# Bonferroni correction hides significant motif combinations

Aika Terada[*] and Jun Sese[*]

*Abstract*— **Three or more motifs often work together, and the effects are essential in cellular machinery. However, the scanning of the associations of motifs is limited to single motifs or pairs, as the result of scanning for combinations of three or more motifs often includes no significant results. Even if we find a combination with a very small raw P-value in the scan, the combination is not significant because the adjusted P-value by a multiple testing correction, such as the Bonferroni correction, is larger than the given significance level. While it is known that the Bonferroni correction is a very conservative correction, a few biological experiments have used more sensitive random permutation based multiple testing correction such as Westfall-Young procedure (WY procedure). In this paper, we show that the Bonferroni correction and its modified procedure are too conservative to find statistically significant combinations consisting of three or more motifs, while the WY procedure can detect them. These results suggest that the statistically significant epistatic effects have been overlooked and motivate us to reanalyze the publicly available datasets.**

## I. INTRODUCTION

To respond to a wide spectrum of environmental and developmental signals, the joint activity of different transcription factors (TFs) is essential [1], [2]. However, the computational scanning of motif combinations is often limited to single motifs or pairs, even though recent computational power could list up combinations of three or more motifs because of the multiple testing correction. When we detect statistically significant motif combinations associated with a gene expression profile, we perform a statistical test for each combinations investigated. These multiple tests cause a number of high false discoveries. For example, when we set the significance level to $\alpha$ = 0.05 and we have 100 tests, the probability that false discoveries occur at least once is $1 - 0.95^{100} = 0.994$, which means 99.4 % probability that we obtain at least one spurious result. To avoid the false discoveries, a multiple testing correction should be performed. In the Bonferroni correction [3], the adjusted P-value is calculated as the product of the raw P-value and the number of tests. The number of tests increases exponentially to the maximum size of the combination. Hence, we cannot find significant combinations by using the Bonferroni correction when combinations of three or more motifs are considered, and very few papers have been published with the higher-order scanning results.

The Bonferroni correction and its variant procedures such as the Holm procedure [4] are often too conservative to be applied to large number of tests despite their simple calculations, and it is known that random permutation based multiple testing correction give us less conservative corrections [5]. However, no one has checked whether the methods using random permutation procedures, such as the Westfall-Young procedure (WY procedure) [6], could detect statistically significant combinations of motifs overlooked by the Bonferroni correction. In this paper, we investigate whether the Bonferroni correction, Holm procedure and WY procedure can identify the combinatorial regulations of motifs. We confirmed that the Bonferroni correction and Holm procedure overlook the combinatorial regulations, while the WY procedure can detect at most a four-motif combination.

## II. MULTIPLE TESTING CORRECTIONS

We compare three frequently used multiple testing procedures: the Bonferroni correction [3] (Bonferroni), Holm procedure [4] (Holm) and Westfall and Young permutation procedure [6] (WY). All of these procedures control the family-wise error rate (FWER), which is the probability that false discoveries occur at least once, to be under a given significance level $\alpha$. Bonferroni controls FWER theoretically, and Holm is an improved method for Bonferroni. WY finds the adjusted significance level based on the distribution computed from the random permutation procedure.

In this section, we introduce the three procedures. We assume that we have $m$ motif combinations to be tested and denote them as $M = \{1, \ldots, m\}$. For a motif combination, we perform a test to investigate whether all of the motifs in the combination are associated with gene expression changes. Let $p_i$ be the raw P-value of the test between the motif combination $i$ and the gene expression changes of the downstream genes. If the $p_i$ is below an adjusted threshold $\delta$, the motif combination is considered regulatory.

### A. Bonferroni Correction

The upper bound of the FWER is calculated as

$$\begin{aligned}\alpha' &= 1 - \Pr\left(\bigcap_{i\in M}\{p_i > \delta\}\right) = \Pr\left(\bigcup_{i\in M}\{p_i \le \delta\}\right) \\ &\le \sum_{i\in M} \Pr(p_i \le \delta) \le m\delta. \qquad (1)\end{aligned}$$

From this inequality, when $\delta$ is set to $\alpha/m$, $\alpha' \le \alpha$. This calculation is the Bonferroni correction.

This correction is often too conservative [7] because the inequality assumes the worst case, in which all of the tests

[*]Dept. of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8550, Japan terada@ss.cs.titech.ac.jp, sesejun@cs.titech.ac.jp

A.T. is supported by JSPS Research Fellowsships for Young Scientists.

This work was partially supported by JSPS KAKENHI Grants 24651227, 25128704 and 25125709 to J.S.

The super computing resource was provided by NIG (ROIS) and HGC (the Univ. of Tokyo).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

are independent to one another (lines 1 and 2 in Equation 1). Especially for combinatorial discovery, many combinations of two motifs have a high correlation with the single motifs, and hence when we use Bonferroni to control the FWER under $\alpha$, the real FWER would be far smaller than $\alpha$.

### B. Holm Procedure

Holm improved the sensitivity of Bonferroni [4] 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 $m$ tests are ordered, $p^{(1)} \leq \ldots \leq p^{(m)}$.

Suppose that an adjusted significance level $\delta$ satisfies $p^{(k)} \leq \delta$. Because $p^{(1)}, \ldots, p^{(k)}$ are significant, $\bigcup_{i \in M} \Pr(p_i \leq \delta)$ in Equation 1 is identical to $\bigcup_{i \in \{k,\ldots,m\}} \Pr(p^{(i)} \leq \delta)$. Hence

$$\begin{aligned} \alpha' &= \Pr\left(\bigcup_{i \in \{k,\ldots,m\}} \{p^{(i)} \leq \delta\}\right) \qquad (2) \\ &\leq \sum_{i \in \{k,\ldots,m\}} \Pr(p^{(i)} \leq \delta) \leq (m-k+1)\delta. \end{aligned}$$

To control the FWER to be under $\alpha$, the Holm procedure uses $\delta = \alpha/(m-k+1)$ for the $k$-th smallest P-value $p^{(k)}$. Holm procedure gradually increases $k$ from 1. When $p^{(k)}$ exceeds $\alpha/(m-k+1)$, $p^{(1)}, \ldots, p^{(k-1)}$ are regarded as significant combinations.

Although the Holm procedure has higher sensitivity than the Bonferroni correction, it is still too conservative to detect combinations of motifs because when $k << m$, $\alpha/(m-k+1)$ is almost $\alpha/m$, which is the adjusted significance level of the Bonferroni correction.

### C. Westfall and Young Permutation Procedure

In contrast to Bonferroni and Holm, WY 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 [8].

If we knew the true set of null hypothesis $M'$ and the distribution of the P-values under $M'$, we could directly compute the FWER for $\delta$ as

$$\alpha' = \Pr\left(\bigcup_{i \in M'} \{p_i \leq \delta\}\right) = \Pr\left(\min_{i \in M'} \{p_i \leq \delta\}\right). \qquad (3)$$

However, we do not know the distribution in advance. Hence, WY estimates it from randomly permuted data.

The procedure of WY is as follows. WY generates the permuted data by randomly shuffling the relationships between the genes and the expression levels. The associations between genes and motifs are held. With the random permuted data, the minimum P-value among all of the tests is computed since FWER depends only on the minimum P-value. Gathering $K$ minimum P-values provides the simulated null distribution of FWER. The $\alpha$ percentile point in the distribution is used as $\delta$.

The weak point of WY is the requirement for the extremely large amount of computational time compared to Bonferroni and Holm, because the enormous tests should be performed to generate an empirical distribution of FWER. In motif combination analysis, however, this weakness is not going to matter as much since Bonferroni, Holm and WY have the common problem that the total number of tests increases by the maximum size of the combination. The statistical significance of WY has a high impact on the following experiments.

## III. PERFORMANCE EVALUATION

We compared the detection power of three FWER controlling methods on yeast and human datasets. In addition, because many biological analyses uses the False Discovery Rate (FDR) as the measure of multiple testing correction, we also compared them with a widely used method to control FDR, Benjamini and Hochberg [9] (BH). Note that FWER is different measure from FDR, and FWER gives a stricter significance level than FDR under the same $\alpha$. Bonferroni and WY are written in Python 2.7. We used R function to correct with the Holm and BH. All experiments were performed on a machine with two 2.3 GHz AMD Opteron processors with 32 GB of RAM running RedHat Linux.

As a statistical measure, a one-sided Fisher's exact test was used. 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.

### A. Datasets

We used two transcriptome datasets. One was a yeast dataset. The binding site positions were generated from Harbison *et al.* [10], and the gene positions from *Saccharomyces* Genome Database [11] were used. We associated a motif with a gene when the TF binding site 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. As gene expression data, we used microarray data observed over 173 different conditions on an average of 5935.6 genes [12]. 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 at least 1.5.

The other dataset was observed in MCF-7 human breast cancer cells. We used MCF-7 breast cancer expression profile [13]. The cells were induced by an ErbB receptor ligand HRG. The gene expression profiles were observed under 28 different conditions by changing the time points and dose levels. We removed genes whose log2 expression levels were less than 4 over all conditions. The numbers of genes was 12,851. A gene was considered up-regulated if the log2 ratio of its expression level in the target sample to that in control sample was at least 1.0. To associate the TF binding motifs

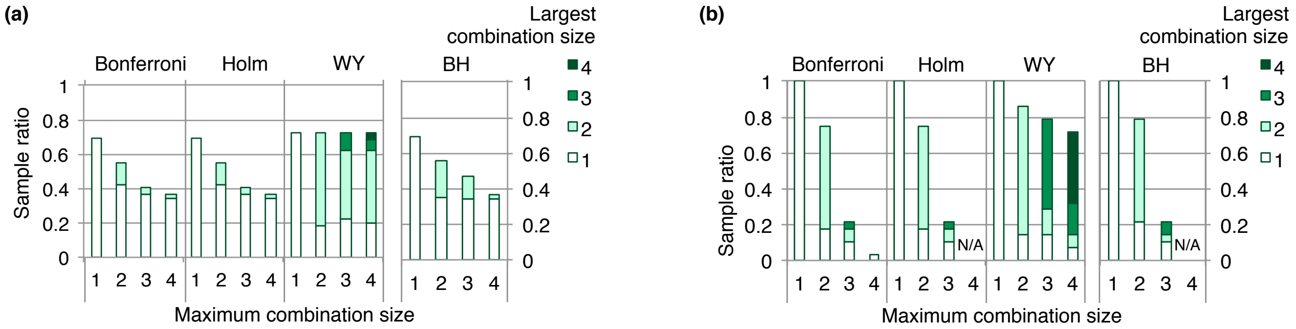

Fig. 1. Sample ratios in which at least one motif combination was deemed significant over all samples. We varied the maximum number of combinations and show that value in the x-axis. The colors indicate the largest size of the detected combinations in each sample. N/A means that the analysis could not be performed due to the high computational resource requirements. (a) The ratios of 173 yeast stress environments. Bonferroni, Holm and BH detected two-motif combinations at most, while WY detected four-motif combinations. (b) The ratios of 28 breast cancer cell samples. Only WY was able to find four-motif combinations.

with genes, we used the value ”motif” in category C3 in the MSigDB [14]. We associated the TF data with the gene expression data through GenBank ID. The number of motifs was 397, and the motifs were associated with an average 217.3 genes.

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. In Fig. 1, the y-axis shows the ratios of the samples over all samples in the dataset (in yeast and human, 173 and 28 samples). We varied the maximum combination size. Within each bar, the colors indicate the largest combination size that was detected in each sample. For example, in Fig. 1(a), 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 two motifs in 12.7% samples.

### *B. Comparing Sensitivity in Yeast Transcriptome Dataset*

Fig. 1 shows the detected ratios of four multiple testing correction methods. 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 $\delta$ extremely small. Bonferroni finds no significant combinations of three or more motif combinations even when we tested for these combinations by Bonferroni. Comparing Holm with Bonferroni, the modification of Holm gets no improvement in the combinatorial discovery. Particularly when Holm was performed for two- to four-motif combination, the detected motif combinations were identical to the result of Bonferroni. This result confirms that Holm has very similar $\delta$ to Bonferroni when the large number of tests are performed.

WY discovers 41 four-motif combinations in eight samples, and the result differs from Bonferroni and Holm. When we consider only single motif, the number of detected samples is closed to the number found by Bonferroni and Holm. However, when we consider two or more combinations, the ratios of detected samples in WY do not change with the maximum combination size, whereas those of Bonferroni and Holm decrease 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 I shows the detected motif combinations under an environment of 90 min exposure to dithiothrietol (DTT). 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 I. 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.* suggest that the response to cell wall damage is induced in yeast cells by DTT exposure [12]. This result confirms that the four-motif combination regulates the downstream genes to respond to DDT exposure environment. By 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 Fig. 1(a). The detection power of BH also dramatically decreases with the maximum combination size, and WY detects more significant combinations more than BH. This result implies that replacing Bonferroni with WY is more useful for detecting statistically significant combinations than replacing Bonferroni with BH.

### *C. Sensitivity in Human Breast Cancer Transcriptome*

Fig. 1(b) compares the numbers of detected samples among Bonferroni, Holm and WY. As in the yeast dataset, the number of detected samples by Bonferroni and Holm dramatically decrease with the maximum combination size. The number of samples detected by WY also decreases

TABLE I

COMBINATIONS DETECTED BY WY IN YEAST CELLS EXPOSED TO DITHIOTHRIETOL. THE ADJUSTED SIGNIFICANCE LEVEL OF WY IS $\delta = 0.000602$. THE ADJUSTED P-VALUES ARE USED FOR BONFERRONI, HOLM AND BH.

| Combination | P-value | Bonf. ($\leq 4$) | Holm ($\leq 4$) | BH ($\leq 4$) |
|---|---|---|---|---|
| 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 | $> 1$ | $> 1$ | $> 1$ |

TABLE II

COMBINATIONS DETECTED BY WY IN BREAST CANCER CELLS. THE ADJUSTED SIGNIFICANCE LEVEL OF WY IS $\delta = 2.58 \cdot 10^{-7}$. THE ADJUSTED P-VALUES ARE USED FOR BONFERRONI, HOLM AND BH.

| Combination | P-value | Bonf. ($\leq 4$) | Holm ($\leq 3$) | BH ($\leq 3$) |
|---|---|---|---|---|
| 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 |

toward the maximum size, but the speed is relatively slower than those of Bonferroni and Holm. Bonferroni finds three-motif combination in maximum. When we investigates combinations of up to four motifs, the adjusted significance level falls below the P-value of the three-motif combinations, and only single-motifs are found. Holm provides the identical results to Bonferroni in our analyses.

WY found four-motif combinations in 39.3% of breast cancer samples. Table II shows the motif combinations detected in 10.0 nM HRG for 30 min. In this sample, six four-motif combinations have statistically significant associations with high expressed genes. Other methods detected none of them. Remarkably, no single motif or motif pair within these four-motif combinations has a statistically significant P-value. Moreover, four combinations of them ({CP2, GATA1, FOXO4, E12}, {OCT, TATA, NFKB, FOXO4}, {LEF1, NFAT, 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 four-motif combinations, is activated in MCF-7 cells [15]. Hence, our result may suggest the collaborative effects of FOXO4. By using WY to detect combinatorial effects, it is possible to find overlooked statistically significant associations.

## IV. CONCLUSION

We showed that the WY procedure discovers the statistically significant combinatorial regulations that were concealed by the over-conservativeness of the Bonferroni and Holm corrections. In applying combinatorial regulation discovery under stress environments in yeast cells, Bonferroni and Holm detected two-motif combinations in 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. In this paper, although we limited the analyses to at most four-motif combinations, our result implies the existence of higher-order statistically significant combinations. Since we investigated only two datasets, statistically significant and biologically meaningful epistatic effects would be found from existing genome-wide datasets by using more sensitive multiple testing correction procedure.

## REFERENCES

[1] T. Ravasi *et al.*, "An atlas of combinatorial transcriptional regulation in mouse and man," *Cell*, vol. 140, no. 5, pp. 744–752, 2010.

[2] B.-K. Lee *et al.*, "Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells." *Genome Res.*, vol. 22, no. 1, pp. 9–24, 2012.

[3] C. E. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilità," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, vol. 8, pp. 3–62, 1936.

[4] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.

[5] S. Dudoit, J. P. Shaffer, and J. C. Boldrick, "Multiple hypothesis testing in microarray experiments," *Statist. Sci.*, vol. 18, no. 1, pp. 71–103, 2003.

[6] P. H. Westfall and S. S. Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Wiley, 1993.

[7] W. S. Noble, "How does multiple testing correction work?" *Nat. Biotechnol.*, vol. 27, no. 12, pp. 1135–1137, 2009.

[8] S. Dudoit and M. J. Van Der Laan, *Multiple Testing Procedures and Applications to Genomics*. Springer, 2007.

[9] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[10] C. T. Harbison *et al.*, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, no. 7004, pp. 99–104, 2004.

[11] J. M. Cherry *et al.*, "SGD: Saccharomyces genome database," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 73–79, 1998.

[12] A. P. Gasch *et al.*, "Genomic expression programs in the response of yeast cells to environmental changes," *Mol. Biol. Cell*, vol. 11, no. 12, pp. 4241–4257, 2000.

[13] T. Nagashima *et al.*, "Quantitative transcriptional control of ErbB receptor signaling undergoes graded to biphasic response for cell differentiation." *J. Biol. Chem.*, vol. 282, no. 6, pp. 4045–4056, 2007.

[14] X. Xie *et al.*, "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals." *Nature*, vol. 434, no. 7031, pp. 338–345, 2005.

[15] 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.*, vol. 77, no. 1, pp. 29–37, 2001.