Genome Biology

CORRESPONDENCE

Observations on shifted cumulative regulation

Chao Ye, Ying Liu and Xuegong Zhang*

Comment on He *et al.*: http://genomebiology.com/2007/8/9/R181

Abstract

A response to Dynamic cumulative activity of transcription factors as a mechanism of quantitative gene regulation by F He, J Buer, AP Zeng and R Balling. *Genome Biol* 2007, **8**:R181.

Studying the collaborative effects of multiple regulators is a key to understanding the basic principles of gene regulation. He *et al.* [1] proposed a shifted cumulative model to dissect combinatorial gene regulation. They discovered significant correlations between the combined expression profiles of regulators and the time series of expression of their target gene. The work highlighted the importance of identifying integrative effects of multiple transcription factors and showed that this identification was possible. We did a series of experiments to study possible combinatorial regulatory mechanisms following their strategy, but we found that the correlation among three genes can increase significantly after time-shifted combination no matter whether there are regulatory relationships. Our observations led to the conclusion that such increases are not sufficient to infer cumulative regulation relations.

We followed the strategy in He *et al.* [1] to generate combined profiles of two regulators in our experiments. Specifically, let τ_i ($0 \le \tau_i \le \tau_{max} < n$, where τ_{max} is the maximum shift and n is the number of time points of the time series) be the time shift between regulator i and the target gene, and let $R_i(t)$ be the expression level of regulator i at time point t. For regulator i, a constrained conversion efficiency C_i ($-1 \le C_i \le 1$) was chosen. Then we calculated the combinatorial profile expression at time point t as:

$$A(t) = \sum_{i=1}^{m} C_i \times R_i(t - \tau_i)$$

where m is the number of regulators ($m = 2$ in our study as we only considered the combination of two regulators). We used the Pearson correlation coefficient (PCC) as the measurement of the correlation between a transcription factor (TF) or the combined profile and their target gene. We adjust τ_i to get the combined profile that has the largest correlation with the target gene. The analysis of He *et al.* [1] indicates that a notable increase in the correlation of a target gene with the combined profile after time-shifting could indicate the existence of collaborative regulation.

We first experimented with the yeast cell-cycle dataset of Cho *et al.* [2] that was analyzed by He *et al.* [1]. We generated five datasets from these data. The first contains 817 two-regulator motifs (two regulators and a common target) in the regulatory network [3] (the original set). (He *et al.* [1] also removed genes not included in the *Saccharomyces* genome database [4] and motifs that had only one target, so their dataset has only 544 motifs.) The other four datasets are randomized datasets used as controls. Random sets 1 and 2 are shuffled from the original set by randomly assigning regulator-target relations among all genes. Random sets 3 and 4 are generated by keeping the structure of regulator-target motifs in the original data but shuffling the genes at random. The PCC improvement is calculated as the PCC of the combined TF profile with the target gene minus the average of PCCs between each profile of the two TFs and target gene. The box-plots in Figure 1 show the distribution of the observed PCC improvement after time-shifting for these five datasets. We can see that most improvement values are between 0.2 and 0.4, and there is no significant difference between the improvements in the original set and those in the random sets. We applied the Wilcoxon rank-sum test to compare the mean improvement for the original set and that for the random sets and did not find a significant difference. We also did the same experiment using the data of Spellman *et al.* [5] and obtained similar results (data not shown).

The cell-cycle data are periodic. We used a mouse liver development dataset [6] to ask whether the above observation is due to the periodic nature of the data, as the liver development data are non-periodic. We selected 169 two-regulator motifs from the regulation network

*Correspondence: zhangxg@tsinghua.edu.cn
MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST and Department of Automation, Tsinghua University, Beijing 100084, PR China

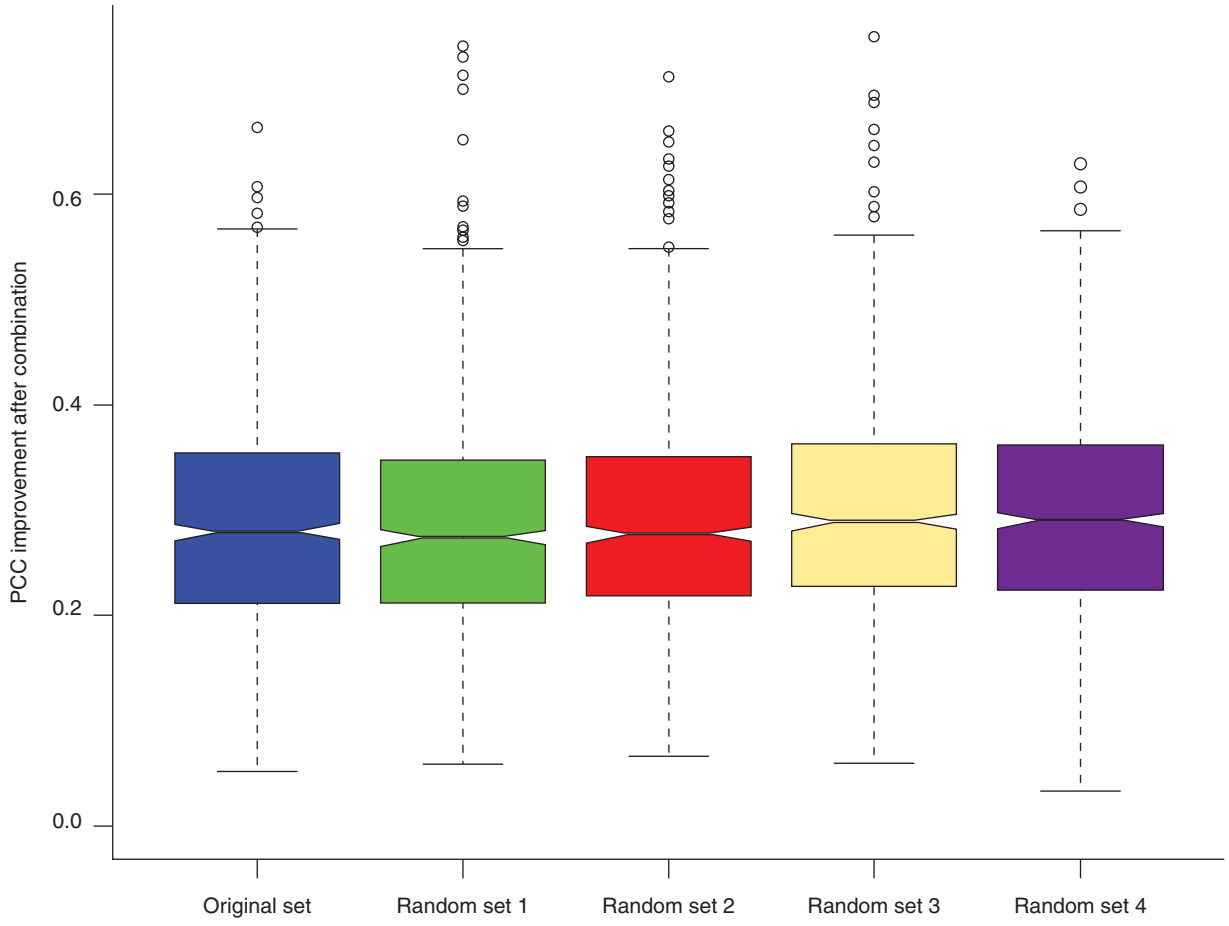


Figure 1. PCC improvements of time-shifted combined profiles with targets in Cho *et al.*'s [2] cell cycle data and randomized data.

generated by gene sets used in Liu *et al.* [7]. We removed some motifs that did not have time series data; this dataset then had 116 motifs. We calculated PCC improvements after the time-shifted combination of TFs. As negative controls, we randomly shuffled the regulation relationship among these motifs, as for the cell-cycle data. Figure 2 shows the box-plots of the PCC improvement of the different groups. It can be seen that, whether or not a gene is the common target of two regulators, there is a noticeable increase in the PCC under the shifted cumulative model. The Wilcoxon rank-sum test supported this observation.

We also used the local clustering coefficient (LC) [8] as the measurement of correlation as in He *et al.* [1] and used the same threshold (LC > 13 as the threshold for significant correlation [8]). The same constraint on the time shift was used as in the original paper [1]. In these experiments, we removed regulator pairs that had only a single target, and also removed genes that were not included in the *Saccharomyces* Genome Database. This gave us 515 two-regulator motifs from the data of Cho *et al.* [2]. (The difference in the number of motifs from the 544 in [1] may be due to an update of the database.) The time shift between two regulators is fixed among their multiple targets. We calculated the LC and counted the number of significant correlations in the original and shuffled data. For the original data, the proportion of significant motifs is 36.12%, close to that observed by He *et al.* [1]. We generated 50 random datasets by shuffling the genes while keeping the structure of the regulation motifs. Figure 3 shows a histogram of the proportion of significant motifs detected for the 50 random datasets. We can see that the proportion observed in the original data is not significantly higher than that in the random data. We also did the same experiments using the data of Spellman *et al.* [5] and of Li *et al.* [6] and observed similar results (data not shown).

One can understand the reason for the above observation using the framework of vector decomposition. Any time series of n points can be treated as a vector in

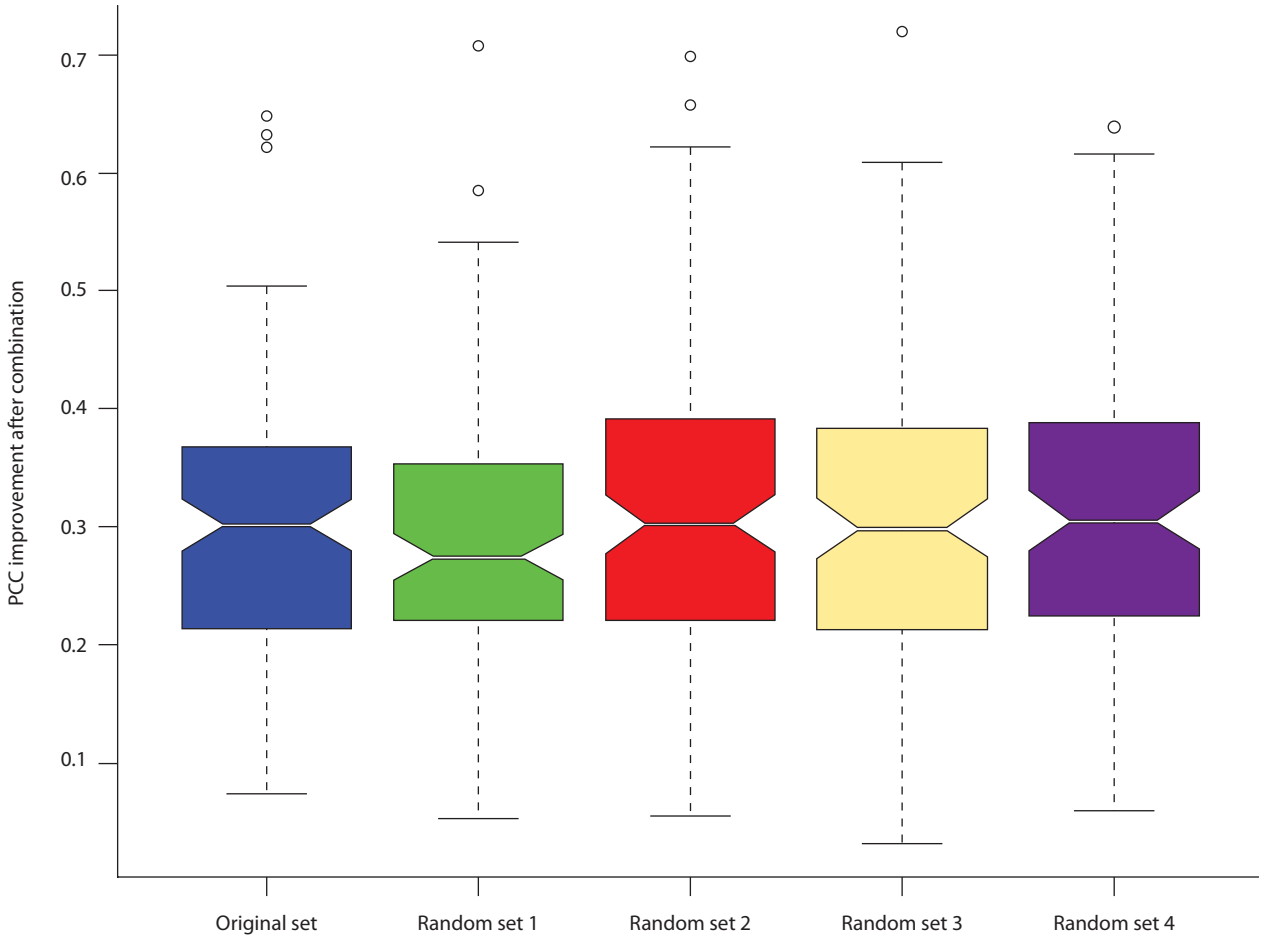


Figure 2. PCC improvements of time-shifted combined profiles with their targets in Li *et al.*'s [6] mouse liver development data and randomized data.

this n-dimensional space so that it can be expressed as a weighted sum of any n linearly independent vectors. When considering two regulators and their target gene, the time-shifting procedure is equivalent to searching through all combinations of two vectors to best represent the target vector. It can be expected that such searching will improve the correlation between the combined profile and the target even if the genes are unrelated. This can also be viewed as an overfitting problem as there are too many parameters in the model. If we can further restrict the number of parameters or their search space by properly introducing extra knowledge or hypotheses, the overfitting problem may be eased or solved.

In conclusion, our experiments illustrate that the observed significant correlation after time-shifting may not be able to be used to infer shifted cumulative regulation. Although we believe that there can be dynamic cumulative regulations in cells, we still need further data and other methods of data analysis to identify such regulations.

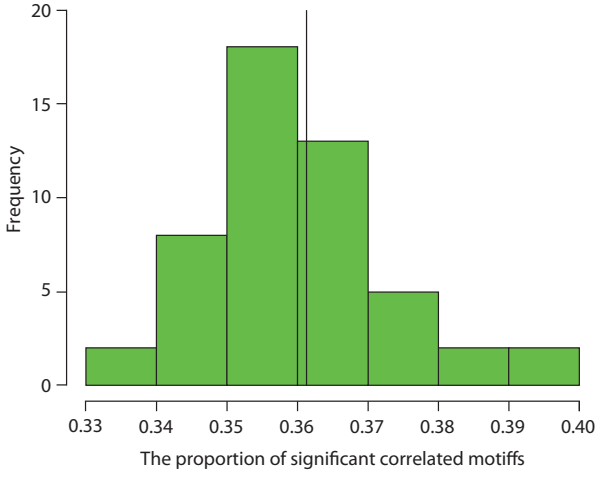
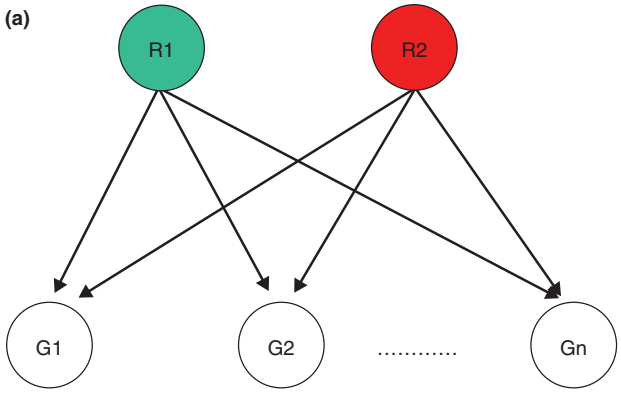
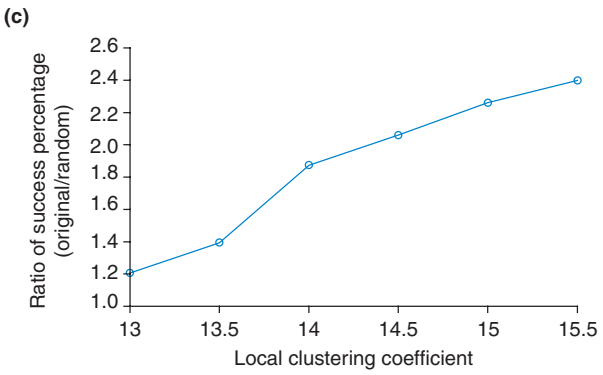


Figure 3. Histogram of the proportion of 'significant motifs' detected in the random data, and the proportion in the original data (indicated by the vertical line at 0.3612).





(b)

(i) $tr_{1,1} = tr_{1,2} = ... = tr_{1,n} = tr_1$

(ii) $tr_{2,1} = tr_{2,2} = ... = tr_{2,n} = tr_2$

(iii) $C_{1,1} = C_{1,2} = ... = C_{1,n} = C_1$

(iv) $C_{2,1} = C_{2,2} = ... = C_{2,n} = C_2$

(v) min $(tr_{1,i}, tr_{2,i}) \leq t_i$

(vi) $tr_{1,i} \leq 10$

(vii) $tr_{2,i} \leq 10$

(viii) $n > 1$

Figure 4. The constraints used in [1] to find biologically meaningful solutions and the ratios of success percentages between the original network and random networks. (a) The transcriptional regulatory structure of a convergence mode with two regulators (R1 and R2) and *n* target genes (G1, G2..., Gn). **(b)** The equations of introducing constraints used in [1]. The symbol tr_i indicates the time when the corresponding regulator *i* starts to function. C_i represents the conversion efficiency of the regulator *i*. C_{ij} and tr_{ij} indicate the values used for the regulator *i* and the target gene *j*. 'min' in equation (v) means the minimal value. **(c)** The ratio of the success percentage at each corresponding threshold in a relatively high score range (≥13) between the original network and random networks. This panel was generated using the same data as used for Figure 2b of [1].

Feng He, Jan Buer, An-Ping Zeng and Rudi Balling respond:

The observations reported by Ye *et al.* above describe the well-known problem of overfitting in computational biology. The experiments carried out by them seem to indicate that the shifted cumulative model reported by us [1] of using combinatorial expression profiles based on the integration of conversion efficiencies and of time delays may not be able to be used to infer shifted cumulative gene regulation.

However, there are essential differences between the experiments carried out by Ye *et al.* and those reported by us. The key difference is that we introduced more constraints in our original paper [1] than they did in their approach. We used a total of eight constraints (Figure 4) in order to limit the potential solution space for the two-regulator convergence modes (for three-regulator models, even more constraints were used).

In Figure 4b, equations (i) and (ii) require that the time when a given regulator starts to function is independent of its different individual target genes in the corresponding convergence mode. Note that the starting time for different individual regulators in a given convergence mode might be distinct from each other. This is also applied to the constraints concerning the conversion efficiency and the latest starting time of different regulators. Equations (iii) and (iv) ensure that the conversion efficiency used for a given regulator is the same for different target genes in the corresponding convergence mode. In addition to restricting our analysis to convergence modes with more than one target gene (equation (viii)), we have also included the requirement that the target genes are not activated (or suppressed) earlier than the time when the regulators start to function (equation (v)). Furthermore, the time when a given regulator starts to function is constrained to be within

one cell cycle (we used ten time points in the data of Cho *et al.* [2], which cover approximately one cycle) by equations (vi) and (vii). We explained all the constraints used in our work in the sections ' Quantification of shifted cumulative regulation of gene expression: principle of the approach' and 'Conversion efficiency and time delay among regulators' of our original paper [1]. All eight equations were used as constraints to optimize correlation between the combinatorial expression profile of the two regulators and the profiles of all their target genes at the same time (defined in paragraph 2 of the section ' Time delay from regulators to target genes' of [1]). The same constraints were also used for randomized networks (see the sections 'Significant difference between results for the original and randomly generated expression data and between results for the original network and randomly generated networks' and 'Multiple hypothesis testing' in [1]).

Ye *et al.* state, 'In these experiments, we removed regulator pairs that have only a single target …', which indicates that they have used the constraint indicated by equation (viii). They also write, 'The time shift between two regulators is fixed among their multiple targets'. This does not necessarily mean that the time when a given regulator starts to function is fixed among the multiple targets. Even if they fixed the time when the given regulator starts to function (indicated by our equation (i) and (ii)), all the other five important constraints (equations (iii), (iv), (v), (vi) and (vii)) out of the eight equations were apparently not used in their approach. It is also not clear whether they have used the same definition of optimal correlation as we did.

After using the eight constraints and the definition of optimal correlation, the success percentage at each corresponding threshold in a relatively high score range is significantly higher in the original network than that in random networks (for details see the section 'Significant difference between results for the original and randomly generated expression data and between results for the original network and randomly generated networks' in [1]). The average ratio of the success percentages between the original network and random networks in the range of significant correlation thresholds (≥13) is 1.865.

In addition, it seems to us from Figures 1 and 2 that Ye *et al.* have mixed the low scores and high scores together, which dilutes the contribution of high scores to the average values. This leads to a loss of information about the proportion of high scores and should not be done. In contrast to Ye *et al.*, we used only the scores in a relatively high range because those high scores might indicate biological relevance and cannot be easily obtained by chance. We therefore successfully reduced the overfitting problem, as shown in Figure 2b,d of the original paper [1].

The overfitting problem is one of the key issues in computational/systems biology and is often not appropriately addressed. In almost all modeling approaches attempts are made to strike a balance between the appropriate number of variables and constraints. We tried to integrate as many constraints as possible to maintain the biological relevance of the model. It seems to us that the inability of Ye *et al.* to derive significant differences between the experimental and random networks is due to the fact they have used far fewer constraints, leading to overfitting.

Published: 27 April 2011

References

1. He F, Buer J, Zeng AP, Balling R: **Dynamic cumulative activity of transcription factors as a mechanism of quantitative gene regulation.** *Genome Biol* 2007, **8:**R181.
2. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2:**65-73.
3. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M: **Genomic analysis of regulatory network dynamics reveals large topological changes.** *Nature* 2004, **431:**308-312.
4. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G: ***Saccharomyces*** **Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO).** *Nucleic Acids Res* 2002, **30:**69-72.
5. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
6. Li TT, Huang J, Jiang Y, Zeng Y, He F, Zhang MQ, Han Z, Zhang X: **Multi-stage analysis of gene expression and transcription regulation in C57/B6 mouse liver development.** *Genomics* 2009, **93:**235-242.
7. Liu Y, Jiang B, Zhang X: **Gene set analysis identifies master transcription factors in developmental courses.** *Genomics* 2009, **94:**1-10.
8. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M: **Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions.** *J Mol Biol* 2001, **314:**1053-1066.

doi:10.1186/gb-2011-12-4-404
Cite this article as: Ye C, *et al.*: Observations on shifted cumulative regulation. *Genome Biology* 2011, **12:**404.