Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 Keller Hall
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 12-002

A pattern mining based integrative framework for biomarker discovery

Sanjoy Dey, Gowtham Atluri, Michael Steinbach, Angus Macdonald,
Kelvin Lim, and Vipin Kumar

February 10, 2012

A pattern mining based integrative framework for biomarker discovery *

Sanjoy Dey
Dept. of Computer Science
University of Minnesota
sanjoy@cs.umn.edu

Michael Steinbach
Dept. of Computer Science
University of Minnesota
steinbach@cs.umn.edu

Vipin Kumar
Dept. of Computer Science
University of Minnesota
kumar@cs.umn.edu

ABSTRACT

Recent advancement in high throughput data collection technologies has resulted in the availability of diverse biomedical datasets that capture complementary information pertaining to the biological processes in an organism. Biomarkers that are discovered by integrating these datasets obtained from a case-control studies have the potential to elucidate the biological mechanisms behind complex human diseases. In this paper we define an interaction-type integrative biomarker as one whose features together can explain the disease, but not individually. In this paper, we propose a *pa*ttern *m*ining based *in*tegrative framework (PAMIN) to discover an interaction-type integrative biomarkers from diverse case control datasets. PAMIN first finds patterns form individual datasets to capture the available information separately and then combines these patterns to find *integrated patterns* (IPs) consisting of variables from multiple datasets. We further use several interestingness measures to characterize the IPs into specific categories. Using synthetic data we compare the IPs found using our approach with those of CCA and discriminative-CCA (dCCA). Our results indicate that PAMIN can discover interaction type patterns that competing approaches like CCA and discriminative-CCA cannot find. Using real datasets we also show that PAMIN discovers a large number of statistically significant IPs than the competing approaches.

1. INTRODUCTION.

Recent advancement in high throughput data collection technologies in bioinformatics has resulted in a dramatic increase in the availability of diverse data sets that capture different perspectives of a biological system pertaining to an organism [28, 13]. These types of data include, but are not limited to, DNA microarrays and RNA seq providing cell process information, Single Nucleotide Polymorphisms (SNPs) representing genetic variations, metabolomics data in terms of proteins and other metabolites, and structural and functional brain data from magnetic resonance imaging (MRI). However, each of these datasets only provide information about a part of the complex biological mechanism [26, 30]. Consolidating the information available in independent data sets has created a real possibility of personalized medicine, i.e., using detailed genetic, genomic, clinical, and environmental information about a person as *biomarkers* for a customized and more effective approach to patient care [20, 29, 4]. Such integrative biomarker discovery requires identifying those features of the data that can distinguish between healthy or low risk subjects (controls) and diseased or high risk subjects (cases) [31, 19, 7].

Integration of multiple datasets for biomarker discovery techniques can be broadly classified into two classes: 1) Predictive models ([11, 22, 34, 12], [18] provides a good survey on several kernel fusion methods) and 2) Feature extraction based biomarker discovery techniques [25, 9, 8]. The goal of the predictive model based approaches is to build classification models with high accuracy, but often such techniques do not yield easily interpretable results. In contrast, biomarkers (that are constructed using a small number of features) can be directly useful in diagnosis, treatment or prevention, but equally as important; they can also provide insights into the underlying nature of the disease or related biomedical processes. Hence we focus only on such techniques in this paper. Feature extraction based techniques including several blind source separation techniques like independent component analysis (ICA) [25] and canonical correlation analysis (CCA) [8] have been developed to find the relationship between variables across the datasets directly. In general, these models look for components in each of the available dataset such that those components have some relationships across multiple datasets. The original ICA framework, which cannot combine multiple dataset directly, has been extended in several ways for integration purpose. Examples include joint ICA [5], parallel ICA [25], group ICA [6]. Each of these has its own model assumptions [8]. On the other hand, canonical correlation analysis(CCA) and its extensions [33, 24] provide a more natural framework for integration where the relationship between different components found from multiple datasets is defined as the inter-subject variabilities. It has been also shown that CCA has less model assumptions than ICA based techniques. Therefore, CCA has been very popular for integrating datasets both integrating multiple neuroscience datasets [9] and also

*(Does NOT produce the permission block, copyright information nor page numbering). For use with ACM_PROC_ARTICLE-SP.CLS. Supported by ACM.

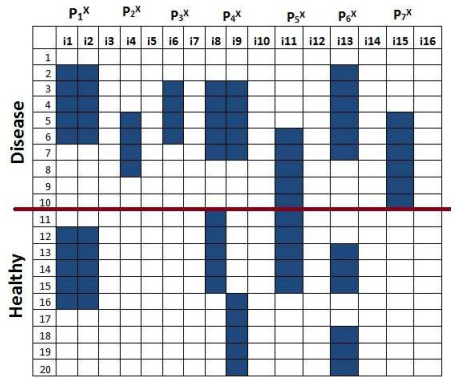
several biological datasets [23]. Multi-set CCA, a generalized CCA which can integrate more than two datasets, has also been applied for integrating fMRI, EEG and structural images recently [10]. Note that these approaches can only find biomarkers whose individual features are discriminative and correlated. However, as we will show in this paper, these techniques are unable to find integrative biomarkers that consist of features that are not discriminative individually or correlated but together they distinguish between the healthy and disease groups. Such biomarkers, referred in the rest of the paper as *interaction-type* integrative biomarkers, are important due to their ability to combine complimentary information from different data sets.

In this paper, we propose a *pa*ttern *m*ining based *in*tegrative framework (PAMIN) to discover interaction type integrative biomarkers from diverse case control datasets. PAMIN first finds patterns form individual datasets to capture the available information separately and then combines these patterns to find *integrated patterns* (IPs) consisting of variables from multiple datasets. We further use several interestingness measures to characterize the IPs into specific categories that are discussed in Section 2. Using synthetic data we compare the IPs found using our approach with those of CCA [9] and discriminative-CCA (dCCA) [33]. Our results indicate that PAMIN can discover interaction type patterns that competing approaches like CCA and discriminative-CCA cannot find. Using real datasets we also show that PAMIN discovers a large number of statistically significant IPs than the competing approaches.

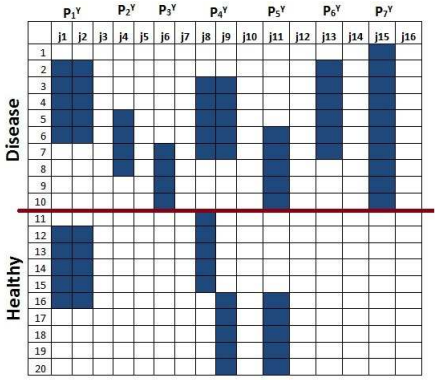
The rest of the paper is organized as follows: Section 2 presents a toy example illustrating the goal of the proposed framework. In section 3, we present our proposed framework. Evaluation and results are presented in section 4. Finally, we conclude with section 5.

2. THE OVERALL GOAL OF THE FRAMEWORK.

In this section, we will define the types of integrated patterns that are relevant to the biomarker discovery problem. Consider two binary matrices X and Y, representing two case-control datasets as shown in figure 1(a) and figure 1(b), respectively. Each of the two datasets has 15 features(represented by columns) and 20 subjects(represented by rows) with equal representations from healthy and diseased groups (separated by a horizontal line). A shaded cell in these data matrices indicates the presence of a feature (i.e., has a value 1) for a corresponding subject and a white cell indicates that a feature has a value 0 for a corresponding subject. We define a pattern as a combination of features including singletons that are associated with a particular set of samples. X and Y have a set of 7 patterns represented as $\{P_i^X\}_{i=1}^7$ and $\{P_i^Y\}_{i=1}^7$, respectively. Notice that some of these individual patterns are over-represented in one group, but not in other (e.g., P_1^X and P_1^Y) and some are equally represented (same number of samples) in both groups (P_3^X and P_3^Y). Among different existing interestingness measures, we use *diffsup* to measure the discrimination power of a pattern, which is defined as the difference between number of samples supported by a pattern in two classes (more formally defined in Section 3).



(a) Synthetic dataset X with imputed patterns



(b) Synthetic dataset Y with imputed patterns

Figure 1: Two synthetic datasets.

We define the notion of an *integrated pattern(IP)* denoted by $IP_{ij} = \{P_i^X, P_j^Y\}$ as one that is formed by combining a pattern found in X and a pattern found in Y. An IP can serve as potential biomarker only if it is over-represented in one of the two groups of subjects (higher discrimination power). For example, $IP_{11} = \{P_1^X, P_1^Y\}$ is not interesting, since both the IP and the constituent patterns ($P_1^X = \{i1, i2\}$ and $P_1^Y = \{j1, j2\}$) in individual datasets are eaqually represented(5 samples) in both classes ($diffsup = 0$). On the other hand, $IP_{22} = \{P_2^X, P_2^Y\}$ where $P_2^X = \{i4\}$ and $P_2^Y = \{j4\}$, is interesting since both IP and constituent patterns are over-represented in disease group with $diffsup = 0.4$. In contrast, $IP_{33} = \{P_3^X, P_3^Y\}$ is not discriminative since it is not represented in any of the groups leading to $diffsup = 0$, although its constituent patterns ($P_3^X = \{i6\}$ and $P_3^Y = \{j6\}$) were discriminative with $diffsup = 0.4$ before integration. In contrast, integrated pattern $IP_{55} = \{P_5^X, P_5^Y\}$ demonstrates totally opposite phenomenon of IP_{33}. Here, the constituent patterns $P_5^X = \{i11\}$ and $P_5^Y = \{j11\}$ are equally represented in the disease group and the healthy group leading to $diffsup = 0$. However, together they cover the same disease subjects but different healthy subjects. So, IP_{55} is over-represented in the disease group.

Among the four IPs described above, only IP_{22} and IP_{55} are discriminative after integration and thus can act as an potential biomarker. We also refer them as discriminative

integrated patterns. Furthermore, these four IPs demonstrate the relationships among the discrimination power of the IP and those of its constituent patterns. Indeed, we define two types of discriminative IPs based on such relationship namely, *coherence-type* and *interaction-type* IP. More specifically, an *coherence-type IP* has same degree of discrimination as similar to that of the constituent individual patterns (e.g., IP_{22}). These patterns are interesting for biomarker discovery in scenarios when the upstream effects like (genetic perturbations) can be validated in downstream effects (like changes in protein abundance in metabolomics data) [31]. Thus, it can potentially illucidate an underlying causal/cascade relationship among different biomarkers coming from individual datasets. In contrast, we also define an *interaction-type IP* as one whose constituent individual patterns have a degree of discrimination that is lower than that of the integrative pattern (e.g.,IP_{55}). So, the discrimination power arises only when diiferent types of markers are integrated.

Lastly, $IP_{44} = \{P_4^X, P_4^Y\}$ represents a special type of discriminative IP where the individual patterns $P_4^X = \{i8, i9\}$ and $P_4^Y = \{j8, j9\}$ represent the with-in dataset interaction for dataset X and Y, respectively. Therefore, it is a potential biomarker and also an *coherence-type integrated pattern* similar as IP_{22}.

In this paper, we aim to discover all discriminative integrated patterns[1] given any number of datasets. Moreover, we will differentiate between two types of IPs: *coherence-type* and *interaction-type* for finding meaningful biomarkers. Note that among these two types of IPs, feature-extraction based biomarker techniques including CCA and its extensions can only find the coherence types of markers. This is illustrated in more details later in the Section 4 in the context of corresponding real-valued datasets.

3. THE GENERIC FRAMEWORK.

In this integration based framework, we aim to find interesting integrated discriminative patterns (IP_{22},IP_{44},IP_{55} of Figure 1) from multiple datasets. A straightforward way to analyze multiple datasets is to combine them into a common matrix format, and then apply discriminative pattern mining on the combined dataset. Unfortunately, differences in the nature of the data creates numerous challenges to taking such an approach. In particular, there can be differences in format (record vs. network), semantics (a genetic sequence vs. a time sequence), type of variables (binary vs. categorical vs. continuous), the number of variables (dimensions), the amount of information present in each dataset, and differences in biases and assumptions of each data set due to differences in experimental designs and protocols. In addition, we will find many patterns with variables coming from only one dataset which are not interesting for integration and thus, have to be filtered in a post-processing step. More generally, putting such disparate datasets together limits our ability to apply the most relevant pattern finding techniques, muddles the underlying semantics of the data, increases computational complexity, and reduces statistical significance of the discovered patterns.

[1]may be refered simply as integrated patterns from now on.

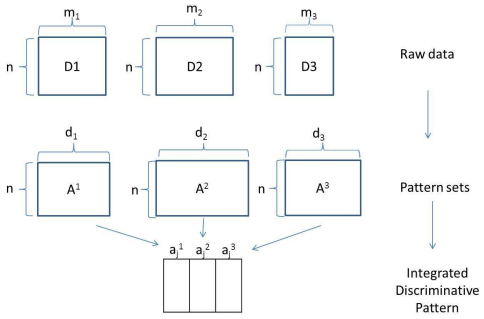


Figure 2: The generic two-step framework for finding integrated discriminative patterns.

To address the above challenges, we propose a two-step approach. The idea is to first find the discriminative patterns from individual datasets respecting the individual properties of the dataset and then combine them into integrated patterns that can distinguish the disease group from the healthy population (Figure 2). In next two subsections, we will describe these two steps in details.

3.1 Finding patterns from individual datasets.

In this subsection, we will first define some notations and then describe how to generate patterns from individual datasets. The definition of a pattern in this step is generic here, as long as the pattern can be associated with the subsets of samples of both classes. For example, the pattern can be either singleton or interaction or sequential pattern. Moreover, any kind of pattern mining technique that seems the most appropriate can be used for this purpose.

Let D be a dataset with a set of m items (variables),$I = \{i_1, i_2, ..., i_m\}$, and n samples from two classes S^+ and S^-. Each sample can be represented as a vector $(\vec{x_i}, y_i)$ for $i = [1, ..., n]$, where $\vec{x_i} \subseteq I$ is a set of items and $y_i \in \{S^+, S^-\}$. The two sets of instances that respectively belong to the two classes S^+ and S^- are denoted by D^+ and D^- such that $|D| = |D^+| + |D^-|$. We define a pattern as $P^D = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ of dataset D, where l is the length of the pattern and $\alpha_i \in I$, $\forall i \in \{1...l\}$. The set of instances from the two classes that contain P are denoted by $D_P^+ \subseteq D^+$ and $D_P^- \subseteq D^-$. The ratio of the samples covered by P in a particular class to the total samples of that class is defined as *RelSup*. For example, $RelSup^+(P^D) = \frac{|D_P^+|}{|D^+|}$ for the positive class S^+.

Definition. The absolute difference of the relative supports of P in D_P^+ and D_P^- is defined originally in [2, 15] and denoted in this paper as diffsup:

$$diffsup(P^D) = |RelSup^+(P^D) - RelSup^-(P^D)| \quad (1)$$

In this step, we will generate all discriminative patterns from each dataset being integrated based on the *diffsup* score. However, relying on *diffsup* only will lead to lot of *driver-passenger* types of pattern from within datasets. A *driver-passenger* pattern is defined as pattern, where at least one

constituent feature has a degree of discrimination that is equal to that of the pattern and at least one constituent feature whose degree of discrimination is quite less than the integrated pattern. For example, pattern $P^X =< i6, i13 >$ from dataset X has diffsup=0.4 which is contributed by i6 only. Note that these types of patterns are not interesting because the pattern does not provide additional discriminative power than its subsets. To filter out such patterns coming from a single dataset, the concept of *improvement* has been used in the literature [3, 16]. Intuitively, *improvement* of a pattern is defined as the gain of the discriminative power over the best discriminative power of any subset.

Definition. For a pattern $P^D = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ in a dataset D, the improvement is defined as

$$improvement(P^D) = diffsup(P^D) - \max_{q^D \subset P^D}(diffsup(q^D)) \quad (2)$$

A pattern P^D is called an *interaction* pattern if its $diffsup(P^D) > \delta$ and $improvement(P^D) > \gamma$, for parameters $\delta > 0$ and $\gamma > 0$.

In this paper, we are mainly interested to retain such improvement type of patterns found from individual datasets to retain the within dataset interactions as demonstrated by $IP_{44} =< \{P_4^X, P_4^Y\} >$ in Figure 1. To find all interaction patterns from a dataset, we use an approach similar to that used in [17, 16]. More specifically, we first mine for the discriminative patterns with $diffsup > \delta$ using SMP [15] and then look at the improvement score of the discriminative patterns with non-negative scores. However, the improvement score is not anti-monotonic in nature [17]. This can lead potentially lead to many missed interactions present across the datasets. For example, a singleton variable from a particular dataset may not be discriminative, but may have interactions with the individual variables or patterns found in other datasets. Therefore, we also retain all the singletons along with the interaction patterns obtained from each datasets. We denote the set of all patterns found from a Dataset D as a pattern set, $PS(D) = \{P_j^D\}_{j=1}^{d} \bigcup I$, where d is the number of patterns found from dataset D and I is the itemset associated with D.

3.2 Finding integrated patterns using the patterns from multiple datasets

In the second step, we will combine the individual patterns found from individual datasets to obtain the final integrated patterns. Suppose we have K heterogeneous datasets $\{D_k\}_{k=1}^{K}$ collected for the same set of n samples. Let, m_k denote the number of items(variables) in dataset D_k. We will have in this stage the pattern sets $PS(D_k)$ found from each of the datasets from the step 1. Let d_k be the number of patterns found from dataset D_k.

For each set of patterns $PS(D_k)$ found from the k^{th} dataset, we define a binary matrix A^k with dimensions $n \times d_k$ for ease of further discussion. Each entry of the binary matrix $\{A^k\}_{ij}$ represents whether the pattern $(P_j^{D_k})$ covers the sample i, for $i = 1...n$ and $j = 1...d_k$. Thus, each column of this matrix, (a_j^k) corresponds to the j^{th} pattern of k^{th} dataset, for $j \in \{1, ..., d_k\}$. Once we have represented the patterns from each dataset in matrix form, we can look for the associations and interactions of these patterns across the datasets to obtain final *integrated patterns*. More formally, an individual integrated pattern IP is given by $IP = \bigcup_{t=1}^{l} P_j^{D_t}$ of length l, where $P_j^{D_t} \in PS(D_t)$ and $t \in [1...K]$. Note that an integrated pattern might not contain patterns from each of the datasets, so $2 \leq l \leq K$.

As discussed earlier, the first criteria of an integrated pattern to be considered as interesting is that it should be discriminative enough, i.e., $diffsup(IP) > \beta$. Then, we aim to differentiate between the two types of IPs: *coherence-type* and *interction-type* IP. We first aim to find the coherence-type pattern by measuring the association of the constituent individual patterns of an IP coming from multiple datasets. In particular, we use *IS* measure to measure such associations. Furthermore, we define another measure called *balance score* to aid the process of finding *interaction-type* IPs. We will first describe these measures in details and then described the algorithm to find both types of IPs.

Objective measures for finding coherence-type IPs. We first aim to measure the association of the constituent patterns of an IP to find the *coherence-type* of patterns. However, the disparate natures of the diverse datasets being integrated introduce additional challenges for measuring associations. For example, the traditional *RelSup* measure may not be appropriate for measuring the associations across multiple datasets. Since, each dataset has their own properties and different degrees of information, having same amount of support may not mean same association for each dataset. Lets consider two IPs: $IP_{22} =< P_2^X, P_2^Y >$ and $IP_{77} =< P_7^X, P_7^Y >$ with same support in disease group from Figure 1. The first integrated pattern has a real association, but the second one is more likely to be statistically insignificant since the individual patterns themselves have higher supports. Another important issue is that the two patterns may contribute unequally to the joint association. For example, for IP_{77}, P_7^X contributes more than P_7^Y, since the former has lower support than the latter ($conf(P_7^X \rightarrow P_7^Y) = 0.66$, but $conf(P_7^Y \rightarrow P_7^X) = 0.5$).

Among the wide variety of interestingness measures [35], we will consider one interestingness measure called IS for assessing associations due to two interestingness properties. First, it can measure the association relative to the baseline supports of constituent markers in each dataset (expected association). Second, IS measure combines the contribution of each individual markers towards the joint associations from the constituent markers by geometric mean [36]. So, if the contribution from any of the two directions(confidence measure) is low then the IS measure is also low. Note that we are looking for disease heterogeneity here and thus interested only in RelSup in the disease (positive) class.

Definition The IS measure between two binary variables A and B is defined as follows:

$$\begin{aligned} IS(A,B) &= \frac{relsup^+(A,B)}{\sqrt{relsup^+(A) \times relsup^+(B)}} \\ &= \sqrt{conf^+(A \rightarrow B) \times conf^+(B \rightarrow A)} \end{aligned} \quad (3)$$

Example $IS(IP_2) = 1.00$, while $IS(IP_7) = 0.57$.

Furthermore, we generalize the original definition of IS measure for integrated patterns from pairs to higher-order integrated pattern for $l \geq 2$, in such a way that the measure becomes anti-monotonic. More formally, the IS measure for an IP with length $l > 2$ can be defined as the minimum of IS measures of all pairwise subsets of IP. The anti-monocity property directly follows from the nature of the min function.

$$IS(IP) = \min_{r,s \in \{1,...,l\}, i \in \{1,...,d_r\}, j \in \{1,...,d_s\}} IS(a_i^r, a_j^s) \quad (4)$$

Objective measure for finding interaction-type IPs If the *IS* measure is low for an IP, then it can be either *interaction-type* or *driver-passenger* type IP. To further explain this scenerio, lets consider two IPs of Figure 1: $IP_{55} = \{P_5^X, P_5^Y\}$ and $IP_{66} = \{P_6^X, P_6^Y\}$. Both of these IPs will have low IS scores. Because, for both of the IPs, the samples covered by the consituent individual patterns are significantly different than the samples covered by the IPs after integration. Among them, IP_{55} is an interaction-type IP and thus is of our interest. In contrast, IP_{66} has $diffsup = 0.6$, where the individual patterns $P_6^X = \{i13\}$ and $P_6^Y = \{i13\}$ already have discrimination power of 0 and 0.6, respectively. These types of *driver-passenger* IPs is not of interest because of the same reason as described in last subsection.

We observed an interesting property of driver-passenger type IPs which is different than both interaction and coherence types of IPs. The diffsups of the constituent patterns of driver-passenger patterns are highly skewed. Thus, we want to make sure that the discriminative power of each individual patterns of an integrated pattern is balanced, rather than skewed. This observation motivates our use of a measure called *balance score*, which is defined as below.

Definition Balance score: For an integrated pattern $IP = \bigcup_{t=1}^{l} P_j^{D_t}$ of length l, where $P_j^{D_t} \in PS(D_t)$ and $t \in [1...K]$, we can represent the diffsups of each pattern $P_j^{D_t}$ as a ***diffsup vector***
$\vec{DV(IS)} =< diffsup(P_j^{D_1}), ..., diffsup(P_j^{D_t}), ..., diffsup(P_j^{D_l}) >$

The *balance score(bs)* is then defined as the cosine similarity ($\cos\theta$, where θ is the angle) between the perfectly balanced vector $\vec{I} =< 1, ..., 1 >$ of length l and $\vec{DV(IS)}$. More formally,

$$bs(IP) = \frac{\sum_{k=1}^{K} diffsup(a_j^k)}{\sqrt{n \times \sum_{k=1}^{K} diffsup(a_j^k)}} \quad (5)$$

For any IP, $0 \leq bs \leq 1$. The larger the bs score, the more balanced the diffsups of the constituent patterns. **Example**: The $IP_{66} =< P_6^X, P_6^Y >$ has a balance score = 0.7, while $bs(IP_{22}) = bs(IP_{44}) = 1$.

Proposition: The balance factor of a driver passenger pattern is low.

To summarize, we want the IP to be both discriminative and balanced as well to filter out the *driver-passenger type* IP. Hence, we use both diffsup and the balance score (bs), $DBS(IP) = diffsup(IP) * bs(IP)$ for assessing the discriminative power of final integrated patterns. Afterward, we use the IS measure to differentiate between the two types of patterns. If the IS measure is high, it will be of coherence-type, otherwise we will conclude it as interaction-type. Now, we describe the algorithm used to mine integrated patterns, such that an integrated pattern IP has both $IS(IP) > \alpha$ and $DBS(IP) > \beta$ given the A^k matrix for each dataset k.

We will use three different pruning criteria to search for such integrated patterns efficiently. This first pruning can be performed based on the anti-monocity property of IS measure. Thus, if an IP of length 2 has $IS(IP) < \alpha$, then any superset of IP with $l > 2$ can be easily pruned.

Another level of pruning is done based on an alternative diffsup formulation suggested by Fang et al. [15], which makes the diffsup measure anti-monotonic. The last pruning criteria for making the algorithm efficient in this stage is to use an upper bound of $DBS(IP)$.

Lemma: For any IP, $DBS(IP) \leq diffsup(IP)$.

Proof: It follows from the definition of DBS **??** and the observation that $0 \leq BS(IP) \leq 1$. Thus, if $diffsup(IP) < \beta$, then we can prune the IP rather than calculating the DBS. The modified PAMIN algorithm is shown in algorithm 1. This algorithm works in the similar to that of the apriori framework [1]. It starts from IPs integrating two datasets and then proceed to integrate higher-order IPs based on the interesting IPs found from the previous stage. The apriori-gen function used in line 19 is same as the original apriori framework.

ALGORITHM 1. *PaminFinal*

1: *Input: A_k, for $k = [1, ..., K]$, parameters α and β*
2: *Output: The set of all discriminative coherence and interaction type IP.*
3: *s=2.*
4: $C_s = \{IP | l(IP) = 2\}$
5: ***repeat***
6: ***for*** $IP \in C_s$ ***do***
7: ***if*** $diffsup(IP) > \beta$ ***then***
8: ***if*** $DBS(IP) > \beta$ ***then***
9: $F_s = F_s \cup \{IP\}$
10: ***if*** $IS(IP) \geq \alpha$ ***then***
11: *Select IS as coherent-type pattern.*
12: ***else***
13: *Select IS as interaction-type pattern.*
14: ***end if***
15: ***end if***
16: ***end if***
17: ***end for***
18: $s = s + 1$
19: $C_s = apriori - gen(F_s)$
20: ***until*** $F_s = \phi$
21: $result = \cup F_s$

end

4. EXPERIMENTS AND RESULTS.

In this section, we will present results for both synthetic datasets and real datasets.

4.1 Datasets.

We generated several synthetic datasets that contain different types of patterns that are similar of those described in section 2. More specifically, we created different types of patterns, i.e., P1-P5 mentioned in the Figure 1 for two real datasets as shown in Figure 3. The first pattern IP_1 is a non-discriminative pattern. For any non-discriminative patterns like these, we generated the data from two identical normal distributions. One the other hand, the discriminative features of IP_2 and IP_3 were created from two different normal distributions with $\mu = 2$ and $\mu = 0$ for two classes. Furthermore they are correlated to the same extent across the two datasets. However, there are some heterogeneous population groups who support the association across the two datasets. For the IP_2, the samples supporting the constituent patterns in individual datasets (P_2^X and P_2^Y) are highly correlated and thus interesting for integrative purposes. On the other hand, the samples supporting the components (P_3^X and P_3^Y) of integrated pattern IP_3 are different and thus, not intersting for integration purpose, although the individual patterns are discriminative by themselves. Once we have generated these types of patterns, we add small amount of white random noise to the data. Moreover, for each of the 5 integrated patterns, we also vary the number of samples that support those patterns totaling 30 integrated patterns. Note that the samples shown in Figure 3 are reordered to represent the homogeneous subgroups.

Beyond this synthetic datasets, we will also evaluate the effectiveness of PAMIN for finding interesting patterns from a real dataset collected from 229 schizophrenic and healthy people. The data has three modalities: functional MRI to measure the functional activity of brain, SNPs to measure the genetic variation and sMRI to capture the structural connectivity of brain. After several preprocessing steps, we extracted 90, 162 and 70 features from each of the datasets respectively. For fMRI dataset, we summarize the activation of each of the 90 brain regions based on their mean correlation with all other 89 brain regions involved in functional network. For details please refer to [27].

4.2 Processing of CCA components to generate pattern.

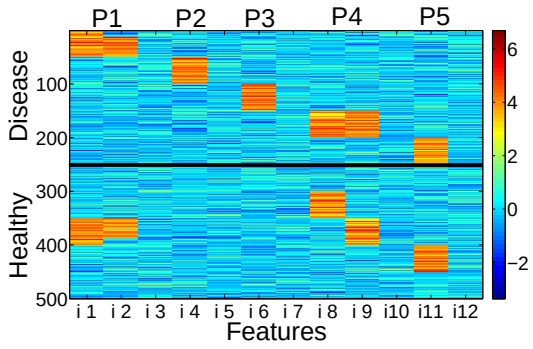
One difficulty with comparing CCA with our approach is that CCA and DCCA combine original features into components by taking a linear combination of the features. Therefore, once we find the discriminative components from CCA, it is difficult to map them back to the original feature space, and thus to the true patterns induced in the synthetic datasets which consist of those original features. To circumvent this problem, we assess the activation level of each original feature for the discriminative canonical component by looking at the coefficints of component maps. More specifically, for each component, we compute the Z-score of the activation levels and then take the variables that exceed a particular Z-score threshold as in [10].

4.3 Results on synthetic dataset.

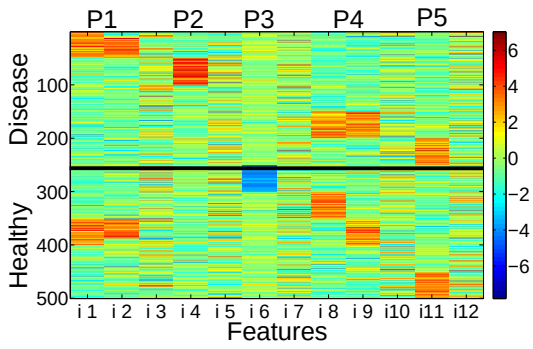
: We investigated how CCA, DCCA and PAMIN work for different cases when applied on the data shown in Figure 3. Note that, there are 5 integrated patterns, among which IP_{11} and IP_{33} are non-discriminative integrated patterns, and IP_{22}, IP_{44}, IP_{55} are the discriminative integrated patterns, as discussed earlier in Section 2.

Figures 4 and 5 illustrate the components obtained from the dataset represented in Figure 3 using the CCA and DCCA techniques, respectively. In these figures, each row represents a component and each column represents the variables in the original datasets, X and Y, respectively. Each entry of these component maps represents the activation profile or contribution of the variable to that component. The higher(red) or lower(blue) the co-efficients are, the better the the association between the component and the corresponding feature. The y-axis of each of these two figures represents the discrimination power of those components in terms of their t-test (-log P value). These two figures yeild the following observations:

- The first non-discriminative integrated pattern IP_{11} containing the first two features of both X and Y datasets are not found by CCA (correponding to component 1 and 10). However, CCA finds IP_{33}, which is represented by the component 11 with p-value $< 10^{-9}$ and 10^{-11}. This IP is not discriminative after integration, although the corresponding individual patterns are discriminative. They are taken by CCA because, CCA looks for overall correlation across the two datasets and the support of the two constituent patterns come from a different subset of samples. Note, this can represent a different type of integrated pattern which can handle population heterogeneity better. Our pattern mining based approach will not find these types of IPs because of the AND type of relationship is used when defining associations.
- The integrated pattern IP_{22} is a discriminative integrated pattern and it is supported by same set of samples across the two datasets. So, both CCA (8th component with $p-value < 10^{-3.4}$ and $< 10^{-3}$ for dataset X and Y respectively) and PAMIN can find it.
- The interactive integrated patterns (represented by IP_{44} and IP_{55}) cannot be found by any of the components of CCA. Althogh the feature of IP_{55} corresponding to across dataset interaction is picked by the last component of CCA, it is not discriminative enough in both of the datasets. We can see that some components pick these patterns based on their common correlations in the disease group, but cannot find the discriminative power; since they look for discrimination power in individual datasets independently. Note that our approach can easily find all these interactive integrated patterns.
- Besides all these observations, we can also see that sometimes a good pattern is merged with a poor non-discriminative pattern into a single component. For example, component 8 in both the datasets not only covers the integrated pattern IP_{22}, but also merges it with the 7th variable from both datasets. This leads to the potential driver-passenger patterns and impedes the interpretability of the obtained biomarkers.

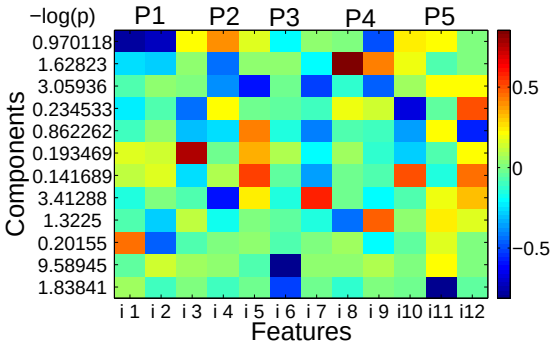


(a) Dataset X

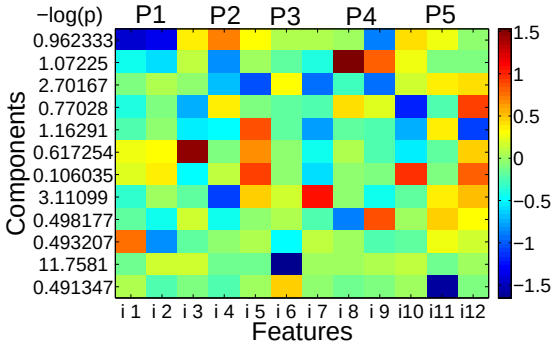


(b) Dataset Y

Figure 3: Two datasets containing different types of patterns of interest.

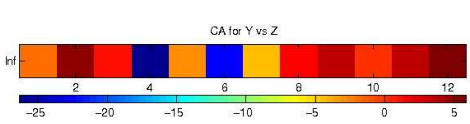


(a) The components selected by CCA for each of the variables of dataset X.

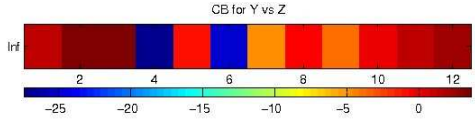


(b) The components selected by CCA for each of the variables of dataset Y.

Figure 4: Integrated patterns recovered by CCA.



(a) The components selected by DCCA for each of the variables of dataset X.



(b) The components selected by DCCA for each of the variables of dataset Y.

Figure 5: Integrated patterns recovered by DCCA.

The findings of DCCA is same as those of CCA. Based on these observations, we conclude that our approach can find interactions both within and across the datasets, which CCA and DCCA based technique will not be able to find. This is because these approaches measure association across the datasets based on the correlation and thus, misses any non-linear types of relationships present in the data including interactions.

4.4 Quantitative analysis on synthetic dataset.

In this section, we will compare PAMIN and CCA more systematically on a synthetic datasets containing 30 integrated patterns. For applying our approach, we first converted the real-valued dataset into a binary format by choosing different thresholds. In particular, we binarized each column by taking the top [10, 20, 30, 40] percentile from both tails of the distribution and then represented each variable by two new variables corresponding to two tails of the disctributions. For detailed evaluation purpose, we compared the precision and recall of all the methods(PAMIN, CCA, and DCCA). Unlike CCA and DCCA, computing precision and recall is easy for PAMIN, since it aims to find the integrated patterns directly. To create an evaluation matrix for CCA and DCCA, we first constructed a matrix $\{E^k\}_{k=1}^{K}$ for each datasets with dimension $S \times C$, where S and $C = \min_{k=1,..K} d_k$) are the number of true integrated patterns in our synthetic dataset and the number of components found from CCA(or DCCA) respectively. Each entry of this matrix ${E^k}_{ij}$ is computed as the Jaccard similarity between the set of variables present in the i-th integrated pattern and the variables selected by the j-th canonical component. Finally, we binarized these K matrices based on a particular threshold (τ) to obtain the final confusion matrix for computing true positives, false positives and false neg-

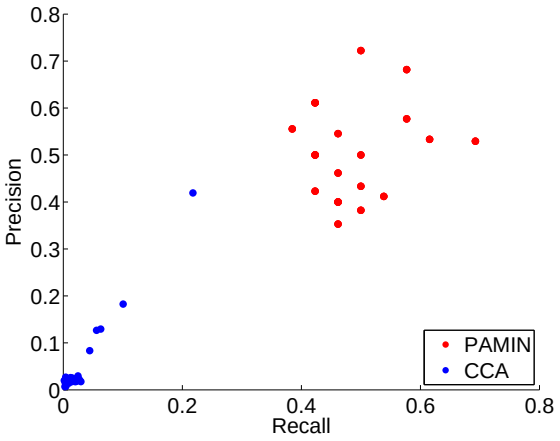
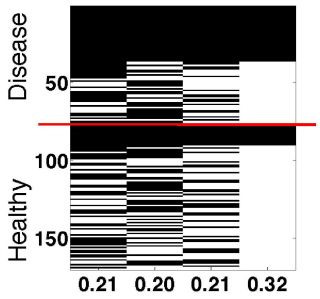


Figure 6: The precision-recall curve for both PAMIN and CCA.

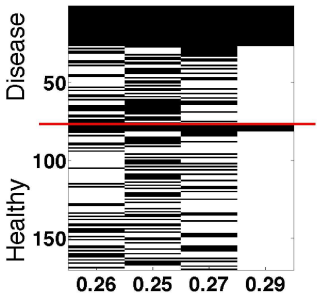
atives. Thus, there are two parameters involved in building the confusion matrix for CCA: the Z-score threshold for selecting the variables and τ. We varied these thresholds and then plotted their impact using precision-recall values. Similarly, PAMIN was varied by using different binarization thresholds (top percentile) for discretizing the data, IS threshold(α) and DBS threshold(β). As shown in figure 6, PAMIN outperforms CCA in terms of both precision and recall. Note that, since the negative class(negative Integrated patterns) is hard to define for our case, we therefore, rely on precision-recall values rather than ROC curves.

4.5 Neuroscience data.

In the neuroscience data, we want to demonstrate the applicability of our method for integrating more than two datasets. More specifically, we integrated three different types of modalities, namely fMRI, sMRI and SNP data. Figure 7 represents two different types of integrated patterns (IP) discovered by the proposed approach. In each figure, the first three columns represent the three different patterns coming from the three datasets and the fourth column represents the integrated pattern after combining them. In Figure 7(a), we show one of the interaction type integrated patterns, where the fMRI (left temporal pole), SNP (rs760761), and sMRI (parsorbitalis) patterns have 0.21, 0.20, and 0.21 discrimination power (measured by *Diffsup*), respectively (indicated along the x-axis of the figure). However, the discrimination power increases by 11% after integration (*Diffsup* $= 0.32$ and *Imp* $= 0.11$). Similarly, Figure 7(b) represents one of the coherence type of integrated patterns where the discriminative power is essentially the same for all three patterns (*Diffsup* $= 0.26, 0.25, 0.27$) and the integrated pattern (*Diffsup* $= 0.29$), and thus there is no significant improvement in terms of the discrimination power (*Imp* $= 0.02$). Both of this patterns are statistically significant with FDR$< 10^{-2}$ (the process of finding statistical significance is described in next paragraph). Furthermore, in this dataset, we have two categories of schizophrenic patients: chronic and first episode. This provides us an opportunity to check the homogeneity of the subspace covered by the two patterns. The patients covered by the individual patterns found in each the three datasets explain subjects of both subcategories. However, the integrated pattern containing all three patterns covers mostly the chronic subgroup (33 out of 38 for the coherence pattern).



(a) The top interacting pattern generated by the proposed approach.



(b) The top associative pattern generated by the proposed approach.

Figure 7: Two sample patterns discovered by the proposed approach.

4.6 Statistical significance of the obtained patters.

Since the real datasets lack the ground truth, it is hard to evaluate the obtained set of patterns. Moreover, the $diffsup$ measure only captures the discriminitative power of the pattern without providing any statistical significance of the obtained pattern, since the sample size is not considered in the formula of diffsup. Furthermore, the search space for pattern mining is exponential, and thus is often criticized to generate numerous hypothesis which can potentially lead to the increased type I error. Therefore, it is important to correct the statistical significance of the obtained patterns for multiple hypothesis testing[38, 37, 21]. Among different available techniques for correcting multiple hypothesis testing, we used the randomization strategy of class label, which has been extensively used in several computational biology applications [32, 14]. In particular, we randomize the class labels of each sample for certain number of times and then repeated the whole pattern mining procedure using the same parameter settings (alpha and beta) for several runs, and then computed the false discovery rate of the DBS scores of the obtained discriminative integrated patterns in compare to the randomized versions.

We also computed the statistical significance of the obtained discriminative components of MCCA, so that we can compare those with the statistical significant integrated patterns. However, as mentioned earlier, the components of CCA are not directly comparable to patterns obtained by PAMIN. To map the components of the MCCA into the original feature space of the dataset, we use the similar approach as described earlier in section 4.2. Once, we map each of the statistical significant component to the original features of each of the three datasets, we took a conservative approach to compute the number of patterns from those statistical significant components. In particular, if there is n_i features mapped from i-th dataset, then thre are $\prod n_i$, for $i = 1, ..., 3$. Figure 8(a) shows that our approach can find more significant patterns than MCCA for same label of statistical signif-

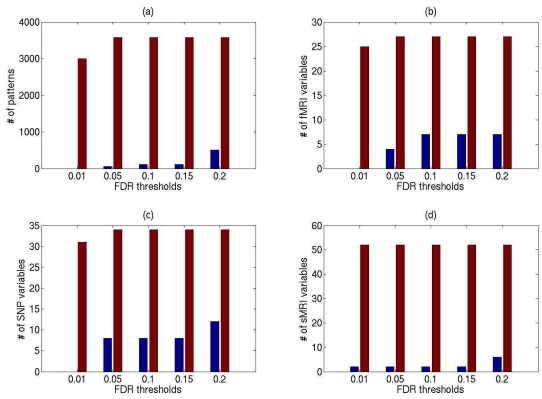


Figure 8: Comparision between MCCA and PAMIN for different statistical significance levels. Subfigure (a) compares total number of patterns and subfigure (b-d) compares the original features covered by those patterns for three datasets: fMRI, SNP, thickness respectively.

icance label (computed by FDR). Then, we also computed the coverage of the obtained patterns and components in terms of number of variables picked from the three datasets (Figure 8)(b-c). Our proposed approach discovers more statistically significant patterns and the coverage of those patterns is also higher than those of MCCA for each datasets. The results are shown only for a particular parameter setting(with IC=40 for MCCA and alpha=0.65, beta=0.25 for PAMIN) because of space limitations, although the observations do not change for other parameter settings. The effect of the parameters of our algorithm: alpha and beta is very obvious leading monotonically increasing number of IPs as any of the parameters decreases.

Figure 9 demonstrates the overlap between the coverages of both algorithm for fMRI dataset. Although we observed good overlap between the coverages of these two algorithms, the similar phenomenon was not observed for SNP and thickness datasets. This may be due to the difficulty of MCCA to handle discrete datasets like SNP. Furthermore, MCCA is an iterative approach and thus, there was lot of variation in the correlation structures for different number of components chosen. We aim to exflore further the coverage issues for other two datasets in future works.

5. CONSLUSION.

In this paper, we proposed a pattern mining based integration framework for finding relationship from multiple heterogeneous datasets with binary class labels. This framework has several advantages. First, this framework is generic enough to integrate any number of heterogenous datasets. Second, this pattern mining based framework can find both within and across dataset interactions that are not discovered by feature-extraction based approaches such as CCA and DCCA.

Given the potential utility of our proposed pattern mining based integration framework, there are some interesting future directions. Currently, the proposed pattern mining framework works only for binary dataset. This needs to be

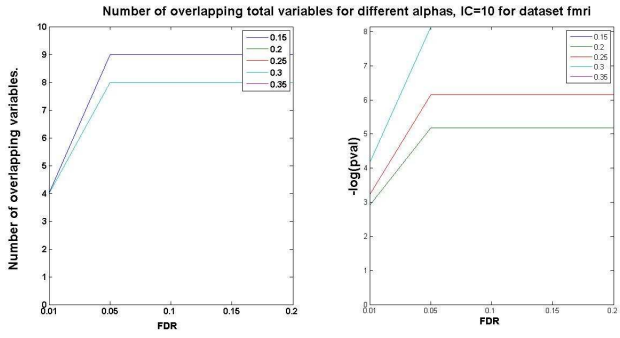


Figure 9: The overlap between the fMRI coverages of MCCA and PAMIN measured by number of variables selected (a) and hypergeometric p-value(b)

extended for real-valued dataset also. Furthermore, we observed an issue with pattern mining based framework that many times patterns are themselves correlated, i.e.,there is redundancy among the patterns within the datasets. Summarizing those related patterns in the first step of the framework is a potential future direction. Another potential future research direction is to develop an efficient algorithm combining both generation and summarization steps of the algorithm, so that many of the redundant patterns can be pruned earlier in the search process.

6. REFERENCES

[1] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994.

[2] S. Bay and M. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[3] R. Bayardo, R. Agrawal, and D. Gunopulos. Constraint-based rule mining in large, dense databases. *Data Mining and Knowledge Discovery*, 4(2):217–240, 2000.

[4] C. Bloss, N. Schork, and E. Topol. Effect of direct-to-consumer genomewide profiling to assess disease risk. *New England Journal of Medicine*, 364(6):524–534, 2011.

[5] V. Calhoun, T. Adali, G. Pearlson, and K. Kiehl. Neuronal chronometry of target detection: fusion of hemodynamic and event-related potential data. *Neuroimage*, 30(2):544–553, 2006.

[6] V. Calhoun, J. Liu, and T. AdalI. A review of group ica for fmri data and ica for joint inference of imaging, genetic, and erp data. *Neuroimage*, 45(1):S163–S172, 2009.

[7] H. Chuang, E. Lee, Y. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3(1), 2007.

[8] N. Correa, T. Adali, Y. Li, and V. Calhoun. Canonical correlation analysis for data fusion and group inferences. *Signal Processing Magazine, IEEE*, 27(4):39–50, 2010.

[9] N. Correa, Y. Li, T. Adali, and V. Calhoun. Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. *Selected Topics in Signal Processing, IEEE Journal of*, 2(6):998–1007, 2008.

[10] N. Correa, Y. Li, T. Adali, and V. Calhoun. Fusion of fmri, smri, and eeg data using canonical correlation analysis. In *ICASSP*, pages 385–388. IEEE, 2009.

[11] A. Daemen and others. Integrating microarray and proteomics data to predict the response on cetuximab in patients with rectal cancer. In *Pacific Symposium on Biocomputing 2008*, page 166, 2008.

[12] T. Diethe, D. Hardoon, and J. Shawe-Taylor. Constructing nonlinear discriminants from multiple data views. *Machine Learning and Knowledge Discovery in Databases*, pages 328–343, 2010.

[13] R. Edgar, M. Domrachev, and A. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data

repository. *Nucleic acids research*, 30(1):207, 2002.

[14] G. Fang, R. Kuang, G. Pandey, M. Steinbach, C. Myers, and V. Kumar. Subspace differential coexpression analysis: problem definition and a general approach. In *Proceedings of the 15th Pacific Symposium on Biocomputing (PSB)*, volume 15, pages 145–156, 2010.

[15] G. Fang, G. Pandey, W. Wang, M. Gupta, M. Steinbach, and V. Kumar. Mining low-support discriminative patterns from dense and high-dimensional data. *Knowledge and Data Engineering, IEEE Transactions on*, (99):1–1, 2009.

[16] G. Fang, W. Wang, B. Oatley, B. Van Ness, M. Steinbach, and V. Kumar. Characterizing discriminative patterns. *Arxiv preprint arXiv:1102.4104*, 2011.

[17] W. W. H. Y. M. S. T. C. W. O. B. V. N. Gang Fang, Majda Haznadar and V. Kumar. High-order snp combinations associated with complex diseases: Efficient discovery, statistical power and functional interactions. *PLoS ONE*, To appear, 2012, http://vk.cs.umn.edu/HSC/.

[18] M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

[19] Y. Guo, J. Li, Y. Chen, L. Zhang, and H. Deng. A new permutation strategy of pathway-based approach for genome-wide association study. *BMC bioinformatics*, 10(1):429, 2009.

[20] M. Hamburg and F. Collins. The path to personalized medicine. *New England Journal of Medicine*, 363(4):301–304, 2010.

[21] S. Hanhijärvi, M. Ojala, N. Vuokko, K. Puolamäki, N. Tatti, and H. Mannila. Tell me something i don't know: Randomization strategies for iterative data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379–388. ACM, 2009.

[22] G. Lanckriet, M. Deng, N. Cristianini, M. Jordan, and W. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 9, page 2. World Scientific Singapore, 2004.

[23] K. Lê Cao, I. González, and S. Déjean. integromics: an r package to unravel relationships between two omics datasets. *Bioinformatics*, 25(21):2855, 2009.

[24] K. Lê Cao, P. Martin, C. Robert-Granié, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics*, 10(1):34, 2009.

[25] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N. Perrone-Bizzozero, and V. Calhoun. Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica. *Human brain mapping*, 30(1):241–255, 2009.

[26] J. Loscalzo, I. Kohane, and A. Barabasi. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular systems biology*, 3(1), 2007.

[27] M. Lynall, D. Bassett, R. Kerwin, P. McKenna, M. Kitzbichler, U. Muller, and E. Bullmore. Functional connectivity and brain networks in schizophrenia. *The Journal of Neuroscience*, 30(28):9477, 2010.

[28] E. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008.

[29] P. Ng, S. Murray, S. Levy, and J. Venter. An agenda for personalized medicine. *Nature*, 461(7265):724–726, 2009.

[30] E. Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461(7261):218–223, 2009.

[31] E. Schadt, J. Lamb, X. Yang, J. Zhu, S. Edwards, D. GuhaThakurta, S. Sieberts, S. Monks, M. Reitman, C. Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710, 2005.

[32] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545, 2005.

[33] T. Sun, S. Chen, J. Yang, and P. Shi. A novel method of combined feature extraction for recognition. In *2008 Eighth IEEE International Conference on Data Mining*, pages 1043–1048. IEEE, 2008.

[34] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. In *Proceedings of the 25th international conference on Machine learning*, pages 1040–1047. ACM, 2008.

[35] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41. ACM, 2002.

[36] P. Tan, M. Steinbach, V. Kumar, et al. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.

[37] G. Webb. Discovering significant patterns. *Machine Learning*, 68(1):1–33, 2007.

[38] G. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71(2):307–323, 2008.