Discovering Combinatorial Disease Biomarkers

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Gang Fang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

VIPIN KUMAR, advisor

August, 2012
Acknowledgements

I am indebted to my advisor Vipin Kumar for his help, advice and support in all aspects of my thesis and research career. Five years ago, I decided to shift from the field of pattern recognition to data mining for a PhD. Vipin offered me such an opportunity in one of the best data mining groups and introduced me to the exciting field of biomedical informatics and computational biology (which I did not know at the beginning and planned to only pursue theoretical data mining research). His passion in doing high-impact work and solving real-world problems initiated and then boosted my interests in advancing data mining algorithms and applying them to challenging problems in computational biology. His expertise helped me accumulate rich research skills and methodology. Beyond that, he also taught me how to think, communicate, present, lead and collaborate. Most importantly, I feel so fortunate to see and learn from his great personality and big vision. His mentorship goes way beyond my PhD, more like a lighthouse in my entire life. This thesis would not have been possible without his advises on research and the flexibility he gave me to pursue different research ideas right from the beginning of my PhD study when most of my ideas were immature.

Great gratefulness to my committee members alphabetically: Rui Kuang, Kelvin Lim, Chad Myers and Brian Van Ness, for their help on this thesis and most of of my PhD projects. Rui’s supervision is critical on the SDC work and on my development of research methodology in general. The collaboration with Kelvin opened the door to fMRI data analysis and multimodal data integration. The discussions with Chad motivated the study of pathway level human genetic interactions and his advise on general research methodology is inspiring. The collaboration with Brian started all the discriminative pattern mining work and his advises are very helpful from a biologist perspective. Special gratefulness to Michael Steinbach, who is certainly a co-advisor
to me and who I sincerely respect. Many thanks also go to the other collaborators from whom I learned the importance of emphasizing biological impact when developing computational methods: Angus McDonald, Mike Miller, Bill Oetting, Tim Church, Chris Wendts and Bonnie Westra.

The folks in the DMBio group made my PhD years more colorful. Gaurav Pandey gave me advises beyond what I could imagine one can get from a labmate or a friend. Rohit Gupta taught me things in research and life that I could hardly learn from anyone else. Great thanks to Wen, without her help, several projects in this thesis would not have been possible. Deep thanks also go to other folks in the group who helped me a lot: Gowtham Atluri, Shyam Boriah, Vanja Paunic, Xiaoeye Liu, Benjamin Oatley, Sanjoy Dey, Sean Landman; other computational biology folks in CSE: Ze Tian, Taehyun Hwang, Wei Zhang, Baryun Hwang, Raamesh Deshpande, Benjamin VanderShuis, Elizabeth Koch and Yungil Kim. Also thank my other friends in CSE: Qiang Fu, Dingcheng Li, Hanhuai Shan, Ruipeng Li and Huahua Wang.

The two summer internships at Pacific Biosciences and the following collaboration are also critical to the development of my research skills and methodology in analyzing second- and third- generation sequencing data. I owe deep gratefulness to Eric Schadt who offered me these opportunities and gave me experienced advices and inspired me on targeting high impact biological problems. I am fortunate to interact with and learn from the smart and nice colleagues at PacBio: Onureena Banerjee, Zhixing Feng, Jason Chin, Khai Luong, Tyson Clark, Andrey Kislyuk, Jim Bullard, Ali Bashir, Patrick Marks, John Eid, Jonas Korlach, Steve Turner. Discussions with the professors in Mount Sinai School of Medicine are also quite inspiring: Andrew Karsaskis, Bin Zhang, Pamela Sklar, Shaun Purcell and Joseph Buxbaum.

I acknowledge NSF, the BICB fellowship and the Walter Barnes Lang fellowship for supporting my PhD research.

Last but the most important, the constant support from my family gave me all the energy and passion for every day. Deep gratefulness and love to my parents, Jichang Fang and Lili Dong, for their unconditional love and support in the past thirty years. The deepest thanks and love to my wife, Wei Shi, who held me up through the hard moments in the past five years and made every single day meaningful.
Dedication

To my grandma, Gao Shuqin, and my grandpa, Dong Haihuan.
Abstract

Many diseases have a genetic component. Some, including many cancers, are caused by a change in the functioning of a gene or a group of genes in a person’s cells. Disease-biomarker discovery seeks to find the association between diseases and a person’s genetic or associated characteristics, such as genes, DNA mutations, methylations, non-coding RNAs, proteins, metabolic products, and biological pathways. These biomarkers, such as the mutations in the BRCA1 and BRCA2 genes that indicate a high risk of breast cancer, can help in understanding the mechanisms causing a disease, and can guide diagnosis, prognosis and treatment. With the recent availability of high-throughput ”-omics” and next-generation sequencing data, biomarker discovery is shifting from hypothesis-driven analysis towards data-driven analysis, which enables the discovery of previously unsuspected genetic associations for a variety of diseases. However, for most diseases, there remains a substantial disparity between the disease risk explained by the discovered loci and the estimated total heritable disease risk based on familial aggregation, a problem that has been referred to as ”missing heritability”. While there are a number of possible explanations for missing heritability, genetic interactions between loci are one potential culprit. Genetic interactions generally refer to two or more genes whose contribution to a phenotype goes beyond the independent effects of the genes and are expected to play an important role in complex diseases. This thesis takes a data mining based approach, specifically discriminative pattern mining, and targets the computational discovery of combinatorial biomarkers associated with complex human diseases from a variety of large scale case control genomic datasets. It addresses several key challenges confronted by existing discriminative pattern mining algorithms: computational complexity, sample heterogeneity due to disease subtypes and lack of statistical power for most real datasets. It also proposes a novel concept to organize discriminative patterns into an interaction network that allows the discovery of high-level structural knowledge, in both global and local scales. Specifically, a general framework is proposed to detect pathway-pathway interaction pairs that are enriched for genetic level interactions from genome wide association datasets. Validations across independent real datasets not only demonstrate the reliability of the proposed framework but also lead to
several interesting biological insights on several complex diseases such as breast cancer and Parkinson’s disease. The data-mining algorithmic contributions in this thesis also hold promise for addressing generic challenges in other domains beyond biology.
# Contents

Acknowledgements i  
Dedication iii  
Abstract iv  
List of Tables xi  
List of Figures xii  

1 Introduction 1  
1.1 Background ................................. 1  
1.2 Challenges and Objectives .......................... 2  
1.3 Contributions and Outline of following chapters ................. 4  
1.4 Relationship to a broader set of data mining and machine learning approaches ........................................ 8  

2 Efficient Discriminative Patterns Mining from Dense and High-dimensional Binary Data 10  
2.1 Introduction .................................. 10  
2.2 Background .................................. 11  
   2.2.1 Contributions of this chapter ....................... 17  
2.3 Basic Terminology and Problem Definition ......................... 18  
2.4 Computational Limitations of Current Approaches ................. 18  
2.5 Proposed Approach ................................ 20
2.5.1 SupMax1: A Simple Measure to Start with .......................... 20
2.5.2 SupMaxK ............................................................... 21
2.5.3 Properties of the SupMaxK Family ................................. 21
2.5.4 Relationship between DiffSup, BiggerSup and the SupMaxK Family 24
2.5.5 SupMaxPair: A Special Member Suitable for High-Dimensional
Data ................................................................. 25
2.6 Framework for Discriminative Pattern Mining .......................... 27
2.7 Experimental Results .................................................... 27
2.7.1 Experiments on Synthetic Data Sets with Varying Density and
Dimensionality ....................................................... 28
2.7.2 Experiments on a Breast Cancer Gene Expression Data Set ... 32
2.7.3 Summary of Results .............................................. 45
2.8 Related work .......................................................... 45
2.9 Conclusions ............................................................. 46

3 Continuous Data, Different Types of Discriminative Patterns and Novel
Algorithms for Efficient Discovery ........................................ 49
3.1 Introduction ............................................................. 49
3.2 Background ............................................................. 50
3.3 Proposed Approach ..................................................... 53
3.3.1 Subspace Differential Coexpression Analysis ..................... 53
3.3.2 Computation Algorithm .......................................... 56
3.3.3 DiffRange: an illustration of SDC ................................. 57
3.4 Experimental Results ................................................... 58
3.4.1 Datasets and Preprocessing ..................................... 58
3.4.2 Pattern Discovery .................................................. 59
3.4.3 Are the discovered subspace differential coexpression patterns statistically significant? .................. 60
3.4.4 How differentially coexpressed are the discovered subspace patterns when measured over full-space? .......... 60
3.4.5 Are the discovered subspace differential coexpression patterns biologically relevant? ................................. 63
4 A Generic Strategy for Enhancing the Statistical Power of Discriminative Pattern mining

4.1 Introduction ................................................. 70
4.2 Background ................................................. 71
4.3 Analysis and Results ........................................ 74
  4.3.1 Three real case-control SNP datasets and a synthetic dataset .... 74
  4.3.2 The Binary Encoding of a SNP and a Combination of SNPs .... 75
  4.3.3 Illustrative Examples of High-order Discriminative SNP Combinations ........................................ 77
  4.3.4 Discovering High-order SNP Combinations that have Strong Association with a Phenotype ........ 78
  4.3.5 The DPM framework has substantially better efficiency and scalability ........................................ 80
  4.3.6 Identifying High-order SNP Combinations with Stronger Association than their subsets ........ 82
  4.3.7 Many high-order patterns are trivial extensions of their smaller subsets ........................................ 84
  4.3.8 Some high-order patterns are highly discriminative beyond univariate and low-order SNP-combinations .... 85
  4.3.9 A Genetic procedure that generally enhance statistical power of discriminative pattern mining ........ 86
  4.3.10 Exploring functional interactions in high-order combinations .. 88
4.4 Discussion .................................................... 94
4.5 Supplementary Methods ..................................... 97
  4.5.1 Three SNP datasets and pre-processing considerations ........ 97
  4.5.2 Simulation of a synthetic case-control SNP dataset ........... 97
  4.5.3 The Apriori framework: efficient combinatorial search with anti-monotonic objective function ........ 98
  4.5.4 The anti-monotonic objective function SupMaxPair .......... 98
6.6 Supplementary Methods ........................................... 143
6.6.1 GWAS SNP datasets, data quality control and balancing population stratification ........................................... 143
6.6.2 Estimating SNP-SNP genetic interactions ....................... 144
6.6.3 Two possible directions for handling redundant SNP-SNP interactions due to linkage disequilibrium .......................... 145
6.6.4 Selecting a less-redundant set of SNPs ............................ 146
6.6.5 Calculating all pairwise SNP-SNP interactions for the SNPs in $S_0$ 146
6.6.6 Estimating pathway-pathway interaction based on SNP-SNP interaction network: a conceptual view ....................... 146
6.6.7 Estimating pathway-pathway interaction based on SNP-SNP interaction network: technical details ......................... 147
6.6.8 Correcting for multiple hypothesis testing ......................... 148
6.6.9 Properly handle overlaps among a set of BPMs .................... 149
6.6.10 Stability of the proposed framework over different parameters and insights on parameter selection .......................... 150

7 Conclusion and Discussion .............................................. 159

References ................................................................. 161
List of Tables

2.1 Number of type-I and type-II discriminative patterns of size-2, 4, 6, 8 and 10. 29
2.2 Details of patterns discovered by CSET at various $BiggerSup$ thresholds. 34
(* Expansion of the set of patterns to patterns of larger sizes could not finish in over 12 hours, and thus, their results are not included here.)
2.3 Details of patterns discovered by SMP at various $SupMaxPair$ thresholds. 35
2.4 Precision-recall results of CSET patterns with $BiggerSup \geq 0.4$ (Pre: Precision, Rec: Recall, Expected Precision for random gene collections is 10.2%, ERec: Expected Recall of Random Gene Collections with the Same Size) 41
2.5 Precision-recall results of SMP patterns with $SupMaxPair \geq 0.18$ (Pre: Precision, Rec: Recall, Expected Precision for random gene collections is 10.2%, ERec: Expected Recall of Random Gene Collections with the Same Size) 42
5.1 Details of each dataset and a summary of the number of different types of discriminative patterns discovered. For column 7 – 9, in addition to the number of discovered patterns, we also show (in the bracket) the number of unique items in the union of the set of patterns to reflect the redundancy among the patterns. $\delta = 0.1$ and $j = 0.05$ are used for all the datasets. 121
List of Figures

2.1 A sample data set with interesting discriminative patterns ($P_1, P_4$) and uninteresting patterns ($P_2, P_3$) ................................................................. 12

2.2 An illustration of the coverage of the space of discriminative patterns by different approaches given the same amount of time. The $t$’s on the y-axis represent the lowest support of the patterns that are respectively covered by the corresponding approaches (represented by boxes), given the same and fixed amount of time. (a) Box $A$, $B$ and $C$ represent the set of patterns discovered by the corresponding approaches in group $A$, $B$ and $C$ respectively. (b) Illustration of the tradeoff between the capability to search low-support discriminative patterns in dense and high-dimensional data and the completeness of the pattern discovery. Boxes $X$, $Y$ and $Z$ represent three conceptual low-support discriminative pattern mining approaches that discover patterns not found by the approaches in groups $A$, $B$ and $C$. Note that, in this figure, the set of interesting discriminative patterns is the same as that in (a), but the corresponding *’s are not shown for the sake of clarity. ................................................................. 15

2.3 An extended version of the data set shown in Figure 2.1 containing fifteen addition items ($i_{16} - i_{30}$) and two patterns $P_5$ and $P_6$, the rest being identical to Figure 2.1. ................................................................. 23
2.4 Nested layers \((L_1, L_2, L_3, \ldots, L_k, L_{k+1}, \ldots, L_{All}, L_{CSET})\) of patterns defined by \(SupMaxK\), and relationship with the complete set of discriminative patterns (layer \(L_{All}\)), and the search space of \(BiggerSup\) used by CSET (layer \(L_{CSET}\)). (PDb stands for "Patterns Discovered by".) \(PDbSupMaxK\) is a subset of \(PDbSupMax(K + 1)\). Note that this figure only shows the subset-superset relationship, while the size of each rectangle does not imply the number of patterns in each set.

2.5 Levels that can be reached by CSET and SMP in the two series of synthetic data sets (varying density and varying dimensionality).

2.6 Histogram of the \(DiffSup\) of individual genes in the patterns discovered only by SMP, but not by CSET.

2.7 Plot of \(NegLogP\) vs. global support for patterns from CSET and SMP, where the support is relative to the whole data set.

2.8 Histogram of \(NegLogP\) values: (a) the maximum \(NegLogP\) for each of the 1000 permutation tests where randomized labels are used by SMP, (b) the top 300 \(NegLogP\) values of the patterns discovered only by SMP but not by CSET.

2.9 Comparison of the distributions of pattern-based precision between (a) the patterns discovered by SMP but not CSET and (b) random generated patterns.

2.10 Histogram of the best enrichment \(NegLogEnrichP\) values w.r.t. the gene sets in MSigDB, for the patterns discovered by SMP but not by CSET. An enrichment p-value is computed only if a pattern and a gene set have at least 2 genes in common.

2.11 Scalability of different discriminative pattern mining algorithms on the gene expression data.
3.1 Illustration of (a) a differential expression (DE) pattern, (b) a differential coexpression (DC) pattern, and (c) a possible mechanism for the occurrence of a DC pattern due to the mutation of a regulator. Note that, while $G0$ is a dominant regulator of $G1-G4$, the latter genes are also regulated by other independent regulators, that are not shown in the two pathway graphs for simplicity. ((a) and (b) taken from Kostka and Spang (2004) [1].)

3.2 Different types of full-space and subspace, differential and non-differential coexpression patterns. Each curve denotes the expression values of a gene on all the samples. The horizontal line separates all the samples into class A and B. Five patterns are illustrated, with a brown line indicating the samples on which the set of genes are coexpressed.

3.3 Patterns discovered with the original class labels (a) and indication of the statistically significant ones (b).

3.4 Illustration of the full-space differential coexpression (correlation difference) for the discovered statistically significant subspace differential coexpression patterns. The dashed lines in (a) – (c) indicate the statistical significance cutoffs for correlation difference for size-2, size-3 and size-4 patterns respectively. The solid line in (a) is the statistical significance cutoff for SDC (0.4620). There are no corresponding lines in (b) and (c) because all the patterns of size 3 and 4 are statistically significant in terms of SDC as discussed in section 3.4.3. Region A contains patterns that are considered significant by both correlation difference and SDC; Region B has patterns that are not significant as full-space patterns, several of which have close-to-zero correlation difference (within the circle); and Region C shows the significant full-space patterns that are not discovered by SDC.

3.5 A statistically significant subspace differential coexpression pattern, for which the minimal pairwise correlation in the normal class is only 0.28. For better visualization, samples are sorted by increasing range of expression ranks separately in the two classes (similar for Figures 3.6(a), 3.6(b) and 3.7(a)).
3.6 Two patterns that are respectively enriched with the $TNF_{\alpha}/NFkB$ and the $WNT$ signaling pathway. .......................... 65

3.7 The union of the three patterns that are all enriched with the $TNF_{\alpha}/NFkB$ signaling pathway. There are ten genes in this combined pattern (all are known cancer related genes), out of which six are included in the $TNF_{\alpha}/NFkB$ signaling pathway (enrichment $p$-value $1.4024 \times 10^{-5}$). 66

4.1 Transforming a toy SNP dataset in categorical representation to the corresponding binary representation.) ................................................. 75

4.2 Visualization of the two synthetic SNP-genotype combinations and their high-order association with the two classes. The two subfigures in the left column are the visualization of the genotypes of 4 SNPs separated by vertical green lines, over the 70 cases and 70 controls separated by a horizontal yellow line. The black color indicates present and the white indicates absent, in the binary format described in the method section. The $\chi^2$ statistic, odds ratios and the $-\log_{10}$ Fisher exact test $p$ value of the two combinations are $(28.7, 21.0, 7.84)$ and $(25.8, 18.6, 7.1)$, respectively. Each subfigure in the right column contains 4 pairs of bars. For each pair, the unfilled bar and the filled bar indicate the minimal and the maximal $\chi^2$ statistics for the size-$i$ ($i \in [1, 4]$) subsets of the combination. The right most pair, both bars are equal since they both denote the $\chi^2$ statistic of the SNP combination itself.) ................................................. 77

4.3 The scalability of SMP with respect to sample size (cases and controls combined). The computational time of SMP increases linearly with the sample size (Note that the x-axis is not linearly spaced)) .............................. 81

4.4 Comparing the $\chi^2$ statistic of each pattern with the maximal $\chi^2$ statistic among all of its subsets. The three subfigures correspond to the three datasets. Each subfigure shows the $\chi^2$ statistic of each pattern and the maximal $\chi^2$ statistic among all of its subsets for all the discovered patterns. The color of a circle indicates the size of the pattern. The red line and the black line in each subfigure show $y = x + 5$ and $y = x - 5$ respectively.) ................................................. 84
4.5 Comparing the $\chi^2$ statistic of each pattern with the maximal $\chi^2$ statistic among all of its subsets. The three subfigures correspond to the three datasets. Each subfigure shows the $\chi^2$ statistic of each pattern and the maximal $\chi^2$ statistic among all of its subsets for all the discovered patterns. The color of a circle indicates the size of the pattern. The red line and the black line in each subfigure show $y = x + 5$ and $y = x - 5$ respectively.

4.6 Comparison between the FDRs without $\chi^2 - jump$ based filtering and the FDRs with $\chi^2 - jump$ filtering for the Lungcancer and Kidney datasets respectively. In both subfigures, each circle represent a SNP combination. There are several circles sitting below the line $y = x$, indicating that they have lower (more significant) FDR when a $\chi^2 - jump$ filtering is applied compared to the case where no $\chi^2 - jump$ is used.

4.7 Functional similarity of within-combination gene pairs in three groups of discriminative SNP combinations and the null distributions(best view in color). This is to reveal the connection between $\chi^2 - jump$ and within-combination functional coherence. The six comparisons, a – f, and the associated ranksum test p-values are also shown.

4.8 Visualization of two SNP-genotype combinations discovered from the Lungcancer and Kidney datasets respectively. The interpretation is similar to the subfigures in Figure 5. The rsnumbers of the five SNPs in the Lungcancer pattern and the four SNPs in the Kidney pattern (all with $MM$ genotype) are shown. The SNPs in the two patterns are mapped to the following two sets of genes, ($SIM1$, $PARP1$, $WT1$, $ABCC1$, $ABCC4$) and ($XRCC4$, $SLC7A7$, $XRCC1$ and $ITGB3$). The $\chi^2$ statistics of the pattern and its subsets are shown in the right subfigures. Their permutation test-based FDRs and odds ratios are also shown. The top enriched molecular interaction network (by Ingenuity Pathway Analysis) is also shown for each pattern, where the shaded nodes are those genes mapped from the SNPs in each pattern.
4.9 Four synthetic discriminative patterns of size-3 6 that we embed in the synthetic dataset as described in the method section, with similar description as Figure 4.2) ........................................... 101

5.1 A sample data set with three discriminative patterns (A, C, D) and an uninteresting (non-discriminative) pattern (B) .................................................. 104

5.2 Comparing the discriminative power and maximal-subset discriminative power of a set of patterns discovered from the UCI Hepatic dataset, with three different measures for discriminative power. Each circle represent a pattern, with its color indicating pattern size (same for Figures 5.3(a), 5.5, 5.6 and 5.9). ................................................................. 106

5.3 Illustration of T2 discriminative patterns on the gene expression dataset (described in section 5.4). (a) The entire set of discovered patterns; (b) visualization of the pattern in a binary matrix format (black indicating 1’s and white representing 0’s similar as Fig. 5.1) with the horizontal yellow line separating the two classes and the vertical green lines separating genes from each other. ......................................................... 110

5.4 Illustration of the mechanism of synergistic interaction in the context of yeast genetic interaction (Figure taken from Costanzo et al [2]). ................. 114

5.5 Illustration of general improvement and synergistic interaction beyond independent additive effect on the gene expression dataset. Best aggregated MI is only computed for those patterns that have positive Imp. ......... 115

5.6 Illustration of T3 and T4 discriminative patterns on the gene expression dataset. Syn is only computed for those patterns that have positive Imp. 117

5.7 A T3 example and a T4 example, both discovered from the M-Survival SNP dataset as described in section 5.4.1 (refer to Fig. 5.3(b) for similar description. ................................................................. 118

5.8 Relationship among the three types of interesting patterns for the same value of $\alpha$ and $j$. ................................................................. 119

5.9 Existence of $T_2 - T_4$ patterns in two representative datasets: (a)-(c) M-Survival (SNP); (d)-(f) Sonar (UCI). In subfigures (c) and (f), synergy is only computed for those patterns that have positive improvement. ........ 122
6.1 A toy example illustrating the concept of between pathway model (BPM) in the context of complex human disease. (a) Two distinct linear pathways, ABC and XYZ, that both contribute to a common biological process that is essential for maintaining the normal health state of a person. The genetic perturbations on the two pathways and their effect on disease risk are also illustrated. (b) The bipartite structure resulting from the functional compensations between the two pathways shown in (a), i.e., the joint genetic perturbation on any pair of genes between the two pathways yields a non-additive synergistic effect on disease risk. (c) The flowchart of the five steps in BridGE. (d) A conceptual illustration of the pathway-pathway pairwise enrichment analysis.

6.2 A toy example demonstrating that each of the five topological measures have different values between an original network (with LD structure) with the corresponding network without the bias of LD structure.

6.3 Illustration of the evaluation strategy used to access the significance of a topological measure for each binarized SNP-SNP interaction network. X-axis: different thresholds for network binarization; Y-axis (left): number of edges in log scale; Y-axis (right): p value of number of edges at each threshold for binarization. Red line corresponds to the real network with 40 different binarization thresholds. Green lines are based on 500 permutations. The blue line shows the empirical p value for each of the binarization threshold.

6.4 Comparing the three networks with different amount of LD structure with respect to each of the five topological measures on each of the two datasets. X-axis: different thresholds for network binarization; Y-axis: p value of a topological measure (indicated by the text on the left of each row) at each threshold for binarization. Blue line corresponds to the network \(N_0\) constructed with 4000 SNPs \(S_{4000}\) with proper removal of LD structure. Green lines corresponds to network \(N_1\) with 2000 SNPs that are in LD with \(S_{4000}\) added to \(N_0\). The red line corresponds to network \(N_2\) with 2000 SNPs that are in LD with \(S_{4000}\) added to \(N_0\).
6.5 Statistical significance of each of the five topological measures on the Parkinson’s disease dataset. X-axis: different thresholds for network binarization; Y-axis: p value of a topological measure (indicated by the text nearby) at each threshold for binarization. Blue curves correspond to the direct significance. Green curves represent the significance conditional number of edges. Red curves represent the significance conditional number of nodes. ................................................................. 134

6.6 Statistical significance of each of the five topological measures on the breast cancer dataset. X-axis: different thresholds for network binarization; Y-axis: p value of a topological measure (indicated by the text nearby) at each threshold for binarization. Blue curves correspond to the direct significance. Green curves represent the significance conditional number of edges. Red curves represent the significance conditional number of nodes. ................................................................. 135

6.7 Detailed information for a significant BPM discovered from the NIA-II cohort. The center of the figure is a bipartite plot two sets of SNPs mapped to the two pathways whose names are shown on the two sides of the plot, respectively. The two groups of horizontal bars (color coded to indicate chromosome) show the $- \log_{10} p$ values (single locus $\chi^2$ test) for the two sets of SNPs. The two dashed lines show the Bonferroni corrected 0.05 p value cutoff, indicating that no single SNPs are significant. Furthermore, none of the SNP-SNP interactions between the two sets of SNPs (plink lines, i.e. those above the global top 1% quantile) are significant (minimum FDR 0.994). The expected and observed number of SNP-SNP interactions between the two pathways and the associated significance are shown on the top of the figure. ................................. 152
6.8 Type I BPMs discovered from the NIA-II cohort and the validation results in the NGRC and NIA-I cohorts. (a) Network representation of the 48 significant pathway-pathway interactions (FDR < 0.05, less than three false positives) discovered from the NIA-II cohort. Each node shows the pathway name with the letter in the bracket indicating the source of the pathway (K: KEGG, R: Reactome, B: Biocarta), and each edge indicates a pathway-pathway interaction (i.e. a BPM). The color of an edge indicates if it is validated in NGRC, NIA-I or neither of the two, as specified on the top of the subfigure. No type I BPM was validated in both NGRC and NIA-I. (b and c) Venn diagrams summarizing the validation of the 32 less-redundant type I BPMs (refer to the main text and supplementary text for details) in the NGRC cohort and the NIA-I cohort, respectively. The green and blue colored ellipses and texts indicate the two validation tests (1 and 2) checking if the set of 32 BPMs discovered from NIA-II are enriched for positive LLR values (test 1) and significant LLR values (test 2) in each of the two validation cohorts. The significance of the enrichment based on hypergeometric tests are given in the texts: numbers in the brackets are the fractions for the background (i.e. all possible pathway pairs) and the discovered (the 32 significant BPMs).

6.9 Type II BPMs discovered from the NIA-II cohort and the validation in the NGRC and NIA-I cohorts. Only the interactions that are validated in at least one of the two independent cohorts are shown in this figure. The complete network is shown in Supplementary Figure 6.12. Refer to the caption for Figure 6.8. Additional description for (a): orange edges indicate the BPMs that are validated in both the NGRC and NIA-I cohorts.

6.10 Overlaps among the 48 type I BPMs (a) and among the 221 type II BPMs (b) discovered from the NIA-II cohort. The rows and columns are ordered by hierarchical clustering so that similar rows and columns are grouped together, using the same order shown in Supplementary File 2.
6.11 Overlaps of top-scoring SNP-SNP interaction pairs (top 1% quantile) between the NIA-II cohort and the NGRC cohort, for each BPM discovered in NIA-II and validated in NGRC: type I BPMs (a) and type II BPMs (b), as measured by Fisher’s exact test p-value (x-axis) and Jaccard similarity (y-axis). Similar summary of overlaps between NIA-II and NIA-I for type I and type II BPMs are shown in (c) and (d), respectively.

6.12 The complete set of Type II BPMs discovered from the NIA-II cohort organized as a pathway-pathway network. Pathway names are only shown for the eight hub nodes (in the green box), while the pathway names for the other nodes can be found by looking up the node ID in Supplementary File 1.

6.13 Consistency of log likelihood ratios for all the 286903 BPMs tested in the NIA-II cohort between different quantile thresholds. (a) type I BPM, q = 1% (y-axis) vs. q = 0.5% (x-axis); (b) type I BPM, q = 1% (y-axis) vs. q = 2% (x-axis); (c) type II BPM, q = 1% (y-axis) vs. q = 0.5% (x-axis); (d) type II BPM, q = 1% (y-axis) vs. q = 2% (x-axis).
Chapter 1

Introduction

1.1 Background

There has been a dramatic increase in the quantity, quality, and types of advanced biomedical information available to individuals and their medical providers. These types of data include, but are not limited to, cell process information provided by microarrays and DNA/RNA sequencing, genetic information in the form of Single Nucleotide Polymorphisms (SNPs) and copy number variations (more generally structural variations), metabolomics data in terms of proteins and other metabolites, and structural and functional brain data from magnetic resonance imaging (MRI). Together with the increasing availability of clinical data from electronic medical records, this abundance of data has created the very real possibility of personalized medicine, i.e., using detailed biomedical, clinical, and environmental information about a person for a customized and more effective approach to patient care [3, 4, 5]. Achieving this goal requires identifying those features of the data that can distinguish not only between healthy or low risk subjects (controls) and diseased or high risk subjects (cases), but also among different subgroups of cases. These features are typically predictive patterns (biomarkers) that are associated with the disease or other phenotype of interest. Simple examples are a SNP that indicates a predisposition for a particular disorder or the presence of a protein or small molecule that signals the presence of cancer. These patterns 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
1.2 Challenges and Objectives

Despite the initial success, the lack of readily available, easy to use, and effective tools and techniques for finding trustworthy and useful markers is limiting progress in medical research and slowing the advent of personalized medicine [6, 7]. A fundamental limitation is that many of the current approaches for biomarker discovery are based on univariate statistical analysis. This may be effective for studying diseases that are driven by a single gene or SNP etc [8]. However, complex diseases, such as cancer and diabetes, are believed to be caused by the interactions of multiple genes or SNPs, as well as environmental factors [8], which can not be captured by traditional univariate analysis. Thus, techniques for extracting meaningful associations must be able to discover combinatorial biomarkers, i.e., combinations of factors that show a significant association with a disease phenotype even when single factors have little or no association. Discovering combinatorial biomarkers is the primary focus of this thesis. We discuss several key challenges for discovering combinatorial biomarkers, and more generally for other domains as well, that call for more effective and novel computational approaches towards the following capabilities:

1. **Computational efficiency**: The need of searching for high order combinatorial biomarkers leads to increased computational complexity, since the number of possible patterns increase exponentially with pattern length. This general challenge exists in many other domains beyond the field of computational biology, such as climate change detection, document analysis, etc. To search for high order combinations from datasets with high density, brute-force approaches can handle only a relatively small number of variables (tens or hundreds), while heuristic approaches risk missing informative combinations. This calls for computational approaches with computational efficiency and scalability.

2. **Statistical power**: In addition to the challenge of computational efficiency, perhaps an even more serious challenge is that of multiple hypothesis testing which results
from the enormous number of potential patterns (hypotheses) and the resulting increased probability of mistaking spurious patterns for real ones. Specifically, even if there is a perfect algorithm that can efficiently search for high order combinations from datasets of any given density and dimensionality, many combinations could be discovered even due to random change, as a consequence of the huge number of combinations tested in the search process compared to the relatively insufficient number of samples available. This calls for computational approaches with statistical power over the huge number of hypothesis tests.

3. Effective handling of sample heterogeneity: Yet another complication is the heterogeneous nature of many datasets. For example, patients with a particular disease may form different subgroups and predictive patterns appropriate for one subgroup may not apply to another. Many existing techniques assume that a biomarker is universally applicable, and thus can not handle the situation where a biomarker may be applicable over only a subset of samples (e.g. patients) that have a unique genetic background or cause of disease. Understanding such sample heterogeneity is of significant value for accurate personalized prognosis, diagnosis and personalized medicine, and thus calls for effective computational approaches that can discover multi-gene biomarkers over sample subgroups.

4. Understanding of the nature of combinatorial interactions: Although a few techniques have been proposed to discover combinatorial biomarkers that are statistically significant, the biological relevance of the biomarkers discovered by different types of combinatorial search algorithms are not well understood. Specifically, some combinatorial biomarkers may represent a gene-gene coexpression network that have a coherent molecular function in a disease mechanism, while some combinatorial biomarkers could reflect the complementarity across several pathways with similar functions that can compensate the loss of biological function on each other. This calls for systematic characterization of different types of combinatorial interactions that often each yields very different domain interpretations.

5. Capability of system-level interpretation: Most existing approaches treat the discovered multi-gene biomarkers as separate candidates for disease prognosis and
diagnosis. This may only provide a fragmented view of the mechanisms of complex diseases such as cancer, since a cancer-related pathway can be large in size and have coherent but heterogeneous activities. This limitation could be part of the reason that most statistically significant biomarkers discovered from one cohort cannot be validated in independent cohorts. Thus, it calls for approaches that can aggregate multi-gene biomarkers into higher-level structures, e.g. pathways or interaction subnetworks, towards a systems level interpretation of the discovered combinatorial biomarkers.

1.3 Contributions and Outline of following chapters

To address the challenges described above, this thesis aims to advance the state of the arts in association analysis techniques and apply them to combinatorial disease biomarker discovery. Association analysis techniques use the basic Apriori framework [9, 10],\textsuperscript{1} that can leverage the anti-monotonicity of the objective functions to systematically prune the exponential search space of high-order combinations and avoid exploring the whole search space. Furthermore, association analysis techniques not only discover interactions of different features, they also naturally identify the set of samples explained by a specific interaction. This allows us to address the challenges of finding complex patterns (patterns of multiple variables) from a large set of variables (high dimensional data) in situations when the pattern may only appear in a subset of the participants (e.g. when the disease is heterogeneous).

While traditional association analysis techniques have the advantages for efficiently handling combinatorial search and discovering sample subgroups, they also have several limitations due to the other challenges discussed in the last section, that prevent them from effectively handling biomedical and genomic data with unique characteristics. First, while association analysis serves a more efficient framework for combinatorial search, existing approaches are not effective for datasets that are high-dimensional (unless they are very sparse) or of high density (unless they are low dimensional). On biomedical and genomic data that have both high density and high dimensionality at

\textsuperscript{1} Here we are referring to the branch of data mining known as ‘association analysis’ and not the collection of statistical techniques known by the same name that are often used to analyze SNP and other biomedical data.
the same time, most existing techniques can only discover combinatorial biomarkers with relatively high support and thus tend to miss informative ones supported by only a small fraction of subjects. Some recent work [11, 12, 13] has proposed to leverage the class label information to help prune the exponential search space (discriminative pattern mining), however, their efficiency and scalability still could not meet the density and dimensionality of some genomic datasets such as single nucleotide polymorphism datasets. This calls for advancements of the state of the art discriminative pattern mining techniques to allow efficient mining of dense and high dimensional datasets. Second, because most existing techniques for discriminative pattern mining are designed for binary data, they can not be directly applied to some biomedical and genomic datasets that naturally have continuous values. While binarization can be performed before pattern discovery, it is not effective when the target type of differentiation can only be revealed from continuous data. This calls for novel discriminative pattern mining techniques that can directly handle continuous data and detect different types of differentiation. Third, the overall statistical power of existing association analysis and discriminative pattern mining techniques, considering multiple hypothesis testing, is insufficient. This is essentially due to the huge number of combinations searched for by the Apriori algorithm, which usually leads to a high false discovery rate. This calls for effective approaches to enhance the statistical power of discriminative pattern mining by properly controlling the number of hypothesis tested in the search process. Third, existing work on discriminative pattern mining usually target classification as the major goal while there is a lack of in-depth understanding of the combinatorial interactions in discriminative patterns. This is important for combinatorial biomarker discovery because different types of interactions may reflect different disease mechanisms that could suggest different treatments. It is also important for other domains because a systematic characterization of discriminative patterns according to different types of interactions could help understand the discriminative structure of a dataset. Last but not least, existing approaches for discriminative pattern mining usually treat different discovered patterns as separate candidates for analysis, but there is a lack of understanding on their inter-relationships among the set of pattern beyond summarization [14]. A integrative network analysis of the set of discovered patterns in terms of global topology and local substructure may provide novel insights on disease mechanisms that are invisible from
individual combinatorial biomarkers.

Below, we summarize the contributions of this thesis in terms of addressing these limitations of association analysis and and maximizing their biological impacts for combinatorial biomarker discovery.

- Chapter 2 aims to address the computational challenge confronted by existing discriminative pattern mining techniques when handling dense and high-dimensional datasets. Specifically, I present a framework named SMP for the efficient search of low-frequency discriminative patterns from binary datasets with both high density and high dimensionality. Experiments on both synthetic data sets and a breast cancer gene expression data set demonstrate that there are low-support patterns that can be discovered using SMP but not by existing approaches. Furthermore, we show that the low-support discriminative patterns that are only discovered using SMP from the cancer gene expression data set are statistically significant and biologically relevant.

- Chapter 3 addresses the challenge confronted by existing discriminative pattern mining techniques on different types of biomedical and genomic data with continuous values. First, I motivate the search for different types of biologically relevant discriminative patterns defined for continuous data. Then, I discuss strategies for efficient pattern mining. Specifically, I present a generalized framework named subspace differential coexpression analysis, for direct discovery from continuous case-control data. The proposed approach is applied to gene expression datasets from three lung cancer studies and discovers differential coexpression patterns supported by different subsets of patients that are enriched for genes from known cancer pathways.

- Chapter 4 pursues the enhancement of the statistical power of discriminative pattern mining techniques. We propose a generic strategy that handles the issue of multiple hypothesis testing by identifying and pruning patterns that are not interesting or useful in the context of a certain domain. Specifically, in combinatorial disease biomarker discovery, the type of discriminative patterns with most interests for biologists are those combinations that provide additional discriminative power over its sub patterns. Only focusing on such patterns in the exponential
search space could simultaneously reduce computational complexity and reduce the overall number of hypotheses tested (thus reduce the risk of finding spurious patterns). Three case-control SNP datasets (from three studies on multiple myeloma, lung cancer and kidney transplant rejection) are used to demonstrate the effectiveness of this strategy for enhancing the overall statistical power of discriminative pattern mining.

• Chapter 5 aims to achieve a systematic understanding of different types of discriminative patterns and the corresponding biological relevance. Specifically, I categorize discriminative patterns according to four types of item interaction and present a comprehensive discussion that defines these four pattern types and investigates their properties and their relationship to one another. In addition, these ideas are explored for a variety of datasets (ten UCI datasets, one gene expression dataset and two genetic-variation datasets). In the results, we first demonstrate the existence of different types of discriminative patterns in real datasets. Some datasets have all the four types of patterns while some datasets are specifically enriched for certain types of patterns. We then demonstrate the characterization framework can help interpret different types of molecular-level interaction underlying different sets of genes and genetic variations. For datasets in other domains beyond biology, we illustrate how pattern characterization can provide novel insights into the discriminative structure of datasets. In Chapter 4, we proposed a generic strategy for enhancing the statistical power of discriminative pattern mining by focusing on a specific type of discriminative patterns with a specific characteristics. We expect this generic strategy can also be used to enhancing the statistical power for discovering each of the four types of patterns characterized in this chapter.

• Chapter 6 proposes a novel concept to organize discriminative patterns into an interaction network that allows the discovery of high-level structural knowledge. On the global scale, we explore the network structure of the constructed pattern networks and discover that they generally have statistical significant topological measures compared to random networks. On the local scale, biologically relevant substructures can be systematically discovered. Specifically, a general framework
is proposed to detect pathway-pathway interaction pairs that are enriched for genetic level interactions from genome wide association datasets. Validations across independent real datasets not only demonstrate the reliability of the proposed framework but also lead to several interesting biological insights on several complex diseases such as breast cancer and Parkinson’s disease.

- In Chapter 7, I conclude the thesis with summary of contributions, limitations and future directions. The data-mining algorithmic contributions in this thesis also hold promise for applications in other domains beyond biology.

1.4 Relationship to a broader set of data mining and machine learning approaches

Many of the challenges described above are being addressed by a broad range of work in statistics, machine learning, bioinformatics, data mining, and many domain specific research efforts. Nonetheless, we believe that the proposed discriminative pattern mining approach offers a unique set of benefits not offered by other approaches. For instance, it has the ability to discover patterns that reflect the heterogeneous nature of the populations and the results (patterns) produced are easily interpreted. To illustrate, for the analysis of neuroimaging data, a wide variety of statistical techniques [15, 16] are applied, e.g., structural equation models and Independent Component Analysis (ICA). Although these approaches have provided many useful results, they and many other multivariate statistical approaches cannot handle heterogeneity. Although many machine learning approaches can create good classification models for heterogeneous data, often they do not yield easily interpretable results. As another illustration, penalized logistic regression based approaches have been extensively studied [17, 18, 19] to search for statistically significant predictive patterns that relate genetic variation to disease. These approaches are computationally efficient, but they provide only a single view of the disease association study by optimizing the global classification accuracy with sparsity constraints [18], whereas typically there are many relatively distinct sets of factors (variables), each of which can contribute additional biological insight. In addition, these approaches normally cannot detect predictive patterns involving an interaction of more
than two factors [17, 18]. However, we wish to note that the pattern mining approach we propose does not replace other techniques already in use in the fields we are investigating, but rather, it provides a complementary set of techniques for finding patterns that these other techniques were not designed to find.
Chapter 2

Efficient Discriminative Patterns Mining from Dense and High-dimensional Binary Data

2.1 Introduction

This chapter aims to address the computational challenge confronted by existing discriminative pattern mining techniques when handling dense and high-dimensional datasets. Specifically, I first demonstrate the necessity of trading off the completeness of discriminative pattern discovery with the efficient discovery of low-support discriminative patterns from such data sets. Then, I present a framework for the efficient search of low-support discriminative patterns from binary datasets with both high density and high dimensionality. Experiments on both synthetic data sets and a breast cancer gene expression data set demonstrate that there are low-support patterns that can be discovered using SupMaxPair but not by existing approaches. Furthermore, we show that the low-support discriminative patterns that are only discovered using SupMaxPair from the cancer gene expression data set are statistically significant and biologically relevant.
2.2 Background

For data sets with class labels, association patterns [10, 9] that occur with disproportionate frequency in some classes versus others can be of considerable value in many applications. Such applications include census data analysis that aims at identifying differences among demographic groups [20, 21] and biomarker discovery, which searches for groups of genes or related entities, that are associated with diseases [22, 23, 24]. We will refer to these patterns as discriminative patterns\(^1\) in this chapter, although they have also been investigated under other names [25], such as emerging patterns [20] and contrast sets [21]. In this chapter, we focus on 2-class problems, which can be generalized to multi-class problems as described in [21].

Discriminative patterns have been shown to be useful for improving the classification performance for data sets where combinations of features have better discriminative power than the individual features [12, 26, 27, 11, 28, 29]. More importantly, as discussed in [21], discriminative pattern mining can provide insights beyond classification models. For example, for biomarker discovery from case-control data (e.g. disease vs. normal samples), it is important to identify groups of biological entities, such as genes and single-nucleotide polymorphisms (SNPs), that are collectively associated with a certain disease or other phenotypes [24, 30, 31]. Algorithms that can discover a comprehensive set of discriminative patterns are especially useful for domains like biomarker discovery, and such algorithms are the focus of this chapter.

The algorithms for finding discriminative patterns usually employ a measure for the discriminative power of a pattern. Such measures are generally defined as a function of the pattern’s relative support\(^2\) in the two classes, and can be defined either simply as the ratio [20] or difference [21] of the two supports, or other variations, such as its information gain [12], Gini index, odds ratio [9] etc. In this chapter, we use the measure that is defined as the difference of the supports of an itemset in the two classes (originally proposed in [21] and used by its extensions [32, 33]). We will refer to this measure as \textit{DiffSup} (formally discussed in Section 2.3). Given a dataset with 0-1 class

---

1 The terms “pattern” and “itemset” are used interchangeably in this chapter.

2 Note that, in this chapter, unless specified, the support of a pattern in a class is relative to the number of transactions (instances) in that class, i.e. a ratio between 0 and 1, which can help handle the case of skewed class distributions.
labels and a \( \text{DiffSup} \) threshold \( r \), the patterns with \( \text{DiffSup} \geq r \) can be considered as valid discriminative patterns.

To introduce some key ideas about discriminative patterns and make the following discussion easier to follow, consider Figure 2.1, which displays a sample dataset\(^3\) containing 15 items (columns) and two classes, each with 10 instances (rows). In the figure, four patterns (sets of binary variables) can be observed: \( P_1 = \{i_1, i_2, i_3\} \), \( P_2 = \{i_5, i_6, i_7\} \), \( P_3 = \{i_9, i_{10}\} \) and \( P_4 = \{i_{12}, i_{13}, i_{14}\} \). \( P_1 \) and \( P_4 \) are interesting discriminative patterns that occur with different frequencies in the two classes, whose \( \text{DiffSup} \) is 0.6 and 0.7 respectively. In contrast, \( P_2 \) and \( P_3 \) are uninteresting patterns with a relatively uniform occurrence across the classes, both having a \( \text{DiffSup} \) of 0.

\(^3\) The discussion in this chapter assumes that the data is binary. Nominal categorical data can be converted to binary data without loss of information, while ordinal categorical data and continuous data can be binarized, although with some loss of magnitude and order information.
Furthermore, $P_4$ is a discriminative pattern whose individual items are also highly discriminative, while those of $P_1$ are not. Based on support in the whole dataset, $P_2$ is a frequent non-discriminative pattern, while $P_3$ is a relatively infrequent non-discriminative pattern.

Note that the discriminative measures discussed above are generally not anti-monotonic as shown by [20, 21, 12]. Take \textit{DiffSup} for instance (while other measures like support ratio, information gain and odds ratio are not anti-monotonic either): although the \textit{DiffSup} of the three items in $P_1$ are 0, 0 and 0.2 respectively, $P_1$ has a \textit{DiffSup} of 0.6 as an itemset. Due to the lack of anti-monotonicity, these measures can not be directly used in an Apriori framework [10] for exhaustive and efficient pattern mining as can be done for measures like support [10], h-confidence [34] etc. To address this issue, many approaches [35, 36, 37, 38, 12] adopt a two-step strategy (denoted as Group A), where first, a frequent pattern mining algorithm is used to find all (closed) frequent patterns that satisfy a certain support threshold minsup either from the whole dataset or from only one of the classes. The patterns found can be further refined using other interestingness measures (e.g. [39, 40, 41]). Then, as post-processing, \textit{DiffSup} is computed for each of these patterns, based on which discriminative patterns are selected. Note that, in general, these two-step approaches can work even with a very low minsup threshold [42, 12] on relatively sparse or low-dimensional datasets.

However, since these approaches ignore class label information in the mining process, many frequent patterns discovered in the first step may turn out to have low discriminative power in the second step. For instance, in Figure 2.1, the relative supports of $P_2$ and $P_3$ in the whole dataset are 0.6 and 0.3 respectively, and will be considered as frequent patterns if the support threshold is 0.2. However, $P_2$ and $P_3$ are not discriminative since they both have a \textit{DiffSup} of 0. In particular, in datasets with relatively high density \footnote{The density of a transaction matrix is the percentage of 1’s in the transaction-by-item matrix} and high-dimensionality, a huge number of non-discriminative patterns like $P_2$ and $P_3$ in Figure 2.1 may exist. Such patterns may meet the minsup threshold and would be discovered in the first step, but would be found as non-discriminative patterns in the second step. If a low minsup is used, a huge number of such patterns can reduce the efficiency of both the two steps as discussed in [11]. In such a situation, the two-step approaches have to use a sufficiently high minsup in order to generate the
complete set of results within an acceptable amount of time, and thus may miss a large number of highly discriminative patterns that fall below the minsup threshold.

A possible strategy for improving the performance of the two-step approaches is to directly utilize the support of a pattern in the two classes for pruning some non-discriminative patterns in the pattern mining stage. Indeed, several approaches have been proposed [21, 12], where the anti-monotonic upper bounds of discriminative measures, such as DiffSup, are used for pruning some non-discriminative patterns in an Apriori-like framework [10]. This strategy, like the two-step approaches, also guarantees to find the complete set of discriminative patterns with respect to a threshold, although in a more efficient manner. However, in datasets with relatively high density and high-dimensionality, there can be a large number of frequent non-discriminative patterns like $P_2$ in Figure 2.1. Such patterns may not be pruned by these approaches because the upper bounds of the discriminative measures may be weak (technical details in Section 2.4). Thus, as illustrated in Figure 2.2(a), these approaches (referred to as group $B$ in the rest of this chapter) are able to discover a larger fraction of the interesting discriminative patterns as compared to the two-step approaches. However, they may still miss a lot of highly discriminative patterns, particularly those at low support levels, given the same fixed amount of time. These low support patterns are supported by a relatively small number of samples but can still be highly discriminative according to their DiffSup value, especially in the case of data sets with skewed class size distributions.

Yet another strategy for discovering a significant subset of the discriminative patterns is to directly use a measure of discriminative power for pruning non-discriminative patterns [43]. As an instance of such an approach, DiffSup can be computed for each candidate pattern $\alpha$, and if $\text{DiffSup}(\alpha) < r$, then $\alpha$ and all its supersets can be pruned in an Apriori-like algorithm [10]. This strategy is computationally more efficient than the two-step approaches, because no patterns with $\text{DiffSup}(\alpha) < r$ are generated during the mining process. However, this improved efficiency comes at the cost of not discovering the complete set of discriminative patterns, since DiffSup is not anti-monotonic [21]. More specifically, the algorithms in this group (referred to as group $C$ in the rest of this chapter), may miss interesting discriminative patterns whose individual items are not discriminative (e.g. $P_1$ in Fig. 2.1). With respect to the coverage of the set of
(a) The limitation of existing approaches

(b) The motivation of the proposed work

Figure 2.2: An illustration of the coverage of the space of discriminative patterns by different approaches given the same amount of time. The t’s on the y-axis represent the lowest support of the patterns that are respectively covered by the corresponding approaches (represented by boxes), given the same and fixed amount of time. (a) Box A, B and C represent the set of patterns discovered by the corresponding approaches in group A, B and C respectively. (b) Illustration of the tradeoff between the capability to search low-support discriminative patterns in dense and high-dimensional data and the completeness of the pattern discovery. Boxes X, Y and Z represent three conceptual low-support discriminative pattern mining approaches that discover patterns not found by the approaches in groups A, B and C. Note that, in this figure, the set of interesting discriminative patterns is the same as that in (a), but the corresponding *’s are not shown for the sake of clarity.

As can be seen from the discussion above, which is summarized in Figure 2.2(a), the current approaches face an inherent trade-off when discovering discriminative patterns from a dense and high-dimensional data set. The approaches in groups A and B face challenges with discovering low-support patterns due to their focus on the complete discovery of discriminative patterns satisfying the corresponding thresholds. On the other hand, the approaches in group C sacrifice completeness for the ability of discovering low-support discriminative patterns. This trade-off is expected to be faced interesting discriminative patterns, the approaches in this group may be able to discover low-support patterns at the expense of missing a large number of interesting patterns, as illustrated by the stars not included in box C in Figure 2.2(a). This observation is also reflected in our experimental results (section 2.7.2).
by any algorithm for this complex problem, particularly due to the restriction of fixed computational time. In such a scenario, an appropriate approach to discover some of the interesting discriminative patterns missed by the current approaches, is to formulate new measures for discriminative power and corresponding algorithms that can progressively explore lower support thresholds for discovering patterns, while trading off completeness to some extent. Such a design is illustrated in Figure 2.2(b), where boxes $X$, $Y$ and $Z$ represent three approaches, which can discover patterns with progressively lower thresholds ($t_x > t_y > t_z$). However, the cost associated with this ability is that of potentially missing some patterns that are at higher support levels. Still, $X$, $Y$ and $Z$ can all discover several patterns that are exclusive to only one of them, and can thus play a complementary role to the existing approaches by expanding the coverage of the set of interesting discriminative patterns.

Corresponding to the motivation discussed above, we propose a family of anti-monotonic measures of discriminative power named $SupMaxK$. These measures conceptually organize the set of discriminative patterns into nested layers of subsets, which are progressively complete in their coverage, but require increasingly more computation for their discovery. Essentially, $SupMaxK$ estimates the $DiffSup$ of an itemset by calculating the difference of its support in one class and the maximal support among all of its size-$K$ subsets in the other class. The smaller the value of $K$, the more effective $SupMaxK$ is for finding low-support discriminative patterns by effectively pruning frequent non-discriminative patterns. Notably, due to the anti-monotonicity property of all the members of $SupMaxK$, each of them can be used in an Apriori-like framework [10] to guarantee the discovery of all the discriminative patterns with $SupMaxK \geq r$, where $r$ is a user-specified threshold. Given the same (limited) amount of time, the members of this family provide a tradeoff between the ability to search for low-support discriminative patterns and the coverage of the space of valid discriminative patterns for the corresponding threshold, as illustrated by the three conceptual approaches $X$, $Y$ and $Z$ in Figure 2.2(b). In particular, we find that a special member with $K = 2$ named $SupMaxPair$, is suitable for dense and high-dimensional data. We have designed a framework, named SMP, which uses $SupMaxPair$ for discovering discriminative patterns. Carefully designed experiments with both synthetic datasets and a cancer gene expression dataset are used to demonstrate that SMP can serve a complementary role
to the existing approaches by discovering low-support yet highly discriminative patterns from dense and high-dimensional data, when the latter fail to discover them within an acceptable amount of time.

2.2.1 Contributions of this chapter

The contributions of this chapter can be summarized as follows:

1. We address the necessity of trading off the completeness of discriminative pattern discovery with the ability to discover low-support discriminative patterns from dense and high-dimensional data within an acceptable amount of time. For this, we propose a family of anti-monotonic measures named SupMaxK that conceptually organize the set of discriminative patterns into nested layers of subsets, which are progressively more complete in their coverage, but require increasingly more computation for their discovery.

2. In particular, SupMaxK with $K = 2$, named SupMaxPair, is a special member of this family that is suitable for dense and high-dimensional data, and can serve a complementary role to the existing approaches by helping to discover low-support discriminative patterns, when the latter fail to discover them within an acceptable amount of time. We designed a framework, named SMP, which uses SupMaxPair for discovering discriminative patterns.

3. A variety of experiments with both synthetic datasets and a cancer gene expression dataset are presented to demonstrate that there are many patterns with relatively low support that can be discovered by SMP but not by the existing approaches. In particular, these experiments rigorously demonstrate that the low-support discriminative patterns discovered only by SMP from the cancer gene expression dataset are statistically significant (via permutation test [44, 45]) and biologically relevant (via comparison with a list of cancer-related genes [46] and a collection of biological gene sets [45] (e.g. pathways)). These are the recognized methods for evaluating the utility of such patterns for applications such as biomarker discovery [45, 22, 47].
2.3 Basic Terminology and Problem Definition

Let $D$ be a dataset with a set of $m$ items, $I = \{i_1, i_2, ..., i_m\}$, two class labels $S_1$ and $S_2$, and a set of $n$ labeled instances (itemsets), $D = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \subseteq I$ is a set of items and $y_i \in \{S_1, S_2\}$ is the class label for $x_i$. The two sets of instances that respectively belong to the class $S_1$ and $S_2$ are denoted by $D_1$ and $D_2$, and we have $|D| = |D_1| + |D_2|$. For an itemset $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ where $\alpha \subseteq I$, the set of instances in $D_1$ and $D_2$ that contain $\alpha$ are respectively denoted by $D_{\alpha}^1$ and $D_{\alpha}^2$. The relative supports of $\alpha$ in classes $S_1$ and $S_2$ are $\text{RelSup}^1(\alpha) = \frac{|D_{\alpha}^1|}{|D_1|}$ and $\text{RelSup}^2(\alpha) = \frac{|D_{\alpha}^2|}{|D_2|}$, respectively. $\text{RelSup}$ is anti-monotonic since the denominator is fixed and the numerator is support of the itemset, which is anti-monotonic.

The absolute difference of the relative supports of $\alpha$ in $D_1$ and $D_2$ is defined originally in [21] and denoted in this chapter as $\text{DiffSup}$:

$$\text{DiffSup}(\alpha) = |\text{RelSup}^1(\alpha) - \text{RelSup}^2(\alpha)|. \quad (2.1)$$

An itemset $\alpha$ is $r$-discriminative if $\text{DiffSup}(\alpha) \geq r$. The problem addressed by discriminative pattern mining algorithms is to discover all patterns in a dataset with $\text{DiffSup} \geq r$.

Without loss of generality, we only consider discriminative patterns for the binary-class problem. Our work can be extended to multiple classes as described in [21].

2.4 Computational Limitations of Current Approaches

As discussed in Section 2.2, in dense and high-dimensional data, the approaches in groups $A$ and $B$ have to use a relatively high threshold in order to provide the complete result within an acceptable amount of time. In this section, we will show that this limitation is essentially due to the ineffective pruning of frequent non-discriminative patterns (e.g. $P_2$ in Figure 2.1). Generally, the approaches in group $B$ is relatively more efficient than those in group $A$, as discussed in Section 2.2. We use the measure
originally proposed in CSET [21] as a representative of group B for this discussion, while a similar discussion also holds for other approaches in group B [12, 48]. In CSET, an upper bound of $\text{DiffSup}$ is defined as the bigger of the relative supports of a pattern $\alpha$ in $D^1$ and $D^2$. In this chapter, we denote it as $\text{BiggerSup}$:

$$\text{BiggerSup}(\alpha) = \max(\text{RelSup}^1(\alpha), \text{RelSup}^2(\alpha)).$$

(2.2)

**Lemma 1** $\text{BiggerSup}$ is anti-monotonic

**Proof 1** Follows from the anti-monotonicity of $\text{RelSup}$ and the property of the $\max$ function.

Since $\text{BiggerSup}$ is an upper bound of $\text{DiffSup}$ [21], and it is also anti-monotonic (Lemma 1), CSET [21] uses $\text{BiggerSup}$ as a pruning measure in a Apriori-like framework, and can discover, given sufficient time and computing resources, the complete set of discriminative patterns (w.r.t a $\text{BiggerSup}$ threshold). However, by using the bigger one to estimate the difference of the two supports, $\text{BiggerSup}$ is a weak upper bound of $\text{DiffSup}$. For instance, if we want to use CSET to search for $0.4 - \text{discriminative}$ patterns in Figure 2.1, $P_3$ can be pruned, because it has a $\text{BiggerSup}$ of 0.3. However, $P_2$ can not be pruned ($\text{BiggerSup}(P_2) = 0.6$), even though it is not discriminative ($\text{DiffSup}(P_2) = 0$). More generally, $\text{BiggerSup}$-based pruning can only prune infrequent non-discriminative patterns with relatively low support, but not frequent non-discriminative patterns. Therefore, in dense and high-dimensional data, where a large number of frequent non-discriminative patterns are expected to exist, CSET with a relatively low $\text{BiggerSup}$ threshold can often fail to produce the complete results in a reasonable amount of time. Thus, CSET has to set the $\text{BiggerSup}$ threshold high and may not discover discriminative patterns at lower support that may be of interest. Similar discussion on the limited ability of pruning frequent non-discriminative patterns also holds for other approaches in groups A and B, i.e., all the two-step approaches, and those based on the information gain upper bound [12], and other statistical metric-based pruning [21, 48].
2.5 Proposed Approach

As shown above, the limitation of existing approaches is essentially the ineffectiveness of pruning frequent non-discriminative patterns. Conceptually, to prune frequent non-discriminative patterns, a new measure should be designed such that a pattern’s support in one class can be effectively limited to a relatively smaller number compared to its support in the other class. In this section, we start with such a measure $\text{SupMax1}$ in Definition 1, and then extend it to a family of measures $\text{SupMaxK}$. Then, we will discuss the relationships between $\text{DiffSup}$, $\text{BiggerSup}$ and $\text{SupMaxK}$. Finally, we will focus on a special member of this family $\text{SupMaxPair}$ that is suitable for high-dimensional data. Note that, for an itemset $\alpha$ two cases can happen: $\text{RelSup}^1(\alpha) \geq \text{RelSup}^2(\alpha)$ or $\text{RelSup}^1(\alpha) < \text{RelSup}^2(\alpha)$. In the following discussion, without loss of generality, we assume $\text{RelSup}^1(\alpha) \geq \text{RelSup}^2(\alpha)$ for simplicity.

2.5.1 $\text{SupMax1}$: A Simple Measure to Start with

**Definition 1** The $\text{SupMax1}$ of an itemset $\alpha$ in $D^1$ and $D^2$ is defined as

$$\text{SupMax1}(\alpha) = \text{RelSup}^1(\alpha) - \max_{a \in \alpha}(\text{RelSup}^2\{\{a\}\}).$$

$\text{SupMax1}$ of an itemset $\alpha$ is computed as the difference between the support of $\alpha$ in $D^1$, and the maximal individual support of the items in $\alpha$ in $D^2$. $\text{SupMax1}$ approximates $\text{DiffSup}$ by using the maximal individual support in $D^2$ to estimate $\text{RelSup}^2(\alpha)$. Clearly, the maximal individual support is quite a rough estimator for $\text{RelSup}^2(\alpha)$, because a pattern can have very low support in class $S_2$ but the items in it can still have very high individual supports in this class. However, an alternative way to interpret $\text{SupMax1}$ is that, a pattern with large $\text{SupMax1}$ has relatively high support in one class and all the items in it have relatively low support in the other class. $P_4$ is such an example whose $\text{SupMax1}$ is $0.9 - \max(0.3, 0.3, 0.3) = 0.6$ as shown in Figure 2.1. Thus, given a $\text{SupMax1}$ threshold, say 0.4, $\text{SupMax1}$ discovers a subset of $0.4 - \text{discriminative}$ patterns but not all, e.g. it will miss patterns like $P_1$ in Figure 2.1, which has relatively high $\text{DiffSup}$ (0.6) but zero $\text{SupMax1}$. 
2.5.2 SupMaxK

Following the rationale of SupMax1, the maximal support of size-\(k\) subsets of a pattern in \(D^2\) can be used to estimate \(\text{RelSup}^2(\alpha)\) instead of using maximal individual support in class \(S_2\) to estimate \(\text{RelSup}^2(\alpha)\). This can provide a better estimation of \(\text{RelSup}^2(\alpha)\). In such a manner, SupMax1 can be generalized into a family of measures SupMax\(K(\alpha)\), which is formally defined in Definition 2. Note that in the following discussion, SupMax\(K\) will be used to refer to this family as well as one of its general members, for the clarity of presentation.

Definition 2 The SupMax\(K\) of an itemset \(\alpha\) in \(D^1\) and \(D^2\) is defined as

\[
\text{SupMax}K(\alpha) = \text{RelSup}^1(\alpha) - \max_{\beta \subseteq \alpha}(\text{RelSup}^2(\beta)), \text{ where } |\beta| = K
\]

So, SupMax\(K\) of an itemset \(\alpha\) is computed as the difference between the support of \(\alpha\) in \(D^1\), and the maximal support among all the size-\(K\) subsets of \(\alpha\) in \(D^2\). Note that, in this chapter, SupMax\(K\) is defined with respect to DiffSup, while similar concept can also be applied to other discriminative measures such as the ratio-based measure [20].

2.5.3 Properties of the SupMaxK Family

In the following subsections, we discuss three properties of the SupMax\(K\) family.

The subset-superset relationship among SupMaxK members

Based on the definition of SupMax\(K\), the following two Lemmas show the relationship among SupMax\(K\) members.

Lemma 2 If we use MaxSup\((\alpha, K)\) to denote the second component of SupMax\(K(\alpha)\), i.e. \(\max_{\beta \subseteq \alpha}(\text{RelSup}(\beta))\) with \(|\beta| = K\), then MaxSup\((\alpha, K)\) is a lower bound of MaxSup\((\alpha, K - 1)\) for integer \(K \in [2, |\alpha|]\)

Proof 2 For every size-(\(K - 1\)) subset of \(\alpha\) (say \(\beta\), \(|\beta| = K - 1\)), there exists a size-\(K\) subset of \(\alpha\) (say \(\beta'\), \(|\beta'| = K\)) such that \(\beta \subset \beta'\), e.g. by adding any \(i\) to \(\beta\), where \(i \in \alpha\) and \(i \notin \beta\). Based on the anti-monotonicity property of RelSup, it is guaranteed that \(\text{RelSup}(\beta') \leq \text{RelSup}(\beta)\). Then, from the properties of the max function,
max_{\beta' \subseteq \alpha} (\text{RelSup}(\beta')) \leq \max_{\beta' \subseteq \alpha} (\text{RelSup}(\beta)). Thus, MaxSup(\alpha, K) is a lower bound of MaxSup(\alpha, K - 1).

**Lemma 3** \( \text{SupMax}(K - 1) \) of an itemset \( \alpha \) is a lower bound of its \( \text{SupMaxK} \), or alternatively \( \text{SupMaxK} \) of an itemset \( \alpha \) is an upper bound of its \( \text{SupMax}(K - 1) \), for integer \( K \in [2, |\alpha|] \)

**Proof 3** Follows directly from Definition 2, Lemma 2.

From Lemma 3, we know that, given the same threshold \( r \) and sufficient time, the set of patterns discovered with \( \text{SupMax}(K - 1) \) in an Apriori framework is a subset of the set of patterns discovered with \( \text{SupMaxK} \). This means that \( \text{SupMaxK} \) can find more and more discriminative patterns as \( K \) increases from 1 (\( \text{SupMax1} \)), to 2 (\( \text{SupMax2} \)), to 3 (\( \text{SupMax3} \)) and so on. The patterns that are discovered by \( \text{SupMaxK} \) but not by \( \text{SupMax}(K - 1) \) are those with \( \text{SupMaxK} \geq r \), but with \( \text{SupMax}(K - 1) < r \). Figure 2.3 shows an extended version of the data set shown in Figure 2.1 containing fifteen addition items (\( i_{16} - i_{30} \)) and two patterns \( P_5 \) and \( P_6 \), the rest being identical to Figure 2.1. In this data set, given the same threshold \( r = 0.4 \), \( \text{SupMax1} \) can find \( P_4 \), but not \( P_1 \) and \( P_3 \), both of which have \( \text{DiffSup} = 0.6 \), but zero \( \text{SupMax1} \); \( \text{SupMax2} \) can find \( P_1 \) in addition to \( P_3 \); furthermore, \( \text{SupMax3} \) can find \( P_3 \) in addition to \( P_4 \) and \( P_3 \). This illustrates that \( \text{SupMax3} \) can find all the patterns found using \( \text{SupMax1} \) and \( \text{SupMax2} \), but not vice versa, as discussed above. Furthermore, \( \text{SupMax10} \) will be able to discover pattern \( P_6 \) in addition to the patterns found using \( \text{SupMax1} \), \( \text{SupMax2} \) and \( \text{SupMax3} \).

**The Exactness of the \( \text{SupMaxK} \) Family**

Lemmas 2 and 3 lead to Theorem 1, which shows the relationship between \( \text{SupMaxK} \) and \( \text{DiffSup} \).

**Theorem 1** \( \text{SupMaxK} \) is a lower bound of \( \text{DiffSup} \), for integer \( K \in [1, |\alpha| - 1] \).

**Proof 4** Since \( \text{DiffSup}(\alpha) \) is equivalent to \( \text{SupMaxK}(\alpha) \) with \( K = |\alpha| \) (We assumed \( \text{RelSup}^1(\alpha) \geq \text{RelSup}^2(\alpha) \) for simplicity earlier this Section), this theorem follows from Lemma 3.
Figure 2.3: An extended version of the data set shown in Figure 2.1 containing fifteen addition items \((i_{16} - i_{30})\) and two patterns \(P_5\) and \(P_6\), the rest being identical to Figure 2.1.

Theorem 1 guarantees that the patterns discovered by any \(\text{SupMaxK}\) members with threshold \(r\) also have \(\text{DiffSup} \geq r\). Therefore, \(\text{SupMaxK}\) members with threshold \(r\) discover only \(r - \text{discriminative}\) patterns.

**The Increasing Completeness of the \(\text{SupMaxK}\) Family**

The \(\text{max}\) function together with the anti-monotonicity of \(\text{RelSup}\) yields the following result about the anti-monotonicity of each member of \(\text{SupMaxK}\).

**Theorem 2** Each member of \(\text{SupMaxK}\) is anti-monotonic.

**Proof** Let \(\alpha \subseteq I\) be an itemset, and \(\alpha' \subseteq I\) be a superset of \(\alpha\), such that \(\alpha' = \alpha \cup \{i\}\), where \(i \in I\) and \(i \notin \alpha\). Firstly, from the anti-monotonicity of \(\text{RelSup}\), we have \(\text{RelSup}^1(\alpha') \leq \text{RelSup}^1(\alpha)\). Then, based on the property of the \(\text{max}\) function, \(\max_{\beta' \subseteq \alpha'}(\text{RelSup}^2(\beta')) \geq \max_{\beta \subseteq \alpha}(\text{RelSup}^2(\beta))\), where \(|\beta| = K\) and \(|\beta'| = K\). Finally, we have the following:

\[
\text{SupMaxK}(\alpha') = \text{RelSup}^1(\alpha') - \max_{\beta' \subseteq \alpha'}(\text{RelSup}^2(\beta')) \\
\leq \text{RelSup}^1(\alpha) - \max_{\beta \subseteq \alpha}(\text{RelSup}^2(\beta)) \\
= \text{SupMaxK}(\alpha).
\]
Based on Theorem 2, given a threshold \( r \), any member of the SupMax\( K \) family can be used within an Apriori-like framework \([10]\) to discover the complete set of patterns with \( \text{SupMaxK} \geq r \). Note that \( \text{SupMaxK} \) could be alternatively defined using the \( \min \) function, thus providing a better estimation of \( \text{DiffSup} \). However, this version of \( \text{SupMaxK} \) will not be anti-monotonic and thus cannot be used in the Apriori framework for the systematic search of discriminative patterns.

Since there are a finite number of discriminative patterns in a dataset given a \( \text{DiffSup} \) threshold, and \( \text{SupMaxK} \) finds more and more discriminative patterns as \( K \) increases (Lemma 3), the set of patterns discovered with \( \text{SupMaxK} \) and threshold \( r \) within an Apriori-like framework is increasingly more complete with respect to the complete set of \( r \)–discriminative patterns.

**Summary of the three properties of the SupMax\( K \) Family**

From the subset-superset relationship among \( \text{SupMaxK} \) members, and the exactness and increasing completeness of the \( \text{SupMaxK} \) family, \( \text{SupMaxK} \) members conceptually organize the complete set of discriminative patterns into nested subsets of patterns that are increasingly more complete in their coverage with respect to \( r \)–discriminative patterns. This yields interesting relationships between \( \text{DiffSup} \), \( \text{BiggerSup} \) and the \( \text{SupMaxK} \) family, which are discussed below.

### 2.5.4 Relationship between DiffSup, BiggerSup and the SupMax\( K \) Family

To understand relationship among \( \text{DiffSup} \), \( \text{BiggerSup} \) and \( \text{SupMaxK} \), Figure 2.4 displays the nested structure of the \( \text{SupMaxK} \) family together with \( \text{DiffSup} \) and \( \text{BiggerSup} \) from the perspective of the search space of discriminative patterns in a dataset. \( L_{\text{All}} \) is the complete set of \( r \)–discriminative patterns given a \( \text{DiffSup} \) threshold \( r \). \( L_{\text{CSET}} \) is the search space explored by CSET in order to find all the patterns in \( L_{\text{All}} \). Note that \( L_{\text{CSET}} \) is a superset of \( L_{\text{All}} \), because \( \text{BiggerSup} \) is an upper bound of \( \text{DiffSup} \). Note that, \( L_{\text{CSET}} \) can be much larger than \( L_{\text{All}} \) for dense and high-dimensional data sets, especially when a relatively low \( \text{BiggerSup} \) threshold is used. In such cases, CSET may not be able to generate complete results within an acceptable amount of time. For
instance, on the cancer gene expression data set used in our experiments, the lowest
$BiggerSup$ threshold for which CSET can produce the complete results within 4 hours
is 0.6. With a lower threshold 0.4, CSET can not produce the complete results within
24 hours.

Members of the $SupMaxK$ family help address this problem with $BiggerSup$ by
stratifying all the $r$-discriminative patterns into subsets that are increasingly more
complete (Set $L_1, L_2, \ldots, L_k, L_{k+1}, \ldots, L_{All}$), as shown in Lemma 3 and the subsequent
discussion, and illustrated in Figure 2.4. However, note that these superset-subset
relationships among $SupMaxK$ members and between $SupMaxK$ and $BiggerSup$ (used
by CSET) hold only when the same threshold is used for $BiggerSup$ and all the $SupMaxK$
members, and unlimited computation time is available. In practice, given the same fixed
amount of time, progressively lower thresholds can be used for $SupMaxK$ members as
$K$ decreases. This tradeoff was illustrated earlier in Figure 2.2(b).

Since the focus of this chapter is on dense and high-dimensional data, another prac-
tical factor should be considered, that is, the computational efficiency of the $SupMaxK$
members. In the next section, we will introduce a special member of the $SupMaxK$
family that is computationally suitable for dense and high-dimensional data.

2.5.5 $SupMaxPair$: A Special Member Suitable for High-Dimensional
Data

In the previous discussion, we showed that as $K$ increases, the set of patterns dis-
covered with $SupMaxK$ and threshold $r$ in an Apriori framework is increasingly more
complete with respect to the complete set of $r$-discriminative patterns. Thus, in
order to discover as many $r$-discriminative patterns as possible, an as large as
possible value of $K$ should be used given the time limit. However, it is worth not-
ing that the time and space complexity to compute and store the second com-
ponent in the definition of $SupMaxK$, i.e. $MaxSup(\alpha, K) = \max_{\beta \in \alpha}(RelSup^2(\beta))$ with
$|\beta| = K$ are both $O(m^K)$ (The exact times of calculation is $\left(\begin{array}{c}M \\ K \end{array}\right)$), where $M$ is the
number of items in the dataset. In high-dimensional data set (large $M$), $K > 2$
is usually infeasible. For instance, if there are 10000 items in the data set ($M =
10000$), even $SupMaxK$ with $K = 3$ will require the computation of the support of
all $\binom{10000}{3} \approx 1.6 \times 10^{11}$ size-3 patterns. Therefore, due to our emphasis on dense and
Figure 2.4: Nested layers \((L_1, L_2, L_3, \ldots, L_k, L_{k+1}, \ldots, L_{All}, L_{CSET})\) of patterns defined by \(\text{SupMaxK}\), and relationship with the complete set of discriminative patterns (layer \(L_{All}\)), and the search space of \(\text{BiggerSup}\) used by CSET (layer \(L_{CSET}\)). (PDb stands for "Patterns Discovered by".) \(\text{PDbSupMaxK}\) is a subset of \(\text{PDbSupMax}(K+1)\). Note that this figure only shows the subset-superset relationship, while the size of each rectangle does not imply the number of patterns in each set.

For high-dimensional data, we will focus on \(\text{SupMaxK}\) with \(K = 2\), i.e., \(\text{SupMaxPair}\), to balance the accurate estimation of \(\text{DiffSup}\) and computational efficiency. Note that, based on the definition of \(\text{SupMaxPair}\), the computational complexity of the second component of \(\text{SupMaxPair}\) (maximal pair-wise support in class \(S_2\)) for an itemset \(\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_l\}\) with size greater than 2 is \(O(l^2)\). However, according to the Apriori framework \([10]\), \(\text{MaxSup}(\alpha, 2)\) only depends on three terms that will have been computed before the computation of \(\text{MaxSup}(\alpha, 2)\) itself: \(\text{MaxSup}(\{\alpha_1, \alpha_2, \ldots, \alpha_{l-1}\}, 2)\) and \(\text{MaxSup}(\{\alpha_1, \alpha_2, \ldots, \alpha_{l-2}, \alpha_l\}, 2)\), and \(\text{MaxSup}(\{\alpha_{l-1}, \alpha_l\}, 2)\), and thus the computational complexity for \(\text{MaxSup}(\alpha, 2)\) is \(O(1)\) per itemset \(\alpha\).

As shown in Figure 2.4, \(\text{SupMaxPair}\) can perform a complete search of the \(r - \text{discriminative}\) patterns in the first two layers, even for a low value of \(r\). Indeed, we will demonstrate in our experimental results on a cancer gene expression data set (Section
that searching these two layers itself can enable SupMaxPair to discover many low-support patterns that may not be discovered by CSET within an acceptable amount of time. Furthermore, these patterns are statistically significant and biologically relevant. Before we discuss these results, we lay out the complete framework that we use for discovering discriminative patterns from dense and high-dimensional data.

2.6 Framework for Discriminative Pattern Mining

In this section, we explain the major steps in the framework used for discriminative pattern mining in our experiments:

- **Step 1**: This is an algorithm-specific step. For example, for SupMaxPair, all the item-pair supports are computed and stored in a matrix, whose \((i, j)\) entry is the item-pair support of item \(i\) and \(j\). The complexity of this step is \(O(nm^2)\), where \(n\) is the number of transactions, and \(m\) is the number of unique items. No such pre-computation has to be done for CSET.

- **Step 2**: The Apriori framework [10] is used in this step for discriminative pattern mining using the anti-monotonic measures BiggerSup and SupMaxPair. For SMP, discriminative patterns are firstly mined from one class and then mined from the other, while CSET discovers patterns once from the whole dataset.

- **Step 3**: To facilitate further pattern processing and pattern evaluation, we selected only the closed itemsets [49] from the complete set of itemsets produced.

For clarity, we refer to the version of this framework where BiggerSup is used for discovering patterns as CSET, while the version using SupMaxPair is referred to as SMP in the subsequent discussion. Our analysis of the quality of the patterns and the computational time requirements are presented with respect to the patterns produced by these complete pipelines.

2.7 Experimental Results

In order to evaluate the efficacy of different discriminative pattern mining algorithms, particularly CSET (a representative of the approaches in group B discussed in Section 2.2) and our proposed algorithm SMP, we designed two sets of experiments. The
first set of experiments utilize synthetic data sets with varying density and dimensionality to study the properties of CSET and SMP. The second set of experiments involve the application of CSET and SMP to a breast cancer gene expression data. The second set aims at a systematic evaluation of the statistical significance and biological relevance of the resultant patterns, thus validating the effectiveness of CSET and SMP for knowledge discovery from real data. All the experiments presented here were run on a Linux machine with 8 Intel(R) Xeon(R) CPUs (E5310 @ 1.60GHz) and 16GB memory.

2.7.1 Experiments on Synthetic Data Sets with Varying Density and Dimensionality

In the first set of experiments, we study the performance of SMP and CSET on synthetic binary data sets whose background can be fully controlled. Specifically, we created two collections of synthetic datasets respectively with (i) varying density and fixed dimensionality, and (ii) varying dimensionality and fixed density. We first describe the approach we used to create these two collections of data sets and then present the performance of SMP compared to CSET.

Methodology for Generating Synthetic Data Sets

Each synthetic data set have two major components: discriminative and non-discriminative patterns. Discriminative patterns are the target of the mining algorithms, while non-discriminative patterns are obstacles. As discussed in Section 2.2, an effective discriminative pattern mining algorithm should be able to prune the non-discriminative patterns at early stage while discovering discriminative patterns.

Ten discriminative patterns each of sizes 2, 4, 6, 8 and 10 were embedded in each synthetic dataset, resulting in a total of 50 discriminative patterns per dataset. To reflect the distribution of different types of discriminative patterns in real data, for each of the five sizes, we randomly determined a number of patterns (out of ten) that can be discovered by CSET but not SMP (type-I), and the remaining patterns that can be discovered by SMP but not CSET (type-II). Specifically, type-I patterns are those that have \( \text{DiffSup} \) greater than 0.2, but \( \text{SupMaxPair} \) below 0.2. As discussed in Section 2.5, SMP can not find type-I patterns due to the fact that \( \text{SupMaxPair} \) is an lower bound of
DiffSup. In contrast, type-II patterns are those that have BiggerSup below the lowest threshold (0.2) that CSET can finish within an acceptable amount of time (we use 4 hours as the representative acceptable amount of time). SMP can find these type-II patterns if it can effectively prune non-discriminative patterns and can search at lower support levels (0.1). Table 2.1 displays the number of type-I and type-II discriminative patterns of different sizes embedded in each of the synthetic datasets. Note that these numbers are kept the same for all the synthetic datasets to ensure that results across different datasets are comparable. Note that in practice, there may be other types of patterns that can be discovered by both CSET and SMP. In this analysis, we do not embed these other types of patterns and focus only on the effectiveness of CSET and SMP for discovering different types of discriminative patterns.

For all the synthetic data sets, we fix the number of samples at 700, in which half are of class 1 and the other half are of class 2. Two collections of datasets were generated as follows.

Varying density with fixed dimensionality: For this collection of data sets, we fix the dimensionality at 4000. After we embed the 50 discriminative patterns, we have the first dataset of density 10%. Next, we keep adding non-discriminative patterns of size 10 and support greater than 0.2, and create four more data sets with densities of 0.13, 0.16, 0.19 and 0.22 respectively.

Varying dimensionality with fixed density: For this collection of data sets, we fix the density of the dataset at 0.2. After we embed the 50 discriminative patterns (density 10%), we further add non-discriminative patterns to make the density equal to 0.2 and use this dataset as the first dataset (the dimensionality is 350). Next, we further add non-discriminative patterns of size-10 and support greater than 0.2 and simultaneously increase the dimensionality of the data set to maintain the density at 0.2. In this way, we create another four data sets with dimensionalities of 500, 2000,
Note that the supporting transactions of both the discriminative and non-discriminative patterns are selected randomly to avoid their combination into patterns of larger sizes. To simulate practical situations, for each data set generated in the above process, we add an additional 10% noise by flipping 10% of the 0’s to 1’s and 1’s to 0’s.

Performance of SMP and CSET on Synthetic Data Sets

For both the collections of datasets, we use a $\text{BiggerSup}$ threshold of 0.2 for CSET and a $\text{SupMaxPair}$ threshold of 0.1 for SMP. These thresholds agree with the definitions of type-I and type-II patterns for the following experiments (Section 2.7.1). The questions we want to answer in these experiments are: Which level of the itemset lattice can CSET and SMP reach when mining these synthetic datasets given the time limit of 4 hours, and correspondingly, how many of the discriminative patterns at each level can be discovered by the two algorithms?

Figures 2.5(a) and 2.5(b) display the levels that CSET and SMP reach on each of the five synthetic data sets of varying density and varying dimensionality respectively. Note that the highest level is 10, which is the size of the largest discriminative and non-discriminative patterns. Several observations can be made from Figure 2.5(a). First, when the density is 10%, both CSET and SMP can reach all the 10 levels. Thus, CSET can discover all the 29 type-I patterns (but none of the type-II patterns) and SMP can discover all the 21 type-II patterns (but none of the type-I patterns). Second, when the density increases to 13%, CSET only reaches level 3 and thus can only discover its 3 type-I patterns of size-2. In contrast, SMP can complete all the 10 levels and discovers all the 21 type-II patterns. Similar observation also holds for densities 0.16 and 0.19. This illustrates that even for reasonably high levels of density, SMP can discover type-II patterns with lower-support that can not be discovered by CSET, even though it can miss some type-I patterns that can be discovered by CSET. Finally, when density increases to 0.22, both SMP and CSET only reach level-2, i.e. CSET discovers its 3 type-I patterns and SMP discovers its 7 type-II patterns. This indicates that for relatively very high levels of density, both CSET and SMP can face challenges in discovering the embedded patterns that they are supposed to discover (i.e. type I patterns for CSET and type II patterns for SMP). However it should be noted that this deterioration in
the performance of SMP is due to the expense of the $O(N^2)$ time complexity in the generation of level-2 candidates. Indeed, even at this density (0.22), SMP can again finish all the 10 levels in only an additional 0.5 hour (total 4.5 hrs). However, CSET is still unable to generate all the level-3 candidates even in another 4 hours (total more than 8 hours). In summary, these results show that SMP is more effective for searching for low-support discriminative patterns on dense datasets.

Similar observations can also be made from Figure 2.5(b). First, at the dimensionality 350, both CSET and SMP can complete all the 10 levels and discover all the patterns they are supposed to find. Second, at dimensionality 500, 2000 and 4000, CSET can only reach up to levels 6, 3 and 2 respectively, while SMP still reaches all the 10 levels. Finally, at dimensionality 6000, both SMP and CSET can only complete level 2. Again, SMP can finish all the 10 levels in another half an hour, but CSET is still generating level-3 candidates in another 4 hours. These results show that SMP is more effective for searching for low-support discriminative patterns from high-dimensional datasets.

From the above experimental results on the two collections of synthetic datasets with varying density and varying dimensionality, we demonstrated the efficacy of SMP for mining low-support discriminative patterns from dense and high-dimensional data sets. Next, we will use a real gene expression data set to study the practical utility of SMP for discovering low-support discriminative patterns.
2.7.2 Experiments on a Breast Cancer Gene Expression Data Set

In the second set of experiments, we used CSET and SMP to discover discriminative patterns from a breast cancer gene expression data set. Only closed patterns are used in these experiments. The details of this data set are provided in Section 2.7.2. We first present a global analysis of these patterns in Section 2.7.2. Subsequently, we perform an extensive statistical and biological evaluation of these patterns, the results of which are presented in Sections 2.7.2 and 2.7.2. In particular, we highlight the statistical significance and biological relevance of low-support patterns discovered by SMP but not CSET, thus illustrating the complementarity that SMP can provide to the existing approaches discussed in Section 2.2.

Dataset description

A breast cancer gene expression data set [50] is used for evaluating the efficacy of discriminative pattern mining algorithms on complex, real data sets. This data set contains the expression profiles of about 25000 genes in 295 breast cancer patients, categorized into two classes corresponding to whether the patient survive the disease (0) or not (1). Using pre-processing methodologies suggested by the authors [51], we only considered 5981 genes that showed evidence of significant up- or down-regulation (at least a two-fold change), and whose expression measurements were accurate (p-value ≤ 0.01) for at least five patients. Furthermore, to make the dataset usable for binary pattern mining algorithms, each column pertaining to the expression of a single gene is split into two binary columns. Since the data has been properly normalized to eliminate between-gene variations in the scale of their expression values, we adopt a simple discretization method, as used in other studies [52, 53]: a 1 is stored in the first column if the expression of the gene is less than −0.2, while a 1 is stored in the second column if the expression of the gene is greater than 0.2. Values between −0.2 and 0.2 are not included, since genes showing an expression around 0 are not expected to be interesting, and may add substantial noise to the data set. The resulting binary data set has 11962 items and 295 transactions, with a density of 16.62%.

For this data set, discriminative pattern mining can help uncover groups of genes that are collectively associated with the progression or suppression of cancer, and our
experiments are designed to evaluate the effectiveness of different algorithms for this task.

**General analysis of the patterns discovered**

We ran CSET and SMP at the lowest parameter thresholds for which they would finish in about 4 hours\(^5\). Only closed patterns are used in our experiments. Due to the weaker pruning of \textit{BiggerSup} and the resulting large number of discriminative patterns, we were forced to use relatively higher thresholds for CSET and restrict the computation to patterns of a limited size to obtain the patterns necessary for our evaluation. Table 2.2 shows that the lowest \textit{BiggerSup} threshold for which CSET can produce the complete results within 4 hours is 0.6. The lowest \textit{BiggerSup} threshold for which CSET can discover size-2 and size-3 patterns within 4 hours is 0.55. At a lower threshold of 0.4, CSET can only discover size-2 patterns before running out of time. In contrast, SMP is

\(^5\) Some time period needed to be chosen for the experiments. The duration of four hours is, although slightly arbitrary, is generally reasonable for most data analysis operations.
<table>
<thead>
<tr>
<th>$BiggerSup$ Threshold</th>
<th>Time (sec)</th>
<th># Closed Patterns</th>
<th>Pattern Size(s)</th>
<th>Highest NegLogP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4*</td>
<td>617</td>
<td>64942</td>
<td>2</td>
<td>12.09</td>
</tr>
<tr>
<td>0.55*</td>
<td>1454</td>
<td>84840</td>
<td>2-3</td>
<td>9.65</td>
</tr>
<tr>
<td>0.6</td>
<td>1558</td>
<td>90637</td>
<td>2-10</td>
<td>8.78</td>
</tr>
</tbody>
</table>

Table 2.2: Details of patterns discovered by CSET at various $BiggerSup$ thresholds. (* Expansion of the set of patterns to patterns of larger sizes could not finish in over 12 hours, and thus, their results are not included here.)

able to run at a much lower $SupMaxPair$ threshold of 0.18 and finds patterns of size as high as 7 in about 40 minutes. See Table 2.3 for the details of the patterns found by SMP at different thresholds. For the evaluation of pattern quality, we combine the patterns discovered by CSET at the 0.4, 0.55, and 0.6 $BiggerSup$ thresholds as the collection of all patterns that can be discovered by CSET, while for SMP, we only use the patterns discovered at the single $SupMaxPair$ threshold 0.18. Indeed, even with this setup that is slightly biased towards CSET, there are still high quality low-support patterns that can only be discovered by SMP, the details of which are provided later.

In addition to analyzing the characteristics of the patterns discovered by SMP and CSET, we also examined the value of DiffSup for each individual gene constituting these patterns. Specifically, Figure 2.6 displays the distribution of the DiffSup of individual genes in the patterns discovered only by SMP at a SupMaxPair threshold of 0.18, but not by CSET. Among the 332 genes covered by these patterns, almost 60% (198) of the genes have DiffSup lower than the 0.18. Based on the discussion of approaches that directly utilize DiffSup or other measures of discriminative power for finding discriminative patterns (group C) in Section 2.2, it can be seen that these approaches can not discover any of these genes, and thus can not discover the patterns that include them. Since one of our major foci is on algorithms that can discover patterns whose individual genes may not be discriminative, we discuss only the results of CSET and SMP, which can find such patterns, in the rest of this section.

**Statistical Evaluation**

There are various ways to evaluate the importance of discriminative patterns. We are interested in patterns that occur disproportionately between the two classes. However,
Table 2.3: Details of patterns discovered by SMP at various $SupMaxPair$ thresholds.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Time (sec)</th>
<th># Closed Patterns</th>
<th>Pattern Size(s)</th>
<th>Highest NegLogP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.18</td>
<td>2401</td>
<td>45982</td>
<td>2-7</td>
<td>12.09</td>
</tr>
<tr>
<td>0.2</td>
<td>1187</td>
<td>21285</td>
<td>2-5</td>
<td>12.09</td>
</tr>
<tr>
<td>0.25</td>
<td>332</td>
<td>3007</td>
<td>2-4</td>
<td>12.09</td>
</tr>
<tr>
<td>0.3</td>
<td>186</td>
<td>283</td>
<td>2-3</td>
<td>12.09</td>
</tr>
</tbody>
</table>

in real world data sets, particularly those with small number of instances in the two classes, even patterns that occur with similar support across classes will show some deviation from perfect balance in data sets with relatively small sample size. Thus, to ensure that the patterns found are not just a result of random fluctuation, a statistical test is commonly used to ensure that any deviation from equal support is statistically significant. In this section, we will perform this type of evaluation for the patterns from CSET and SMP.

We use the Fisher exact test [54] for this evaluation, whose result is a p-value (probability). If the p-value is below some user defined threshold, e.g., 0.05 or 0.01, then the pattern is regarded as authentic. Note that p-values are often expressed as their negative $log_{10}$ value for convenience (the higher this $-log_{10}$ value (denoted as $NegLogP$), the more reliable the discriminative pattern is expected to be). We will refer to this measure as $NegLogP$. If there are multiple patterns, the $NegLogP$ threshold needs to be adjusted. By using a randomization test, as discussed below, we were able to determine that a $NegLogP$ of 8 is unlikely to arise from a random pattern. We give the technical details of this a bit later.

In Figure 2.7, we show plots of $NegLogP$ vs. global support for the patterns discovered by both CSET and SMP. For CSET, patterns discovered by using $BiggerSup$ thresholds 0.4, 0.55, and 0.6 were combined as described in Section 2.7.2, while for SMP, a 0.18 threshold was used. Several conclusions can be drawn from this figure. First, CSET finds more patterns than SMP, particularly for patterns with higher support (the ones with support greater than 0.4). This is not surprising since SMP sacrifices completeness to find lower support patterns. Second, CSET finds many patterns with $NegLogP$ less than 2, while all the patterns discovered by SMP have $NegLogP$ higher than 2. This demonstrates the exactness of $SupMaxPair$ (Theorem 1), i.e. because $SupMaxPair$ is
Figure 2.7: Plot of $\text{NegLogP}$ vs. global support for patterns from CSET and SMP, where the support is relative to the whole data set.

an lower bound of $\text{DiffSup}$, all the patterns discovered with $r$ are $r$–*discriminative*. Last and the most importantly, SMP finds many patterns at low support level that are not found by CSET, especially the ones with $\text{NegLogP}$ higher than the significance threshold 8. Also, these patterns are constituted by many genes that are not covered by the patterns discovered by CSET, as will be discussed in Section 2.7.2.

We now come back to the details of how we determined a significance threshold for $\text{NegLogP}$, both for the completeness of the above discussion and to further illustrate the quality of the patterns found by SMP but not found by CSET. Because of the issues of low sample size and high-dimensionality for data sets used for problems such as biomarker discovery, many patterns may be falsely associated with the class label. This raises the multiple-hypothesis testing problem [55, 44], which are addressed by various approaches, such as Bonferroni correction [56], false discovery rate control [57] and permutation test [56, 58, 44]. Permutation tests based on row-wise, column-wise and swap randomization [44] have been used to assess the statistical significance of the results of unsupervised pattern discovery and clustering algorithms. While in the context of labeled transactions, class-label permutation tests [45, 30] are often an effective option. In this approach, a reference distribution for evaluation measures like $\text{NegLogP}$ is generated by randomly shuffling the class labels (permutations). Specifically, for each
iteration, the class labels are randomly shuffled and reassigned to patients, discriminative patterns are found, and the $\text{NegLogP}$ values are computed for these patterns using the same method as for the patterns discovered with the true labels. The $\text{NegLogP}$ values from the random runs can be used to generate an empirical distribution for the $\text{NegLogP}$ values, which can be displayed as a histogram as in Figure 2.8. (Sometimes only the extreme (maximum) $\text{NegLogP}$ values are used as in this figure.) If a $\text{NegLogP}$ of a pattern derived from the true labels falls outside the main concentration of $\text{NegLogP}$ values from the random labels, then the $\text{NegLogP}$ very likely indicates a discriminative pattern with a "more than random" variation from equal frequency across classes.

Figure 2.8 summarizes the results of such a permutation test for the dataset being used in these experiments. The right hand side shows the top 300 $\text{NegLogP}$ of the patterns discovered only by SMP but not by CSET, while the left hand side displays the maximum $\text{NegLogP}$ for each of the 1000 permutation tests where randomized labels are used for pattern mining. We observe that the $\text{NegLogP}$ values with random labels rarely exceed 8 (less than 8.72 in each of the 1000 permutation tests). Thus, we can use 8 as a relaxed threshold for significance, since only a few percent of the random patterns are above this value. The $\text{NegLogP}$ values of the top-300 patterns discovered by SMP but not by CSET with true label are much higher (all larger than 9.67). In contrast, only 34 patterns discovered by CSET have a $\text{NegLogP}$ greater than 8. This shows that SMP can discover additional statistically significant low-support patterns. In the next section, we illustrate the biological significance of these patterns and how they can be used to discover cancer-related genes.

**Biological Relevance of Patterns based on a list of Cancer-related Genes**

There are various ways to determine the biological relevance of discriminative patterns. Since the application we consider is that of discovering biomarkers for cancer, we measured the biological relevance of the patterns using a list of about 2400 human genes known to be involved in the induction, progression and suppression of various types of cancers [46]. Of these 2400 genes, 611 were included in the set of 5981 genes in our processed gene expression data set. If the discriminative patterns found by CSET and SMP, which are just small sets of genes, tend to disproportionately contain these 610 cancer related genes as opposed to the non-cancer related genes, then this indicates that
these patterns contain information that may be of significance to a biological researcher. To make this idea concrete for the purposes of evaluation, two evaluation approaches were designed.

1. **Pattern-based Biological Relevance**: For each pattern generated by CSET or SMP, we matched the genes in the pattern with the set of 611 validated cancer genes, giving us a measure of the ‘precision’ of the pattern. For instance, if a pattern contains 3 genes, of which 2 are found to match the list of cancer genes, then the precision of this pattern is $\frac{2}{3} = 66.67\%$. Note that if a pattern with $N$ genes is randomly chosen from our set of 5981 genes, one would expect a precision of $\frac{N \times (611/5981)}{N} = 10.2\%$.

2. **Gene Collection-based Biological Relevance**: Since patterns may overlap with each other (pattern redundancy), and do not directly show how many cancer genes can be discovered by SMP in addition to CSET, we also designed a gene collection-based evaluation methodology. Here we collect the set of genes covered by all the patterns discovered by CSET(SMP), and compare this set of genes with the set of 611 validated cancer genes just as for pattern-based evaluation. For instance, if a set of 100 patterns covers 300 genes, of which 50 are found to match...
the list of cancer genes, then the precision of the set of patterns is \(\frac{50}{300} = 16.67\%\) and the recall is \(\frac{50}{611} = 8.18\%\). To compare, if we select 300 genes randomly from the 5981 genes, then the expected precision is \(\frac{300 \times (611/5981)}{300} = 10.2\%\), and the expected recall is \(\frac{300 \times (611/5981)}{611} = 5.02\%\).

This section details the results obtained from these evaluation methodologies.

**Brief Preview of Results:** From the pattern-based biological relevance evaluation, we observed that CSET can discover patterns with good precision at relatively high support level, while SMP can further discover good quality patterns at relatively low support level, among which, there are some patterns with 100% precision with respect to the cancer gene list. From the gene collection-based biological relevance evaluation, we observed that both the techniques discovered substantially more cancer genes than expected by random chance, especially among the higher \(NegLogP\) patterns. In particular, SMP was able to discover more cancer genes as compared to CSET due to its ability of discovering low-support patterns. This result further indicates the potential usefulness of recovering low-support patterns and discovering biomarkers that may be examined and utilized by the biology community. The following discussion provides additional details of these results.

**Results from Pattern-based Relevance:** Figure 2.9(a) shows the distribution of pattern-based precision of those patterns discovered only by SMP but not by CSET. For comparison, we generated a sequence of size-\(k\) patterns exactly according to the sizes of the patterns corresponding to Figure 2.9(a). The distribution of precision of these random patterns is shown in Figure 2.9(b). We can make the following observations from a comparison of Figure 2.9(a) and 2.9(b): (i) these patterns that are discovered exclusively by SMP include many that have a relatively high precision. Specifically, about 200 patterns have precisions above 0.6, among which there are 18 with a precision of 100%; (ii) the pattern-based precision of randomly generated patterns is mostly (about 1500 times) 0, and sometimes (about 300 times) fall into the range of 0.2 and 0.3, but rarely (less than 20) go beyond 0.4, and never go beyond 0.8. Interestingly, some of the SMP patterns with 100% precision play similar roles in cancer processes.
Figure 2.9: Comparison of the distributions of pattern-based precision between (a) the patterns discovered by SMP but not CSET and (b) random generated patterns.

**Results from Gene collection-based Relevance:** To investigate how many cancer genes can be discovered using CSET and SMP, we summarized the gene collection-based evaluation results for them in Tables 2.4 and 2.5 respectively. These tables include the number of cancer genes discovered, precision, recall, and expected recall for randomly selected group of genes of the same size. Note that, the expected precision for a random collection of genes is 10.2% as calculated earlier, and thus we do not include this in these tables. The following observations can be made from these tables.

1. Both CSET and SMP usually find very precise patterns for reasonably high levels of the \textit{NegLogP} measure, and this precision is much higher than that expected from a set of randomly selected gene collection of the same size (10.2%). Similarly, the recall values for the genes covered by these patterns are much higher than those expected from the same type of randomly selected gene collection, as shown by a comparison with the last column of these tables.

2. For similar values of cancer gene discovery precision, SMP generally finds more cancer genes than CSET. For instance, at a precision of about 25%, the recall of CSET is only 0.5% (3 cancer genes), while SMP has a recall 4.3% (26 cancer genes).
<table>
<thead>
<tr>
<th>NegLogP Threshold</th>
<th>Patterns</th>
<th># Genes Covered</th>
<th># Cancer Genes</th>
<th>Pre (%)</th>
<th>Rec (%)</th>
<th>ERec (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>66.7</td>
<td>0.3</td>
<td>0.052</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>66.7</td>
<td>0.3</td>
<td>0.052</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>66.7</td>
<td>0.3</td>
<td>0.052</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>12</td>
<td>3</td>
<td>25.0</td>
<td>0.5</td>
<td>0.21</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
<td>31</td>
<td>7</td>
<td>22.6</td>
<td>1.1</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 2.4: Precision-recall results of CSET patterns with $\text{BiggerSup} \geq 0.4$ (Pre: Precision, Rec: Recall, Expected Precision for random gene collections is 10.2%, ERec: Expected Recall of Random Gene Collections with the Same Size)

Note that the highlight of the second observation is not that SMP discovers more cancer genes, but that SMP can discover cancer genes from discriminative patterns with low-support in addition to the ones discovered by CSET, thus indicating the complementarity of SMP to existing approaches like CSET. Because of such complementarity, even if SMP discovered less cancer genes than CSET, SMP still complement CSET as long as additional genes are exclusively discovered by SMP. Indeed, from the specific example in the second observation, at least 23 cancer genes are discovered by SMP in addition to CSET.

**Biological Relevance of Patterns based on Biological Gene Sets**

An alternative way of evaluating the biological relevance of the patterns discovered only by SMP but not by CSET is to estimate how well they capture the 5452 known biological gene sets (e.g. pathways) in the Molecular Signature database [45]\(^6\) (MSigDB). MSigDB is a widely used collection of gene groups containing genes with similar biological functions. The methodology we adopt for this evaluation is one of calculating the enrichment of one pattern with these gene groups. This enrichment is measured as the probability of a random pattern of the same size having the same or better annotations by a given gene group by random chance, and the lower this probability the more enriched a pattern is with a given gene group. Specifically, for a pattern of size $k$ and a gene set of size $m$ which share $x$ common genes, we use the hypergeometric cumulative

Table 2.5: Precision-recall results of SMP patterns with $SupMaxPair \geq 0.18$ (Pre: Precision, Rec: Recall, Expected Precision for random gene collections is 10.2%, ERec: Expected Recall of Random Gene Collections with the Same Size)

<table>
<thead>
<tr>
<th>NegLogP Threshold</th>
<th># Patterns</th>
<th># Genes Covered</th>
<th># Cancer Genes</th>
<th>Pre (%)</th>
<th>Rec (%)</th>
<th>ERec (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>50.0</td>
<td>0.3</td>
<td>0.067</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>7</td>
<td>3</td>
<td>42.9</td>
<td>0.5</td>
<td>0.12</td>
</tr>
<tr>
<td>10</td>
<td>200</td>
<td>36</td>
<td>11</td>
<td>30.6</td>
<td>1.8</td>
<td>0.60</td>
</tr>
<tr>
<td>9</td>
<td>541</td>
<td>57</td>
<td>17</td>
<td>29.8</td>
<td>2.8</td>
<td>0.95</td>
</tr>
<tr>
<td>8</td>
<td>1502</td>
<td>103</td>
<td>26</td>
<td>25.2</td>
<td>4.3</td>
<td>1.72</td>
</tr>
</tbody>
</table>

distribution function \(^7\) to compute the probability that there are greater or equal to \(x\) common genes between the pattern and the gene set by random chance given that the total number of genes in the data set is \(N\) [59]. The $-\log$ value of this probability can be considered as an enrichment score between a pattern and a gene set (denoted by $NegLogEnrichP$), and the larger this score, the more significant the biological relevance of the pattern. For each pattern, we use the best $NegLogEnrichP$ with the 5452 gene sets as a measure of its biological relevance.

Instead of directly applying the above enrichment methodology to all the patterns that are discovered only by SMP but not by CSET, we first select a subset in which no pairs of patterns have greater than 25% overlap of genes. This selection helps reduce the effect of the redundancy between these patterns on the enrichment results. The resultant set has 37 patterns. Figure 2.10 shows the distribution of the best $NegLogEnrichP$ values of these 37 patterns with respect to the gene sets in MSigDB. It can be observed that more than half of the patterns (20) have at least two genes overlapping with one or more gene sets, and some patterns even have a $NegLogEnrichP$ value as high as 8 (original p-value as low as $10^{-8}$). Interestingly, some of the patterns in this collection are enriched with several gene sets that are clearly related to breast cancer such as $BREAST-DUCTAL-CARCINOMA-GENES$ ($NegLogEnrichP = 8.02$) and $BREAST-CANCER-PROGNOSIS-NEG$ ($NegLogEnrichP = 6.73$), as well as several gene sets that are related to general cancer-related biological processes such as the cell-growth-related gene set $IRITANI-ADPROX-LYMPH$ [45] ($NegLogEnrichP = 6.67$)

\(^7\) \(p(x|k, m, N) = 1 - \sum_{i=0}^{x-1} \binom{N-i}{k-i} \binom{m}{i} \binom{(N-k)}{(N-x)}\)
and the proliferation-related gene set $HOFFMANN-BIVSBII-BI-TABLE2[45]$ (NegLogEnrichP = 6.15). These results further support the biological relevance of the patterns discovered only by SMP but not by CSET, and thus demonstrate the benefits of using SMP to search for low-support discriminative patterns in addition to existing approaches.

**Comparison of the scalability of the algorithms**

In section 2.7.1, we compared the effectiveness of CSET and SMP for discovering low-support patterns from synthetic datasets with varying density and dimensionality. In this part of the study, we test the scalability of CSET and SMP with varying thresholds on the real gene expression data. In addition, we also test the FP-Close (FPC) [60] algorithm (plus pattern selection) as the baseline as used by other studies [11, 28]. Note that, as mentioned in Section 2.7.2, the gene expression data set was discretized with $\pm 0.2$ as thresholds, into a binary matrix with density 16.62% and dimension 11962, to preserve most of the information in the data. This dataset is quite dense, due to which CSET can only generate complete results at a threshold larger than 0.6. In order to obtain a more complete picture of the scalabilities of FPC, CSET and SMP, we discretized the gene expression dataset using $\pm 0.3$ as the discretization threshold in this section, which yields a binary matrix with density 8.71%.

Figure 2.11 shows the results of this comparisons. The X-axis in this plot is the
Figure 2.11: Scalability of different discriminative pattern mining algorithms on the gene expression data

threshold used for discriminative pattern mining, while the Y-axis denotes the log_{10}(run-time in seconds) value. Note that run-times are recorded for any algorithm only if it can produce output within four hours. The relative minsup threshold used in FPC is defined on the whole dataset (both classes), while BiggerSup for CSET and SupMaxPair for SMP take into account the support in each of the classes individually. Therefore, for a fair comparison, FPC’s minsup is adjusted according to the size ratio of the two classes (divided by the percentage of the majority class in the whole dataset (0.74)) and then plotted together with BiggerSup and SupMaxPair.

Several observations can be made from these plots: (i) the FPC-based two-step approach can search for discriminative patterns at high support levels (above 0.55), (ii) by using BiggerSup, CSET is able to search at slightly lower support levels (above 0.5) compared to FPC; and for the same threshold, CSET is more efficient than FPC, and (iii) SupMaxPair can explore pattern space with substantially lower support levels (0.1 – 0.3). Thus, FPC and CSET can be used to discover patterns at higher thresholds, while SMP is able to find lower support patterns missed by the other approaches.
2.7.3 Summary of Results

Based on the experimental results on both the synthetic datasets and the cancer gene expression data set presented in this section, we have demonstrated that on dense and high-dimensional data, there are patterns with relatively low support that can only be discovered by $SupMaxPair$ but not by the existing approaches. Specifically, on the cancer gene expression data set, the low-support discriminative patterns discovered only by SMP are statistically significant and biologically relevant.

We also did another set of experiments for studying how well the members of $SupMaxK$ approximate $DiffSup$ as $K$ increases. We selected several UCI datasets [61], on which all the discriminative patterns (given a relatively low $DiffSup$ threshold) can be discovered and used for the study. The experimental results show that: (i) $SupMax1$ generally provides very poor approximation of $DiffSup$; (ii), the approximation is improved substantially when $K$ goes to 2, i.e. $SupMaxPair$; (iii) when $K$ is increased further to 3 and 4, the computation time increases exponentially, but the approximation improves much slower compared to the improvement obtained when $K$ goes from 1 to 2. These experimental results indicate that $SupMaxPair$ provides a good balance between the approximation of $DiffSup$ and the computational expense. The detailed results are discussed as a supplementary material (available at http://vk.cs.umn.edu/SMP/).

2.8 Related work

Over the past decade, many approaches have studied discriminative pattern mining and related topics. Dong and Li [20] defined emerging patterns (EP) as itemsets with a sufficiently large growth rate (support ratio) between two classes. A two-step algorithm is proposed to discover EPs, which first finds frequent itemsets with the Max-Miner algorithm [62] for each of the two classes, and then compares these itemsets to find EPs. Emerging pattern were the first formulation of discriminative patterns and have been extended further to several special cases such as jumping emerging patterns [63] and minimal emerging patterns [64, 13]. Here, the discriminative power of a pattern is measured with support ratio [20], or simply with the two supports of the pattern in the two classes and two corresponding thresholds [64]. As discussed in [21], these emerging pattern mining algorithms must mine the data multiple times given a certain threshold.
for support ratio (or two thresholds for the two supports). In [21], a new formulation of discriminative patterns, contrast sets (CSETs), is proposed along with an algorithm to mine them. CSET is the first technique that formulates discriminative pattern mining within an Apriori-like framework [10, 62], in which different pruning measures can be used to perform a systematic search on the itemset lattice [10]. In [65], contrast set mining is shown to be a special case of a more general task, namely rule learning, where a contrast set can be considered as an antecedent of a rule whose consequent is a group. Notably, CSET has also been used in some biomedical applications [33]. The upper bounds of statistical discriminative measures have also been studied for discriminative pattern mining e.g. information gain [12], $\chi^2$-test [21] and several others [48].

Next, we also briefly discuss other research work related to discriminative pattern mining, although they are not the focus of the chapter. Many existing approaches have studied the use of frequent patterns in classification. Associative classifiers [35, 36, 37, 38, 42] are a series of approaches that focus on the mining of high-support, high-confidence rules that can be used in a rule-based classifier. Cheng et. al. [12] recently conducted a systematic evaluation of the utility of frequent patterns in classification. Several pattern-based classification frameworks have also been proposed, in which a small number of discriminative patterns are selected, which can achieve comparable classification accuracy with respect to the whole set of discriminative patterns [11, 28, 66, 29]. Discriminative pattern mining from multiple classes has been studied in [21, 13, 33], while mining complex discriminative patterns has been studied in [64]. Although traditional pattern summarization approaches [67] can be adopted to control the redundancy among discriminative patterns, closeness and redundancy are specially studied for in the context of discriminative patterns respectively in [68] and [69].

### 2.9 Conclusions

In this chapter, we addressed the necessity of trading off the completeness of discriminative pattern discovery, with the ability to discover low-support discriminative patterns from dense and high-dimensional data within an acceptable amount of time. For this, we proposed a family of anti-monotonic measures of discriminative power named $\text{SupMax}_K$
that conceptually organize the set of discriminative patterns into nested layers of subsets, and are progressively more complete in their coverage, but require increasingly more computation for their discovery. Given the same and fixed amount of time, the \textit{SupMaxK} family provides a tradeoff between the ability to search for low-support discriminative patterns and the coverage of the space of valid discriminative patterns for the corresponding threshold. In particular, \textit{SupMaxK} with \(K = 2\) named \textit{SupMaxPair}, is a special member of this family that is suitable for dense and high-dimensional data. We designed a framework, named SMP, which uses \textit{SupMaxPair} for discovering discriminative patterns from dense and high-dimensional data. A variety of experiments on both synthetic datasets and a breast cancer gene expression dataset demonstrated that there are patterns with relatively low support that can be discovered using SMP but not by the existing approaches. In particular, the low-support discriminative patterns discovered only by SMP from the gene expression dataset are statistically significant and biologically relevant. In summary, SMP can complement existing algorithms for discovering discriminative patterns by finding patterns with relatively low support from dense and high-dimensional data sets that other approaches fail to discover within an acceptable amount of time. Thus, in practice, it is recommended that CSET and other existing approaches should be used to discover medium-to-high support patterns from such data sets within an acceptable amount of time, and then SMP could be used to further discover low-support discriminative patterns that existing approaches may not discover.

Our work can be extended in several directions. As discussed in Section 2.5.4, the members of \textit{SupMaxK} induce a hierarchy of subsets of the complete set of discriminative patterns. This hierarchy motivates further research that focuses on the mining of discriminative patterns from the other layers that are not covered by \textit{SupMaxPair}. It is also interesting to study the quality of the discriminative patterns in the different layers of this hierarchy, which may provide insights into different priorities for discriminative pattern mining from these layers. Note that, the use of measures from the \textit{SupMaxK} family is only one possible method for trading off the completeness of pattern discovery with the ability to discover low-support discriminative patterns from high-dimensional data. Indeed, other approaches that adopt a different strategy for handling this trade-off are also possible and should be studied. Also, most existing discriminative pattern
mining algorithms (as well as SMP) are designed for binary data, and have to rely on
discretization for continuous data. It will be useful to design approaches that can di-
rectly handle continuous data for discriminative pattern mining, as has been done for
discovering patterns in an unsupervised manner[70].
Chapter 3

Continuous Data, Different Types of Discriminative Patterns and Novel Algorithms for Efficient Discovery

3.1 Introduction

This chapter aims to addresses the challenge confronted by existing discriminative pattern mining techniques on different types of biomedical and genomic data with continuous values. First, I motivate the need of searching for different types of biologically relevant discriminative patterns defined for continuous data. Then, I discuss strategies for efficient pattern mining. Specifically, I present a general framework named subspace differential coexpression analysis, for direct discovery from continuous case-control data. This approach can be used to adapt a family of biclustering algorithms to obtain their corresponding differential versions that can directly discover differential coexpression patterns. The proposed approach is applied to gene expression datasets from three lung cancer studies discovered differential coexpression patterns supported by different subsets of patients that are enriched for genes from known cancer pathways, and some are enriched with the target gene sets of cancer-related microRNA and transcription
3.2 Background

Diseases are often caused by perturbations in networks of genes or their products that are working together to keep a cell in a healthy state. DNA microarrays are one of the most popular technologies for studying these perturbations and understanding their effect on the expression of genes at a large scale, and eventually linking them to diseases. The genome-wide expression profiles of many types of diseases, particularly tumors, have been analyzed, and several associations have been identified between gene expression profiles and phenotypes corresponding to different stages of cancer [71]. Traditional analysis of gene expression data for this task focuses on the identification of (groups of) genes with substantially different expression values (up- or down-regulated) across sample-classes of interest, commonly known as differentially expressed (DE) genes (or patterns) [72]. An example of such a group of differentially expressed genes is shown in Figure 3.1(a), where these genes have significantly higher expression levels in the disease class than in the control class.

However, given that diseases are often caused by the disruption of a system, or network, of genes, identifying only the individual differentially expressed genes may not be adequate for discovering the underlying mechanisms of all the diseases. An important example of such mechanisms is the dysregulation of signaling pathways in cancer [73]. A complementary view for studying these mechanisms is provided by a differential coexpression pattern (DC) [74, 75, 1, 76], which is defined as a set of genes that have substantially different levels of coherence of their expression profiles in the two sample-classes, i.e., highly coherent in one class, but not in the other. An example of a DC pattern is shown in Figure 3.1(b), where the constituent genes are either all up-, down-, or neutrally-regulated for each sample in the control group (shown by the vertical streaks), but they do not follow any particular trend in the disease group. Biologically, a differential coexpression pattern may indicate the disruption of a regulatory mechanism possibly caused by the deregulation of a pathway [1] or a mutation of a transcription factor [77, 78], among other mechanisms. Figure 3.1(c) illustrates one of these mechanisms, where the mutation of a regulator causes the disruption of the normal activity
of a pathway. Specifically, $G_0$ is a dominant regulator of $G_1$-$G_4$ that leads to the coordinated and hence coherent expression of all of them. However, once mutated in the disease state, $G_0$ is unable to regulate these genes, and their regulation may be taken over by other independent regulators that may only be active in the disease state. Now, since the regulation of $G_1$-$G_4$ is independent, they are no longer coordinated, thus leading to the disruption of the coherence of their expression. Therefore, DC patterns can serve as biomarker candidates for some diseases, e.g. cancer and its subtypes [1, 79, 78], as well as for differentiating between evolutionarily-related species [77]. Furthermore, at a representation level, a DE pattern can be considered as a connected subgraph of a coexpression network, which is intact in one sample-class but not connected in the other. Such a convenient representation of these patterns can be very useful for their visualization and understanding, and we present some examples of this in Section 3.4.

Owing to their definition, differential coexpression patterns cannot be discovered via univariate analysis, since the coherence of expression values of a group of genes has to be measured collectively. Corresponding to this need, several techniques for identifying DC patterns have been proposed in the literature, the first ones of which only searched for
gene-pairs with sufficiently different correlations (or other statistical measures) between the two classes [74, 75, 80, 81]. Extending this to larger groups of genes, differential coexpression has also been studied in the context of clustering [82, 77] and coexpression networks [83, 84, 85, 76, 86, 79], where a cluster or a sub-network of genes is considered differentially coexpressed if they collectively have different pairwise coexpression across the sample-classes of interest. Some algorithms also employ differential coexpression measures collectively for a set of genes [1, 87], instead of only pairwise coexpression measures. Recently, some studies have adopted a related but different perspective and have proposed methodologies for identifying differentially coexpressed gene-pathway pairs [88] and pathway-pathway pairs [89].

Despite the differences in the methodologies adopted by these approaches for finding DC patterns, a feature common to all of them is that the coexpression of a set of genes is measured over all the samples in each of the two classes, i.e., over the full space of samples. For instance, the example shown in Figure 3.1(b) is a full-space DC pattern. However, as pointed out for the discovery of differentially expressed genes [90, 91, 92], the causes of diseases as well as the population affected by them, are often heterogeneous in nature. In such a scenario, full-space approaches may not always be appropriate and may ignore patterns that cover only a subset of the samples in each class, i.e., subspace patterns. For instance, a set of genes may only be coexpressed over 60% of the samples in the normal class, and may not be even slightly coexpressed over any of the samples in the disease class, thus qualifying to be a valid subspace pattern. However, this pattern may not be uncovered if the discovery algorithm requires the constituent genes to be coexpressed over all the samples in the normal class. Indeed, even if a pattern can be discovered by both full-space approaches and subspace approaches, the latter can better indicate the subgroup of samples on which the pattern is coexpressed, and thus may allow further study of the different causes of diseases and different demographics among subgroups of samples, which may potentially help personal diagnosis and treatment. These challenges call for the design of new approaches that can discover patterns that only show differential coexpression over subsets of the samples in the two classes, and can also indicate these subsets as a companion to the patterns. Interestingly, similar challenges faced by traditional clustering approaches have motivated the design of a variety of biclustering algorithms [93, 94].
In this chapter, we address these challenges by extending differential coexpression analysis to enable the discovery of subspace DC patterns. We define these patterns as sets of genes that are coexpressed over a relatively large percent of the samples in one class, but in a much smaller percent of samples in the other class. Following this definition, we propose a general approach based upon association analysis framework[10] that allows exhaustive\(^1\) yet efficient discovery of subspace differential coexpression patterns. This approach can be used to adapt a family of biclustering algorithms that have antimonotonicity[95, 96, 97, 98, 70] to obtain their corresponding differential versions that can directly discover differential coexpression patterns. Specifically, we illustrate the features of our approach by extending a recently developed biclustering algorithm [70]. Experiments using this approach on lung cancer datasets demonstrate the existence of subspace differential coexpression patterns in real-life data. Permutation tests demonstrate the statistical significance for a large number of discovered subspace patterns, many of which can not be discovered if they are measured over all the samples in each of the classes. Interestingly, some discovered patterns also have a significant overlap with known cancer pathways, and some are enriched with the target gene sets of a cancer-related microRNA and a cancer-related transcription factor. These results suggest that subspace DC patterns may aid in developing new understanding about the mechanisms underlying cancer and other diseases.

3.3 Proposed Approach

In this section, we first extend differential coexpression analysis to subspace patterns, then we will describe a general approach for the discovery of subspace differential coexpression patterns.

3.3.1 Subspace Differential Coexpression Analysis

A subspace differential coexpression pattern is a set of genes that are highly coexpressed in a relatively large percent (not necessarily all) of samples in one class, but in a much smaller percent of samples in the other class. We formulate the problem of subspace

\(^1\) Given a threshold, an exhaustive search guarantees to discover all the patterns w.r.t. that threshold. Different from brute-force search, exhaustive search may avoid exploring the whole search space by pruning a large number of patterns that are guaranteed to disqualify the threshold.
differential coexpression pattern discovery as follows. Let \( D \) be a gene expression dataset with a set of \( p \) genes, \( G = \{g_1, g_2, \ldots, g_p\} \), and two classes of samples, \( A \) and \( B \), which can be considered as cases and controls of size \( N_A \) and \( N_B \), respectively, i.e., \( A = \{a_1, a_2, \ldots, a_{N_A}\} \) and \( B = \{b_1, b_2, \ldots, b_{N_B}\} \). Let \( \Psi \) be a coexpression measure for a set of genes \( \alpha (\alpha \subseteq G) \). To illustrate, this measure could be a test as to whether the minimum of the pairwise correlation of the expression profiles of the genes in \( \alpha \) is above a particular threshold. We use \( A_\Psi(\alpha) (B_\Psi(\alpha)) \) to denote the subset of samples in \( A \) (\( B \)) on which \( \alpha \) is coexpressed, i.e., \( A_\Psi(\alpha) \subseteq A \) and \( B_\Psi(\alpha) \subseteq B \). The two ratios, \( \frac{|A_\Psi(\alpha)|}{|A|} \) and \( \frac{|B_\Psi(\alpha)|}{|B|} \) are respectively the percentage of samples in \( A \) and \( B \) on which \( \alpha \) is coexpressed. They are denoted as \( R^A_\Psi(\alpha) \) and \( R^B_\Psi(\alpha) \), respectively. The absolute difference of these two ratios can be used to measure the subspace differential coexpression of \( \alpha \):

**Definition 3** Subspace Differential Coexpression (SDC)

\[
SDC^\Psi(\alpha) = |R^A_\Psi(\alpha) - R^B_\Psi(\alpha)|
\]  

(3.1)

Given a threshold \( d \), a set of genes \( \alpha (\alpha \subseteq G) \) is called \( d \)-differentially coexpressed if \( SDC^\Psi(\alpha) \geq d \). Then, the problem of subspace differential coexpression pattern discovery with reference to a threshold \( d \) can be formulated as discovering all the \( d \)-differentially coexpressed patterns.

We will explain our approach for addressing this problem using Figure 3.2, which shows a number of types of subspace and full-space, differentiating and non-differentiating, coexpression patterns. Figure 3.2(a) shows a conceptual example of a differential full-space pattern, while Figure 3.2(b) shows a conceptual example of a differential subspace pattern. Figures 3.2(c) and 3.2(d) are examples of non-differential patterns. Although Figure 3.2(e) is a differential full-space pattern, it contains a redundant gene, i.e., the dashed curve.

Given Definition 3, an effective mining algorithm is expected to discover patterns like (a), (b) and (e) in Figure 3.2, but not the patterns that are equally coexpressed in the two classes (as shown in Figure 3.2(c) and (d)). However, if we take a further look at pattern (e), we can observe that, although the four genes together have differential coexpression, two genes are coexpressed in both of the two classes (the two red curves). Only one of the two coexpressed genes is enough to form a differential coexpression pattern with the other two genes (the two blue curves). Indeed, including both genes
Figure 3.2: Different types of full-space and subspace, differential and non-differential coexpression patterns. Each curve denotes the expression values of a gene on all the samples. The horizontal line separates all the samples into class A and B. Five patterns are illustrated, with a brown line indicating the samples on which the set of genes are coexpressed.

A common property of the five patterns in Figure 3.2 is that, they are all coexpressed in a large percent of samples in class A. In the meanwhile, the common property of the three patterns that are expected to be pruned, namely (c),(d) and (e), is that they all have at least one pair of genes that are coexpressed in a large percent of samples in class B. Motivated by these two observations, we refine the target of subspace differential coexpression pattern mining as those sets of genes that are coexpressed in a relatively large percent of samples in one class, while all of the pairs of genes in the set are coexpressed in a much smaller percent of samples in the other class. Mathematically, we define a measure for this refined criteria as follows:
Definition 4  Refined Definition of Subspace Differential Coexpression ($\widetilde{SDC}$) (Assume $R_A^\Psi(\alpha) \geq R_B^\Psi(\alpha)$)

$$\widetilde{SDC}^\Psi(\alpha) = R_A^\Psi(\alpha) - \max_{i,j \in \alpha}(R_B^\Psi(\{i,j\})) \quad (3.2)$$

$\widetilde{SDC}$ is computed as the difference between the percent of samples in class $A$ on which $\alpha$ is coexpressed and the maximal percent of samples in class $B$ on which a size-2 subset of $\alpha$ is coexpressed. A large value for $\widetilde{SDC}$ indicates that a set of genes, $\alpha$ is coexpressed on a much larger percent of samples in class $A$ compared to the coexpression of any size-2 subset of $\alpha$ in class $B$. Therefore, given a proper threshold, $d$, $\widetilde{SDC}$ can differentiate interesting subspace differential coexpression patterns like patterns (a) and (b) from uninteresting patterns like patterns (c) – (e).

Mathematically, for some coexpression measures, $\widetilde{SDC}$ has another property called antimonotonicity, which basically means that $\widetilde{SDC}(\alpha)$ is guaranteed to be no less than the $\widetilde{SDC}$ of any superset of $\alpha$. For $\widetilde{SDC}(\alpha)$ to have the antimonotonicity property, it is sufficient that the coexpression measure used to define $\widetilde{SDC}(\alpha)$ is antimonotonic (A formal proof is given in Fang et al. [99]). Indeed, the coexpression measures used in several existing association-based and subspace clustering based biclustering algorithms have this property [70, 95, 96, 97, 98]. This antimonotonicity property guarantees that, given a threshold, $d$, $\widetilde{SDC}$ can be used in a systematic yet efficient pattern mining framework, namely Apriori [10], to discover all and only the patterns with $\widetilde{SDC} \geq d$. We briefly describe the computational algorithm for this approach in Section 3.3.2.

3.3.2 Computation Algorithm

The Apriori framework is essentially a bottom-up exhaustive combinatorial search framework initially designed for association analysis on binary data. Different from brute-force search, given an antimonotonic measure $M$ and a threshold $m$, the Apriori search algorithm can avoid exploring the whole search space of all sets of items (genes in our case) by pruning a large number of candidates that are guaranteed to disqualify the threshold based on the antimonotonicity of $M$.

The process of searching patterns with $\widetilde{SDC} \geq d$ in the Apriori framework can be viewed as the generation of a level-wise pattern tree. Every level of the tree contains patterns with the same number of genes. If the level is increased by one, the pattern
size (number of genes in each pattern) is also increased by one. Every pattern has a branch (sub-tree) which contains all the supersets of this pattern. The search is breadth-first. We first check all the patterns at the second level, since the elemental component of differential coexpression analysis is a pair of genes. If a pattern does not satisfy the user-specified $\widetilde{SDC}$ threshold $d$, the whole branch corresponding to this pattern can be pruned without the need of further checking. This is guaranteed by the antimonotone property of $\widetilde{SDC}$ measures [99]. Following this approach, the pattern tree grows level-by-level until all the qualified patterns have been discovered. This algorithm is systematic yet efficient for handling large-scale datasets. Note that, in Definition 4, it is assumed that $R^\Psi_A(\alpha) \geq R^\Psi_B(\alpha)$. In practice, the algorithm will be run twice, one time to find patterns for which $R^\Psi_A(\alpha) \geq R^\Psi_B(\alpha)$, and the other to find patterns for which $R^\Psi_B(\alpha) \geq R^\Psi_A(\alpha)$. Use of the general measure $\widetilde{SDC}$ in the Apriori framework allows the effective pruning of non-differential coexpression patterns like (c) and (d), and also controls gene redundancy in patterns like (e). $\widetilde{SDC}$ also provides the antimonotonicity that allows exhaustive yet efficient discovery of differential coexpression patterns like (a) and (b) (Figure 3.2) in the Apriori framework. We will use $\widetilde{SDC}$-Apriori to denote the approach of using the general measure $\widetilde{SDC}$ in the Apriori algorithm.

The coexpression measures used in several existing association-analysis-based and subspace-clustering-based biclustering studies have the antimonotonicity property [70, 95, 96, 97, 98] and can be adapted to yield their corresponding differential versions that can directly discover differential coexpression patterns. Because of the complementarity of biclustering algorithms (i.e. they may discover patterns in common with each other, as well as some unique to their formulation), their corresponding differential versions are also complementary to each other.

### 3.3.3 DiffRange: an illustration of $\widetilde{SDC}$

In this chapter, we shall use a specific instance of this approach based on a recently proposed antimonotonic coexpression measure, namely range-support [70]. This measure is intended for the discovery of constant-row biclusters\(^2\) [94] in the Apriori framework.

\(^2\) In a constant-row bicluster, the set of genes have similar expression values on each condition/sample.
Conceptually, a range-support pattern is a set of genes that are coexpressed (the expression value of the set of genes fall within a close range) over a set of conditions in a gene expression data matrix. Let \( \text{RangeSup}_A^r(\alpha) \) denote the range-support of \( \alpha \) in class \( A \) (an instantiation of \( R^r_A(\alpha) \)), i.e. the percentage of samples in class \( A \) that fall within the predefined range threshold \( r \). From Definition 4(\( \tilde{SDC} \)), the corresponding differential range-support measure \( \text{DiffRange} \) (Differential Range-support) can be adapted:

**Definition 5** Given a range threshold \( r \), the \( \text{DiffRange} \) of a subset of genes \( \alpha \) (\( \alpha \subseteq G \)) on class \( A \) and \( B \)

\[
\text{DiffRange}(\alpha) = \text{RangeSup}_A^r(\alpha) - \max_{i,j \in \alpha} (\text{RangeSup}_B^r(\{i, j\})) \tag{3.3}
\]

### 3.4 Experimental Results

In this section, we describe the experimental design for the analysis of the subspace differential coexpression patterns discovered by \( \text{DiffRange} \). The, we present experimental results which demonstrate that the proposed general approach discovers statistically significant and biologically relevant subspace differential coexpression patterns in real-life data.

#### 3.4.1 Datasets and Preprocessing

In the experiments, three lung cancer datasets [100, 101, 102] are used, which are all generated with Affymetrix microarrays.\(^3\) To have a larger sample size for better illustration of the existence of subspace patterns and of their statistical significance\(^4\), we combined the three datasets resulting in 102 samples with lung cancer and 67 normal samples (total 169 samples). Across the three datasets, 8787 genes are common. We preprocessed the three datasets with RMA-normalization [103]. Additional cross-platform normalization algorithms [104, 105] were also tested and gave similar results, so only RMA normalized results are included here. The effect of different normalization methods on differential coexpression pattern mining will be studied in future work.

---

\(^3\) The first two use platform HG-U95A, while the other uses platform HG-U133A

\(^4\) The patterns discovered from the three datasets separately are not statistically significant in the permutation test (refer to Section 3.4.3 for details), due to the low sample size of each individual datasets.
Instead of normalized gene expression data, we used rank-converted values, i.e., the expression values are converted to expression ranks ranging from 1 to 169 (number of samples) separately for each gene (similar as used in Spearman’s rank correlation). Our analysis shows that rank-conversion can allow the discovery of patterns containing genes with different ranges of expression values but still showing differential coexpression. Thus, we focus on the analysis of the patterns discovered on rank-converted data, on which more patterns are discovered. The patterns discovered on the data with expression value are presented on our website. Note that, rank-transformation is especially useful for DiffRange, since it is based on the biclustering algorithm [70] designed to find constant-row patterns [94]. Such rank-transformation may not be required in the \( \tilde{SDC} \)-Apriori framework for other biclustering algorithms that are able to find coherent additive, coherent multiplicative or coherent evolution biclusters [94].

Higgins et al.[46] collected a list of genes that are shown to be related to cancer. Out of the 8787 genes in the dataset, 1975 are on the cancer gene list. In the following experiments, we analyze the subspace patterns discovered on these 1975 genes (rank-converted data, and denoted as dataset \( D' \)), because cancer genes are more likely to have disregulated patterns and based on the existing knowledge of these cancer genes, the evaluation on the patterns discovered from these genes can better illustrate the biological relevance of subspace differential coexpression patterns. Note that although these 1975 genes are known to be related to cancer, the subspace differential coexpression patterns discovered on them can provide new insights about their relationship with cancer, e.g., by identifying the interactions among individual cancer genes.

### 3.4.2 Pattern Discovery

With parameters \( r = 0.2 \), and \( d = 0.2 \), DiffRange is used in the Apriori framework to discover subspace DC patterns on \( D' \). Most patterns are of size-2 (gene-pairs), but there are also size-3 and size-4 patterns (no larger size patterns are discovered for the selected parameters). To control the redundancy of genes among size-3 and size-4 patterns,\(^7\)

---

5 In this chapter, we union the two lists respectively downloaded in October 2008 and June 2009, with a total of 2622 genes.

6 In the rank-converted data, this means \( k \) genes have coherent expression if the rank difference of their expression is less than 20% of the 169 samples, i.e. 33.

7 Note that in the discussion of pattern (e) in Fig. 3.2, the redundancy is within a pattern rather than among the patterns like here.
we order them by decreasing $SDC$ value and sequentially select a subset of the patterns in which none of the pairs of patterns have greater than 25% overlap of genes. This compact set has 95 patterns (88 size-3 patterns and 7 size-4 patterns). Figure 3.3(a) shows the size and SDC value for each discovered DC pattern.

### 3.4.3 Are the discovered subspace differential coexpression patterns statistically significant?

Due to the issues of low sample size and high-dimensionality for data sets used for problems such as biomarker discovery, many patterns may be falsely associated with the class label by random chance, especially when a large number of combinations of genes are searched. This raises the multiple-hypothesis testing problem [55]. In this chapter, a permutation test is used to evaluate the statistical significance of the discovered subspace DC patterns. Specifically, the original class labels are randomly shuffled 1000 times. For each random labeling, the same $DiffRange$ parameters ($r = 0.2$ and $d = 0.2$) are used to discover a set of patterns. With the 1000 randomized labels, only size-2 patterns are discovered, with SDC values in the range of $[0, 0.4620]$. Figure 3.3(b) indicates this range by overlaying a double-arrow on top of the patterns discovered with the original class labels (as shown in Figure 3.3(a)). Considering 0.4620 as a statistical significance cutoff, Ellipse A indicates the 560 statistically significant gene-pairs whose SDC value was never exceeded by any random pattern. In addition, since there are no size-3 and size-4 patterns discovered with randomized labels in the permutation test, the 88 size-3 patterns (Ellipse B) and 7 size-4 pattern (Ellipse C) are also considered statistically significant. Note that, although a differential coexpression pattern can be highly coexpressed in either the cancer class or the normal class [75, 81], all the statistically significant patterns in the three ellipses are highly coexpressed in the normal class while less coexpressed in the cancer class.

### 3.4.4 How differentially coexpressed are the discovered subspace patterns when measured over full-space?

In this experiment, we measure the full-space differential coexpression for the statistically significant subspace patterns selected based on the above permutation test, i.e.,
the 560 gene-pair patterns and the 88 size-3 patterns and the 7 size-4 patterns. We will show that there are subspace patterns that have close-to-random differential co-expression when considered as full-space patterns. A variety of full-space differential coexpression measures are proposed in existing work as discussed in section 3.2. As used in several studies [74, 82, 1], we will use the correlation difference of a pattern between the two classes for illustration purpose. For a gene-pair pattern, correlation difference is just the difference of the two correlations respectively in the two classes. For a pattern of size greater than 2, we compute the difference between the average pair-wise correlation in the normal and cancer class to measure the correlation difference.

The three subfigures in Figure 3.4 plot the correlation difference and SDC for the statistically significant size-2, size-3, and size-4 patterns, respectively. The three dashed lines indicate the statistical significance cutoff of correlation difference for size-2, size-3 and size-4 patterns (0.9361, 0.5176 and 0.4953), respectively, which is also decided via permutation test. For the gene-pair patterns (Figure 3.4(a)), several observations can be made: (i) some patterns are considered statistically significant by both correlation difference and SDC (region A); (ii) some gene-pairs are considered significant only by SDC but not by correlation difference (region B). Among these patterns, several pairs have close-to-zero correlation difference (within the circle), which means they show very...
Figure 3.4: Illustration of the full-space differential coexpression (correlation difference) for the discovered statistically significant subspace differential coexpression patterns. The dashed lines in (a) – (c) indicate the statistical significance cutoffs for correlation difference for size-2, size-3 and size-4 patterns respectively. The solid line in (a) is the statistical significance cutoff for SDC (0.4620). There are no corresponding lines in (b) and (c) because all the patterns of size 3 and 4 are statistically significant in terms of SDC as discussed in section 3.4.3. Region A contains patterns that are considered significant by both correlation difference and SDC; Region B has patterns that are not significant as full-space patterns, several of which have close-to-zero correlation difference (within the circle); and Region C shows the significant full-space patterns that are not discovered by SDC.

Little differential coexpression when considered as full-space patterns; and (iii) there are also 801 gene-pairs that are only considered significant in terms of correlation difference but not by SDC (region C). This is as expected since many factors can affect the discovery of DC patterns, e.g. different coexpression measures, different mining algorithms, and the parameters used in the algorithms. Our highlight is the existence of subspace differential coexpression patterns that show close-to-random differential coexpression when considered as full-space patterns. Similar observation can also be made in Figures 3.4(b) and 3.4(c) which respectively plot the correlation difference for the size-3 and size-4 patterns.

In Figure 3.5, we illustrate a subspace DC pattern with very small correlation difference (0.19). This pattern is coexpressed in only 58% of the normal samples \(^8\), and the minimal pairwise correlation of the genes in this pattern over all the normal samples

\(^8\) This is with respect to the DiffRange parameters used to discover the patterns, \(r = 0.2\). Similar for Figures 3.6(a) and 3.6(b)
Figure 3.5: A statistically significant subspace differential coexpression pattern, for which the minimal pairwise correlation in the normal class is only 0.28. For better visualization, samples are sorted by increasing range of expression ranks separately in the two classes (similar for Figures 3.6(a), 3.6(b) and 3.7(a)).

is only 0.28. For this pattern, it is not reasonable to assume that the genes are coexpressed on all the normal samples. Furthermore, discovering the pattern as a subspace DC pattern can explicitly show the subgroup of samples on which the three gene show coexpression, i.e., the 39 normal samples and the 3 cancer samples in the cancer class. This allows further analysis of the difference between the 39 normal samples with the pattern and the 28 without it (e.g., different demographic characteristics), which may help personalized diagnosis and treatment.

The existence of subspace patterns that show small and insignificant differential coexpression when considered as full-space patterns demonstrates the potential usefulness of subspace differential coexpression analysis. Next, we will evaluate the biological relevance of the discovered subspace patterns.

### 3.4.5 Are the discovered subspace differential coexpression patterns biologically relevant?

Quantitatively, two enrichment experiments are used to evaluate the biological relevance of the discovered subspace differential coexpression patterns: (i) enrichment with ten known cancer-related signaling pathways ⁹ [46], (ii) enrichment with the 5452 gene

⁹ [http://cbio.mskcc.org/CancerGenes/Select.action](http://cbio.mskcc.org/CancerGenes/Select.action)
sets in the Molecular Signature database (MSigDB)\(^{10}\) [45]. Since patterns of size-2 are difficult to assess in terms of enrichment, we perform the two biological evaluations only for the 95 patterns of size 3 or 4. Briefly, (i) six patterns have an overlap of 2 or more genes with one of the ten known cancer-related pathways, (ii) in the MSigDB enrichment, 40 patterns have enrichment p-value less than 0.001, among which five have p-value less than 0.0001.

Note that, due to the limited knowledge about differentially coexpressed patterns, the current stage of differential coexpression pattern mining is still hypothesis generation rather than hypothesis verification, as discussed in Kostka and Spang[1]. Indeed, since all the 95 patterns are statistically significant in the permutation test, and all the genes contained in the 95 patterns are known cancer-related genes [46] (as described in section 3.4.1), they can be considered as hypotheses that may lead to new understanding of the interactions among them, and of the relationship between differential coexpression and cancer mechanism. Therefore, in addition to the above standard enrichment analyses, we will illustrate and discuss several interesting patterns that are enriched with known cancer pathways, or target sets of cancer-related microRNAs and transcription factors.

Figure 3.6 displays two patterns that are enriched with the $TNF\alpha/NF\kappa B$ signaling pathway and the $WNT$ signaling pathway respectively. Several observations can be made from these two figures. Firstly, they both show strong differential coexpression, i.e. they are both highly coexpressed in the normal class, and much less coexpressed in the cancer class. Secondly, both patterns are subspace differential coexpression patterns, i.e., they show coexpression in only 67% and 64% of the normal samples respectively. Similar to the pattern shown in Figure 3.5, these two patterns are coexpressed in only about two-third of the normal samples. Discovering them as subspace patterns also points out the subgroup of samples covered by them. This allows further study of the different causes of diseases and the different demographics among subgroups of samples. Finally, both the $TNF\alpha/NF\kappa B$ and $WNT$ signaling pathways have been shown to be related to lung cancer [106, 107]. Discovering the differential coexpression patterns enriched with these pathways may shed new light on the understanding of the two pathways and their relationships to cancer mechanism.

\(^{10}\) Specifically, MSigDB contains 386 positional gene sets, 1892 curated gene sets, 837 motif gene sets, 883 computational gene sets, and 1454 GO gene sets. http://www.broadinstitute.org/gsea/msigdb/
(a) The pattern that is enriched with the TNFα/NFκB signaling pathway (enrichment p-value 0.0011).

(b) The pattern that is enriched with the WNT signaling pathway (enrichment p-value 0.0042).

Figure 3.6: Two patterns that are respectively enriched with the TNFα/NFκB and the WNT signaling pathway.

Among the six patterns that are enriched with at least one cancer pathway, three are enriched with the TNFα/ NFκB pathway. In Figure 3.7(a), the union of the three patterns, containing ten genes, are plotted. All the ten genes are known cancer-related genes [46]. Out of the ten genes, six overlap with the TNFα/NFκB pathway (enrichment p-value $1.4024 \times 10^{-5}$). This may suggest that the other four genes may also participate in or interact with this cancer pathway [88]. Figure 3.7(b) shows the enriched IPA\textsuperscript{11} subnetwork, containing 8 of the ten genes (the 8 genes are shaded). Interestingly, IPA also shows that the connecting components (NFκB complex, ERK, Mapk and Cyclin E) are also known to be related to cancer.

\textsuperscript{11} Ingenuity Pathway Analysis: http://www.ingenuity.com/
Specifically among those patterns that are not considered significant by correlation difference (Region Bs in Figure 3.4), some are enriched with the target gene sets of cancer-related microRNAs and transcription factors. For example, the first two genes in the pattern \((PIK3C2B, TSC22D1, AKAP12)\) are among the set of target genes of miR-101 (p-value 0.001), a small non-coding RNA that regulates gene expression. miR-101 was shown to be down-regulated in cancer [108]. This agrees with the loss of coexpression of its target genes \((PIK3C2B, TSC22D1)\) in the cancer class. Furthermore, the miR-101 targets are enriched for several signaling pathways, and the third gene \(AKAP12\) is a known regulator of protein kinase A (PKA), a central signaling pathway involved in cell growth and proliferation. This may lead to the differential coexpression of \((PIK3C2B, TSC22D1)\) and \(AKAP12\) together as a DC pattern. In another example, the first two genes in the pattern \((ETV4, PTHLH, CBX5)\) are among the set of genes with promoter regions \([-2kb, 2kb]\) around the transcription start site containing the motif \(TTACGTAA\) which matches the binding site for the transcription factor ATF2 (p-value \(2.5119 \times 10^{-4}\)). Mutations of ATF2 was shown to be related to cancer [109], which agrees with the loss of coexpression of its target genes \((ETV4, PTHLH)\) in the cancer class. In addition, the ATF2 targets show enrichment for transcription regulation (repression), and \(CBX5\) is component of heterochromatin, an epigenetic factor in the regulation of gene expression. This may lead to the differential coexpression of \((ETV4, PTHLH)\) and \(CBX5\) together as a DC pattern.
3.5 Conclusions

In this chapter, we studied methods to identify disease-related change of coexpression subnetworks, i.e. differential coexpression analysis. Specifically, we extended differential coexpression analysis to subspace patterns and proposed an approach based upon association analysis framework[10] that allows exhaustive yet efficient discovery of subspace differential coexpression patterns. This approach can be used to adapt a family of biclustering algorithms to obtain their corresponding differential versions that can directly discover differential coexpression patterns. We illustrated the general approach on a recently-developed biclustering algorithm, and presented the results of experiments on lung cancer datasets using this algorithm. The results showed the existence of meaningful subspace differential coexpression patterns in real-life data. Permutation tests demonstrated the statistical significance for a large number of discovered patterns, many of which can not be discovered if they are measured over all the samples in each of the classes. Interestingly, some discovered patterns also have a significant overlap with known cancer pathways, and some are enriched with the target gene sets of a cancer-related microRNA and a cancer-related transcription factor. These results suggest that subspace DC patterns may aid in developing new understanding about the mechanisms underlying cancer and other diseases.

3.6 Limitations and Future Work

In this section, we discuss several limitations of the proposed approach, possible solutions and future work.

1. **Size of patterns**: Due to the fixed thresholds imposed on $\tilde{SDC}$ in the Apriori framework, there may be some larger patterns that do not satisfy the thresholds and are split into smaller ones. This limitation of association analysis is usually addressed by pattern summarization [67], in which smaller size patterns are merged into larger ones under some criteria. For example, the size-10 pattern in Figure 3.7 is obtained by merging three smaller patterns as described in Section 3.4.5. More sophisticated summarization approaches [67] can be exploited in future work.

2. **Enhancing scalability**: The scalability of the approach depends on the mining
algorithm, as well as the permutation test. Generally, the algorithm itself takes about ten minutes for 2000 genes, several hours for 4000 genes and more than a day for all the 8787 genes\(^\text{12}\), which is acceptable. However, the real challenge comes from the permutation test in which the mining algorithm is called 1000 times, the total time of which is unacceptable on all the 8787 genes. Thus, to have a comprehensive evaluation of the discovered patterns, we limited the pattern discovery and the follow-up statistical and biological analysis to the subset of genes that are known to be related to cancer. In future work, for the efficiency of the mining algorithm, more effective pruning schemes should be studied together with preprocessing procedures such as standard deviation based gene filtering\(^\text{13}\). For the scalability in the context of permutation test, efficiency could possibly be improved by reusing the calculation over the large number of permutations as studied by Zhang et al. [110]

3. **Modifying other biclustering algorithms**: In this chapter, \textit{DiffRange} is presented as an illustration of the general approach, \textit{\hspace{1pt}SDC-Apriori}, for modifying a biclustering algorithm to its differential version. As discussed in Section 3.3.1, \textit{\hspace{1pt}SDC-Apriori} can also be applied to modify other biclustering algorithms [95, 96, 97, 98] with the antimonotonicity property, and their corresponding differential versions are expected to complement \textit{DiffRange} for discovering differential coexpression patterns.

4. **Differential biclustering**: Differential coexpression patterns can essentially be considered as biclusters that exist mostly in one class but not in the other. Indeed, such type of biclusters have already been observed in several studies [111, 112, 95], where a set of biclusters are discovered in the first step and then the ones that are unique to a single class are selected in the second step. Such a two-step approach can also be used to discover differential coexpression patterns. However, the general approach proposed in this chapter, \textit{\hspace{1pt}SDC-Apriori}, can be considered as an initial effort towards a more general differential biclustering problem, where

\(^{12}\) The experiments presented here were run on a Linux machine with Intel(R) Xeon(R) CPU (E5310 @ 1.60GHz) and 16GB memory

\(^{13}\) A gene with small variation across samples is less likely to constitute a differential coexpression pattern.
more efficient discovery of differential biclusters are possible by making use of class labels within the biclustering process. Similar problems can also be formulated as differential/discriminative co-clustering and differential/discriminative subspace clustering in the data mining community.

5. **Pattern-based classification**: Since a subspace differential coexpression pattern explicitly captures the subgroups of samples it covers, it will also be interesting to investigate the predictive power of subspace differential coexpression patterns in a pattern-based classification framework [27, 12], where the combination of traditional differentially expressed genes and subspace differential coexpression patterns may provide more accurate disease diagnosis.
Chapter 4

A Generic Strategy for Enhancing the Statistical Power of Discriminative Pattern mining

4.1 Introduction

This chapter pursues the enhancement of the statistical power of discriminative pattern mining techniques. We propose a generic strategy that handles the issue of multiple hypothesis testing by identifying and pruning patterns that are not interesting or useful in the context of a certain domain. Specifically, in combinatorial disease biomarker discovery, the type of discriminative patterns with most interests for biologists are those combinations that provide additional discriminative power than any of its sub patterns. Only focusing on such patterns in the exponential search space could simultaneously reduce computational complexity and reduce the overall number of hypotheses tested (thus reduce the risk of finding spurious patterns). Three case-control SNP datasets (from three studies on multiple myeloma, lung cancer and kidney transplant rejection) are used to demonstrate the effectiveness of this strategy for enhancing the overall statistical power of discriminative pattern mining.
4.2 Background

Genotype-phenotype association studies, from both targeted and genome-wide data, have contributed to our ability to identify genetic variants that are associated with disease. Although an increasing number of studies have found single-nucleotide polymorphisms (SNPs) that have statistically significant association with diseases, most of them either have small effects on disease risk [113, 114, 115] or often explain only a small part of the population [116, 117, 118, 18]. Thus, there has been increased interest in discovering combinations of SNPs that are strongly associated with a phenotype even if each SNP has little or even no individual effect [119, 120, 121, 17, 122]. Our goal is to discover and study such combinations of SNPs to complement existing approaches for univariate analysis or pathway/network enrichment-based approaches that are built upon univariate statistics [30, 123, 124, 125]. In particular, as pursued by [126, 127, 128, 129, 130, 131, 132], we focus on discovering SNP combinations, especially high-order ones beyond size 2, that are strongly associated with a phenotype and yield information on interpretable statistical and functional interactions.

There are two challenges in finding SNP combinations that are highly associated with a phenotype from a large number of SNPs. The first arises from the combinatorial nature of the problem, i.e. there are exponentially increasing number of combinations as the order goes higher. This is even more problematic if a large number of permutation tests are used to correct for multiple hypothesis tests [45, 30, 133]. Given a GWAS dataset with hundreds of thousands of SNPs, even the examination of pair-wise combinations of SNPs is computationally challenging [132], and requires efficient enumeration algorithms [134, 132, 135, 136] or specialized hardwares [137, 138]. Finding higher order SNP combinations [139, 126] is far more computationally expensive and is out of reach for GWAS datasets. Hence, existing methods mostly explore higher order SNP combinations with datasets that only have tens or few hundreds of SNPs. These methods adopt either brute-force or heuristic-based greedy search. Brute-force approaches such as multifactor dimensionality reduction (MDR [126]), or the combinatorial partitioning method (CPM [128]) can guarantee the completeness of the search, which is important in detecting SNP combinations with weak marginal effects [131]. However, these brute-force approaches can handle only a relatively small number of SNPs (tens or
hundreds) [126, 140, 127]. The scalability of recent approaches [141] has been improved to allow searching for size 3 combinations from about 600 SNPs within two hours. However, it is still not capable of efficiently handling focus studies that have thousands of SNPs [142, 143], especially for higher order combinations. Greedy search strategies [43, 144, 145, 146, 147, 148, 149, 150, 151], although more computationally efficient than brute-force approaches, risk missing significant SNP combinations [132, 17, 122], and rarely discover high-order combinations beyond size 3 [126], and only from datasets containing tens or hundreds of SNPs covering a even smaller number of genes.

The second challenge is that existing approaches for high-order SNP combination searches lack statistical power. Specifically, due to the use of statistics with high degree of freedom [152, 139] and the huge number of hypothesis tested with often limited sample sizes, many high-order combinations of SNPs can be strongly associated with a disease phenotype by random chance, resulting in a high false discovery rate [133]. Some existing approaches [145, 153, 121] use biological pathways or molecular interaction networks as constraints to reduce the number of hypotheses to test and make the interpretation easier. Essentially, a set of SNPs are considered for an association test only if the SNPs are located around the genes that are on a common pathway or interact with each other. A common limitation of such constraint-based approaches is that, they may miss novel SNP combinations that are not on known pathways or interaction subnetworks due to the incompleteness of biological knowledge. Thus, it calls for a quantitative evaluation on trade off between the reduction of search space and the risk of missing informative SNP combinations, and also calls for alternative constraints that are not limited by existing biological knowledge.

In this chapter, we aim to address both the above challenges.

To improve computational efficiency, we leverage the discriminative pattern mining framework (DPM, originally proposed [10, 99] in the data mining community for mining market basket data) to efficiently search for high-order SNP combinations from SNP datasets in focused studies with thousands of SNPs. The computational efficiency and scalability of DPM is enhanced by the systematic pruning of the combinatorial search space with anti-monotonic objective functions. A unique advantage of anti-monotonicity-based search over brute-force search is that it can avoid exploring the whole search space (all combinations of SNP genotypes) by pruning a large number
of candidates that cannot lead to a sufficiently strong association with a phenotype [154, 99]. We demonstrate that DPM has substantially improved efficiency and scalability on a synthetic and three real datasets with several thousands of SNPs. We observe that most high-order combinations are trivial extensions of their subsets which are not interesting but consume most of the total computation time, however, there are indeed high-order combinations that have discriminative power significantly beyond singleton SNP or low-order SNP combinations.

To improve the statistical power, we study the effect of a generic approach for enhancing the statistical power of discriminative pattern mining, i.e. requiring an increase in discriminative power for a combination over its subsets. We demonstrate that this constraint can reduce the number of hypothesis tests dramatically and thus enable the discovery of significant combinations that would have been missed otherwise.

The improved computational efficiency and statistical power further enables the discovery of significant high-order SNP combinations from the three real datasets and then allows the exploration of functional interactions in high-order SNP combinations. Specifically, we study the functional interactions among the genes covered in high-order SNP combinations with an integrated human functional gene network. We find a positive connection between the increase of discriminative power of a SNP combination over its subsets and the functional coherence among the genes covered in the combination. Such an observation is beyond the disease-specific functional interactions studied by existing work that are based on datasets covering a small number of genes [126] and is supported by the multiple real datasets used in the chapter. In addition to this disease-independent biological insight, we also interpret several high-order combinations discovered from the lung cancer [MIM: 211980] dataset and the dataset for studying rejection after kidney transplant, which provide novel insights beyond univariate or low-order SNP-combination analysis. More generally, we find that many significant associations are combinations of common variations that occur in small fractions of population. This suggests an alternative direction for the exploration of the genetics of rare diseases, where the current focus is mainly on analyzing individually rare variations.
4.3 Analysis and Results

4.3.1 Three real case-control SNP datasets and a synthetic dataset

We use three SNP datasets designed for studying different types of disease phenotypes: (i) short (less than one year) vs. long (greater than three years) survival of multiple myeloma [MIM: 254500] patients [142] (denoted as Survival), (ii) acute rejection [MIM: N/A] (within in six months) vs. non-rejection (within eight years) after kidney transplant [131] (denoted as Kidney), (iii) lung cancer [MIM: 211980] vs. non-lung cancer (both heavy smokers) [143] (denoted as Lungcancer). The three datasets were all collected with a chip [142] targeting 3444 SNPs in 983 genes, representing cellular functions and pathways that may influence disease severity at diagnosis, toxicity, progression or other treatment outcomes. Previous analyses on these three datasets did not reveal statistically significant single SNPs after correcting for multiple hypothesis testing, and this study aims to explore if there are significant (after correcting for multiple hypothesis testing) associations between combinations of SNPs and disease phenotypes, especially high-order combinations (with size greater than 2) that have stronger association beyond single SNPs or low-order combinations.

Preprocessing and quality control steps are described in the method section. More information on these datasets can be found in the original chapters. All the datasets are available from the Eastern Cooperative Oncology Group (ECOG) through requests to the operations office (http://www.ecog.org/, accessed 2012 Feb 20). In addition to the three real datasets, We also use a synthetic dataset with 70 cases and 70 controls, 2172 SNPs without differentiation between the cases and controls, and four synthetic high-order SNP combinations of size 3, 4, 5 and 6 respectively, that are associated with case-control groupings. (See the methods section for simulation details.)

Note that, the above four datasets have much larger number of SNPs (ranging from 2172 to 3428) than the datasets used in previous studies on high-order SNP interactions (tens or hundreds of SNPs). With these four datasets, we will show that the proposed framework is substantially more efficient and scalable than existing approaches. Although the proposed approach could not directly handle datasets with more than 10,000 SNPs due to the intrinsic computational complexity of high-order SNP combination search, it is worth noting that tag SNP selection [155] techniques can be used
to first obtain a set of less redundant SNPs before the use of the proposed approach. In this way, genome-wide studies with hundreds of thousands of SNPs could also be analyzed.

### 4.3.2 The Binary Encoding of a SNP and a Combination of SNPs

We use a binary coding scheme of SNP genotypes, where we create three binary columns for each SNP (Figure 4.1). For a single SNP \( X \) with three genotypes (homozygous minor (mm), heterozygous (Mm) and homozygous major (MM)), we create three binary variables as \( X = mm \), \( X = mM \) and \( X = MM \), each of which is represented as a binary variable indicating if a person’s genotype for SNP X is \( mm \), \( mM \) or \( MM \) respectively. Figure 4.1 illustrates the transformation from categorical encoding to binary encoding. Note that, this is a lossless transformation because it can be mapped back to the original SNP genotypes without ambiguity. As will be shown later, the use of this binary coding is to enable the efficient traversal of the combinatorial search space in the discriminative pattern mining (DPM) framework used in the chapter. Although the number of columns increases to three times of the original number of SNPs, we show the DPM framework has substantially better efficiency and scalability than existing approaches that directly search from the categorical SNP variables. It is worth noting that, binary encoding was also leveraged in [141], where the authors commented that, while binary coding may have somewhat weaker power, it does allows the use of efficient enumeration algorithms.
and the discovery of biologically interesting SNP combinations.

Based on the binary coding for each SNP genotype, a combination of SNPs is essentially a combination of SNP genotypes. For example, for three SNPs $X$, $Y$ and $Z$, a combination might be \{X = mm, Y = mM, Z = MM\}. Such a combination is also called a *pattern* in this chapter, where we use the terms "pattern", "combination" and "SNP combination" interchangeably. Following the traditional setup in discriminative pattern analysis, a pattern is said to be *present* in a subject only if the subject’s genotypes match all the SNP genotypes in the pattern, and *absent* otherwise. Thus, a combination of SNP genotypes (multiple SNPs, each contributing one of its genotype) is also encoded as a binary variable (*present* or *absent*). Again, we use this setup to allow DPM to efficiently perform the search of combinatorial pattern space. The frequency of a pattern (the percentage of subjects in which a pattern is present, also called *support*) has a mathematical property named anti-monotonicity, which can be leveraged by DPM to prune most of the combinatorial search space and only investigate those patterns that are more likely to have strong association with a disease phenotype [10, 99] (see methods section).

With this binary encoding of a SNP combination, a $\chi^2$ test of the association between any combination and a binary phenotype has a fixed degree of freedom of 1 [141] and is independent of the size of the combination. Here, the goal is to test the association between the present and absent of the SNP combination, under the binary encoding, and a binary phenotype. Note that, other statistical measures can also be used for similar purpose. This also implies that the proposed framework can handle datasets with imbalanced number of cases and controls. The degree of freedom being 1 is an important advantage for high-order SNP combination analysis because most real datasets have a limited number of samples that are insufficient for estimating the association between a combination of larger size and a disease phenotype if the statistical measure in use has a degree of freedom increasing with the size of a combination. The fixed degree of freedom also allow the direct comparison of the statistics (e.g. $\chi^2$ statistic or others) of SNP combinations of different sizes, which is important for quantifying the gain of discriminative power of a SNP combination with respect to its subsets. For example, the size-3 combination \{X = mm, Y = mM, Z = MM\} has three size-2 subsets: \{Y = mM, Z = MM\}, \{X = mm, Z = MM\} and \{X = mm, Y = mM\}. 
Figure 4.2: Visualization of the two synthetic SNP-genotype combinations and their high-order association with the two classes. The two subfigures in the left column are the visualization of the genotypes of 4 SNPs separated by vertical green lines, over the 70 cases and 70 controls separated by a horizontal yellow line. The black color indicates present and the white indicates absent, in the binary format described in the method section. The $\chi^2$ statistic, odds ratios and the $-\log_{10}$ fisher exact test p value of the two combinations are (28.7, 21.0, 7.84) and (25.8, 18.6, 7.1), respectively. Each subfigure in the right column contains 4 pairs of bars. For each pair, the unfilled bar and the filled bar indicate the minimal and the maximal $\chi^2$ statistics for the size-$i$ ($i \in [1, 4]$) subsets of the combination. The right most pair, both bars are equal since they both denote the $\chi^2$ statistic of the SNP combination itself.)

4.3.3 Illustrative Examples of High-order Discriminative SNP Combinations

After describing the above binary encoding of a SNP combination, we first illustrate two examples of high-order SNP combination shown in Figure 4.2 ($P_A$ and $P_B$, generated with the method developed in [156]) before presenting the efficient search algorithm. $P_A$ is a pattern containing four SNPs (separated by vertical green lines) over 70 cases and 70 controls, which are separated by a horizontal yellow line (cases top, controls bottom). The black color indicates presence (1’s) and the white indicates absence (0’s) of one of
the three genotypes of a SNP. The χ² statistic, odds and the −log₁₀ fisher exact test p-value of the synthetic combination (as a binary encoded single variable as described above) are (28.7, 21.0, 7.84). The subfigure in the right column contains 4 pairs of bars. For each pair, the left bar (unfilled) and the right bar (filled) indicate the minimal and the maximal χ² statistics for the size-i (i ∈ [1, 4]) subsets of the combination. For the right most pair, both bars are equal since they both denote the χ² statistic of the SNP combination itself. As shown, the χ² statistic of P_A is higher than all of its subsets, which makes P_A interesting because it provides predictive power beyond that of its subsets. Thus, it is important to discover this high-order pattern as a highly confident predictive rule with an odds ratio of 21, rather than discover its subsets.

Similar to P_A, pattern P_B in Figure 4.2 also has high discriminative power in terms of χ² statistic, odds ratios and the −log₁₀. However, in contrast to P_A, pattern P_B is actually less discriminative than one of its size-2 subset (the first two SNP columns), as reflected by the drop in the χ² statistic in the right subfigure. Later in this section, we will differentiate these two types of SNP combinations and show that SNP combinations like P_A provide more information for the functional interactions among the genes in a SNP combination, while the high discriminative power of patterns like P_B are trivial consequences of their highly differentiating subsets. Figure 4.9 shows four high-order SNP combinations of size-3 to size-6 (generated with [156]) that we embedded in the synthetic dataset described earlier, all having higher discriminative power than their subsets. Indeed, such interesting high-order SNP combinations also exist in real datasets for studying complex diseases such as cancer, as will be shown in the result section.

4.3.4 Discovering High-order SNP Combinations that have Strong Association with a Phenotype

With the two discriminative SNP combinations shown in Figure 4.2 and the additional examples in Figure 4.9, we now describe how to leverage the discriminative pattern mining (DPM) framework to efficiently search for high-order SNP combinations that have strong association with a disease phenotype. The DPM mining framework was originally proposed in the data mining community to efficiently enumerate combinations of variables and identify those that are highly predictive [20, 21]. DPM builds upon a general search strategy called Apriori [10], which leverages the anti-monotonicity of
a special type of objective functions for efficient enumeration of high-order variable combinations (see methods for details). Conceptually, with an objective function that is anti-monotonic, a SNP combination satisfies a threshold on the objective function only if all its subsets satisfies the threshold. In another word, if a combination does not pass a threshold on the objective function, all of its supersets can be pruned in the search space and it is guaranteed that no larger combination that satisfies the threshold would be missed. This is the key difference between Apriori-based combinatorial search and brute-force combinatorial search.

In this study, we leverage a recently developed anti-monotonic objective function \textit{SupMaxPair} \cite{99} and use it in the Apriori framework to efficiently search for SNP combinations that are discriminative between cases and controls. \textit{SupMaxPair} captures the association between a SNP combination and a binary disease phenotype (see the methods section), i.e. the higher \textit{SupMaxPair}, the stronger the SNP combination is associated with the phenotype. The Apriori framework using \textit{SupMaxPair} as the objective function is called SMP \cite{99} and has the advantage of handling dense and high dimensional data, which addresses the key challenge in discovering high-order combinations from SNP datasets, i.e. a fixed high density of 33\% as a result of the binary encoding of each SNP (Each SNP is represented with three binary columns and the genotype of a sample for each SNP is represented by a 1 in one of the three columns (assuming there is no missing value). Thus, one third of the matrix values are 1’s (a density of 33\%)) and a large number of SNPs (high dimensionality). This advantage owes to the effective use of phenotype information in the searching process \cite{99} and is the essence of SMP’s better efficiency and scalability over other DPM algorithms.

It is worth noting that Ma et al. \cite{141} is the first that leverages an Apriori-based algorithm \cite{60} (FPC) for the efficient enumeration of SNP combinations. However, FPC does not make use of phenotype information to optimize the search process and thus is much less efficient and less scalable than SMP, as has been shown in \cite{99} on differential gene expression analysis and will also be demonstrated on SNP datasets in the result section of this study. SMP is part of the framework we implement for this study and is available on the chapter website (http://vk.cs.umn.edu/IHSC/, accessed 2012 Feb 20).
4.3.5 The DPM framework has substantially better efficiency and scalability

We compare the DPM framework with two representative existing tools for high-order SNP combination discovery: MDR [126] (http://www.epistasis.org/software.html) and the framework presented in [141] (denoted as FPC in this chapter). For MDR, we used the Java version (http://sourceforge.net/projects/mdr/) and used the standard coding, in which each SNP is represented by a categorical value with three possible values (genotypes). For DPM and FPC, we use the binary coding. FPC requires an input for the parameter $\text{minsup}$ (the minimum frequency of a pattern in the set of cases and controls combined). For comparison purpose, we set a five-hour maximal runtime allowance (Though arbitrary, some threshold needs to be selected for comparison purpose) for all the three techniques. Experiments presented here were run on a Linux machine with 10 Intel(R) Xeon(R) CPUs (2.00GHz) and 100GB memory.

In the synthetic dataset (described in the method section), there are 2172 SNPs. The three frameworks need to search through size-2, size-3, size-4, size-5 and size-6 combinations in order to discover the four embedded patterns of size-3 to 6. After five hours, MDR was still enumerating size-3 SNP combinations, and thus failed to identify the embedded size-4, size-5 and size-6 patterns. FPC could reach size-6 within five hours, but only with a $\text{minsup}$ threshold of 0.9 (With a $\text{minsup}$ threshold of 0.8, FPC could not finish even in 24 hours.), which is so high that none of the four synthetic patterns were discovered (the frequency of the four embedded patterns are all below 0.25.). In contrast, the run time of SMP on the synthetic dataset is around 4 minutes with a $\text{SupMaxPair}$ threshold of 0.15. The threshold of 0.15 was chosen such that all the four embedded synthetic SNP combinations can be discovered. At lower threshold, additional discriminative SNP combinations can be discovered (if they exist), but it will take more computational time. In practice, one should use a threshold as low as possible while the computational time is still acceptable (usually decided after some tests). In addition, given a fixed $\text{SupMaxPair}$ threshold and a fixed number of SNPs, the patterns discovered from a dataset with larger sample size are expected to be more statistically significant in term of false discovery rate. Therefore, given a certain statistical significance cutoff, a lower $\text{SupMaxPair}$ threshold should be used for datasets with larger sample sizes while the computational time is still acceptable.
Figure 4.3: The scalability of SMP with respect to sample size (cases and controls combined). The computational time of SMP increases linearly with the sample size (Note that the x-axis is not linearly spaced)

The discovered SNP combinations are of size 2 to 10, including all the four embedded patterns. The substantially better efficiency of SMP is also observed on the three real datasets, which have 2755 – 3428 SNPs. The substantially better efficiency and scalability of SMP over FPC and MDR is due to the effective use of phenotype information in SMP for pruning combination candidates that are less likely to form a larger discriminative pattern as discussed in the method section (refer to [99] for further details). Indeed, the efficiency of the proposed framework (search as high as size-10 combinations from thousands of SNPs within one hour) is superior to not just MDR and FPC, but also to several other existing approaches which can discover up to size-3 SNP combinations from datasets with hundreds of SNPs [157, 140, 141, 147, 127]. Furthermore, we designed an experiment to test the scalability of SMP with respect to the sample size. We vary the sample size (cases and controls combined) from 140 to 5600 in seven steps (140, 280, 420, 560, 1400, 2800 and 5600) as shown in Figure 4.3 (see method section for the details of data simulation). The first four steps representing one, two, three and four times of the samples in the first synthetic dataset (used in the comparison with MDR and FPC), respectively. The last three steps correspond to a much larger samples sizes in several thousands that represent the number of samples in most GWAS studies. The running time shown on the y-axis of Figure 4.3 shows that
the computational time of SMP increases approximately in a linear manner with respect to the sample size (recall that the x-axis is not linearly spaced). This agrees with the theoretical time complexity of Apriori-based searching algorithms [10] and indicates that SMP is able to handle datasets with much larger number of samples than the three real datasets used in this chapter.

Note that, the synthetic datasets used above (to demonstrate the better efficiency and scalability of DPM over MDR and FPC) are representatives of the three real datasets used in the chapter. For datasets with smaller number of SNPs (e.g. tens or hundreds of SNPs), MDR and FPC (as well as other similar approaches) have been compared with other approaches [130, 141] and demonstrated to be scalable (mostly up to size 3 combinations). In this study, we have the specific focus on datasets with thousands of SNPs such as the three real datasets or datasets of tag SNPs selected from genome-wide studies, and we are particularly interested in high-order interaction (its mathematical and statistical properties as well as functional insights). Therefore, we will only use DPM in the rest of the analyses.

4.3.6 Identifying High-order SNP Combinations with Stronger Association than their subsets

Among the set of discovered SNP combinations discovered by DPM, some have better discriminative power than their corresponding subsets (like $P_A$ in Figure 4.2) while some have similar or lower discriminative power (like $P_B$ in Figure 4.2). A simple way to quantify the increase of discriminative power of a SNP combination over its subsets is to take a difference between the discriminative power of a SNP combination itself and the best discriminative power among all of its subsets. With the $\chi^2$ statistic as the measure for discriminative power, this difference (denoted as $\chi^2_{jump}$) for a pattern $\alpha$ can be formally written as below. Note that, the $\chi^2$ statistics of patterns of different sizes all have the degree of freedom of 1 based on the binary encoding of a SNP combination presented earlier in this section. Also note that, among the thresholds we used for $SupMaxPair$ in the chapter, the lowest is 0.15. This implies that the minimum frequency of any discovered SNP combination is 15% of the number of cases or controls (refer to the definition of $SupMaxPair$ in the method section). Thus, the estimation of $\chi^2$ statistic for any SNP combination would be based on a frequency of at least 15% of the number
of cases or controls, even for high-order combinations.

\[ \chi_{\text{jump}}^2(\alpha) = \chi^2(\alpha) - \max_{\alpha' \subset \alpha} (\chi^2(\alpha')). \]  

(4.1)

With the above definition, the \( \chi_{\text{jump}}^2 \) of the two patterns shown in Figure 4.2 are 14.4 and -6.1 and the four patterns in Figure 4.9 all have positive \( \chi_{\text{jump}}^2 \) values (47.7, 14.4, 6.2 and 4.0 respectively). Indeed, \( \chi_{\text{jump}}^2 \) is not a new concept and similar measures based on other statistics for discriminative power (instead of \( \chi^2 \) statistic) have been studied in data mining literature [158]. More generally, existing measures of epistasis and genetic interaction [146, 152] which capture the difference between the joint statistic between a SNP combination and the linear (or independent) addition of its subsets, could be used for the same purpose as well. However, they are not suitable for high-order combinations analysis due to their increasing degrees of freedom and computational expense as combination size increases, which thus requires an increasing number of samples for accurate estimation. In contrast, \( \chi_{\text{jump}}^2 \) or similar measures based on other statistics have the advantage of a fixed degree of freedom (1) and thus are more practical for measuring the association between high-order combinations and a phenotype. Furthermore, the requirement of epistasis measures is more restrictive than measures like \( \chi_{\text{jump}}^2 \) because the former only captures non-additive effect while the latter targets the general combined effect including both linear and non-linear combinations. Indeed, as will be shown in the result section, both linear and non-linear high-order combinations exist in real datasets, and both can be highly discriminative with respect to a disease phenotype and thus are of great interest.

Intuitively, it would be ideal if an algorithm like SMP can directly differentiate combinations with positive and negative \( \chi_{\text{jump}}^2 \) values and then prune the ones with negative values as early as possible in the searching process. However, this is a non-trivial task because the \( \chi_{\text{jump}}^2 \) does not have the antimonotonicity property (crucial for the efficient enumeration of high-order combinations using the Apriori strategy [10]) and thus some combinations with large positive \( \chi^2 \) would be missed if they have subsets with negative \( \chi_{\text{jump}}^2 \). Therefore, in this study, we use SMP to first discover a set of discriminative combinations and then apply a \( \chi_{\text{jump}}^2 \) based filtering as a separate step.
4.3.7  Many high-order patterns are trivial extensions of their smaller subsets

We ran DPM on the three real datasets (with \(\text{SupMaxPair} = 0.2\), the lowest threshold that DPM can finish within 0.5 hour) and produced a set of SNP combinations from each dataset. With the three sets of discovered patterns, we first study a key mathematical property of high-order patterns, that is, if these combinations provide additional insights beyond their subsets. Specifically, for each combination, we calculate its \(\chi^2\) statistic and \(\chi^2_{\text{jump}}\), and summarize the results in Figure 4.4, with the three subfigures corresponding to the three datasets. Each subfigure shows the \(\chi^2\) statistic of each pattern and the maximal \(\chi^2\) statistic among all of its subsets, for all the discovered patterns. The \(\chi^2\) statistic thresholds of +5 and −5 are indicated by a red line and a black line respectively, in each subfigure. Clearly, many large size patterns have negative \(\chi^2\) jump, which indicates that many high-order patterns are trivial extensions of their smaller subsets (such as pattern \(P_B\) in Figure 4.2). They are not interesting or at least not informative for either enhancing the predictive power of a pattern or exploring functional interactions among the patterns in a SNP. Note that, +5 and −5 are used as two threshold of \(\chi^2_{\text{jump}}\) in Figure 4.4 just for visualization purpose, while different thresholds are studied in the separate experiments.
Figure 4.5: Comparing the $\chi^2$ statistic of each pattern with the maximal $\chi^2$ statistic among all of its subsets. The three subfigures correspond to the three datasets. Each subfigure shows the $\chi^2$ statistic of each pattern and the maximal $\chi^2$ statistic among all of its subsets for all the discovered patterns. The color of a circle indicates the size of the pattern. The red line and the black line in each subfigure show $y = x + 5$ and $y = x - 5$ respectively.

4.3.8 Some high-order patterns are highly discriminative beyond univariate and low-order SNP-combinations

We also note that there are indeed several high-order combinations that provide higher discriminative power than any of their corresponding subsets. Specifically, in the datasets, Kidney and Lungcancer, there are tens of size-4 and size-5 patterns above the line of $y = x + 5$. These patterns may indicate high-order functional gene interactions whose joint genetic variations result in a stronger association with the disease phenotypes than singletons and lower-order combinations. Again, $+5$ and $-5$ are used as two threshold of $\chi^2_{jump}$ in Figure 4.4 just for visualization purpose, while different thresholds are studied in separate experiments. The observation that only a small fraction of high-order patterns have large $\chi^2_{jump}$ values motivates the design of targeted search algorithms that specifically look for patterns with large $\chi^2_{jump}$ in addition to high $\chi^2$. However, this is a non-trivial task as discussed in the method section.

Many patterns with high $\chi^2_{jump}$ (e.g. above the line of $y = x + 5$) in the three datasets have $\chi^2 - statistics$ greater than 20, which corresponds to a low p-value of $10^{-7}$. However, because a huge number of hypotheses were tested in the SMP search, we need to correct for multiple hypothesis testing. We use a permutation-test based approach (see methods section) to estimate unbiased and reliable false discovery rates (FDRs) for the patterns discovered and shown in Figure 4.4 (methods section).
Figure 4.5 shows the $\chi^2$ statistics and FDRs for the patterns with $\chi^2_{\text{jump}}$ above 5 (different parameters for $\chi^2_{\text{jump}}$ are studied in separate experiments), with a layout similar to Figure 4.4. The circles with similar color are clustered together, which results from the size-specific permutation tests which estimate the FDR of a size-$k$ pattern from the null distribution built with only the random patterns of size $k$ (see method). We observe that there are several significant patterns with FDR (w.r.t. $\chi^2$) below 0.25 discovered from the datasets Kidney (up to size-4) and Lungcancer (up to size-5). Note that in Figure 4.5, we only consider the patterns with high $\chi^2_{\text{jump}}$ (above the line of $y = x + 5$). We will present a separate experiment that illustrates the benefit to statistical power of using $\chi^2_{\text{jump}}$ based filtering where we try different thresholds of $\chi^2_{\text{jump}}$.

To better understand the effect of sample size on the FDRs of the patterns discovered from the real datasets. We designed an experiment with the same synthetic datasets used in the scalability test (Figure 4.3). Specifically, we examine the effect of sample size on the FDRs of the four embedded synthetic SNP combinations of sizes 3, 4, 5 and 6, respectively. The key observation is that, although the FDRs of embedded patterns are expected to be more significant when the sample size increases, all the four synthetic patterns have perfect FDR ($< 0.002$, i.e. no better patterns were found in any of the 500 permutations), when the sample size is above 200. This indicates that the sample sizes in the two real datasets (Lungcancer and Kidney) are expected to be good enough for high-order SNP combination search. However, when the sample size is below 200, two of the four embedded real patterns (the size-4 one and size-5 one) can not be discovered with significant FDRs. This is also consistent with our observation on the other real datasets (Survival), on which no significant SNP combinations were discovered. Therefore, this new experiments helped the understanding of the effect of sample sizes on FDR and also support the statistical reliability of the patterns discovered from the two real datasets.

### 4.3.9 A Genetic procedure that generally enhance statistical power of discriminative pattern mining

Here, we present the results studying a generic strategy for enhancing the statistical power of discriminative pattern mining, i.e. by enforcing a proper threshold of $\chi^2_{\text{jump}}$. The statistical power is indirectly measured by the number of combinations and unique
Figure 4.6: Comparison between the FDRs without $\chi^2$ – \textit{jump} based filtering and the FDRs with $\chi^2$ – \textit{jump} filtering for the Lungcancer and Kidney datasets respectively. In both subfigures, each circle represent a SNP combination. There are several circles sitting below the line $y = x$, indicating that they have lower (more significant) FDR when a $\chi^2$ – \textit{jump} filtering is applied compared to the case where no $\chi^2$ – \textit{jump} is used.)

SNPs discovered with respect to a specific false discovery rate of 0.25. (Although somewhat arbitrary, a cutoff is needed. We choose a relatively high FDR threshold as in [45] because, for high-order SNP combination discovery which is still at its early stage, the research focus is more about hypothesis generation instead of hypothesis verification).

**Use of $\chi^2_{jump}$ based filtering generally improves statistical power**

In Figure 4.5, the FDRs are estimated only with those patterns having sufficiently high $\chi^2_{jump}$. Here, we study whether using a $\chi^2_{jump}$ based pattern filtering improves the statistical power of the framework. Figure 4.6 (each circle represent a SNP combination) compares the FDRs without $\chi^2_{jump}$ based filtering (x-axis) and the FDRs with $\chi^2_{jump}$ filtering (y-axis) for the Lungcancer (left subfigure) and Kidney (right subfigure) datasets. We tried three different thresholds for $\chi^2_{jump}$ (0, 3 and 5) and found that the results are similar, which suggest the essential effect of the filtering is to eliminate those patterns with low negative $\chi^2_{jump}$ values. The figures shown here are based a threshold of 5 for $\chi^2_{jump}$. We use these two datasets for this comparison because there are more high-order combinations with high $\chi^2_{jump}$ discovered from them (up to size-4 and size-5).
and because none of the pattern discovered from the other dataset (Survival) have FDR (w.r.t. $\chi^2$) below 0.25. In both subfigures, there are several circles sitting below the line $y = x$, indicating that these patterns have lower (more significant) FDR (w.r.t. $\chi^2$) when a $\chi^2_{jump}$ filtering was applied compared to the case where $\chi^2_{jump}$ was not used. Specifically, there are seven combinations in the right subfigure (the red ones indicated by the arrow) which have an insignificant FDR (0.5) when no $\chi^2_{jump}$-based filtering was applied, but low FDRs (around 0.2) when a $\chi^2_{jump} = 5$ filtering was used.

This comparison demonstrates that $\chi^2_{jump}$ can enhance the statistical power of discriminative SNP-combination discovery and potentially discover SNP combinations that would have been missed. This can be explained as follows: for a real pattern $P$ of size-k and a high $\chi^2_{jump}$, the use of $\chi^2_{jump}$ filters out random patterns in the permutation tests that have high discriminative power but are trivial extensions of its subsets, which would otherwise penalize the statistical significance of $P$. Essentially, the use of $\chi^2_{jump}$ based filtering provides a better estimation of the statistical significance of a pattern with high $\chi^2_{jump}$ by estimating a more reasonable null distribution.

As discussed earlier, $\chi^2_{jump}$ is just one of many possible measures that quantitatively describes the increment of discriminative power of a pattern with respect to its subsets. Specifically, the $\chi^2$-statistic can be replaced by other measures of discriminative power, or the difference can also be replaced by measures for statistical epistasis [146, 152]. The observations from the above comparison, where $\chi^2_{jump}$ is used as a representative, supports the use of these measures to improve the statistical power of discriminative SNP combination discovery.

### 4.3.10 Exploring functional interactions in high-order combinations

Existing work that studies functional interactions in SNP combinations mostly focuses pairs of loci [159, 160, 161, 162, 127]. The few studies that explored functional interactions in high-order combinations are mostly based on SNP datasets that cover a small number of genes [126]. In addition, these studies only focus on one or a few top ranked combinations discovered from a single dataset and thus only reveal disease-specific functional interactions [126, 148]. In this study, before interpreting the top high-order SNP combinations, we first explore functional interactions in SNP combinations from a more
general perspective. The aim is to exploit some common insights on functional interactions in discriminative SNP combinations consistent across multiple datasets which may provide some guidance for future studies.

**Positive connection between $\chi^2_{jump}$ and within-pattern functional coherence**

Specifically, we study how the increase of discriminative power of a SNP combination over its subsets is related to the functional coherence of the genes covered by the combination. For this purpose, we divide all the discriminative patterns discovered by SMP into three groups, i.e. those having $\chi^2_{jump}$ values in $[+5, +\infty)$, $(-5, +5)$ and $(-\infty, -5]$ (denoted as $GP_1$, $GP_2$ and $GP_3$ respectively) and study the relative functional coherence of the patterns in the three groups. To measure the functional coherence of a SNP combination, we first obtain the set of genes covered by the combination by assigning a SNP to its closest gene, and then determine the functional similarity between each unique pair of genes covered by the combination using a human functional network integrated from a comprehensive set of resources [163]. Essentially, such an estimation decomposes the functional coherence of a set of genes covered by a SNP combination into the functional similarities of the set of unique gene pairs. We prefer this approach to a GO enrichment analysis [59] because: 1) the former can provide more detailed functional insights on gene-gene interactions within high order combinations, and 2) the latter is usually applicable to gene sets that are of sizes larger than the high-order SNP combinations discovered in this study (size-3, 4 or 5, Figure 4.5). With the decomposition-based approach for each SNP combination, we can get three distributions of gene-gene functional similarities for the three groups of SNP combinations $GP_1$, $GP_2$ and $GP_3$ respectively, where each distribution contains the functional similarities of the union (unique) of the within-pattern gene pairs from all the patterns in one of the three groups. In addition to the three distributions, we also generate a null distribution ($R_1$) by repeating the following procedure 100 times: we randomly sample gene pairs from the set of genes covered in the corresponding dataset as many as the number of gene pairs in $GP_1$, while fixing the number of times each unique gene occurs with respect to $GP_1$. Because we binarize the human functional network [163] at 0.5 (The corresponding network has a density of 5%) to make the size of the network efficient to manage). It is worth noting that the following results are consistent across different cutoff values for the functional
Figure 4.7: Functional similarity of within-combination gene pairs in three groups of discriminative SNP combinations and the null distributions (best view in color). This is to reveal the connection between $\chi^2$ – jump and within-combination functional coherence. The six comparisons, a – f, and the associated ranksum test p-values are also shown.)

network (0.5, 0.6, 0.7 and 0.8).

Figure 4.7 summarizes the comparison among the four distributions in term of the fraction of functional similarities above 0.5 and the p-values of the ranksum tests for ($GP_1$ vs. $GP_3$) and ($GP_1$ vs. $R_1$). The comparisons are done on the Kidney and Lungcancer dataset but not on Survival because there are significant SNP combinations (with FDR (w.r.t. $\chi^2$) below 0.25) discovered on the former two but not the latter as shown in Figure 4.5. A key observation is that $GP_1$ has higher within-pattern functional similarity than both $R_1$ and $GP_3$. This is reflected by the consistently higher fraction of within-pattern gene pairs with functional similarity scores above 0.5 in $GP_1$ than in $R_1$ and $GP_3$. The relative order among $GP_1$, $GP_3$ and $R_1$ is significant (ranksum test p values as shown in the figure) and consistent on the datasets Kidney and Lungcancer as well as the combined. (Datasets Kidney and Lungcancer cover the same set of genes and thus have the same null distribution of gene-pair functional similarity. Therefore, we combine the each of the four sets of gene pairs ($GP_1$, $GP_2$, $GP_3$ and $R_1$) from the two datasets to increase the sample size and allow a more reliable estimation of p value.) This observation provides a novel positive connection between the increase of discriminative power of a SNP combination over its subsets and the functional coherence among the genes covered by the combination. Essentially, this set of observations suggest that $\chi^2_{jump}$ not only improves the statistical power of the discriminative SNP-combinations
search framework (as shown in earlier), it is also indicative on the biological relevance of the genes covered by discriminative SNP combinations. The fact that $GP_3$ has the lowest fraction of functional scores above 0.5 further supports that a $\chi^2_{jump}$-based filtering is helpful and important for further exploration of functional insights from discriminative SNP combinations. The results are consistent across different cutoff values for the functional network (0.5, 0.6, 0.7 and 0.8).

It is worth noting that, $GP_1$ and $GP_2$ have about the same fraction of functional scores above 0.5 on both Kidney and Lungcancer and the ranksum tests between them are insignificant (ranksum test) on both datasets. This suggests that the genes covered by a SNP combination with $\chi^2_{jump}$ around zero also tend to be functionally related. This may be explained by existing study on positive yeast genetic interactions [164] where multiple genetic perturbations targeted on a single pathway are often found to have similar effect as the genetic perturbation of just one gene in the pathway. In contrast, the SNP combinations with $\chi^2_{jump}$ highly above zero ($GP_1$) may correspond to the genes that are involved with multiple pathways that have compensation with each other, or correspond to the genes on a single pathway but with dosage effect [164]. To our knowledge, this set of analysis is the first exploring the connection between discriminative power of SNP combinations and functional interactions from a general perspective across multiple datasets.

**Specific interpretation of two patterns discovered from Datasets Lungcancer and Kidney**

Beyond the above general biological insights, we also find that several high-order patterns with high $\chi^2_{jump}$ that are biologically interesting with respective to the complex diseases, e.g. size-5 patterns in the Lungcancer dataset and size-4 patterns in the kidney dataset. Figure 4.8 illustrates two examples with descriptions similar to Figure 4.2: a size-5 pattern discovered from Lungcancer with an odds ratio of 11.15, an p-value of $10^{-8}$ and a false discovery rate of 0.20, and a size-4 pattern discovered from Kidney with an odds ratio of 6.31, an p-value of $10^{-9.14}$ and a false discovery rate of 0.12. It is interesting that the two patterns are both more discriminative than their subsets. Furthermore, we also found that the synergy, a measure of statistical epistasis capturing non-additive interactions [152], of the Lungcancer pattern is positive, indicating a
Figure 4.8: Visualization of two SNP-genotype combinations discovered from the Lungcancer and Kidney datasets respectively. The interpretation is similar to the subfigures in Figure 5. The rsnumbers of the five SNPs in the Lungcancer pattern and the four SNPs in the Kidney pattern (all with MM genotype) are shown. The SNPs in the two patterns are mapped to the following two sets of genes, (SIM1, PARP1, WT1, ABCC1, ABCC4) and (XRCC4, SLC7A7, XRCC1 and ITGB3). The $\chi^2$ statistics of the pattern and its subsets are shown in the right subfigures. Their permutation test-based FDRs and odds ratios are also shown. The top enriched molecular interaction network (by Ingenuity Pathway Analysis) is also shown for each pattern, where the shaded nodes are those genes mapped from the SNPs in each pattern.

probable interaction beyond additive effect.

The five SNPs in the Lungcancer pattern are mapped to the five genes that are closest (chromosome location) to them respectively, SIM1 [MIM: 63128], PARP1 [MIM: 173870], WT1 [MIM: 607102], ABCC1 [MIM: 158343] and ABCC4 [MIM: 605250]. Four out of the five genes (the latter four) are previously known to be associated with cancer, with the latter three being associated with lung cancer specifically [165, 166, 167]. SIM1 has been shown to interact with ARNT [MIM: 126110], which binds to Aryl Hydrocarbon Receptor (AHR [MIM: 600253]), and the AHR pathway has been recently shown to be activated upon binding of various exogenous chemicals from cigarette smoke and might link to lung cancer risk [168]. PARP1 is a poly(ADP-ribose) polymerases-1,
Involved with DNA repair and has been associated with both better survival in non-small cell lung cancer, as well as with increased risk of lung cancer [169]. PARP1 is becoming an important target for cancer therapy, as inhibitors of PARP have low toxicity [170]. There was also a group that showed that transcriptional activation of PARP1 leads to in-silica malignant transformation of human bronchial epithelial cells [171]. WT1 (Wilms tumor 1) has been shown to be a critical regulator of senescence and proliferation downstream of oncogenic KRAS signaling [172], and KRAS [MIM: 190070] is one of the most frequently mutated human oncogenes. ABCC1 and ABCC4 are ATP-binding cassette genes, sub-family C, involved with multi-drug resistance [173] so their association with lung cancer here might have something to do with therapy. Discovering these five SNPs together as a highly predictive combination with an odds ratio of 11.15, and a large $\chi^2_{\text{jump}}$ of 7.9 may provide novel insights on their combined effects (beyond their separate effects) on their association with lung cancer. The top molecular interaction network (using Ingenuity Pathway Analysis (http://www.ingenuity.com, accessed 2012 Feb 20)) for this Lungcancer pattern is also shown. This molecular subnetwork is associated with cell death and cell cycle function with the top enriched disease being cancer, and therefore supports the functional interaction among the set of genes and their joint association with the risk of lung cancer. The other pattern in Figure 4.8 (size-4) is discovered on the Kidney dataset (XRCC4 [MIM: 194363], SLC7A7 [MIM: 603593], XRCC1 [MIM: 194360] and ITGB3 [MIM: 173470]). The four gene corresponding to the four SNPs are also enriched with a molecular interaction network with annotations closely related to transplant after kidney transplant: organismal injury and abnormalities, cellular movement, cellular-mediated immune response and cellular development.

It is worth noting that such statistically significant and biologically relevant discriminative SNP combinations are mostly high-order combinations of common variants (those SNPs with high allele frequencies). While the current focus in the exploration of the genetics of rare diseases is mostly on individually rare variants, these high-order SNP combinations indicate that common variants could also be the cause of rare diseases because combinations of common variants can be a rare composite variant.
4.4 Discussion

We presented a computational framework for searching high-order SNP combinations with strong disease association from case-control datasets with thousands of SNPs. The framework is substantially more efficient and scalable than existing techniques that usually handle tens of or hundreds of SNPs and mostly up to size-3 combinations. We further showed that, while most high-order combinations are trivial extensions of their subsets, there are indeed high-order combinations in real datasets and they have stronger associations with some disease phenotypes beyond single SNPs and low-order SNP combinations. We also evaluated the effect of a generic strategy for enhancing the statistical power of high-order SNP combination search: filtering out SNP combinations with lower or similar discriminative power than their subsets. Further leveraging the improved statistical power of this framework, we explored the functional interactions within the SNP combinations discovered from three real case-control datasets and revealed a positive connection between the increase of discriminative power of a SNP combination over its subsets and the functional coherence among the genes covered by the combination. Last but not least, we investigated two representative high-order SNP combinations (one of size-5 and the other of size-4) discovered from a lung cancer case-control dataset and a kidney transplant-rejection case-control dataset respectively, and showed that the genes covered by the two patterns are enriched with molecular interaction networks that are highly relevant to the risk of lung cancer and risk of rejection after kidney transplant, respectively. These results demonstrate the ability of our approach to find statistically significant and biologically relevant high-order, patterns, but we likely find only a subset of all possible SNP patterns of interest. In particular, some interesting patterns could be eliminated during the discriminative pattern mining step or in the $\chi^2_{\text{jump}}$ filtering step. Other existing approaches may discover some of these missed patterns, but likely miss many of the high-order patterns we find. Thus, what we provide is a well-founded and efficient (even though not complete) approach to pattern discovery in SNP datasets.

Given that there has been a lack of tools for higher-order combination analysis due to computational and statistical challenges, the proposed framework is expected to help discover novel genotype-phenotype associations missed by existing approaches that mostly take the route of univariate analysis, pathway/network enrichment analyses
that are based on univariate statistics, or epistasis analysis of low-order SNP combinations. In addition to the proposed framework itself, some general observations made in this study could also help the development of other computational techniques that search for high-order SNP combination and exploit functional insights, namely: 1) two strategies for enhancing statistical power to cope with multiple hypothesis testing in the combinatorial search could be leveraged by other approaches, 2) the observed positive connection between the increase of discriminative power of a combination beyond its subsets and the within-pattern functional coherence, both of which may guide more comprehensive exploration of functional insights of high-order interaction, and 3) the observation that many significant associations are rare combinations of common variations, which suggests an alternative direction to explore the genetics of rare diseases for which current focus is on individually rare variations.

The three real datasets used in this chapter represent a type of studies that have a different perspective from the typical disease-control designs used in most genome-wide association studies (GWAS). Specifically, the case-control designs used in the three studies are the short vs. long survival of multiple myeloma patients (all received the same treatment), acute rejection after kidney transplant (all received the same treatment) and patients with lung cancer and normal subjects (all heavy smokers). Studies with such or similar designs enforce strict additional criteria in sample selection and thus normally have much fewer samples compared to most GWAS studies. Given the limited sample sizes, the three studies adopted a SNP chip that targets a set of SNPs selected on the basis of biological candidacy in order to have better statistical power. Therefore, we expect the proposed framework to help other studies that also use targeted SNP chips to search for high-order SNP combinations that provide insights beyond univariate or lower-order analysis.

The proposed framework is able to efficiently search high-order combinations for focused studies with thousands of SNPs, but not directly suitable for focus studies with even more SNPs (e.g. tens of thousands) or genome-wide data. However, note that, this limitation is not specific to the proposed approach but to high-order interaction discovery in general, because there it is computationally infeasible to search for high-order interactions directly from genome-wide SNP datasets. After all, the state of the art methods for discovering high-order interactions could only handle less than a
thousand SNPs as reviewed in the chapter. Nevertheless, a practical solution to handle genome-wide datasets is to apply the current framework on a subset of SNPs selected by some prioritization strategy [18], e.g. adopt tag SNP selection [155] techniques to first obtain a set of less redundant SNPs, or only search for high-order interactions involving those that have sufficient marginal effects as done in [145, 146], or only search for high-order interactions among the SNPs within a certain category based on prior biological knowledge, e.g. a pathway or a genomic region, etc.

There are several possibilities for future work. First, we used a binary encoding for SNP-genotype combinations which differentiates the present of all the SNP genotypes in a pattern in a subject from the mismatch of any one genotype, but not further distinguish different numbers of mismatches. A more generalized encoding [174] that reflect the numbers of mismatches can be incorporated into the DPM framework and further explored. Second, the current study only assigns a SNP to the closest gene when exploring the functional similarity within a SNP combination, and thus may ignore the effect of a SNP on affected genes located far from the SNP (e.g. long distance cis-regulation or trans-regulation). In future work, one could integrate SNP data with gene expression data (when available) to map eQTLs before studying functional interactions within a SNP combination [175]. Third, because the current framework cannot automatically handle datasets with a large imbalance of race or gender between cases and controls, we only analyzed datasets with balanced or slightly imbalanced populations by requiring a large minimum differentiation threshold and only considered autosomal SNPs in order to avoid trivial discoveries. To make the current framework more widely applicable, we could select a subset of cases and controls to enforce a balance of population structure based on genome-wide autosomal clustering [134] or we could explore some generalization approaches that have been used to allow MDR to automatically handle confounding factors and continuous traits [127]. Last but not least, although this study focused on the discovery of high-order combinations from SNP datasets, a similar framework could also be applied for discovering combinations of other formats of genetic variations such as copy number variations or epigenetic variations such as DNA methylation, or even more generally across different types of (epi)genetic variations.
4.5 Supplementary Methods

4.5.1 Three SNP datasets and pre-processing considerations

We carefully checked the race and gender information in the three datasets to make sure the high-order combinations are not due to spurious allelic association as suggested by [176, 177]. Specifically, the subjects in the first two datasets are all Caucasian descendants and the last dataset contains both Caucasian and African American samples with an 9% imbalance between the cases and controls. We require the minimum differentiation between cases and controls to be 15% in all the SNP-combination search and analysis, in order to avoid the discovery of trivial difference due to population substructure, and we only consider SNPs from autosomes to remove the effect of gender imbalance. As shown in the result section, the comprehensive functional analysis on the discovered SNP patterns also supports that the discovered SNP combinations are functionally related to the disease instead of confounding factors such as gender and race. SNPs with more than 5% missing values are also removed.

4.5.2 Simulation of a synthetic case-control SNP dataset

We first used Hap-Sample simulator (http://www.hapsample.org, accessed 2012 Feb 20) to simulate genotype data with the 3404 SNPs from a recent study on multiple myeloma [142] as input, out of which 2172 SNPs are included in Hap-Sample. The synthetic dataset contains 70 cases and 70 controls (randomly generated from the HapMap project [178]). Note that this genotype dataset by itself does not contain disease-associated loci. Therefore, as a proof of concept we further embedded four synthetic high-order SNP combinations of size 3, 4, 5 and 6 respectively, that are associated with the case-control grouping, as shown in Figure 4.9 (with similar description as those shown in Figure 4.2). To study the scalability of SMP with respect to sample size (summarized in Figure 4.3), we further generated another 6 synthetic datasets with sample size (cases and controls combined) from 280 to 5600 in seven steps (280, 420, 560, 1400, 2800 and 5600). In each of the additional six datasets, we first use Hap-Sample simulator to generate SNP genotypes for more samples. Then, we embedded the same four synthetic patterns as done in the first data but increasing the number of samples while maintaining the frequency of each SNP genotype in the cases and controls. All the synthetic datasets
are available from the supplementary website.

4.5.3 The Apriori framework: efficient combinatorial search with anti-monotonic objective function

The Apriori framework is essentially a bottom-up exhaustive combinatorial search framework initially designed for association analysis on binary data. It first searches all the size-2 combinations and then moves up to size-3 and so on. Different from brute-force search, the Apriori framework leverages the antimonotonicity of the objective function for pruning the combinatorial search space. Specifically, an objective function $F$ is anti-monotonic if the following equation holds:

$$\forall \alpha' \subset \alpha, F(\alpha) \leq F(\alpha'),$$

(4.2)

where $\alpha$ is any combination of SNPs with the binary encoding described in the result section. An anti-monotonic objective function can be used in the Apriori algorithm to efficiently traverse the combinatorial search space without the need of visiting all the nodes in the search space, because as soon as $F(\alpha')$ is found to be disqualified with respect to a threshold ($t$) (i.e. $F(\alpha') \leq t$), Apriori can prune all the supersets of $\alpha'$ without missing any combination with an $F$ value greater than $t$, given that the antimonotonicity of $F$ guarantees that $F(\alpha) \leq F(\alpha') \leq t, \forall \alpha \supset \alpha'$. Further details on the optimized implementation of the Apriori framework can be found in [10].

4.5.4 The anti-monotonic objective function $SupMaxPair$

Given a case-control dataset, the $SupMaxPair$ of a SNP combination $\alpha$ (with the binary encoding described in the result section) is defined in [99] as below (assuming the combination is more frequent in the cases; similarly for the other situation when the combination is more frequent in the controls):

$$SupMaxPair(\alpha) = Sup_{cases}(\alpha) - \max_{\{i,j\} \in \alpha}(Sup_{controls}(\{i,j\}))$$

(4.3)

where $Sup(X)$ denotes the frequency (in percentage) of a SNP combination in a set of samples, cases or controls as shown in the subscript. So, $SupMaxPair(\alpha)$ is defined
as the difference between the frequency of a SNP combination in the cases and the maximal frequency of its size-2 subsets in the controls. An objective function defined in this way not only captures the frequency difference of a SNP combination between the cases and controls, but also has the antimonotonicity property, because the difference between an anti-monotonic function (frequency $\text{Sup}(X)$) and a monotonic function $\text{max}$ is anti-monotonic (refer to [99] for the formal proof).

Using $\text{SupMaxPair}$ in the Apriori framework guarantees the discovery of all the SNP combinations that show at least some frequency differentiation between cases and controls on the size-2 level, as controlled by a threshold on $\text{SupMaxPair}$. Therefore, if a size-5 combination does not show any differentiation until size-3 (or size-4, size-5) $\text{SupMaxPair}$ would miss it. As shown by a recent theoretical study [156], the possibility that a high-order (size-$k$) combination with strong differentiation shows zero differentiation in all of its subsets decreases dramatically when $k$ increases (generally become impossible for $k$ greater than 5). Therefore, in practice, we can use a threshold on $\text{SupMaxPair}$ as low as possible (computationally more and more expensive) as long the computational time is still acceptable, in order to minimize the chance of missing interesting high-order interactions.

4.5.5 Permutation Tests and Estimation of the False Discovery Rate (FDR)

Because of the large number of high-order SNP combinations tested in the search process, correction for multiple hypotheses testing is needed for a reasonable estimation of the statistical significance of the discovered SNP combinations. We follow the widely used empirical permutation-based approach (e.g. as used in [45]) to estimate false discovery rates (FDRs). Specifically, we first apply the proposed algorithm to the data with the original case-control grouping to get a set of discriminative patterns which are called the real patterns. Next, in each permutation test, we randomly shuffle the grouping of subjects into cases and controls while maintaining the original sample-size ratio between cases and controls, and then use SMP with exactly the same setup as for the original case-control grouping to discover a set of patterns. If $\chi^2_{\text{jump}}$ based filtering and gene set-based constraints are used for the original case-control grouping, the same procedures are also applied in each permutation in order to have an unbiased correction.
of multiple hypothesis testing. We repeat the permutation tests 100 times and get 100 lists of discriminative patterns which are called the random patterns. For each pattern (both real and random ones), we compute a $\chi^2$ statistic. The false discovery rate (FDR) of a real pattern (with respect to a $\chi^2$ statistic of $c$ and of size-$k$) is then calculated as follows: if there are $m$ real patterns of size-$k$ with $\chi^2$ greater than $c$ and there are $n$ random patterns of size-$k$ with $\chi^2$ statistic greater than $c$, then the false discovery rate is $n/(100 \times m)$. Note that, the run with real case-control label and each of the runs with randomized case-control label test the same number of hypotheses even though different number of combinations were pruned in the searching process.

Note that, in the above permutation based FDR computation, the estimation of FDR is specific to the pattern size. The use of size-specific FDR is motivated by the fact that it is harder and harder for a combination to provide additional discriminative power than all of its subsets as size increases. That is, given the same threshold of $\chi^2_{jump}$, it is less likely to discover a larger combination than to discover a smaller one. This is supported by the observations made in Figure 4.4 as well as our recent work in [156] from a more theoretical perspective. In addition, this is in accord with the observations made by Ma et al. [141]. Therefore, we chose to estimate size-specific FDRs to better reflect the statistical significance of patterns of larger sizes. It is worth noting that estimating FDRs for combinations of different sizes separately might also increase the risk of discovering false positives. While one conservative approach is to do a second round of correction on multiple hypothesis testing over different combination sizes, we highlight the potential of discovering novel biological insights from a hypothesis generation perspective in this study. Indeed, the independent functional analyses presented in the result section with the discovered high-order combinations do support that the genes covered by the discovered combinations have significant functional relationship compared to the carefully controlled null distribution. Ma et al. [141] also proposed to include the subsets of a pattern for estimating its null distribution in addition to the random patterns discovered in the permutation tests. However in this chapter, we estimate FDRs only with the random patterns discovered in the permutation tests because we directly enforce the requirement that a pattern has a sufficiently larger $\chi^2_{jump}$ than its subsets.
Figure 4.9: Four synthetic discriminative patterns of size-3 6 that we embed in the synthetic dataset as described in the method section, with similar description as Figure 4.2)
Chapter 5

A Systematic Characterization of Different Types of Discriminative Patterns and Their Relevance in Different Application Domains

5.1 Introduction

This chapter aims at a systematic understanding of different types of discriminative patterns and the corresponding biological relevance. Specifically, I categorize discriminative patterns according to four types of item interaction and presents a comprehensive discussion that defines these four pattern types and investigates their properties and their relationship to one another. In addition, these ideas are explored for a variety of datasets (ten UCI datasets, one gene expression dataset and two genetic-variation datasets). In the results, we first demonstrate the existence of different types of discriminative patterns in real datasets. Some datasets have all the four types of patterns while some datasets are specifically enriched for certain types of patterns. We then demonstrate the characterization framework can help interpret different types of molecular-level interaction underlying different sets of genes and genetic variations. For datasets in other domains beyond biology, we illustrate how pattern characterization can provide novel
insights into the discriminative structure of datasets. In Chapter 4, we proposed a
generic strategy for enhancing the statistical power of discriminative pattern mining
by focusing on a specific type of discriminative patterns with a specific characteristics.
We expect this generic strategy can also be used to enhancing the statistical power for
discovering each of the four types of patterns characterized in this chapter.

5.2 Background

For data sets with class labels, association patterns [10, 9] that occur with disproporti-
ionate frequency in some classes versus others, can be of considerable value. We will refer to
them as discriminative patterns [12, 11, 28, 179, 99, 25] in this chapter, although these
patterns have also been investigated under various names, such as emerging patterns
[20], contrast sets [21] and supervised descriptive rules [25]. Discriminative patterns
have been shown to be useful for improving the classification performance [12, 26, 27]
and for discovering sample subgroups [25, 33].

The algorithms for finding discriminative patterns usually employ a measure for the
discriminative power of a pattern. Such measures are generally defined as a function
of the pattern’s relative support$^1$ in the two classes, and can be defined either simply
as the ratio [20] or difference [21] of the two supports, or other variations, such as
information gain [12], Gini index, or odds ratio [9] etc.

To introduce some key ideas about discriminative patterns and make the following
discussion easier to follow, we use the measure that is defined as the difference of the
supports ($\text{DiffSup}$) of an itemset in the two classes (originally proposed in [21] and
used by its extensions [32, 33]). Consider Figure 5.1, which displays a sample dataset$^2$
containing 15 items (columns) and two classes, each with 10 instances (rows). In the
figure, four patterns (sets of binary variables) can be observed: $A = \{i_1, i_2, i_3\}$, $B =
\{i_5, i_6, i_7\}$, $C = \{i_9, i_{10}\}$ and $D = \{i_{12}, i_{13}, i_{14}\}$. $A$, $C$ and $D$ are discriminative patterns
whose $\text{DiffSup}$ is 0.6, 0.5 and 0.7 respectively. In contrast, $B$ is not discriminative with
a relatively uniform occurrence across the classes ($\text{DiffSup} = 0$).

---

$^1$ Note that, in this chapter, unless specified, the support of a pattern in a class is relative to the
number of transactions (instances) in that class, i.e. a ratio between 0 and 1.

$^2$ The discussion in this chapter assumes that the data is binary. Nominal categorical data can be
converted to binary data without loss of information, while ordinal categorical data and continuous data
Although $A$, $C$ and $D$ are all considered to be discriminative because of their large $\text{DiffSup}$, several observations can be made about their different characteristics. First, one of the two items in $C$ has an individual $\text{DiffSup}$ of 0.6 ($i_{10}$), while the other item ($i_9$) has a $\text{DiffSup}$ of 0. Given that $C$ itself has a $\text{DiffSup}$ of 0.5, it is obvious that the discriminative power of the pattern is mainly driven by $i_{10}$, while $i_9$ serves as a “passenger”. Such driver-passenger effects result from the fact that measures for discriminative power such as $\text{DiffSup}$ only capture the joint discrimination of a pattern but ignore the specific contribution from the items in the pattern $^3$. Second, in contrast to $C$, the $\text{DiffSup}$ values of the three individual items in $A$ are 0, 0.1 and 0.2, respectively, which are much lower than the $\text{DiffSup}$ of $A$ itself (0.6). This suggests that the items in $A$ have an incremental effect in their joint discriminative power. Third, in contrast to both $C$ and $A$, the three items in $D$ have $\text{DiffSup}$ values (0.6, 0.7, 0.6), which are very similar to that of the pattern itself (0.7). Thus, the three items in $D$, as well as their can be binarized, although with some loss of magnitude and order information.

$^3$ Such driver-passenger patterns often result when a discriminating, low-support item is combined with a high-support, non-discriminating item. Similar issue exists in frequent pattern mining where a relatively low support item can form trivial patterns with many high-support items.
combination, show a coherent behavior in their ability to differentiate between class 1 and class 2.

Patterns $A$, $C$ and $D$ have shown some of the characteristics of different types of interactions\(^4\). Indeed, some characteristics of such interactions have been discussed and studied. In particular, we can consider the discriminative power of a pattern as the confidence of an association rule by considering the class label as a special item. Then, the difference between the confidence of an association rule and the confidences of its subsets has been explored in the association rule mining community. Specifically, Bayardo et al. [158] proposed a measure called improvement as the difference between the confidence of an association rule (e.g. $Conf(X \rightarrow Y)$) and the maximal confidence of its simplifications (i.e. $max\{Conf(X' \rightarrow Y) \mid X' \subset X\}$). The association rules that have positive improvement are called productive in [180] and are considered to be more desirable than those rules with negative improvements. Similar approaches have also been proposed in the context of discriminative patterns. Garriga et al. [68] studied the closeness of discriminative patterns and proposed to remove a discriminative pattern (e.g. $X$ differentiates class 1 from class 2 by having higher support in class 1 than in class 2) if the support of $X$ is identical to any subset of $X$ in class 2, because such patterns are guaranteed to have non-positive improvements.

To illustrate the concept of improvement and prepare for the following discussion, Figure 5.2 compares the discriminative power of a pattern with the best discriminative power of all its subsets, for all the frequent patterns ($\text{minsup} = 10\%$) in the Hepatic dataset (UCI [61]). Three measures are used in the subfigures (a), (b) and (c) respectively: support difference ($\text{DiffSup}$), $\chi^2$− statistic, and mutual information. The red line indicates $y = x$, which separates the patterns that have positive improvement from those that have negative improvement. A common observation consistent across the three subfigures is that, most patterns have at least one subset having higher discriminative power (negative improvement). In contrast, a small proportion of patterns have much higher discriminative power compared with their subsets (positive improvement). This contrast indicates that, although some combinations of items have a reasonably high joint association with a class variable, the actual amount of improvement can vary

\(^4\) In this chapter, we use interaction to denote the relationship among the items in an itemset, and we use pattern to denote the concept of itemset and used interchangeably with itemset.
Figure 5.2: Comparing the discriminative power and maximal-subset discriminative power of a set of patterns discovered from the UCI Hepatic dataset, with three different measures for discriminative power. Each circle represents a pattern, with its color indicating pattern size (same for Figures 5.3(a), 5.5, 5.6 and 5.9).

greatly from pattern to pattern.

As shown by existing studies [158, 180] as well as by Figure 5.2, adding constraints on improvement can reduce the number of interesting association and discriminative rules substantially. However, the current study of the different types of interactions in discriminative patterns is lacking in the following respects:

1. The type of interaction captured by improvement is only one of several interesting types of interactions. In other words, a discriminative pattern could have an interesting interaction even if it has a close-to-zero or even a negative improvement value. For example, pattern $D$ shown in Figure 5.1 does not have an improvement of discriminative power compared to that of its subsets, and may simply be due to the existence of multiple redundant discriminative features. However, such coherent differentiation of three items may still be interesting in certain domains. Specifically, in the field of differential gene-expression module discovery, a discriminative pattern like $D$ may indicate a functional module or protein complex. A specific example will be given in section 5.3.2.

2. Even for the type of interactions captured by improvement, a further understanding of the improvement in discriminative power is possible. For example, a large improvement can either result from an independent additive aggregation of several items with separate (unrelated) association with a class variable, or a synergistic aggregation beyond the independent addition. Differentiating these different
types of interactions (Section 5.3) can be useful for biomedical informatics because they generally lead to very different types of interpretation for a disease-genetic association [164]. More generally, for other real-life applications, understanding different types of positive improvements can help us understand the discriminative structure of a dataset.

Aiming at a systematic understanding on the different types of interactions that are not captured by existing work, we motivate, formulate and design comprehensive experiments on the characterization of discriminative interactions from a general perspective of the discriminative pattern mining community. The major contributions of the chapter are:

1. We categorize discriminative patterns into four groups based on the following types of interactions: (i) driver-passenger, (ii) coherent, (iii) independent additive and (iv) synergistic beyond independent addition.

2. We present and discuss the properties and utility of the four interaction types we define. We also discuss the relationship of the four pattern types to one another.

3. We design comprehensive experiments on various types of real datasets including ten UCI datasets, a gene expression dataset and two genetic variation (SNP) datasets. The results demonstrate the existence, characteristics and statistical significance of the different types of patterns. They also illustrate how pattern characterization can provide novel insights into discriminative pattern mining and the discriminative structure of different datasets.

The rest of the chapter is organized as follows. In Section 5.3, we discuss different types of interactions and define four types of discriminative patterns. In section 5.4, we describe the datasets and experimental results. Related work on discriminative pattern mining is briefly summarized in Section 5.5, followed by conclusions and future work in Section 5.6.
5.3 Different Types of Interactions and the Corresponding Groups of Discriminative Patterns

In this section, we describe four types of item interactions and categorize discriminative patterns into four groups correspondingly. We also investigate their properties and their relationship to one another.

First we describe some terminologies that will be used through the rest of the section.

Let \( D \) be a dataset with a set of items, \( I = \{i_1, i_2, ..., i_{|I|}\} \), two class labels \(+\) and \(-\), and a set of \( n \) labeled instances (transactions), \( D = \{(x_i, y_i)\}_{i=1}^{n} \), where \( x_i \subseteq I \) is a set of items and \( y_i \in \{+,-\} \) is the class label for \( x_i \). The two sets of instances that respectively belong to the class \(+\) and \(-\) are denoted by \( D^+ \) and \( D^- \), and we have \(|D| = |D^+| + |D^-|\). For an itemset \( \alpha \subseteq I \), the set of instances in \( D \), \( D^+ \) and \( D^- \) that contain \( \alpha \) are denoted by \( D_{\alpha}, D^+_\alpha \) and \( D^-_\alpha \) respectively. Let \( p_{\alpha}, p^+_\alpha \) and \( p^-_\alpha \) be support of \( \alpha \) in \( D \), \( D^+ \) and \( D^- \) respectively, all relative to the entire set of transactions, i.e. \( \frac{|D_{\alpha}|}{|D|}, \frac{|D^+_\alpha|}{|D|} \) and \( \frac{|D^-_\alpha|}{|D|} \). Let \( p^+ \) and \( p^- \) be \( \frac{|D^+|}{|D|} \) and \( \frac{|D^-|}{|D|} \) respectively.

We use mutual information (MI) as representative measure for discriminative power among many others such as the support ratio, support difference and \( \chi^2 \)-statistic shown in section 5.2. This is because \( MI \) is based on information theory, which makes one of the interaction measures to be presented later easy to interpret. The \( MI \) between an itemset \( \alpha \) and the class variable \( C \) is computed as follows:

\[
MI(\alpha; C) = \sum_{c \in \{+,-\}} \left( p^c_{\alpha} \log \left( \frac{p^c_{\alpha}}{p_{\alpha}} \right) + q^c_{\alpha} \log \left( \frac{q^c_{\alpha}}{q_{\alpha}} \right) \right), \tag{5.1}
\]

where \( q_{\alpha}, q^+_\alpha \) and \( q^-_\alpha \) are \( 1 - p_{\alpha}, 1 - p^+_\alpha \) and \( 1 - p^-_\alpha \) respectively. Note that, in this chapter, \( MI \) is always normalized by the entropy of the class variable \( (H(C)) \), after which, it ranges from 0 to 1.

5.3.1 Driver-passenger Interaction (T1)

Pattern \( C \) shown in section 5.2 is an illustration of discriminative patterns with a driver-passenger interaction, where the driver and the passenger are both a single item in the pattern. More generally, any discriminative pattern with a subset having similar discriminative power as the entire pattern while another disjoint subset in the pattern
showing weak discriminative power are considered to have a *driver-passenger* interaction. Formally, we define the discriminative patterns with this type of interaction (T1) as follows:

**Definition 1:** An itemset $\alpha$ is a T1 discriminative pattern if the following criteria are met together for $\delta > 0$, $j > 0$, $\epsilon > 0$:

1. $MI(\alpha, C) > \delta$,
2. $\exists \alpha' \subset \alpha, |MI(\alpha, C) - MI(\alpha', C)| < j$, 
3. $\exists \alpha'' \subseteq (\alpha - \alpha'), MI(\alpha'', C) < \epsilon$.

Criterion (a) is a general requirement of the discriminative power of an itemset, which will also be used in the definition of the other types of discriminative patterns. Criteria (b) and (c) require the existence of at least one driver and at least one passenger in $\alpha$, respectively. Similar to $C$ in Figure 5.1, T1 discriminative patterns are generally not interesting because the passengers are included in a pattern as a purely mathematical consequence rather than an interpretable relationship with the other items in the pattern. Thus, in the rest of the chapter, we will focus on the other three types of interactions that can serve as evidence of meaningful relationship among the items in a pattern.

### 5.3.2 Coherent Interaction (T2)

The illustrative pattern $D$ in Figure 5.1 represents a type of interaction in which every item in a pattern is contributing with a discriminative power similar to that of the entire pattern. We call this a *coherent interaction*, and refer to patterns having this type of interaction T2 patterns.

**Definition 2:** An itemset $\alpha$ is a T2 discriminative pattern if the following criteria are met together for $\delta > 0$, $j > 0$:

1. $MI(\alpha, C) > \delta$,
2. $\text{incoherence}(\alpha) < j$,
3. $\forall i \in \alpha, \text{direction}(i) = \text{direction}(\alpha)$.

The *incoherence* in criterion (b) is calculated as the range\(^5\) of values in $\{MI(\alpha, C)\} \cup$

---

\(^5\) Difference between the maximal and the minimal value.
Figure 5.3: Illustration of T2 discriminative patterns on the gene expression dataset (described in section 5.4). (a) The entire set of discovered patterns; (b) visualization of the pattern in a binary matrix format (black indicating 1’s and white representing 0’s similar as Fig. 5.1) with the horizontal yellow line separating the two classes and the vertical green lines separating genes from each other.

\{MI(i, C)|i \in \alpha\}. Criteria (a) and (b) capture the unique property of this type of coherent interaction, i.e. each individual item in a pattern has similar (controlled by \(j\)) discriminative power as the pattern itself. Given that \(MI\) does not indicate the direction of the differentiation (i.e. a pattern or an item can be either more frequent in class + or more frequent in class −), criterion (c) is further used to make sure that all the items in a pattern have the same differentiating directionality as the pattern itself. Figure 5.3(a) illustrate the existence of T2 discriminative patterns with a real gene expression dataset. Each circle represents a pattern. The circles above the horizontal line meet criterion (a), and the circles on the left of the vertical line meet the criterion (b). Criterion (c) is implicitly enforced in the generation of the figure. The circles in the upper-left corner are T2 discriminative patterns. Note that the definition of different type of interactions is with respect to the specified parameters (here \(\delta = 0.1\) and \(j = 0.05\)), rather than a clear-cut separation. With different parameter values, different set of patterns will be considered to have a certain interaction.
The essential difference between T1 and T2 discriminative patterns is that T1 patterns include passengers (guaranteed by the criterion (c) in Definition 1), while T2 patterns do not include passengers (guaranteed by the criterion (b) in Definition 2). This difference is what distinguishes T1, an uninteresting type of discriminative pattern, from T2, a potentially interesting type of discriminative pattern. Specifically, if a dataset has many T2 discriminative patterns, we can speculate that it contains features that are discriminative and correlated with each other. Such correlation may either be due to the existence of multiple discriminative features that are redundant with each other (uninteresting), or may correspond to a functional module or protein complex that is associated with a disease in the context of differential gene-module discovery. For instance, figure 5.3(b) illustrates a T2 pattern discovered from the gene expression dataset of a study on breast cancer [51] (section 5.4.1). The genes in the pattern \(\text{MI}(\alpha, C) = 0.10\) and \(\text{incoherence}(\alpha) = 0.02\) demonstrate similar type of differentiating effect as \(C\). Discovering such patterns rather than the individual items separately could provide valuable insights towards the understanding of gene interactions in complex diseases. Indeed, three genes in the pattern (\(\text{BIRC}5, \text{Contig}38901\) and \(\text{Contig}41413\)) have been associated with breast cancer specifically\(^6\), and the other one (\(\text{CCNB}1\)) was identified as a general tumor-related gene [181]. These facts suggest that the genes in the pattern may correspond to a functional module or protein complex.

5.3.3 Independent-Additive Interaction and Synergistic Interaction beyond Independent Addition (T3 and T4)

In addition to coherent interaction, another type of interesting interaction in biomedical and genetic domains is a pattern containing a set of items (e.g. genes) that has better discriminative power than any of its subsets. Pattern A illustrates such an example, i.e. the three individual items are not discriminative by themselves while they have a 100% prediction confidence as a combination.

As discussed in section 5.2, this type of interaction can be captured by existing measures such as \textit{improvement}, which is defined to be the difference between the discriminative power (e.g. MI) between a pattern and its best subset. However, a deeper

\(^6\) www.genecards.org
understanding of the characteristics of the improvement in discriminative power is possible. For example, for pattern $A$ in Figure 5.1, the large improvement can either result from an independent additive aggregation of several items with separate (unrelated) association, or a synergistic aggregation beyond the independent addition. Differentiating these different types of interactions is important because they generally lead to very different types of interpretation of a disease association.

Next, we will first discuss two different types of improvement interactions and then define another two types of discriminative patterns accordingly.

**Differentiating two types of improvement interactions**

Bayardo et al. [158] defined *improvement* in the context of association rule mining based on the *confidence* of a rule. We first rewrite the *improvement* ($\text{Imp}$) in the context of discriminative pattern mining based on MI as below:

$$\text{Imp}\,C(\alpha) = \text{MI}(\alpha, C) - \max_{\alpha' \subset \alpha} (\text{MI}(\alpha', C)).$$

(5.4)

To ease the motivation of different types of *improvement*, we consider the following equation for a pair of items $\alpha = \{i_a, i_b\}$.

$$\text{Imp}\,C(\alpha) = \text{MI}(\alpha, C) - \max(\text{MI}(i_a, C), \text{MI}(i_b, C)).$$

(5.5)

which is essentially the amount of additional information about the class variable $C$ that can be provided by the two items as a combination, compared to the information that each item can provide (the bigger one). This additional amount of information can either result from an independent additive aggregation of several items with separate (unrelated) association, or a synergistic aggregation beyond the independent addition.

D. Anastassiou [152] applied a measure called *synergy* (originally used in neuroscience literature [182]) to discover gene-gene interactions that are beyond the independent addition of all possible partitions of its subsets. In this chapter, we leverage it to characterize discriminative patterns from a more general perspective.

We start from the following equation for calculating the synergy computation between a size-2 pattern $\alpha = \{i_a, i_b\}$ and a class variable $C$,
\[ \text{Syn}^C(\alpha) = MI(\alpha, C) - (MI(i_a, C) + MI(i_b, C)), \] (5.6)

which is calculated as the amount of additional information about the class variable \( C \) that can be provided by the two items as a combination, compared to the information that each of the item can provide independently (sum of the two individual MIs). Compared to Equation 5.5, the essential difference between \textit{improvement} and \textit{synergy} is that, \textit{improvement} is with respect to the bigger MI of the two, while \textit{synergy} is compared to the summation of MIs of the two. Indeed, the summation of the mutual information of two items is used in information theory to represent the combined effect of two items with independent association with a class variable \[152\]. Thus, \textit{synergy} can be leveraged to refine the discriminative patterns with positive \textit{improvement}, based on the characteristics of an improvement.

In order to provide an intuitive understanding, Figure 5.4 illustrates an underlying mechanism of synergistic interaction in the context of yeast genetic interaction. Two distinct pathways are shown in the figure, i.e. \( A \rightarrow B \rightarrow C \) and \( X \rightarrow Y \rightarrow Z \), which impinge on a common biological process that is essential to the survival of a yeast cell (the wild type). Due to parallel structure, the two pathways can compensate for the loss of the other, and thus a genetic perturbation (natural variations) on either of the two pathways separately (e.g. perturbation only in \( A \)) fails to cause any observable defects in cell survival. However, the simultaneous perturbations in \( A \) and \( Y \) disrupt both pathways and result in the lethality the cell. In this example, \( A \) and \( Y \) have a synergistic interaction with respect to the class label (survival or not) of a cell.

Equation 5.7 gives the general definition of \textit{synergy} for an itemset \( \alpha \) beyond pairs (also defined in \[152\]).

\[ \text{Syn}^C(\alpha) = MI(\alpha, C) - \max_{\text{all partitions into } \{S_i\}} \sum_i MI(S_i, C), \] (5.7)

where a partition is defined as a collection \( \{S_i\} \) of disjoint subsets \( S_i \) whose union is \( \alpha \). For example, for a size-3 pattern \( \alpha = \{i_a, i_b, i_c\} \),
Figure 5.4: Illustration of the mechanism of synergistic interaction in the context of yeast genetic interaction (Figure taken from Costanzo et al [2]).

\[ S_{\text{syn}}^C(\alpha) = MI(\alpha, C) - \max \left\{ \begin{array}{l} MI(i_a, C) + MI(i_b, C) + MI(i_c, C) \\ MI(i_b, C) + MI(\{i_a, i_c\}, C) \\ MI(i_c, C) + MI(\{i_a, i_b\}, C) \\ MI(i_a, C) + MI(\{i_b, i_c\}, C) \end{array} \right\} \quad (5.8) \]

This generalized definition is consistent with the intuition that \textit{synergy} is the additional amount of information about a class variable provided by an integrated discriminative power compared with what can be best achieved after breaking the pattern into components by the sum of the contributions of these components. The partition of the set of factors that is chosen in this formula is the one that maximizes the sum of the amounts of mutual information connecting the subsets in that partition with the class variable, and we will refer to it as the \textit{best aggregated MI}. Note that, the computational complexity of \textit{synergy} for an itemset of size \( n \) is \( O(B(n)) \), where \( B(n) \) is the \( n^{\text{th}} \) Bell number\(^7\), which increases in a dramatically fast speed. In practice, to avoid unnecessary computations, we actually only need to compute \textit{synergy} (as well as \textit{best aggregated MI}) for those patterns with positive \textit{improvement}, which is much more efficient to compute, i.e \( O(n) \).

Given the definition of \textit{improvement} and \textit{synergy}, it is easy to notice that \textit{synergy} is guaranteed to be larger than \textit{improvement} (follows from the fact that MI is non-negative.

\(^7\) en.wikipedia.org
Figure 5.5: Illustration of general improvement and synergistic interaction beyond independent additive effect on the gene expression dataset. Best aggregated MI is only computed for those patterns that have positive Imp.

Proof omitted). Essentially, synergy is a more restrictive measure specifically for capturing interaction beyond independent addition. Figure 5.5 compares how improvement and synergy capture the interaction of discriminative patterns discovered from a gene expression dataset (described in section 5.4.2). Figure 5.5(a) shows the MI and best subset MI of the discriminative patterns as a scatter plot. The horizontal dashed line indicates the cutoff values for MI, and the other dashed line (representing $y = x$) separate the patterns with MI higher than best subset MI (positive improvement) with those that have negative improvement. As shown, there are quite a few patterns above both the horizontal line and $y = x$, with size ranging from 2 to 5. In contrast to Figure 5.5(a), the x-axis in Figure 5.5(b) is best aggregated MI instead of best subset MI. Corresponding to this difference, there are far fewer discriminative patterns (all of size-2) that are above both the horizontal line (high discriminative power) and $y = x$ (positive synergy) at the same time. This contrast is as expected given our discussion above that synergy is a more restrictive type of interaction beyond the independent additive effect, and is guaranteed to be no more than improvement for any pattern.
Defining two different types of discriminative patterns with positive improvement

With synergy, we can divide all discriminative patterns with positive improvement into two groups, i.e. those that have negative synergy and those that have positive synergy. Alternatively, the two groups can also be defined as those patterns that have positive improvement (including both positive and negative synergy) and those that specifically have positive synergy. We take the latter route, given its simplicity in term of the definitions as shown below. Note that the observations made from both routes are essentially the same.

**Definition 3**: An itemset $\alpha$ is a T3 discriminative pattern if the following criteria are met together for $\delta > 0$, $j > 0$:

\[
\begin{align*}
(a) & \quad MI(\alpha, C) > \delta, \\
(b) & \quad Imp^C(\alpha) > j.
\end{align*}
\]  
(5.9)

**Definition 4**: An itemset $\alpha$ is a T4 discriminative pattern if the following criteria are met together for $\delta > 0$, $j > 0$:

\[
\begin{align*}
(a) & \quad MI(\alpha, C) > \delta, \\
(b) & \quad Syn^C(\alpha) > j.
\end{align*}
\]  
(5.10)

For illustration, we note that pattern A in Figure 5.1 is a T4 discriminative pattern, with $MI(A) = 0.39$, individual item $MI$’s 0.007, 0.008, and 0.029, respectively. The $M$-improvement is 0.18, while the synergy is 0.17.

If a dataset has many T3 discriminative patterns, we can speculate that it contains features that complement each other for higher discriminative power in their association with the class variable. Further, if there are also many T4 discriminative patterns, it is expected that some features have synergistic cooperative effect beyond independent addition. In contrast, if a dataset has few or no T3 discriminative patterns, the discriminative features, if they exist in the dataset, are expected to be either correlated with each other (T2) or not form high-order combinations at all, i.e., have a very low joint frequency to pass the support threshold.

Figure 5.6 shows the two sets of patterns: T3 (upper right region in Figure 5.6(a)) and T4 (upper right region in Figure 5.6(b)) respectively, both with $\delta = 0.1$ and $j = 0.05$. 
Figure 5.6: Illustration of T3 and T4 discriminative patterns on the gene expression dataset. Syn is only computed for those patterns that have positive Imp.

Note that, there are only two patterns (size-2) that have synergy greater than $j = 0.05$. This again indicates that the synergistic interaction in T4 patterns is rare. However, as will be shown in section 5.4.2, these two T4 patterns (even very rare) are statistically significant after correcting for multiple hypothesis testing to control type I error (false discover rate < 0.01), and thus can be of significant interest in the biomedical domain. After all, $j = 0.05$ is an arbitrary threshold that is used to illustrate the concept. In fact, there are many other discriminative patterns with positive synergy (even though they are below 0.05) as shown in Figure 5.6(b), which may also be interesting to specific domains.

Figure 5.7 illustrates two example patterns for T3 and T4 respectively. In Figure 5.7(a), the individual MIs of the two SNPs are 0.054 and 0.04 respectively. As a combination, it has a MI of 0.107, which is almost the same as the sum of the two individual MIs (a low synergy of 0.013), indicating a independent additive effect and thus a T3 pattern. In contrast, the two SNPs in Figure 5.7(b) have a high synergy of 0.108, indicating a large cooperative effect beyond independent addition. Indeed, the two genes that the two SNPs are located on, MSH6 and DPYD are known to code proteins that
SLCO1A2  rs4337089
XRCC4   rs2075685
MI: 0.107; MI-Imp: 0.053; Synergy: 0.013;
MI Individual: 0.054, 0.040

MSH6  rs2020911
DPYD  rs1520663
MI: 0.160; MI-Imp: 0.134; Synergy: 0.108;
MI Individual: 0.026, 0.026

Figure 5.7: A T3 example and a T4 example, both discovered from the M-Survival SNP dataset as described in section 5.4.1 (refer to Fig. 5.3(b) for similar description.

have the following functions\textsuperscript{8}: (i) recognizing mismatched nucleotides and (ii) catabolizing two specific types of nucleotides (uracil and thymidine), respectively. The fact that they have a synergistic interaction agrees with their closely related functions and potential compensation for each other as illustrated in Figure 5.4.

5.3.4 The relationships among the four different types of interactions

In this subsection, we discuss the relationships among different types of interactions and relate other types of interactions to the four defined interactions in order to have a systematic understanding about item interactions in discriminative patterns.

Figure 5.8 shows the three interesting types of discriminative patterns (T2, T3 and T4) in the context of all discriminative patterns using a Venn diagram. The outermost circle contains all the discriminative patterns with $MI > \delta$. The set of T3 discriminative patterns is a superset of the set of T4 patterns based on the Definitions 3 and 4 (with the same $\alpha$ and $j$) and the fact that synergy is always no more than improvement for any pattern. The set of T2 discriminative patterns is disjoint with the set of T3 patterns, when the same value of $j$ is used in Definitions 2 and 3. Specifically, for any given value of $j$, criterion (b) in Definition 2 and criterion (b) in Definition 3 can not be met at the

\textsuperscript{8} www.genecards.org
The next natural question is the nature of the discriminative patterns that are not any of the three types (T2, T3 and T4), i.e. the region represented by the gray background color. Indeed, they can all be considered to be in one of the two possible cases: either (i) T1 patterns with the driver-passenger interaction or (ii) the patterns, each of which can be considered as a combination\(^9\) of T2 and T3 patterns. Due to the limit of space, the proof of this is available on the chapter website.

Note that, the goal of characterizing discriminative patterns with different types of interactions is to identify different types of interesting discriminative patterns, which are specifically $T2 - T4$ in the context of this chapter. It is worth noting that we do not exclude the possibility that the patterns in the gray region ($T1$ patterns, or combinations of $T2$ and $T3$ patterns) may also be interesting in some specific domains even though they are not considered as such in this chapter. Thus the focus of this chapter is to initiate a study of the item interactions in discriminative patterns, rather than identifying the all possible types of interesting item interactions in discriminative patterns.

\(^9\) For example, if $R$ is a $T2$ pattern and $Q$ is a $T3$ pattern, then $R \cup Q$ is a combination of $T2$ and $T3$ pattern.
5.3.5 Correction for Multiple Hypothesis Tests

As discussed by recent work [180, 183, 184], an association pattern mining task (e.g. frequent patterns, discriminative patterns) essentially conducts a large number of hypothesis testing. Thus, in order to control type I error (due to the multiple hypothesis testing), corrections on the significance of the discovered patterns is necessary. Among different approaches for correcting multiple hypothesis testing, the randomization based approaches [183] are non-parametric and thus more reliable in term of not introducing bias. Randomization frameworks have been extensively explored in the context of frequent pattern mining and clustering [184]. For discriminative pattern mining, a special type of randomization procedure is needed, in which the randomization is performed by shuffling the class labels for the samples. For the details of the randomization and the calculation of corrected p-value or false discovery rate (FDR), refer to [99, 179, 45]. In section 5.4.2, we will show that many of the discovered $T_2 - T_4$ patterns are statistically significant after correcting for multiple hypothesis tests.

5.4 Experiments

In this section, we use a variety of real datasets to demonstrate the existence, properties and statistical significance of different types of discriminative patterns that we characterized in section 5.3. We also show how the characterization can provide novel insights into discriminative pattern mining and the discriminative pattern structure of different datasets, beyond those provided by current approaches that focus mostly on pattern-based classification and subgroup discovery.

5.4.1 Data Sets

We use the following three different types of real datasets, with details summarized in Table 5.1 and detailed pre-processing steps described on the chapter website:

Ten UCI datasets [61] with a variety of dimensions and densities (Table 5.1).

A gene expression dataset on breast cancer [50] (pre-processed as suggested in [51] and binarized as done in [99, 52]. We denote this dataset by Breast(GEP).

Two single-nucleotide polymorphism (SNP) datasets: SNP profile captures the genetic variations of a person at single-nucleotide resolution, which are commonly
Table 5.1: Details of each dataset and a summary of the number of different types of discriminative patterns discovered. For column 7–9, in addition to the number of discovered patterns, we also show (in the bracket) the number of unique items in the union of the set of patterns to reflect the redundancy among the patterns. $\delta = 0.1$ and $j = 0.05$ are used for all the datasets.

For each dataset, we first discover a set of discriminative patterns with existing algorithms. Specifically, for the dense and high-dimensional datasets (Breast (GEP), the two SNP datasets, Chess (UCI) and Hypo (UCI)), we leverage the SMP algorithm proposed in [99] to discover discriminative patterns ($\text{SupMaxPair} = 0.2$). For the other sparse or low-dimensional datasets we simply use FPC [60] with $\text{minsup} = 10\%$, because SMP may miss some high-support patterns although it is efficient on discovering discriminative patterns from dense and high-dimensional data [99].

For each set of discovered patterns (only closed itemsets), we apply the criteria of the three types of discriminative patterns ($T2 - T4$) presented in section 5.3 and get the number of patterns for each type. Figures 5.9 illustrate the existence of T2, T3 and T4 discriminative patterns in the representative SNP dataset (subfigures (a)-(c)), and the representative UCI datasets (subfigures (d)-(f)). Note that, the similar set of figures for the gene expression dataset can be found in section 5.3.3, i.e. Figures 5.3(a), 5.6(a) and
Several observations about each type of interactions can be made from Table 5.1 and Figure 5.9.

1: T2 discriminative patterns are common in most UCI datasets and the gene expression dataset, but not in the SNP datasets: On one hand, this indicates that the UCI datasets and the gene expression dataset have features that are discriminative and correlated with each other. On the other hand, the fact that the SNP datasets do not have T2 patterns indicates that the discriminative SNPs are not correlated with each other. In addition, column 7 in Table 5.1 indicates that the across-pattern redundancy is high in T2 patterns, i.e. the number of unique items is generally much smaller than the number of patterns.

2: T3 discriminative patterns exist in about half of the UCI datasets and all three of the biological datasets: These datasets are expected to contain discriminative features that are complementary to each other in their improved discriminative
power as a pattern. In contrast, the other datasets that have very few or no $T_3$ discriminative patterns, the discriminative features, if they exist in the dataset, are expected to be either correlated with each other ($T_2$) such as Cleve (UCI) or simply do not contain interesting feature combinations (independent association with the class variable) such as Mushroom and Waveform. The fact that the three biological datasets have many $T_3$ discriminative patterns is consistent with common knowledge that complex diseases involve the cooperation of multiples genes. This is especially true for the two SNP datasets, where there are no $T_2$ discriminative patterns but many $T_3$ discriminative interactions. In addition, column 8 shows that the across-pattern redundancy in $T_3$ patterns is lower than in $T_2$ patterns, because the number of unique items is generally similar as the number of patterns.

3: $T_4$ discriminative patterns exist in all three of the biological datasets and only one UCI dataset (Sonar): First, $T_4$ pattern is rare because $T_4$ is based on the most restrictive type of interaction (synergy). Nevertheless, the fact that the gene expression and SNP datasets contain many $T_4$ discriminative patterns indicates the relatively higher complexity in genetic datasets compared to the common UCI datasets. In addition, column 9 shows that the across-pattern redundancy in $T_4$ patterns is similar as in $T_3$ patterns, which are both lower than in $T_2$ patterns.

4: The number of $T_2−T_4$ patterns is much smaller compared to the overall number of discriminative patterns: The last column in Table 5.1 shows the fraction of discriminative patterns that are either $T_2$, $T_3$ or $T_4$ (the three interesting types). Except for the two SNP datasets, the fractions are generally very low, which indicate that many discriminative patterns with good discriminative power are not interesting from the perspective of the interestingness considered in this chapter. The extreme cases are the Mushroom and Waveform datasets, which do not contain any of the three types of patterns. This indicates that the discriminative features are neither uncorrelated with each other nor complementary to each other in these two datasets, i.e. independently discriminative features. This observation indicates that the actual number of interesting patterns is much more manageable compared to the huge number of patterns that are generally encountered without a detailed characterization.

5: $T_2−T_4$ patterns generally have smaller size compared to the entire set of discriminative patterns: From the color of the circles in the figures, $T_2−T_4$
patterns are generally of size 2 – 6. This is in contrast to the wider range of sizes for the entire set of discriminative patterns, which can be as high as 14. This agrees with the observations made in the recent work on constraint-based generation of high-order discriminative patterns [187]. Specifically, Steinbach et al. observed that the larger (size) an itemset becomes, the harder it is for the itemset to meet the constraints for a discriminative pattern, when the constraints are not only on the discriminative power of the pattern but also the improvement of the discriminative power. This also suggests that the computational complexity of discriminative pattern mining could be less than expected given that too large patterns tend to be uninteresting in term of the meaningful relationships scoped in this chapter.

6: $T_2 - T_4$ patterns discovered from all the datasets are statistically significant. In the columns 7 – 9 in Table 5.1, all the $T_2 - T_4$ patterns are statistically significant with $FDR < 0.01$ after correcting for multiple hypothesis tests to control type I error (method discussed in section 5.3.5). Specifically for the three biological datasets, the characterization of those statistically significant gene or SNP combinations can assist the further biological interpretations, and reveal novel insights to the mechanisms of complex diseases.

The above comprehensive observations illustrate the existence, characteristics and statistical significance of the different types of patterns. They also illustrate how the proposed framework can provide novel insights into discriminative pattern mining and the discriminative pattern structure of different datasets.

5.5 Related Work

Over the past decade, many approaches have studied discriminative patterns and related topics. The most relevant related work was discussed earlier in Section 5.2. Among other work focusing on mining discriminative patterns, the most relevant ones are [63, 64, 13, 48]. Many existing approaches also used discriminative pattern for classification [35, 12, 11, 28, 66, 29]. Additional related chapters in the area include [13, 33, 64, 68, 69, 99]. We also refer the readers to a comprehensive survey on discriminative patterns by Novak et al. [25].
5.6 Conclusion

In this chapter, we categorized discriminative patterns into four groups based on item interactions: (i) driver-passenger, (ii) coherent, (iii) independent additive and (iv) synergistic beyond independent addition. The coherent, additive, and synergistic patterns are of practical importance, with the latter two representing a gain in the discriminative power of a pattern over its subsets. Synergistic patterns are most restrictive, but perhaps the most interesting since they capture a cooperative effect that is more than the sum of the effects of the individual items in the pattern. The experiments provided a number of insights into the nature of discriminative patterns in various real datasets and the characteristics of the different types of patterns.

Particularly worth noting is that all types $T^2 - T^4$ patterns were significant in all the datasets for which we evaluated pattern significance. While this needs to be investigated further, we believe that this is mostly due to the pruning of a large number of patterns that are not likely to be of interest. Without such pruning, the number of patterns is typically very large, as is typical in most types of association analysis, and thus, the FDR of the resulting patterns tends to be low unless the patterns are very strong since FDR depends very heavily on the number of patterns being considered. We are hopeful that this observation will allow discriminative pattern mining to be more effectively used for a wide variety of applications, both in the biomedical area and beyond.

Several further directions can be explored in the future. (1) The four types of patterns defined in the chapter are mainly based on the building-block measure mutual information to make the presentation consistent and easy to follow, and other statistical measures can also be explored as building-block measures or specifically for a certain type of pattern. For instance, the logistic regression-based measure studied in [146] can be leveraged as an alternative to synergy. (2) Other types of interactions can be explored especially those that may be interesting to specific domain but are considered as non-interesting in the context of this chapter. (3) From the computational perspective, it is also interesting to design mining algorithms that can directly search for a particular type of discriminative patterns, which is expected to be much faster given a more specific definition, whose anti-monotonic properties can be leveraged as additional pruning constraints.
Chapter 6

Discriminative Pattern Network: Global Topology and Local Structures

6.1 Introduction

This chapter proposes a novel approach to organize discriminative patterns into an interaction network that allows the discovery of high-level structural knowledge. On the global scale, we explore the network structure of the constructed pattern networks and discover that they generally have statistically significant topological measures compared to random networks. On the local scale, biologically relevance substructures can be systematically discovered. Specifically, a general framework is proposed to detect pathway-pathway interaction pairs that are enriched for genetic level interactions from genome wide association datasets. Validations across independent real datasets not only demonstrate the reliability of the proposed framework but also lead to several interesting biological insights on several complex diseases such as breast cancer and Parkinson’s disease. It is worth noting that the network-level analysis of discriminative patterns is related to the pattern summarization problem studied in the data mining community but provides different insights. Specifically, existing work on pattern summarization
explores the overlaps within a set of patterns and aims to select a subset of less redundant patterns that could represent the original set. While the discriminative pattern network analysis proposed in this chapter also organizes a set of discriminative patterns as a network that reflect relationships among different patterns, the ultimate goal is to discover significant network topology and high level substructures with strong domain relevance. Therefore, in contrast to pattern summarization that still generate a list of isolated patterns, the approach proposed in this section can provide network-level insights beyond any individual patterns.

6.2 Background

Genome-wide association studies (GWAS) have been increasingly successful at identifying single-nucleotide polymorphisms (SNPs) with statistically significant associations for a variety of diseases [18]. However, for most diseases, there remains a substantial disparity between the disease risk explained by the discovered loci and the estimated total heritable disease risk based on familial aggregation, a problem that has been referred to as "missing heritability" [116, 118, 117]. While there are a number of possible explanations for missing heritability, genetic interactions between loci are one potential culprit [116, 118, 117]. Genetic interactions (used interchangeably with epistasis in this paper [19]) generally refer to two or more genes whose contribution to a phenotype goes beyond the independent effects of the genes (e.g. additive or multiplicative) [17, 188]. Genetic interactions are expected to play an important role in complex diseases [116, 118, 117], and in fact, there have been a number of recent reports of confirmed interactions between genetic variants in different disease contexts. Recent work also demonstrated through simulation studies that genetic interactions may explain a substantial portion of the heritability unexplained by single-locus genome-wide association studies [189], but also illustrated the major challenge in detecting them, i.e., the same study estimated that approximately half a million subjects would be needed to detect significant genetic interactions under reasonable assumptions [189], which is clearly far beyond the cohort sizes of typical GWAS studies.

Recent genome-wide reverse genetic screens in model organisms have produced rich insights into the prevalence and nature of genetic interactions [164, 190], and provide
Figure 6.1: A toy example illustrating the concept of between pathway model (BPM) in the context of complex human disease. (a) Two distinct linear pathways, ABC and XYZ, that both contribute to a common biological process that is essential for maintaining the normal health state of a person. The genetic perturbations on the two pathways and their effect on disease risk are also illustrated. (b) The bipartite structure resulting from the functional compensations between the two pathways shown in (a), i.e., the joint genetic perturbation on any pair of genes between the two pathways yields a non-additive synergistic effect on disease risk. (c) The flowchart of the five steps in BridGE. (d) A conceptual illustration of the pathway-pathway pairwise enrichment analysis.

clues that could benefit the discovery of disease-associated genetic interactions in the GWAS context. Specifically, a global analysis of the yeast genetic interaction network suggested that genetic interactions tend to occur in larger structures connecting across functionally compensatory modules (pathways or protein complexes) as opposed to exist as isolated instances. For example, suppose there are two distinct pathways, ABC and XYZ (Figure 6.1), both of which contribute to a common biological process that is essential for maintaining a healthy (non-disease) state. Due to their functional redundancy, each pathway can compensate for the loss of the other, and thus, a genetic perturbation on either of the two pathways independently (e.g. genetic variation only
in A) would not cause any observable defects. However, joint perturbations in A and Y would disrupt both pathways and result in an increased disease risk. Importantly, the same phenotypic outcome could be achieved by several other paired genetic perturbations as well (e.g. $A - X$, $A - Z$, $B - X$, $B - Y$, $B - Z$, as summarized in Figure 6.1). This model for how genetic interactions should naturally arise, termed the "between pathway model" (BPM), has been widely observed in yeast genetic interaction networks [164, 190]. In fact, a recent study estimated that more than 70% of negative genetic interactions derived from fitness phenotypes in yeast occur in such structures of size 3 by 3 or larger (9 total interactions), suggesting this type of local clustering of interactions is the rule rather than the exception [191]. Other work in yeast exploited this structure to derive genetic interaction networks from phenotypic variation in a yeast recombinant inbred line population [192]. This expected structural property of genetic interactions has important practical implications in the human GWAS context. Specifically, although SNP-level pairwise tests of genetic interaction are substantially statistically under-powered, significant pathway-level interactions may be detectable if one were to leverage the knowledge that pairwise SNP interactions tend to cluster into larger substruc

We will explore the statistical properties and biological relevance of the SNP-SNP interaction networks from two directions. In the first direction, we characterize the global topology of the networks and compare these topological properties of the network constructed with the original case-control grouping information with the random networks that are constructed with randomized case-control grouping information, to access the statistical significance of the topology of the original SNP-SNP interaction network. In the second direction, we search for biologically relevance local network substructures, e.g. bipartite subgraphs, cliques, that could reflect functional interactions between biological modules. Through, these two scales of analyses (global and local), we want to demonstrate the important of study the network structures of discriminative patterns to obtain insights beyond those provided by a set of discriminative patterns when they are treated and analyzed as separate candidates.
6.3 Global Topology of Discriminative Pattern Networks

A recent work by Hu et al. [193] studied the statistical significance of several topological measures of statistical epistasis network. However, there are several key confounding factors that may have biased the analysis in this work. First, the study used a targeted chip that only contains a set of 2000 SNPs instead of a genome-wide association study. The major concern is that population stratification could not be stably estimated from such targeted studies. Consequently, it is unclear if the significance of the statistical epistasis network observed in the study is due to imbalanced population structure. Second, the SNP-SNP network is constructed with all the SNPs genotyped in the chip where the linkage disequilibrium (LD) structures among the SNPs could introduce bias in measurement of the topology of SNP-SNP epistasis network. Third, the inter-relationship among the three topological measures used in the study (number of edges, number nodes and the size of the largest connected component) is not explored, which could lead to over estimated statistical significance of the epistasis network. Specifically, size of the largest connected component could be significant simply due to a trivial consequence of the significant number of edges or nodes. Last but not least, the results in the study is based on a single threshold for network binarization. Thus, it is unclear if the significance of the epistasis network is stable over a range of different binarization thresholds.

6.3.1 Methods and experimental design

In this work, we aim to explore the above limitations on two genome-wide association datasets with an expanded set of topological measures, i.e. adding clustering coefficient and network efficiency. We first study the effect of linkage disequilibrium structures in SNP datasets on the properties of the SNP-SNP interaction network and demonstrate that we need to remove the LD bias in order to more accurately capture the topology of the SNP-SNP interaction network. Figure 6.2 provides a toy example showing that, each of the five topological measures have different values between an original network (with LD structure) with the corresponding network without the bias of LD structure.

The topology of a SNP-SNP network varies with different threshold used for binarization. To access the statistical significance of each of the five topological measures for
each binarization, we shuffle the case-control grouping in the original case-control data and reconstruct random SNP-SNP networks with the exact same procedures as applied to the original network and estimate a empirical p value for the original network. Figure 6.3 illustrate this evaluation strategy for the number of edges on a Parkinson dataset [194] as described in the method section.

### 6.3.2 Experimental results

We first study the effect of LD structure on topological measures. We use two genome wide association datasets, one from a study on Parkinson’s disease [194] and the other from a study on breast cancer [195]. For each dataset, we first select a subset of 4000 SNPs with stringent LD removal thresholds and then we create two additional datasets with 2000 and 4000 SNPs added (the SNPs are selected from those in linkage with the original 4000 SNPs). We denote these three networks as $N_0$, $N_1$ and $N_2$. Figure 6.4 summarize the results comparing the three network with respect to each of the five topological measures on each of the two datasets.
Figure 6.3: Illustration of the evaluation strategy used to access the significance of a topological measure for each binarized SNP-SNP interaction network. X-axis: different thresholds for network binarization; Y-axis (left): number of edges in log scale; Y-axis (right): p value of number of edges at each threshold for binarization. Red line corresponds to the real network with 40 different binarization thresholds. Green lines are based on 500 permutations. The blue line shows the empirical p value for each of the binarization threshold.

A couple of observations can be made. First, the statistical significance of a SNP-SNP interaction network changes over different binarization thresholds. This suggests that the statistical significance of a SNP-SNP interaction network is not stable over different binarization thresholds and an reliable interpretation should be made considering the variations of the significance rather than based on a single view of the network. Second, the statistical significance of a SNP-SNP interaction network for differs among the three networks for most binarization thresholds. Specifically, the significance of the $N_0$ could be either higher or lower than the corresponding $N_1$ and $N_2$. When a topological measure in $N_0$ is significant while that of the corresponding $N_1$ and $N_2$ network is not, it indicates that the significance of the SNP-SNP interaction network is only visible after properly controlling the LD bias. In contrast, when a topological measure in $N_0$ is not significant but that of the corresponding $N_1$ and $N_2$ network is, it indicates that the significance of the SNP-SNP interaction network is over estimated because of the LD
Figure 6.4: Comparing the three networks with different amount of LD structure with respect to each of the five topological measures on each of the two datasets. X-axis: different thresholds for network binarization; Y-axis: p value of a topological measure (indicated by the text on the left of each row) at each threshold for binarization. Blue line corresponds to the network ($N_0$) constructed with 4000 SNPs ($S_{4000}$) with proper removal of LD structure. Green lines corresponds to network $N_1$ with 2000 SNPs that are in LD with $S_{4000}$ added to $N_0$. The red line corresponds to network $N_2$ with 2000 SNPs that are in LD with $S_{4000}$ added to $N_0$.

bias. This demonstrates the importance of removing bias introduced by LD structure in the evaluation of the topology of a SNP-SNP interaction network.

Next, we study the global topology of SNP-SNP interaction networks with the two genome-wide association datasets, after properly handling LD structure (see Methods for details) and also explore the dependency of the significance among different topological measures. For number of edges (and nodes), we first estimate the direct p value comparing to randomized networks and then we fix the number of nodes (and edges) between the original network and the randomized networks and estimate a conditional p value. For the size of the largest connected component, clustering coefficient and network efficiency, we compute three p values for each, i.e. direct p value, p value
Figure 6.5: Statistical significance of each of the five topological measures on the Parkinson’s disease dataset. X-axis: different thresholds for network binarization; Y-axis: p value of a topological measure (indicated by the text nearby) at each threshold for binarization. Blue curves correspond to the direct significance. Green curves represent the significance conditional number of edges. Red curves represent the significance conditional number of nodes.

Figure 6.5 summarize the results on the Parkinson’s disease dataset. Figure 6.6 summarize the results on the breast cancer dataset. Direct (non conditional) p value, p value conditioned on the number of nodes and p value conditioned on the number of edges are colored with blue, red and green, respectively. Several consistent observations can be made from the results obtained from both datasets.

1. Number of edges: In both datasets, the unconditional p values are not significant. However, p values conditioned on number of nodes are significant over several different binarization thresholds. This indicates that, although random networks could have similar number of edges as the original network under the same binarization threshold, these edges in random networks generally involve a larger number of nodes (SNPs). That is consistent with the biological mechanism assumed in the introduction of this chapter, i.e. the between pathway model tends to result in SNP-SNP interactions that cluster into bipartite substructures. In such substructures, many edges can exist between the common two sets of nodes,
Figure 6.6: Statistical significance of each of the five topological measures on the breast cancer dataset. X-axis: different thresholds for network binarization; Y-axis: p value of a topological measure (indicated by the text nearby) at each threshold for binarization. Blue curves correspond to the direct significance. Green curves represent the significance conditional number of edges. Red curves represent the significance conditional number of nodes.

and thus tend to have a larger number of edge with the same number of nodes. The interpretation not only demonstrate the importance of calculating conditional p value but also suggests the significance of the number of edges in real SNP-SNP interaction networks do exist but only detectable when conditioned on the number of nodes (consistent between the two datasets).

2. Number of nodes: In the PD dataset, number of nodes have significant unconditional p value as well as p value conditioned on number of edges. It is worth noting that, such significance only exist at the binarization thresholds on which the number of edges are not significant. On one hand, this suggests SNP-SNP interaction networks could have significance on different topological measures with respect to different binarization thresholds. On the other hand, it again demonstrate the additional insights provided by a systematic evaluation of the SNP-SNP interaction network over a wide range of binarization thresholds. On the BC dataset, number of nodes did not have either significant unconditional p value or p value conditioned on number of edges. This illustrate the unexpected fact, a topological
measure should not expected to have significance on all type of diseases because different diseases could have different mechanism and epistasis may only be involved in the mechanism of some of the diseases.

3. Size of largest connected component: On both datasets, the original SNP-SNP interaction network has significantly bigger largest connected component than the random networks in term of both unconditional p value and p value conditioned on number of nodes (also moderately significant in terms of p values conditioned on number of edges). However, the ranges of binarization thresholds on which the significance exists are different among the three p values. This suggests that the proper interpretation of the significance of a SNP-SNP interaction network should be clearly coupled with the conditions made for the significance estimation.

4. Clustering coefficient: On both datasets, while the significance of clustering coefficient in terms of all the three types of p values are moderate, it is worth noting that the p value conditioned on number of nodes is consistently lower than the p value conditioned on number of edges, which is consistently lower than the unconditional p value. This demonstrates the importance of the conditional analysis and also agrees with the between pathway model as discussed above, i.e. the original network tend to have SNP-SNP interactions that cluster together such that the neighbors of a node tend to connect to each other (the property captured by clustering coefficient).

5. Network efficiency: On both datasets, the significance of network efficiency are strong in terms of all the three types of p values, it is worth noting that the p value conditioned on number of nodes is consistently lower than the p value conditioned on number of edges and the unconditional p value. This demonstrates the importance of the conditional analysis and also agrees with the between pathway model as discussed above, i.e. the original network tend to have SNP-SNP interactions that cluster together such that average distance among all pairs of nodes are smaller (the property captured by network efficiency).
6.4 Significant Local Structures in Discriminative Pattern Network and their Domain Relevance

While the above exploration on the global topology demonstrate the significance of SNP-SNP interaction networks, further insights are needed to better understand the functional organization of disease-associated SNP-SNP interactions. In this section, we target the discovery of significant local network substructures that are biologically interesting, specifically the between pathway model that reflect compensations between pathways under genetic mutations.

6.4.1 Methods and experimental design

We developed a method (BridGE, Bridging Genes with Epistasis) for explicitly searching for larger structures of SNP-SNP interactions from GWAS datasets (i.e. between pathway models or BPMs, also used interchangeably with pathway-pathway interactions), guided by established sets of genes belonging to characterized pathways or gene modules. Specifically, although many pairs of loci do not have statistically significant interactions when considered separately, they can be collectively significant if they form a dense bipartite substructure between two functionally coherent sets of genes (Figure 6.1). Briefly, our approach involves the following main components (Figure 6.1, see Methods for details): (1) Quality control and balancing of population substructure between the cases and controls to avoid bias from population stratification [196]; (2) Clustering of loci in linkage disequilibrium (LD) that would otherwise result in trivial bipartite substructures. A subset of independent SNPs (denoted as S0) are selected by sampling from each LD block; (3) SNP-SNP interactions are scored using a information-theoretic measure that captures genetic interactions (see Methods for details), and a lenient threshold (global top q% quantile) is applied to derive a low-confidence, high-coverage interaction network. The results presented in the paper is based on q=1, while the stability of the proposed framework is tested over different q values in the supplementary material. (4) Pairs of pathways from predefined gene sets are tested for enrichment of SNP-SNP pair interactions with a likelihood ratio test. For the resulting log likelihood ratios (LLRs), we use positive and negative signs to differentiate the cases when two pathways are over and under enriched for top SNP-SNP interactions; (5) Both case-control permutations
and SNP-pathway assignment permutations [124] are used in a hybrid manner to assess the global significance of discovered pathway pairs and correct for multiple hypothesis testing. The entire pipeline is applied twice: the first time using only SNP pairs with higher frequency of joint minority homozygous genotypes in cases (to discover "type I" pathway-pathway interactions, i.e. associated with increased risks of disease) and the second time using only SNP pairs with higher frequency of joint minority homozygous genotypes in controls (to discover "type II" pathway-pathway interactions, i.e. associated with reduced risks of disease).

6.4.2 Experimental results

The ability of our approach to discover genetic interactions was tested by applying it to a genome-wide association study of Parkinson’s disease (PD) from the second phase of the NIA study [194] (denoted as NIA-II), consisting of 519 patients and 519 controls (individuals of European ancestry) after balancing the population substructure (see Methods section for the details of data processing). 758 gene sets representing established pathways or functional modules were derived by combining all human pathways from KEGG [197], Biocarta [198] and Reactome [199] and filtering based on set size (minimum: 10 genes, maximum: 200 genes). After permutation testing to establish global significance considering multiple hypothesis testing (see Methods for details), BridGE reported 48 significant type I pathway-pathway interactions (false discovery rate (FDR) < 0.05, i.e. less than three false positives) and 221 significant type II pathway-pathway interactions (false discovery rate (FDR) < 0.05, i.e. less than six false positives). Refer to Supplementary File 1 for the two lists of BPMs.

One interesting type I pathway-pathway interactions was between the KEGG TGF-signaling pathway and the KEGG Parkinson’s Disease gene set (Figure 6.7). We observed 802 SNP-SNP interactions across these two pathways, which was 1.4-fold higher than the expected number (585.6) given the 1% density of the SNP-SNP interaction network (a log likelihood ratio of 36.8 with a permutation based p value of $< 5e-6$, a permutation based FDR of 0.03). In contrast to the significance of this BPM, the large majority of SNPs mapped to these two pathways were not significant based on single-locus tests (Figure 6.7). Furthermore, none of the individual SNP-SNP interactions between the two pathways were significant when tested independently (minimum
FDR of 0.994). Thus, this significant pathway-pathway interaction would have been missed based on traditional univariate tests or epistasis tests that focus on individual SNP pairs. The KEGG Parkinson’s disease pathway involved in this BPM includes several key genes that have been individually associated with Parkinson’s disease: alpha-synuclein (SNCA), ubiquitin carboxy-terminal hydrolase L1 (UCHL1), parkin (PARK2), DJ-1 (PARK7), and PTEN-induced putative kinase 1 (PINK1). The other pathway involved in this interaction, KEGG TGF- signaling pathway is also related to PD, and the genetic interaction between the two pathways is plausible as supported by existing knowledge that they both can prevent neurodegeneration from two different molecule processes. Specifically, on one hand, alpha-synuclein in the K-PD pathway has been shown to function as a molecular chaperone in the formation of SNARE complexes that can prevent neurodegeneration. On the other hand, TGF-s are neuroprotective factors and organizers of injury responses and have been linked to neuronal phenotypes and neurodegeneration through genetic perturbation of mice. The original GWA study specifically noted no significant evidence for epistasis was observed. However, our approach identified the significant genetic interactions between a pathway that has been associated with Parkinson’s disease and another one with clear relevance. This significant interaction suggests that, epistasis between several SNPs linking these two processes may contribute to additional disease risk. Interestingly, this observation is consistent with findings in the yeast genetic interaction study, where it was reported that genes that participate in large numbers of genetic interactions tend to have phenotypic effects even as single mutants.

In addition to the BPMs that are clearly relevant to PD as we discovered earlier, there are several other interesting pathway-pathway interactions that are less obvious but still closely related to PD. For example, in the type I pathway-pathway network (Figure 6.8), another interaction involves the Parkin pathway in Biocarta and the systematic lupus erythematosus in KEGG. While the Parkin pathway containing 12 genes is a subset of the bigger KEGG Parkinson’s disease pathway and thus clearly related to PD, there is also plausible association between the systematic lupus erythematosus gene set (SLE) and PD. Specifically, SLE contains genes that are significantly up-regulated in the blood mononuclear cells from lupus patients compared to healthy persons, and
the immune system in lupus patients can harm different types of cells and tissues including the central nervous system. Other pathways in Figure 6.8 that are relevant to PD include: the ubiquitin mediated proteolysis pathway (its role in the pathogenesis of neurodegenerative disorder has been discussed in recent work), mitochondrial tRNA aminoacylation pathway (Aminoacyl tRNA synthetases have been linked to PD), fatty acid metabolism pathway (dietary intakes of fat has been associated with increased risk of PD) etc. It is worth noting that, there are also several pathways related to HIV in Figure 6.8. Although seemingly irrelevant, in fact, the neurologic manifestations of AIDS have been found to lead to neuropathology and neurodegeneration in human. Not only do many type I BPMs are relevant to PD, we also found some interesting type II pathway-pathway interactions (Figure 6.9). Specifically, an interesting interaction involves the IL1R pathway and the golgi associated vesicle biogenesis. On one hand, Interleukin-1 (IL-1) is a pro-inflammatory cytokine with important role in neurodenegenerative events and it signals primarily through the type 1 IL-1 receptor. On the other hand, defective sequestration of dopamine into vesicles is a key event in the demise of dopaminergic neurons in Parkinson’s disease. Therefore, the joint perturbations on the signal transduction through IL1R and vesicle biogenesis could possibly reduce the extent of neurodegeneration and thus reduce the risk of PD, which is consistent with the significant genetic interaction discovered between these two pathways.

Given the above overall summary and specific interpretations of the significant pathway-pathway interactions discovered from the NIA-II cohort, we further checked whether the same BPMs can be validated on two independent cohorts (all are Americans of European ancestry). Because of pathway overlaps, there are also redundancy within the discovered type I and type II BPMs. Such redundancy needs to be properly handled to avoid bias in the validation analysis. Refer to the Supplementary text for a detailed discussion and the method with which we selected 32 and 111 less-redundant BPMs from the sets of type I and type II BPMs for validation, respectively. The first validation cohort is from the NeuroGenetics Research Consortium (denoted as NGRC [200], 1947 cases and 1947 controls) and the second validation cohort is from the first phase of the NIA study [201] (denoted as NIA-I, 244 cases and 244 controls). For type I BPMs, likelihood ratio tests were performed for all 286,903 possible pairs of
the candidate pathways tested in NIA-II, now on the NGRC cohort and the NIA-I cohort, respectively. We designed two specific validation experiments for the 32 type I BPMs. The first experiment (denoted as validation test 1, green colored in Figures 6.8) estimates if the 32 type I BPMs discovered from NIA-II are enriched for BPMs with positive LLR values (over enrichment of top SNP-SNP interactions) in each validation cohort. The second experiment (denoted as validation test 2, blue colored in Figures 6.8 estimates if the 32 BPMs are further enriched for BPMs with significant LLR values (Bonferonni corrected likelihood ratio test p value < 0.05, with respect to the overall 32 tests) in each validation cohort. In the NGRC cohort, the 32 BPMs had a 1.7 fold enrichment for BPMs with positive LLRs (hypergeometric p value 1.69e-4) and a 4.1 fold enrichment for BPMs with significant LLRs (hypergeometric p value 2.41e-4). In the NIA-I cohort, the 32 BPMs had a 1.9 fold enrichment for BPMs with positive LLRs (hypergeometric p value 3.42e-7) but not significantly enriched for BPMs with significant LLRs (hypergeometric p value 0.09). This could be possibly due to the smaller sample size in NIA-I (488) compared to the sample sizes in NIA-II (1038) and NGRC (3894). Similar summary for the validation of the 111 type II BPMs is shown in Figure 6.9, which show that the 111 BPMs are significantly enriched for both BPMs with positive LLRs and also enriched for BPMs with significant LLRs, in both validation cohorts.

We further checked if the top-scoring SNP-SNP interaction pairs (top 1% quantile) in each type I (or type II) BPM discovered in NIA-II and validated in NGRC were similar between the two cohorts. As summarized in Supplementary Figure 6.11, the overlaps were significant (Bonferonni corrected Fisher’s exact test p-value < 0.05) for most BPMs but small (Jaccard similarity ranges from 0.1 to 0.18). This suggests that the same pathway-pathway interaction may be reflected by different sets of SNP-SNP interactions in different cohorts, or alternatively, that the power for reliably pinpointing specific locus pairs is simply too low (a key motivation of this study). In either case, our results suggest that if we hope to understand genetic interactions based on currently typical cohort sizes in GWA studies, our best hope is to analyze such combinations at pathway-pathway or module-module level rather than at the level of between specific genome loci. The BPMs discovered by our approach serve as a starting point for dissecting which variants and combinations may functionally contribute to disease risk.
6.5 Discussion

To our knowledge, this work is the first reported systematic approach for discovering human disease-specific pathway-pathway genetic interactions from genome-wide association data. The key motivation for the approach, derived from the extensive analysis of genetic interaction networks in yeast, is that genetic interactions tend to occur between functionally compensatory modules rather than between isolated pairs of genes. Therefore, we hypothesized that pairwise SNP interactions tend to cluster into larger substructures and designed a computational framework for detecting significant pathway-pathway genetic interactions. Results on a Parkinson’s disease GWAS dataset confirmed that larger genetic interaction structures indeed exist and can be discovered by BridGE, and we showed that a significant number of these pathway-pathway genetic interactions were confirmed in two independent cohorts. We believe our method generalizes beyond the Parkinson’s disease context, although its effectiveness will depend on the disease of interest and the size and nature of the specific cohort. Several directions should be explored in future work. First, we used a specific measure for measuring SNP-SNP interactions and there are many other possibilities [146, 19]. Further study of the effect of using different measures within the proposed framework would be worthwhile. Second, in this study, we focus on just one type of genetic interaction substructure, between-pathway models, while other local structures are also expected and could also provide interesting results (e.g. within-pathway structures [202, 164]). Third, the proposed framework currently depends on a set of pre-defined gene sets which may prevent the discovery of genetic interactions between sets of genes that are not part of well-characterized pathways. Developing methods that could relax the dependency on known pathways by leveraging other genomic data (e.g. protein-protein or protein-DNA interactions, gene co-expression) while still constraining the search to expected network structures would be a fruitful direction.
6.6 Supplementary Methods

6.6.1 GWAS SNP datasets, data quality control and balancing population stratification

Three case-control SNP datasets are used in this study, all from genome-wide association studies on Parkinson's disease. The cohort denoted as NIA-II is from [194] (dbGaP Study Accession: phs000003.v1.p1, platform ILLUMINA HumanHap550v3.0, 677 cases, 538 controls, 561466 SNPs). The cohort denoted as NGRC (2000 cases, 1986 controls, 1012895 SNPs) is part of the study [200] (dbGaP Study Accession: phs000196.v2.p1, platform: ILLUMINA HumanOmni1 Quad v1-0 B). The cohort denoted as NIA-I is from [201] (dbGaP Study Accession: phs000003.v1.p1, platform: Illumina Infinium I and HumanHap300, 270 cases, 271 controls, 408000 SNPs).

Data quality control was performed with PLINK [134] with the following parameters for each of the three datasets: 0.02 as the maximum genotyping missingness for each sample (--mind), 0.01 as the maximal genotyping missingness (--geno), 0.05 as the minimum minor allele frequency (--maf) and $1 \times 10^{-6}$ as the threshold for controlling Hardy-Weinberg equilibrium (--hwe) for each SNP. After these filtering, 291015, 730288 and 357235 SNPs are left in the three datasets (NIA-II, NGRC and NIA-I), respectively. Then, we took the 159270 SNPs that are common among the three datasets, and finally selected 49014 SNPs that can be mapped (a SNP is mapped to all the genes that overlap with a $+/-50\, \text{kb}$ window around the SNP) to at least one of the 6744 genes in the collection of pathways to be used in the pathway-pathway interaction search.

Only Caucasian subjects were considered and the subjects with age of onset below 20 were also removed as done in recent meta analysis. Duplicate subjects were kept in just one cohort in the following order: NIA-I, NIA-II and NGRC so that we can keep as many samples as possible for the smaller cohorts. Furthermore, the cohort in a case-control GWAS study may have imbalanced proportions of race or gender, which could lead to spurious allelic associations as discussed by [176, 177]. Even for the studies that focus on a single race and only consider autosomal SNPs, spurious allelic associations can still be discovered due to unknown population substructure [196]. Therefore, recent GWAS analyses have suggested the use of a special procedure to ensure balanced population structure between cases and controls [134]. Specifically, all
the subjects are clustered into groups of size 2, each containing one case and one control that are from the same sub-population (based on pairwise identity-by-state distance and the corresponding statistical test). We use the implemented procedure in PLINK [134]. After removing low-quality subjects and balancing population stratification, there are 519, 1947 and 244 pairs of cases and controls in the three cohorts (NIA-II, NGRC and NIA-I), respectively.

6.6.2 Estimating SNP-SNP genetic interactions

We use \{MM, Mm, mm\} to denote the three genotypes of each SNP, i.e., majority homozygous, heterozygous and minority homozygous, respectively. We assume allele \(M\) shows dominance to allele \(m\) and use a binary encoding for each SNP, i.e. \(mm\) genotypes being 1’s and non-\(mm\) genotypes being 0’s. There are four possible genotype combinations for a pair of SNPs \((S_i, S_j)\): 11, 10, 01 and 00. Mutual information can be calculated between each of these four genotype combinations and the binary case-control indicator variable \(C\), i.e.

\[
MI((S_i, S_j)_{11}, C), MI((S_i, S_j)_{01}, C), MI((S_i, S_j)_{10}, C) \text{ and } MI((S_i, S_j)_{00}, C).
\]

After normalization by \(H(C)\) (the entropy of \(C\)), these four mutual information values range from −1 to +1. We define the following measure \(SSI\) (SNP-SNP Interaction) between \(S_i\) and \(S_j\):

\[
SSI_C(S_i, S_j) = MI((S_i, S_j)_{11}, C) - \max \left\{ F(MI((S_i, S_j)_{01}, C)), F(MI((S_i, S_j)_{10}, C)), F(MI((S_i, S_j)_{00}, C)) \right\}
\]

where

\[
F(MI((S_i, S_j)_{a_1a_2}, C)) = \begin{cases} 
MI((S_i, S_j)_{a_1a_2}, C), & \text{if criterion } D \text{ is met} \\
0, & \text{otherwise}
\end{cases}
\]

Criterion \(D\) is met if the \(a_1a_2\) genotype combination of the SNP pair \(S_i\) and \(S_j\) has the same direction of differentiation as 11, i.e. they are both more frequent in cases or both more frequent in controls. The interpretation of a positive \(SSI\) value, e.g. 0.2, is that the 11 genotype of the two SNPs provide 20% more information about the disease status than the maximal information provided by any of the other three genotypes.
with the same direction of differentiation. Thus, the larger $SSI$ a SNP pair has, the more information the 11 genotype provides about the disease status than the other three genotypes. Although there are several different possible measures for statistical epistasis [152, 146] as reviewed in [19, 17]. We chose the above $SSI$ measure because it explicitly captures the interaction between the minority homozygous genotypes of two SNPs. This directly corresponds to the between pathway model we assumed in the paper as illustrated in Figure 6.1. It will be worthwhile exploring the use of other measures for genetic interaction in BriGE in future work.

### 6.6.3 Two possible directions for handling redundant SNP-SNP interactions due to linkage disequilibrium

As motivated in the main text, bias from linkage disequilibrium (LD) needs to be removed before pathway-pathway enrichment of SNP-SNP interactions. Two directions can be pursued towards this goal: 1) removing the LD bias before calculating pairwise SNP-SNP interactions; and 2) removing the LD bias after calculating pairwise SNP-SNP interactions.

The first direction is more likely to miss informative SNP-SNP interactions than the second direction because it only considers a subset of all the SNPs selected based on SNP-SNP correlation. However, the first direction has much better scalability and computational efficiency given that the target datasets are genome-wide case-control SNP datasets. It is worth noting that a biclustering algorithm pursuing the second direction was designed in [192] to summarize a yeast SNP-SNP interaction network into an LD-LD network. The algorithm takes the SNP-SNP interaction matrix as an input and searches for two sets of SNPs that are consecutive in terms of genome location on a single chromosome and have a statistically significant number of across-set SNP-SNP interactions based on a hypergeometric test. The algorithm was applied on a yeast SNP-SNP interaction network (originally constructed in [146]) with 1977 SNPs where LD effect was assumed to not span more than 60 SNPs for computational reasons.

1 This implies that it is computationally infeasible to remove the LD bias after

---

1 We applied the algorithm (obtained from the authors of [192]) to the human SNP datasets used in this paper and observed that the algorithm could handle about 1500 SNPs with a threshold of $\delta$ below 60) but not beyond. For example, on a data set with 2000 SNPs, the program did not finish in two days with a $\delta = 100$. 

calculating pairwise SNP-SNP interactions, given that the LD blocks in human GWAS datasets could span thousands of SNPs. Therefore, in this paper, we pursue the first direction.

6.6.4 Selecting a less-redundant set of SNPs

We used a procedure in PLINK [134] to select a subset of less-redundant SNPs from a GWAS dataset, i.e. –indep-pairwise 50 5 0.1. Specified with this setup, PLINK searches each window of 50 SNPs with a sliding step of 5 SNPs, and select a subset of SNPs with pairwise $r^2$ below 0.1 within each sliding window. After this procedure, 14528 SNPs were left in $S_0$. As double check, we found that the highest $r^2$ between any pair of SNPs within any window of 1Mb is 0.16, demonstrating that the LD has been effectively controlled within each chromosome. By using a stringent $r^2$ threshold of 0.1, we may have ignored some informative SNPs. However, we decided to take this conservative approach to minimize the bias from LD structures in this initial step towards discovering disease-specific genome-wide pathway-pathway genetic interactions.

6.6.5 Calculating all pairwise SNP-SNP interactions for the SNPs in $S_0$

Some existing work, such as [146] proposes only computing interactions for the SNP pairs with sufficiently high marginal effects in order to enhance statistical power. In contrast, we chose to compute $SSI$ for all pairs of SNPs and create the entire networks, because we do not want to risk missing SNP pairs that have weak (or even no) marginal effect but a strong combined effect [150]. More importantly, as discussed in the introduction, the highlight of this paper is how to relax the requirement of statistical significance of individual edges for discovering module-level genetic interactions that are missed by existing approaches.

6.6.6 Estimating pathway-pathway interaction based on SNP-SNP interaction network: a conceptual view

Essentially, for each pair of pathways, we want to test if the number of SNP-SNP interactions between them is significantly higher than expected due to random chance
(with respect to the density of the SNP-SNP network $N_{s_0}$: q%). Conceptually, this is related to the existing approaches designed for gene set enrichment analysis (GSEA) originally applied on case-control gene expression datasets [203, 45] and then adapted to GWAS SNP datasets [30, 124]. However, there are two key differences. First, traditional approaches for GSEA start from univariate statistics of genes or SNPs, while our approach is built up on non-additive interactions between pairs of SNPs that could have little or no single locus association with a disease phenotype. Second, approaches for GSEA target the enrichment of single gene/SNP associations in each individual pathway while our approach explores the enrichment of SNP-SNP interactions across each pair of pathways. Beyond the conceptual difference, the pairwise pathway enrichment analysis based on SNP-SNP interactions is much more computationally challenging, and thus we choose to binarize the SSI values (top q% quantile) into binary interactions to make follow up computation efficient and scalable. Specifically, we use a likelihood ratio test for the enrichment analysis (next paragraph).

### 6.6.7 Estimating pathway-pathway interaction based on SNP-SNP interaction network: technical details

Given a collection of pathways $P$, we first map the SNPs in $S_0$ to each pathway by first mapping SNPs to genes (a SNP is mapped to all the genes that overlap with a $+/-50kb$ window around the SNP). Then, we map genes to pathways based on gene-pathway membership in the collection of 758 pathways with sizes ranging from 10 to 200. For each pair of pathways in $P$, say $P_i$ and $P_j$, we first remove the overlapping SNPs. Then, we denote the number of unique SNP-SNP pairs between $P_i$ and $P_j$ as $n_{ij}$ and the number unique SNP-SNP interactions with respect to $N_{s_0}$ (the k%-density binary SNP-SNP interaction network) as $k_{ij}$. We can then test if the number of observed interaction between $P_i$ and $P_j$ is significantly higher than the expected number (calculated based on the global density of the SNP-SNP interaction network and the sizes of the $P_i$ and $P_j$). A straightforward way is to use a hypergeometric test.

However, the hypergeometric test can be very time consuming when the latter two numbers are very large. Therefore, we use an alternative approach based on a standard likelihood ratio test [204] that is much more computationally efficient. Specifically, we assume the number of unique SNP-SNP interactions among a set of SNP pairs follow a
binomial distribution with rate of \( r \). We want to test if the rate for the set of SNP pairs between \( P_i \) and \( P_j \) is significantly different than the set of SNP pairs in \( S_0 \) excluding those between \( P_i \) and \( P_j \). This can be tested with a standard likelihood ratio test in which we first estimate a maximal likelihood (\( L_0 \)) assuming the two rates are the same, i.e., all the SNP pairs follows a common rate of \( r_0 \). We then estimate a maximal likelihood (\( L_1 \)) assuming the two rates are different. The likelihood ratio \( \frac{L_1}{L_0} \) captures the difference between the two rates, and can be calculated much more efficiently with matrix operation.

We calculate a log likelihood ratio (LLR) for all pairwise pathways in \( P \) and denote this pathway-pathway matrix as \( LLR_0 \). Because we are specifically interested in the pathway pairs that have significantly more SNP-SNP interactions compared to the background, we flip the LLRs of the pathway pairs with smaller numbers of SNP-SNP interactions than expected to their corresponding negative values, to differentiate over- and under-enrichment. Although these LLRs can be directly mapped to \( p \) values, we choose to estimate the \( p \) values empirically based on the null distribution from the permutation tests (see next paragraph) in order to have an unbiased estimation.

6.6.8 Correcting for multiple hypothesis testing

Because a huge number of pathway pairs are tested in the search of significant BPMs, corrections are needed for multiple hypothesis tests. SNP-pathway membership permutation and sample permutation are two common strategies for estimating empirical null distribution in pathway-based analysis of genome-wide association data [124]. In SNP-pathway membership permutation, the mapping between SNPs and pathways are permuted while fixing the number of SNPs mapped to each pathway. In sample permutation, the original case-control groups are randomly shuffled over the entire set of samples in a dataset. SNP-pathway membership permutation has the advantage of being much more computational efficient because the SNP-SNP interaction network constructed with the case-control SNP dataset stays unchanged. In contrast, sample permutation is much more consuming because SNP-SNP interaction network needs to be recomputed every time the case-control grouping is permuted. Although having better computational efficiency, SNP-pathway membership permutation has its own caveat
as widely discussed in existing work [124]: while it maintains the size of each pathway, it might break the LD structure within each pathway and introduce bias in the null distribution estimation. However, because a subset of LD-reduced SNPs have been selected with a stringent $r^2$ threshold of 0.1 before the search of pathway-pathway interactions, the bias introduced by SNP-pathway membership permutation is expected to be low in the context of this work. In fact, several studies have demonstrated the use of SNP-pathway membership shuffling with an even less stringent $r^2$ threshold of 0.2 [124].

In this paper, we designed a hybrid permutation approach. First, we did a large number of SNP-pathway membership permutations leveraging its better efficiency (200000 times). After each permutation, the exact same pipeline for pathway-pathway interaction estimation discovery was repeated and an LLR is computed for each pair of pathways. The number of times a BPM's LLR got exceeded by its corresponding permutation LLRs is maintained over the 200000 permutations. With these counts, p values and false discovery rates (FDR) can be estimated with the method described in [45]. We used an FDR cutoff of 0.05 to select significant BPMs based on SNP-pathway membership permutation: 48 type I BPMs and 221 type II BPMs were discovered. Then, 1000 sample permutations were applied only to the 48 and 221 BPMs respectively and sample permutation based FDRs are estimated. We found that all the significant BPMs discovered by SNP-pathway permutation remain significant based on sample permutation with respect to the same FDR threshold 0.05. The sample permutation can be considered as a post-discovery sanity check to filter out BPMs that were claimed to be significant by SNP-pathway permutation simply due to the bias introduced when within-pathway LD structure was broken. The hybrid permutation strategy described here led to the 48 type I and 221 type II BPMs presented in the main text.

6.6.9 Properly handle overlaps among a set of BPMs

In addition to the interaction between TGF-β pathway and the Parkinson's disease pathway discussed in the main text, BridGE also revealed another significant interaction involving the TGF-β pathway, i.e. with another interacting partner, the Biocarta Parkin pathway (Figure 6.8). In fact, nine of the twelve genes in the Biocarta Parkin pathway overlap with the 133 genes in the KEGG Parkinson's disease pathway (hypergeometric
distribution based p value < 6.6e−13). Thus, the Biocarta Parkin pathway is essentially a subset of the KEGG Parkinson’s disease pathway. This illustrates the existence of redundancy between BPMs due to pathway overlaps and it needs to be considered when we interpret a set of significant BPMs. Specifically, in the validation analysis, redundancy in a set of BPMs would introduce bias when testing if the number of BPMs that can be validated in an independent cohort is significant.

We chose to interpret pathway-pathway overlaps together with the detected BPMs (post discovery) instead of selecting a subset of pathways with less overlap beforehand, because the latter approach risk missing pathways involved in important functional compensations that cannot be well represented by other overlapping pathways. Although the overlaps across different pathways imply the dependency among different BPMs being tested, the use of SNP-sample permutation tests and sample permutation tests (as will be described in the next paragraph) preserves the dependency across different permutations and thus can provide a reliable estimation of false discovery rates.

For the 48 type I BPMs, we first calculated a pairwise overlap matrix, where the overlap is defined as the number of common SNP-SNP interactions between two BPMs divided by the smaller one of the number of SNP-SNP interactions in the two BPMs. Supplementary Figure 6.10 show the pairwise overlaps among type I and type II BPMs, respectively, where the rows and columns are ordered by hierarchical clustering so that similar rows and columns are grouped together. We binarize the overlap matrix with a threshold of 0.5, i.e. if two BPMs have overlap greater than 0.5, they are considered being redundant with each other. The binarized matrix corresponds to a network with 32 connected components (refer to Supplementary File 2 for the component each BPM is assigned to). For each connected component containing more than one BPMs, we randomly chose one BPM as the representative in the validation experiments and got the 32 less-redundant BPMs listed in Supplementary File 2. Similarly, 111 less-redundant BPMs were selected from the 221 type II BPMs.

6.6.10 Stability of the proposed framework over different parameters and insights on parameter selection

The proposed framework requires a user-specified parameter, i.e. the quantile threshold $q$ for binarizing the SNP-SNP SSI matrix into an interaction network. The results
discussed in the main text are based on \( q = 1\% \). Here, we perform a consistency check across three different values for \( q \) (0.5%, 1% and 2%). The results summarized in Supplementary Figure 6.13 show that, while variations exit as expected, the LLR p values for the same BPM are highly correlated with each other across the three \( q \) values (scatter). The ideal threshold for \( q \) should be a value that include as many true SNP-SNP interactions as possible while simultaneously excluding as many false SNP-SNP interactions as possible. This is not practical because the true SNP-SNP interactions are unknown. A too small \( q \) may filter out many true SNP-SNP interactions that do not rank high enough and a too large \( q \) may include many false SNP-SNP interactions. Both cases could affect the enrichment of top SNP-SNP interaction for a BPM. Another complexity is that different complex diseases may have different fractions of genetic heretability coming from genetic interaction, making it hard to provide a fixed suggestion on setting a threshold for \( q \). In practice, we suggest trying several different values and then checking which \( q \) leads to the most significant set of BPMs. This certainly adds another level of multiple hypothesis testing and requires proper adjustment. However, a relatively high FDR cutoff (e.g. 0.25 and 0.33) could still be of interest after adjustment for testing different \( q \) thresholds, because the ultimate validation would still be from independent datasets.
Figure 6.7: Detailed information for a significant BPM discovered from the NIA-II cohort. The center of the figure is a bipartite plot two sets of SNPs mapped to the two pathways whose names are shown on the two sides of the plot, respectively. The two groups of horizontal bars (color coded to indicate chromosome) show the \(-\log_{10} p\) values (single locus $\chi^2$ test) for the two sets of SNPs. The two dashed lines show the Bonferroni corrected 0.05 p value cutoff, indicating that no single SNPs are significant. Furthermore, none of the SNP-SNP interactions between the two sets of SNPs (plink lines, i.e. those above the global top 1% quantile) are significant (minimum FDR 0.994). The expected and observed number of SNP-SNP interactions between the two pathways and the associated significance are shown on the top of the figure.
Figure 6.8: Type I BPMs discovered from the NIA-II cohort and the validation results in the NGRC and NIA-I cohorts. (a) Network representation of the 48 significant pathway-pathway interactions (FDR < 0.05, less than three false positives) discovered from the NIA-II cohort. Each node shows the pathway name with the letter in the bracket indicating the source of the pathway (K: KEGG, R: Reactome, B: Biocarta), and each edge indicates a pathway-pathway interaction (i.e. a BPM). The color of an edge indicates if it is validated in NGRC, NIA-I or neither of the two, as specified on the top of the subfigure. No type I BPM was validated in both NGRC and NIA-I. (b and c) Venn diagrams summarizing the validation of the 32 less-redundant type I BPMs (refer to the main text and supplementary text for details) in the NGRC cohort and the NIA-I cohort, respectively. The green and blue colored ellipses and texts indicate the two validation tests (1 and 2) checking if the set of 32 BPMs discovered from NIA-II are enriched for positive LLR values (test 1) and significant LLR values (test 2) in each of the two validation cohorts. The significance of the enrichment based on hypergeometric tests are given in the texts: numbers in the brackets are the fractions for the background (i.e. all possible pathway pairs) and the discovered (the 32 significant BPMs).
Figure 6.9: Type II BPMs discovered from the NIA-II cohort and the validation in the NGRC and NIA-I cohorts. Only the interactions that are validated in at least one of the two independent cohorts are shown in this figure. The complete network is shown in Supplementary Figure 6.12. Refer to the caption for Figure 6.8. Additional description for (a): orange edges indicate the BPMs that are validated in both the NGRC and NIA-I cohorts.
Figure 6.10: Overlaps among the 48 type I BPMs (a) and among the 221 type II BPMs (b) discovered from the NIA-II cohort. The rows and columns are ordered by hierarchical clustering so that similar rows and columns are grouped together, using the same order shown in Supplementary File 2.
Figure 6.11: Overlaps of top-scoring SNP-SNP interaction pairs (top 1% quantile) between the NIA-II cohort and the NGRC cohort, for each BPM discovered in NIA-II and validated in NGRC: type I BPMs (a) and type II BPMs (b), as measured by Fisher’s exact test p-value (x-axis) and Jaccard similarity (y-axis). Similar summary of overlaps between NIA-II and NIA-I for type I and type II BPMs are shown in (c) and (d), respectively.
Figure 6.12: The complete set of Type II BPs discovered from the NIA-II cohort organized as a pathway-pathway network. Pathway names are only shown for the eight hub nodes (in the green box), while the pathway names for the other nodes can be found by looking up the node ID in Supplementary File 1.
Figure 6.13: Consistency of log likelihood ratios for all the 286903 BPMs tested in the NIA-II cohort between different quantile thresholds. (a) type I BPM, q = 1% (y-axis) vs. q = 0.5% (x-axis); (b) type I BPM, q = 1% (y-axis) vs. q = 2% (x-axis); (c) type II BPM, q = 1% (y-axis) vs. q = 0.5% (x-axis); (d) type II BPM, q = 1% (y-axis) vs. q = 2% (x-axis).
Chapter 7

Conclusion and Discussion

This thesis targeted the computational discovery of combinatorial disease biomarkers associated with complex human diseases from a variety of large scale case control genomic datasets. It took a data mining approach, specifically discriminative pattern mining, that are inherently designed for efficient combinatorial search and subgroup discovery. It addressed several key challenges confronted by existing techniques when handling biomedical and genomic datasets, namely, computational efficiency, statistical power, biological relevance, systems-level interpretation and data integration. It also proposed novel frameworks for network-level analysis of discriminative pattern that can provide novel insights from both global and local scale that go beyond those provided by individual patterns. Validations across independent datasets not only demonstrate the reliability of the proposed approaches in this thesis but also lead to several interesting biological insights on several complex diseases such as breast cancer and Parkinson’s disease. The data-mining algorithmic contributions in this thesis also hold promise for addressing generic challenges in other domains beyond biology.

Several directions should be explored in future work. First, there is critical need and potential benefit from in-depth comparison and understanding of different types of statistical measures for capturing genetic interaction from population level data. Second, different types of combinatorial biomarkers and the corresponding local network structures should also be explored and could provide additional insights that can broadening the scope of this thesis. Third, both Chapters 4 and 6 talked about frameworks
that currently depends on a set of pre-defined gene sets which may prevent the discovery of genetic interactions between sets of genes that are not part of well-characterized pathways. Developing methods that could relax the dependency on known pathways by leveraging other genomic data (e.g. protein-protein or protein-DNA interactions, gene co-expression) while still constraining the search to expected network structures would be a promising direction. Last but not least, in order to better understand the complex mechanisms of human diseases, additional types of genomic data should be further integrated, including proteomic data, metabolic data, second and third-generation sequencing data, as well as the data from other clinical domain such as neuroimaging data, electronic medical records etc.
References


[12] Hong Cheng, Xifeng Yan, Jiawei Han, and Chih-Wei Hsu. Discriminative frequent pattern analysis for effective classification. In *Proceedings of International Conference on Data Engineering*, pages 716–725, 2007.


[28] Wei Fan, Kun Zhang, Hong Cheng, Jing Gao, Xifeng Yan, Jiawei Han, Philip S. Yu, and Olivier Verscheure. Direct mining of discriminative and essential graphical and itemset features via model-based search tree. In Proceeding of the ACM SIGKDD international conference on knowledge discovery in databases, pages 230–238, 2008.

Proceedings of the 15th ACM SIGKDD international conference on Knowledge

genomewide association studies. The American Journal of Human Genetics,

[31] E. Segal, N. Friedman, N. Kaminski, A. Regev, and D. Koller. From signatures
to models: understanding cancer using microarrays. Nature Genetics, pages S38–
S45, 2005.

[32] Petra Kralj, Nada Lavrac, Dragan Gamberger, and Antonija Krstacic. Contrast
set mining for distinguishing between similar diseases. In AI in Medicine in Europe

ology for contrast set mining through subgroup discovery. Journal of Biomedical

[34] H. Xiong, P.N. Tan, and V. Kumar. Hyperclique pattern discovery. Data Mining

[35] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule min-
ing. In Proceeding of the ACM SIGKDD International Conference on Knowledge

[36] W. Li, J. Han, and J. Pei. CMAR: Accurate and efficient classification based on
multiple class-association rules. In Proceedings of the IEEE International Confer-

[37] X. Yin and J. Han. CPAR: Classification based on predictive association rules. In
Proceedings of the SIAM International Conference on Data Mining, pages 331–

[38] G. Cong, K.L. Tan, A.K.H. Tung, and X. Xu. Mining top-k covering rule groups
for gene expression data. In Proceedings of the 2005 ACM SIGMOD internationa


[67] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15:55–86, 2007.


