Discriminative pattern mining and its applications in bioinformatics

Xiaoqing Liu, Jun Wu, Feiyang Gu, Jie Wang and Zengyou He

Abstract

Discriminative pattern mining is one of the most important techniques in data mining. This challenging task is concerned with finding a set of patterns that occur with disproportionate frequency in data sets with various class labels. Such patterns are of great value for group difference detection and classifier construction. Research on finding interesting discriminative patterns in class-labeled data evolves rapidly and lots of algorithms have been proposed to specifically address this problem. Discriminative pattern mining techniques have proven their considerable value in biological data analysis. The archetypical applications in bioinformatics include phosphorylation motif discovery, differentially expressed gene identification, discriminative genotype pattern detection, etc. In this article, we present an overview of discriminative pattern mining and the corresponding effective methods, and subsequently we illustrate their applications to tackling the bioinformatics problems. In the end, we give a general discussion of potential challenges and future work for this task.

Key words: discriminative pattern mining; contrast sets; emerging patterns; subgroup discovery

Introduction

Pattern mining contributes a set of patterns that characterize the large complex data sets in concise and informative forms. One of the most well-known tasks is ‘frequent pattern mining’ [1–5], which aims at finding all sets of items with frequent occurrences in a given data set. An archetypical example of frequent patterns could be the products that are often purchased together, such as bread and milk, beer and chips. Furthermore, the comparative investigation of discrimination between different data sets has also become an indispensable work, which provides us novel insights into how to efficiently and effectively tackle complex problems in the real world such as group difference detection and classifier construction [6, 7]. Therefore, this study focuses on a task of pattern mining technique dubbed ‘discriminative pattern mining’.

The discovery of distinguishing characteristics and humanly interpretable differences between data sets with class labels is one of the most important objectives in data mining. Discriminative pattern mining aims at finding a set of interesting patterns that occur with disproportionate frequency in some classes versus others. For example, the US census data from 1979 to 1992 show that

\[ P(\text{occupation} = \text{sales} \mid \text{PhD}) = 2.7\% \]
while $P(\text{occupation} = \text{sales} | \text{Bachelor}) = 15.8\%$ [8]. Then (occupation = sales) is a discriminative pattern if we consider the education levels as class labels. In fact, those biological entities such as differentially expressed genes or proteins in patients and healthy persons, are also examples of discriminative patterns.

During the recent years, the task of discriminative pattern mining has drawn widespread attention in both data mining and machine learning communities. Research on discriminative patterns evolves rapidly under several nonuniform definitions such as contrast sets [9], emerging patterns [10] and subgroups [11, 12]. According to [9, 13], contrast set mining aims at discovering patterns that capture prominent frequency differences across different user-defined groups of subjects. Emerging pattern mining detects patterns that capture the frequency growth change from one class to another [10, 14]. While subgroup discovery tries to find population subgroups that are statistically most interesting given a population of individuals and a property of those individuals that we are interested in, e.g. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest [11, 12]. Overall, though such patterns are investigated under different names, they are equivalent as owners of discriminative power across classes in essence. Accordingly, we refer to all these patterns as discriminative patterns in this article.

Discriminative patterns can exactly capture the differences between data sets, which contributes to building high powerful classifiers and characterizing different classes. Moreover, such patterns have been shown to be of considerable value in a wide range of applications such as the patient risk group detection in medicine, the discovery of overexpressed genes in microarray data analysis and the identification of distinguishing features in customer relationship management [15]. Hence, the identification of such patterns is a valuable work in the need of practical issues.

Many algorithms and software frameworks have been devised for efficiently discovering discriminative patterns using supervised or unsupervised learning technology, under exhaustive or heuristic search strategy. All of these techniques can be divided into different categories from different points of view. First, we can explore the task of discriminative pattern detection from two perspectives according to their target objectives for investigation: one is called local pattern mining or individual pattern mining [13, 16–20]; the other is global pattern mining or pattern set mining [21–24]. The main difference between local and global pattern mining is that the former imposes the constraints at the level of individual patterns, whereas the latter applies the constraints to the overall set of patterns [25]. Contrast to global pattern discovery, local pattern mining evaluates every pattern separately under no consideration of the relationships between each other. Accordingly, one recognized drawback of local pattern mining is that the generated set of patterns is generally too large and complicated to be directly used without postprocessing. Second, to conduct the mining process, some researches adopt some feasible thresholds of the constraints such as information gain [6] and odds ratio [26] to measure the discriminative power and then filter out those insignificant patterns. Whereas other researches focus on finding $k$ patterns with the strongest discriminative power [12, 24, 27]. In addition, it is difficult to produce an integrated set of significant discriminative patterns especially in very large data sets. Some methods [28–30] employ the exhaustive search strategy to detect an optimal set of interesting patterns under the given constraints, while some methods [31–33] choose to report a good enough but not necessarily optimal result by the heuristic search strategy.

All these technologies proposed for discriminative pattern mining can be readily applied to solving some bioinformatics problems. They provide great insights into identifying biologically significant patterns with discriminative knowledge across different biological data. The typical applications in bioinformatics include the phosphorylation motif discovery [34, 35], differentially expressed gene identification [16, 36, 37], discriminative genotype detection [38, 39], etc. For example, equivalent to discovering discriminations between data sets, the discovery of phosphorylation motifs pursues a goal of finding a set of motifs that appear more frequently in the phosphorylated peptide set than that in the unphosphorylated peptide set [34, 35]. Upon this issue, the data sets can be translated into the inputs of discriminative pattern discovery methods by considering the property of (un)phosphorylation as class labels. The significant motifs with statistical interestingness can represent the differences between these two classes, which is equivalent to the discriminative patterns. Hence, we refer to phosphorylation motif discovery as an example task of discriminative pattern mining in the field of bioinformatics. In addition, another representative application of discriminative pattern discovery is to analyze gene expression data for cancer detection [16, 36, 40]. It offers an effective tool for explicitly modeling and characterizing different stages of cancers, which thus facilitates biologists to achieve a better understanding of the function of genes and the pathogenesis of cancer. Ultimately, it is much easier to conduct cancer diagnosis and find the optimal treatment plan.

One book [41] and several instructive reviews [15, 42] that summarize recent advances on discriminative pattern mining are available in the literature. However, [42] just lists several applications in bioinformatics without detailed discussion and [15] presents a unifying framework without any relevant applications. Moreover, all of them focus on the discussion of mining discriminative patterns on the local level, but put little attention on the global pattern mining. Hence, this article provides a complementary study to the existing research with a straightforward introduction of discriminative pattern mining and presentation of the current technologies from different perspectives. Some successful applications of discriminative pattern mining techniques in the field of bioinformatics with detailed analysis are presented as well.

In this introductory article, we first give a brief introduction of the standard definition and formulation of discriminative pattern mining. We then review a number of representative and popular open-source approaches. Subsequently, we concentrate on the discussion of some typical bioinformatics applications. To conclude, we summarize our contributions and highlight the expectations and the challenges of the use of these techniques to address biological problems.

**Preliminaries**

Discriminative patterns have been investigated under different terminologies (contrast sets, emerging patterns, subgroups) in different communities. Nevertheless, the definitions and the comprehension with respect to such patterns still remain scattered. In this section, we formalize and generalize these specific definitions to propose a uniformly straightforward formulation for illustrating the discovery of discriminative patterns. Then we discuss some frequently used statistical measures for evaluating the discriminative power.
Basic terminology and problem formulation

Let \( D \) be a data set consisting of a certain number of transactions with \( n \) class labels, which is denoted by \( D = D_1 \cup D_2 \cup \ldots \cup D_n \). The universe of all possible items that occur in the transactions is denoted as \( I = \{i_1, i_2, \ldots, i_n\} \), which only includes discrete values. A pattern \( p = \{i_1, i_2, \ldots, i_k\} \subseteq I \) is defined as a \( k \)-pattern because it is composed of \( k \) items. \( p \) is a sub-pattern of one pattern \( q \) if \( p \subseteq q \) (\( q \subseteq I \)), in which situation \( q \) is called a super-pattern of \( p \). A ‘transaction’ in \( D_j (j = 1, \ldots, n) \) is a pair \( T = (\text{tid}, \text{trans}) \), where tid is the unique identifier and trans is a subset of \( I \). We say a transaction \( T \) contains the pattern \( p \) under the condition of \( p \subseteq \text{trans} \). Thus, the set of transactions that contain \( p \) in \( D_j \) is denoted by \( \mathcal{D}_j(p) = \{T : T \in \mathcal{D}_j, p \subseteq \text{trans}\} \). Every transaction that contains the pattern \( p \) must also contain its sub-patterns, but may not contain its super-patterns such as \( q \). Moreover, the ‘support’ of a pattern is defined as the percentage of the transactions that contain this pattern in a data set. Accordingly, the ‘support’ of \( p \) over \( D \) is denoted by:

\[
\text{sup}(p, D) = \frac{|\mathcal{D}(p)|}{|D|},
\]

where \(|\cdot|\) denotes the size of a data set. \( p \) is frequent if its ‘support’ value is no less than a given threshold \( \theta_{\text{sup}} \) (\( 0 \leq \theta_{\text{sup}} \leq 1 \)). The problem of frequent pattern mining which aims at finding a set of frequent patterns can now be formulated as finding:

\[
\mathcal{F} = \{p \subseteq I : \text{sup}(p, D) \geq \theta_{\text{sup}}\}.
\]

A pattern is closed in a data set if there exists no super-pattern with the same frequency in the same data set. A frequent pattern is called maximal if it is not a subset of any other frequent pattern.

We use \( \Pi \) to denote a pattern set consisting of \( s \) patterns, i.e., \( \Pi = \{p_1, p_2, \ldots, p_s\} \). The transaction set containing the entire pattern set \( \Pi \) is defined as the union over the individual transaction sets which each pattern corresponds to as follows:

\[
\mathcal{D}(\Pi) = \mathcal{D}(p_1) \cup \mathcal{D}(p_2) \cup \ldots \cup \mathcal{D}(p_s).
\]

Similar to the individual pattern, the ‘support’ of the pattern set is:

\[
\text{sup}(\Pi, D) = \frac{|\mathcal{D}(\Pi)|}{|D|}.
\]

\( \Pi \) is said to be a frequent pattern set if \( \text{sup}(\Pi, D) \) is no less than the given threshold \( \theta_{\text{sup}} \).

For illustration purpose, consider Figure 1, which is a sample data set containing 20 transactions from two classes, where tid ranges from 1 to 20. For the transaction with \( \text{tid} = 5 \), \( \text{trans} = \{i_1, i_2, i_3, i_6, i_7\} \). In this figure, \( D = D_1 \cup D_2 \), \( I = \{i_1, i_2, \ldots, i_{10}\} \), and three patterns can be observed: \( p_1 = \{i_1, i_2, i_3\} \), \( p_2 = \{i_5, i_6, i_7\} \) and \( p_3 = \{i_8, i_9\} \). The possible pattern sets are \( \Pi_1 = \{p_1, p_2\} \), \( \Pi_2 = \{p_2, p_3\} \), \( \Pi_3 = \{p_1, p_3\} \) and \( \Pi_4 = \{p_1, p_2, p_3\} \). We just take these patterns and pattern sets as examples to illustrate the basic terminology of patterns, even though many other patterns such as \( \{i_1, i_2\} \) are also contained in this data set. Suppose \( \theta_{\text{sup}} = 0.3 \), the frequent patterns and pattern sets in \( D_1 \) include \( p_1, p_2, \Pi_1, \Pi_2, \Pi_3, \Pi_4 \), while those in \( D_2 \) include \( p_1, \Pi_1, \Pi_2, \Pi_3, \Pi_4 \).

To discover all frequent patterns, one could enumerate every possible pattern and evaluate its frequency with the given threshold \( \theta_{\text{sup}} \). For example, the widely used Apriori algorithm [1] adopts a level-wise search to reduce the search space based on the Apriori property, which describes that if a pattern is frequent then all of its sub-patterns must also be frequent. This method first scans the data set and selects all the 1-patterns \( i \). Then, the set of candidate frequent 2-patterns could be generated as \( \{i_1, i_2\}, \{i_1, i_3\}, \{i_1, i_6\}, \{i_1, i_7\}, \{i_2, i_3\}, \{i_2, i_6\}\) according to the definition of frequent pattern. Subsequently, only one frequent 3-pattern can be obtained, that is \( \{i_1, i_2, i_3\} \).

Figure 1. A sample data set. The sample data set contains 20 transactions with two classes, in which 10 items make up 3 patterns. Rows correspond to items, and columns correspond to transactions. A shadowed cell indicates that an item is included in a transaction.
Based on considering the frequency of one possible pattern (set), discriminative pattern mining usually adopts some statistical measures to evaluate the discriminative power of each individual pattern or the whole pattern set. Only the patterns or the pattern sets that are able to pass the user-specified significance tests can be considered as significant ones.

On the local level, where the main search objective is the individual pattern, a wide variety of evaluation measures are available, such as the ‘growth rate’ [10], the ‘difference of the two supports’ [9, 28], ‘information gain’ [6], ‘odds ratio’ [26], etc. For example, the absolute difference of supports is referred to as $\text{DiffSup}$ with the following definition:

$$\text{DiffSup}(p, D_i, D_j) = |\text{sup}(p, D_i) - \text{sup}(p, D_j)|, (1 \leq i \neq j \leq n). \quad (5)$$

In the context of individual discriminative pattern mining, the pattern $p$ is a significant discriminative pattern if its $\text{DiffSup}$ is no less than a given threshold $\theta_{\text{sup}}$.

For example, consider Figure 1 again, we adopt $\text{DiffSup}$ as the significance measure with $\theta_{\text{sup}} = 0.2$ to conduct the analysis of discriminative pattern mining. Accordingly, $p_1$ and $p_2$ are 3-patterns whose $\text{DiffSup}$ values are 0.4 and 0, respectively. While $p_1$ is a 2-pattern with a $\text{DiffSup}$ value of 0.2. Thus, $p_1$ and $p_3$ are significant discriminative patterns with valid occurrence distinctions in the two classes, which possess strong discriminative power with $\text{DiffSup}$ values no less than $\theta_{\text{sup}}$. However, discovering significant patterns in this way possibly results in a redundant and complex result set in some instantiations, which needs to be heuristically post-processed for further use. To address this problem, some methods [32, 43-46] also impose another constraint such as ‘support’ as the primary constraint to reduce the number of potential discriminative patterns in one specific class, especially when conducting the discovery in very large data sets. These methods usually first explore the data in one class to collect the most frequent patterns as candidates, and then they adopt the significance measures to prune the uninteresting patterns. For instance, $p_3$ will be filtered out if we also use a frequency threshold of $\theta_{\text{sup}} = 0.3$ due to its infrequency both in $D_1$ and in $D_2$.

Instead of evaluating each individual pattern as introduced above, some methods [24, 25, 47, 48] also apply the significance tests to the entire pattern set on the global level. They formalize the problem of finding discriminative patterns as global pattern mining or pattern set mining. The main idea of this task is to find a set of significant patterns by incorporating the dependencies and the similarities between different patterns. More precisely, a pattern set is considered to be significant on condition that it satisfies some user-defined global constraints by computing the ‘accuracy’ or ‘area’ under the receiver operating characteristic curve [25, 47], using the ‘diversity’ or ‘redundancy’ to measure the overlap between different patterns or transactions [23, 47, 49]. Additionally, to measure the overall discriminative power, one simple and convenient approach is to directly adopt local significance measures for global significance evaluation. Suppose a task of discriminative pattern set mining over Figure 1 directly uses the combination of local and global statistical significance evaluations. Then, the global significance value of the pattern set $\Pi_4 = \{p_1, p_2, p_1\}$ in Figure 1 is equivalent to 0.2 because its support values are 0.7 in $D_1$ and 0.5 in $D_2$, respectively. Hence, $\Pi_4$ is globally significant with $\theta_{\text{sup}} = 0.2$. However, taking the local evaluation into consideration, this pattern set should be filtered out because $p_2$ is insignificant. Another pattern set $\Pi_3 = \{p_1, p_3\}$ can pass the global significance test with a significance of 0.4, which is larger than the given threshold. Additionally, because $p_1$ and $p_3$ are also significant on the local level, $\Pi_3$ will be reported as an interesting pattern set (This plain example is just used to show how some researches evaluate one pattern set on the global level, and it excludes considering other constraints such as redundancy and diversity, etc. Concrete and detailed introduction of the pattern set mining will be presented in the following section).

Overall, we provide a general definition of the problem of discriminative pattern mining with the clearly stated input and output.

- **Input:** A data set $D$ with $n$ class labels, the user-defined parameters such as the significance threshold $\theta_{\text{sup}}$ or the expected number of interesting discriminative patterns to be discovered $\theta_{\text{k}}$.

- **Output:** A set of discriminative patterns $R$, which satisfies at least one of the following requirements, i.e. $R = R_1$ or $R = R_2$:

$$R_1 = \{p \subseteq I \mid |\text{Top}(\pi(p))| = \theta_{\text{k}} \text{ or } \text{Compare}(\pi(p), \theta_{\text{sup}}) = \text{true}\}, \quad (6)$$

where $\pi$ presents a local significance evaluation function on each individual pattern, Top is a function which extracts a certain number of the most interesting patterns and Compare is true if the significance value of $p$ satisfies the comparison constraints (such as $<, \leq, >, \geq, =, \neq$) with respect to the given threshold $\theta_{\text{sup}}$ according to the definition.

$$R_2 = \{\Pi \subseteq R_1 \mid |\rho(\Pi)| = \theta_{\text{k}} \text{ or } \text{Compare}(\rho(\Pi), \theta_{\text{sup}}) = \text{true}\}. \quad (7)$$

where $\rho$ specifies constraints that have to be satisfied by the overall pattern set.

### Statistical measures for discriminative patterns

In addition to the measures introduced in the former section, a wide range of statistical measures can be used to evaluate the discriminative power of discriminative patterns. Though some of these measures may be proposed for other tasks, they can be adapted to handle the discriminative pattern discovery problem as well. Considering that there exist numbers of measures, we just present a non-exhaustive review by listing some popular ones for explanation.

For the sake of illustration, we will present the introduction of statistical measures with an example of two-class problem, which can be extended to deal with three or more classes as introduced in [9]. Moreover, a valuable survey of several applicable measures to the multi-class problems has been provided in [50]. Given a pattern $p$ in a data set, for instance $p_1$ in Figure 1, we can construct a generic contingency table as shown in Table 1. In this table, $\Phi$ denotes the absence of the pattern $p$. $t_{ij}$ $(i = 1, 2)$ represents the number of transactions that contain $p$ in $D_i$. Similarly, $t_{0i}$ corresponds to the number of transactions that exclude $p$ in $D_i$. Note that $\text{sup}(p, D_i) = \frac{t_{1i}}{|D_i|}$ and $\text{sup}(\Phi, D_i) = 1 - \frac{t_{1i}}{|D_i|}$. Table 2 lists a number of measures for evaluating the discriminative power of the individual pattern $p$. Notably, [51] has

<table>
<thead>
<tr>
<th>Table 1. A contingency table for a potential discriminative pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
</tr>
<tr>
<td>$\Phi$</td>
</tr>
<tr>
<td>Sum</td>
</tr>
</tbody>
</table>
Table 2. Some popular measures for evaluating the discriminative power of individual pattern in two-class problems, which can be extended to deal with multi-class problems

<table>
<thead>
<tr>
<th>No</th>
<th>Measures</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \text{DiffSup}(p, D_1, D_2) =</td>
<td>\text{sup}(p, D_1) - \text{sup}(p, D_2)</td>
</tr>
<tr>
<td>2</td>
<td>( \text{GrowthRate}(p, D_1, D_2) = \frac{\text{sup}(p, D_1)}{\text{sup}(p, D_2)} )</td>
<td>[10, 35]</td>
</tr>
<tr>
<td>3</td>
<td>( \text{WRAcc}(p, D_1, D_2) = \frac{t_{11} + t_{12}}{t_{11} + t_{12} + t_{21} + t_{22}} )</td>
<td>[12, 15, 52, 53]</td>
</tr>
<tr>
<td>4</td>
<td>( q_p(p, D_1, D_2) = \frac{t_{11}}{t_{11} + t_{12}} ) (g is a user-defined parameter)</td>
<td>[15, 33]</td>
</tr>
<tr>
<td>5</td>
<td>( \text{OddsRatio}(p, D_1, D_2) = \frac{\text{sup}(p, D_1)}{1 - \text{sup}(p, D_1)} )</td>
<td>[34, 51, 54]</td>
</tr>
<tr>
<td>6</td>
<td>( \text{Gain}(p, D_1, D_2) = \sum_{p \in P} \log \left( \frac{\text{sup}(p, D_1)}{\text{sup}(p, D_2)} \right) )</td>
<td>[45]</td>
</tr>
<tr>
<td>7</td>
<td>( \text{SupMaxK}(p, D_1, D_2) = \sup_{p \in P} \left( \text{sup}(p, D_1) - \text{sup}(p, D_2) \right) )</td>
<td>[28]</td>
</tr>
<tr>
<td>8</td>
<td>( \text{MutualInformation}(p, D_1, D_2) = \sum_{i=1}^{\text{num_classes}} \sum_{j=1}^{\text{num_classes}} t_{ij} \log \frac{t_{ij}}{t_{i}.t_{.j}} )</td>
<td>[59]</td>
</tr>
<tr>
<td>9</td>
<td>( \chi^2 = \sum_{i=1}^{\text{num_classes}} \sum_{j=1}^{\text{num_classes}} \frac{(t_{ij} - E_{ij})^2}{E_{ij}} )</td>
<td>[9]</td>
</tr>
<tr>
<td>10</td>
<td>( p-value = \sum_{i=0}^{\text{num_classes}} \binom{t_{ij}}{i} \frac{t_{ij}!}{i! \cdot (t_{ij} - i)!} \frac{t_{ij}!}{i! \cdot (t_{ij} - i)!} )</td>
<td>[67]</td>
</tr>
</tbody>
</table>

Table 3. Some popular measures for evaluating the discriminative power of the pattern set

<table>
<thead>
<tr>
<th>No</th>
<th>Measures</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \text{ItemOverlap}(I) = \frac{1}{</td>
<td>I</td>
</tr>
<tr>
<td>2</td>
<td>( \text{TransOverlap}(I) = \frac{1}{</td>
<td>I</td>
</tr>
<tr>
<td>3</td>
<td>( \text{SumDist}(I) = \sum_{p_i \in I} \sum_{p_j \in I}</td>
<td>D(p_i) \cup D(p_j)</td>
</tr>
<tr>
<td>4</td>
<td>( \text{tile}(p_i, D) = {</td>
<td>T, t_k</td>
</tr>
<tr>
<td></td>
<td>( \text{area}(I) = |\text{tile}(p_i, D) \cup \text{tile}(p_j, D) | \cup \ldots \cup \text{tile}(p_n, D)| )</td>
<td>[47, 55]</td>
</tr>
<tr>
<td>5</td>
<td>( \text{accuracy}(I) =</td>
<td>D(I)</td>
</tr>
<tr>
<td>6</td>
<td>( \text{Laplace}(I) =</td>
<td>D(I)</td>
</tr>
</tbody>
</table>

The equivalence relationship between ‘GrowthRate’ and ‘relative risk’, ‘DiffSup’ and ‘(absolute) risk difference’, OddsRatio and ‘risk ratio’, respectively. Thus, we just provide one calculation formula of these equivalent pairs.

The first eight measures in Table 2 are positively correlated with the discriminative power. That is, the patterns with larger values are generally considered to possess higher discriminative power. On the contrary, smaller values of the last two measures indicate higher discriminative power.

DiffSup and GrowthRate, respectively, assess the absolute difference and ratio of the frequency of one pattern between two classes, while OddsRatio is the ratio of odds of the pattern in one class to that in the complementary class. The weighted relative accuracy (WRAcc) and the generalization quotient \( q_p \) are two popular measures used for evaluating the discriminative power of subgroups. The well-known WRAcc consists of two components: the relative size of a subgroup (generality) and the difference of the ratio of positive transactions containing the subgroup and the ratio of positives in the original data set (relative accuracy). Gain evaluates the discriminative power of patterns by measuring the information gained by including this pattern for distinguishing the target class from other classes. MutualInformation measures the information shared by the occurrences of one discriminative pattern and the class labels of a given data set, which can be thought of as the reduction in uncertainty about the class labels given the knowledge of the discriminative pattern occurrences. Moreover, MutualInformation can be easily extended to process the continuous values. SupMaxK is a family of monotonous interestingness measures for discriminative power, which can be used to prune the search space. They serve a complementary role to the other measures with the ability of discovering discriminative patterns with extremely low frequency especially from dense and high-dimensional data sets. The chi-square statistic \( \chi^2 \) is used to assess the interestingness of patterns in the context of contrast set mining, where \( E_{ij} \) is the expected frequency count in cell \( ij \) of the contingency table. The \( p \)-value derived from Fisher exact test denotes the probability of finding a table where the investigated pattern is more positively associated with the class label.

Note that though the measures \( \chi^2 \), DiffSup and MutualInformation are effective in measuring the discriminative power of one pattern between different classes, they cannot specify the class where this pattern occurs more frequently. On the contrary, the other measures can clearly specify the target class that the pattern will be overrepresented. This difference between these available measures might be of great importance in some particular applications. Additionally, it has been shown that some of the measures are closely related to each other with reasonable deduction [15, 41]. For instance, for a fixed data set \( D \), WRAcc increases monotonically with the support difference and vice versa. SupMaxK is a lower bound of DiffSup so that if \( \text{SupMaxK} \geq \theta_{\text{diff}} \) then \( \text{DiffSup} \geq \theta_{\text{diff}} \) for one pattern. And GrowthRate is proportional to \( q_p \) with \( q = 0 \). These facts explain why some of the interestingness measures are equal alternatives in the real-world applications.

Some statistical measures for evaluating the entire pattern set \( I \) are listed in Table 3. In general, the pattern sets with smaller values of the first two measures possess higher...
discriminative power, whereas the pattern sets with larger values of the last four measures are deemed to be more significant. ItemOverlap and TransOverlap are used to measure the similarity between the discriminative patterns or the transaction sets they are contained in, respectively. Likewise, SumDist collects the symmetric differences between the transaction sets of all pairwise patterns. Instead of the summation, another alternative is to bound the minimum or maximum value over the differences of all pairwise transaction sets. Moreover, the ‘area’ of a pattern set is first proposed in the context of large tile mining [55]. ‘Accuracy’ is defined as the percentage of instances that are correctly covered by a pattern set. Because the size of a given data set is a constant, the calculation of accuracy is equivalent to calculating the coverage difference [56]. Some probability estimates, like the ‘Laplace’ or the m-estimate, can also be used to evaluate the statistical significance of a pattern set [52]. More extensive discussion with respect to such measures can be found in [56–58].

Notice that different significance measures of various types may correspond to different properties such as monotonicity and anti-monotonicity. Furthermore, [23] has demonstrated the feasibility of directly adjusting the relevant principles and constraints of local pattern mining for global pattern discovery. Accordingly, this finding makes it possible for us to adopt the effective measures at the level of individual pattern discovery to measure the global significance in a natural way.

When investigating the discriminative patterns, an important question is how to select an appropriate measure in some specific practical situations. Furthermore, [59] presents an interesting formulation to divide the discriminative patterns into several categories with respect to their different kinds of discriminative power. Notably, the efficacy of one discrimination measure may differ with the distinction of target objectives, data types and discriminative pattern categories. Therefore, choosing appropriate measures for discriminative power evaluation sometimes needs domain knowledge and clear recognition of the nature of problems. In this article, we will not detailedly explain the correspondence relationship between different measures and applications. Please refer to [41, 50] for more details.

**Algorithms and software frameworks**

**Problem statement**

With the explicit formulation of discriminative pattern mining provided in the former section, the main objective of this task is specified as finding a set of interesting discriminative patterns with a desired number of \( \theta_0 \), otherwise, all the patterns must adhere to strict local or global significance constraint with a user-defined threshold of \( \theta_{sig} \). Discriminative patterns are easy to understand and can provide some useful information, which helps explain group differences across class-labeled data. This property is especially important for applications in some specific domains. Nonexperts without domain knowledge could intuitively interpret technical concepts with the analysis of discriminative patterns. Furthermore, the adoption of discriminative patterns can achieve better performance than that of frequent patterns when solving the same problems. For example, as pointed in [6], discriminative patterns are more useful for building high-quality classifiers. However, some additional advantages and challenges are raised in the discovery of discriminative patterns compared with the frequent pattern detection.

First, discriminative patterns are used to show differences between different classes, while frequent patterns present the property with respect to just one class. Discriminative patterns can be considered as an extension of frequent patterns, which are more complex than frequent patterns. Discriminative pattern mining is a relatively more computationally challenging problem compared with frequent pattern mining. On the other hand, as different statistical measures are available to evaluate one discriminative pattern, the induction of rules describing discriminative patterns can be considered as a multi-objective problem. Accordingly, the multi-objective evolutionary algorithms can be adapted to address the problem of discriminative pattern mining [60, 61].

Second, two well-known algorithms for discovering frequent patterns, Apriori [1] and FP-growth [3], respectively, traverse the pattern space in a breadth-first manner and a depth-first manner. Apriori iteratively enumerates and checks all frequent patterns based on an important principle dubbed Apriori property, which describes that if a pattern is frequent then all of its sub-patterns must also be frequent. However, this anti-monotonicity principle does not hold for discriminative patterns over most discriminative measures, which implies that there exists no correlation between one pattern and its sub-patterns with respect to discriminative power. Hence, it is infeasible for discriminative pattern discovery methods to directly use the Apriori framework by replacing the frequency constraint with statistical measures for discriminative power. The Apriori-based approaches have to postprocess frequent patterns or enumerate every possible pattern for discriminative pattern discovery. FP-growth discovers frequent patterns without candidate generation by constructing a frequent pattern tree (called FP-tree). It adopts a divide and conquer strategy and has been shown more efficient than Apriori. Alternatively, one can design an approach for mining discriminative patterns based on FP-growth by either casting discrimination constraints into frequent pattern mining process to discover significant patterns in a single step, or post-processing the frequent patterns with statistical measures for discriminative power in two steps. Furthermore, some efficient and effective methods based on FP-growth [30, 62] have been proposed to discover significant discriminative patterns.

Third, patterns with low frequency but high discriminative power can well describe the differences between different classes, even though it occurs infrequently in the classes. They are also of great value in some specific domains such as bioinformatics and medical informatics, which implies that mining these patterns is practically necessary for scientific research. Discovering significant discriminative patterns with relatively low frequency generally needs high computation cost. However, the pattern discovery methods involving the frequency evaluation possibly ignore these patterns, which results in an incomplete and biased understanding of data.

Notably, this challenging problem of discovering discriminative patterns has received a lot of attention in recent years, and a wide range of software tools and algorithms that use various strategies are available in the literature. Table 4 presents some commonly used and effective techniques with clearly related publications and available implementations. Additionally, the relatively big number of such available algorithms makes a detailed discussion of each technique complex and less feasible. Therefore, to obtain more scientific insights into these accomplishments, we just present some frequently used techniques in this section.
and shift our coherent discussion along the following dimensions:

1. **Mining discriminative patterns on the local level.** Some methods [13, 16–20] usually use measures such as the ones listed in Table 2 to evaluate the discriminative power of each individual pattern, then generate the significant ones as the outcome. Methods of this kind usually conduct the discovery under the following two strategies:
   - **Threshold-based strategy (TB):** Report the patterns that can pass the significance tests with given thresholds, and filter out the insignificant ones according to the definition.
   - **Top-ranking strategy (TR):** Generate a certain number of top-ranked patterns according to their statistical significance.

2. **Mining discriminative patterns on the global level.** Some methods [21–24] perform global statistical significance tests on the entire sets of potential discriminative patterns in addition to considering each individual pattern. The TB and TR are also commonly adopted on the global level. Contrast to local pattern discovery, the global pattern mining algorithms are also concerned with the joint merit and correlation between each pair of patterns. The methods under this formulation can be classified into two categories as follows:
   - **Investigation over all patterns (OA):** Conduct the global significance test on the overall pattern set, and generate the interesting pattern sets with strong discriminative power between different class-labeled data sets.
   - **Investigation over pairwise patterns (PW):** Evaluate the correlation relationship such as ‘redundancy’ and ‘diversity’ between two patterns, then aggregate all combinations of pairwise patterns in some way such as choosing their summation, upper bound or lower bound as the final significance value of the whole pattern set.

3. **Generating exact or approximate sets of discriminative patterns.** Based on exhaustive search, some methods [13, 28–30, 69] provide strong guarantee on the precision of discriminative pattern sets under the user-specified constraints. On the contrary, some methods [31–33, 52] use heuristic strategy to find good yet not necessarily optimal result sets, which possibly contain a certain number of redundant patterns as well as ignore many potentially significant ones.

### Local discriminative pattern mining

On the local level, where considering that every pattern exists independently, all relevant user-specified constraints and evaluations are imposed on each possible individual pattern separately. To discover interesting patterns, one can define the alternative statistical measures as local constraints to conduct the local significance test. More precisely, the methods designed for local pattern discovery generally traverse and check the whole space of potential patterns, apply the local significance test to each generated pattern, subsequently select the significant ones and remove the insignificant ones [19, 28, 29, 64]. Figure 2 presents the brief process of generic discriminative pattern discovery algorithms from the local perspective.

For a given class-labeled data set that consists of extremely high cardinality of items, an explosive number of patterns can be found due to the exponential combinations among items. Hence, conducting a complete search over all the patterns is a time-consuming work. In addition, a large proportion of these patterns are actually meaningless in the real-world applications, which even possibly increase lots of complexity for subsequent study. Accordingly, setting a reasonable threshold to filter out those less significant ones is a considerable choice. Furthermore, the threshold-based approaches [17, 28, 29, 64] have demonstrated their value in the task of pattern mining, which achieve significant reduction of the search space over some particular thresholds. However, selecting an appropriate threshold is practically difficult. This is because most pattern mining algorithms would generate numbers of meaningless patterns if the thresholds are not strict and tight enough, especially when using some insufficient constraints. On the other hand, no patterns will be reported if the constraints
are too restrictive. To address this issue, an available approach is to rank all the possible patterns according to their statistical significance, and then report a certain number of top-ranked patterns [19, 30, 65].

In the context of local discriminative pattern mining, many effective techniques and strategies have been proposed. Some approaches use the common ‘step-wise’ strategy to conduct the search, which first enumerate a set of candidate patterns, and subsequently post-process such patterns to select a smaller pattern set. Typical example of this perspective is to generate the interesting discriminative patterns with respect to the statistical significance based on a complete set of frequent patterns [6, 43–45]. In addition, some anti-monotonic discriminative measures like SupMaxK [28] can help simplify the search procedure. Therefore, the methods can adopt these anti-monotonic measures in the existing Apriori [1] framework for exhaustive and efficient mining. Another strategy is to directly report significant discriminative patterns without generating any candidates. One typical method is DDPMine [32], which searches discriminative patterns directly by constructing a compact FP-tree with the original data. Experiments demonstrate that the direct mining approaches can tackle the efficiency issues raised in the step-wise approaches, whereas possibly fail to guarantee the completeness and accuracy of the reported pattern sets in some situations.

Overall, one recognized drawback of local discriminative pattern mining is that the generated set of patterns is generally too large and complicated to be used directly without further processing for real applications. Because current technologies offer narrow help on the postprocessing process, this drawback makes the task of local discriminative pattern discovery more complicated and time-consuming. Moreover, the independent local constraint on each individual pattern usually results in many variants of essentially equivalent patterns due to the high cardinality of items and the correlation among items. These equivalent patterns might distinguish different classes by covering the same set of transactions, which implies that they share the same level of statistical significance. Practically, keeping all these equivalent patterns is not necessary in some applications and it will increase the degree of redundancy of the result set. Notice that the adoption of the closedness constraint can help eliminate such kind of redundancy by ensuring that a pattern has no super-patterns with the same frequency. Thus, one alternative strategy is to generate closed patterns for a more compacted representation instead [47, 74, 75].

Global discriminative pattern mining
Rather than investigating individual pattern separately, global discriminative pattern mining is concerned with finding a
condensed set of significant and non-redundant patterns. At
the global level, the patterns can be interpreted as the items
that are defined at the local level, and a pattern set is essentially
a disjunction of these patterns. Thus, it becomes possible to
specify global constraints on the whole pattern set. Figure 3 pro-
vides the brief process of general global discriminative pattern
discovery algorithms.

The study of global pattern discovery provides novel insights
into solutions to the redundancy problem raised by local pat-
tern mining algorithms [24, 25, 47, 48]. Contrast to local pattern
mining, the pattern set mining also considers the correlation re-
lationship such as the similarity and diversity between different
patterns. So some local evaluations are alternatives to reduce
the search space and prune the redundancy on the global level.

To date, global pattern mining specifies constraints to ensure
that the whole pattern sets are globally significant, whereas
such final pattern sets are allowable to contain some patterns
which are insignificant and deserve pruning in the local discov-
erly. Moreover, it has been demonstrated that the technique of
global pattern mining generally outperforms some existing local
approaches with a trade-off between the efficiency of execution
and the non-redundancy of reported pattern sets.

For the purpose of finding interesting pattern sets, one feas-
able approach is to formulate the discovery of pattern set as a
global optimization problem with user-specified significance
constraints. For instance, by specifying the optimization prob-
lem as finding a pattern set of size $k$, the approach in [47]
imposes significance constraints on the whole pattern set and
its constituent patterns simultaneously and searches for the significant pattern set in one step. In addition, another approach is to directly post-process the local significant patterns, which is similar to the step-wise strategy used in the local discovery. The corresponding algorithms first adopt some local significance constraints to test individual patterns independently, afterward apply global constraints to eliminating redundant patterns. For instance, one possible post-processing method is to perform greedy sequential covering as introduced in [76].

At each iteration, the method selects the best local patterns and removes all training transactions that contain this pattern. Then, it repeats the same procedure until meeting a stopping criterion, such as all the expected number of patterns are found or all transactions are covered by the target pattern set. However, generating interesting pattern sets with such kind of methods cannot ensure the accuracy of the final result, even though the local pattern set is generated exactly in the first step.

Compared with that in the local pattern discovery, exhaustive search is more difficult and infeasible on the global level due to the enormous candidate sets resulted from large data sets. However, TB and TR are practical to restrict the search space at the global level as well. The huge pattern enumeration space motivates some researchers to choose heuristic search strategy for an approximate discovery instead. For example, some algorithms use a beam search to iteratively discover the significant patterns sets in a breadth-first way. They borrow a local pattern mining algorithm to find the most interesting patterns at each iterative loop, and then use these patterns to generate candidate pattern sets on the current beam. Subsequently, the most significant pattern set is selected according to the overall statistical significance. This work continues until no more candidates can be generated, after which all pattern sets with the strongest discriminative power will be maintained in the final result. Particularly, DSSD (Diverse subgroup discovery) proposed in [24] is a typical example that implements the pattern set mining by extending this heuristic strategy.

The global significance can be obtained in two alternative ways. One is to apply the statistical testing to the entire pattern sets, in which each pattern is one component element and the target investigation objective is the collection of these patterns [24]. The consequent significance value denotes the overall discriminative power of the collection of inclusive patterns under the discussion. Another is to first take individual pairs of different patterns as research objectives to evaluate pairwise discriminative power, then combine the statistical significance values of all pairs to generate a consensus value as the overall significance for the whole pattern set [47, 48]. This strategy can be interpreted as evaluating localized discriminative power of pairs of distinct patterns first, then aggregating the regional significance over two patterns for the comprehensive assessment of the global significance with respect to the entire pattern set.

Generating exact or approximate sets of discriminative patterns

With the same formulation of discriminative pattern mining problem on a given data set, some methods [28-30, 69] focus on finding an exact set that contains all significant patterns, while some methods [31–33, 52] take flexible heuristic strategies to generate as many interesting patterns as possible.

More precisely, the exact methods often generate and check every possible pattern or pattern set that exists in the data sets to achieve the maximum involvement of significant ones. For instance, if a method under the step-wise strategy uses an exhaustive search to generate a complete set of candidates in the first step, subsequently adopts a rigorous post-processing approach with integrity assurance, then this method must be able to report an optimal pattern set with respect to its own definition. Similarly, the direct mining algorithms can also provide strong guarantee on the precision of discriminative pattern sets by traversing the whole space of possible patterns at the local level or all pattern sets at the global level. In addition, when the cumbersome exhaustive search is not possible or the requirement on accuracy is not so rigorous, it is feasible to sample a representative set of significant patterns instead of producing a vast amount of all patterns. This motivates some algorithms to adopt heuristic strategies to produce an approximate set of discriminative patterns for some particular studies. For example, sampling algorithms in [31] pursue a goal of directly generating sets of significant patterns from the pattern space according to given interestingness measures such as frequency, squared frequency, area and discriminativity. The good enough yet not necessarily optimal pattern sets possibly contain a certain number of redundant patterns as well as miss many potential significant ones. However, the heuristic-based methods have been demonstrated their superiority of efficiency to the exhaustive mining algorithms when generating an equal number of patterns. Hence, the heuristic strategy is a considerable alternative to exhaustive search when the given data set is too dense and huge.

In general, the methods under different strategies can achieve different and comparative performance under the same settings. Some methods possibly possess relatively superior capability of detecting sets of discriminative patterns, whereas they may also involve some disadvantages such as containing redundant patterns, or not covering all expected patterns. Accordingly, making an optimal determination on selecting the most suitable method is very difficult in different practical situations. The combination of practical consideration and technical analysis is indispensable in essence. To cope with a specific problem, one can first categorize this problem according to the characteristics of different categories as discussed above, and then select a desired method that falls into this category. For example, if the users focus on finding interesting, non-redundant and representative patterns, the global pattern discovery methods are preferred than the methods designed for local pattern discovery. More precisely, the method PattCP, cop [25, 47] and pattset_comet [48] cannot due to their heuristic search. If the requirement on the completeness of significant pattern sets is very high, PattCP_cop is superior to DSSD and pattset_comet. In contrast, as the methods under heuristic strategy generally perform more efficiently than exhaustive search methods when generating the same number of significant patterns, the heuristic-based methods are better if the users put more emphasis on the efficiency of discovery. Similar analysis can be achieved when mining discriminative patterns on the local level as well. On one hand, BSD (the bitset based subgroup discovery algorithm) [64] and NMEEF-SD (Non-dominated Multi-objective Evolutionary Algorithm for Extracting Fuzzy rules in Subgroup Discovery) [71] have been demonstrated to have superior efficiencies to the other methods such as Apriori-SD [29], SD-Map [30] and SD-Map⁺ [69] for conducting the local exhaustive search. Moreover, the methods M⁺bT (Model-based Search Tree) [19] and SMP [28] can discover significant patterns with extremely low support within an acceptable amount of time. On the other hand, the heuristic-based approaches can produce a
number of significant patterns without traversing the whole pattern space, which indicates that less running time is needed with such methods. Particularly, Rules for Contrast Sets (RCS) [60] and Magnum Opus [13, 73] are derived from association rule discovery, which implies that other rule discovery methods might also be able to find discriminative patterns. Overall, to address complex pattern mining problems, it is feasible to combine different methods to conduct the discovery in an iterative way and gather all their advantages to improve the joint performance under the suitable conditions.

Bioinformatics applications

The techniques proposed for discriminative pattern mining have shown great use in solving a wide range of bioinformatics problems. First, discriminative patterns can provide valuable insights into identifying the biological differences between various biological data sets. These differences are of great importance for us to better understand the essence of biological problems and further develop more effective solutions. The investigation of discriminative pattern mining opens a new door to in-depth analysis of biologically relevant knowledge. Second, the problem of discriminative pattern mining has been investigated extensively in the past few decades. As a result, many algorithms and software tools are publicly available. This implies that if one bioinformatics problem can be modeled as a discriminative pattern mining problem, there will be many existing algorithms and tools that can be directly used to address this bioinformatics problem. Third, discriminative pattern mining was rooted from data mining society, in which a great deal of efforts are devoted to advancing the efficiency of pattern discovery. A majority of the relevant approaches potentially possess fast running speed and the ability of handling large-scale data [19, 28, 30, 62, 69, 71]. Hence, an important advantage of efficiency can be obtained when applying the methods of discriminative pattern discovery to dealing with the problems in bioinformatics compared with other statistical methods. Moreover, the methods for discriminative pattern mining have demonstrated their efficiency and effectiveness to address a broad range of bioinformatics problems, some of which are even superior to the state-of-the-art methods under the same conditions [16, 34–40]. Last but not least, many algorithms use the exhaustive search to discover a complete set of significant patterns according to their definitions, which can potentially increase the chance of identifying all interesting patterns with biological significance.

In this section, we describe several typical applications of the discriminative pattern mining techniques in bioinformatics with detailed analysis. One of the greatest challenges is how to select a complete set of actually significant and nonredundant patterns that are of really biological significance. Most of significance constraints in discriminative pattern mining are defined without considering biological knowledge, which might result in a certain number of redundant patterns that lack real biological relevance. Then, the direct applications of pattern discovery approaches to complex problems in bioinformatics usually require some additional considerations to quantify biological significance. Therefore, bridging the gap between statistical measures for discriminative power and the actual biological significance is still an open problem.

Phosphorylation motif discovery

Protein phosphorylation is one of the most frequent post-translational modification (PTM) events for the regulation and maintenance of most biological processes. This event is of great interest in numerous key cellular processes such as the regulation of cellular activities. Phosphorylation motifs represent common amino acids aligned upstream and downstream of the phosphorylation sites. The discovery of phosphorylation motifs pursues a goal of finding a set of motifs that occur with disproportionate frequency in two coordinate sequence data sets: one is the phosphorylated peptide set and the other is the unphosphorylated peptide set [34, 35, 77–80]. The identified interesting motifs are all overrepresented ones in the phosphorylated peptide set relative to the unphosphorylated peptide set (Because many phosphorylation sites are still unrecognized, the unphosphorylated peptide set might contain some peptides that actually can be phosphorylated as well).

To detect all interesting phosphorylation motifs, the relevant methods typically assess two aspects of each motif including frequency and statistical significance, which derive from the same definitions as those in discriminative pattern mining. The threshold-based strategy is commonly used in phosphorylation motif discovery: only those frequent motifs that can pass the significance threshold are reported to the user. Hence, such motifs with significant frequency differences can provide useful value in explaining the discrimination between the phosphorylated and unphosphorylated peptide sets, which is equivalent to the efficacy of discriminative patterns in data mining. Moreover, the given data sets in phosphorylation motif discovery can be divided into two classes by considering the property of (un)phosphorylation as class labels. As a consequence, three mapping relationships between these two tasks can be observed. First, by taking the position-specific amino acid as set as the collection of all possible items, the given peptides in phosphorylation motif discovery correspond to the transactions in discriminative pattern mining. Second, the investigated motifs can be considered as the target objectives when mining discriminative patterns on the local level. Last, the frequency and statistical significance of phosphorylation motifs can be, respectively, calculated by the support and statistical measures used for assessing the discriminative power as listed in Table 2. Particularly, considering phosphorylation motifs are generally overrepresented in the phosphorylated peptide set, the statistical measures except for $\chi^2$, DiffSup and MutualInformation in Table 2 are more appropriate to assess the interestingness of phosphorylation motifs. Therefore, the task of phosphorylation motif discovery can be considered as an example of local discriminative pattern mining in bioinformatics.

The problem of phosphorylation motif discovery has been widely explored and several effective algorithms have been proposed based on data mining techniques. Two effective approaches, Motif-All [34] and C-Motif [35], are representative ones which tackle the problem of mining phosphorylation motifs from the viewpoint of mining discriminative patterns. To measure the statistical significance of one potentially interesting phosphorylation motif, Motif-All uses OddsRatio while C-Motif uses ‘relative risk’ (i.e. GrowthRate), which are, respectively, defined in Table 2. More precisely, Motif-All uses the frequency test to mine a set of frequent motifs as candidates first, and then adopts statistical significance evaluation to filter out the insignificant ones and reports the significant ones, whereas C-Motif conducts these two assessments in a single stage to directly generate the significant motifs. In general, compared with other existing motif discovery methods [77–80], Motif-All can discover the largest number of interesting motifs whereas they often contain a large amount of redundant motifs whose discriminative power mainly derives from their subsets, while
C-Motif can achieve a relatively good trade-off between the high coverage and the nonredundancy of significant motif set. In addition, [79] claims to report the maximal phosphorylation motifs that are not subsumed by motifs with more fixed amino acids, which is consistent with the concept of maximal patterns. Figure 4 illustrates the workflow of phosphorylation motif discovery with the techniques of discriminative pattern mining. For a possible motif, the frequency is measured by calculating its support value, and its statistical significance can be obtained by any measures provided in Table 2.

Figure 4. The workflow of phosphorylation motif discovery with the techniques of discriminative pattern mining. For a possible motif, the frequency is measured by calculating its support value, and its statistical significance can be obtained by any measures provided in Table 2.

To conclude, discriminative pattern mining provides a group of effective tools for the task of phosphorylation motif detection, which also helps facilitate the reduction of the number of phosphorylation motifs returned to the biologists. Generally, the reported motifs are theoretically useful in explaining the phosphorylation mechanism, and actually have the potential to represent important biological knowledge. However, as not all the methods proposed for discriminative pattern mining can guarantee the completeness of their discovery, there are possibly many statistically significant motifs that remain undetected. In addition, the detection approaches based on discriminative pattern mining could ignore the biological meaning of each motif and the interplays between different motifs, which probably report a complex collection containing some redundant motifs with little use for further research. Therefore, directly applying discriminative pattern mining techniques to the discovery of phosphorylation motifs might increase some attached challenges in biological research.

Identifying differentially expressed genes

The expression level of genes can be different in various biological conditions, cell cycle states, tissues and subjects. Identifying differentially expressed genes is one of the most popular topics in bioinformatics, which captures interesting information about distinction in different types of tissues such as the cancerous and normal tissues [81–84]. Such genes are useful for us to get a better understanding of the complex mechanisms with respect to some incurable diseases.

The advances in relevant technologies make it possible to measure the expression levels of thousands of genes simultaneously with a single experiment, rendering an enormous amount of high-dimensional data. Accordingly, traditional methods for differentially expressed gene identification become less feasible to handle such high-dimensional gene expression data effectively. So it is considerable to apply some existing efficient methods derived from other fields such as machine learning and data mining to conduct the comparative analysis of gene expression across different conditions. Furthermore, considering the differentially expressed genes possess the power of discriminating and classifying two different classes, discriminative pattern mining can be used to tackle the problem of identifying significant genes with different expression.

Some discriminative pattern-based algorithms have already been proposed [16, 36, 37, 40]. The method in [16] uses the Fisher exact test to generate a P-value, while the other methods in [36, 37, 40] directly adopt GrowthRate to measure the discriminative power of genes. In fact, other statistical measures listed in Table 2 are alternatives for assessing the significance of each individual gene set as well. These discriminative pattern-based algorithms create some versatile and efficient tools for elucidating the genes with discriminative expression and illustrating their applications in biology. Because gene expression data are continuous, data discretization is a necessary preprocessing step to apply discriminative pattern analysis. The discretization method [85] and its implementation [86] are used in [36, 37, 40] to partition gene expression levels into a number of suitable intervals with the idea of entropy minimization. Importantly, the experimental results on the colon tumor data set demonstrate that the discretization step actually removes many noisy genes and effectively explores the remaining most
discriminatory genes [36]. Though the colon tumor data set contains 2000 genes, only 35 genes are partitioned into two disjoint intervals, while there is no cut point in the remaining 1965 genes. This implies that only 1.75% of the genes are considered as the most discriminatory genes and the others can be regarded as irrelevant indicators. Hence, genes that are potentially the most discriminative are selected and a large number of noisy genes are removed, which helps narrow down the search for significantly differentially expressed genes.

The methods of discriminative pattern discovery handle the problem of identifying differentially expressed genes with several assumptions. First, they formulate the combinations of intervals of gene expression levels of possibly relevant genes as the discriminative patterns by considering each gene as an individual constituent item of patterns. This provides potential opportunity to discover some groups of different genes that work together to discriminate different tissues. Second, they directly adopt the significance tests for measuring the discriminative power of patterns to quantify the differential expression level of genes. Last, they conduct the identification of interesting genes commonly with the techniques of local discriminative pattern mining rather than those of global pattern mining. The main reason is that discovering significant patterns at the global level is a relatively new formulation to the field of data mining and bioinformatics, which still involves some problems and requires more attention with respect to the practical applications.

The identification of differentially expressed genes plays vital roles in many biological applications, ranging from the selection of disease-related biomarkers to the discovery of disease diagnosis and treatment development [36, 37, 81, 83]. The methods in [36, 40] only focus on the discriminative patterns with the maximum frequency difference between different tissues, and experiments demonstrate that they can successfully discover groups of significant genes that are correlated to specific disease states. Furthermore, the discovered genes provide useful information on distinguishing normal cells from cancerous cells, based on which more accurate classification can be obtained than the support vector machine [87] and a clustering-based method [88].

However, there exist some questions if the discriminative pattern mining methods are applied to identifying differentially expressed genes in a straightforward manner. For instance, the investigations directly using the statistical measures for discriminative patterns might produce some complex groups of genes, which are difficult for further use owing to the lack of biological interpretation of the meaning and the correlation of genes. This might result in a certain degree of redundancy: including numbers of biologically meaningless genes while excluding some significant ones. To address this issue, adopting the global pattern discovery techniques is an available choice, which is formulated with some local constraints on the relationship between different pairs of patterns in addition to the global evaluation. Therefore, it is highly feasible to generate differentially expressed genes with global pattern discovery, and take some additional biological investigation into account if necessary.

In addition, some methods of local discriminative pattern mining take the step-wise strategy, which first generate frequent patterns in one class and then assess their statistical significance to remove the insignificant ones [6, 43–45]. Directly applying these methods to discovering differentially expressed genes can result in that a certain number of genes with low occurrences in one class are missed. However, these genes might also possess relatively high discriminative power and provide significant information of discrimination between different tissue types. They are potential to be of considerable importance in some practical situations such as cancer detection [37]. Hence, from the biological perspective, generating the genes with low frequency is actually required as well. However, detecting these less frequent genes could increase the challenges of applications of pattern mining algorithms in high-dimensional gene expression data [37, 89]. The discovery procedure becomes very time-consuming due to the exponentially increasing search space. In general, even though the adoption of discriminative pattern mining contributes to identifying differentially expressed genes greatly, it is insufficient to offer much detailed and understandable analysis for the biological research in some situations.

Identifying discriminative genotype patterns

Genetic factors strongly influence susceptibility to common complex diseases and related quantitative traits. One of the greatest challenges for human genetics is to identify and characterize the disease-relevant genes, which is of great importance for clinical and scientific applications. Recent advances in high-throughput technologies generate large-scale data collections on genomic variation, rendering the availability of identifying the genetic variants and their closed association with some specific diseases and other complex traits. A good knowledge of the mapping relationship between DNA sequence variants and susceptibility to diseases can facilitate us to propose innovative and far-reaching insights to deal with the problems of disease diagnosis, prevention and elegant treatment. An increasing number of approaches have been proposed to discover discriminative genotype patterns which possess strong discriminative power of classifying different phenotype classes [90, 91]. More importantly, the techniques of local discriminative pattern mining have proven their great value for this task [38, 39]. The approach in [39] illustrates that the statistical measures $\chi^2$, OddsRatio and the P-value derived from the Fisher exact test are available to assess the discriminative power of one discriminative genotype pattern. The method in [38] computes an exact hypergeometric score to generate a P-value to measure the significance of one potential discriminative genotype pattern. Moreover, other statistical measures defined in Table 2 are also available.

Note that the majority of DNA sequence variants are single nucleotide polymorphisms (SNPs). A popular problem of identifying discriminative genotype patterns is to generate allelic combinations of SNPs, which are significantly associated with a phenotype of interest whereas each member SNP therein may have little effect. Different combinations of SNPs might have different mapping consequences at the phenotypic level, which provides useful information for separating a certain group of phenotypes from others. These SNPs are routinely analyzed as markers for diagnosis in genomics, bioinformatics and many other disciplines.

Traditional approaches to finding significant sets of SNPs involve several challenges such as low computational efficiency, incomplete coverage of expected SNPs and deficient statistical constraints. These disadvantages are magnified and more complex when addressing the problem of high-order SNP combination identifications especially in dense and high-dimensional data sets. The so-called high-order combinations of SNPs are generally considered as the ones with larger sizes, which
Consist of at least more than two SNPs. Accordingly, there exists a huge hypothesis space to explore and the search space even increases along with the exponential possibilities of high-order combinations due to the enormous number of existing SNPs. Therefore, the methods with brute-force strategy are not capable of efficiently completing this difficult task, while those with heuristic strategy successfully enhance the search efficiency without guarantee on the completeness of their results. Actually, the effective approaches in [92–95] have demonstrated their ability in tackling the problem within the limited time, yet fail to efficiently produce an optimal set of significant SNPs.

To address this issue, the techniques of discriminative pattern mining provide a potential opportunity to achieve a trade-off between the efficiency of discovery and the quality of results. Additionally, the methods derived from discriminative pattern mining have already shown superior performance to the traditional methods under the same conditions [38, 39]. The method in [38] uses greedy search to generate significant SNP combinations whereas provides no guarantee on the completeness. In contrast, the method in [39] claims to provide a complementary method based on discriminative pattern mining to the existing approaches and focuses on discovering high-order combinations in case-control data sets. This method also combines some gene-set constraints with an effective framework of discriminative pattern mining designed in [28]. Experiments demonstrate that this method outperforms the statistical methods in [92, 93] under the same setting. However, directly applying discriminative pattern discovery to identifying discriminative genotype patterns also involves the problem of filling the gap between statistical significance and biological interestingness.

Some other applications
The techniques of discriminative pattern mining can also be applied to addressing some other complex problems in bioinformatics. According to the introduction of three specific biological problems in the former section, discriminative patterns can be readily transformed into the biological concepts such as phosphorylation motifs, differentially expressed genes and discriminative genotype patterns. Accordingly, we can deduce that the techniques of discriminative pattern mining can also be used to generate other biological patterns only if they possess equivalent ability of capturing significant differences between different classes. As a consequence, discriminative pattern mining is also of interest for other comparative analysis in bioinformatics. For example, discriminative pattern mining techniques can be further applied to identifying PTM crosstalk motifs [96] and tissue-selective transcription factor binding sites [97] as well.

In addition, discriminative pattern mining algorithms are of great use for the problems in other domains. For example, in medical informatics, several successful applications of discriminative pattern mining have been reported such as distinguishing similar diseases [18, 98] and interpreting positron emission tomography (PET) scans [99]. The techniques of discriminative pattern mining have also been used to perform the identification of activating structural features associated with toxicity [100, 101].

Conclusion
This article provides a straightforward introduction of discriminative pattern mining, and bridges the gap between the fields of data mining and bioinformatics by investigating how standard discriminative pattern mining techniques can be applied to solving a wide range of biological problems. We formally present a uniform definition of discriminative pattern mining and some frequently used statistical measures for discriminative power. A number of popular and effective open-source algorithms are also presented.

Discriminative pattern mining techniques have demonstrated their great value in tackling numbers of difficult problems in bioinformatics. They offer some effective tools to discover significant biological patterns from biological data sets. Furthermore, they can also provide valuable insights for us to obtain a better understanding of the underlying nature of biological problems and further developing more effective solutions. However, there are still several challenges.

First, the local pattern mining methods have drawn lots of attention and many local discriminative pattern-based methods in bioinformatics have already been proposed, whereas mining patterns at the global level is a relatively new formulation in data mining and bioinformatics with few applications. In addition, the local discovery often involves many redundant patterns with difficulties in direct interpretation and application, whereas the global discriminative pattern mining generally imposes some constraints to evaluate the correlation relationship between different patterns for the purpose of reporting more preferred and nonredundant sets of significant patterns. This provides an important information about the potential advantage of adopting global pattern discovery methods to deal with bioinformatics problems over using local discovery approaches, as the interplays of patterns can change the interestingness of individual patterns.

Second, directly applying discriminative pattern mining to generating biological patterns can result in several problems. In data mining, one pattern is considered to be significant on condition that it can satisfy some user-defined constraints mostly from the mathematical and statistical perspectives. However, such a definition as well as the constraints may not be suitable for the patterns in biological data. Adopting the significance investigation designed for discriminative pattern mining possibly cannot guarantee to help us find a complete set of actually interesting patterns with biological significance. Some additional consideration and domain knowledge are necessary. Therefore, it is still a challenging task to produce a certain number of really interesting biological patterns in the context of bioinformatics.

Third, the frequently used TB and TR can help reduce the search space and save the computational cost, which are defined under the assumption that the patterns with high significance are more valuable in practical situations. However, such a definition may raise some questions when dealing with the bioinformatics problems. For example, the genes with relatively low significance values might be useful to distinguish different tissues as well. Hence, such genes are possibly necessary in bioinformatics study but they may be filtered out by discriminative pattern mining approaches. In contrast, some genes with high statistical significance are reported to the biologists even if they might have little practical effect. Therefore, more difficulties will be involved in selecting an appropriate value for either the significance threshold or the expected number of significant patterns in bioinformatics applications.

Last, more efforts should be devoted to improving the efficiency and completeness of the algorithms for generating significant patterns especially from dense and high-dimensional data for both data mining and bioinformatics communities.
Key Points

- Discriminative pattern mining aims at finding a set of interesting patterns that occur with disproportionate frequency in some classes versus others, such as differentially expressed genes or proteins in patients and healthy persons.
- Discriminative pattern mining techniques can be readily applied to solving some bioinformatics problems and they provide great insights to identifying biologically significant patterns with discriminative knowledge across different biological data sets.
- We elucidate several typical applications of the discriminative pattern mining techniques in bioinformatics with detailed analysis, including phosphorylation motif discovery, differentially expressed gene identification, discriminative genotype detection, etc.

Funding

This work was partially supported by the Natural Science Foundation of China under Grant No. 61003176 and the Fundamental Research Funds for the Central Universities of China (DUT14QY07). The comments and suggestions from the anonymous reviewers greatly improved the paper.

References


---

**Discriminative pattern mining and its applications | 899**