Quantitative modeling and data analysis of SELEX experiments

Marko Djordjevic,[1,2,*] and Anirvan M. Sengupta[3]

[1]*Department of Physics, Columbia University, New York, NY 10027*
[2]*Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210*
[3]*Department of Physics and BioMaPS Institute, Rutgers University, Piscataway, NJ 08854*

[*]*Corresponding author: Marko Djordjevic, Mathematical Biosciences Institute, The The Ohio State University, Columbus, OH 43210 Phone: (614) 292-6159, FAX: (614) 247-6643; E-mail: mdjordjevic@math.ohio-state.edu*

November 18, 2005

Abstract

SELEX (Systematic Evolution of Ligands by Exponential Enrichment) is an experimental procedure that allows extracting, from an initially random pool of DNA, those oligomers with high affinity for a given DNA-binding protein. We address what is a suitable experimental and computational procedure to infer parameters of transcription factor-DNA interaction from SELEX experiments. To answer this, we use a biophysical model of transcription factor-DNA interactions to quantitatively model SELEX. We show that a standard procedure is unsuitable for obtaining accurate interaction parameters. However, we theoretically show that a modified experiment in which chemical potential is fixed through different rounds of the experiment allows robust generation of an appropriate data set. Based on our quantitative model, we propose a novel bioinformatic method of data analysis for such modified experiment and apply it to extract the interaction parameters for a mammalian transcription factor CTF/NFI. From a practical point of view, our method results in a significantly improved false positive/false negative trade-off, as compared to both the standard information theory based method and a widely used empirically formulated procedure.

Keywords: SELEX, protein-DNA interactions, transcription factor binding sites, weight matrix, CTF/NFI

1 Introduction

One of the most important issues in molecular biology is to understand regulatory mechanisms that control gene expression. Gene expression is often regulated by proteins, called transcription factors (TFs), which bind to short (6 to 20 base pairs) segments of DNA [1]. To understand a regulatory system one needs a detailed knowledge of both TFs and their binding sites in

a genome. Binding sites of a given TF share a common sequence pattern [2], which is often represented by a consensus sequence. However, TF binding sites are often highly degenerate, so it is not possible to reliably detect TF binding sites in a genome by using just the consensus sequence [3]. As an alternative, position-weight matrices (PWMs) [2, 4, 5] have been used to search for TF binding sites, with demonstrable advantage over consensus sequence based methods [3].

The most widely used method to construct a PWM originates from information-theoretic considerations [5, 6]. To distinguish such weight matrices from those constructed by other methods, we will further call them information-theoretic weight matrices (see also [7]). To build these weight matrices, one usually starts from a known collection of aligned binding sites and calculates the corresponding matrix elements as the logarithm of the ratio of probability to observe a given base at a given position in a collection of binding sites, compared to the probability of observing the base in the genome as a whole [3]. However, despite the obvious advantages of using such PWMs over the consensus sequence, the majority of PWMs provide a low level of both sensitivity and specificity [8]. In particular, there tends to be a large number of false positives in searches using most PWMs [3, 8, 9].

In general, two problems may lead to the low sensitivity and specificity of PWMs. First, the information-theoretic method may not be the most appropriate one. It does not properly incorporate saturation in binding probability, as shown by [7], and an alternative method of weight matrix construction[1], based on the biophysical model of TF-DNA interaction, was developed. The method reduced the number of false positives and resulted in the explicit appearance - and determination - of the binding threshold. Additionally, there are probably problems with the collection of binding sites used to construct the weight (energy) matrix [8] because, first, the collection of binding sites is most often obtained from a database, and is likely assembled under diverse and ill-characterized conditions [7]. Second, for most TFs, only a few binding sites are available [10, 11], making the amount of data insufficient for determining parameters of TF-DNA interaction (i.e. weight matrix).

As an alternative to using binding sites assembled in biological databases, SELEX (Systematic Evolution of Ligands by Exponential Enrichment) experiments [12, 13] can be suitable for generating an appropriate dataset under controlled (uniform) conditions. Additionally, a recent experimental advance [14] which combines SELEX with SAGE (Serial Analysis of Gene Expression) [15], allows an efficient generation of a large number of binding sites for a given TF. In this paper, we ask the question: What is an appropriate experimental and computational procedure for inferring parameters of TF-DNA interaction from SELEX experiments? In particular, we will address the following two issues: 1) How should SELEX experiments be designed in order to generate a dataset suitable for determining parameters of TF-DNA interactions? 2) How should a correct analysis of data from a suitable experiment be done? To address those questions, we will use a biophysical model of TF-DNA interactions to quantitatively model SELEX experiments. We will incorporate this model in the novel bioinformatic method of data analysis that we will subsequently develop.

The outline of this paper is as follows. In Section 2, we will review SELEX experiments and point to the potential problems in the experimental procedure from the viewpoint of con-

[1]Weight matrices constructed by the method given in [7] were denoted energy matrices. This emphasizes that weights in the matrix correspond to the estimates of contribution to the binding (free) energy due to the presence of a certain base at a certain position in the binding site.

structing the appropriate weight matrix. In Section 3 we will quantitatively model the SELEX experiments. Based on this model, we will show that there is a range of experimental parameters for which the energy matrix cannot be inferred by using the standard SELEX procedure. However, we will show that a modified experiment allows a robust generation of the appropriate data set. In Section 4 we will propose a novel bioinformatic method of data analysis for modified SELEX experiments, and apply it to the data obtained in the experiment by Roulet *et al.* [14]. In Subsection 4.2 we will show that our method leads to a significant improvement in the false positive/false negative trade-off compared to the standard methods of data analysis. Quantitative analysis that supports Sections 3 and 4 is presented in Appendices B-F. Finally, in Section 5 we will summarize our results, compare them with some widely held views and put our work in the context of future research.

2 From SELEX to weight matrix

SELEX is a method in which a large number of oligonucleotides (DNA, RNA or unnatural compounds) can be rapidly screened for specific sequences that have high binding affinities and specificities toward the given protein target [12]. For the explanation of the applications of SELEX procedure, one should refer to some of the review papers [13, 16].

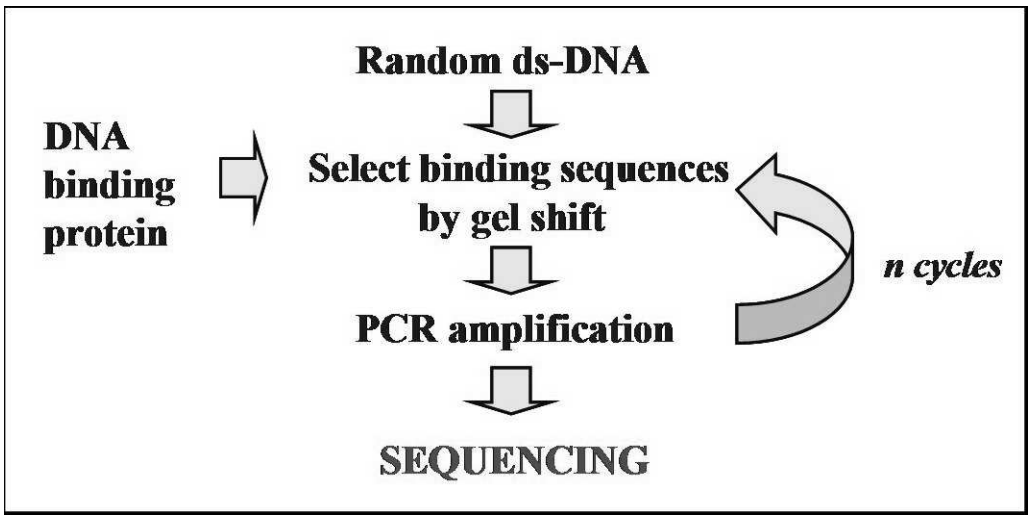

Figure 1: Scheme of SELEX experiment procedure. After n cycles of protein binding, selection and amplification, a certain number of DNA sequences are extracted and sequenced. Note that ds-DNA stands for double-stranded DNA.

The scheme of the widely used SELEX experiment procedure is shown in Fig. 1. The experiment is usually performed as follows. In the first step, a library of random oligonucleotides is synthesized. Protein is then mixed with the oligonucleotides library. Oligonucleotides that are bound by proteins are then separated from those that are not bound (e.g. by gel shift), which is called a selection step. Selected oligonucleotides are then amplified by the Polymerase Chain Reaction (PCR) [17], which is called the amplification step. One cycle of TF binding, selection and amplification is called a SELEX round. The SELEX rounds are repeated several times [12, 13], and some number of sequences (typically from 20 to 50) are extracted and sequenced from the final round. This procedure is successful in identifying the strongest binding

sites in the initial pool of random sequences, as demonstrated by experiments reported in the literature (e.g. [12]), as well as by the numerical studies [18, 19]. We will further call the widely used experimental procedure described above the *high stringency* SELEX, for reasons that will become apparent later. Next, in a typical data analysis the sequences selected in the last round of SELEX are used as a training set, from which elements of the information-theoretic weight matrix for a given TF are constructed [20, 21]. This weight matrix can then, in principle, be used to search for TF binding sites in a genome.

Is the standard experimental and data analysis procedure outlined above really suitable to successfully infer a correct weight matrix? Looking at the literature, it appears that this procedure fails often in practice. For example, in the SELEX experiment performed by Cui *et al.* [21], around 50 binding sites for LRP TF, selected in the last SELEX round, were extracted and sequenced. Binding sites were then used to construct an information-theoretic weight matrix, and the binding dissociation constants of the extracted sequences were then experimentally measured. However, the correlation between the dissociation constants and the information scores (i.e. the weight matrix scores) was quite poor, and accordingly, Cui *et al.* [21] comment that "the poor correlation for the data described here seems surprising...". Further, in [9] a comprehensive comparison between the weight matrices obtained from the eight available SELEX experiments with *E. coli* TFs and the corresponding weight matrices constructed from natural binding sites was performed. In seven out of those eight cases, large discrepancies between the matrices derived from natural sites and those derived from SELEX were reported. Therefore, obtaining good weight matrices from the standard SELEX procedure appears to be more an exception than a rule.

Why does the procedure described above appear to fail in many cases? The first possibility is that the assumption of additivity in TF-DNA interactions [22], on which the weight matrix representation is based, may not be suitable. This is, however, not likely, since this approximation has proved to be very good in many cases [20, 22, 23, 24]. Alternatively, the reason for the failure might be that the analysis was done by using the information theory based method. As discussed in the Introduction, the information theory based procedure does not properly incorporate saturation in binding probability. Therefore, it is not surprising that this is not the most appropriate method for correctly inferring the energy matrix. In this paper, we will develop a method of data analysis based on the biophysical model of protein-DNA interactions.

However, it seems rather surprising that the information theory based method is the only reason for the apparent problems with the weight matrices inferred from SELEX discussed above. There may be a systematic problem with the high stringency SELEX procedure when it comes to generating a data set suitable for inferring weight/energy matrices. With regard to this, it is apparent that two possible problems may arise. First, it may be that the noise in the dataset is too large, i.e. that many of the extracted sequences are too weak or are non-specific binders (for a discussion on nonspecific binding see Appendix B). Second, if the extracted sequences consist of only the strongest binding sites, the inferred weight matrix elements will come with large errors. To observe this, it is useful to take the limit in which only the sequence corresponding to the consensus binding site is extracted, where it is obvious that the energy matrix cannot be obtained from such information. For a more detailed statistical analysis of this issue refer to [14]. Therefore, our first goal is to address possible systematic problems with the experimental procedure by quantitatively modeling the SELEX experiments. We also incorporate this model in the bioinformatic method of data analysis that we develop in

Section 4.

3 A quantitative model of SELEX

Our model of SELEX is based on the biophysical view of TF-DNA interactions, which was used in a few recent papers (see e.g. [7, 25, 26]). For completeness and for introducing the notation, we briefly review a biophysical model of TF-DNA interaction in Appendix A. We start this section by extending this model in several ways, as to make it suitable for modeling of SELEX.

First, we take into account non-specific binding of a TF to DNA. As we show in Appendix B, the binding probability (Eq. (9)) is (approximately) modified in the following way due to the non-specific binding (see Eqs. (20) and (24)):

$$p\left(S\right) \approx \frac{1}{\exp\left(E\left(S\right)-\mu\right)+1} + c_{ns} = f\left(E\left(S\right)-\mu\right) + c_{ns}, \tag{1}$$

where $E(S)$ is binding (free) energy[2] of TF to a DNA sequence S, μ is chemical potential (see Appendix A), while c_{ns} depends on the threshold of non-specific binding E_{ns} (see Eq. (23)). Note that we scale all energies with k_BT. From Eq. (1) follows that non-specific binding cannot be distinguished from the so-called background partitioning. Background partitioning [12] is an effect that, during the selection step (Section 1), it is not possible to perfectly separate sequences that are bound by protein from those that are not bound. Due to that, in each round of SELEX, some DNA sequences *not* bound by TF are also selected with probability c_b. A combined effect of non-specific binding and background partitioning can be described by Eq. (1), where $c_{ns} \to c_{ns} + c_b$ (further in the text, we denote $c = c_{ns} + c_b$). Although c itself is likely small (e.g. $c_b \sim 0.1\%$ in [12]), the effect of non-specific binding/background partitioning can be considerable, since in SELEX experiments DNA is typically in large excess over protein, so only a small fraction of all DNA sequences are typically (specifically) bound by TF.

Second, we take into account that the length of DNA sequences in SELEX experiments is usually larger than the length of the TF binding site. For example, in the experiment by Roulet *et al.* [14], which will be the subject of our analysis in Section 4, the sequence length is $l = 25$ bp long, while the length of the binding site for CTF/NFI TF, studied in the experiment, is $L = 15$ bp. In Appendix C we discuss the modification of the binding probability due to the fact that l is greater than L. We show that this effect can be approximately accounted for by modifying the distribution of binding energies from $\rho(E)$ (see Eq. (12)) to $\rho_M\left(E\right)$ given by Eq. (28). Additionally, we note that the model has to take into account that the support E_S of the energy distribution is finite (see Eq. (31)), with "bottom of the band" determined by the energy of the strongest binder in the pool of random DNA sequences.

We note that we neglect stochastic effects in our model. This is generally justified by the fact that the typical length L of TF binding site is 20bp or less, while the total number of sequences used in SELEX is typically $N \sim 10^{15}$ [13], so each possible DNA sequence of length L is present in about $N/4^L \sim 10^3$ copies. Additionally, we note that high fidelity RNA polymerase is used in SELEX experiments, so mutations of sequences during the PCR amplification can be generally neglected. In that respect, SELEX is different from the so-called *in vitro* evolution

[2]For brevity, from now on we will refer to the free energy of binding simply as "binding energy". In chemical literature, the commonly used notation for this quantity would be $\Delta G(S)$ rather than $E(S)$.

experiments [27, 28], where low fidelity RNA polymerase and larger number of PCR rounds are used to (purposefully) introduce mutations and generate strong binding sequences that do not exist in the initial, small ($N \sim 10^5$), random DNA pool.

3.1 High stringency SELEX

In this subsection, we model the high stringency SELEX procedure, based on the (extended) model of TF-DNA interactions described above. We introduce equations that allow us to determine the position of chemical potential $\mu^{(n)}$ and energy distribution of selected oligos $\rho_M^{(n)}(E)$, as a function of the number of performed SELEX rounds n:

$$\rho_M^{(n)}(E) \sim \left[f\left(E - \mu^{(n)}\right) + c \right] \rho_M^{(n-1)}(E) \tag{2}$$

and

$$\begin{aligned} p_t &= p_f^{(n)} + p_b^{(n)} = K \exp\left(\mu^{(n)}\right) + \\ &+ \int \rho_M^{(n-1)}(E)\ f\left(E - \mu^{(n)}\right) dE + d_t\ \exp(\mu^{(n)} - E_{ns}) \end{aligned} \tag{3}$$

In the equations above, p_t, $p_f^{(n)}$ and $p_b^{(n)}$ are respectively total, free, and bound concentrations of protein and d_t is the total amount of DNA. Eq. (2) connects energy distributions of selected sequences from nth and (n-1)th round. Eq. (3) is mass conservation law, and it determines the position of $\mu^{(n)}$. $\rho_M^{(n-1)}(E)$ in Eq. (3) is normalized to d_t. Note that all energies are rescaled by k_BT. The equations above are solved recursively, i.e. we first solve them for $n = 1$, then increase n by one etc. Note that $\rho_M^{(0)}(E)$ is $\rho_M(E)$ given by Eq. (28) (Appendix C). We also note that, since in SELEX experiments DNA is typically in (large) excess over protein ($d_t \gg p_t$), most of the protein is bound to DNA and $p_f^{(n)}$ (i.e. $K \exp\left(\mu^{(n)}\right)$ term in Eq. (3)) can be neglected compared to $p_b^{(n)}$.

In general, Eqs. (2) and (3) have to be solved numerically; however the main features can be understood qualitatively. Let us assume that total amount of protein and the total amount of DNA (after each amplification step) are kept constant in each round of the experiment (as e.g. in [12]). It is evident that as n increases, the (average) affinity of selected oligos will also increase, which leads to the increase in the amount of $p_b^{(n)}$ and decrease of $p_f^{(n)}$. Therefore, both $\mu^{(n)}$ (note that $\mu^{(n)} = \log(p_f^{(n)}/K)$) and the maximum $E_m^{(n)}$ of energy distribution $\rho_M^{(n)}(E)$ move to stronger binding energies with the increase in the number of performed SELEX rounds. (Many experiments are performed in a way that the total amount of protein $p_t \equiv p_t^{(n)}$ is decreased from one SELEX round to the next (e.g. see [29]). It is evident that the previous conclusion about the decrease of $\mu^{(n)}$ holds in this case as well.) A limit in which Eqs. (2) and (3) can be solved analytically is analyzed in Appendix D. In particular, Eqs. (34) and (38) quantitatively support the discussion above.

In further analysis, we take the following values of parameters: $\chi_{l_e} = 4\, k_BT$ (see Appendix C), $p_t = 10nM$, $d_t = 10\mu M$ and $E_S = -4.3\chi_{l_e}$, while $E_{ns} = -2.0\chi_{l_e}$ (see Appendix B). The assumed values of p_t and d_t are typical for SELEX experiments [12], while (typical) $L \in \{6, ..., 20\}$ bp (base pairs) leads to E_S in the interval from $-3.5\chi_{l_e}$ to $-7\chi_{l_e}$ (see Eq. (31)). Values of χ_{l_e} and E_{ns} are expected to differ from one TF to the other, but the assumed values are

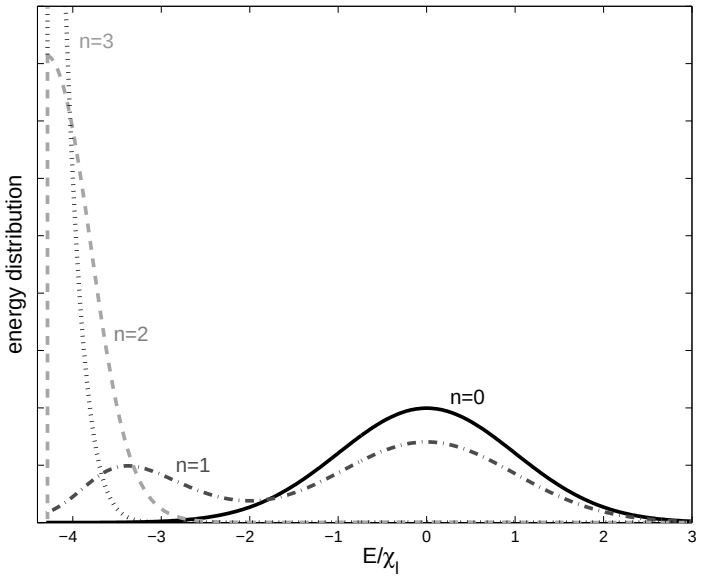

Figure 2: Change of $\rho_M^{(n)}(E)$ for a different number of performed rounds n, for a high stringency SELEX experiment. Peaks centered around zero correspond to random DNA binders, while the left hand corner of the figure corresponds to the highest affinity binder.

likely inside the realistic range [25]. Figure 2 shows $\rho_M^{(n)}(E)$ numerically obtained from Eqs. (2) and (3), with the parameter values stated above. It is useful to observe how the signal to noise ratio changes with n, where the noise is the number of selected random binders (corresponding to peak centered around zero), while the signal is the number of selected specific binders (for signal to noise ratio in the limit of unsaturated binding see Appendix D and Eq. (36)). For $n = 1$, we see that there is a small number of specific binders (note the small peak centered at $E/\chi_{l_e} \approx -3.5$) compared to the number of selected random sequences, so that signal to noise ratio is low. Because of the high noise to signal ratio, it is not possible to infer (correct) energy matrix from such a dataset. For $n = 2$, random binders are completely eliminated, so the problem with noise does not exist anymore. However, another problem emerges, i.e. since energy distribution of selected oligos has reached support E_S, (only) the strongest binding sites are selected. As discussed in Section 2, a correct energy matrix cannot be obtained from such a sequence set. For $n = 3$, we select the sequence set with even stronger binding affinities, etc. We note that the sharp cuts in distributions $\rho^{(n)}(E)$ at E_S (see Fig. 2) are the consequence of the approximation that we use for $\rho_M^{(0)}(E)$ (see Eq. (32)). In reality, $\rho_M^{(n)}(E)$ becomes discrete when one approaches E_S, which obviously does not change the conclusions inferred from Fig. 2.

We solved Eqs. (2) and (3) for different parameter values. As shown by the example above, for a range of realistic parameter values the appropriate choice for the total number of performed rounds n does not exist[3]. On the other hand, for some (other) parameters, for which selected DNA sequences have an acceptable signal to noise ratio and $E_m^{(n)}$ does not reach the highest affinity binders, the optimal choice of n does exist. However, we note that in practice, it is very hard to reliably predict such n, i.e. to decide when to stop the experiment, because a studied TF has *a priori* unknown binding parameters (i.e. χ_{l_e} and E_{ns}). Therefore, even in the case when the appropriate choice of n exists, it cannot, in practice, be simply calculated from

[3]In the analysis presented here, we used that p_t is constant in each round of SELEX. If p_t is decreased from one round to the next, as it is, in practice, done in many experiments (e.g. [29]), it is evident that this problem is even more pronounced.

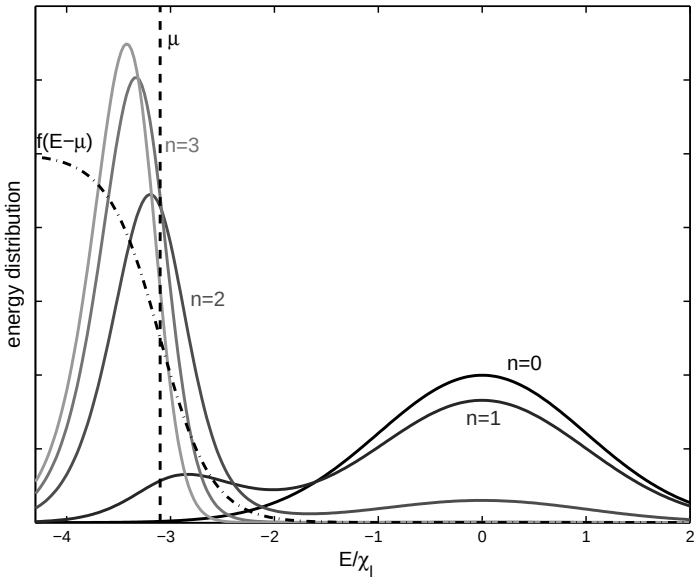

Figure 3: Change of $\rho_M^{(n)}(E)$ for different number of performed rounds n, for a SELEX experiment in which the chemical potential μ is fixed. Note that once the maximum of $\rho_M^{(n)}(E)$ reaches μ, it further moves very slowly toward the higher binding affinities. $\mu = -3.1\chi_{l_e}$, while the other parameters are the same as in Fig. 2.

Eqs. (2) and (3). From this, it follows that the high stringency SELEX procedure is, in practice, unsuitable for inferring the parameters of TF-DNA interaction. In the next section, we discuss modification of the SELEX procedure that allows a robust generation of the appropriate set of binding sequences.

3.2 SELEX with fixed selection stringency

Let us now assume that instead of moving toward the stronger binding energies, the chemical potential $\mu^{(n)} = \mu$ is constant in each round of the experiment. In such a case, from Eq. (2) it follows that the energy distribution is given by the following (simple) expression:

$$\rho_M^{(n)}(E) \sim f^n(E-\mu)\,\rho_M^{(0)}(E) \tag{4}$$

As we show in Appendix E, from Eq. (4) it follows that $E_m^{(n)} = \overline{E}_{l_e} - n\chi_{l_e}^2$ in the first rounds of experiment, when $n < (\overline{E}_{l_e} - \mu)/(\chi_{l_e})^2$. Thus, in the first few rounds, the maximum of energy distribution $E_m^{(n)}$ rapidly moves to the higher affinities. However, once $E_m^{(n)}$ reaches μ, it further drifts very slowly toward stronger binding energies (Eq. (42)). Figure 3 is equivalent to Fig. 2 with the difference that $\mu^{(n)} = \mu$ is fixed throughout the experiment. We see that for $n = 1$, we have the same situation as in the high stringency experiment, i.e. the signal to noise ratio is too low. For $n = 2$, non-specific binders are eliminated, similarly as in the high stringency experiment. However, the important difference is that instead of reaching the strongest binders (i.e. E_S), $E_m^{(2)}$ is close to μ. For $n > 2$, $E_m^{(n)}$ drifts very slowly toward the higher binding energies, remaining in the proximity of μ, and consequently does not reach E_S. More precisely, in Appendix E we show that $E_m^{(n)}$ asymptotically approaches $\mu - 2k_BT$.

The procedure described above has a significant practical advantage compared to the high stringency experiment discussed in the previous subsection. Since $E_m^{(n)}$ remains essentially fixed for larger n, one can perform more rounds (say 4 or 5), thus being sure that random binders

are eliminated, without the risk that only the strongest sequences will be selected. Since the procedure tolerates the whole range of n (in the above example $n > 2$), we call it robust.

The next issue is how the constraint of fixed chemical potential can be experimentally implemented. To our knowledge, all but one of the performed experiments correspond to high stringency SELEX. However, in the experiment done by Roulet *et al.* [14], the SELEX experiment was modified by inclusion of the radiolabeled sequence (probe) of moderate binding affinity E^*. The concentration of the DNA, added to the reaction mixture as a competitor to the radiolabeled probe, was in each round adjusted, so that a fixed fraction of the probe is bound by CTF/NFI TF in each SELEX round. From this it follows that $f\left(E^* - \mu^{(n)}\right) = const$, leading to $\mu^{(n)} = const$. Therefore the procedure in [14], provides a practical solution for fixing chemical potential through different rounds of experiment.

We also note that the analysis above gives a practical criterion at what n the experiment should stop. The procedure can be completed when random binders are eliminated and $E_m^{(n)}$ has reached μ, at which point we have a "quasi-saturation" (see Fig. 3 and Eq. (42)). Since at quasi-saturation the total amount of DNA (adjusted and directly observed by experimentalists) ceases to significantly change from one round to the next, this gives a practical criterion at what n the experiment should end.

4 SELEX data analysis

In this section we present a bioinformatic method for the data analysis of fixed stringency ($\mu = const$) SELEX experiments. We will first briefly present the basic idea behind our method. We will then introduce our novel algorithm for constructing energy matrix in Subsection 4.1. In Subsection 4.2 we will apply the algorithm to the experimental data, and compare the results with both the information theory based method and a widely used empirically formulated procedure, MatInspector [30].

Figure 3 and Appendix E show that the maximum of energy distribution for oligos selected in the final rounds of SELEX has to be in the vicinity of the chemical potential. It follows that the majority of sequences extracted from SELEX are in the saturated regime, i.e. bound with the probability close to one (see Appendix A). In [7] we showed that the information theory method is appropriate to use when sequences are in the unsaturated regime, but that this method does not properly incorporate saturation in the binding probability. In the context of SELEX experiments considered here, the information theory based method would be appropriate to use only if the majority of oligos were in the exponential tail of the binding probability $f(E - \mu)$. Since this does not happen in the fixed stringency SELEX experiments, we will in this Section devise a method which uses correct binding probability. Additionally, as described in Section 3 and Appendix C, we have extended a biophysical model of TF-DNA interactions to take into account that the length of TF binding site is typically shorter than the lengths of DNA sequences in SELEX, which we will also incorporate into our new procedure.

A key point in the implementation of our method will be to obtain TF-DNA interaction parameters through a maximum likelihood procedure. We will infer the initially unknown parameters by maximizing the probability that the extracted set of DNA sequences is observed as the outcome of the experiment. The probability of extracting the given set of DNA sequences will be calculated by taking into account the correct TF-DNA binding probability (which

properly describes the saturation effects) and by appropriately modifying affinity distribution of DNA oligos to account for difference in lengths between TF binding sites and SELEX DNA sequences. The set of equations resulting from varying this probability with respect to the unknown parameters will be then numerically solved to compute the elements of the energy matrix.

4.1 Estimating the energy matrix

In this subsection, we introduce a novel algorithm appropriate for data analysis of fixed stringency SELEX experiments. Let us assume that after n rounds of SELEX, set A, which contains n_S sequences $S^{(j)}$ ($j \in (1,..,n_S)$), has been extracted and sequenced. As summarized above, we will infer the unknown energy matrix by the maximum likelihood procedure. The probability of observing sequences from set A, but no other sequences from the initial DNA pool, is given by:

$$\exp(\Lambda) = \prod_{S \in A} \gamma\, f^n\left(E\left(S\right) - \mu\right) \prod_{S' \notin A} \left[1 - \gamma\, f^n\left(E\left(S'\right) - \mu\right)\right] \tag{5}$$

Terms f^n in the equation above account for n selection processes. The factor γ is a "sampling probability" and it absorbs all extraction and amplification events (there are n of them), as well as the final sequencing after the n^{th} round. Probabilities of extraction, amplification and sequencing are all assumed to be independent of sequence S, so γ does not depend on S either. We also note that we have neglected non-specific binding/background partitioning in Eq. (5), since we assume that n in Eq. (5) is large enough, so that non-specific binders are eliminated by the n^{th} round (see Fig. 3). Further, the sum over unbound sequences S' can be approximated in terms of the binding energy distribution:

$$\begin{aligned} \prod_{S' \notin A} \left[1 - \gamma\, f^n\left(E\left(S'\right) - \mu\right)\right] &\approx \exp[-\gamma \sum_{S' \notin O} f^n\left(E\left(S'\right) - \mu\right)] \\ &\approx \exp[-\gamma \int \rho_M\left(E\right)\; f^n\left(E - \mu\right)\; dE] \end{aligned} \tag{6}$$

In the above equation, we use $\rho_M(E)$ (see Eq. (28)) instead of $\rho(E)$ (see Eq. (12)) to account for the fact that the length of used DNA sequences is typically larger than the length of the TF binding site (see Appendix C). Similarly, we approximate $f(E(S) - \mu)$ in the first term of Eq. (5) by $f(E(s_M) - \mu)$ (see Appendix C), where s_M is the TF binding site of length L, with the maximal binding energy on l long sequence S. In practice, the set of binding sites s_M can be identified by an unsupervised search of the set of sequences S for L long statistically overrepresented motifs (e.g. by using the Gibbs search algorithm [31]). We also note that in the first (approximate) equality in Eq. (6), we used $\gamma \ll 1$, which is justified by the fact that the number of DNA sequences with binding energies below μ is typically much larger than the number of sequences n_S extracted from the last round of the experiment. For example, if we assume $\mu = -3\chi_{l_e}$ (as in Fig. 3), $n_S \sim 10^3$ [14] and typical $N \sim 10^{15}$ [13], we have $\gamma = n_S/[\int f^n(E-\mu)\rho_M(E)dE] \sim 10^{-9}$.

Since changing the overall scale of energy corresponds to multiplying energy scores for all binding sites with the (same) constant, for bioinformatic purposes, i.e. for TF binding site identification, it is not necessary to determine the overall scale of energy (see also [7]). The

natural quantity to scale all energies is width χ (in units of k_BT) of energy distribution for an ensemble of random oligos of length L (see [7] as well as Eqs. (12-14) in Appendix A). With such scaling, and provided that zero of energy is set to coincide with the mean $\overline{E}$ of energy distribution in the ensemble of random sequences (see Eq. (13) in Appendix A), $E(S)/\chi$ directly gives the estimate of significance of the given energy score. That is, the probability that a random DNA sequence will have a stronger binding energy than $E(S)/\chi$ is given by

$$\int_{-\infty}^{E(S)/\chi} \rho(E)dE \approx \frac{1}{2}[1 - erf(-\frac{E(S)/\chi}{\sqrt{2}})] \sim exp(-E(S)^2/(2\chi^2)) \tag{7}$$

Here $\rho(E)$ is the energy distribution for the set of random oligos (Eq. (12) in Appendix A), $erf(x)$ is the error function, and the last approximation is valid for $|E(S)/\chi| \gg 1$. Also note that if we consider the energy matrix as an vector in $4L$ dimensional space, χ is equal to the norm of this vector (see Eq. (15) in Appendix A), so rescaling with χ corresponds to normalizing the energy matrix to unit "length". We further use the notation $\widetilde{\epsilon}_{i,\alpha} = \epsilon_{i,\alpha}/\chi$.

Additionally, since maximum $E_m^{(n)}$ (i.e. the mean within the gaussian approximation of $\rho_M^{(n)}(E)$) has to be close to μ (see Subsection 3.2 and Fig. 3), we impose the constraint that

$$\mu/\chi = \sum_{S\in A} E(S) / (n_S \chi) = \sum_{i,\alpha} \widetilde{\epsilon}_{i,\alpha} S^*_{i,\alpha} \tag{8}$$

where $S^*_{i,\alpha} = \sum_j S^{(j)}_{i,\alpha}/n_S$. In order to obtain $\widetilde{\epsilon}_{i,\alpha}$ we maximize Λ (defined by Eqs. (5), (6) and (8)) with respect to $\widetilde{\epsilon}_{i,\alpha}$ and γ. Variation of Λ with respect to $\widetilde{\epsilon}_{i,\alpha}$ and γ then leads to the set of equations which are given in Appendix F. Those equations can be numerically solved to obtain $\widetilde{\epsilon}_{i,\alpha}$.

4.2 Application of the algorithm

To demonstrate our method, we use it to analyze the data from the experiment by Roulet *et al.* [14]. In addition to the modification discussed in Subsection 3.2, Roulet *et al.* combined SELEX with the SAGE protocol [15], which allowed them to sequence a large number of DNA oligos. A large dataset provides an obvious advantage for a precise estimation of energy parameters. In particular, a total of 4 SELEX rounds were performed, and approximately 880, 960, 1200, 6900 and 230 sequences were obtained from rounds 0, 1, 2, 3, and 4 respectively (note that round 0 refers to the initial, completely random, DNA pool).

We use about 230 sequences obtained after the 4^{th} round of experiment, to estimate $\widetilde{\epsilon}_{i,\alpha}$. We first search those 230 sequences S, to identify CTF/NFI binding sites s_M that correspond to $L = 15$ bp long statistically overrepresented motifs. The unsupervised search was performed by using a Gibbs search based algorithm [32], and the obtained set of binding sites s_M was used to determine $\widetilde{\epsilon}_{i,\alpha}$ by numerically solving Eq. (47). The obtained energy matrix $\widetilde{\epsilon}_{i,\alpha}$ is given in Appendix F (Table 1). We will further call $\widetilde{\epsilon}_{i,\alpha}$ the finite T energy matrix, to emphasize that, contrary to the QPMEME algorithm [7], the optimization of Λ (Eq. (5)) is not done in $T \to 0$ limit. We shift columns of $\widetilde{\epsilon}_{i,\alpha}$[4], so that the mean of the energy distribution $\rho(E)$ (given by Eq. (13)), for the ensemble of random oligos of length L is zero.

[4]Note that a provisional base independent constant can be added to each column of weight (energy) matrix, which corresponds to shifting zero of the weight matrix scores.

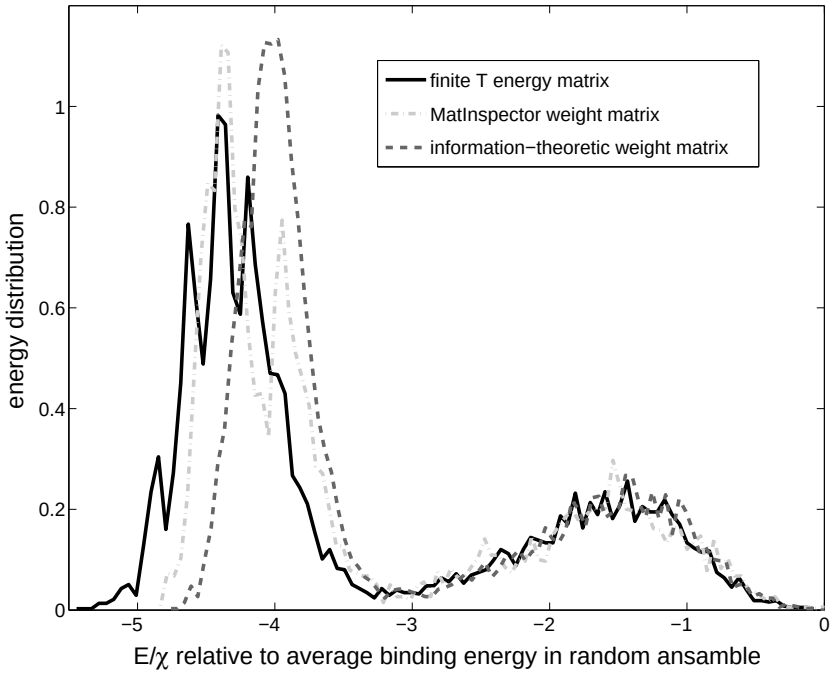

Figure 4: Energy distributions obtained by using the finite T energy matrix, the information-theoretic weight matrix and the MatInspector weight matrix are compared. Energy distributions are computed for more than 6000 binding sites extracted from the 3^{rd} round of SELEX [14]. Zero on the horizontal axis, corresponds to the mean value of binding energy in random ensemble. Note that maximum of energy distribution for non-specific binders is displaced relative to zero (i.e. positioned at around -1.5), caused by the fact that the length of the DNA sequences (25bp) is larger than the length of CTF/NFI TF binding site (15 bp). Actually, the energy distribution of non-specific binders, matches well with $\rho_M(E)$ calculated in Appendix F (Eq. (28)).

We next compare our method with both the information theory based method and a widely used empirically formulated procedure, MatInspector [30]. For this purpose, we construct both the information theory weight matrix $w_{i,\alpha}$ (see e.g. [33]) and the MatInspector weight matrix $w^{MI}_{i,\alpha}$ (see [30]) from the same set of binding sites s_M that we used to compute our finite T energy matrix. The obtained matrices are given and compared in Appendix F. We normalize $w_{i,\alpha}$ and $w^{MI}_{i,\alpha}$ and choose the zeros of weight matrix scores (i.e. "energies") for both matrices in the same way as for $\widetilde{\epsilon}_{i,\alpha}$. The three matrices are then used to compute the corresponding energy distributions for more than 6000 sequences extracted from the 3^{rd} round of SELEX, and their comparison is shown in Fig. 4. We see that there is a noticeable difference in the estimates of energies obtained by the three weight matrices.

To compare performance of the methods, we will infer false positive/false negative trade-off for the three matrices. The fraction of false negatives, for certain threshold E/χ, can be readily estimated by computing cumulatives of the distributions shown in Fig. 4. Further, the fraction of false positives can be calculated by computing the corresponding cumulative of the energy distribution of random oligos (see Eq. (12) in Appendix A). Therefore, with our choice for the zero of energy, it is evident that $\exp(-E^2/(2\chi^2))$ (on the abscissa of Fig. 5) is proportional to

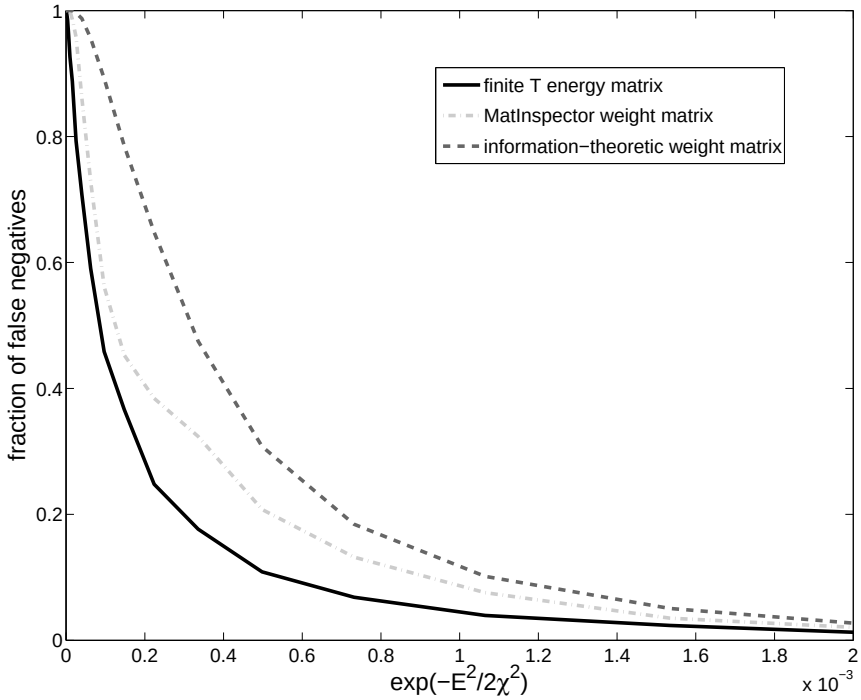

Figure 5: Comparison of the DET curves for the finite T energy matrix, the information theory weight matrix and the MatInspector weight matrix. Note that E/χ is measured relative to the average binding energy in random ensemble, so that $\exp(-E^2/(2\chi^2))$ on the horizontal axis is proportional to the number of false positives. False negative fraction is inferred from Fig. 4, by calculating cumulatives of the corresponding energy distributions.

the number of false positives. More precisely, if one searches a random DNA sequence with total length N_D, the total number of binding sites with energy scores below E/χ is approximately $(N_D/\sqrt{2\pi})\int_{-\infty}^{E/\chi}\exp(-x^2/2)dx \sim N_D\exp(-E^2/(2\chi^2))$ (see Eq. (12) and Eq. (7)), where the last approximation is valid for $|E/\chi| \gg 1$.

The above estimates of false positives and false negatives can be used to obtain the Detection Error Trade-off (DET) curves (see [7, 34]), which are shown in Fig. 5[5]. We see that over the entire range of fixed false negative values on the vertical axis, fraction of false positives for the finite T energy matrix is few times lower compared to the false positive rate for the information-theory weight matrix. In the case of the MatInspector weight matrix, the finite T energy matrix has about two times smaller false positive rate for false negative fractions smaller than 40%, while for larger false negative fractions the two DET curves come close to each other. We note that the meaningful range, in which one is likely to operate in practice, is the one with a smaller fraction of false negatives (i.e. 40% or below), where the finite T energy matrix clearly shows better performance.

Figure 5 shows that, while our method outperforms both MatInspector and the information theory method, it appears that MatInspector shows a better false positive/false negative trade-off compared to the information theory method. We however note that MatInspector is an

[5]In the construction of the DET curve for MatInspector we did not use the so-called "core similarity", which can optionally be used as a second threshold in the method [30]. The reasons are that the use of core similarity is left as an "optional feature at the discretion of the user" [30], which was not used in the example searches in [30]. Additionally, in [30] it is not stated how to choose the value of the second threshold, which makes it hard to fairly compare a method with two (arbitrary) thresholds with the methods that use only one threshold.

empirically formulated procedure, i.e. it is not founded on either statistical considerations, such as the information theory method, or on a (bio)physical model of TF-DNA interactions, such as the method presented here. We therefore believe that the performance of MatInspector would change from one experiment to the other, and that a better performance of MatInspector compared to the information theory method may not prove to be systematic. On the other hand, since our method is based on a correct physical model of SELEX experiments, we expect that it will systematically produce reliable estimates of the interaction parameters. In any case, fixed stringency SELEX experiments, which will be performed in future, will provide opportunities to more throughly test the performance of the algorithms.

Finally, we discuss how the above comparison of the three algorithms may be affected by the possible presence of noise in the input data used to construct the three weight matrices. We obtained the set of (putative) binding sites through a heuristic Gibbs sampling algorithm, which is in principle not guaranteed to find the true binding sites. So if there are misassignments in the input data, the question is whether the better performance of our method could be the result of higher robustness with regard to noise, rather than higher accuracy for error-free data. To asses the amount of noise in the input data we performed a self-consistency check. We scored all DNA sequences selected in the 4^{th} round of the experiment with the finite T energy matrix and classified all binding sites above score -3.4 as specific binders and all below -3.4 as non-specific binders (note that -3.4 corresponds to the saddle of the bimodal distribution shown in Fig. 4). We find that only one in the 175 large set of input binding sites obtained by the Gibbs algorithm is not contained in the list of 184 specific binders classified with the finite T energy matrix. This indicates that the noise level in the input data is likely quite small. The apparent low number of misassignments in the output of Gibbs alignment is likely a consequence of the low number of random sequences extracted from 4^{th} round of SELEX[6], and the fact that the length of SELEX sequences is quite short (25 bp).

5 Conclusion and Outlook

In this paper we modeled SELEX experiments and proposed a novel method of data analysis. Our analysis showed that for a certain realistic range of parameters, the suitable solution for the number of rounds that should be performed does not exist at all. We argued that even for the parameters for which the solution exists, it is very hard to find such a solution in practice. However, we showed that the modification of the standard SELEX procedure in which chemical potential is fixed [14] robustly selects sequences that allow one to successfully determine the energy matrix. We next proposed a novel method for inferring energy matrix from the sequences extracted from a SELEX experiment with fixed selection stringency. Contrary to the widely used information theory weight matrix method, our procedure correctly represents saturation in binding probability. As an example of the procedure, we estimated the energy matrix for CTF/NFI TF by analyzing the data from the experiment by Roulet *et al.* [14]. We demonstrated that our energy matrix leads to a significantly better false positive/false negative trade-off.

Finally, we compared the results of our analysis with some widely held views. It is generally well understood that doing too large a number of SELEX rounds leads to too strong a selection.

[6]Notice that Fig. 4 shows that there is a quite small amount of noise even in the 3^{rd} round of experiment. The noise is further reduced in the 4^{th} round.

It is, however, widely believed that the problem can be solved by doing only few selection cycles. For example, in the recent SELEX experiment by Kim *et al.* [35], it was noted that "...we used fewer rounds of selection than a conventional SELEX to avoid the enrichment of just a few high affinity winners". However, as shown by our analysis, this can lead to another problem, i.e. if too few SELEX rounds are performed, too large a number of random sequences are likely to be selected. Moreover, we found that the problems of over-selection and too high amount of noise are, in practice, very hard to reconcile within the standard SELEX procedure and that a modified experiment with a fixed chemical potential has to be performed instead. From the aspect of data analysis, we showed that the commonly used information theory based method, widely believed to be well founded in both statistics and thermodynamics [36], is not appropriate for analysis of data for SELEX experiments with fixed selection stringency.

In the context of future research, we believe that the analysis presented here, together with the experimental methods introduced in [14], open a perspective to apply high-throughput, fixed stringency, SELEX experiments for a large number of different transcription factors. This would provide a reliable method for detection of transcription factor binding sites, and would facilitate the comprehensive understanding of gene regulation.

Acknowledgments

This work was supported by NIH grant GM67794 (to Boris Shraiman). Final parts of this work were supported by NSF under Agreement No. 0112050 and NSF grant MCB-0418891. We are grateful to Boris Shraiman for useful discussions, and to Andrew Millis, Harmen Bussemaker and Istok Mendas for critical reading of the manuscript. M.D. acknowledges the hospitality of the Kavli Institute for Theoretical Physics where part of this work was done and is grateful to Valerie Parsegian for reading the manuscript.

Glossary

Weight matrix for a binding pattern of length L is defined as a matrix of numbers $w_{i,\alpha}$ where $i \in [1, 2, ...L]$ and $\alpha \in [A, T, C, G]$. The score of the sequence $\alpha_1 \dots \alpha_L$ is given by $w_{1,\alpha_1} + w_{2,\alpha_2} + \cdots + w_{L,\alpha_L}$.

Sensitivity is defined as TP/(TP+FN), where TP is the number of true positives and FN is the number of false negatives. In the context of a TF binding site search, a true positive (TP) arises when an algorithm correctly classifies a true binding site as such, while a false negative (FN) arises when an algorithm classifies a true binding site as a non-binding site.

Specificity is defined as TN/(TN+FP), where TN is the number of true negatives and FP is the number of false positives. In the context of TF binding site search, a true negative arises when an algorithm correctly classifies a true non-binding site as such, while a false positive (FP) arises when a search algorithm classifies a true non-binding site as a binding site.

Serial Analysis of Gene Expression (SAGE) is a method for comprehensive analysis of gene expression patterns. In the context of this paper, a part of the SAGE protocol can be used to link together oligomers extracted from SELEX in order to form longer DNA molecules that can be efficiently sequenced.

Gel shift is a technique used to separate free DNA molecules from DNA molecules that are in complex with protein, based on the fact that protein-DNA complexes migrate more slowly through gel under the influence of an electric field.

Polymerase Chain Reaction (PCR) is an experimental technique that allows one to produce a large number of copies of any fragment of DNA. In principle, the number of DNA molecules is doubled in each round of PCR, so there is an exponential increase in the number of molecules with the number of performed PCR rounds.

Dissociation constant for sequence S is equal to the concentration of (free) TF for which there is 50% probability that a DNA molecule S will be bound by the TF. The relationship between the dissociation constant and the binding energy $E(S)$ is given by $K_D(S) = K \exp(E(S))$, where K is a proportionality constant.

Maximum likelihood estimation is a statistical method used to estimate unknown parameters of a (known) probability distribution. The basic principle is to draw a sample from the distribution, calculate the probability that this sample is observed and then determine the unknown parameters such that this probability is maximal.

A A biophysical model of TF-DNA interaction

Let us consider an experiment where a certain number of identical DNA oligomers with sequence S and length L (equal to the length of the TF binding site) are mixed into a solution with some concentration of TF. It can be shown (see e.g. [7]) that the equilibrium probability $p(S)$ that a DNA sequence S is bound by TF is given by:

$$p(S) = \frac{1}{1 + \exp(E(S) - \mu)} = f(E(S) - \mu) \tag{9}$$

In the above equation, p_f is the concentration of free TF, μ is the chemical potential, while K is a multiplicative constant related with counting number of quantum states in a box (e.g. see [37]).

Chemical potential μ is set by the free TF concentration in the solution:

$$\mu = \log(p_f/K). \tag{10}$$

Note that in the above equations all energies are rescaled by $k_B T$. The form of the binding probability $f(E(S) - \mu)$ in Eq. (9) corresponds to the Fermy-Dirac distribution (see e.g. [37]). If the binding energy $E(S)$ of a sequence S is well below μ, then $f(E(S) - \mu)$ is close to one and the sequence S is almost always bound by TF. We will further call sequences with binding energy $E(S)$ which corresponds to this limit saturated. On the other hand, if $E(S)$ is well above μ, the sequence S is rarely bound, with probability given by the Boltzmann distribution $f(E(S) - \mu) \approx \exp(-(E(S) - \mu))$. As shown in [7], the information theory weight matrix procedure assumes Boltzmann distribution of binding probability, and is, therefore, not appropriate whenever saturation of binding occurs.

Further, we need an expression for $E(S)$. The most simple model of TF-DNA interaction, which we use in this paper, assumes that the interaction of a given base with the factor does not depend on the neighboring bases:

$$E(S) \approx \epsilon \cdot S = \sum_{i=1}^{L} \sum_{\alpha=1}^{4} \epsilon_i^\alpha S_i^\alpha, \tag{11}$$

where $S_i^\alpha = 1$, if base α is at the position i and $S_i^\alpha = 0$ otherwise. ϵ_i^α is the interaction energy with the nucleotide α at the position $i = 1, ..., L$ of the DNA string [20], and ϵ is called energy matrix. The simple parameterization given by Eq. (11) provides a very good approximation in many cases [20, 23, 22, 24], although there are examples where binding at some positions in the binding site shows dependence on dinucleotide pairs [38, 39, 40].

We further need to compute the energy distribution $\rho(E)$ for an ensemble of randomly generated oligonucleotides. Such ensemble corresponds to the random set of DNA sequences used in the first round of SELEX experiment. In [7] it was shown that in the first approximation $\rho(E)$ is given by a gaussian:

$$\rho(E) \approx \exp(-(E - \overline{E})^2/2\chi^2)/\sqrt{2\pi\,\chi^2} \tag{12}$$

with

$$\overline{E} = \sum_{i=1}^{L} \overline{\epsilon}_i \tag{13}$$

and

$$\chi^2 = \sum_{i=1}^{L} \sum_{\alpha=1}^{4} p_\alpha (\epsilon_i^\alpha - \overline{\epsilon}_i)^2 \tag{14}$$

where $\overline{\epsilon}_i = \sum_{\alpha=1}^{4} p_\alpha(\epsilon_i^\alpha)$.

As noted in footnote 4, each column of an energy matrix can be shifted for a provisional base independent value, and a convenient choice would be to set $\overline{\epsilon}_i = 0$, so that $\overline{E} = 0$ and

$$\chi^2 = \sum_{i=1}^{L} \sum_{\alpha=1}^{4} p_\alpha (\epsilon_i^\alpha)^2 \ . \tag{15}$$

From the above equation follows that χ is equal to the norm of the energy matrix ϵ_i^α, with "metric" given by the background single base frequencies p_α.

We further note that $\rho(E)$ is well approximated by Eq. (12) in the proximity of maximum $E = \overline{E}$, however, away from the maximum deviations from gaussianity appear. In fact, the support of $\rho(E)$ is finite with the "bottom of the band" $E_S = \sum_i \min_\alpha \epsilon_{i,\alpha}$ (while the "top" is given by $\sum_i \max_\alpha \epsilon_{i,\alpha}$). In this paper we work with the Gaussian approximation to $\rho(E)$, but we also introduce a cut in the distribution to account for the fact that the support E_{bb} is finite (see Appendix C).

Further extensions of the TF-DNA interaction model, necessary for our modeling of SELEX experiments, are given in Appendix B and Appendix C.

B Non-specific binding of TF to DNA

We here assume that a given TF can bind to DNA in two conformations. The first conformation results in the sequence specific interaction, with the interaction energy $E(S)$. The second conformation results in the sequence independent (non-specific) interaction, with the interaction energy E_{ns} [2]. E_{ns} is called the threshold for non-specific binding. We consider a reversible reaction of binding of the TF to a DNA sequence S, where TF can bind with S in two conformations. Sequence specific, and sequence non-specific reactions can, respectively, be represented by:

$$[TF] + [S] \Leftrightarrow [TF - S]_s \tag{16}$$

and

$$[TF] + [S] \Leftrightarrow [TF - S]_n \tag{17}$$

Here, $[TF]$ is concentration of free TF, $[S]$ is concentration of sequence S that is *not* in the complex with protein, while $[TF - S]_s$ and $[TF - S]_n$ are concentrations of TF that is bound to the TF in the sequence specific, and in the sequence non-specific conformation respectively. In the equilibrium, the following relations hold:

$$K \exp(E(S)) = \frac{[TF]\ [S]}{[TF - S]_s} \tag{18}$$

$$K \exp(E_{ns}) = \frac{[TF]\ [S]}{[TF - S]_n} \tag{19}$$

From Eqs. (18) and (19) we have that the probability that a sequence S is bound by the TF is given by:

$$p(S) = \frac{[TF - S]_s + [TF - S]_n}{[S] + [TF - S]_s + [TF - S]_n} = \frac{a}{b \exp(E(S)) + 1} + c_{ns}. \tag{20}$$

In the equation above,

$$a = \frac{1}{1 + \exp(\mu - E_{ns})} \tag{21}$$

$$b = \exp(-\mu)\ [1 + \exp(\mu - E_{ns})] \tag{22}$$

and

$$c_{ns} = \frac{1}{1 + \exp(E_{ns} - \mu)} \tag{23}$$

where $\mu = \log([TF]/K)$. By comparing Eq. (20), with Eq. (9) the quantity b can be identified as the "effective" fugacity in the presence of non-specific binding. From Eq. (22) then follows that, non-specific binding, in principle, shifts μ toward more negative values. In practice,

however, we most often have a case in which the amount of TF is much less than the amount of DNA. For example, even for a pleiotropic TF such as CRP in *E. coli*, the total number of CRP molecules is much less than the total length of the genome [41], while in SELEX experiments protein is typically in large excess over DNA. It is then obvious that μ has to be significantly below E_{ns}, since all DNA sequences would, otherwise, have to be bound with high probability. Therefore, we in practice have that $\exp(\mu - E_{ns}) \ll 1$, so from Eqs. (21) and (22) follows that $a \approx 1$, $b \approx \exp(-\mu)$ and $c_{ns} \approx \exp(-(E_{ns} - \mu))$, so we have that:

$$p(S) \approx \frac{1}{\exp(E(S) - \mu) + 1} + c_{ns} \tag{24}$$

Therefore, the effect of non-specific binding enters through c_{ns}, which is determined by the position of μ relative to E_{ns}. We note that values of E_{ns} were not experimentally quantitated [25], so one does not know what range of values E_{ns} can take. Unless E_{ns} is positioned in the strong binding tail of $\rho(E)$, non-specific binding will not have a large effect. As discussed in Section 3, non-specific binding is in SELEX experiments essentially indistinguishable from the background partitioning effect.

C Binding of TF to a longer DNA sequence

In Appendix A, we derived the binding probability under the assumption that the length l of DNA sequence is equal to the length L of the TF binding site. However, for DNA sequences used in SELEX, l is typically larger than L. For example, in the SELEX experiment performed by Roulet *et al.* [14] (see Subsection 3.2) $l = 25$ bp, while $L = 15$ bp for CTF/NFI TF. It is straightforward to obtain that the probability that sequence with length l will be bound is given by:

$$p(S) = \frac{\exp(\mu) \sum_{i=1}^{l-L} \exp(-E(s_i))}{1 + \exp(\mu) \sum_{i=1}^{l-L} \exp(-E(s_i))} \tag{25}$$

In the equation above, the sequences s_i are L long binding sites, corresponding to all possible $l - L$ frame shifts in which the TF can bind to a sequence S, while μ is chemical potential (see Appendix A). We assume that $l < 2L$, which is typically the case in SELEX experiments, so that two or more TF molecules cannot simultaneously bind to the sequence S. Note that all quantities in Eq. (25) are rescaled by $k_B T$. For $l_e = l - L$ that is not too large, which is typically the case in the experiments, the expression $\sum_{i=1}^{l_e} \exp(-E(s_i))$ can be approximated by taking into account only the contribution from the strongest binding site s_M, where $E_M(S) = E(s_M) = \min\{E(s_i), i \in (1, l_e)\}$. With this approximation, Eq. (25) simplifies to:

$$p(S) \approx f(E_M(S) - \mu) = \frac{1}{1 + \exp(E_M(S) - \mu)} \tag{26}$$

which is the Fermi Dirac probability encountered before (see Appendix A).

From Eq. (26) follows that binding of a TF to a sequence S with $l > L$ is (approximately) equivalent to the binding to the sequence s_M with length L. Let us now look at the first round of SELEX, where the TF is mixed with a large number of randomly generated sequences of

length l. In order to make the problem equivalent to the one in which $l = L$, instead of density of states $\rho(E)$ (see Appendix A), we have to use $\rho_M(E)$ defined as the number of sequences S, for which $E_M(S)$ has energy between E and $E + dE$.

To calculate $\rho_M(E)$, we neglect correlations in binding energies $E(s_i)$ (see Eq. (25)) of l_e binding sites that belong to the same sequence S. This is in general well justified, unless sequence S consists of the long repeat. In particular, the validity of this approximation was confirmed by numerically testing Eq. (28) below. Based on this, $\rho_M(E)$ can be calculated by generating sets of l_e values of E from distribution $\rho(E)$ and retaining only the strongest binding energy from each set. It is straightforward to see that $\rho_M(E')$ can be obtained from $\rho(E)$ by:

$$\rho_M(E') = l_e \frac{d\Phi(E')}{dE'} \left(1 - \Phi(E')\right)^{l_e - 1}, \tag{27}$$

where $\Phi(E')$ is the cumulative distribution given by $\Phi(E') = \int_{-\infty}^{E'} \rho(E)\, dE$. The expression on the right hand side of Eq. (27) is, therefore, equal to the probability that $l_e - 1$ values of E generated from $\rho(E)$ are above E', while one of them (the strongest binding energy) is between E' and $E' - dE'$.

By plotting $\rho_M(E)$, given by Eq. (27) we see that $\rho_M(E)$ can be approximated by Gaussian:

$$\rho_M(E) \approx \exp\left(\left(E - \overline{E}_{l_e}\right)^2 / 2\chi_{l_e}^2\right), \tag{28}$$

with

$$\overline{E}_{l_e} = \overline{E} - a(l_e)\, \chi \tag{29}$$

and

$$\chi_{l_e} = b(l_e)\, \chi, \tag{30}$$

where $\overline{E}$ and χ are respectively the mean value and the standard deviation for $\rho(E)$ (see Appendix A), while $a(l_e)$ and $b(l_e)$ are respectively monotonically increasing and decreasing functions of l_e. Functions $a(l_e)$ and $b(l_e)$ can be calculated numerically (e.g. $a(10) = 1.4$ and $b(10) = 0.3$). Numerical analysis shows that approximately, $a(l_e) \approx 0.6 \log(l_e)$, while $b(l_e) \approx 1/\sqrt{l_e}$.

Finally, we have to take into account that $\rho_M(E)$ has the finite support, where "bottom of the band" E_S is determined by the energy of the strongest binder in the random pool of DNA sequences, and can be approximated by:

$$(4^L) \int_{-\infty}^{E_S} \rho(E)\, dE \sim 1, \tag{31}$$

where $\rho(E)$ is normalized to 1.

In Eq. (31), we assumed that the total number of sequences N (more precisely $l_e N$) in the DNA pool is larger than 4^L, so that all possible sequences of length L are present. This is most often the case in practice, since typically $L < 20$, while $N \sim 10^{15}$ [13], so that $4^L \ll N$.

To take into account the finite support of $\rho_M(E)$, we make a simple approximation and introduce a sharp cut in the distribution $\rho_M(E)$, i.e. we take that

$$\rho_M(E) \sim \theta(E - E_s) \exp\left(\left(E - \overline{E}_{l_e}\right)^2 / 2\chi_{l_e}^2\right), \tag{32}$$

where $\theta(E - E_s)$ is unit step (Heaviside) function. We note that the top of the band is finite as well, however we do not include it in Eq. (32), since energy distribution of selected oligos moves toward higher binding affinities in SELEX (see Subsection 3.1). In reality, $\rho_M(E)$ becomes discrete when we approach E_s, however a simple approximation given by Eq. (32) is sufficient for the purpose of our model.

D High stringency SELEX in the limit of unsaturated binding

In this Appendix, we look at the limit in which the binding probability $f(E - \mu)$ in Eqs. (2) and (3) can be approximated by the Boltzmann factor $\exp(\mu - E)$. In this limit, those equations can be solved analytically. The above approximation is valid if all selected binding sites are unsaturated in each SELEX round, i.e. if $\mu^{(k)} < E_S$ $(\forall k \in (1, .., n))$, where n is the number of performed SELEX rounds. In this limit, Eq. (2) gives:

$$\rho_M^{(n)}(E) \sim (\exp(-E) + \exp(-E_{ns}))^n \, \rho_M(E), \tag{33}$$

where, for simplicity of the notation, we assume that all noise comes from non-specific binding, i.e. that $c \equiv c_{ns}$ (see Eqs. (1)). If we use the gaussian approximation for $\rho_M(E)$ (see Eq. (28)), from Eq. (33) follows that $\rho_M^{(n)}(E)$ has n peaks, which are centered at positions $E_{m,k}^{(n)} = -k\chi_{l_e}^2$, $(k \in (1, ..n))$. We here shifted zero of energy, so that it coincides with $\overline{E}_{l_e}$ (see Eq. (29)). From Eq. (33) it is obvious that n^{th} peak $E_{m,k=n}^{(n)} \equiv E_m^{(n)}$ contains only sequence specific binding sites, while $k = 0$ peak (centered at zero) corresponds exclusively to non-specific (i.e. random) binders. Therefore, in the limit considered here, the maximum of specifically selected binding sites (the leading maximum) is positioned at:

$$E_m^{(n)} = -n\,\chi_{l_e}^2. \tag{34}$$

From Eq. (34) follows that $E_m^{(n)}$ rapidly moves to higher binding energies. For example, for the realistic parameter values of $\chi_{l_e} = 2$ and $E_S = -5\chi_{l_e}$ (corresponding to $L = 12$, see Eq. (31)), $E_m^{(n)}$ reaches E_S after (only) three SELEX rounds.

The "intensity" $I_k^{(n)}$ of k^{th} peak, i.e. the number of binding sites corresponding to the peak, is:

$$\begin{aligned} I_k^{(n)} &\sim \exp(-E_{ns}(n-k)) \int \exp(-Ek)\,\rho_M(E) \\ &\approx \exp(-E_{ns}(n-k))\exp(k^2\chi_{l_e}^2/2) \end{aligned} \tag{35}$$

From Eq. (35) is straightforward to obtain that the conditions $I_k^{(n)} > I_n^{(n)}$ and $I_k^{(n)} > I_0^{(n)}$ cannot be simultaneously satisfied for $k \in (2, ..., n-1)$, for any parameter values. Therefore, for any given n, either 0^{th} or n^{th} peak have the maximal intensity. Further, it is sensible to define "signal to noise" ratio $\nu^{(n)}$ as the ratio of the number of specific binders corresponding to the n^{th} peak and the number of non-specific binders corresponding to the 0^{th} peak. From Eq. (35) follows:

$$\nu^{(n)} = \exp(n^2\chi_{l_e}^2/2 + n\,E_{ns}), \tag{36}$$

so signal to noise ratio necessarily increases with increasing n.

We next want to derive for what parameter values is the unsaturated limit, which we analyze in this Appendix, valid. We use the self-consistency condition:

$$\mu^{(j)} < E_S \, (\forall j \in (1, ...n)), \tag{37}$$

with $\mu^{(j)}$ calculated from Eq. (3) by using $f(E-\mu) \approx \exp(\mu - E)$:

$$\exp(\mu^{(j)}) = \frac{p_t/d_t}{\exp[(j-1/2)\chi_{l_e}^2\,]\,\alpha(j, \chi_{l_e}, E_{ns}) + \exp(-E_{ns})} \tag{38}$$

where,

$$\alpha(j, \chi_{l_e}, E_{ns}) = \frac{1 + \sum_{k=0}^{j-2} \binom{j-1}{k} [\exp(-(k+j+1)\chi_{l_e}^2/2 - E_{ns}]^{j-k-1}}{1 + \sum_{k=0}^{j-2} \binom{j-1}{k} [\exp(-(k+j-1)\chi_{l_e}^2/2 - E_{ns}]^{j-k-1}}. \tag{39}$$

We note that the condition (37) (with $\mu^{(j)}$ given by Eqs. (38) and (39)) is satisfied for the subset of parameter values that are inside the realistic range. For example, the condition (37) holds for $p_t = 10\,nM$, $d_t = 10\,\mu M$ (see e.g. [12]), $\chi_{l_e} = 2$, $E_{ns} = -2\chi_{l_e}$, $E_S = -5\chi_{l_e}$, and for all n until $E_m^{(n)}$ reaches E_S (i.e. for $n = 1, 2, 3$).

We finally discuss how $\mu^{(n)}$ depends on the parameter values (for further discussion, we let $j \to n$ in Eq. (38)). From Eq. (38) is straightforward to show that $\mu^{(n)}$ necessarily decreases, with increasing n. Further, the following interpretation can be assigned to the terms on the right hand side of Eq. (38). It is obvious that $\mu^{(n)}$ increases, with the increase of (total) protein to (total) DNA ratio p_t/d_t. The term $\exp[(n-1/2)\chi_{l_e}^2\,]$ accounts for the fact that the (average) binding energy of sequences corresponding to the n^{th} peak increases with n (see Eq. (34)), which results in the decrease of the concentration of *free* TF and consequently in the decrease of $\mu^{(n)}$. From Eq. (39) is can be noticed that $\alpha(n, \chi_{l_e}, E_{ns}) \leq 1$. This term, therefore, leads to the increase of $\mu^{(n)}$, which may be explained by the fact that for $n > 1$ there are $n-1$ peaks of $\rho_M^{(n)}(E)$ "generated" by non-specific binding. All those peaks have smaller (mean) binding energy compared to the n^{th} peak, which results in the smaller amount of the specifically bound TF. Finally, the term $\exp(-E_{ns})$ accounts for the fact that some amount of TF is non-specifically bound by DNA sequences, which leads to the decrease of $\mu^{(n)}$. The dependence of $\mu^{(n)}$ from E_{ns} is, therefore, quite complicated, since increase in non-specific binding decreases $\mu^{(n)}$ through $\exp(-E_{ns})$, but increases it through $\alpha(n, \chi_{l_e}, E_{ns})$. Similarly, the increase of χ_{l_e} has the opposite effects on $\mu^{(n)}$ through terms $\exp[(n-1/2)\chi_{l_e}^2\,]$ and $\alpha(n, \chi_{l_e}, E_{ns})$.

E Maximum of energy distribution in SELEX with fixed selection stringency

We want to determine how the position of maximum $E_m^{(n)}$ of $\rho_M^{(n)}(E)$ (see Eq. (4)), changes with the number of performed SELEX rounds n. We first shift zero of energy to coincide with $\overline{E}_{l_e}$ (see Eqs. (28) and (29)). Position of maxima of $\rho_M^{(n)}(E)$ is given by:

$$\frac{d\rho_M^{(n)}(E)}{dE} \, |_{E=E_m^{(n)}} = 0 \tag{40}$$

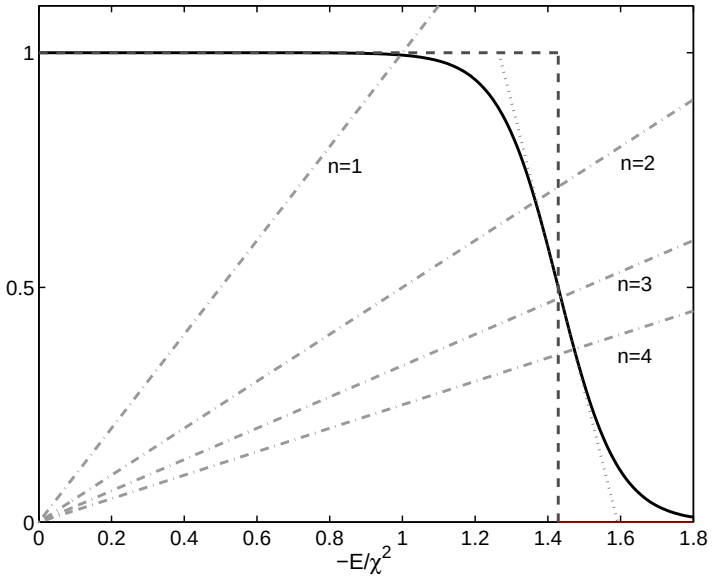

Figure 6: Full lines and dashed lines present $1 - f(E - \mu)$ and unit step approximation to $1 - f(E - \mu)$ respectively. Dash-doted lines present family of lines $\phi_n(E) = -E/(n\chi_{l_e}^2)$, which is here plotted for $n = 1, ..., 4$. The parameter values on this figure are $\chi_{l_e} = 3.5$ and $\mu = -5\chi_{l_e}$. Positions of $E_M^{(n)}$ are determined by the intersections of dash-dot lines and solid curve. Note that, for the parameters we choose, $E_M^{(1)} = -\chi_{l_e}^2$, while $E_M^{(n)} \approx \mu$ for $n \geq 2$.

From Eq. (4) and the equation above, we obtain:

$$[1 - f(E_m^{(n)} - \mu)] = -\frac{E_m^{(n)}}{n\,\chi_{l_e}^2} \tag{41}$$

The equation above can be solved graphically, i.e. positions of $E_m^{(n)}$ (for different n) are determined by the intersections of the family of lines $\phi_n(E) = -E/(n\chi_{l_e}^2)$ and the curve $[1 - f(E - \mu)]$. If we approximate $f(E - \mu)$ by a step $\theta(\mu - E)$ (see the dashed line in Fig. 6), we obtain that $E_m^{(n)} = -n\,\chi_{l_e}^2$ for $n < -\mu/\chi_{l_e}^2$ and $E_m^{(n)} = \mu$ otherwise, which is shown in Fig. 6. Correction, accurate up to the next order in $1/\chi_{l_e}^2$ can be found by linearizing $f(E - \mu)$ around $E = \mu$ (see dotted line in Fig. 6). We then obtain:

$$E_m^{(n)} = \frac{(\mu - 2)}{1 + 4/(\chi_{l_e}^2\, n)} \tag{42}$$

if $n > (-\mu + 2)/\chi_{l_e}^2$, and $E_m^{(n)} = -n\,\chi_{l_e}^2$ otherwise. Note that, for n large enough, i.e. $n \gg 4/\chi_{l_e}^2$, $E_m^{(n)} \to (\mu - 2)$. Since in SELEX experiments μ is typically positioned in the tail of the energy distribution of random binders, we expect $\mu \gg 1$ (see also the comment below Eq. (42)), so $E_m^{(n)}$ for $n > -\mu/\chi_{l_e}^2$ is well approximated by $E_m^{(n)} \approx \mu$ (see also Fig. 6 and Fig. 3).

F Computing energy matrix from extracted SELEX sequences

We determine the energy matrix $\widetilde{\epsilon}$, and the parameter γ by maximizing the likelihood function Λ (see Eqs. (5) and (6)), subject to the constraints:

$$\sum_{i,\alpha} p_\alpha \, \widetilde{\epsilon}_{i,\alpha}^2 = 1 \tag{43}$$

and

$$\sum_{\alpha} p_\alpha \, \widetilde{\epsilon}_{i,\alpha} = 0 \; (\forall i) \tag{44}$$

The constraint Eq. (43) follows from $\widetilde{\epsilon}_{i,\alpha} = \epsilon_{i,\alpha}/\chi$ and $\chi^2 = \sum_{i,\alpha} p_\alpha \epsilon_{i,\alpha}^2$ (see Eq. (14)), while the constraint Eq. (44), shifts columns of $\widetilde{\epsilon}_{i,\alpha}$ (see the footnote 4 in Subsection 4.2) so that $\overline{\epsilon}_i = 0$ and consequently $\overline{E} = 0$ (see Eqs. 12 and 13). As discussed in Section 4 and Appendix C, we approximate the binding probability $f(E(S) - \mu)$ in Λ by using the strongest binding site s_M with length L, on each sequence S ($S \in A$) of length l. To simplify notation, we further in this Appendix use $s_M \equiv s$. If we use μ/χ from Eq. (8), variation of Λ with respect to $\widetilde{\epsilon}_{i,\alpha}$ and γ leads to:

$$\begin{aligned} \frac{\partial \Lambda}{\partial \widetilde{\epsilon}_{i,\alpha}} \;=\; & -(n\chi) \sum_{s \in A} [1 - f\left(E\left(s\right) - \mu\right)] \left(s_{i,\alpha} - s^*_{i,\alpha}\right) - \\ & -(n\chi)\, \gamma \int f^n \left(E - \mu\right) \left[1 - f\left(E - \mu\right)\right] \rho_M \left(E\right) dE \, s^*_{i,\alpha} - \\ & -2\alpha \, p_\alpha \widetilde{\epsilon}_{i,\alpha} - \lambda \, p_\alpha = 0, \end{aligned} \tag{45}$$

$$\frac{\partial \Lambda}{\partial \gamma} = \frac{n_S}{\gamma} - \int f^n \left(E - \mu\right) \rho_M \left(E\right) dE = 0, \tag{46}$$

where α and λ are the Lagrange multipliers associated with the constraints Eq. (43) and Eq. (44) respectively, while $\rho_M\left(E\right)$ is given by Eq. (28). Eliminating α, λ and γ, using respectively Eqs. (43), (44) and (46), leads to the equation that implicitly determines $\widetilde{\epsilon}_{i,\alpha}$:

$$\widetilde{\epsilon}_{i,\alpha} = \frac{\frac{p_\alpha^{-1}}{n_s} \sum_{s \in A} [1 - f\left(E\left(s\right) - \mu\right)] \left(s_{i,\alpha} - s^*_{i,\alpha}\right) + \left(s^*_{i,\alpha}/p_\alpha - 1\right) \left(1 - \nu^{(n+1)}\right)}{\frac{1}{n_s} \sum_{s \in A} [1 - f\left(E\left(s\right) - \mu\right)] \left(E\left(s\right) - \mu\right) + \mu \left(1 - \nu^{(n+1)}\right)}, \tag{47}$$

where

$$\nu^{(n+1)} = \frac{\int f^{n+1} \left(E - \mu\right) \rho_M \left(E\right) dE}{\int f^n \left(E - \mu\right) \rho_M \left(E\right) dE}. \tag{48}$$

From Eq. (2) can be observed that $\nu^{(n+1)}$ is equal to the fraction of DNA that is in complex with protein in the $(n+1)^{th}$ round of SELEX, when non-specific binding is small (i.e. n large enough, as discussed in Section 4.1).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.13	0.23	0.24	0.13	-0.07	-0.16	0.03	0.02	0.00	0.05	0.18	0.19	0.24	-0.54	-0.31
T	-0.29	-0.54	0.24	0.23	0.18	0.01	-0.02	0.02	0.03	-0.17	-0.07	0.21	0.24	0.22	0.15
C	-0.04	0.20	0.24	0.24	-0.32	0.07	-0.03	0.00	0.03	0.07	0.21	-0.64	-0.72	0.13	0.21
G	0.20	0.11	-0.72	-0.61	0.21	0.09	0.02	-0.04	-0.06	0.04	-0.33	0.24	0.24	0.19	-0.05

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.07	0.28	0.26	-0.15	-0.15	-0.11	0.04	0.04	0.01	0.01	0.18	-0.02	0.26	-0.39	-0.23
T	-0.22	-0.40	0.26	0.23	0.21	0.02	-0.03	0.02	-0.01	-0.11	-0.12	0.08	0.26	0.29	0.07
C	-0.05	0.18	0.26	0.49	-0.25	0.05	-0.01	-0.01	0.03	0.07	0.21	-0.55	-0.79	-0.04	0.26
G	0.20	-0.05	-0.79	-0.56	0.19	0.04	0.00	-0.04	-0.03	0.03	-0.27	0.50	0.26	0.15	-0.10

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	0.06	0.19	0.29	0.16	-0.03	-0.04	0.01	0.01	0.00	0.01	0.11	0.21	0.29	-0.48	-0.16
T	-0.16	-0.50	0.29	0.23	0.11	0.01	-0.01	0.01	0.00	-0.05	-0.01	0.23	0.29	0.19	0.07
C	0.01	0.18	0.29	0.23	-0.18	0.02	0.00	0.00	0.01	0.02	0.11	-0.69	-0.86	0.12	0.10
G	0.09	0.12	-0.86	-0.63	0.10	0.02	0.00	-0.01	-0.01	0.02	-0.21	0.25	0.29	0.17	-0.01

Table 1: Upper table: The finite T energy matrix for CTF/NFI transcription factor. Middle table: The information-theoretic weight matrix. Lower table: The MatInspector weight matrix.

In practice, we typically do not know the exact value of χ, so we fix it to some reasonable value and then (numerically) solve Eq. (47) with respect to $\widetilde{\epsilon}_{i,\alpha}$. We can estimate the reasonable range of χ values, if we adopt the so-called "two state" model [2], in which (a same) penalty in binding energy ϵ_0 is assigned for each nucleotide that does not match the consensus sequence. Since one or two hydrogen bonds are formed per contact of the TF surface with a prefered nucleotide (energy of a hydrogen bond is $\sim k_B T$), ϵ_0 can be estimated to be $(1 \sim 3)\, k_B T$ [25]. From Eq. (14) follows that $\chi \sim \sqrt{L}\epsilon_0$, and with $L = 12$ for CTF/NFI (we are ignoring the contribution of the 3bp spacer to the binding energy), we obtain that χ is expected to take values from 3 to 12. Numerical solutions of Eq. (47) show that for χ in this range, the quantity $\widetilde{\epsilon}_{i,\alpha}$ depends weakly on the imposed value of χ. More precisely, we tested that $\widetilde{\epsilon}_{i,\alpha}$, which corresponds to solving Eq. (47) with (different) values of χ in the indicated range, leads to negligible differences in energy distribution and DET curve for finite T energy matrix (data now shown), which justifies solving the equation without the knowledge of the exact value of χ. The energy matrix $\widetilde{\epsilon}_{i,\alpha}$ obtained by our method (corresponding to $\chi = 5$) is given in Table 1.

Finally, we use Table 1 to discuss the differences between the weight matrix parameters estimated by the three different methods. In Fig. 7A histograms of energy levels corresponding to the finite T energy matrix and to the information-theoretic weight matrix are shown together. Similarly Fig. 7B compares the matrix parameters corresponding to the finite T energy matrix and the MatInspector weight matrix. The figure directly indicates which parameters are most different between the two matrices. For example, there is $\sim$ 150% difference between the information-theoretic and our matrix at the position 2G, $\sim$ 100% difference at the positions 4C and 5A, etc. Similarly, comparison with MatInspector weight matrix in Fig. 7B shows that there are significant differences (from 40% to 100%) between the two matrices at the positions 1, 5 and 6 while the matrices mostly agree with each other at the positions 2,3 and 4. Since

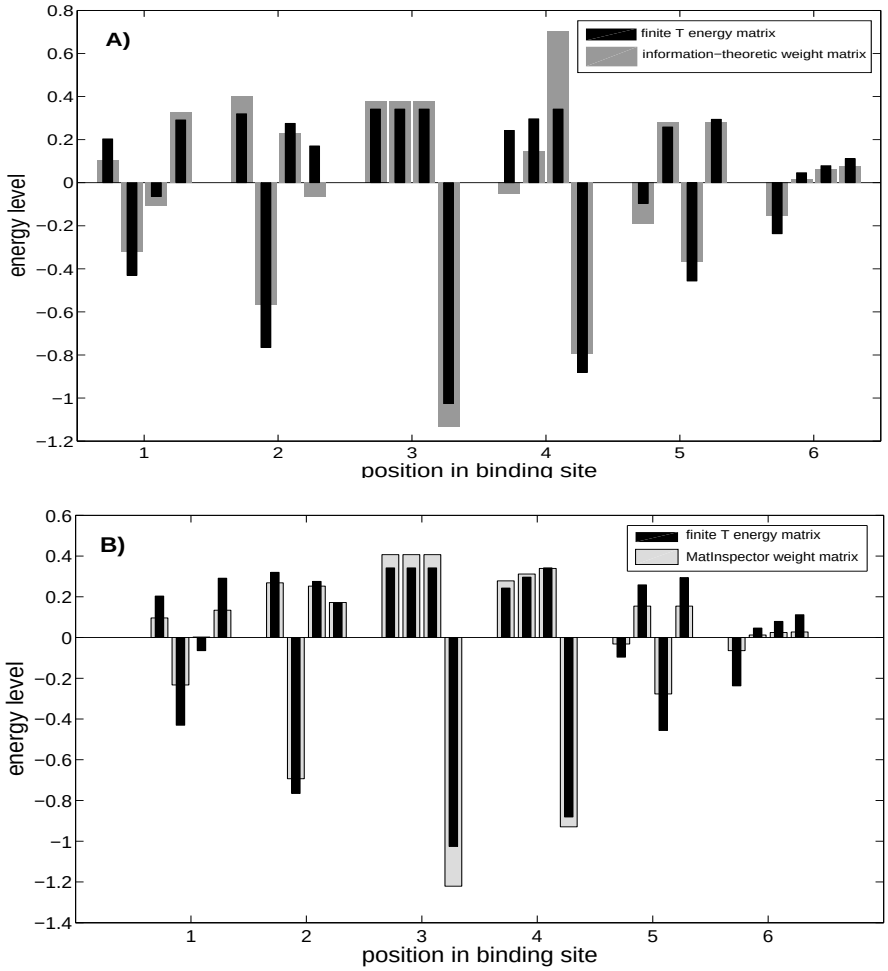

Figure 7: Comparison of the finite T matrix elements with A) the information theory weight matrix and B) the MatInspector matrix elements. CTF/NFI binds DNA as a homodimer, and recognizes two 6 bp long palindrome symmetric motifs, separated by a 3 bp spacer. Consequently, matrices in Table 1 were appropriately symmetrized and only the first six positions are shown. Four bars at each position on the horizontal axis correspond to A, T, C and G respectively, while heights of the bars correspond to the values of the matrix elements.

positions 2, 3 and 4 contribute more to the binding energy than positions 1, 5 and 6, the smaller difference between the DET curves of the finite T and the MatInspector matrices (as compared to the difference between the finite T and the information theory DET curves) can be attributed to the localization of the matrix differences at the less conserved positions. We note that, while the differences between the individual matrix elements are generally not very large, when binding site scores are calculated the individual differences add up to produce significant differences in e.g. the false positive/false negative trade-off shown in Fig. 5.

References

[1] B. Lewin, *Genes VII*, Oxford Univ. Press, Oxford, UK (2000).

[2] O. G. Berg, and P. H. von Hippel, Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters, *J. Mol. Biol.* **193**, 723 (1987).

[3] G. D. Stormo, DNA binding sites: representation and discovery, *Bioinformatics* **1**, 16 (2000).

[4] G. D. Stormo, T. D. Schneider, L. Gold and A. Ehrenfeucht, Use of "Perceptron" algorithm to distinguish translation initiation sites, *Nucl. Acids Res.* **10**, 2997 (1982).

[5] R. Staden, Computer Methods to locate signals in nucleic acid Sequences, *Nucl. Acids Res.* **12**, 505 (1984).

[6] T. D. Schneider, G. D. Stormo, L. Gold and A. Ehrenfeucht, Information content of binding sites on nucleotide sequences, *J. Mol. Biol.* **188**, 415 (1986).

[7] M. Djordjevic, A. M. Sengupta and B. I. Shraiman, A biophysical approach to transcription factor binding site discovery, *Genome Res.* **13** (11), 2381 (2003).

[8] K. Frech, K. Quandt and T. Werner, Finding protein-binding sites in DNA sequences: the next generation, *Trends Biochem. Sci.* **22**, 103 (1997).

[9] K. Robinson, A. M. McGuire and G. M. Church, A comprehensive Library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome, *J. Mol. Biol.* **284**, 241 (1998).

[10] E. Wingender *et al.*, The TRANSFAC system on gene expression regulation, *Nucleic Acids Res* **29**, 281 (2001).

[11] H. Salgado *et al.* RegulonDB (version 4.0): Transcriptional Regulation, Operon Organization and Growth Conditions in Escherichia coli K-12, *Nucleic Acids Res.* **32**, 303 (2004).

[12] C. Tuerk and L. Gold, Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA polymerase, *Science* **249**, 505 (1990).

[13] Gold L., Oligonucleotides as Research, Diagnostic and Therapeutic Agents, *J. Biol. Chem.* **270**, 13581 (1995).

[14] E. Roulet, S. Busso, A. A. Camargo, A. J. G. Simpson, N. Mermod and P. Bucher, High throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites, *Nature Biotech.* **20**, 831 (2002).

[15] V. E. Valculescu, L. Zhang, B. Vagelstein and K. W. Kinzier, Serial Analysis of Gene Expression, *Science* **270**, 484 (1995).

[16] L. Gold, D. Brown, Y.Y. He, T. Shtatland, B. S. Singer and Y. Wu, From oligonucleotide shapes to genomic SELEX: Novel biological regulatory loops, *Proc. Natl. Acad. Sci. USA* **94**, 5964 (1997).

[17] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. H. Horn, H. A. Erlich, N. Arnheim, Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia, *Science* **230**, 1350 (1985).

[18] D. Irvine, C. Tuerk and Gold L., SELEXION-Systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis, *J. Mol. Biol.* **222**, 739 (1991).

[19] B. Vant-Hull, A. P. Baez, R. H. Davis and L. Gold, The Mathematics of SELEX Against Complex Targets, *J. Mol. Biol.* **278**, 579 (1998).

[20] G. D. Stormo, and D. S. Fields, Specificity, free energy and information content in protein-DNA interactions, *Trends Biochem. Sci.* **3**, 109 (1998).

[21] Y. Cui, Q. Wang, G.D. Stormo and J.M. Calvo, A Consensus Sequence for Binding of Lrp to DNA, *J. Bacteriol.* **177**, 4872 (1995).

[22] P. V. Benos, M. L. Bulyk and G. D. Stormo, Additivity in protein-DNA interactions: how good an approximation is it?, *Nucleic Acids Res.* **30**, 4442 (2002).

[23] Y. Takeda, A. Sarai and V. M. Rivera, Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments, *Proc. Natl. Acad. Sci. USA* **86**, 439 (1989).

[24] A. Sarai and Y. Takeda, Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences assymetrically, *Proc. Natl. Acad. Sci. USA* **86**, 6513 (1989).

[25] U. Gerland, D. J. Moroz and T. Hwa, Physical constraints and functional characteristics of transcription factor-DNA interaction, *Proc. Natl. Acad. Sci.* **99**, 12015 (2002).

[26] A. M. Sengupta, M. Djordjevic and B. I. Shraiman, Specificity and robustness of transcription control networks, *Proc. Natl. Acad. Sci. USA* **99**, 2072 (2002).

[27] B. Dubertret, S. Liu, Q. Ouyang, and A. Libchaber, Dynamics of DNA-protein interaction deduced from in vitro DNA evolution, *Phys. Rev. Lett.* **86**, 6022 (2001).

[28] W. Peng, U. Gerland, T. Hwa and H. Levine, Dynamics of competitive evolution on a smooth landscape, *Phys. Rev. Lett.* **90**, 088103 (2003).

[29] Y. He, P. G. Stockley, and L. Gold, In Vitro Evolution of the DNA Binding Sites of Escherichia coli Methionine Repressor MetJ, *J. Mol. Biol.* **255**, 55 (1996).

[30] K. Quandt, K. Frech, H. Karas, E. Wingender and T. Werner, MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *Nucleic Acids Res.* **23**, 4878 (1995).

[31] C. E. Lawrence, S. F. Altschul, M. S. Bogouski, J. S. Liu, A. F. Neuwald and J. C. Wooten, Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment, *Science* **262**, 208 (1993).

[32] X. Liu, D. L. Brutlag, J. S. Liu, BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes, *Pac Symp Biocomput.*, 127 (2001).

[33] G. Z. Hertz and G. D. Stormo, Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, *Bioinformatics* **15**, 563 (1999).

[34] A. Martin *et al.*, The DET curve in assessment of detection task performance, *EuroSpeech* **4**, 1895 (1997).

[35] S. Kim, H. Shi, D. Lee and J. T. Lis., Specific SR protein-dependent splicing substrates identified through genomic SELEX, *Nucl. Acids Res.* **31**, 1955 (2003).

[36] J. W. Fickett and W. W. Wasserman, Discovery and modeling of transcriptional regulatory regions, *Curr. Opin. Biotechnol.* **1**, 19 (2000).

[37] Ryogo Kubo, Statistical Mechanics - An advanced course with problems and solutions, *North-Holland Publishing Company*, Amsterdam (1965).

[38] T. K. Man and G. D. Stormo, Non-independence of Mnt repressor operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay, *Nucleic Acids Res.* **15**, 2471 (2001).

[39] M. L. Bulyk, P. L. Johnson and G. M. Church, Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors, *Nucleic Acids Res.* **30**, 1255 (2002).

[40] R. A. O'Flanagan, G. Paillard, R. Lavery and A. M. Sengupta, Non-additivity in protein-DNA binding, *Bioinformatics* **21**, 2254 (2005).

[41] W. B. Anderson, A. B. Schneider, M. Emmer, R. L. Perlman and I. Pastan, Purication of and properties of the cyclic adenosine 3'-5'-monophosphate receptor protein which mediates cyclic adenosine 3'-5' monophosphate-dependent gene transcription in Escherichia coli, *J. Biol. Chem.* **246**, 5929 (1971).