BMC Genomics

BioMed Central

Research

Open Access

Biomarker discovery across annotated and unannotated microarray datasets using semi-supervised learning

Cole Harris*[1] and Noushin Ghaffari[1,2]

Address: [1]Exagen Diagnostics, Inc. Houston, TX, USA and [2]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA

Email: Cole Harris* - charris@exagen.com; Noushin Ghaffari - nghaffari@tamu.edu

* Corresponding author

from IEEE 7th International Conference on Bioinformatics and Bioengineering at Harvard Medical School
Boston, MA, USA. 14–17 October 2007

Published: 16 September 2008

BMC Genomics 2008, **9**(Suppl 2):S7 doi:10.1186/1471-2164-9-S2-S7

This article is available from: http://www.biomedcentral.com/1471-2164/9/S2/S7

Abstract

The growing body of DNA microarray data has the potential to advance our understanding of the molecular basis of disease. However annotating microarray datasets with clinically useful information is not always possible, as this often requires access to detailed patient records. In this study we introduce GLAD, a new Semi-Supervised Learning (SSL) method for combining independent annotated datasets and unannotated datasets with the aim of identifying more robust sample classifiers.

In our method, independent models are developed using subsets of genes for the annotated and unannotated datasets. These models are evaluated according to a scoring function that incorporates terms for classification accuracy on annotated data, and relative cluster separation in unannotated data. Improved models are iteratively generated using a genetic algorithm feature selection technique.

Our results show that the addition of unannotated data into training, significantly improves classifier robustness.

Background

The introduction of DNA microarray technology in 1995 [1] has likely resulted in a huge volume of as yet undiscovered and potentially medically useful knowledge within gene expression profiles. This new bank of information has motivated researchers to develop new techniques for extracting this knowledge, and relating it to externally obtained sample information. For experiments aimed at answering a clinical question, such information might include patient disease stage, or response to a particular drug. The cost of producing adequately annotated datasets has been a barrier to the widespread application of microarray technology in medicine.

Based on the nature of the datasets, a variety of machine learning techniques, including supervised learning algorithms such as classification, and unsupervised learning algorithms such as clustering, have been applied. Clustering techniques [2] are applied to the datasets for assigning samples to their corresponding group solely based on

similar expression levels. Supervised algorithms on the other hand classify [3] samples according to their externally determined class.

None of the standard supervised and unsupervised techniques are appropriate for datasets with some unlabeled samples; Semi-supervised algorithms can address these situations.

Related work

Blum and Mitchell [4] introduced the *co-training* algorithm for improving the sample classification performance when there are few labeled samples and many unlabeled samples. The co-training algorithm assumes that there are two independent sets of features available, such that each feature set is good enough to train a good classifier. The algorithm incorporates an iterative classification of samples from the unlabeled data using two naive Bayes classifiers designed from the independent features sets. In a demonstration of their technique aimed at web page classification, the addition of unlabeled samples decreased classification error relative to classification using only labeled data.

In a subsequent study, Nigam and Ghani [5] further examined the performance of the co-training algorithm and specifically its sensitivity to the independence of the feature sets. Their results confirm that when there is natural split of the features sets, co-training outperforms the other approaches such as expectation-maximization (EM). In the situation that such a split is not available, a random assignment of features into two sets still performs better than using only one feature set. They also introduced the *co-EM* algorithm, a hybrid that iteratively updates the unlabeled data labels using EM. Li et al. [6] proposed a Semi-Supervised Learning (SSL) algorithm for heterogeneous datasets having both labeled and unlabeled samples. Their example data were comprised of DNA microarray expressions and phylogenic reconstructions, with class labels corresponding to gene function. Their work may be considered a form of co-training in that two distinct datasets from a common set of samples (genes) is equivalent to a single dataset with two distinct sets of features. As with the above approaches, independent models are developed for each dataset. They show that minimizing the disagreement in predictions between these models leads to improved accuracy, and introduced a *co-updating* technique for iteratively improving prediction concordance.

Recently, Qi et al. [7] introduced a Bayesian Semi-Supervised approach termed BGEN (Bayesian GENeralization). The BGEN method trains a kernel classifier using both labeled and unlabeled data. Their example data consisted of expression profiles of wild type and mutant *C. elegant* embryos and identified enriched genes, with a small subset of genes labeled according to involvement in development of cell lineage. BGEN predictions were more accurate than predictions from either K-means clustering or SVM classification.

Table 1: Dataset details

	Dataset	Genes before mapping	Genes after mapping	Samples
AML – ALL				
Labeled	AML-ALL 1 [3]	7129	6002	Train: 1-ALL (27) 2-AML (11) Test: 1-ALL (20) 2-AML (14)
Unlabeled	AML-ALL 2 [8]	12582	6002	1-ALL (24) 2-AML (28) 3-MLL (20: deleted)
CML				
Labeled	CML 1 [9]	22283	22283	1-no cytogenetic response to imatinib (15) 2-cytogenetic response to imatinib (30)
Unlabeled	CML 2 [10]	22283	22283	1-Aggressive (10) 2-indolent (9)
DLBCL				
Labeled	DLBCL 1 [11]	7129	1117	1-DLBCL (32: cured, 26: fatal or refractory) 2-FL (19: deleted)
Unlabeled	DLBCL 2 [12]	44928	1117	1-DLBCL (176) 2-MLBCL (34: deleted)

In this paper we propose the Genetic Learning Across Datasets concept (GLAD), and demonstrate an implementation that enables feature selection across unlabeled and labeled datasets. GLAD algorithms are distinct from previous approaches of semi-supervised learning in that the datasets analyzed may have very different statistical distributions, such as would arise in datasets collected independently by labs using different measurement technology. Additionally, a subset of labeled examples is not required for each dataset. As many available datasets will not have the desired annotation for any samples, this method extends the usability of the limited number of adequately annotated microarray datasets.

Methods

Datasets

We conducted three experiments, each addressing a different cancer diagnostic problem: ALL/AML differential diagnosis, prediction of response to imatinib in CML, and prediction of outcome in DLBCL. In each experimental group, two microarray gene expression datasets were selected. If available, labels were removed from one of the component datasets, thus creating a combined dataset with both labeled and unlabeled subsets.

All datasets were produced using Affymetrix GeneChips, and in two cases the labeled and unlabeled datasets were collected with different Affymetrix GeneChips. This required mapping of features between the chips in order to identify a common set of features between the chips. GenBank Gene Accession Numbers were used to generate the common features. Table 1 provides additional details on these datasets.

Demonstrations

For this study we implemented a GLAD algorithm as a wrapper technique for feature selection. A Genetic Algorithm (GA) is used for generating a population of relevant feature subsets. For a given subset, a model is computed from the labeled data and separately for the unlabeled data. Linear Discriminant Analysis (LDA) and K-means (K = 2) cluster algorithms were used for these two data types. A unique two-term scoring function was derived to independently score the labeled and unlabeled data models. An overall score is computed as a weighted average of the two terms as shown below.

$$Score = w \times Score_{\text{labeled}} + (1\text{-}w) \times Score_{unlabeled}$$

We defined the labeled data model score as the standard leave-one-out-cross-validation accuracy for the labeled training samples.

The unlabeled data model score consists of two terms: A cluster separation term and a consistent proportion term.

$$Score_{unlabeled} = \frac{\sum_{i \neq j} \left| C_i - C_j \right|}{\sum_{i \neq j} \left| C_i - C_j \right| + \sum_i \frac{1}{N_{C_i}} \sum_j \left| x_{ij} - C_i \right|} - \sqrt{\frac{1}{n_c} \sum_i \left(\pi_i - \pi_{\exp_i} \right)^2}$$

$C_i \equiv$ centroid of cluster i

$\pi_i \equiv$ proportion of data in cluster i

$\pi_{\exp_i} \equiv$ expected proportion in cluster i

$N_{C_i} \equiv$ number of datapoints in cluster i

$n_c \equiv$ number of clusters

The cluster separation term is given by a modified ratio of the inter-cluster distance to the mean cluster size. The consistent proportion term, is defined as the RMS difference between the sorted actual and expected class priors. The class priors may be estimated from the labeled data, or may be available externally.

For each experiment, we did the following:

1. Iterate GLAD algorithm on labeled training data only.

2. Iterate GLAD algorithm on labeled and unlabeled training data.

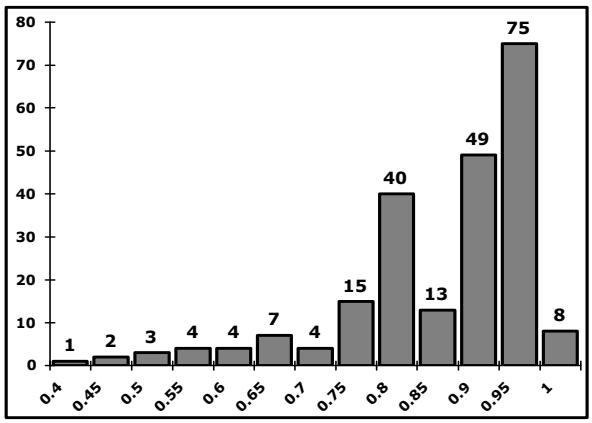
3. Compare model accuracy on test data across generated populations of models.

In these experiments, GLAD was run for 100 iterations with a population size of 5000, and a subset size of 3 features.

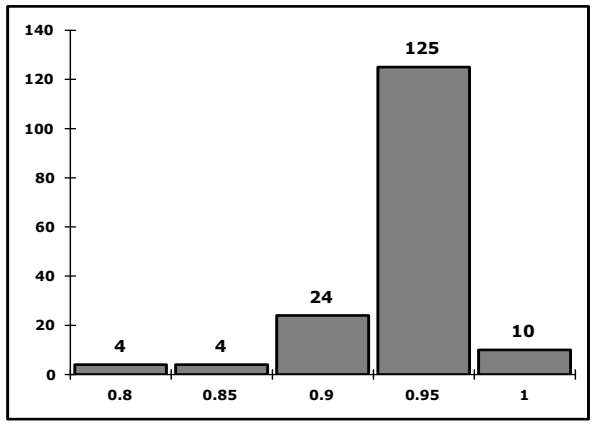
Results and discussion

In the first evaluation of GLAD performance, test data classification accuracy was compared between models identified using only labeled data and models using both labeled and unlabeled data. Figure 1 shows the results for three cancer groups. The top 5% of the model populations were used to generate these histograms. For each cancer set, the two histograms can be compared graphically. As is evident in figure 1, adding unlabeled samples increased the mean accuracy of the models significantly.

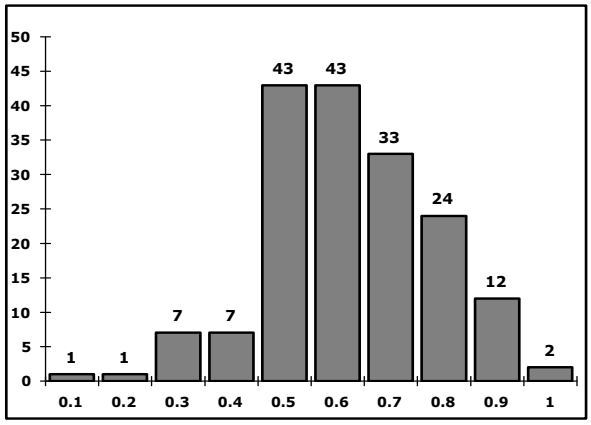
Figure 2 displays the improvements of the classification accuracies for the population of unique classifiers in each cancer group. The output of GLAD has 5000 models, each comprised of 3 genes, with some duplication of classifiers expected. For testing the classification accuracy on the independent set, only unique classifiers were used. Figure 2 compares the performance of the unique classifiers on



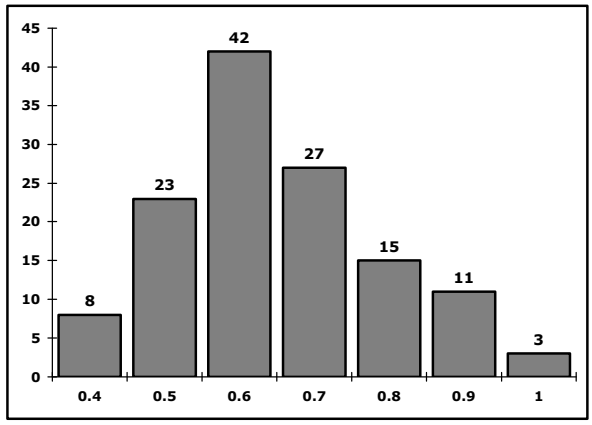
1.a. using only labeled AML-ALL, top 225 classifiers (5%)



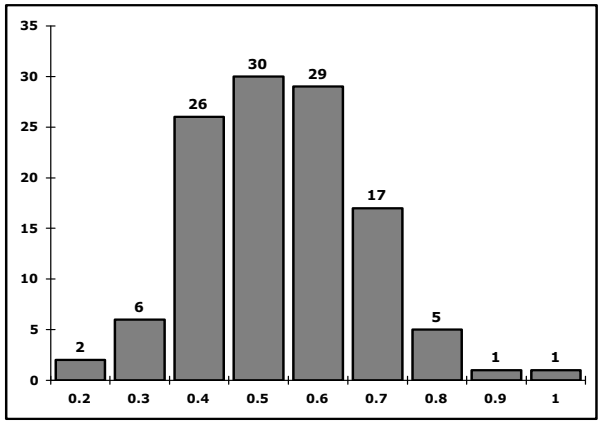
1.b. using labeled plus unlabeled AML-ALL, top 167 classifiers (top 5%)



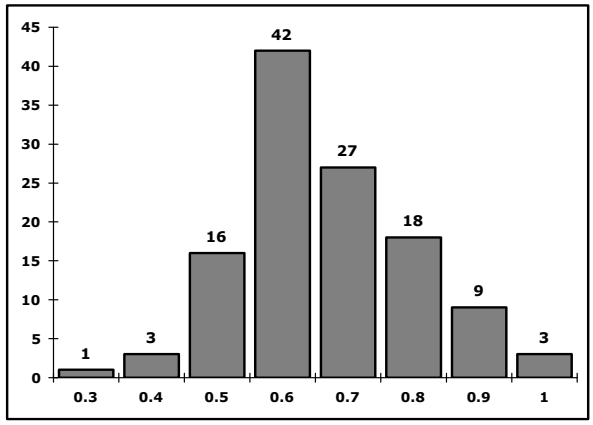
1.c. using only labeled CML, top 173 classifiers (5%)



1.d. using labeled plus unlabeled CML, top 129 classifiers (top 5%)

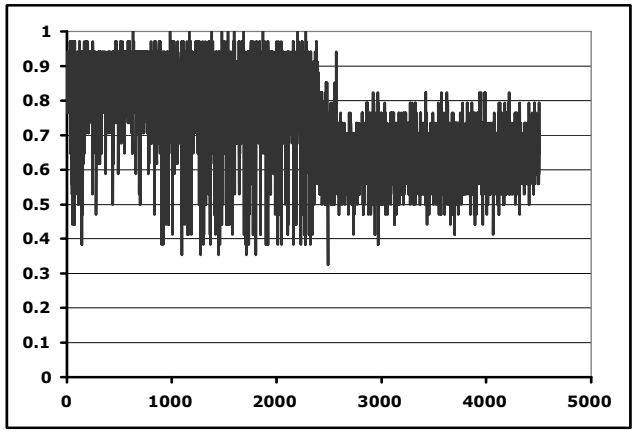


1.e. using only labeled DLBCL, top 117 classifiers (5%)

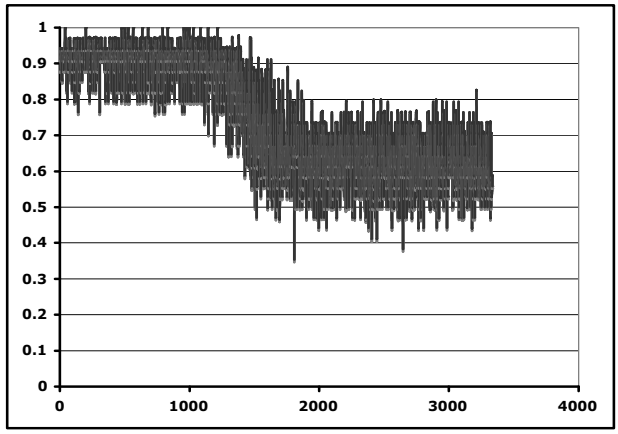


1.f. using labeled plus unlabeled DLBCL, top 119 classifiers (top 5%)

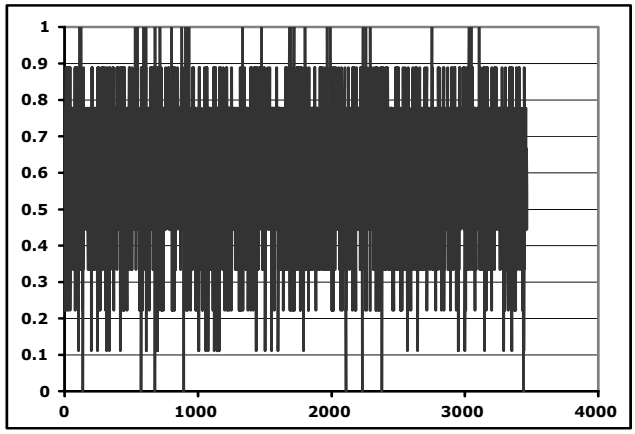
Figure 1
This figures shows the improvement of the classification by adding unlabeled samples into the experiments.



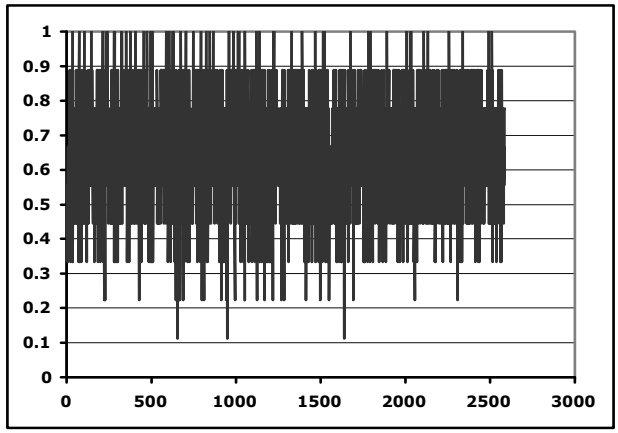
2.a. AML-ALL entire 4504 classifiers, using only labeled samples



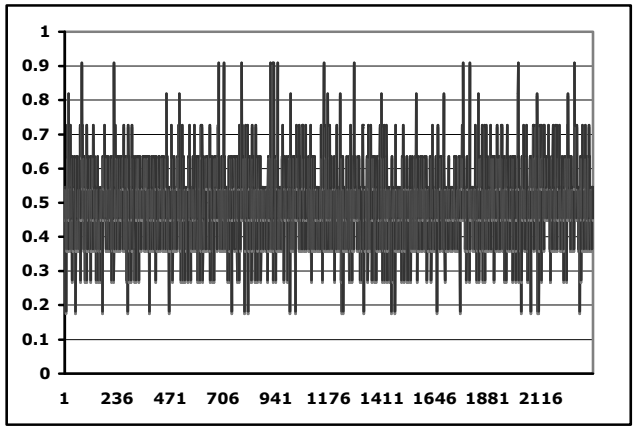
2.b. AML-ALL entire 3336 classifiers, using labeled plus unlabeled samples



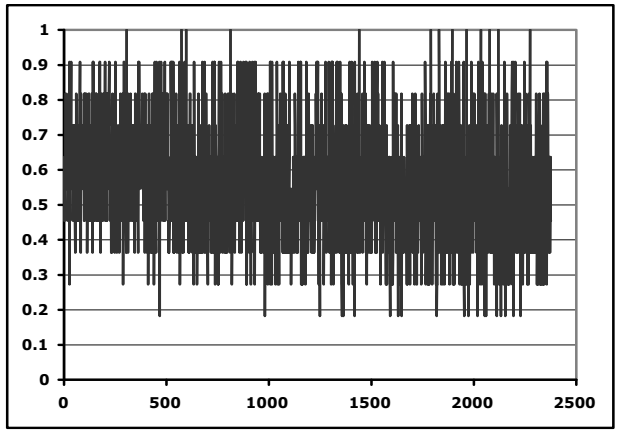
2.c. CML entire 3466 classifiers, using only labeled samples



2.d. CML entire 2587 classifiers, using labeled plus unlabeled samples



2.e. DLBCL entire 2344 classifiers, using only labeled samples



2.f. DLBCL entire 2377 classifiers, using labeled plus unlabeled samples

Figure 2
Comparing the performance of the entire unique classifiers on the testing set for two approaches: 1 – using only labeled samples 2 – using labeled plus unlabeled samples.

Table 2: Improvements by adding unlabeled samples for AML-ALL

Dataset: AML-ALL	Unique Classifiers	Min	Max	Average
only labeled	4504	32.35%	100.00%	73.46%
labeled + unlabeled	3336	35.29%	100.00%	75.14%

the testing set for two approaches: 1 – using only labeled samples 2 – using labeled plus unlabeled samples. The results of the all three cancer groups are improved by adding unlabeled samples to the training sets. Tables 2, 3, 4 show the improvements in more detail. In the AML-ALL group, for the top 1000 classifiers, the accuracy range using only labeled samples is ~40% to 100%. The addition of unlabeled samples increases the range from 70% to 100%. In CML experiments, adding unlabeled samples increases the minimum accuracy from 0% to 11.11%. Combining labeled and unlabeled sample for DLBCL increases the maximum accuracy from 90% to 100%.

Conclusion

In this study we proposed a new technique for concurrently mining labeled and unlabeled datasets. This method supplements standard supervised learning with clustering of data lacking clinical annotation to estimate the predictive power of gene subsets. The performance of our algorithm was evaluated in comparison with supervised learning only on microarray data from three different cancer types. Our results show that adding unlabeled samples can increase the accuracy of classification significantly.

Competing interests

CH and NG were employees of Exagen Diagnostics during the course of this research and the preparation of this manuscript. Additionally, CH owns stock in Exagen Diagnostics.

Authors' contributions

CH devised and implemented the GLAD algorithm, and contributed to the final preparation of the manuscript. NG ran the experiments, interpreted the results, composed early draft versions of the manuscript and contributed to the final preparation of the manuscript. Both authors read and approved the final manuscript.

Table 3: Improvements by adding unlabeled samples for CML

Dataset: CML	Unique Classifiers	Min	Max	Average
only labeled	3466	0.00%	100.00%	59.34%
labeled + unlabeled	2587	11.11%	100.00%	65.57%

Table 4: Improvements by adding unlabeled samples for DLBCL

Dataset: DLBCL	Unique classifiers	Min	Max	Average
only labeled	2344	18.18%	90.91%	49.67%
labeled + unlabeled	2377	18.18%	100.00%	55.79%

Acknowledgements

We thank Exagen Diagnostics for support in conducting this research and presenting these results.

This article has been published as part of *BMC Genomics* Volume 9 Supplement 2, 2008: IEEE 7th International Conference on Bioinformatics and Bioengineering at Harvard Medical School. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/9?issue=S2

References

1. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270:**467-470.
2. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *PNAS* 1998, **95:**14863-14868.
3. Golub TR, *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
4. Blum A, Mitchell TM: **Combining labeled and unlabeled data with co-training.** *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* 1998:92-100.
5. Nigam K, Ghani R: **Analyzing the Effectiveness and Applicability of Co-training.** *Ninth International Conference on Information and Knowledge Management (CIKM-2000)* 2000:86-93.
6. Li T, Zhu S, Li Q, Ogihara M: **Gene Functional Classification by Semi-supervised Learning from heterogeneous data.** *Proceedings of The 18th Annual ACM Symposium on Applied Computing (SAC 2003)-Bioinformatics Track* 2003:78-82.
7. Qi Y, *et al.*: **Semi-supervised analysis of gene expression profiles for lineage-specific development in the Caenorhabditis elegans embryo.** *Bioinformatics* 2006, **22(14):**e417-423.
8. Armstrong SA, *et al.*: **MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia.** *Nature Genetics* 2002, **30:**41-47.
9. Frank O, *et al.*: **Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients.** *Leukemia* 2006, **20:**1400-7.
10. Yong ASM, *et al.*: **Molecular profiling of CD34+ cells identifies low expression of CD7, along with high expression of proteinase 3 or elastase, as predictors of longer survival in patients with CML.** *Blood* 2006, **107:**205-12.
11. Shipp MA, *et al.*: **Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning.** *Nature Medicine* 2002, **8:**68-74.
12. Savage KJ, *et al.*: **The molecular signature of mediastinal large B-cell lymphoma differs from that of other diffuse large B-cell lymphomas and shares features with classical Hodgkin lymphoma.** *Blood* 2003, **102:**3871-9.