Epigenome Prediction of Gene Expression using a Dynamical System Approach

James Brunner[†]

Center for Individualized Medicine Microbiome Program, Mayo Clinic
Rochester, MN 55901, USA
[†]E-mail: brunner.james@mayo.edu
www.mayo.edu

Jacob Kim

Department of Biological Sciences, Columbia University
New York, NY 10027, USA
www.columbia.edu

Timothy Downing

The Henry Samueli School of Engineering, University of California Irvine
Irvine, CA 92697, USA
www.uci.edu

Eric Mjolsness

Departments of Computer Science and Mathematics, University of California Irvine
Irvine, CA 92697, USA
www.uci.edu

Kord M. Kober[‡]

Department of Physiological Nursing and
Bakar Computational Health Sciences Institute, University of California San Francisco
San Francisco, CA 94143, USA
[‡]E-mail: kord.kober@ucsf.edu
www.ucsf.edu

Gene regulation is an important fundamental biological process. The regulation of gene expression is managed through a variety of methods including epigentic processes (e.g., DNA methylation). Understanding the role of epigenetic changes in gene expression is a fundamental question of molecular biology. Predictions of gene expression values from epigenetic data have tremendous research and clinical potential. Dynamical systems can be used to generate a model to predict gene expression using epigenetic data and a gene regulatory network (GRN). Here we present a novel stochastic dynamical systems model that predicts gene expression levels from methylation data of genes in a given GRN.

Keywords: Epigenetic Modification, Gene Regulatory Networks, Piecewise Deterministic Markov Process

1. Introduction

Gene regulation is an important fundamental biological process. It involves a number of complex sub-processes that are essential for development and adaptation to the environment (e.g., cell differentiation[1] and response to trauma[2]). Understanding gene expression patterns has broad scientific[3] and clinical[4] potential, including providing insight into patterns of fundamental molecular biology (e.g., gene regulatory networks) and a patient's response to disease (e.g., HIV infection[5]) or treatment (e.g., chemotherapy-induced peripheral neuropathy[6]).

The regulation of gene expression is managed through a variety of methods, including transcription, post-transcriptional modifications, and epigenetic processes.[7] One epigenetic process, DNA methylation,[8] occurs primarily at the cytosine base of the molecule that is adjacent to guanine (i.e., CpG site). DNA methylation of promoter and gene body regions can act to regulate gene expression by repressing (e.g.,[9]) or activating (e.g.,[10]) transcription.

Understanding the role of epigenetic changes in gene expression is a fundamental question of molecular biology and predicting gene expression from epigenetic data is an active area of research. Predictions of gene expression values from epigenetic data have tremendous research and clinical potential. For example, DNA is inexpensive to collect and is easy to store. It offers both genetic (i.e., genotype) and epigenetic (i.e., methylation status) information in a stable format. This information is obtainable from a large number of ongoing and previously existing biobanks. Gene expression (i.e., RNA), however, is more difficult and more expensive to obtain. Given the unique type of information that gene expression can provide (i.e., the presence and quantity of the functional product of a gene), it will be very useful and economical if gene expression values could be reliably predicted from methylation information.

A dynamical systems model can predict gene expression using epigenetic data and a gene regulatory network (GRN) by simulating hypothesized mechanisms of transcriptional regulation. Such models provide predictions based directly on these biological hypotheses. We develop an interaction network model[11] that depends on epigenetic changes in a GRN.

One major advantage to the dynamical systems approach is obtaining a distribution of gene expression beyond just expression status. Furthermore, a stochastic dynamical system provides us with a distribution of gene expression estimates, representing the possibilities that may occur within the cell.

Recent studies have developed models to predict gene expression levels with deep convolutional neural networks[12] from genome sequence data and with regression models from methylation data.[13] Previous studies have developed models to predict expression status (e.g., on/off or high.low) with gradient boosting classifiers from histone modification data,[14] and with machine learning classification methods from methylation data[15] and from methylation and histone data combined.[16] To our knowledge, there are no studies that have taken a dynamical systems approach to predicting gene expression from methylation data and a GRN.

Given the opportunity presented by dynamical systems approaches and the potential practical utility, we present a novel stochastic dynamical systems model for predicting gene expression levels from epigenetic data for a given GRN.

2. Methods

2.1. *Assumptions*

Our model assumes fundamentally that transcription of DNA is (relatively) fast and done at a linear rate determined by the bound or unbound state of transcription factor binding sites. We assume that binding and unbinding of transcription factors is a (relatively) slow and stochastic process, with propensity proportional to availability of transcription factor. Our model is built on the hypothesis that the propensity of transcription factor binding is influenced non-linearly by epigenetic modification of the binding site. We assume that "background" transcription (i.e. transcription occurring when no transcription factor is bound) is only enough to allow down regulation to exist. In model training and testing, genes with missing expression data are assigned 'zero' expression values.

2.2. *Model Equations*

We model gene regulation using a piecewise-deterministic Markov process (PDMP) as introduced in Davis 1984[17] (see also[18,19]) given by the equations:

$$B_i(t) = B_i(0) + Y_1^i\left(\int_0^t (1 - B_i(\tau))\lambda_i \frac{\mu_i}{\mu_i + (\alpha_i)^{\nu_i}}(\boldsymbol{\kappa}_i \cdot \boldsymbol{g})d\tau\right) - Y_2^i\left(\int_0^t \hat{\lambda}_i B_i(\tau)d\tau\right) \tag{1}$$

and

$$\frac{d}{dt}g_j = \gamma_j + (\boldsymbol{\phi}_j \cdot \boldsymbol{B}) - d_j g_j \tag{2}$$

where $B_i(t) \in \{0,1\}$, is a boolean random variable representing the bound/unbound state of a binding site region of DNA and g_i is the transcript amount the genes modeled. Equation (1) is given as the sum of two Poisson jump processes $Y_1^i(h_1^i(t))$ and $Y_2(h_2^i(t))$ which take values in $\mathbb{Z}_{\geq 0}$, and are piecewise constant between randomly spaced discrete time points (the binding and unbinding events).[20]

The propensities $h_1^i(t)$ and $h_2^i(t)$ are taken to be linear functions of the available transcription factors, which is assumed to be the same as the transcript variables g_j. We take the values $\kappa_{ij} \in \{0,1\}$; these parameters along with the set of ϕ_{ji} represent the structure of the underlying gene regulatory network.

We include the term

$$\frac{\mu_i}{\mu_i + (\alpha_i)^{\nu_i}} \tag{3}$$

to represent the impact of epigenetic modification on the binding propensity of transcription factors. In this term, α_i is the measured epigenetic modification to a transcription factor binding site (e.g. percentage of methylated bases). Equation (3) is a sigmoidal function which is either strictly increasing or strictly decreasing depending on the sign of ν_i. If $\nu_i > 0$, then this term decreases, implying that epigenetic modification decreases transcription factor binding. Conversely, if $\nu_i < 0$, the model implies that epigenetic modification increases transcription factor binding.

Finally, we use a linear ODE for the value of the transcripts g_j. We take $\phi_{ji} \in \{-1,0,1\}$ based on the structure of the underlying gene regulatory network. We include baseline transcription γ_j and decay d_j. Because we use a linear ODE in Eq. (2), we can solve exactly between jumps of $\boldsymbol{B}$.

2.3. *Master Equation and Equilibrium Distribution*

It is common practice in the study of reaction networks modeled as stochastic jump processes to represent the process using so called "chemical master equation",[21,22] which is the Kolmogorov forward equation for the jump process. The generator for a PDMP can be defined (see Azaïs 2014[23] for details). We define a density $P(B^i, \boldsymbol{g}(t), t) = P^i(\boldsymbol{g})$, $i = 1, ..., |\boldsymbol{B}|$ for each possible state $\boldsymbol{B}^i$ of $\boldsymbol{B}$ such that $\sum_{i=1}^{|\boldsymbol{B}|} P^i(\boldsymbol{g}) = P(\boldsymbol{g})$ is the probability distribution for the vector $\boldsymbol{g}$, and each P_i satisfies

$$\begin{aligned}\frac{dP^i(\boldsymbol{g},t)}{dt} &= \sum_{j=1}^{|\boldsymbol{g}|} \left(\gamma_j + (\phi_j \cdot \boldsymbol{B}^i) - d_j g_j\right) \frac{\partial P^i(\boldsymbol{g},t)}{\partial g_j} \\ &\quad + \sum_{(j:\|\boldsymbol{B}^j - \boldsymbol{B}^i\|_1 = 1)} \sum_{k=1}^{|\boldsymbol{B}|} \left[B^i_k (1 - B^j_k) \lambda_k \frac{\mu_k}{\mu_k + (\alpha_k)^{\nu_k}} (\boldsymbol{\kappa}_k \cdot \boldsymbol{g}) + \hat{\lambda}_k B^j_k (1 - B^i_k) \right] P^j(\boldsymbol{g},t) \\ &\quad - \sum_{k=1}^{|\boldsymbol{B}|} \left[(1 - B^i_k) \lambda_k \frac{\mu_k}{\mu_k + (\alpha_k)^{\nu_k}} (\boldsymbol{\kappa}_k \cdot \boldsymbol{g}) + \hat{\lambda}_k B^i_k \right] P^i(\boldsymbol{g},t). \qquad (4)\end{aligned}$$

2.4. *Model Sampling & Equilibrium Distribution Estimation*

We can estimate the equilibrium distribution for the PDMP using kernel density estimation (KDE) with a Gaussian kernel. To compute the marginal distributions on the various gene variables, which we can estimate with a 1 dimensional kernel. In a realization of the process, binding or unbinding events occur at times t_i such that $[t_0, t_1) \sqcup [t_1, t_2) \sqcup \cdots \sqcup [t_{n-1}, T) = [t_0, T)$ and we may compute $\boldsymbol{g}(t)$ exactly in each interval. The estimation is then

$$f_i(x) = \sum_{k=0}^{n-1} \int_{t_k}^{t_{k+1}} \frac{1}{\sqrt{2\pi}h} \exp\left(- \frac{\left(x - \left[e^{-d_i(t-t_k)}(g_i(t_k) - S^k_i) + S^k_i\right]\right)^2}{2h^2} \right) dt \qquad (5)$$

where $S^k_i = \frac{1}{d_i}(\gamma_i + \sum_j \phi_{ij} B_j)$. In Fig. 1, we show schematically how $f_i(x)$ is estimated from a realization of the process $\boldsymbol{B}(t), \boldsymbol{g}(t)$.

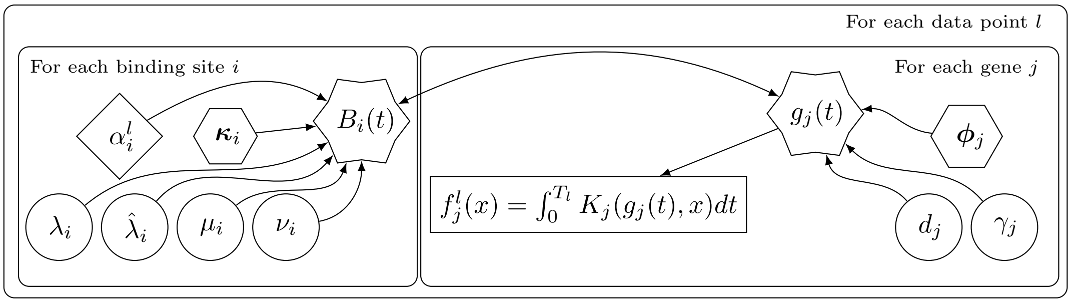

Fig. 1. Plate diagram of the process to estimate the marginal PDF $f_j(x)$ of each gene's transcript level according to our model. Parameters in diamonds are read from data, parameters in hexagons are determined by the structure of the network, parameters in circles must be fit to the model by maximizing likelihood over a training data set, and parameters in stars are the state variables of the dynamical model. Notice that the dynamical model implies that the state variables depend on each other, meaning this network of dependence is *not acyclic*. The kernel $K_j(x, y)$ used to estimate likelihood is Gaussian kernel in only component j (i.e., it is the Guassian kernel after orthogonal projection onto dimension j).

2.5. *Model Parameter Estimation*

While the parameters κ_{ij}, ϕ_{ji} and γ_j are determined by the structure of the underlying gene regulatory network, assumed to be known, and the epigenetic parameter α_i is assumed measurable, we must still estimate the remaining parameters $\lambda_i, \hat{\lambda}_i, \mu_i, \nu_i$ and d_j. We estimate these parameters using a negative log-likelihood minimization procedure using stochastic gradient descent. Sample paths used to estimate the gradient of the likelihood are generated using a modified form of Gillespies algorithm which handles time-dependent jump propensities.[24] This procedure involves approximating the gradient of the map from parameter set to log-likelihood so that we can use a gradient descent method. In Fig. 2, we give a schematic representation of how $\hat{L}_D$ is estimated from a set of realizations of the model, each realization corresponding to a single data sample.

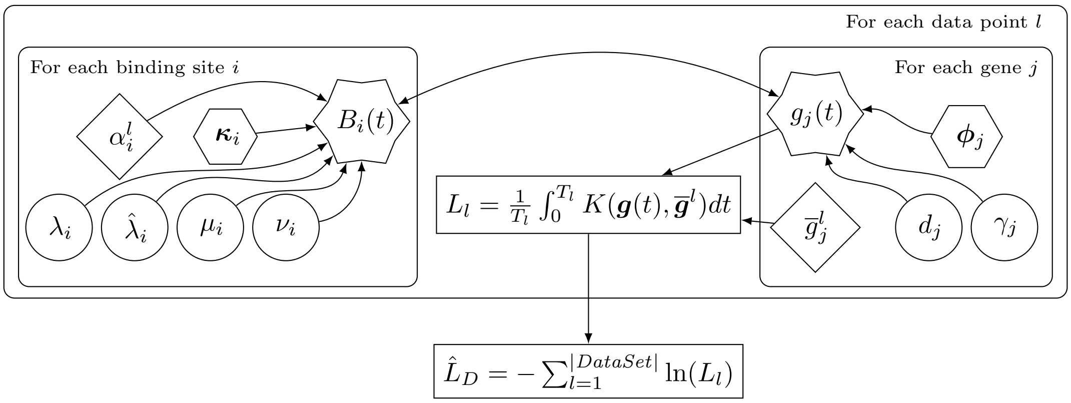

Fig. 2. Plate diagram of the process to estimate total likelihood of a data set according to our model. Parameters in diamonds are read from data, parameters in hexagons are determined by the structure of the network, parameters in circles must be fit to the model by maximizing likelihood over a training data set, and parameters in stars are the state variables of the dynamical model. Notice that the dynamical model implies that the stat variables depend on each other, meaning this network of dependence is *not acyclic*. The kernel $K(x, y)$ used to estimate likelihood is Gaussian.

2.5.1. *Gradient Estimation*

In order to estimate $\nabla \hat{L}_{D,\boldsymbol{\omega}}$ for use in optimization, we can use the generator of the system. To do this, we must rewrite the likelihood estimate $L_{\overline{\boldsymbol{g}},\boldsymbol{\alpha}}(\boldsymbol{\theta})$ as a sum of partial likelihoods $L^i_{\overline{\boldsymbol{g}},\boldsymbol{\alpha}}$ defined for each possible state of the vector $\boldsymbol{B}$. Then, because we are estimating the likelihood at an equilibrium distribution, we may use Eq. (4) and assume that $\frac{dP^i}{dt} = 0$ to compute each $\nabla L^i_{\overline{\boldsymbol{g}},\boldsymbol{\alpha}}$. See supplemental file for a derivation of $\nabla L^i_{\overline{\boldsymbol{g}},\boldsymbol{\alpha}}$ using this method.

2.6. *Creation of model from known GRN*

We use a gene regulatory network assumed to be known to create a model of gene regulation that includes transcription factor binding dynamics. To do this, we associate binding sites with the genes that they regulate, and use these associations to create a bipartite graph. Due to the available level of detail in the data set, we rely only on associations between binding sites and targets. We therefore assume that any regulator of a given target binds with every site associated with that target.

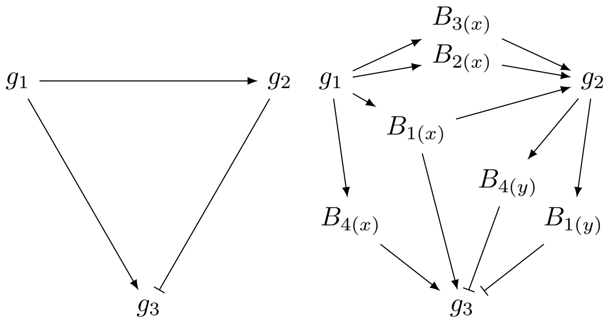

Fig. 3. (left) Underlying gene regulatory network. (right) Bipartite network of the underlying GRN.

In order to capture the effects of different regulating transcription factors that may bind to the same site, we create "duplicate" variables that represent the same binding site but bound with different transcription factors. The result is that we have only one transcription factor for each binding site variable. As shown in Fig. 3, each edge of a graph representing the original gene regulatory network is replaced by a path from transcription factor to binding site to target gene. Note that the model as described in Section 2.2 allows a more general topology, and our procedures for parameter fitting and model analysis do not depend on this construction.

3. Model Evaluation

3.1. *Gene Regulatory Network*

Gene to gene interactions ($\boldsymbol{\kappa}$) were defined using the Discriminant Regulon Expression Analysis (DoRothEA) framework.[25] Transcription factor (TF) to target interactions were defined using the DoRothEA highest confidence interactions and scored are just 1 or -1 for upregulating and downregulating, respectively. Binding site to target edges ($\boldsymbol{\phi}$) were defined by CpG methylation sites which were associated with changes in transcript expression (eCpG).[26] eCpG probes were identified for genes with which a change in expression was associated with a change in methylation state in participants from the Multi-Ethnic Study of Atherosclerosis (MESA). These CpG probes were then classified into a status which described the geographic relationship between CpG probe and the associated gene (e.g., "TRANS", "Promoter", "TSS", "Distal"). The binding site regions used in this study for a gene were defined by the proximal status classifications of the MESA data (i.e., 'Promoter' or 'TSS') and were scored as the mean of all of CpG probes in the status class. For evaluation, we identified a set of genes previously identified as deferentially expressed in individuals with PTSD as compared to controls.[27] Of these, we identified 278 regulator to target maps (kappa) which we then used to identify other targets to include. The final set included 512 gene to gene relationships comprised of 303 unique target genes. A GRN was built using these 303 genes as input producing a final network with 74 genes with 65 sites (network shown in supplemental file). Of these 74 genes, 29 had sufficient regulatory information (i.e., an associated binding site and transcription factor) for which parameters could be estimated and expression distributions generated.

3.2. *Model Training and Testing Data*

Matched epigenetic and gene expression data were obtained from the Grady Trauma Project (GTP) study. Methylation and gene expression was measured from whole blood from African American participants. Methylation data were obtained from the NCBI Gene Expression Omnibus (GEO) (GSE72680) and measured using the Illumina HumanMethylation450

BeadChip. Methylation status was quantified as a beta score. A total of 19,258 eCpG probes were used as input. Beta scores for CpG sites within the same region for a gene (i.e., ‘Promoter’ or ‘TSS’) were aggregated together as the average. Probes were merged in 1,885 regions. Regions that were not detected in any sample (i.e., “NaN”) were removed. Gene expression data were obtained from GEO (GSE58137) measured with the Illumina HumanHT-12 expression beadchip. Two batches of non-overlaping samples were quantified using two versions of the beadchip (V3.0: n=243 participants, mean expression intensity = 189.96, IQR = 49.88 to 106.60; V4: n=106 participants, mean expression intensity = 321.58, IQR = 88.85 to 139.78). Intensity scores were log2-transformed. Gene expression probes were first annotated to ENTREZ ID and then annotated to the symbol using the HUGO annotation.[28] Genes with multiple gene expression measurements (i.e., multiple probes) were represented by the last one in the list when processed. The final merged datasets contained n=243 samples for the participants measured for expression on the V3.0 beadchip and n=97 for the participants measured for expression on the v4.0 beadchip.

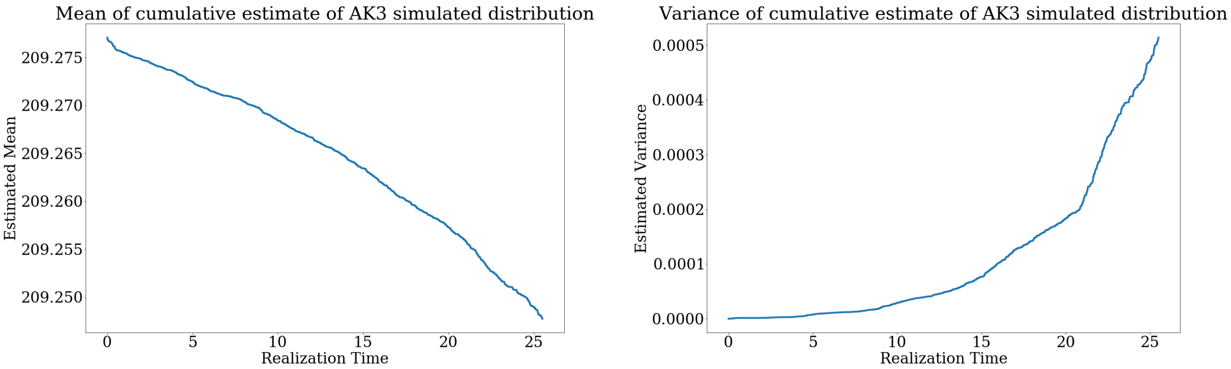

Fig. 4. The mean and variance over time for simulation of gene AK3 for sample 5881.

Matched data from participants measured for expression on the v3.0 beadchip (n=243) were used for evaluation. This primary dataset was split into training and testing datasets, containing 80%/20% (n=195, n=48 samples, respectively). To avoid the impact of a particular split, we repeated the shuffle process 100 times. For each split of the data, parameter estimation was performed on the training set and equilibrium distributions of the predicted expression levels were generated using the testing set. We evaluated the performance of the method as the difference between the measured and predicted value for each gene.

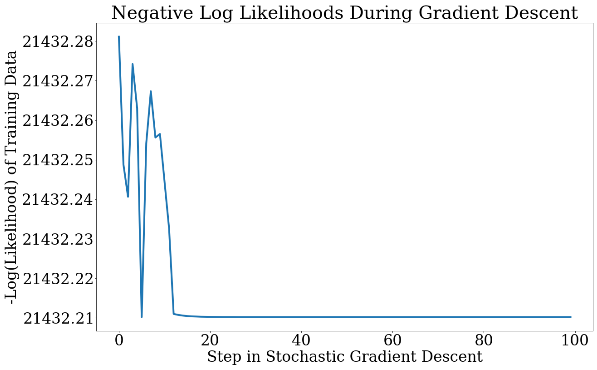

Fig. 5. Negative log-likelihood estimate of training data set for split 99x at each iteration of maximum-likelihood gradient descent procedure.

3.3. *Estimation of equilibrium distribution*

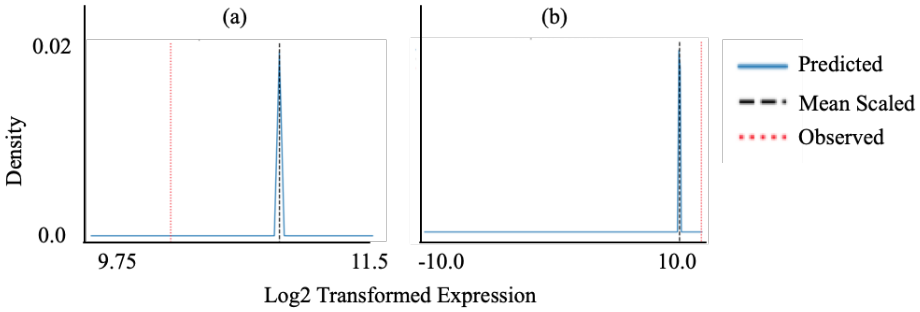

Fig. 6. Equilibrium distribution plots for CCM2 for individual (a) 6110 in shuffle 43, and (b) 6742 in shuffle 77.

Table 1. Average root mean square errors for each gene across the 100 shuffles with fitted parameters and random parameters.

Gene Symbol	Average RMSE fitted parameters	Average RMSE random parameters	Difference	Relative performance
AHR	2.196	6.113	3.918	2.785
AK3	0.757	10.459	9.702	13.810
ALOX5	2.172	6.797	4.625	3.129
BAG3	1.276	7.251	5.976	5.685
BAK1	1.637	8.114	6.478	4.958
CCM2	0.876	10.842	9.966	12.377
CD19	0.925	9.389	8.464	10.148
CD4	1.232	8.982	7.750	7.290
CTSH	1.262	8.278	7.016	6.561
CXCR5	4.425	7.444	3.020	1.683
CYP27A1	0.886	8.535	7.650	9.636
FBXO32	1.163	8.939	7.777	7.689
FCER1A	1.536	9.926	8.391	6.463
GSTM1	0.792	11.160	10.369	14.101
ICAM4	2.101	7.583	5.482	3.609
IRF1	3.357	10.636	7.280	3.168
LDHA	2.693	10.870	8.178	4.037
LTA4H	3.193	11.846	8.653	3.710
MT1X	1.222	9.468	8.247	7.749
NR1D2	0.932	7.585	6.654	8.142
OAS1	1.162	9.794	8.633	8.431
RPL39L	1.211	7.995	6.785	6.602
RRM2B	1.085	7.658	6.573	7.057
SCP2	4.598	7.354	2.757	1.600
SLC20A1	1.877	10.684	8.807	5.691
SREBF1	1.784	6.582	4.799	3.691
SURF6	1.107	9.236	8.130	8.347
VWA5A	2.744	7.310	4.567	2.664
ZNF654	3.254	8.015	4.762	2.464

In order to conserve computational resources, we used as a stopping condition for sample path simulation the number of random numbers that were needed for simulation to some time point. This was number was set to 2000 random draws. To evaluate if the gradient descent process was being ended too early, one additional shuffle was performed ("99x") which used an existing shuffle ("99") but with parameters allowing for less constrictive stopping conditions maxsteps=10000 StoppingCondition=100. While simulating, we saved mean and variance estimates for the distribution for intermediate time-points in order to approximate the rate of convergence to an equilibrium distribution. We use a finite difference estimate for the rate of change of estimated mean and variance to determine if the stopping condition that we used is appropriate. We can see from Fig. 4 that longer simulation time may be necessary for accurate estimates of equilibrium.

3.4. *Performance evaluation*

To evaluate the performance of our fitting procedure on gene expression predictions we generated predictions using a randomly generated parameter set. To avoid the impact of a particular split, we randomly generated parameters for each of the previously generated splits (i.e., the 100 splits previously used for training and testing). Ten random estimates were generated for each shuffle giving 1000 predictions for each gene generated using random parameters. We evaluated the performance of the parameter fitting method as the ratio of root-mean-square error of the predicted value given by randomly chosen and fitted parameters:

$$Relative\ performance = \frac{Average\ RMSE(random\ parameters)}{Average\ RMSE(fitted\ parameters)}.$$

4. Results

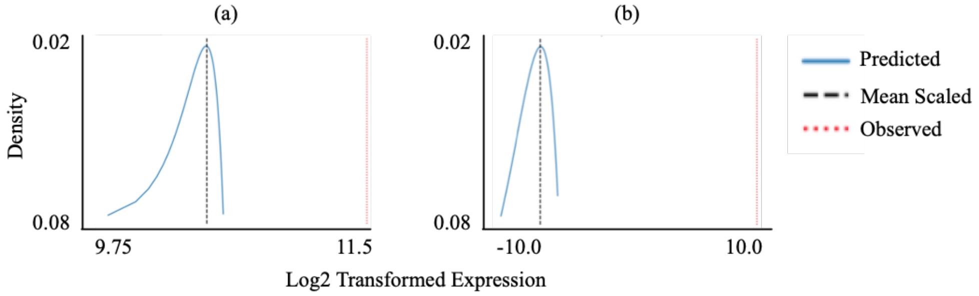

Fig. 7. Equilibrium distribution plots generated from random parameters for AK3 for individual (a) 6522 for random parameter set 0 in shuffle 0, and (b) 8331 random parameter set 7 in shuffle 13.

4.0.1. *Estimated parameters*

Parameters were estimated for all genes using the procedure detailed in Section 2.5. Figure 5 shows the likelihood of the training dataset during the course of the gradient descent procedure.

4.0.2. *Equilibrium distributions*

Equilibrium distributions were generated for 29 genes for which we had sufficient connectivity information available. Example equilibrium distributions for the CCM2 gene for two samples from different shuffles are shown in Fig. 6. Average root mean square errors for each gene across the 100 shuffles is reported in Table 1. We observed the highest performance for AK3 (average RSME = 0.757, Fig. 8) and lowest for SCP2 (average RSME = 4.598). In this evaluation, our model biases towards underestimating the expression levels (see supplemental file). Relative to a random set of parameters, our fitted parameters improved the predictions by a factor of 1.6 to over 14 (Table 1). For example, the predicted expression distribution for AK3 generated from random parameters are show in Fig. 9 and the predicted versus observed values are shown in Fig. 7.

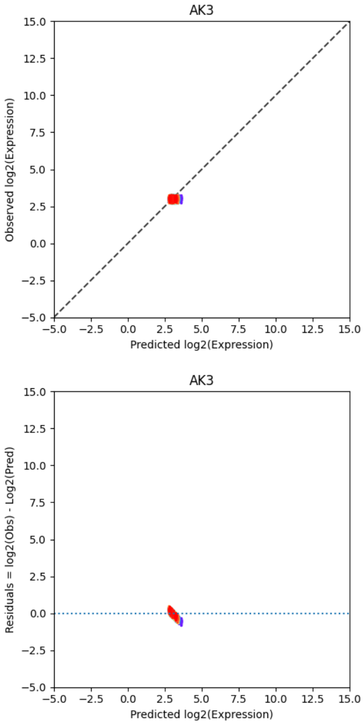

Fig. 8. (Top) Predicted versus observed expression values and (Bottom) residuals for the test samples for all 100 shuffles for AK3. Each shuffle is colored.

5. Discussion

In this study, we demonstrate that gene expression levels can be accurately predicted from methylation state of a promoter region and a GRN. Our model successfully uses quantitative data describing epigenetic modification of transcription factor binding sites to generate a probability distribution which describes the possible level of transcript. To our knowledge, this is the first study to develop and evaluate a stochastic dynamical systems model predicting gene expression levels from epigenetic data for a given GRN.

By using a dynamical systems approach, our model generates an estimation of gene expression given DNA methylation based on the mechanistic hypothesis of differential binding affinity of a transcription factor caused by epigenetic modification. Furthermore, the dynamical systems approach allows study of complex regulatory networks, including those which contains cycles. This method has the potential for broad practical usage. DNA is broadly available in banked tissue in the absence of RNA. By measuring methylation from these specimens, RNA expression value can be estimated. In addition to predicting RNA expression from specimens, our model can also be used to evaluate for functional effects of changes in methylation states at particular sites on gene expression levels.

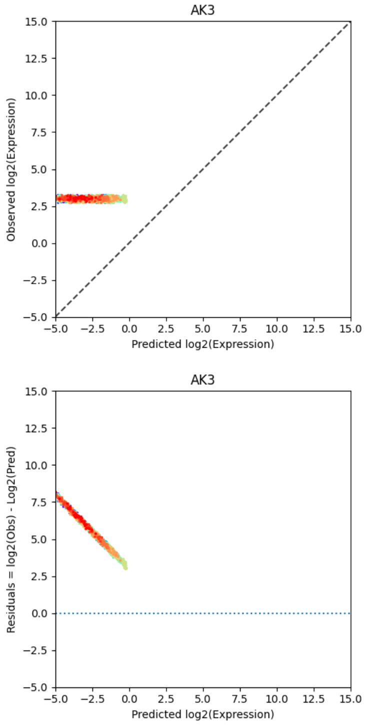

Fig. 9. (Top) Predicted versus observed expression values and (Bottom) residuals for the test set generated from 10 sets of random parameters for all 100 shuffles for AK3. Each shuffle and parameter set is colored.

Overall we see good performance in the predictions of the model with fitted parameters (e.g., Fig. 8) and dramatic improvements to prediction relative to a randomly generated set of parameters (e.g., Fig. 9). Although the predictions are somewhat dependent on the selection of training data, and in some cases, the observed value and predicted value are not well represented by the equilibrium distribution the overall prediction of the occurrence of gene expression remains accurate (i.e., on/off or low/high). Poor predictions could be the result of slow convergence in the maximum-likelihood parameter estimation. In order to estimate parameters, we must iteratively generate model predictions for a range of parameter values. This is done by generating equilibrium distributions in a method similar to Markov-Chain Monte Carlo (MCMC) sampling.

The estimated fit of the model to training data improved over iterations of the procedure (e.g., Fig. 5). However, the mean and standard deviations from the equilibrium distributions do not converge as quickly as we would like (e.g., Fig. 4). This slow convergence, and the necessity for repeated estimations, mean that computational time is a limiting factor. Future analyses should simulate longer to identify the appropriate cut offs given the data used, and thus improve the fit of the model parameters.

We were able to accurately predict gene expression despite the limited size of our GRN. We expect that model predictions will improve with more regulatory information. Although we queried for 302 genes, our transcription factor to target and binding site reference data produced a gene regulatory network with 74 genes, of which 29 had sufficient regulatory information to be predicted. In particular, dynamical systems should perform well in gene regulatory networks with complex topologies, including those with loops. We were unable to evaluate a more complicated GRN from all reference regulatory data due to computational constraints. As such, we evaluated with the toy model of PTSD data which was acyclic.

While the use of a stochastic dynamical system has distinct advantages over more statistically-driven methods, there are of course limitations to the approach as well. In particular, our model is based on the assumption that epigenetic modification effects the propensity of the random process of transcription factor binding and unbinding. Furthermore, our model assumes that DNA transcription is a comparatively fast (and so approximated as deterministic) process that depends on transcription factor binding. Finally, our model implicitly assumes that RNA translation to protein product is immediate. In addition to the fundamental limitations of the model, we also limit the scope of our testing to linear production of DNA transcript, depending on transcription factor binding status. Future efforts will be focused on improving the prediction accuracy, improving prediction robustness across folds, and evaluating across other gene regulatory networks and gene sets.

In conclusion, we have developed a dynamical systems approach which accurately predicts gene expression from methylation state and a GRN. Our results support the idea that methylation patterns of cis-promoter regions are associated with gene expression levels. Advances in gene regulatory information will continue to improve the predictions of our model by providing more structure to the GRNs. In addition, we have a route forward to develop optimizations which can step up the ease of use and scaling of our approach. Finally, our approach is broadly accessible and can be used for a diverse array of research projects which have DNA samples available but in which gene expression data are missing or limited or in studies evaluating the functional effects of changes in methylation state on gene expression.

6. Code availability

Supplemental files, including further mathematical details, are available at `https://doi.org/10.5281/zenodo.3970970`. All code are available at `https://github.com/kordk/stoch_epi_lib`.

7. Acknowledgments

This project was initially conceived as an interdisciplinary project as part of the "Short Course in Systems Biology - a foundation for interdisciplinary careers" at the Center for Complex Biological Systems at the University of California Irvine held Jan. 21 - Feb. 8, 2019 in Irvine, CA (NIH GM126365). This study was funded by the National Cancer Institute (NCI, CA233774). Preprint of an article submitted for consideration in Pacific Symposium on Biocomputing 2020 World Scientific Publishing Company `https://www.worldscientific.com/`.

References

1. Reik W. *Stability and flexibility of epigenetic gene regulation in mammalian development.* Nature, vol. 447(7143):pp. 425–32, 2007.
2. Cobb JP, *et al. Application of genome-wide expression analysis to human health and disease.* Proc Natl Acad Sci U S A, vol. 102(13):pp. 4801–6, 2005.
3. King MC *et al. Evolution at two levels in humans and chimpanzees.* Science, vol. 188(4184):pp. 107–116, 1975.

4. Singh KP, *et al. Mechanisms and Measurement of Changes in Gene Expression.* Biol Res Nurs, vol. 20(4):pp. 369–382, 2018.
5. Bosinger SE, *et al. Gene expression profiling of host response in models of acute HIV infection.* J. Immunol., vol. 173(11):pp. 6858–6863, 2004.
6. Kober K, *et al. Differential Methylation and Expression of Genes in the Hypoxia Inducible Factor 1 (HIF-1) Signaling Pathway Are Associated With Paclitaxel-Induced Peripheral Neuropathy in Breast Cancer Survivors and with Preclinical Models of Chemotherapy-Induced Neuropathic Pain.* Mol Pain, vol. 16:p. 1744806920936502, 2020.
7. Stephens KE, *et al. Epigenetic regulation and measurement of epigenetic changes.* Biol Res Nurs, vol. 15(4):pp. 373–381, 2013.
8. Razin A *et al. DNA methylation and gene function.* Science, vol. 210(4470):pp. 604–610, 1980.
9. Eden S *et al. Role of DNA methylation in the regulation of transcription.* Curr Opin Genet Dev, vol. 4(2):pp. 255–9, 1994.
10. Spruijt CG *et al. DNA methylation: old dog, new tricks?* Nat Struct Mol Biol, vol. 21(11):pp. 949–54, 2014.
11. Anderson DF, *et al. On classes of reaction networks and their associated polynomial dynamical systems.* Journal of Mathematical Chemistry, 2020.
12. Agarwal V *et al. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks.* Cell Rep, vol. 31(7):p. 107663, 2020.
13. Zhong H, *et al. Predicting gene expression using DNA methylation in three human populations.* PeerJ, vol. 7:p. e6757, 2019.
14. Ebert P, *et al. Epigenome-based prediction of gene expression across species.* bioRxiv, 2018.
15. Klett H, *et al. Robust prediction of gene regulation in colorectal cancer tissues from DNA methylation profiles.* Epigenetics, vol. 13(4):pp. 386–397, 2018.
16. Li J, *et al. Using epigenomics data to predict gene expression in lung cancer.* BMC Bioinformatics, vol. 16 Suppl 5:p. S10, 2015.
17. Davis MH. *Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models.* Journal of the Royal Statistical Society: Series B (Methodological), vol. 46(3):pp. 353–376, 1984.
18. Zeiser S, *et al. Simulation of genetic networks modelled by piecewise deterministic Markov processes.* IET systems biology, vol. 2(3):pp. 113–135, 2008.
19. Crudu A, *et al. Hybrid stochastic simplifications for multiscale gene networks.* BMC systems biology, vol. 3(1):p. 89, 2009.
20. Anderson DF *et al. Stochastic analysis of biochemical systems*, vol. 1. Springer, 2015.
21. Anderson DF, *et al. Product-form stationary distributions for deficiency zero chemical reaction networks.* Bulletin of mathematical biology, vol. 72(8):pp. 1947–1970, 2010.
22. Wang Y, *et al. Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent.* BMC systems biology, vol. 4(1):p. 99, 2010.
23. Azaïs R, *et al. Piecewise deterministic Markov processrecent results.* In *ESAIM: Proceedings*, vol. 44, pp. 276–290. EDP Sciences, 2014.
24. Anderson DF. *A modified next reaction method for simulating chemical systems with time dependent propensities and delays.* The Journal of chemical physics, vol. 127(21):p. 214107, 2007.
25. Garcia-Alonso L, *et al. Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer.* Cancer Research, vol. 78(3):pp. 769–780, 2018.
26. Varley KE, *et al. Dynamic DNA methylation across diverse human cell lines and tissues.* Genome Res., vol. 23(3):pp. 555–567, 2013.
27. Breen MS, *et al. PTSD Blood Transcriptome Mega-Analysis: Shared Inflammatory Pathways across Biological Sex and Modes of Trauma.* Neuropsychopharmacology, vol. 43(3):pp. 469–481, 2018.

28. Yates B, *et al. Genenames.org: the HGNC and VGNC resources in 2017*. Nucleic Acids Res, vol. 45(D1):pp. D619–D625, 2017.