

UC Merced

UC Merced Previously Published Works

Title

Inferring gene regulatory networks using transcriptional profiles as dynamical attractors

Permalink

<https://escholarship.org/uc/item/1v06p0wc>

Journal

PLOS Computational Biology, 19(8)

ISSN

1553-734X

Authors

Li, Ruihao

Rozum, Jordan C

Quail, Morgan M

et al.

Publication Date

2023

DOI

10.1371/journal.pcbi.1010991

Peer reviewed

RESEARCH ARTICLE

Inferring gene regulatory networks using transcriptional profiles as dynamical attractors

Ruihao Li¹, Jordan C. Rozum², Morgan M. Quail¹, Mohammad N. Qasim¹, Suzanne S. Sindi³, Clarissa J. Nobile^{4,5}, Réka Albert^{6,7*}, Aaron D. Hernday^{4,5*}

1 Quantitative and Systems Biology Graduate Program, University of California, Merced, Merced, California, United States of America, **2** Department of Systems Science and Industrial Engineering, Binghamton University (State University of New York), Binghamton, New York, United States of America, **3** Department of Applied Mathematics, University of California, Merced, Merced, California, United States of America, **4** Department of Molecular Cell Biology, University of California, Merced, Merced, California, United States of America, **5** Health Sciences Research Institute, University of California, Merced, Merced, California, United States of America, **6** Department of Physics, Pennsylvania State University, University Park, University Park, Pennsylvania, United States of America, **7** Department of Biology, Pennsylvania State University, University Park, University Park, Pennsylvania, United States of America

* rza1@psu.edu (RA); ahernday@ucmerced.edu (ADH)



OPEN ACCESS

Citation: Li R, Rozum JC, Quail MM, Qasim MN, Sindi SS, Nobile CJ, et al. (2023) Inferring gene regulatory networks using transcriptional profiles as dynamical attractors. *PLoS Comput Biol* 19(8): e1010991. <https://doi.org/10.1371/journal.pcbi.1010991>

Editor: Tim Beißbarth, University Medical Center Göttingen, GERMANY

Received: March 2, 2023

Accepted: July 19, 2023

Published: August 22, 2023

Copyright: © 2023 Li et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The mRNA and csRNA sequencing data can be accessed on GEO (GSE217461 and GSE217383). Our evolutionary algorithm, the datasets used in this study, and the results are available on a GitHub repository at <https://github.com/UJCM-RuihaoLi/GeneRegulatoryNetworkInference>. We have also used Zenodo to assign a DOI to the repository: <https://doi.org/10.5281/zenodo.7553611>.

Funding: This work was supported by the NSF (<https://www.nsf.gov/>) IIS1814405 and

Abstract

Genetic regulatory networks (GRNs) regulate the flow of genetic information from the genome to expressed messenger RNAs (mRNAs) and thus are critical to controlling the phenotypic characteristics of cells. Numerous methods exist for profiling mRNA transcript levels and identifying protein-DNA binding interactions at the genome-wide scale. These enable researchers to determine the structure and output of transcriptional regulatory networks, but uncovering the complete structure and regulatory logic of GRNs remains a challenge. The field of GRN inference aims to meet this challenge using computational modeling to derive the structure and logic of GRNs from experimental data and to encode this knowledge in Boolean networks, Bayesian networks, ordinary differential equation (ODE) models, or other modeling frameworks. However, most existing models do not incorporate dynamic transcriptional data since it has historically been less widely available in comparison to “static” transcriptional data. We report the development of an evolutionary algorithm-based ODE modeling approach (named EA) that integrates kinetic transcription data and the theory of attractor matching to infer GRN architecture and regulatory logic. Our method outperformed six leading GRN inference methods, none of which incorporate kinetic transcriptional data, in predicting regulatory connections among TFs when applied to a small-scale engineered synthetic GRN in *Saccharomyces cerevisiae*. Moreover, we demonstrate the potential of our method to predict unknown transcriptional profiles that would be produced upon genetic perturbation of the GRN governing a two-state cellular phenotypic switch in *Candida albicans*. We established an iterative refinement strategy to facilitate candidate selection for experimentation; the experimental results in turn provide validation or improvement for the model. In this way, our GRN inference approach can expedite the development of a sophisticated mathematical model that can accurately describe the structure and dynamics of the *in vivo* GRN.

MCB1715826 to R.A., National Institutes of Health (NIH) National Institute of Allergy and Infectious Diseases (NIAID) (<https://www.niaid.nih.gov>) R15AI37975 to A.D.H., and NIH National Institute of General Medical Sciences (NIGMS) (<https://www.nigms.nih.gov/>) R35GM124594 to C.J.N.. This work was also supported by the Kamangar family in the form of an endowed chair to C.J.N. The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: Dr. Clarissa J. Nobile is a cofounder of BioSynthesis, Inc., a company developing inhibitors and diagnostics of biofilm formation. All other authors declare that no competing interests exist.

Author summary

The establishment of distinct transcriptional programs, where specific sets of genes are activated or repressed, is fundamental to all forms of life. Sequence-specific DNA-binding proteins, often referred to as regulatory transcription factors, form interconnected gene regulatory networks (GRNs) which underlie the establishment and maintenance of specific transcriptional programs. Since their discovery, many modeling approaches have sought to understand the structure and regulatory behaviors of these GRNs. The field of GRN inference uses experimental measurements of transcript abundance to predict how regulatory transcription factors interact with their downstream target genes to establish specific transcriptional programs. However, most prior approaches have been limited by the exclusive use of “static” or steady-state measurements. We have developed a unique approach which incorporates dynamic transcriptional data into a sophisticated ordinary differential equation model to infer GRN structures that give rise to distinct transcriptional programs. Our model not only outperforms six other leading models, it also is capable of accurately predicting how changes in GRN structure will impact the resulting transcriptional programs. These notable features of our model, in conjunction with experimental validation of our predictions in real-world scenarios, contribute to an advancement in the field of gene regulatory network inference.

Introduction

Gene regulatory networks (GRNs) are comprised of interactions between sequence-specific DNA-binding proteins, or transcription factors (TFs), and their respective regulatory target genes [1]. The characteristics of GRN stability and adaptability underlie the ability of cells to maintain homeostasis [2], respond to environmental variables [3], develop into multicellular organisms [4], and make cell fate decisions [5]. Inferring the architecture of GRNs based on experimental datasets, also known as the “inverse problem” [6], is important to understanding these cellular processes (see [7, 8] for examples). The advent of high-throughput “omics” techniques [9] has dramatically accelerated the pace by which researchers can obtain these experimental data sets for GRN reverse engineering [10]. A commonly used high-throughput technique is RNA sequencing, which effectively identifies and counts the number of transcripts present for each RNA species, and thus generates a transcriptional profile of the cell or tissue being assayed. With multiple bulk or single-cell transcriptional profiles measured at different time points, or in different cell types, the genes that are upregulated and downregulated can be determined and be used to further infer the logic of the GRNs that underlie those regulatory changes [11]. Although transcriptional profiles are informative and have been widely used to study biological processes of interest, they do not directly reflect the regulatory status of genes (i.e., whether a gene is activated or repressed) [12], since some mRNAs are highly stable and can accumulate in cells, while others are degraded.

In the past twenty years, many modeling approaches have been developed to infer GRN architectures using “omics” data [9, 13–15]. GRN inference models can be broadly categorized into three distinct categories, based on the algorithms and hypotheses they employ (see reviews: [16–24]): (i) data-driven static models, which do not simulate the biological processes such as transcription or translation, but hypothesize that interacting genes have correlated expression and use the correlations to infer GRN architecture [25, 26]; (ii) discrete models, which simulate the time evolution of discrete variables that qualitatively describe the activity of

genes [27, 28]; and (iii) continuous dynamical models which simulate the dynamics of gene expression processes in a quantitative manner based a set of linear [29] or non-linear [30, 31] ordinary differential equations (ODEs).

Dynamical models suffer from the “curse of dimensionality”, i.e., the problem of a state space or parameter space growing exponentially with the number of genes considered. One approach to dealing with this challenge when building dynamical models is to focus on dynamics near attractors [32–34], the states toward which a kinetic system tends to evolve and converge. This approach has been successfully applied in Boolean network inference of GRNs [34–38], in which binary variables, 1 and 0, define the state of a gene as “on” or “off”, respectively. Here, we focus on extending the attractor matching approach from Boolean models to ODE models, which can simulate continuous gene expression levels. Specifically, we infer an ODE model of a GRN whose attractors match the experimentally measured attractor states. The primary challenge for implementing the attractor matching approach in network inference lies in the absence of experimentally determined kinetic data for the GRN [23, 39]. This includes crucial information such as transcription rates and mRNA degradation rates, which, unlike time-series transcriptional profiles that represent a trajectory, play pivotal roles in shaping the overall dynamics of the GRN system, specifically the global vector field, and are instrumental in determining the positions and basins of the attractors [12]. As a result, most ODE models must rely on fitting procedures to estimate kinetic parameters [40–44]. This strategy differs significantly from the application of attractor dynamics, where measured kinetic parameters are already known and are used to find and match the attractors. Thus, there are few GRN inference techniques available that can effectively leverage kinetic data when it is available. Although the attractor matching approach has not been applied to ODE-based models for the purpose of GRN inference, previous work has explored the potential of attractor-matching strategies in ODEs [45, 46], and several software tools have been developed in this area. For instance, FOS-GRN [47] and Netland [48] can reconstruct multi-attractor kinetic landscapes with ODEs and user-defined parameters. In addition, there are several studies [49–55] exploring how ODE models of GRNs can be steered from one attractor to another; many of these techniques have Boolean analogues, which led us to more closely examine attractor matching inference strategies for ODEs.

In this work, we have developed a GRN inference approach that extends the application of the attractor matching strategy from a Boolean model to an ODE-based model by incorporating measurements of mRNA synthesis and degradation. Our model can simulate genetic regulatory processes with a novel parameterization framework that is built using combinatorial logic operators and Hill functions and can reveal the correlation between a GRN architecture and its dynamics in terms of fixed-point attractors. While precise estimation of the Hill coefficient is challenging due to the fact that similar dynamics are obtained for a wide range of values, our procedure effectively captures the overall regulatory structure without relying on definitive values. Additional measurements can enhance parameter estimation, and our framework allows for the integration of more advanced techniques if desired. The inferred input parameter estimates are decoupled from the evolutionary algorithm, providing users with the freedom to substitute parameters obtained through alternative and more sophisticated methods. We have tested our model using both *in silico* data and experimental data generated from a real-life GRN constructed using synthetic biology [56]. Our results show that the use of kinetic parameters and the application of attractor dynamics can significantly improve the inference performance of ODE-based GRN models. Furthermore, since the model simulates GRN dynamic systems in a quantitative manner, it can also predict stable-state transcriptional profiles when given a GRN architecture and kinetic parameters. Using this model, we are able to estimate, for the first time, the similarities between an unknown real-world GRN of an

organism and the inferred GRN model based on predictions of the steady state transcriptional profiles that result under perturbations that were not incorporated into the inference process.

Methods

GRN model architecture

Drawing on the conventions of early work, we depict the GRN architecture as a directed graph consisting of nodes representing both TF proteins and their respective coding genes, and edges representing interactions among these nodes. Non-TF coding genes are not considered in our model. For example, in a simple three gene GRN architecture (Fig 1), the nodes A, B, and C represent three distinct TF proteins and the genes that encode them, while the connecting lines indicate physical interactions between the TF protein(s) and their respective regulatory target genes. In our framework, two types of interactions exist between TFs and genes: activating or inhibitory, represented by pointed or blunt arrows, respectively. As shown in Fig 1B and 1C, we denote these TF-gene relationships by an adjacency matrix (named *AM*), which uses 1 or -1 to indicate the activating or inhibitory interactions, and 0 to indicate the absence of an interaction between a given TF and target gene. In addition to the TF-gene interactions, TF-TF interactions may also exist, and are represented by the logic operators AND, OR, or NOT, which qualitatively indicate how multiple TFs are aggregated to affect a common target gene. The qualitative logic of these TF-TF interactions is organized in a protein coordination matrix, or a set of logic gates (denoted by *LG*), whose Boolean values are assigned to decide

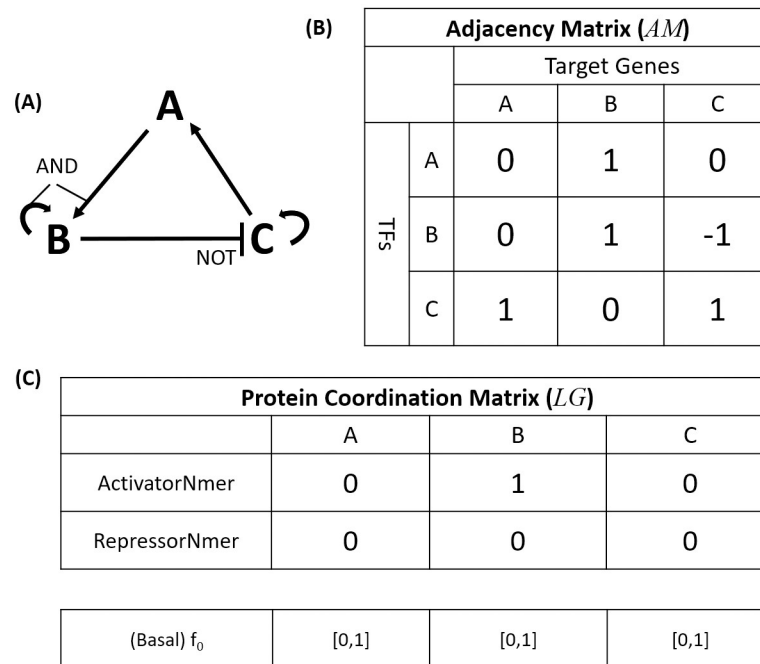


Fig 1. Depiction of a hypothetical GRN architecture. (A) Schematic of a simple GRN in which A and B cooperatively activate B, C activates A and itself, and B represses C in a manner that can override the self-activation of C. (B) The network topology table represents the direct activating, inhibiting, and null connections by 1, -1, and 0, respectively. (C) The protein coordination parameters are assigned to each gene in the genome and qualitatively describe the coordination between each gene's regulatory TFs. 'ActivatorNmer' decides whether the activators of a gene work independently (0) or cooperatively (1); 'RepressorNmer' decides whether the repressors of a gene work independently (0) or cooperatively (1). f_0 determines the basal expression level of a gene and whether its activators or repressors outcompete the other.

<https://doi.org/10.1371/journal.pcbi.1010991.g001>

whether the activators or repressors of a gene work synergistically or independently. If a gene in the model has fewer than two TFs, its protein coordination parameters do not affect the GRN dynamics. In this case, we fix the protein coordination parameters to 0. We use f_0 , which is bounded between 0 and 1, to represent the basal expression level of a gene when no TF acts upon it. The overall GRN architecture, or A_{net} , can be expressed by $AM_{n \times n}$, with $LG_{2 \times n}$ and f_0 as the features of the network kinetics, where n denotes the number of genes in the GRN. In this simplified GRN architecture, the activators or repressors of a gene work either synergistically or independently (Fig 1C). This approach greatly reduces the complexity compared to fully general network logic gates ($2N$ binary degrees of freedom per node), but leaves many degrees of freedom (N^2 trinary degrees of freedom) in the GRN topology. For example, a GRN consisting of 3 genes has 3^9 possible configurations in AM and 2^6 in LG, and therefore $3^9 \cdot 2^6 = 1,259,712$ possible A_{net} . Also, the f_0 , bounded between 0 and 1, can vary independently of the A_{net} , creating a larger inverse problem. The qualitative character of the GRN architecture will systematically be made quantitative, as we describe in subsequent sections.

Overview of the GRN dynamic system

A list of symbols and parameters used in the GRN dynamic system is given in Table 1. We assume that the diffusion and binding processes of TFs, which happen much faster than transcription and translation, are instantaneous. We assign a V_{max} and a V_{min} to each gene, which represent the potential highest and lowest production rates of mRNA transcripts, respectively. We assume linear decay of the proteins and mRNA, and linear translation. These assumptions give rise to Eqs 1 and 2:

$$\begin{aligned} \frac{d[R]_i}{dt} &= V_{i,min} + (V_{i,max} - V_{i,min}) \cdot f_{A_{net}}([P]_{*,i}, \Theta) - D_{mRNA} \cdot [R]_i \\ \Theta &= \{f_0, T, k\} \\ A_{net} &= \{AM, LG\}, \end{aligned} \quad (1)$$

Table 1. Parameter table for the GRN dynamic system.

Symbol	Description	Unit
$[R]_i$	number of mRNA transcripts for each gene	dimensionless
$[P]_i$	number of protein copies encoded by each gene	dimensionless
$V_{i,max}$	maximal rate of transcription for each promoter	nucleotides/second
$V_{i,min}$	minimal rate of transcription for each promoter	nucleotides/second
$f_{0,i}$	basal transcription rate for each promoter: a percentage of the $V_{i,max}$	percentage
$V_{i,trl}$	rate of translation for each protein	amino acids/second
$T_{i,j}$	also known as K_A , the protein abundance producing half occupation	dimensionless
k_i	Hill coefficient	dimensionless
A_{net}	the architecture of a GRN, including the adjacency matrix and the protein coordination matrix	dimensionless
AM	the adjacency matrix of a GRN	dimensionless
LG	the protein coordination matrix of a GRN	dimensionless
$D_{i,mRNA}$	rate of degradation for each mRNA	1/second
$D_{i,protein}$	rate of degradation for each protein	1/second
$I_{n \times m}$	the input matrix that contains m steady-state transcription profiles in the length of n genes	dimensionless

<https://doi.org/10.1371/journal.pcbi.1010991.t001>

$$\frac{d[P]_i}{dt} = V_{trl} \cdot [R]_i - D_{protein} \cdot [P]_i. \tag{2}$$

In these equations, $[P]_{*,i}$ represents the concentrations of the TFs that regulate the i^{th} gene in the GRN, and $f_{A_{net}}$ is the regulation function, which describes how TFs regulate a gene. A_{net} embodies the structure of a GRN, encompassing the adjacency matrix (AM) and the protein coordination matrix (LG) that govern its behavior. Other symbols in Eqs 1 and 2 are defined in Table 1. The regulation function $f_{A_{net}}$ is a continuous function given in Eq 9. Typically, Hill functions [57–59] and sigmoid functions [60, 61] form the building blocks of the regulation function. We use Hill functions (Eqs 3 and 4) to formulate the regulation function $f_{A_{net}}$:

$$S_{act,i}([P]_i) = \frac{[P]_i^{k_i}}{[P]_i^{k_i} + T_i^{k_i}}, \tag{3}$$

$$S_{rep,i}([P]_i) = 1 - S_{act,i}([P]_i). \tag{4}$$

In Eqs 3 and 4, T is the protein abundance producing half occupation (disassociation constant of the TF binding) and k_i is the Hill coefficient (effective cooperativity). Eq 3 describes the regulatory contribution of an activator, and Eq 4 serves the same role for a repressor. The effect of multiple TFs on the transcription rate, as shown in Fig 2A and 2B, is determined by the choice of LG according to the combinatorial Hill functions equations below (Eqs 5–8):

$$C_{IA,i}([P]_i) = 1 - \prod_{i=1}^{numTF} (1 - S_{act,i}([P]_i)), \tag{5}$$

$$C_{IR,i}([P]_i) = \prod_{i=1}^{numTF} (1 - S_{act,i}([P]_i)), \tag{6}$$

$$C_{SA,i}([P]_i) = \prod_{i=1}^{numTF} S_{act,i}([P]_i), \tag{7}$$

$$C_{SR,i}([P]_i) = 1 - \prod_{i=1}^{numTF} S_{act,i}([P]_i). \tag{8}$$

$C_{IA,i}$, $C_{IR,i}$, $C_{SA,i}$ and $C_{SR,i}$ denote the combinatorial Hill functions for independent activators, independent repressors, synergistic activators, and synergistic repressors, respectively.

With the A_{net} and the abundances of the TFs, we define the regulation function $f_{A_{net}}$ by Eq 9, which applies combinatorial Hill function formulas and f_0 to represent the gene regulations:

$$f_{A_{net}}([P]_{*,i}, \Theta) = f_0 + f_0 \cdot (C_A - 1) \cdot (1 - C_R) + (1 - f_0) \cdot C_A \cdot C_R. \tag{9}$$

Here, $[P]_{*,i}$ denotes the effective abundance of the TFs regulating the i^{th} gene. For notational convenience, we define $C_{A,i} = C_{SA,i}$ if activators are synergistic and $C_{A,i} = C_{IA,i}$ if they are independent, and similarly for $C_{R,i}$. How the parameters affect the shape of $f_{A_{net}}$ is shown in Fig 2C and 2D. The shapes of the regulation function $f_{A_{net}}$ for two activators/repressors in the dependent versus independent cases are shown in Fig A in S1 Text.

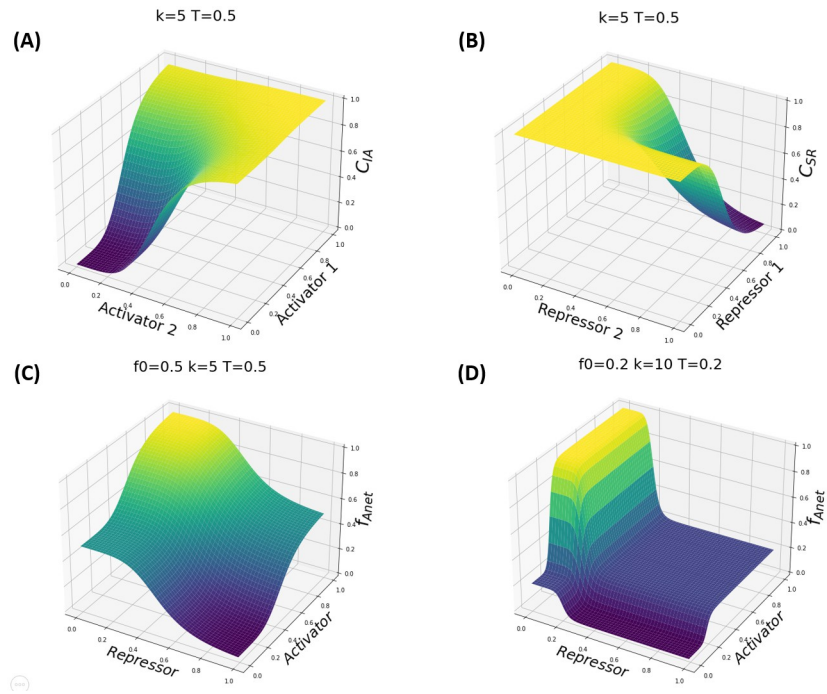


Fig 2. Demonstration of the combinatorial Hill functions (A and B), and the regulation function $f_{A_{net}}$ (C and D) under the regulation of an activator and a repressor. In the top two panels (A and B), two combinatorial Hill functions are plotted; in (A) two activators work independently to activate a target gene, while in (B), two repressors work synergistically. In the bottom two panels (C and D), the dependence of the regulatory function $f_{A_{net}}$ on the activating and repressing combinatorial Hill functions is plotted for two example cases. In (C), $f_{A_{net}}$ achieves the basal transcription rate fraction of 0.5 when there is a lack of both activator and repressor, or when both are present. Activation (resp. inhibition) occurs when the activator (resp. repressor) is abundant, and the repressor (resp. activator) is scarce. In (D), The Hill coefficient, k , determines the steepness of the regulation function; The basal expression level, f_0 , controls the position of the middle plane and can slide between 0 and 1. The threshold T decides the TF abundance that will trigger the activation or repression.

<https://doi.org/10.1371/journal.pcbi.1010991.g002>

GRN architecture inference

With the deterministic GRN dynamic model constructed above, we propose to infer the A_{net} using experimentally-derived transcriptional profiles and mRNA production rates. Specifically, we consider transcriptional profiles of cells in the exponential growth phase under defined and mixed culture conditions. We therefore assume that the resulting transcriptional profiles represent a steady-state transcriptional output of the GRN. By incorporating experimentally determined transcription, translation, and degradation rates, we simulate the GRN dynamics and determine whether a given A_{net} can accurately reproduce the observed attractors. To search for the optimal A_{net} for a particular GRN, we utilize a modified evolutionary algorithm [62, 63] to iteratively refine the A_{net} parameters until the predicted network attractors converge upon the experimentally measured ones. The main step-by-step processes of our iterative computational and experimental strategy are presented in Fig 3.

Since A_{net} and the values of some system parameters are often unknown in practice, we will make use of the measurable kinetic parameters (including V_{max} , V_{trb} , D_{mRNA} , and $D_{protein}$) and the steady-state transcriptional profiles to estimate the unknown parameters (V_{min} , T , k , and f_0) for a given network architecture (See Table 1). First, the V_{min} value for a gene is estimated by the minimal expression level of the gene across all the samples. Second, without other prior knowledge, we must assume that when the genes are under TF regulation, they have the same

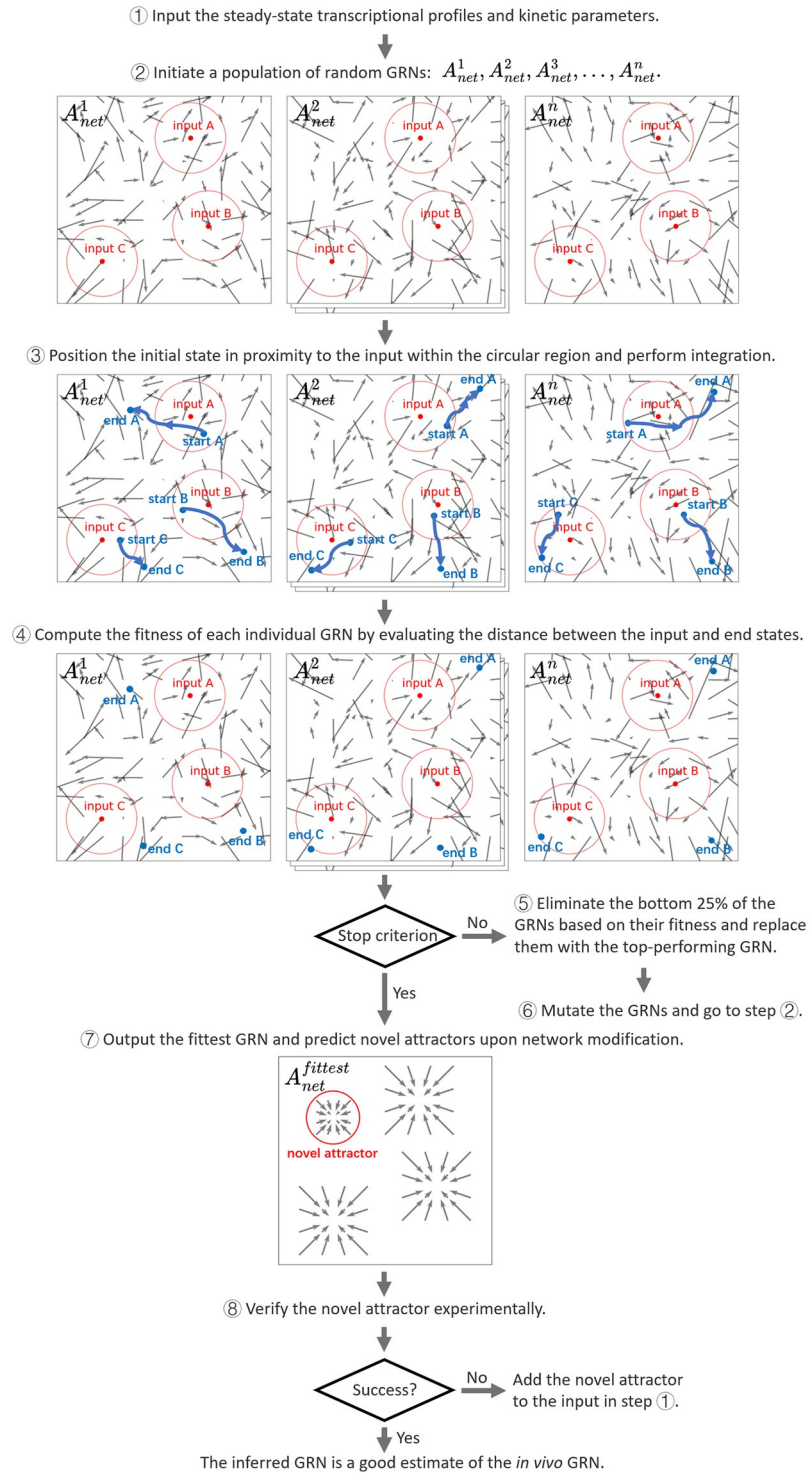


Fig 3. A flowchart illustrating the step-by-step processes of our iterative computational and experimental strategy to infer GRNs and predict novel attractors.

<https://doi.org/10.1371/journal.pcbi.1010991.g003>

chance to be activated or inhibited. Hence, the T of the TFs is calculated by their average expression levels using Eq 2. Third, we assume that the input transcription profiles include fully activated genes, and estimate k by invoking the steady state assumption for transcriptional profiles that include maximal expression. The last unknown parameter, f_0 , is obtained by solving Eq 1, assuming the inputs are at steady state and averaging across inputs. The equations for solving these parameters can be found in Eqs A-G in S1 Text. Our framework can incorporate more advanced parameter estimates if they exist, but in the absence of direct measurements, we rely on the available data to assign reasonable values to unknown parameters. The assumptions made are designed to prevent bias and select values that are unlikely to produce qualitatively different dynamics. Therefore, these parameters should be interpreted as qualitatively reasonable rather than quantitatively definitive.

The modified evolutionary algorithm that we developed to efficiently search the GRN architecture space is illustrated in Algorithm 1. First, we randomly create a population of A_{net} , each of which has the same initial fitness score. Second, we update the estimates of the unmeasured parameters for each A_{net} . If an A_{net} has a f_0 less than 0 or greater than 1, a penalty will be added to the fitness score for A_{net} . Third, we examine each independent self-activating edge in an A_{net} . If any input with the self-activating gene does not have a companion input state with that gene active, we construct such a state for use as an initial condition to test whether such a state will converge into a measured attractor. When the system is initiated in such a state, the input transcriptional profile suggests that the self-activating gene should eventually be inactivated by other TFs. By including these additional initial conditions, we penalize networks that are highly stable for all or nearly all initial states. Next, we add a mild perturbation to each of the initial states and numerically integrate the deterministic GRN dynamical system using the Runge-Kutta 4th order method [64, 65] to find the final steady states. Since we assume that the measured transcriptional profiles represent attractors, these initial states should converge to the corresponding attractor states. If the system state ends up far away from the attractors, the current A_{net} cannot generate the anticipated attractors. Although oscillations are possible in the GRN dynamic system, we cannot assess them because the input transcriptional profiles are at steady states. As a result, a fixed and exceedingly high penalty is imposed on each A_{net} that generates oscillatory behavior. We use a normalized Manhattan distance between the attractors and the final states as a metric:

$$AttractorDistance_j = \frac{1}{n} \cdot \sum_{i=1}^n \frac{|state_{i,j} - I_{i,j}|}{\max(I_{i,*})}, \quad (10)$$

where $state_{i,j}$ is the i^{th} gene expression level of the j^{th} final state obtained by running the GRN dynamic system, $I_{i,j}$ is the i^{th} value of the j^{th} input state, m is the total number of attractors, and n is the total number of genes.

Autoregulation tends to produce excessive attractors not belonging to the input. We compensate for this by considering additional initial states. If one of these additional initial states belongs to the input, the final state will be compared against the initial state. Otherwise, the final state will be compared against the closest attractor in the input. If the initial condition does not converge to a steady state, we apply a penalty to the fitness of the network. We define the fitness of each A_{net} in the current population by the reciprocal of the minimal average “AttractorDistance” for all initial conditions, subtracting off the penalties for steady state convergence failure and for unrealistic estimation of f_0 .

After the attractor distance is calculated, we create the next generation population by randomly mutating each A_{net} by a certain Hamming distance, and numerically solve the new ODEs constructed by the mutated A_{net} to obtain the new minimal average

“AttractorDistance”. If the new minimal average “AttractorDistance” is less, we keep the mutated population. Otherwise, the mutated population will be abandoned, and we will return to the former population. Finally, we update the fitness of each A_{net} according to their “AttractorDistance”. We sort the population by fitness in descending order and then eliminate the last 20% of individuals and duplicate the fittest 20% to restore the population number. By iteratively running the algorithm, we can obtain the fittest A_{net} in the last generation as the output. Alternative distance metrics such as Euclidean distance and Jaccard distance exist, but we chose Hamming distance due to its simplicity, intuitive interpretation, and efficiency for equally sized strings. Moreover, the Hamming distance is less sensitive to outliers than the Euclidean distance, ensuring that a single poorly fit attractor does not unduly influence the algorithm. In this study, unless explicitly specified, the default setting for the population size of A_{net} in the evolutionary algorithm was 100, and the algorithm was run for 800 generations. The probability distribution used to mutate the A_{net} was a uniform distribution (i.e. the probabilities for an edge to be mutated into 0, 1, and -1 were equal). The population size and the number of generations were excessively high for the *in silico* and *in vivo* tests, and therefore the results were not sensitive to these parameters unless they were set too low (≤ 50 for the population size and ≤ 400 for the number of generations).

Because the sampling and mutating steps in the algorithm are stochastic, the output A_{net} can be different each time. We draw a consensus GRN architecture using Eq H in [S1 Text](#) with the assumption that a particular regulatory connection (i.e., an entry in A_{net}) occurring at a higher frequency among a group of inferred GRN architectures is more significant than one occurring at a lower frequency. In all results presented here, unless otherwise specified, we conducted 30 independent and identical GRN inference processes and obtained a consensus GRN architecture by averaging the fittest A_{net} architectures, weighting each by its fitness.

In addition to regular transcriptional profiles, EA can also incorporate data from genetically engineered strains. For instance, if the input transcriptional profiles are from knockout or overexpression strains, we can set the mRNA abundance of the inoperative genes to zero or maximal expression level during the numerical integration. If a regulatory connection is removed by genetic engineering (e.g., by disrupting a TF binding site in a promoter), we can fix the corresponding entry in A_{net} as zero to eliminate the effect of the disrupted regulatory connection. Furthermore, we can use genome-wide binding data, such as chromatin immunoprecipitation (ChIP) data, to guide the mutation of A_{net} during inference; if a physical binding interaction exists between a TF and a gene, it is likely to represent a regulatory connection. Therefore, we can lower the probability of the corresponding entry being mutated to zero. Conversely, if the ChIP data indicates that a network of interest is sparse, we can incorporate a mutation step with an 80% probability of the selected edges being mutated to 0. The accommodation of different data types allows the model to integrate more available data and perform better. Although EA can enforce network sparsity in the network initiation and mutation steps, we do not use any sparsity regularization in this study because we are not focusing on general sparse biological networks but instead on regulatory networks among a small group of closely related TFs. Such networks can be much denser than large GRNs among less closely related TFs.

Algorithm 1 Evolutionary Algorithm

- 1: Randomly initiate a population $\mathbf{A}_{net}^0 : \{A_{net}^1, A_{net}^2, \dots, A_{net}^N\}$;
- 2: Each individual has an initial fitness: $\frac{1}{N}$
- 3: Calculate f_0 for each A_{net}
- 4: **for** i in $1 : m$ **do**

```

5:  initial system state  $\leftarrow I_{n,i}$ 
6:  update  $f_0$  for each  $A_{net}$  and add penalty when  $f_0 \notin [0, 1]$ 
7:  if  $A_{net}$  has an independent self activating edge and the initial
    system state in which the self activating gene has been set to its
    maximal expression  $\notin I_{n,m}$  then
8:    append the modified state described above to  $I_{n,m}$ 
9:  end if
10:  initial system state  $\leftarrow$  initial system state + mild random
    perturbation
11:  for each  $A_{net}^j$  do
12:    integrate GRN ODEs with  $A_{net}^j : \{AM_j, LG_j, f_{0j}\}$ 
13:  end for
14:  record last_states $_{i,j}$ 
15: end for
16: Calculate  $\min(\text{AttractorDistance}_{A_{net}^0})$ 
17: for generation from 1 to x do
18:  for each  $A_{net}^j$  in  $A_{net}^0$  do
19:     $AM_j^t \leftarrow$  mutate  $AM_j$  by x Hamming distance
20:     $LG_j^t \leftarrow$  mutate  $LG_j$  by x Hamming distance
21:  end for
22:  for i in 1: m do
23:    initial system state  $\leftarrow I_{n,i}$ 
24:    update  $f_0$  for each  $A_{net}$  and add penalty when  $f_0 \notin [0, 1]$ 
25:    if  $A_{net}$  has an independent self activating edge and the initial
        system state in which the self activating gene has been set to
        its maximal expression  $\notin I_{n,m}$  then
26:      append the modified state described above to  $I_{n,m}$ 
27:    end if
28:    initial system state  $\leftarrow$  initial system state + mild random
        perturbation
29:    for each  $A_{net}^j$  in  $A_{net}^t$  do
30:      integrate GRN ODEs with  $A_{net}^j : \{AM_j^t, LG_j^t, f_{0j}^t\}$ 
31:    end for
32:    record last_states $_{i,j}$ 
33:    if last_states $_{i,j}$  is a fixed point attractor then
34:      record Attractor $_{i,j}$ 
35:    else
36:      add a penalty to  $A_{net}^j$ 
37:    end if
38:  end for
39:  Calculate  $\min(\text{AttractorDistance}_{A_{net}^t})$ 
40:  if  $\min(\text{AttractorDistance}_{A_{net}^t}) < \min(\text{AttractorDistance}_{A_{net}^0})$  then
41:     $A_{net}^0 \leftarrow A_{net}^t$ 
42:  else
43:     $A_{net}^0$  remains
44:  end if
45:  for each  $A_{net}^j$  in  $A_{net}^0$  do
46:    fitness $_{A_{net}^j} = \text{fitness}_{A_{net}^j} + \frac{1}{\text{AttractorDistance}_{A_{net}^j}}$ 
47:  end for
48:  Sort  $A_{net}^0$  by fitness in descending order:  $\{A_{net}^{(1)}, A_{net}^{(2)}, \dots, A_{net}^{(N)}\}$ 
49:   $A_{net}^0 \leftarrow \{A_{net}^{(1)}, A_{net}^{(2)}, \dots, A_{net}^{(0.8N)}, A_{net}^{(1)}, \dots, A_{net}^{(1)}\}$ 
50: end for
51: Return the  $A_{net}^{(1)}$  in  $A_{net}^0$ 

```

Model networks for validation

It is extremely challenging to directly determine the complete and comprehensive composition and structure of “real-world” GRNs in living organisms [18, 66]. Therefore, the use of experimental data in GRN inference can be problematic when it comes to validating the outcome of GRN model predictions, since one can rarely, if ever, be certain that the experimental data provides a complete picture of the real-world GRN structure. For this reason, it has become common practice in the field of GRN inference to utilize *in silico* (i.e., computer generated) datasets for method validation, which can provide gene expression data that is directly predicted based on a hypothetical “source” GRN model [57–59, 67].

We used both *in silico* and biologically observed GRN instances to evaluate EA. The *in silico* instance consists of five arbitrarily generated toy GRN ODEs (Fig 4A), each of which has at least 9 different fixed-point attractors and no oscillations. We believe this selection for multiple fixed-point attractors and no oscillations led to the emergence of a large number of autoregulatory interactions. The A_{net} architectures for these systems are regarded as the reference GRN architectures, which will be used as answer keys against which the inferred GRN architectures will be compared. The *in silico* fixed-point attractors for each A_{net} are generated by SynTReN, a commonly used benchmark generator for GRN inference [57]. SynTReN employs an alternative ODE modeling framework, distinct from the one we utilize, to generate gene expression

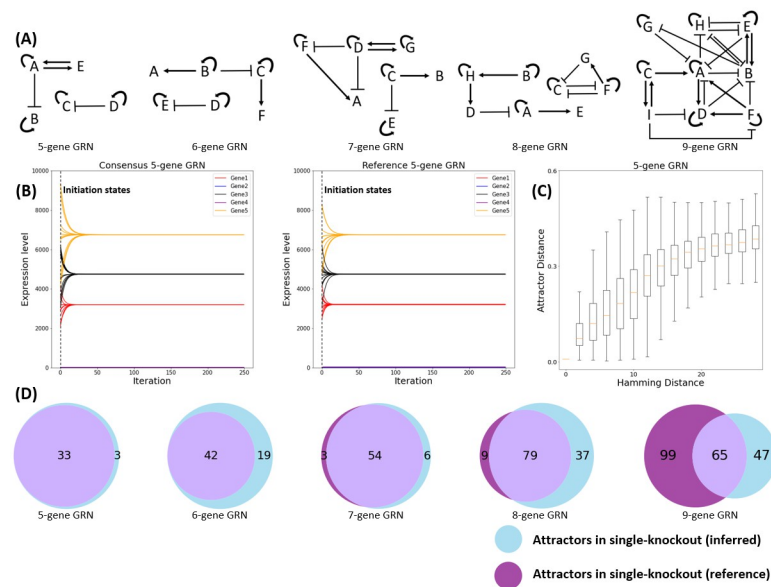


Fig 4. (A) Five GRN architectures were arbitrarily generated as references in the *in silico* test. They have 5–9 genes and at least 9 different fixed-point attractors and no oscillations. The pointed and blunt arrows represent activating and repressing regulatory interactions, respectively. (B) GRN dynamics when initiated near a fixed-point attractor of a reference GRN. The consensus GRN was inferred by the attractors of the 5-gene *in silico* reference GRN. The initial states were obtained by the attractor position plus a uniform distributed random variable by Eq K in S1 Text. The perturbation power was set to 0.1. The dynamics showed similarly good agreement in other reference GRNs. (C) Positive correlation between A_{net} similarity (Hamming distance on the horizontal axis) and attractor profiles similarity (attractor distance on the vertical axis). Each column in the box plot contains 1000 random A_{net}^{mut} mutated from the A_{net}^{ref} consisting of 5 genes. Similar strong correlation has been observed in all other reference GRNs (see Fig C in S1 Text). (D) *in silico* attractors prediction result summary. Two attractors considered matched have an attractor distance less than 0.16 (a cutoff below which a simple null model has a less than 5% chance of producing matched attractors; see Table C in S1 Text). Overall, the single-knockout reference GRNs produced 384 fixed-point attractors and the single-knockout inferred GRNs produced 385. Of these attractors, 273 were matched. No attractors were matched in a random GRN.

<https://doi.org/10.1371/journal.pcbi.1010991.g004>

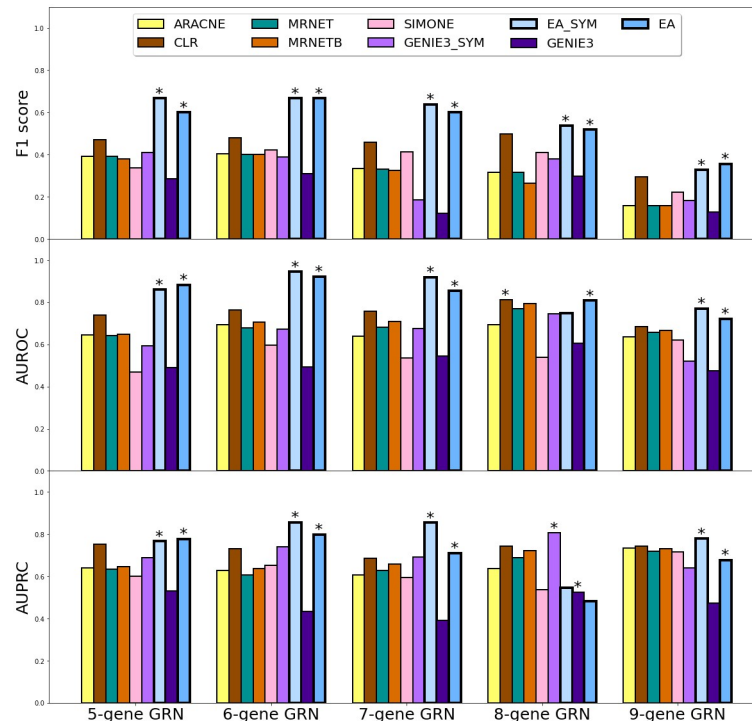


Fig 5. The *in silico* test comparison result in F1 score (upper panel), AUROC (middle panel), and AUPRC (bottom panel). The F1 scores are calculated using a threshold cutoff of 0.5 for all models. Best performances are marked by asterisks for symmetric and asymmetric methods.

<https://doi.org/10.1371/journal.pcbi.1010991.g005>

data using the network structure and initial conditions (Eq J in S1 Text). This is done so that the results (Fig 5) are not biased by generating state data using the ODE framework we aim to evaluate. Considering that noise generally exists in the experimentally derived transcriptional profiles, we added a Gaussian distributed noise with a standard deviation of 0.2 to the *in silico* input transcription profiles. The signal to noise ratio was approximately 14 dB. The kinetic parameters used for the *in silico* tests were identical to the parameters used to generate the simulated data, and they were assigned by the values experimentally measured in *Escherichia coli* [68–71] (See Table A in S1 Text).

To test our model against an *in vivo* GRN instance, we used experimental data derived from a synthetic GRN engineered in *Saccharomyces cerevisiae* by Cantone et al. [56]. Consisting of 5 genes and a variety of regulatory interactions (Fig 6A and Table 2), the GRN can switch amongst 10 distinct stable states in response to the overexpression of each individual gene in two different carbon sources, galactose and glucose. These stable states were measured by quantitative PCR (qPCR) and converted to absolute expression levels. The promoter strengths, which indicate the rates of transcription initiation for each gene in the GRN, have been estimated by a stochastic optimization algorithm from steady-state gene expression data measured by qPCR [56]. Other kinetic parameters used in the *in vivo* test are provided by Table A in S1 Text. The A_{net} transcription profiles, and kinetic parameters for the *in silico* and *in vivo* tests can be found in S1 and S2 Data, and Table A in S1 Text.

We also tested our model using transcriptional profiles derived from a set of 12 wild-type and targeted TF deletion strains of *Candida albicans*. All strains used in this study are described in Table B in S1 Text and are derived from SN156, which is a commonly used derivative of the SC5314 strain that is used widely in *C. albicans* studies [72, 73]. All of the *C.*

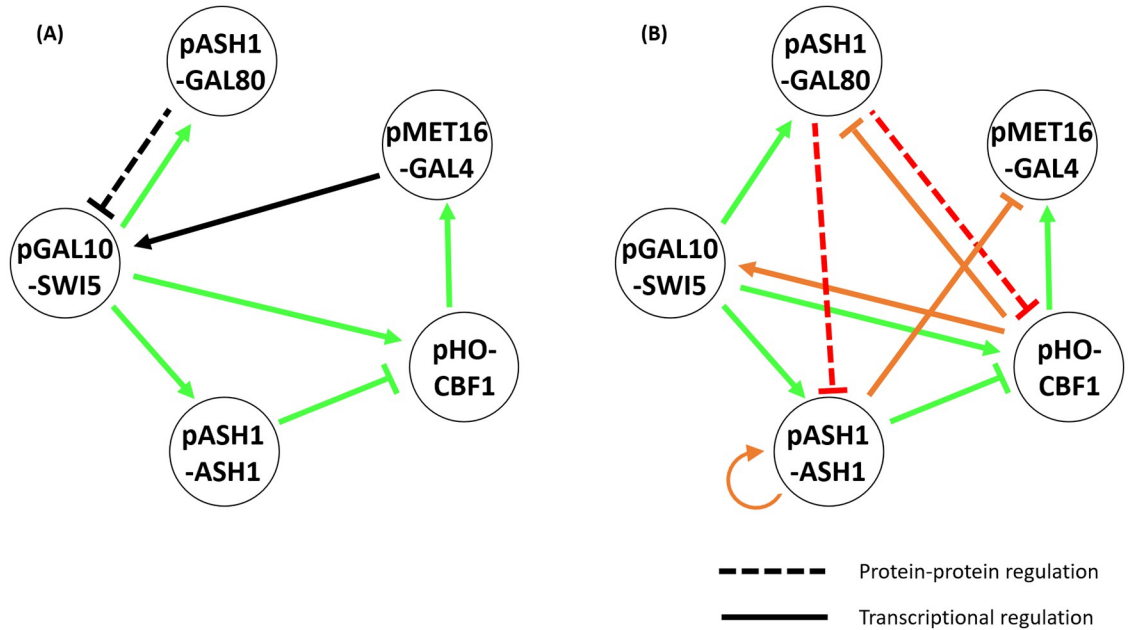


Fig 6. (A) The schematic diagram of the *S. cerevisiae* synthetic circuit. Solid lines represent direct transcriptional regulation and dotted lines indicate indirect transcriptional regulation mediated by a protein-level activation or inhibition of a transcription factor. Edges accurately inferred by EA are labeled in green, otherwise in black. Gal80 protein can inhibit *SWI5* transcription by preventing Gal4-mediated activation of target genes in the absence of galactose. Modified from the original paper [56]. (B) The schematic diagram of the inferred circuit. Additional inferred edges that are not present in the original design but are supported by previously published literature are labeled in orange. Inferred inhibitory edges indicated in red represent putative mechanisms for indirect repression of *SWI5* by Gal80, but are not supported by known mechanisms of Gal80 function as discussed in the text.

<https://doi.org/10.1371/journal.pcbi.1010991.g006>

albicans single TF deletion strains used in this study were reported previously [73]. TF double deletion strains were generated using CRISPR-mediated genome editing to delete the *WOR1* coding sequence as described by Nguyen et al [74]. Steady-state transcript levels were determined using the 3' quant seq RNA sequencing methodology as described by Moll et al [75]. Briefly, *C. albicans* cells were harvested from mid to late-log cultures and total RNA was isolated using the RiboPure RNA Purification kit. cDNA libraries were prepared using the QuantSeq 3' mRNA-Seq Library Prep kit from Lexogen and multiplexed in pools of 96 libraries. Single-end 100bp reads were obtained using an Illumina HiSeq4000 instrument. The resulting de-multiplexed sequencing reads were trimmed and aligned using STAR Aligner [76] to

Table 2. Experimental evidence for regulatory associations in the synthetic circuit.

	Cbf1	Ash1	Gal4	Gal80	Swi5
pHO_CBF1	Inhibition [83]	Inhibition [84, 85]			Activation [86, 87]
pGAL10_SWI5	Activation [83]	[82]	Activation [88, 89]	Inhibition [90]	
pMET16_GAL4	Activation [83, 91]	[82]			
pASH1_GAL80	Activation [83]	[82]			Activation [84]
pASH1_ASH1	Activation [83]	[82]			Activation [84]

Column names are the TF proteins and row names are the promoters followed by their open reading frames. Orange shaded boxes indicate potential regulatory associations found in YEASTRACT+; blue shaded boxes are experimental evidence found by microarray and/or Northern blot experiments; pink shaded boxes are experimental evidence found by both microarray/Northern blot and ChIP/EMSA experiments.

<https://doi.org/10.1371/journal.pcbi.1010991.t002>

obtain raw read counts for each transcript genome-wide. The promoter strengths of each gene in the network were determined using capped small RNA (csRNA) sequencing [77]. This method enables the isolation of short nascent mRNA transcripts, rather than full-length mRNAs, and thus provides an instantaneous snapshot of the level of transcriptional activity at each transcriptional start site, genome-wide. Briefly, we enriched for nascent small, capped, RNA molecules from total RNA extracted from mid-log phase *C. albicans* cultures and prepared sequencing ready libraries using the small RNA library preparation kit from New England Biolabs. The resulting libraries were multiplexed and 16 indexed libraries were pooled prior to sequencing on an Illumina HiSeq4000 instrument. Sequencing data were analyzed using HOMER [77]. The mRNA and csRNA sequencing data can be accessed on GEO (GSE217461 and GSE217383). Our evolutionary algorithm, the datasets used in this study, and the results are available on a GitHub repository at <https://github.com/UCM-RuihaoLi/GeneRegulatoryNetworkInference>. We have also used Zenodo to assign a DOI to the repository: <https://zenodo.org/record/7553611>.

To account for noise in the experimentally derived transcriptional profiles we measured the “average replicate distance” which describes the average pairwise attractor distance between each of the three biological replicates for each genotypic/phenotypic combination. This metric was then used to determine whether a given GRN model prediction was considered successful, with the basic premise that a successful GRN prediction should yield a transcriptional profile that lies within the average noise range of 60% (Table 3) in the experimentally derived

Table 3. *C. albicans* strains transcriptional profiles prediction results.

Genotypes	Phenotype	Prediction trial 1	Prediction trial 2	Average replicate distance (noise range)	Control
Wildtype and single TF deletion strains					
wildtype	White	0.087	0.084	0.130 (42%)	0.329(57%)
wildtype	Opaque	–	0.506	0.231 (107%)	0.352(59%)
Δ/Δ <i>wor1</i>	White	0.119	0.131	0.156 (58%)	0.281(50%)
Δ/Δ <i>wor2</i>	White	0.114	0.093	0.252 (89%)	0.295(60%)
Δ/Δ <i>wor3</i>	White	0.146	0.144	0.152 (49%)	0.310(47%)
Δ/Δ <i>wor3</i>	Opaque	–	0.319	0.463 (144%)	0.286(54%)
Δ/Δ <i>wor4</i>	White	0.128	0.120	0.107 (42%)	0.310(53%)
Δ/Δ <i>efg1</i>	White	0.198	0.182	0.227 (78%)	0.314(56%)
Δ/Δ <i>efg1</i>	Opaque	–	–	0.120 (42%)	0.303(57%)
Δ/Δ <i>ahr1</i>	White	0.153	0.164	0.229 (72%)	0.319(62%)
Δ/Δ <i>ahr1</i>	Opaque	0.269	0.254	0.204 (94%)	0.279(55%)
Δ/Δ <i>czf1</i>	White	0.150	0.166	0.237 (81%)	0.340(45%)
Δ/Δ <i>czf1</i>	Opaque	0.327	0.353	0.282 (100%)	0.287(60%)
Δ/Δ <i>ssn6</i>	Opaque	0.254	0.366	0.182 (89%)	0.314(98%)
Δ/Δ <i>rbf1</i>	Opaque	0.136	0.139	0.334 (168%)	0.295(46%)
Double TF deletion strains					
Δ/Δ <i>ssn6</i> Δ/Δ <i>wor1</i>	White	0.094	0.097	0.051 (20%)	0.263(99%)
Δ/Δ <i>rbf1</i> Δ/Δ <i>wor1</i>	White	0.154	0.155	0.118 (49%)	0.282(54%)

The transcriptional profile(s) of a specific genotype was left out in each prediction. Predictions whose attractor distances are no greater than 0.16 or the average replicates distance of the experimental data are indicated by green shaded boxes. The average replicate distance for the biological replicate samples is indicated by the decimal values while the corresponding noise range is indicated in parentheses as percentages. Predictions that show no attractor (represented by dashes) or attractors exceeding the cutoff are indicated by unshaded boxes. Highly variable samples are indicated by gray shaded boxes. For the control, the average attractor distances of random GRNs that produced fixed-point attractors are indicated by the decimal values, while the percentage of random GRNs that produced fixed-point attractors is indicated in parentheses.

<https://doi.org/10.1371/journal.pcbi.1010991.t003>

transcriptional profiles. Since some of the experimental replicates had high variance, we also included an attractor distance threshold of 0.16. This threshold was selected based on the performance of a null model, which samples from a uniform distribution with the upper and lower limits as the maximal and minimal expression levels for each gene. For transcriptional profiles with five genes or more, the null model has a 5% chance or less to generate a profile below this cutoff of 0.16 (See Table C in [S1 Text](#)).

Results

Consensus GRNs converge upon attractors of reference GRNs

We propose that a successfully inferred GRN should accurately reproduce the experimentally derived attractor states from which it was derived. Therefore, we initiated dynamic simulations using the inferred consensus GRN and the reference GRN at random initial states around their attractor positions. Specifically, a uniformly-distributed perturbation was added to the initial states by Eq K in [S1 Text](#). The magnitude of perturbation was determined by a user-adjustable parameter known as “perturbation power”. Without prior knowledge of the range of the actual basin of attraction, the estimation for the perturbation power should be conservative. In our implementation, the default value for the perturbation power was set to 0.1. As anticipated, the results showed that their expression levels converged upon the same attractor ([Fig 4B](#)).

GRN architectures and attractor profiles are strongly coupled

One key hypothesis of EA is that a GRN’s architecture can be revealed by its attractors. To test this hypothesis under normal circumstances, we arbitrarily generated five *in silico* GRN architectures as test subjects ([Fig 4A](#)) and examined the correlation between their A_{net} and attractor profiles. Specifically, we randomly mutated the A_{net} of the reference GRN architectures and observed how their attractor profiles change accordingly. The difference between the mutant GRN architectures ($A_{net}^{mut} = \{AM^{mut}, LG^{mut}\}$) and the reference architectures ($A_{net}^{ref} = \{AM^{ref}, LG^{ref}\}$) is measured by the Hamming distance, and the difference between their attractor profiles is measured by the attractor distance given by [Eq 10](#). As shown in [Fig 4C](#), when the A_{net}^{mut} becomes more different from A_{net}^{ref} , its attractor profiles tend to be further away from the reference. This general trend between A_{net} and attractor profiles justifies the search strategy of our algorithm, whereby the A_{net}^{ref} is inferred by gradually mutating A_{net} and improving the distance between the population’s attractor profiles and those of the reference GRN. Based on our observations, network mutations tend to be more efficient when considering the slopes between the Hamming distance and the attractor distance of these five *in silico* GRNs, whereas crossover operations appear to be more suitable for cases with smaller slopes. The significant fluctuation in the y-axis (attractor distance) was anticipated to have a minimal effect on the inference process as the evolutionary algorithm gives precedence to identifying networks with smaller attractor distances. We performed sensitivity tests on key kinetic parameters and perturbed them by 50% to evaluate their impact on the correspondence between network structure and attractors. The results, depicted in [Fig B in S1 Text](#), showed that perturbations in V_{max} and D_{mRNA} had a more significant effect compared to V_{trl} and $D_{protein}$, potentially due to violations of the steady-state assumption. Consequently, we identified V_{max} and the attractors as essential inputs for the model, while other parameters can be estimated based on the steady-state assumption. In addition, [Fig 4C](#) shows that the attractor distance can reach zero before the Hamming distance goes to zero, indicating that identical attractors can be generated by distinct GRNs. Therefore, even though the reference

GRN can be approached by our searching strategy following the trend, it is unlikely to be eventually obtained by an individual architecture in the output, thereby motivating the examination of a consensus GRN.

Our algorithm correctly infers the GRN architecture in the *in silico* test and outperforms other models

We tested the ability of our model, and several other existing models, to infer GRN architectures using the attractor profiles generated by the five *in silico* reference GRNs depicted in Fig 4A. To avoid biasing the results in favor of our algorithm, we generated ODEs for these five topologies using SynTReN [57]. We generated simulated transcriptional profiles from the attractors of these ODEs using a global search strategy and utilized them as the input for the *in silico* test. For each instance, we used the attractors as input and ran 600 iterations in algorithm 1. We performed 30 independent and identical inference processes and obtained a consensus A_{net} by weighted averaging of edge frequencies. We compared the inferred consensus A_{net} to reference A_{net} using common machine learning metrics, including the F1 score, area under Receiver Operating Characteristic (ROC) and Precision and Recall (PR) curves. We compared our evolutionary algorithm method, to six widely used benchmark methods, namely ARACNE [25], CLR [26], GENIE3 [78], MRNET [79], MRNETB [80] and SIMONE [81]). These methods had fixed hyperparameters and did not depend on initialization. They exhibited significantly shorter runtimes compared to EA. The key factors that affect the runtime of EA include the size of the network (i.e. the number of genes), the number of input transcriptional profiles, the population size, and the ODE integration step size. The evolutionary algorithm allowed us to apply a parallel computing technique to shorten the time. While these methods only took minutes, EA took approximately 150 core-hours to infer the 5-gene network and 300 core-hours to infer the 9-gene network. This discrepancy in runtime can be attributed to EA's need to solve ODEs for each individual A_{net} in every iteration of the evolutionary algorithm to achieve stable states. Amongst these methods, only EA and GENIE3 can infer directed networks with asymmetric adjacency matrices, which can differentiate the regulating gene and the target gene (Table D in S1 Text). Therefore, we symmetrized the inferred networks of EA and GENIE3 as EA_SYM and GENIE3_SYM by making all edges undirected. Furthermore, all the examined methods, except for EA, were unable to infer self-regulatory edges. As a result, the diagonal elements in their inferred adjacency matrices were set to zero. In this evaluation, we only considered true/false positives/negatives for edges between different nodes, which ensures that the self-loops did not provide any advantage or disadvantage to any of the methods in this test. As presented by Fig 5, EA performed generally better than the other algorithms. Moreover, the protein coordination parameters and basal transcription rate parameters of EA converged well to the ground truth (see S2 Data). To account for the variation in self-regulations among different methods, we conducted an additional comparison using a set of reference GRNs that specifically excluded self-regulatory edges. (Fig D in S1 Text). This set of reference GRNs puts our algorithm at a disadvantage because other methods cannot generate false positives for autoregulation, leading to automatic correctness for the diagonal predictions. GENIE3 performed the best out of all methods we considered. Nevertheless, despite the penalty this test imposes on our EA method, it was still among the most competent ones (Fig E in S1 Text). Additionally, we observed that when the scale of the *in silico* GRN increases from five to nine genes, it becomes more difficult to infer the AM. We believe that additional attractor profiles are needed to reveal additional stable states of large-scale GRNs and to compensate the curse of dimensionality brought by its bigger state space volume.

Our algorithm can predict novel attractors produced by genetically engineered GRNs

Because the *in silico* inferred and reference GRNs are similar in both structure and associated attractor profiles, we anticipated that attractors predicted by a mutated form of the inferred GRN should closely resemble the attractors produced by an identical mutation in the true GRN. To test this hypothesis, we systematically “deleted” each of the individual genes in the five *in silico* reference GRNs and found their new attractors by searching through the state space. These new attractors are unknown to the inferred GRNs because they had not been used in the GRN inference process. We performed the same gene deletions in the inferred GRNs to see if they could accurately predict the new attractors of the mutated reference GRNs. The attractors of the mutated reference GRNs were generated by SynTReN, while the attractors of the mutated inferred GRNs were predicted by our algorithm. We applied a systematic global search to find attractors and used a random GRN as a control. We found that the single knockout reference GRNs had altogether 384 attractors (combining attractors from all knockouts across all five reference GRNs) and the single knockout inferred GRNs had a combined total of 385 attractors. Of these attractors, 273 (71.1% of the reference GRN attractors and 70.9% of the inferred GRN attractors) were matched. The random GRN showed 32 attractors, none of which were matched. See [Fig 4D](#) for a summary of these results and Tables H-M in [S1 Data](#) for the complete results.

Our algorithm revealed unintended edges in an engineered *S. cerevisiae* GRN

To examine how well the GRN dynamic model produced by our algorithm simulates experimentally derived gene expression and to what extent it is robust to measurement noise, we tested EA using experimental data derived from an engineered synthetic GRN in *S. cerevisiae* [56]. This engineered GRN consists of seven activating or inhibitory edges and five genes, some of which are under control of non-native promoters ([Fig 6A](#)). Using the experimentally measured promoter strengths and ten distinct steady state gene expression profiles, derived from strains which individually overexpress each of the five genes in galactose and glucose, our model inferred the GRN shown in [Fig 6B](#). In glucose, Gal80 blocks Gal4 from activating *SWI5*, while galactose can inactivate Gal80 and Gal4 is free to activate *SWI5*.

Our algorithm correctly identified five of the six transcriptional regulatory edges present in the original design of the engineered GRN (see comparison result in Table N in [S1 Data](#)). In addition, our algorithm predicted two additional edges related to protein-protein interactions and four that were not intended in the original design of the engineered GRN, but for which there is experimental evidence in the literature (see [Table 2](#)). We believe the missing transcriptional regulation of *SWI5* by Gal4 can be explained as follows. First, as shown in [Fig 4C](#), the same set of attractors can be produced by different GRNs. In this case, the activation of *SWI5* by a feedback loop via *CBF1* and *GAL4* is replaced by a more direct activation by *CBF1* only. Second, the difference in the regulatory effects of Gal80 is related to how protein-protein interactions are encoded in our ODE framework. Special care must be taken in interpreting protein-protein interactions in the context of the inferred network produced by our algorithm. Our algorithm does not incorporate explicit protein-protein interactions, such as the interaction between Gal80 and Gal4, which leads to the downregulation of *SWI5* and furthermore *ASH1* and *CBF1*. Thus, in our inferred network, the inhibitory edge from *GAL80* to *SWI5* is not present. Instead, this protein inhibition is incorporated into the regulatory function for the targets of Swi5. Specifically, the Swi5 protein activates *CBF1* and *ASH1* transcription, but the protein Gal80 interferes with this activation. Therefore, at the mRNA level, increased *GAL80* transcription does not directly decrease *SWI5* mRNA production; rather it decreases *ASH1*

and *CBF1* transcription. Thus, the inhibitory effect of Gal80 on the Swi5 protein is represented as the two inhibitory edges from *GAL80* to *ASH1* and *CBF1*. With this consideration in mind, only the self-activation of *ASH1*, the inhibition of *GAL4* by *ASH1*, the activation of *SWI5* by *CBF1*, and the inhibition of *GAL80* by *CBF1* represent regulatory effects that are not present in the intended synthetic system. We found previous experimental evidence for all these interactions in the literature (see [Table 2](#)).

Furthermore, while investigating the source of these additional edges, we observed that certain elements of the experimentally derived transcriptional profiles did not appear to be consistent with the intended design of the engineered GRN as described by Cantone et al. [56]. Specifically, Cbf1 was intended to serve as the sole activator of *GAL4*, which in turn was meant to serve as the sole activator of *SWI5*. This would imply that, at steady state, *SWI5* should be expressed if and only if Cbf1 is elevated and *GAL4* is expressed. The experimentally derived transcriptional profiles contradict this. They indicated that at steady state, *SWI5* was activated even when *GAL4* was not expressed. Cantone et al. [56] argued that *GAL4* is transiently expressed during an early phase of the experimental protocol, and that the Gal4 protein could persist to activate *SWI5* even after *GAL4* mRNA levels drop. However, this argument contradicts the steady state assumption of the transcriptional data and furthermore does not explain why *GAL4* mRNA levels were low when *CBF1*, which was intended to activate *GAL4*, was overexpressed.

We speculate that these discrepancies between the intended engineered GRN and the experimentally derived data may be explained by unintended regulatory interactions that modify the GRN structure and dynamics. By performing a systematic search on each TF-promoter pair in the intended engineered GRN using YEASTRACT+ [82], we uncovered support for this hypothesis. Specifically, we found experimental evidence from microarray, Northern blot, ChIP, and electrophoretic mobility shift assay (EMSA) experiments, supporting the idea that Cbf1 and Ash1 proteins regulate more than their intended target genes in the circuit. In fact, all four promoters within the circuit can be responsive to Cbf1 and Ash1 ([Table 2](#)).

Our inferred GRN predicted additional regulatory interactions beyond those that were intended in the synthetic regulatory network [56], and we identified experimental support for these putative regulatory interactions ([Table 2](#)). We conclude that the inferred GRN may have identified actual regulatory associations that impacted the experimentally derived transcriptional profiles, thus allowing our inferred GRN to accurately reproduce the experimentally measured attractor states and resolve the conflict between the intended GRN and the experimentally derived transcriptional profiles. This conclusion is supported by the observation that the attractors reproduced by our inferred GRN have 25.8% of the attractor distance of the mathematical model built by Cantone et al. [56] ([Table O in S1 Data](#)). Furthermore, the experimental transcriptional profiles showed that *SWI5* was repressed during overexpression of *GAL80* in both galactose and glucose, which was inconsistent with the intended GRN. The attractors produced by our model showed a consistent result: *SWI5* was suppressed when *GAL80* was overexpressed in glucose media, and it was expressed when galactose inactivated Gal80. Our model also explains the low expression of *GAL4* under *CBF1* overexpression: when *CBF1* was overexpressed, *ASH1* was activated by Cbf1 by two feed-forward loops (one via *SWI5* and the other via *GAL80*), and Ash1 in turn inhibited *GAL4*, lowering its expression. Together these results strongly suggest that our evolutionary algorithm approach to model construction can provide significant insight into the structure and regulatory dynamics of “real world” *in vivo* GRNs.

Modeling the white-opaque switch GRN in *C. albicans*

To expand beyond our model testing using data derived from “known” *in silico* and engineered *in vivo* GRNs, we next applied our algorithm to infer and simulate the dynamics of a

naturally occurring GRN which controls reversible differentiation between two distinct cell types—white and opaque—in the human fungal pathogen *C. albicans*. The white and opaque cell types are heritably maintained for hundreds of generations and the frequency of stochastic switching between these two cell types is controlled by a complex, highly interwoven series of transcriptional regulatory interactions [72]. The white and opaque cell types differ in the expression of approximately 18% of all genes in the *C. albicans* genome, thus providing two very distinct attractor states for the underlying GRN. To model the white-opaque GRN, we utilized transcriptional profiles derived from wildtype white and opaque cells, along with a series of strains that lack one or more of the TFs controlling the switch (See Table B in [S1 Text](#)). These additional TF deletion strains serve to provide additional steady-state attractors to further constrain the GRN structures. The majority of these strains can switch reversibly between the white and opaque cell types, thus providing two distinct attractor states per strain, with the exception of those TF deletion strains that are “locked” in one cell type or the other. In total, we obtained RNAseq data for seventeen distinct genotypic/phenotypic combinations including two wildtype strains, thirteen single TF deletion strains, and two double TF deletion strains. Each of the deleted TFs is known to impact the frequency of switching between the white and opaque cell types, and is known or predicted to impact the transcriptional profile of the resulting white and/or opaque cell types.

We first tested the ability of our evolutionary algorithm to predict the “unknown” transcriptional profiles produced by the GRNs of the wildtype and single TF deletion strains by omitting the transcriptional profile(s) of a specific genotype from the training dataset and allowing the model to predict the omitted transcriptional profile(s). Transcriptional profiles from the two double TF deletion strains were excluded from all training sets and were reserved as final test subjects for a “fully trained” version of the model developed using the full complement of fifteen wildtype and single TF deletion strain transcriptional profiles as the training dataset. If the attractor distance between the predicted and omitted transcriptional profile(s) was below the average replicate distance, or a cutoff of 0.16, the prediction would be considered successful. The cutoff of 0.16 was selected by the null model, which has less than a 2% chance of generating a result below this cutoff for eight and nine gene networks (See Table C in [S1 Text](#)). Since the null model produces transcriptional profiles by simply picking a value between the maximal and minimal expression levels, while the GRN dynamic system generates transcriptional profiles by numerically solving the differential equations, potential discrepancies may exist between the two. To rule out potential discrepancies due to the GRN dynamic system, we also generated 10,000 random GRNs as a control group and performed the same predictions on the omitted transcriptional profiles. Generally, half of the random GRNs produced fixed-point attractors, while the other half did not reach a steady state. Both the null model and the control GRNs showed similar distribution on their attractor distances and had an average of approximately 0.3.

Overall, nine out of the fifteen omitted wildtype and single TF deletion strain transcriptional profiles were successfully predicted by our model ([Table 3](#)). Of these nine successful predictions, eight had an average attractor distance of less than 0.16, and the last one ($\Delta/\Delta\text{efg1}$; [Table 3](#)) had an attractor distance above 0.16 but below the average replicate distance, meaning that the predicted transcriptional profiles were within the range of noise in the experimentally derived transcriptional profiles for the *EFG1* deletion strain. The six remaining prediction results showed either attractors exceeding the cutoff, or no attractor at all (indicated by dashes). We note that several of the experimentally derived transcriptional profiles had unusually high variability, as indicated by high average replicate distance values ($\Delta/\Delta\text{wor3}$ opaque, $\Delta/\Delta\text{czf1}$ opaque, and $\Delta/\Delta\text{rbf1}$ opaque; [Table 3](#)). This high variability suggests excessive noise in the RNAseq libraries, or multiple states/oscillations existing in these specific cell types, either

of which would violate the model assumption of a single stable-state transcriptional profile and make it challenging to evaluate the prediction. If we exclude these highly variable samples, the success rate of the model predictions increases to 66.7%. These highly variable transcriptional profiles also indicate that overfitting is likely mitigated in EA due to the limited experimentally derived attractors and the incredibly large search space. In practice, EA was often underfitted and failed to find a network that can perfectly reproduce all input attractors. If there is excessive noise in one of the attractors in the input set, EA may struggle to fit it accurately. This is because accurately fitting the majority of the other attractors yields a much higher fitness score than fitting this noisy attractor.

Next, we applied all fifteen of the wildtype and single TF deletion strain transcriptional profiles as training data to infer a consensus “fully trained” GRN. This consensus fully trained GRN was derived from thirty inferred GRN architectures and then used to predict the transcriptional profiles for two distinct double TF deletion strains. Since more attractors were used in the input, we anticipated that this consensus fully trained GRN should have a higher predictive power than the partially trained model. Both double TF deletion predictions were successful (Table 3), indicating that the transcriptional profiles produced by the consensus fully trained GRN closely mirror the experimentally derived transcriptional profiles for these two strains. In addition, we observed a range of 0.16 to 0.40 (in terms of attractor distance) in the differences between the single knockout samples ($\Delta/\Delta wor1$, $\Delta/\Delta rbf1$, and $\Delta/\Delta ssn6$) and the double knockout samples. In Table 3, we observe that the model predictions for the double-knockout samples were closer to the ground truth (0.097 and 0.155 in terms of attractor distance) compared to any of the single knockout samples. This suggests that some collective information, beyond what was solely present in the single knockout samples, played an important role in the double TF deletion predictions. Given the predictive accuracy of the consensus fully trained GRN, we next asked whether the underlying architecture, or adjacency matrix, of the inferred GRN also closely resembled the experimentally determined binding interactions between these regulators and their respective coding genes, as previously reported [72, 92, 93]. The GRN architectures inferred by the fully trained model are relatively diverse, with an average success rate of approximately 50% in predicting the experimentally determined TF-gene binding interactions observed in the ChIP data (Fig F in S1 Text). This discrepancy is not entirely unexpected given that our *in silico* testing demonstrated that multiple distinct GRN structures, covering a wide range of hamming distances, are capable of producing virtually identical transcriptional profiles, or attractor distances (Fig 4C).

Given the enormous number of potential GRN architectures in the search space, and the fact that distinct GRNs, which produce identical attractors cannot be differentiated based purely on transcriptional profiles, we asked whether incorporating TF binding constraints could enable the model to converge upon an architecture that more closely resembles the experimental ChIP data while simultaneously reproducing accurate transcriptional profiles. To bias the model toward the GRN architecture observed in the experimental data, we included a TF binding probability function in our evolutionary algorithm. Briefly, this function alters the probability of an edge being created or removed in the adjacency matrix, thus biasing the inferred GRNs towards the experimentally determined architecture. However, if the resulting GRNs fail to converge upon the experimental attractors, the evolutionary algorithm would ultimately converge upon a distinct GRN structure if needed to fit the transcriptional profiling data. We applied all seventeen of the transcriptional profiles used above, plus the previously published *in vivo* TF-DNA binding data, to infer “directed” GRNs. On average, the individual directed GRNs retained approximately 90% of the experimentally determined TF binding interactions while also reproducing most of the experimentally derived transcriptional profiles (Fig F in S1 Text). The consensus directed GRN, constructed by the high-

frequency edges of the individual directed GRNs, accurately reproduced thirteen out of the seventeen experimentally observed transcriptional profiles and eighty out of the eighty-one physical binding interactions between each of the regulatory TFs and their respective coding genes. The transcriptional profiles that the consensus directed GRN failed to incorporate were wildtype opaque, $\Delta/\Delta wor3$ opaque, $\Delta/\Delta ahr1$ opaque, and $\Delta/\Delta ssn6$ opaque, most of which had relatively high variability in their biological replicates (see full report for both *in silico* and *in vivo* prediction tests and inferred GRNs in [S1 Data](#)). Together these results indicate that it is indeed possible to converge upon a GRN structure that closely mirrors the experimentally determined TF-DNA binding data for the white-opaque switch, while accurately producing many of the same attractor states observed *via* RNAseq. However, this data also suggests a high degree of redundancy or potential for plasticity within the white-opaque GRN, thus compromising the ability of our model to infer the observed GRN structure based solely on transcriptional profiling data.

Discussion

In this work, we extended the attractor-matching strategy from a Boolean model to an ODE-based model by incorporating transcriptional kinetic parameters. We consider transcriptional profiles of stable cell types as fixed-point attractors [94] in the mRNA state space, and search for the GRN architecture that produces these attractors. We found in the *in silico* simulation that GRN architectures are significantly correlated with the attractors they produce. This correlation supports the logic of applying the attractor-matching approach to GRN inference. The ability of EA to infer “unknown” GRNs has been validated using both simulated datasets derived from “known” *in silico* GRNs and *in vivo* test datasets from an engineered GRN in *S. cerevisiae*. EA outperformed six other leading GRN inference methods when applied to the *in silico* attractors generated by SynTREN ODEs. In the *in vivo* test, EA not only successfully identified five of the six intended transcriptional edges, but also revealed some unintended edges that might account for the inconsistency between the designed GRN and experimentally derived transcriptional profiles. In addition to inferring GRN architecture based on transcriptional profiles, EA can also predict the effects of genetic perturbation on the inferred GRN. As a proof of principle, we used the inferred GRNs generated during our *in silico* model testing to then predict the unknown attractors that would be produced upon genetic perturbation of the original reference GRN (i.e., by deletion of each TF). The inferred *in silico* GRNs successfully predicted 71.1% of the attractors produced by the reference GRNs using the identical knockout strains (Fig 4D), indicating that EA can effectively capture GRN behavior based on transcriptional profiles. This result further suggests that EA can be used to generate testable predictions on the behavior of *in vivo* GRNs. Specifically, we envision the application of this approach to a hybrid computational and *in vivo* experimental process whereby GRNs are inferred based on *in vivo* transcriptional profiles, the inferred GRNs are perturbed *in silico* to generate “mutant” transcriptional profiles, and the accuracy of inferred GRNs are ultimately assessed by comparing predicted versus observed transcriptional profiles generated using *in silico* versus *in vivo* mutant strains. The accuracy of the inferred GRN could thus be supported if the predicted and experimentally measured transcriptional profiles converge. If not, the *in vivo* mutant strain and the resulting experimentally derived attractors could reveal a new pattern of GRN dynamics that had not been covered by the initial input attractors, and would thus complement the original wildtype attractors to further refine the inferred GRN. In this manner, it should be possible to iteratively refine predicted GRNs until they approximate the *in vivo* results.

As a proof of principle, we applied this iterative computational and experimental strategy to infer the GRN governing the white-opaque switch in *C. albicans*. We first used a dropout

strategy to infer GRNs based on a subset of available data and tested the ability of the inferred GRNs to predict the transcriptional profiles that were omitted from the training data. This approach led to an overall success rate of 66.7%, which approaches the 71.1% success rate observed in our *in silico* testing. Next, we demonstrated that a “fully trained” GRN inferred from all fifteen of the wildtype and single gene deletion strain profiles was successful in predicting the transcriptional profiles of two distinct “unknown” double TF knockout strains that were omitted from our training data. This result demonstrates that the inference of a GRN using a set of known attractors can bring insight into attractors that exist biologically but have not yet been measured in the lab. Although the inferred white-opaque GRNs accurately predicted most if not all of the dropped-out transcriptional profiles, they did not fully converge upon the TF localization patterns that we have observed in *in vivo* genome-wide TF localization experiments. To further constrain the white-opaque GRN, we inferred the “directed” GRNs with all seventeen available experimentally measured transcriptional profiles and included ChIP data that biases the GRN architecture towards the TF localization pattern observed *in vivo*. This consensus directed GRN accurately reproduced 76% of the RNAseq-derived transcriptional profiles and converged upon 99% of the ChIP-derived TF binding interactions.

There are potential pitfalls that can impact our approach. First, regulatory elements other than TFs, such as non-coding RNA molecules, post-translational modifications, and chromatin modifiers/remodelers, can also influence the behavior of a GRN of interest. Their regulatory effects can lead to false compensatory TF regulations and make the inferred network converge less often. Second, as observed in our *C. albicans* GRN modeling, noise in the experimental data can lead to a “fuzzy” target for prediction and compromise the ability of the approach to fit the transcriptional profiles into a GRN. Furthermore, while RNAseq data derived from a particular genotypic/phenotypic state is assumed to represent a fixed-point attractor, this is not necessarily the case. Multiple stable states or oscillatory transcriptional outputs could exist within a population of cells that appear to be phenotypically homogeneous, thus bulk RNA sequencing could average out single-cell heterogeneity and underlying GRN dynamics. These limitations could lead to inferred GRNs that simulate biased or non-existent targets. In future research, there is potential to extend our framework to accommodate time-series transcriptional profiles. In the case of non-periodic systems, we can assess the fitness of the network by measuring the distance between the model-derived trajectory and the curve formed by connecting the experimental data points in the time series. However, for oscillatory/periodic systems, a major challenge arises in accurately identifying limit cycles, a type of closed trajectory in phase space, from time-series data, particularly when the data are noisy. To accurately capture the dynamics of these systems, it may be necessary to employ nonparametric regression, nonlinear time-series analysis, or machine learning algorithms, which can be computationally demanding. Overcoming this challenge would significantly advance the application of EA to periodic biological systems. Third, EA assumes that the highest expression levels for each gene have been observed in the input transcriptional profiles and utilizes them to estimate the unknown parameters. Potential bias in parameter estimation can occur if this assumption is not satisfied. Fourth, as the scale of the network increases, more attractors are required to accurately infer the underlying network structure. However, due to the limited availability of attractors and the escalating computational costs, EA is only practical for small-scale networks comprising several dozen genes. We propose that EA is better suited for inferring regulatory networks of TFs. By focusing solely on the TFs, we can considerably reduce the size of the network that needs to be inferred using EA. We can then extend the network’s size by treating the TFs as hub genes that are significantly correlated with other non-TF genes in a biologically important module. By expressing the non-TF genes’ expression patterns through a

linear combination of the TFs, we can construct a TF-TF regulatory network that identifies interactions among the biological modules and includes the regulation of downstream target genes in the network using a much faster algorithm, such as LASSO ([95]). Alternatively, we intend to utilize evolutionary algorithms, similar to the one described in this study, to enhance existing “base-line” regulatory networks within a discrete dynamics framework. This approach enables the handling of much larger networks by employing a multi-pronged strategy where more scalable methods are initially employed, and more computationally intensive methods, like the algorithm presented in this research, are utilized to achieve higher fidelity in large networks. Moreover, EA has simplified the way that multiple activators or inhibitors regulate a target gene, either independently as monomers or cooperatively as a polymer, but *in vivo* TFs could have more complex and sophisticated forms of incorporation than modeled in EA. These and likely other confounding factors have the potential to adversely impact the process of GRN inference and can cause reduced accuracy in predicting unknown transcriptional profiles.

The most significant challenge in GRN inference is perhaps the inherent functional redundancy and plasticity of real-world GRNs. This was apparent in our *in silico* testing where we observed that GRNs differing in as many as ten regulatory interactions can produce qualitatively similar transcriptional profiles (Fig 4C). Similarly, we observed that most of the attractors produced by the *C. albicans* white-opaque GRN could be reproduced, and “unknown” attractors predicted, even when the inferred GRN does not closely match the experimentally determined GRN architecture (Fig F in S1 Text). These results are consistent with the idea that GRN structures can evolve while maintaining the same overall output, which is also supported by experimental evidence [96]. For example, Tsong et al. [97] identified a set of sexual differentiation genes that are negatively regulated in *S. cerevisiae*, but are believed to have been positively regulated in an ancestral fungal species. In this example, the overall output of the transcriptional circuit remains the same, despite significant changes in GRN architecture. Our work provides a mathematical foundation for the idea that GRN architecture has plasticity and evolves [98–100] under selective pressure [101]. Thus, experiments performed under a specific set of experimental conditions may fail to reveal some of the evolutionary pressures that have constrained the behavior of real-world GRNs under distinct environmental conditions. While the impact of these unobserved evolutionary pressures on GRN architecture and logic could be revealed by extensive measurements of GRN output in an array of different environmental conditions, we hypothesize that the iterative model refinement strategy that we propose here may represent an efficient alternative strategy.

Despite these potential limitations and challenges, we have shown that EA outperforms competing GRN inference tools when applied to *in silico* datasets and has a unique set of capabilities that can provide insights into the inner workings of *in vivo* GRNs. In future iterations of our GRN inference approach, we can incorporate other types of interactions between TFs that are not independent as assumed by default. For instance, to consider the fact that Gal80 can only perform gene regulation by binding to Gal4, we can add the following rule to the algorithm: if a target gene is regulated by Gal80 but not by Gal4, the regulation of Gal80 on this gene will be voided. Interactions between metabolites and TFs, such as IPTG deactivating the lac repressor, can also be incorporated into the approach by adding similar rules. In this manner, EA can flexibly integrate more detailed biological information beyond sequencing data and better simulate complex biological systems. We have demonstrated that our GRN inference approach can successfully narrow down the number of potential GRN structures for a given transcriptional program using only a relatively small number of transcriptional profiles. For example, there are 3^{81} potential adjacency matrices and 2^{18} protein coordination matrices for a GRN with nine genes in our framework, not to mention the

number of possible values of f_0 s bounded between 0 and 1. In total this creates a GRN search space in excess of 1.2×10^{44} distinct configuration options. In the *in silico* test of a nine-gene network, the GRN inferred by EA was at the $5.6 \times 10^{-13\text{th}}$ percentile in a network quantity distribution based on the Hamming distance to the reference GRN, whereby most of the incorrect networks were eliminated. Therefore, EA can effectively select candidate GRNs that can be further examined by experimentation. Based on these results, we believe that this model is a valuable companion to experimental approaches for deciphering the structure and logic of *in vivo* GRNs.

Supporting information

S1 Text. Contains additional details regarding the parameter estimation and methods employed in this study.

Fig A. Positive correlation between A_{net} similarity (Hamming distance on the horizontal axis) and attractor profiles similarity (attractor distance on the vertical axis). Each column in the box plots (A-E) contains 1000 random A_{net}^{mut} mutated from the 5 A_{net}^{ref} consisting of 5–9 genes. **Fig B.** Demonstration of multiple-TF regulation. The x-axis and y-axis are the protein concentration of two activators or repressors, and the z-axis shows the outcome of the regulation function. (A) Two independent activators. (B) Two independent repressors. (C) Two synergistic activators. (D) Two synergistic repressors. **Fig C.** Sensitivity tests for the kinetic parameters, (A) rates of transcription, (B) mRNA degradation, (C) translation, and (D) protein degradation. Each of these parameters was perturbed by 50% of their original values and used to generate the correlation between A_{net} similarity (Hamming distance on the horizontal axis) and attractor profiles similarity (attractor distance on the vertical axis). **Fig D.** Five GRN architectures were arbitrarily generated as references in the *in silico* test. They have five-nine (A-E) genes and no self-regulatory edges. The pointed arrows represent activating and the blunt arrows represent repressing regulatory interactions. **Fig E.** The non-autoregulation *in silico* test comparison results in F1 score (upper panel), AUROC (middle panel), and AUPRC (bottom panel). The F1 scores are calculated using a threshold cutoff of 0.5 for all models. Best performances are marked by asterisks for symmetric and asymmetric methods. **Fig F.** Accuracy distributions of the fully trained and directed GRNs determined by the ChIP data in *C. albicans*. Each distribution contains 30 GRN samples. The fully trained GRNs were solely inferred by the transcriptional profiles while the directed GRNs were also constrained by the ChIP data. Performing equally well on reproducing the transcriptional profiles, the direct GRNs showed a significant increase compared to the fully trained GRNs. **Fig G.** A stacked histogram displays the distribution of edge variances across 30 independent inference runs, showing the number of edges for each variance category. Prediction of dropout transcriptional profiles in *C. albicans*. **Fig H.** A dropout strategy was utilized to infer GRNs based on a subset of available data and assessed the predictive capability of the inferred GRNs for transcriptional profiles that were deliberately excluded from the training dataset. The initial states were configured to correspond to the omitted transcriptional profiles. **Table A.** Kinetic parameters used for *in silico* and real-life tests. **Table B.** *C. albicans* strains used in this study. **Table C.** Probabilities of cumulative attractor distance by the null model. Table C in S1 Text shows the probabilities of cumulative attractor distances produced by a null model. For each gene, the null model randomly picks a value in a continuous uniform distribution $U([R]_{i,min}, [R]_{i,max})$, where $[R]_{i,min}$ and $[R]_{i,max}$ are the minimal and maximal expression levels of the i^{th} gene. **Table D.** Comparison of inference software features. **Table E.** The *in silico* test result for protein coordination matrix. **Table F.** The *in silico* test result for f_0 . **Table G.** Variance of the outcome networks in the *in silico* test.

(PDF)

S1 Data. Table A. Ground truth for the 5-gene *in silico* network. **Table B.** Ground truth for the 6-gene *in silico* network. **Table C.** Ground truth for the 7-gene *in silico* network. **Table D.** Ground truth for the 8-gene *in silico* network. **Table E.** Ground truth for the 9-gene *in silico* network. **Table F.** Model comparison results for the *in silico* GRNs comprising autoregulation. **Table G.** Model comparison results for the *in silico* GRNs comprising no autoregulation. **Table H.** Overall Summary of Predicting Unknown *in silico* Attractors. **Table I.** Knockout attractors prediction results of the 5-gene networks. **Table J.** Knockout attractors prediction results of the 6-gene networks. **Table K.** Knockout attractors prediction results of the 7-gene networks. **Table L.** Knockout attractors prediction results of the 8-gene networks. **Table M.** Knockout attractors prediction results of the 9-gene networks. **Table N.** Model comparison results for the synthetic GRN built in *S. cerevisiae*. **Table O.** Comparison of attractors between our model and Cantone et al.'s model [56] for the synthetic GRN constructed in *S. cerevisiae*. **Table P.** The consensus “directed” *C. albicans* GRN inferred by our model using all 17 attractors. **Table Q.** Comparison between *C. albicans* experimental transcriptional profiles and attractors inferred by the consensus “directed” GRN.
(XLSX)

S2 Data. Contains the *in silico* and *in vivo* datasets used to perform GRN inference, including transcriptional profiles, gene lengths, promoter strengths, and TF-DNA binding data, and the detailed results of each individual inferred GRNs in this study.
(XLSX)

Acknowledgments

We gratefully acknowledge computing time on the Multi-Environment Computer for Exploration and Discovery (MERCED) cluster at the University of California, Merced. We especially extend our appreciation to Dr. David Ardell who provided valuable advice on the development of our approach. We also thank all Hernday and Nobile lab members for providing insight on the project.

Author Contributions

Conceptualization: Ruihao Li, Suzanne S. Sindi, Clarissa J. Nobile, Réka Albert, Aaron D. Hernday.

Data curation: Ruihao Li, Morgan M. Quail, Mohammad N. Qasim.

Formal analysis: Ruihao Li, Jordan C. Rozum, Morgan M. Quail, Mohammad N. Qasim, Suzanne S. Sindi, Clarissa J. Nobile, Réka Albert, Aaron D. Hernday.

Funding acquisition: Clarissa J. Nobile, Réka Albert, Aaron D. Hernday.

Investigation: Ruihao Li, Jordan C. Rozum, Morgan M. Quail, Mohammad N. Qasim, Suzanne S. Sindi, Clarissa J. Nobile, Réka Albert, Aaron D. Hernday.

Methodology: Ruihao Li, Jordan C. Rozum, Mohammad N. Qasim, Suzanne S. Sindi, Réka Albert, Aaron D. Hernday.

Project administration: Aaron D. Hernday.

Resources: Aaron D. Hernday.

Software: Ruihao Li.

Supervision: Aaron D. Hernday.

Validation: Ruihao Li, Jordan C. Rozum, Réka Albert, Aaron D. Hernday.

Visualization: Ruihao Li.

Writing – original draft: Ruihao Li.

Writing – review & editing: Ruihao Li, Jordan C. Rozum, Suzanne S. Sindi, Clarissa J. Nobile, Réka Albert, Aaron D. Hernday.

References

1. Ristevski B. A survey of models for inference of gene regulatory networks. *Nonlinear Analysis: Modelling and Control*. 2013; 18(4):444–465. <https://doi.org/10.15388/NA.18.4.13972>
2. Rué P, Martínez Arias A. Cell dynamics and gene expression control in tissue homeostasis and development. *Molecular systems biology*. 2015; 11(2):792. <https://doi.org/10.15252/msb.20145549> PMID: 25716053
3. Bui TT, Selvarajoo K. Attractor concepts to evaluate the transcriptome-wide Dynamics Guiding Anaerobic to Aerobic State transition in *Escherichia coli*. *Scientific reports*. 2020; 10(1):1–14. <https://doi.org/10.1038/s41598-020-62804-3> PMID: 32246034
4. Davidson EH, Erwin DH. Gene regulatory networks and the evolution of animal body plans. *Science*. 2006; 311(5762):796–800. <https://doi.org/10.1126/science.1113832> PMID: 16469913
5. Enver T, Pera M, Peterson C, Andrews PW. Stem cell states, fates, and the rules of attraction. *Cell stem cell*. 2009; 4(5):387–397. <https://doi.org/10.1016/j.stem.2009.04.011> PMID: 19427289
6. Kauffman S. A proposal for using the ensemble approach to understand genetic regulatory networks. *Journal of theoretical biology*. 2004; 230(4):581–590. <https://doi.org/10.1016/j.jtbi.2003.12.017> PMID: 15363677
7. Madhanshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Medicine*. 2012-5-1; 4(5):41. <https://doi.org/10.1186/gm340> PMID: 22548828
8. Moore D, de Matos Simoes R, Dehmer M, Emmert-Streib F. Prostate cancer gene regulatory network inferred from RNA-Seq data. *Current genomics*. 2019; 20(1):38–48. <https://doi.org/10.2174/1389202919666181107122005> PMID: 31015790
9. Hawe JS, Theis FJ, Heinig M. Inferring interaction networks from multi-omics data—a review. *Frontiers in genetics*. 2019; 10:535. <https://doi.org/10.3389/fgene.2019.00535> PMID: 31249591
10. Natale JL, Hofmann D, Hernández DG, Nemenman I. Reverse-engineering biological networks from large data sets. *arXiv preprint arXiv:170506370*. 2017;.
11. Kaderali L, Radde N. Inferring gene regulatory networks from expression data. In: *Computational Intelligence in Bioinformatics*. Springer; 2008. p. 33–74.
12. Larsen SJ, Röttger R, Schmidt HHHW, Baumbach J. *E. coli* gene regulatory networks are inconsistent with gene expression data. *Nucleic acids research*. 2019; 47(1):85–92. <https://doi.org/10.1093/nar/gky1176> PMID: 30462289
13. Siahpirani AF, Chasman D, Roy S. Integrative Approaches for Inference of Genome-Scale Gene Regulatory Networks. In: *Gene Regulatory Networks*. Springer; 2019. p. 161–194.
14. Omony J. Biological network inference: A review of methods and assessment of tools and techniques. *Annual Research & Review in Biology*. 2014; p. 577–601. <https://doi.org/10.9734/ARRB/2014/5718>
15. Liu W, Rajapakse JC. Fusing gene expressions and transitive protein-protein interactions for inference of gene regulatory networks. *BMC systems biology*. 2019; 13(2):37. <https://doi.org/10.1186/s12918-019-0695-x> PMID: 30953534
16. Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems*. 2009; 96(1):86–103. <https://doi.org/10.1016/j.biosystems.2008.12.004> PMID: 19150482
17. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*. 2008; 9(10):770–780. <https://doi.org/10.1038/nrm2503> PMID: 18797474
18. Ristevski B. Overview of computational approaches for inference of microRNA-mediated and gene regulatory networks. In: *Advances in Computers*. vol. 97. Elsevier; 2015. p. 111–145.
19. Lee WP, Tzou WS. Computational methods for discovering gene networks from expression data. *Briefings in bioinformatics*. 2009; 10(4):408–423. PMID: 19505889

20. Mercatelli D, Scalambra L, Triboli L, Ray F, Giorgi FM. Gene regulatory network inference resources: A practical overview. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*. 2020; 1863(6):194430. <https://doi.org/10.1016/j.bbagr.2019.194430> PMID: 31678629
21. Gardner TS, Faith JJ. Reverse-engineering transcription control networks. *Physics of life reviews*. 2005; 2(1):65–88. <https://doi.org/10.1016/j.plrev.2005.01.001> PMID: 20416858
22. Banf M, Rhee SY. Computational inference of gene regulatory networks: approaches, limitations and opportunities. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*. 2017; 1860(1):41–52. <https://doi.org/10.1016/j.bbagr.2016.09.003> PMID: 27641093
23. Delgado FM, Gómez-Vela F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artificial intelligence in medicine*. 2019; 95:133–145. <https://doi.org/10.1016/j.artmed.2018.10.006> PMID: 30420244
24. Omranian N, Nikoloski Z. Computational Approaches to Study Gene Regulatory Networks. In: *Plant Gene Regulatory Networks*. Springer; 2017. p. 283–295.
25. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In: *BMC bioinformatics*. vol. 7. Springer; 2006. p. 1–15.
26. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol*. 2007; 5(1):e8. <https://doi.org/10.1371/journal.pbio.0050008> PMID: 17214507
27. Wang RS, Saadatpour A, Albert R. Boolean modeling in systems biology: an overview of methodology and applications. *Physical biology*. 2012; 9(5):055001. <https://doi.org/10.1088/1478-3975/9/5/055001> PMID: 23011283
28. Gan X, Albert R. General method to find the attractors of discrete dynamic models of biological systems. *Physical Review E*. 2018; 97(4):042308. <https://doi.org/10.1103/PhysRevE.97.042308> PMID: 29758614
29. Gardner TS, Di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*. 2003; 301(5629):102–105. <https://doi.org/10.1126/science.1081900> PMID: 12843395
30. Wu S, Cui T, Zhang X, Tian T. A non-linear reverse-engineering method for inferring genetic regulatory networks. *PeerJ*. 2020; 8:e9065. <https://doi.org/10.7717/peerj.9065> PMID: 32391205
31. Ma B, Fang M, Jiao X. Inference of Gene Regulatory Networks Based on Nonlinear Ordinary Differential Equations. *Bioinformatics*. 2020;. <https://doi.org/10.1093/bioinformatics/btaa032> PMID: 31950997
32. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*. 2010; 107(14):6286–6291. <https://doi.org/10.1073/pnas.0913357107> PMID: 20308593
33. De Smet R, Marchal K. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology*. 2010; 8(10):717–729. <https://doi.org/10.1038/nrmicro2419> PMID: 20805835
34. Munoz S, Carrillo M, Azpeitia E, Rosenblueth DA, Griffin A. A Tool for Symbolic Inference of Synchronous Boolean Molecular Networks. *Frontiers in genetics*. 2018; 9:39. <https://doi.org/10.3389/fgene.2018.00039> PMID: 29559993
35. Ghaffarizadeh A, Podgorski GJ, Flann NS. Applying attractor dynamics to infer gene regulatory interactions involved in cellular differentiation. *Biosystems*. 2017; 155:29–41. <https://doi.org/10.1016/j.biosystems.2016.12.004> PMID: 28254369
36. Rodriguez A, Crespo I, Androsova G, del Sol A. Discrete logic modelling optimization to contextualize prior knowledge networks using PRUNET. *PloS one*. 2015; 10(6):e0127216. <https://doi.org/10.1371/journal.pone.0127216> PMID: 26058016
37. Poret A, Guziolowski C. Therapeutic target discovery using Boolean network attractors: improvements of kali. *Royal Society open science*. 2018; 5(2):171852. <https://doi.org/10.1098/rsos.171852> PMID: 29515890
38. Borriello E, Walker SI, Laubichler MD. Cell phenotypes as macrostates of the GRN dynamics. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*. 2020; 334(4):213–224. <https://doi.org/10.1002/jez.b.22938> PMID: 32157818
39. Fard AT, Ragan MA. Quantitative Modelling of the Waddington Epigenetic Landscape. In: *Computational Stem Cell Biology*. Springer; 2019. p. 157–171.
40. Youseph ASK, Chetty M, Karmakar G. Reverse engineering genetic networks using nonlinear saturation kinetics. *Biosystems*. 2019; 182:30–41. <https://doi.org/10.1016/j.biosystems.2019.103977>
41. McGoff KA, Guo X, Deckard A, Kelliher CM, Leman AR, Francey LJ, et al. The Local Edge Machine: inference of dynamic models of gene regulation. *Genome biology*. 2016; 17(1):214. <https://doi.org/10.1186/s13059-016-1076-z> PMID: 27760556

42. González J, Vujačić I, Wit E. Inferring latent gene regulatory network kinetics. *Statistical applications in genetics and molecular biology*. 2013; 12(1):109–127. PMID: [23744300](#)
43. Ehsan Elahi F, Hasan A. A method for estimating Hill function-based dynamic models of gene regulatory networks. *Royal Society open science*. 2018; 5(2):171226. <https://doi.org/10.1098/rsos.171226> PMID: [29515843](#)
44. Villaverde AF, Banga JR. Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *Journal of the Royal Society Interface*. 2014; 11(91):20130505. <https://doi.org/10.1098/rsif.2013.0505> PMID: [24307566](#)
45. Aguilar-Hidalgo D, Lemos MC, Córdoba A. Evolutionary dynamics in gene networks and inference algorithms. *Computation*. 2015; 3(1):99–113. <https://doi.org/10.3390/computation3010099>
46. Zhou JX, Bruschi L, Huang S. Predicting pancreas cell fate decisions and reprogramming with a hierarchical multi-attractor model. *PloS one*. 2011; 6(3):e14752. <https://doi.org/10.1371/journal.pone.0014752> PMID: [21423725](#)
47. Davila-Velderrain J, Villarreal C, Alvarez-Buylla ER. Reshaping the epigenetic landscape during early flower development: induction of attractor transitions by relative differences in gene decay rates. *BMC systems biology*. 2015; 9(1):1–14. <https://doi.org/10.1186/s12918-015-0166-y> PMID: [25967891](#)
48. Guo J, Lin F, Zhang X, Tanavde V, Zheng J. NetLand: quantitative modeling and visualization of Waddington's epigenetic landscape using probabilistic potential. *Bioinformatics*. 2017; 33(10):1583–1585. <https://doi.org/10.1093/bioinformatics/btx022> PMID: [28108450](#)
49. Jin S, Wu FX, Zou X. Domain control of nonlinear networked systems and applications to complex disease networks. *Discrete & Continuous Dynamical Systems-B*. 2017; 22(6):2169. <https://doi.org/10.3934/dcdsb.2017091>
50. Wang LZ, Su RQ, Huang ZG, Wang X, Wang WX, Grebogi C, et al. A geometrical approach to control and controllability of nonlinear dynamical networks. *Nature communications*. 2016; 7(1):1–11. <https://doi.org/10.1038/ncomms11323> PMID: [27076273](#)
51. Mochizuki A, Fiedler B, Kurosawa G, Saito D. Dynamics and control at feedback vertex sets. II: A faithful monitor to determine the diversity of molecular activities in regulatory networks. *Journal of theoretical biology*. 2013; 335:130–146. <https://doi.org/10.1016/j.jtbi.2013.06.009> PMID: [23774067](#)
52. Rozum JC, Albert R. Identifying (un) controllable dynamical behavior in complex networks. *PLoS computational biology*. 2018; 14(12):e1006630. <https://doi.org/10.1371/journal.pcbi.1006630> PMID: [30532150](#)
53. Zañudo JGT, Yang G, Albert R. Structure-based control of complex networks with nonlinear dynamics. *Proceedings of the National Academy of Sciences*. 2017; 114(28):7234–7239. <https://doi.org/10.1073/pnas.1617387114> PMID: [28655847](#)
54. Liu YY, Slotine JJ, Barabási AL. Controllability of complex networks. *nature*. 2011; 473(7346):167–173. <https://doi.org/10.1038/nature10011> PMID: [21562557](#)
55. Mochizuki A. Theoretical approaches for the dynamics of complex biological systems from information of networks. *Proceedings of the Japan Academy, Series B*. 2016; 92(8):255–264. <https://doi.org/10.2183/pjab.92.255> PMID: [27725468](#)
56. Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, et al. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell*. 2009; 137(1):172–181. <https://doi.org/10.1016/j.cell.2009.01.055> PMID: [19327819](#)
57. Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, Verschoren A, et al. SynTRen: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC bioinformatics*. 2006; 7(1):43. <https://doi.org/10.1186/1471-2105-7-43> PMID: [16438721](#)
58. Mendes P, Sha W, Ye K. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*. 2003; 19:ii122–ii129. <https://doi.org/10.1093/bioinformatics/btg1069> PMID: [14534181](#)
59. Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*. 2011; 27(16):2263–2270. <https://doi.org/10.1093/bioinformatics/btr373> PMID: [21697125](#)
60. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, et al. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology*. 2006; 7(5):R36. <https://doi.org/10.1186/gb-2006-7-5-r36> PMID: [16686963](#)
61. Di Camillo B, Toffolo G, Cobelli C. A gene network simulator to assess reverse engineering algorithms. *Annals of the New York Academy of Sciences*. 2009; 1158(1):125–142. <https://doi.org/10.1111/j.1749-6632.2008.03756.x> PMID: [19348638](#)
62. Morishita R, Imade H, Ono I, Ono N, Okamoto M. Finding multiple solutions based on an evolutionary algorithm for inference of genetic networks by S-system. In: *The 2003 Congress on Evolutionary Computation, 2003. CEC'03.* vol. 1. IEEE; 2003. p. 615–622.

63. Tominaga D, Koga N, Okamoto M. Efficient numerical optimization algorithm based on genetic algorithm for inverse problem. In: Proceedings of the 2nd Annual Conference on Genetic and Evolutionary Computation; 2000. p. 251–258.
64. Butcher JC. Numerical methods for ordinary differential equations. John Wiley & Sons; 2016.
65. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*. 2020; 17:261–272. <https://doi.org/10.1038/s41592-019-0686-2> PMID: 32015543
66. Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. *Development*. 1997; 124(10):1851–1864. <https://doi.org/10.1242/dev.124.10.1851> PMID: 9169833
67. Camacho D, VERA LICONA P, Mendes P, Laubenbacher R. Comparison of reverse-engineering methods using an in silico network. *Annals of the New York Academy of Sciences*. 2007; 1115(1):73–89. <https://doi.org/10.1196/annals.1407.006> PMID: 17925358
68. Klumpp S, Hwa T. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences*. 2008; 105(51):20245–20250. <https://doi.org/10.1073/pnas.0804953105> PMID: 19073937
69. Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan LK, et al. Sizing up metatranscriptomics. *The ISME journal*. 2013; 7(2):237–243. <https://doi.org/10.1038/ismej.2012.94> PMID: 22931831
70. Guet CC, Bruneaux L, Min TL, Siegal-Gaskins D, Figueroa I, Emonet T, et al. Minimally invasive determination of mRNA concentration in single living bacteria. *Nucleic acids research*. 2008; 36(12):e73–e73. <https://doi.org/10.1093/nar/gkn329> PMID: 18515347
71. Maurizi M. Proteases and protein degradation in *Escherichia coli*. *Experientia*. 1992; 48(2):178–201. <https://doi.org/10.1007/BF01923511> PMID: 1740190
72. Hernday AD, Lohse MB, Fordyce PM, Nobile CJ, DeRisi JL, Johnson AD. Structure of the transcriptional network controlling white-opaque switching in *Candida albicans*. *Molecular microbiology*. 2013; 90(1):22–35. <https://doi.org/10.1111/mmi.12329> PMID: 23855748
73. Lohse MB, Ene IV, Craik VB, Hernday AD, Mancera E, Morschhäuser J, et al. Systematic genetic screen for transcriptional regulators of the *Candida albicans* white-opaque switch. *Genetics*. 2016; 203(4):1679–1692. <https://doi.org/10.1534/genetics.116.190645> PMID: 27280690
74. Nguyen N, Quail MM, Hernday AD. An efficient, rapid, and recyclable system for CRISPR-mediated genome editing in *Candida albicans*. *MSphere*. 2017; 2(2):e00149–17. <https://doi.org/10.1128/mSphereDirect.00149-17> PMID: 28497115
75. Moll P, Ante M, Seitz A, Reda T. QuantSeq 3' mRNA sequencing for RNA quantification. *Nature methods*. 2014; 11(12):i–iii. <https://doi.org/10.1038/nmeth.f.376>
76. Dobin A, Gingeras TR. Mapping RNA-seq reads with STAR. *Current protocols in bioinformatics*. 2015; 51(1):11–14. <https://doi.org/10.1002/0471250953.bi1114s51> PMID: 26334920
77. Duttke SH, Chang MW, Heinz S, Benner C. Identification and dynamic quantification of regulatory elements using total RNA. *Genome research*. 2019; 29(11):1836–1846. <https://doi.org/10.1101/gr.253492.119> PMID: 31649059
78. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS one*. 2010; 5(9):1–10. <https://doi.org/10.1371/journal.pone.0012776> PMID: 20927193
79. Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology*. 2007; 2007:1–9. <https://doi.org/10.1155/2007/79879> PMID: 18354736
80. Meyer P, Marbach D, Roy S, Kellis M. Information-Theoretic Inference of Gene Networks Using Backward Elimination. In: *BioComp*; 2010. p. 700–705.
81. Chiquet J, Smith A, Grasseau G, Matias C, Ambroise C, Simone. Statistical inference for modular networks. *Bioinformatics*. 2009; 25(3):417–418. <https://doi.org/10.1093/bioinformatics/btn637> PMID: 19073589
82. Monteiro PT, Oliveira J, Pais P, Antunes M, Palma M, Cavalheiro M, et al. YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic acids research*. 2020; 48(D1):D642–D649. <https://doi.org/10.1093/nar/gkz859> PMID: 31586406
83. Moxley JF, Jewett MC, Antoniewicz MR, Villas-Boas SG, Alper H, Wheeler RT, et al. Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. *Proceedings of the National Academy of Sciences*. 2009; 106(16):6477–6482. <https://doi.org/10.1073/pnas.0811091106> PMID: 19346491
84. Parnell EJ, Yu Y, Lucena R, Yoon Y, Bai L, Kellogg DR, et al. The Rts1 regulatory subunit of PP2A phosphatase controls expression of the HO endonuclease via localization of the Ace2 transcription

- factor. *Journal of Biological Chemistry*. 2014; 289(51):35431–35437. <https://doi.org/10.1074/jbc.M114.611715> PMID: 25352596
85. Carrozza MJ, Florens L, Swanson SK, Shia WJ, Anderson S, Yates J, et al. Stable incorporation of sequence specific repressors Ash1 and Ume6 into the Rpd3L complex. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*. 2005; 1731(2):77–87. <https://doi.org/10.1016/j.bbaexp.2005.09.005> PMID: 16314178
 86. McBride HJ, Yu Y, Stillman DJ. Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation. *Journal of Biological Chemistry*. 1999; 274(30):21029–21036. <https://doi.org/10.1074/jbc.274.30.21029> PMID: 10409653
 87. Yarrington RM, Goodrum JM, Stillman DJ. Nucleosomes are essential for proper regulation of a multi-gated promoter in *Saccharomyces cerevisiae*. *Genetics*. 2016; 202(2):551–563. <https://doi.org/10.1534/genetics.115.183715> PMID: 26627840
 88. Phenix H, Morin K, Batenchuk C, Parker J, Abedi V, Yang L, et al. Quantitative epistasis analysis and pathway inference from genetic interaction data. *PLoS computational biology*. 2011; 7(5):e1002048. <https://doi.org/10.1371/journal.pcbi.1002048> PMID: 21589890
 89. Dutoit R, Dubois E, Jacobs E. Selection systems based on dominant-negative transcription factors for precise genetic engineering. *Nucleic acids research*. 2010; 38(19):e183–e183. <https://doi.org/10.1093/nar/gkq708> PMID: 20702421
 90. Reimand J, Vaquerizas JM, Todd AE, Vilo J, Luscombe NM. Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic acids research*. 2010; 38(14):4768–4777. <https://doi.org/10.1093/nar/gkq232> PMID: 20385592
 91. Cormier L, Barbey R, Kuras L. Transcriptional plasticity through differential assembly of a multiprotein activation complex. *Nucleic acids research*. 2010; 38(15):4998–5014. <https://doi.org/10.1093/nar/gkq257> PMID: 20392822
 92. Lohse MB, Johnson AD. Identification and characterization of Wor4, a new transcriptional regulator of white-opaque switching. *G3: Genes, Genomes, Genetics*. 2016; 6(3):721–729. <https://doi.org/10.1534/g3.115.024885> PMID: 26772749
 93. Hernday AD, Lohse MB, Nobile CJ, Noiman L, Laksana CN, Johnson AD. Ssn6 defines a new level of regulation of white-opaque switching in *Candida albicans* and is required for the stochasticity of the switch. *MBio*. 2016; 7(1):e01565–15. <https://doi.org/10.1128/mBio.01565-15> PMID: 26814177
 94. Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*. 1969; 22(3):437–467. [https://doi.org/10.1016/0022-5193\(69\)90015-0](https://doi.org/10.1016/0022-5193(69)90015-0) PMID: 5803332
 95. Omranian N, Eloundou-Mbebi JM, Mueller-Roeber B, Nikoloski Z. Gene regulatory network inference using fused LASSO on multiple data sets. *Scientific reports*. 2016; 6(1):20533. <https://doi.org/10.1038/srep20533> PMID: 26864687
 96. Dalal CK, Johnson AD. How transcription circuits explore alternative architectures while maintaining overall circuit output. *Genes & Development*. 2017; 31(14):1397–1405. <https://doi.org/10.1101/gad.303362.117> PMID: 28860157
 97. Tsong AE, Tuch BB, Li H, Johnson AD. Evolution of alternative transcriptional circuits with identical logic. *Nature*. 2006; 443(7110):415–420. <https://doi.org/10.1038/nature05099> PMID: 17006507
 98. Nosedal I, Mancera E, Johnson AD. Gene regulatory network plasticity preyears a switch in function of a conserved transcription regulator. *Elife*. 2017; 6:e23250. <https://doi.org/10.7554/eLife.23250> PMID: 28327289
 99. Baker CR, Booth LN, Sorrells TR, Johnson AD. Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. *Cell*. 2012; 151(1):80–95. <https://doi.org/10.1016/j.cell.2012.08.018> PMID: 23021217
 100. Erwin DH. Chapter Thirteen—Evolutionary dynamics of gene regulation. In: Peter IS, editor. *Gene Regulatory Networks*. vol. 139 of *Current Topics in Developmental Biology*. Academic Press; 2020. p. 407–431. Available from: <https://www.sciencedirect.com/science/article/pii/S0070215320300326>.
 101. Sorrells TR, Johnson AD. Making sense of transcription networks. *Cell*. 2015; 161(4):714–723. <https://doi.org/10.1016/j.cell.2015.04.014> PMID: 25957680