

論文 / 著書情報
Article / Book Information

題目(和文)	組合せ効果の発見に向けた多重検定補正と生命科学への応用
Title(English)	Multiple Testing Correction for Discovering Combinatorial Effects with Application to Biology
著者(和文)	寺田愛花
Author(English)	Aika Terada
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9559号, 授与年月日:2014年3月26日, 学位の種別:課程博士, 審査員:瀬々 潤,佐藤 泰介,秋山 泰,関嶋 政和,杉山 将
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9559号, Conferred date:2014/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	要約
Type(English)	Outline

Multiple Testing Correction for Discovering Combinatorial Effects with Application to Biology

Aika Terada



Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

February 2014

Submitted in partial fulfillment of
the requirements for the degree of
Doctor of Engineering

Abstract

To make advancement in biology, it is essential to discover collaborative effects, such as interactions between transcription factors and multiple DNA mutations. However, few combinations have been discovered with biological experiments because investigation of the higher-order effects needs to check intractable number of combinations. For example, when combinations of up to k items are considered, the number of investigated combinations increases exponentially to k . The scanning of the large number of combinations may be possible with recent computational resources but extremely unlikely due to multiple testing correction.

In this thesis, we first show how existing multiple testing procedures, such as Bonferroni correction and Westfall-Young procedure, hide the higher-order significant combinations that have already been discovered experimentally. Bonferroni correction is classical procedure but widely used in practical, however, it is known that Bonferroni correction is conservative when tests have a high correlation to each other. In combinatorial effects discovery, most tests depend on one another. Therefore, it is hopeless to detect these effects using Bonferroni correction. Unlike Bonferroni correction, Westfall-Young procedure considers the dependence among tests by estimation of the distribution of statistical score with permutation test. Westfall-Young procedure has a higher detection power than Bonferroni correction, however, it requires intractable amount of computational time even when a small number of tests are performed. In combinatorial effect discovery, it is difficult to apply Westfall-Young procedure because a huge number of tests are performed. To overcome the problems, we propose two multiple testing procedures, called *Limitless Arity Multiple-testing Procedure (LAMP)* and *Fast Westfall-Young procedure (FastWY)*. Both procedures control the Family-Wise Error Rate (FWER), which is the probability that at least one false discovery occurs.

LAMP is an algorithm to improve the sensitivity of Bonferroni correction. The problem with the Bonferroni correction is that the bound on FWER is proportional to the number of tests with the assumption that all of the tests increase the FWER. However, not every test increases the FWER and the property can be used to improve the sensitivity of Bonferroni correction. By using an efficient branch-and-bound algorithm, LAMP counts the exact number of testable combinations and calibrates the

Bonferroni factor to the smallest possible value. In the yeast and human breast cancer transcriptome datasets, whereas LAMP discovers statistically significant combinations of as many as four and eight binding motifs, respectively, Bonferroni correction cannot detect both combinations.

FastWY is an accelerated algorithm of Westfall-Young procedure which estimates the null distribution via a permutation test. To estimate the null distribution, Westfall-Young procedure only uses the minimum P-value on each permuted dataset. Therefore, we propose an efficient branch-and-bound algorithm, FastWY, which is expanded procedure of LAMP to find the minimum P-value in a permuted dataset. Our demonstrations using yeast transcriptome datasets show that FastWY showed the significance of the combination less than 30 minutes. In contrast, Westfall-Young procedure requires more than 100 hours even if the combination size is limited up to four motifs.

Both LAMP and FastWY greatly contribute to combinatorial effect discovery because there have been no multiple testing procedure that can consider higher-order effects. The statistically significant combinations found by our procedures can motivate biologists to confirm whether these combinations function in cells.

Acknowledgments

I express my deepest gratitude to my academic supervisor, Associate Professor Jun Sese, who has supervised my work for six years. He has provided valuable advice and support to me throughout my work. My experience under him will become the basis for my future life for as a scientist. I also warmly thank Dr. Koji Tsuda and Dr. Mariko Okada-Hatakeyama. Dr. Koji Tsuda gave me incisive suggestions and advice that were indispensable for this thesis. Dr. Mariko Okada-Hatakeyama supported the results of this study with biological knowledge and experiments. I also thank the members of my dissertation committee, Professor Yutaka Akiyama, Professor Taisuke Sato, Associate Professor Masakazu Sekijima, and Associate Professor Masashi Sugiyama for extremely helpful comments regarding this dissertation.

Furthermore, I appreciate the wonderful days spent with all of the members of Sese Laboratory. I am especially thankful for working with Ms. Akiko Izawa and Ms. Megumi Yamada. I have had a cheerful time with them even after we moved to Tokyo Institute of Technology from Ochanomizu University. I am also grateful to Dr. Satoru Akama for the frequent and fruitful discussions. Mrs. Setsuko Mizoguchi, who is the secretary of my laboratory, has helped me tremendously in completing my paperwork. I will never forget the pleasant time with these members who provided me with moral support.

My research projects were financially supported by Minato Discrete Structure Manipulation System Project and Japan Society for the Promotion of Science (JSPS) Research Fellowships for Young Scientists. Without these supports, I could not have afforded to work my study and to live a life.

Finally, but most importantly, I would like to offer heartfelt thanks to my dearest parents. They have made innumerable contributions to my wellbeing throughout my life, and I feel deep gratitude for their genuine love. Their thoughtfulness and kindness will always be remembered.

Contents

1	Introduction	3
1.1	Background	3
1.2	Overview of Multiple Testing Procedures	4
1.2.1	Theoretically FWER Controlling Procedures	5
1.2.2	Empirically FWER Controlling Procedures	7
1.2.3	Theoretically FDR Controlling Procedures	8
1.2.4	Empirically FDR Controlling Procedures	9
1.3	Contributions of This Thesis	9
2	Combinatorial Discovery and Multiple Testing Procedure	11
2.1	Introduction	11
2.2	Significant Combination	11
2.3	Procedures to Control Family-Wise Error Rate	12
2.3.1	Bonferroni Correction	13
2.3.2	Holm Procedure	14
2.3.3	Westfall and Young Procedure	15
2.4	Procedures to Control False Discovery Rate	16
2.4.1	Benjamini and Hochberg procedure	18
2.5	Performance Evaluation for Combinatorial Regulation Discovery . . .	18
2.5.1	Experimental Datasets	19
2.5.2	Sensitivities in Yeast Transcriptome Dataset	21
2.5.3	Sensitivities in Human Breast Cancer Transcriptome	23
2.6	Conclusion	25
3	Thoretical Multiple Testing Procedure for Combinatorial Effects Discovery	33
3.1	Introduction	33
3.2	Properties to Count Testable Combinations Efficiently	34
3.2.1	Lower Bound and Monotonicity	36
3.2.2	Enumeration of Testable Combinations with a FIM Algorithm	40

3.3	LAMP Algorithm	41
3.4	Time Complexity of LAMP	44
3.5	Experimental Results	45
3.5.1	Comparison of Detection Powers	46
3.5.2	Comparing Power to Control the Ratio of False Discoveries . .	49
3.5.3	Running Time	50
3.6	Biological Confirmation of Detected Motif Combinations	51
3.6.1	Yeast Transcriptome dataset	51
3.6.2	Human Breast Cancer Cells Dataset	63
3.7	Conclusion	67
4	Empirical Multiple Testing Procedure for Combinatorial Effects Discovery	71
4.1	Introduction	71
4.2	Strategy to Efficiently Find the Minimum P-value	72
4.3	Algorithm of FastWY	73
4.3.1	Calculation for a Permuted Dataset	73
4.3.2	Enumerating the Significant Combinations	74
4.3.3	Time Complexity of FastWY	75
4.4	Performance Comparison between FastWY and WY-procedure	76
4.4.1	Datasets and Experimental Settings	76
4.4.2	Performance upon Increasing the Number of Items	77
4.4.3	Performance upon Increasing the Number of Samples	77
4.5	Running Time to Find an Optimal λ	78
4.6	Biological Confirmation of Detected Combinations	79
4.7	Conclusion	88
5	Conclusion and Future Work	89
5.1	Conclusion	89
5.2	Future Work	90
	Bibliography	91

List of Figures

1.1	Illustration of a multiple testing procedure.	5
2.1	Illustration of a salient three-item combination.	12
2.2	Illustration of how Bonferroni correction computes the upper bound of FWER.	14
2.3	Illustration of how Holm procedure computes the upper bound of FWER when k -th smallest P-value is less than or equal to δ	15
2.4	Illustration of the association between FWER and the probability distribution of the minimum P-value.	17
2.5	Sample ratios in which at least one motif combination was deemed significant over all samples in yeast dataset. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5	27
2.6	The number of motif combinations was deemed significant in 173 yeast stress environments of yeast dataset. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5	28
2.7	Sample ratios in which at least one motif combination was deemed significant over all samples in 27 EGF-induced human breast cancer cell samples. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5	29
2.8	The number of motif combinations was deemed significant in 27 EGF-induced human breast cancer cell samples. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5	30
2.9	Sample ratios in which at least one motif combination was deemed significant over all samples in 28 HRG-induced human breast cancer cell samples. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5	31
2.10	The number of motif combinations was deemed significant in 28 HRG-induced human breast cancer cell samples. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5	32

3.1	Relationships between a threshold δ , the minimum P-value $f(x)$ for each combination, testable and untestable combinations, and the FWER bound.	35
3.2	Illustration of the properties of testable combinations.	37
3.3	Illustration of the relationship between δ and $g(\delta)$	42
3.4	Illustration of the algorithm of the LAMP.	44
3.5	Time complexity of LAMP when Fisher's exact test is used.	45
3.6	Sample ratios in which at least one motif combination is significant over 173 yeast stress environments (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.	53
3.7	The number of statistically significant motif combinations over 173 yeast stress environments (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.	54
3.8	Sample ratios in which at least one motif combination was deemed significant in 27 EGF-induced human breast cancer cell samples. (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.	55
3.9	Sample ratios in which at least one motif combination was deemed significant in 28 HRG-induced human breast cancer cell samples. (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.	56
3.10	The number of statistically significant motif combinations in 27 EGF-induced human breast cancer cell samples. (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.	57
3.11	The number of statistically significant motif combinations in 28 HRG-induced human breast cancer cell samples. (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.	58
3.12	Comparison of the empirical FWERs.	59
3.13	Comparison of the average numbers of false discoveries.	59
3.14	Comparison of the average numbers of false discoveries.	59
3.15	Threshold for the frequency of combinations and the number of combinations enumerated.	60
3.16	Threshold for the frequency of combinations and running time for enumerate frequent combinations.	60
3.17	Statistical significant of the four-motif combination in a yeast heat shock condition.	61

3.18	Statistical significant of the eight-motif combination in the human breast cancer cells treated with 0.5 nM EGF for 15 min.	65
3.19	Statistical significant of the seven-motif combination in the human breast cancer cells treated with 1.0 nM EGF for 60 min.	65
3.20	Statistical significant of the five-motif combination in the human breast cancer cells treated with 10.0 nM HRG for 30 min.	66
3.21	The expression levels of TFs that bind to the motifs in Figure 3.18. .	68
3.22	The expression levels of TFs that bind to the motifs in Figure 3.19. .	69
3.23	The expression levels of TFs that bind to the motifs in Figure 3.20. .	69
4.1	Illustration of the properties of candidate combinations. (a) Candidate and non-candidate combinations. (b) Relationship between a combination and the number of objects that have all of the items in I . . .	80
4.2	Illustration of FastWY to find the minimum P-value in each permuted dataset.	81
4.3	Illustration of FastWY to enumerate the combinations whose P-values are less than δ (Step 3).	82
4.4	Time complexity of FastWY when Fisher's exact test is used.	83
4.5	Running time of FastWY and WY-procedure with increasing numbers of items.	84
4.6	The number of combinations tested P-values of FastWY and WY-procedure with increasing numbers of items.	84
4.7	Running time of FastWY and WY-procedure with increasing numbers of samples.	85
4.8	The number of combinations tested P-values of FastWY and WY-procedure with increasing numbers of samples.	85
4.9	Threshold for the frequency of combinations and the number of combinations enumerated in yeast dataset.	86

List of Tables

1.1	Overview of existing multiple testing procedures.	6
1.2	Contribution of this thesis.	10
2.1	Numbers of errors committed when m tests are performed.	18
2.2	Statistics of datasets.	20
2.3	Threshold and the average number of up-regulated genes.	20
2.4	Relationships between the arity limit and the number of tests.	21
2.5	A contingency table for a combination I	21
2.6	Combinations detected by WY in yeast cells exposed to dithiothrietol.	23
2.7	Combinations detected by WY in breast cancer cells.	25
3.1	Statistics of the number of significant combinations of Yeast dataset.	47
3.2	Statistics of the number of significant combinations of EGF induced sample.	48
3.3	Statistics of the number of significant combinations of HRG induced sample.	49
3.4	Statistically significant motif combinations under a yeast heat shock condition.	62
3.5	Statistically significant motif combinations under a yeast raffinose medium condition.	62
3.6	The list of TFs that can bind to the motifs in Figure 3.18.	67
3.7	The list of TFs that can bind to the motifs in Figure 3.19.	68
3.8	The list of TFs that can bind to the motifs in Figure 3.20.	68
4.1	Performances of FIM algorithm with increasing genes in Yeast data	78
4.2	Statistically significant motif combinations detected by FastWY under the nitrogen limitations condition in yeast cells.	87

Chapter 1

Introduction

1.1 Background

To make advancements in biology, genomics and medical science, it is essential to discover collaborative effects, such as interactions between transcription factors (TFs) and multiple DNA mutations. Examples of combinatorial effects that have been observed through biological experiments include the following.

Combinatorial Regulations by TFs: To respond to a wide spectrum of environmental and developmental signals, the joint regulation of gene expression through the binding of different TFs is crucial component of cellular machinery. Combinatorial effects by pairs of TFs have been identified in several organisms [52, 100, 114, 102, 40, 72, 27]. Moreover, functional combinations of more than two TFs are commonly observed in both mammals [65, 76] and plants [7, 103] to respond to various signals.

Combinations of TFs to Induce Cell Reprogramming: The reprogramming of differentiated cells into a pluripotent state might be useful for improving medical treatment. Induced Pluripotent Stem (iPS) cells can be generated from mouse fibroblasts and somatic cells through the introduction of four TFs (*Oct3/4*, *Sox2*, *c-Myc* and *Klf4*) [90, 63]. In human cells, the same combination and a different combination of four TFs (*OCT4*, *SOX2*, *NANOG* and *LIN28*) induce reprogramming of adult human somatic cells [89] and human somatic cells [111], respectively.

Joint Effects of Several Amino Acid Mutations: Discovering associations between phenotypes and amino acid mutations is an important problem in biology, and it is often necessary to consider the joint effects of multiple mutations. For example, to develop drug resistance, each mutation of HIV protein is not equally important. However, the sequential acquisition of mutations exhibits the resistance [26, 120].

Collaborative Effects of Histone Tail Modifications: Histone tail modifications such as methylation, acetylation and phosphorylation are implicated in influencing gene transcription. The collaborative effects of two or more histone mod-

ifications are considered to play an important role in fundamental mechanisms in eukaryotic cells, such as cell fate decisions and both normal or pathological development [88, 49, 5, 104, 21].

To detect these combinatorial effects, computational analyses are useful because investigation of the higher-arity effects (effects with many factors) requires examining an intractable number of combinations. The methods are mainly categorized into two types depending on the strategy employed. One strategy is the prioritization of factors using machine learning methods, such as SVM [64, 77, 22, 2], LASSO [68, 35] and Bayesian analysis [121, 115]. When there is a primary factor that triggers the joint effect, this strategy is expected to detect the collaborators. However, when the simultaneous activities of two or more factors are required, this strategy might overlook the collaborative effects. Another strategy is to screen all pairs of factors using a statistical score [72, 4, 122, 113, 80], a regression analysis [51, 98] or other analysis. Although this strategy provides much better results than the single screening process, the higher-arity effects cannot be detected. Using current computational power, a comprehensive screening of the combinations of three or more factors may be possible. A crucial problem, however, is how to assess the statistical significance of the discovered combinations.

When combinations of up to k items are considered, the number of tested combinations increases exponentially to k . Such a large number of tests causes serious false discoveries problem, which is a well-known problem as multiple tests. Figure 1.1 shows an example of multiple tests. When the P-value is less than $\delta = 0.05$, the item is regarded as significant. If one test is performed, the probability that the result is a false discovery is 0.05 (the illustration at the upper left). However, when 10 tests are performed, the probability of at least one false discovery occurring becomes $1 - (1 - 0.05)^{10} = 0.401$, which means that at least one spurious result is obtained with 40.1% probability (the illustration at the upper right). Using the adjusted δ , the probability of false discovery can be controlled to keep it under a threshold α . For example, if we set δ to 0.005, the probability is $1 - (1 - 0.005)^{10} = 0.0489$, even after we perform 10 tests. We should use such adjustment, called multiple testing correction, to statistically evaluate the outcomes [71]. For instance, Nature magazine demands the multiple testing correction in their guidelines (http://www.nature.com/nature/authors/gta/Statistical_checklist.doc).

1.2 Overview of Multiple Testing Procedures

Multiple testing correction methods have been proposed to prevent false discoveries [15, 46, 106, 11]. The widely used multiple testing procedures can be categorized into two types according to their measures of how often false discovery occurs. One measure is the probability that at least one false discovery occurs, called *Family-Wise*

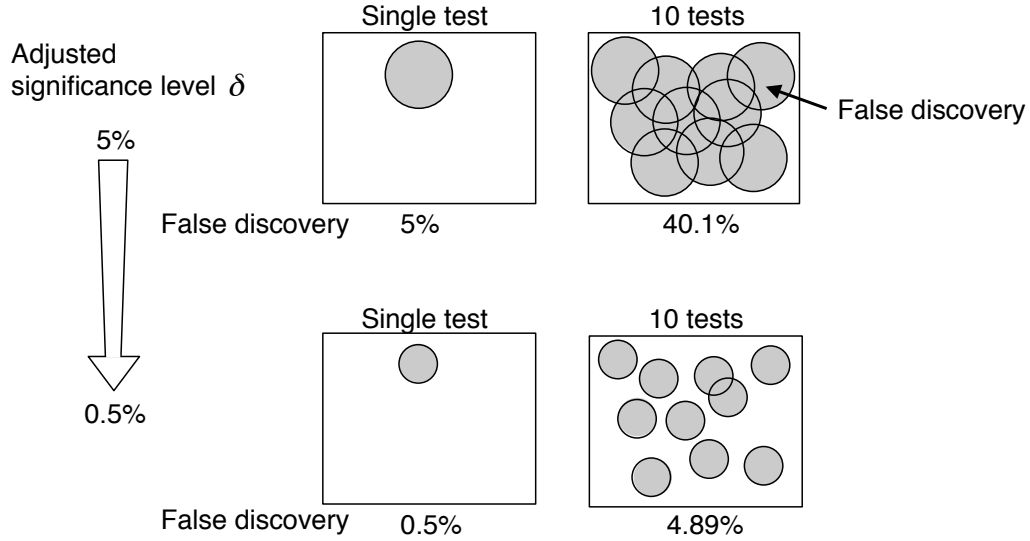


Figure 1.1: Illustration of a multiple testing procedure.

Error Rate (FWER). Another measure is the expected value of the proportion of false discoveries among the tests that are regarded as significant, called *False Discovery Rate (FDR)*. Both FWER and FDR controlling methods are categorized into two strategies, which are summarized in Table 1.1. One strategy is theoretical determination of the adjusted significance level. The other is an empirical distribution based strategy, which yields a higher sensitive correction than the theoretical correction. Here, we provide an overview of the multiple testing procedures and their software implementations. The theoretical background will be described in Chapter 2.

1.2.1 Theoretically FWER Controlling Procedures

We here summarize theoretically FWER controlling procedures. At first, we introduce Bonferroni correction [15], which is classical procedure but widely used in practical, then show the improved procedures.

Bonferroni Correction

One popularly used procedure is Bonferroni correction [15], which theoretically guarantees that FWER is controlled under α . Given α and the number of tests m , δ is calibrated to α/m . Because the procedure is very simple and fast, it is used in variously studies, e.g., Gene Ontology (GO) enrichment analysis [16, 9, 60], functional MRI [29, 39], and genome-wide association studies (GWAS) [53, 82, 107]. Moreover,

Table 1.1: Overview of existing multiple testing procedures.

	Theoretical	Empirical
FWER	Bonferroni [15], Holm [46] Shaffer [78], Šidák [101] Simes [81]	Westfall and Young [106]
FDR	Benjamini and Hochberg [11] Benjamini and Yekutieli [12]	Yekutieli and Benjamini [110] Efron [32] Storey and Tibshirani [87]

GeneSpring, which is a software released by Agilent Technologies Inc. for statistical analysis of microarrays, offers Bonferroni correction for multiple testing procedure, and `p.adjust` function on the R statistical programming language performs this adjustment. However, it is known that Bonferroni correction is conservative when tests have a high correlation to each other because this correction computes the upper bound of FWER with the assumption that each of the m tests is independent of each other. In combinatorial effects discovery, most tests depend on one another, therefore, it is hopeless to detect these effects using Bonferroni correction.

Holm Procedure

To improve the detection power of a single-step procedure such as Bonferroni correction, Holm proposed a step-down procedure that considers sequential detection [46]. To describe the motivation of Holm procedure, we suppose that the P-values of m tests are ordered $p^{(1)} \leq p^{(2)} \leq \dots \leq p^{(m)}$. If the smallest P-value $p^{(1)}$ is less than or equal to $\delta = \alpha/m$, our multiple testing problem has been reduced by one test, and the second smallest P-value $p^{(2)}$ is compared with $\alpha/(m-1)$. In general, we calibrate the adjusted threshold for $p^{(k)}$ as $\delta^{(k)} = \alpha/(m-k+1)$, start at $k=1$ and stop as soon as $p^{(k)} \leq \delta^{(k)}$. As with Bonferroni correction, Holm procedure is used in GeneSpring software and it is provided via `p.adjust` function in R. The procedure requires a small amount of time and is expected to improve the conservativeness of Bonferroni correction when a small number of tests are performed. However, when we have a large number of tests, this procedure not only requires a large amount of time to order the P-values, it also does little to improve the sensitivity of Bonferroni correction. For example, for large m , there are slight differences between α/m and $\alpha/(m-1)$. Although Holm procedure has been further improved using logically related detection, as described by Shaffer [78], these remedies might not be possible when a large number of tests are performed. Therefore, Holm and its improved procedures have inadequate sensitivity to discover combinatorial effects because they produce an intractable number of tests. Moreover, the problem of computing time is unresolved.

Other Bonferroni-like Procedures

To improve the conservativeness of Bonferroni correction, many procedures have been proposed that use certain assumption about the dependence structure of the test statistics for theoretical computation of the upper bound of FWER [44, 45, 47, 74, 101, 81]. These procedures are more sensitive than Bonferroni correction, but lack generality because the upper bound computations hold only under the assumptions concerning the dependence structure of the test statistics, such as the assumption of a joint distribution that satisfies Šidák's or Simes' Inequality, and we are uninformed about the test statistics in combinatorial effect discovery. Hence, we will propose a general procedure for combinatorial effect discovery that only uses Boole's Inequality as well as Bonferroni correction in Chapter 3.

1.2.2 Empirically FWER Controlling Procedures

Theoretically FWER controlling procedures such as Bonferroni correction are often conservative because they assume an arbitrary dependence structure between test statistics. An empirically FWER controlling procedure, such as Westfall-Young procedure [106] and other procedures [30, 73, 31], can consider the dependence among tests to compute the upper bound of FWER. If we know the distribution of test statistics when the test must result in insignificance (null distribution), δ is determined by the α -percentile point of the distribution. However, we cannot know the distribution a priori. Therefore, Westfall and Young proposed a method that empirically estimates the distribution via a permutation test. This procedure has high detection power [30, 34, 62], and it has been widely used in practice [37, 24, 107]. For example, GeneSpring, which is a software package released by Agilent Technologies, Inc., for statistical analysis of microarray data, offers Westfall-Young procedure for multiple testing correction, and Bioconductor project (<http://www.bioconductor.org>), which provides open source and open developmental software for the analysis of genomic data on the R statistical programming language, has implemented Westfall-Young procedure as `mt.maxT` function [37]. In haplotype analysis, Becker and Knapp applied the Westfall-Young procedure and reported a substantial gain in power over Bonferroni correction [8]. However, the procedure requires a greater amount of computing time for the permutation test, even when a small number of tests are performed. In combinatorial effect discovery, it is difficult to apply the Westfall-Young procedure because a huge number of tests are performed. Zhang *et al.* have proposed the methods for accelerating Westfall-Young procedure in epistasis detection [118, 117, 116], but these methods are specialized for the discovery of statistically significant combinations of two factors. Thus, in Chapter 4, we will propose a fast algorithm that enables us to apply Westfall-Young procedure to the discovery of higher-arity combinatorial effects.

1.2.3 Theoretically FDR Controlling Procedures

FDR controlling procedures are often preferred over FWER controlling procedures because FWER is too stringent in many situations, including genome scans for locating traits [105, 87], psychiatric research [58], and the analysis of microarray experiments [32]. Here, we introduce a widely used procedure proposed by Benjamini and Hochberg [11].

Benjamini-Hochberg Procedure

Benjamini and Hochberg [11] provided a step-up method for controlling FDR under the assumption that the test statistics are independent. As with Holm procedure, the m P-values are ordered $p^{(1)} \leq p^{(2)} \leq \dots \leq p^{(m)}$. Benjamini-Hochberg procedure computes the threshold for the i -th smallest P-values as $\delta^{(i)} = i\alpha/m$ and finds $k = \max\{i : p^{(i)} \leq \delta^{(i)}\}$. The set of tests whose P-values are $p^{(1)} \leq p^{(2)} \leq \dots \leq p^{(k)}$ is considered significant. In their article [11], Benjamini and Hochberg showed that this procedure controls FDR to at most α when the test statistics are independent.

Benjamini-Hochberg procedure is widely used because the detection power is greater than the theoretically controlling FWER controlling procedures, such as Bonferroni correction, and it is fast when the number of tests is not large. For example, GeneSpring software and `p.adjust` function in R provides this procedure. In addition, DAVID [48] (<http://david.abcc.ncifcrf.gov>), which is an annotation resource for large gene lists, uses Benjamini-Hochberg procedure for multiple testing correction in its functional annotation charts.

Although Benjamini-Hochberg procedure has the potential to detect more factors as significant than FWER controlling procedures, it is difficult to detect combinatorial effects with it because of the assumption of independence among test statistics [12]. To improve the conservativeness, modified procedures to theoretically control FDR have been proposed [12, 10, 38]. In particular, Benjamini and Yekutieli [12] established FDR control for test statistics with arbitrary joint distributions, such as positive regression dependence, by using the simple modifications of Benjamini-Hochberg procedure. Benjamini-Yekutieli procedure uses $\delta'^{(i)} = i\alpha/mC(m)$ where $C(m) = \sum_{i=1}^m 1/i$. As well as Benjamini-Hochberg procedure, Benjamini-Yekutieli procedure is also used by `p.adjust` function in R. Unfortunately, there is not much evidence about what form of dependence must be used between the statistical scores in combinatorial effects discovery. Therefore, making any assumptions about the form of dependence has no theoretical basis. Additionally, both Benjamini-Hochberg procedure and Benjamini-Yekutieli procedure require great computational resources to order the P-values when a large number of tests are performed.

1.2.4 Empirically FDR Controlling Procedures

The empirical FDR controlling procedures are classified into two approaches. One approach is the random sampling based approach, and the other is Bayesian approach. Yekutieli and Benjamini [110] proposed FDR controlling multiple testing procedure that uses resampling-based adjusted P-value to account for certain types of dependence structure among test statistics. However, this procedure requires a large amount of time to simulate the distribution of the statistical score. A number of Bayesian approaches have been proposed to control FDR in microarray data analysis [32, 97, 84, 86, 85, 87, 59, 70]. These procedures require an assumption for the distribution of statistical score, however we cannot find the distribution a priori. Therefore, it is difficult to generalize these approaches.

1.3 Contributions of This Thesis

In this thesis, we first show how existing multiple testing procedures hide the higher-arity significant combinations that have already been discovered experimentally. Then, to overcome the problem, we propose two novel procedures whose contributions are shown in Table 1.2. One theoretically adjusts the significance level, and the other uses an empirical distribution for the given dataset. We focus here on FWER because FWER is more tightly bound than FDR, and the significant results based on the FWER would also be found by using FDR. This thesis is organized as described below.

1. In Chapter 2, we present an overview of the widely used multiple testing procedures, and we reveal how these procedures cannot detect combinations as significant in combinatorial regulation discovery. We first define combinatorial effects by considering this thesis in Section 2.2. In Sections 2.3 and 2.4, we explain the theoretical background of widely used multiple testing procedures, and we then investigate the problem of the procedures on combinatorial effects discovery by applying them to yeast and human transcriptome datasets in Section 2.5.
2. In Chapter 3, we propose a method to theoretically control FWER, called *Limitless Arity Multiple-testing Procedure (LAMP)*. A method to theoretically control FWER, Bonferroni correction, assumes that all of the tests increase the FWER. In Section 3.2, we show that not every test increases FWER and the properties can be used to improve the sensitivity of Bonferroni correction. In Section 3.3, we propose an efficient branch-and-bound algorithm, LAMP, which calibrates the Bonferroni factor to the smallest possible value. We also experimentally

Table 1.2: Contribution of this thesis.

	FWER controlling procedures	
	Theoretical	Empirical
Single items (1 - 100 tests)	Bonferroni correction [15] Holm procedure [46]	Westfall-Young procedure [106]
Pair items (1000 - 10^6)	LAMP [93] (Chapter 3)	FastANOVA [118] FastChi [119], COE [117] FastWY [94] (Chapter 4)
More than three items ($10^5 - \infty$)	LAMP [93] (Chapter 3)	FastWY [94] (Chapter 4)

demonstrate that LAMP outperforms other theoretical multiple testing procedures in Section 3.5.

3. In Chapter 4, we propose a method to empirically control FWER, called *Fast Westfall-Young procedure* (*FastWY*). We accelerate an empirically controlling FWER procedure, Westfall-Young procedure [106] because it usually requires an enormous amount of computational time even when a small number of tests are performed. Westfall-Young only uses the minimum P-value on each permuted dataset to estimate the null distribution. Therefore, we expand the algorithm of LAMP to find the minimum P-value in a permuted dataset in Section 4.3. In Section 4.4, we compare the performance of FastWY to that of Westfall-Young procedure by applying the procedures to yeast.

This thesis is based on the following publications of the author [93, 94, 95].

Chapter 2

Combinatorial Discovery and Multiple Testing Procedure

2.1 Introduction

To understand the complex biological systems, it is crucial to investigate higher-arity combinatorial effects, such as joint activities of three or more TFs [65, 76, 7, 103], is crucial. However because multiple testing correction hides the significant effects even though they have been discovered by biological experiments, scanning for the combinations by computational analysis is limited to single factors or pairs.

In this chapter, we show how existing multiple testing procedures hide statistically significant combinations. To clarify the problem, we first define the statistically significant combinations targeted in this thesis in Section 2.2. Then, the widely used multiple testing procedures are introduced in Sections 2.3 and 2.4. The procedures can be categorized into groups according to the indexes used to measure how often the false discovery occurs. One measure is FWER which is the probability that at least one false discovery occurs, and the other is FDR which is the expected value of the proportion of false discoveries among the tests that are regarded as significant. In each section, we first formulate the indexes, and then we show the widely used multiple testing procedures. Finally, we apply these procedures to yeast and human transcriptome datasets, and we discuss their problems in combinatorial effects discovery in Section 2.5.

This chapter is based on the following publication of the author [94].

2.2 Significant Combination

In this section, we define a combinatorial effect and formulate a statistically significant combination.

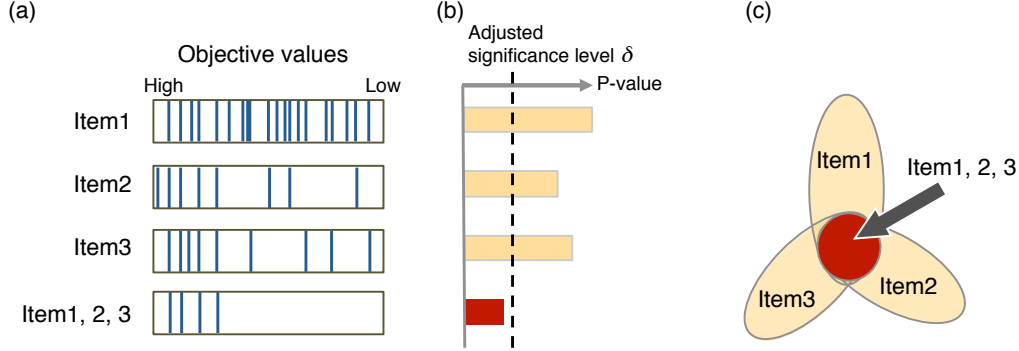


Figure 2.1: Illustration of a salient three-item combination.

Figure 2.1(a) shows the relationship between the dataset and the combination to be tested. Suppose that we have N objects and M items. Each object has a value y_i ($i \in \subseteq \{1, \dots, N\}$) and is associated with a set of items $I_i \subseteq \{1, \dots, M\}$. In the first three boxes of Figure 2.1(a), the objects are ordered by value, and the blue bars indicate that the object is related by to each item. For a given set of items I , we partition N objects into $x(I)$ related objects and $N - x(I)$ unrelated objects depending on whether they are related by all of the items of the combination or not (the bottom box on Figure 2.1(a)). For example, when we identify a combination of TFs that is related to highly expressed genes, I and y_i are given by the set of TFs and the gene expression level, respectively. The P-value of I can be calculated by a statistical test, such as Fisher's exact test and Mann-Whitney U test, and if the P-value is below a threshold δ , the combination is considered significantly associated with the objects division. We call such combination as a *significant combination*.

Figures 2.1(b) and (c) illustrate a salient three-item combination. Although any single item is not significant, the three-item combination is significant. If combinations of up to k items (k -item combinations) are considered for the detection of such higher-arity combinations, the number of hypotheses m increases exponentially with k . Hence, a multiple testing correction such as Bonferroni correction and Benjamini-Hochberg procedure must be used to avoid false discoveries [71]. In the following sections, we describe the principles of the widely used multiple testing procedures, and then we apply them to higher-arity combination discovery.

2.3 Procedures to Control Family-Wise Error Rate

FWER is the probability that at least one false discovery occurs,

$$FWER = Pr(V > 0), \quad (2.1)$$

where V is the number of false discoveries. The widely used FWER controlling procedures can be categorized into two approaches. One approach is a theoretical approach that computes the theoretical upper bound of FWER, for instance, Bonferroni correction [15] and Holm procedure [46]. To compute the upper bound, this approach uses the assumption that all of the tests are independent of each other. The other approach is an empirical approach that estimates the probability distribution of P-values from the null hypothesis, for instance, Westfall-Young procedure [106]. Unlike theoretical approaches, empirical approaches consider the dependence among tests by estimation of the distribution of statistical score with permutation test.

In this section, we first introduce two theoretical approaching procedures, which are called Bonferroni correction [15] and Holm procedure [46], and then introduce an empirical approaching procedure, which is called Westfall-Young procedure [106]. To simply description, we explain the procedures with an example that m items are scanned to enumerate significant items.

2.3.1 Bonferroni Correction

Suppose that we have m items to be tested. Let p_i be the raw P-value of item $i \in \{1, 2, \dots, m\}$. To control FWER under α , we use an adjusted threshold δ . If p_i is below the adjusted threshold δ , the item is regarded as significant. The Bonferroni correction calculates the upper bound of the FWER as

$$\begin{aligned}
 FWER &= 1 - \Pr \left(\bigcap_{i \in \{1, \dots, m\}} \{p_i > \delta\} \right) \\
 &= \Pr \left(\bigcup_{i \in \{1, \dots, m\}} \{p_i \leq \delta\} \right) \\
 &\leq \sum_{i \in \{1, \dots, m\}} \Pr(p_i \leq \delta) \\
 &\leq m\delta.
 \end{aligned} \tag{2.2}$$

Figure 2.4 provides an illustration of this correction. In Figure 2.4(a), the total of the gray area indicates FWER for $m = 5$. Bonferroni correction assumes that each test is independent of the others as shown in Figure 2.4(b). Thus, the upper bound of FWER is computed as $m\delta$. Based on this inequality, when δ is set to α/m , FWER becomes at most α . This calculation is the Bonferroni correction.

Pseudo code for Bonferroni correction is shown in Algorithm 1. When p_i and the number of tests are given, Bonferroni correction is performed for each statistical test. If $p_i \leq \alpha/m$, then the result of test i is significant. When we scan m items,

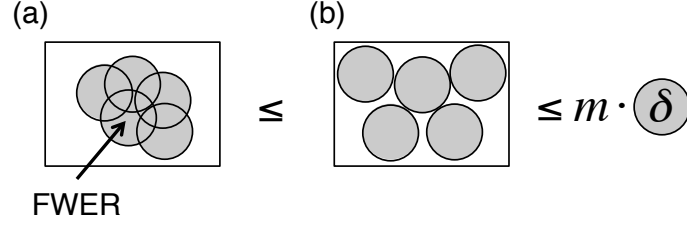


Figure 2.2: Illustration of how Bonferroni correction computes the upper bound of FWER.

Algorithm 1 Bonferroni correction

Input: P-value p_i , the number of tests m , FWER upper limit α

Output: decide whether the test results in significant or non-significant

- 1: **if** $p \leq \alpha/m$ **then**
 - 2: significant
 - 3: **else**
 - 4: non-significant
 - 5: **end if**
-

Algorithm 1 is repeated m times. Therefore, the computational time grows linearly with m .

This correction is often too conservative [71] because the inequality assumes the worst case in which all of the tests are independent of one another (lines 3 and 4 in Equation 2.2). Especially for combinatorial discovery, many combinations of two items have a high correlation between the single items. Therefore, when we use Bonferroni correction to control the FWER under α , the real FWER would be far smaller than α .

2.3.2 Holm Procedure

Holm improved the sensitivity of Bonferroni correction [46] by considering that all tests whose P-value are less than a significant test are significant. To describe the principle of the Holm procedure, we suppose that the P-values of m items are ordered, $p^{(1)} \leq \dots \leq p^{(m)}$.

Suppose that δ satisfies $p^{(k)} \leq \delta$. Because $p^{(1)}, \dots, p^{(k)}$ are significant, $\bigcup_{i \in \{1, \dots, m\}} \Pr(p_i \leq$

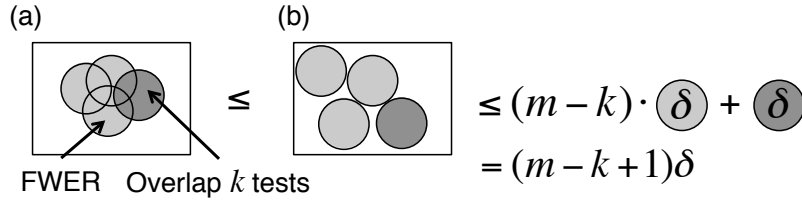


Figure 2.3: Illustration of how Holm procedure computes the upper bound of FWER when k -th smallest P-value is less than or equal to δ .

δ) in Equation 2.2 is identical to $\bigcup_{i \in \{k, \dots, m\}} \Pr(p^{(i)} \leq \delta)$. Hence,

$$\begin{aligned}
 FWER &= \Pr \left(\bigcup_{i \in \{k, \dots, m\}} \{p^{(i)} \leq \delta\} \right) \\
 &\leq \sum_{i \in \{k, \dots, m\}} \Pr(p^{(i)} \leq \delta) \\
 &\leq (m - k + 1)\delta.
 \end{aligned} \tag{2.3}$$

To control the FWER to be under α , the Holm procedure calibrates the adjusted threshold $\delta^{(k)} = \alpha / (m - k + 1)$ for the k -th smallest P-value $p^{(k)}$. As shown in Algorithm 2, Holm procedure gradually increases k from 1. When $p^{(k)}$ exceeds $\alpha / (m - k + 1)$, $p^{(1)}, \dots, p^{(k-1)}$ are regarded as significant combinations.

Although the Holm procedure has higher sensitivity than the Bonferroni correction, it remains too conservative to detect combinations of items because when $k \ll m$, $\alpha / (m - k + 1)$ is almost α / m , which is the adjusted significance level of the Bonferroni correction. Moreover, it requires the computational resource to order P-values.

2.3.3 Westfall and Young Procedure

In contrast to Bonferroni correction and Holm procedure, Westfall-Young procedure generates the distribution of the FWER using the random permutation procedure, and its great advantage is not using the independence between tests. This procedure generally has a higher sensitivity than Bonferroni correction [31].

If we knew the probability distribution of the statistical scores of m items when all of the m items should not be detected as significant, we could directly compute

Algorithm 2 Holm procedure**Input:** ordered P-values $p^{(1)} \leq p^{(2)} \leq \dots \leq p^{(m)}$, FWER upper limit α **Output:** enumerate significant items

```

1: for  $k = 1$  to  $m$  do
2:    $\delta \leftarrow \alpha / (m - k + 1)$ 
3:   if  $P^{(k)} > \delta$  then
4:      $\delta \leftarrow \alpha / (m - k)$ 
5:     break
6:   end if
7: end for
8: return the set of items whose  $p_i \leq \delta$ 

```

the FWER for δ as

$$FWER = \Pr \left(\bigcup_{i \in \{1, \dots, m\}} \{p_i \leq \delta\} \right) = \Pr \left(\min_{i \in \{1, \dots, m\}} \{p_i \leq \delta\} \right). \quad (2.4)$$

Therefore, if we knew the probability distribution of the minimum P-value of m items, the α -percentile point of the distribution represents the adjusted significance level δ . However, we do not know the distribution in advance. Hence, Westfall-Young procedure estimates the distribution from randomly permuted data.

A pseudo code of Westfall-Young procedure is shown in Algorithm 3. N objects and m items to be tested are given. Westfall-Young procedure generates the permuted data by randomly shuffling the relationships between the object i and the value y_i in line 3 of Algorithm 3. The associations between objects and samples are held. With the random permuted data, the minimum P-value among all of the tests is computed in line 4 since FWER depends only on the minimum P-value. Gathering K minimum P-values provides the simulated null distribution of FWER. Line 7 calculates δ by the α percentile point in the distribution.

Although the sensitivity of Westfall-Young procedure is higher than that of Bonferroni correction and Holm procedure, it requires a large amount of computational time compared to Bonferroni correction and Holm procedure, because a large number of tests should be performed to generate an empirical distribution of FWER.

2.4 Procedures to Control False Discovery Rate

In many situations, control of FWER can lead to unduly conservative procedures. In current areas of application of multiple testing procedures, such as gene expression analysis and genome-wide association studies, thousands of tests are performed

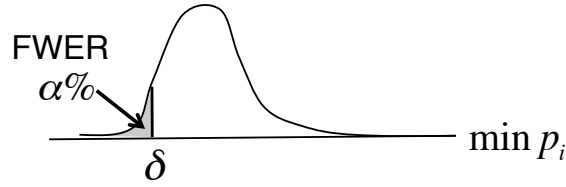


Figure 2.4: Illustration of the association between FWER and the probability distribution of the minimum P-value.

Algorithm 3 Westfall-Young procedure

Input: N samples and the values $\{y_1, \dots, y_N\}$, the relations between samples and m items, FWER upper limit α , the number of iteration K ,

Output: enumerate significant itemsets.

// Step 1: estimate null distribution

1: $A \leftarrow \{\}$ // Store the minimum P-values

2: **for** $k = 1$ to K **do**

3: Generate a permuted dataset

4: $p \leftarrow \min p_i$ for $I \in \{1, \dots, m\}$ // The minimum P-value

5: Add p to A

6: **end for**

// Step 2: calibrate the threshold δ

7: $\delta \leftarrow \max_k A[k]$, where $A[k] < \lceil \alpha K \rceil$ -th value in A

// Step 3: Enumerate all itemsets that P-value are less than δ

8: return the set of items whose $p_i \leq \delta$ for $i \in \{1, \dots, m\}$

simultaneously and a fairly small proportion of items is expected to be detected as significant. In this context, multiple testing procedures to control FDR that tolerate some false discoveries have been proposed [11, 12, 36, 54, 57]. Benjamini and Hochberg [11] formulated a criterion of the ratio of false discoveries. Table 2.4 summarizes a number of errors committed. Let R , V and S be the number of tests that regard as significant, the number of tests for which the null hypothesis is true (false discovery), and S be the number of tests for which the (true discovery). The quantity Q is defined as

$$Q = \begin{cases} V/R & \text{for } R > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (2.5)$$

V , S and Q are unobservable values. FDR is defined as the expected value $E(Q)$. In this section, we introduce a classical article of the FDR controlling method proposed by Benjamini and Hochberg [11].

Table 2.1: Numbers of errors committed when m tests are performed.

		Decision		
		Non-significant	Significant	Total
Actual	False	U	V	m_0
	True	T	S	$m - m_0$
Total		$m - R$	R	m

2.4.1 Benjamini and Hochberg procedure

Benjamini-Hochberg procedure controls FDR to be less than or equal to α by assuming that P-values of m_0 tests have a uniform distribution and by assuming that the tests are independent. We suppose that the m P-values are ordered as $p^{(1)} \leq \dots \leq p^{(m)}$ and that an adjusted significance level δ satisfies $p^{(k)} \leq \delta < p^{(k+1)}$. If we know the number of tests whose actual results are false m_0 , the expected value of the number of false discoveries is computed as $m_0\delta$ because we assume that the P-values of m_0 tests have a uniform distribution. Therefore, the false discovery proportion Q is computed as $m_0\delta/k$, and FDR is computed as

$$E(Q) = \frac{m_0\delta}{k} \leq \frac{m\delta}{k}.$$

To control FDR to be under α , Benjamini-Hochberg procedure calibrates the adjusted threshold $\delta^{(k)} = k\alpha/m$ for the k -th smallest P-value $p^{(k)}$. As shown in Algorithm 4, Benjamini-Hochberg procedure gradually decreases k from m . When $p^{(k)}$ falls below $k\alpha/m$, $p^{(1)}, \dots, p^{(k)}$ are regarded as the significant items.

Benjamini-Hochberg procedure may have higher sensitivity than theoretically FWER controlling procedures such as Bonferroni correction and Holm procedure. However, when many tests correlate with each other, this procedure is conservative because of the assumption of independence among tests. In significant combination discovery, many combinations of two items have high correlations between the single items. Hence, when we use Benjamini-Hochberg procedure to control FDR under α , the real FDR would be less than α .

2.5 Performance Evaluation for Combinatorial Regulation Discovery

In this section, to compare the abilities, we apply the often used multiple testing procedures, three FWER controlling methods introduced Section 2.3 (Bonferroni correction (Bonferroni), Holm procedure (Holm) and Westfall-Young procedure

Algorithm 4 Benjamini-Hochberg procedure

Input: ordered P-values $p^{(1)} \leq p^{(2)} \leq \dots \leq p^{(m)}$, FDR upper limit α **Output:** enumerate significant items

```

1: for  $k = m$  to 1 do
2:    $\delta \leftarrow k\alpha/m$ 
3:   if  $P^{(k)} \leq \delta$  then
4:     break
5:   end if
6: end for
7: return the set of items whose  $p_i \leq \delta$ 

```

(WY)) and an FDR controlling method in Section 2.4.1 (Benjamini-Hochberg procedure (BH)), for identifying joint activity of different TFs with yeast and human transcriptome datasets. Note that FWER is a different measure from FDR, and FWER offers a stricter significance level than FDR under the same α . Bonferroni and WY are written in Python 2.7. We used the R function to correct with the Holm and BH. All the experiments were performed on a machine with two 2.3 GHz AMD Opteron processors with 32 GB of RAM running RedHat Linux.

2.5.1 Experimental Datasets

To detect combinatorial effects of TFs, we generated datasets that were observed from yeast and human breast cancers, and we summarize the statistics in Table 2.2 and 2.3. The objects and items correspond to genes and motifs, respectively. In our experiment, N genes were divided into up-regulated genes and un-regulated genes by applying a threshold because up-regulation by external signals has to be distinguished from natural fluctuations. If the expression level of gene i is higher than the threshold value, y_i is set to 1 (up-regulated), otherwise, it is set to 0 (un-regulated). We investigated how our results changed with different thresholds.

In the yeast dataset, the binding site positions were generated from Harbison *et al.* [43], and the gene positions from *Saccharomyces* Genome Database [23] were used. We associated a motif with a gene when the TF binding site was located between 800 bp upstream and 50 bp downstream from the transcription start sites of the gene. The integrated data contained 102 types of motifs. The motifs were associated with an average of 30.1 genes. For gene expression data, we used microarray data observed over 173 different conditions on an average of 5935.6 genes [33]. A gene was considered up-regulated if the log2 ratio of its expression level in the target environment to its expression level in the control environment was greater than or equal to a threshold value. We used three different thresholds: 0.5, 1.0 and 1.5.

Human breast cancer dataset was generated from the MCF-7 breast cancer ex-

Table 2.2: Statistics of datasets.

	Yeast	EGF	HRG
# of datasets	173	27	28
# of motifs (M)	102	397	397
Avg. # of genes (N)	5935.6	12,773	12,851
Avg. $x(I_i)$	30.1	215.4	217.3

Table 2.3: Threshold and the average number of up-regulated genes.

	Expression threshold		
	0.5	1.0	1.5
Yeast	990.861	363.624	171.358
EGF	2359.815	1226.667	777.037
HRG	4191.593	2340.148	1276.333

pression profile (accession no. GSE6462; normalized using the trimmed mean method with GCOS 1.2) [67], and the relationships between motifs and genes were provided by the MSigDB [108]. The expression data contain samples induced by two different ErbB receptor ligands, EGF and HRG. We removed genes whose log2 expression levels were less than 4 over all conditions on each ligand. The numbers of genes in EGF- and HRG-induced samples were 12,773 and 12,851, respectively. A gene was considered up-regulated if the log2 ratio of its expression level in the target sample to that in the control sample was at least a threshold. Similar to the yeast dataset, we used three different thresholds: 0.5, 1.0 and 1.5. To associate TF binding motifs with genes, we used the value “motif” in category C3 in the MSigDB [108]. The data contained a few different quality scores for each motif. To remove ambiguity, we merged associated genes over the different qualities for each motif. We associated the TF data with gene expression data through GenBank ID. The number of motifs was 397, and the motifs were associated with an average of 215.4 and 217.3 genes for EGF and HRG, respectively.

As a statistical measure, a one-sided Fisher’s exact test was used. Given a motif combination I , we partitioned N genes in two ways as shown in Table 2.5. n indicates the number of highly expressing genes. We define $x(I)$ and $a(I)$ as the number of samples that contain a combination I and the number of genes that both exit I in upstream region and are highly expressed, respectively. The P-value of one-sided

Table 2.4: Relationships between the arity limit and the number of tests.

Arity limit	# of tests	
	Yeast	EGF and HRG
1	102	397
2	5,253	79,003
3	176,953	10,428,793
4	4,426,528	1,029,883,108

Table 2.5: A contingency table for a combination I .

	$y_i = 1$	$y_i = 0$	Total
$I_i \supseteq I$	$a(I)$	$x(I) - a(I)$	$x(I)$
$I_i \not\supseteq I$	$n - a(I)$	$N - x(I) - n + a(I)$	$N - x(I)$
Total	n	$N - n$	N

Fisher's exact test of I can be calculated as

$$P(I) = \sum_{a_i=a(I)}^{a_{\max}} \frac{\binom{n}{a_i} \binom{N-n}{x(I)-a_i}}{\binom{N}{x(I)}}, \quad (2.6)$$

where $a_{\max} = \min \{x(I), n\}$.

The FWER is controlled to be under 0.05 by the Bonferroni, Holm and WY. For BH, the FDR is controlled to be under 0.05. We estimated null distributions from 1,000 permuted data in WY. We varied the maximum number of motifs of the tested combinations from one to four to check the operation of each procedure.

2.5.2 Sensitivities in Yeast Transcriptome Dataset

To compare the detection abilities of the procedures in FWER, we computed the number of samples in which at least one motif combination was significant. Figure 2.5 shows the detected ratios of four multiple testing correction methods. In Figure 2.5, the y-axis represents the ratios of the samples over 173 samples in the yeast dataset. Within each bar, the colors indicate the largest combination size that was detected in each sample. For example, in Figure 2.5(c), the white and light green portions of the second left bar (2 of Bonferroni) indicate that the largest combinations found by Bonferroni contained one motif in 42.2% samples and pairs of motifs in 12.7% samples.

The detected ratio of Bonferroni decreases with the increasing maximum combination size because the number of tested motif combinations m increases exponentially with the maximum combination size, which makes the adjusted threshold δ extremely small. Bonferroni found no significant combinations of four motif combinations even when we tested for four combinations by Bonferroni in Figure 2.5(a)-(c). In particular, in Figure 2.5(c), Bonferroni detected only single or pairs of motifs. Comparing Holm with Bonferroni, the modification of Holm resulted in no improvement in the combinatorial discovery. Especially when Holm was performed for two- to four-motif combinations, the detected motif combinations were identical to the result of Bonferroni. This result confirms that Holm has a very similar δ to as Bonferroni when a large number of tests are performed.

WY discovered 41 four-motif combinations in eight samples, and the result differed from Bonferroni and Holm. When we considered only a single motif, the number of detected samples was closed to the number found by Bonferroni and Holm. However, when we considered two or more combinations, the ratios of the detected samples in WY did not change with the maximum combination size, whereas those of Bonferroni and Holm decreased dramatically. These results suggest that Bonferroni and Holm overlooks higher-order motifs due to their conservative corrections.

The combinations detected only by WY contain biologically known and important results. Table 2.6 shows the detected motif combinations under an environment of 90 min exposure to dithiothrietol (DTT). The adjusted significance level of WY is $\delta = 0.000602$. The adjusted P-values $p_i \cdot \delta$ are used for Bonferroni, Holm and BH. A four-motif combination (MBP1, STE12, SWI4, SWI6) was found in the environment. Surprisingly, no single motif or two-motif combination within the four-motif combination had a statistically significant P-value because these combinations do not appear in Table 2.6. This result indicates the importance of considering the effects of larger motif combinations. The combination regulates 10 genes in total (*CRH1*, *ENV9*, *FKS3*, *GIC2*, *MNN5*, *PET54*, *SCW10*, *SRL1*, *SWE1* and *TOS2*) and four of them (*CRH1*, *FKS3*, *SCW10* and *SRL1*) are annotated as cell wall organization (Gene Ontology GO:0071555). Gasch *et al.* suggests that the response to cell wall damage is induced in yeast cells by DTT exposure [33]. This result confirms that the four-motif combination regulates the downstream genes to respond to DDT exposure environment. Using Bonferroni and Holm, these interesting combinations may be overlooked due to their overly conservative correction.

In many biological data analyses, FDR is also used as a multiple testing correction. We compared the samples detected by WY with those of BH and showed it in Figure 2.5. The detection power of BH also dramatically decreases with the maximum combination size, and WY detected significant combinations from more samples than BH. To compare the sensitivity in each samples, we computed the number of motif combinations was deemed significant in 173 yeast stress environments of the yeast

Table 2.6: Combinations detected by WY in yeast cells exposed to dithiothrietol.

Combination	P-value	Bonf. (≤ 4)	Holm (≤ 4)	BH (≤ 4)
HAC1	0.000238	> 1	> 1	> 1
HSF1, HAC1	0.000238	> 1	> 1	> 1
YAP7	0.000361	> 1	> 1	> 1
MBP1, STE12, SWI4, SWI6	0.000401	> 1	> 1	> 1
MBP1, STE12, SWI6	0.000546	> 1	> 1	> 1
MBP1, STE12, SWI4	0.000546	> 1	> 1	> 1

dataset and showed it in Figure 2.6. Only when the statistical tests are performed for single motifs, BH detected more motifs as significant motifs than WY. When we checked combinations of two or more motifs, WY detected more combinations as significant combination than BH. In particular, in Figure 2.6(c), BH detected no combinations of three or four motifs as significant combination as well as the results of Bonferroni and Holm, while WY discovered 80 three-motif combinations and 41 four-motif combinations when we tested combinations of up to four motifs. This result implies that replacing Bonferroni with WY is more useful for detecting statistically significant combinations than replacing Bonferroni with BH.

2.5.3 Sensitivities in Human Breast Cancer Transcriptome

Figures 2.7(a)-(c) report the numbers of samples detected by using Bonferroni, Holm, WY and BH in 27 EGF-induced human breast cancer cell samples. N/A means that the analysis could not be performed due to the high computational resource requirements to order the P-values of m tests. As in the yeast dataset, the number of detected samples by Bonferroni and Holm dramatically decreases with the arity limit. The number of samples detected by WY also decreases toward the maximum size, but it decreases more slowly than it does Bonferroni and Holm. The results of BH procedure were similar to those of Bonferroni and Holm. When the arity limit was larger than two, the sample ratios of BH were less than WY. When combinations up to two motifs were checked, BH procedure detected more combinations than WY (Figures 2.8(a)-(c)). However, when three-motif combinations were tested, the numbers of detection by BH markedly decreased, whereas those numbers by WY increased as shown in Figures 2.8(a) and (b). As the result, WY detected more higher-arity combinations than BH.

Figure 2.9 compares the numbers of detected samples among Bonferroni, Holm and WY in 28 HRG-induced human breast cancer cell samples. N/A means that the analysis could not be performed due to the high computational resource requirements to

order the P-values of m tests. As in the yeast dataset, the number of samples detected by Bonferroni and Holm dramatically decreased with the maximum combination size. The number of samples detected by WY also decreased toward the maximum size, but the speed of the decrease was slower than those of Bonferroni and Holm. In all Figures 2.9(a)-(c), Bonferroni found three-motif combination in maximum. When we investigated combinations of up to four motifs, the adjusted significance level fell below the P-value of the three-motif combinations, and only single-motifs were found. Holm provided identical results to Bonferroni in our analyses.

By comparing the detected sample ratio between WY and BH in Figure 2.9(a)-(c), we see that the detection power of BH decreases as the maximum combination size increases, and it decreases more rapidly than WY. WY detected significant combinations from more samples than BH when combinations of two or more motifs were tested. For example, when combinations of up to three motifs were tested in Figure 2.9(b), BH discovered the combinations in 21.4% of samples and detected three-motif combinations in 7.1% samples, while WY discovered combinations in 78.6% of samples and found three-motif combinations in 50% of samples. Moreover, BH could not be used to detect combinations of up to four motifs in our analyses because the high computational resources needed to order the P-values of more than 10^9 tests. To compare the sensitivity in each sample, we computed the number of motif combinations that was deemed significant in 28 HRG-induced human breast cancer cell samples, and the results are shown in Figure 2.10. When the statistical tests were performed for single motifs or pairs of motifs, BH identified more motifs as significant than WY. When we checked combinations of three or four motifs, WY identified more combinations as significant than BH. As Figure 2.10(b) shows in particular, BH detected only 8 three-motif combinations, whereas WY discovered 51 three-motif combinations when we tested combinations of up to three motifs. These results suggest that we overlook many significant combinations and combinatorial effects of higher-order motifs even when we use the FDR controlling procedure, BH, instead of a theoretically controlling FWER controlling procedure such as Bonferroni and Holm.

In Figure 2.9(b), WY detected four-motif combinations in 39.3% of breast cancer samples when combinations of up to four motifs were tested. Table 2.7 shows the motif combinations detected in the samples exposed to 10.0 nM HRG for 30 min. The adjusted significance level of WY is $\delta = 2.58 \cdot 10^{-7}$. The adjusted P-values $p_i \cdot \delta$ are used for Bonferroni, Holm and BH. N/A means that the combination could not be tested because Holm and BH required high computational resources to order the P-values of m tests. In this sample, six four-motif combinations had statistically significant associations with high expressed genes. Other methods detected none of these combinations. Remarkably, no single motif or motif pair within these four-motif combinations has a statistically significant P-value. Moreover, four of the combinations ($\{\text{CP2, GATA1, FOXO4, E12}\}$, $\{\text{OCT, TATA, NFkB, FOXO4}\}$, $\{\text{LEF1, NFAT,}$

Table 2.7: Combinations detected by WY in breast cancer cells.

Combination	P-value	Bonf. (≤ 4)	Holm (≤ 3)	BH (≤ 3)
CP2, E12, FOXO4, GATA1	$3.58 \cdot 10^{-9}$	> 1	N/A	N/A
FOXO4, NFKB, OCT, TATA	$8.99 \cdot 10^{-9}$	> 1	N/A	N/A
MAZ	$1.90 \cdot 10^{-8}$	> 1	0.198	0.198
E12, FREAC2, PAX4, STAT5B	$4.85 \cdot 10^{-8}$	> 1	N/A	N/A
FREAC2, NFAT, PAX4, STAT5B	$8.96 \cdot 10^{-8}$	> 1	N/A	N/A
FREAC2, PAX4, STAT5B	$9.81 \cdot 10^{-8}$	> 1	> 1	0.405
AP4	$1.17 \cdot 10^{-7}$	> 1	> 1	0.405
LEF1, NFAT, PAX4, STAT5B	$1.26 \cdot 10^{-7}$	> 1	N/A	N/A
E12, OCT1, STAT5A, TATA	$1.29 \cdot 10^{-7}$	> 1	N/A	N/A
AP4, MAZ	$2.05 \cdot 10^{-7}$	> 1	> 1	0.496
STAT5B, SP1	$2.38 \cdot 10^{-7}$	> 1	> 1	0.496

PAX4, STAT5B} and {TATA, E12, OCT1, STAT5A}) are found only when we test for the combinations of up to four motifs because no ternary combinations within these four-motif combinations have significance. FOXO4, which is contained in the two of the four-motif combinations, is activated in MCF-7 cells [55]. Hence, our result may suggest the collaborative effects of FOXO4. Using WY to detect combinatorial effects, it is possible to find overlooked statistically significant associations.

2.6 Conclusion

In this chapter, we introduced four multiple testing procedures that are often used to uncover biological problems: Bonferroni, Holm, WY, and BH procedures. Bonferroni and Holm control FWER by theoretically computing the upper bound of FWER. Different from them, WY uses the permutation test to estimate the probability distribution of P-values of null distribution, and calibrate the adjusted significance level from the estimated distribution to control FWER. BH controls FDR apart from other three procedures. In applying combinatorial regulation discovery under stress environments in yeast cells, Bonferroni and Holm detected two-motif combinations as a maximum, whereas WY detected 41 four-motif combinations. The combinatorial effects of the four-motif combinations are supported by biological knowledge. We also applied these procedures to the human breast cancer transcriptome. As in the yeast data analysis, Bonferroni and Holm overlooked four-motif combinations that are detected by WY. WY detected 23 four-motif combinations. Even BH could not detect the motif combinations that were detected in yeast and human breast cancer datasets.

These results imply that Bonferroni correction, Holm procedure and BH conceal statistically significant combinatorial regulations due to their over conservativeness.

Our results imply the existence of higher-order statistically significant combinations. However, it is difficult to apply existing multiple testing procedures to discover these combinations because these procedures conservatively adjust the significance level or requires a large amount of computational time or memory space. In the rest of this thesis, we propose efficient multiple testing procedures that can practically enumerate statistically significant motif combinations for any size of combinations considered. In Chapter 3, we propose a theoretically FWER controlling algorithm that computes the upper bound more exactly than Bonferroni. In Chapter 4, we propose an efficient branch-and-bound algorithm to perform WY to enumerate statistically significant motif combinations.

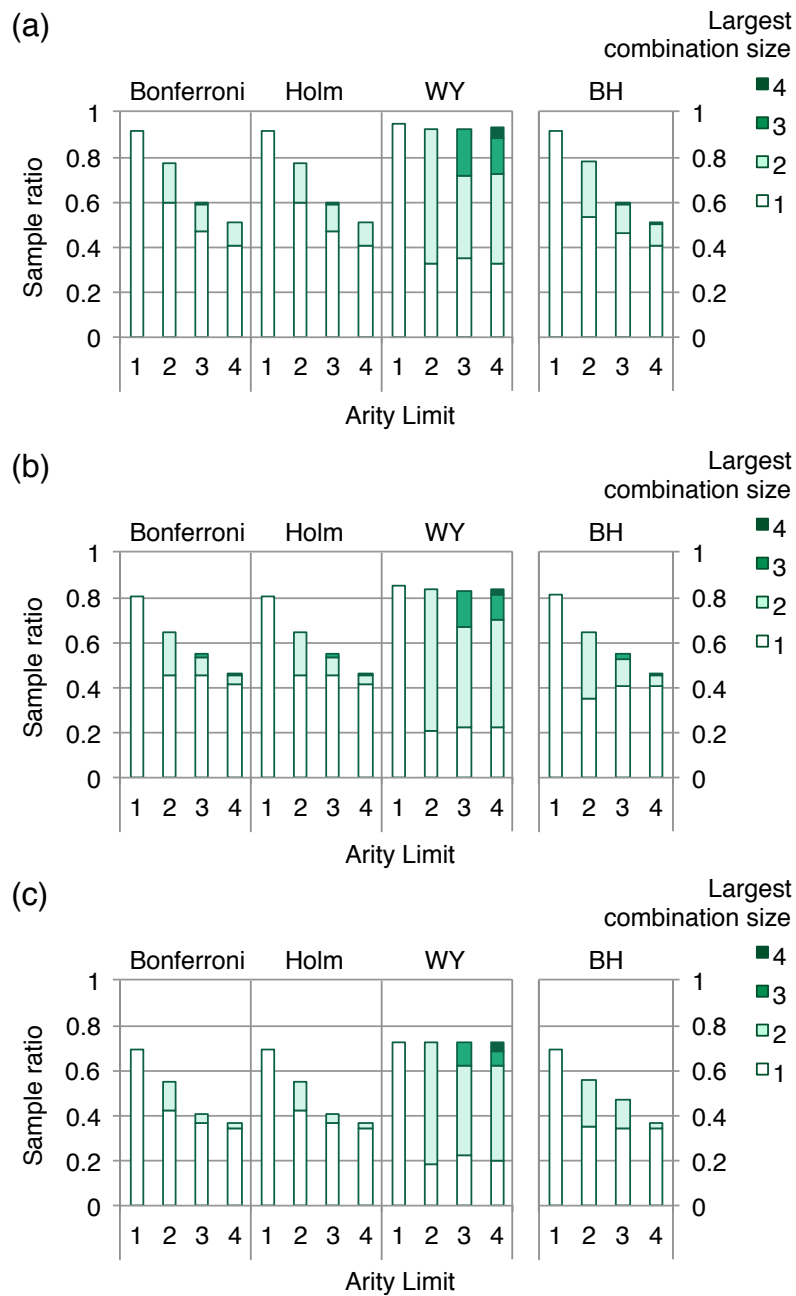


Figure 2.5: Sample ratios in which at least one motif combination was deemed significant over all samples in yeast dataset. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5 .

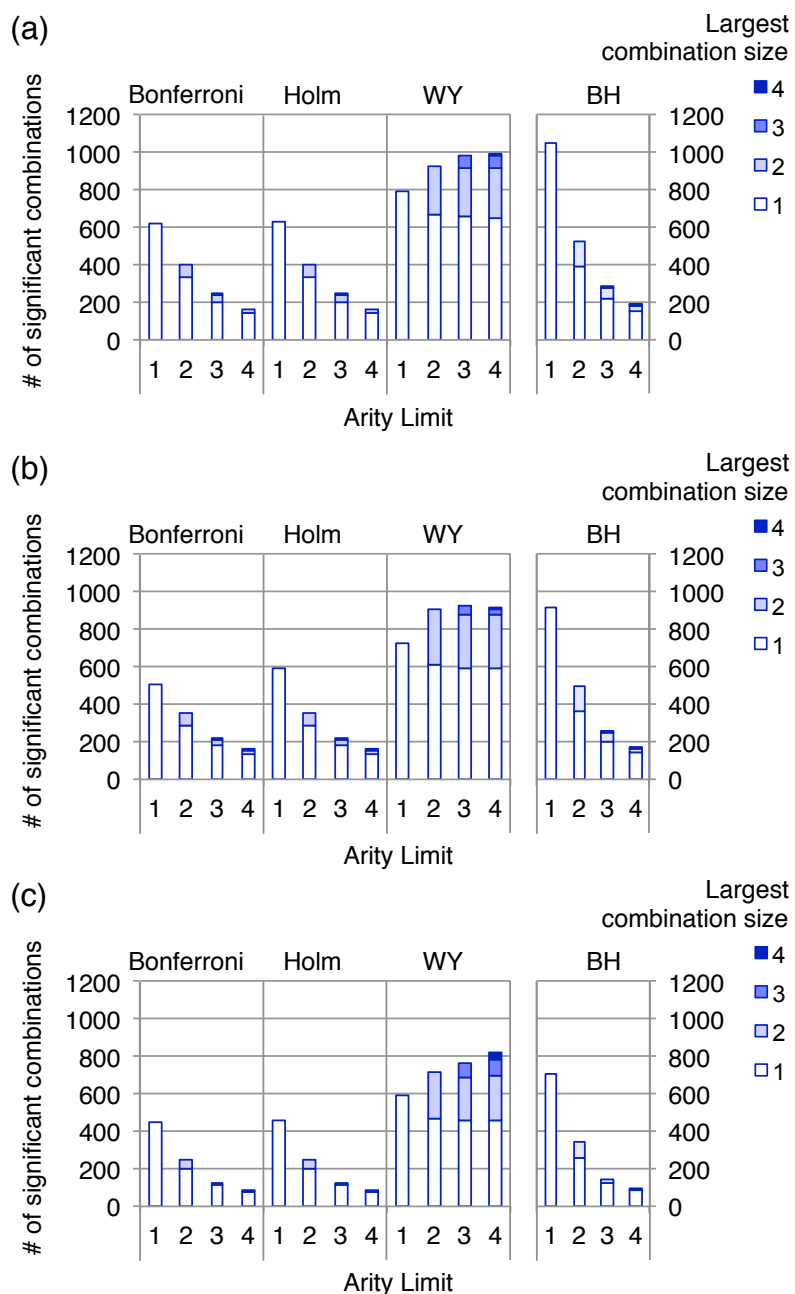
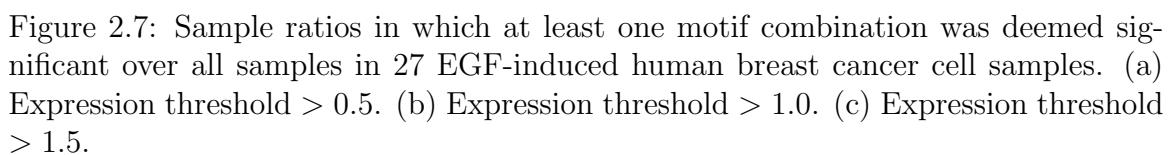


Figure 2.6: The number of motif combinations was deemed significant in 173 yeast stress environments of yeast dataset. (a) Expression threshold > 0.5. (b) Expression threshold > 1.0. (c) Expression threshold > 1.5.



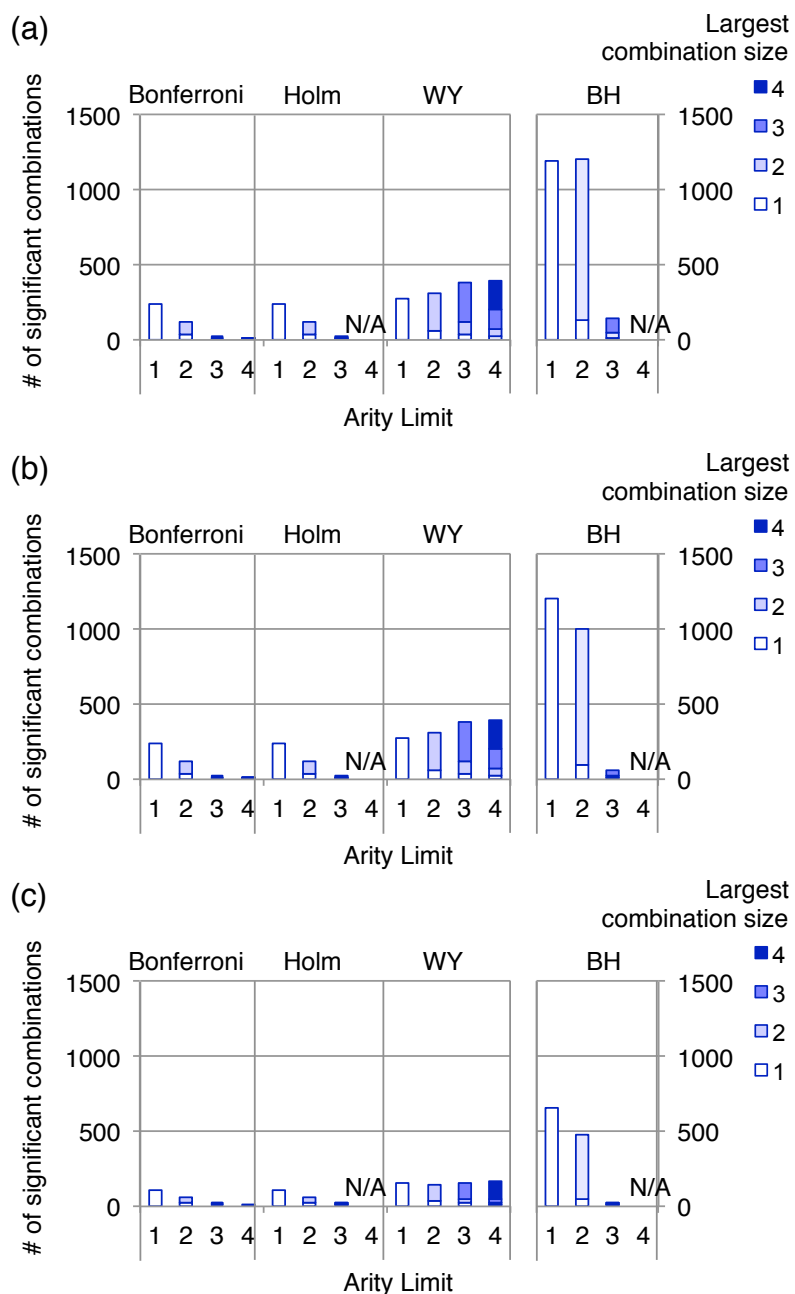


Figure 2.8: The number of motif combinations was deemed significant in 27 EGF-induced human breast cancer cell samples. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5 .

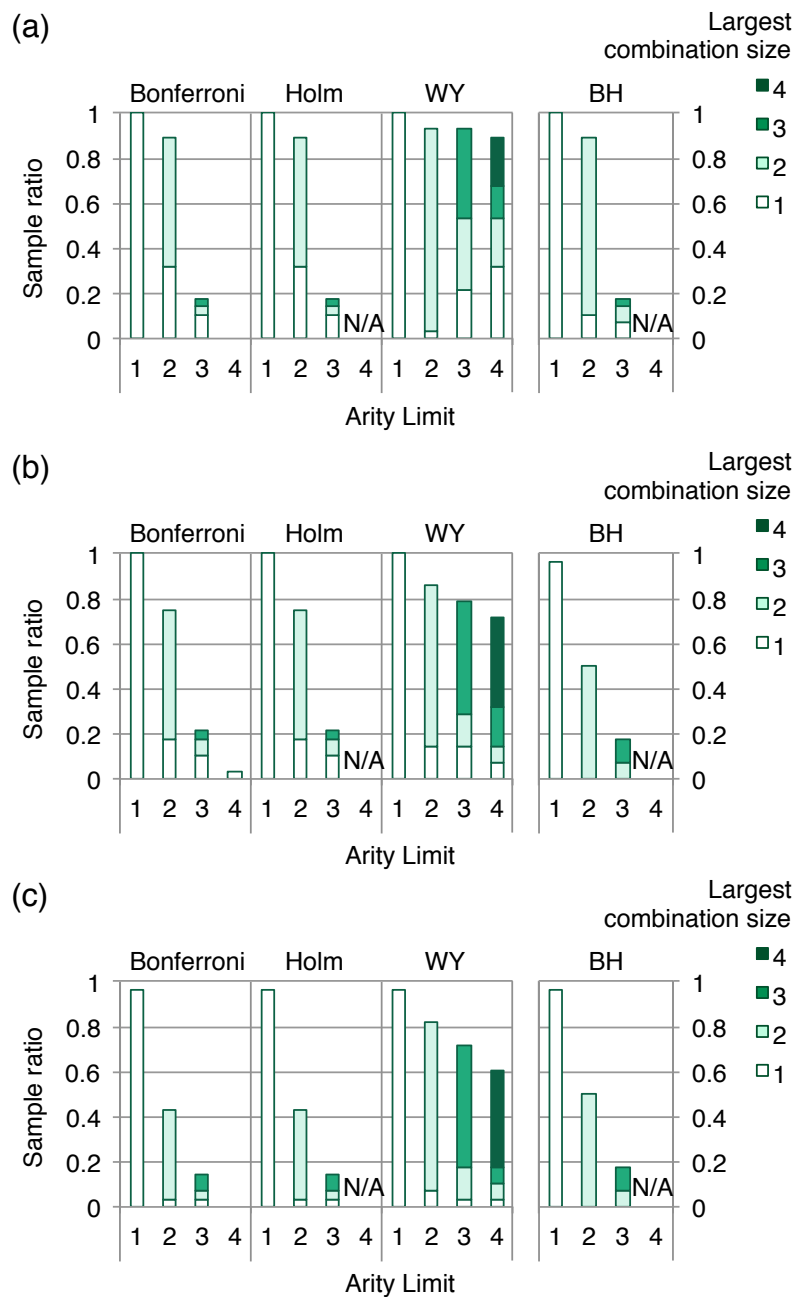


Figure 2.9: Sample ratios in which at least one motif combination was deemed significant over all samples in 28 HRG-induced human breast cancer cell samples. (a) Expression threshold > 0.5 . (b) Expression threshold > 1.0 . (c) Expression threshold > 1.5 .

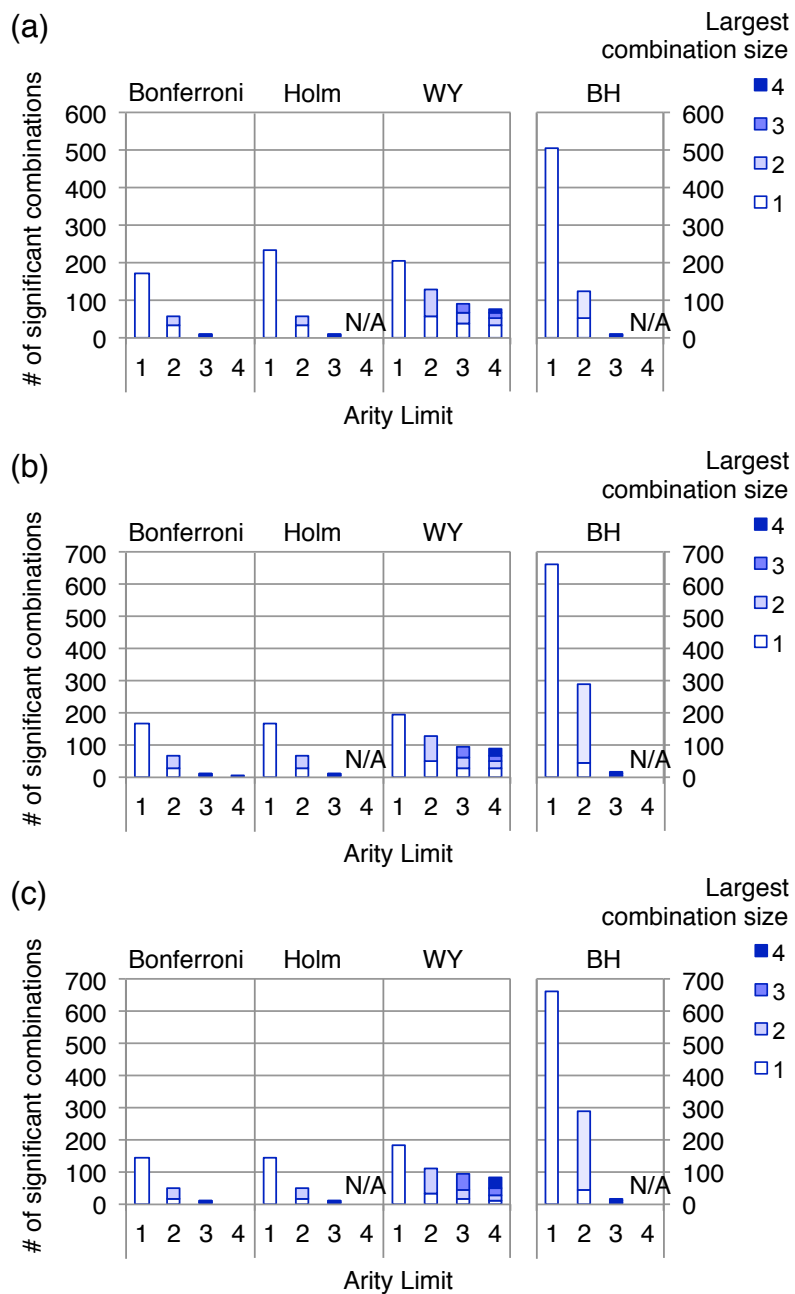


Figure 2.10: The number of motif combinations was deemed significant in 28 HRG-induced human breast cancer cell samples. (a) Expression threshold > 0.5. (b) Expression threshold > 1.0. (c) Expression threshold > 1.5.

Chapter 3

Theoretical Multiple Testing Procedure for Combinatorial Effects Discovery

3.1 Introduction

Bonferroni correction [15], which is a widely used multiple testing procedure to control FWER at most α , computes the adjusted threshold $\delta = \alpha/m$ when we have m tests. Therefore, it is unlikely that a test result is significant by using Bonferroni correction when a large number of tests are performed. However, when any size of combinations of k motifs are considered, the number of tested combinations is $2^k - 1$ and increases exponentially to k . Therefore, the discovery of combinations of higher-arity item combinations (combinations of many items) is extremely unlikely even with more sensitive corrections [101, 46, 44].

The problem with the Bonferroni correction is that the bound on FWER is proportional to the number of tests. However, as Tarone showed in his pioneering paper [91], not every test increases the FWER. Thus, the Bonferroni factor can be improved by deliberately excluding such tests. In this chapter, we propose an efficient branch-and-bound algorithm named *Limitless Arity Multiple testing Procedure (LAMP)* to count the exact number of “testable” combinations and calibrate the Bonferroni factor to the smallest possible value. LAMP lists up significant combinations without any limit, while FWER is rigorously controlled under the threshold.

In Section 3.2, we introduce two properties for a statistical test: the existence of a non-zero lower bound of P-value and decrease of the lower bound value to the number of objects containing combination. The properties enable us to apply a *frequent itemset mining (FIM)* algorithm to count testable combinations. In Section 3.3, we propose LAMP that can be applicable to any statistical tests as long as it satisfies

the properties. In Section 3.5, we compare LAMP with Bonferroni correction using legacy yeast expression data and human breast cancer transcriptome datasets.

This chapter is based on the following publication of the author [93].

3.2 Properties to Count Testable Combinations Efficiently

We will propose an efficient algorithm to enumerate the statistically significant combinations that have higher sensitivity than Bonferroni correction in Section. 3.3. In this section, we introduce properties used in our algorithm.

As shown in Section. 2.3.1, Bonferroni correction conservatively computes the upper bound of FWER as

$$\begin{aligned} FWER &\leq \sum_{i \in \{1, \dots, m\}} \Pr(p_i \leq \delta) \\ &\leq m\delta. \end{aligned} \tag{3.1}$$

Bonferroni correction assumes that all of m tests have the possibility to increase FWER and computes the upper bound of FWER as $m\delta$. However, as Tarone showed in his pioneering paper [91], not every test increases FWER. Thus, the Bonferroni factor may be improved by deliberately excluding such tests. Figure 3.1 illustrates a set of tests excluded. In Figure 3.1(a), the bars are the P-values of combination. Tarone found that the P-value of a test i cannot be smaller than f_i , and the lower bound is plotted by the red point on the bar. If $f_i > \delta$, the combination can never be detected as significant combination. This type of combination is called “untestable” (combinations in the white arrow region), and the other combination is called “testable” (combinations in the red arrow region). Because $\Pr(p_i \leq \delta)$ in Equation 3.1 is 0, FWER does not increase with the addition of untestable combinations. Therefore, the FWER bound can be rewritten as the following,

$$\begin{aligned} FWER &\leq \sum_{i \in \{1, \dots, m\}} \Pr(p_i \leq \delta) \\ &\leq \sum_{j \in \{i | f_i \leq \delta\}} \Pr(p_j \leq \delta) \\ &\leq m'\delta, \end{aligned} \tag{3.2}$$

where $m' = |\{i | f_i \leq \delta \text{ for } i \in \{1, \dots, m\}\}|$, i.e., the number of testable combinations. Hence, the FWER bound can be tightened from δm to $\delta m'$ as illustrated in Figure 3.1(b). This tightening leads to a dramatic reduction of the Bonferroni factor and enables us to discover statistically significant combinations without any arity limit.

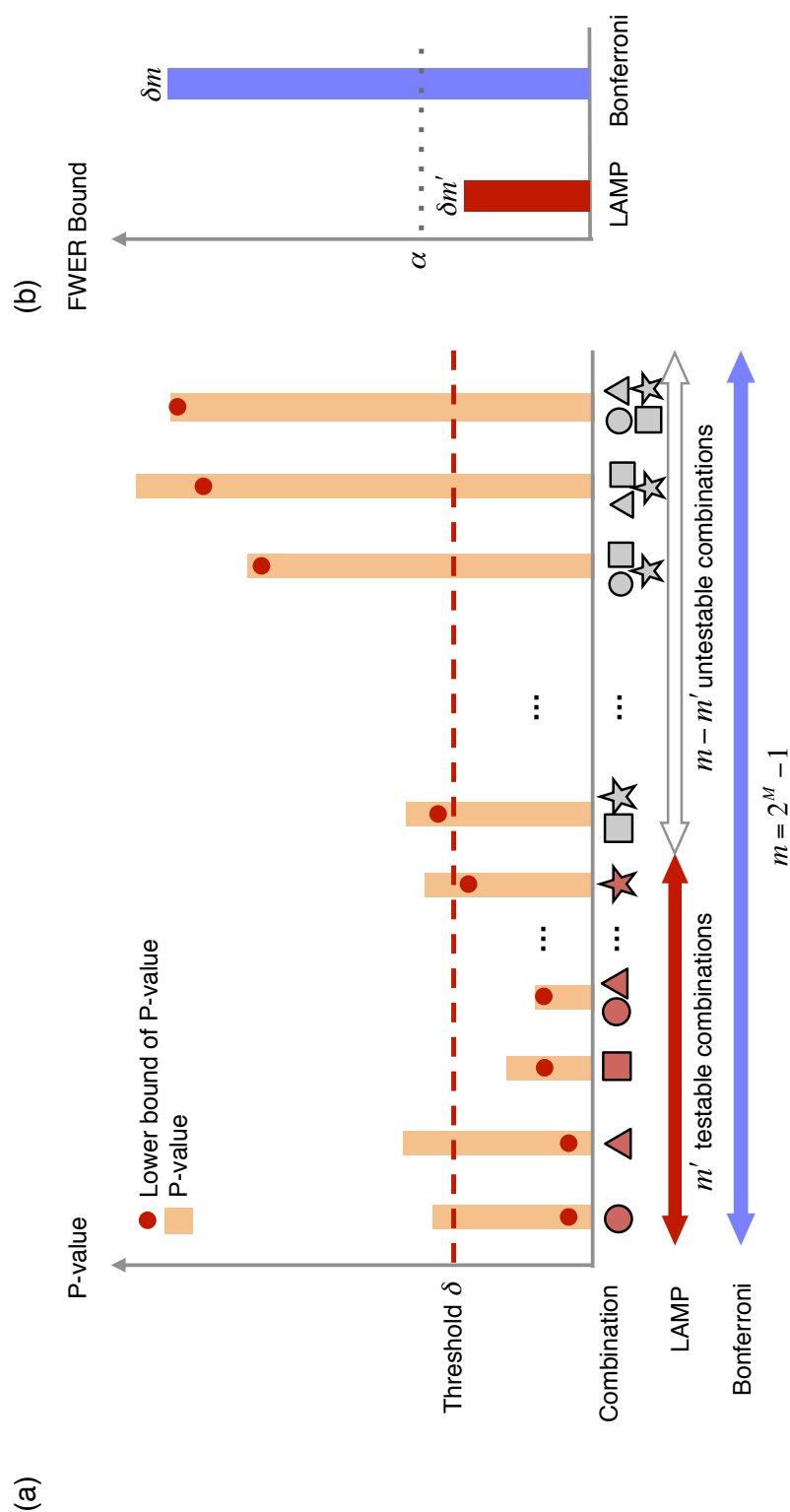


Figure 3.1: Relationships between a threshold δ , the minimum P-value $f(x)$ for each combination, testable and untestable combinations, and the FWER bound.

However, there are two problems when we apply this approach to enumerate statistically significant combinations. One problem is how to efficiently list up testable combinations on δ . The other problem is how to determine δ . Figure 3.2 illustrates the important properties to solve these problems. A red point in the above figure and a blue point in the bottom figure represents the lower bound of P-value of the combination f and the number of objects related to the combination x , respectively. In Section 3.2.1, we show that f only depend on x , and then prove the inverse correlation between $f(x)$ and x . The inverse correlation enables us to solve the first problem. Fixing a lower bound threshold δ , it is possible to determine $\lambda = f^{-1}(\delta)$. Hence, the testable combinations that satisfy $f(x) \leq \delta$ are identical to the set of combinations relating to λ or more objects. Fortunately, such combination can efficiently enumerate by using *Frequent Itemset Mining* (FIM) algorithm [1, 41, 99]. In Section 3.2.2, we introduce the properties for FIM algorithm, and then show an overview of the algorithm. For the second problem, we will propose an algorithm to find the optimal threshold in Section 3.3.

3.2.1 Lower Bound and Monotonicity

We define the lower bound of P-values as the monotonic function. Here, we show the properties for two popular statistical tests: Fisher's exact test for contingency tables, and Mann-Whitney U test for numerical data.

Fisher's Exact Test

Using one-sided Fisher's exact test, the P-value of the combination I is defined as Equation 2.6 by

$$P(I) = \sum_{a_i=a(I)}^{\min\{x,n\}} \frac{\binom{n}{a_i} \binom{N-n}{x(I)-a_i}}{\binom{N}{x(I)}}. \quad (3.3)$$

If we regard that $a(I)$ is a variable, the minimum value of $P(I)$ is achieved at $a_i = \min\{x, n\}$. To use two-sided Fisher's exact test, the lower bound is achieved at the smaller value between the two edge points of $\max\{0, n+x-N\}$ and $\min\{n, x\}$.

Property 1. *For a combination relating x objects, the lower bound of the P-value is described as*

$$f(x) = \begin{cases} \binom{n}{x} / \binom{N}{x} & \text{for } x \leq n, \\ 1 / \binom{N}{n} & \text{otherwise.} \end{cases} \quad (3.4)$$

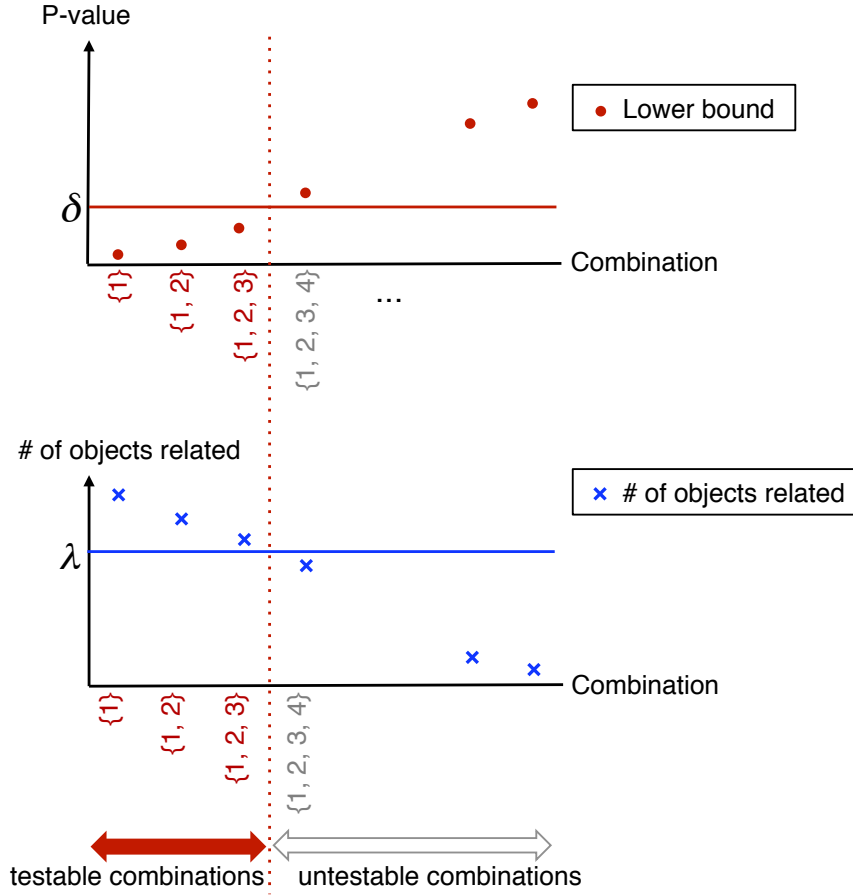


Figure 3.2: Illustration of the properties of testable combinations.

Proof. Suppose I is related by $x = x(I)$ objects. Since Equation 3.3 is formed as the sum of positive terms, the minimum value of $P(I)$ is achieved when the number of the terms is maximal, which is at the edge point $\min\{x, n\}$.

For $x \leq n$, $\min\{x, n\}$ becomes x . Hence, $P(I)$ achieves its smallest value as

$$\binom{n}{x} \bigg/ \binom{N}{x},$$

which is equal to Equation 3.4.

For $x > n$, $\min\{x, n\}$ becomes n , and hence, the smallest $P(I)$ becomes

$$f^*(x) = \binom{N-n}{x-n} \bigg/ \binom{N}{x}.$$

We prove that $f^*(x) \geq f^*(n) = 1 / \binom{N}{n}$.

$$\begin{aligned} \binom{N}{x} &= \frac{x+1}{N-x} \frac{N-x}{x+1} \binom{N}{x} \\ &= \frac{x+1}{N-x} \binom{N}{x+1} \end{aligned}$$

Similarly,

$$\binom{N-n}{x-n} = \frac{x-n+1}{N-x} \binom{N-n}{x-n+1}.$$

Based on these equations,

$$\begin{aligned} f^*(x) &= \binom{N-n}{x-n} / \binom{N}{x} \\ &= \left\{ \frac{x-n+1}{N-x} \binom{N-n}{x-n+1} \right\} / \left\{ \frac{x+1}{N-x} \binom{N}{x+1} \right\} \\ &= \frac{x-n+1}{x+1} \left\{ \binom{N-n}{x-n+1} / \binom{N}{x+1} \right\} \\ &= \frac{x-n+1}{x+1} f^*(x+1) \end{aligned}$$

Because $(x-n+1)/(x+1) \leq 1$, $f^*(x+1) \geq f^*(x)$. Hence, for $x > n$, $f^*(x) < f^*(n)$, and $f^*(x)$ is larger than $1 / \binom{N}{n}$.

Therefore, $f(x)$ is the lower bound of the P-value. □

The following Property 2 confirms the monotonicity of $f(x)$.

Property 2. $f(x)$ decreases to x .

Proof. For $x \leq n$, the following equation holds between x and $x-1$:

$$\begin{aligned} \binom{n}{x} &= \frac{n(n-1) \cdots (n-x+2)(n-x+1)}{x!} \\ &= \frac{n-x+1}{x} \frac{n(n-1) \cdots (n-x+2)}{(x-1)!} \\ &= \frac{n-x+1}{x} \binom{n}{x-1}. \end{aligned}$$

Similarly,

$$\binom{N}{x} = \frac{N-x+1}{x} \binom{N}{x-1}.$$

Based on these equations,

$$\begin{aligned} f(x) &= \binom{n}{x} / \binom{N}{x} \\ &= \left\{ \frac{n-x+1}{x} \binom{n}{x-1} \right\} / \left\{ \frac{N-x+1}{x} \binom{N}{x-1} \right\} \\ &= \left\{ \frac{n-x+1}{x} / \frac{N-x+1}{x} \right\} \left\{ \binom{n}{x-1} / \binom{N}{x-1} \right\} \\ &= \frac{n-x+1}{N-x+1} f(x-1) \end{aligned}$$

From $n \leq N$, $(n-x+1)/(N-x+1) \leq 1$. Hence, $f(x) \leq f(x-1)$ for $x \leq n$.

For $x > n$, $f(x)$ is constant.

Hence, $f(x)$ is a monotonically decreasing function. $\square$

Lower Bound of Mann-Whitney U Test

Here, we prove that Mann-Whitney U test satisfies these the following two properties: the existence of a non-zero lower bound of P-value and decrease of the lower bound to the number of related objects.

Suppose that we have N objects. Each object i relates a set of item $I_i \subseteq \{1, \dots, M\}$ and has a value y_i . Given a combination I , x objects relate all of items in I , and the set of objects is defined as S . Let s_j be the value rank of object j . The Mann-Whitney U statistic of I is defined as

$$U = x(N-x) + \frac{x(x+1)}{2} - \sum_{j \in S} s_j. \quad (3.5)$$

U is approximates a normal distribution with a mean μ and variance σ as $x(N-x)/2$ and $\{x(N-x)(N+1)/12\}^{1/2}$, respectively. By checking the corresponding statistics table for a normal distribution, the P-value of U can be computed.

The lower bound of the P-value is achieved when the values of objects in S are larger than the others, or they are smaller than the others.

Property 3. *For the combination related in x objects, the lower bound of the P -value computed by Mann-Whitney U test is achieved at*

$$f(x) = \begin{cases} 1 / \binom{N}{x} & \text{for } x \leq N/2, \\ 1 / \binom{N}{\lfloor N/2 \rfloor} & \text{otherwise.} \end{cases} \quad (3.6)$$

Proof. The P -value is achieved at the lower bound when values of x objects in S are larger than the others, or they are smaller than the others. The probability to appear the case is described as $1 / \binom{N}{x}$, and takes the minimum value when $x = \lfloor N/2 \rfloor$. $\square$

This value decreases with an increase of x .

Property 4. *$f(x)$ in Equation 3.6 decreases to x .*

Proof. $1 / \binom{N}{x}$ decreases with increase of x for $1 \leq x \leq N/2$, and takes the minimum value when $x = \lfloor N/2 \rfloor$. Therefore, $f(x)$ decreases to x . $\square$

3.2.2 Enumeration of Testable Combinations with a FIM Algorithm

The following Properties 5 and 7 enable us to use the threshold λ on $x(I)$ instead of δ , and FIM algorithm for listing the testable combinations with δ .

Property 5. *For combinations I and $I \subseteq I'$, $x(I) \geq x(I')$.*

Proof. Given an itemset I that includes k items and that is related by $x(I)$ objects. Let i be an item which is not in I and let us generate I' by adding i to I . The number of objects relating all items in I' is smaller or equal to that of I . $\square$

Property 6. *For combinations I and $I' \supseteq I$, $f(x(I)) \leq f(x(I'))$.*

Proof. Because $x(I) \geq x(I')$, $f(x(I)) \leq f(x(I'))$. $\square$

The monotonicity of f is useful for listing all combinations within the testable combinations.

Property 7. *Assuming that $f(\lambda) \leq \delta < f(\lambda - 1)$, the testable combinations with δ are identical to the combinations related by λ or more objects.*

Proof. When $x(I) \geq \lambda$ for a combination I , I is the testable combination because $f(x(I)) \leq f(\lambda)$ from Property 2. Otherwise, I is the untestable combination with δ . Hence, Property 7 holds. $\square$

We can use λ in FIM algorithm to list testable combinations with λ with Property 7. Since FIM algorithm requires a large amount of computing time for a small λ , we start from a large λ and gradually decrease λ to find the optimal δ .

Overview of Frequent Itemset Mining Algorithm

We here provide an overview of FIM algorithms. The frequent itemset mining problem is to list all patterns that appear frequently in a dataset. One typical application of this problem is to find sets of items (itemsets) that customers are likely to purchase on a given trip to the store. The itemset that satisfies the condition is called a *frequent itemset*. In this study, we used the algorithm to list up combinations, which has the same meaning as itemset, that are associated with λ objects.

The FIM algorithm generates the combinations from lower-arity itemsets. It generates all single items first, and items that are related λ or more objects are selected. Let $\mathbf{L}_1$ be the set of the selected items. In LAMP, all items in the $\mathbf{L}_1$ are testable because their lower bounds of the P-value are less than the candidate P-value. Then, the algorithm generates 2-item combinations by combining two items in $\mathbf{L}_1$ and selects 2-item combinations that are contained by λ or more objects. Since the number of relating objects of a 2-item combination is smaller or equal to both single items in the 2-item, we generate the 2-item only from the items in $\mathbf{L}_1$. 2-item combinations that target λ or more genes are stored in $\mathbf{L}_2$.

Let $\mathbf{L}_k$ be the list of k -item combinations that are related by λ or more objects. From the list, $\mathbf{L}_{k+1}$ can be generated. The FIM algorithm repeats this process until $\mathbf{L}_k$ relates no combinations. By combining all $\mathbf{L}_k$, all combinations that are related by λ or more objects are listed without the generation of entire combinations.

Based on this idea, the first algorithm to efficiently list up the frequent itemsets called *Apriori* was introduced by Agrawal and Shrikant [1]. Han *et al.* [41] introduced more efficient algorithm under large memory environment to list up frequent itemsets, called FP-tree. It generates a tree of items in frequent itemsets, and then lists up the frequent items using the tree. Uno *et al.* [99] proposed a memory efficient algorithm, called LCM. We used the Uno's implementation to find frequent sets of motifs.

3.3 LAMP Algorithm

In Section 3.2, we showed that the testable combinations that satisfy $f(x) \leq \delta$ are identical to the set of combinations relating to λ or more objects. Such combinations can be efficiently enumerated by using FIM algorithm [1, 41, 99], however it is

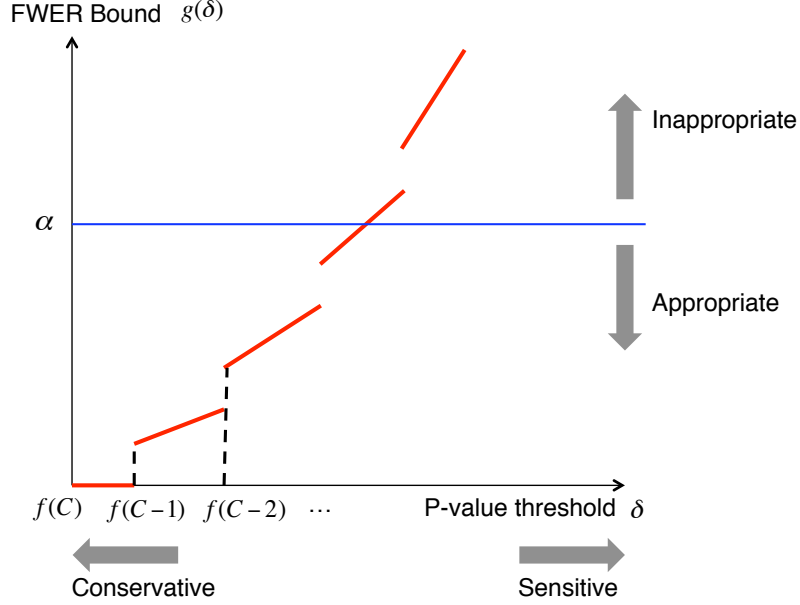


Figure 3.3: Illustration of the relationship between δ and $g(\delta)$.

unknown how to determine λ to calibrate δ to control FWER under α . Therefore, in this section, we propose an algorithm called *Limitless Arity Multiple-testing Procedure* (LAMP) which finds optimal λ and efficiently computes δ .

Fixing a P-value threshold δ , as shown in Equation 3.2, FWER is bounded by $g(\delta) = m'\delta$, where m' is the number of testable combinations. Because the domain of x is discrete, $f^{-1}(\delta)$ is defined as λ such that $f(\lambda) \leq \delta < f(\lambda - 1)$. Because of discontinuities of $f^{-1}(\delta)$, the FWER bound is a step-wise linear function of δ (Figure 3.3). In Figure 3.3, the red lines show $g(\delta)$ and form a piece-wise linear function. The gaps are located at $f(\lambda)$, where $\lambda = C, C - 1, \dots, 1$. α is the predetermined significance level. The intersection point of the two lines indicates the optimal correction.

The P-value threshold δ must be determined so that the FWER bound is calibrated to a predetermined level α . To this aim, we construct the function $g(\delta)$ by drawing line segments in Figure 3.3 one by one from bottom to top (Figure 3.4). To draw the first segment, FIM algorithm is performed to find the combinations that are related by C or more objects. If the number of such combinations is m_C , the line segment is described as

$$g(\delta) = \delta m_C, f(C) \leq \delta < f(C - 1). \quad (3.7)$$

FIM algorithm is repeated by decrement the parameter one by one, and the procedure stops if one of the end points of the line segment goes beyond α . Algorithm 5

Algorithm 5 LAMP

Input: N objects with M items, the relations between objects and items, FWER upper limit α

Output: enumerate the significance combinations

```

1:  $\mathcal{I} \leftarrow \{\}$ ,  $\lambda \leftarrow \min\{x_{\max}, n\}$ ,  $m_\lambda \leftarrow 0$ ,  $\delta \leftarrow 0.0$ .
   // compute the optimal  $\delta$ 
2: while  $\lambda \geq 1$  do
3:    $\mathcal{I} \leftarrow$  combinations that relate to  $\lambda$  or more objects (frequent itemsets)
4:    $m_\lambda \leftarrow |\mathcal{I}|$ ,  $\delta \leftarrow \alpha/m_\lambda$ 
5:   if  $f(\lambda)m_\lambda \leq \delta < f(\lambda - 1)m_\lambda$  then
6:     break
7:   else if  $f(\lambda)m_\lambda > \alpha$  then
8:      $\lambda \leftarrow \lambda + 1$ 
9:     break
10:  else
11:     $\lambda \leftarrow \lambda - 1$ 
12:  end if
13: end while
   // Enumerate the significant combinations
14:  $\mathcal{I} \leftarrow$  combinations that relate to  $\lambda$  or more objects
15:  $m_\lambda \leftarrow |\mathcal{I}|$ ,  $\delta \leftarrow \alpha/m_\lambda$ 
16: return the set of combinations  $I$  where  $P(I) \leq \delta$  for  $I \in \mathcal{I}$ 

```

shows the whole procedure. Because FIM algorithm requires a large amount of computing time for a small λ , LAMP starts from a large value of $\lambda = \{x_{\max}, n\}$, where $x_{\max} = \max\{x(I) \text{ for } I = \{i\} \text{ where } i = 1, \dots, M\}$ (line 1 of Algorithm 5). The FIM algorithm lists combinations that are related by λ or more objects (Fig 3.4(a)). The number of combinations is m_λ . When δ is varied between $f(\lambda)$ and $f(\lambda - 1)$, the FWER bound is calculated between $f(\lambda)m_\lambda$ and $f(\lambda - 1)m_\lambda$. Because the FWER bound $f(\lambda - 1)m_\lambda$ is smaller than α , m_λ is not sufficiently large to be a corrected significance level (Fig 3.4(b)). Therefore, we replace λ with $\lambda - 1$ (line 11 of Algorithm 5). When λ is reduced, m_λ becomes large. If $f(\lambda - 1)m_\lambda$ is still smaller than α , we change λ to a smaller value and repeat the process (Figures 3.4(c), (d)). The procedure stops if one of the end points of the line segment goes beyond α (line 5 of Algorithm 5 and Figures 3.4(e), (f)). Finally, λ is gradually decreased until the intersection point of $f(\lambda)$ and δ is found as shown in Fig. 3.3.

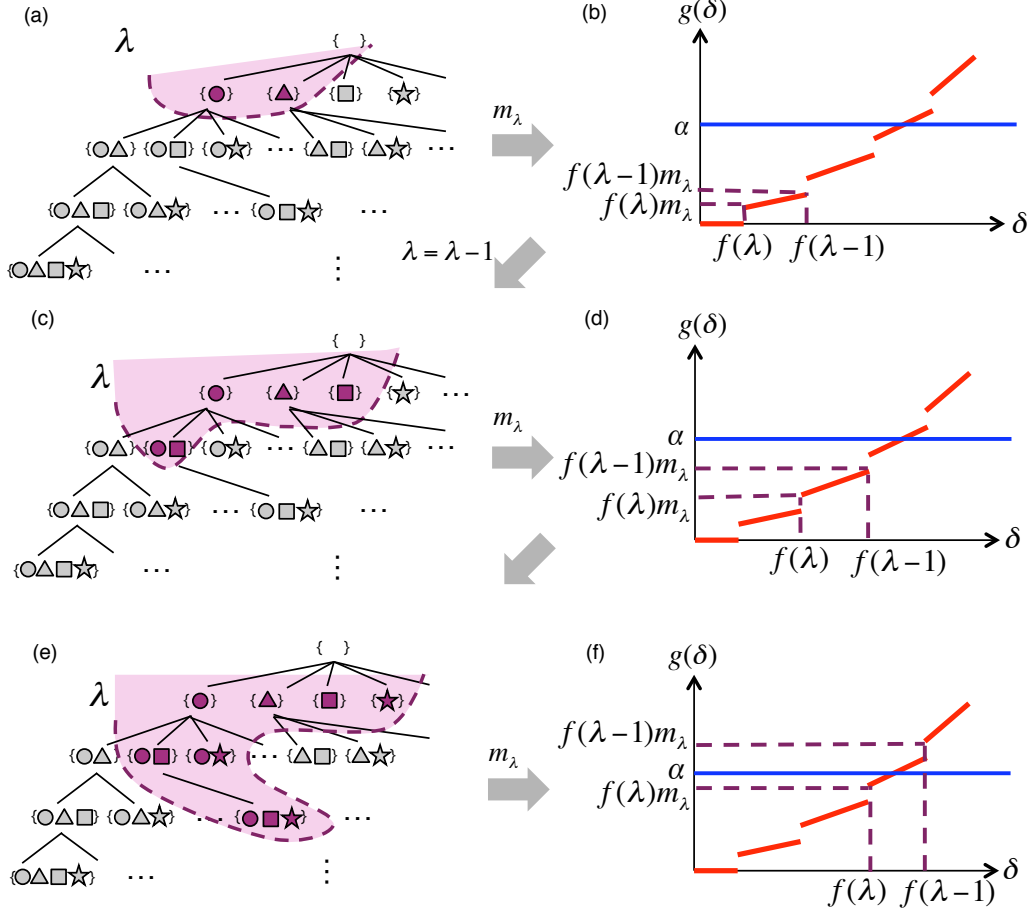


Figure 3.4: Illustration of the algorithm of the LAMP.

3.4 Time Complexity of LAMP

In this section, we present the time complexity of LAMP. Figure 3.5 shows an overview of LAMP algorithm and indicates the time complexity when Fisher's exact test is used to compute the P-value.

The running time of LAMP is divided into two parts: finding the adjusted significance level δ and enumerating the significant combinations from the testable itemsets. Given λ , LAMP enumerates the combinations that have frequencies larger than or equal to λ . Since this enumeration is an NP-hard problem [109], it requires $O(2^M N)$. After the enumeration, LAMP computes $f(\lambda)$ to check the FWER for δ . When we use Fisher's exact test, $f(\lambda)$ is defined by Equation 3.4, and the computational com-

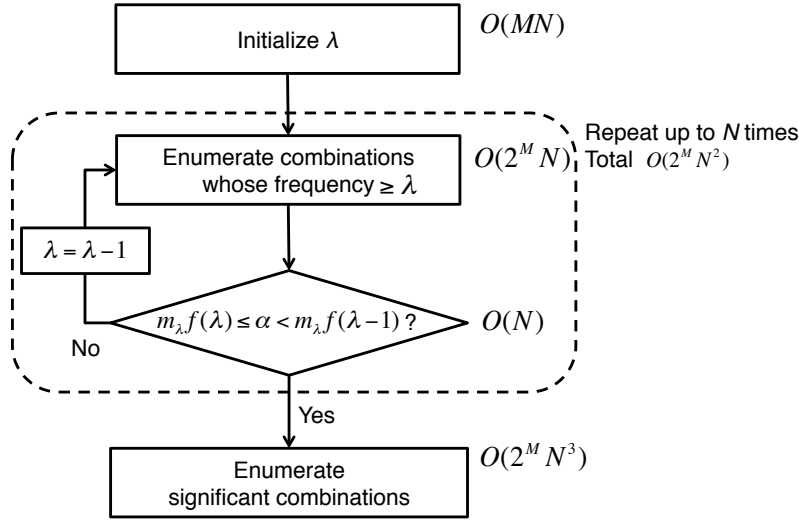


Figure 3.5: Time complexity of LAMP when Fisher's exact test is used.

plexity is $O(N)$ to compute $\binom{n}{\lambda}$ and $\binom{N}{\lambda}$. Because this process is repeated up to N times, the total time becomes $O(\sum_{\lambda=1}^N (2^M N)) = O(2^M N^2)$.

After the computation of δ , we enumerate the set of combinations whose P-values are less than or equal to δ from the set of testable combinations. The time complexity of Fisher's exact test (Equation 3.3) is $O(N^3)$. Because the test is performed up to $2^M - 1$ times, the total time for the enumeration of significant combinations is $O(2^M N^3)$.

In total, when Fisher's exact test is used, $O(MN) + O(2^M N^2) + O(2^M N^3) = O(2^M N^3 + MN)$ is required for LAMP in the worst case. For the other statistical test such as Mann-Whitney U test, the time complexity becomes at least $O(2^M N)$ because enumerating combinations with frequencies larger than or equal to λ is independent of the selection of the statistical test. Although the time complexity increases exponentially to M in the worst case, the computational time is feasible in most cases because the number of testable combinations is much smaller than $2^M - 1$ and efficient algorithms can be used to enumerate the combinations [1, 41, 99].

3.5 Experimental Results

We examined the performance of LAMP on the real biological data that were used datasets in Section 2.5: the yeast and human breast cancer dataset. One-sided Fisher's exact test was employed, and the FWER is controlled under $\alpha = 0.05$

by the Bonferroni method and LAMP. The FWER is defined with respect to each environment separately. All programs are written in Python 2.7 except for FIM algorithm in line 3 in Algorithm 5. For FIM program, we use LCM [99] due to its fast calculation. All experiments were conducted on a machine with 2 AMD Opteron processors at 2.3 GHz with 32 GB of RAM running RedHat Linux.

3.5.1 Comparison of Detection Powers

Here, we compare the detection powers of LAMP with that of Bonferroni correction.

Comparison with Yeast Transcriptome Dataset

The number of testable combinations m' was optimized for each environment, and we summarize them in Table 3.1. For any threshold, the values of m' of LAMP are smaller than those of Bonferroni correction. Thus, LAMP detected larger and more combinations than Bonferroni correction. For example, in the result of threshold 1.5, the value of m' was in the range of 233 to 702 when the maximum arity was not limited, whereas the original Bonferroni factor m was 5,253 when maximum arity was two, which indicates that LAMP has a higher sensitivity than the Bonferroni correction. A total of 326 motif combinations were discovered, which is 1.70-fold higher than the number of motif combinations that were discovered using the Bonferroni correction (with a maximum arity of two).

Figures 3.6 and 3.7 shows the sample ratios in which at least one motif combination is significant and the number of statistically significant motif combinations over 173 yeast stress environments. We varied the arity limit and computed the number of samples in which at least one sample was deemed significant for the specified arity limit. In Figures 3.6 and 3.7, each bar indicates the sample ratio that a significant combination was detected and the number of statistically significant motif combinations for a specific arity limit, respectively. Within the bar, the colors indicate the largest arity. Comparing the sample ratio in Figure 3.6, LAMP found at least one motif combination from at least half the samples even when the arity limit was removed for any threshold. However, the number of samples detected using Bonferroni correction decreased with the increasing arity limit, and Bonferroni correction found significant combinations from only a few samples when the arity limit was removed. LAMP discovered at most four-motif combinations, whereas Bonferroni correction found at most two-motif combinations. In Figures 3.7, LAMP found more than 200 motif combinations even when the arity limit was removed for any expression thresholds for Fisher's exact test, whereas Bonferroni correction found 9 combinations under any threshold.

Table 3.1: Statistics of the number of significant combinations of Yeast dataset.

Threshold (log2 ratio)		0.5	1.0	1.5	2.0
Avg. # of up-regulated genes		990.86	363.62	171.36	95.29
# of detected samples	LAMP	157	135	113	95
	Bonferroni (≤ 2)	134	111	95	78
	Bonferroni (≤ 3)	103	95	71	57
# of significant combination	LAMP	513	449	326	235
	Bonferroni (≤ 2)	308	264	192	137
	Bonferroni (≤ 3)	187	176	108	82
Stats. of LAMP	Maximum arity	3	3	4	3
	Avg. of m'	243.30	366.43	446.30	498.34
	Min. of m'	132	183	233	290
	Max. of m'	554	556	702	703

Comparison with Human Breast Cancer Transcriptome Dataset

The number of testable combinations m' was optimized for each condition, and we summarize them in Table 3.2 for EGF- and Table 3.3 for HRG-induced cells, respectively. For any threshold for gene expression levels, the values of m' of LAMP is less than 7,247,034 for EGF and 7,635,557 for HRG, while that of Bonferroni correction is 10,428,793 when combinations up to three motifs were tested. Thus, the sensitivity of LAMP was higher than that of Bonferroni correction. For example, when we used 1.0 as the expression level threshold, our method discovered a total of 43 and 23 motif combinations in the EGF- and HRG-induced cells, respectively, including one eight-motif combination, while the Bonferroni correction (with a maximum arity of three) detected only 19 and 7 combinations, respectively, and found at most one three-motif combination.

Figures 3.8 and 3.9 show that sample ratios in which at least one motif combination is significant in 27 EGF- and 28 HRG-induced human breast cancer cell samples. Each bar indicates the sample ratio that a significant combination was detected and the number of statistically significant motif combinations for a specific arity limit, respectively. Within the bar, the colors indicate the largest arity. In all results in Figures 3.8 and 3.9, LAMP detected combinations as significant more than Bonferroni correction when the same arity limit was used for both procedures. When the arity limit becomes larger, the sample ratio of Bonferroni correction sharply decreases due to an exponentially increasing number of combinations to be tested, while that of LAMP decreases more gradually. These results indicate that LAMP has a higher sensitivity than the Bonferroni correction. For example, in Figure 3.8(b), LAMP

Table 3.2: Statistics of the number of significant combinations of EGF induced sample.

Threshold (log2 ratio)		0.5	1.0	1.5	2.0
Avg. # of up-regulated genes		2359.81	1226.67	777.04	491.26
# of detected samples	LAMP	17	20	16	10
	Bonferroni (≤ 2)	21	23	22	16
	Bonferroni (≤ 3)	8	14	13	6
# of significant combination	LAMP	46	43	23	24
	Bonferroni (≤ 2)	78	70	41	38
	Bonferroni (≤ 3)	20	19	14	8
Stats. of LAMP	Maximum arity	5	8	7	8
	Avg. of m'	1,742,870	3,669,596	5,180,881	7,169,767
	Min. of m'	1,459,869	2,700,718	3,750,336	5,238,103
	Max. of m'	2,700,718	3,750,336	5,238,103	7,247,034

detected at most one motif combinations in 74.1% samples and found an eight-motif combination when maximum arity was not limited, whereas Bonferroni detected motif combinations in only 51.9% samples combinations even when maximum arity was three and found at most three-motif combinations.

To compare the sensitivities for each condition, we count the total number of significant combinations in 27 EGF- and 28 HRG- induced human breast cancer cell samples, and present the results in Figures 3.10 and 3.11. Each bar indicates the number of statistically significant motif combinations for the specified arity limit. Within the bar, the colors indicate the largest arity. In Figure 3.10, whereas Bonferroni correction detected at most three combinations even when up to five-motif combinations were considered (Figure 3.10(b)), LAMP found at least 15 combination when the maximum arity was not limited (Figure 3.10(d)). In both Figures 3.10(b) and (d), eight-motif combinations were detected by LAMP, which could not found by Bonferroni correction due to its conservativeness. In Figure 3.11, Bonferroni correction detected no combination as significant when up to five- or more motif combinations were considered, while LAMP found at least nine combinations including one five-motif combination and two six-motif combinations when the maximum arity was not limited (Figure 3.11(d)). There results indicate that LAMP can detect significant combinations from each condition more than Bonferroni correction.

Table 3.3: Statistics of the number of significant combinations of HRG induced sample.

Threshold (log2 ratio)		0.5	1.0	1.5	2.0
Avg. # of up-regulated genes		4189.54	2337.36	1277.00	758.68
# of detected samples	LAMP	18	12	10	7
	Bonferroni (≤ 2)	25	21	12	12
	Bonferroni (≤ 3)	5	6	4	1
# of significant combination	LAMP	26	23	16	11
	Bonferroni (≤ 2)	40	43	26	19
	Bonferroni (≤ 3)	5	7	7	1
Stats. of LAMP	Maximum arity	6	5	5	6
	Avg. of m'	579,531	1,790,220	3,800,657	5,561,464
	Min. of m'	359,149	1,174,198	2,874,627	3,982,295
	Max. of m'	1,174,108	2,874,627	5,542,428	7,635,557

3.5.2 Comparing Power to Control the Ratio of False Discoveries

To demonstrate that LAMP improves the detection of combinatorial regulations, we computed empirical FWERs from randomly permuted datasets and compared the FWERs with the Bonferroni correction. In permutation, gene expression levels were randomly shuffled. If at least one false discovery occurred in a permuted dataset, the dataset was counted as a family-wise error. An estimate of the empirical FWER is computed as the fraction of family-wise errors among the 1,000 datasets. The significance level α was set to 0.05. This procedure is repeated 10 times, and Figure 3.12 reports the average of the estimated values and their standard deviation. Figure 3.12(a) shows the empirical FWERs with the heat shock environment of yeast datasets in which the largest combination was detected by LAMP. The expression threshold was 1.5. The x - and y -axis indicate the arity limits and empirical FWERs, respectively. Although the Bonferroni correction never found significant combinations from the permutation data with a maximum arity of three or more, the average FWER computed by LAMP was 0.013, which demonstrates that the detection power of LAMP is potentially higher than that of the Bonferroni correction. We also computed the empirical FWER using human cancer cell data in which the largest eight-motif combination was detected by LAMP: 0.5 nM EGF for 15 min and the expression threshold to 1.0. Figure 3.12(b) presents the results. As for the yeast data analysis, the results showed indicated that the detection power of LAMP is higher

than that of the Bonferroni correction.

FDR is often employed instead of FWER to improve the sensitivity of statistical tests [11, 87]. With permuted expression data in Figures 3.12(a) and (b), we measured the average number of discoveries using the procedures by Benjamini-Hochberg (BH) [11] and Storey and Tibshirani (ST) [87], where FDR is controlled under 0.05. In addition, LAMP was applied to the same datasets, where FWER is controlled under 0.05. Figures 3.13(a) and (b) report the results. Because any procedure controlling FWER also controls FDR [54], LAMP controls FDR under 0.05 as well. For arity limits larger than two, the numbers of discoveries of BH and ST were smaller by orders of magnitude than that of LAMP, suggesting that these methods are still too conservative for an extremely large number of tests.

3.5.3 Running Time

We checked the efficiency of computational time of LAMP with the permuted datasets in Figures 3.12(a) and (b). The average computational time was computed over 1,000 permuted datasets, and this process was repeated 10 times. Figures 3.14(a) and (b) report the average of the times and their standard deviation. Note that the y-axis is on a log scale. Both figures indicate that LAMP is sufficiently fast compared with the other correction methods.

As shown in Figure 3.14(a), LAMP required less than one second to analyze the yeast data even without an arity limit, whereas the computing time of Bonferroni correction increases exponentially in the number of tests. For the human data in Figure 3.14(b), LAMP required 2.3 hours without an arity limit, which was faster than the Bonferroni method (maximum arity ≤ 3). The estimated computing time for the analysis of all of the possible combinations of the yeast data using Bonferroni method is more than 10^{20} years. Thus, the branch-and-bound algorithm dramatically reduced the computing time to enumerate significant combinations. LAMP is also faster than other sensitive corrections such as Šidák procedure [101] and Holm [46] procedure. These corrections require listing and sorting an intractable number of tests when high-arity combinations are considered. LAMP avoids this calculation using an FIM technique. Another approach for avoiding false discoveries in multiple testing is the estimation of the error rate with random sampling-based methods [30, 106, 96, 54]. This approach would require an extraordinarily large amount of time because a large number of hypotheses must be investigated to detect high-arity combinations.

Although LAMP requires $O(2^M N)$ time to enumerate the combinations relating to λ or more objects, this task can be efficiently performed by using an itemset mining algorithm [1, 41, 99], which can enumerate the combinations in $O(m_\lambda MN)$, where m_λ represents the number of combinations enumerated. We investigated the performance of FIM algorithm upon decreasing λ by using human breast cancer dataset, in which

the largest combination was detected (0.5 nM EGF-induced MCF-7 cells for 15 min), and the results are shown in Figure 3.15 and Figure 3.16. Figure 3.15 shows the number of combinations enumerated when λ was varied. When λ becomes small, m_λ increases exponentially. For example, m_7 , m_6 and m_5 were 1.40, 2.04 and 2.98 times larger than the optimal value of LAMP ($m_8 = 7,500,674$), respectively. Figure 3.16 shows the running time for FIM algorithm as λ was decreased. Similar to Figure 3.15, the running time grows exponentially when λ becomes small. Therefore, LAMP starts from a large λ and gradually decreases λ to find the optimal δ .

3.6 Biological Confirmation of Detected Motif Combinations

In this section, we confirm the higher-arity combinations that detected by LAMP in biological knowledge.

3.6.1 Yeast Transcriptome dataset

Our analysis of the yeast gene expression dataset yielded statistically significant but previously uncharacterized higher-arity motif combinations. Two four-motif combinations, which is the largest arity combination in our experiment with the yeast dataset, were detected in two stress environment when the expression threshold for Fisher's exact test was set to 1.5. We list the detected combinations in these environments in Tables 3.4 and 3.5. The bold numbers are significant after the correction.

One four-motif combination was observed in the heat shock environment (shift from 25 to 37 degrees) as shown in Table 3.4. The optimal Bonferroni factor m' in LAMP is 167, while the Bonferroni factors are 5,253, 176,953, and 4,426,528 for maximum arities of 2, 3, and 4, respectively. LAMP detected six-motif combinations, including a four-motif combination, while the Bonferroni method found three singletons when the arity was limited to two and two singletons when the limit was set to three or four, failing to find any combinatorial regulations. The members of the four-motif combination are bound by *PHD1*, *SUT1*, *SOK2* and *SKN7*, respectively [43]. Figure 3.17 shows the statistical significance of the four-motif combination. Surprisingly, no single motif in the four-motif combination had a statistically significant P-value. According to the literature, *PHD1*, *SUT1* and *SKN7* are recruited by *Tup1* [69, 42], a repressor under growth conditions, and the other gene *SOK2* is a candidate *Tup1* recruiter [42], which suggests that all of the TFs bound to the motif jointly regulate genes within the *Tup1*-characterized pathway. Yu *et al.* [112] performed a comprehensive scanning of two-motif combinations with the expression data that includes our data. They predicted a part of this four-motif combination (*PHD1* and *SKN7*) indeed,

but failed to discover the whole combination. As shown in Figure 3.17(a), seven genes are found to have the four-motif combination in their regulatory regions, of which five genes (*GAT2*, *GID8*, *HAP4*, *MSN4* and *MGA1*) are up-regulated. All of the proteins translated from the five genes are located in the nucleus, and three of them are related to RNA polymerase II according to Saccharomyces Genome Database [23], which suggests that this motif combination plays a role in the upstream regulation of the heat shock response.

Another four-motif combination was found on raffinose medium as shown in Table 3.5. The optimal Bonferroni factor m' in LAMP is 548 while the Bonferroni factors are 5,253, 176,953, and 4,426,528 for maximum arities of 2, 3, and 4, respectively. The four-motif combination consists of the binding motifs of the heme activators *HAP1*, 2, 3 and 5 (adjusted p-value = 0.00754). Yu *et al.* [112] also detected a part of this combination (*HAP3* and 5), but overlooked the higher-arity combination.

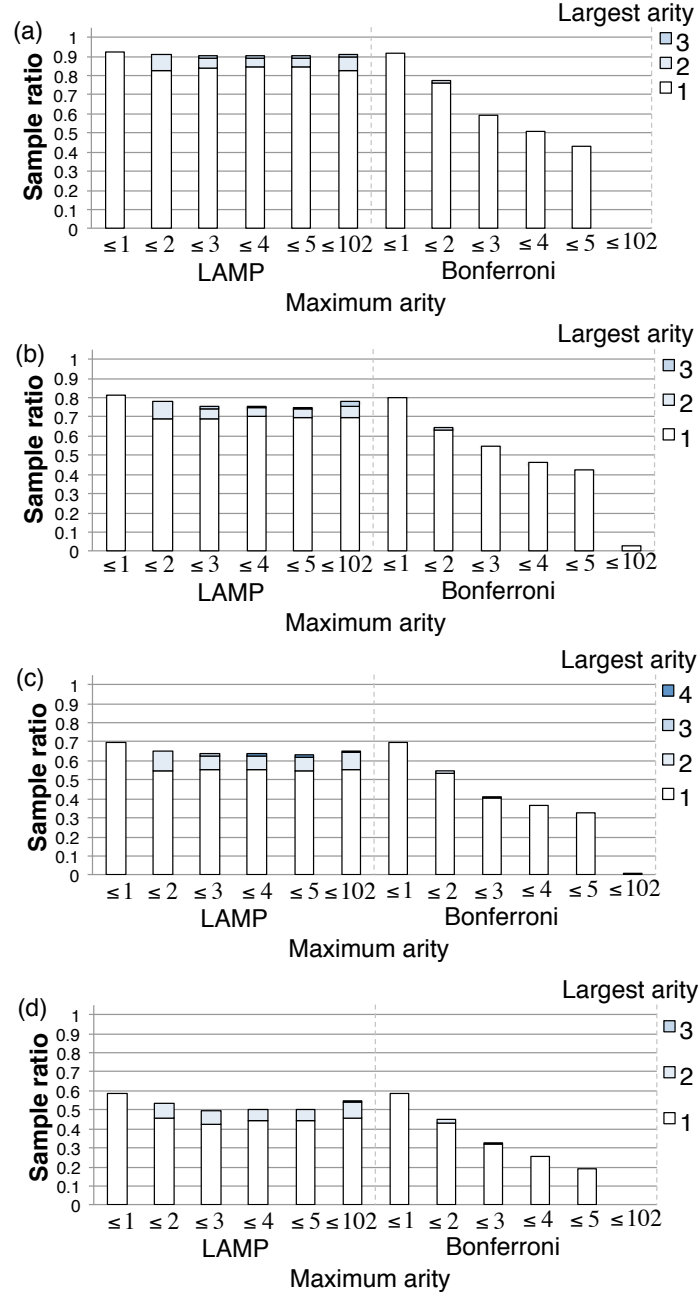


Figure 3.6: Sample ratios in which at least one motif combination is significant over 173 yeast stress environments (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.

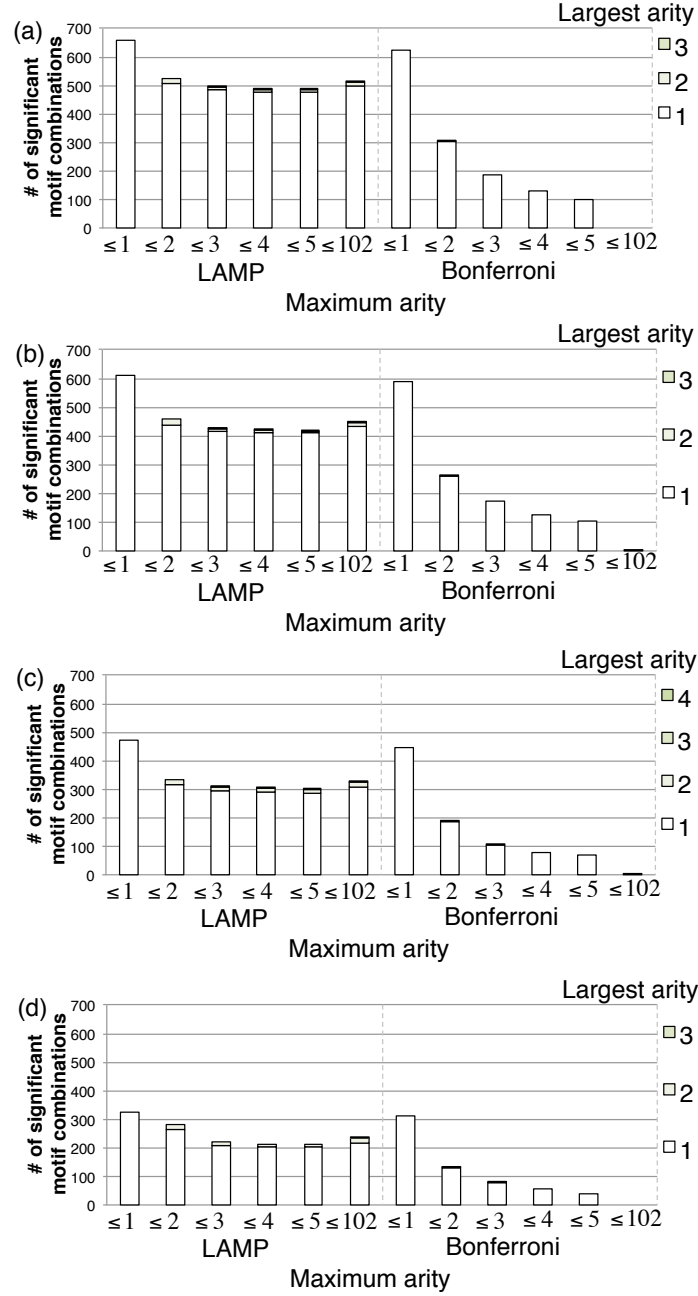


Figure 3.7: The number of statistically significant motif combinations over 173 yeast stress environments (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.

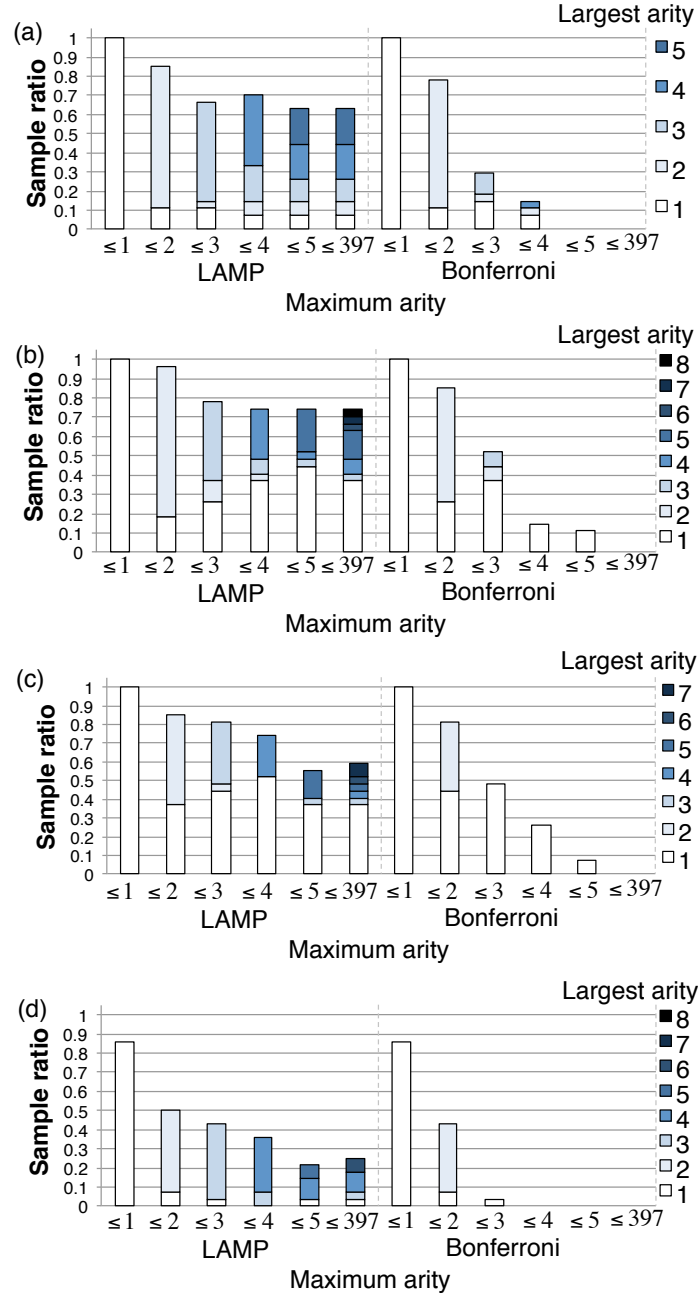


Figure 3.8: Sample ratios in which at least one motif combination was deemed significant in 27 EGF-induced human breast cancer cell samples. (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.

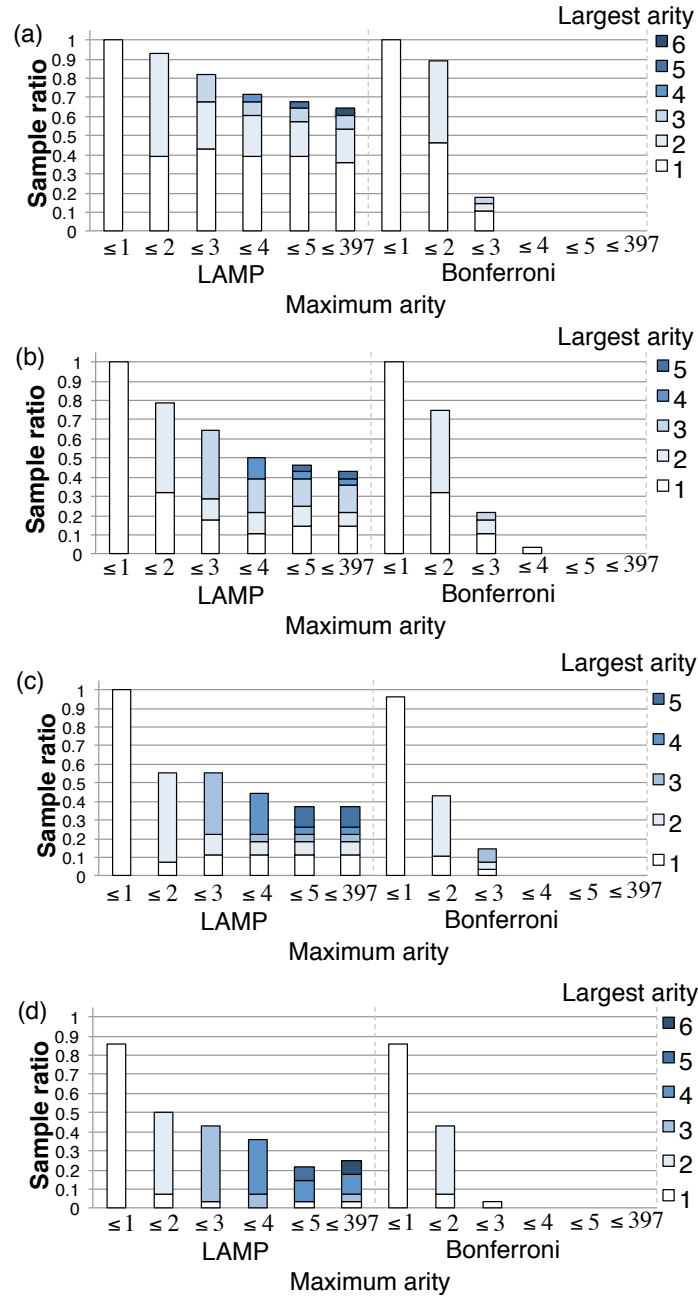


Figure 3.9: Sample ratios in which at least one motif combination was deemed significant in 28 HRG-induced human breast cancer cell samples. (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.

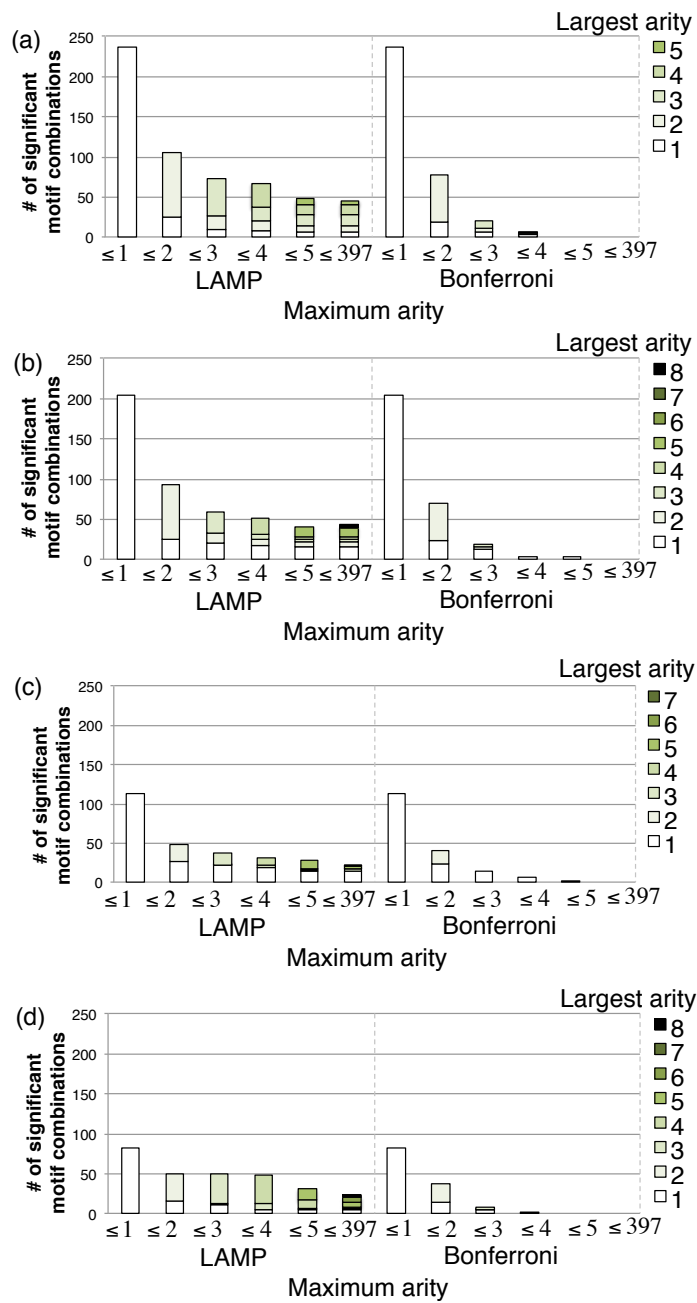


Figure 3.10: The number of statistically significant motif combinations in 27 EGF-induced human breast cancer cell samples. (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.

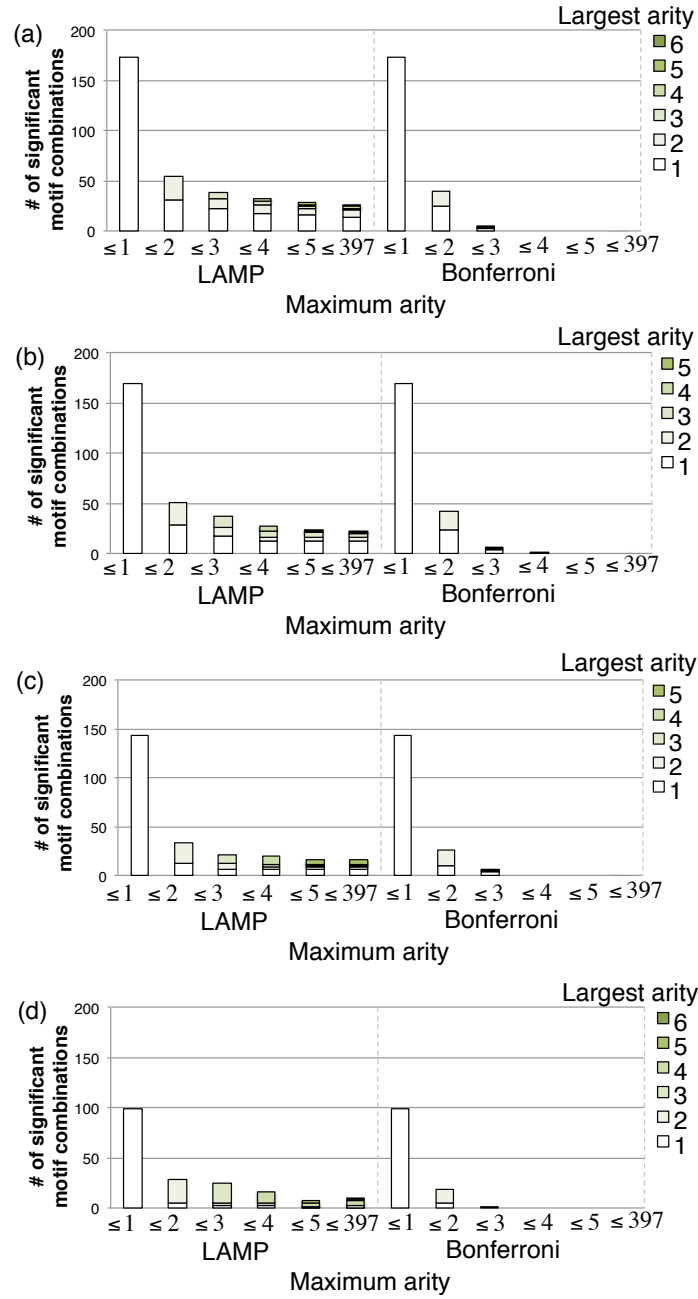


Figure 3.11: The number of statistically significant motif combinations in 28 HRG-induced human breast cancer cell samples. (a) Expression threshold = 0.5. (b) Expression threshold = 1.0. (c) Expression threshold = 1.5. (d) Expression threshold = 2.0.

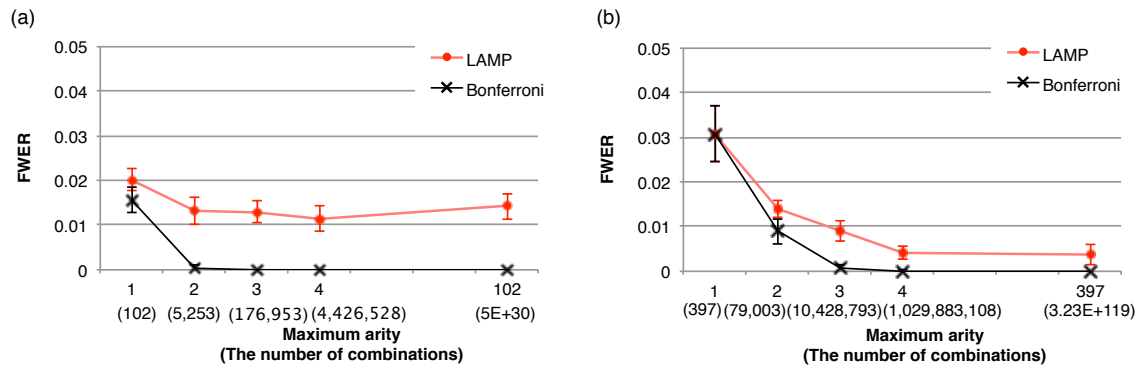


Figure 3.12: Comparison of the empirical FWERs.

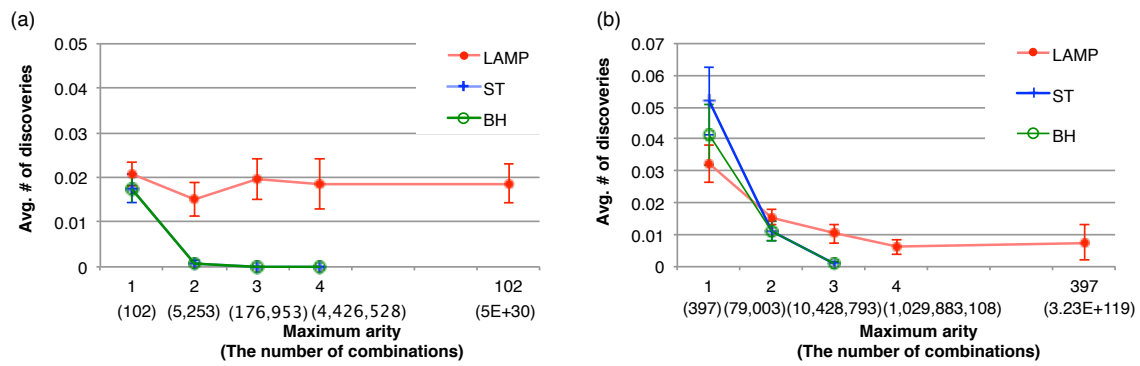


Figure 3.13: Comparison of the average numbers of false discoveries.

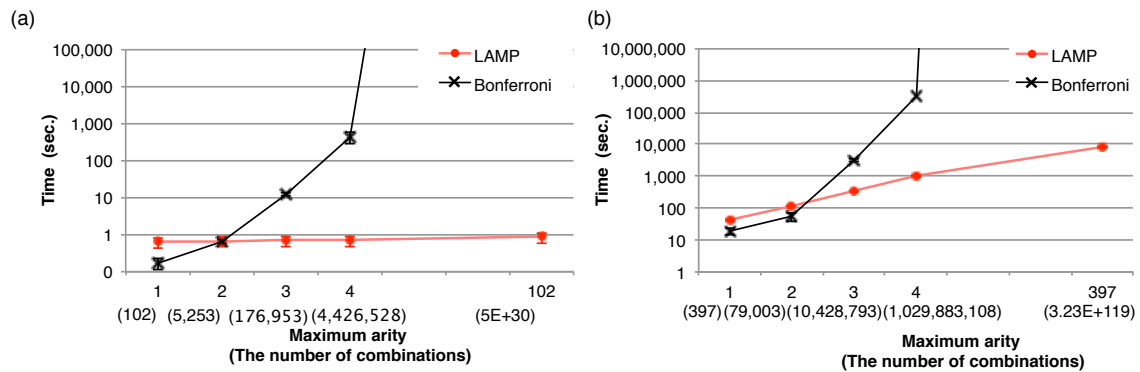


Figure 3.14: Comparison of the average numbers of false discoveries.

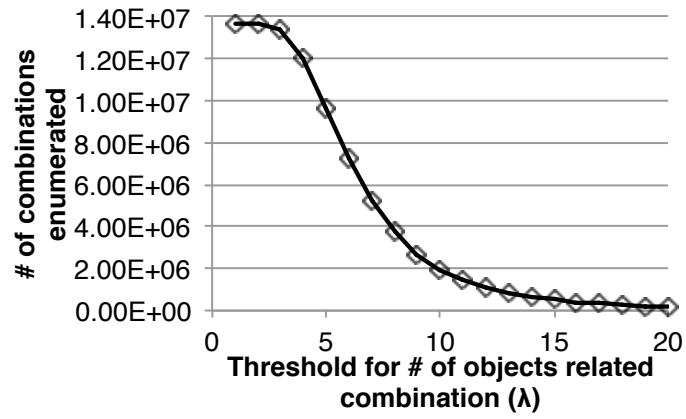


Figure 3.15: Threshold for the frequency of combinations and the number of combinations enumerated.

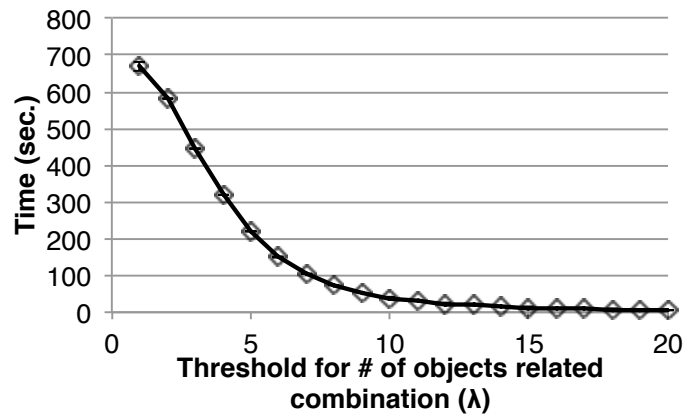


Figure 3.16: Threshold for the frequency of combinations and running time for enumerate frequent combinations.

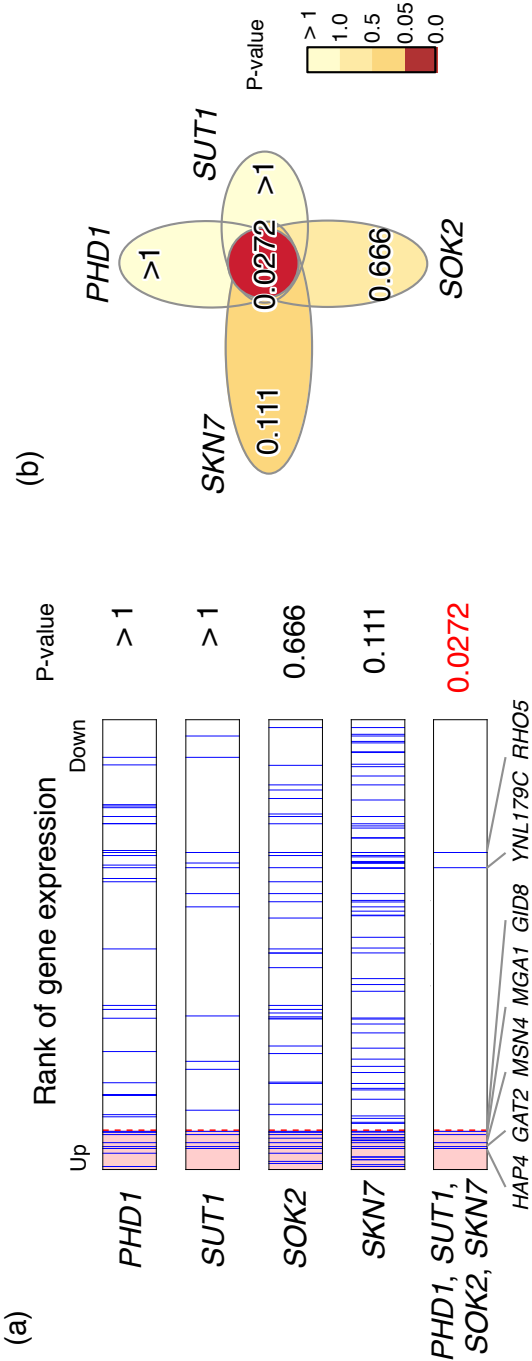


Figure 3.17: Statistical significant of the four-motif combination in a yeast heat shock condition.

Table 3.4: Statistically significant motif combinations under a yeast heat shock condition.

Motif combination	# of target genes	Adjusted P-values		
		LAMP	Bonferroni correction	
			Max. arity ≤ 2	≤ 3 ≤ 4
HSF1	69	$4.41 \cdot 10^{-24}$	$7.64 \cdot 10^{-23}$	$2.57 \cdot 10^{-21}$ $6.44 \cdot 10^{-20}$
MSN2	21	$3.73 \cdot 10^{-11}$	$6.46 \cdot 10^{-10}$	$2.18 \cdot 10^{-08}$ $5.45 \cdot 10^{-07}$
MSN4	24	0.000532	0.00923	0.311 > 1
MSN2, MSN4	65	0.00839	0.1454014	> 1 > 1
SKO1	6	0.00839	0.145	> 1 > 1
SNT2	18	0.0192	0.333	> 1 > 1
PHD1, SUT1, SOK2, SKN7	7	0.0272	-	- > 1
PHD1, SKN7	11	0.0410	0.7114498	> 1 > 1

Table 3.5: Statistically significant motif combinations under a yeast raffinose medium condition.

Motif combination	# of target genes	Adjusted P-values		
		LAMP	Bonferroni correction	
			Max. arity ≤ 2	≤ 3 ≤ 4
HAP4	27	0.00143	0.0137 0.461	> 1
HAP1, HAP2, HAP3, HAP5	2	0.00754	-	> 1
HAP1	45	0.0116	0.111006	> 1 > 1
HAP1, HAP5	3	0.0226	0.216	> 1 > 1

3.6.2 Human Breast Cancer Cells Dataset

Although both epidermal growth factor (EGF) and heregulin (HRG) belong to the ErbB family receptor ligands [25], EGF induces the proliferation of MCF-7 human breast cancer cells while HRG induces their differentiation (the accumulation of lipid droplets). A transcriptomics study [67] suggested the existence of unidentified mechanisms by which the same family ligands produce different patterns of early responsive gene expression. Using LAMP to reanalyze the gene expression profile of breast cancer cells that were treated with these two ligands, we identified statistically significant motif combinations that are related to proliferation- and differentiation-specific pathways.

The largest eight-motif combination was found when we set the expression threshold to 1.0. Figures 3.18- 3.20 show the statistically significant high-arity combinations in EGF- and HRG-induced human breast cancer cells.

The largest eight-motif combination was found after 15 min of treatment with 0.5 nM EGF (adjusted P-value = 0.0137). Surprisingly, no single motif had an adjusted P-value lower than 1 (Figure 3.18(a)). Furthermore, none of the combinations with arity below six were statistically significant (Figure 3.18(b)). According to existing knowledge, the eight-motif combination is highly relevant to the feature of MCF-7 cells and induced ligand type. A ChIP-seq study confirmed that some TFs bind to C/EBP, OCT1 and FOXO4 motifs in the estrogen receptor (ER) positive cells [19], and MCF-7 cells are indeed ER-positive. Because the TF has to be expressed to be functional, we check the expression levels of the TFs known to bind the motifs. Table 3.6 and Figure 3.21 show that the list of TFs that can bind to the motif in the eight-motif combination and expression levels of them, respectively. In Figure 3.21, each line corresponds to one motif and shows the expression levels of a TF that can bind to the motif. When single motif is recognized by multiple TFs such as *C/EBP* and *FOXO4*, we selected the TF whose expression level under control (0 min) among TFs that can bind to the motif is the highest one. The name on each line is the motif name, and the TF name is in parentheses. For example, since *CEBPB* can bind to the C/EBP-motif (Table 3.6) and the expression of *CEBPB* is higher than that of *C/EBP* at 0 min, the expression levels of *CEBPB* are described in Figure 3.21 as a blue line. When the TF has multiple probes in Affimetrix chip, the probe that has the highest expression level at 0 min is used. The motif CTTTAAR is not shown because no TF bound to the motif is identified yet. Among the TFs known to bind to the motifs, *CEBPB*, *GR*, *JUN*, *FOXA1*, and *FOXK2* are substantially expressed. The transcription factor *GR* is highly regulated in MCF-7 cells by estrogen treatment [55]. High expression of *JUN* increases biologic aggressiveness of MCF-7 cells [83].

The second largest combination (adjusted P-value = 0.0209) was a seven-motif combination identified in the cells treated 1.0 nM EGF for 60 min. As shown in Figure 3.19(a), no single motif has a significant p-value except TATA (adjusted P-

value = 0.0462), which is known to be activated in the MCF-7 cells [20]. As with the eight-motif combination, we show the list of TFs that can bind to the motif in the seven-motif combination and the expression levels in Table 3.7 and Figure 3.22, respectively. Among the expressed TFs binding to the motifs, *MEF2* is of particular interest, because it works with *NFAT* [61]. The combination regulates 10 genes in total (*BDNF*, *CHD2*, *EYA1*, *HOXA2*, *HAS2*, *IL1RAPL1*, *PDZRN4*, *PHOX2B*, *PPARGC1A* and *RBFOX1*), five of which are related to differentiation (*BDNF*, *EYA1*, *HOXA2*, *PHOX2B* and *PPARGC1A*), and two of which (*HOXA2* and *PHOX2B*) are annotated as negative regulators of differentiation (Gene Ontology GO:0045665). In addition, Figure 3.22(c) shows that expression levels of the 10 downstream genes in the EGF-induced cells are higher than those measured in the HRG-induced cells for the first 60 min. These results imply that the combination plays a negative regulatory role for differentiation in response to EGF.

A five-motif combination was found after 30 min of treatment with 10 nM HRG which is shown in Figures 3.20, 3.23 and Table 3.8. The combination targets 11 genes (*BDNF*, *CYR61*, *EGR2*, *EHF*, *ETV1*, *HOXB3*, *HOXB6*, *ID2*, *IL1RAPL1*, *PDZRN4* and *SOX5*). Their average expression levels were higher than those observed in the EGF-induced cells, which suggests that the combination committed the cell toward an HRG-specific regulatory pathway.

LAMP identified several high-arity combinations that would have been missed by single motif or motif pair scans, providing a deeper understanding of the ligand-specific pathways. About a half of the genes targeted by the discovered high-arity combinations are related to the regulation of transcription, LAMP may contribute to capture key regulations that determine the cell fate. Notice that the biological views presented here are only hypothetical. Extensive biological investigation is necessary to confirm the hypotheses.

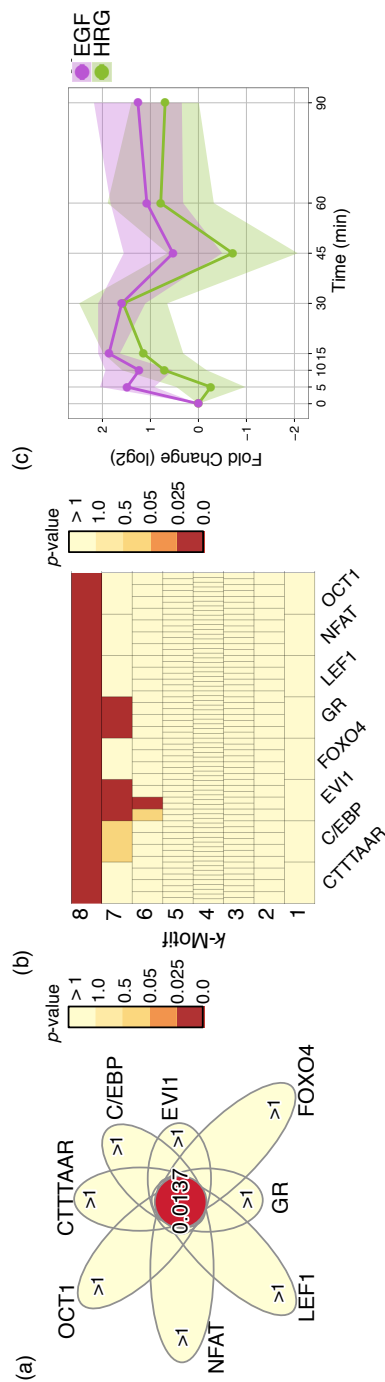


Figure 3.18: Statistical significant of the eight-motif combination in the human breast cancer cells treated with 0.5 nM EGF for 15 min.

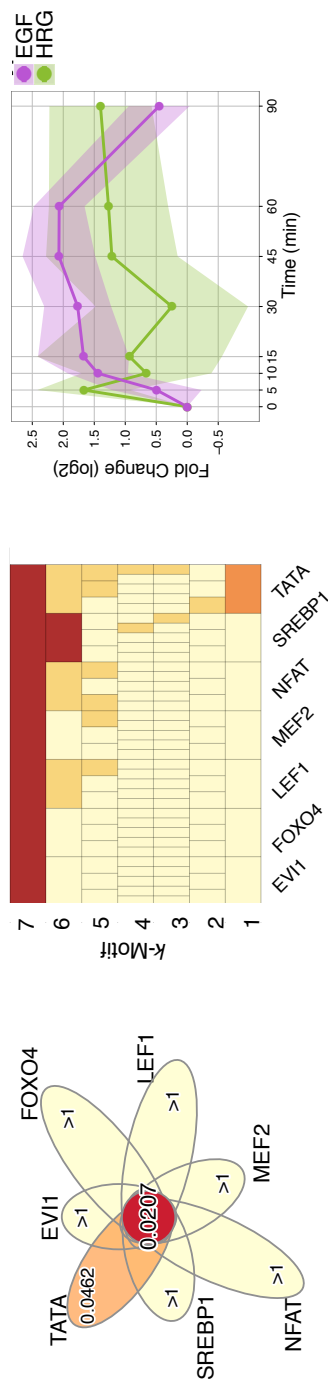


Figure 3.19: Statistical significant of the seven-motif combination in the human breast cancer cells treated with 1.0 nM EGF for 60 min.

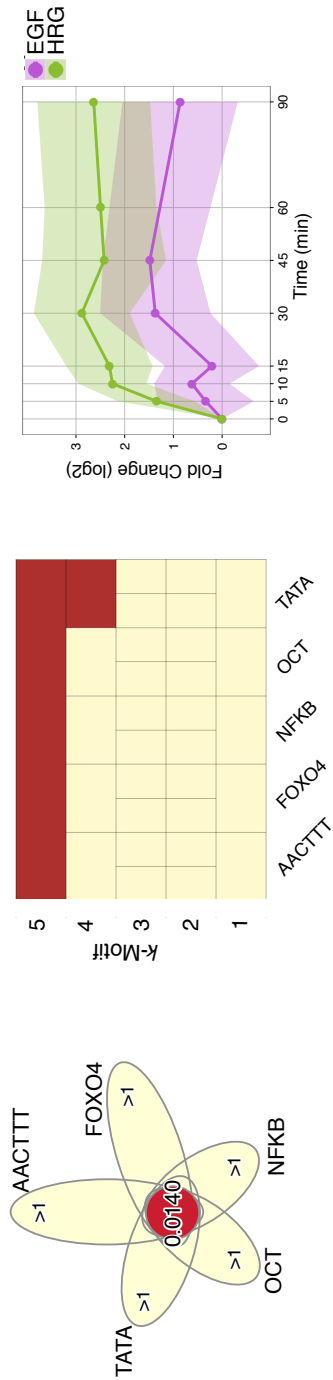


Figure 3.20: Statistical significant of the five-motif combination in the human breast cancer cells treated with 10.0 nM HRG for 30 min.

Table 3.6: The list of TFs that can bind to the motifs in Figure 3.18.

Motif	TFs that bind to the motif
CTTTAAR	-
C/EBP	<i>C/EBP, CEBPB</i>
EVI1	<i>EVI1</i>
FOXO4	<i>FOXA1, FOXA2, FOXO1, FOXO4, HMGA1</i>
GR	<i>GR</i>
LEF1	<i>LEF1</i>
NFAT	<i>ELOF1, FOXK2, NFATC1, NFATC2, NFATC3, NFATC4</i>
OCT1	<i>GATA1, JUN, NFYA, NFYB, NFYC, OCT1, OCT2, SP1</i>

3.7 Conclusion

In this chapter, we proposed LAMP, which counts the exact number of testable combinations using an FIM algorithm. LAMP calibrated the adjusted significance level to a larger value than Bonferroni correction and could compute FWER closer to the upper limit. We applied LAMP to the yeast and human breast cancer transcriptome datasets and discovered statistically significant combinations of as many as four and eight binding motifs, respectively.

Although LAMP offered improvement over the conservativeness of Bonferroni correction, a gap remains between the estimated FWER and the given upper limit (Figures 3.12(a) and (b)) because LAMP assumes independence among tests in Equation 3.2 as well as the other procedures that theoretically control FWER. A solution for theoretical approach, LAMP possibly may be improved by incorporating the method by Roth [75], which uses a non-integer correction factor. Another solution for this problem is to apply an empirical multiple testing procedure, e.g., Westfall-Young procedure [106] and its improved procedures [30, 62]. We will propose an algorithm to apply Westfall-Young procedure for significant combinations discovery in Chapter 4. In addition, as Gilbert [38] showed in his paper, the Tarone modification [91] is also useful to improve Benjamini-Hochberg procedure [11] which is the widely used FDR controlling procedure. Hence, LAMP would be expanded to control FDR instead of FWER to improve sensitivity of statistical tests.

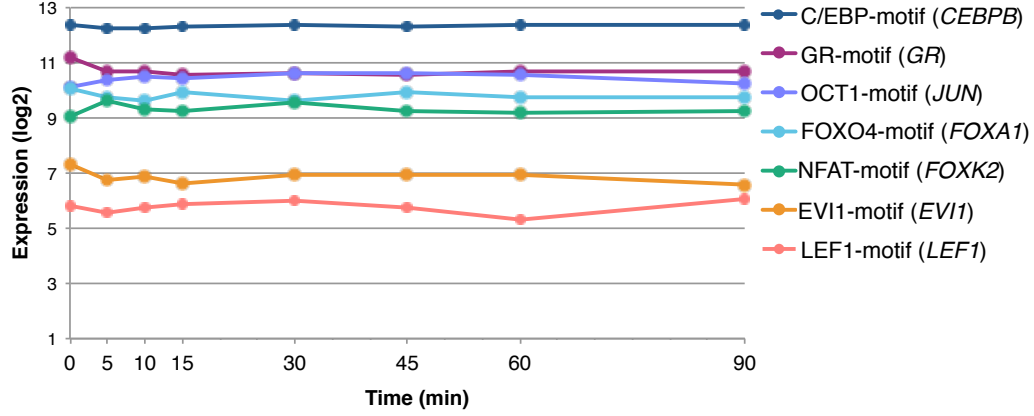


Figure 3.21: The expression levels of TFs that bind to the motifs in Figure 3.18.

Table 3.7: The list of TFs that can bind to the motifs in Figure 3.19.

Motif	TFs that bind to the motif
EVI1	<i>EVI1</i>
FOXO4	<i>FOXA1, FOXA2, FOXO1, FOXO4, HMGA1</i>
LEF1	<i>LEF1</i>
MEF2	<i>MEF2, MEF2D</i>
NFAT	<i>ELOF1, FOXK2, NFATC1, NFATC2, NFATC3, NFATC4</i>
SREBP1	<i>SREBP1</i>
TATA	<i>GTF2B, HINFP, RXRA, TBP, THRB, TMF1</i>

Table 3.8: The list of TFs that can bind to the motifs in Figure 3.20.

Motif	TFs that bind to the motif
AACTTT	-
FOXO4	<i>FOXA1, FOXA2, FOXO1, FOXO4, HMGA1</i>
NFKB	<i>NFKB1, NFKB2, REL, RELA</i>
OCT	<i>CREB1, GATA1, JUN, NFYA, NFYB, NFYC, POU5F1, POU5F1B, SP1</i>
NFAT	<i>ELOF1, FOXK2, NFATC1, NFATC2, NFATC3, NFATC4</i>

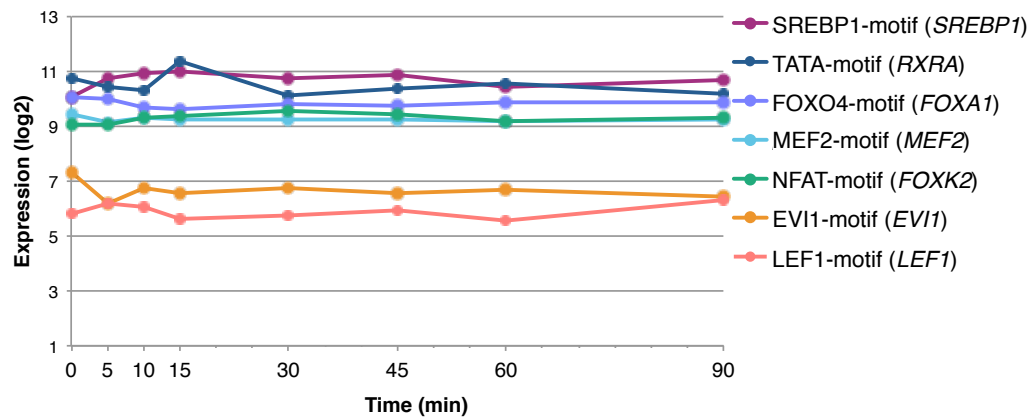


Figure 3.22: The expression levels of TFs that bind to the motifs in Figure 3.19.

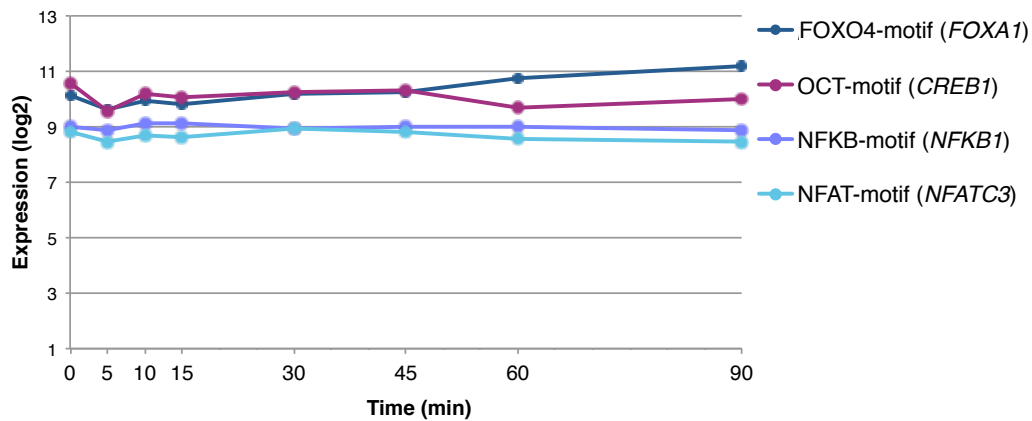


Figure 3.23: The expression levels of TFs that bind to the motifs in Figure 3.20.

Chapter 4

Empirical Multiple Testing Procedure for Combinatorial Effects Discovery

4.1 Introduction

Empirical FWER controlling multiple testing procedures, such as Westfall-Young procedure (WY-procedure) [106], generate a null distribution from thousands of randomly permuted datasets and determine δ based on the distribution. It is known that WY-procedure has a higher detection power than Bonferroni correction [15], and improved methods of it [30, 62] have been widely used in practice [24, 107]. While the high sensitivity may allow us to discover statistically significant combinations, a large amount of computing time is required in the permutation tests, which prohibits the application of WY-procedure to significant combination discovery. For example, when WY-procedure is used to find significant combinations with 1,000 permutation tests for 100 items, the statistical tests are performed $1000 \cdot (2^{100} - 1) \approx 10^{33}$ times because $2^{100} - 1$ tests are performed to find the minimum P-value in each permuted dataset.

In this chapter, we propose an efficient algorithm for performing WY-procedure, called *FastWY*, to enumerate statistically significant combinations. FastWY can be applied to the statistical tests that satisfy the two properties described in Section 3.2: the existence of a non-zero lower bound of the P-value and the decrease of the lower bound of the P-value to $x(I)$. In Section 4.3, we propose a basic branch-and-bound algorithm of FastWY. In Section 4.4, we demonstrate that FastWY achieves orders of magnitude acceleration over WY-procedure using yeast dataset, and we show that the significance level can be successfully calibrated even when we apply FastWY to these datasets without an arity limit.

This chapter is based on the following publication of the author [95].

4.2 Strategy to Efficiently Find the Minimum P-value

As described in Section. 2.3.3, WY-procedure uses the probability distribution of only the minimum P-values to compute the corrected significance level. If we would find the minimum P-value among all of the combinations in each permuted dataset, WY-procedure could be applied for significant combinations discovery. Unfortunately, this task is NP-hard [66]. Therefore, in this section, we introduce a pruning strategy for combinations to solve the problem in practice. Because minimum P-value calculations of thousands of permuted datasets are required to generate the empirical distribution of minimum P-values, the strategy has a large impact on the acceleration of the entire computing time.

Figure 4.1 illustrates the important properties of our algorithm. In Figure 4.1(a), a light red bar indicates the P-value of a combination. A red point represents the lower bound of the P-value of the combination f , which was introduced in Section 3.2.1. Assume that we have a threshold t for the lower bound (the red solid line). Combinations whose lower bounds are at most t are defined as *candidate combinations* (combinations in the red arrow region). The other combinations are defined as *non-candidate combinations* (combinations in the white arrow region). When a candidate combination has the P-value smaller than t , such as the combination $\{1, 2, 3\}$, all of the non-candidate combinations such as $\{1, 2, 3, 4\}$ never achieve the minimum P-value because their lower bounds for the P-value are larger than t . Hence, we can avoid P-value calculation of non-candidate combinations to find the minimum P-value.

However, there are two problems with this approach. One problem is how to efficiently enumerate the candidate combinations with t . The other problem is how to determine t . To solve these problems, we adopt a similar strategy to LAMP that utilizes an FIM algorithm to enumerate candidate combinations with t . The lower bound of the P-values of a popularly used statistical test, such as Fisher's exact test and Mann-Whitney U test, monotonically increases upon adding an item to a combination (Property 6), and has one-to-one correspondence to the number of objects that are related to all of the items in the combination (Figure 4.1(b)). This relation enables us to convert t to λ . With λ , we can efficiently list combinations relating to λ or more objects using FIM algorithm. Because the frequent itemsets correspond to the candidate combinations, we can accelerate the findings of the candidate combinations. For the second problem, we use λ instead of t and propose an algorithm in which λ decreases gradually to find the minimum P-value in each permuted dataset.

4.3 Algorithm of FastWY

In this section, we propose an algorithm called FastWY to efficiently find the adjusted significance level δ for maintaining $\text{FWER} \leq \alpha$ using randomly permuted datasets. For each permuted dataset, FastWY discovers the minimum P-value among all of the combinations using properties in Section 3.2. Gathering these values provides us with a null distribution of the minimum P-values, in which the α percentile value can be used as the adjusted significance level for multiple testing.

Algorithm 4.2 shows the procedure of FastWY. FastWY includes three steps.

Step 1: Estimation of a null distribution (lines 1 to 16).

Step 2: Calculation of the δ (lines 17 and 18).

Step 3: Listing combinations whose P-values are less than or equal to δ (line 19).

The overall structure of FastWY is the same as that of WY-procedure, whose algorithm is written in Algorithm 3 in Section 2.3.3. The main difference is in Step 1. we introduce a branch-and-bound method which is similar to that of LAMP. Step 1 is performed on K different permuted datasets and generates K minimum P-values. Based on the P-values, the adjusted significance level δ is determined in Step 2. In Step 3, the significant combinations for δ are listed.

4.3.1 Calculation for a Permuted Dataset

In Step 1, for a permuted dataset, we compute the minimum P-value among all of the tests. When $f(\lambda)$ is larger than the minimum P-value, it is impossible for any of the non-candidate combinations to achieve the minimum P-value. Therefore, given that p is the smallest P-value of the candidate combinations and $f(\lambda) > p$, p is the minimum P-value of all of the possible combinations. One solution to find λ whose $f(\lambda) > p$ might perform the binary search to find the optimal λ . However, since a small λ requires a large amount of computing time for FIM algorithm, the strategy will require the large amount of computing time. We therefore start from a large value of λ and gradually decrease λ while $f(\lambda) \leq p$. First, λ is assigned $\min\{x_{\max}, n\}$, where $x_{\max} = \max_{i \in \{1, \dots, M\}} \{x(\{i\})\}$. FIM algorithm lists candidate combinations on λ . FastWY computes the smallest P-value p among the candidate combinations. Then, FastWY compares p with $f(\lambda)$. If $p > f(\lambda)$, a non-candidate combination can achieve a P-value smaller than p , and hence, we replace λ with $\lambda - 1$ and repeat the procedure. A decrease in λ causes an increase in $f(\lambda)$ from the inverse correlation (Properties 2 and 4), which brings $f(\lambda)$ closer to p . If $p \leq f(\lambda)$, then p is the minimum P-value.

Figure 4.2 illustrates Step 1. First, we start from $\lambda = \min\{x_{\max}, n\}$, where $x_{\max} = \max_{i \in \{1, \dots, M\}} \{x(\{i\})\}$ (line 4 in Algorithm 4.2). FIM algorithm lists candidate combinations on λ (the blue region in Figure 4.2(a), lines 6 to 11 in Algorithm 4.2). FastWY computes the smallest P-value p among the candidate combinations (the

Algorithm 6 FastWY

Input: N objects with M items and the relations between objects and items. FWER upper limit α . The number of iteration K .**Output:** Enumerate the significance combinations.

```

// Step 1: Estimate null distribution
1:  $\mathcal{I} \leftarrow \{\}, A \leftarrow \{\}, \lambda_{\min} \leftarrow n$ 
2: for  $k = 1$  to  $K$  do
3:   Generate the permuted dataset
4:    $p \leftarrow 1.0, \lambda \leftarrow \min\{x_{\max}, n\}$ 
   // Iteration the procedure illustrated in Figure 4.2
5:   while  $f(\lambda) < p$  do
6:     if  $\lambda < \lambda_{\min}$  then
7:        $I \leftarrow$  combinations related  $\lambda$  or more objects
8:       Add  $I$  to  $\mathcal{I}$ 
9:        $\lambda_{\min} \leftarrow \lambda$ 
10:    end if
11:     $\mathcal{I}' \leftarrow$  extract combinations related  $\lambda$  objects from  $\mathcal{I}$ 
12:     $p \leftarrow \min P(I)$  for  $I \in \mathcal{I}'$ 
13:     $\lambda \leftarrow \lambda - 1$ 
14:  end while
15:  Add  $p$  to  $A$ 
16: end for
// Step 2: Calibrate the threshold  $\delta$ 
17: Sort  $A$  in ascending order
18:  $\delta \leftarrow \max_k A[k]$ , where  $A[k] < \lceil \alpha K \rceil$ -th value in  $A$ 
// Step 3: Enumerate all combinations that P-value are less than  $\delta$ 
19: Output combinations  $I$  where  $P(I) \leq \delta$  for  $I \in \mathcal{I}$ 

```

dotted red line in Figure 4.2(b)), and then, FastWY compares p with $f(\lambda)$ (the solid red line in Figure 4.2(b), line 12 in Algorithm 4.2). If $p > f(\lambda)$, a non-candidate combination can achieve a P-value smaller than p , and hence, we replace λ with $\lambda - 1$ and repeat the procedure (Figures 4.2(c) and (d)). Because there is the inverse correlation, the decrease of λ increases in $f(\lambda)$ and it brings $f(\lambda)$ closer to p . If $p \leq f(\lambda)$, p is the minimum P-value (Figures 4.2(e) and (f)).

4.3.2 Enumerating the Significant Combinations

In Step 3, the significant combinations for δ are listed, and Figure 4.3 illustrates the algorithm. The step is similar to the minimum P-value findings in Step 1 except using δ instead of p . First, we start from $\lambda = \min\{x_{\max}, n\}$, where $x_{\max} =$

$\max_{i \in \{1, \dots, M\}} \{x(\{i\})\}$. We extract combinations $\mathcal{I}'$ related λ objects from the stored results of FIM algorithm, which is computed in Step 1. The P-value $P(I)$ is computed for all $I \in \mathcal{I}'$, and the combination is regarded as significant if $P(I) \leq \delta$ (the dotted red line in Figure 4.3(a) and (b)). If $\delta \geq f(\lambda)$, non-candidate combinations can achieve a P-value smaller than δ , and hence we replace λ with $\lambda - 1$ and repeat the procedures (Figures 4.3(c) and (d)). Because $f(\lambda)$ has the inverse correlation to λ (Properties 2 and 4), which brings $f(\lambda)$ closer to δ . If $\delta < f(\lambda)$, p is the minimum P-value (Figures 4.2(e) and (f)), all of the combinations whose P-values are less than or equal to δ are listed, and they are statistically significant combinations after multiple testing correction. Unlike p in Step 1, δ never changes in the Step 3.

4.3.3 Time Complexity of FastWY

Figure 4.4 shows the procedure of FastWY and that of the time complexity when Fisher's exact test is used to compute P-value. The running time of FastWY is divided into two parts: calculation of the adjusted significance level δ and enumeration of the significant combinations whose P-value is less than δ .

To calculate δ , FastWY repeatedly finds the minimum P-value from each permuted dataset. Also, the enumeration of candidate combinations and the computation of the minimum P-value in the enumerated combinations are repeated up to N times in each permuted dataset. On fixing λ , FastWY first enumerates the candidate combinations that have frequencies larger than or equal to λ . Because this enumeration belongs to NP-hard problem [109], it requires $O(2^M N)$. Then, FastWY finds the minimum P-value p in the enumerated combinations. The time complexity of Fisher's exact test (Equation 3.3) is $O(N^2)$ and the test is repeated up to $2^M - 1$ times, therefore, $O(2^M N^3)$ is required to find the minimum P-value for each λ . When we use Fisher's exact test, $f(N)$ is defined as Equation 3.4 and the computation complexity is $O(\lambda)$ to compute $\binom{n}{\lambda}$ and $\binom{N}{\lambda}$. These procedures are repeated up to N times, and therefore $O(MN) + O(\sum_{\lambda=1}^N (2^M N + 2^M N^3 + N)) = O(2^M N^4 + MN)$ is required in each permuted dataset. Because FastWY repeats the calculation for K permuted datasets, the time complexity to compute δ is $O(K(2^M N^4 + MN))$.

After the estimation of the null distribution, FastWY computes α -percentile point of the distribution in $O(K \log K)$. In the enumeration of the significant combinations, the number of tests is up to $2^M - 1$. Because each Fisher's exact test requires $O(N^3)$, the time complexity to enumerate the significant combinations is $O(2^M N^3)$. Therefore, in total, $O(K(2^M N^4 + MN + \log K))$ is required for FastWY in the worst case.

Although the time complexity is exponential growing to M in the worst case, in practice, the computational time is quite tractable. Supposing there are m_λ candi-

date combinations, the efficient branch-and-bound algorithms [1, 41, 99] enables us to rewrite the time complexity to enumerate combinations whose frequency is up to λ as $O(m_\lambda MN)$. Moreover, finding the minimum P-value from enumerated combinations and enumerating the significant combinations are also denoted by $O(m_\lambda N^3)$. Because the number of candidate combinations m_λ is often greatly smaller than $2^M - 1$, computational time of FastWY is much faster than that in the worst case.

4.4 Performance Comparison between FastWY and WY-procedure

In this section, to demonstrate the computational efficiency of FastWY, we apply WY-procedure and FastWY for identifying the joint activity of different TFs with yeast dataset.

4.4.1 Datasets and Experimental Settings

We compare the performance of FastWY and WY-procedure. We used the expression level of a gene in the nitrogen limitations condition for 30 minutes which have been obtained from Gasch *et al.* [33]. The expression threshold was set to 1.5. The relationships between motifs and genes are observed in Harbison *et al.* [43]. Since there is no standard method to find the optimal threshold to discriminate between genes that are highly expressed and those that are not [17], we used an arbitrary threshold.

For the statistical measure, a one-sided Fisher's exact test was used. We set α to 0.05. To compare the computing time, K is set to 100 for FastWY and WY-procedure. In our experiments, FastWY can find the minimum P-values from all permuted datasets even when the sizes of all combinations are considered. However, WY-procedure requires a limit to the number of items in a combination because it takes more than 10^{20} years. To verify the scalability, we generate datasets of various sizes by random selection from yeast dataset. All the analyses were repeated 5 times to calculate the means and variances of the values.

All programs are written in Python 2.7 except for FIM algorithm in line 7 in Algorithm 6. For the FIM program, we used LCM [99] due to its fast calculation. All the experiments were tested on a machine with 2 AMD Opteron processors at 2.3 GHz with 32 GB of RAM running RedHat Linux.

4.4.2 Performance upon Increasing the Number of Items

First, we check the performance upon increasing the number of items (Figure 4.5). $WY \leq 2$ and $WY \leq 3$ show WY-procedure for up to two and three items, respectively. Each plot reports the average running time and the standard deviation. Note that the y-axis is on a log scale. Figure 4.5 shows that FastWY can compute the adjusted significance level δ orders-of-magnitude faster than with WY-procedure in the yeast dataset. The computation time of FastWY increases slightly to M , whereas that of WY-procedure increases exponentially to M . Therefore, FastWY succeeds in accelerating WY-procedure.

For an increasing number of items, our strategy works more effectively. The entire number of combinations to be tested increases exponentially to M , while the slope of the computation time of FastWY is much lower than exponential in Figure 4.5. Figure 4.6 shows the total number of combinations which are calculated the P-values in the yeast dataset analyses. The slopes in Figure 4.6 is similar to the slopes in Figure 4.5, respectively. Therefore, the slopes in Figure 4.5 implies that our pruning technique works well to reduce the computation time for FastWY. Thus, FastWY is faster than WY-procedure.

4.4.3 Performance upon Increasing the Number of Samples

Next, we check the performance with an increasing number of objects. Figure 4.7 shows the running time with varying N . As in Figure 4.5, the y-axis is on a log scale. Figure 4.7 indicates that the running time of FastWY is less than that of WY-procedure for up to two items ($WY \leq 2$) independent of N . Therefore, FastWY succeeds in accelerating WY-procedure and enables us to apply the empirical multiple testing procedure for significant combinatorial discovery.

The computation time of FastWY is increased to N in Figure 4.7 because FastWY tends to finish the computation at a small λ/N for a dataset with large N . We summarize the values of the smallest λ in the analyses of 100 permuted datasets and λ/N in Table 4.1. FastWY continues its calculation until λ decreases to 1, independent of N . The ratio of λ/N decreases with increasing N in the dataset. Given a small λ/N , FIM algorithm requires a large amount of time to enumerate the combinations with frequencies larger than λ because the number of listed combinations increases. Therefore, FastWY requires a large amount of time to analyze a dataset with large N . However, FastWY is much faster than WY-procedure and enables us to test all of the combinations without an arity limit.

We check the pruning power of combinations of FastWY. We compare the total numbers of candidate combinations in FastWY with the number of combinations whose P-values are computed in WY-procedure. For M items, WY-procedure computes all possible combinations, and the procedure computes P-value for $2^M - 1$

Table 4.1: Performances of FIM algorithm with increasing genes in Yeast data

# of genes	λ_{min}	λ_{min}/N
1000	1.0 ± 0.0	0.001 ± 0.0
2000	1.0 ± 0.0	$5.0 \cdot 10^{-4} \pm 0.0$
3000	1.0 ± 0.0	$3.33 \cdot 10^{-4} \pm 0.0$
4000	1.0 ± 0.0	$2.5 \cdot 10^{-4} \pm 0.0$
5000	1.0 ± 0.0	$2.0 \cdot 10^{-4} \pm 0.0$
6054	1.0 ± 0.0	$1.65 \cdot 10^{-4} \pm 0.0$

combinations. Figure 4.8 shows the numbers of combinations tested with each algorithm. For $M = 102$ in Figure 4.8, the average value in five repetitions of FastWY (68,996.6) is 7.613 times smaller than $WY \leq 2$ (525,300). These values are consistent with the results in Figure 4.7. By introducing the pruning technique, FastWY can reduce the number of P-value calculations by orders of magnitude. Thus, FastWY can list the significant combinations for all combination sizes, whereas WY-procedure cannot finish the computation in practical time even when the combinations size is limited to four.

4.5 Running Time to Find an Optimal λ

Here, we describe how FastWY spends time finding an optimal λ and discuss whether other searching algorithms, such as a binary search, could be used instead of gradually decreasing the value. We recorded the average optimal values of λ in 100 permuted datasets and recorded the total running time for FIM algorithm. FastWY was applied to yeast dataset five times to compute the variance.

In the yeast dataset, FastWY started from $\lambda = x_{min} = 221 \pm 0.0$ and finished at $\lambda = 1.088 \pm 0.0172$. FastWY repeated FIM algorithm 219.912 times on average by gradually decreasing λ , and spent 3.434 ± 0.090329 seconds which was 21.110% of the total amount of time. In yeast dataset, it might be possible to reduce the running time for finding λ by other searching algorithm because FIM algorithm required only less than 0.1 seconds under $\lambda = 1$. However, this reduction of running time can be applied only when the branch-and-bound algorithm to enumerate candidate combinations efficiently works. When FIM algorithm enumerates m_λ combinations, the computational time for it rises in the proportion of m_λ . Figure 4.9 shows the number of combinations enumerated when λ was varied. When λ becomes small, m_λ increases exponentially. Therefore, FastWY starts from the largest λ and gradually decreases the value to avoid the exponentially growing running time to M for the enumeration of candidate combinations.

4.6 Biological Confirmation of Detected Combinations

We verified the significant combinations in both the yeast dataset. We used $\alpha = 0.05$ and $K = 10,000$ for the calculation. FastWY requires 1,533.57 seconds in the yeast dataset.

In the yeast dataset, the adjusted significance level δ is 0.000490. Table 4.2 summarizes the 15 combinations that were detected as significant. The bold numbers represent significant after the correction. The largest combination contains four motifs, CBF1, MET4, MET31, and MET32, and the P-value of this combination is 0.000195. FastWY showed the significance of the combination about 1,533.57 seconds. In contrast, WY-procedure requires more than 100 hours even if the combination size is limited up to four because it needs P-value calculation for more than $4 \cdot 10^9$ combinations. Moreover, this four-motif combination could not be detected as significant by using Bonferroni correction [15], which is often used theoretically in multiple testing procedure, because of its conservativeness. The motifs CBF1, MET4, MET31, and MET32 are bound by *CBF1*, *MET4*, *MET31*, and *MET32*, respectively [43]. Because *MET4* lacks DNA-binding ability, the coordinator such as *CBF1*, *MET31* and *MET32* are required when *MET4* becomes active [13, 18]. In microorganisms, limiting amounts of important nutrients, such as nitrogen and sulfur often lead to the induction of metabolic systems that allow more efficient incorporation of the nutrients [92, 6, 14, 28], and *MET4* activates sulfur metabolism by transcriptional regulation [56]. The result of FastWY implies that the nitrogen depletion activates the *MET4* by interacting *CBF1*, *MET31* and *MET32* for inducing the sulfur metabolism.

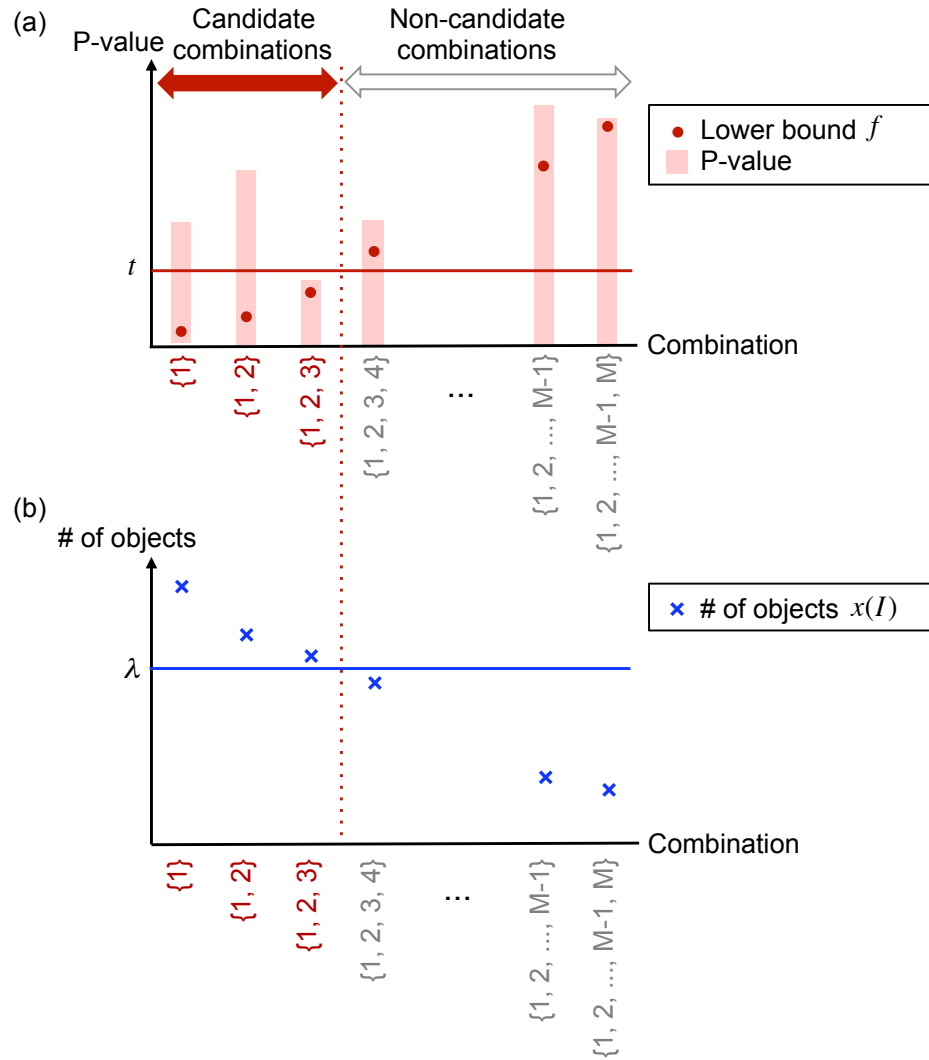


Figure 4.1: Illustration of the properties of candidate combinations. (a) Candidate and non-candidate combinations. (b) Relationship between a combination and the number of objects that have all of the items in I .

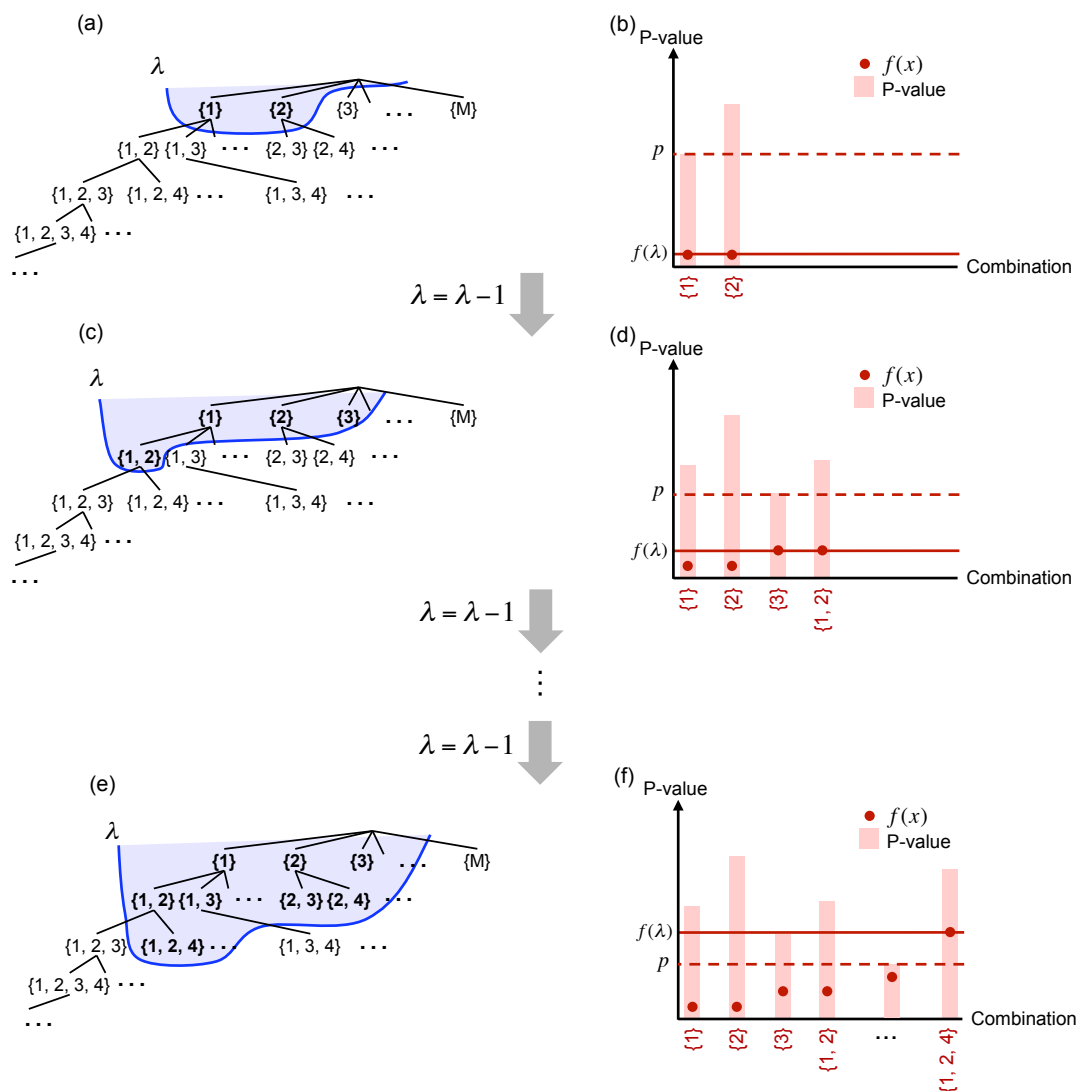


Figure 4.2: Illustration of FastWY to find the minimum P-value in each permuted dataset.

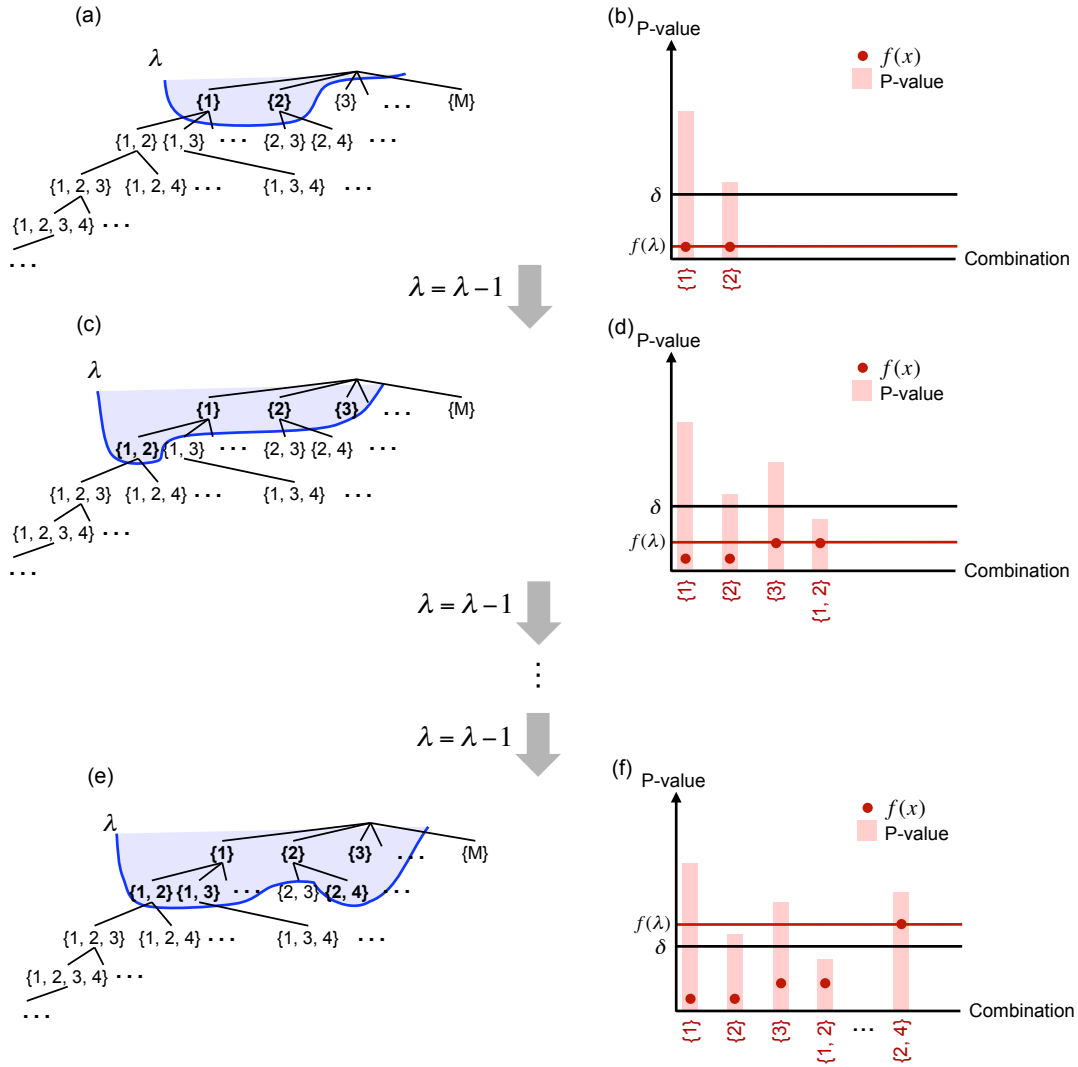


Figure 4.3: Illustration of FastWY to enumerate the combinations whose P-values are less than δ (Step 3).

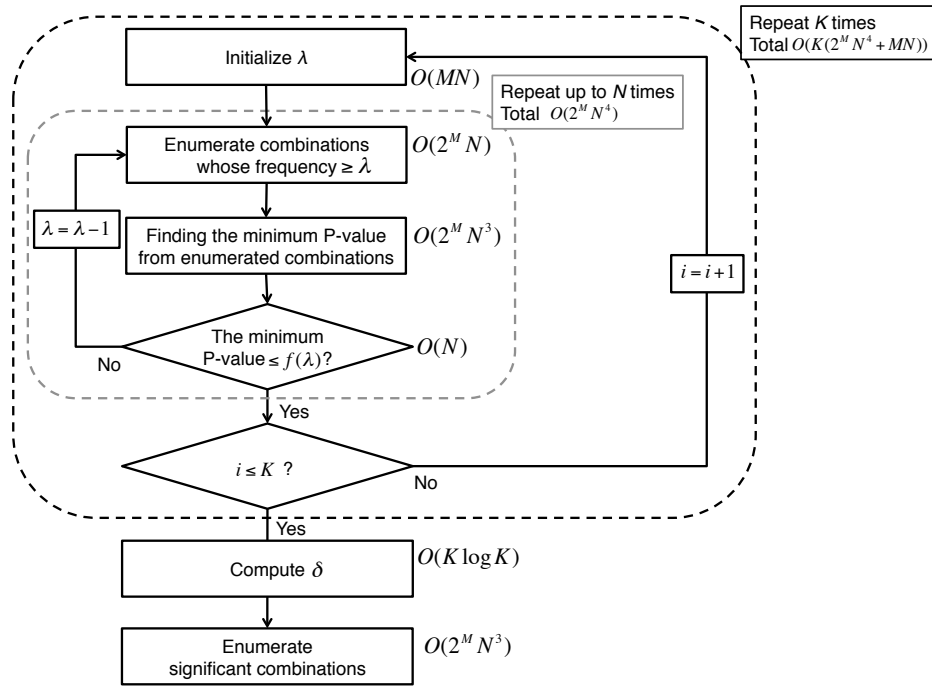


Figure 4.4: Time complexity of FastWY when Fisher's exact test is used.

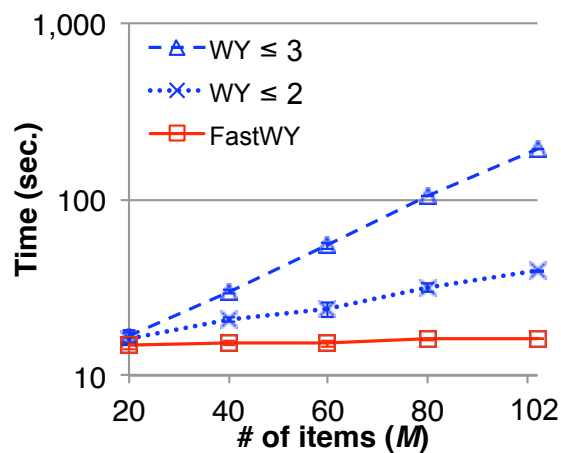


Figure 4.5: Running time of FastWY and WY-procedure with increasing numbers of items.

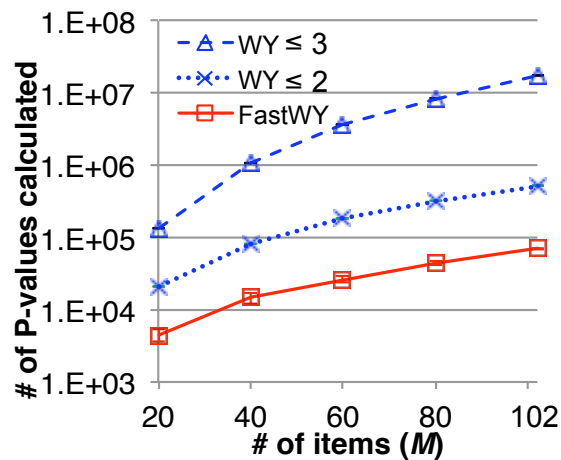


Figure 4.6: The number of combinations tested P-values of FastWY and WY-procedure with increasing numbers of items.

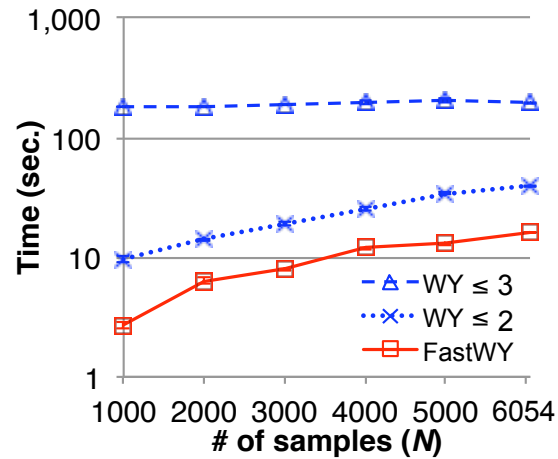


Figure 4.7: Running time of FastWY and WY-procedure with increasing numbers of samples.

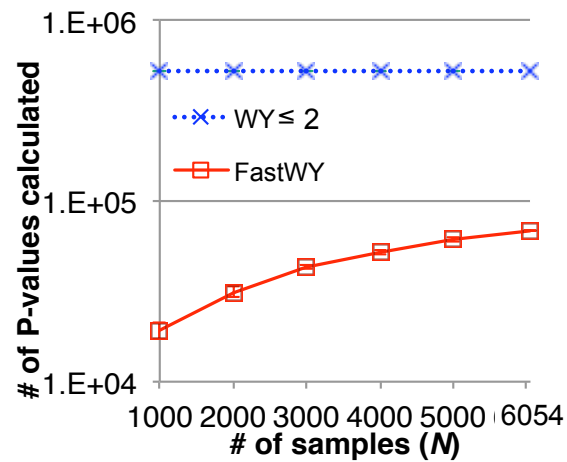


Figure 4.8: The number of combinations tested P-values of FastWY and WY-procedure with increasing numbers of samples.

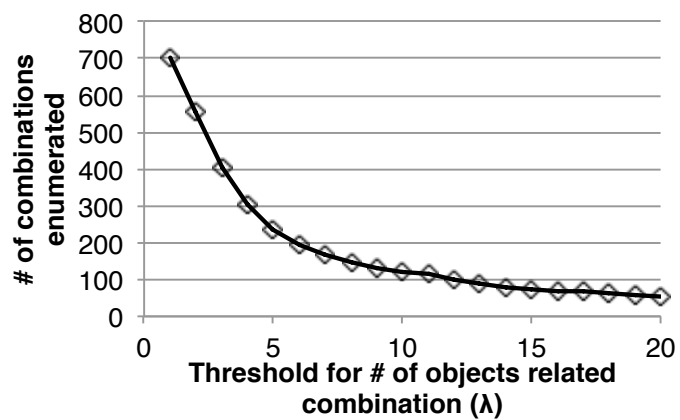


Figure 4.9: Threshold for the frequency of combinations and the number of combinations enumerated in yeast dataset.

Table 4.2: Statistically significant motif combinations detected by FastWY under the nitrogen limitations condition in yeast cells.

Motif combination	# of target genes	Raw P-value	Bonferroni correction		
			Max. arity ≤ 2	≤ 3	≤ 4
GCN4	127	$8.41 \cdot 10^{-17}$	$4.42 \cdot 10^{-13}$	$1.49 \cdot 10^{-11}$	$3.72 \cdot 10^{-10}$
GLN3	33	$4.84 \cdot 10^{-15}$	$2.54 \cdot 10^{-11}$	$8.57 \cdot 10^{-10}$	$2.14 \cdot 10^{-8}$
CBF1, MET32	8	$6.32 \cdot 10^{-8}$	0.000332	0.0112	0.280
BAS1	24	$8.92 \cdot 10^{-8}$	0.000469	0.0158	0.395
MET32	13	$1.23 \cdot 10^{-7}$	0.000645	0.0217	0.544
DAL82, GLN3	10	$4.45 \cdot 10^{-7}$	0.00234	0.0788	> 1
GCN4, GLN3	8	$3.39 \cdot 10^{-6}$	0.0178	0.600	> 1
YAP7, GCN4	9	$7.40 \cdot 10^{-6}$	0.0389	> 1	> 1
CBF1	151	$9.70 \cdot 10^{-6}$	0.0509	> 1	> 1
YAP7	65	$2.14 \cdot 10^{-5}$	0.112	> 1	> 1
DAL82	25	$2.35 \cdot 10^{-5}$	0.123	> 1	> 1
GLN3, YAP7	3	$5.00 \cdot 10^{-5}$	0.263	> 1	> 1
MET4	7	$5.85 \cdot 10^{-5}$	0.307	> 1	> 1
CBF1, MET4, MET31, MET32	4	0.000195	N/A	N/A	> 1
GAT1	10	0.000321	> 1	> 1	> 1

4.7 Conclusion

In this chapter, we proposed an efficient algorithm to accelerate WY-procedure to list statistically significant combinations. Our algorithm uses the property that distribution of FWER only depends on the minimum P-values of permuted dataset, and we developed the method to accurately find the minimum value. We also introduced the method to reduce the duplicate computation among permuted dataset. We showed that our algorithm can successfully calibrate the significance level for combinatorial regulation discovery even when all possible combinations are considered. The computation time is faster than that of WY-procedure when up to three item combinations are computed, showing that the algorithm achieved orders-of-magnitude acceleration to WY-procedure.

Although we succeeded in accelerating WY-procedure, improvement of the computational speed of our algorithm would be required. For example, the drug combination detection problem requires investigation of combinations of 10,000 drugs, and the discovery of SNPs combinations in medical science requires the consideration of combinations of more than three million positions. These problems yield datasets containing a larger number of items than we applied. One solution would be the development of FIM algorithm suitable for our problem. We ran an existing FIM algorithm on every λ , but some overlaps were found between them. Reducing duplication may accelerate the running time of our algorithm. Another solution is fast computation of the FWER distribution. Instead of using the exact minimum value from each permuted dataset, estimation of the value would be enough to adjust the significance level in practice.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Although the important combinatorial effects have been discovered in various research such as combinatorial regulation by TFs [65, 76, 7, 103], joint effects of amino acid mutations [26, 120] and SNPs [107, 3, 79, 50], multiple testing procedures have hidden these combinatorial effects which have already been experimentally verified. To overcome the problems associated with multiple testing correction in combinatorial effect discovery, we first clarified how existing multiple testing procedures hide the significant combinations. The results are listed below.

- The theoretically FWER controlling procedures, such as Bonferroni correction [15] and Holm procedure [46] conservatively calibrate the adjusted threshold because of the assumption of independence between tests. In combination discovery, because many of the tests are highly correlated with one another, many significant combinations are hidden by Bonferroni correction and Holm procedures in our experiments.
- The empirically approaches, such as Westfall-Young procedure [106], can take into account the dependence between tests by empirical estimation of the null distribution. However, this approach requires a large amount of time to estimate the null distribution, even when a small number of tests are performed. The scanning of combinations requires performing an intractable number of tests. Therefore, the procedure cannot be applied the procedure for combinatorial effect discovery.

To overcome these problems, we proposed two multiple testing procedures that can detect significant combinations without any limit. One procedure is LAMP, which theoretically calibrates the adjusted threshold. Because not every test increases

FWER [91], LAMP counts the exact number of testable combinations by using FIM algorithm and calibrates the Bonferroni factor to the smallest possible value. The other procedure is FastWY, which empirically determines the adjusted threshold by using a permutation test. By expanding the branch-and-bound algorithm of LAMP, we succeeded in accelerating Westfall-Young procedure and applying FastWY to combinatorial effects discovery for any arity combinations. Both LAMP and FastWY greatly contribute to combinatorial effect discovery because there have been no multiple testing procedure that can consider higher-arity effects. The statistically significant combinations found by our procedures can motivate biologists to confirm whether these combinations function in cells.

5.2 Future Work

In this thesis, we focused on controlling FWER. However the control of FDR instead FWER often is used in biological problems [71, 30], therefore, FDR controlling procedure for combinatorial effects would be required. Additionally, we focused on finding significant joint effects only, but it would be interesting to decompose the effects to multilevel interaction effects. On the application side, although our experiments were performed to discover combinatorial activities of TFs in this thesis, our procedure could be applied to other problem such as GWAS. However, such application would require improvements in the computational speed of our algorithm because the number of items, such as SNPs, would be substantially larger than the numbers of items in the datasets that we used.

Bibliography

- [1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.
- [2] Aaron Arvey, Phaedra Agius, William Stafford Noble, and Christina Leslie. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.*, 22(9):1723–1734, 2012.
- [3] Susanna Atwell, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, Tina T Hu, Rong Jiang, N Wayan Muliyati, Xu Zhang, Muhammad Ali Amer, Ivan Baxter, Benjamin Brachi, Joanne Chory, Caroline Dean, Marilyne Debieu, Juliette de Meaux, Joseph R Ecker, Nathalie Faure, Joel M Kniskern, Jonathan D G Jones, Todd Michael, Adnane Nemri, Fabrice Roux, David E Salt, Chunlao Tang, Marco Todesco, M Brian Traw, Detlef Weigel, Paul Marjoram, Justin O Borevitz, Joy Bergelson, and Magnus Nordborg. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, 465(7298):627–631, 2010.
- [4] Nilanjana Banerjee and Michael Q Zhang. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, 31(23):7024–7031, 2003.
- [5] Artem Barski, Suresh Cuddapah, Kairong Cui, and TY Roh. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–837, 2007.
- [6] P Baudouin-Cornu, Y Surdin-Kerjan, P Marlière, and D Thomas. Molecular evolution of protein atomic composition. *Science*, 293(5528):297–300, 2001.
- [7] Antoine Baudry, Marc A Heim, Bertrand Dubreucq, Michel Caboche, Bernd Weisshaar, and Loïc Lepiniec. TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant J.*, 39(3):366–380, 2004.

- [8] Tim Becker and Michael Knapp. A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am J Hum Genet.*, 75(4):561–570, 2004.
- [9] Jeremy Bellay, Gowtham Atluri, Tina L Sing, Kiana Toufighi, Michael Costanzo, Philippe Souza Moraes Ribeiro, Gaurav Pandey, Joshua Baller, Benjamin VanderSluis, Magali Michaut, Sangjo Han, Philip Kim, Grant W Brown, Brenda J Andrews, Charles Boone, Vipin Kumar, and Chad L Myers. Putting genetic interactions in context through a global modular decomposition. *Genome Res.*, 21(8):1375–1387, 2011.
- [10] Yoav Benjamini and Henry Braun. John W. Tukey’s contributions to multiple comparisons. *Ann Stat.*, 30(6):1576–1594, 2002.
- [11] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B*, 57(1):289–300, 1995.
- [12] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann Stat.*, 29(4):1165–1188, 2001.
- [13] PL Blaiseau, AD Isnard, Y Surdin-Kerjan, and D Thomas. Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol Cell Biol.*, 17(7):3640–3648, 1997.
- [14] Viktor M Boer, Johannes H de Winde, Jack T Pronk, and Matthew D W Piper. The genome-wide transcriptional responses of *Saccharomyces cerevisiae* grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur. *J Biol Chem.*, 278(5):3265–3274, 2003.
- [15] Carlo Emilio Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.
- [16] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.
- [17] Jan Budczies, Frederick Klauschen, Bruno V Sinn, Balázs Gyrfy, Wolfgang D Schmitt, Silvia Darb-Esfahani, and Carsten Denkert. Cutoff Finder: a comprehensive and straightforward Web application enabling rapid biomarker cutoff optimization. *PLoS One*, 7(12):e51862, 2012.

- [18] Emilio Carrillo, Giora Ben-Ari, Jan Wildenhain, Mike Tyers, Dilon Grammentz, and Traci A Lee. Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30. *Mol Biol Cell.*, 23(10):1928–42, 2012.
- [19] Jason S Carroll, Clifford a Meyer, Jun Song, Wei Li, Timothy R Geistlinger, Jérôme Eeckhoute, Alexander S Brodsky, Erika Krasnickas Keeton, Kirsten C Fertuck, Giles F Hall, Qianben Wang, Stefan Bekiranov, Victor Sementchenko, Edward a Fox, Pamela a Silver, Thomas R Gingeras, X Shirley Liu, and Myles Brown. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet.*, 38(11):1289–97, 2006.
- [20] V Cavallès, P Augereau, and H Rochefort. Cathepsin D gene is controlled by a mixed promoter, and estrogens stimulate only TATA-dependent transcription in breast cancer cells. *Proc Natl Acad Sci USA.*, 90(1):203–7, 1993.
- [21] Howard Cedar and Yehudit Bergman. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet.*, 10(5):295–304, 2009.
- [22] Chao Cheng, Roger Alexander, Renqiang Min, Jing Leng, Kevin Y. Yip, Joel Rozowsky, K.-K. Koon-Kiu Yan, Xianjun Dong, Sarah Djebali, Yijun Ruan, Carrie a. Davis, Piero Carninci, Timo Lassman, Thomas R. Gingeras, R. Guigo, Ewan Birney, Zhiping Weng, Michael Snyder, Mark Gerstein, and Roderic Guigó. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, 22(9):1658–1667, 2012.
- [23] J. Michael Cherry, Caroline Adler, Catherine Ball, Stephen A. Chervitz, Selina S. Dwight, Erich T. Hester, Yankai Jia, Gail Juvik, TaiYun Roe, Mark Schroeder, Shuai Weng, and David Botstein. SGD: *Saccharomyces* genome database. *Nucleic Acids Res.*, 26(1):73–79, 1998.
- [24] Vivian G Cheung, Richard S Spielman, Kathryn G Ewens, Teresa M Weber, Michael Morley, and Joshua T Burdick. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437(7063):1365–9, 2005.
- [25] Ami Citri and Yosef Yarden. EGF-ERBB signalling: towards the systems level. *Nat Rev Mol Cell Biol*, 7(7):505–516, 2006.
- [26] J H Condra, W A Schleif, O M Blahy, L J Gabryelski, D J Graham, J C Quintero, A Rhodes, H L Robbins, E Roth, and M Shivaprakash. In vivo emergence of HIV-1 variants resistant to multiple protease inhibitors. *Nature*, 374(6522):569–571, 1995.

- [27] Michael D Conkright, Ernesto Guzmán, Lawrence Flechner, Andrew I Su, John B Hogenesch, and Marc Montminy. Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol Cell*, 11(4):1101–1108, 2003.
- [28] Michael K Conway, Douglas Grunwald, and Warren Heideman. Glucose, nitrogen, and phosphate repletion in *Saccharomyces cerevisiae*: common transcriptional responses to different nutrient signals. *G3 (Bethesda)*, 2(9):1003–1017, 2012.
- [29] R W Cox and J S Hyde. Software tools for analysis and visualization of fMRI data. *NMR Biomed*, 10(4-5):171–178, 1997.
- [30] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C Boldrick. Multiple hypothesis testing in microarray experiments. *Statist Sci.*, 18(1):71–103, 2003.
- [31] Sandrine Dudoit and Mark J Van Der Laan. *Multiple Testing Procedures and Applications to Genomics*. Springer Series in Statistics. Springer, 2007.
- [32] Bradley Efron, Robert Tibshirani, John Storey, and Virginia Tusher. Empirical Bayes Analysis of a Microarray Experiment. *JASA*, 96(456):1151 – 1160, 2001.
- [33] Audrey P Gasch, Paul T Spellman, Camilla M Kao, Orna Carmel-Harel, Michael B Eisen, Gisela Storz, David Botstein, and Patrick O Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–4257, 2000.
- [34] Youngchao Ge, Sandrine Dudoit, and Terence P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12(1):1–77, 2003.
- [35] Geert Geeven, Ronald E van Kesteren, August B Smit, and Mathisca C M de Gunst. Identification of context-specific gene regulatory networks with GEMULA—gene expression modeling using Lasso. *Bioinformatics*, 28(2):214–221, 2012.
- [36] Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *Ann Statist.*, 32(3):1035–1061, 2004.
- [37] Robert Gentleman, Vincent J. Carey, Wolfgang Huber, Rafael A. Irizarry, and Sandrine Dudoit, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Statistics for Biology and Health. Springer-Verlag, New York, 2005.

- [38] Peter B Gilbert. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *J R Stat Soc Ser C Appl Stat.*, 54(1):143–158, 2005.
- [39] S Gold, B Christian, S Arndt, G Zeien, T Cizadlo, D L Johnson, M Flaum, and N C Andreasen. Functional MRI statistical software packages: a comparative analysis. *Hum Brain Mapp.*, 6(2):73–84, 1998.
- [40] M S Halfon, A Carmena, S Gisselbrecht, C M Sackerson, F Jiménez, M K Baylies, and A M Michelson. Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell*, 103(1):63–74, 2000.
- [41] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, volume 29, pages 1–12. ACM Press, 2000.
- [42] Sean E Hanlon, Jason M Rizzo, Deirdre C Tatomer, Jason D Lieb, and Michael J Buck. The stress response factors Yap6, Cin5, Phd1, and Skn7 direct targeting of the conserved co-repressor Tup1-Ssn6 in *S. cerevisiae*. *PLoS One*, 6(4):12, 2011.
- [43] Christopher T Harbison, D Benjamin Gordon, Tong Ihn Lee, Nicola J Rinaldi, Kenzie D Macisaac, Timothy W Danford, Nancy M Hannett, Jean-bosco Tagne, David B Reynolds, Jane Yoo, Ezra G Jennings, Julia Zeitlinger, Dmitry K Pokholok, Manolis Kellis, P Alex Rolfe, Ken T Takusagawa, Eric S Lander, David K Gifford, Ernest Fraenkel, and Richard A Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.
- [44] Yosef Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.
- [45] Burt S Holland and Margaret DiPonzio Copenhaver. An Improved Sequentially Rejective Bonferroni Test Procedure. *Biometrics*, 43(2):417–423, 1987.
- [46] Sture Holm. A simple sequentially rejective multiple test procedure. *Scand J Stat.*, 6(2):65–70, 1979.
- [47] Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386, 1988.
- [48] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.*, 4(1):44–57, 2009.

- [49] T Jenuwein and C D Allis. Translating the histone code. *Science*, 293(5532):1074–1080, 2001.
- [50] U. Johanson. Molecular Analysis of FRIGIDA, a Major Determinant of Natural Variation in Arabidopsis Flowering Time. *Science*, 290(5490):344–347, 2000.
- [51] Tony Kam-Thong, Benno Pütz, Nazanin Karbalai, Bertram Müller-Myhsok, and Karsten Borgwardt. Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs. *Bioinformatics*, 27(13):i214–21, 2011.
- [52] O V Kel, A G Romaschenko, A E Kel, E Wingender, and N A Kolchanov. A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res.*, 23(20):4097–4103, 1995.
- [53] Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, Michael B Bracken, Frederick L Ferris, Jurg Ott, Colin Barnstable, and Josephine Hoh. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.
- [54] Edward L Korn, James F Troendle, Lisa M McShane, and Richard Simon. Controlling the number of false discoveries: application to high-dimensional genomic data. *J Stat Plan Inference.*, 124(2):379–398, 2004.
- [55] A V Krishnan, S Swami, and D Feldman. Estradiol inhibits glucocorticoid receptor expression and induces glucocorticoid resistance in MCF-7 human breast cancer cells. *J Steroid Biochem Mol Biol.*, 77(1):29–37, 2001.
- [56] L Kuras, H Cherest, Y Surdin-Kerjan, and D Thomas. A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, Met4 and Met28, mediates the transcription activation of yeast sulfur metabolism. *EMBO J.*, 15(10):2519–2529, 1996.
- [57] E. L. Lehmann and Joseph P. Romano. Generalizations of the familywise error rate. *Ann Stat.*, 33(3):1138–1154, 2005.
- [58] L Mallet, B Mazoyer, and J L Martinot. Functional connectivity in depressive, obsessive-compulsive, and schizophrenic disorders: an explorative correlational analysis of regional cerebral metabolism. *Psychiatry Res.*, 82(2):83–93, 1998.
- [59] E. Manduchi, G. R. Grant, S. E. McKenzie, G. C. Overton, S. Surrey, and C. J. Stoeckert. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics*, 16(8):685–698, 2000.

- [60] D Marbach, S Roy, F Ay, P E Meyer, R Candeias, T Kahveci, C A Bristow, and M Kellis. Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.*, 22(7):gr.127191.111–, 2012.
- [61] Timothy A McKinsey, Chun Li Zhang, and Eric N Olson. MEF2: a calcium-dependent regulator of cell division, differentiation and death. *Trends Biochem Sci.*, 27(1):40–47, 2002.
- [62] Nicolai Meinshausen, Marloes H. Maathuis, and Peter Bühlmann. Asymptotic optimality of the WestfallYoung permutation procedure for multiple testing under dependence. *Ann Stat.*, 39(6):3369–3391, 2011.
- [63] Alexander Meissner, Marius Wernig, and Rudolf Jaenisch. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat Biotechnol.*, 25(10):1177–81, 2007.
- [64] Manuel Middendorf, Anshul Kundaje, Chris Wiggins, Yoav Freund, and Christina Leslie. Predicting genetic regulatory response using classification. *Bioinformatics*, 20 Suppl 1:i232–40, 2004.
- [65] Hisashi Miura, Yasuhiro Tomaru, Misato Nakanishi, Shinji Kondo, Yoshihide Hayashizaki, and Masanori Suzuki. Identification of DNA regions and a set of transcriptional regulatory factors involved in transcriptional regulation of several human liver-enriched transcription factor genes. *Nucleic Acids Res.*, 37(3):778–792, 2009.
- [66] Shinichi Morishita and Jun Sese. Traversing Itemset Lattices with Statistical Metric Pruning. In *Proceedings of the 19th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS '00)*, pages 226–236, 2000.
- [67] Takeshi Nagashima, Hidetoshi Shimodaira, Kaori Ide, Takashi Nakakuki, Yuki-taka Tani, Kaoru Takahashi, Noriko Yumoto, and Mariko Hatakeyama. Quantitative transcriptional control of ErbB receptor signaling undergoes graded to biphasic response for cell differentiation. *J Biol Chem.*, 282(6):4045–4056, 2007.
- [68] Leelavati Narlikar, Noboru J Sakabe, Alexander A Blanski, Fabio E Arimura, John M Westlund, Marcelo A Nobrega, and Ivan Ovcharenko. Genome-wide discovery of human heart enhancers. *Genome Res.*, 20(3):381–392, 2010.
- [69] F Ness, S Bourot, M Régnacq, R Spagnoli, T Bergès, and F Karst. SUT1 is a putative Zn[II]2Cys6-transcription factor whose upregulation enhances both sterol

- uptake and synthesis in aerobically growing *Saccharomyces cerevisiae* cells. *Eur J Biochem.*, 268(6):1585–95, 2001.
- [70] M A Newton, C M Kendzierski, C S Richmond, F R Blattner, and K W Tsui. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol.*, 8(1):37–52, 2001.
 - [71] William S Noble. How does multiple testing correction work? *Nat Biotechnol.*, 27(12):1135–1137, 2009.
 - [72] Yitzhak Pilpel, Priya Sudarsanam, and George M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet.*, 29(2):153–159, 2001.
 - [73] Katherine S. Pollard and Mark J. van der Laan. Choice of a null distribution in resampling-based multiple testing. *J Stat Plan Inference.*, 125(1):85–100, 2004.
 - [74] Dror M Rom. A Sequentially Rejective Test Procedure Based on a Modified {B}onferroni Inequality. *Biometrika*, 77(3):663–665, 1990.
 - [75] Arthur J Roth. Multiple comparison procedures for discrete test statistics. *J Stat Plan Inference.*, 82(1-2):101–117, 1999.
 - [76] Jenny Schlesinger, Markus Schueler, Marcel Grunert, Jenny J Fischer, Qin Zhang, Tammo Krueger, Martin Lange, Martje Tönjes, Ilona Dunkel, and Silke R Sperling. The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs. *PLoS Genet.*, 7(2):e1001313, 2011.
 - [77] Sebastian J Schultheiss, Wolfgang Busch, Jan U Lohmann, Oliver Kohlbacher, and Gunnar Rätsch. KIRMES: kernel-based identification of regulatory modules in euchromatic sequences. *Bioinformatics*, 25(16):2126–2133, 2009.
 - [78] J. P. Shaffer. Modified sequentially rejective multiple test procedures. *JASA.*, 81(395):826–831, 1986.
 - [79] Chikako Shindo, Maria Jose Aranzana, Clare Lister, Catherine Baxter, Colin Nicholls, Magnus Nordborg, and Caroline Dean. Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of Arabidopsis. *Plant Physiol.*, 138(2):1163–73, 2005.
 - [80] Yuichi Shiraishi, Mariko Okada-Hatakeyama, and Satoru Miyano. A rank-based statistical test for measuring synergistic effects between two gene sets. *Bioinformatics*, 27(17):2399–2405, 2011.

- [81] R J Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.
- [82] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, Beverley Balkau, Barbara Heude, Guillaume Charpentier, Thomas J Hudson, Alexandre Montpetit, Alexey V Pshezhetsky, Marc Prentki, Barry I Posner, David J Balding, David Meyre, Constantin Polychronakos, and Philippe Froguel. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.
- [83] L M Smith, S C Wise, D T Hendricks, A L Sabichi, T Bos, P Reddy, P H Brown, and M J Birrer. cJun overexpression in MCF-7 breast cancer cells produces a tumorigenic, invasive and hormone resistant phenotype. *Oncogene*, 18(44):6063–6070, 1999.
- [84] J D Storey. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol.*, pages 479–498, 2002.
- [85] J D Storey, J E Taylor, and D Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Statist Soc B*, 66(1):187–205, 2004.
- [86] John D. Storey. The positive false discovery rate: a Bayesian interpretation and the q -value. *Ann Stat.*, 31(6):2013–2035, 2003.
- [87] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA.*, 100(16):9440–5, 2003.
- [88] B D Strahl and C D Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–5, 2000.
- [89] Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–72, 2007.
- [90] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, 2006.
- [91] R E Tarone. A modified Bonferroni method for discrete data. *Biometrics*, 46(2):515–522, 1990.

- [92] D W Tempest, J L Meers, and C M Brown. Synthesis of glutamate in *Aerobacter aerogenes* by a hitherto unknown route. *Biochem J.*, 117(2):405–7, 1970.
- [93] Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. Statistical significance of combinatorial regulations. *Proc Natl Acad Sci USA.*, 110(32):12996–13001, 2013.
- [94] Aika Terada and Jun Sese. Bonferroni correction hides significant motif combinations. In *13th IEEE International Conference on BioInformatics and Bio-Engineering*. IEEE, 2013.
- [95] Aika Terada, Koji Tsuda, and Jun Sese. Fast Westfall-Young Permutation Procedure for Combinatorial Regulation Discovery. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2013.
- [96] James F. Troendle. A permutational step-up method of testing multiple outcomes. *Biometrics*, 52(3):846–859, 1996.
- [97] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA.*, 98(9):5116–21, 2001.
- [98] Masao Ueki and Heather J. Cordell. Improved Statistics for Genome-Wide Interaction Analysis. *PLoS Genet.*, 8(4):e1002625, 2012.
- [99] Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. LCM : An efficient algorithm for enumerating frequent closed item sets. In *In Proceedings of Workshop on Frequent itemset Mining Implementations (FIMI’03)*, 2003.
- [100] P R van Ginkel, K M Hsiao, H Schjerven, and P J Farnham. E2F-mediated growth regulation requires transcription factor cooperation. *J Biol Chem.*, 272(29):18367–74, 1997.
- [101] Zbynek Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc.*, 62(318):626–633, 1967.
- [102] J Wang, K Ellwood, A Lehman, M F Carey, and Z S She. A mathematical model for synergistic eukaryotic gene activation. *J Mol Biol.*, 286(2):315–25, 1999.
- [103] Rongchen Wang, Peizhu Guan, Mingsheng Chen, Xiujuan Xing, Yali Zhang, and Nigel M Crawford. Multiple regulatory elements in the Arabidopsis NIA1 promoter act synergistically to form a nitrate enhancer. *Plant Physiol.*, 154(1):423–432, 2010.

- [104] Zhibin Wang, Chongzhi Zang, Jeffrey A Rosenfeld, Dustin E Schones, Artem Barski, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Weiqun Peng, Michael Q Zhang, and Keji Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet.*, 40(7):897–903, 2008.
- [105] J I Weller, J Z Song, D W Heyen, H A Lewin, and M Ron. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics*, 150(4):1699–706, 1998.
- [106] Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley, New York, 1993.
- [107] Juliane Winkelmann, Barbara Schormair, Peter Lichtner, Stephan Ripke, Lan Xiong, Shapour Jalilzadeh, Stephany Fulda, Benno Pütz, Gertrud Eckstein, Stephanie Hauk, Claudia Trenkwalder, Alexander Zimprich, Karin Stiasny-Kolster, Wolfgang Oertel, Cornelius G Bachmann, Walter Paulus, Ines Peglau, Ilonka Eisensehr, Jacques Montplaisir, Gustavo Turecki, Guy Rouleau, Christian Gieger, Thomas Illig, H-Erich Wichmann, Florian Holsboer, Bertram Müller-Myhsok, and Thomas Meitinger. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nat Genet.*, 39(8):1000–6, 2007.
- [108] Xiaohui Xie, Jun Lu, E J Kulbokas, Todd R Golub, Vamsi Mootha, Kerstin Lindblad-Toh, Eric S Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005.
- [109] Guizhen Yang. The complexity of mining maximal frequent itemsets and maximal frequent patterns. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)*, page 344, New York, New York, USA, 2004. ACM Press.
- [110] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Stat Plan Inference.*, 82(1-2):171–196, 1999.
- [111] Junying Yu, Maxim A Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L Frane, Shulan Tian, Jeff Nie, Gudrun A Jonsdottir, Victor Ruotti, Ron Stewart, Igor I Slukvin, and James A Thomson. Induced pluripotent stem cell lines derived from human somatic cells. *Science*, 318(5858):1917–20, 2007.

- [112] Xueping Yu, Jimmy Lin, Tomohiro Masuda, Noriko Esumi, Donald J Zack, and Jiang Qian. Genome-wide prediction and characterization of interactions between transcription factors in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, 34(3):917–27, 2006.
- [113] Xueping Yu, Jimmy Lin, Donald J Zack, and Jiang Qian. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.*, 34(17):4925–36, 2006.
- [114] C H Yuh, H Bolouri, and E H Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279(5358):1896–902, 1998.
- [115] Jing Zhang, Tingjun Hou, Wei Wang, and Jun S Liu. Detecting and understanding combinatorial mutation patterns responsible for HIV drug resistance. *Proc Natl Acad Sci USA.*, 107(4):1321–6, 2010.
- [116] Xiang Zhang, Shunping Huang, Wei Sun, and Wei Wang. Rapid and robust resampling-based multiple-testing correction with application in a genome-wide expression quantitative trait loci study. *Genetics*, 190(4):1511–20, 2012.
- [117] Xiang Zhang, Feng Pan, Yuying Xie, Fei Zou, and Wei Wang. COE: a general approach for efficient genome-wide two-locus epistasis test in disease association study. *J Comput Biol.*, 17(3):401–15, 2010.
- [118] Xiang Zhang, Fei Zou, and Wei Wang. FastANOVA: an Efficient Algorithm for Genome-Wide Association Study. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD’08)*, 2008.
- [119] Xiang Zhang, Fei Zou, and Wei Wang. FastChi: an efficient algorithm for analyzing gene-gene interactions. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 528–39, 2009.
- [120] Y M Zhang, H Imamichi, T Imamichi, H C Lane, J Falloon, M B Vasudevachari, and N P Salzman. Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites. *J Virol.*, 71(9):6662–70, 1997.
- [121] Yu Zhang and Jun S Liu. Bayesian inference of epistatic interactions in case-control studies. *Nat Genet.*, 39(9):1167–73, 2007.

- [122] Zhou Zhu, Jay Shendure, and George M Church. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.*, 15(6):848–55, 2005.