4 The Inverse Problem

Up to now, our consideration of the influences on nucleosome positioning has been mostly restricted to effects arising from competitive binding -either between the nucleosomes themselves, or between nucleosomes and transcription factors- and the statistical positioning that arises from this interaction. In Chapter 2 we also introduced one experimentally-motivated mechanism for remodeller action that was shown to be particularly important at low histone densities.

In addition to these effects, however, another ever-present influence on nucleosome positioning is the inherent binding affinity of different locations along the DNA encoded in its sequence. We refer to this as the ‘energy landscape’ of the genome: regions where the landscape is low are attractive to nucleosome formation, while regions of high potential are repulsive.

Omission of this effect in Chapter 2 was justified by taking averages over many different genes throughout the yeast genome, with alignment to +1 nucleosome barrier. There, the assumption was that the sequence-encoded features over many genetic loci across the genome would average out to an approximately flat landscape, and that there is insignificant *systematic* trend, genome-wide, in the landscape relative to +1 nucleosomes. Although this latter assumption was not proven, the resulting analysis was predicated to some degree on Occams Razer. As Bertrand Russel put it: “Whenever possible, substitute constructions out of known entities for inferences to unknown entities.”[77]. With this advice in mind, we assert that if sequence-dependent landscape features are not *necessary* to reproduce gene-averaged experimental data, then an appropriately parsimonious model should not include them. At the single-gene level, however, landscape features play a much more prominent role.

Thus, we return to this omission and address the effects on nucleosome positioning from the sequence-encoded landscape itself, a subject in which a great deal of work has already been invested. For example, it is known that in *Saccharomyces cerevisiae*, certain sequences of DNA, such as Poly(dA:dT) tracts are associated with nucleosome-depleted regions, perhaps due to their inherent stiffness against bending[24, 78], though this has been shown not to be consistent with other species of yeast, such as *Schizosaccharomyces pombe*[79]. Indeed, there is also evidence supporting the additional stiffness of the AA/TT bond at the bp level while the stiffnesses of other pair bonds seem not to obey any simple rule[80]. This suggests that the effective histone affinity of a given stretch of DNA is actually determined by a complex interplay of many factors.

Kaplan *et al* have developed a computational model to predict histone affinity from an arbitrary sequence across the yeast genome[73], while Gossette and Lieb have shown that under artificially reduced histone conditions, sequence effects become increasingly important in determining which nucleosomes are preserved, and which are not[81]. From a more theoretical perspective, Chereji *et al* have proposed a means for inferring the landscape based on a presumed 10-bp periodicity in linker length[82, 83], while Teif and Rippe have proposed a model for predicting nucleosome positions based on sequence and remodeling activity[84]. For these and many more reasons, the influence of the energetic landscape cannot be excluded from discussion of nucleosome positioning at the single-gene level.

An important caveat here is that we are referring to an *effective* landscape in the sense that it represents the net influence of all non-translationary-invariant nucleosome positioning effects. The elasticity-dependence of DNA on sequence is a fairly intuitive mechanical dimension of the energy landscape, as discussed above. Another, less obvious, component of the landscape, however, comes in the form of active remodellers that target specific loci for remodelling and repositioning of nucleosomes.[85, 86, 87] The SWI/SNF family of remodellers, for example, tends to target acetylated histones and serves to clear histones from the nucleosome free region. The ISWI family of remodellers, on the other hand, activates 'sliding' at the interface between nucleosomes and linker DNA; ISWI2 activity is focused at the +1 nucleosome while ISWI1 tends to have greater influence on the spacing of the nucleosomes downstream[25]. Naturally, then, changes in remodelling activity lead to altered nucleosome positioning that will be reflected in the effective landscape of a given region of DNA. The following calculation of this landscape is not intended to evince the precise mechanism responsible for preferred nucleosome positioning, but rather to capture the global effective specificity of nucleosome positioning arising –directly or indirectly– from the genetic sequence, all things considered. One may hope that mechanistic understanding can then be gleaned from controlled changes in the conditions under which the landscape is determined.

4.1 Introducing the Potential Landscape

For explicit inclusion of sequence-dependence in our lattice model, we consider 4 related quantities which can be subdivided into analogous pairs. The energetic landscape, $V(x)$, is the focus of this chapter and captures the potential energy inherent in a one-dimensional particle[1] being bound to a substrate at position x in isolation –i.e. without any energetic costs incurred from neighboring particles. The latter are explicitly included in the two-body interaction potential $\phi(x, x')$ between a pair of neighboring particles at x and $x' > x$.

The density of particles at a position x is defined as $n_1(x)$, while the two-body pair density $n_2(x, x')$ describes the density of nearest-neighbor pairs at x and x'. The analogy between the two pairs of quantities $[n_1(x), n_2(x, x')]$ and $[V(x), \phi(x, x')]$ relates to one- and two-body effects and either pair can be inferred from the other, as will be shown below. To do so however, we must take note of two simplifications implicit in this starting point.

First, there is, in principle, no reason *a priori* why we must restrict ourselves to one- and two-body effects. For example, a so-called 'three body' potential $\phi_3(x, x', x'')$ and 'triplet density' $n_3(x, x', x'')$ could be introduced to represent nucleosome clustering due to, for example, the influences of higher-order chromatin structure that extend beyond nearest-neighbors. Similar higher-order interactions and distributions could be extended to arbitrarily high order, with corresponding complexity. Restricting nucleosome interactions to nearest-neighbors, however,

[1] throughout this chapter, we will refer more generally to 'particles' in one-dimension –nucleosomes along DNA being one such particle

is in line with the mechanistic assumptions about neighbor interactions made in previous chapters, and renders the problem analytically tractable.

Secondly, it is conceivable that the neighbor interaction ϕ could itself exhibit some position dependence that is not captured purely by the difference in position between the neighboring particles. Based on the physical argumentation from Chapter 2, however, we assume a translationally invariant competitive binding interaction, with unwrapping cost ε per bp

$$\phi(x, x') = \phi(x' - x) = \phi(\Delta x). \tag{4.1}$$

For the form of ϕ in Eq. 4.1, later in this chapter we will simply take $\phi(\Delta x) = v(\Delta x)$ from Chapter 2, but for the moment we will use ϕ to underscore the generality of the following argument. To some degree, both of these simplifications are, again, motivated by the fact that $V(x)$ and $\phi(\Delta x)$ represent 'known constructions', as there is no apparent need to invoke a more complex set of interactions.

4.2 The 1-D Inverse Problem

Converting between 1- and 2- body particle densities and 1- and 2- body energetic potentials for nearest-neighbor interacting particles along a 1-D axis has been considered[88, 89, 90]. This work by Percus provides the fundamental tools for the calculations in this chapter; the derivation below from Eq.s 4.2 through 4.15 represents recapitulation of the relevant aspects of this work, with some added commentary and a minor correction midway, though the conclusion is identical. We follow this reasoning explicitly since an understanding of this methodology is necessary for the application that will follow.

For a 1-D lattice of size L, we assume a corresponding basis set of vectors $|i\rangle$ (with length L) that denote the position of particle x_i. That is to say, if particle x_n is at position m, then the m-th elements of $|n\rangle$ is 1 and all others are zero. This leads to the unconstrained boundary condition vector

$$|J\rangle = \sum_{n=0}^{L} |n\rangle. \tag{4.2}$$

Hence, if all particle interactions and positions were energetically neutral, our partition function for a single particle would be simply $Z = \langle J|\, J\rangle$. For a particular particle (say, the i-th particle $\langle i|\, J\rangle = 1$) We define $\phi(1, 2) = \langle 1|\phi|2\rangle$ to be the energy of interaction between the two particles $x_1 < x_2$ and thus

$$\langle 1|w|2\rangle = \mathrm{e}^{[-\beta\phi(1,2)]}\Theta(x_2 - x_1) \tag{4.3}$$

where Θ is the heaviside function, and β is inverse temperature. Again, our application allows for the more restricted $\phi(1, 2) \to \phi(\Delta_{1,2})$, but the more general form is left in place. The diagonal matrix e contains the weights of particle positions in isolation

$$\langle 1|e|2\rangle = \mathrm{e}^{[-\beta V(x_1)]}\delta(x_2 - x_1) \tag{4.4}$$

where the diagonal elements of e are simply the Boltzmann factors of $V(x)$. Assuming a given particle number N, we define a canonical partition function Q_N. For $N = 1$, the lone particle has no neighbors with which to interact, and thus a partition sum of

$$Q_1 = \langle J|\, e \,|J\rangle\,, \tag{4.5}$$

accounting for all possible positions of the particle. Naturally then, the partition sum for two particles involves two Boltzmann weights for specific binding in addition to a single interaction matrix

$$Q_2 = \langle J|\, ewe \,|J\rangle\,. \tag{4.6}$$

Likewise, for arbitrary N,

$$Q_N = \langle J|\, (ew)^{N-1}e \,|J\rangle\,, \tag{4.7}$$

Assuming $N > 0$, otherwise $Q_0 = 1$ corresponding to the empty state.

In the *grand canonical* ensemble, however, we must sum over the set of all possible numbers of particles N. To accomplish this, we augment the Boltzmann matrix of specific binding with an overall chemical potential μ that represents the basic non-specific binding energy for nucleosomes to be formed *anywhere*. Hence, we replace the previous matrix e with $z = \mathrm{e}^{[\beta\mu]}e$, yielding:

$$\langle 1|\, z \,|2\rangle = \mathrm{e}^{[\beta(\mu - V(x_1))]}\delta(x_2 - x_1). \tag{4.8}$$

The grand canonical partition function is then:

$$Z = \sum_{N=0}^{L} \mathrm{e}^{[\beta N\mu]}Q_N = 1 + \sum_{N=0}^{L} \langle J|\, (zw)^{N-1}z \,|J\rangle\,. \tag{4.9}$$

At this point, we exploit the matrix generalization of the geometric series (or, the Neumann series)

$$\sum_{k=0}^{n} A^k = (I - A)^{-1}(I - A^{n+1}), \tag{4.10}$$

assuming $(I - A)$ is invertible, and the eigenvalues of A are less than unity. In our case, $k \to N - 1, A \to zw$, and $A^L \to 0$ for large L (i.e. the probability of having an entire array stacked with particles *everywhere* is vanishingly small) and truncation artifacts of this series can be neglected. Hence

$$Z = 1 + \sum_{N=1}^{L} \langle J|\, (zw)^{N-1}z \,|J\rangle = 1 + \langle J|\, (I - zw)^{-1}z \,|J\rangle\,, \tag{4.11}$$

which differs slightly from Eq. 2.6 of Ref. [90]. By extending this argument in either direction from a designated particle, the 1−particle density distribution $n(1)$, is seen to be:

$$\begin{aligned} n(1)Z &= \langle J|\,(I-zw)^{-1}z\,|1\rangle\,\langle 1|\,z(I-wz)^{-1}\,|J\rangle \\ &= \langle J|\,(I-zw)^{-1}\,|1\rangle\,z(x_1)\,\langle 1|\,(I-wz)^{-1}\,|J\rangle \end{aligned} \tag{4.12}$$

Likewise, the 2−particle neighbor function is given by

$$n_2(1,2)Z = \langle J|\,(I-zw)^{-1}\,|1\rangle\,z(x_1)\,\langle 1|\,w\,|2\rangle\,z(x_2)\,\langle 2|\,(I-wz)^{-1}\,|J\rangle\,. \tag{4.13}$$

Eq.'s 4.12 and 4.13 suffice to carry out what could be termed the 'forward' calculation. That is to say, the determination of the 'effects' ($n_1(x)$, and $n_2(x,x')$) from the 'causes' ($V(x)$ and $\phi(x,x')$). The above description also suffices to present a qualitatively clear understanding of the relevance of this work. The remaining calculations to derive the analogous 'inverse' calculations (i.e. inferring the energetic interactions based on the observed 1- and 2-body density patterns) can be found in the original Ref. [90]. The results of this calculation, are:

$$\begin{aligned} \beta\left[u(x_1)-\mu\right] &= \ln\left[1-\langle J|\,(I-n_2n^{-1})n\,|J\rangle\right]-\ln\left[\langle J|\,(I-n_2n^{-1})\,|J\rangle\right] \\ &- \ln\left[n(x_1)\right]-\ln\left[\langle 1|\,(I-n^{-1}n_2)\,|J\rangle\right] \qquad (4.14) \\ \beta\phi(x_1,x_2) &= \ln\left[\langle 1|\,(I-n^{-1}n_2)\,|J\rangle\right]+\ln\left[\langle J|\,(I-n_2n^{-1})\,|2\rangle\right] \\ &- \ln\left[\langle 1|\,(n^{-1}n_2n^{-1})\,|J\rangle\right]-\ln\left[1-\langle J|\,(I-n_2n^{-1})n\,|J\rangle\right] \qquad (4.15) \end{aligned}$$

Eq.'s 4.12–4.15 define the forward and reverse calculations for the two pairs of quantities discussed above. In principle, they sufficiently constrain the landscape and two-body interaction potential provided the full statistics of particle positioning can be obtained. However, for nucleosomes, such statistical data is often incomplete in practice, and in the following section we explore how the above analysis can be practically applied to nucleosome data.

4.3 Application to Nucleosome Positioning Data

The preceding has shown the feasibility of interchangeably calculating the pair of 1- and 2-body energetic quantities from corresponding density quantities and vice-versa. Given the availability of MNase positioning data, and the neighbor interactions described in previous chapters, it may seem that the remaining quantities of interest can be easily calculated.

However, while experimental techniques provide n_1 up to a normalization factor, the process of read-collection omits all information related to $n_2(x,x')$. Since many cells are used in the process, it is impossible to differentiate which mononucleosomes had previously been adjacent to one-another on the same genome. Conversely, the theoretical discussion from Chapter 2 provides a description of neighbor-interactions $\phi(x) \to v(x)$, but omits any consideration of the potential landscape, V. Thus, both the forward and reverse calculations above are missing one half of the necessary input.

In the face of this conundrum, an iterative scheme to calculate V was developed. This scheme relies on taking the theoretical description of ϕ given in Chapter 2 (i.e. the SoNG potential $v(\Delta x)$) along with successive candidate functions for V nominated iteratively to calculate n_1 theoretically until deviation from experimental data is minimized.

In this scheme then, the landscape value $V(x)$ is a free parameter at every value of x, and convergence at a solution for $V(x)$ requires optimization over an L-dimensional parameter space. For a typical system of L in the thousands, an exhaustive mesh comparison over all $V(x)$ values would be computationally infeasible, and so an optimized search algorithm is required.

High-dimensional numeric optimization is a field that encompasses many different techniques; one of the most standard[91] of these is the Nelder-Mead, or 'amoeba' search algorithm[92]. This method has been adopted for high-dimensional numeric optimization successfully in a variety of fields. Its power lies in its ability to infer regional trends in the function's slope from sample evaluations without needing to explicitly calculate the gradient of the function at any particular point. For these reasons, it is well-suited to our goal of determining $V(x)$.

4.3.1 The Amoeba Method

We seek a numeric scheme to numerically minimize a functional f of the euclidean distance between the nucleosome read counts observed experimentally $\Omega(x)$, and the density predicted theoretically from Eq. 4.12, $n_1(x)$, with optimization over the set of parameters $V(x)$ for all positions x. We define the functional subject to minimization explicitly as:

$$f([V(x)], \phi(\Delta x), \Omega(x)) = \sum_x [\Omega(x) - \alpha n_1(x)]^2, \tag{4.16}$$

where α is an unknown normalization constant associated with the number of cells sampled in the experiment, and is assigned analytically

$$\alpha = \left(\sum_x n_1(x)\Omega(x) \right) / \left(\sum_x n_1(x)^2 \right) \tag{4.17}$$

(See Ref. [36], SI). We assume that f is a smoothly-varying function in the space of $V(x)$, and while $n_1(x)$ is written merely as a function of x, dependence on V and ϕ, as described above, is assumed. Clearly, f depends on the experimental read counts $\Omega(x)$, as well as the set of parameters $V(x)$ to be optimized, and the neighbor interaction $\phi(\Delta x)$ which we assume to be the SoNG potential $v(\Delta x)$ from Chapter 2.

To implement the amoeba search algorithm, a set of $L + 1$ candidate functions $V_i(x)$ are generated (each of which correspond to their own unique energetic landscape), representing vector positions throughout this $L-$dimensional parameter space[2], and are each considered as possible solutions for optimization. A convenient choice for initial candidates is given by

$$V_i(x) = \begin{cases} 0 \,\forall\, x & \text{if } i = 0 \\ \pm 1 \cdot \xi\, \delta_{ix} & \text{if } i \in [1, L] \end{cases} . \tag{4.18}$$

In Eq. 4.18, the factor ξ represents an intuitive guess at the characteristic length scale over which $n_1(x)$ changes significantly with $V(x)$, and can be arbitrarily assigned a value of $1k_BT$ without significantly affecting the final results. With V_0 representing a baseline 'flat' energetic landscape, each other candidate landscape is initially shifted by either $+\xi$ or $-\xi$ (randomly, with equal probability) from zero at a single position. $\overline{V}$ is the centroid of these points in the L-dimensional space, and the L vectors $V_i - V_0$ form a linearly independent basis, the correct combination of which yields the point in the space which minimizes the functional f.

To determine that combination correctly, the initial candidate points V_i are ordered by their corresponding f_i values. V_l is then the point with the lowest f value, while V_h corresponds to the highest, and is iteratively replaced with transformed positions to reduce its f value. These transformations are described in more detail using the flow-chart in Appendix A.2, Fig. A.1 and are repeated until eventually all candidate points converge to the same position in parameter space with the same f value within numeric tolerance. A full description of this algorithm as well as the criteria for convergence can be found in Ref. [92].

4.3.2 An Illustrative Minimal Example

Even with the description in A.2, and the accompanying diagram, Fig. A.1, the application of the amoeba method to our landscape problem may still seem unintuitive. For this reason, it is helpful to consider a minimal illustrative example where $L = 2$ –an extremely simplified scenario.

In this case, the two degrees of freedom to the potential V are denoted $V(x)$, and $V(y)$ respectively, as shown in Fig. 4.1, and within this space, the optimum solution which leads to the correct density exactly **E** is sought (E for 'exact' is a useful mnemonic). To that end, three initial guesses are made in the 2-D parameter space of V_x, V_y in Fig. 4.1(b): points **A**, **B**, and **C**, color-coded red, blue and green in the figure respectively. Each of these points, however, also represents a candidate landscape sketched in Fig. 4.1(a). In this figure, **A** represents V_0 from Eq. 4.18 and is the default, or 'flat', featureless landscape, while **B** and **C** contain random linear displacements from **A** at positions x and y respectively. Since **C** has the highest scalar functional value in Fig. 4.1(b) it is reflected about the centroid to point **C***. Thereafter, **B** has the next poorest fit and will also be substituted. The algorithm will continue to make iterative substitutions until all candidates converge at **E**.

Obviously, calculating a genomic landscape requires considering far more than $L = 2$ adjacent positions (typical genomic segments under consideration range in sizes of approximately $L \approx 10,000$) and the dimensionality of such parameter space searches is not conducive to visualization. Nevertheless, the same principle applies: the set of $L + 1$ vertices form a simplex

[2] bp are treated as lattice sites here, ignoring the option of coarse-graining

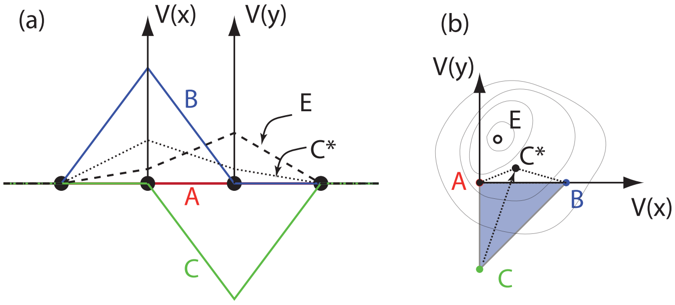

Figure 4.1: Schematic of a highly reduced inverse-landscape algorithm with only two values $V(x)$ and $V(y)$. (a), Various potential landscapes corresponding to the potential at these two positions are sketched: an initially flat landscape, **A**, along with two landscapes with random displacements at the two positions, **B**, and **C**. The optimal, or 'Exact' solution is the dashed line **E**, and two axes 'V(x)' and 'V(y)' indicate the degrees of freedom open to the system. All of these profiles correspond to a point in the 2-D plane in figure (b). Here, a simplex of $N + 1 = 3$ points is shown, lines of constant f are sketched in transparency. Points are drawn only approximately to scale. Since **C** has the highest f value, it is substituted with **C*** before the fitness of each points is reevaluated; upon the next iteration, **B** will be substituted for a new point **B***. Eventually, all points will converge to the exact solution **E** within numeric tolerance.

in the $L-$dimensional space which iteratively converge to a single position representing the global minimization in f.

As a proof of this principle, an intermediate test case is proposed in which the exact solution **E** is known, but 'hidden' from the algorithm, along a segment for $L = 100$. For this confirmation see appendix Fig. A.2.

4.3.3 The 601 Sequence

We may now proceed with application of this algorithm to actual experimental data using a sequence known particularly for its high-histone affinity[93, 94], which has come to be referred to as the 601 sequence. Nucleosomes reconstituted *in vitro* on plasmid DNA segments with histones and DNA mixed in mass ratio 1:1 and 0.5:1 respectively were sequenced, and the resulting positioning data was made available for the collaborative analysis above[95]. Naturally, the overall densities of nucleosome positioning differed, depending on histone mass concentration. However, assuming the above analysis is valid, the underlying potential landscape ought to be independent of histone density. For this application, two interactions $\phi(\Delta x)$were considered: (1) neighboring nucleosome interactions exactly as obtained in the SoNG model, as described in Eq. 2.5 from Chapter 2, and (2) the neighbor-potential from Eq. 2.5 in addition to a slight linear attractive component to account for possible clustering due to higher-order chromatin structure or histone-tail interactions. The results of the fits are shown in Fig. 4.2.

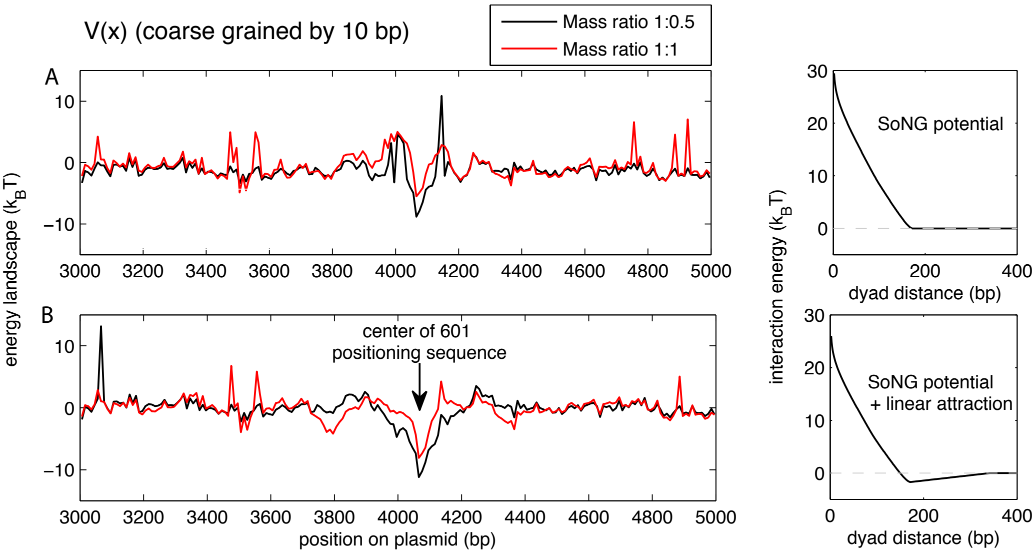

Figure 4.2: Landscape potential inferred from plasmid 601 sequence at different histone concentrations assuming two different neighbor interactions: (A) the SoNG model alone, above, as well as (B) the SoNG interaction in addition to a slight attractive potential to capture possible clustering effects, below. Both potentials are sketched to the right.

4.4 Discussion

It bears repeating that this landscape is an effective one: it incorporates direct effects from DNA due to, e.g. elasticity, as well as indirect specific sequence effects from targeted remodelling when such factors are present. In the data used in Fig. 4.2 there were no remodellers acting on the substrate, however analogous experiments were carried out with remodellers re-introduced and resulting density profiles were compared to the previous as a control. Since different classes of remodellers are known to target different segments of the gene[25], it was proposed that the change in density patterns that can be observed upon addition of different remodellers[95] could be translated to an effective energetic potential -or gradient thereof- in the various segments of the DNA.

Although characteristic changes in density patterns were clearly observed under the influence of remodellers, no additive potential could be associated with each, and further study is needed to make definitive claims about the degree to which $V(x)$ is determined from active remodelling.

One avenue for further development of this technique, however, would be the change in the effective nucleosome potential under changing cell conditions that affect the binding of Transcription factors at known loci. As observed in Chapter 3, competing proteins such as specifically-binding transcription factors influence the binding of histones, and such an effect would be included in the measure of the energetic nucleosome landscape. Such a change in the effective potential could be attractive or repulsive, and the relative change could serve as an indication of relative changes in TF binding activity at particular loci under changing conditions.