2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

An Improved Ratio-based (IRB) Batch Effects Removal Algorithm for Cancer Data in a Co-analysis Framework

Shuchu Han
Computer Science Department
Stony Brook University (SUNY)
Stony Brook, New York 11794
Email: shhan@cs.stonybrook.edu

Hong Qin
Computer Science Department
Stony Brook University (SUNY)
Stony Brook, New York 11794
Email: qin@cs.stonybrook.edu

Dantong Yu
Computational Science Center
Brookhaven National Lab
Upton, New York 11973
Email: dtyu@bnl.gov

***Abstract*—Ratio-based algorithms are proven to be effective methods for removing batch effects that exist among microarray expression data from different data sources. They are outperforming than other methods in the enhancement of cross-batch prediction, especially for cancer data sets. However, their overall power is limited by: (1) Not every batch has control samples. The original method uses all negative samples to calculate the subtrahend. (2) Microarray experimental data may not have clear labels, especially in the prediction application, the labels of test data set are unknown. In this paper, we propose an Improved Ratio-Based (IRB) method to relieve these two constraints for cross-batch prediction applications. For each batch in a single study, we select one reference sample based on the idea of aligning *probability density functions (pdfs)* of each gene in different batches. Moreover, for data sets without label information, we transfer the problem of finding reference sample to the dense subgraph problem in graph theory. Our newly-proposed IRB method is straightforward and efficient, and can be extended for integrating large volume microarray data sets. The experiments show that our method is stable and has high performance in tumor/non-tumor prediction.**

I. Introduction

In this digital era, we have been obtaining much more biological experiment data than before. Consequently, biological scientists have collected and built many genomic knowledge database by taking the advantage of today's information technology. These large database, for example, NIH GEO [1], inSilicoDb [2], and ArrayExpress [3], not only share many experiments data from different independent studies, but also provide computing tools for researchers to analyze data. The approach of integrative analyzing multiple microarray gene expression datasets is proved to be a robust way for extracting biological information from genomic datasets [4]. Comparing with "meta-analysis" [5] which combines analysis results from many small-sized independent datasets, integrative analysis shows higher statistical relevance of results from one integrated large size dataset [6]. Nevertheless, combining or merging microarray expression data from different data sources suffers from the so-called *batch effects* [7] which is still a challenging and difficult problem in computational biology research.

Batch effects are different from bias and noise. They are systematical unwanted variations existing among batches from different sources [7]. Many research works have been proposed in past decade to learn their math properties, and try to reduce its impacts in microarray data analysis. Lazar et al. [6] documented a comprehensive survey about existing batch effect removal methods. In all those methods, ratio-based methods are proved to have high prediction performance by Luo et al. [8]. Moreover, ratio-based methods have low computational cost which is demanding for integrating large volume data sets. However, ratio-based methods require each batch of data to have a group of *reference samples*, which could be either control samples or negative (non-tumor) samples.

GENESHIFT is another batch effect removal method proposed by Lazar et al. [9]. It is a nonparametric algorithm and assumes that samples in different batches are from same population, which means they will have same distributions. By this assumption, GENESHIFT reduces the batch effect by aligning the *pdfs* of each gene's expression values crossing different batches. It has the same expression value model as ratio-based methods. However, It does not have a clear mathematical definition/operation about how the batch effects are neglected or removed. In this paper, we propose an *Improved Ratio-based (IRB)* method for batch effect removal by taking the advantages of both ratio-based methods and GENESHIFT. The main contributions of our works are listed as follows:

- We show that it is better if the *pdfs* of genes are estimated from negative (non-tumor) samples instead of all samples for cancer data sets(§ IV-C).
- We propose a co-analysis framework (§ IV-D) to find reference samples for ratio-based algorithms. We define *matching score* for searching best reference samples for labeled data samples. We also propose a greedy algorithm for obtaining the local optimal solution.
- For unlabeled data samples, we convert the reference samples searching problem to the dense subgraph problem in graph theory (§ IV-D) and design a searching algorithm based on bipartite graph to solve it.
- We propose an improved ratio-based method (IRB) (§ IV-E) by using one sample in each batch as the subtrahend comparing with original method which use many samples. We also evaluate the prediction performance over two real cancer data sets.

In this paper, we represent different batch data as $X^k, k \in$

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.47

Symbol	Meaning
X^k	X: one batch; k: batch id;
X^k_{ij}	expression value of i_{th} row and j_{th} column;
$\hat{X}^k_{ij}$	expression value after batch effect removal;
b^k_{ij}	batch effect of value at (i, j) in batch k;
ϵ^k_{ij}	noise;
P_i, Q_i	*pdfs* of gene i in batch P and Q;
$G(V, E)$	graph G with vertices V and edge set E;
S	vertices of subgraph;
$e[S]$	number of edges induced by S;

TABLE I. FREQUENTLY-USED MATHEMATICAL NOTATIONS.

$\{1, \cdots, K\}$, where k is the batch ID. Each batch data has m rows and n columns. The rows represent genes(feature), and the columns represent samples. Moreover, We assume that all batches have been log-transformed and preprocessed for background correction, normalization and summarization by using either MAS5 [10], RMA [11], fRMA [12] or other preprocessing tools.

II. BACKGROUND AND RELATED WORK

Batch Effect Removal. The surveys [13], [6] give detailed comparison and analysis about existing batch effect removal algorithms. The most popular ones include, but not limited to: Batch Mean-Centering (BMC) [14],Gene Standardization [15],Ratio-based methods [8], Scaling relative to reference dataset [16], Empirical Bayes method [17], Cross-Platform Normalization (XPN) [18], Distance-Weighted Discrimination [19], Singular Value Decomposition based method [20], Surrogate Variable Analysis [21],GENESHIFT [9], Remove Unwanted Variation,2-step (RUV-2) [22] and etc. These methods can be divided into two groups: location-scale (LS) methods and matrix-factorization(MF) methods. LS methods assume a statistical model for the location (mean) and scale (variance) of the data within the batches and proceed to adjust the batches in order to agree with these methods. MF algorithms assume that the data variation corresponding to batch effects is independent to the biological variable of interest and can be captured by a small set of factors which can be estimated through certain matrix factorization methods.

Ratio-Based Methods. Ratio-based methods [8] shift the expression value of each gene based on a set of reference samples in each batch. It is designed with two versions: *Ratio-A* and *Ratio-G*. *Ratio-A* uses arithmetic mean value as subtrahend while Ratio-G uses geometric mean value. They assume that expression value of each gene in reference samples are subjected to the same amount of batch effects as in the other samples in same batch. Then the batch effects can be removed by subtracting the mean of those reference samples. Assuming that there are r reference samples in batch X^k, method *Ratio-A* and *Ratio-G* can be described as:
Ratio-A: Arithmetic mean ratio-based method:

$$\hat{x}^k_{ij} = x^k_{ij} - \frac{1}{r}\sum_{l=1}^{r} x^k_{il}; \tag{1}$$

Ratio-G: Geometric mean ratio-based method:

$$\hat{x}^k_{ij} = x^k_{ij} - \left(\prod_{l=1}^{r} x^k_{il}\right)^{1/r}. \tag{2}$$

GENESHIFT. GENESHIFT is a high quality nonparametric method. It first estimates genewise *pdfs* for each batch using the Parzen-Rosenblatt density estimation method [23]. Secondly, it estimates the offset term by finding the best match between two *pdfs*. This algorithm processes two batch data at one time. Assume P_i and Q_i are the *pdfs* of gene i in studies of batch X and batch Y. The algorithm put P_i as being fixed, and slides Q_i step by step across the range where P_i is estimated. In each step, the algorithm computes the inner product between P_i and part of Q_i, which lays in the range where the densities are estimated as follows:

$$M(t) = P_i * Q_i = \sum_{j=1}^{d} P_i(j) W^t_{Q_i(j)}, \tag{3}$$

where d is number of sampling ticks of *pdf* and $W^t_{Q_i(j)}$ is given by:

$$W^t_{Q_i(j)} = \begin{cases} \omega Q^t_i, & \text{for } Q^t_i \text{ in window} \\ 0, & \text{otherwise,} \end{cases}$$

with $\omega = 1$ a rectangular window defined on the support of P_i and Q^t_i is part of Q_i found in the *pdfs* estimation range at step t. The best matching between P_i and Q_i is given by $max(M)$ and the offset term is obtained by subtracting from the initial position of $Q_i(b_{ref})$, the best matching position $(b_{max(M)})$ is:

$$\delta = b_{ref} - b_{max(M)}.$$

By setting the reference position to 0, the offset term becomes $\delta = -b_{max(M)}$.

Dense Subgraph. Dense subgraph extraction is a classic problem in Graph theory [24]. The algorithms of solving this problem have been applied to biological networks research [25] [26] [27]. Here, we want to extract a densest subgraph from defined bipartite graph. We wish the extracted subgraph has high quality and concise. To archive this goal, we apply the latest technique described in [28] to extract the *optimal quasi-clique* which is a high quality dense subgraph.

Given a graph $G(V, E)$, find a subset of vertices $S^* \subseteq V$ such that $f_\alpha(S^*) = e[S] - \alpha\binom{|S|}{2} \leq f_\alpha(S)$ for all $S \subseteq V$. The resulted set S^* is called *optimal quasi-clique* of G. We use the recommend value $\alpha = 1/3$ in this paper.

III. DATA

We use two real world cancer data sets to validate our proposed algorithms.

a) Lung Cancer Dataset.: The lung cancer dataset comprises three data sets hybridized on two different Affymetrix platforms. The first lung cancer data set (GSE19804) contains 120 samples of tumor and adjacent normal tissue samples hybridized on Affymetrix HGU133plus2 expression arrays. The second data set (GSE19188) contains 94 tumor and 62 adjacent normal tissue samples hybridized on Affymetrix HGU133plus2 expression arrays. The third lung cancer data

set (GSE10072) contains 58 tumor samples and 49 normal tissues samples consists of a mix of independent controls and tumor adjacent tissues hybridized on Affymetrix HGU133A expression array.

Type	Name	NT	T	Platform
Train	GSE19804	60	60	GPL570
	GSE19188	62	94	GPL570
Test	GSE10072	49	58	GPL96

TABLE II. LUNG CANCER DATASET. **NT**: NON-TUMOR; **T**: LUNG TUMOR;

b) Iconix Dataset.: We use the Iconix dataset (GSE24417) from Microarray Quality Control Phase II(MAQC-II) microarray gene expression data ([8].) The Iconix dataset is a toxicogenomic data set provide by Iconix Bioscience (Mountain View,CA,USA). It aims at evaluating hepatic tumor induction by non-genotoxic chemicals after short-time exposure. The training set consists of 216 samples treated for 5 days with one of 76 structurally and mechanistically diverse non-genotoxic hepatocarcinogens and non-hepatocarcinogens. The test set consists of 201 samples treated for 5 days with one of 68 structurally and mechanistically diverse non-genotoxic hepatocarcinogens. Gene expression data were profiled using the GE Codelink microarray platform. The separation of the training set and the test set was based on the time when the microarray data were collected, also the different batches. The detail data set information is listed as follows.

Type	Batch	NT	T	Date
Train	B1	17	24	11/6/01-12/10/01
	B2	87	17	12/11/01-02/25/02
	B3	39	32	3/20/02-7/18/02
Test	B4	91	18	07/22/02-12/4/02
	B5	53	39	4/3/03-9/28/04

TABLE III. INFORMATION OF THE ICONIX DATASET; **NT**: NON-TUMOR; **T**: TUMOR;

IV. ALGORITHM

In this section, we are presenting the Improved Ratio-based (IRB) method. Comparing to the original ratio-based method, we solve the problem of finding reference samples. Instead of finding reference samples in each batch separately, IRB selects reference samples by taking all batches into consideration at the same time. The outline of this section is as follows. First, the expression value model of microarray data sets is described. Second, we define the reference samples searching problem formally. Third, we describe the assumption used in our method. Finally, we introduce a co-analysis framework for finding reference samples in labeled and unlabeled data sets separately.

A. Expression Value Model

In general, batch effect comes with multiplicative and additive form. After log-transform, these batch effects are both represented as additive terms. We assume that the expression value of feature i in sample j of batch X^k can be expressed in the following general form:

$$x_{ij}^k = x_{ij}^{'} + b_{ij}^k + \epsilon_{ij}^k, \tag{4}$$

where $x_{ij}^{'}$ is the actual feature value. b_{ij}^k is the batch effect term and ϵ_{ij}^k represents noise.

Moreover, we use the same genewise density estimation method as GENESHIFT algorithm which is Parzen-Rosenblatt density estimation method[23].

B. Problem Definition

As mentioned in the abstract, we only want to find one reference sample for each batch. The searching guideline is following the principle of GENESHIFT algorithm: *the inner product of each gene's pdf in different batches are maximized after integration*. Before giving the formal definition of our problem, we first define the matching score of two batches:

Definition 1: Given two batches that have same number m of genes(or features), and with *pdfs* $\mathbf{P}$ and *pdfs* $\mathbf{Q}$ respectively. Let $\mathbf{P}_i$ ($\mathbf{Q}_i$)is the *pdf* of i_{th} gene, the matching score of them is defined as:

$$M(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{m} < \mathbf{P}_i, \mathbf{Q}_i > . \tag{5}$$

Now, our problem can be defined formally as:

Problem 1: Given K batches of microarray expression dataset $X^k : m \times n(k)$,m genes,$n(\cdot)$ samples,$k \in \{1, \cdots, K\}$, with estimated *pdfs*:

$$\vec{\mathbf{P}} = [\mathbf{P}^1(\vec{x}), \mathbf{P}^2(\vec{x}), \cdots, \mathbf{P}^K(\vec{x})],$$

where $\mathbf{P}^k$ is the vector of *pdfs* for genes $\vec{x}$ in batch X^k. $\mathbf{P}^k$ is a $m \times 1$ vector where each $P_i^k, i \in \{1, \cdots, m\}$ represents the *pdf* of i-th gene.The problem is to find K offset samples x_{offset}^k within each batch respectively:

$$\vec{\mathbf{x}}_{offset} = [\vec{x}_{offset}^1, \vec{x}_{offset}^2, \cdots, \vec{x}_{offset}^K],$$

such that the total matching score of *pdfs* after shifting by its offset samples respectively archives maximum:

$$\underset{\vec{\mathbf{x}}_{offset}}{\operatorname{argmax}} \sum_{i=1}^{K} \sum_{j \neq i, j=1}^{K} M(\mathbf{P}^i(\vec{x} - \vec{x}_{offset}), \mathbf{P}^j(\vec{x} - \vec{x}_{offset})), \tag{6}$$

In the above problem, $\mathbf{x}_{offset}^k$ is a specific sample in batch k. If we don't limit $\mathbf{x}_{offset}^k$ to be a specific sample in the batch and let it be a regular offset vector, the problem 1 can be seen as a generalized version of GENESHIFT which takes two batches at the same time and shift *pdfs* of every gene separately from one batch to another batch. The reason we put this constrain here is that the batch effect term b^k in equation (4) can be neglected by subtracting a sample, and this sample inherits the batch effect term with its true signal value. The advantage of applying this constrain is that we obtain a clear math explanation about how the batch effects are removed.

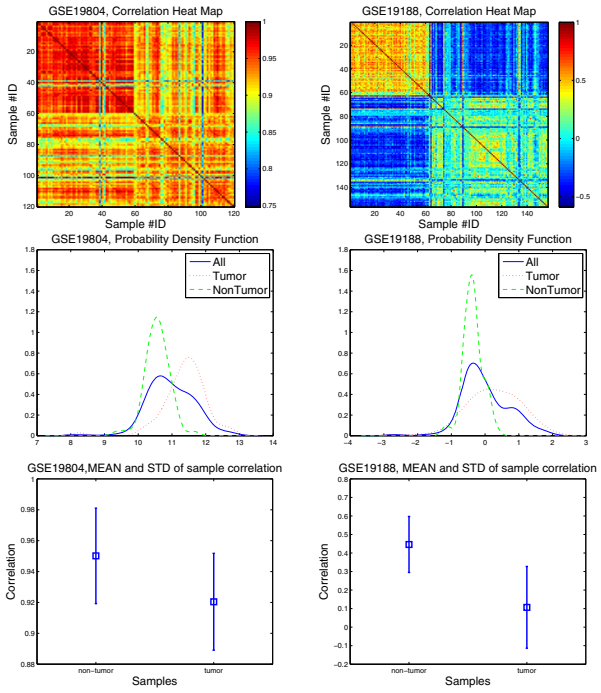

Fig. 1. Left: GSE19804; Right: GSE19188; Top row: correlation(PCC) heat map, samples are sorted from non-tumor to tumor samples; Middle row: *pdf* of a random gene(GeneBank ID:U48705). Bottom row: correlation values distribution

C. Assumption

In GENESHIFT, the authors assume that the expression of each gene from two different experiments (batches) can be represented accurately enough through the expression of that gene across all population if the number of samples in two microarray Gene Expression (MAGE) experiments is sufficiently high. By this assumption, a consequence conclusion is that the *pdfs* of each gene should be similar in all experiments. However, as we observed from above cancer data sets, the average similarity among the non-tumor(negative) samples is higher than the tumor(positive) samples, as show by figure 1. We then argue that the similarity *pdfs* assumption of GENESHIFT holds for cancer data sets only if the *pdfs* are estimated from non-tumor samples but not from all. This argument is not only based on the observation but also based on the fact that tumors with similar histopathological appearance can follow significantly different clinical courses [29]. The assumption of IRB now can be described as following:

Assumption 1: The *pdf* of a gene have similar distribution in all experiments iif the *pdf* is estimated from non-tumor samples.

D. Co-analysis Framework

In this section, we propose a co-analysis framework to find the reference samples both for labeled and unlabeled data samples. For all ratio-based methods, we need reference samples to calculate the subtrahend. Original ratio-based methods use average of all negative samples or median of them. As for our method, we only use one reference sample for each batch. Comparing to the original ratio-based methods that find reference sample independently, we take all batches into consideration at the same time. Our co-analysis framework can be described as following from labeled data sets to unlabeled data sets.

c) Labeled Data Sets.: For example, the training data sets have clear labels of samples. To find the reference samples for them, we need to solve the optimization problem (1). However, the properties like convexity or non-convexity of objective function in problem (1) are uncertain. Because (1)the objection function cumulates all matching scores of genes that show very different *pdfs*; (2)the *pdf* curve could be either convex or non-convex.

To solve this problem, we propose a greedy algorithm as described in algorithm 1. Our algorithm first selects an *anchor* batch that has the largest number of non-tumor samples and shifts its geometric median to axis origin. Secondly, for the rest of batches, we calculate the best offset vector for each of them according to this *anchor* batch. In the last step, we search a sample from each batch that has the smallest Euclidean distance to this offset vector and treat it as the reference sample that we are looking for. At the first step, we shift the geometric

Algorithm 1: FindingReferenceSamples

input : Microarray experiments data:
$X : \{X^k : m \times n\}, k \in \{1, \cdots, K\}$ with labels.
output: Reference samples:
$\vec{\mathbf{x}}_{offset} = [\vec{x}^1_{offset}, \vec{x}^2_{offset}, \cdots, \vec{x}^K_{offset}]$.

begin
  Find anchor batch x^{anchor};
  Shift x^{anchor} by GM_{approx};
  for *Batch* $X^k, k \neq anchor$ **do**
    for *Each gene* $g_i, i \in \{1, \cdots, m\}$ **do**
      Estimates the *pdf* across batches: pdf^k_i;
      Calculate the offset term δ^k_i;
    end
    Find the closest sample $\hat{x}^k_{offset}$ to δ^k;
  end
  Return $\vec{\mathbf{x}}_{offset} = [\vec{x}^1_{offset}, \vec{x}^2_{offset}, \cdots, \vec{x}^K_{offset}]$.
end

median of *anchor* batch to axis origin in order to place the median of *pdfs* of all genes around the axis origin as close as possible. However, the geometric median is not only difficult to compute but also not necessary to be an experiment sample that inherits batch effect. To solve this dilemma, we choose the sample that nearest to the geometric median as a substitute. We call this sample *approximate geometric median (GM) sample*: GM_{approx}. and the definition is as:

$$GM_{approx} = \underset{y \in X}{\operatorname{argmin}} \sum_{x_j \in X \setminus y} \|x_j - y\|_2, \tag{7}$$

where the parameter δ controls the width of neighborhoods. Our greedy algorithm now can be described as Algorithm 1.

d) Unlabeled Data Sets.: For these data sets, the tumor/non-tumor labels are unknown but the batch labels are clear. We estimate the non-tumor samples of a unlabeled batch by using dense subgraph extraction algorithms. We first build a bipartite similarity graph between the known non-tumor samples and all unlabeled samples. The Pearson Correlation Coefficient (PCC)

metric, represented as $sim(\cdot)$, is used. After that, we extract a dense subgraph, called *optimal quasi-clique*, from the built graph. The nodes of the resulted subgraph that belong to the unlabeled side are treated as non-tumor samples. The algorithm of building the bipartite graph is described by algorithm 2.

Algorithm 2: BuildBipartiteGraph

input : Non-tumor samples: L, unlabeled samples: R, User specified threshold θ
output: A unweighted undirected bipartite graph $G(V, E)$, where $L, R \subseteq V$.

```
begin
    Calculate the similarity sim(l, r),where l ∈ L, r ∈ R;
    for Each pair (l, r) do
        if sim(l, r) ≥ θ then
            Add one edge to E for nodes pair (l, r);
        end
    end
    Remove the nodes with zero degree;
    Return G(V, E);
end
```

The user-specific value θ will affect the output of our algorithm as the input is a completed weighted graph. In our experiments, we use the value that equals to half of the highest similarity value.

We use the *GreedyOQC* algorithm introduced in [28] to extract the *optimal quasi-clique*. An illustration of the algorithm output is as following:

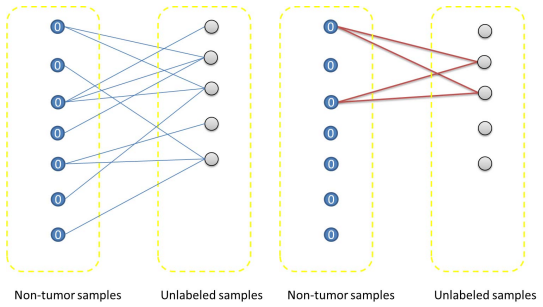

Fig. 2. Left: Input bipartite graph; Right: extracted *optimal quasi-clique*; Blue nodes: known non-tumor samples, Gray nodes: unlabeled samples.

E. Improved Ratio-based Method

Once we have reference sample for each batch, it's straightforward to modify the original ratio-based method and obtain our proposed IRB method as following:

$$\hat{x}_{ij}^k = x_{ij}^k - x(i)_{offset}. \tag{8}$$

The overall **IRB** algorithm can be described by algorithm 3.

V. Validation

In this section, we demonstrate and validate our proposed co-analysis framework by using the Lung cancer dataset. Results

Algorithm 3: IRB

input : Labeled data sets: $X : \{X^k\}, k \in \{1, \cdots, K\}$ with labels;
Unlabeled data set: Y;
output: Data sets with batch effect removed: $\hat{X}$ and $\hat{Y}$;

```
begin
    FindingReferenceSample(X), obtain x⃗_offset;
    Shift all X by x⃗_offset, obtain X̂;
    BuildBipartiteGraph(X,Y), and extract optimal quasi-clique;
    Estimate the offset of Y;
    Find reference sample y⃗_offset;
    Shift Y by y⃗_offset ;
    Return shifted X̂ and Ŷ;
end
```

of each step are presented here to better show the details of our proposed algorithm.

For Lung cancer dataset, we have three batches from two different gene chip platforms. The batch GSE19188 is selected as *anchor batch* since it has the largest number of non-tumor samples. The approximate geometric median sample is GSM475732. The difference of *pdfs* before and after shifting (applying IRB method) shows as figure 4.

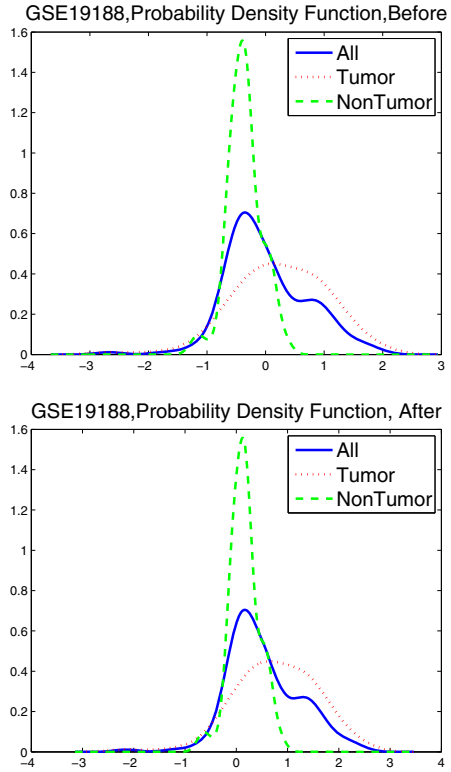

Fig. 4. Difference of gene U48705 *pdf* before (top) and after (bottom) applying IRB by reference sample GSM475732.

Now we calculate the reference sample for second batch

Fig. 3. Resulted *optimal quasi-clique* of Lung Cancer dataset. $G = (|V| = 35, |E| = 287)$. The top two row lists the estimated(fake) non-tumor samples found by *GreedyOQC*.

GSE19804 according to *anchor batch* and the changing of *pdf* is as figure 5.

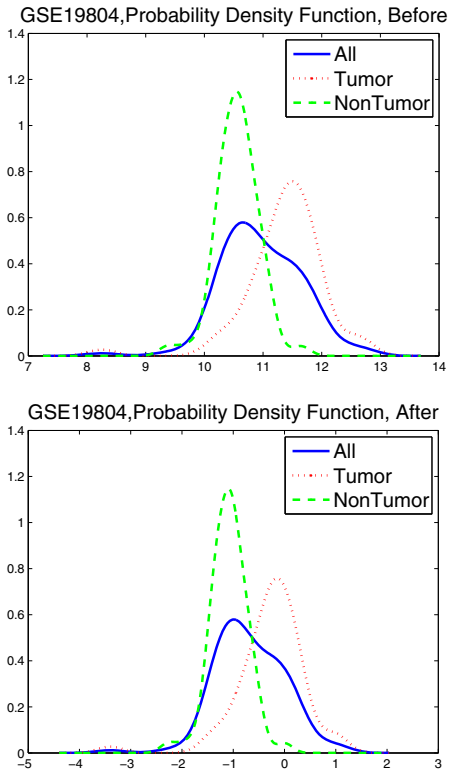

Fig. 5. *pdf* difference of gene U48705. *pdfs* before (top) and after (bottom) applying IRB.The value offset is -10.4113.

For test data GSE10072, we build the bipartite graph and find the resulted *optimal quasi-clique* as figure V. The constructed bipartite graph has 173 nodes and 747 edges. The output *optimal quasi-clique* shows as figure V and it has 35 nodes and 287 edges. Among them, 18 nodes are samples of GSE10072 and the real labels of them are non-tumor samples. The changes of *pdfs* of GSE10072 is as figure 6.

To check the quality of batch effect removal, we show the correlation heat map and clustering dehendragraph here. As we can see, the correlation values among different batches are enhanced and more smooth. The correlation heat map before and after batch effect removal is:

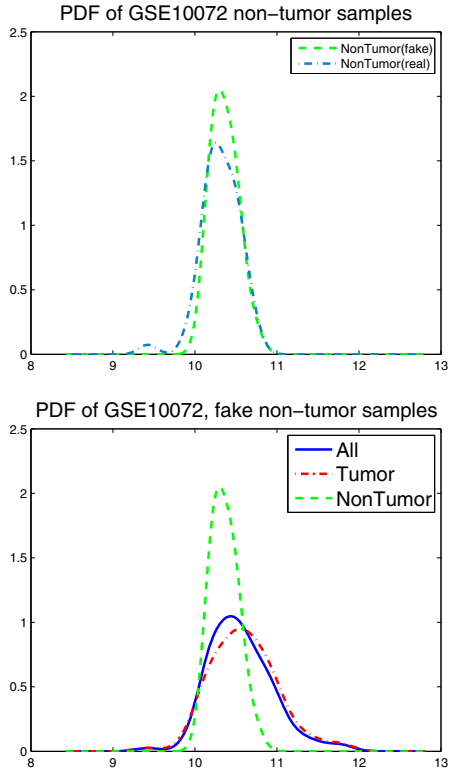

Fig. 6. *pdf* of GSE10072 by estimated(fake) non-tumor samples

VI. Results

In this section, we examine the prediction performance of our proposed algorithm comparing to original ratio-based methods and GENESHIFT. We use *Support Vector Machine*(SVM) algorithm with penalty $C = 1$, which is the setting in [8] except that we omit feature selection here. Accuracy and *Matthews correlation coefficient(MCC)* are used for our measurements.

The prediction performance of Lung Cancer data is summarized by following table: As the results show, GENESHIFT has the best prediction accuracy but Ratio-A and IRB have the better MCC scores.

Also, we compare the prediction performance of Iconix data set in table VI. The results show that IRB obtain the best accuracy and MCC scores.

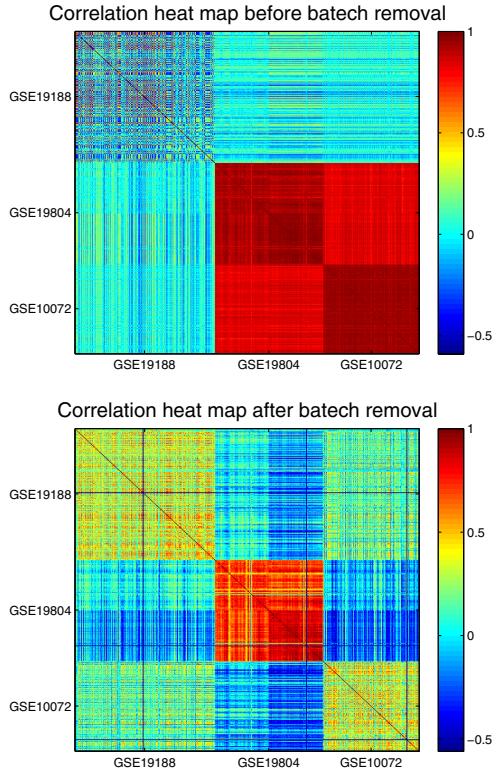

Fig. 7. Correlation heat map of Lung Cancer data. Top: original data. Bottom: after batch effect removal by IRB.

Classifier	Method	Accuracy	MCC
SVM(C=1)	Ratio-G	0.45	Nan
SVM(C=1)	Ratio-A	0.9629	**0.9813**
SVM(C=1)	GENESHIFT	**0.9723**	0.9803
SVM(C=1)	IRB	0.9623	**0.9813**

TABLE IV. PREDICTION PERFORMANCE OF LUNG CANCER DATASET

Classifier	Method	Accuracy	MCC
SVM(C=1)	Ratio-G	0.72	0.1
SVM(C=1)	Ratio-A	0.71	0.01
SVM(C=1)	GENESHIFT	0.68	0.04
SVM(C=1)	IRB	**0.73**	**0.15**

TABLE V. PREDICTION PERFORMANCE OF ICONIX DATASET

By above two experiment results, we can see that IRB method always has higher prediction performance than others. This means that IRB is a stable batch effect removal algorithm.

VII. DISCUSSION AND CONCLUSIONS

Batch effects removal has been a challenging research problem in computational biology while integrating large volume microarray data sets. Batch effects have the characteristics that neither have a clear mathematical description nor have an unique way to evaluate the performance of batch effects removal. In this paper, we propose a new algorithm to generalize the idea of GENESHIFT, which is the latest batch effect removal algorithm and a non-parametric method.

Our contribution is two-fold. First, we have solved the problem of finding reference samples for ratio-based methods from labeled data sets to unlabeled sets. The proposed co-analysis framework aligns the density function of non-tumor samples of each batch as much as possible. Comparing with the original ratio-based method which processes the batch effect less adequately, our framework takes all batches into consideration at the same time. Moreover, we applied the latest algorithm for dense subgraph problem from graph theory to solve the problem of finding reference samples for unlabeled data sets. The motivation of using the graph algorithm is that the non-tumor samples are much more similar to each other than tumor samples.

Second, our algorithm has the advantage of lowering the computational cost of both ratio-based method and GENESHIFT method. Comparing with several other batch effect removal methods, this property is valuable while integrating large volume of microarray datasets. The *GreedyOQC* has complexity $\mathrm{O}\big(|V|+|E|\big)$ for graph $G(V,E)$.

In summary, the newly-devised IRB algorithm solves the reference sample finding problem of the original ratio-based method. It inherits the characteristic of GENESHIFT that has little negative impact on the data distortion (only on samples). As a non-parametric method, it is stable and has high performance in prediction applications for cancer data sets. It has low computational cost and can be easy adapted to large volume data applications.

ACKNOWLEDGMENT

we gratefully thank all the anonymous reviewers for their constructive suggestions towards this paper's improvement. This research is supported in part by NSF grants IIS-0949467, IIS-1047715, and IIS-1049448. It is also supported by United States Department of Energy, Contract No. DE-AC02-98CH10886 to Brookhaven National Laboratory.

REFERENCES

[1] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository," *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 2002.

[2] J. Taminau, D. Steenhoff, A. Coletta, S. Meganck, C. Lazar, V. de Schaetzen, R. Duque, C. Molter, H. Bersini, A. Nowé *et al.*, "insilicodb: an r/bioconductor package for accessing human affymetrix expert-curated datasets from geo," *Bioinformatics*, vol. 27, no. 22, pp. 3204–3205, 2011.

[3] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway *et al.*, "Arrayexpress updatean archive of microarray and high-throughput sequencing-based functional genomics experiments," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D1002–D1004, 2011.

[4] D. R. Rhodes and A. M. Chinnaiyan, "Integrative analysis of the cancer transcriptome," *Nature genetics*, vol. 37, pp. S31–S37, 2005.

[5] S. Ma, "Integrative analysis of cancer genomic data," 2009.

[6] C. Lazar, S. Meganck, J. Taminau, D. Steenhoff, A. Coletta, C. Molter, D. Y. Weiss-Solís, R. Duque, H. Bersini, and A. Nowé, "Batch effect removal methods for microarray gene expression data integration: a survey," *Briefings in bioinformatics*, vol. 14, no. 4, pp. 469–490, 2013.

[7] A. Scherer, *Batch effects and noise in microarray experiments: sources and solutions*. John Wiley & Sons, 2009, vol. 868.

[8] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong *et al.*, "A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data," *The pharmacogenomics journal*, vol. 10, no. 4, pp. 278–291, 2010.

[9] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, D. Y. Weiss Solis, C. Molter, R. Duque, H. Bersini, and A. Nowé, "Geneshift: A nonparametric approach for integrating microarray gene expression data based on the inner product as a distance measure between the distributions of genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 10, no. 2, pp. 383–392, 2013.

[10] E. Hubbell, W.-M. Liu, and R. Mei, "Robust estimators for expression analysis," *Bioinformatics*, vol. 18, no. 12, pp. 1585–1592, 2002.

[11] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[12] M. N. McCall, B. M. Bolstad, and R. A. Irizarry, "Frozen robust multiarray analysis (frma)," *Biostatistics*, vol. 11, no. 2, pp. 242–253, 2010.

[13] C. Chen, K. Grennan, J. Badner, D. Zhang, E. Gershon, L. Jin, and C. Liu, "Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods," *PloS one*, vol. 6, no. 2, p. e17238, 2011.

[14] A. H. Sims, G. J. Smethurst, Y. Hey, M. J. Okoniewski, S. D. Pepper, A. Howell, C. J. Miller, and R. B. Clarke, "The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis," *BMC medical genomics*, vol. 1, no. 1, p. 42, 2008.

[15] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proceedings of the National Academy of Sciences*, vol. 98, no. 1, pp. 31–36, 2001.

[16] K.-Y. Kim, S. H. Kim, D. H. Ki, J. Jeong, H. J. Jeong, H.-C. Jeung, H. C. Chung, and S. Y. Rha, "An attempt for combining microarray data sets by adjusting gene expressions," *Cancer Research and Treatment*, vol. 39, no. 2, pp. 74–81, 2007.

[17] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[18] A. A. Shabalin, H. Tjelmeland, C. Fan, C. M. Perou, and A. B. Nobel, "Merging two gene-expression studies via cross-platform normalization," *Bioinformatics*, vol. 24, no. 9, pp. 1154–1160, 2008.

[19] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, and J. S. Marron, "Adjustment of systematic microarray data biases," *Bioinformatics*, vol. 20, no. 1, pp. 105–114, 2004.

[20] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10 101–10 106, 2000.

[21] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS genetics*, vol. 3, no. 9, p. e161, 2007.

[22] J. A. Gagnon-Bartsch and T. P. Speed, "Using control genes to correct for unwanted variation in microarray data," *Biostatistics*, vol. 13, no. 3, pp. 539–552, 2012.

[23] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, pp. 1065–1076, 1962.

[24] V. E. Lee, N. Ruan, R. Jin, and C. Aggarwal, "A survey of algorithms for dense subgraph discovery," pp. 303–336, 2010.

[25] B. H. Junker and F. Schreiber, *Analysis of biological networks*. John Wiley & Sons, 2008, vol. 2.

[26] H. Hu, X. Yan, Y. Huang, J. Han, and X. J. Zhou, "Mining coherent dense subgraphs across massive biological networks for functional discovery," *Bioinformatics*, vol. 21, no. suppl 1, pp. i213–i221, 2005.

[27] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang *et al.*, "Topological structure analysis of the protein–protein interaction network in budding yeast," *Nucleic acids research*, vol. 31, no. 9, pp. 2443–2450, 2003.

[28] C. Tsourakakis, F. Bonchi, A. Gionis, F. Gullo, and M. Tsiarli, "Denser than the densest subgraph: extracting optimal quasi-cliques with quality guarantees," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*. ACM, 2013, pp. 104–112.

[29] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.