Feature Selection with Exponential Model for Classification on Diffuse Large B-Cell Lymphoma

Winston H.-M Hsu, Weiwei Fortino, Zhenyu Peng

Electrical Engineering & Computer Science Departments
Columbia University
{winston@ee., wf98@, zp2103@}columbia.edu

ABSTRACT

In this project, we present a statistical feature selection approach, called exponential model or maximum entropy model that can systematically select the most significant features from high-dimension microarray data. The model utilizes a family of weighted, exponential functions to account for the contributions from different features. The Kullbak-Leibler divergence measure is used in an optimization procedure to iteratively estimate the model parameters, and automatically select the optimal features. When tested on B-cell lymphoma microarray data, the proposed feature selection approach coupled with a neural net classifier achieves significant performance improvement over Fisher criteria score. Especially these selected genes also catch essential genomic signatures from biological interpretations. We also aggressively suggest some hypothesized genes essential to DLCL subtype classifications.

1. INTRODUCTION

One of the obstacles to cancer treatment is to diagnose accurately and to target specific therapy for distinct tumor types. The current clinic diagnosis methods have been developed a lot from simple morphology observation to histological microscopic appearance of pathology cytogenetic abnormality. However, the diverse responses of patients with same diagnosis indicate the heterogeneity. The molecular classification becomes necessary and possible with the development of microarray profiling technology (also called as gene chip).

The small chip can record genomic or expression information of thousands of genes simultaneously. On the spotted cDNA chip, the gene expression profiling can present a wealth of data containing information not only about the absolute expression of a large amount of genes but also about the potential interaction and regulation among genes. The powerful application has involved the study of all fields of biological and medical science, from cancer classification to prognostic prediction and from tissue development to therapeutic methods development. It's especially useful in cancer study to include classification and etiology of cancer.

Feature selection has been successfully used on discovering the subcategories of diffuse large B-cell lymphoma [4]. Diffuse large B-cell lymphoma (DLCL) is one of the most common non-Hodgkin's lymphoma. There are two types of lymphoma: Hodgkin's and non-Hodgkin's lymphoma (NHL). More than 90% lymphoma patients were diagnosed with NHL. But there are ten subcategories within NHL according to WHO classification standard. Also patients with DLCL respond to same treatments quite differently, implying the heterogeneity in this group. The discovery of new subcategories will definitely help physicians develop different treatment regimes for patients belonging to different subcategory.

Whatever chips used in the research have such a high dimension that they are not clinically usable because (1) The more genes the chip covers, the more expensive to prepare the chip; (2) the hybridization condition can not satisfy all the genes so that the result fluctuates even for the same sample; (3) the analysis is very difficult due to high computational cost. In theory it's not necessary to go through all the genes since one type of cancer usually disrupt one or more pathways which affect cell proliferation, programmed cell death (apoptosis), cell cycle controlling, cell differentiation and so on. The expression levels of most genes from cancer cells are similar to those from normal cells. The ideal chip should contain much fewer genes, which are signatures for a certain type of cancer. How to select these informative signature genes from high dimensional microarray data becomes critical for clinical usage of microarray technology. The reduced features could help develop a smaller microarray chip with relevant clones, reduce computational cost and improve accuracy.

There are some challenges for feature selection. First, the data has high dimensions. There are some promising works [2] trying to solve these problems. Especially, biological samples are usually limited. Besides, high feature dimensions often incur problems, such as curse of dimension, during developing a classifier with limited training data samples. Meanwhile, there are always some missing values in the data and data are mislabeled due to the complexity of biology.

In this project, we experiment a statistical feature selection approach trying to find those most "informative" clones/genes that best approximate the empirical sample distributions. Comparing with the popular feature selection filter method using Fisher criteria score, the proposed approach has significant improvement on diffuse large B-cell lymphoma microarray data. The selected features also catch those essential genomic signatures.

2. FEATURE SELECTIONS

Generally, feature selection process could be referenced as filter and wrapper approaches. A wrapper method identifies those salient features from the construction of a classifier. For filter method, statistic approaches are used to find these feature subsets. In this experiment, we investigate Fisher criteria score and also propose the exponential model as the selection process. The experiment result shows that it gains quite high performance even with rather small feature sets.

2.1. Data Set

We have the primary and raw data from [3][4], in which there are 133 samples, with 8062 to 18432 spots on the chip. The number of unique genes in each batch is 7680. And the effective number of genes used in the training and testing set is 6692. Here we skip those chips with too many damaged spots and finally have 96 samples. 42 of them are diffuse large B-cell lymphoma with 21 GC B-like DLCL and 21 activated B-like DLCL. The rest are non-DLCL including different types of tissues and follicular lymphoma, B- chronic lymphocyte leukemia.

Raw data files for each array containing all measured values representing the fluorescent ratios of Cy5, used to label each sample, to the fluorescent ratio of Cy3, used to label reference probes. The fluorescent ratios were calibrated independently for each chip by applying a scaling factor, which was achieved by shifting the median value on the chip to zero. The use of the reference probes makes the data from different chips comparable.

2.2. Fisher Criteria Score

The Fisher criteria score is motivated by the Gaussian model. The score $f_c(i)$ for the i'th feature is defined as,

$$f_c(i) = \frac{|u_i^+ - u_i^-|}{s_i^+ + s_i^-}, \tag{1}$$

where u_i^+, u_i^- are means of feature i on positive and negative training samples and s_i^+, s_i^- are their correspondent standard deviations. Basically, the score is to find those features that discriminate the most from the positive and negative samples and get normalized by their standard deviations. Such a score is expected to work well when the data is normally distributed in each class of samples. On the other hand, if the data is not normally distributed, this score can fail. For example, an asymmetric distribution of values in one of the classes can skew the estimation of the variance and lead to inaccurate score.

2.3. Exponential Model

The exponential model constructs an exponential, log-linear function that fuses multiple features to approximate the posterior probability of an event (i.e., story boundary) given the current observations, as in equation (2). It was initially invented in natural language processing society [5] and recently used for fusing multi-modal features [6]. The construction process includes two main steps – parameter estimation and feature induction. Here we adopt induced features as those "informative" features for further classification.

The exponential posterior probability is denoted as $q(b|x)$, where $b \in \{0,1\}$ is a random variable corresponding to the presence or absence of the special event based on the observation x. From x we compute a set of binary variables: $f_i(x,b) = 1_{\{g_i(x)=b\}} \in \{0,1\}$, $i = 1...K$. $1_{\{\cdot\}}$ is an indication function; g_i is a predictor associating with the output label of the i'th feature. f_i equals 1 if the prediction of predictor g_i equals b, and equals 0 otherwise; K is the total number of predictors or features, and in this case is the number of effective genes (6692). In this experiment we just let $g_i(x) = 1_{\{x(i)>T\}}$, meaning that the feature value clone i in sample x is larger than a threshold T.

Given a training data set $\tilde{D} = \{(x_j, b_j)\}$, we construct a linear exponential function as follows

$$q(b|x) = \frac{1}{Z_l(x)} e^{\sum_i l_i \cdot f_i(x,b)} \tag{2}$$

, where $\sum_i l_i f_i(x,b)$ is a linear combination of binary features with real-valued parameters l_i. $Z_l(x) = e^{\sum_i l_i \cdot f_i(x,0)} + e^{\sum_i l_i \cdot f_i(x,1)}$ is a normalization factor to ensure equation (2) is a valid conditional probability distribution. Basically, l_i controls the contribution of i'th feature in estimating the posterior probability of the current candidate point being a story boundary.

The parameters $\{l_i\}$ are estimated by minimizing Kullbak-Leibler divergence measure computed from $\tilde{D}$ that has empirical distribution $\tilde{p}(b,x)$. The optimal estimation is,

$$q_l^* = \operatorname*{argmin}_l D(\tilde{p} \,\|\, q_l) \tag{3}$$

$D(\cdot \,\|\, \cdot)$ is the divergence which is defined as:

$$D(\tilde{p} \,\|\, q_l) = \sum_x \tilde{p}(x) \sum_{b \in \{1,0\}} \tilde{p}(b \mid x) \log \frac{\tilde{p}(b \mid x)}{q_l(b \mid x)}. \tag{4}$$

During the feature selection process, the system selects those features contributing to the framework the most, step by step. Given a set of candidate features C and an initial exponential model q, the model can be refined by adding a new feature $g \in C$ with weight a .

$$q_{a,g}(b \mid x) = \frac{e^{a\,g(x,b)} q(b \mid x)}{Z_a(x)}, \tag{5}$$

where $Z_a(x) = \sum_{b \in \{0,1\}} e^{a\,g(x,b)} q(b \mid x)$ is the normalization factor. A greedy induction process is to select the feature that has the largest improvement in terms of gain, or divergence reduction, $G_q(g)$, defined in equation (6). Intuitively, the feature contributing the most divergence reduction is the most "informative" feature about the classification work.

$$G_q(g) = \sup_a \left(D(\tilde{p} \,\|\, q) - D(\tilde{p} \,\|\, q_{a,g}) \right) \tag{6}$$

3. EXPERIMENT RESULTS AND DISCUSSIONS

We conducted 2 classification experiments. The first task is to classify DLCL subtypes, including GC-DLCL and activated-DLCL. There are 33 GC-DLCL and 30 activated-DLCL samples. The second task is for DLCL discrimination and there are 63 positive samples and 30 negative ones.

3.1. K-fold cross-validation

A classical neural net with 10 hidden layers and back-propagation training is used for each classification work on different number of feature sets. We interpret the classification performance in terms of classification error rate. Since the data set is quite limited, we adopted *K-fold cross-validation* [1] for performance evaluation, where samples are randomly divided into K approximately equal-size subsets. For each of the subsets, the remaining $K-1$ subsets are combined to form a training set and the resulting classifier's error rate is estimated on the reserved testing subset. A weighted average of the K error rates is used and is weighted for the test set size. At the meanwhile, the entire procedures iterate for R times and the results are averaged. Here we choose $K = 5$ and $R = 10$.

For each classification work, we train a neural net classifier with different feature sets selected with Fisher criteria score and exponential model. We attempt to understand what's the impact of feature numbers on the classification performance and how those feature selection approach works. To compare the significance of the feature selection, we add a random feature selection as a *null* approach to compare the other two sophisticated approaches. The feature configuration varies on the number of features used in the classifier. Here the numbers of features used from feature set 1 to 8 are 5, 10, 20, 25, 50, 75, 100 and 200.

3.2. Classification Results

The result of neural net classification for DLCL subtypes is shown in Figure 1. Apparently, increasing the number of the features would increase the classification performance, and the exponential model outperforms the other approaches. We observed that the classification performance using exponential model are almost independent of the number of features. It might imply that some of the "informative" or "important" features have been caught in the first feature configuration, even 5 features only. It would be interesting if we could further conduct some biological experiment to confirm the hypothesis. Meanwhile, the Fisher criteria score is only slightly better than the random selection approach. We might assert that the distribution of these features is not Gaussian style, which is the main assumption of the approach. The classification result of DLCL vs. the rest is shown in Figure 2. The exponential model still outperforms the other approaches. In some configurations, the random approach even performs better than the Fisher criteria score, which might imply that this approach does not catch those discriminative features.

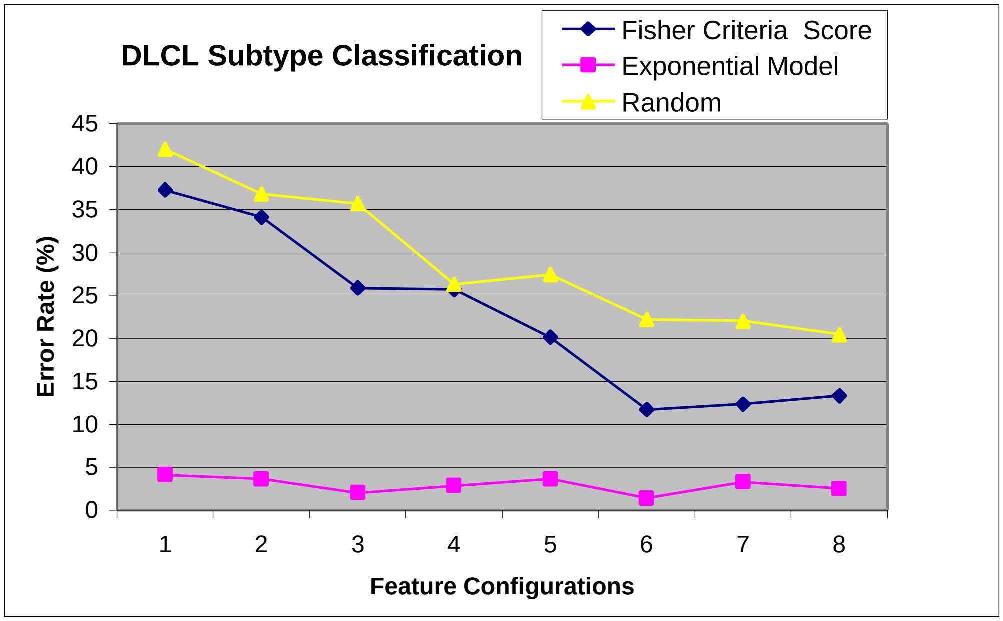

Figure 1: DLCL subtype classification error rate on different feature configurations, in which 5, 10, 20, 25, 50, 75, 100 and 200 features are used.

In these two experiments, the exponential model significantly outperforms Fisher criteria score. It might be because that: (1) Fisher criteria score assumes those features conform to Gaussian distributions. However, in the real case, this does not hold. (2) The selected features in the exponential model could increase the empirical likelihood or reduce the divergence of the constructed model and take into consideration the label of the samples. On the contrast, Fisher criteria score simply calculates the mean and standard deviations of the samples based on the Gaussian assumption. (3) The exponential model picks up those features that contribute to the constructed model with a greedy method. If there are some duplicated features that provide the same discriminative capability, only one of them is selected in this iteration. The next feature selected is the next one that would improve current model the most. Each feature selected would compensate the others. However, in Fisher criteria score, these duplicated features will be selected simultaneously since they get the same scores. Thus the exponential model would have more compensated features constrained on the same feature numbers but the other approach might include those duplicated features that provide similar discriminative capability.

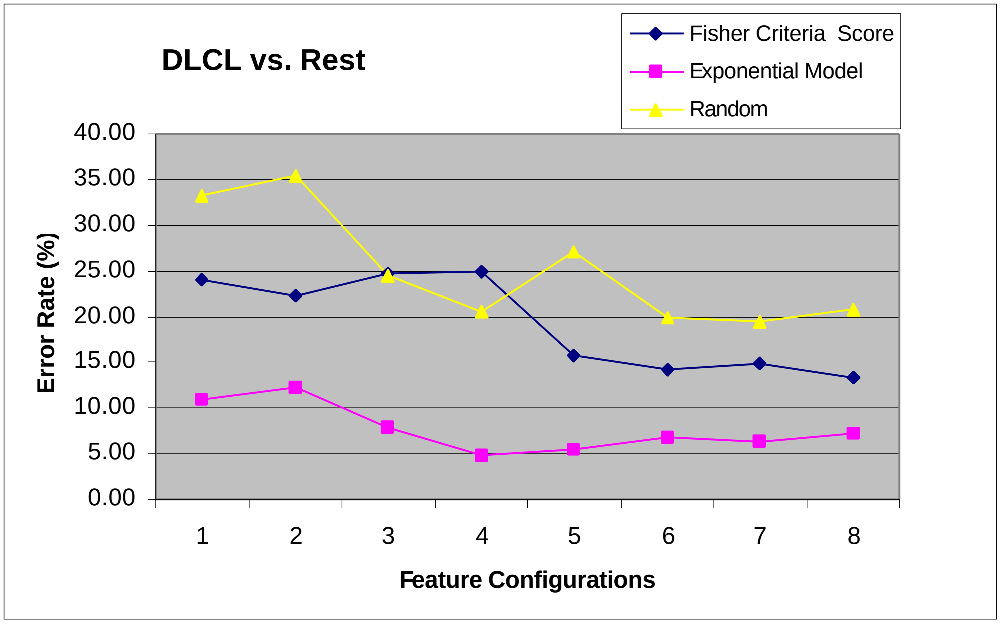

Figure 2: DLCL vs. rest classification error rate on different feature configurations, in which 5, 10, 20, 25, 50, 75, 100 and 200 features are used.

Comparing with these two approaches, the exponential model is much more computationally expensive. However, it delivers quite salient feature selection from a large number of features by measuring its contribution to the gradually constructed classifier. It's still worthwhile if the reduced feature set is further used for building a cost-effective and small microarray chip for detecting certain diseases.

3.3. Selection of informative genes

The selected features by the exponential model from classifying DLCL and non-DLCL reflect genes involved in different pathways, which could possibly induce cancers.

In Table 1, the genes were grouped by their functions. It implies the different pathways that DLCL could disrupt comparing with FL, B-CLL and other samples representing differentiation stages of B cells and different tissues. Note that there are more than 50% unknown genes since the microarrays were made from cDNA clones. The advantage is that it may allow us to discover new genes but the disadvantage is that it leaves gaps to analyze the known genes. Table 1 lists genes from the top 200 features. To our surprise the exponential model gives the informative genes for subgroups in the mixed non-DLCL group. It's very useful since it's not unusual that a sample is mislabeled in biological experiments. The classification also implies potential new classes.

The individual genes in Table 1 infer the potential pathways in which DLCL and non-DLCL could get involved. DLCL samples show varied expression for DNA repair proteins but FL and B-CLL presents uniform expression for these proteins. For example, all B-CLL has low expression of Topoisomerase I and HHR23B. APEX has a low expression level for most FL, B-CLL as well as normal tonsil cell and resting peripheral B cells. Various levels of these proteins from DLL samples prove the heterogeneity of DLCL cells. It's not a surprise that a lot of genes related to T cell growth, which is expressed higher in normal tonsil cells.

Table 1: Informative Genes from classifying DLCL and non-DLCL

Categories	Genes and Annotation
DNA repair	APE X(DNA alkylation repair protein) HHR23B (XPC p58 subunit, damage recognition and nuclear excision repair) RCC1 (regulator of chromosome condensation) Topoisomerase I (uncoil DNA for DNA repair) LAF-4 (lymphoid nuclear protein) XPC (nucleotide excision repair protein)
Cell origin and stage	Immunoglobulin lambda light chain (expression in resting/activated T cell) MHC class II DQ alpha (resting/activated T cell) IL-2/IL-4/Il-7/IL-9/IL-15 receptor common gamma chain (T cell) T cell lymphoma invasion and metastasis 1 IL-15 (T cell growth factor) CD37 (mature B cells)
Cytokine signaling pathway	Transforming growth factor (TGF), beta 3 TGF beta 2 IL-10 receptor IL-11 and CD68 FAS-associated factor 1 SOCS box-containing proteins Germinal Center (GC) kinase and GC kinase related protein kinase LIMK1 (LIM-kinase I) cyclin D2 Protein phosphatase 2A Interferon (IFN) alpha receptor and IFN -related protein c-fos Cadherin-5 MAP kinase kinase 6 STAT proteins BCL-2 (Germinal cell)

Normal cells transform into tumor cells usually by the imbalance of cell proliferation and cell death. The cytokines are signals for hematopoietic cells to grow and differentiate at the right speed at the right time on the right place. Lymphoma tumor cells have one or more disrupted cytokine signaling pathways, shown here by a lot of kinases involved in cytokine signaling pathways. It's believed that low-growth lymphoma like FL and B-CLL are resulted from inhibition of apoptosis while aggressive DLCLs are characterized by enhanced proliferation activity [7]. The major pathways shown here are the MAPK kinase pathway and JAK/STAT signaling pathway, which are the major pathways for B lymphocyte proliferation and maturation. There might be cross talk somewhere in the cells. The under expressed death signal Fast in B-CLL and FL is consistent with the theory of the disruption of apoptosis for these two types of cancer. But under expressed SOCS protein and IL-10 in B-CLL imply other mechanisms. In the JAK/STAT kinase pathway, SOCS proteins can inhibit the kinase activity of JAK by physical binding [8]. Also IL-1 directly inhibits STAT-1 dependent early response gene transcription induced by IFN [8]. Besides antiapoptosis, the stimulation on JAK/STAT is one possible reason for B-CLL and FL, different from DLCL. For the MAPK kinase, DLCL samples presented higher expression levels than other samples, suggesting the increased cell proliferation in the DLCL samples. It's noticeable that BCL-2, over expressed in some of FL and most B-CLL, stands out. BCL-2 is a key anti-apoptotic gene, strongly implying that the disruption of apoptosis pathway through BCL-2 is the reason how tumor cells in FL and B-CLL escape the destiny of death. Most DLCLs have low level of BCL-2 but much higher than germinal cells, indicating the potential mediating functions of BCL-2 in DLCL.

Table 2: Informative Genes from classifying GC B-like DLCL and Activated B-like DLCL

Syndecan-2 (heparin sulfate proteoglycan core protein),vinculin, cadherin, integrin alpha 2,cathepsin B
RAN (GTP-binding nuclear protein, regulating nuclear cytoplasmic transport)
IRF-7 (IFN regulatory factor-7)
PKC (Protein kinase C) alpha
p27kip1 (cyclin kinase inhibitor)
chemokines (eotaxin,TECK)
CD71 (transferrin receptor)
TTK dual specificity kinase
TRAF3
BCL-9
WNT-2
MAPK kinases
CD45
IL-4
JAW1 (lymphoid-restricted membrane protein)
Bfl-1 (Bcl-2 related protein)

Contrary to the low error rate of classification between subgroups of DLCL, the selected top 200 genes are different from the literature [4] except for JAW1. The reasons could be possibly: (1) our selection is global and they chose genes from germinal center first then applied the classifier. Actually we already selected genes like GC kinase, cyclin D2 during classifying DLCL and non-DLCL. (2) The reproducibility of microarray. Golub et al. [10] used Affymetrix oligonucleotide microarray covering approximately 6000 genes to classify the 58 DLCL samples. The 13 selector genes were selected by supervised learning algorithm, only 3 out of 13 were in the lymphochip but not in the top 50 predicator genes selected from lymphochip. (3) The existence of possible 3rd type of DLCL [9].

Even so, from our genes we can still find that they are involved in the similar pathways implied by genes selected by Alizadeh et al. [4]. It's believed that BCL-6 is the GC marker even though its levels vary in GC B-like DLCL. A high expression level of BCL-6 contributes to maintain terminally differentiated cells, which should die later. p27Kip1 is an inhibitor of BCL-6. The low expression of p27Kip1 definitely will release BCL-6, preventing cell from apoptosis [10]. Similarly Bf1-1, a BCL-2 related protein may play similar role like the signature gene FLIP for activated B-like DLCL and inhibits apoptosis of tumor cells. TRAF3 and TTK are involved in TNF (tumor necrosis factor) induced apoptosis like NF-?B [11]. BCL-9 was a potential oncogene and the role has not been defined. But it might somehow be associated with BCL-2 [12].

From Figure 1, the low error rate at the top 5 features strongly implies that the top 5 genes are the most important features, which grasp the nature of the distinction between the two classes. But unfortunately 3 out of top 5 are unknown genes. One is Syndecan-2, which is known now associated with inflammation. But with a lot of other cell adhesion proteins, Syndecan-2 may play an important role with cell-cell interaction and start signal transduction pathway. RAN is a GTP/GDP binding protein, regulating the nuclear cytoplasmic transport [13]. In DLCL the regulation of transcription factors binding with DNA in the nuclei is the down regulated by the signaling pathway. It's very possible that the regulation by RAN of the importing of the transcription factors plays a key role in tumor genesis.

4. CONCLUSIONS

The exponential model successfully identifies features discriminating the DLCL from non-DLCL even though the non-DLCL group contains varied samples while Fisher criteria score provides poor discrimination due to its assumption of normal distribution of data samples. The selected top 200 genes are quite informative about the pathways in which different lymphoma get involved. The correlation of these genes is quite high even though there are a lot of unknown genes.

The features from exponential model between two subgroups of DLCL presents an amazing low error rate even with features as few as 5 genes. Though the selected features don't match literatures of other methods, they are possible to disclose some new mechanisms related to tumor genesis, which need more biological experiments to confirm.

Biological data are generally not error-free due to experiment processes or physical constrains, such as the spot defects on the microarray chips. Classification or feature selection approaches that could deal with or tolerate missing values are also necessary for these kinds of works.

As investigating the primary data samples, we found that microarray chips are subject to errors that might affect the completeness of genes. Most vendors use spot duplication to solve the problem. It would be more efficient and interesting if we could approximate the error model of microarray chips and try to put those duplications with specific spatial layouts that could balance error rate, spot numbers or costs. This is not only a research topic but also helps realize the technology into clinical treatments.

REFERENCES

[1] J. Kent Martin, D.S. Hirschberg, “Small Sample Statistics for Classification Error Rates I: Error Rate Measurements,” Tech. Rep. 96-21, Department of Information and Computer Science, UC Irvine, 1996.

[2] Eric P. Xing and Michael I. Jordan and Richard M. Karp, “feature Selection for High-Dimensional Genomic Microarray Data,” Proc. 18th International Conf. on Machine Learning, 2001.

[3] National Institutes of Health, “Lymphoma/Leukemia Molecular Profiling Project, ” URL: http://llmpp.nih.gov/lymphoma/

[4] Ash A. Alizadeh, et al., “Distinct Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling,” Vol. 403, Nature, Feb. 2000.

[5] A. Berger, S. D. Pietra, and V. Pietra, "A Maximum Entropy Approach to Natural Language Processing," Computational Linguistics, (22-1), March 1996.

[6] Winston Hsu, Shih-Fu Chang, “A Statistical Framework for Fusing Mid-level Perceptual Features in News Story Segmentation,” IEEE International Conference on Multimedia & Expo (ICME) 2003.

[7] Margariat Sanchez-Beato et al., "Cell Cycle Deregulation in B-cell Lymphomas" Blood, 101,2003.

[8] Katsuhiko Ishihara et al, "Molecular basis of the cell specificity of cytokine action", Biochimica et Biophysica Acta 1592, 281-296, 2002.

[9] C. Schwaenen et al "DNA microarray analysis in malignant lymphomas", Annals of Hematology, 2003.

[10] Hosokawa Y, Maeda Y and Seto M. "Target genes downregulated by the BCL-6/LAZ3 oncoprotein in mouse Ba/F3 cells,” Biochem. Biophys. Res. Commun., 283, 563–568,2001

[11] H. Ah-Kim et al. "Tumor necrosis factor alpha enhances the expression of hydroxyl lyase, cytoplasmic antiproteinase-2 and a dual specificity kinase TTK in human chondrocyte-like cells," Cytokine 12: 142-50, 2000.

[12] A. Sarris and R. Ford, " Recent advances in the molecular pathogenesis of lymphomas," Current Opinion Oncology, 11: 351-363, 1999.

[13] K. Weiss, "Regulating access to the genome: nucleocytoplasmic transport throughout the cell cycle,” Cell 112: 441-451, 2003.