Assessment of the feasibility of a k-mer based model for binding of the transcription factor E2-2 in humans

Maitagorri Schade
*Department of Physics, Massachusetts Institute of Technology**
(Dated: June 3, 2009)

In this paper a probabilistic model of transcription factor binding, based on sets of short sequence chunks or k-mers influencing the likelihood of binding, proposed by Wunderlich et al., is explored experimentally. Analyzing data of the human transcription factor E2-2 and using a Monte Carlo simulation to find a potential set of probabilistic motifs, I find that the statistical significance of the best sets is not sufficient to support the model for binding of E2-2.

I. MOTIVATION

As Wunderlich and Mirny describe in [1], for specific, purely functional transcription factor binding sufficient information in the binding motif is needed. The information content required depends on the length of the genome: $I_{\min} = \log_2(N)$ for a genome of length N to specify a single target site. In prokaryotes, a typical motif contains 19.8 bits of information, close to $I_{\min} = 22.2$ bits required for complete specificity. In eukaryotes on the other hand, motifs are typically shorter and the genome is longer, yielding an average 12.1 bits per motif, much less than $I_{\min} \simeq 30$ bits. This indicates that eukaryotic TFs function in a very different manner from prokaryotic ones.

Wunderlich et al. suggest a clustering model with several low-specificity binding sites which are enriched in functional binding regions, leading to clustered binding of TFs in such regions. As opposed to a classical motif, a relatively well-defined binding region described by a density matrix, this is a more generalized model of smaller sequences of length k, or k-mers, that show up with a much higher frequency in binding than in non-binding regions. This model should reproduce the motif where one is present but could describe a broader range of situations.

If this model is indeed correct, we expect to see several short motifs (on the order of 5 bp) strongly enriched in regions where a TF binds preferentially. Given a set of sequences with known TF affinity we can thus refute or corroborate the above hypothesis. More specifically, if Δn is the difference in average occurrence of a set of motifs and σ is the standard deviation, we expect $\frac{\Delta n}{\sigma} \gg 1$ for the best set of k-mers if the hypothesis holds true - this allows us to test the feasibility of the model.

It has been shown by A.Tafvizi that the transcription factor Zfx does show significant enrichment, or clustering, of certain k-mers in regions of strong binding, as shown in fig.1.[2] In the following I investigate the feasibility of describing the binding of human E2-2 with a clustering-based model.

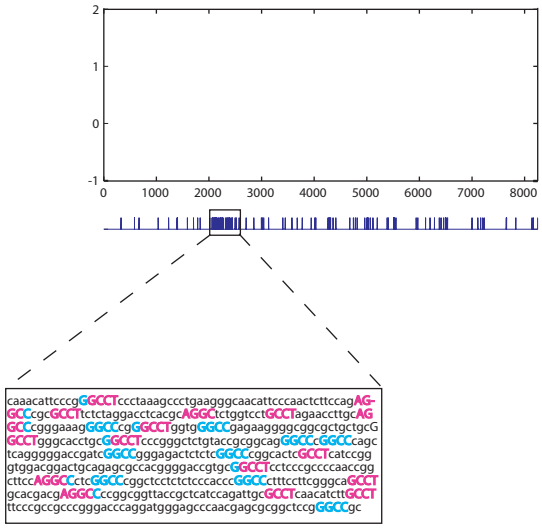

FIG. 1: A diagram of the frequency of occurrence of two k-mers along the genome, aligned with a plot of binding strength at different points of the sequence. The enlarged box shows part of the genome with target k-mers highlighted. [a]

[a]image courtesy of A.Tafvizi and L.Mirny

II. DATA AND METHODS

Data The data at hand is the result of ChIP-chip experiments with the transcription factor E2-2 on the human genome.[3] It consists of ∼378,000 sequence chunks, each consisting of 1000bp and tagged with a measure of binding affinity to E2-2, the so-called xbar. The 0.5% (∼2000) of sequences with the strongest binding are used as a bound or positive set, 2% (∼8000) of the remaining sequences are used as background.

Theory First, for each sequence the number of occurrence of every k-mer for $k = \{4, 5, 6\}$ is counted (I take into account reverse compliments, using a modified version of fasta2matrix [4], dealing with a total of 2728 k-mers). The range of k is chosen to have k-mers that are longer than 3bp for statistical significance and not longer than the typical motif size of 6bp.This gives two matrices, M_{pos} for the positive set and M_{bg} for the background, with the rows corresponding to different sequences and the columns corresponding to different k-mers.

*Electronic address: mschade@mit.edu

In order to be able to quantify the "goodness" of any single motif I introduce a measure of statistical significance of the enrichment of a given k-mer in a sequence:

$$z_i = \frac{n_i - \langle n_{\rm bg}\rangle}{\sigma_{\rm bg}}, \tag{1}$$

where n stands for the number of occurrence of the k-mer in the ith sequence, and $\sigma_{\rm bg}$ stands for the standard deviation in the background.

Now we can describe a set of k-mers as a vector $\mathbf{j}$, containing one element for each k-mer taken into account, 1 for k-mers in the set and 0 for k-mers not in the set. For a given set $\mathbf{j}$ of k-mers, let m_i be the sum of occurrences of all k-mers in the set in the ith sequence. Then $\mathbf{m}_{\rm pos} = \mathbf{j}\cdot \mathrm{M}_{\rm pos}$ and $\mathbf{m}_{\rm bg} = \mathbf{j}\cdot \mathrm{M}_{\rm bg}$. The significance of enrichment of $\mathbf{j}$ in the ith sequence is then

$$z_i(j) = \frac{\langle m_i\rangle - \langle m_{\rm bg}\rangle}{\sigma_{\rm bg}}, \tag{2}$$

and the overall significance of enrichment in the positive set versus the background is

$$z(j) = \frac{\langle m_{\rm pos}\rangle - \langle m_{\rm bg}\rangle}{\sqrt{\sigma^2_{\rm pos} + \sigma^2_{\rm bg}}}. \tag{3}$$

Assuming that a set $\mathbf{j}_0$ is responsible for the binding of the TF, the statistical measure is correlated to the affinity of the TF to sequence i and can thus be used to define a binding energy $E(i) \propto -z_i(\mathbf{j}_0)$ for the ith sequence. In this case, we expect the overall $z(\mathbf{j})$ to be maximized for $\mathbf{j}_0$.

Monte Carlo In order to find the set of k-mers of the correct size, since it is impossible to exhaustively sample the 2^{2728} different possible constellations given by all 4-6-mers, I perform a Monte Carlo simulation to find the optimal set of k-mers.

Since it does not seem feasible to have a motif set with larger order of magnitude than 10, we introduce a weight lambda against infinite addition of further mers to get a modified z:

$$z_{\rm mod}(j) = z(j) - \lambda \cdot N \tag{4}$$

where N is the size of the motif set.

In addition, we also introduce a temperature τ which allows for random fluctuations against the gradient.

The scheme for the algorithm used is outlined below:

1. pick initial set [generate $\mathbf{j_0}$ with 10 random elements]
2. explore first step [generate $\mathbf{j_1}$ by randomly switching 10 elements of $\mathbf{j_0}$]
3. calculate goodness of each set [$z_0 = z_{\rm mod}(\mathbf{j_0})$, $z_1 = z_{\rm mod}(\mathbf{j_1})$]
4. decide whether to take step
   [if $z_mod(j_1) > z_mod(j_0)$:
   accept [$\mathbf{j_0} = \mathbf{j_1}$]
   else:
   if random number between 0 and 1 $< \mathrm{e}^{\frac{z_1 - z_0}{\tau}}$:
   accept [$\mathbf{j_0} = \mathbf{j_1}$]]
5. explore next step [$\mathbf{j_1} = \mathbf{j_0}$ with between 3-10 elements randomly switched]
6. go back to 3.

The value of λ is set by testing the algorithm against blowing up of the number of kmers: the smallest lambda that ensures a stable final number of k-mers included is chosen. The value of τ is set by testing various values, using the largest one for which the algorithm still consistently converges. This algorithm is first tested starting with a set containing the 10 k-mers with the best individual z-scores, then run 100 times for 200,000 iterations on truly random initial sets.

III. RESULTS

I find that the Monte Carlo algorithm always converges to a similar set of k-mers, containing an assortment out of ∼25 different k-mers. The final number of k-mers included varies between 8 and 12, as expected centered around 10. The maximum value of z is always around 0.66. The maximum z ever found is 0.67, for the following set of kmers:

{AACT, AGGAA, CACGAA, CCACC, CGAC, CAGCAA, CAGGAA, CGGAC, ACTTCG}

All of these 9 k-mers also appear in other optimal outputs of the algorithm.

$z = 0.67$ is significantly less than 1 and does certainly not meet the significance criterion given above ($z \gg 1$). Furthermore, the histogram of $E \propto -z_i$ for both the bound set of sequences and the background for this k-mer-set is given in fig.2. Given the fact that our $\mathbf{j}$ was selected for a maximum difference between the average energy in these two sets, the difference between the two distributions is not sufficient to corroborate our hypothesis.

IV. DISCUSSION OF RESULTS

The fact that the motif set with the highest difference in enrichment only reaches a statistical significance $z < 1$ indicates that our k-mer based model of transcription factor binding does not describe E2-2 in humans well. More refined versions of k-mer-set finding, such as weighting each k-mer included with a weight between 0 and 1, may

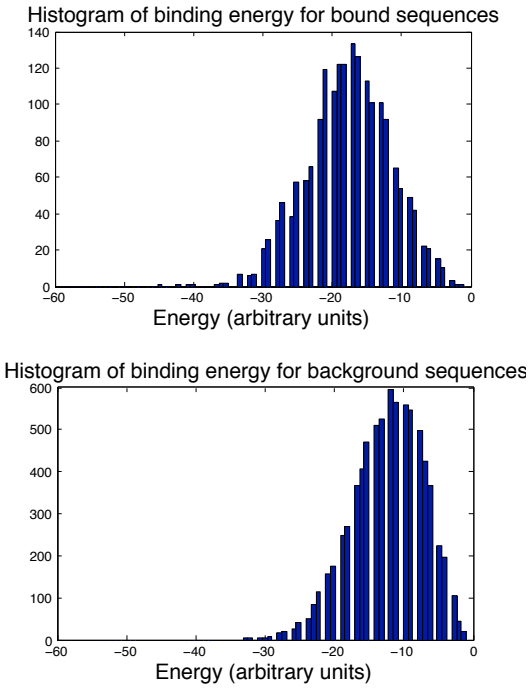

FIG. 2: Histograms of binding energy for "bound" sequences and background.

lead to slightly improved results, but will not yield drastic improvements of statistical significance and thus not change the feasibility of the model. Including more k-mers into the motif set (on the order of hundreds) would amount to an overfitting of the data and is therefore not desirable.

Since it has been shown that in Zfx an approach similar to mine yields a strong correlation between binding strength and frequency of an optimized set of k-mers, it is possible that the model is still useful in some cases. Zfx has a binding domain of about 60bp distributed over 13 zinc fingers, whereas E2-2 has a shorter and more localized binding domain.[5] It is thus possible that the proposed model of binding is much more useful in TFs with large and "fuzzy", i.e. not very localized, sites of binding.

I thus conclude that the probabilistic clustering model for transcription factor binding proposed above is not applicable to E2-2, possible because of a failure of the model or because the model only applies to certain types of TFs. It still remains uncertain how human TFs in general find their unique binding sites on a genome that is too long for their typical motif size.

[1] Z. Wunderlich, L. Mirny. "Fundamentally different strategies of gene regulation revealed by analysis of binding motifs"
[2] A. Tafvizi. Work in progress, Mirny Lab, Harvard-MIT HST.
[3] Fraenkel et al. "Transcriptional regulatory code of a eukaryotic genome", Nature 431, 99-104
[4] Gupta et al. "Supplementary data for 'Predicting human nucleosome occupancy from primary sequence'", PLoS Computational Biology. 4(8):e10000134, 2008
[5] According to Leonid Mirny, Harvard-MIT HST.