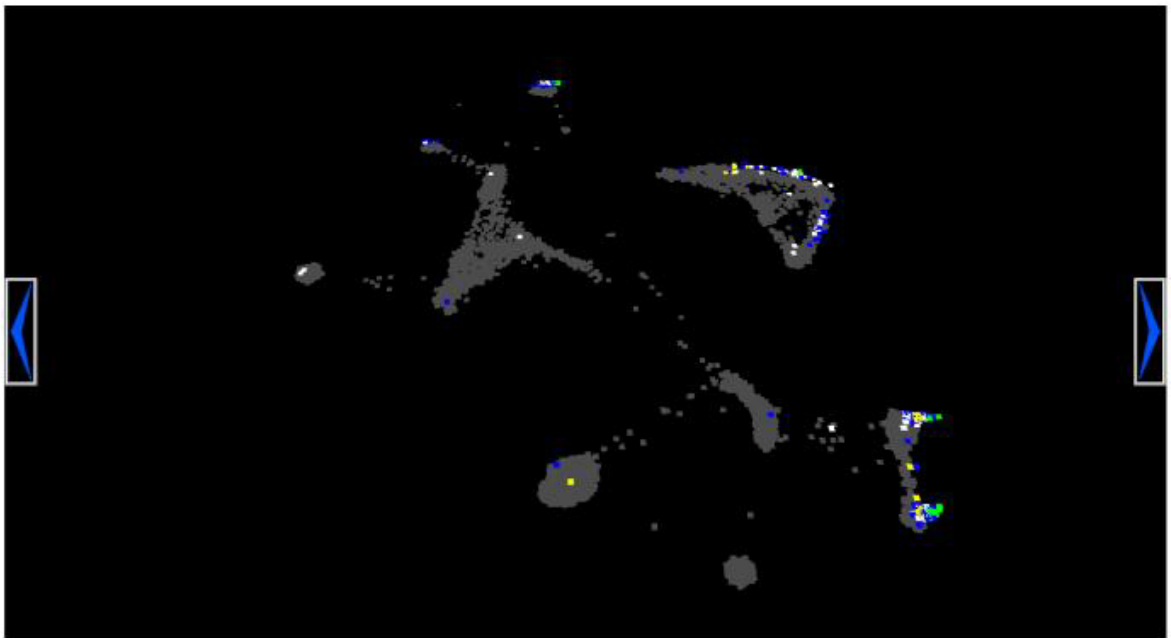

Figure 4. Co-localization of ALL vs. AML gene lists in a gene expression map. The genes that characterize ALL versus AML samples are shown, with a different color for each of the methods used to obtain them (green for Bayesian Networks, yellow for discriminant analysis, blue for Fuzzy logics and white for SVM). Very similar lists will be co localized, while lists with bigger variation will be further apart. A computer program called VxInsight was used to visualize the spatial distribution of the genes, resulting in a display in which genes with a high correlation are placed near to each other on a three-dimensional terrain map wherein the z-axis denotes the density of genes within an area.

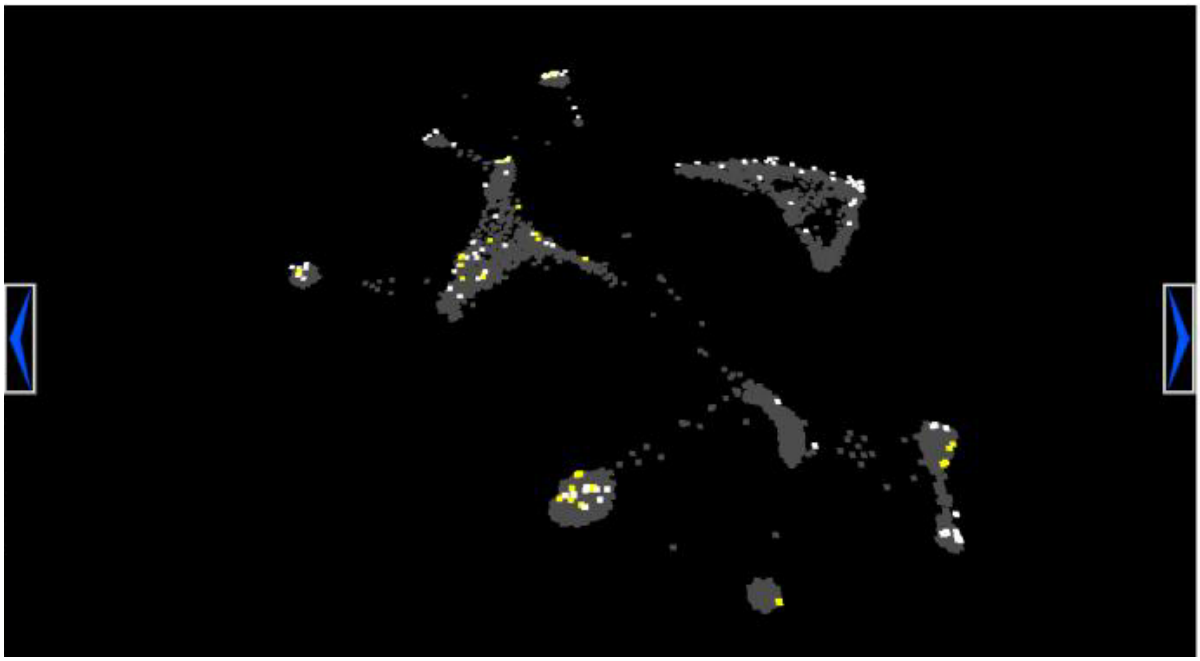

Figure 5. Visualization of outcome (remission vs. failure) gene lists in a gene expression map. The genes that characterize remission versus failure are shown, with a different color for each one of the methods used to obtain them (white for SVM and yellow for Bayesian Networks). Very similar lists will be co-localized, while lists with bigger variation will be further apart.

We developed a second method to compare and visualize many gene lists simultaneously. In this approach, each gene is considered to be a point in patient-space, where each dimension corresponds to a different patient. Since there were 12,625 genes and 126 patients, this spatial representation had 12,625 points (samples) in a 126 dimensional space. Of the 12,625 genes we only considered about 600 that occurred in the different gene lists, reducing our problem to 600 genes in 126 dimensions. Furthermore, because we were mainly interested in how the genes compared as discriminators, and not how their actual expression levels compared, we projected the genes onto the 126 dimensional unit sphere in patient-space, as shown in Figure 6. Geometrically, this corresponds to comparing "directions" of the genes in the various gene lists as opposed to their "magnitudes".

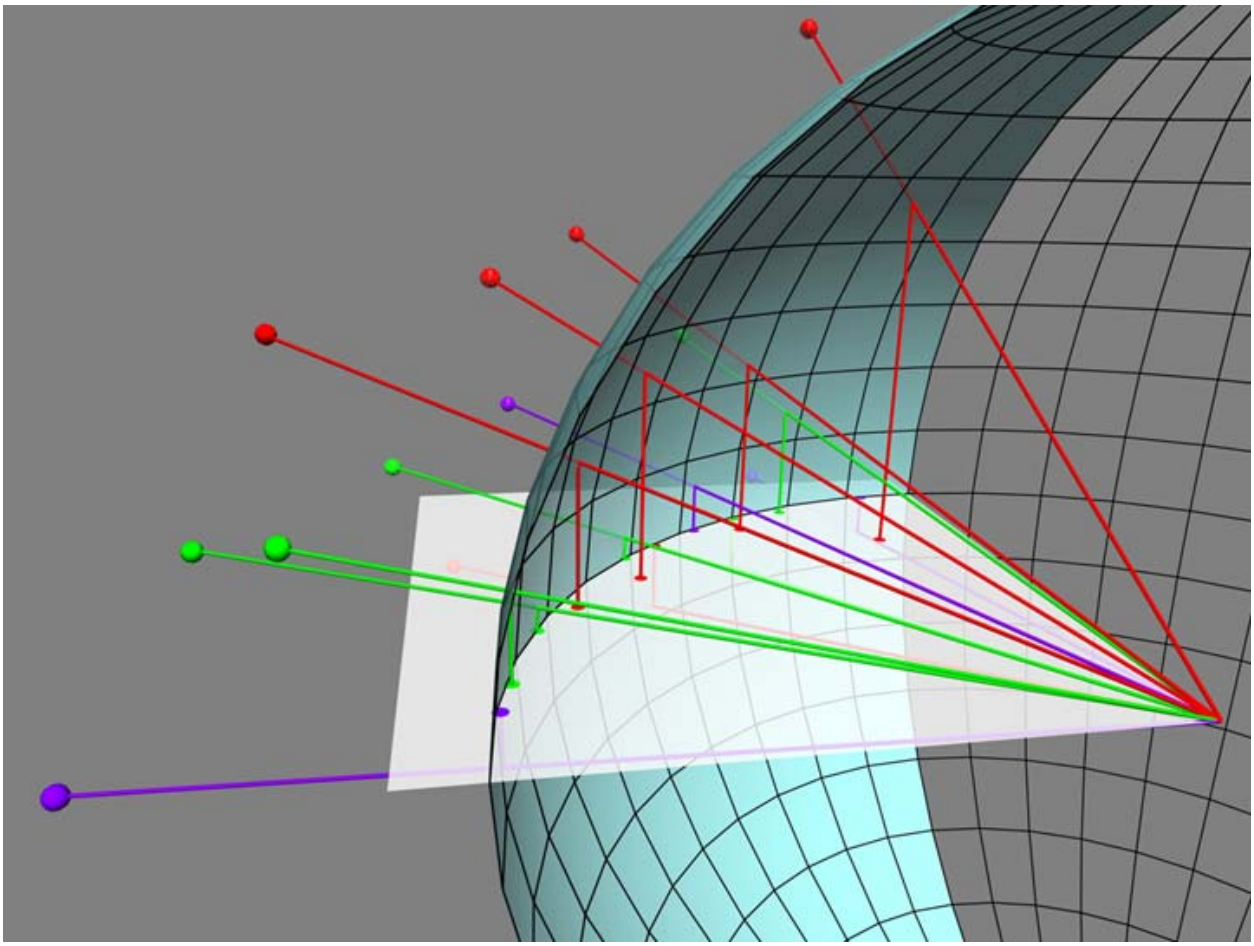

Figure 6. Gene lists projections onto the "126 dimensional unit" sphere in patient space. This is an artificial depiction of the sphere method used to visualize gene lists. The plane spanned by the first two principal components is shown intersecting the sphere, and each gene is shown as a point. The method of projecting from the ambient dimensions to the principal component plane is illustrated by first following a given point back to the sphere and then to the plane via the vertical lines.

In order to understand this visualization is it useful to imagine a sphere with a plane passing through the origin. The sphere corresponds to the unit sphere (the sphere with radius one centered at the origin) in the patient space and the plane corresponds to the plane determined by the first two principal components. The first principal component points in the radial direction of

the sphere and the second principal component is tangential to the sphere at the sphere's intersection with the first principal component. It is precisely the first two dimensions that are shown in Figure 7. The vector representing a particular gene will intersect the unit sphere, and will be near the arc of the sphere (unit circle) in the plane if it lies in the first two principal components. To the extent that the gene lies out of the plane, the projection of the intersection back down onto the plane will lie further inside the arc. The distribution of these projections onto that principal component plane suggests how a given method of gene selection identifies important genes.

Figure 7. ALL vs. AML gene lists comparison. The gene lists that characterize ALL versus AML are shown, with a different color for each of the methods used to obtain them. In distinguishing infant ALL from infant AML we found that most of the genes in the list were co-localized in our representative visualization. Compare this plot with the results shown in Figure 8.

One of the main observations that can be made is the division of the gene lists above and below the center of the plot (i.e. divided by the 2nd principal component). This division is especially noticeable in the Bayesian and discriminant analysis gene lists and is due to the fact that these methods are univariate gene selection methods. The univariate methods rank and subsequently select genes as isolated values, and hence obtain gene lists that are in some sense very

redundant. In contrast, the NeuroFuzzy and MSV methods are multivariate and tend to select gene lists that are less redundant and hence not entirely determined by the first two principal components.

It is evident from Figure 7 that the gene lists selected for the ALL/AML problem are related. Unfortunately, it is equally obvious that the gene lists selected for the remission/failure problem are unrelated, as shown using the same analysis in Figure 8.

Figure 8. Remission vs. Failure gene lists comparison. The gene lists that characterize remission versus failure are shown, with a different color for each of the methods used to obtain them. It can be seen in this figure that distinguishing remission from failure is a difficult task.

In summary, when distinguishing infant ALL from infant AML we found that most of the list were co-localized in our representative visualization (see Figure 4 and 7). When distinguishing remission from failure, on the other hand, we could not arrive at a satisfactory conclusion (Figures 5 and 8). Understanding the relationships between these gene lists was important as we evaluated their implications, although the lists alone were not sufficient. We wanted to understand the mechanisms of these genes in the context of leukemia. The next section discusses how we explored the biology of the genes using one specific gene list.

Class discovery in infant leukemia

The unsupervised force-directed clustering method previously described with respect to gene clusters, can also be used to cluster patients. When applied to the infant data using the similarity of gene expression profiles between patients, we found the existence of three major groups (as shown in Figure 9A), hereafter denoted clusters A, B, and C. We searched for genes with different expression patterns across these three groups using analysis of variance (ANOVA). This method was applied to order all of the genes with respect to different expressions between the groups as shown in Figure 9B. The strengths of these gene lists were studied using statistical bootstrapping. The results suggested that the three groups represented well-separated patient subclasses. Analysis of the genes that characterize each one of these clusters revealed patterns that implied different characteristics with potential clinical relevance. In particular, the three distinct expression profiles are unrelated to type labels or cytogenetics, but are instead characterized by genes predominantly expressed and probably related to three independent disease initiation mechanisms.

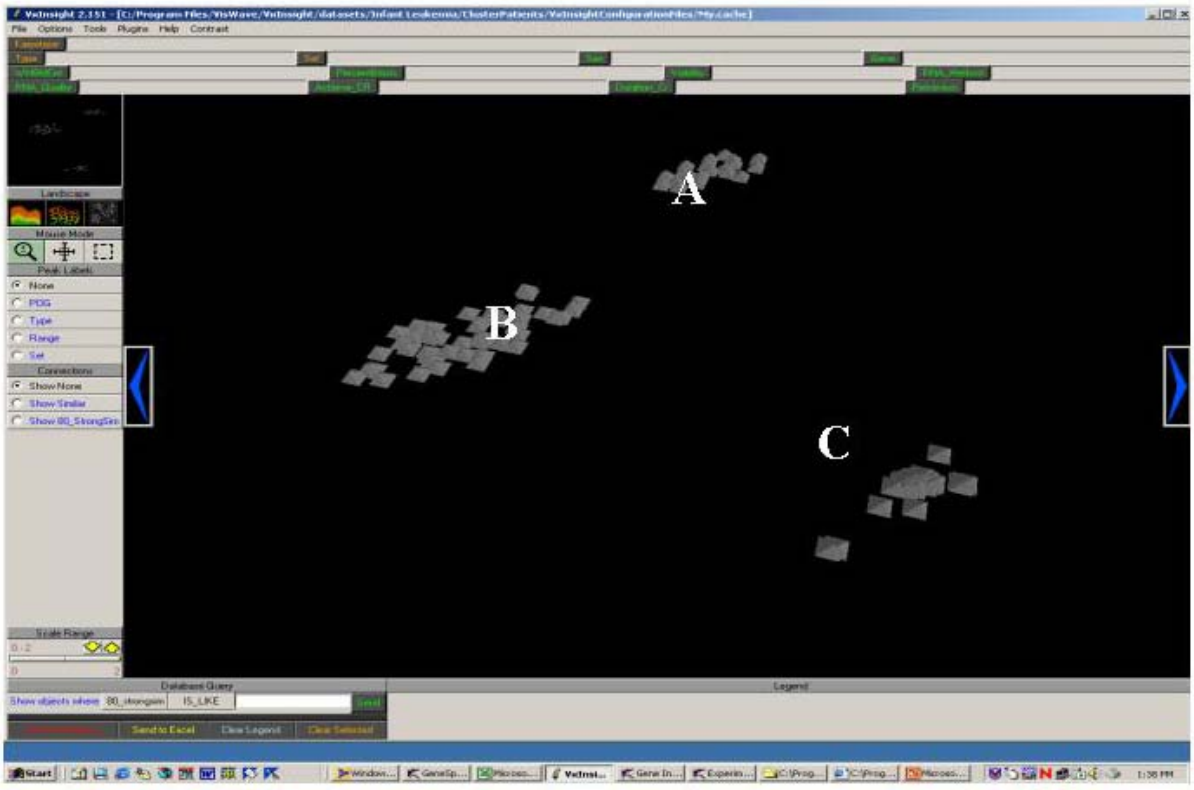

Figure 9A. Cluster-by-patients representation. Results of the force directed algorithm applied to the infant dataset. The VxInsight program constructs a mountain terrain over the clusters such that the height of each mountain represents the number of elements in the cluster under the mountain, A (n=20), B (n=52) and C (n=54). The force-directed clustering algorithm coupled with the VxInsight visualization tool suggested the existence of three clusters of infant patients separated by their gene expression patterns, and not correlated to the traditional clinical labels of morphology: ALL vs. AML, or cytogenetics: MLL rearrangement vs. not

Figure 9B. Gene expression "heat map" of the 126 infant samples (left hand side). Panel B shows the expression levels of the top 89 genes that distinguish the three subgroups of infant leukemia (right hand side) cluster A, B and C; as shown, also, in Figure 9A. Each column represents an infant leukemia sample and each row represents the relative expression for a particular gene across the samples. Gene expression above the mean, below the mean, and around the mean is shown in shades of red, green and black, respectively.

Remarkably, the performance of the supervised, class predictor algorithms improved once the classifiers were conditioned within the A, B, and C clusters (see Table 2). We were particularly interested in the stability, or sensitivity to change in the data, of the rank ordering for these genes. We have previously studied gene list stability by adding increasing amounts of white noise to the gene similarities (Davidson et al, 2001). However, we now believe that statistical bootstrapping, using the actual data, is a better approach (Efron, 1979). Figure 10 outlines how the original data was processed to generate an ordered gene list, and then how 100 random resamplings (with replacement) from that original data were created. These additional data sets yield another 100 gene lists so that each gene in the original list can be annotated to show the range of positions assigned to it across the 100 lists from the bootstrap study. Note that the top ranking genes, as identified by the original measurements, are generally very near the top of the ordered list of 12,625 probes. The number one gene has an average rank order of 5.3, and none of the average rank orders are below 47, thus increasing our confidence in the stability of the reported list. On the other hand, Table 3 shows the list associated with the more difficult problem of separating remission and failure. In this case, while the genes are relatively high compared to the total of 12,625, the average ranking is much less stable than observed for the AML/ALL distinction. The bootstrap method described above addresses list stability, but can be extended to address the null hypothesis, namely, there is no significant difference in gene expression between the two classes being contrasted. By testing this hypothesis we can compute a p-value for the gene's significance.

Table 2. Overall Success Rates of Class Predictors After Including the A, B, and C Cluster Distinctions

Task #	Description	Bayesian Net			SVM			Fuzzy Inference			Discriminant Analysis		
		r	C.I.	p-value	R	C.I.	p-value	r	C.I.	p-value	r	C.I.	p-value
2	Remission. vs. Fail	.568	[.39, .73]	.256	.622	[.45, .78]	.094	.405	[.25, .58]	.906	.568	[.39, .73]	.256
7	Remission. vs. Fail in VX-GA	.714	[.29, .96]	.226	.714	[.29, .96]	.226	.857	[.42, .00]	.062	.714	[.29, .96]	.226
8	Remission. vs. Fail in VX-GB	.688	[.41, .89]	.105	.563	[.30, .80]	.401	.563	[.30, .80]	.401	.438	[.20, .70]	.772
9	Remission. vs. Fail in VX-GC	.714	[.42, .92]	.090	.714	[.42, .92]	.089	.500	[.23, .77]	.604	.500	[.23, .77]	.604
OnVx	R/F Conditioned on VX-Groups	.703	[.53, .84]	.010**	.649	[.47, .80]	.049*	.595	[.42, .75]	.162	.514	[.34, .68]	.500

r = Estimate of the success rate of the class predictor.
C.I. = 95% confidence interval of the success rate of the class predictor.
p-value = p-value of hypothesis test [2] (see text).
* means that r > 0.5 at significance level $D = 0.05$.
** means that r > 0.5 at significance level $D = 0.01$.

Figure 10. Gene list stability exploration. A schematic showing the bootstrap process, where the original data was resampled to create 100 new datasets each of which were processed in exactly the same manner as the original data, to produce the associated 100 new gene lists. The stability of the original data is assessed from the bootstrap distribution; see the text for a more detailed description.

Table 3. Top 24 genes that characterize ALL vs. AML samples, derived from ANOVA.

Order	ANOVA_F	ORF	Contrast	% R R W V W U D S D Y J R U G H U	Description
1	160.82	40103_at	852.39	[1<=[1<=**5.3** {<=16} <=18] p<=0.001	villin 2
2	134.75	39689_at	-822.75	[1<=[1<=**10.4** {<=26} <=32] p<=0.002	cystatin C amyloid angiopathy and cerebral hemorrhage
3	134.62	1230_g_at	-817.22	[1<=[1<=**12.9** {<=46} <=49] p<=0.004	cisplatin resistance associated
4	130.95	39062_at	-930.76	[1<=[1<=**12.4** {<=41} <=42] p<=0.009	protective protein for beta-galactosidase (galactosialidosis)
5	128.94	36766_at	-1389.66	[1<=[1<=**14.4** {<=42} <=46] p<=0.004	ribonuclease RNase A family 2 liver eosinophil-derived neurotoxin
6	124.26	38269_at	794.69	[1<=[1<=**14.8** {<=47} <=60] p<=0.005	protein kinase D2
7	123.69	41523_at	-689.50	[1<=[1<=**14.2** {<=40} <=44] p<=0.007	RAB32 member RAS oncogene family
8	123.18	36938_at	-1003.98	[1<=[1<=**14.0** {<=36} <=40] p<=0.003	N-acylsphingosine amidohydrolase acid ceramidase
9	119.83	40432_at	-918.53	[1<=[1<=**17.1** {<=42} <=47] p<=0.002	glucosamine (N-acetyl)-6-sulfatase (Sanfilippo disease IIID)
10	111.60	36879_at	-968.73	[1<=[1<=**18.6** {<=70} <=73] p<=0.005	endothelial cell growth factor 1 platelet-derived
11	109.22	36889_at	-756.72	[1<=[1<=**20.0** {<=58} <=68] p<=0.002	Fc fragment of IgE high affinity I receptor for gamma polypeptide precursor
12	106.12	1096_g_at	1000.03	[1<=[2<=**20.8** {<=47} <=54] p<=0.007	CD19 antigen
13	101.60	38363_at	-1152.58	[1<=[3<=**26.1** {<=69} <=75] p<=0.008	TYRO protein tyrosine kinase binding protein
14	101.57	38604_at	1032.19	[1<=[7<=**23.6** {<=43} <=48] p<=0.002	neuropeptide Y
15	100.80	37398_at	-824.41	[1<=[4<=**27.0** {<=67} <=77] p<=0.005	platelet/endothelial cell adhesion molecule CD31 antigen
16	100.22	41221_at	-744.57	[1<=[2<=**24.1** {<=56} <=66] p<=0.004	phosphoglycerate mutase 1 brain
17	99.00	40310_at	-625.93	[1<=[2<=**35.4** {<=78} <=123] p<=0.005	toll-like receptor2
18	94.81	35926_s_at	-1584.41	[1<=[3<=**30.7** {<=68} <=79] p<=0.004	leukocyte immunoglobulin-like receptor subfamily B with TM and ITIM domains
19	94.51	39581_at	-736.47	[1<=[2<=**32.0** {<=94} <=96] p<=0.009	cystatin A stefin A
20	93.87	39994_at	-929.11	[1<=[2<=**35.0** {<=92} <=99] p<=0.010	chemokine C-C motif receptor 1
21	89.53	35012_at	-888.77	[1<=[3<=**34.9** {<=76} <=86] p<=0.007	myeloid cell nuclear differentiation antigen
22	87.53	40282_s_at	-904.35	[1<=[3<=**36.0** {<=85} <=111] p<=0.008	adipsin/complement factor D precursor
23	85.87	39593_at	-907.59	[1<=[4<=**47.4** {<=142} <=167] p<=0.009	fibrinogen-like 2
24	85.71	33856_at	-604.82	[1<=[3<=**35.2** {<=86} <=98] p<=0.013	CAAX box 1

Table 3. The 24 genes, out of 12,625, with the largest F-scores by ANOVA to differentiate between ALL and AML samples. Note that these F-scores are only used for ranking, while stability is investigated by bootstrapping (see Figure 10). The average order across the bootstraps is shown for both an upper 95% confidence band, and for the 95% confidence band surrounding the average ranking, which is the bold number. The reported p-value is derived from another bootstrap as described in the text.

The bootstraps described above resample from the two categories being contrasted. For example, the bootstrap AML cases will be drawn from AML patients, and ALL cases are drawn from the ALL patients. However, under the null hypothesis that there is no difference in gene expression between these two cases (AML or ALL) the bootstrap should not distinguish between the cases when resampling. Hence, to test the null hypothesis, the samples are drawn randomly from either type to investigate how rare the actual observation would be in the absence of a real distinction between AML and ALL gene expressions. To compute the significance (p-value) for a gene, we generate 10,000 such bootstraps, and observe the fraction of times the gene ranked at or above the list order found with the real data.

These statistics have been very valuable by allowing us to avoid investing large amounts of effort into genes that are unlikely to be significant. However, a great proportion of the investigative effort still involves reading existing papers and other text about each one of the genes in the lists. We recognized that this text processing had become a bottleneck in our research. As a result, we investigated how Natural Language Processing (NLP) could be employed to help us, as described in the following section.

Gene List Exploration Environment

The next step, in the traditional exploratory analysis of microarray data, is the very labor and knowledge intensive work of learning everything that is known about these genes, especially with respect to disease and biological pathways. We collaborated with computational linguists to build a knowledge-mining tool, which we regularly use in our analysis. The first implementation of our Gene List Exploration Environment (GLEE program) consists of a simple interface that speeds up our search through text about genes identified by any of our approaches. A demonstration version of GLEE, together with user documentation, is available from Computing Research Laboratory web site:

KWWS DLDL.D QPVX HGX

The input to the system is a list of gene identifiers from Affymetrix translated by the program to the equivalent OMIM gene identifier (See Figure 11, and further details at the OMIM web site:

KWWS ZZZ QFEL QOP QLK JRY H)QWUH] TXHU\ I

As shown in the Figure 12, the relevant OMIM text is retrieved and re-ordered to match the criteria that we use for evaluating genes. This automated retrieval and reordering also employs text summarization. We are presently in the process of extending GLEE to use a subset of the NCI Enterprise Vocabulary Services, EV, which is a first step toward a more knowledge-based tool that will be implemented with semantic networks. Because so much of our knowledge about the functions, localizations, and clinical impacts of genes is encoded in published literature, and because the effort to incorporate that knowledge is so labor and knowledge intensive we believe the application of NLP to our specific needs is a critical, and a still largely missing tool for genomic and proteomic investigations.

Figure 11.The Gene Literature Exploration Environment (GLEE) interface is configured as a web server, which handles document and query management, and a web browser that provides the user interface.

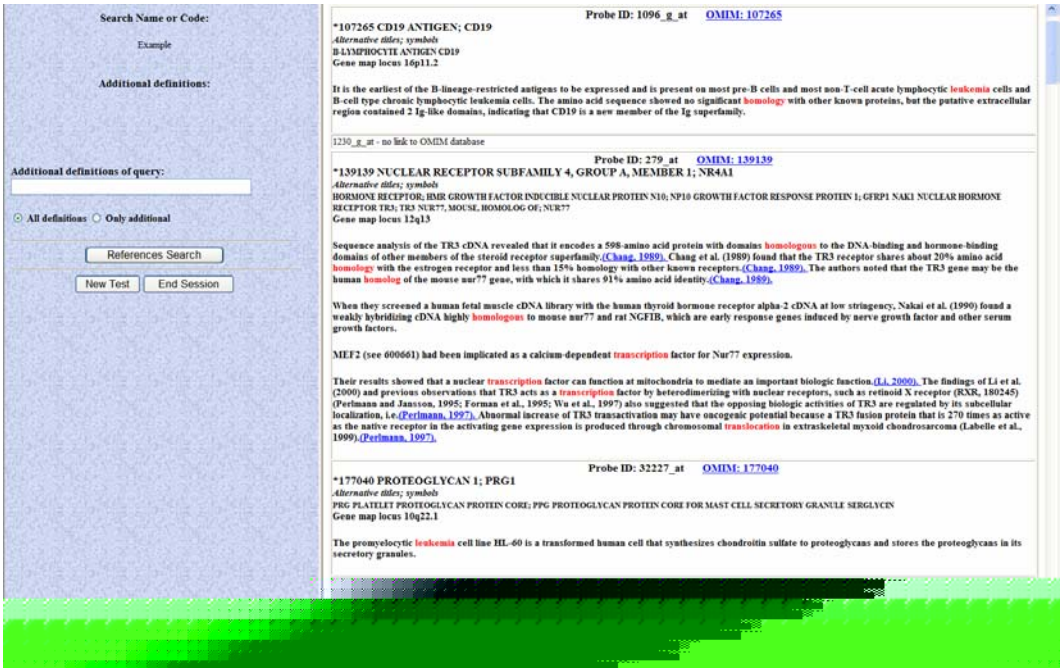

Figure 12.Output of the GLEE program. Summarized and reordered annotations of a set of genes. Note that this is just the first page of annotations; further annotations are available by scrolling down in the browser.