Submitted to the Annals of Applied Statistics
arXiv: math.PR/0000000

A MODEL FOR SEQUENTIAL EVOLUTION OF LIGANDS BY EXPONENTIAL ENRICHMENT (SELEX) DATA

By Juli Atherton[*,†,‡,§,**], Nathan Boley[*,¶,‖] Ben Brown[¶,‖] Nobuo Ogawa[¶,††] Stuart M. Davidson[¶,††] Micheal B. Eisen[¶,††] Mark Biggin[¶,††] and Peter Bickel[‡,‖]

University of California Berkeley [‖]*, McGill University* [**] *and Lawrence Berkeley National Laboratory*[††]

A Systematic Evolution of Ligands by EXponential enrichment (SELEX) experiment begins in round one with a random pool of oligonucleotides in equilibrium solution with a target. Over a few rounds, oligonucleotides having a high affinity for the target are selected. Data from a high throughput SELEX experiment consists of lists of thousands of oligonucleotides sampled after each round. Thus far, SELEX experiments have been very good at suggesting the highest affinity oligonucleotide but modeling lower affinity recognition site variants has been difficult. Furthermore, an alignment step has always been used prior to analyzing SELEX data.

We present a novel model, based on a biochemical parametrization of SELEX, which allows us to use data from all rounds to estimate the affinities of the oligonucleotides. Most notably, our model also aligns the oligonucleotides. We use our model to analyze a SELEX experiment containing double stranded DNA oligonucleotides and the transcription factor Bicoid as the target. The results of this SELEX experiment are used in combination with in vivo DNA binding data to improve detection of putative recognition sites for Bicoid in the genome of Drosophila melanogaster.

1. Introduction.

Transcription factors are proteins that regulate gene transcription of DNA by binding to DNA sequence motifs within the genome. Mapping these DNA recognition sequences, and determining the relationship between DNA sequence and transcription factor binding affinity, is central

[*] Joint first authorship

[†] Corresponding author

[‡] Supported by NIH−R01GM075312

[§] Supported also by NSERC grant RGPIN 356107-2009

[¶] The in vitro and in vivo DNA binding data were funded by the U.S. National Institutes of Health (NIH) under grant GM704403 (to MDB and MBE). Work at Lawrence Berkeley National Laboratory was conducted under Department of Energy contract DE-AC02-05CH11231

AMS 2000 subject classifications: Primary 92B15, 62P10; secondary 92B05

Keywords and phrases: SELEX, Transcription factor binding

to understanding the regulation of gene expression. Transcription factors comprise approximately 8% of the genes encoded in the human genome. A comprehensive understanding of the behavior of these proteins will aid in our understanding of key developmental processes including body patterning, brain development, and tissue specification.

One assay, known as Systematic Evolution of Ligands by EXponential enrichment (SELEX) measures the affinity of transcription factor binding to DNA. SELEX was introduced in the 1990's by (Tuerk & Gold 1990) and (Ellington & Szostak 1990). It has been used in a number of genomic studies (e.g. (Kim et al. 2003) and (Freede & Brantl 2004)) and for the purposes of drug discovery (e.g. (Guo et al. 2008) and (Ng et al. 2006)). In genomic studies, SELEX has been used to identify the highest affinity recognition sequences for target proteins.

Analytical methods and algorithms for analyzing SELEX data have been developed by (Djordjevic & Sengupta 2006), (Djordjevic 2007), and most recently by (Zhoa et al. 2009). We have developed analytical methods and algorithms that can be applied to extant SELEX data sets. Like (Zhoa et al. 2009) our approach also starts with the Djordjevic model but diverges sharply from (Djordjevic & Sengupta 2006), (Djordjevic 2007), and (Zhoa et al. 2009) as we shall describe.

The model presented in this paper is the result of a collaboration to analyze the results of SELEX experiments for many different transcription factors from the Berkeley Drosophila Transcription Network Project (BDTNP). It has been validated by comparison with in vivo enrichment in Chromatin ImmunoPrecipitation on chip (ChIP-chip) experiments. For the application of this paper we present results from a single transcription factor, Bicoid. We have chosen to explain the results from Bicoid in detail because it has been studied extensively in the literature and we have multiple replicates of both the SELEX experiments and the ChIP-chip experiments.

1.1. *The SELEX Assay.* A typical SELEX experiment begins in round one with a solution of random double stranded DNA oligonucleotides and a target protein. In the application presented in this paper, the oligonucleotides are 16 base pairs long sequences and are flanked by additional DNA sequences.

The oligonucleiotides react with the transcription factor and eventually a dynamic equilibrium is reached where the concentrations of bound oligonucleotides, unbound oligonucleotides and unbound target are constant. After equilibrium is reached, the oligonucleotides are separated from the solution. Next, polymerase chain reaction (PCR) is performed on the collected oligonucleotides. PCR chemically amplifies the quantity of DNA present in a way that does not significantly change the frequency distribution of oligonucleotides. At this point, a sample is taken for sequencing, and the remaining oligonucleotides are entered into round two. The main steps for round one of SELEX are depicted in Figure 1.

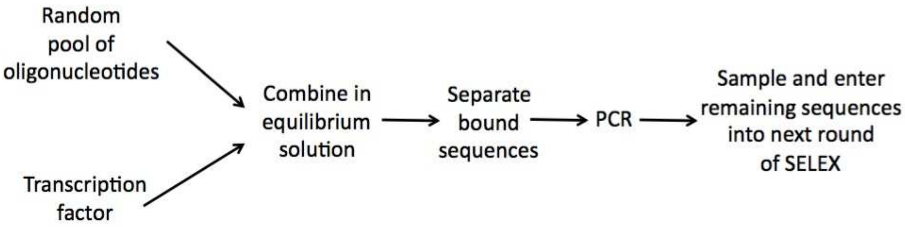
Random pool of oligonucleotides
Transcription factor
Combine in equilibrium solution
Separate bound sequences
PCR
Sample and enter remaining sequences into next round of SELEX

Fig 1. *The main experimental steps for round one of a SELEX experiment.*

Round two of SELEX proceeds exactly as round one, except that the initial pool of oligonucleotides is the set of bound oligonucleotides from round one which went through PCR but were not sequenced. Thereafter, the assay proceeds as before: the oligonucleotides react with the transcription factor and, after equilibrium is reached, the bound oligonucleotides are selected and PCR is performed. A sample is taken for sequencing and the remaining oligonucleotides are entered into round three. These steps are repeated for as many rounds as the experimenter desires.

We observe the outcome of a SELEX experiment by sequencing the oligonucleotides that are sampled at the end of each round. That is, after performing the assay, the results are a list of sequenced oligonucleotides (see Figure 2) and usually meta-data such as, the SELEX round in which each oligonucleotide was sequenced, the concentration of unbound transcription in a particular round, and/or the temperature at which the experiment was performed.

We are interested in modeling the affinity of oligonucleotides that bind in a sequence specific manner to the target. Specific binding involves hydrogen bonding, van der Waals interactions, and other short-range forces. Sequence independent binding also occurs. This is due in part because oligonucleotides bind weakly via electrostatic forces, see (von Hipple 2007), and because a small percent of DNA will non-specifically associate with the bead or non-DNA binding surfaces of the target. Thus even, oligonucleotides that do not bind to the target specifically can be present in later rounds.

Our model has three features which separate it from (Zhoa et al. 2009) and (Djordjevic & Sengupta 2006).

TCCCATTAATCCCACC	2
GGTGTCGGTTTAAGCG	2
CTGATTAATCCGAGTG	1
TGAGATTCCATACCCT	1
TGTGAGGATATGTTTC	1
TGGGGTTGGATTAAAG	1
GGATTAGGGTTAAGCA	1
GACCCCGGCCTAATCC	1
GGTAATCTCGGGATTA	1
TGGACGGATTACGCGG	1

FIG 2. *Example of first 10 sequences (out of 1324 sequences) and their frequencies collected after the third round of a SELEX experiment for the transcription factor Bicoid.*

1. We propose that binding occurs at a unique subsequence of each oligonucleotide which is not determined in advance.
2. We provide for the possibility of non-specific binding which is unrelated to the biochemical process.
3. We permit information from all rounds to contribute to our likelihood.

The first modification permits us to align while simultaneously fitting our model. That is, unlike previous models for SELEX, an alignment step is not required prior to using our model. The third modification allows us to use data from all SELEX rounds. Like (Zhoa et al. 2009), we build into our model the possibility of basing inference on only a sample of sequences in each round.

1.2. *Binding Sites Within Oligonucleotides.* Recall the double stranded DNA in our SELEX experiment contains a random insert of length k surrounded by flanking sequences. We refer to the random component of length k as an oligeonucleotide S.

One of the primary difficulties in the analysis of SELEX is that the target protein may bind to an oligonucleotide in one of many possible configurations. Furthermore the binding site of the target is of length l and is typically less than k. In the application of this paper the oligonucleotides are $k = 16$ base pairs long, but Bicoid binds to sites that are subsequences of length $l = 10$ consecutive basepairs.

We refer to such a subsequence as a *binding site* b. To illustrate this point, in Figure 3, we show an oriented double stranded DNA and its seven distinct binding sites. Each of the $k - l + 1$ binding sites has at most four possible sequence names associated to it, by orientation and strandedness. Once the transcription factor has bound to a binding site, we refer to the bound state as a *binding configuration*.

```
3'   GTTTATAATCCGCGTC    5'
     CAAATATTAGGCGCAG

1       GTTTATAATC
2        TTTATAATCC
3         TTATAATCCG
4          TATAATCCGC
5           ATAATCCGCG
6            TAATCCGCGT
7             AATCCGCGTC
```

FIG 3. *Possible binding sites in an oligonucleotide of length 16 which has a high affinity for Bicoid.*

It is important that we specify a binding site by a sequence name. For example, binding site 4 is represented in Figure 3 by `TATAATCCGC`, but could also be represented by its reverse complement, `GCGGATTATA`. Please see Appendix A.2 for further discussion.

2. The Model. Before presenting our model, we introduce some chemistry which will aid in the parameterization of our model. We focus on a quantity called the change in *Gibbs free energy*, ΔG, of a reaction. In Section 2.1 we explain how ΔG can be estimated from a SELEX experiment. In Section 2.2 we express our likelihood in terms of ΔG. Finally, in Section 2.3, we present a parametrization of ΔG in terms of the nucleotide sequence of a binding site.

2.1. *Chemical Concepts.* The concepts introduced here can be found in (Atkins 1998). We begin by considering many copies of a single oligonucleotide species S in solution with a transcription factor TF. Furthermore we assume that S and TF always bind in the same configuration.

When S and TF are entered into solution with one another they will react to form the product $TF\colon S$. We call this the *forward reaction*. The product $TF\colon S$ will also disassociate into S and TF; we call this the *backward reaction*. The following chemical equation,

$$TF + S \rightleftharpoons TF : S$$

represents these reactions. The solution is said to be in *dynamic equilibrium* when the forward rate of reaction equals the backward rate of reaction. A dimensionless physical constant quantifying the dynamic equilibrium is the *equilibrium constant* K. Our interest in K is that it relates directly to the change in Gibbs free energy, ΔG, for the reaction. The quantity ΔG quantifies the affinity of S for TF. Hence, in Section 2.2, we parameterize our SELEX model in terms of ΔG.

Letting R represent the ideal gas constant and T the temperature in Kelvins, we have

$$K = \exp\left(-\frac{\Delta G}{RT}\right). \tag{2.1}$$

As we shall see below, K is unidentifiable without meta data.

The forward rate of reaction is proportional to the product of concentrations of the reactants. The *forward rate constant*, k_f, is the proportionality constant. Hence,

$$\text{Forward rate} = k_f[S][TF] \tag{2.2}$$

and similarly

$$\text{Backward rate} = k_b[TF\colon S]. \tag{2.3}$$

At equilibrium, equating (2.2) and (2.3) gives the following expression for the equilibrium constant K.

$$K = \frac{k_f}{k_b} = \frac{[TF : S]}{[TF][S]} \tag{2.4}$$

We can think of K as an expected value where the ‘concentrations’ are averages over time and space. In principle, we can use the *observable* concentrations $[\widehat{S}]$, $[\widehat{TF}]$ and $[\widehat{TF} : S]$ to estimate the theoretical physical quantity K and in turn ΔG (via (2.1)). However, we do not have direct access to these quantities and instead must make further estimates.

In SELEX, we have multiple oligonucleotide species in solution. We use S_i to represent the i^{th} species that is in solution. At dynamic equilibrium, the probability of any copy of S_i being bound at a particular instant is equal to the expectation of the fraction of S_i that is bound at that instant. We define this expectation to be $t(S_i)$, which we can write in the same spirit as K as:

$$t(S_i) = \frac{[TF : S_i]}{[TF : S_i] + [S_i]}. \tag{2.5}$$

At this point we make three assumptions concerning specific binding.

1. All members of the same oligonucleotide type bind at the same subsequence. We refer to this subsequence as the binding site.
2. This subsequence is assumed to be of fixed length and independent of the oligonucleotide type in which it is contained.
3. The binding site for each oligonucleotide type is that subsequence which has maximum affinity according to the proposed model.

These assumptions correspond to the hypothesis that the ΔG for a binding site is independent of nucleotide bases surrounding the binding site, that all members of the same oligonucleotide type bind in exactly the same way, and that there is no variability in the subsequence chosen as a binding site. They are implicit in all previous approaches since candidate binding sites within oligos are found through alignment to a consensus sequence.

Given these assumptions we can use (2.1), (2.4) and (2.5) to write

$$t(S_i) = \frac{[TF]\exp(\frac{-\Delta G(S_i)}{RT})}{1+[TF]\exp(\frac{-\Delta G(S_i)}{RT})}. \tag{2.6}$$

where $\Delta G(S_i) \equiv \Delta G(b(S_i))$ and $b(S_i)$ maximizes $\Delta G(b)$ among all b of the length l we have specified contained in S_i.

2.2. *Modeling SELEX.* We define $t_r(S_i)$ to be the conditional probability that a particular molecule of the species S_i is bound at the end of round r given that it is present at the beginning of round r. Formally,

$$t_r(S_i) \equiv P[S_i \text{ bound at the end of } r \mid \text{it is present in } r]. \tag{2.7}$$

Physically, (2.7) is the expectation of the fraction of S_i that is bound at equilibrium in round r. This is precisely the quantity $t(S_i)$, as defined in (2.6). Then,

$$\widehat{t_r}(S_i) = \frac{\widehat{[TF]}_r\exp(\frac{-\Delta G(S_i)}{RT})}{1+\widehat{[TF]}_r\exp(\frac{-\Delta G(S_i)}{RT})} \tag{2.8}$$

is an estimate of $t_r(S_i)$.

We intend to pursue more sophisticated models in which, although only one $b \subset S$ is bound at any moment, our second and third assumptions fail and a suboptimal site can bind instead of the optimal one. Our formulation here resembles that of (Zhoa et al. 2009). The thermodynamic formulation of both models include competitive binding between oligonucleotides S_i. An important difference is that we search all possible binding sites of each S_i for the optimal site. Thus our model takes alignment into account implicitly.

The structure of t_r reveals that the $\Delta G(b)$s are not directly identifiable without knowledge of $[TF]_r$. This is because t_r is unchanged by rescaling all the $\Delta G(b)$s and $[TF]_r$ by the same constant. However, with the given data, we can always estimate

$$\Delta\Delta G(b) = \Delta G(b) - \Delta G(b_o)$$

where b_o is a reference binding site such as a consensus sequence. Of course, if we have meta data such as $[TF]_r$ we can estimate $\Delta G(b)$.

Next we express the distribution of bound sequences in terms of (2.8). We first assume that each sequence is present in an initial concentration C_0 in round zero. We then make the assumption that each PCR step replicates each molecule of S_i A_r times on average in round r. Then, after R rounds of selection, the concentration of S_i in solution is

$$[S_i] = C_0 \prod_{r=1}^{R} A_r t_r(S_i).$$

Dividing the total concentration of S_i after round R by the total amount of all sequences after round R gives an estimate of the frequency distribution of bound sequences at the end of round R. Formally,

$$P_R(S_i) = P[\, S_i \text{ is sequenced in round } R\,] = \frac{\prod_{r=1}^{R} t_r(S_i)}{\sum_{allS_j} \prod_{r=1}^{R} t_r(S_j)}. \tag{2.9}$$

We note that this description of the SELEX assay fails to account for any variance generated during amplification by PCR. It also fails to correct for the case in which zero oligonucleotides of a particular species are bound in round r. That is, we do not treat SELEX as a birth and death process. However, the large oligonucleotides counts makes this a reasonable approximation. For instance, in the data we study in Section 3, each species was present in approximately 65,000 copies in round zero.

It is possible for oligonucleotides to make it though the selection step via a variety of mechanisms, including non-sequence mediated protein-DNA interaction (non-specific binding), DNA-DNA interactions, or DNA-apparatus interactions (experimental error). We account for such sequences in our model, and refer to the effects that result in their selection collectively as *Junk Binding*. If c_J is a constant between 0 and 1, then we can modify our equations to allow for junk binding as follows:

$$t_r(c_J, S_i) = \left((1 - c_J)t_r(S_i) + c_J\right).$$

Finally, we model the portion of the bound DNA that is taken from the pool after the PCR step and sequenced as a simple random sample with replacement. Under this model, if we observe each of the n unique species S_i exactly l_{ir} times in round r, $r = 1, \dots, R$, then, drawing on (2.9), the likelihood is expressed as

$$L(\Delta G | l_{11}, \dots, l_{nR}) = \prod_{r=1}^{R} \left(\prod_{i=1}^{n} P_r(S_i)^{l_{ir}} \right). \tag{2.10}$$

Details of the numerical optimization of (2.10) is discussed in Appendix B.

2.3. *Binding Model.* The *binding model* is the relationship between the actual DNA sequence of a binding site b and the free energy. So far we have formulated our model in complete generality with respect to the binding model. The most widely applied model is an additive one. Such a model assumes that each basepair of DNA makes some contribution to the total binding affinity independent of all other basepairs in the binding site. Representing the nucleotide base pair at position j in b as o_j, we write

$$\Delta G(b) = \sum_{j=1}^{l} \sum_{t \in \{A,C,G,T\}} \lambda_{jt} \varepsilon_t(o_j) \tag{2.11}$$

where

$$\varepsilon_t(o_j) = \begin{cases} 1 & \text{if } o_j = t \\ 0 & \text{otherwise} \end{cases},$$

l is the length of the binding site and λ_{jt} are parameters to be estimated.

It is important to note that our additive model (2.11) does not correspond to a Position Weight Matrix (PWM). It permits considerable dependence between positions and multiple modes well separated in hamming distance. The primary reason for this is that we assume that the binding of an oligonucleotide is determined by a smaller binding site. If we group the oligonucelotides by the binding sites that give minimal free energy we see that the distribution of binding probabilities over oligonucleotides is a mixture of probability distributions each of which could be characterized by PWM.

A minor reason why we are not dealing with an independence model even when the oligonucelotide and binding site coincide is that the effects of single based pair members of binding sites are assumed to add on the log odds scale, rather than log probabilities.

3. Application and Model Validation. The Berkeley Drosophila Transcription Network Project (BDTNP) has generated SELEX and ChIP-chip data for Bicoid. ChIP-chip data measures the genome wide relative levels of occupancy for a single protein of interest. We used the BDNTP ChIP-chip data and a simple, non-parametric method to validate and compare our Bicoid motifs with a motif derived from MEME and several motifs from the literature (Bailey et al. 2006), (Segal et al. 2006) and (Berman et al. 2004).

The ChIP-chip experiments identified thousands of genomic regions to which Bicoid binds. This data has been shown to provide a quantitative measure of relative occupancy. That is, regions can be assigned a score, and those scores have been shown to be reproducible between biological replicates (Li et al. 2008) and (MacArthur et al. accepted). From these and other observations, the authors concluded that the high scoring regions correspond to those with the highest net occupancy of bound factor.

Because of the complexity of intracellular processes, a binding model alone does not provide enough information to predict the results of ChIP-chip experiment. For instance, without additional data, we have no way of modeling the inhibitory affect of chromatin structure. However, we can still use the identified binding regions to test the validity of our SELEX model and data.

If a binding model is identifying true in-vivo binding sites, then we expect the number of high affinity sites predicted by our model to be higher near ChIP-chip peaks. Roughly, we compared the binding models by measuring the enrichment of identified binding sites as compared to the genomic background. There were several variables that we controlled for; we explain the method in detail in Appendix C. We plotted the results of this analysis for ours and competing motifs in Figure 4.

4. Conclusion. The model presented here attempts to infer a comprehensive map of the sequence specific binding affinities between double stranded DNA and a transcription factor from a SELEX experiment. There exist a variety of assays, including ChIP-chip, that attempt to measure the average binding behavior of a protein in a population of cells. However, only assays like SELEX can provide precise models of protein/DNA interactions for downstream models of transcriptional control. We conclude with some observations about SELEX experiments in Section 4.1 and specific comments about our model in Section 4.2.

4.1. *General Comments About SELEX.* Two points that caused the most difficulty in our analysis are the unknown amount of transcription factor and the alignment. As discussed, not knowing $[TF]_r$ causes an identification problem in our model. It is important to note that although, we may be

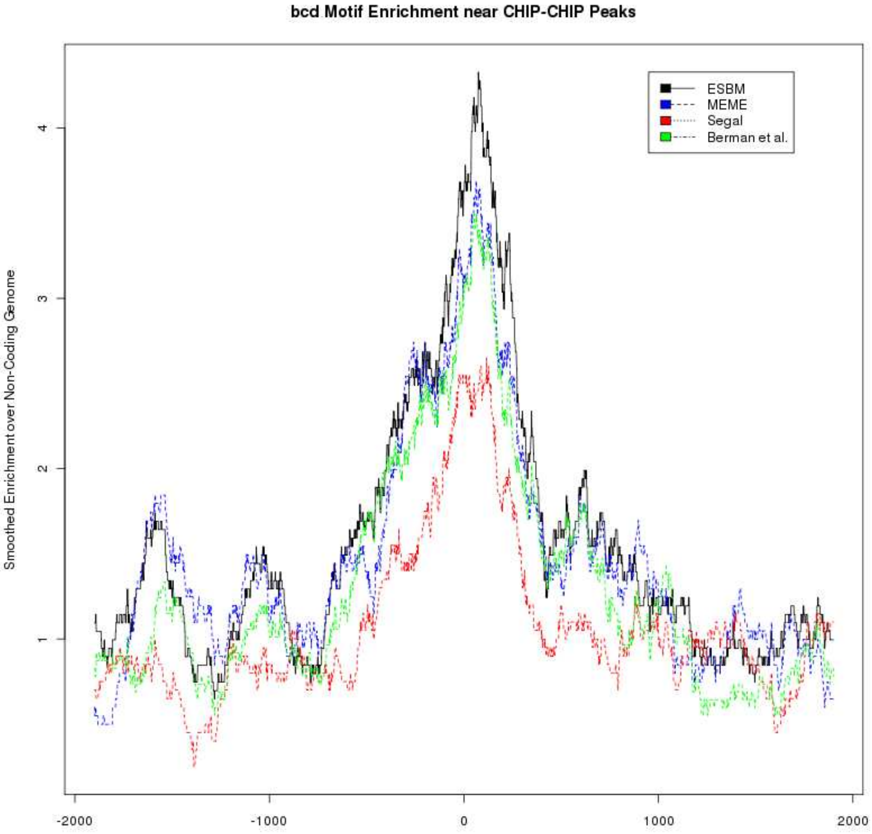

Fig 4. *Enrichment of predicted binding sites for several models at ChIP-chip binding sites. The legend is as follows: ESBM represents the model discussed in this paper, MEME represents (Bailey et al. 2006), Segal represents (Segal et al. 2006) and Berman et al. represents (Berman et al. 2004). The fixed parameters (as described in Appendix C) for the analysis of the ChIP-chip data are $n_p = 100$, $w_s = 4000$, $n_s = 100$, and $s_t = 0.999$.*

able to measure the amount of bound TF, it still may not be possible to estimate the *active* concentration of unbound TF because we do not necessarily know how much of the TF that was added was well folded. Therefore from an experimental perspective, measuring $[TF]_r$ can be quite difficult. We have overcome the identifiability problem caused by not knowing the $[TF]_r$ by working with $\Delta\Delta G$ instead of ΔG with remarkable success as illustrated by our results in Section 3. Of course, if $[TF]_r$ is known we could estimate ΔG directly.

The second difficulty has been dealt with by previous authors using a pre-alingment step prior to estimating the free energy. Our model presents a novel approach which aligns the sequences while estimating the free energy. It is possible that the difficulty of alignment may also be addresses experimentally by a serialization of the SELEX assay. An initial SELEX assay could be conducted with long oligonucleotides. Analysis of this data, under our model via maximum likelihood, would provide insight into the length of the binding site of the protein of interest. A second assay could then be performed with oligonucleotides of the length discovered in the first. Many other variations of the SELEX assay are discussed in (Stoltenburg et al. 2007). We are also optimistic that our maximum energy binding site model can be extended to allow multiple binding configurations.

4.2. *Optimization of our Model.* With regards to the model presented in this paper, a point not yet discussed is the difficulty of maximizing the likelihood. If the amount of transcription factor is "small" then the denominator in (2.8) can be approximated by one, the likelihood (2.10) simplifies, and the optimization between each alignment step can be reformulated as a convex optimization problem. However, since we avoid making this simplification the optimization is involved. Please see Appendix B for more discussion.

For $k = 16$ the number of oligonucleotide types in the initial random pool is $2^{15} + 4^{15}$. It is infeasible to to include all oligonucleotide types in the denominator of (2.9). We estimate the denominator using Monte Carlo and take a simple random sample of oligonucleotides by selecting nucleotide base pairs from a uniform distribution. As discussed in the paper we are assuming in our likelihood that all oligonucleotide types are present in each round. In reality some oligonucleotides are 'lost' between rounds. We suggest that sequencing from both the bound and unbound oligonucleotides after each round may be useful for deciding which oligonucleotides types to include in the denominator.

4.3. *General Comments.* Some final short comments concerning our SELEX model are:

1. Although we have used a simple additive model for ΔG, preliminary results suggest that the predictive power of our model can be increased substantially by allowing for basepair dependencies in the energy model. This is a point of future research.
2. Regarding the junk binding term we found that adding C_J into the model helps to identify the motifs of some transcription factors but not for all of them.
3. (Djordjevic 2007) and (Roulet et al. 2002) suggest that using data with a range of affinities improves inference from SELEX experiments. Having the flexibility to use data from as many rounds of the SELEX experiment as we desire provides us with a nice mix of medium to high affinity sequences from which to make our inference about ΔG.

We conclude by pointing out that the simple thermodynamic and binding assumptions on which our model is based is a crude approximation to the complex process which actually occurs. In Section 3, we have shown that our model provides surprisingly good predictions of genomic regions which are enriched for a given transcription factor in in vivo ChIP-chip experiments. This is surprising since in vivo, not only is a given transcription factor competing with others but also a single stretch of genome rather than many oligonucleotides is the venue of the reactions. These predictions are much better than ones obtained using PWMs and other methods.

APPENDIX A: IDENTIFIABILITY

There are three types of lack of identifiability in the SELEX model. The first type has already been discussed in Section 2.2. The other two types are discussed below.

A.1. Identifiability in the Additive Binding Model. Physically, we are able to identify the total binding affinity of a binding configuration but not the contributions of the individual basepairs. To solve this, we choose to fix the energy of the highest affinity basepair in each position except one to be zero. Then, the value of the first position's highest energy basepair is interpretable as the binding affinity of the 'consensus sequence', or the modeled highest affinity binding site. Some care is needed in ensuring that this constraint does not interfere with whatever optimization algorithm is chosen - such concerns are discussed in the code's comments.

A.2. Identifiability of the Binding Site Names. The third identifiability problem is briefly addressed in Section 1.2. It is present in any binding model which represents binding sites by their sequences. For any segment

b of a double stranded DNA sequence there are four possible names. To ensure that the paramterization is physically meaningful, all of the binding sites must be represented by the same sequence. For example, Bicoid has a high affinity for sequences that contain the subsequence $TAATCC$. As can be seen in Figure 3 it is possible to align the full sequences by the subsequences that are closest to $TAATCC$ in the hamming sense. If, for instance, one were to name half of the subsequences by $TAATCC$ and half by $ATTAGG$ then the likelihood would not optimize properly. This being said, it is irrelevant which name is chosen, as long as it is consistent. For instance, the subsequence $TAATCC$ could also be called $CCTAAT$, $ATTAGG$ or $GGATTA$. For the binding model presented in Section 2.3, the likelihood will be symmetric with four identical modes, each corresponding to a different naming scheme for the strongest binding site. Which of the names our code chooses is chosen, arbitrarily, to be the one with the consensus sequence that is first alphabetically.

APPENDIX B: NUMERICAL OPTIMIZATION

There are substantial computational and algorithmic difficulties in fitting the model. Standard optimization techniques are often ineffective because the likelihood surface is neither convex nor differentiable. In particular, the lack of continuous derivatives makes gradient descent methods like Broyden-Fletcher-Goldfarb-Shanno (BFGS) (Nocedal & Wright 2006) unstable. In addition, the lack of convexity means that line search methods (Nelder & Mead 1965) tend to become trapped in local maxima. In view of these considerations, we have had success using downhill simplex methods (Powell 1964) from a large set of random starting locations. This method is, empirically, stable; we have provided a software tool see, Supplement A, which implements this method. The Bicoid motif was discovered with this tool.

APPENDIX C: DESCRIPTION OF CHIP-CHIP COMPARISON

We chose the n_p highest scoring regions identified through ChIP-chip. For each of those regions, the authors of (MacArthur et al. accepted) defined a "peak", to be a single point in the genome where the local signal achieves its maximum. Around each peak, we examined a symmetric interval of fixed size $2w_s$. Within each of these intervals we evaluated the relative affinity, under some model (e.g. our fitted SELEX model), of each subsequence of length l as determined by the model in question. We utilize these scores, after some additional calculation, to compare various models of binding affinity for the same factor. Our approach involves setting a threshold-score above which we consider a particular genomic subsequence to be a potential binding site, or "hit".

Some models, such as ours, attempt to assign physically meaningful scores to each subsequence, whereas others assign scores based on estimated probabilities or background frequencies. In order obtain a common scale for any given set of models, we identify each score, for each model, with its frequency of occurence in the non-coding genome at large. In other words, we simulate to obtain null distributions. We do this as follows.

For a given model, we sample n_s intervals of size $2w_s$ from the non-coding mappable genome that do not overlap regions identified by ChIP-chip. Within each of these intervals, we evaluate the score of each subsequences, generating n_s samples of emprical null distributions of scores. We now set a threshold on probability, α, e.g. 0.01. For each of the n_s samples, we now have one estimate of the score, s_α that corresponds to α. We use the median of the empirical distribution as our estimate, call it $\widehat{s}_\alpha$.

For each scored subsequence around each of the n_p peaks, we consider a position to be a hit if its score is greater than $\widehat{s}_\alpha$. In this way, we obtain n_p binary vectors which record each position at which a hit begins. We align these intervals at the peaks in the 5'-3' direction and sum across them. This generates a vector of counts recording, with respect to the position of peaks, how many of the n_p intervals had a hit at each relative position. We smooth these counts with a 200bp moving average[1], and then divide the result by the expected number of hits under a uniform null, $n_p(1 - s_t)^{-1}$.

ACKNOWLEDGEMENTS

Thanks to John Atherton, Stephanie Atherton and Alex Glazer for helpful discussions regarding physical chemistry.

SUPPLEMENTARY MATERIAL

Supplement A: Code for SELEX model (http://encodestatistics.org/SELEX). The code for the SELEX model used in the application of this paper is available at the above url.

REFERENCES

[Atkins 1998] Atkins, P., (1998). Physical chemistry. *6th edition.* W.H. Freedman and Company, New York.

[Bailey et al. 2006] Bailey, T. L., Williams, N., Misleh, C., and Li, W. W., (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, **34**: 369–373.

[1] The 200bp is motivated by the fact that in the ChIP-chip assay proteins bind to DNA fragments of roughly 200 bps

[Berman et al. 2004] Berman, B. P., Pfeiffer, B. D., Laverty, T. R., Salzberg, S. L., Rubin, G. M., Eisen, M. B., and Celniker, S. E., (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biology*, **5**(9): R61.

[Djordjevic et al. 2003] Djordjevic, M., Sengupta, A. M., and Shraiman, B. I., (2003). A biophysical approach to transcription factor binding site discovery. *Genome Research*, **13**: 2381–2390.

[Djordjevic & Sengupta 2006] Djordjevic, M., and Sengupta, A. M., (2006). Quantitative modeling and data analysis of SELEX experiments. *Physical Biology*, **3**: 13–28.

[Djordjevic 2007] Djordjevic, M., (2007). SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering*, **24**: 179–189.

[Ellington & Szostak 1990] Ellington, A.D., and Szostak, J.W., (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**: 818–822.

[Freede & Brantl 2004] Freede, P., and Brantl, S., (2004). Transcriptional repressor CopR: Use of SELEX to study the *copR* operator Indicates that evolution was directed at maximal binding. *Journal of Bacteriology*, **186**(18): 6254–6264.

[Guo et al. 2008] Guo, K., Paul, A., Schichor, C., Ziemer, G., and Wendel, H. P., (2008). CELL-SELEX: Novel perspectives of aptamer-based therapeutics. *International Journal of Molecular Sciences*, **9**: 668–678.

[Gupta 2008] Gupta, P.K., (2008). Ultrafast and low-cost DNA sequencing methods for applied genomics research. *Proceedings of the National Academy of Sciences India. Section B, Biological Sciences*, **78**(2): 91–102.

[Kim et al. 2003] Kim, S., Shi, H., Lee, D., and Lis, J. T., (2003). Specific SR protein-dependent splicing substrates identified through genomic SELEX. *Nuclei Acids Research*, **31**(7): 1955–1961.

[Li et al. 2008] Li, X.-Y., MacArthur, S., Bourgon, R., Nix, D.A., Pollard, D.A., Iyer V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., Chu, H.C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S.E., Knowles, D.W., Gingeras, G., Speed, T.P., Eisen, M.B., and Biggin, M.D., (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biology*, **6**(2): e27.

[MacArthur et al. accepted] MacArthur, S., Li, X.Y., Li, J., Brown, J.B., Chu, H.C., Zeng, L., Grondona, B.P., Hechmer, A., Simirenko, L., Stapleton, M., Bickel. P.J., Biggin, M.D., and Eisen, M.B., (accepted). Functionally distinct transcription factors show a quantitative continuum of binding and function to highly overlapping sets of thousands of genomic regions in the Drosophila melanogaster blastoderm. *Genome Biology*, 1596769202507849.

[Nelder & Mead 1965] Nelder, J. A., and Mead, R., (1965). A simplex method for function minimization. *Computer Journal*, **7**: 308–313.

[Ng et al. 2006] Ng, E. W. M., Shima, D. T., Calias, P., Cunningham, E. T. Jr., and Guyer, D. R., (2006). Pegaptanib, a targeted anti-VEGF aptamer for ocular vascular disease. *Nature reviews drug discovery*, **5**: 123–132.

[Nocedal & Wright 2006] Nocedal, J., and Wright, S., (2006). Numerical Optimization. *2nd edition.* Springer-Verlag, Berlin.

[Powell 1964] Powell, M.J.D., (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Computer Journal*, **7**(2): 155–162.

[Roulet et al. 2002] Roulet., E., Busso., S., Canargo, A. A., Simpson, A. J. G., MERMOD, N., AND BUCHER, P., (2002). High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nature Biotechnology*, **20**: 831–835.

[Segal et al. 2006] SEGAL., E., SADKA., T., SCHROEDER., M., UNNERSTALL., U., AND GAUL, U., (2006). Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, **451**: 535–540.

[Stoltenburg et al. 2007] STOLTENBURG, R., REINEMANN, C., AND STREHLITZ, B., (2007). SELEX - A (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomolecular Engineering*, **24**: 381–403.

[Tuerk & Gold 1990] TUERK, C., AND GOLD, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**: 505–510.

[von Hipple 2007] VON HIPPLE, P. H., (2007). From "simple" DNA-Proteirn Interactions to the macromolecular machines of gene expression. *Annal Reviews of Biophysics & Biomolecular Structure*, **36**: 79–105.

[Zhoa et al. 2009] ZHOA, Y., GRANAS, D., AND STORMO, G.D., (2009). Inferring binding energies from selected binding sites. *PLoS Computational Biology*, **5**(12).

JULI ATHERTON
DEPT OF EPI, BIOSTATS AND OCC HEALTH
MCGILL UNIVERSITY
E-MAIL: Juli.Atherton@mcgill.ca

NATHAN BOLEY
BEN BROWN
PETER BICKEL
DEPT OF STATISTICS
UNIVERSITY OF CALIFORNIA BERKLEY
E-MAIL: npboley@gmail.com
ben@newton.berkeley.com
bickel@stat.berkeley.edu

NOBUO OGAWA
STUART DAVIDSON
MIKE EISEN
MARK BIGGIN
GENOMICS DIVISION
LAWRENCE BERKELEY NATIONAL LABORATORY
E-MAIL: nobogw@gmail.com
stuartd@horizoncable.com
mbeisen@gmail.com
mdbiggin@lbl.gov
URL: http://bdtnp.lbl.gov/Fly-Net/