BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 17 2011, pages 2399–2405
doi:10.1093/bioinformatics/btr382

Systems biology

Advance Access publication June 23, 2011

A rank-based statistical test for measuring synergistic effects between two gene sets

Yuichi Shiraishi[1,*], Mariko Okada-Hatakeyama[2] and Satoru Miyano[1]

[1]Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo 108-8639 and [2]Laboratory for Cellular Systems Modeling, RIKEN Research Center for Allergy and Immunology, Yokohama 230-0045, Japan

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Due to recent advances in high-throughput technologies, data on various types of genomic annotation have accumulated. These data will be crucially helpful for elucidating the combinatorial logic of transcription. Although several approaches have been proposed for inferring cooperativity among multiple factors, most approaches are haunted by the issues of normalization and threshold values.

Results: In this article, we propose a rank-based non-parametric statistical test for measuring the effects between two gene sets. This method is free from the issues of normalization and threshold value determination for gene expression values. Furthermore, we have proposed an efficient Markov chain Monte Carlo method for calculating an approximate significance value of synergy. We have applied this approach for detecting synergistic combinations of transcription factor binding motifs and histone modifications.

Availability: C implementation of the method is available from http://www.hgc.jp/~yshira/software/rankSynergy.zip.

Contact: yshira@hgc.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 22, 2011; revised on June 14, 2011; accepted on June 19, 2011

1 INTRODUCTION

One of the important goals of genome biology is to unravel the transcriptional regulatory mechanism. Among various statistical approaches, one common approach is to check non-random association between gene expression patterns and the presence of various factors close to the transcription start sites of genes. Although studies have mostly been performed on the relationships between expression patterns and transcription factor binding motifs, recent advances in high-throughput technologies have made it possible to consider other genome annotations such as *in vivo* transcription factor binding sites [e.g. (Carroll *et al.*, 2005)], histone modification [e.g. (Barski *et al.*, 2007; Wang *et al.*, 2008)], nucleosome positioning [e.g. (Ozsolak *et al.*, 2007; Segal *et al.*, 2006)] and open chromatin regions [e.g. (Giresi *et al.*, 2007)].

Among the statistical methods used for testing relationships between functional genome elements and gene expression, Gene Set Enrichment Analysis [GSEA; (Subramanian *et al.*, 2005)] has several merits over other statistical approaches such as the chi-squared test and Fisher's exact test. Because GSEA measures the significance of the gene set on the basis of the rank of the expression level, threshold values do not need to be set in advance for determining upregulated and downregulated genes. As discussed in Subramanian *et al.* (2005), this threshold-independent property is crucially helpful because variations in expression caused by changes in the biological environment are often modest, and the list of statistically significant genes is usually difficult to reproduce if the experiments are performed in different laboratories. In addition, because GSEA has a non-parametric framework, the validity of probabilistic assumptions such as the normality of noises need not be considered.

However, GSEA has one limitation. GSEA in its present form cannot be directly used to infer the combinatorial logic of multiple factors. Many biological studies have shown that gene expression in some cases is induced only when multiple factors come in place, and many pairs of transcription factors that work in synergy have been identified, such as E2F and NF-Y [(Caretti *et al.*, 2003; van Ginkel *et al.*, 1997)], TATA and CREB Conkright *et al.* (2003) and estrogen receptor and FOXA1 Carroll *et al.* (2005). Because of the accumulation of various types of genome annotation data, statistical methods for detecting combinations of multiple factors functioning in synergy for transcription are becoming increasingly important.

Several approaches for identifying cooperativity among transcription factors have been proposed. One popular approach [e.g. (Banerjee and Zhang, 2003; Pilpel *et al.*, 2001; Zhu *et al.*, 2005)] is to screen synergistic pairs of two transcription factor binding motifs on the basis of expression coherence. If genes with both the motifs have expression patterns that are significantly coherent than those of genes with only one motif, then we can conclude that these two factors work in synergy. This approach was successfully used to identify many cooperative combinations of motifs, especially those related to cell cycle. However, it requires many types of samples for determining coherence among genes, which is not the case in general problems. Furthermore, the choice of threshold for coherence is another difficult issue.

Another approach [e.g. (Beer and Tavazoie, 2004; Das *et al.*, 2004; Middendorf *et al.*, 2004; Segal and Sharan, 2005)] is to treat the subject as a predicting problem of gene expression from sequence motifs, and examine whether the interaction term of two factors is important. However, these approaches usually make several assumptions on probabilistic models such as the linearity of the effects of relevant factors and their interaction terms, and the validity

[*]To whom correspondence should be addressed.

of assumptions are usually difficult to investigate. In addition, the choice of normalization method considerably affects the results.

In this article, we propose a novel rank-based statistical test for measuring the synergistic effects of two factors. Similar to GSEA, our method just utilizes the ranked list of genes and is independent from threshold determination and normalization. First, we have newly defined a measure of synergy between two factors on the bases of ranks. Although this measure is hard to calculate when many genes are involved, we have provided an efficient approach for obtaining the approximate value by using a Markov chain Monte Carlo method.

2 RANK-BASED ENRICHMENT ANALYSIS

In this section, we will briefly describe how the significance value for each gene set can be obtained. Here, gene sets are, for example, genes included a specific pathway or having a certain transcription factor binding motif in their promoter sequences. As discussed in Irizarry *et al.* (2009), GSEA is not the only solution for determining whether the distribution of the genes included in a specific set is different from that of the genes not included in the set. Here, we have used a rather simple approach.

The method introduced here is based on the intuitive idea illustrated in Figure 1. When the overall expression values within a gene set tend to be significantly high or low, then that gene set is considered to have significant functions in transcription regulation. This idea is brought into shape via Wilcoxon rank-sum test, where the significance of each gene set is quantified by the gap between the obtained rank sum and the expected rank sum under the null hypothesis.

Suppose we have gene the expression profile of N genes and the target gene set. By ordering genes according to the absolute expression level or the amount of change after treatment, we can obtain a set of ranks for the gene set, which is denoted by $\mathcal{Q}^* \subset \{1,\cdots,N\}$. The number of the set $\mathcal{Q}^*$ is denoted by $|\mathcal{Q}^*|$, and the rank sum for the set $\mathcal{Q}^*$ is denoted by $S(\mathcal{Q}^*) = \sum_{e\in\mathcal{Q}^*} e$. For example, in the case of gene set 1 in Figure 1, $\mathcal{Q}^* = \{3,9,16,25,30,36,37,42\}$, $|\mathcal{Q}^*| = 8$ and $S(\mathcal{Q}^*) = 198$.

The significance of the target gene set, $T(\mathcal{Q}^*)$, is obtained by normalizing the rank sum as

$$T(\mathcal{Q}^*) = (\mu_{\mathcal{Q}^*} - S(\mathcal{Q}^*))/\sigma_{\mathcal{Q}^*},$$

where $\mu_{\mathcal{Q}^*} = |\mathcal{Q}^*|(N+1)/2$ and

$$\sigma_{\mathcal{Q}^*} = \sqrt{|\mathcal{Q}^*|(N - |\mathcal{Q}^*|)(N+1)/12}.$$

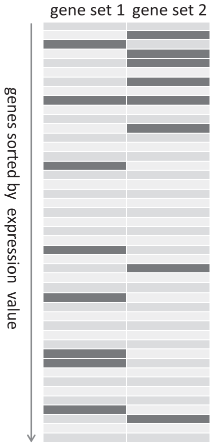

Fig. 1. Conceptual diagram of rank-based enrichment analysis: gene set 2 mainly includes genes with large amounts of expressions. Therefore, gene set 2 seems to be significant.

This approach takes advantage of the fact that the rank sum is asymptotically normally distributed, i.e. $T(\mathcal{Q}^*) \sim \mathrm{N}(0,1)$, under the hypothesis that the rank of each gene in the gene set is uniformly distributed. The significance values of gene sets correspond to observation values from the standard normal distribution. Therefore, we can easily obtain the corresponding P-value from the significance value by consulting percentile tables [e.g. $T(\mathcal{Q}^*) = 3$ is converted to 0.0027 in P-value]. Furthermore, unlike other methods such as Fisher's exact test, it is not necessary to decide the threshold value in this method. This approach takes the variation pattern of entire genes into account, and thus, this method has higher statistical power even in experimental conditions where the gene expression level shows only modest changes.

3 NEW DEFINITION OF RANK-BASED SYNERGY BETWEEN TWO GENE SETS

In this subsection, we will give a definition of synergy on the basis of the rank-based enrichment analysis method described in the previous section.

3.1 How can synergy be defined on the basis of the rank sets ?

Examples of the configurations of the ranked gene list in two gene sets is described in Figure 2. Using these illustrative examples, let us determine in which case a synergistic effect exists between two gene sets.

First, the bias of genes belonging to both gene sets can be considered as an appropriate measure for the significance of synergy. In Case 1, there seems to be no synergy between gene sets 1 and 2, because no bias is observed in the genes within the common sets. In contrast, because the intersection of two gene sets has a significant upward bias in Case 2, there seems to be a cooperative effect between the gene sets.

Second, and more importantly, the bias of each gene set has to be taken into account for determining whether or not the bias of the common set is significant. In Case 3, the common set is biased upward. However, this bias is not surprising because two gene sets are already biased, and therefore, we cannot say that there is a synergistic effect.

3.2 Definition of rank-based synergy

We adopt the rank sum of the intersection of two gene sets as the test statistics. Then, the reference distribution of the test statistics—taking into account the bias of each gene set—is necessary. The approach taken in this article is to list all the configurations where the biases of gene sets 1 and 2 are the same as those of the target configuration (the bias of each gene set is measured again using the rank sum), and collect the rank sum of the common sets for each configuration as an element from the reference distribution. An

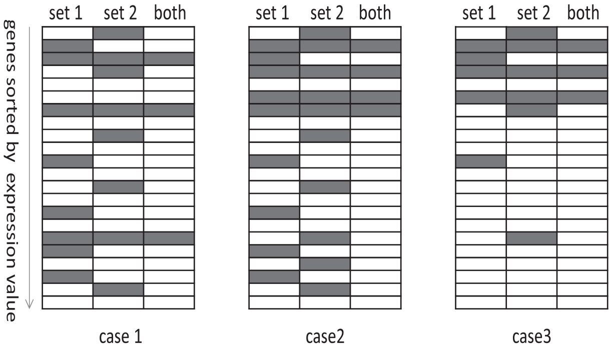

Fig. 2. Illustrative examples for rank-based synergy between two gene sets. Each cell indicates genes. Red cells indicate the genes within gene sets 1, 2 or both.

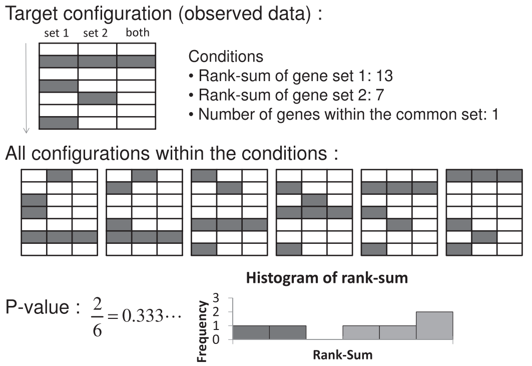
Target configuration (observed data) :
set 1
set 2
both
Conditions
• Rank-sum of gene set 1: 13
• Rank-sum of gene set 2: 7
• Number of genes within the common set: 1
All configurations within the conditions :
Histogram of rank-sum
P-value : $\frac{2}{6} = 0.333\cdots$
Frequency
3
2
1
0
Rank-Sum

Fig. 3. Illustrative examples for determining the rank-based synergy between two gene sets.

illustrative example of this procedure is shown in Figure 3. In the reminder of this subsection, we will describe the detailed procedure with rigorous mathematical notations.

Suppose the goal is to measure the synergistic effect between gene sets 1 and 2. Let $(\mathcal{Q}_1^*, \mathcal{Q}_2^*) \in 2^{\{1,\ldots,N\}} \times 2^{\{1,\ldots,N\}}$ denote the rank set pair of the target. In the case of Figure 3, $(\mathcal{Q}_1^*, \mathcal{Q}_2^*) = (\{2,4,7\},\{2,5\})$. The concrete procedure for calculating the amount of synergy between gene sets 1 and 2 is as follows:

(1) List all the configurations $(\mathcal{Q}_1, \mathcal{Q}_2) \in 2^{\{1,\ldots,N\}} \times 2^{\{1,\ldots,N\}}$ that satisfy the following requirements:
   - The numbers of sets $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are the same as those for the target configuration. Therefore, $|\mathcal{Q}_1| = |\mathcal{Q}_1^*|$ and $|\mathcal{Q}_2| = |\mathcal{Q}_2^*|$.
   - The rank sums of sets $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are the same as those for the target configuration. Therefore, $S(\mathcal{Q}_1) = S(\mathcal{Q}_1^*)$ and $S(\mathcal{Q}_2) = S(\mathcal{Q}_2^*)$.
   - The cardinality of the intersection of two gene sets is the same as that for the target configuration. Therefore, $|\mathcal{Q}_1 \cap \mathcal{Q}_2| = |\mathcal{Q}_1^* \cap \mathcal{Q}_2^*|$.

   The family of rank set pairs satisfying the above conditions is represented as

$$\begin{aligned} C(\mathcal{Q}_1^*, \mathcal{Q}_2^*) = \big\{(\mathcal{Q}_1, \mathcal{Q}_2) \big| \; & |\mathcal{Q}_1| = |\mathcal{Q}_1^*|, |\mathcal{Q}_2| = |\mathcal{Q}_2^*|, \\ & S(\mathcal{Q}_1) = S(\mathcal{Q}_1^*), S(\mathcal{Q}_2) = S(\mathcal{Q}_2^*), \\ & |\mathcal{Q}_1 \cap \mathcal{Q}_2| = |\mathcal{Q}_1^* \cap \mathcal{Q}_2^*| \big\}. \end{aligned}$$

(2) Construct the histogram of the rank sum of the intersection of two gene sets $S(\mathcal{Q}_1 \cap \mathcal{Q}_2)$ from all the configurations listed above.

(3) The P-value of the target configuration is the ratio of the number of configurations for which the rank sum of the intersection is smaller than that of the target, among all the configurations satisfying the conditions above. Therefore

$$\begin{aligned} & P(\mathcal{Q}_1^*, \mathcal{Q}_2^*) \\ & = \frac{|\{(\mathcal{Q}_1, \mathcal{Q}_2) | \; S(\mathcal{Q}_1 \cap \mathcal{Q}_2) \le S(\mathcal{Q}_1^* \cap \mathcal{Q}_2^*)\} \cap C(\mathcal{Q}_1^*, \mathcal{Q}_2^*)|}{|C(\mathcal{Q}_1^*, \mathcal{Q}_2^*)|}. \end{aligned}$$

In fact, the number of configurations meeting the conditions exponentially increases with increase in the number of genes, and thus, it is not possible to list all the configurations. Therefore, we need to resort to a sampling approach from the population so that an approximate distribution of the rank sum can be obtained.

4 SAMPLING CONFIGURATIONS VIA MARKOV CHAIN MONTE CARLO METHOD

In this section, we provide an approach for uniformly sampling the rank set pairs included in the family $C(\mathcal{Q}_1^*, \mathcal{Q}_2^*)$. Let π denote the uniform distribution on $C(\mathcal{Q}_1^*, \mathcal{Q}_2^*)$. Since, this is a sampling problem under considerably complex conditions, we use an approach based on Markov Chain Monte Carlo methods. Here, we introduce a simple Metropolis–Hastings Markov chain (Hastings, 1970) whose stationary distribution is π.

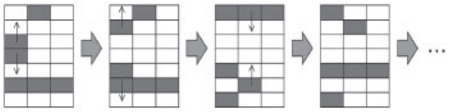

Fig. 4. Example of the proposed Markov chain Monte Carlo procedure. In each step, two genes are moved so that the rank sum remain the same.

Suppose we have a rank set pair $(\mathcal{Q}_1^{(t)}, \mathcal{Q}_2^{(t)}) \in C(\mathcal{Q}_1^*, \mathcal{Q}_2^*)$. $(\mathcal{Q}_1^{(t+1)}, \mathcal{Q}_2^{(t+1)})$ is stochastically determined by the following rule:

(1) Choose $\mathcal{Q}_1^{(t)}$ or $\mathcal{Q}_2^{(t)}$ randomly for updating. In what follows, we assume that $\mathcal{Q}_1^{(t)}$ was selected here. The case for $\mathcal{Q}_2^{(t)}$ is easily derived by slightly changing notations.

(2) Randomly select two ranks $q_a, q_b \in \mathcal{Q}_1^{(t)}$.

(3) List the set of integers $\big\{c \; \big| (\mathcal{Q}_1'(q_a, q_b, c), \mathcal{Q}_2^{(t)}) \in C(\mathcal{Q}_1^*, \mathcal{Q}_2^*)\big\}$, where $\mathcal{Q}_1'(q_a, q_b, c) = \big(\mathcal{Q}_1^{(t)} \setminus \{q_a, q_b\}\big) \cup \{q_a - c, q_b + c\}$. Select one integer c' randomly from the above set.

(4) Set $(\mathcal{Q}_1^{(t+1)}, \mathcal{Q}_2^{(t+1)}) = (\mathcal{Q}_1'(q_a, q_b, c'), \mathcal{Q}_2^{(t)})$.

The above updating rule is illustrated in Figure 4.

The following proposition guarantees that each sequential output $\big\{(\mathcal{Q}_1^{(t)}, \mathcal{Q}_2^{(t)})\big\}$ of the chain constructed above is asymptotically a sample from the target distribution π.

Proposition 1. *Let K denote the transition matrix of the Markov process described above. If the chain is irreducible, then, for every initial distribution μ,*

$$\lim_{t \to \infty} \left\| \mu K^t - \pi \right\|_{TV} = 0,$$

where $\|\cdot\|_{TV}$ denotes the total variation distance.

Proof. Because the chain is Metropolis–Hastings Markov Chain, it is only necessary to prove that the chain is irreducible and aperiodic [e.g. the Theorem 7.4 in Robert and Casella (2004)]. Aperiodicity holds because the chain can remain in the same state. ■

In fact, the Markov chain proposed in this section slightly compromises theoretical validity. In rare cases where the set of ranks is densely populated, irreducibility, which means that the Markov chain can traverse all the configurations within $C(\mathcal{Q}_1^*, \mathcal{Q}_2^*)$, does not always hold; thus, the outputs obtained from the above procedure may lead to inaccurate P-value estimates. In such cases, more complicated updating rules are required. However, in most practical cases, we can obtain samples uniformly from $C(\mathcal{Q}_1^*, \mathcal{Q}_2^*)$. Since the description of a Markov chain with a rigorous guarantee is rather technical, we have included it in Supplemental Material.

In the statistical community, Markov chain Monte Carlo approach has been extensively used for sampling distributions of test statistics in various settings [(Aoki and Takemura, 2005, 2010; Besag and Clifford, 1989; Diaconis and Sturmfels, 1998; Guo and Thompson, 1992; Smith *et al.*, 1996)]. However, most approaches focuses on contingency tables, and our approach can shed new light on the field of Monte Carlo statistical test.

5 NUMERICAL EXPERIMENTS ON SYNTHETIC DATA

In this section, we will evaluate the following issues through a comparison study of numerical experiments.

(1) The behavior of the proposed method under null hypotheses.

(2) The effectiveness of the proposed method compared with that of other methods.

5.1 Generative models

We use two probabilistic models. The common platform is the following equation:

$$\log y_i = (1-z_i)u_{i,1} + z_i u_{i,2},$$
$$u_{i,1} \sim \mathrm{U}(-3,0),\ u_{i,2} \sim \mathrm{U}(0,3),$$

where $i = 1,\ldots,N$ is the index of genes, y_i is the expression level of the i-th gene and $\mathrm{U}(a,b)$ is the uniform distribution with support $[a,b]$. Suppose that each gene is a member of none, either or both two gene sets and

$$x_{i,j} = \begin{cases} 1 & \text{the } i\text{-th gene is a member of the gene set } j \\ 0 & \text{the } i\text{-th gene is not a member of the gene set } j \end{cases},$$

for $j = 1,2$.

The difference between the two models is the generative mechanisms of $z_i, \in \{0,1\}$. In the first model, z_i is generated independent identically for all i as follows:

$$\Pr(z_i = 1) = \begin{cases} 0.6 & x_{i,1} = 1 \\ 0.5 & \text{otherwise} \end{cases}.$$

Since $x_{i,2}$ does not influence the expression, there is no synergy between two gene sets in this model.

For the second model, we adopt the following rule:

$$\Pr(z_i = 1) = \begin{cases} 0.75 & x_{i,1} = 1 \text{ and } x_{i,2} = 1 \\ 0.6 & x_{i,1} = 1 \text{ and } x_{i,2} = 0 \\ 0.5 & \text{otherwise} \end{cases}.$$

$x_{i,2}$ by itself has no effect on the expression level. However, in the case of $x_{i,1} = 1$, $x_{i,2}$ increases the probability $\Pr(z_i = 1)$, which generates a certain type of synergy between two factors.

5.2 Behavior of the proposed method under null hypotheses

We investigated the behavior of the P-value obtained using our approach for the first model. For each trial, we randomly generated the dataset $\{x_{i,1}, x_{i,2}, y_i\}_{1 \le i \le N}$ and obtained the p-value via the proposed method. In this study, $x_{i,1}$ and $x_{i,2}$ were generated according to Bernoulli trial with probability 0.2; $x_{i,1}$ and $x_{i,2}$ become 1 with probability 0.2 for each i independently. We performed this trial for 10 000 times.

Figure 5 shows that the P-values are distributed nearly uniformly, Thus, the P-value obtained via our approach reflects the extremeness in an appropriate manner under a null hypothesis.

5.3 Comparison to other methods

For data generated by the second model, the power of the proposed approach was compared to a basic likelihood-ratio test. For the likelihood-ratio test, we preprocessed the expression values in several ways (taking logarithms, assigning ranks in decreasing order and using unchanged values). We checked the significance of the interaction term in two-way analysis of variance models, which was performed classically for seeing a synergistic effect. See Supplementary Material for the detailed models of the likelihood-ratio test. We generated the dataset as described in the previous subsection and performed both tests 10 000 times.

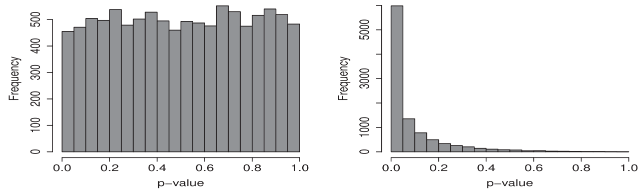

Fig. 5. Histogram of P-values obtained using our method for the first (**A**) and the second (**B**) model.

Table 1. The number of reported significance via a likelihood-ratio tests with preprocessing (U: unchanged, L: logarithm, R:rank) and proposed method in the second model

Test normalization	Likelihood ratio			Proposed
	U	L	R	—
Percent significance ($\alpha = 0.01$)	12.13	22.59	22.61	32.30
Percent significance ($\alpha = 0.05$)	27.72	45.96	45.96	59.82
Percent significance ($\alpha = 0.10$)	38.45	58.90	58.75	73.35

The level of significance is denoted by α.

The results are shown in Table 1. Our method has stronger statistical power compared to likelihood-ratio tests in this example. This may be because the likelihood-ratio test assumes a model that differs from reality. In contrast, the proposed approach is non-parametric and has no assumption of probabilistic models.

Likelihood-ratio tests greatly depend on the choice of normalization methods. The number of true positives decreased when performing the likelihood-ratio test with unchanged values. However, since the proposed approach is just designed for rank-based values, normalization methods need not be considered. Although many sophisticated approaches using probabilistic models have been proposed, the above observation clearly shows that their results are heavily influenced by the choice of normalization methods.

6 APPLICATION TO BIOLOGICAL DATA

6.1 Combinatorial effects of sequence motifs in a breast cancer cell

In many types of breast cancer, malfunction of the estrogen receptor (ER) often leads to aberrations in downstream transcriptional machinery and subsequent several biological processes such as the cell cycle and apoptosis. Therefore, studying the estrogen triggered transcriptional mechanism is crucially important. In this article, we applied our method to gene expression data for MCF-7 (estrogen receptor-positive human breast cancer cell line) stimulated via 17β-estradiol (E2, one of the most common forms of estrogen). The data we used are available at Gene Expression Omnibus (GEO, accession number GSE8597). Using transcription factor binding motifs as gene sets, we tried to identify pairs of synergistically working transcriptional factors.

Table 2. Pairs of motifs that are estimated to have synergistic effects

Motif 1	Motif 2	$\|\mathcal{Q}_1^*\|$	$\|\mathcal{Q}_2^*\|$	$\|\mathcal{Q}_1^*\cap\mathcal{Q}_2^*\|$	$T(\mathcal{Q}_1^*)$	$T(\mathcal{Q}_2^*)$	$P(\mathcal{Q}_1^*,\mathcal{Q}_2^*)$	q-value
V$NFY_01	V$E2F_03	604	593	50	3.895 448 433	3.122 547 947	0.000 948	0.109
V$ARP1_01	V$GATA_C	496	253	22	0.630 436 436	0.335 085 869	0.001 506	0.109
V$MYCMAX_01	V$CMYB_01	575	950	73	0.056 034 006	0.75 829 155	0.005 424	0.226
V$IK2_01	V$ARNT_01	352	646	28	1.838 481 609	0.802 912 513	0.006 232	0.226
V$OCT1_06	V$CDP_01	228	274	23	0.446 725 297	0.858 727 667	0.009 598	0.278
V$HAND1E47_01	V$TAL1BETAITF2_01	314	346	25	1.50 372 417	0.280 598 046	0.013 571	0.287
V$MYCMAX_01	V$ARNT_01	575	646	213	0.056 034 006	0.802 912 513	0.014 795	0.287
V$NFY_Q6	V$NMYC_01	1248	1005	128	2.59 687 386	1.826 344 064	0.016 903	0.287
V$NFY_01	V$CMYB_01	604	950	58	3.895 448 433	0.75 829 155	0.017 827	0.287
V$BRN2_01	V$CMYB_01	281	950	26	0.002 775 692	0.75 829 155	0.019 981	0.290

The pairs for which the common set includes <20 genes ($|\mathcal{Q}_1^*\cap\mathcal{Q}_2^*|<20$) have not been included in this list.

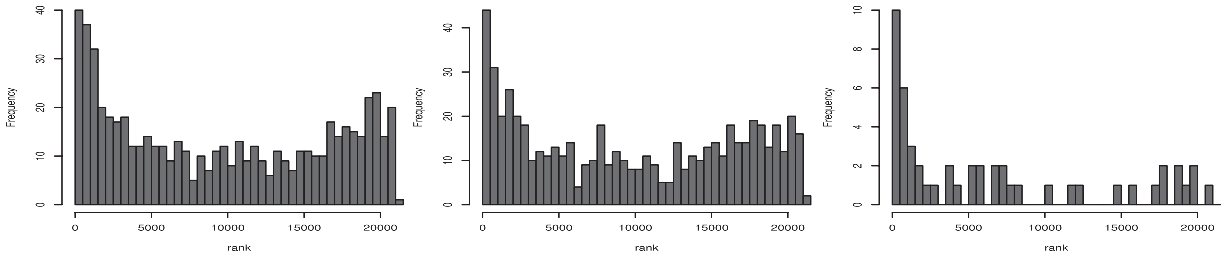

Fig. 6. Histograms of the ranks of genes within the gene set for V$NFY_01 (**A**), V$E2F_03 (**B**) and both (**C**), respectively.

6.1.1 Experimental methodology The information on transcription start sites and sequence motifs was extracted from the UCSC Genome Browser (hg18) RefSeq Gene track and transcription factor binding site (TFBS) conserved track, respectively.

For each gene, we checked whether the corresponding promoter sequence covering 1000 bp upstream and 1000 bp downstream included each TFBS motif. For genes that had multiple promoter sequences, concatenated sequences were used. Next, the gene set for each motif was determined as the set of genes that have at least one in their promoter sequences.

To avoid the redundancy of multiple probe set IDs, the probe set ID whose expression profile had the maximum variance was associated for each gene. We calculated t-statistics for each gene by measuring the difference between the expression levels of samples collected 24 h after treatment and control samples. Gene are ranked in descending order of t-statistics.

6.1.2 Results First, using the method described in Section 2, we extracted the 42 gene sets whose enrichment score was positive (Supplementary Table S1). After filtering the pairs whose common sets include <20 genes ($|\mathcal{Q}_1^*\cap\mathcal{Q}_2^*|<20$) for a robust inference, we measured the synergistic effect of each combination of these gene sets by using the proposed method. Then, to account for multiple hypothesis testing, false discovery rate (FDR) q-values are calculated (Supplementary Material for the detailed procedure). The significant synergistic combinations are listed in Table 2.

E2F and NF-Y, whose cooperativity have been verified in many published studies [(Caretti *et al.*, 2003; van Ginkel *et al.*, 1997)], are reported to be highly synergistic. Although the single effects of these motifs are already significant, their cooperativity provides additional benefits (Fig. 6). Furthermore, the motifs for E2F and NF-Y were also enriched in upregulated genes in tumor samples from patients with breast cancer (Niida *et al.*, 2009). These observations confirms that the combinatorial functions of these two transcription factors are closely related to transcription mechanism in breast cancer.

6.2 Combinatorial effects of modifications

Histone tail modifications such as methylation, acetylation and phosphorylation play an important role in chromatin structure. Although combinations of histone modifications are considered to play a role in chromatin–DNA interactions and subsequent gene expression [(Jenuwein and Allis, 2001; Strahl and Allis, 2000)], the precise mechanisms underlying this involvement remain unclear. Using the data for 39 histone modification in human CD4+ T cell (Barski *et al.*, 2007; Wang *et al.*, 2008) as gene sets, we tried to detect the pairs of synergistic histone modifications.

6.2.1 Experimental methodology Summary BED files of histone methylation and acetylation data were obtained from http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.aspx and http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellacetylation.aspx. Information on the transcription start site and promoter of each gene was obtained as described in the previous subsection.

Histone modifications on each gene were identified by the same way as (Wang *et al.*, 2008). The modification on a promoter was deemed significant when the tag count exceeded a threshold, which was determined by a P-value of 10^{-7} under the Poisson distribution fitted to the genome-wide tag density. Then, the gene set for each modification was obtained as the set of genes that have the corresponding modification pattern in their promoter sequences. Gene expression values for CD4+ T cells were obtained from the BioGPS atlas (http://biogps.gnf.org/downloads/). Filtering pairs whose common set includes <20 genes and FDR q-value calculation were performed as described in the previous subsection.

6.2.2 Results Estimated P-values for synergistic effects among histone modifications are summarized in Figure 7. More than half of the total modifications (from the top of the figure until to H3K79me3) are included in a group; most of these combinations show anti-synergistic effects. Since the information that the gene is associated with an additional modification does not aid in expecting higher expression within this set, we call this set as the 'redundant effect set'. On the other hand, some modifications such as H2AK5ac and H3R2me1 seem to have synergistic effects with several members in the redundant effect set (Table 3).

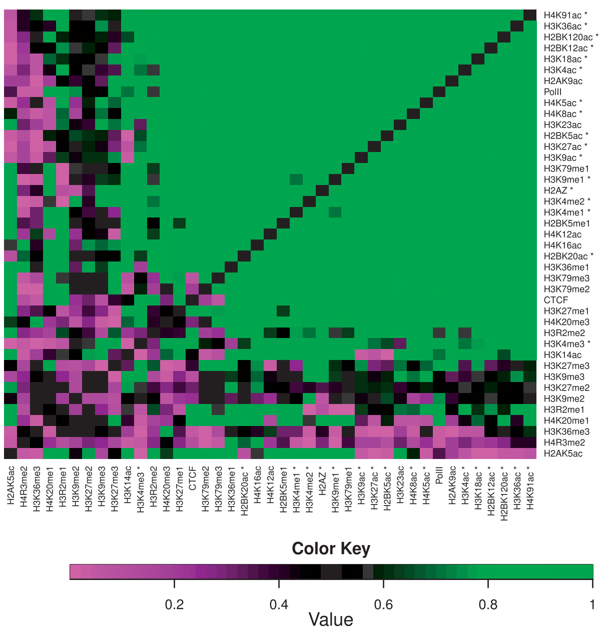

Fig. 7. Heatmap of P-values for synergy among histone modifications. Asterisks are marked on modifications belonging to the backbone module reported in a previous study (Wang *et al.*, 2008).

The 'backbone' module consisting of 17 modifications was identified in a previous study (Wang *et al.*, 2008). These modifications tend to colocalize, and the genes associated with this module tend to have higher expression than others. The redundant effect set includes most members of the backbone module. This overlap can be explained as follows: since modifications within the backbone module are highly correlated with each other, it is hypothesized that there exists a latent factor that is responsible for simultaneous modifications within the backbone module and subsequent gene expression. One modification in the backbone module is highly a predictive indicator for the activation of that factor, and additional information on other modifications may be redundant. Furthermore, the fact that H2AK5ac and H3R2me1 synergize with different members of the redundant effect set suggests the existence of more complicated machinery in the backbone module.

7 DISCUSSION

In this article, we proposed a novel non-parametric framework for measuring the synergy between two gene sets. The choice of normalization method and threshold values are issues that need not be considered for this approach. Application to sequence motifs and histone modifications data implies that our method is helpful for unraveling the gene expression mechanisms.

However, our approach still has several limitations. First, we did not take into account the positional relationships between two factors in this study. Several studies have reported the importance of positional relationships [e.g. (Beer and Tavazoie, 2004)], therefore, methods that can use this information should be devised. Second, although this is not just confined to our approach, identifying gene sets itself is a very difficult problem. The members of gene sets

Table 3. Pairs of histone modifications that are estimated to have synergistic effects

Modification 1	Modification 2	$\lvert\mathcal{Q}_1^*\rvert$	$\lvert\mathcal{Q}_2^*\rvert$	$\lvert\mathcal{Q}_1^*\cap\mathcal{Q}_2^*\rvert$	$T(\mathcal{Q}_1^*)$	$T(\mathcal{Q}_2^*)$	$P(\mathcal{Q}_1^*,\mathcal{Q}_2^*)$	q-value
H2AK5ac	H3K18ac	495	5718	430	11.07589463	49.01058367	0.004152	0.040
H3K14ac	H3K4me3	608	8328	525	11.95224806	57.21041866	0.004987	0.040
H3K9me1	H3R2me1	2839	347	140	20.85202126	−1.245034162	0.006042	0.113
H3K79me1	H3R2me1	404	347	58	11.18114935	−1.245034162	0.008093	0.113
H2AK5ac	H3K27ac	495	5812	412	11.07589463	54.35996815	0.010263	0.113
H3K4me2	H3R2me1	4657	347	152	35.50551836	−1.245034162	0.011246	0.126
H3K4me2	H4R3me2	4657	113	26	35.50551836	−2.239648018	0.014021	0.131
H3K9me1	H4R3me2	2839	113	28	20.85202126	−2.239648018	0.014194	0.131
H2AK5ac	H2BK12ac	495	4007	431	11.07589463	42.16088064	0.018313	0.132
H2BK5ac	H3K14ac	5556	608	491	53.93100775	11.95224806	0.018649	0.132
H2AK5ac	H3K4me3	495	8328	404	11.07589463	57.21041866	0.019089	0.137
H3K14ac	H3K9ac	608	7090	521	11.95224806	59.07935402	0.029839	0.137
H3K4me3	H4K20me1	8328	449	284	57.21041866	12.0094889	0.033885	0.137
H2AK5ac	H2BK5ac	495	5556	416	11.07589463	53.93100775	0.034928	0.193
H2AK5ac	H4K5ac	495	4283	403	11.07589463	40.11553114	0.035292	0.193

The pairs for which the common set includes <20 genes ($|\mathcal{Q}_1^*\cap\mathcal{Q}_2^*| < 20$) have been removed from the list.

vary greatly depending on the threshold of similarity to the position weight matrix or the number of reads collected via ChIP-Seq. We need to continue developing efficient identification methods.

Furthermore, there are still a lot of things to consider about the definition of synergistic measure. Recently, using the information theoretic measure of synergy [e.g. (Anastassiou, 2007)], a framework for identifying combinatorial rules of gene sets were proposed in (Park *et al.*, 2010) though their approach is different from that of this article because it requires a lot of samples. However, investigating theoretical relationships between the definition of synergy in this article and the information theoretic synergistic measure is an important future task.

Currently, several large projects are generating a large amount of data on various functional genome elements [e.g. (Birney *et al.*, 2007)]. Developing statistical methods for integrative understanding of the transcription mechanism will become increasingly crucial. We believe that measuring the cooperativity of two factors is the first important step for elucidating the combinatorial logic of various functional elements, and our method will make an important contribution to solving this problem.

ACKNOWLEDGEMENT

We thank Takeshi Nagashima, André Fujita and Yoshiaki Tanaka for their helpful discussion and comments.

Funding: Cell Innovation Project and Systems Cancer Project of the Ministry of Education, Culture, Sports, Science and Technology. The super-computing resource was provided by Human Genome Center, Institute of Medical Science, University of Tokyo.

Conflict of Interest: none declared.

REFERENCES

Anastassiou,D. (2007) Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.*, **3**, 83.

Aoki,S. and Takemura,A. (2005). Markov chain Monte Carlo exact tests for incomplete two-way contingency tables. *J. Stat. Comput. Simul.*, **75**, 787–812.

Aoki,S. and Takemura,A. (2010) Markov chain Monte Carlo tests for designed experiments. *J. Stat. Plan. Inference*, **140**, 817–830.

Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Beer,M. A. and Tavazoie,S. (2004). Predicting gene expression from sequence. *Cell*, **117**, 185–198.

Besag,J. and Clifford,P. (1989) Generalized Monte Carlo significance tests. *Biometrika*, **76**, 633–642.

Birney,E. *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Caretti,G. *et al.* (2003) Dynamic recruitment of NF-Y and histone acetyltransferases on cell-cycle promoters. *J. Biol. Chem.*, **278**, 30435–30440.

Carroll,J.S. *et al.* (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, **122**, 33–43.

Conkright,M.D. *et al.* (2003) Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness. *Mol. Cell*, **11**, 1101–1108.

Das,D. *et al.* (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.

Diaconis,P. and Sturmfels,B. (1998) Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.*, **26**, 363–397.

Giresi,P.G. *et al.* (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.

Guo,S.W. and Thompson,E.A. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, **48**, 361–372.

Hastings,W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Irizarry,R.A. *et al.* (2009) Gene set enrichment analysis made simple. *Stat. Methods Med. Res.*, **18**, 565–575.

Jenuwein,T. and Allis,C.D. (2001) Translating the histone code. *Science*, **293**, 1074–1080.

Middendorf,M. *et al.* (2004) Predicting genetic regulatory response using classification. *Bioinformatics*, **20**(Suppl. 1), i232–i240.

Niida,A. *et al.* (2009) Gene set-based module discovery in the breast cancer transcriptome. *BMC Bioinformatics*, **10**, 71.

Ozsolak,F. *et al.* (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, **25**, 244–248.

Park,I. *et al.* (2010) Inference of combinatorial Boolean rules of synergistic gene sets from cancer microarray datasets. *Bioinformatics*, **26**, 1506–1512.

Pilpel,Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Robert,C.P. and Casella,G. (2004) *Monte Carlo Statistical Methods*. 2nd edn., Chapter 7. Springer, New York, p. 274.

Segal,E. and Sharan,R. (2005) A discriminative model for identifying spatial cis-regulatory modules. *J. Comput. Biol.*, **12**, 822–834.

Segal,E. *et al.* (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.

Smith,P.W.F. *et al.* (1996) Monte Carlo exact tests for square contingency tables. *J. R. Stat. Soc. A*, **156**, 309–321.

Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

van Ginkel,P.R. *et al.* (1997) E2F-mediated growth regulation requires transcription factor cooperation. *J. Biol. Chem.*, **272**, 18367–18374.

Wang,Z. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Zhu,Z. *et al.* (2005) Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.*, **15**, 848–855.