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

FACTORS AND IN TRIPLETS OF

ABSTRACT

Regulatory factors act cooperatively to control gene expression. We present Loregic, a novel computational method, to identify and characterize regulatory cooperativity using logic-circuit models from gene expression and regulatory networks. We study the logic operations of functional regulatory modules consisting of multiple RFs with common target genes, and specifically focus on two RFs co-regulating a target gene ~~as a triplet~~ using a two-input-one-output logic gate model. Next, using binarized gene expression data, we score how well each triplet matches each of all 16 (2^4) possible logic gates. A high score implies a strong operation between the two regulatory factors to target following the corresponding logic gate. We first apply Loregic to the yeast transcription factors (TFs), and validate their logic co-operations using TF deletion data. We then integrate the ChIP-seq data from ENCODE and the RNA-seq expression data from TCGA to look at logic co-operations among TFs, TFs binding to enhancers and miRNAs in cancer cell lines. In addition, we use Loregic to discover indirect binding between TFs, when their sequence motifs are absent from the target gene promoter region. Finally, we predict the TF logics of the network motif, feed-forward loops in which two TFs have regulatory relationships. In summary, Loregic provides a valuable computational tool (https://github.com/gersteinlab/Loregic) to reveal logic operations in gene regulation, and can be extended to analyze cooperativity amongst other regulatory elements such as long non-coding RNAs.

Contact: pi@gersteinlab.org

1 INTRODUCTION

Gene expression is a complex process that is controlled by regulatory factors (RFs) on multiple dimensions. For example, from a spatial perspective, multiple transcription factors bind to the promoter region of their target gene (Hardison and Taylor, 2012; Neph, et al., 2012), while the regulatory network controls gene expression during embryo development in a temporal dimension (Peter and Davidson, 2011). Due to the process complexity, the majority of regulatory factors work cooperatively, rather than independently, to determine the correct gene expression outcome in various cell types. For example, at transcriptional level, the gene expression can be controlled by various factors such as transcription factors

Daifeng Wang 7/8/14 10:18 PM
Deleted: circuits

Daifeng Wang 7/8/14 10:18 PM
Deleted: (RFs)

Daifeng Wang 7/8/14 10:18 PM
Deleted: framework

Daifeng Wang 7/8/14 10:18 PM
Deleted: RFs'

Daifeng Wang 7/8/14 10:18 PM
Deleted: circuits

Daifeng Wang 7/8/14 10:18 PM
Deleted: RFs

Daifeng Wang 7/8/14 10:18 PM
Deleted: test

Daifeng Wang 7/8/14 10:18 PM
Deleted: method using

Daifeng Wang 7/8/14 10:18 PM
Deleted: cell cycle data. We find that co-operative logic gates (e.g. AND gate) are better at describing the

Daifeng Wang 7/8/14 10:18 PM
Deleted:) regulatory activity. We find that target gene expression fold-changes in yeast

Daifeng Wang 7/8/14 10:18 PM
Deleted: experiments support the predicted logic-gate like cooperation of TFs. Next, we look at cooperativity among TFs and miRNAs in human leukemia cells using

Daifeng Wang 7/8/14 10:18 PM
Deleted: K562,

Daifeng Wang 7/8/14 10:18 PM
Deleted: Acute Myeloid Leukemia (AML) samples in

Daifeng Wang 7/8/14 10:18 PM
Deleted: . We find that overall AML-related TFs (

Daifeng Wang 7/8/14 10:18 PM
Deleted: genes) such as MYC do not cooperate with other TFs or miRNAs. However, the logic gates assessment highlights some interactions betwee ... [1]

Daifeng Wang 7/8/14 10:18 PM
Deleted: our method

Daifeng Wang 7/8/14 10:18 PM
Deleted: framework

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Hardison and Taylor, 2012; ... [2]

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Peter and Davidson, 2011)

*To whom correspondence should be addressed.

2. THESE

CSPS INTRO + EMAIL

(TFs), histone modifications, enhancers (distal TFs), and non-coding RNAs. Thus, beyond individual factor behaviors such as binding, activating or repressing targets, ones should look at gene regulatory functions from the point of view of large order groups. More and more experimental and computational studies have indeed shown that commonly regulatory factors work together rather than individually to regulate transcription. Those approaches study TF-TF relationships from various aspects 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 (Banerjee and Zhang, 2003; Hardison and Taylor, 2012; Karczewski, et al., 2011). Also, TFs cooperate with other factors (e.g. miRNAs) to co-regulate gene expression (Gerstein, et al., 2012; Poos, et al., 2013). However, those previous efforts have focused solely on the identification of the wiring relationships between TFs (e.g. co-binding, co-association and co-expression) leaving untouched the cooperative patterns of TFs that drive the biological functions behind the wiring diagrams. Similar to an electronic circuit, where SIM wiring different elements (e.g. as resistors, capacitors, etc) as electronic units can generate various (Rabaey, et al., 2003) electrical functions, connecting diverse regulatory factors as functional modules will result in different biological functions. Thus, beyond identifying the wiring relationships among individual regulatory factors, it is necessary to study their cooperative patterns, and further regulatory functional modules driven by those cooperative patterns.

IS IT NEC TO SAY THIS ?

Regulatory factors control gene expression in a discrete way, as such, in numerous cases gene regulation can be regarded as a logic process (Albert and Othmer, 2003; Das, et al., 2009; Mangan and Alon, 2003; Peter and Davidson, 2011; Peter, et al., 2012; Shmulevich and Dougherty, 2007; Tu, et al., 2013; Xie, et al., 2011). While DNA sequence motifs follow the combinatorial logic (AND, OR and NOT) to match gene expression patterns (Beer and Tavazoie, 2004), TFs can still connect with binding TFs via protein-protein interactions and control gene expression without binding directly to regulatory sequence elements (Farnham, 2009; Neph, et al., 2012). Moreover, combinatorial logics are much more numerous than the three simple logic operations (AND, OR and NOT) (Mangan and Alon, 2003). For example, there are 16 logic gates for any two-input-one-output scenario (including all possible logic combinations between positive and negative regulators). As such, in order to capture all possible combinatorial co-operations between TFs and other regulatory factors we need a more complex model. Previous studies took advantage of binarized regulatory data provided by perturbation experiments such as TF knock-out and Boolean model to capture this logic processing, especially for logic combinatorial effects of different TFs working together (Somogyi and Sniegoski, 1996). The simple binary operations in the Boolean

Daifeng Wang 7/8/14 10:18 PM
Deleted: Among them, TFs play key roles in transcription regulation. Previous

Daifeng Wang 7/8/14 10:18 PM
Deleted: TFs

Daifeng Wang 7/8/14 10:18 PM
Deleted: Hardison and Taylor, 2012; Karczewski, et al., 2011)

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Gerstein, et al., 2012; Poos, et al., 2013)

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Rabaey, et al., 2003)

Daifeng Wang 7/8/14 10:18 PM
Deleted: TFs

Daifeng Wang 7/8/14 10:18 PM
Deleted: further

Daifeng Wang 7/8/14 10:18 PM
Deleted: the

Daifeng Wang 7/8/14 10:18 PM
Deleted: of TFs

Daifeng Wang 7/8/14 10:18 PM
Deleted: activity

Daifeng Wang 7/8/14 10:18 PM
Deleted: Das, et al., 2009

Daifeng Wang 7/8/14 10:18 PM
Deleted: Mangan and Alon, 2003

Daifeng Wang 7/8/14 10:18 PM
Deleted: Peter and Davidson, 2011

Daifeng Wang 7/8/14 10:18 PM
Deleted: Peter, et al., 2012

Daifeng Wang 7/8/14 10:18 PM
Deleted: Shmulevich and Dougherty, 2007

Daifeng Wang 7/8/14 10:18 PM
Deleted: Tu, et al., 2013

Daifeng Wang 7/8/14 10:18 PM
Deleted: Xie, et al., 2011

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Farnham, 2009; Neph, et al., 2012)

Daifeng Wang 7/8/14 10:18 PM
Deleted: the

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Mangan and Alon, 2003)

Daifeng Wang 7/8/14 10:18 PM
Deleted: RFs

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Somogyi and Sniegoski, 1996)

THINK THIS COULD BE SMD DIFF

SAME WIRING DIFF FUNC

model are also computationally efficient. However, previous efforts focused only on a small set of genes, missing the genome-wide identification and characterization of logic operations in gene regulation. Thus our study gives a comprehensive analysis of all possible regulatory logic operations in from a genome-wide perspective.

TFs along with other regulatory factors interact with each other to form regulatory networks, which can be modeled as directed networks and structured in a hierarchical way with top, middle and bottom levels (Bhardwaj, et al., 2010; Bhardwaj, et al., 2010; Gerstein, et al., 2012). The feed-forward loops (FFLs), consisting of two RFs, one regulating another along with a common target, are a common hierarchically structured motif found in regulatory networks, and can be described by different logic gates according to known positively (activator) or negatively (repressor) regulating factors (Mangan and Alon, 2003). The known knowledge that RFs are activators or repressors, however, is insufficient. Moreover, a same RF could switch between activator and repressor to different targets. Thus, our method is designed to identify RF logics to various targets without prior knowledge about activators and repressors.

NOT SURE! SEE CONNECT

In this paper, we developed a novel computational method, Loregic, which integrates gene expression and regulatory data, and characterizes logic operations of gene regulatory factors at genome-wide scale using logic-circuits models. Loregic classifies individual regulatory factors into functional modules according to the regulatory network, and look at how modular genes act functionally as logic circuits. We apply our method to study regulatory factors (TFs and micro-RNAs) in yeast and human cancer datasets.

OVER VIEW OF METH. 2.

2 RESULTS

& HOW MANY W. INCON

2.1 Yeast TFs are cooperative during cell cycle

We identified ~39k TF-TF-target triplets from 176 different TFs using TF-target assignments in (Harbison, et al., 2004; Jothi, et al., 2009). We used Loregic to characterize their TF-TF-target logics during yeast cell cycle across 59 time points (see Methods). We found 4126 TF-TF-target triplets with consistent logic gates (Fig. 3A). Among those, we found that AND (i.e., “T=RF1*RF2”), “T=~RF1*RF2”, and “T=RF1*~RF2” logic gates, have more triplets matched than all the others. The AND triplets mean that both TFs have to be present to activate the expression of their target gene (see interpretations for other logic gates in Fig. S1). After matching all triplets against logic gates, we were able to check how consistent logic gates change for the triplets with the same RF1 and RF2. For example (Fig. 3B), we grouped RF1-RF2 pairs into three categories (Fig. 3B): 1) the triplets with RF1 and RF2 are almost consistent with logics, and have homogeneous consistent logics gate (e.g., top table); 2) the

STILL CONFUSED

CSDS PRELIM RES ON YEAST

Daifeng Wang 7/8/14 10:18 PM
Deleted: ….and structured in a hierar ... [3]

Daifeng Wang 7/8/14 10:18 PM
Deleted:

Daifeng Wang 7/8/14 10:18 PM
Moved down [1]: <#>MATERIAL ... [4]

Daifeng Wang 7/8/14 10:18 PM
Deleted: Loregic is a novel computati ... [5]

Daifeng Wang 7/8/14 10:18 PM
Moved down [2]: Loregic algorithr ... [6]

Daifeng Wang 7/8/14 10:18 PM
Deleted: corresponding binarized ger ... [7]

Daifeng Wang 7/8/14 10:18 PM
Moved down [3]: Step 4: Match th ... [8]

Daifeng Wang 7/8/14 10:18 PM
Deleted: ; ... [9]

Daifeng Wang 7/8/14 10:18 PM
Moved down [4]: <#>**Gene expr** ... [10]

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Cho, et al., 1998; Spellma ... [11]

Daifeng Wang 7/8/14 10:18 PM
Moved down [5]: . We combined a ... [12]

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Harbison, et al., 2004; Jotl ... [13]

Daifeng Wang 7/8/14 10:18 PM
Moved down [6]: , and found 3901 ... [14]

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Consortium, 2011; Djebali ... [15]

Daifeng Wang 7/8/14 10:18 PM
Moved down [7]: . Because TFs ca ... [16]

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Yip, et al., 2012)

Daifeng Wang 7/8/14 10:18 PM
Moved down [8]: .

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Mussel, et al., 2010)

Daifeng Wang 7/8/14 10:18 PM
Moved down [9]: . This method us ... [17]

Daifeng Wang 7/8/14 10:18 PM
Deleted: A logic gate with two input ... [18]

Daifeng Wang 7/8/14 10:18 PM
Moved down [10]: With 2^4 differe ... [19]

Daifeng Wang 7/8/14 10:18 PM
Deleted: (5+4+5+4)/20=0.9. ... [20]

Daifeng Wang 7/8/14 10:18 PM
Moved down [12]: , given a triplet ... [21]

Daifeng Wang 7/8/14 10:18 PM
Formatted ... [22]

Daifeng Wang 7/8/14 10:18 PM
Deleted: predictions…ssignments in ... [23]

triplets with RF1 and RF2 are almost consistent with logics, but have inhomogeneous consistent logics gates; i.e., consistent logic gates are different for targets (e.g., middle table); 3) the triplets with RF1 and RF2 do not have consistent logics gates; i.e., most are inconsistent with logic gates (e.g., bottom table).

Daifeng Wang 7/8/14 10:18 PM
Deleted: varied logics across

CSDS VALIDATION PRELIM

Daifeng Wang 7/8/14 10:18 PM
Deleted: logic co-operations across targets

2.2 Deleting TFs with cooperative logic gates gives rise to significantly higher fold changes of target gene expression

The yeast TF knockout experiments gave us fold changes in gene expression as a result of deleting a single TF (Hu, et al., 2007; Reimand, et al., 2010). If a target gene is regulated by two cooperative TFs in an “AND” relationship, deletion of either TF may corrupt the cooperativity and that impacts gene expression. For example, for the triplets with high significant scores at “AND” gate, 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 non-cooperative TFs such as “T=RF1” or “T=RF2” gates, i.e., one of TFs (dominate TF) fully determines target gene expression, we found that target genes are more affected (down-regulated) by the removal of the dominant TFs rather than by deleting the other TFs (*t-test p-value* < 0.05 for T=RF1, <0.005 for T=RF2).

BOTH TFS REQUIRED

GIVE US A WAY OF VALIDATING LOGIC

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Hu, et al., 2007; Reimand, et al., 2010)

2.3 Logic operations between TF-TF, miRNA-TF and distTF-TF across targets in Acute Myeloid Leukemia

Next, we applied Loregic to analyze the human leukemia datasets. We identified 50865 TF-TF-target triplets from ChIP-seq experiments for 70 TFs in ENCODE K562 cell line (Consortium, 2011; Djebali, et al., 2012; Gerstein, et al., 2012), and also 821 distTF-TF-target triplets, where distTFs were predicted to bind distal regulatory regions such as enhancers of targets in (Yip, et al., 2012). Moreover, because miRNAs and TFs have been found to co-regulate common target genes (Cheng, et al., 2011; Gerstein, et al., 2012), we studied their logic co-operations. We obtained 222 miRNAs that have highly confident interactions with their targets in K562 cell line (Chen, et al., 2014). Thus, integrating miRNA- and TF-target pairs in K562, we identified 56944 miRNA-TF-target triplets. The gene/miRNA expression datasets used comprise ~20k protein-coding genes across 197 samples and 705 miRNAs across 188 samples in TCGA Acute Myeloid Leukemia (AML). 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 consistent logic gates found from TF-TF-target triplets, miRNA-TF-target triplets and distTF-TF-target triplets. In the case of TF-TF-target triplets, we randomly assigned TFs as RF1 and RF2 and observed that the numbers of consistent logic gates between the complementary gates (e.g.

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Consortium, 2011; Djebali, et al., 2012; Gerstein, et al., 2012)

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Yip, et al., 2012)

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Cheng, et al., 2011; Gerstein, et al., 2012)

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Chen, et al., 2014)

L2.

"T=RF1+~RF2" vs. "T=~RF1+RF2", "T=RF1" vs. "T=RF2", etc.), are roughly equal (Fig. 4A). The OR gate is most among consistent logic gates, where either RF1 or RF2 can activate the target expression. But for miRNA-TF-target and distTF-TF-target triplets, where RF1=miRNA or distTF and RF2=TF, we noticed differences between complementary gates (Figs. 4B and 4C). The most consistent logic gate is "T=RF2" gate, which suggests that TFs binding to promoters (RF2) can determine the target expressions without being influenced by the presence of miRNAs or distTFs.

2

Daifeng Wang 7/8/14 10:18 PM
Deleted: gate is the OR gate

Daifeng Wang 7/8/14 10:18 PM
Deleted: expressions

Daifeng Wang 7/8/14 10:18 PM
Deleted: /

2.4 AML-related TFs (including MYC) solely determine target expressions

The transcription factor, MYC has been found to universally amplify target gene expressions in lymphocytes (Nie, et al., 2012), implying that it does not require cooperation from other TFs in order to preform its regulatory function. We identified 2153 MYC-TF-target (i.e., RF1=MYC, RF2=other TFs, T=target) triplets with 67 other TFs, and found that 905 out of 2153 triplets can be assigned significantly high scores (s=1) for one logic gate. The two most enriched consistent logic gates among the 905 ones were "T=MYC" (133 triplets, hypergeometric test < $4.3*10^{-27}$) and "OR" (T=MYC+TF) (211 triplets, hypergeometric test < $1.1*10^{-21}$) (Fig. 5A). "T=MYC" indicates that the target gene expression is solely determined by MYC, while "T=MYC+TF" means that either MYC or TF can regulate the target's expression. However, both scenarios suggest that ~~MYC is able to control the target expressions~~ without requiring the presence of other TFs. These results support the recent finding that MYC plays a universal amplifier role in gene expression. Next we analyzed all the triplets associated with AML-related TFs (i.e., RF1=AML-related TFs, RF2=non-AML related TFs, T=targets) from cancer gene datasets (Forbes, et al., 2011), and found that the most enriched consistent logic gates are "T=RF1" and "T=~RF1" (Fig. 5B). We did not find any enrichment for these two gates in triplets containing only non-AML TFs. Therefore, this suggests that the AML-related TFs activate or repress target expressions by themselves.

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Nie, et al., 2012)

TURN
ON

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Forbes, et al., 2011)

2.5 Prediction of indirect binding from cooperative TF motifs analysis

We studied TF promoter motifs in the target genes promoter regions (1000 bps (yeast) and 5000 bps (human) upstream of TSS) (DebRoy, 2013; Lawrence, 2014; Li, 2014; Pages, 2014). 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 with consistent "AND" gates (see examples in Fig. 6), 348 have one TF whose motifs is not present in the target promoters (364 out of 1100 for "T=RF1*~RF2", 377 out of 1095 for "T=~RF1*RF2"). Similarily, in the human leukemia dataset, we found that out of 888 TF-TF-target triplets with consistent "AND" gates, 71 have one TF whose motifs

Daifeng Wang 7/8/14 10:18 PM
Deleted: (DebRoy, 2013; Lawrence, 2014; Li, 2014; Pages, 2014)

is not present in the target promoters. For example (Fig. S2), the triplet: RF1=USF2, RF2=NFYB, T=YPEL1 has a consistent "AND" gate, and both TFs have motifs in the YPEL1 promoter region (see Fig. S2 for IGV visualizations). 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 has a dominant "AND" gate. However, USF2 and NFE2 are connected through protein-protein interactions, and consequently NFE2 is regulating NBPF1 through indirect binding (Neph, et al., 2012). As such, we suspect that those TFs with absent motifs (as above) can potentially regulate targets through indirect binding by cooperating (through protein-protein interactions) with directly bound TFs (Biddie, et al., 2011; Farnham, 2009; Gordan, et al., 2009; Neph, et al., 2012; Zhao, et al., 2012).

2.6 Logic gates for feed-forward loops (FFLs)

CONSIS CIRCUIT LOGIC ?

Feed-forward loops (FFLs) are RF1-RF2-T triplets where RF1 is also regulating RF2. FFLs are found to be important patterns in regulatory networks, with many following the logics (Mangan and Alon, 2003). For the yeast cell cycle, we found that 659 FFLs have consistent logic gates. Two enriched consistent logic gates among FFLs are "AND" (162 FFLs, hypergeometric test $<1.3*10^{-3}$) and "T=RF1" (159 FFLs, hypergeometric test $<7.5*10^{-5}$). It has been shown that these two logic gates that also match the logics for coherent type 1 FFL (e.g. RF1 activates RF2, both of which activate the target) are more abundant that other logic gates (Mangan and Alon, 2003). Then we investigated the FFLs in human leukemia TF-TF-T triplets, and found that the two most abundant consistent logic gates are 'T=RF1" (1306 FFLs, hypergeometric test $<3.4*10^{-9}$) and "T=RF1+~RF2" (1765 FFLs, hypergeometric test $<1.7*10^{-5}$), both of which correspond to the coherent type 4 FFL (RF1 down-regulates RF2 and RF2 down-regulates target but RF1 activates target). This suggests that the master TFs, (RF1s) of FFLs in leukemia, aims to activate the targets, but due to the gene down regulation action from the second TF, (RF2s,), RF1s simultaneously down-regulate RF2s to activate the target. Moreover, we did not find any enriched logic gates among the triplets that do not form FFLs in both yeast and human.

2.

2.

2.7 miRNAs and c-Myc double down-regulate to each other

MYC and miRNAs have been found to down-regulate each other by forming double down-regulatory FFLs in leukemia (Tao, et al., 2014). We identified 1805 miRNA-MYC-target triplets with 117 miRNAs, and 1143 out of these 1805 triplets have consistent logic gates. Then, out of the 1143 triplets, 446 match "T=MYC" (hypergeometric test $< 2.5*10^{-124}$), and 201 match "T=~miRNA+MYC" (hypergeometric test $< 4.1*10^{-25}$). These two most enriched logic gates, also match the logic for the coherent type 4 FFL as previously shown in (Mangan and Alon, 2003). This implies that miRNAs repress target ex-

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Neph, et al., 2012)

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Biddie, et al., 2011; Farnham, 2009; Gordan, et al., 2009; Neph, et al., 2012; Zhao, et al., 2012)

Daifeng Wang 7/8/14 10:18 PM
Deleted: logic gates arithmetics

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Mangan and Alon, 2003)

Daifeng Wang 7/8/14 10:18 PM
Deleted:

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Mangan and Alon, 2003)

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Tao, et al., 2014)

Daifeng Wang 7/8/14 10:18 PM
Deleted: Out

Daifeng Wang 7/8/14 10:18 PM
Deleted: are

Daifeng Wang 7/8/14 10:18 PM
Deleted: matching

Daifeng Wang 7/8/14 10:18 PM
Deleted: logics

Daifeng Wang 7/8/14 10:18 PM
Deleted: (Mangan and Alon, 2003)

pressions, while MYC activates it and simultaneously down-regulates miRNAs. We also found that there were 56 triplets matching "T=~miRNA*MYC", and 16 triplets matching "T=~miRNA", two logics matching coherent type 2 FFL. This result suggests that miRNAs repress both MYC and target expressions, while MYC aims to activate the targets. In short, those matched logic gates support that the miRNAs and MYC form indeed ~~a double-negative regulatory loop in leukemia.~~

Daifeng Wang 7/8/14 10:18 PM
Deleted:

TAKES AS INPUT

3 MATERIALS AND METHODS

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [1]

CSDS AIM

Loregic inputs a regulatory network (regulatory factors and their target genes) along with binarized gene expression datasets across multiple samples. The binarized gene expressions (on or off) are resulting from regulations in the network. The inputs can be chosen from different resources to meet user interests. In this paper, we used BoolNet (Mussel, et al., 2010) to obtain binarized gene expressions, but ones can also input their customized binarized expression datasets. Loregic tries to describe each regulatory module (triplet) consisting of two regulator factors and one common target gene using a particular type of logic; i.e., the binarized gene expression changes in the triplet across samples highly match a particular two-input-one-output logic gate. If Loregic is able to find such a logic gate, we claim that the triplet is consistent with logics, and refer to the gate as the consistent logic gate for the triplet, and give the corresponding consistency score. If not, we claim that the triplet is inconsistent with logic gates; i.e., the cooperativity of two RFs may not be described by logics. In this paper, we reveal ~~Loregic's capabilities,~~ analyzing transcription factors, micro-RNAs (miRNAs) and their target genes. In details, Loregic algorithm comprises of five steps (Figure 1):

WD

DESC

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [2]

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

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

Step 3: Given a particular triplet (RF1, RF2 and T) query the binarized gene expression data;

Step 4: Match the triplet's gene expressions against all possible two-in-one-out logic gates based on the binary values;

Step 5: Find the consistent logic gate(s) that best matches the expressions and calculate the consistency score. Test the score significance against random effects;

Repeat Step 3-5 for all triplets in the regulatory network.

Daifeng Wang 7/8/14 10:18 PM
Formatted: Indent: First line: 0.12"

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [3]

After determining the triplets consistent with logics, we can map them to other features of regulatory networks, and see how logics enrich in those features. Though we can relate any regulatory features (Discussion), we focus on two important ones in this paper: 1) Feed-forward loops (FFLs), a common

PARTIC GATE LOGIC ARE ENRIC.

network motif with RF1-RF2-T and RF1 also regulating RF2, where we predict FFL logics; 2) TF promoter sequence motifs, where we use the predicted TF logics to infer the potential indirect bindings.

3.1 Gene expression, transcription factor and miRNA datasets

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [4]

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') (Cho, et al., 1998; Spellman, et al., 1998). We combined all three datasets (5581 genes and 59 time points), and standardized gene expressions for each time point. For gene regulation in yeast, we used the transcription factors with their target genes identified in (Harbison, et al., 2004; Jothi, et al., 2009), and found 39011 TF-TF-target triplets.

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [5]

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [6]

For the study of gene expression in human leukemia, we obtained RPKM expressions in RNA-seq for ~20k protein-coding genes (705 miRNAs) across 197 (188) samples with Acute Myeloid Leukemia (AML) from The Cancer Genome Atlas (TCGA) Data Portal (https://tcga-data.nci.nih.gov/tcga/). We standardized log(RPKM+1) across genes for each sample. We identified 50865 TF1-TF2-target (i.e., RF1=TF1, RF2=TF2, T=target gene) triplets using ChIP-seq data in ENCODE K562 cell line (Consortium, 2011; Djebali, et al., 2012; Gerstein, et al., 2012), and 56944 miRNA-TF-target (i.e., RF1=miRNA, RF2=TF, T=target gene) triplets using confident miRNA-targets for human K562 cell line in (Chen, et al., 2014). Because TFs can also bind to the distal regulatory regions such as enhancers (here denoted as 'distTF'), we also included 821 distTF-TF-target (i.e., RF1=distTF, RF2=TF, T=target gene) triplets. The distTFs were obtained from (Yip, et al., 2012).

REAL METH.

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [7]

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [8]

3.2 Converting gene expression changes over conditions to Boolean values

Previous Boolean models normally converted the gene expression to 1 or 0 based on whether its expression values are greater (1) or not (0) than an imposed threshold. This method, however, is subjective to the selected threshold, which may vary depending on genes or datasets. Moreover, the gene expression varies dynamically over conditions if their regulators express differently. As such there can be different thresholds for highly or lowly expressed genes. Thus, we converted gene expressions to Boolean values (1 or 0) using BoolNet (Mussel, et al., 2010). This method uses K-means clustering to group genes into co-expression modules, and discretizes gene expressions to binary values from co-expressed modular patterns across time points (yeast) or AML patients (human).

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [9]

3.3 Mapping and scoring a RF-RF-T triplet to 16 logic gates

A logic gate with two inputs (RF1, RF2) and one output (T) can be determined by a combination of four (RF1, RF2, T) binary vectors, v_1=(RF1=0, RF2=0, T), v_2= (RF1=0, RF2=1, T), v_3= (RF1=1, RF2=0, T), and v_4=(RF1=1, RF2=1, T) with specific values (0 or 1) for T, also known as a truth table. With 2^4 different combinations of T values, we obtain 16 different logic gates (Fig. S1), where '~' denotes NOT (negative regulation), '*' denotes AND and '+' denotes OR logic operations. Given a RF1-RF2-T triplet, we find output T (0 or 1) for each of four input combinations of RF1 and RF2, and find the logic gate(s) whose truth tables matches best the four outputs as follows. Suppose that we have a triplet with m binary vectors in total. For a given logic gate g, we define the gate consistency score of the triplet, $S(g)=(n_1 + n_2 + n_3 + n_4)/m$, where n_i as number of vectors matching $v_i(g)$ with i=1,2,3,4.

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [10]

2, 1

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [11]

For example (Fig. 2), suppose a triplet with RF1=TF1, RF2=TF2, and T=target, has m=20 binary vectors after conversion. 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 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. 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/expected triplet, because for the given input the majority of cases has zero as the output value. Similarly, when RF1=1 and RF2=0, T=0 is chosen (magenta) because it appears more than T=1. Finally, when RF1=1 and RF2=1, T=1 is chosen (orange) because it appears four times but T=0 appears only once. Combining the outputs chosen for four different input combinations of RF1 and RF2, we obtain the triplet's truth table, and find that it matches the AND logic gate. As such we define the AND gate as consistent logic gate for this triplet, and calculate its consistency score. This score is equal to number of the vectors matching AND logic gate over the total number of vectors, i.e., $S(\text{AND})=(n_1 + n_2 + n_3 + n_4)/m = (5+4+5+4)/20=0.9$.

CSDS
AIM

ND

3.4 Testing consistent logic gate significances of triplets by randomizing their targets

In order to test the consistent logic gate identified not due to chance, given a triplet of (RF1, RF2, T), we calculate its significances over the 16 logic gates' scores as follows. We suppose that it matches the k^{th} logic gate, G_k. We replace the target gene, T by a randomly selected gene N times (e.g., N=1000), and define its significance score, as $p(G_k)$=(number of matched logic gate=G_k)/N. Thus, a high significance score implies that random effects may cause the matched logic gate. In this paper, we select the consistent logic gates within top 2% of consistency and significance scores.

Daifeng Wang 7/8/14 10:18 PM
Moved (insertion) [12]

Daifeng Wang 7/8/14 10:18 PM
Formatted: Font:Times New Roman

Daifeng Wang 7/8/14 10:18 PM
Formatted: Normal, Justified

Daifeng Wang 7/8/14 10:18 PM
Formatted: Font:Times New Roman, 18 pt

4 DISCUSSION

Loregic is a computational method using logic-circuit models to characterize the cooperativity among regulatory factors such as transcription factors and miRNAs by integrating gene expression and regulatory networks. Loregic can be further extended to study coordination among other regulatory elements such as splicing factors, long non-coding RNAs and so on through availability of high quality expression (e.g., RNA-seq, small RNA-seq), and regulation (e.g., ChIP-seq, CLIP-seq, DNase-seq) datasets.

Daifeng Wang 7/8/14 10:18 PM
Deleted: used

Daifeng Wang 7/8/14 10:18 PM
Deleted: cooperativity using logic-circuit models and

Daifeng Wang 7/8/14 10:18 PM
Deleted: In this paper, we focus on the logic operations among transcription factors and miRNAs.

Loregic is capable to relate triplet logics to any features of regulatory networks. Here, we map the logic-consistent triplets to two regulatory features including feed-forward loops (FFLs), a common network motif reflecting the hierarchical structure of regulatory networks. A straightforward future work is to find enrichments of the logic consistent triplets in hierarchical layers, and identify logic co-operations between and among regulatory factors at different hierarchical layers; e.g., top, middle and bottom layers found in (Bhardwaj, et al., 2010; Bhardwaj, et al., 2010; Gerstein, et al., 2012).

We test Loregic using 2-RFs-1-target triplets and particularly focusing on TF/miRNA-TF-target triplets. We highlight that Loregic could be also used to analyze the regulatory modules with multiple RFs and multiple target genes as long as there is enough expression data support (2^N samples, N is number of RFs in module). For those regulatory modules with N_1 RFs and N_2 targets, we have 2^{N1} input combinations and 2^{N2} output combinations, and we calculate the consistency scores associated with corresponding logic gates with N_1-input and N_2-output.

We convert the gene expression to Boolean values by comparing co-expression patterns across samples. Using a significance test, we are able to use binarized expression values, even for noisy datasets (e.g. yeast microarrays) and thus reduce the noise effect. Loregic is also compatible with other discretization methods, and is able to use any binarized gene expression data input.

OTHER TFS

We find that some triplets didn't have strong consistency and significance scores for any logic gates, indicating that the regulatory cooperativity between those RFs might be random. Another explanation is that the target gene expression might be driven by other stochastic biological processes, rather than deterministic ones, and thus cannot be simply explained as logic operations. Moreover, the target gene can be regulated by more than two RFs, thus we might need the higher-order logic circuit models with multiple inputs (>2) as discussed above to capture the RF logics to the target.

Daifeng Wang 7/8/14 10:18 PM
Deleted: found

Daifeng Wang 7/8/14 10:18 PM
Deleted: processes

Daifeng Wang 7/8/14 10:18 PM
Deleted: regulation

To our knowledge, Loregic is the first computational method to systematically characterize the regulatory cooperativity using logic-circuit models. It has a wide variety of applications for the study of regulatory mechanisms, and can help build the gene regulatory panoramagram.

ACKNOWLEDGEMENTS

Funding: National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Albert, R. and Othmer, H.G. (2003) 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.

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

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

Bhardwaj, N., Kim, P.M. and Gerstein, M.B. (2010) Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators, *Science signaling*, **3**, ra79.

Bhardwaj, N., Yan, K.K. and Gerstein, M.B. (2010) 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.

Biddie, S.C.*, et al.* (2011) Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding, *Mol Cell*, **43**, 145-155.

Chen, D.*, et al.* (2014) Dissecting the chromatin interactome of microRNA genes, *Nucleic Acids Res*, **42**, 3028-3043.

Cheng, C.*, et al.* (2011) Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data, *PLoS computational biology*, **7**, e1002190.

Cho, R.J.*, et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle, *Molecular cell*, **2**, 65-73.

Consortium, E.P. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE), *PLoS biology*, **9**, e1001046.

Das, D., Pellegrini, M. and Gray, J.W. (2009) A primer on regression methods for decoding cis-regulatory logic, *PLoS computational biology*, **5**, e1000269.

DebRoy, H.P.a.P.A.a.R.G.a.S. (2013) Biostrings: String objects representing biological sequences, and matching
algorithms, *R package version 2.28.0*.

Djebali, S.*, et al.* (2012) Landscape of transcription in human cells, *Nature*, **489**, 101-108.

Farnham, P.J. (2009) Insights from genomic profiling of transcription factors, *Nat Rev Genet*, **10**, 605-616.

Forbes, S.A.*, et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer, *Nucleic acids research*, **39**, D945-950.

Gerstein, M.B.*, et al.* (2012) Architecture of the human regulatory network derived from ENCODE data, *Nature*, **489**, 91-100.

Gordan, R., Hartemink, A.J. and Bulyk, M.L. (2009) Distinguishing direct versus indirect transcription factor-DNA interactions, *Genome Res*, **19**, 2090-2100.

Harbison, C.T.*, et al.* (2004) Transcriptional regulatory code of a eukaryotic genome, *Nature*, **431**, 99-104.

Hardison, R.C. and Taylor, J. (2012) Genomic approaches towards finding cis-regulatory modules in animals, *Nature reviews. Genetics*, **13**, 469-483.

Hu, Z., Killion, P.J. and Iyer, V.R. (2007) Genetic reconstruction of a functional transcriptional regulatory network, *Nature genetics*, **39**, 683-687.

Jothi, R.*, et al.* (2009) Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture, *Molecular systems biology*, **5**, 294.

Karczewski, K.J.*, et al.* (2011) Cooperative transcription factor associations discovered using regulatory variation, *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 13353-13358.

Lawrence, M.C.a.H.P.a.P.A.a.S.F.a.M.M.a.D.S.a.M. (2014) GenomicFeatures: Tools for making and manipulating transcript centric annotations, *R package version 1.12.4*.

Li, H.P.a.M.C.a.S.F.a.N. (2014) AnnotationDbi: Annotation Database Interface, *R package version 1.22.6*.

Mangan, S. and Alon, U. (2003) 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.

Mussel, C., Hopfensitz, M. and Kestler, H.A. (2010) BoolNet--an R package for generation, reconstruction and analysis of Boolean networks, *Bioinformatics*, **26**, 1378-1380.

Neph, S.*, et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints, *Nature*, **489**, 83-90.

Nie, Z*., et al.* (2012) c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells, *Cell*, **151**, 68-79.

Pages, H. (2014) BSgenome: Infrastructure for Biostrings-based genome data packages, *R package version 1.28.0*.

Peter, I.S. and Davidson, E.H. (2011) Evolution of gene regulatory networks controlling body plan development, *Cell*, **144**, 970-985.

Peter, I.S., Faure, E. and Davidson, E.H. (2012) 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.

Poos, K*., et al.* (2013) How microRNA and transcription factor co-regulatory networks affect osteosarcoma cell proliferation, *PLoS computational biology*, **9**, e1003210.

Rabaey, J.M., Chandrakasan, A.P. and Nikoli*c, B. (2003) *Digital integrated circuits : a design perspective*. Prentice Hall electronics and VLSI series. Pearson Education, Upper Saddle River, N.J.

Reimand, J*., et al.* (2010) Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets, *Nucleic acids research*, **38**, 4768-4777.

Shmulevich, I. and Dougherty, E.R. (2007) *Genomic Signal Processing*. Princeton series in applied mathematics. Princeton University Press, Princeton.

Somogyi, R. and Sniegoski, C.A. (1996) Modeling the complexity of genetic networks: Understanding multigenic and pleiotropic regulation, *Complexity*, **1**, 45-63.

Spellman, P.T*., et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol Biol Cell*, **9**, 3273-3297.

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.

Tu, S., Pederson, T. and Weng, Z. (2013) Networking development by Boolean logic, *Nucleus*, **4**, 89-91.

Xie, Z*., et al.* (2011) Multi-input RNAi-based logic circuit for identification of specific cancer cells, *Science*, **333**, 1307-1311.

Yip, K.Y*., et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors, *Genome biology*, **13**, R48.

Zhao, Y*., et al.* (2012) Improved models for transcription factor binding site identification using nonindependent interactions, *Genetics*, **191**, 781-790.