Bioinformatics, 38(1), 2022, 173–178
doi: 10.1093/bioinformatics/btab577
Advance Access Publication Date: 12 August 2021
Original Paper

OXFORD

Systems biology

Using high-throughput multi-omics data to investigate structural balance in elementary gene regulatory network motifs

Alberto Zenere [1], Olof Rundquist[2], Mika Gustafsson[2] and Claudio Altafini[1,*]

[1]Division of Automatic Control, Department of Electrical Engineering, Linköping University, SE-58183 Linköping, Sweden and [2]Bioinformatics, Department of Physics, Chemistry and Biology, Linköping University, SE-58183 Linköping, Sweden

*To whom correspondence should be addressed.

Associate Editor: Lenore Cowen

Received on February 15, 2021; revised on July 4, 2021; editorial decision on July 23, 2021; accepted on August 10, 2021

Abstract

Motivation: The simultaneous availability of ATAC-seq and RNA-seq experiments allows to obtain a more in-depth knowledge on the regulatory mechanisms occurring in gene regulatory networks. In this article, we highlight and analyze two novel aspects that leverage on the possibility of pairing RNA-seq and ATAC-seq data. Namely we investigate the causality of the relationships between transcription factors, chromatin and target genes and the internal consistency between the two omics, here measured in terms of structural balance in the sample correlations along elementary length-3 cycles.

Results: We propose a framework that uses the a priori knowledge on the data to infer elementary causal regulatory motifs (namely chains and forks) in the network. It is based on the notions of conditional independence and partial correlation, and can be applied to both longitudinal and non-longitudinal data. Our analysis highlights a strong connection between the causal regulatory motifs that are selected by the data and the structural balance of the underlying sample correlation graphs: strikingly, $> 97\%$ of the selected regulatory motifs belong to a balanced subgraph. This result shows that internal consistency, as measured by structural balance, is close to a necessary condition for 3-node regulatory motifs to satisfy causality rules.

Availability and implementation: The analysis was carried out in MATLAB and the code can be found at https://github.com/albertozenere/Multi-omics-elementary-regulatory-motifs.

Contact: claudio.altafini@liu.se

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

1 Introduction

One of the trends in the field of gene regulatory network (GRN) inference is to increase the inference power of the data by combining multiple omics techniques. For instance, in recent years the integration of RNA sequencing (RNA-seq) and Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) data has given promising results, see Ackermann *et al.* (2016), Calderon *et al.* (2019) and Ramirez *et al.* (2017).

This integration can be carried out in different ways. Some studies use a two-step approach, where for instance ATAC-seq is used to obtain a large set of candidate interactions and then RNA-seq is used to prune this set and to identify a reliable subset of high-confidence transcription factor (TF)-target gene interactions, see e.g. Miraldi *et al.* (2019) and Johnson *et al.* (2020). Alternatively, many studies analyze the correlation between chromatin peaks and target genes, see e.g. Hendrickson *et al.* (2018) and Starks *et al.* (2019). Unlike these studies, we propose to consider simultaneously three layers of transcription: TFs, chromatin peaks and target genes. The first and third level are quantified via RNA-seq and the second via ATAC-seq. In other words, we consider not only the correlation between peak and target gene, but also between peak-TF and TF-target gene. We show that this method can be used to identify new cross-layers elementary regulatory motifs involving TFs, chromatin peaks and target genes. A necessary condition for performing this analysis is to have paired ATAC-seq and RNA-seq data, as it is in our case.

More precisely, we decided to focus on two classes of three-node regulatory motifs, formed by transcription factors, chromatin peaks and target genes. The regulatory motifs we have chosen to work with are the chains and forks shown in Figure 1, because they encode conditional independence relationships: two nodes in these regulatory motifs become independent once their values are conditioned on that of the third node (Christopher, 2006). Such conditional

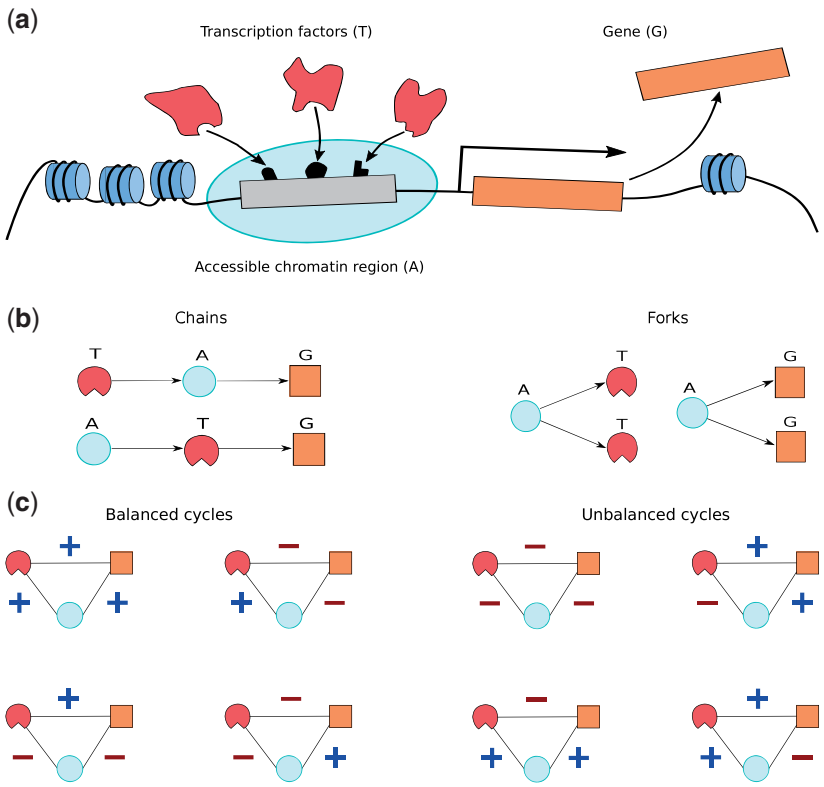


Fig. 1. Workflow of this article. (**a**) Schematic depiction of two key events that lead to gene expression: (1) the chromatin region around the promoter is loose and accessible to TFs binding, and (2) available TFs bind to specific DNA sequences in the promoter region of the gene. (**b**) Possible regulatory motifs. Which event precedes the other is still under investigation, thus several causal three-node regulatory motifs can be associated to represent the regulatory interactions between TFs, chromatin regions and target genes. (**c**) Balanced and unbalanced cycles corresponding to the undirected graph of $A \to T \to G$ (i.e. chromatin → TF → gene). Plus and minus signs denote positive and negative correlation values between the corresponding nodes

independence can be explored in a systematic way using partial correlation (Baba *et al.*, 2004). Partial correlation has been used extensively in the context of causal inference in GRNs to investigate gene-gene interactions, see Opgen-Rhein and Strimmer (2007) and Yiming *et al.* (2014). Here, instead we use sample partial correlations as a tool to screen all possible three-node regulatory motifs, based on a computational map of all possible interactions among TFs, chromatin peaks and target genes we constructed. Only chains and forks that pass the partial correlation test are considered as 'selected' regulatory motifs based on the data.

Another concept that has been used in biological networks is structural balance (hereafter simply denoted balance), see Facchetti *et al.* (2013), Iacono *et al.* (2010) and Mangan and Alon (2003). Notice that, in the context of signed networks, balance is synonym to coherence, although the latter assumes different meanings in other fields (e.g. Cadzow and Solomon, 1987). Balance is associated to signed cycles, in particular a cycle is balanced if it has an even number of negative edges and unbalanced otherwise. In previous studies (Facchetti *et al.*, 2013; Iacono *et al.*, 2010; Mangan and Alon, 2003) the focus was on counting balanced motifs in a given biological network, and the common result was that balanced motifs were enriched over unbalanced ones. Here, balance is instead associated to the sample correlations of triplets of nodes that belong to different omics, which form our elementary regulatory motifs. Interestingly, in our analysis we also find a similar property: the triplets of correlations selected by the data for our chain and fork regulatory motifs tend to be enriched for balanced triangles, while the percentage of unbalanced triangles is significantly lower than in random data, suggesting that the notion of balance can be observed in experimental data, even when these span different omics.

We have gathered four publicly available datasets of paired RNA-seq and ATAC-seq experiments on human immune cells, see Table 1. Datasets A and B represent time-series of primary human naïve CD4$^+$T during early T-helper type 1 differentiation (Magnusson *et al.*, 2019). The difference between A and B is that in the latter the activation was performed in the presence of progesterone. Datasets C and D are time-series experiments on human monocyte-derived dendritic cells under infection with HIV-1, where the latter serves as mock experiment (Johnson *et al.*, 2020). For details, see the corresponding publications.

Table 1. List of paired RNA-seq and ATAC-seq datasets used in this study

Index	Cell type	Availability and reference
A	Human Th1	E-MTAB-7775, E-MTAB-10444, (Magnusson *et al.*, 2019)
B	Human Th1	E-MTAB-10423, E-MTAB-10444
C	Human DC	GSE125817 (Johnson *et al.*, 2020)
D	Human DC	GSE125918 (Johnson *et al.*, 2020)

2 Materials and methods

2.1 TF-peak-target gene map

Assume mRNA expression levels of transcription factors (T) and target genes (G) have been measured with RNA-seq, while the accessibility of chromatin regions (A), also called peaks, has been quantified by ATAC-seq.

ATAC-seq data can also be used to build interaction maps between A—G and A—T. More precisely, each peak was mapped to the closest gene, with the constraint that its TSS must be located within a maximum distance of 3000 base pairs (bp) from either side

of the peak edges, see e.g. Corces *et al.* (2016), Fullard *et al.* (2018) and Wu *et al.* (2018). In addition, whenever such a target gene was found, we have also associated the peak to every gene whose TSS was situated within a distance of ± 5000 from the TSS of the aforementioned (i.e. closest) gene, as done e.g. in Yu *et al.* (2015). Footprinting and motif analysis was then performed to associate each peak to a list of potential TFs binding events. See Supplementary Materials and Methods for more details. The result is two sets of interactions: between chromatin regions and target genes (A–G), and between TFs and chromatin regions (T–A). From there, we can retrieve a third set of interactions, between TFs and target genes (T–G), by connecting TFs and target genes that share at least one common chromatin region in the computational templates A–G and A–T. Altogether, the three combined interaction mappings form what we call a *multi-omics TF-peak-target gene map*.

Such mapping typically contains a significant amount of false positives, as highlighted in Yan *et al.* (2020). In this work, we address the issue by combining the notions of dynamical correlation, partial correlation and balance, which we now introduce.

2.2 Dynamical correlation

Calculating correlation coefficients in longitudinal studies requires appropriate tools to take into account the dependency between (often irregularly spaced) time points as well as latent factors, see Yule (1926) and Granger (2007). Failing to do so will introduce bias in the correlation coefficients and create false connections between the data. One of the approaches to render the data normally distributed is to use the notion of dynamical correlation. In particular we focus on the definition introduced by Opgen-Rhein and Strimmer (2006) and reviewed in Supplementary Materials and Methods. From now on the adjective 'dynamical' will be implicitly assumed when dealing with correlation or partial correlation.

2.3 Partial correlation

A partial correlation reflects the strength of a linear relationship between two variables after controlling for potential effects coming from other variables. The concept has received wide attention in different fields, such as GRN inference (Opgen-Rhein and Strimmer, 2007; Yiming *et al.*, 2014; Zampieri *et al.*, 2008) and brain functional connectivity (Reid *et al.*, 2019). We denote the partial correlation coefficient between the variables X and Y given Z with $\mathbf{R}(X, Y|Z)$, which is expressed in formula by

$$\mathbf{R}(X, Y|Z) = \frac{\mathbf{R}(X, Y) - \mathbf{R}(X, Z)\mathbf{R}(Y, Z)}{\sqrt{(1 - \mathbf{R}(X, Z)^2)(1 - \mathbf{R}(Y, Z)^2)}}. \quad (1)$$

In particular, partial correlations can be used to test causal interactions in the data. To illustrate its usefulness, consider the simplest case of three variables: X, Y and Z. Assume X, Y and Z are part of a regulatory chain, for instance X regulates Z, which in turn regulates Y: $X \to Z \to Y$, see Figure 1b, left. This common regulatory motif is characterized by the fact that the dependence between X and Y is mediated by Z and that X and Y become independent once we 'project away' the information due to Z (Baba *et al.*, 2004). More formally, if we consider X, Y and Z as (Gaussian) random variables, the joint probability distribution of the regulatory motif $X \to Z \to Y$ factorizes as $p(X, Y, Z) = p(X)p(Z|X)p(Y|Z)$ where $p(X)$ is the probability distribution of the variable X and $p(Z|X)$ is the conditional probability distribution of Z given X. Conditioning over Z and using Bayes rule

$$p(X, Y|Z) = \frac{p(X, Y, Z)}{p(Z)} = p(X|Z)p(Y|Z)$$

shows that once conditioned on Z, the joint probability between X and Y factorizes, i.e. X and Y are conditionally independent given Z: $X \perp Y|Z$. Technically we have that X and Y are conditionally independent given Z when the residuals are uncorrelated. In practice we can setup a test using the sample partial correlation $\mathbf{R}(X, Y|Z)$ and consider as conditional independence the following condition:

$$X \perp Y|Z \iff |\mathbf{R}(X, Y|Z)| < \theta_1,$$

where θ_1 is a threshold calculated in Supplementary Materials and Methods.

A similar observation can be made for forks, $X \leftarrow Z \to Y$, see Figure 1b, right. In fact the apparent correlation between X and Y disappears once we control for the effects of the common regulator Z. This regulatory motif is also characterized by conditional independence.

2.4 Structural balance

Given three variables X, Y and Z let us compute their pairwise correlations $\mathbf{R}(X, Y)$, $\mathbf{R}(X, Z)$ and $\mathbf{R}(Y, Z)$. These three correlations form an undirected cycle of length three (i.e. a triangle). We say that such a cycle is balanced if $\mathbf{R}(X, Y) \cdot \mathbf{R}(X, Z) \cdot \mathbf{R}(Y, Z) > 0$. In the following section, balance will be used as a test of internal consistency among the variables involved in the basic chain and fork regulatory motifs.

3 Results

3.1 Elementary gene regulatory motifs and their conditional independence

The approach we follow in this article is to break down the complexity of GRNs by analyzing elementary causal regulatory motifs. In particular, we start our analysis by modeling the interplay between TF and chromatin accessibility, which leads to gene expression. We show that it can be represented as two regulatory motifs, $T \to A \to G$ and $A \to T \to G$.

Chromatin accessibility at the promoter region can enable (or amplify) the effect of TFs on gene expression. Consider the example of a gene with a unique transcriptional activator: it is plausible to assume that the rate of its transcription depends on the state of the TF binding region, and that the opening (closing) of the chromatin surrounding it is reflected in a higher (lower) ratio between gene transcription and TF availability. The opposite happens for a TF which is a transcriptional inhibitor. In terms of causal graphs, we can associate this example with the chain regulatory motif $T \to A \to G$, where the relationship between T and G is mediated by A. As discussed in Section 2.3, chain regulatory motifs are characterized by a conditional independence. Denoting with $\mathbf{R}(T, G|A)$ the sample partial correlation between T and G conditioned on A, then T and G are considered conditionally independent given A if $|\mathbf{R}(T, G|A)| < \theta_1$. When this condition is satisfied we say that the regulatory motif $T \to A \to G$ is *selected by the data*, i.e. that the data provide a (statistically significant) evidence in support of the existence of the regulatory motif. To check for spurious conditionally independent results caused by correlations close to zero before conditioning, we discarded the cases where $|\mathbf{R}(T, G)| < \theta_0$; here, θ_0 is the threshold obtained when the number of controlled variables is set to zero. This procedure was repeated systematically on the $\sim 4 \cdot 10^5$ (T, A, G) triplets present in our interaction map. For each of the four datasets we consider in this study, ~5–15% of the chain regulatory motifs were selected for a total of ~1–2 regulatory motifs per target gene. The results of this analysis are summarized in Table 2.

Alternatively, the interplay between TF and chromatin accessibility can be represented by the regulatory motif $A \to T \to G$. In fact, chromatin accessibility does not lead to gene expression unless a suitable TF binds, and we can argue that the concentration of TF amplifies the effect of chromatin accessibility (for instance due to the presence of stable TF binding to the promoter region), thus leading to the alternative chain model $A \to T \to G$. Also in this case, the conditional independence encoded in this chain can be tested using partial correlation. Interestingly, the two regulatory motifs selected by the data almost never contain simultaneously the same (A, T, G) triplet (the overlap is significantly low as measured by a hypergeometric test on the contingency table of Table 3, (P-value = $< 10^{-16}$). Selecting different (A, T, G) triplets is significant, since it suggests that the two regulatory motifs are non-equivalent and supports the decision of taking both into account.

Table 2. Overview of the datasets. (Upper) We report the total number of regulatory motifs (and the percentage of balanced ones) present in the TF-peak-target map. (Middle) Next, we test if each regulatory motif is characterized by a statistically large balance ratio (see Section 3.2 for details on how the statistical test was built); fold change indicates the ratio between the value observed in the data and the mean of the null distribution. (Lower) Lastly, we report the number of regulatory motifs that pass the conditional independence test described in Supplementary Materials and Methods and how many of them belong to an unbalanced cycle.

	Number of regulatory motifs in the data (of which balanced)		
Dataset	Chains	$T_1 \leftarrow A \rightarrow T_2$	$G_1 \leftarrow A \rightarrow G_2$
A	408088 (71%)	9134221 (72%)	7736 (77%)
B	367308 (67%)	8227783 (67%)	7100 (72%)
C	309675 (63%)	10686804 (65%)	3456 (65%)
D	255324 (70%)	9004018 (68%)	3419 (74%)
	Enrichment of balanced regulatory motifs: *P*-value, fold change		
	Chains	$T_1 \leftarrow A \rightarrow T_2$	$G_1 \leftarrow A \rightarrow G_2$
A	$< 10^{-16}$, 1.11	$< 10^{-16}$, 1.11	$< 10^{-16}$, 1.19
B	$3.60 \cdot 10^{-7}$, 1.04	$1.16 \cdot 10^{-7}$, 1.04	$< 10^{-16}$, 1.11
C	not significant	not significant	$2.50 \cdot 10^{-3}$, 1.02
D	$< 10^{-16}$, 1.08	$< 10^{-16}$, 1.07	$< 10^{-16}$, 1.16
	Number of selected regulatory motifs (of which unbalanced)		
	$A \rightarrow T \rightarrow G$	$T \rightarrow A \rightarrow G$	$T_1 \leftarrow A \rightarrow T_2$	$G_1 \leftarrow A \rightarrow G_2$
A	21138 (4)	19272 (3)	419330 (32)	298 (0)
B	26573 (13)	12627 (12)	290724 (191)	184 (0)
C	37856 (440)	23427 (422)	808439 (13855)	187 (2)
D	38435 (324)	15154 (309)	519882 (11838)	202 (3)

Note: Since $A \rightarrow T \rightarrow G$ and $T \rightarrow A \rightarrow G$ correspond to the same undirect graph we use the more general term 'Chains' to denote (A, T, G) triplets.

Table 3. Contingency table between the number of selected T → A → G and A → T → G regulatory motifs in dataset A

		T →A →G	
		Selected	Non-selected
A →T →G	Selected	302	20 840
	Non-selected	18 973	367 973

Note: See Supplementary Results for the contingency tables of datasets B, C and D.

A gene is normally regulated by multiple TFs, and associated with multiple ATAC-seq peaks. In fact, footprinting analysis reveals that up to 100 TFs can interact with the same promoter; moreover a single chromatin region can be associated with multiple target genes. To model this massive co-regulation we used other elementary three-node regulatory motifs, like the forks shown in Figure 1b. In particular we decided to focus on the regulatory motifs $T_1 \leftarrow A \rightarrow T_2$ and $G_1 \leftarrow A \rightarrow G_2$. In dataset A, for example, the number of such regulatory motifs is 9 134 221 and 7736, of which 419 362 and 298 were selected by a partial correlation test similar to the one described above, see Table 2.

3.2 Structural balance as a data consistency criterion

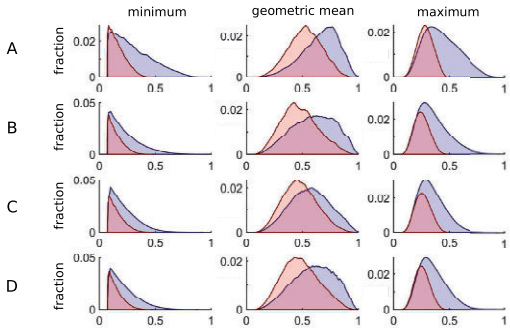
In this work balance assumes the meaning of an intrinsic test of compatibility between the regulatory interactions in the data. For instance, if for a triplet $T \rightarrow A \rightarrow G$ the sample correlations $\mathbf{R}(T, A)$ and $\mathbf{R}(A, G)$ are both positive, suggesting that we have two activatory regulations $T \xrightarrow{+} A$ and $A \xrightarrow{+} G$, then we expect that also the edge between T and G has positive correlation. Proceeding in this way means associating to the chain regulatory motif $T \rightarrow A \rightarrow G$ an undirected cycle, formed by the branches T—A—G and T—G and checking if the triangle (T, A, G) has balanced correlations. When this does not happen, then our data shows internal inconsistency, i.e. the signs of the three correlations $\mathbf{R}(T, A), \mathbf{R}(A, G)$ and $\mathbf{R}(T, G)$ are incompatible. A similar construction can be carried out for the other regulatory motifs mentioned above and we can then proceed to checking the balance (i.e. internal consistency) of the resulting triangles, see Figure 1c.

It is interesting to observe that the data appears to be significantly consistent, as measured by the percentage of balanced cycles in the network. To retrieve the null distribution of the percentage of balanced cycles, we used a bootstrapping approach. Namely, we generated a population of 50 000 triplets of Gaussian random signals, having the same number of time-points as the data. Thereafter we extracted 10 000 sub populations, comprising 10 000 triplets each, and we calculated their balance ratio, thus leading to the null distribution. Balanced regulatory motifs appear to be significantly over-represented in the data; as can be seen in Table 2, both chain and fork regulatory motifs are enriched for balance in almost all the datasets.

Not only balanced triplets are over-represented in the data, they also consist of edges corresponding to the correlations in the network with the highest absolute values. To formalize this observation, we have associated each triplet to scalar measures that quantify the magnitude of the corresponding correlations. We have chosen three measures: *geometric mean*, *minimum* and *maximum*; although similar results can be obtained using other measures, such as mean and harmonic mean. A Kolmogorov-Smirnov test reveals that the distribution of each measure differs significantly (every P-value is $< 10^{-16}$) between balanced and unbalanced regulatory motifs, where the former show higher average values, as seen in Figure 2.

3.3 Structural balance is a necessary condition for conditional independence

The categorization of triplets into balanced and unbalanced sheds light also on the conditional independence of the variables involved. As can be seen in Figure 3, the distributions of partial correlation



Dataset	regulatory motif category	Scalar measure mimimum	geometric mean	maximum
A	balanced	0.30	0.43	0.66
	unbalanced	0.16	0.28	0.52
B	balanced	0.23	0.35	0.60
	unbalanced	0.14	0.25	0.47
C	balanced	0.21	0.34	0.56
	unbalanced	0.15	0.26	0.47
D	balanced	0.22	0.34	0.60
	unbalanced	0.15	0.25	0.48

Fig. 2. For each dataset, we gather all the regulatory motifs in Figure 1, then for each regulatory motif we calculate minimum, geometric mean and maximum of its three correlations. Blue denotes the distributions obtained in the balanced cycles, red the unbalanced. In the table below we summarize the mean of each scalar measure, computed separately in the balanced and unbalanced case

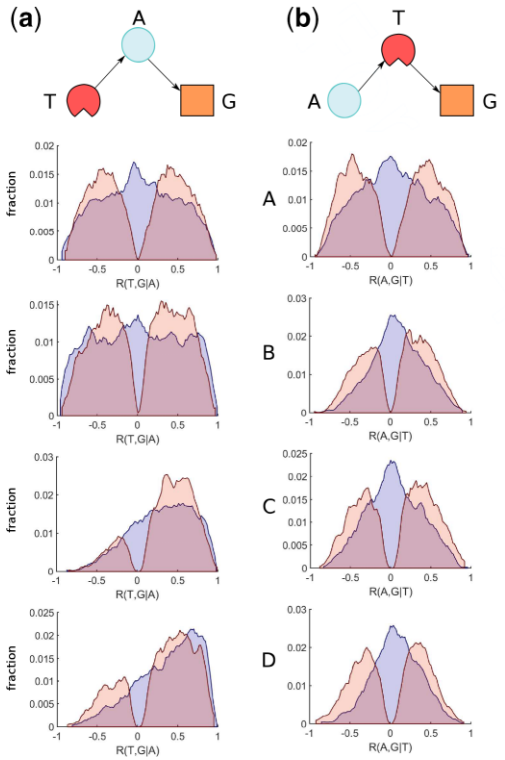


Fig. 3. (**a**) $T \to A \to G$ regulatory motif and corresponding distribution of $\mathbf{R}(T, G|A)$, divided in balanced (blue) and unbalanced (red) cycles. The balanced and unbalanced distributions are normalized with respect to their total count independently. (**b**) Similar analysis for the $A \to T \to G$ regulatory motif and the corresponding distribution of $\mathbf{R}(A, G|T)$

values in chains differ significantly between balanced and unbalanced cycles. In particular the latter distributions are characterized by a 'drop' around zero, meaning that *unbalanced cycles rarely lead to conditional independence*. A similar observation holds for fork regulatory motifs as well, see Supplementary Results. What stands out from the analysis is that balance is 'almost' a necessary condition for conditional independence. Strikingly, for all four datasets, $>$ 97% of the selected chain and fork regulatory motifs belong to a balanced cycle. The enrichment of balance among selected regulatory motifs is statistically confirmed by a hypergeometric test that compares the balance ratio among the selected regulatory motifs and among all the regulatory motifs in the network (P-value $< 10^{-16}$).

3.4 Balanced and selected regulatory motifs are conserved under different cell stimuli

Datasets A and B come from the same cell type under partially similar stimuli. Both datasets have been generated from Th cells differentiated under Th1 polarizing conditions, with the difference that for dataset B the Th1 polarization was done in presence of progesterone. Accordingly, they are characterized by similar TF-peak-target gene mappings: $\sim$50% of $A \to T \to G$ and $T \to A \to G$, $\sim$ 40% of $T_1 \leftarrow A \to T_2$ and $\sim$80% of $G_1 \leftarrow A \to G_2$ regulatory motifs are shared by the two datasets. When we focus on this pool of common regulatory motifs we observe that a significant portion is balanced in both datasets. More precisely, there is a mild but significant overlap between the regulatory motifs that are balanced in A and those that are balanced in B, see Table 4. Interestingly, the relationship becomes stronger when we look at those regulatory motifs (except $A \to T \to G$) that are selected in dataset A and B.

A similar comparison can also be carried out between datasets C and D, see Supplementary Results, leading to similar conclusions.

4 Discussion

In this article, we consider two alternative chain models to represent the interplay that exists between TFs and chromatin modeling in regulating gene expression, differing for the causality direction between A and T. Although the precise mechanisms are still unclear, several studies have showed that the regulation can happen in both directions: TFs affects chromatin accessibility and viceversa (Li *et al.*, 2007; Li and Leonard, 2018; Stadhouders *et al.*, 2018). Hence we decided to consider both $A \to T \to G$ and $T \to A \to G$ as distinct plausible regulatory motifs. In our case, the two sets of (A, T, G) triplets that fit the conditional independence hypothesis for these regulatory motifs are significantly disjoint. This is in accordance with the notion that in some physiological situations chromatin remodeling precedes TF binding whereas in other situations it is the TF binding that leads to chromatin remodeling (Choukrallah and Matthias, 2014).

In this work, we use balance as a consistency criterion. In the context of biological networks, multiple studies have already highlighted that GRNs are enriched for balanced patterns (Facchetti *et al.*, 2013; Mangan and Alon, 2003) and altogether tend to be close to monotone systems (Ma'ayan *et al.*, 2008). However the application of these ideas to sample correlations multi-omics data in particular has never been explored before, at least in the authors' knowledge. Indeed, the observation that combined RNA-seq and ATAC-seq data is predominantly balanced provides evidence that it is for the most part internally consistent. It is interesting to couple this observation with the fact that $>$ 97% of selected (i.e. conditionally independent) regulatory motifs were found to belong to a balanced cycle. Conditional independence is associated to low correlation values upon conditioning, thus it may be surprising that unbalanced cycles (characterized by lower correlation values) rarely lead to conditional independence.

We have also observed that the peaks that belong to chain regulatory motifs selected by the data are, on average, closer to the TSS

Table 4. To test if there exists a relationship between which regulatory motifs are balanced (resp. selected) in dataset A and B we performed a hypergeometric test that compares the ratio of balanced (resp. selected) regulatory motifs in dataset A with the same quantity but when we restrict only to regulatory motifs that are also balanced (resp. selected) in dataset B

	Relationship between 'balanced in A' and 'balanced in B' (*P*-value, FC)	Relationship between 'selected in A' and 'selected in B' (*P*-value, FC)
$A \to T \to G$	$< 10^{-16}$, 1.03	not significant
$T \to A \to G$	$< 10^{-16}$, 1.03	2.22×10^{-9}, 1.32
$T_1 \leftarrow A \to T_2$	$< 10^{-16}$, 1.02	$< 10^{-16}$, 1.36
$G_1 \leftarrow A \to G_2$	$2.04 \cdot 10^{-11}$, 1.03	0.04, 1.64

Note: FC indicates the fold change of the latter with respect to the former quantity.

of the corresponding target gene (see Supplementary Section S2.4). From a biological perspective, this suggests that the regulation of gene transcription is primarily mediated by the remodeling of chromatin in near proximity of the TSS.

Another application of the ideas presented in this article it to use conditional independence to identify relevant TF-target interactions from the data. A thorough analysis has been performed in Supplementary Section S2.5, which shows that conditional independence highlights relevant interactions supported by the literature.

Lastly, it should be noted that the techniques presented in this article can readily be applied to non-longitudinal data. In fact, chains and forks are also characterized by conditional independence in that case, and dynamical correlation reduces to standard correlation in the case of steady-state data and multiple replicates (i.e. non-longitudinal data). Conceptually, the same remark can be made regarding single cell (sc) data, the only difference being that correlations must necessarily be computed across different cells. However, the limited depth of the currently available methods, especially for scATAC-seq (Chen *et al.*, 2019), poses serious technical limitations.

Funding

This work was supported by the Swedish Foundation for Strategic Research [SB16-0011].

Conflict of Interest: none declared.

Data availability

The data underlying this article are available in ArrayExpress under accession numbers E-MTAB-7775, E-MTAB-10423 and E-MTAB-10444 and in Gene Expression Omnibus under accession numbers GSE125817 and GSE125918.

References

Ackermann,A.M. *et al.* (2016) Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Mol. Metab.*, **5**, 233–244.

Baba,K. *et al.* (2004) Partial correlation and conditional independence as measures of conditional independence. *Aust. N. Zeal. J. Stat.*, **46**, 657–664.

Cadzow,J.A. and Solomon,O.M. (1987) Linear modeling and the coherence function. *IEEE Trans. Acoustics Speech Signal Process.*, **35**, 19–28.

Calderon,D. *et al.* (2019) Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat. Genet.*, **51**, 1494–1505.

Chen,H. *et al.* (2019) Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.*, **20**, 1–25.

Choukrallah,M.A. and Matthias,P. (2014) The interplay between chromatin and transcription factor networks during B cell development: who pulls the trigger first? *Frontiers Immunol.*, **5**, 1–11.

Christopher,M.B. (2006) *Pattern Recognition and Machine Learning*. Springer, Berlin.

Corces,M.R. *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, **48**, 1193–1203.

Facchetti,G. *et al.* (2013) A rate-distortion theory for gene regulatory networks and its application to logic gate consistency. *Bioinformatics (Oxford, England)*, **29**, 1166–1173.

Fullard,J.F. *et al.* (2018) An atlas of chromatin accessibility in the adult human brain. *Genome Res.*, **28**, 1243–1252.

Granger,C.W. (2007) Spurious regressions in econometrics. *Companion Theor. Econometr.*, **2**, 557–561.

Hendrickson,D.G. *et al.* (2018) Simultaneous profiling of DNA accessibility and gene expression dynamics with ATAC-seq and RNA-seq. *Methods Mol. Biol.*, **1819**, 317–333.

Iacono,G. *et al.* (2010) Determining the distance to monotonicity of a biological network: a graph-theoretical approach. *IET Syst. Biol.*, **4**, 223–235.

Johnson,J.S. *et al.* (2020) A comprehensive map of the monocyte-derived dendritic cell transcriptional network engaged upon innate sensing of HIV. *Cell Rep.*, **30**, 914–931.e9.

Li,P. and Leonard,W.J. (2018) Chromatin accessibility and interactions in the transcriptional regulation of T cells. *Front. Immunol.*, **9**, 1–8.

Li,B. *et al.* (2007) The Role of chromatin during transcription. *Cell*, **128**, 707–719.

Ma'ayan,A. *et al.* (2008) Proximity of intracellular regulatory networks to monotone systems. *IET Syst. Biol.*, **2**, 103.

Magnusson,R. *et al.* (2019) A validated strategy to infer protein biomarkers from RNA-Seq by combining multiple mRNA splice variants and time-delay. *bioRxiv* doi: 10.1101/599373.

Mangan,S. and Alon,U. (2003) Structure and function of the feed-forward loop network motif. *Proc. Natl. Acad. Sci. USA*, **100**, 11980–11985.

Miraldi,E.R. *et al.* (2019) Leveraging chromatin accessibility for transcriptional regulatory network inference in T helper 17 cells. *Genome Res.*, **29**, 449–463.

Opgen-Rhein,R. and Strimmer,K. (2006) Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *Revstat*, **4**, 53–65.

Opgen-Rhein,R. and Strimmer,K. (2007) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, **1**, 37.

Ramirez,R.N. *et al.* (2017) Dynamic gene regulatory networks of human myeloid differentiation. *Cell Syst.*, **4**, 416–429.e3.

Reid,A.T. *et al.* (2019) Advancing functional connectivity research from association to causation. *Nat. Neurosci.*, **22**, 1751–1760.

Stadhouders,R. *et al.* (2018) Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nat. Genet.*, **50**, 238–249.

Starks,R.R. *et al.* (2019) Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. *Epigenet. Chromatin*, **12**, 1–16.

Wu,J. *et al.* (2018) Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature*, **557**, 256–260.

Yan,F. *et al.* (2020) From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.*, **21**, 22.

Yiming,Z. *et al.* (2014) Biological network inference using low order partial correlation. *Methods*, **69**, 266–273.

Yu,G. *et al.* (2015) ChIP seeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.

Yule,G.U. (1926) Why do we Sometimes get Nonsense-Correlations between Time-Series?–A Study in Sampling and the Nature of Time-Series. *J. R. Stat. Soc.*, **89**, 1.

Zampieri,M. *et al.* (2008) Discerning static and causal interactions in genome-wide reverse engineering problems. *Bioinformatics*, **24**, 1510–1515.