Connecting quantitative regulatory-network models to the genome

Yue Pan1,2,*, Tim Durfee3, Joseph Bockhorst4 and Mark Craven2,1

1Department of Computer Sciences, 2Department of Biostatistics and Medical Informatics, 3Department of Genetics, University of Wisconsin, Madison, WI 53706 and 4Department of Electrical Engineering and Computer Science, University of Wisconsin, Milwaukee, WI 53211, USA

ABSTRACT

Motivation: An important task in computational biology is to infer, using background knowledge and high-throughput data sources, models of cellular processes such as gene regulation. Nachman et al. have developed an approach to inferring gene-regulatory networks that represents quantitative transcription rates, and simultaneously estimates both the kinetic parameters that govern these rates and the activity levels of unobserved regulators that control them. This approach is appealing in that it provides a more detailed and realistic description of how a gene’s regulators influence its level of expression than alternative methods. We have developed an extension to this approach that involves representing and learning the key kinetic parameters as functions of features in the genomic sequence. The primary motivation for our approach is that it provides a more mechanistic representation of the regulatory relationships being modeled.

Results: We evaluate our approach using two Escherichia coli gene-expression data sets, with a particular focus on modeling the networks that are involved in controlling how E.coli regulates its response to the carbon source(s) available to it. Our results indicate that our sequence-based models provide predictive accuracy that is better than similar models without sequence-based parameters, and substantially better than a simple baseline. Moreover, our approach results in models that offer more explanatory power and biological insight than models without sequence-based parameters.

Contact: ypan@cs.wisc.edu

1 INTRODUCTION

A central challenge in computational biology is to develop detailed and accurate models of the regulatory, signaling and metabolic networks for a wide variety of cell types. The state of the art for this challenging task involves encoding known molecular entities and relationships in such models as well as inferring relationships from high-throughput data sources. The work presented here is focused on inferring gene-regulatory networks from experimental data sources such as genomic sequences and expression data from microarrays. In particular, we present a novel approach that involves representing key parameters of regulatory relationships as functions of genomic sequence features.

A wide variety of model representations have been used for regulatory-network inference task, including Boolean networks (Ideker et al., 2001; Tanay and Shamir, 2001), differential equations (Chen et al., 1999), logistic functions (Weaver et al., 1999) and probabilistic models (Friedman et al., 2000; Pe’er et al., 2001; Hartemink et al., 2001, 2002; Ong et al., 2002; Yoo and Cooper, 2002; Yoo et al., 2002; Segal et al., 2003a,b; Tamada et al., 2003; Nachman et al., 2004; Nariai et al., 2005; Noto and Craven, 2005; Sachs et al., 2005), among others. Probabilistic models, such as Bayesian networks (Pearl, 1988) are appealing for this task because they can, in part, account for the uncertainty inherent in available data, and the non-deterministic nature of many of the interactions in a cell. A Bayesian network consists of two components: a qualitative one (the structure) in the form of a directed acyclic graph whose nodes correspond to the random variables, and a quantitative component consisting of a set of conditional probability distributions (CPDs). The nodes in such a network represent the presence/absence or the numeric quantities of certain gene products in a cell. The abundance of molecules of interest can be represented using either continuous or discrete values. The CPDs can be represented using tables (Friedman et al., 2000), tree structures (Noto and Craven, 2005) or linear Gaussian models (Friedman et al., 2000).

In contrast to these standard CPD representations, Nachman et al. (2004) devised an approach that represents quantitative transcription rates, and simultaneously estimates both the kinetic parameters that govern these rates and the activity levels of unobserved regulators that control them. This approach is appealing in that it provides a more detailed and realistic description of how a gene’s regulators influence its level of expression than do the standard CPD representations.

We have developed an extension to this approach that involves representing and learning the key kinetic parameters as functions of features in the genomic sequence. For example, instead of estimating a set of free parameters representing the binding affinity of given transcription factor (TF) to the promoter regions in which it binds, our method represents these parameters as a function of the relevant DNA sequence in the promoter regions, and learns the coefficients in this function. The primary motivation for our approach is that it provides a more mechanistic representation of the regulatory relationships being modeled. It offers more explanatory power than the models inferred by the approach of Nachman et al. in that it represents why a given kinetic parameter has the value it does, and it enables one to predict how certain changes in the genomic sequence might affect gene regulation.
Our approach also differs from that of Nachman et al. in that we explicitly represent RNA polymerase (RNAP) as a regulator. This is important when modeling bacterial networks, because a primary mechanism for gene regulation in bacteria is the recruitment of RNAP to specific promoters via its association with different sigma factors (Moooney et al., 2005), small RNAs (Willkomm and Hartmann, 2005) and small molecules such as ppGpp (Magnussen et al., 2005). We investigate and evaluate our approach in the context of trying to elucidate the networks that are involved in controlling how E. coli regulates its growth and other activities in response to the carbon source(s) available to it (Liu et al., 2005).

2 APPROACH

In this section, we first review the quantitative network model devised by Nachman et al. We then discuss how we have extended this method to represent certain key parameters in terms of sequence features, and other important aspects of our approach.

2.1 A kinematic model of transcriptional regulation

To model gene-regulatory networks in a quantitative, realistic way, we extend a model for regulator–regulatee dependencies introduced by Nachman et al. Their regulation model, as illustrated in Figure 1, is based on a regulation function that describes the transcription rate of a target gene as a function of the concentration of active regulators. Since the data being modeled, gene-expression profiles, are typically measured on populations of cells, the transcription rates represent average rates over a large cell population. Assuming that regulator binding and disassociation reactions are being modeled at steady state, they use a non-linear Michaelis–Menten form to describe transcription rates as a function of the concentration of active regulators. In the simplest case of a single activator, the regulation function is:

\[ g(H; \beta, \gamma) = \beta \frac{\gamma H}{1 + \gamma H} \]

where \( H \) denotes the concentration of active regulator protein, \( \beta \) is the maximum transcription rate the gene can achieve and \( \gamma \) is \( k_d / k_a \), the ratio of the association and disassociation constants for the regulator. Figure 1 depicts this case.

To illustrate the form of the \( g \) function in the general case of multiple regulators, consider a gene that has two regulators, with activity levels \( H_1 \) and \( H_2 \). Depending on whether no regulator, \( H_1 \), \( H_2 \) or both are bound to the promoter, we can distinguish four possible binding-site fractions, denoted \( S^{--}, S^{1-}, S^{-2}, S^{12} \). By solving the steady-state equations, we can determine the distribution over the possible binding fractions:

\[
S^{--} = \frac{1}{Z}, \quad S^{1-} = \gamma_2 H_2 / Z \\
S^{12} = \gamma_1 H_1 / Z, \quad S^{12} = \gamma_1 H_1 H_2 / Z
\]

where \( Z = (1 + \gamma_1 H_1)(1 + \gamma_2 H_2) \) is a normalizing constant. Therefore, the regulation function in this case is:

\[ g(H_1, H_2; \beta, \gamma_1, \gamma_2) = \beta \times (\alpha^{--} S^{--} + \alpha^{1-} S^{1-} + \alpha^{12} S^{12} + \alpha^{12} S^{12} H_1 S^{12} H_2) \]

where, \( \alpha \) is a vector of parameters indicating the ‘productive’ binding states that lead to transcription. For example, for two independent activators, Nachman et al. would set \( \alpha^{--}, \alpha^{1-}, \alpha^{12}, \alpha^{12} \) to 0, 1, 1, 1, respectively, reflecting that transcription occurs whenever at least one activator is bound. For a more realistic model, where different promoter states may result in different rates of transcription, they allow the \( \alpha \) parameters to take real values.

In this kinematic model of regulation, RNAP itself is not represented as a factor. Moreover, neither the \( \alpha \) nor \( \gamma \) parameters are related to the regulators’ binding-site sequence features. We extend the approach of Nachman et al. by explicitly considering the role of RNAP in transcription regulation, and by linking the binding strength (\( \gamma \)) and productivity (\( \alpha \)) parameters to features of the regulators’ binding sites.

2.2 Considering RNAP as regulator

One important extension of the approach of Nachman et al. is that we explicitly take into account the effects of RNAP in addition to those of TFs. Most computational approaches for modeling gene regulation do not represent the effects of RNAP in gene regulation, instead assuming that changes in gene expression are primarily due to changes in TF activities. As discussed in the Introduction Section, however, representing RNAP is a crucial consideration in modeling bacterial systems like E.coli. First, in E.coli, RNAP is thought to be limiting under most, if not all, conditions (Ishihama et al., 1976; Bremer and Dennis, 1996). Second, experiments have shown that E.coli RNAP distribution is dynamic and dramatically influenced by cell growth conditions (Cabrera and Jin, 2003), strongly arguing that RNAP availability is an important factor for the
differential gene expression observed under different conditions. Third, promoter studies (Aoyama et al., 1983; Hawley and McClure, 1983; Harley and Reynolds, 1987; Jacquet and Reiss, 1990) have demonstrated that RNAP has different binding affinities and initiation kinetics at different promoters, heavily depending on their primary sequence. Therefore, it is imperative that RNAP also be considered to more accurately model gene regulation in organisms like E.coli.

In our models, we treat RNAP as an essential activator. The degree of activation, for a particular gene, is determined by the available RNAP concentration, the binding affinity of RNAP to the gene’s promoter and its productivity when it binds to the promoter, in addition to the effects of other TFs. For example, a low expression level for a given gene could be explained by a low available RNAP concentration, low binding affinity of RNAP to the promoter or minimal productivity of the RNAP-only bound state, as well as an absence of activators. Hence, our models can capture some regulation scenarios that others cannot.

To account for RNAP as an essential activator for every gene in our model, we modify the regulation function as follows. For the case in which the given gene is regulated by one TF in addition to RNAP the regulation function is:

\[ g(H_1, H_2 : \alpha, \beta, \gamma_1, \gamma_2) = \beta \times (a^{H_1} \cdot \gamma_1 + b^{H_2} \cdot \gamma_2) \]

where, \( H_1 \) is the available RNAP concentration and \( H_2 \) is the concentration of an active TF. If the TF is a repressor, we set \( a^{H_1} \) to 0. If it is an activator, we constrain \( a^{H_1} \) to be larger than \( b^{H_2} \). We use this equation to reflect that RNAP is required to bind in order for any promoter state to have non-zero productivity. This approach is general and is easily extended to handle more regulators and different fractions of RNAP (e.g. ppGpp-bound and unbound fractions).

### 2.3 Modeling regulator binding-strength and productivity using genomic-sequence features

The regulation functions of Nachman et al., described in Section 2.1, involve one \( \gamma_{k,r} \) parameter for each pair of regulator \( r \) and regulated gene \( k \). Additionally, each gene has one \( \alpha \) parameter for each binding state that leads to transcript ion. The primary contribution of our work is an approach that involves representing the \( \gamma \)'s and \( \alpha \)'s as independent parameters, but instead as functions of the relevant sequences in promoter regions.

Figure 2 illustrates the basic idea of the approach. In the case of the \( \gamma \) parameters, we learn one function for each regulator. This function takes as input the known or putative binding-site sequence for the regulator in a given promoter region, and it outputs a predicted binding strength.

The rationale for such models comes from the assumption that binding strength is largely dependent on how a regulator’s DNA-binding domain interacts with the nucleotides at the binding site. For most regulators, different binding-site sequence features lead to different binding strengths, assuming other factors relevant to binding are the same.

Protein–DNA interaction analysis (Benos et al., 2002) shows that additivity is an appropriate first approximation for protein–DNA interaction. Binding affinity can be roughly attributed to the summation of contribution from each nucleotide at the binding site.

We represent our sequence-based binding models using linear functions. We encode the sequence composition of each binding site using a ‘1 of 4’ representation for each position. For example, a 4 nt sequence of ‘ACGT’ would be encoded as a binary-feature vector of [1000 0100 0010 0001]. Our models therefore link the sequence to a \( \gamma_{k,r} \) value as follows:

\[ \gamma_{k,r} = \sum w_{r,i} \times f_{k,i} + c_r \]

Where, \( \gamma_{k,r} \) is the binding strength of regulator \( r \) to the promoter of gene \( k \), \( f_{k,i} \) is the \( i \)th binary-feature value in the feature vector for gene \( k \), \( w_{r,i} \) is the weight associated with this feature and \( c_r \) is a constant unique for each regulator.

For the case of representing the binding affinity of RNAP to a given promoter sequence, we use a somewhat simpler linear model. In particular, we take advantage of the known consensus sequence for \( \sigma^{35} \) promoters in E.coli. The form of this linear model is:

\[ \gamma_{k,RNAP} = w_1 \times \text{sim}(k, -35\text{con}) + w_2 \times \text{sim}(k, -10\text{con}) \]

Where, \( \gamma_{k,RNAP} \) represents the binding strength of RNAP to gene \( k \), \( \text{sim}(k, -35\text{con}) \) is the similarity of the \(-35\) region of gene \( k \)'s promoter to the \(-35\) consensus sequence, \( \text{sim}(k, -10\text{con}) \) is the similarity measure for \(-10\) region to the \(-10\) consensus sequence and \( w_1 \) and \( w_2 \) are learnable weights indicating the importance of the two hexamer regions. We define the similarity function, \( \text{sim} \), as the number of nucleotides in the promoter region that match the corresponding consensus sequence.

We use this simpler linear form for RNAP for two reasons. First, it allows us to take advantage of the experimental evidence that RNAP binding strength is closely related to the
consensus sequence (Kobayashi et al., 1990). Second, it may help to simplify the learning process which is trying to simultaneously induce two sequence-based models for RNAP: this model for $\gamma_R$,RNAP and an $\alpha$ model for the productivity of RNAP when it is bound alone. By imposing this bias on the $\gamma_R$,RNAP model, the $\alpha$ is better able to capture the sequence characteristics that are important for the subsequent steps in transcription initiation, after binding.

RNAP is a special regulator in our models. Its effect on transcription depends not just on the binding strength of a given promoter, but also how the sequence influences the isomerization and elongation steps in transcription initiation. We represent the productivity, $\alpha$, of the promoter state when only RNAP is bound using a linear function similar to those we use for $\gamma$ parameters. The set of sequence features used for this RNAP productivity model includes the nucleotide composition at each position in the $-35$ and $-10$ regions, the length of the spacer between $-35$ and $-10$ hexamers, the GC content of the spacer and GC content of the region between the $-10$ hexamer and the transcription start site (TSS).

Currently, we do not use these linear models to represent the productivity of promoter states when TFs are bound. This situation is more complex since it involves not only protein-DNA interactions but also protein-protein interactions. We use free $\alpha$ parameters, as in the models of Nachman et al. for these states.

For genes that are in the same operon, we use one set of $\gamma_R,\alpha$ and $\beta\gamma$ parameters to represent the regulation function of all the genes in the operon.

2.4 Accounting for sequence uncertainty

The sequence-based $\gamma$ and $\alpha$ models described above assume that the relevant binding sites are known. In some cases, however, there is uncertainty about the locations of these binding sites because they have not been experimentally determined. This is especially the case for many RNAP binding sites.

In E.coli, $\sigma^{70}$ promoters are characterized by the well-known consensus sequence and structure: the $-35$ region (TTGACA consensus hexamer) and $-10$ region (TATAAT consensus hexamer) and the spacer region between them. However, for most operons in E.coli, the TSSs are unknown. Even for operons with known TSSs, the exact positions of $-35$ and $-10$ hexamers are still uncertain because the length of spacer and the distance between the $-10$ region and TSS vary from operon to operon.

In order to account for the uncertainty in binding-site location, we use the procedure illustrated in Figure 3. The starting point for this procedure is a model that represents the known (experimentally determined) binding sites for the regulator. Using such a model along with assumptions about how many binding sites for the regulator might occur in a given promoter region, we can calculate a distribution over possible starting positions of the regulator's binding site. From this starting-position distribution, we can then calculate a distribution over the base composition at each binding-site position. For example, suppose that our binding-site model for RNAP predicts that a given gene's $-10$ region starts with nucleotide ‘A’ at position p1 with probability 0.7 and starts with nucleotide ‘T’ at position p2 with probability 0.3, then the base distribution for the first nucleotide of $-10$ region is $[0.7 \ 0 \ 0.3]$. These distribution vectors are used in lieu of the binary-feature vectors as input to the sequence-based $\gamma$ and $\alpha$ models described in Section 2.3.

In the work reported here, we use this procedure to represent the predicted binding sites of RNAP. Our predictions are made using a hidden Markov model (HMM) we previously developed that represents various sequence elements of E.coli intergenic regions (Bockhorst et al., 2003b). Specifically, given a promoter sequence, we use this HMM to compute the joint probability of the locations of the TSS and the $-10$ and $-35$ hexamers: $Pr(C_{-35}, C_{-10}, C_{TSS}|X_1, \ldots, X_d)$. Here $C_{-35}$, $C_{-10}$ and $C_{TSS}$ represent the coordinates of the $-35$ and $-10$ hexamers, and the TSS, respectively. $X_1, \ldots, X_d$ represents the bases of the promoter sequence.

To calculate this joint probability, we employ a procedure based on posterior decoding (Durbin et al., 1998), making the assumption that each promoter region contains exactly one RNAP binding site. We do this calculation in two steps. The first step involves computing a distribution over what can be thought of as the ‘boundaries’ of the core promoter. This is defined by a distribution over the locations of the $-35$ hexamer and the TSS given the upstream sequence: $Pr(C_{-35}, C_{TSS}|X_1, \ldots, X_d)$. The second step involves computing $Pr(C_{-10}|C_{-35}, C_{TSS}, X_{-35}, \ldots, X_{TSS})$, a distribution over the coordinates of the $-10$ hexamer given the ‘boundary’ coordinates of the promoter and the sequence of bases falling within these coordinates. Here $X_{-35}$ represents the first position in the $-35$ hexamer, and $X_{TSS}$ represents the coordinate of the TSS. Within the promoter submodel of our intergenic HMM,
Section 2.1. Currently, we use a fixed regulation structure based on the non-linear kinematic regulation functions described in Section 2.1. The bottom layer represents the target genes regulated by their regulators through relationships between TFs/RNAP and their target genes. Genes gene, the weights in the contrast, the parameters to be learned include leads to transcription. In our sequence-based models, in these parameters, the learner must simultaneously infer the parameters for other binding states. In both types of models, the known quantities in the learning process are gene expression profiles that common regulatory mechanisms are being employed across carbon sources, at least some of which respond to the alternative carbon sources. These genes can be grouped into More importantly, there is a group of 154 up-regulated genes, of affected genes increases inversely with the growth rate. There are several interesting features that make this data set interesting: glucose > glycerol and succinate > alanine > acetate > proline. These compounds result in dramatically different growth rates in the order: glucose > glycerol and succinate > alanine > acetate > proline. At an OD600 of 0.2, total RNA from each culture was processed and hybridized to Affymetrix E.coli Antisense GeneChips. Expression profiles from the five alternative carbon sources were then compared to the glucose controls to identify differentially expressed genes in each culture. There are several interesting features that make this data set a prime candidate for modeling. First, there are far more up-regulated than down-regulated genes and the number of affected genes increases inversely with the growth rate. More importantly, there is a group of 154 up-regulated genes, representing 117 known and predicted operons (Bockhorst et al., 2003a), that are induced in a nested-subset manner across the alternative carbon sources. These genes can be grouped into three nested subsets (Liu et al., 2005). These results suggest that common regulatory mechanisms are being employed across carbon sources, at least some of which respond to the
growth rate. One candidate mechanism involves regulation of RNAP itself by ppGpp (‘magic spot’). ppGpp levels rise in response to many stimuli which decrease the growth rate, including carbon downshifts. ppGpp elicits a transcriptional switch whereby growth-promoting genes (e.g. stable RNAs) are down-regulated and auxiliary pathways for adapting to new environments are up-regulated. Therefore, a quantitative understanding of how this and other general mechanisms are integrated as carbon source quality varies would shed light on the crucial issue of how cells coordinate transcription with growth rate. As we explicitly represent the available RNAP as a regulator in the network, our model can potentially capture regulation mechanisms like the redistribution of RNAP due to the change in ppGpp levels.

Our model for this data set represents these 154 up-regulated genes, the concentrations of available RNAP and seven TFs and the regulatory relationships that exist among the TFs and the regulated genes. The TFs in the model are selected such that each one regulates at least three genes in the network. The binding-site locations for the TFs are collected from the EcoCyc (Keseler et al., 2005) and RegulonDB (Salgado et al., 2006) databases. In cases where the operon structure is not known for a gene, predictions made by our previously developed models are used (Bockhorst et al., 2003b). The binding-site locations for RNAP are predicted as described in Section 2.4.

The second data set we use includes a set of 302 CRP-regulated genes and their expression measurements from 50 Affymetrix microarray experiments (T. Durfee and F.R. Blattner, unpublished data, available upon request). From the original 50 experiments, we average the measurements under identical experimental conditions (i.e. replicates) to get an expression profile consisting of 23 values for each gene. The number of TFs represented in this network is 14. As with the carbon-shift data set, we incorporate known TFs that regulate at least three genes in the network.

3.2 Evaluating model accuracy

Our data sets consist of ‘gene-condition’ data points describing the expression levels of each gene in each experimental condition. Thus we have \( K \times J \) samples in a data set, where \( K \) is the number of genes and \( J \) is the number of conditions. To assess the extent to which a model captures the underlying regulatory mechanisms, we use a standard 10-fold cross-validation method to evaluate a model’s predictive ability on held-out test data. We randomly partition a data set into training and test sets, such that the test set consists of the expression measurements of a subset of genes in a subset of conditions, and the training set includes the rest of the expression values. We learn a model from a training set and test its ability to generalize to the held-aside test set; i.e. its ability to predict the expression levels for the gene-condition pairs in the test set.

To evaluate the real-valued expression-level predictions, we use relative error \( \epsilon_p \), where \( p \) is the observed expression level and \( p \) is the predicted expression level for each test case. Note that expression levels are directly proportional to transcription rates in our system as described in Section 2.6.

In addition to learning our models with sequence-based parameters, we consider models that are the same as ours except that the \( \gamma \) and \( \alpha \) parameters are free and unlinked to the genomic sequence. This baseline is effectively the approach of Nachman et al. Another baseline we consider is a simple method that simply predicts that a gene’s expression level for some held-aside test case is its average expression level in training set. Table 1 shows the average relative error for all three methods and both data sets. The top two panels in Figure 4 show the cumulative error distributions for the three methods for both data sets. Each point in a cumulative error curve represents the percentage of test cases that have relative error less than or equal to the associated \( x \)-coordinate. From these results, we can see that the regulatory network models have predictive accuracy that is superior to the baseline. This outcome indicates that the models are capturing some of the real properties of the regulator kinetics and inferring the unmeasured regulator concentrations with reasonable accuracy. Moreover, the results indicate that the sequence-based models provide better predictive accuracy than the models with the free \( \gamma \) and \( \alpha \) parameters. We test the significance of these differences in cumulative error using the Kolmogorov-Smirnov test. The differences in cumulative error between the sequence-based and free-\( \gamma/\alpha \) models are significant at \( P = 0.019 \) for the carbon-shift data set and \( P = 0.001 \) for the CRP-regulon data.

In order to test whether the predictive gain of the sequence-based models is coming from the sequence-based \( \gamma \) functions alone or the sequence-based \( \alpha \) functions alone, we conduct two ‘lesion’ tests. Each lesion test involves running our approach with either the sequence-based binding (\( \gamma \)) or productivity (\( \alpha \)) functions replaced with free parameters (i.e. \( \gamma \) parameters are sequence-based but \( \alpha \) parameters are not, and vice versa). Table 1 and the bottom two panels in Figure 4 show that running our model with both parameter types represented as sequence-based functions results in more accurate predictions than with only one parameter type represented in this way.

3.3 Evaluating model explanatory power

In addition to making accurate predictions (which a ‘black-box’ model may do well), we also wish to gain insight into the underlying regulation properties represented by our learned models. We argue that models with more explanatory power can better help biologists understand observed regulation phenomena and potentially aid in predicting the outcomes of biological manipulations, such as promoter-sequence mutations.

Toward this end, we first evaluate the learned RNAP concentration profile from the carbon-shift data set, for which we have some prior biological knowledge. Figure 5 shows the relative concentration inferred from the model with sequence-based \( \gamma \) and \( \alpha \) parameters, and from the model with free \( \gamma \) and \( \alpha \) parameters. Although both models infer somewhat similar RNAP profiles for glycerol, succinate and alanine relative to glucose, they are dramatically different for the relative levels of acetate and proline. Previous studies (Ishihama et al., 1976; Ishihama, 1981) show that the available RNAP levels are highly similar in growth culture with acetate and proline as the carbon source. The almost 10-fold concentration
increase inferred by the model with free $\gamma$ and $\alpha$ parameters therefore is not in agreement with current knowledge. The RNAP concentration profiles inferred by our sequence-based models are more accurate in this respect.

We also assess the biological significance of the weights learned from our RNAP productivity model. Since we do not have ground truth to directly consult, we compare our learned weights to the weights learned by a regression model trained on a whole-genome in vitro transcription data set produced by the Burgess lab at University of Wisconsin–Madison (K. Zhao, M. Liu, F. Blattner, T. Durfee and R. Burgess, publication in preparation). In these assays, RNA products from transcription reactions using MG1655 genomic DNA and purified *E.coli* RNAP were hybridized to Affymetrix *E. coli* GeneChips. The resulting signal intensity for any given operon is a reflection of the intrinsic strength of the corresponding promoter as no TFs were present to influence RNAP-promoter interactions. The experiment was done in two biological replicates. Normalized signal intensity values for each gene were averaged and genes with intensities $> 9$ (log$_2$ scale) were chosen for further analysis. This cutoff was used to ensure that the selected genes were unambiguously transcribed in the experiment. Next, to limit subsequent analysis to only those genes with experimentally

Table 1. Average relative error of various models on both data sets.

<table>
<thead>
<tr>
<th>Model</th>
<th>Carbon-shift</th>
<th>CRP-regulon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence-based $\gamma$ and $\alpha$</td>
<td>0.309</td>
<td>0.507</td>
</tr>
<tr>
<td>Sequence-based $\gamma$, free $\alpha$</td>
<td>0.354</td>
<td>0.621</td>
</tr>
<tr>
<td>Sequence-based $\alpha$, free $\gamma$</td>
<td>0.368</td>
<td>0.535</td>
</tr>
<tr>
<td>Free $\gamma$ and $\alpha$</td>
<td>0.535</td>
<td>0.851</td>
</tr>
<tr>
<td>Baseline</td>
<td>1.306</td>
<td>2.386</td>
</tr>
</tbody>
</table>

Fig. 4. Cumulative error distributions of various models. Panels on the left and right show 10-fold cross-validation results on carbon-shift data set and CRP-regulon data set, respectively. The top panels show error distributions for the sequence-based models, free-$\gamma$ and $\alpha$ models and baseline. The bottom panels show error distributions for the lesion experiments, with the sequence-based models shown for reference purposes.
defined promoter sequences, the gene subset was compared
against data culled from the EcoCyc database (Keseler et al.,
2005) to identify those with known TSSs. In cases where an
operon contains more than one gene, only the first gene in that
operon was chosen for the analysis. This procedure resulted in
82 promoter sequences.

From this data set we induce various regression models,
using the WEKA library (Witten and Frank, 2005), that map
the promoter sequences to their associated in vitro expression
levels. We represent the promoter sequences using features
extracted by the HMM described in Section 2.4. The features
used to represent these promoters are the same as the ones used
in the RNAP productivity model. Among the best regression
models induced in this study is one based on a linear regression
method that minimizes mean squared error. Figure 6 shows a
scatter plot of expression values predicted by this model versus
the measured expression values. Although the accuracy of the
model is not exceptional, it does have some predictive power.
The correlation coefficient between the predicted and measured
values is 0.65.

We compare the weights from this in vitro model against the
weights learned in the RNAP α model when it is trained on the
carbon-shift data set. The key distinction between these two
models is that whereas the former is given a direct training
signal (in vitro expression levels), the latter must infer its
weights in the context of a more sophisticated model that is
(i) taking into account the role of TFs, (ii) inferring activity
levels of these regulators and (iii) modeling expression in a
range of in vivo conditions. Gene expression levels in E.coli are
determined primarily by the rate of transcription initiation
which involves multiple kinetic steps, including RNAP binding,
open complex formation and promoter clearance. Weights
learned from the in vitro data should in principle capture the
composite impact of a base at a position on the whole process.
Weights in the RNAP α function in our network model, on
the other hand, are expected to represent the same composite
effect minus the effect on binding, which is modeled by the γ
parameter. By comparing the RNAP α model to the in vitro
model, we believe that we can garner some evidence indicating
whether or not our RNAP α model is learning meaningful
weights.

The most contributing base (the one with highest weight) at
each position is shown for both models in Figure 7. We observe
the high similarity of the most contributing bases in −35 region
(5 out of 6) and some level of similarity in −10 region (2 out of
6). These results demonstrate that the weights inferred for the
RNAP α function in our network model have a reasonably high
correspondence to the in vitro weights, and therefore are likely
to have some biological significance.

Figure 7 also shows the standard consensus sequence for σ^70
promoters. Interestingly, neither our RNAP α model nor
the in vitro model learns weights in correspondence with the
consensus. Furthermore, we have evaluated ‘similarity with the
consensus’ as an approach to predicting the in vitro expression

![Fig. 5. Inferred available RNAP concentrations in carbon shift
experiments. Values are scaled such maximum concentration is
one unit.](image)

![Fig. 6. Agreement between predicted and measured expression values
on in vitro transcription profiling data set. The predicted values are
from a linear regression model.](image)

![Fig. 7. The most contributing base at each position of the E.coli promoter −35 and −10 regions versus promoter consensus. The row labeled
In vitro model represents the bases with the largest weights at each position in the linear regression model trained on the in vitro transcription data.
The row labeled Network alpha model represents the bases with the largest weights at each position in the RNAP productivity function, which is
learned in the regulatory-network model. The row labeled Consensus shows the standard E.coli σ^70 consensus sequence.](image)
values and found that it does not predict nearly as well as the linear regression model considered here. One possible explanation for these results is that the consensus sequence characterizes the sequence properties that are more important for sequence binding than for the other steps involved in transcription initiation.

4 CONCLUSION

We have presented our extension to a state-of-the-art approach for inferring quantitative gene-regulatory network models. Our method represents and learns the key kinetic parameters of transcription regulation as functions of features in the genomic sequence. The primary motivation for our approach is to offer more explanatory power than models without these sequence-based functions. Our experiments on two *E. coli* expression data sets indicate several key results. First, models with sequence-based kinetic parameters are more accurate than models without. Second, the models learned with sequence-based kinetic parameters produced more biologically sound activity profiles for active RNAP. Third, the sequence-based $\alpha$ function learned for RNAP captured some aspects of the binding site that are in agreement with a model trained directly on controlled *in vitro* data.

Some limitations of our current models are that we (1) do not learn sequence-based $\alpha$ parameters for promoter states when TFs are bound, (2) do not represent presumably different fractions of RNAP and separate $\gamma$ and $\alpha$ functions for these fractions and (3) have not investigated structured learning in the context of these models. We are currently developing extensions to our approach to address these limitations, which we believe will lead to even more realistic and accurate models.

**ACKNOWLEDGEMENTS**

The authors thank Keith Noto for helpful comments on a draft of this article. This research was supported by the BACTER Institute at UW-Madison, DOE-GTL training grant DE-FG02-04ER25627 (YP), NIH/NLM training grant 5T15LM007359 (TD), NIH/NLM training grant 5T15LM005359 (TD), NIH/NLM training grant 5T15LM005359 (JB), NSF grant IIS-0093016 (MC).

**Conflict of Interest:** none declared.

**REFERENCES**


Bockhorst,J. et al. (2003b) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, 19, (Suppl. 1), i34–i43.


