Specificity and Pleiotropy in *E.coli* Transcription

Jared Markowitz
Massachusetts Institute of Technology Department of Physics, Cambridge, MA 02139, USA

(Dated: May 19, 2007)

Transcription factors play a crucial role in gene expression through their interactions with regulatory DNA elements. Binding between factors and DNA sequences is governed by the balance between sequence accuracy requirements and the need for robustness against mutations. Sengupta *et al.* [1] have used these conditions to derive an analytical relation between the specificity of a given factor and its pleiotropy, or number of binding sites on a genome. In this article we apply the weight matrix formalism as outlined by Stormo and Fields [2] to find the relation between factor/DNA binding specificity and pleiotropy in the *E. coli* genome. We find that the pleiotropy in this case is well predicted by a random sequence for low specificity and may converge to the analytical prediction from [1] for high specificities.

PACS numbers:

I. Introduction

Transcription factors are DNA-binding proteins that affect promotor activity in regions upstream of the regulatory response element (RE). In general transcription factors do not have unique DNA binding sequences, but rather can bind to a small set of similar sequences. Two requirements compete to determine this group of potential binding sites. First, the factor must have a significantly higher affinity for binding to the correct sites than to other (non-specific) sites so acceptable sequences must be exceedingly rare. On the other hand, transcription should be relatively robust with respect to mutation, so factors should be able to account for small changes in their specific DNA counterparts. Sengupta *et al.* [1] show that these two conditions lead to a quantitative relation between the number of potential binding sites (pleiotropy) and the binding sequence specificity. In this work we test this relation using some of the known binding sites on the *E. coli* genome.

The number of potential binding sites in a genome for a given transcription factor can be estimated given a set of experimentally verified binding sequences. The procedure used here follows that presented by Stormo and Fields [2] and involves comparison of the average information content of the observed binding sites with that of other sequences of the same length through weight matrices. It relies on the assumptions of independent nucleic binding energies within a sequence and an overall random genome. The information content method can be applied to both real and synthetic genomes.

This paper is organized as follows. In section II we describe transcription factor binding and the effect of mutations. We follow the steps taken in [1] to arrive at the predicted relation between pleiotropy and specificity. In Section III we describe the information-based approach to binding site prediction given in [2]. We apply this method to known binding sites in the *E. coli* genome in Section IV and discuss our findings in Section V.

II. Transcription Factors, DNA, and Mutations

Transcription is moderated by the binding of transcription factors to the regulatory response elements (REs) of genes. Several factors, each bound to multiple sites, work together to activate/repress gene expression. The binding of factors to REs is governed by the binding energy of the pair and the concentration of the factor. The binding energy of an RE and a factor can be expressed as

$$E(x) = \mathbf{x} \cdot \epsilon \equiv \sum_{i=1}^{L} \sum_{\beta=1}^{4} \epsilon_i^\beta x_i^\beta \tag{1}$$

where ϵ_i^β is the binding energy of the protein with base β at position $i = 1...L$ along the DNA strand and x_i^β is 1 if the base at position i is β and 0 if not. The Central Limit Theorem applied to a sum of L random variables implies that the density of states $\rho(E)$ will have a gaussian form. The equilibrium probability of a sequence x to bind to a factor is given by

$$f(E(x)) = (e^{(E(x)-\mu)/kT} - 1)^{-1} \tag{2}$$

where the chemical potential μ is determined by the factor concentration. In most cases the binding energy scale is large compared to kT, so $f(E)$ can be modeled by a step function. This distribution essentially sets an upper bound (μ) on the interaction energy between an RE and a factor for binding to occur. The location of the chemical potential with respect to the normal distribution of interaction energies determines the binding specificity $\sigma = -\ln(\nu)$, where ν is the fraction of random sequences of length L that will bind to the protein. The specificity must generally be large for regulation to occur efficiently, implying that ν is in the extreme low energy tail of the interaction energy distribution.

The interaction energy of a given RE and protein factor can be altered by mutations, possibly leading to a change

in binding status. If a bound factor becomes unbound or a spurious factor becomes bound, the transcription process could fail. The mutation process is tractable analytically if one assumes a continuous interaction energy, which is a good assumption in the limit of large L and small binding energies. Sengupta *et al.* [1] have derived a diffusion relation that governs the population of bound states in this limit. Defining $n(E,t)$ as the number of bonds with energy between E, E+dE they find

$$\partial_t n(E,t) = \partial_E^2 n(E,t) + \partial_E[En(E,t)] \tag{3}$$

where the time scale is determined by the point mutation rate. The boundary condition $n(E,t)|_{E=\mu} = 0$ implements the hard limit on the binding sites in the limit where f(E(x)) can be approximated as a step function. Note that without this boundary condition we would again obtain $n(E) \approx \exp(-E^2)$, i.e. a gaussian distribution. To determine the net change in $n(E,t)$, one must solve [3] on both sides of the boundary condition. To the left of μ, $n(E,t) \approx \exp(-\kappa_l t) n_\infty(E)$ after long times. Here $\kappa_l = \sigma/2$, so most bound sequences are clustered near the transition point μ. To the right of the boundary condition, $n(E,t) \approx \exp(-\kappa_{sp} t) n_\infty(E)$ after long times, where $\kappa_{sp} \approx (\pi/2)\sigma e^{-\sigma}$. Note that the κ values are the probability per mutation of diffusion across the border, so that although $\kappa_s p << \kappa_l$ the contributions from spurious sequences diffusing into the binding region and binding sequences mutating out of the binding region may be comparable as there are many more sites with $E > \mu$. We can find the relation between the expected number of binding sites and the specificity by setting the probabilities of diffusion in each direction equal and thereby requiring equilibrium. This yields

$$n/N_{cR} \approx \pi\sigma e^{-\sigma} \tag{4}$$

where N_{cR} is the number of non-binding sequences (cR stands for cis regulatory), n pleiotropy, and σ is the specificity. This prediction can be tested by identifying binding sites on a whole genome and comparing with the a random sequence of same base frequency to obtain σ.

III. Weight Matrices, Information Content and Binding Sites

Given a set of known transcription binding sites, we would like to be able to determine their information content and use them to predict other binding sites. We start by defining the binding constant

$$K_{eq} = \frac{[\mathbf{T} \cdot X_i]}{[\mathbf{T}][X_i]} \tag{5}$$

of the reaction $\mathbf{T} + X_i \leftrightarrow \mathbf{T} \cdot X_i$, where T is a protein, X_i is a DNA sequence, and $\mathbf{T} \cdot X_i$ is a bound system. Normalizing over the set of X_i in the genome we obtain the specific binding constants K_s and free energies $\Delta G_s = -RT\ln(K_s)$. This allows us to find the partition function of the genome, which will simply be its length Γ. The probability that a particular protein is bound to a given sequence will then be K_s/Γ.

Experimentally, it would be nearly impossible to determine K_s for all X_i, due to the sheer magnitude of potential sequences. However, we can get around this difficulty by studying only binding sites and applying the weight matrix technique. Given a set of known binding sites S_i of length L we want our weight matrix $\overrightarrow{\mathbf{W}}$ to approximate the matrix $\overrightarrow{\mathbf{G}}$, where $\overrightarrow{\mathbf{G}}$ is a 4xL array of the ΔG_s values for each nucleotide at each position in the sequence. We need two assumptions to make this feasible. The first assumption is that each site in the sequence contributes independently to the total binding energy. This allows our presumption of a matrix that depends only on single nucleotides and also allows us to determine the information in the sites given their frequencies. Specifically, the information will be $\sum_{b,j} f(b,j)\Delta G_s(b,j) \equiv \overrightarrow{\mathbf{f}} \cdot \overrightarrow{\mathbf{G}}$. The second assumption is that the nucleotides are distributed randomly across the genome. This assumption is obviously not strictly true, but in most cases is a good approximation. It allows us to estimate the partition function as the length Γ as mentioned above. Now the approximation for $\overrightarrow{\mathbf{G}}$ (modulo a sign convention) should be that which maximizes the probability that all of the observed sites are bound. This gives [3]:

$$\overrightarrow{\mathbf{W}}(b,j) = \ln\left(\frac{f(b,j)}{p(b)}\right), \tag{6}$$

where $f(b,j)$ is the frequency of the base b at position j in the sample sequences and $p(b)$ is the frequency of the base b in the entire genome. Note that the larger the sample, the less likely that $f(b,j)$ is biased. The average information content of the observed sets is defined by

$$I_{seq} = \sum_j \sum_b f(b,j)\ln\left(\frac{f(b,j)}{p(b)}\right) = \overrightarrow{\mathbf{f}} \cdot \overrightarrow{\mathbf{W}}. \tag{7}$$

This information content is a measure of the average specific binding energy for the observed binding sites. As such, it gives a characteristic binding energy threshold for determining whether the protein in question will bind to a given sequence. Specifically, one can add the terms in the weight matrices corresponding to a trial sequence and compare with the information content of the observed binding sites to determine whether the transcription factor will bind to the trial sequence. This method was applied to estimate the pleiotropy and specificity of several factors on the *E. coli* genome.

IV. Application to *E. coli*

We applied the weight matrix methodology to study the binding of 37 different transcription factors to the *E. coli* genome[12]. The known binding sequences were obtained from RegulonDB [4] and were chosen from a larger set as the factors where at least 10 binding sites were observed. The objective was to test the pleiotropy-specificity relation derived by Sengupta *et al.* [1] by counting the number of potential binding sites for each transcription factor in the *E. coli* genome and then comparing the result to the number of potential binding sites of a randomly generated sequence.

The pleiotropy of a given factor to the *E. coli* genome was estimated as follows. First the weight matrix and average information content of the observed binding sites were calculated according to (6) and (7). The negative of the average information content was taken as the upper bound of the interaction energy for binding, or the chemical potential in the model of Section II. This is clearly an approximation as we are taking the average binding energy of observed sites as an upper bound. However the same approximation was used for both the pleiotropy and the specificity estimates and so should not alter the results significantly. The *E. coli* genome was then searched for sequences of the same length as the observed sites (L) that satisfied $E_{seq} < \mu$, where

$$E_seq = \vec{\mathbf{x}} \cdot \vec{\mathbf{W}} \equiv \sum_{j=1}^{L} \sum_{b=1}^{4} W(b,j) x_j^b. \qquad (8)$$

Here again x_j^b is 1 if the base at position j is b and 0 if not. Note that the genome was searched in both directions and in both strands to find all possible binding sites. The number of candidate sequences from this search was taken as an estimate of the pleiotropy of the protein-genome pair under the assumption that nearly all potential binding sites are occupied.

The specificity in each case was estimated in much the same manner as the pleiotropy. Recalling that the specificity $\sigma = -\ln(\nu)$ where ν is the fraction of random sequences of length L that will bind to the protein allows us to estimate σ by feeding random sequences with the *E. coli* base frequencies to the weight matrix. A potential site is identified as above, i.e. when its E_{seq} is less than the negative of the average information content of the observed sites.

The resulting pleiotropy and specificity values were plotted and compared to the results of Sengupta *et al.* (see Figures 1 and 2). In both cases a linear relationship between the pleiotropy n and the fraction of random sequences that bind to the factor ν were observed. This linear relation was predicted by mutation considerations, but also holds for random sequences. Indeed our results appear to be more consistent with the random energy model than with the predicted mutation-dependent behavior, although there is generally some excess above the random energy prediction. This may still be consistent with the result of Sengupta *et al.*, as their study was geared more toward pairs with high specificity while ours examined more of the low specificity region. It appears from both studies that the mutation-motivated behavior is more closely followed for high specificities and then decays toward the random energy model at low specificities as more binding sites are allowed.

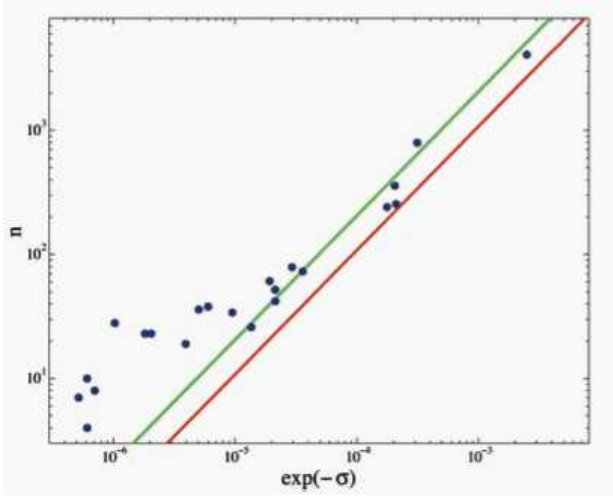

FIG. 1: Plot of the number of (predicted) response elements (n) in *E. coli* genome as a function of binding specificity. Figure reproduced from Sengupta *et al.* [1].

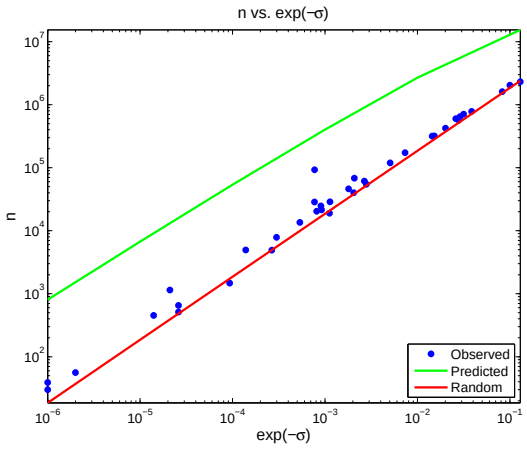

FIG. 2: Plot of the number of (predicted) response elements (n) in *E. coli* genome as a function of binding specificity in the current study. Note that the predicted linear trend is observed, but that sequences appear more random than predicted.

There are a few noteworthy differences between the work of Sengupta *et al.* and that presented here. The offset observed on the n axis between the two plots comes partly from the fact that we studied the full *E. coli* genome, parsing approximately 35 times as much data per factor as in [1]. We also used different databases for the known binding sites, which should produce slightly

different results. Most importantly, the two algorithms for determining the binding threshold energy were different [13]. We used the traditional weight matrix approach, while a "quadratic programming" algorithm designed to reduce the number of false positives was used in [1].

V. Conclusions

In this article we have examined the role of mutations in protein/DNA binding. Specifically we used weight matrices to study the relation between specificity and pleiotropy in the *E. Coli* genome for 37 different transcription factors. Our results were approximately consistent with a random energy model, although we did observe a slight excess likely due to the mutation considerations described in [1]. We have left plenty of room for further study in this matter; obvious follow up studies include a trial with the algorithm used in [1] on our data set, trials at higher specificity, and trials on other genomes. Mutations play a critical role in the functioning of transcription factors, so further studies certainly have merit.

[1] Sengupta, A., Djordjevic M., and Shraiman B. (2002) Specifity and robustness in transcription control networks. *Proc. Natl. Acad. Sci. USA.* **99**, 2072-2077.

[2] Stormo, G.D. and Fields, D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.,* **23**, 109-113.

[3] Heumann, J.M., Lapedes, A.S. and Stormo, G.D. (1994) in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 188-194, AAAI Press.

[4] Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions Nucleic Acids Res. 2006 Jan 1;34(Database issue):D394-7.

[5] O.G. Berg and P.H. von Hippel, J. Mol. Biol. **193**, 723 (1987).

[6] P.H. von Hippel and O.G. Berg, Proc. Natl. Acad. Sci. USA **83**, 1608 1986.

[7] B. Derrida, Phys. Rev. B **24**, 2613 (1981).

[8] U. Gerland, J. D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. USA **99**, 12015 (2002).

[9] Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics,* **16**, 16-23.

[10] Y. Takeda, A. Sarai, and V. M. Rivera, Proc. Natl. Acad. Sci. USA **86**, 439 (1989).

[11] Wikipedia Contributors, 'Transcription Factor', *Wikipedia, The Free Encyclopedia,* 9 May 2007, 01:15 UTC, $< http://en.wikipedia.org/w/index.php?title = Transcription_factoroldid = 129401372 >$ [accessed 19 May 2007].

[12] The source code is available in http://www.ligo.mit.edu/ jaredm/8592/paper/final.

[13] Unfortunately the website describing the algorithm of Sengupta *et al.* has been taken down, so it is difficult to compare the two algorithms.