Inferring gene regulatory networks using transcriptional profiles as dynamical attractors

Ruihao Li[1], Jordan C. Rozum[2], Morgan M. Quail[1], Mohammad N. Qasim[1], Suzanne S. Sindi[3], Clarissa J. Nobile[4,7], Réka Albert[5,6] and Aaron D. Hernday[4,7]

[1]Quantitative and Systems Biology Graduate Program, School of Natural Sciences, University of California, Merced, CA, USA
[2]Department of Systems Science and Industrial Engineering, Binghamton University (State University of New York), NY, USA
[3]Department of Applied Mathematics, School of Natural Sciences, University of California, Merced, CA, USA
[4]Department of Molecular Cell Biology, School of Natural Sciences, University of California, Merced, CA, USA
[5]Department of Physics, Pennsylvania State University, PA, USA
[6]Department of Biology, Pennsylvania State University, PA, USA
[7]Health Sciences Research Institute, University of California, Merced, CA, USA

Address correspondence to: Aaron D. Hernday, ahernday@ucmerced.edu or Réka Albert, rza1@psu.edu

$$V_{i,min} - \min(I_{i,*}) \cdot D_{i,mRNA} = 0 \tag{S1}$$

$$V_{i,min} = \min(I_{i,*}) \cdot D_{i,mRNA} \tag{S2}$$

$$T_{*,i} = \frac{\frac{1}{2} \cdot (max(I_{i,*}) + min(I_{i,*})) \cdot V_{i,trl}}{D_{i,protein}} \tag{S3}$$

$$\frac{[P]^k}{[P]^k + T^k} = q \tag{S4}$$

$$k = \frac{\log \frac{1-q}{q}}{\log \frac{T_i}{[P]}} \tag{S5}$$

$$[P]_{max} = \frac{V_{trl} \cdot max(I_{i,*})}{D_{protein}} \tag{S6}$$

$$f_0 = \frac{\frac{D_{i,mRNA} \cdot I_{i,j} - V_{i,min}}{V_{i,max} - V_{i,min}} - C_A \cdot C_R}{C_A + C_R - 2 \cdot C_A \cdot C_R} \tag{S7}$$

$$H(S^1, S^2) = \frac{1}{N} \sum_{i=1}^{N} |S_i^1 - S_i^2| \tag{S8}$$

where S^1 and S^2 are the two strings rewritten by GRN architectures, $H(\cdot)$ is the Hamming distance function, and N is the length of the strings.

Table S1: 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 proteins	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

Table S2: Kinetic parameters used for *in silico* and real-life tests

Kinetic parameter	Derived values	Reference
mRNA elongation rate	4.8 nt./s	[1]
Ribosome elongation rate	8 aa./s	[2]
mRNA degradation rate	0.0067/s	[3]
Protein degradation rate	0.00796/s	[4]

Table S3: The *in silico* test result for protein coordination matrix

in silico GRN instances	Hamming distance	Percentile	Accuracy	Precision	Recall
5-gene GRN	1.00(5.00)	1.08%(62.37%)	0.90(0.50)	1.00(0.82)	0.90(0.50)
6-gene GRN	3.00(6.00)	7.30 %(61.29%)	0.75(0.50)	1.00(1.00)	0.75(0.50)
7-gene GRN	3.00(7.00)	2.87%(60.47%)	0.79(0.50)	0.73(0.76)	0.79(0.50)
8-gene GRN	1.5(8.00)	0.12%(59.82%)	0.88(0.50)	0.88(0.88)	0.88(0.50)
9-gene GRN	4.00(9.00)	1.54%(59.27%)	0.78(0.50)	0.60(0.65)	0.78(0.50)

Values in parenthesis show the results of random LG. Accuracy, precision, and recall are calculated by weighted average (averaging the support-weighted mean per label).

Table S4: The *in silico* test result for $f0$

in silico GRN instances	Average $f0$	Std $f0$
5-gene GRN	[0.001, 0.016, 0.013, 0.042, 0.001]	8.28e-2
6-gene GRN	[0.023, 0.043, 0.018, 0.07, 0.004, 0.028]	1.42e-1
7-gene GRN	[0.039, 0.045, 0.046, 0.052, 0.045, 0.015, 0.019]	1.28e-1
8-gene GRN	[0.028, 0.029, 0.031, 0.228, 0.021, 0.036, 0.074, 0.3]	1.84e-1
9-gene GRN	[0.006, 0.021, 0.031, 0.27, 0.098, 0.054, 0.086, 0.095, 0.035]	1.14e-1

The f_0s are 0 for all *in silico* reference GRNs.

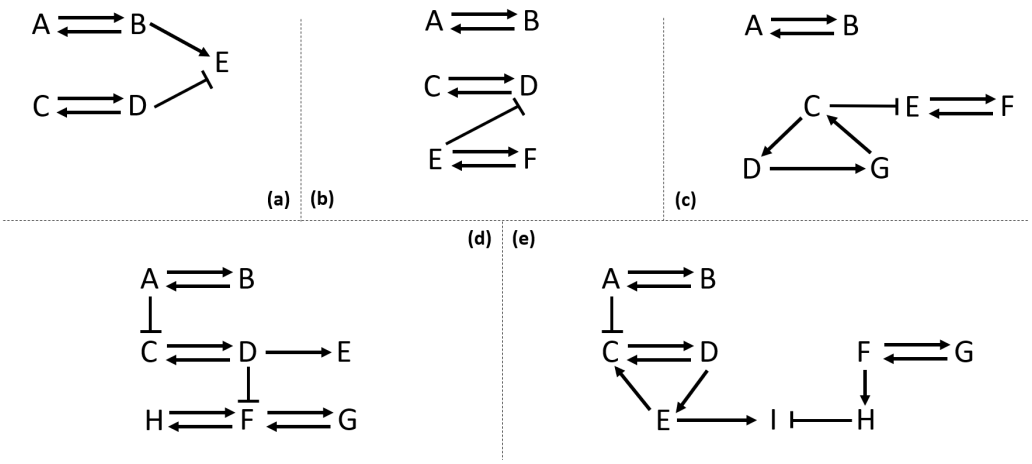

Figure S1: Five GRN architectures were arbitrarily generated as references in the *in silico* test. They have five-nine (a-e) genes and no self-regulatory edges. The pointed arrows represent activating and the blunt arrows represent repressing regulatory interactions.

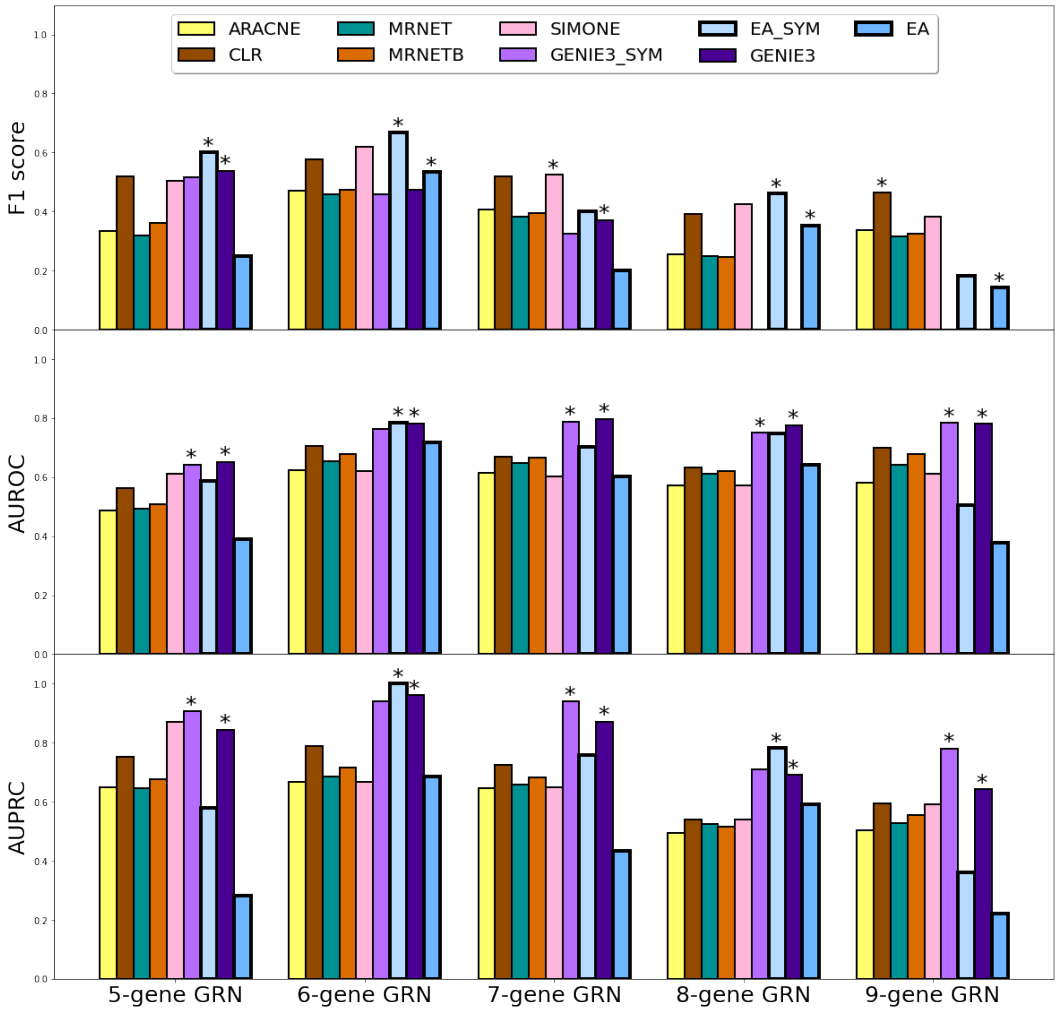

Figure S2: The non-autoregulation *in silico* test comparison results in F1 score (upper panel), AUROC (middle panel), and AUPRC (bottom panel). The F1 scores are calculated using a threshold cutoff of 0.5 for all models. Best performances are marked by asterisks for symmetric and asymmetric methods.

Table S5: Probabilities of cumulative attractor distance by the null model

	Attractor distances				
Number of genes	≤ 0.1	≤ 0.16	≤ 0.2	≤ 0.3	≤ 0.4
5 genes	0.61%	5.03%	12.50%	45.10%	73.39%
6 genes	0.29 %	3.61 %	10.53%	44.77%	74.11%
7 genes	0.14 %	2.63 %	9.00%	44.55 %	74.68%
8 genes	0.07%	1.95%	7.77%	44.36%	75.08%
9 genes	0.03%	1.46%	6.75%	44.23%	75.42%

Table S5 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 $\mathbf{U}([R]_{i,min}, [R]_{i,max})$, where $[R]_{i,min}$ and $[R]_{i,max}$ are the minimal and maximal expression levels of the i^{th} gene.

Table S6: *C. albicans* strains used in this study

Description	AHY	TF	Genotype	Reference
a/Δ wildtype	304	WT	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm^{434} IRO1/iro1Δ::imm^{434} arg4::hisG/arg4::hisG*	[5]
Δ/Δwor1	856	Wor1	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm^{434} IRO1/iro1Δ::imm^{434} arg4::hisG/arg4::hisG Δorf19.4884(wor1)::C.a.HIS1/Δorf19.4884(wor1)::C.a.LEU2*	[5]
Δ/Δwor2	736	Wor2	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm^{434} IRO1/iro1Δ::imm^{434} arg4::hisG/arg4::hisG Δorf19.5992(wor2)::C.a.HIS1/Δorf19.5992(wor2)::C.a.LEU2*	[5]
Δ/Δwor3	850	Wor3	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm^{434} IRO1/iro1Δ::imm^{434} arg4::hisG/arg4::hisG Δorf19.467(wor3)::C.a.HIS1/Δorf19.467(wor3)::C.a.LEU2*	[5]
Δ/Δwor4	861	Wor4	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm^{434} IRO1/iro1Δ::imm^{434} arg4::hisG/arg4::hisG Δorf19.6713(wor4)::C.a.HIS1/Δorf19.6713(wor4)::C.a.LEU2*	[5]

Δ/Δ*efg1*	836	Efg1	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm*^{434} *IRO1/iro1Δ::imm*^{434} *arg4::hisG/arg4::hisG Δorf19.610(efg1)::C.a.HIS1/Δorf19.610(efg1)::C.a.LEU2*	[5]
Δ/Δ*ahr1*	812	Ahr1	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm*^{434} *IRO1/iro1Δ::imm*^{434} *arg4::hisG/arg4::hisG Δorf19.7381(ahr1)::C.a.HIS1/Δorf19.7381(ahr1)::C.a.LEU2*	[5]
Δ/Δ*czf1*	784	Czf1	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm*^{434} *IRO1/iro1Δ::imm*^{434} *arg4::hisG/arg4::hisG Δorf19.3127(czf1)::C.a.HIS1/Δorf19.3127(czf1)::C.a.LEU2*	[5]
Δ/Δ*ssn6*	801	Ssn6	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm*^{434} *IRO1/iro1Δ::imm*^{434} *arg4::hisG/arg4::hisG Δorf19.6798(ssn6)::C.a.HIS1/Δorf19.6798(ssn6)::C.a.LEU2*	[5]
Δ/Δ*rbf1*	793	Rbf1	*a/ΔMTLalpha::ARG4 C.m.LEU2/Δleu2 C.d.HIS1/Δhis1 URA3/ura3D::imm*^{434} *IRO1/iro1Δ::imm*^{434} *arg4::hisG/arg4::hisG Δorf19.5558(rbf1)::C.a.HIS1/Δorf19.5558(rbf1)::C.a.LEU2*	[5]
Δ/Δ*wor1* Δ/Δ*ssn6*	1355	Wor1 Ssn6	*a/Δalpha C.m.LEU2/Δleu2 C.d.HIS1/his1Δ URA3/ura3D::imm*^{434} *IRO1/iro1Δ::imm*^{434} *arg4::hisG/arg4::hisG ΔMTLalpha::ARG4 Δorf19.6798(ssn6)::C.a.HIS1/Δorf19.6798(ssn6)::C.a.LEU2 Δwor1/Δwor1*	
Δ/Δ*wor1* Δ/Δ*rbf1*	1354	Wor1 Rbf1	*a/Δalpha C.m.LEU2/leu2Δ C.d.HIS1/his1Δ URA3/ura3D::imm*^{434} *IRO1/iro1Δ::imm*^{434} *arg4::hisG/arg4::hisG ΔMTLalpha::ARG4 Δorf19.5558(rbf1)::C.a.HIS1/Δorf19.5558(rbf1)::C.a.LEU2 Δwor1/Δwor1*	

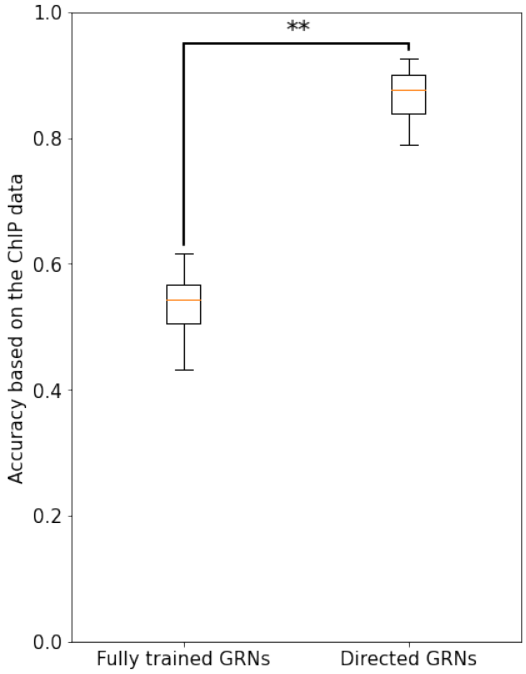

Figure S3: Accuracy distributions of the fully trained and directed GRNs determined by the ChIP data in *C. albicans*. Each distribution contains 30 GRN samples. The fully trained GRNs were solely inferred by the transcriptional profiles while the directed GRNs were also constrained by the ChIP data. Performing equally well on reproducing the transcriptional profiles, the direct GRNs showed a significant increase compared to the fully trained GRNs.

References

[1] S. Klumpp and T. Hwa, "Growth-rate-dependent partitioning of RNA polymerases in bacteria," *Proceedings of the National Academy of Sciences*, vol. 105, no. 51, pp. 20 245–20 250, 2008, ISBN: 0027-8424 Publisher: National Acad Sciences.

[2] C. C. Guet, L. Bruneaux, T. L. Min, *et al.*, "Minimally invasive determination of mrna concentration in single living bacteria," *Nucleic acids research*, vol. 36, no. 12, e73–e73, 2008.

[3] M. A. Moran, B. Satinsky, S. M. Gifford, *et al.*, "Sizing up metatranscriptomics," *The ISME journal*, vol. 7, no. 2, pp. 237–243, 2013.

[4] M. Maurizi, "Proteases and protein degradation inescherichia coli," *Experientia*, vol. 48, no. 2, pp. 178–201, 1992.

[5] M. B. Lohse, I. V. Ene, V. B. Craik, *et al.*, "Systematic genetic screen for transcriptional regulators of the candida albicans white-opaque switch," *Genetics*, vol. 203, no. 4, pp. 1679–1692, 2016.