Genome Informatics 21: 126-137 (2008)

INFERRING DIFFERENTIAL LEUKOCYTE ACTIVITY FROM ANTIBODY MICROARRAYS USING A LATENT VARIABLE MODEL

JOSHUA W.K. HO[1,6] RAJEEV KOUNDINYA[2] TIBÉRIO S. CAETANO[5,6]
joshua@it.usyd.edu.au raj@anatomy.usyd.edu.au tiberio.caetano@nicta.com.au
CRISTOBAL G. DOS REMEDIOS[2] MICHAEL A. CHARLESTON[1,3,4]
crisdos@anatomy.usyd.edu.au mcharleston@it.usyd.edu.au

[1] *School of Information Technologies, The University of Sydney, NSW 2006, Australia*
[2] *Bosch Institute, The University of Sydney, NSW 2006, Australia*
[3] *Sydney Bioinformatics, The University of Sydney, NSW 2006, Australia*
[4] *Centre for Mathematical Biology, The University of Sydney, NSW 2006, Australia*
[5] *RSISE, Australian National University, ACT 2601, Australia*
[6] *NICTA, Australia*

Recent development of cluster of differentiation (CD) antibody arrays has enabled expression levels of many leukocyte surface CD antigens to be monitored simultaneously. Such membrane-proteome surveys have provided a powerful means to detect changes in leukocyte activity in various human diseases, such as cancer and cardiovascular diseases. The challenge is to devise a computational method to infer differential leukocyte activity among multiple biological states based on antigen expression profiles. Standard DNA microarray analysis methods cannot accurately infer differential leukocyte activity because they often fail to take the cell-to-antigen relationships into account. Here we present a novel latent variable model (LVM) approach to tackle this problem. The idea is to model each cell type as a latent variable, and represent the class-to-cell and cell-to-antigen relationships as a LVM. Once the parameters of the LVM are learned from the data, differentially active leukocytes can be easily identified from the model. We describe the model formulation and assumptions which lead to an efficient expectation-maximization algorithm. Our LVM method was applied to re-analyze two cardiovascular disease datasets. We show that our results match existing biological knowledge better than other methods such as gene set enrichment analysis. Furthermore, we discuss how our approach can be extended to become a general framework for gene set analysis for DNA microarrays.

Keywords: antibody microarray; latent variable model; Bayesian network; EM algorithm

1. Introduction

Leukocytes (white blood cells) play a critical role in the human immune system. Several subtypes of leukocytes exist, including granulocytes, lymphocytes (T, B and NK cells), monocytes and others. These leukocyte subtypes can be characterized by different subsets of cell surface proteins, called cluster of differentiation (CD) antigens. The activity (in terms of absolute cell count, or density of expressed CD antigens) of each leukocyte subtype is associated with inflammation, particularly in

cardiovascular diseases [10, 15]. Therefore efficiently quantifying leukocyte activity is important. Our laboratory has been developing a cell-captured antibody microarray platform that enables concurrent quantification of many CD antigens [1, 2]. This array platform has been successfully used to identify changes in the immunophenotype of various human diseases, such as leukemia [1, 2], heart failure [8, 9], and coronary artery disease [4].

Standard DNA microarray analysis methods, such as differential expression (DE) analysis, clustering and classification, are used to analyze these antigen expression profiles. However, essentially none of them can directly infer differential activity of leukocyte subpopulations as they only focus on mining "interesting" antigen expression patterns. Currently we rely on manual inspection of the list of DE antigens and their associated leukocyte subtypes to infer cellular activity. This approach is subject to human bias, and does not scale to analyzing larger expression profiles. Therefore the challenge is to devise a computational method that can accurately infer differential leukocyte activity from a set of antigen expression profiles.

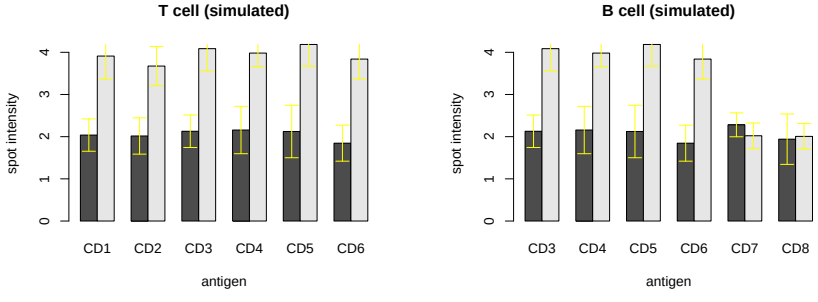


Fig. 1. The mean expression values of a simulated "toy" dataset. The dark and light bars represent antigen expression from normal individuals and diseased patients respectively.

We initially tackled this problem by gene set enrichment analysis (GSEA) [19]. In this case, gene sets correspond to leukocyte subpopulations. However, we soon discovered that the small number of genes and gene sets, and the large amount of overlap among gene sets leads to incorrect inference of leukocyte activity. To illustrate these problems, we constructed a simple "toy" dataset of two cell types - T and B cells, where each expresses six CD antigens, and four of these are expressed by both cell types (Figure 1). T cells were simulated to have elevated activity in the diseased patients, while B cells activity remain unchanged. GSEA indicated that neither T nor B cells was significantly enriched in DE antigens, based on the false discovery rate (FDR) of 0.70 and 0.69 respectively. Since the FDR calculated by GSEA is related to the distribution of enrichment score of all the gene sets, a large overlap between the two gene sets renders both gene sets insignificant. These statistical problems are likely to be shared by other gene set analysis methods which

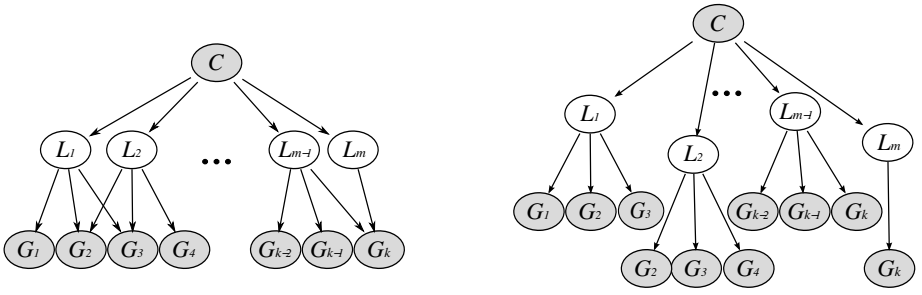


Fig. 2. An exemplary latent variable model (LVM). The shaded nodes are observed variables while the clear nodes represent latent variables. (a) The full model. (b) The effective decomposition of the full model.

use a hypothesis-testing approach [12]. Clearly we need an alternative approach that takes the cell-to-antigen relationships into account.

In this paper, we present a probabilistic graphical modeling approach to solve the problem. The main idea is to encode the observed data (antigen expression and class labels) and unobserved entities (leukocyte activity) as a special type of Bayesian network, called a latent variable model (LVM). The structure of the LVM is determined by the cell-to-antigen relationships, and the model parameters can be learned from the data (see Figure 2(a) for an exemplary LVM). Once the model parameters are learned, differential cellular activity can easily be obtained by performing probabilistic inference on the model.

2. Methods

2.1. *Model specification*

Our LVM consists of a set of variables $\mathbf{X} = \{C, \mathbf{L}, \mathbf{G}\}$, which includes an observed class variable, C, a set of m latent variables, $\mathbf{L} = \{L_1, L_2, ..., L_m\}$, and a set of k observed antigen expression variables, $\mathbf{G} = \{G_1, G_2, ..., G_k\}$. In this paper, the latent variables represent the cellular activity status. These variables are connected as a Bayesian network such that each L_i is the immediate parent of a subset of $\mathbf{G}$ corresponding to the antigens that are expressed by that cell type, and each L_i is an immediate child of C (Figure 2(a)). For convenience, we denote the set of $w(X)$ parents and $d(X)$ descendents of any variable X as $\pi(X) = \{\pi_1^X, \pi_2^X, ..., \pi_{w(X)}^X\}$ and $\delta(X) = \{\delta_1^X, \delta_2^X, ..., \delta_{d(X)}^X\}$. We further denote the set of possible realizations, or the *state space*, of each variable X to be $S(X)$. Each node X is associated with a conditional probability distribution (CPD), which is the probability distribution of X given the state of its parents, *i.e.*, $P(X|\pi(X))$. Using the standard Bayesian network approach, the joint probability distribution (JPD) of the LVM can be decomposed as the product of local CPDs:

$$p(\mathbf{X}) = p(C)\left[\prod_{i=1}^{m} p(L_i|C)\right]\left[\prod_{i=1}^{k} p(G_i|\pi(G_i))\right] \quad (1)$$

We observe that the terms $p(G_i|\pi(G_i))$ do not generally decompose because each G_i can have many parents. Full parameterization of this CPD can result in a large computational burden during parameter estimation. Therefore we introduce an assumption here to further simplify the JPD:

$$p(G_i|\pi(G_i)) = p(G_i|\pi_1^{G_i}, \pi_2^{G_i}, ..., \pi_{w(G_i)}^{G_i}) = p(G_i|\pi_1^{G_i})p(G_i|\pi_2^{G_i})...p(G_i|\pi_{w(G_i)}^{G_i}) \quad (2)$$

The above assumption decomposes the poly-tree structure of the LVM into a tree by effectively duplicating those G_i with $|\pi(G_i)| > 1$ (Figure 2(b)). This leads to the following effective decomposition of the JPD:

$$p(\mathbf{X}) = p(C)\prod_{i=1}^{m}\left[p(L_i|C)\prod_{j=1}^{d(L_i)} p(\delta_j^{L_i}|L_i)\right] \quad (3)$$

The LVM is associated with a set of model parameters, which are used to specify the CPDs. We model $P(C)$ as a multinomial distribution, where $S(C)$ is the set of distinct class labels. Since $P(C)$ is the relative frequency of each distinct class label, it can be directly estimated from the dataset. We model $P(L_i|C)$ as a binary variable where $S(L_i) = \{\text{inactive}, \text{active}\}$. Each $P(L_i|C)$ is associated with $2|S(C)|$ parameters, each specifying the probability of L_i being active or inactive in each of the $|S(C)|$ classes. $P(G_i|\pi(G_i))$ is modeled as a Gaussian distribution with means $\{\mu_{G_i,1}, .., \mu_{G_i,w(G_i)}\}$, and variances $\{\sigma^2_{G_i,1}, .., \sigma^2_{G_i,w(G_i)}\}$.

One major consequence of the decomposition of the JPD in Equation 3 is that data of some nodes are duplicated. In general, such duplication of data may lead to bias in parameter learning. To alleviate this problem, we down-play the contribution of each duplicated antigens G_i by scaling up the set of variances $\{\sigma^2_{G_i,1}, .., \sigma^2_{G_i,w(G_i)}\}$. The basic idea is that antigens that are expressed by more than one cell type should have higher expression variability compared to antigens that are expressed by only one cell type. Therefore, we fix the variance of antigen expression per cell to be proportional to the number of parents, *i.e.*, $\sigma^2_{i,j} = w(G_i)^r \times \sigma^2$, where we use $r = 3$ in this study since it works well in practice. The more parents G_i has, the higher is its expression variance. Using this formulation, the set of parameters that have to be estimated from the model is $\Theta = \{\theta_{L_1|C}, ..., \theta_{L_m|C}, \theta_{G_1|\pi(G_1)}, ..., \theta_{G_k|\pi(G_k)}\}$ where $\theta_{L_i|C} = p(L_i|C)$, and $\theta_{G_i|\pi(G_i)} = \{\mu_{G_i,1}, .., \mu_{G_i,w(G_i)}\}$.

2.2. *Parameter learning using EM algorithm*

Here we describe an efficient algorithm to obtain an approximate maximum likelihood estimate (MLE) of the LVM parameters from data. Since our model contains latent variables, we learn parameters by the expectation maximization (EM)

approach [5]. The main idea of the EM algorithm is to iteratively calculate the expected distribution of the latent variables (E-step) and then use the results from E-step to re-estimate the MLE (M-step). Since the log likelihood of the model increases after each iteration of E- and M-step, the algorithm terminates when the expected log-likelihood of the model converges.

Given the observed data of array u, the E-step finds the expected distribution of the latent variables, $q_u(\mathbf{L}|C, \mathbf{G}, \Theta^{(t-1)})$, based on the current parameters. Using the effective decomposition in Equation 3, we can decompose q_u as well:

$$\begin{aligned} q_u(\mathbf{L}|C, \mathbf{G}, \Theta^{(t-1)}) &= p(L_1, L_2, ..., L_m|G_1, G_2, ..., G_k, C) \\ &= \frac{p(L_1, L_2, ..., L_m, G_1, G_2, ..., G_k, C)}{\sum_{S(\mathbf{L})} p(L_1, L_2, ..., L_m, G_1, G_2, ..., G_k, C)} \\ &= \frac{p(L_1|C)p(\delta(L_1)|L_1)}{\sum_{S(L_1)} p(L_1|C)p(\delta(L_1)|L_1)} ... \frac{p(L_m|C)p(\delta(L_m)|L_m)}{\sum_{S(L_m)} p(L_m|C)p(\delta(L_m)|L_m)} \\ &= q_u(L_1|C, \mathbf{G}, \Theta^{(t-1)}) ... q_u(L_m|C, \mathbf{G}, \Theta^{(t-1)}) \end{aligned}$$

Such decomposition implies that expected distribution of $\mathbf{L}$ is the product of the expected marginal distribution of each L_i, which can be computed by:

$$q_u(L_i|C, \mathbf{G}, \Theta^{(t-1)}) = \frac{p(L_i|C) \prod_{j=1}^{d_i} p(\delta_j^{L_i}|L_i)}{\sum_{S(L_i)} p(L_i|C) \prod_{j=1}^{d(L_i)} p(\delta_j^{L_i}|L_i)}$$

The M-step re-estimates Θ using the set of q_u calculated from the previous E-step. Given the antigen expression data $\{e_{1,1}, ..., e_{n,k}\}$, and the class labels $\{c_1, ..., c_n\}$, where n is the number of arrays and k is the number of antigens, we can calculate the MLEs as follows:

$$p(L_i|C) = \frac{\sum_{u=1}^{n} q_u(L_i|C, \mathbf{G})}{n} \qquad \mu_{L_i,j} = \frac{\sum_{u=1}^{n} e_{u,j} \times q_u(L_i|C, \mathbf{G})}{\sum_{u=1}^{n} q_u(L_i|C, \mathbf{G})}$$

Since there is no known efficient way to obtain the best initial parameters, we turn to a heuristic approach. The idea is to iteratively try out different random initial parameters. We select the parameter set that produces a model with the highest likelihood score after two iterations of EM. In general, the more random initial parameter sets being tested, the higher chance of finding the optimal one.

2.3. *Model analysis*

Once the model parameters are estimated, differential cellular activity can be obtained by inspecting the set of $P(L_i|C)$. To quantify the extent of differential cellular activity, we use total correlation $C_{\text{tot}}(L_i, C)$ [20] to measure the extent of dependency between L_i and C. Total correlation can be calculated by:

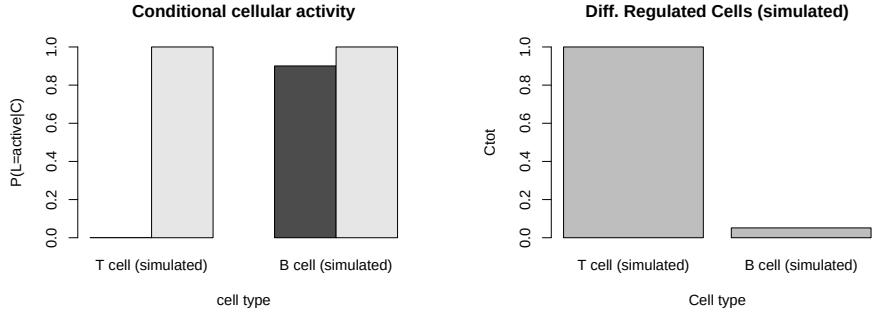
$$C_{\text{tot}}(L_i, C) = \sum_{l \in S(L_i)} \sum_{c \in S(C)} p(l, c) \log \left[\frac{p(l, c)}{p(l)p(c)} \right]$$

where $p(L_i)$ and $p(C)$ are the marginal distributions of L_i and C respectively. If L_i and C are statistically independent, C_{tot} becomes 0. In general, a higher C_{tot} implies that L_i is more strongly dependent on C, and therefore more differentially active.

3. Results

3.1. *Analysis of the toy example*

Here we analyze the toy example described in the introduction (Figure 1) using our LVM approach. We performed 20 iterations of search heuristics (described in Section 2.2) to obtain the initial values, then performed 20 iterations of EM procedures. By inspecting the set of $P(L_i = \text{active}|C)$, we observe an increase in T cell activity in patients with disease compared to healthy individuals (Figure 3(a)). The C_{tot} of T and B cells are 1.0 and 0.052 respectively (Figure 3(b)), which correctly implies that T cell is the only cell type that is differentially active. Moreover, we observed that the total correlation results converge after the first two EM iterations, implying that our results are stable.



(a) (b)

Fig. 3. Results of LVM analysis of the simulated toy example shown in Figure 1. (a) A plot showing the probability of each cell type being active in each condition. The dark and light bars represent the healthy and diseased individuals respectively. (b) The C_{tot} of T and B cells.

3.2. *Re-analysis of two cardiovascular disease datasets*

Two cardiovascular disease datasets [4, 9] were re-analyzed using our LVM approach. All data were generated in our laboratory using an 82 spot antibody array platform.

In the original studies, only peripheral blood mononuclear cells (PBMCs), which include T cells (T), natural killer cells (NK), B cells (B) and monocytes (M), were investigated. The set of CD antigens being expressed by each leukocyte subpopulation is shown in Table 1. The set of CD antigens that are not expressed by any PBMCs are also listed here under the category *Others*, and should be regarded as a negative control for the analysis since it should not be differentially active. After data filtering and normalization (discussed in the original studies), the datasets were analyzed by our LVM approach. For each dataset, we performed 100 iterations of heuristic search to obtain the initial parameters, then performed 20 iterations of EM procedures to obtain the model parameters.

Table 1. A list of all CD antigens expressed by each type of PBMC.

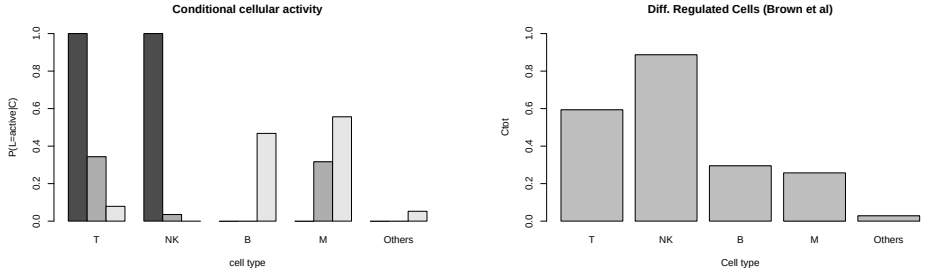
Leukocyte	CD antigens[a]
T cell (T)	TCR a/b TCR g/d CD1a CD2 CD3 CD4 CD5 CD7 CD8 CD9 CD11a CD11b CD11c CD16 CD25 CD28 CD29 CD31 CD37 CD38 CD43 CD44 CD45 CD45RA CD49d CD49e CD52 CD54 CD56 CD57 CD60 CD62L CD80 CD86 CD95 CD102 CD103 CD120a CD122 CD126 CD128 CD130 CD134 CD154
B cell (B)	CD1a CD2 CD5 CD9 CD11a CD11b CD11c CD19 CD20 CD21 CD22 CD23 CD24 CD25 CD29 CD31 CD32 CD37 CD38 CD40 CD44 CD45 CD45RA CD45RO CD49d CD52 CD54 CD62L CD77 CD79a CD79b CD80 CD86 CD95 CD102 CD120a CD122 CD126 CD130 CD138 HLA-DR l FMC7 k
Monocyte (M)	CD1a CD4 CD9 CD11a CD11b CD11c CD13 CD14 CD15 CD16 CD29 CD31 CD32 CD33 CD36 CD37 CD38 CD40 CD43 CD44 CD45 CD45RA CD45RO CD49d CD49e CD52 CD54 CD60 CD61 CD62L CD64 CD65 CD80 CD86 CD88 CD95 CD102 CD120a CD122 CD126 CD128 CD130 HLA-DR
Natural Killer (NK)	CD2 CD7 CD8 CD11a CD11b CD11c CD16 CD25 CD29 CD31 CD38 CD43 CD44 CD45 CD45RA CD45RO CD49d CD49e CD52 CD56 CD57 CD62L CD95 CD102 CD120a CD122 CD128 CD130
Others	CD10 CD34 CD41 CD42a CD62E CD62P CD66c CD71 CD117 CD135 CD235a

Note: [a]These relationships were extracted from the official poster of the Eight International Workshop on Human Leukocyte Differentiation Antigens.

Brown *et al.* [4] studied two major coronary artery diseases (CAD): stable angina pectoris (SAP), and unstable angina pectoris (UAP). The dataset consists of antigen expression profiles from 15 SAP patients, 19 UAP patients and 19 healthy donors. Brown *et al.* manually mapped 19 DE antigens with the leukocytes that express them, and concluded that the observed patterns support a drop in T cell activity and an elevation in monocyte activity. Our results support their conclusion. Additionally we observe a drop in NK cell activity in CAD patients (Figure 4(a)-(b)). Unlike the original analysis by Brown *et al.* [4], we excluded granulocytes from our analysis since they are not PBMC. As noted by Brown *et al.*, the presence of granulocytes specific CD antigens may be an experimental artefact.

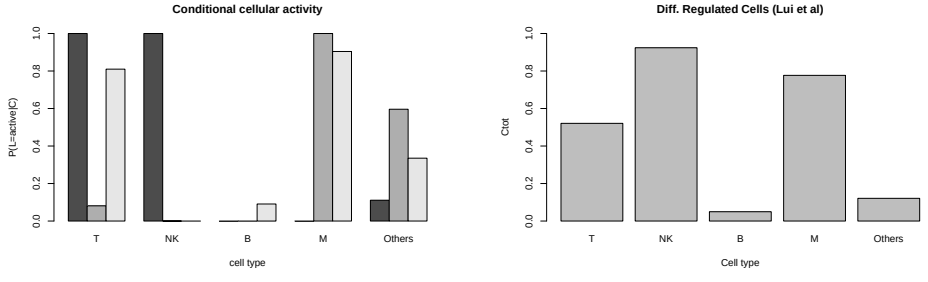
Lui *et al.* [9] studied two major aetiologies of heart failure (HF): ischemic heart disease (IHD), and idiopathic dilated cardiomyopathy (IDCM). Their dataset consists of antigen expression profiles from 22 IHD patients, 15 IDCM patients and 19

healthy donors. Our results (Figure 5(a)-(b)) show that HF patients have decreased NK cell activity and elevated monocyte activity. Further, we found that T cells are down-regulated in IHD patients but not in IDCM patients.



(a) (b)

Fig. 4. The LVM analysis result of Brown *et al.*'s data. (a) The conditional cellular activity plot. (b) The C_{tot} of various leukocyte populations.



(a) (b)

Fig. 5. The LVM analysis result of Lui *et al.*'s data. (a) The conditional cellular activity plot. (b) The C_{tot} of various leukocyte populations.

In general, our approach indicates that there are decreased T and NK cell activity and increased monocyte activity in cardiovascular patients compared to healthy donors. An increase in monocyte count is known to be linked to various cardiovascular conditions [13, 14, 21]. In our arrays, all CD antigens in NK cells represented in our arrays are also expressed by other leukocytes in this study (primarily because NK cells are a sub-lineage of T cells). None of the original studies found differential activity of NK cells, since their changes are attributed to other classes of leukocytes. However, our model detected a strong signal for decrease in NK cells activity in both CAD and HF compared to healthy donors. This drop in NK cell activity is supported by the literature [7]. T cell activity is down-regulated in CADs and IHD, but not in IDCM. This is again consistent with previous findings which link

decreased T cell count with myocardial infarction [3]. Our results correctly indicate no differential activity for the *Others* category in both studies.

In addition to our LVM analysis, we performed GSEA [19] on the two datasets. We used version 2 of the Java GSEA program [23]. Default parameters were used for all analyses. Only half of those true differentially active leukocyte subtypes (according to known biology and visual inspection of the data) are considered significantly enriched with DE antigens by GSEA (Table 2). The significant enrichment of B cells in Lui *et al.*'s dataset contradicts the results from manual data inspection and known biological knowledge. The results indicate that our LVM approach is superior to GSEA in terms of identifying biologically meaningful differential leukocyte activities. We note that general conclusion holds even when a nominal P-value is used to determine statistical significance.

Table 2. Results from GSEA. Gene sets with FDR$\leq$ 0.25 are deemed significant (in **bold**).

Analysis	Up-regulated in control (FDR)	Up-regulated in disease (FDR)
control vs. SAP	**T (0.11)** , B (0.66), NK (0.49)	M (0.33), Others (0.87)
control vs. UAP	T (0.27), NK (0.54)	M (0.62), Others (0.95), B (0.9)
control vs. IHD	**T (0.051)** , **B (0.17)**, **NK (0.17)**	M (0.64), Others (0.63)
control vs. IDCM	**T (0.25)**, **B (0.15)**, **NK (0.15)**	M (0.34), Others (0.67)

4. Discussion

There has been a great interest in applying probabilistic graphical modeling (PGM) techniques to analyzing microarray data. Applications of PGM include pathway discovery [17], regulatory gene modules discovery [16], inferring alternative splice variants [18], and inferring gene network structures [6]. One advantage of PGM is that it allows structural information (relationships between variables) and systems dynamics (expression values) to be integrated under a simple yet theoretically sound framework.

There are two main contributions in this paper. The first is the application of PGM to the inference of differential leukocyte activity using antigen expression profiles. The re-analysis of the two real datasets clearly demonstrates the applicability of our approach to discover biological knowledge. With an increasing number of arrayed antibodies and more reliable experimental protocols, this cell-captured antibody array technology should become increasingly useful in both basic biological investigations and clinical diagnostic applications.

To demonstrate the merit of our approach, let us consider the mean expression value of all the CD antigens expressed by T cell in the Brown *et al.* dataset as an example (Figure 6). The changes in expression patterns across all these antigens differ a lot since many antigens are expressed by other leukocytes. We notice that the expression patterns of those cell specific antigens are much more informative in elucidating the cellular activity. However, removing those antigens expressed by

multiple leukocytes is not desirable since some leukocytes do not express, or express only one or two, cell specific CD antigens (like NK cells in this study). Therefore our LVM model provides a general framework for such inference.

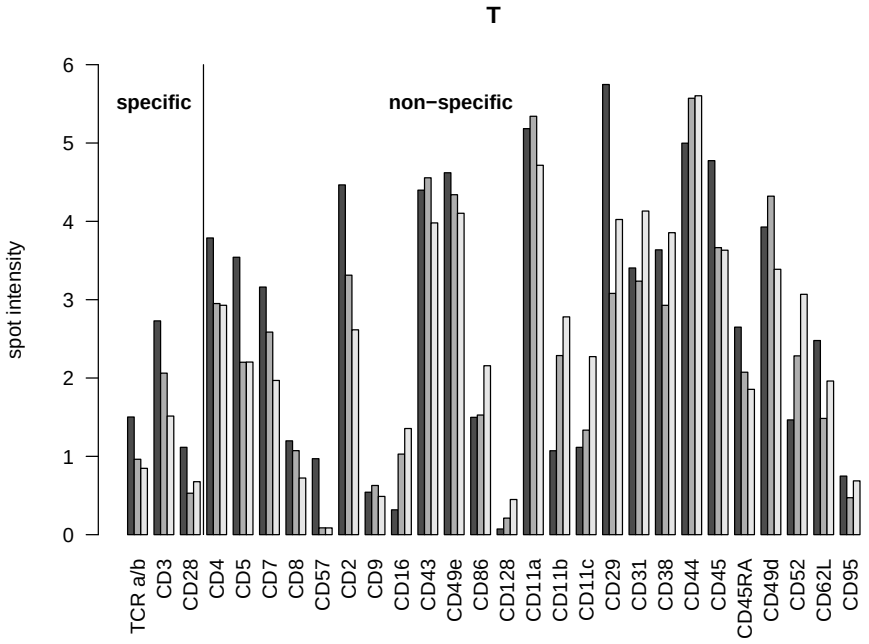


Fig. 6. The mean antigen expression levels of the CD antigens associated with the T cells from the Brown *et al.* dataset. The CD antigens in the barplots are sorted according to the number of different cell types that express it. The antigens on the left of the vertical line represent T cell specific CD antigens. The dark, gray and light bars represent healthy donors, SAP and UAP patients respectively.

Our second contribution is to introduce a novel LVM approach for microarray gene set analysis. Our model is similar to the hierarchical naïve Bayes model proposed by Zhang *et al.* [22], except that our LVM consists of strictly one level of latent variables, and our LVM network structure is known *a priori*. Since the network structure of the LVM is given by biological knowledge, our method eliminates the need to perform computationally expensive structural learning.

In this work we also present a computationally efficient method to learn the conditional probabilities associated with the latent variables. The computational efficiency is achieved by the product assumption in Equation 2, which leads to the decomposition of the JPD (Equation 3). To avoid losing the antigen overlapping information, we made the second assumption that antigens which are expressed by multiple cell types have higher expression variability. This assumption effectively gives more weight to cell type specific CD antigens. As a result, the antigen over-

lapping information is retained without increasing the computational complexity in parameter learning. The effectiveness of our approach is demonstrated by the analyses of a simulated and two real datasets.

We propose that our LVM approach can be used as a general framework for finding differentially expressed gene sets in DNA microarrays. Since the initial publication of GSEA [11, 19], many gene set analysis methods emerged [12]. All of them use a hypothesis testing approach to define interesting gene sets. However, as indicated by our toy example, the correctness of the results depends on meeting a set of assumptions which may be biologically or technically unrealistic. Our LVM approach is not based on hypothesis testing, so the aim of our method is not to find significantly differentially expressed gene sets, but to map the gene expression profiles into the hidden gene set expression space.

In general, there are many possible formulations of the CPDs in our model. We are currently investigating the CPD formulation that is most suitable for general gene set analysis. Moreover, we will investigate the use of other learning techniques to achieve more robust estimates of the model parameters. Nonetheless, this paper presents a conceptually new approach to perform gene set analysis.

Acknowledgement

JWKH is supported by an Australia Postgraduate Award and a NICTA Research Project Award. We thank Angus Brown and Rodney Lui for providing the antibody microarray data.

References

[1] Belov, L., de la Vega, O., dos Remedios, C.G., Mulligan, S.P., and Christopherson, R.I., Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. *Cancer Res.*, 61:4483-4489, 2001.

[2] Belov, L., Huang, P., Barber, N., Mulligan, S.P., and Christopherson, R.I., Identification of repertoires of surface antigens on leukemias using an antibody microarray, *Proteomics*, 3:2147-2154, 2003.

[3] Blum, A., Sclarovsky, S., Rehavia, E., and Shohat, B., Levels of T-lymphocyte subpopulations, interleukin-1 beta, and soluble interleukin-2 receptor in acute myocardial infarction, *Am. Heart J.*, 127:1226-1230, 1994.

[4] Brown, A., Lattimore, J.-D., McGrady, M., Sullivan, D., Dyer, W., Braet, F., and dos Remedios, C.G., Stable and unstable angina: Identifying novel markers on circulating leukocytes. *Proteomics Clin. Appl.*, 2:90-98, 2008.

[5] Dempster, A.P., Laird, N.M., and Rubin. D.B., Maximum likelihood from incomplete data via the EM algorithm, *J. R. Statist. Soc. B.*, 39:1-38, 1977.

[6] Friedman, N., Inferring cellular networks using probabilistic graphical models. *Science*, 303:799-805, 2004.

[7] Jonasson, L., Backteman, K., and Ernerudh, J., Loss of natural killer cell activity in patients with coronary artery disease. *Atherosclerosis*, 183:316-321, 2005.

[8] Lal, S., Lui, R., Nguyen, L., Macdonald, P.S., Denyer, G., and dos Remedios, C.G., Increases in leukocyte cluster of differentiation antigen expression during cardiopul-

monary bypass in patients undergoing heart transplantation, *Proteomics*, 4:1918-1926, 2004

[9] Lui, R., Macdonald, P.S., Hayward, C., and dos Remedios, C.G., Proteomics analysis of leukocyte membrane proteins from human heart failure patients using an antibody microarray platform, *J. Mol. Cell. Cardiol.*, 42:S146, 2007.

[10] Madjid, M., Awan, I., Willerson, J.T., and Casscells, S.W., Leukocyte count and coronary heart disease. *J. Am. Coll. Cardiol.*, 44:1945-1956, 2004.

[11] Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrle, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., and Groop. L.C., PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nat. Genet.*, 34:267-273, 2003.

[12] Nam, D., and Kim, S.-Y., Gene-set approach for expression pattern analysis, *Brief. Bioinform.*, 9:189-197, 2008.

[13] Nasir, K., Gaullar, E., Navas-Acien, A., Criqui, M.H., and Lima, J.A.C., Relationship of monocyte count and peripheral arterial disease: Results from the national health and nutrition examination survey 1999-2002, *Arteroscler. Thromb. Vasc. Biol.*, 25:1966-1971, 2005.

[14] Olivares, R., Ducimetiére, P., and Claude, J.R., Monocyte count: A risk factor for coronary heart disease, *Am. J. Epidemiol.*, 137:49-53, 1993.

[15] Ommen, S.R., Hodge, D.O., Rodeheffer, R.J., McGregor, C.G.A., Thomson, S.P., and Gibbons, R.J., Predictive power of the relative lymphocyte concentration in patients with advanced heart failure, *Circulation*, 97:19-22, 1998.

[16] Segal, E., Shapira, M., Regev, A., Pe 弹 r, D., Botstein, D., Koller, D., and Friedman, F., Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat. Genet.*, 34:166-176, 2003.

[17] Segal, E., Wang, H., and Koller, D., Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19:i264-i272, 2003.

[18] Shai, O., Morris, Q.D., Blencowe, B.J., and Frey, B.J., Inferring global levels of alternative splicing isoforms using a generative model of microarray data, *Bioinformatics*, 22:606-613, 2006.

[19] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, R.S., and Mesirov., J.P., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *Proc. Natl. Acad. Sci. U.S.A.*, 102:15545-15550, 2005.

[20] Watanabe, S., Information theoretical analysis of multivariate correlation, *IBM J. Res. Dev.*, 4:66-82, 1960.

[21] Zalai, C.V., Kolodziejcyk, M.D., Pilarski, L., Christov, A., Nation, P.N., Lundstrom-Hobman, M., Tymchak, W., Dzavik, V., Humen, D.P., William, K., Jablonsky, G., Pflugfelder, P.W., Brown, J.E., and Lucas, A., Increased circulating monocyte activation in patients with unstable coronary syndromes, *J. Am. Coll. Cardiol.*, 38:1340-1347, 2001.

[22] Zhang, N.L., Nielsen, T.D., and Jensen, F.V., Latent variable discovery in classification models, *Artif. Intell. Med.*, 30:283-299, 2004.

[23] `http://www.broad.mit.edu/gsea/`