2014

ISSN : 0974 - 7435 Volume 10 Issue 7

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(7), 2014 [2245-2255]

A Sequence model based phenotype structure discovery algorithm

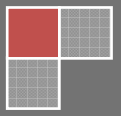
Yu-Hai Zhao*, Ying Yin
College of Information Science & Engineering, Northeastern University, Shenyang 110819, (CHINA)
E-mail : zhaoyuhai@ise.neu.edu.cn

ABSTRACT

Phenotype structure discovery is one of the most important problem in microarray data analysis. The goal is to (1) find groups of samples corresponding to different phenotypes (such as disease or normal), and (2) for each group of samples,find the representative expression pattern that distinguishes this group from others. Different from the existing singleton discriminability based approach and combination discriminability-based approach, we present a novel method in this paper. Based on the proposed g^*-sequence model, an efficient algorithm, namely FINDER, is developed to mine the optimal phenotype structure from a given dataset. Further, several effective pruning strategies are designed to improve the efficiency. The experiments conducted on both synthetic and real microarray datasets show that the phenotype structures discovered by FINDER are of both statistical and biological significance. Moreover, FINDER is 2~3 orders of magnitude faster than the alternatives.

KEYWORDS

Data mining; Bioinformatics; Microarray data.



INTRODUCTION

Advanced microarray technologies have made large amounts of gene expression profiles available. Analyzing microarray data is essential for understanding the gene functions, gene regulation, cellular process, and subtypes of cells[1–3].

An important task in microarray data analysis is phenotype structure discovery[4]. Given a microarray dataset of *m* samples and *n* genes, a phenotype structure refers to a group of "blocks" (or submatrices), each of which consists of a subset of samples and a subset of genes such that: (1) the samples from all the blocks make up a partition of *m* samples, and the samples in a block correspond to a phenotype (such as a disease subtype); and (2) the gene expression pattern within a block can be used as the signature to distinguish this group of samples from others[5]. The genes in a signature may suggest the potential biomarkers related to the disease. In particular, phenotype structure discovery is an unsupervised learning problem. It is more challenging than the problem of biomaker selection with known class labels[4,6].

Most existing phenotype structure discovery methods fall into two major categories, i.e. singleton discriminability based and combination discriminability based[4,6]. The former selects top-ranked genes according to their individual discriminative power to the target classes[6]. Obviously, this over simplifies the complex relationship among genes due to the gene independence assumption. The latter aims to find a subset of genes of the high combinatorial discriminative power. However, it just take into account the co-occurrence of genes. This often leads to a large number of selected genes, as poses crucial challenge for biologists to interpret and validate the results.

In this paper, we model the discriminative genes from a new perspective by exploiting their ordered gene expression values. Compared with the existing models, our model is more robust to noise. Figure 1 illustrates our basic motivation by an real example from Prostate cancer gene expression dataset[8].

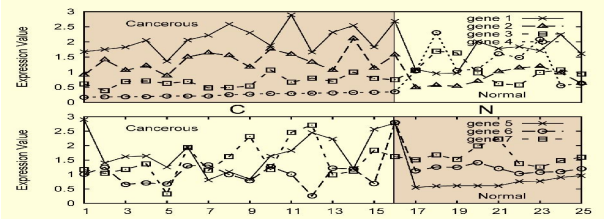


Figure 1 : A real example from the Prostate cancer dataset

Figure 1 consists of two subfigures. In the top subfigure, 4 genes are expressed over 25 samples. Samples 1~16 are cancerous (labeled as '*C*') and samples 17~25 are normal (labeled as '*N*'). In the bottom subfigure, another set of 3 genes are expressed over the same set of samples. The existing singleton or combination discriminability-based methods cannot distinguish the two phenotypes. Since most genes are of similar average expression values in the two phenotypes, they will not be selected by the singleton approach. Moreover, all genes are expressed in both phenotypes. Thus, the combination approach based on the co-occurrence of genes will not select them either. Both of the methods ignore the hidden interrelation among genes. In the top subfigure, the gene order over the samples of cancerous phenotype '*C*' is always $gene_4 \prec gene_3 \prec gene_2 \prec gene_1$. Such order is disturbed in normal phenotype '*N*'. In the bottom subfigure, the gene order in normal phenotype '*N*' is $gene_5 \prec gene_6 \prec gene_7$, while in cancerous phenotype '*C*' such order does not exist. Based on the ordered expression values, the disease phenotypes (the two shadowed "blocks") are well identified.

In biology community, discriminative sequential patterns involving the ordered gene expression values have been shown effective in distinguishing phenotypes[7,9]. Such patterns have an intuitive biological interpretation. Complex diseases often involve the cooperation of multiples genes. These genes work together as a system to keep the cell in a specific state, e.g., disease or normal. In such a

state, some special interrelationship among genes will exhibit. Once such relationship is disrupted, the state may change, e.g., from normal to disease.

In this paper, we propose a novel phenotype structure discovery method by profitably exploiting the ordered gene expression values. Our contributions are summarized as follows.

(1) A *g**-sequence model is devised. It introduces the *significant chain* to ensure the robustness of the proposed model, and enables to identify highly discriminative signatures with only a small number of genes.

(2) A novel sequence dissimilarity metric, namely *projection divergence*, is proposed. By this metric, the difference between a pair of blocks (submatrices) can be quantified based on the signatures features of the blocks.

(3) An efficient algorithm, FINDER, is developed to find the optimal phenotype structure. By incorporating the cross projection into a progressive exploring framework, candidate phenotype structures are searched in a quality-guaranteed way.

The rest of this paper is organized as follows. In Section 1, we introduce some preliminaries and give the problem statement. Section 2 details our solution. Experimental analysis is given in Section 3. Finally, section 4 concludes this paper.

THE PRELIMINARY

In this section, we first introduce some basic concepts useful for further discussion, and then formalize the problem to be addressed in this paper.

g*-sequence

A microarray dataset D is an $m \times n$ matrix, with m samples $S=\{s_1, s_2,\ldots, s_m\}$ and n genes $G=\{g_1, g_2,\ldots,g_n\}$. A real value d_{ij} in D represents the expression value of gene g_j on sample s_i. An example microarray dataset of 4 samples and 9 genes is shown in TABLE 1. Microarray data are often noisy. We introduce the concept of equivalent dimension group which represents a set of genes with similar expression values.

TABLE 1 : An example Microarray dataset

Sample	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
s_1	103	68	76	48	71	101	55	50	83
s_2	35.5	20.1	28.7	17.2	13.2	23.8	13.5	15.8	30
s_3	5.7	6.7	9	5	10.3	10	15.2	5.2	8.7
s_4	32	53	79	43	35	72	105	38	68

Definition 1. Given an expression matrix D of a sample set, $S=\{s_1, s_2,\ldots, s_m\}$, and a gene set, $G=\{g_1, g_2,\ldots,g_n\}$, if for a grouping threshold δ, $\delta \geq 0$, and some sample $s_i \in S$, there exists a subset, G', of genes holding both Eq.(1) and Eq.(2), we say G' is an equivalent dimension group, or an *EDG* for short, of the sample s_i.

$$\max_{g_j, g_{j'} \in G'} |d_{ij} - d_{ij'}| < \delta \times \min_{g_j \in G'} d_{ij} \quad (1)$$

$$\forall g_i, g_j \in G', \min_{g_{j'} \in G'} \left|d_{ij} - d_{ij'}\right| < \min_{g_{j''} \in (G - G')} \left|d_{ij} - d_{ij''}\right| \quad (2)$$

Eq. (1) limits the maximum difference between any pair of expression values in an EDG. Eq. (2) guarantees that a gene is always grouped with its closest neighbor. We call a gene satisfing Eq. (1) but not Eq. (2) a *breakpoint*.

Due to the highly noisy, considering close values as ordered is impractical in the context of microarray data analysis. An EDG encloses a group of genes with the similar expression values together. Thus, the sequences of genes in which any pair of genes are not in the same EDG is robust to noise w.r.t the group threshold δ. Moreover, this shortens the maximum size of the sequences such that the computing time is also greatly reduced.

For a sample s_i, a sliding window approach can be used to find all EDGs. First, all genes are sorted by their expression values in ascending order. Second, we slide a window from left to right. The size of every window is initially determined by Eq. (1), and then refined by Eq. (2). If a breakpoint is encountered, the next window starts from the first breakpoint. Otherwise, start from the position immediately right to the current left-end of the window.

s_1: $(g_4\ g_8\ [g_7 <g_2\ g_5)\ g_3]\ g_9>\{g_6\ g_1\}$
s_2: $(g_5\ [g_7\ g_8 <g_4)\{g_2]\|\ g_6>g_3\ g_9\}\ g_1\|$
s_3: $(g_4\ g_8\ g_1\ [g_2)\ g_9\ g_3]<g_6\ \{g_5>g_7\}$
s_4: $(g_1\ g_5\ [g_8\ g_4)<g_2\ ]g_9\ \{g_6\ g_3>\ g_7\}$

Figure 2 : g^*-sequences for the samples in TABLE 1, δ=0.5

If δ=0.5, the sequences of EDGs corresponding to every sample in TABLE 1 are shown in Figure 2, each of which is called as a g*-sequence. Specially, for a given sample s_i, the corresponded g*-sequence is denoted as $\$_i$, and the i-th *EDG* is denoted as EDG_i. Given a g*-sequence $\$_i$, $R(x, y)$ is a binary relation for a pair of genes x and y. $R(x, y)$ is TRUE if there exists an *EDG* in $\$_i$ containing both x and y. Otherwise, $R(x, y)$ is FALSE.

Definition 2: Given two g*-sequences $\$_i$ and $\$_j$, if $\forall x,y\in\$_i$, $R(x,y)$ always holds the same value for both $\$_i$ and $\$_j$, we say $\$_i$ is a subsequence of $\$_j$, denoted as $\$_i\sqsubseteq\$_j$. In particular, if $\forall x, y\in\$_i$, $R(x, y)$ is always FALSE in $\$_i$ and $\$_j$, we say $\$_i$ is a significant chain of $\$_j$. Further, $\$_i$ is closed if there is no $\$_i'$ s.t. $\forall\$_j$, $\$_i\sqsubseteq\$_i'\sqsubseteq\$_j$.

Suppose that $\$_i=(g_8<g_2g_5)g_3> g_6$ and $\$_j=(g_8g_2<g_5)g_3> g_6$. Then, for $\$_1$ in Figure 2, $\$_i\sqsubset\$_1$ but $\$_i\not\sqsubset\$_1$. Moreover, $g_8g_3g_6$ is a significant chain of $\$_1$. A significant chain ensures that there is a significant difference between the expression values of any pair of genes within it. In particular, $g8g3g6$ is a closed significant chain.

Phenotype structure

Next, we introduce how to quantify the quality of a phenotype structure based on the g^*-sequences model.

Definition 3: Suppose that m g^*-sequences $\$_i$ ($i\in[1,m]$) are partitioned into k disjoint subsets $set_1,set_2,...,set_k$. A subsequence $\$ is a signature of subset set_l ($l\in[1,k]$), iff: (1) $\forall\$_x\in set_l$, $\$\sqsubseteq\$_x$, and (2) $\forall\$_y\notin set_l$, $\$\not\sqsubset\$_y$. In particular, if $\forall\$_x\in set_l$, $\$ is a significant chain of $\$_x$, we call $\$ a p-signature of set_l.

Suppose that the four g^*-sequences are partitioned into two disjoint subsets, $set_1=\{\$_1, \$_2\}$ and $set_2=\{\$_3,\$_4\}$. According to Definition 3, $\$=g_7(g_6g_1)$ is a signature of set_1, $g7g6$ and $g7g1$ are two p-signatures of set_1.

Given a p-signature p_i and a sample s, the *projection* of p_i on s, denoted as $p_i|s$, refers to the sequence of all genes in p_i permuted according to their relative orders in $\$. If a pair of genes in p_i has a reverse relative order in $p_i|s$, we call it a *reverse pair*. Given p_i and $p_i|s$, for a gene x, if it is at the k-th locus in p_i and at the j-th locus in $p_i|s$, we call $|k–j|$ the *distortion* of x between p_i and $p_i|s$, denoted as $dist_x(p_i,s)$. For example, if $p_i=g_3g_4g_6$, and s is s_1 in Tab. 1, then $p_i|s=g_4g_3g_6$ and (g_3g_4) is a reverse pair.

Definition 4: Given a p-signature p_i and a sample s, the projection divergence of p_i and $p_i|s$, denoted as $PD(p_i, p_i|s)$, is

$$PD(p, p \mid s) = \sum_{\substack{x,y \in p \\ x \neq y}} \varphi(x,y)[dist_x(p,s) + dist_y(p,s)] \quad \textbf{(3)}$$

where $\varphi(x,y) = \begin{cases} 1, \ if \ (x,y) \text{ is a reverse pair} \\ 0 \ , \text{otherwise} \end{cases}$ **(4)**

PD takes into account the interrelationship among genes when computing the difference, as is quite different from some commonly used sequence distance metrics (e.g. edit distance *ED*[10]). Continuing the previous example where $p_i = g_3g_4g_6$ and s is s_1 in Tab. 1, since there is only one reverse pair in p_i, i.e. (g_3g_4), then $PD(p_i, p_i|s) = 1\times[1+1] = 2$.

Below is a quality measure for a candidate phenotype structure based on *PD*.

Definition 5: For a microarray dataset *D*, let $\hat{I} = \{set_1, set_2, \ldots, set_k\}$ be a partition of the *m* samples and $Ç = \{p_1, p_2, \ldots, p_k\}$ be a set of p-signatures, where p_i is a p-signature of set_i. A *phenotype structure* in *D* refers to the collection of all submatrices $\{(set_i, p_i)\}$. Its quality function is defined as follows

$$Q(\hat{I}, Ç) = \frac{1}{C_k^2} \sum_{i=1}^{k} \sum_{j=i+1}^{k} B(i, j) \quad \textbf{(5)}$$

where $B(i,j) = \dfrac{\sum_{\forall s \in set_j} PD(p_i, p_i \mid s) + \sum_{\forall s \in \$_i} PD(p_j, p_j \mid s)}{|set_i| + |set_j|}$ **(6)**

$|set_i|$(or $|set_j|$) denotes the number of samples in set_i or set_j.

Let $D_i = \{d_{x,y} | s_x \in set_i, g_y \in p_i\}$ be the projected submatrix of set_i on p_i. $B(i,j)$ evaluates the mutual difference between two submatrixes D_i and D_j. Larger $B(i,j)$ indicates larger mutual difference between D_i and D_j. Thus, $Q(\hat{I}, Ç)$ measures the average pairwise difference between submatrices.

Consider the example in TABLE 1. Suppose that the samples are partitioned into $set_1 = \{\$_1, \$_2\}$ and $set_2 = \{\$_3, \$_4\}$ with p-signatures $p_1 = g_7g_1$ and $p_2 = g_1g_6$, respectively. The corresponding $Q(\hat{I}, Ç)$ can be calculated as follows: First, $p_1|s_3 = p_1|s_3 = g_1g_7$, $p_2|s_1 = p_2|s_2 = g_6g_1$. Then, according to Definition 4, we have $\sum_{\forall s \in set_2} PD(p_1, p_1 \mid s) = 2 + 2 = 4$, $\sum_{\forall s \in set_2} (p_2, p_2 \mid s) = 0 + 0 = 0$. Since $|set_1| + |set_2| = 2$, $B(1,2) = \dfrac{4+0}{2+2} = 1$. Thus, $Q(\hat{I}, Ç) = B(1,2) = 1$.

The problem statement

Given an expression matrix *D* of *m* samples and *n* genes, and a grouping threshold δ, our goal is to find the phenotype structure with the largest quality score $Q(\hat{I}, Ç)$. To filter out the blocks with too few or too many samples, we introduce Min_s and Max_s to limit the minimum and the maximum number of samples in a submatrix.

THE FINDER ALGORITHM

FINDER consists of three major steps: (1) trivial g^*-sequences identifying; (2) phenotype structure discovery; and (3) refinement.

Trivial g^*-sequences identifying

A subsequence $\$ is *trivial* if it is common to all *m* samples. Clearly, a trivial sequence cannot be selected as a p-signature of a specific phenotype. Thus, the genes involved in the trivial subsequences can be ignored. However, it is intractable to exhaustively enumerate all trivial subsequences. The following theorem states that the search space of trivial subsequences can be dramatically reduced.

Theorem 1: The genes covered by all trivial g^*-sequences are just as that covered by all closed trivial significant chains.

Proof: Let $\$ be a trivial g^*-sequence and $x \in \$$ be a gene not covered by any significant chain of $\$. According to the sliding window method discussed in Section 1, there must be another gene $y \in \$$ such that either xy or yx form a significant chain, which contradicts the assumption. Hence the proof. ■

Theorem 2 indicates that, instead of testing all trivial g^*-sequences, we only need to consider the closed trivial significant chains, the lengths of which are usually much shorter than that of the original g^*-sequences. As a result, the search space is greatly reduced.

Phenotype structure discovery

A block (or submatrix) is the basic element of a phenotype structure, which consists of a subset of samples and the corresponding p-signature. The basic idea of the phenotype structure discovery method proposed in this paper can be described as follows. First, generate the candidate p-signatures. Then, derive the corresponding blocks from the candidate p-signatures. Finally, find the block combination of the largest $Q(\hat{I}, \hat{C})$ by testing various block combinations.

According to Definition 3, a p-signature must be a significant chain. Thus, a naive candidate p-signature generating method is to check all significant chains, which, however, is infeasible in practice. The following theorem states that the candidate p-signatures can only result from the closed significant chains.

Theorem 4: Let $(\hat{I}, \hat{C})$ and $(\hat{I}', \hat{C}')$ be two candidate phenotype structures, where $\hat{I} = \{set_1, set_2, \ldots, set_k\}$, $\hat{I}' = \{set'_1, set'_2, \ldots, set'_k\}$, $\hat{C} = \{p_1, p_2, \ldots, p_k\}$, $\hat{C}' = \{p'_1, p'_2, \ldots, p'_k\}$. If $\forall i$, i $(1 \le i \le k)$, $p_i \sqsupset p'_i$ and p'_i is closed, then $Q(\hat{I}', \hat{C}') \ge Q(\hat{I}, \hat{C})$.

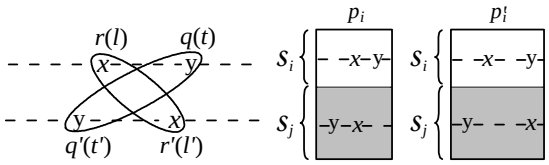


Figure 3 : $PD(p'_i, p'_i \mid s) \ge PD(p_i, p_i \mid s)$

Proof: We prove the theorem by Figure3, where the shadowed blocks are the projections of p_i and p'_i on all samples in set_j. For a sample s in set_j, the two dashed lines denote p_i and $p_i|s$ (or p'_i and $p'_i|s$), where (x, y) is a reverse pair. The position of x in p_i (resp. $p'_i|s$) is indicated by r (resp. r'), and that of y in p_i (resp. $p'_i|s$) is indicated by q (resp. q'). Similarly, the position of x in p'_i (resp. $p'_i|s$) is indicated by l (resp. l'), and that of y in p'_i (resp. $p'_i|s$) is indicated by t (resp. t'). Then, $[dist_x(p'_i, s) + dist_y(p'_i, s)] - [dist_x(p_i, s) + dist_y(p_i, s)] = (l' - l + t - t') - (r' - r + q - q') = [(l' - t') - (r' - q')] + [(t - l) - (q - r)]$. Since p_i, p'_i, $[(t - l) - (q - r)] \ge 0$. Likewise, since $p_i|s \sqsupset p'_i|s$, $[(l' - t') - (r' - q')] \ge 0$. Therefore, the preceding formula is no less than 0. Extending the conclusion to any reverse pair in p_i, we have $PD(p'_i, p'_i|s) \ge PD(p_i, p_i|s)$. Moreover, since s is any sample in set_j, we have that $\sum_{\forall s \in \$_j} PD(p'_i, p'_i \mid s) \ge \sum_{\forall s \in \$_j} PD(p_i, p_i \mid s)$. Similarly, $\sum_{\forall s \in \$_i} PD(p'_j, p'_j \mid s) \ge \sum_{\forall s \in \$_i} PD(p_j, p_j \mid s)$. Thus, $Q(\hat{I}', \hat{C}') \ge Q(\hat{I}, \hat{C})$. ■

Theorem 4 ensures that we can generate all candidate p-signatures at low cost. A phenotype structure is a combination of blocks. Thus, the next step is, for each candidate p-signature, to find a sample set of this p-signature as a candidate block, and then, select the best combination as the final phenotype structure by testing the block combinations. Clearly, it is intractable to enumerate all block combinations. In this section, we develop two heuristic methods to tackle the problem.

Aggressive Greed: This approach is inspired by the intuitive idea that the best individuals constitute the best combination. Concretely, according to the value of $\frac{\sum \forall s \in S - set_i PD(p_i, p_i \mid s)}{|S - set_i|}$, the block whose p-

signature is of the maximum average *PD* to its projections on the remaining samples is selected as the first block. The remaining blocks are selected based on the value of $\Sigma B(i,j)$, where j is the index of the block to be selected and i is the index of any block having been selected. The block with the maximum average difference on $B(i,j)$ will be selected.

In this approach, each block will be examined just once. That is, once block i is determined in step i, it will remain unchanged in the whole process. Although this approach may be of an advantage in terms of efficiency, it heavily depends on the quality of the first selected block. If a bad block is selected in the first, the remaining selections will be based on this block.

Progressive Greed: This method allows to update a previously selected block by a new block if such an update can improve the quality of the block combination. During the search, for the current sample set X, we derive the most distinctive block (set_i, p_i) from it. Then, remove set_i from the complete sample set S and search the remaining sample set $S–set_i$ to seek the next block (set_j, p_j) such that $B(i, j)$ is maximum while $s_i \cap s_j$ is minimum. This ensures to select the block with the maximum average difference and the minimum overlap with the selected blocks. The process proceeds recursively until every sample is assigned to a block. When such a block combination is obtained, it is considered as a candidate. Instead of immediately returning this candidate as the result, we track back to the sample set X and continue searching the remaining combinations containing X to generate new candidates in a similar way. During the process, we always keep track of the current best result and its quality score Q_{best}. Once a new candidate is generated, we compare its quality score, Q_c, with Q_{best}. If $Q_c > Q_{best}$, update Q_{best} to Q_c; otherwise, remain Q_{best} and the related information. Experimental results show that this method greatly improves the quality of the results due to the quality-guaranteed block updating way.

Refinement

FINDER uses Min_s as a terminal condition to stop the block combination test. A small number of samples may not be assigned to any block. Such a case can be dealt with by reassigning those samples according to certain criterion.

In this paper, we address the problem by breaking every current p-signature into some smaller fragments. Then, a sample is reassigned by combining the decisions from all fragments. The process is treated as a voting based on *PD* and the cross-projection. That is, for a sample s to be reassigned, we project the fragments of every block onto $\$_i$ and compute the average projection distance PD_{avg}. Finally, s is assigned to the block with minimum PD_{avg}. Next, a top-down recursive process is given to break a p-signature into the smaller fragments.

Suppose that p_i is a closed p-signature. We first generate all its immediate sub-patterns, p_{i1}, p_{i2}, …, p_{in}, by removing a single item from p_i, respectively. We then compare the supports of p_i and p_{ix} for all $x \in [1,n]$. If the support of p_{ix}, i.e. the number of samples containing p_{ix}, is larger than that of p_i, i.e. $supp(p_{ix}) > supp(p_i)$, we remove p_{ix} and all its immediate sub-patterns from considering. Otherwise, we recursively continue the process for p_{ix}. The patterns that can not be further reduced are left as the final fragments.

PERFORMANCE EVALUATION

In this section, we study the performance of FINDER by evaluating its efficiency and effectiveness. The algorithms are coded in C++. All experiments are conducted on a 2.0-GHz HP PC with 1G memory running Window XP. Both real and synthetic datasets are used in the experiments. The real datasets are colon tumor[9], ALL-AML leukemia[7] and Hereditary Breast Cancer (HBC)[23]. TABLE 2 shows the statistics of these three datasets. The synthetic datasets are generated by a specific data generator in [?]. Unless otherwise specified, the default parameters setting for FINDER are δ=0.3, Min_s=0.3, Max_s=0.5.

TABLE 2 : The information of three real microarray datasets

dataset	# sample	# gene	class1 : # class1	class2: # class2	class3: # class3
Colon	62	2000	negative:40	positive:22	N/A
Leukemia	38	5000	B-ALL:19	T-ALL:8	AML:11
HBC	22	3326	BRCA1:7	BRAC2:8	Sporadic:7

Efficiency

In this section, we evaluate the efficiency of FINDER by studying how response time varies with respect to *#sample* and *#gene*, where the synthetic datasets are used. Since no previous work can be directly applied to the problem setting in this paper, we implemented a naïve two-step method as the baseline method. First, all candidate *p*-signatures are mined using BIDE[12], one of the state-of-the-art closed sequence mining algorithm; Second, do an exhaustive combination test over all derived blocks. Two greedy strategies proposed in this paper are also implemented, which are called A-FINDER (aggressive approach) and P-FINDER (progressive approach), respectively.

As Figure 4 shows, the running time of the three phenotype structure discovery algorithms becomes longer as *#sample* and *#gene* increases. This is because larger *#sample* may lead to more sample combinations to be tested and the increasing of *#gene* makes the number of EDGs in every *g**-sequence larger. Note that FINDER is *two or three orders of magnitude* faster than the naive method. This confirms the efficiency of the proposed algorithm.

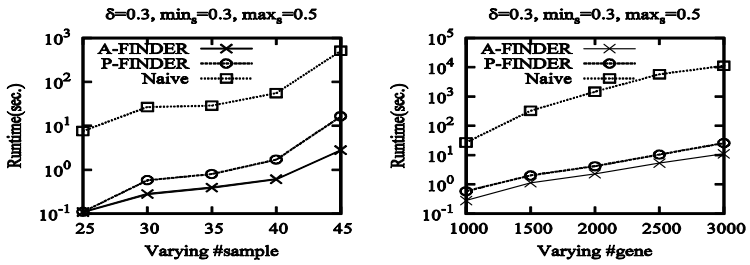


Figure 4 : Scalability

Effectiveness

In this section, we evaluate the effectiveness of FINDER in terms of statistical and biological significance. In the statistical sense, we use *p*-value. In the biological sense, we show some interesting results discovered from the Leukemia dataset, and explain them based on GENE database of NCBI.

Statistical significance

A *p*-value indicates the probability that a phenotype structure is formed by chance. We use the hypergeometric distribution to calculate the *p*-value for each block of a phenotype structure. Specifically, it is computed as follows:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{m-M}{t-i}}{\binom{m}{t}} \quad \textbf{(9)}$$

In the above equation, *m* is the total number of samples in a given dataset, and *M* is the number of samples annotated to a particular phenotype. Eq.(9) calculates the probability that seeing at least *k* samples annotated to that particular phenotype in randomly chosen *t* samples. This approach is widely used to evaluate the statistical significance of the result in many existing tools, such as Gene Ontology and GO TermFinder. A smaller *p*-value indicates a stronger statistical significance. If most of the blocks of a phenotype structure are of small *p*-values, the phenotype structure is unlikely formed by chance.

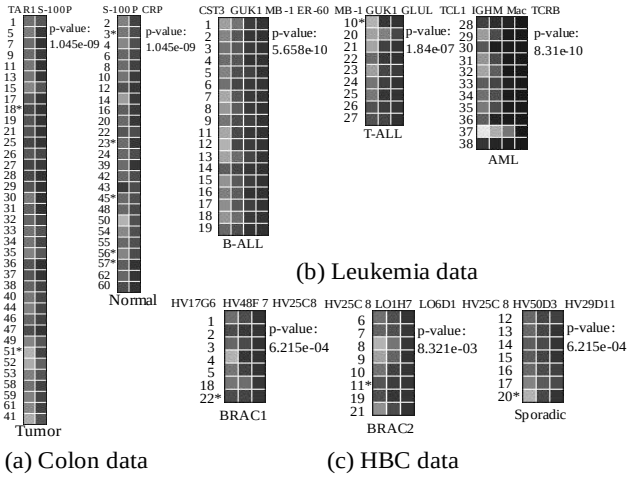


Figure 5 : The result visualization

To show the power of the ordered gene expression values in the phenotype structure discovery more clearly, we visualize the phenotype structures discovered from the three real datasets in Figure 5(a)~5(c), where the strength of gene expression is mapped into the darkness of color. The stronger the gene expresses, the darker the color is. The gene orders (p-signatures) and the sample labels are given at the top and the left of every block, respectively. '*' marks the samples not properly grouped. Clearly, in each block of a phenotype structure, the mapped expression values are always from lightness to darkness. The order among genes can be used to discover the phenotype structures of statistical significance.

Biological significance

In this section, we present some interesting results discovered by FINDER from the Leukemia dataset[7] and show that FINDER is able to find not only the genes identified by the existing methods, but also some important genes ignored by the existing methods.

TABLE 3 lists all genes involved in the phenotype structure discovered from the Leukemia dataset. If a gene is ranked within top-100 by two or more commonly used statics, it is marked with '*'. As shown in TABLE 3, genes MB-1, CST3 and MacMarcks are top-ranked genes by all eight methods. They are also discovered by FINDER. Indicated by GENE, a searchable database of genes in NCBI, MB-1 gene encodes the Ig-alpha protein of the B-cell antigen component. It is a sensitive and specific reagent for B-lineage blasts that will aid in the classification of B-cell precursor ALL and in the identification of biphenotypic leukemia presenting as AML[14]; CST3 encodes the most abundant extracellular inhibitor of cysteine proteases, which is found in high concentrations in biological fluids and is expressed in virtually all organs of the body. A mutation in this gene is associated with amyloid angiopathy (e.g. AML); MacMarcks gene is proven to be immune-related[15]. Tumor is often immune-related, thus it is biologically plausible to find MacMarcks in the phenotype structure of Leukemia. Genes IGHM and TCL1 are identified by two and five methods in TABLE 3, respectively. As GENE states, IGHM is the antigen recognition molecule of B cells; TCL1 is activated in T-cell leukemias by translocations and inversions that juxtapose it to regulatory elements of T-cell receptor genes, and activation of TCL1 in mature T-cells causes T-cell leukemia in humans[16]. Immunologic processes have been well studied by Yunji's mathematical model about the macrophage activation. A novel network model and framework are established[17-19].

For the genes without '*', extensive biological evidences indicate that these genes are also related to leukemia. For example, TCRB is ranked outside top-100 in TABLE 3. However, TCRA is reported by five methods in TABLE 3[13]. From the gene description in the Leukemia dataset[7], we know that the two are both T-cell receptors. They have very similar function. Moreover, GENE database confirms that chromosomal abnormalities involving TCRB are closely associated with T-cell lymphomas. Also, we find two other interesting cases. That is, the gene sequence <MB-1 GUK1

GLUL> identifies T-ALL phenotype with precision=88.9% and recall=100%, and the gene sequence <CST3 GUK1 MB-1 ER-60> identifies B-ALL phenotype with precision=100% and recall=94.7%. It is the order among genes, which is ignored by singleton or combination discriminability based methods, that enables FINDER to discover the statistical significant phenotype structures with higher accuracy and fewer genes. Moreover, such order may provide a possible explanation to some diseases from a new point of view. For example, due to the small p-value, it is statistically reasonable to infer that the cause of T-ALL may be that gene GLUL expresses more than gene GUK1 and gene GUK1 expresses more than gene MB-1 in an individual.

TABLE 3 : The genes discovered from Leukemia dataset

gene	RANK							
	T-test	**Information gain**	**Sum of variances**	**Twoing-rule**	**Gini-index**	**Sum minority**	**Max minority**	**ID SVM**
MB-1*	4	18	26	26	26	41	34	21
CST3*	49	4	3	3	3	2	2	4
MacMarcks*	19	38	29	29	29	21	13	27
TCL1*	42	30	61	61	61	>100	>100	>100
IGHM*	69	>100	>100	>100	>100	>100	83	>100
TCRB	>100	>100	>100	>100	>100	>100	>100	>100
GUK1	>100	>100	>100	>100	>100	>100	>100	>100
GLUL	>100	>100	>100	>100	>100	>100	>100	>100
ER-60	>100	>100	>100	>100	>100	>100	>100	>100

CONCLUSION

In this paper, we model the phenotype structure discovery problem from a sequence perspective. Different from the existing methods, the proposed g*-sequences model uses the ordered gene expression values as the discriminative signatures. It enables to find highly accurate phenotype structure with a small number of genes. Further, we develop two progressive exploring strategy to tackle the proposed problem. Extensive experimental results on real and synthetic datasets show that our method dramatically improves the accuracy of the discovered phenotype structure (in terms of statistical and biological significance). Moreover, FINDER is 2~3 orders of magnitude faster than the alternative methods.

REFERENCES

[1] Sami Hocine, Pascal Raymond, Daniel Zenklusen, Jeffrey A.Chao, Robert H.Singer; Single-molecule analysis of gene expression using two-color RNA labeling in live yeast, Nature Methods, **10**, 827-836 **(2013)**.

[2] Fred A.Wright, Patrick F.Sullivan, Andrew I.Brooks, Fei Zou et al; Heritability and genomics of gene expression in peripheral blood, Nature Genetics, **46**, 430–437 **(2014)**.

[3] D.W.Scott, G.W.Wright, P.M.Williams, C.J.Lih, W.Walsh et al; Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin embedded tissue, Blood, **123**, 1214–1217 **(2014)**.

[4] Yuhai Zhao, Guoren Wang, Xiang Zhang, Jeffrey Xu Yu, Zhanghui Wang; Learning Phenotype Structure Using Sequence Model, IEEE Trans.Knowl.Data Eng, **26(3)**, 667-681 **(2014)**.

[5] J.R.Nevins, A.Potti; Mining gene expression profiles: expression signatures as cancer phenotypes, Nature Reviews Genetics, **8(8)**, 601–609 **(2007)**.

[6] Yang Zhao, Defu Cheng Bin Su; Global Sliding Mode Control for Electro-Hydraulic System Considering Disturbance Observing Strategy, The Scientific World Journal, **2014(2014)**, **(2014)**.

[7] T.R.Golub, D.K.Slonim, P.Tamayo et al.; Molecular classification of cancer: class discovery and class prediction by gene Expression monitoring, Science, **286**, 531–537 **(1999)**.

[8] J.Luo, D.J.Duggan, Y.Chen et al; Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling. Cancer Res, **61(12)**, 4683–8 **(2001)**.

[9] U.Alon, N.Barkai, D.A.Notterman et al.; Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, PNAS, **96(12)**, 6745–6750 **(1999)**.

[10] Yuhai Zhao, Baiyou Qiao, Tianliang Lin, Guoren Wang; An Algorithm for Clustering Co-Regulated Genes based on General Similarity. Journal of Northeastern University (Natural Science), **30(11)**, 1558-1561 **(2009)**.

[11] Hedenfalk, D.Duggan, Y.Chen et al.; Gene-expression Profiles in hereditary breast cancer, New England Journal of Medicine, **344(8)**, 539–548 **(2001)**.

[12] Jianyong Wang, Jiawei Han, Chun Li; Frequent Closed Sequence Mining without Candidate Maintenance, IEEE Trans.Knowl.Data Eng., **19(8)**, 1042-1056 **(2007)**.

[13] Y.Su, T.M.Murali, V.Pavlovic, M.Schaffer, S.Kasif; Rankgene:identification of diagnostic genes based on expression data, Bioinformatics, **19(12)**, 1578–1579 **(2003)**.

[14] V.Buccheri, B.Mihaljevic; mb-1: a new marker for b-lineage lymphoblastic leukemia, Blood, **82(3)**, 853–857 **(1993)**.

[15] Yang Zhao; Study on Predictive Control for Trajectory Tracking of Robotic Manipulator, Journal of Engineering Science and Technology Review; **7(1)**, 45-51 **(2014)**.

[16] Y.Pekarsky, C.Hallas, C.M.Croce; The role of tcl1 in humant-cell leukemia, Oncogene, **20(40)**, 5638–5643 **(2001)**.

[17] Y.Wang et al.; Mathematical modeling and stability analysis of macrophage activation in left ventricular remodeling post-myocardial infarction, BMC Genomics, **13(l6)**, S21 **(2012)**.

[18] Y.Wang et al.; A conceptual cellular interaction model of left ventricular remodeling post MI: dynamic network with exit-entry competition strategy, BMC Systems Biology, **4(1)**, S5 **(2010)**.

[19] Y.Wang et al.; A conceptual cellular interaction model of left ventricular remodeling: dynamic network with exit-entry evolution strategy, The FASEB Journal, **(2010)**.