Loregic: A Logic-circuit method to characterize cooperativity of regulatory factors

ABSTRACT

Regulatory factors act cooperatively to control gene expression. Leveraging on the vast amount of functional genomics data available, it is now possible to conduct a comprehensive and systematical analysis of regulatory factors' functional cooperativity. We present Loregic, a novel computational method that integrates gene expression and regulatory network data, to identify and characterize the cooperativity of regulatory elements using logic-circuit models. Loregic is freely available as a general-purpose tool via https://github.com/gersteinlab/Loregic. We describe the basic regulatory triplet consisting of two regulatory factors (RFs) acting on a common target, using a two-input-one-output logic gate model. We use binarized gene expression data, to score the agreement between a triplet's cross-sample expression and the idealized expression pattern of each 16 possible logic gates. A high score suggests a strong cooperativity between two RFs to target following the corresponding logic gate pattern. To demonstrate Loregic's versatility, we apply it to yeast cell cycle and human cancer datasets. In yeast, we validate our method using data from transcription factor (TF) knockout experiments and we are able to predict logical cooperation among TFs. In human, we integrate ENCODE ChIP-Seq and TCGA RNA-Seq expression data to study cooperativity between and among TFs and micro-RNAs. We find that the oncogenic TFs such as MYC, can be modeled as acting independently from other TFs, but antagonistically with micro-RNAs. Finally, we explore Loregic's applicability to other regulatory features. As such as we use it for 1) discovery and classification of indirectly bound TFs, and 2) prediction of logical operations in feed-forward loops, a special type of regulatory triplets in which one TF regulates both the target gene and the other TF.

Contact: pi@gersteinlab.org

1 INTRODUCTION

The rapidly increasing amount of high throughput sequencing data offers novel and diverse resources to probe molecular functions and activities on genome scale. Integrating and mining these various large-scale datasets is both a central priority and a great challenge for the field of functional genomics. Reaching these goals necessitates the development of specialized computational tools.

*To whom correspondence should be addressed.

Daifeng Wang 8/17/14 10:51 PM
Deleted: gene

Daifeng Wang 8/17/14 10:51 PM
Deleted: characteristics with

Daifeng Wang 8/17/14 10:51 PM
Deleted: of all

Daifeng Wang 8/17/14 10:51 PM
Deleted: the regulatory activities of the

Daifeng Wang 8/17/14 10:51 PM
Deleted: the method's applicability

Daifeng Wang 8/17/14 10:51 PM
Deleted: use Loregic

Daifeng Wang 8/17/14 10:51 PM
Deleted: study

Daifeng Wang 8/17/14 10:51 PM
Deleted: factors (TFs) regulatory activity, and validate the results using

Daifeng Wang 8/17/14 10:51 PM
Deleted: experimental datasets. Additionally we investigate the

Daifeng Wang 8/17/14 10:51 PM
Deleted: operations

Daifeng Wang 8/17/14 10:51 PM
Deleted: , and miRNAs in

Daifeng Wang 8/17/14 10:51 PM
Deleted: cancer. For this

Daifeng Wang 8/17/14 10:51 PM
Deleted: .

Daifeng Wang 8/17/14 10:51 PM
Deleted: in acute myeloid leukemia,

Daifeng Wang 8/17/14 10:51 PM
Deleted: miRNAs. Next

Daifeng Wang 8/17/14 10:51 PM
Deleted: the algorithm's

Daifeng Wang 8/17/14 10:51 PM
Deleted: We

Daifeng Wang 8/17/14 10:51 PM
Deleted: Loregic

Daifeng Wang 8/17/14 10:51 PM
Deleted: the

Daifeng Wang 8/17/14 10:51 PM
Deleted: . We also predict

Daifeng Wang 8/17/14 10:51 PM
Deleted: triplet

Daifeng Wang 8/17/14 10:51 PM
Deleted: In summary, Loregic is a valuable computational tool that describes the complex process of gene regulation in terms of logic operations. The present method can be further extended to a ... [1]

Gene expression is a complex process that achieves both spatial and temporal control through the coordinated action of multiple regulatory factors [1, 2, 3]. These regulatory factors affecting gene expression take several forms, such as transcription factors, which directly or indirectly bind DNA at promoters and enhancers of their target genes, and non-coding RNAs (e.g. microRNAs) [4, 5]. RFs can act as activators or repressors, but ultimately, the target gene expression is determined by combining the effects of multiple regulatory factors. As a large amount of genomic data has become available, it is possible to systematically study the genomic functions of various RFs and see how they interact with each other in order to regulate the target gene expression.

In the past decade, an increasing number of experimental and computational studies have focused on analyzing links between RFs, from various biological characteristics such as protein-protein interactions, sequence motifs in *cis*-regulatory modules, TF binding sites, co-associations of TFs in binding sites, and co-expressions of TF target genes [1, 5-8]. However, they focused solely on the identification of the wiring relationships (e.g. co-binding, co-association, and co-expression), leaving untouched the cooperative patterns among RFs that drive the biological functions behind the wiring diagrams. In this study, we use data derived from ChIP-Seq and RNA-Seq experiments to predict the cooperative patterns between RFs as they co-regulate the expression of target genes. On a genome-wide scale ChIP-Seq provides regulatory information about wiring between RFs and targets, while RNA-Seq provides gene expression data. By combining these two data types we are able to go beyond the regulatory activities of individual RFs and investigate the relationships between larger order RFs groups.

Cells achieve tremendous diversity in their gene expression programs, in large part due to cooperation among RFs, which may individually act as activators or repressors [9]. While the individual activity of many RFs remains to be characterized, their combined actions determine the expression pattern of their target gene. Here, we seek to systematically describe RF cooperation using logic models. At a high level, the gene regulatory network can be regarded as an electronic circuit. Therefore, we can build on the vast electronics knowledge base to draw useful insights for understanding and probing biological regulation. For example we can apply regulatory combinatorics (a key design principle in the field of electronics [10]) in the study of gene regulation using logic gate models. A logic gate is a discrete, high-level functional module that describes in an objective manner the relationship between a system's input and output elements. By applying logic functions to study the TF interactions in *E. coli* and *S. cerevisiae* in [11], the authors found that the logic gate is a simple but useful framework for understanding regulatory cooperativity among RFs. While this model is not able to capture very complex regulatory patterns, that

Daifeng Wang 8/17/14 10:51 PM
Deleted: (RFs)

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Hardison and Taylor, 2012; Neph, et al., 2012)

Daifeng Wang 8/17/14 10:51 PM
Deleted: , (Peter and Davidson, 2011).

Daifeng Wang 8/17/14 10:51 PM
Deleted: (TFs),

Daifeng Wang 8/17/14 10:51 PM
Deleted: microRNAs).

Daifeng Wang 8/17/14 10:51 PM
Deleted: datasets

Daifeng Wang 8/17/14 10:51 PM
Moved down [1]: In this study, we use data derived from ChIP-Seq and RNA-Seq experiments to predict the cooperative patterns between RFs as they co-regulate the expression of target genes. On a genome-wide scale ChIP-Seq provides regulatory information about wiring between RFs and targets, while RNA-Seq provides gene expression data.

Daifeng Wang 8/17/14 10:51 PM
Deleted: TF-TF or TF-miRNA wiring relationships

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Banerjee and Zhang, 2003; Gerstein, et al., 2012; Hardison and Taylor, 2012; Karczewski, et al., 2011; Poos, et al., 2013)

Daifeng Wang 8/17/14 10:51 PM
Deleted: focus

Daifeng Wang 8/17/14 10:51 PM
Deleted: between TFs

Daifeng Wang 8/17/14 10:51 PM
Moved (insertion) [1]

Daifeng Wang 8/17/14 10:51 PM
Deleted: Cells achieve tremendous diversity in their gene expression programs, in large part by leveraging on RFs capacity to act as repressors and activators. As such, at

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Rabaey, et al., 2003)) in the study of gene regulation using logic gate models. A logic gate is a discrete high-level functional module

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Mangan and Alon, 2003),

Daifeng Wang 8/17/14 10:51 PM
Deleted: gates are

Daifeng Wang 8/17/14 10:51 PM
Deleted: biological

may be characterized by continuous models [12, 13], it is computationally efficient, and comprehensive enough to be meaningful and to accurately describe a large variety of regulatory networks on a genome-wide scale in multiple organisms. Here, we present a computational method that streamlines the process of inferring logical cooperative relationships among RFs without requiring any prior information regarding their individual activity (as activators or repressors). We successfully apply our algorithm towards developing a comprehensive map of gene regulation.

In numerous cases, gene regulation can be regarded as a logical process, described by a logic gate model, where RFs are the input variables and the target gene expression is the output [3, 11, 14-21]. For example, DNA sequence motifs have been found to work together following standard combinatorial logic (AND, OR and NOT) to match gene expression patterns [22]. By contrast, TFs can indirectly control gene expression without binding to regulatory sequence elements but rather connecting with other bound TFs through protein-protein interactions [2, 23]. As such, in order to describe this process we need a more complex logic pattern. In this respect, we use general logic-circuit models to describe the logic operations for regulatory modules, consisting of multiple RFs and their common target genes.

The three basic logic operators, AND, OR, and NOT, can be combined in a variety of ways to describe all possible logical operations [11]. However, for simplicity it is useful to consider each operation independently. For any two-input-one-output scenario there are 16 possible logic gates (including all possible logic combinations between positive and negative regulators) (See Methods). These logic gates represent a useful and systematic framework for describing complex interactions between RFs and targets. Previous studies took advantage of binarized regulatory data (provided by perturbation experiments, such as TF knock-outs) and Boolean models in order to capture the logic processes that describe the TFs interactions [24]. The simple binary operations in the Boolean model are computationally efficient for large-scale datasets. However, previous efforts focused only on a small set of transcription factors and target genes, missing patterns from genome-wide identification and characterization of logic operations in gene regulation. In addition, numerous other important regulatory factors such as micro-RNAs (miRNAs) and TFs distally bound to target enhancer regions, have been overlooked in previous regulatory analyses.

By combining the activity of RFs and their respective targets on a genome-wide scale a bigger picture emerges, the gene regulatory network. To better understand this network we explore the interactions among its various components and features. Mathematically, it can be modeled as a directed network with a hierarchical structure comprising of top, middle, and bottom layers [5, 25, 26]. Previous studies have

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Garg, et al., 2009; Karlebach and Shamir, 2008)

Daifeng Wang 8/17/14 10:51 PM
Deleted: Therefore, here

Daifeng Wang 8/17/14 10:51 PM
Deleted: tool

Daifeng Wang 8/17/14 10:51 PM
Deleted: relationship

Daifeng Wang 8/17/14 10:51 PM
Deleted: biological

Daifeng Wang 8/17/14 10:51 PM
Deleted: .

Daifeng Wang 8/17/14 10:51 PM
Deleted: this method

Daifeng Wang 8/17/14 10:51 PM
Formatted: Font color: Text 1

Daifeng Wang 8/17/14 10:51 PM
Formatted: Font color: Text 1

Daifeng Wang 8/17/14 10:51 PM
Deleted: ,

Daifeng Wang 8/17/14 10:51 PM
Formatted: Font color: Text 1

Daifeng Wang 8/17/14 10:51 PM
Deleted: Actually, in

Daifeng Wang 8/17/14 10:51 PM
Formatted: Font color: Text 1

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Albert and Othmer, 2003; Das, et al., 2009; Fenno, et al., 2014; Mangan and Alon, 2003; Peter and Davidson, 2011; Peter, et al., 2012; Shmulevich and Dougherty, 2007; Siuti, et al., 2013; Tu, et al., 2013; Xie, et al., 2011)

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Beer and Tavazoie, 2004).

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Farnham, 2009; Neph, et al., 2012)

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Mangan and Alon, 2003).

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Somogyi and Sniegoski, 1996).

Daifeng Wang 8/17/14 10:51 PM
Deleted: the

Daifeng Wang 8/17/14 10:51 PM
Moved down [2]: To our knowledg... [2]

Daifeng Wang 8/17/14 10:51 PM
Deleted: A gene regulatory network ... [3]

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Bhardwaj, et al., 2010; Bha... [4]

shown that the middle levels RFs play important roles in gene regulation. Another feature of gene regulatory networks is the network motif. A common motif is the feed-forward loop (FFL), which consist of two RFs acting on a common target, while one RF regulates the other. FFLs can be classified into eight types based on the combination of the two RFs acting as activators and/or repressors. Previous studies in yeast [11] looked at a small set of FFLs, and have shown that they interact following logic operations. Thus, it is interesting to investigate how the logic operations associate with various regulatory network features.

In this paper, we present a novel computational method, Loregic, which integrates gene expression and regulatory data to characterize RFs on a genome-wide scale using logic-circuit models. Loregic classifies individual regulatory factors into functional modules (i.e., regulatory triplets) and reveals how modulat genes act functionally as logic circuits. We apply Loregic to study regulatory factors (TFs and miRNAs) in yeast cell cycle and human cancer datasets. We also illustrate our method's applicability to predict logic cooperation for two regulatory features: indirectly bound TFs and FFLs.

2 RESULTS

Loregic takes as inputs two types of data: a regulatory network (defined by RFs and their target genes) and a binarized gene expression dataset across multiple samples. The binarized gene expression data (1 – on and 0 – off) is simple but useful in representing the network RFs' activity on target genes. The inputs can be chosen from different resources to meet the user's needs. In this paper, we used BoolNet [27] to obtain binarized gene expressions. Loregic describes each regulatory module (triplet) using a particular type of logic gate; i.e. the gate that best matches the binarized expression data for that triplet across all samples. Loregic scores the agreement between the triplet's cross-sample expression and the idealized expression pattern of each 16 possible logic gates using Laplace's rule of succession (see Methods). A high score implies a strong co-operation between the activities of the two RFs on the target as described by the matched logic gate. If such a logic gate is found, we define a "logic-gate-consistent" or "gate-consistent" triplet as the triplet consistent with the respective logic gate. In the case that no best-matching logic gate is found (e.g. all logic gates score low, or there are tied scores between multiple logic gates), we define the triplet as inconsistent with all logic gates. This "negative" result suggests that the two-input-one-output model cannot appropriately describe the gene regulation, perhaps due to the fact that more RFs are involved and thus a more complex model should be required (see Discussion). In this

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Mangan and Alon, 2003)

Daifeng Wang 8/17/14 10:51 PM
Deleted: our method

Daifeng Wang 8/17/14 10:51 PM
Deleted: use Loregic results

Daifeng Wang 8/17/14 10:51 PM
Deleted: TF

Daifeng Wang 8/17/14 10:51 PM
Deleted: relationships by classifying logic operations for

Daifeng Wang 8/17/14 10:51 PM
Deleted: . Finally, we investigate the TF logic co-operations for human cancer

Daifeng Wang 8/17/14 10:51 PM
Deleted: In this paper, we used BoolNet (Mussel, et al., 2010)

Daifeng Wang 8/17/14 10:51 PM
Deleted: logic

Daifeng Wang 8/17/14 10:51 PM
Moved (insertion) [3]

Daifeng Wang 8/17/14 10:51 PM
Deleted: Next, Loregic test the statistical significance of the found logic gate by computing a gate consistency score.

Daifeng Wang 8/17/14 10:51 PM
Moved up [3]: A high score implies a strong co-operation between the activities of the two RFs on the target as described by the matched logic gate.

paper, we evaluate Loregic's capabilities to analyze transcription factors, miRNAs and their target genes. In detail, our method comprises of five steps (Figure 1):

Step 1: Input of gene regulatory network consisting of regulatory factors and their target genes;

Step 2: Identification all RF1-RF2-T triplets where RF1 and RF2 co-regulate the target gene T;

Step 3: Query of binarized gene expression data for any given triplet;

Step 4: Matching of the triplet's gene expressions against all possible two-input-one-output logic gates based on the binary values;

Step 5: Finding the matched logic gate if the triplet is gate-consistent, and calculating the consistency score;

Finally, **Steps 3-5** are repeated for all triplets in the regulatory network and all logic-gate-consistent triplets are identified.

The gate-consistent triplets can be further mapped onto other regulatory features (see Discussion). In this paper we describe two applications leveraging on the logic-gate-consistent triplet data: 1) prediction of logic operations for 1) indirectly bound TFs and 2) feed-forward loops.

2.1 Applications

We study Loregic's ability to characterize gene regulation in both small and complex biological systems. In particular we analyze two model datasets: yeast cell cycle and human cancer.

Yeast (*S. Cerevisiae*) is a small but well-studied biological system. The large variety of publicly available gene regulation and expression data makes it an ideal model organism to test and validate our algorithm. As an example, we use Loregic to predict logic cooperations among yeast TFs. We validate our results using data from genome-wide TF knockout experiments.

By contrast, human cancers are complex biological systems. Here we use Loregic to study the acute myeloid leukemia (AML), a quickly progressing cancer with low (five-year) survival rates (<25%), causing over 10,000 deaths in USA in 2014 [28]. We extracted gene regulatory network data from the ENCODE leukemia cell line, K562, and gene and miRNA expression datasets for AML from TCGA. Using Loregic, we predicted functional cooperation between and among TFs and miRNAs in AML.

2.1.1. Yeast TFs are cooperative during cell cycle

We used Loregic to characterize the TF-TF-target logics during the yeast cell cycle (see Methods) and found 4,126 TF-TF-target triplets that are gate-consistent (Fig. 3A). Among the gate-consistent triplets, we found that "T=RF1*RF2"(i.e., AND gate), "T=~RF1*RF2", and "T=RF1*~RF2" logic gates, have

Daifeng Wang 8/17/14 10:51 PM
Deleted: Repeating **Step**

Daifeng Wang 8/17/14 10:51 PM
Deleted: find

Daifeng Wang 8/17/14 10:51 PM
Formatted: Indent: First line: 0.1"

Daifeng Wang 8/17/14 10:51 PM
Deleted: **First,**

Daifeng Wang 8/17/14 10:51 PM
Formatted: Normal, Justified

Daifeng Wang 8/17/14 10:51 PM
Deleted: test our method by analyzing the

Daifeng Wang 8/17/14 10:51 PM
Deleted: -

Daifeng Wang 8/17/14 10:51 PM
Deleted: data. Next, we apply Loregic to

Daifeng Wang 8/17/14 10:51 PM
Deleted: data from acute myeloid leukemia (AML), and uncover the logic operations that govern the regulatory factors in this more complex system. ... [5]

Daifeng Wang 8/17/14 10:51 PM
Formatted: Font:Times New Roman, 12 pt

Daifeng Wang 8/17/14 10:51 PM
Formatted: Indent: First line: 0.1"

Daifeng Wang 8/17/14 10:51 PM
Deleted: We

Daifeng Wang 8/17/14 10:51 PM
Deleted: of

Daifeng Wang 8/17/14 10:51 PM
Formatted: Font:Times

Daifeng Wang 8/17/14 10:51 PM
Moved down [4]: <#>Yeast TFs are cooperative during cell cycle

Daifeng Wang 8/17/14 10:51 PM
Deleted: We identified 39,011 TF-TF-target triplets from 176 different TFs using TF ... [6]

Daifeng Wang 8/17/14 10:51 PM
Formatted: Font:Times New Roman

Daifeng Wang 8/17/14 10:51 PM
Moved (insertion) [5]

Daifeng Wang 8/17/14 10:51 PM
Moved (insertion) [4]

Daifeng Wang 8/17/14 10:51 PM
Deleted: yeast regulatory networks fr ... [7]

Daifeng Wang 8/17/14 10:51 PM
Deleted: across 59 time points

Daifeng Wang 8/17/14 10:51 PM
Deleted:). We

more triplets matched than all other gates, where '~' and '*' represent the NOT and AND logic operators respectively. Having randomly assigned TFs as RF1 and RF2, the "T=~RF1*RF2" and "T=RF1*~RF2" logic gates are symmetric. The AND gate triplets indicate that both TFs are required in order to activate the expression of their target gene (see discussion of other logic gates in Fig. S1). After matching all triplets against logic gates, we looked at variations in matched logic gates for a particular type of triplets (RF1, RF2, X), that share regulatory factors (RF1 and RF2) but have distinct targets (T=X) (Fig. 3B). As a result we were able to distinguish three categories for this triplet group: 1) "homogenous" gate-consistent triplets – matching the same logic gate across all targets (e.g., top table); 2) "inhomogeneous" gate-consistent triplets – matching different logic gates across all targets (e.g., middle table); and 3) non gate-consistent triplets, i.e. triplets inconsistent with all logic gates across all targets (e.g., bottom table).

2.1.2. Logic operations between TF-TF, miRNA-TF, and distTF-TF across targets in acute myeloid leukemia

We characterized TF-TF, miRNA-TF, and distTF-TF logic operations by integrating ENCODE and TCGA AML datasets using Loregic. Fig. 4 shows the distributions of gate-consistent TF-TF-target, miRNA-TF-target and distTF-TF-target triplets across all possible logic gates.

To test the relative importance of the TFs, miRNAs, and distTFs as regulators in the RF1-RF2-target triplet, we randomly assigned TFs as RF1 and RF2, and looked at the variations between symmetrical logic gate pairs (e.g. T=RF1+~RF2 vs T=~RF1+RF2 or T=RF1 vs T=RF2) in terms of matched triplets. We found no significant differences for the TF-TF-target triplet (Fig. 4A). However miRNA-TF-target and distTF-TF-target triplets told another story (Figs. 4B and 4C), suggesting that miRNAs and distTFs (as RF1) interact with TFs (as RF2) following different regulatory logics. For these scenarios, the "T=RF2" gate matches more triplets than any other gate, suggesting that in general promoter-binding TFs are the dominant regulators of target expression without being influenced by the presence of miRNAs or distTFs. Also, we found that for these cases, the gate-consistent TF-TF-target triplets preferentially match the 'OR' gate (2505 triplets).

2.2 Validations

We assessed the biological relevance of the insights gained by using logic circuit models to characterize gene regulation by comparing our results with experimental observations described in literature for yeast and human regulatory factors.

Daifeng Wang 8/17/14 10:51 PM
Deleted: i.e. homogeneous

Daifeng Wang 8/17/14 10:51 PM
Deleted: , i.e. inhomogeneous,

Daifeng Wang 8/17/14 10:51 PM
Moved down [6]: <#>Deleting TFs that form cooperative logic gates gives rise to significantly higher fold changes of target gene expression

Daifeng Wang 8/17/14 10:51 PM
Deleted: We use yeast genome-wide TF knockout experiments to validate the TF logic from gate-consistent triplets. The yeast TF knockout experiments give information regarding fold changes in gene expression as a result of deleting a single TF (Hu, et al., 2007; Reimand, et al., 2010).

Daifeng Wang 8/17/14 10:51 PM
Moved down [7]: Using these knockout datasets, we found that if a target gene is regulated by two cooperative TFs in an "AND" relationship, and thus it is most likely that the presence of both TFs is required to turn on the target gene (Fig. S1), the deletion of either TF impacts the target expression. For example, analyzing 871 AND-consistent triplets, we found that deleting either of their TFs gave rise to considerably down-regulated target genes, i.e., negative expression fold changes (*t-test p-value* =0.068). For the triplets consistent with non-cooperative gates such as "T=RF1" or "T=RF2" (i.e., only one TF controls the target regulation), we found that the target gene is more affected (down-regulated) by the ... [8]

Daifeng Wang 8/17/14 10:51 PM
Deleted:
We identified 50,865 TF-TF-target triplets triplets from ChIP-seq experiments for 70 TFs in the K562 cell line (Consortium, 201 ... [9]

Daifeng Wang 8/17/14 10:51 PM
Moved up [5]: We extracted gene regulatory network data from the ENCODE leukemia cell line, K562, and gene and miRNA expression datasets for AML from TCGA.

Daifeng Wang 8/17/14 10:51 PM
Deleted: Using Loregic, we predicted functional co-operations between and among TFs and miRNAs in AML.

Daifeng Wang 8/17/14 10:51 PM
Deleted: AML

Daifeng Wang 8/17/14 10:51 PM
Deleted: We identified 50,865 TF-TF-target triplets from ChIP-seq experiments for 70 TFs in the K562 cell line (Consortium, 2011; Djebali, et al., 2012; Gerstein, et al., 2 ... [10]

Daifeng Wang 8/17/14 10:51 PM
Formatted: Indent: First line: 0.1"

2.2.1. Deleting TFs that form cooperative logic gates gives rise to significantly higher fold changes of target gene expression

We used yeast genome-wide TF knockout experiments to validate the TF logic from gate-consistent triplets. The yeast TF knockout experiments give information regarding fold changes in gene expression as a result of deleting a single TF [35, 36]. Using these knockout datasets, we found that if a target gene is regulated by two cooperative TFs in an "AND" relationship, and thus it is most likely that the presence of both TFs is required to turn on the target gene (Fig. S1), the deletion of either TF impacts the target expression. For example, analyzing 871 AND-consistent triplets, we found that deleting either of their TFs gave rise to considerably down-regulated target genes, i.e., negative expression fold changes (*t-test p-value* =0.068). For the triplets consistent with non-cooperative gates such as "T=RF1" or "T=RF2" (i.e., only one TF controls the target regulation), we found that the target gene is more affected (down-regulated) by the removal of the dominant RF (i.e., RF1 for "T=RF1" consistent triplets, RF2 for "T=RF2" consistent triplets) than the removal of the other one (*t-test p-value* < 0.0004 for 811 triplets consistent with "T=RF1" or "T=RF2").

2.2.2. AML-related TFs play a dominant role in regulating target gene expression

The cancer-related TFs play key roles in gene regulation. For example, the transcription factor MYC has been found to universally amplify target gene expressions in lymphocytes [37], implying that it does not require cooperation from other TFs in order to perform its regulatory function. We identified 2,153 MYC-TF-target triplets (i.e., RF1 is MYC, RF2 is chosen from other TFs from ENCODE, and T is target), and found that 905 of them are gate-consistent. The two most enriched logic gates are "T=RF1" (133 triplets, hypergeometric test < $4.3*10^{-27}$) and "T=RF1+RF2 (OR)" (211 triplets, hypergeometric test < $1.1*10^{-21}$) (Fig. 5A). "T=RF1" with RF1 being MYC suggests that in general the presence of a highly expressed MYC is necessary and sufficient for high target gene expression. "T=RF1+RF2" with RF1 being MYC and RF2 being other TFs suggests that the presence of either MYC or TF is sufficient for regulating target expression. However, both scenarios indicate that MYC turns on target expression without requiring the presence additional TFs. These results support the recent findings that MYC plays a universal amplifier role in gene expression.

Next we analyzed all the triplets associated with AML-related TFs, where RF1 is chosen from AML-related TFs, RF2 is chosen from non-AML related TFs, and T is their common target. The AML-related TFs were also identified as AML cancer genes [38]. We found that "T=RF1" and "T=~RF1" (Fig. 5B) are

Daifeng Wang 8/17/14 10:51 PM
Moved (insertion) [6]

Daifeng Wang 8/17/14 10:51 PM
Moved (insertion) [7]

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Nie, et al., 2012),

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Forbes, et al., 2011).

the most enriched matched logic gates for these TFs. However, we did not find any enrichment for these two gates in triplets containing only non-AML TFs. Therefore, this result suggests that the AML-related TFs play a dominant role in regulating target expression.

2.3 Loregic applications for other regulatory features

2.3.1. Classification of logic-gate-consistent triplets with indirectly bound TFs

TFs can regulate target genes without binding directly to target regulatory regions, but forming protein-protein interactions with already bound TFs [2]. We suggest that evaluating the logic cooperation of TF pairs along with the analysis of promoter motifs, can give insights regarding the TF binding activity. We studied TF promoter motifs in target promoter regions (1,000 bps in yeast and 5,000 bps in human upstream of the transcription start site) [39-42]. We identified numerous TFs with no motifs (<80% PWM similarity) in target promoter regions, even though the logic gate assessment predicted that cooperation between the two RFs is required in order to control the target gene expression. Out of 948 yeast TF-TF-target triplets consistent with "T=RF1*RF2" (AND gate) (see examples in Fig. 6), 348 have one TF whose motif is not present in the target's promoter region (the same happens for 364 out of 1,100 for "T=RF1*~RF2" and 377 out of 1,095 for "T=~RF1*RF2", a symmetric logic gates pair). Similarly, in the human leukemia dataset, we found that from 888 TF-TF-target triplets consistent with "AND" gates, 71 have one TF whose motif is not present in the target's promoter. For example (Fig. S2), the triplet of RF1=USF2, RF2=NFYB, T=YPEL1 is consistent with "AND" gate, and both TFs have motifs in the YPEL1 promoter region. By contrast, the triplet of RF1=USF2, RF2=NFE2, T=NBPF1, does not have an NFE2 motif in NBPF1's promoter region, even though it is also consistent with "AND" gate. However, USF2 and NFE2 are connected through protein-protein interactions, and consequently NFE2 is regulating NBPF1 through indirect binding [2]. As such, we suspect that those TFs with absent motifs (as above) can potentially regulate by cooperating with directly bound TFs through protein-protein interactions, a phenomenon that has been previously observed [2, 23, 43-45]. Moreover, we further classified those triplets with indirectly bound TFs using their matched logic gates, and identified the indirectly bound TFs cooperating with bound TFs to regulate their targets in a logical way.

2.3.2. Logic gates for feed-forward loops

Feed-forward loops (FFLs) are RF1-RF2-T triplets in which RF1 also regulates RF2. FFLs have been found to be important motifs in regulatory networks, with many interacting by following logic operations [11]. We apply Loregic to find the logic operations that characterize the FFLs from a genome-wide

Daifeng Wang 8/17/14 10:51 PM
Deleted: <#>Applications

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Neph, et al., 2012).

Daifeng Wang 8/17/14 10:51 PM
Deleted: co-operation

Daifeng Wang 8/17/14 10:51 PM
Deleted: (DebRoy, 2013; Lawrence, 2014; Li, 2014; Pages, 2014).

Daifeng Wang 8/17/14 10:51 PM
Deleted: for the

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Neph, et al., 2012).

Daifeng Wang 8/17/14 10:51 PM
Deleted: in

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Biddie, et al., 2011; Farnham, 2009; Gordan, et al., 2009; Neph, et al., 2012; Zhao, et al., 2012)

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Mangan and Alon, 2003).

perspective in both yeast cell cycle and human cancer. For the yeast regulatory network, we found that from a total 5707 FFLs, 659 are gate-consistent triplets. Out of these, 162 FFLs are consistent with the 'AND' gate (hypergeometric test $<1.3*10^{-3}$), and 159 are consistent with 'T=RF1' (hypergeometric test $<7.5*10^{-5}$) making them the dominant logic gates for yeast FFL. These results match previous experiments that have shown that the majority of FFLs are of the so-called coherent type 1, in which RF1 activates RF2, and both activate the target [11].

Next, we looked at FFLs from human leukemia TF-TF-T triplets (23385 FFLs in total), and found that the two most abundant matched logic gates are 'T=RF1" (1,306 FFLs, hypergeometric test $<3.4*10^{-9}$) and "T=RF1+~RF2" (1,765 FFLs, hypergeometric test $<1.7*10^{-5}$). Both gates match the logics of the coherent type 4 FFL, where RF1 down-regulates RF2, RF2 down-regulates target, and RF1 activates target as described in [11]. This suggests that the master TF (RF1) of the FFL aims to activate the target, but due to the gene down-regulation action from the secondary TF (RF2), it must simultaneously down-regulate RF2s. Moreover, we did not find any enriched logic gates among the triplets that do not form FFLs in both yeast and human.

2.3.3. miRNAs and MYC down-regulate each other

MYC and miRNAs have been found to down-regulate each other by forming double down-regulatory FFLs in leukemia [46]. We identified 1,805 miRNA-MYC-target triplets with 117 miRNAs, 1,143 of which are gate-consistent. From these triplets, 446 match "T=RF2" when RF2 is MYC (hypergeometric test $< 2.5*10^{-124}$), and 201 match "T=~RF1+RF2" when RF1 is a miRNA and RF2 is MYC (hypergeometric test $< 4.1*10^{-25}$). These two dominant logic gates also match the logic for the coherent type 4 FFL as described in [11]. As expected, these results imply that miRNAs repress target gene expressions, while MYC activates it and simultaneously down-regulates the miRNAs. We also found 56 gate-consistent miRNA-MYC-target triplets matching "T=~RF1*RF2" when RF1 is a miRNA and RF2 is MYC, and 16 triplets matching "T=~RF1" with RF1 being a miRNA. These two logics match the coherent type 2 FFL[11]. This result suggests that miRNAs repress both the expression of both MYC and the target gene, while MYC activates the target. In short, these matched logic gates support the notion that the miRNAs and MYC form a double-negative regulatory loop in this system.

3 DISCUSSION

Loregic is a multi-purpose computational method that uses logic-circuit models to characterize the cooperativity among regulatory factors such as TFs and miRNAs by integrating gene expression and regu-

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Mangan and Alon, 2003).

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Mangan and Alon, 2003).

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Tao, et al., 2014).

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Mangan and Alon, 2003).

Daifeng Wang 8/17/14 10:51 PM
Deleted: .

latory network data. Given the multitude of appropriate high quality expression (e.g., RNA-seq, small RNA-seq), and regulation (e.g., ChIP-seq, CLIP-seq, DNase-seq) datasets available, Loregic can be further used to study cooperations among other regulatory elements such as splicing factors, long non-coding RNAs and so on. To our knowledge, the present study describes for the first time the use of 16 logic operations to perform a comprehensive genome-wide analysis of regulatory triplets (including miRNAs, proximally and distally bound TFs).

In our analysis, we found triplets inconsistent with all the logic gates. There are several potential explanations for such cases. First, the cooperative patterns of two RFs might follow a more complex mechanism, perhaps one that depends on timing or the phosphorylation state of the RF, which our model does not take into account. Second, the target gene might be regulated by more than two RFs, and thus a higher-order logic circuit model with multiple inputs (>2) as discussed above might be required to capture the RF-target logic. Finally, the target gene expression may also be impacted by stochastic signals, which may necessitate more advanced models [12].

We tested Loregic using two-RFs-one-target triplets, focusing on scenarios where the RFs are either two TFs or one TF and one miRNA. However, we can extend Loregic, to analyze regulatory modules with multiple RFs and multiple target genes using higher-order logic circuit model discussed as above if there is enough supporting data. Loregic is also compatible with other discretization methods including any custom-made binarized gene expression data as input.

One of Loregic functionalities is relating triplet logics to any set of regulatory network features. Here, we map the logic-gate-consistent triplets to two regulatory features: promoter sequence motifs and feed-forward loops. Loregic's results can also be directly applied to differentially assess the abundance of various types of logic gates among other gene regulatory features. For example, a potential future application is finding enrichments of logic-gate-consistent triplets in hierarchical layers, and identifying logic cooperations between and among RFs at different hierarchical layers in the network; e.g., top, middle and bottom layers, which may potentially help understand cooperations among even larger order regulatory groups [5, 25, 26].

In summary, Loregic systematically characterizes genetic regulatory cooperativity using logic-circuit models. This algorithm is widely applicable for the study of regulatory mechanisms and to the assembly of the gene regulatory panoramagram.

Daifeng Wang 8/17/14 10:51 PM
Moved (insertion) [2]

Daifeng Wang 8/17/14 10:51 PM
Deleted: .

Daifeng Wang 8/17/14 10:51 PM
Deleted: also

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Garg, et al., 2009).

Daifeng Wang 8/17/14 10:51 PM
Deleted: long as

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Bhardwaj, et al., 2010; Bhardwaj, et al., 2010; Gerstein, et al., 2012)

Daifeng Wang 8/17/14 10:51 PM
Deleted: characterize

4 MATERIALS AND METHODS

4.1 Gene expression, transcription factor and miRNA datasets

We analyzed the gene expression in yeast using three well-studied cell-cycle datasets: 1) alpha-factor time course with 18 time points (0, 7', … , 119'); 2) cdc15 time course with 24 time points (10', 30', … , 290') and 3) cdc28 time course with 17 time points (0, 10', … , 160') [47, 48]. We combined all three datasets (5,581 genes and 59 time points) and normalized gene expressions for each time point by centering the mean to zero (i.e., standardization). For gene regulation in yeast, we used 176 transcription factors with their target genes identified in [29, 30], and found 39,011 TF-TF-target triplets.

In the study of gene expression in human leukemia, we obtained RNA-seq RPKM expressions from The Cancer Genome Atlas Data Portal [49] for 19,798 protein-coding genes and 705 miRNAs across 197 and 188 AML samples, respectively. For each sample, we standardized the log(RPKM+1) across all genes. We identified 50,865 TF1-TF2-target triplets using ChIP-seq data (70 TFs) from ENCODE K562 cell line [5, 31, 32] and 821 distTF-TF-target triplets, where distTFs were predicted to bind distal regulatory regions in [33]. Thus by integrating miRNA- and TF-target pairs in K562, we were able to identify and 56,944 miRNA-TF-target triplets, in which Rf1 is a microRNA, RF2 is a TF, and target is a gene co-regulated by that miRNA and that TF, using the confident miRNA-targets for human K562 cell line as described in [34]. For the differential logic gate enrichment analysis of promoter-bound transcription factors versus enhancer-bound transcription factors (distTFs), we obtained the distTFs data from [33] and identified 821 distTF-TF-target triplets.

4.2 Converting gene expression changes over conditions to Boolean values

In this paper, we binarized the gene expression levels to Boolean values 1 and 0 to represent high or low gene expression, respectively, using BoolNet. Loregic is also compatible with user-input, customized binary gene expression data [27]. BoolNet assigns Boolean values to expression data on the basis of modular co-expression patterns by *k*-means clustering across inputted samples such as time points (yeast) or AML samples (human), and therefore accounts for differences in the dynamic ranges of expression among genes in the input data. In our yeast study, samples input to BoolNet were drawn from different time points whereas in our AML study, input samples were taken from different patients. After conversion, there are in total 79% zeros and 21% ones in yeast binarized expression data (42% zeros, 58% ones in human).

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Cho, et al., 1998; Spellman, et al., 1998)

Daifeng Wang 8/17/14 10:51 PM
Deleted: the

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Harbison, et al., 2004; Jothi, et al., 2009)

Daifeng Wang 8/17/14 10:51 PM
Deleted: (https://tcga-data.nci.nih.gov/tcga/)

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Consortium, 2011; Djebali, et al., 2012; Gerstein, et al., 2012)

Daifeng Wang 8/17/14 10:51 PM
Deleted: ,

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Chen, et al., 2014).

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Yip, et al., 2012)

Daifeng Wang 8/17/14 10:51 PM
Deleted: (Mussel, et al., 2010).

4.3 Mapping and scoring a RF1-RF2-T triplet to 16 logic gates

Mathematically, a logic gate can be described by the truth table that lists the outputs of the logic gate for each allowed combination of inputs. For a two-input-one-output logic gate, each of the two input variables we have two possible values 1 or 0, thus the truth table will contain 4 binary two-element vectors representing all the possible combinations of the two input variables i.e., $v_1=(0,0), v_2=(0,1), v_3=(1,0), v_4=(1,1)$, where v_i is the vector representing i^{th} input combination i=1,2,3,4. Given the fact that there are 4 possible combinations of input variables, the truth table output will be a four elements vector, with each element having two possible values 0 or 1. Thus there are 2^4 possible combinations of 0 and 1 for the output vector, in other words for any two-input-one-output equations there are 16 possible truth tables. The 16 different truth tables correspond to 16 logic gates as shown in Fig. S1. The three basic logic operations, AND ("*"), NOT ("~") and OR ("+") are used to express all the 16 possible logic gates. However, for simplicity we are going to consider each logic gate separately.

We denote $f^g(v_i)$ the function to obtain the output value from the i^{th} input vector v_i in the logic gate g, with i=1,2,3,4. For example for the AND logic gate we have:

$$f^{\mathrm{AND}}(v_1) = f^{\mathrm{AND}}(0,0) = 0*0 = 0$$
$$f^{\mathrm{AND}}(v_2) = f^{\mathrm{AND}}(0,1) = 0*1 = 0$$
$$f^{\mathrm{AND}}(v_3) = f^{\mathrm{AND}}(1,0) = 1*0 = 0$$
$$f^{\mathrm{AND}}(v_4) = f^{\mathrm{AND}}(1,1) = 1*1 = 1$$

In our model, the two RFs (RF1, RF2) in a regulatory triplet, serve as inputs, while the common target gene T is the output (the result of the f^g acting on the (RF1, RF2) binary vector).

For m samples, we denote $\vec{x}$, $\vec{y}$ and $\vec{z}$ as the m-dimension binary vectors containing m binarized expression values for RF1, RF2, and T respectively. Loregic identifies the logic gate whose truth table best matches the input/output data as follows. For the input vector v_i, we denote as $m_i = \sum_{j=1}^{m} \mathrm{I}(x(j)=v_i(1))\mathrm{I}(y(j)=v_i(2))$ the number of samples ($x(j)$, $y(j)$) matching v_i, where I(.) is indicator function, $x(j)$ and $y(j)$ are j^{th} elements of $\vec{x}$ and $\vec{y}$, with j=1,2,…,m, and i=1,2,3,4. Thus we have $m=m_1+m_2+m_3+m_4$. Second, given a logic gate g, we denote

Daifeng Wang 8/17/14 10:51 PM
Deleted: $v_1=(0,0), v_2=(0,1), v_3=(1,$

Daifeng Wang 8/17/14 10:51 PM
Deleted: $f^{AND}(v_1)$

Daifeng Wang 8/17/14 10:51 PM
Deleted: f^{AND}

Daifeng Wang 8/17/14 10:51 PM
Deleted: f^{AND}

Daifeng Wang 8/17/14 10:51 PM
Deleted: f^{AND}

Daifeng Wang 8/17/14 10:51 PM
Deleted: f^{AND}

Daifeng Wang 8/17/14 10:51 PM
Deleted: f^{AND}

Daifeng Wang 8/17/14 10:51 PM
Deleted: f^{AND}

Daifeng Wang 8/17/14 10:51 PM
Deleted: $\vec{x}$, $\vec{y}$

Daifeng Wang 8/17/14 10:51 PM
Deleted: $\vec{z}$

Daifeng Wang 8/17/14 10:51 PM
Deleted: $m_i = \sum_{j=1}^{m} \mathrm{I}(x(j)=v_i(1))\mathrm{I}(y(j)$

Daifeng Wang 8/17/14 10:51 PM
Deleted: $\vec{x}$

Daifeng Wang 8/17/14 10:51 PM
Deleted: $\vec{y}$

$$n_i = \sum_{j=1}^{m} (1 - |z(j) - f^g(x(j), y(j))|) \mathrm{I}(x(j) = v_i(1)) \mathrm{I}(y(j) = v_i(2))$$

as the number of $z(j)$ target binary samples matching the logic gate g output $f^g(v_i)$, for the v_i input vector. Next we calculate $s_i^g = (1+n_i)/(2+m_i)$ as the succession probability matching the v_i of g by Laplace's rule of succession [50]. This is an effective way to simply but rigorously penalizes logic-gate assignments that were distinguished from alternative logic gates on the basis of only a small number of observations. As such given the binarized expression data, $\bar{x}, \bar{y}$, and $\bar{z}$ for the (RF1, RF2, T) triplet, the consistency score for the logic gate g, $C^g(\bar{x}, \bar{y}, \bar{z})$ is given by the product of the succession probabilities for four input types, $s_1^g, s_2^g, s_3^g, s_4^g$ as follows:

$$C^g(\bar{x}, \bar{y}, \bar{z}) = \prod_{i=1}^{4} s_i^g(\bar{x}, \bar{y}, \bar{z})$$

$$\text{where } s_i^g(\bar{x}, \bar{y}, \bar{z}) = \frac{1 + \sum_{j=1}^{m} (1 - |z(j) - f^g(x(j), y(j))|) \mathrm{I}(x(j) = v_i(1)) \mathrm{I}(y(j) = v_i(2))}{2 + \sum_{j=1}^{m} \mathrm{I}(x(j) = v_i(1)) \mathrm{I}(y(j) = v_i(2))}.$$

Finally, we choose the logic gate with the highest consistency score as the best matched logic gate for the analyzed triplet. Note that according to Laplace's rule of succession, if there is no data available for a triplet, then $m=n_i=m_i=0$, and for each logic gate the consistency score by the succession rule is 1/2*1/2*1/2*1/2=1/16, which is the probability of a random guess from 16 logic gates. In order to identify potentially spurious logic gate assignments (i.e. not due to chance) for any gate-consistent triplets (RF1, RF2, T), we calculate a permutation score for each triplet over the 16 logic gates as follows: We suppose that the triplet matches the k^{th} logic gate, g_k. We replace the target gene, T by a randomly selected gene N times (here we use N=1000), and define its permutation score, as $p(g_k)$=(the number of replacement triplets that can be identified as gate-consistent with matched g_k)/N. A high permutation score implies that random effects may cause the matched logic gate. In this paper, we only keep the gate-consistent triplets with permutation scores less than 0.1.

Test Case [[maybe moved to caption]]: In Fig. 2, we exemplify the calculation of consistency score for the (TF1, TF2, T) triplet where RF1 is TF1, RF2 is TF2, and T is their common target gene, across a dataset of 20 samples. Thus after the conversion there are m=20 binary vectors. There are 5 vectors with RF1=0 and RF2=0, all of which have output of T=0 (red). Thus, when RF1=0 and RF2=0, the output of

Daifeng Wang 8/17/14 10:51 PM
Deleted: $n_i = \sum_{j=1}^{m} (1 - |z(j) - f^g(x(j), y(j))|) \mathrm{I}(x($

Daifeng Wang 8/17/14 10:51 PM
Deleted: $s_i^g = (1+n_i)/(2+m_i)$ as the succession probability matching the v_i of g by Laplace's rule of succession (Feller, 1968).

Daifeng Wang 8/17/14 10:51 PM
Deleted: $\bar{x}, \bar{y}$, and $\bar{z}$

Daifeng Wang 8/17/14 10:51 PM
Deleted: $C^g(\bar{x}, \bar{y}, \bar{z})$

Daifeng Wang 8/17/14 10:51 PM
Deleted: $s_1^g, s_2^g, s_3^g, s_4^g$

Daifeng Wang 8/17/14 10:51 PM
$C^g(\bar{x}, \bar{y}, \bar{z}) = \prod_{i=1}^{4} s_i^g(\bar{x}, \bar{y}, \bar{z})$

$1 + \sum_{j=1}^{m} (1 -$

where $s_i^g(\bar{x}, \bar{y}, \bar{z}) =$

Deleted:

Daifeng Wang 8/17/14 10:51 PM
Deleted: suppose a

Daifeng Wang 8/17/14 10:51 PM
Deleted: with

Daifeng Wang 8/17/14 10:51 PM
Deleted: being one TF,

Daifeng Wang 8/17/14 10:51 PM
Deleted: being another TF,

Daifeng Wang 8/17/14 10:51 PM
Deleted: being

Daifeng Wang 8/17/14 10:51 PM
Deleted: has

Daifeng Wang 8/17/14 10:51 PM
Deleted: after conversion

this triplet is more likely to be 0 (T=0), so (RF1=0, RF2=0, T=0) is chosen as the most suitable triplet-logic gate match, and its succession probability s_1=(5+1)/(5+2)=6/7 with n_1=5 and m_1=5. Next, there are 5 vectors with RF1=0 and RF2=1, four of which have output of T=0 (green), and one of which has output of T=1. We choose (RF1=0, RF2=1, T=0) as the most common triplet with its succession probability s_2=(4+1)/(5+2)=5/7 with n_2=4 and m_2=5, because for the given input the majority of cases have zero as the output value. Similarly, when RF1=1 and RF2=0, T=0 is chosen (magenta) because it appears more than T=1, and its succession probability s_3=(5+1)/(5+2)=6/7 with n_3=5 and m_3=5. Finally, when RF1=1 and RF2=1, T=1 is chosen (orange) because it appears four times but T=0 appears only once, and its succession probability s_4=(4+1)/(5+2)=5/7 with n_4=5 and m_4=5. Combining the outputs chosen for four different input combinations of RF1 and RF2, we obtain the triplet's truth table, and find that it best matches the AND logic gate. As such we define the this triplet is consistent with AND gate, and calculate its consistency score; i.e., $C(\text{AND})=s_1 * s_2 * s_3 * s_4 = 0.37$.

ACKNOWLEDGEMENTS

Funding: National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

1. Hardison, R.C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nature reviews. Genetics* **13**, 469-483 (2012).
2. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).
3. Peter, I.S. & Davidson, E.H. Evolution of gene regulatory networks controlling body plan development. *Cell* **144**, 970-985 (2011).
4. Cheng, C. et al. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS computational biology* **7**, e1002190 (2011).
5. Gerstein, M.B. et al. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
6. Banerjee, N. & Zhang, M.Q. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic acids research* **31**, 7024-7031 (2003).
7. Karczewski, K.J. et al. Cooperative transcription factor associations discovered using regulatory variation. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 13353-13358 (2011).
8. Poos, K. et al. How microRNA and transcription factor co-regulatory networks affect osteosarcoma cell proliferation. *PLoS computational biology* **9**, e1003210 (2013).
9. Whitington, T., Jolma, A. & Taipale, J. Beyond the balance of activator and repressor. *Science signaling* **4**, pe29 (2011).
10. Rabaey, J.M., Chandrakasan, A.P. & Nikoli*c, B. Digital integrated circuits : a design perspective, Edn. 2nd. (Pearson Education, Upper Saddle River, N.J.; 2003).
11. Mangan, S. & Alon, U. Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11980-11985 (2003).
12. Garg, A., Mohanram, K., Di Cara, A., De Micheli, G. & Xenarios, I. Modeling stochasticity and robustness in gene regulatory networks. *Bioinformatics* **25**, i101-109 (2009).

13. Karlebach, G. & Shamir, R. Modelling and analysis of gene regulatory networks. *Nature reviews. Molecular cell biology* **9**, 770-780 (2008).
14. Albert, R. & Othmer, H.G. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. *Journal of theoretical biology* **223**, 1-18 (2003).
15. Shmulevich, I. & Dougherty, E.R. Genomic Signal Processing. (Princeton University Press, Princeton; 2007).
16. Das, D., Pellegrini, M. & Gray, J.W. A primer on regression methods for decoding cis-regulatory logic. *PLoS computational biology* **5**, e1000269 (2009).
17. Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R. & Benenson, Y. Multi-input RNAi-based logic circuit for identification of specific cancer cells. *Science* **333**, 1307-1311 (2011).
18. Peter, I.S., Faure, E. & Davidson, E.H. Predictive computation of genomic logic processing functions in embryonic development. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 16434-16442 (2012).
19. Tu, S., Pederson, T. & Weng, Z. Networking development by Boolean logic. *Nucleus* **4**, 89-91 (2013).
20. Siuti, P., Yazbek, J. & Lu, T.K. Synthetic circuits integrating logic and memory in living cells. *Nature biotechnology* **31**, 448-452 (2013).
21. Fenno, L.E. et al. Targeting cells with single vectors using multiple-feature Boolean logic. *Nature methods* **11**, 763-772 (2014).
22. Beer, M.A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185-198 (2004).
23. Farnham, P.J. Insights from genomic profiling of transcription factors. *Nat Rev Genet* **10**, 605-616 (2009).
24. Somogyi, R. & Sniegoski, C.A. Modeling the complexity of genetic networks: Understanding multigenic and pleiotropic regulation. *Complexity* **1**, 45-63 (1996).
25. Bhardwaj, N., Kim, P.M. & Gerstein, M.B. Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. *Science signaling* **3**, ra79 (2010).
26. Bhardwaj, N., Yan, K.K. & Gerstein, M.B. Analysis of diverse regulatory networks in a hierarchical context shows consistent tendencies for collaboration in the middle levels. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 6841-6846 (2010).
27. Mussel, C., Hopfensitz, M. & Kestler, H.A. BoolNet--an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* **26**, 1378-1380 (2010).
28. http://www.cancer.gov/ (
29. Jothi, R. et al. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. *Molecular systems biology* **5**, 294 (2009).
30. Harbison, C.T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).
31. Consortium, E.P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology* **9**, e1001046 (2011).
32. Djebali, S. et al. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).
33. Yip, K.Y. et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome biology* **13**, R48 (2012).
34. Chen, D. et al. Dissecting the chromatin interactome of microRNA genes. *Nucleic Acids Res* **42**, 3028-3043 (2014).
35. Hu, Z., Killion, P.J. & Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. *Nature genetics* **39**, 683-687 (2007).
36. Reimand, J., Vaquerizas, J.M., Todd, A.E., Vilo, J. & Luscombe, N.M. Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. *Nucleic acids research* **38**, 4768-4777 (2010).
37. Nie, Z. et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**, 68-79 (2012).
38. Forbes, S.A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research* **39**, D945-950 (2011).
39. DebRoy, H.P.a.P.A.a.R.G.a.S. Biostrings: String objects representing biological sequences, and matching algorithms. *R package version 2.28.0* (2013).
40. Pages, H. BSgenome: Infrastructure for Biostrings-based genome data packages. *R package version 1.28.0* (2014).
41. Lawrence, M.C.a.H.P.a.P.A.a.S.F.a.M.M.a.D.S.a.M. GenomicFeatures: Tools for making and manipulating transcript centric annotations. *R package version 1.12.4* (2014).
42. Li, H.P.a.M.C.a.S.F.a.N. AnnotationDbi: Annotation Database Interface. *R package version 1.22.6* (2014).
43. Biddie, S.C. et al. Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell* **43**, 145-155 (2011).
44. Zhao, Y., Ruan, S., Pandey, M. & Stormo, G.D. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics* **191**, 781-790 (2012).
45. Gordan, R., Hartemink, A.J. & Bulyk, M.L. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res* **19**, 2090-2100 (2009).

Deleted: and… Othmer, H.G. ... [12]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [11]
Daifeng Wang 8/17/14 10:51 PM
Deleted: Banerjee, N. and Zhang, M ... [13]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [14]
Daifeng Wang 8/17/14 10:51 PM
Deleted: and… Tavazoie, S. (200 ... [15]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [16]
Daifeng Wang 8/17/14 10:51 PM
Deleted: and… Gerstein, M.B. ... [17]
Daifeng Wang 8/17/14 10:51 PM
Deleted: and… Gerstein, M.B. ... [18]
Daifeng Wang 8/17/14 10:51 PM
Deleted: Biddie, S.
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [19]
Daifeng Wang 8/17/14 10:51 PM
Deleted: *et al.* (2011) Transcription ... [20]
Daifeng Wang 8/17/14 10:51 PM
Deleted: Gerstein, M.B., *et al.* (2012 ... [21]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [22]
Daifeng Wang 8/17/14 10:51 PM
Deleted: .,
Daifeng Wang 8/17/14 10:51 PM
Deleted: (2009)…Genomic analy ... [24]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [23]
Daifeng Wang 8/17/14 10:51 PM
Deleted: Karczewski, K.J.,
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [25]
Daifeng Wang 8/17/14 10:51 PM
Deleted: (2011) Cooperative transcr ... [26]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [27]
Daifeng Wang 8/17/14 10:51 PM
Deleted: *the National Academy of Sc* ... [28]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [29]
Daifeng Wang 8/17/14 10:51 PM
Deleted: *reviews. Molecular cell b* ... [30]
Daifeng Wang 8/17/14 10:51 PM
Moved (insertion) [8] ... [31]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [32]
Daifeng Wang 8/17/14 10:51 PM
Deleted: Lawrence, ... [33]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [34]
Daifeng Wang 8/17/14 10:51 PM
Deleted: and analysis …f Boolean n ... [35]
Daifeng Wang 8/17/14 10:51 PM
Moved up [8]: et al.
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [36]
Daifeng Wang 8/17/14 10:51 PM
... [37]
Daifeng Wang 8/17/14 10:51 PM
... [38]
Daifeng Wang 8/17/14 10:51 PM
Formatted ... [39]
Daifeng Wang 8/17/14 10:51 PM
... [40]
Daifeng Wang 8/17/14 10:51 PM
... [41]

46. Tao, J., Zhao, X. & Tao, J. c-MYC-miRNA circuitry: a central regulator of aggressive B-cell malignancies. *Cell Cycle* **13**, 191-198 (2014).
47. Cho, R.J. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular cell* **2**, 65-73 (1998).
48. Spellman, P.T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-3297 (1998).
49. https://tcga-data.nci.nih.gov/tcga/ (
50. Feller, W. An introduction to probability theory and its applications, Edn. 3rd. (Wiley, New York; 1968).

Daifeng Wang 8/17/14 10:51 PM
Deleted: .,

Daifeng Wang 8/17/14 10:51 PM
Formatted: Indent: Left: 0", Hanging: 0.5"

Daifeng Wang 8/17/14 10:51 PM
Deleted: (1998)

Daifeng Wang 8/17/14 10:51 PM
Formatted: Font:Not Italic

Daifeng Wang 8/17/14 10:51 PM
Deleted: ,

Daifeng Wang 8/17/14 10:51 PM
Deleted: ,

Daifeng Wang 8/17/14 10:51 PM
Deleted: .

Daifeng Wang 8/17/14 10:51 PM
Deleted: Tao, J., Zhao, X. and Tao, J. (2014) c-MYC-miRNA circuitry: a central regulator of aggressive B-cell malignancies, *Cell Cycle*, **13**, 191-198. ... [42]