Version 1.0 as of April 18, 2008
Primary Authors: David J. Schwab

Positioning Ambiguity in 1D Hard Core Lattices

(Dated: April 18, 2008)

One dimensional statistical mechanics has recently found applications in biology, in areas ranging from nucleosome positioning to transmembrane protein structure prediction to sequence alignment. In this paper, we address the kinetic accessibility and stability of the ground state of systems with hard core particles in a random energy landscape. We introduce the deviation of the ground state from the "quenched" configuration, i.e. that formed by the sequential filling of the lowest energy sites, as a measure of positioning ambiguity and hence potential kinetic traps in finding the true ground state. Both configurations are calculated exactly, and we find that they differ by an extensive amount for *all* non-zero densities due to geometric frustration. We then address the different functional roles certainty and ambiguity in positioning may play in transmembrane proteins and nucleosomes.

PACS numbers:

The problem of placing hard rods on a line is an old problem in statistical mechanics. It was first addressed in the 1930s by Tonks [1] who calculated the density of a $1D$ hard rod gas at fixed pressure and showed that no phase transition can occur. The equilibrium problem was later taken up by Percus, who extensively studied the effects of, among other things, an arbitrary external potential using a continuum density functional approach [2]. The solution for the particle density from the Percus equation turns out to be a formidable task [3]. On a lattice, however, one can of course use a simple transfer matrix approach to solve for the density. In fact, one can do even better and derive a discrete version of the Percus equation that can be efficiently iterated and has a simple physical interpretation [4].

In recent years, the tools developed by statistical mechanists for $1D$ problems have found use in biology, which provides a playground for one-dimensional physics. The clearest example is DNA. Many proteins, such as RNA polymerase and transcription factors, move along DNA searching for their specific target location, i.e. a particular sequence of base pairs. This is called direct read-out. Others, notably nucleosomes, feel the DNA sequence through the varying stiffness of dinucleotide sequences (indirect read-out) [5]. There is also the DNA (or protein) sequence itself. During evolution, stretches of DNA are cut and moved around, flipped, sections inserted, sections removed, etc. A fundamental problem is then to find matching sequences from different species' genomes that have the same origin, despite this random shuffling. Sequence alignment is one of the great successes of the interaction between biology and $1D$ statistical mechanics [6] and the basic mathematical structure is the same as that of finding the ground state of hard-core particles.

Another fruitful application of $1D$ statistical mechanics is the prediction of α–helix locations in transmembrane proteins. The formation of helices that bridge across the cellular membrane is driven by the hydrophobicity of the amino acids that make up the protein chain. Sections of $\sim$ 20 particularly hydrophobic amino acids are more likely to fold into an α–helix spanning the cell membrane, thereby creating a heterogeneous $1D$ effective "energy" landscape governing the location of helix formation.

Our main interest in this paper is quantifying the kinetic accessibility and stability of the $T = 0$ ground state of one dimensional hard-core particles to infer properties of the biological systems which they represent. The failure of proteins to find their correct folded configuration can result in diseases ranging from cystic fibrosis to Creutzfeldt-Jakob disease, and the incorrect positioning of nucleosomes along DNA could potentially induce dramatic changes in gene expression [4]. Rather than studying the actual dynamics of $1D$ systems, we will attempt to gain information through a different approach. We first define the quenched configuration, ρ_Q, to be that formed by the sequential filling of the lowest energy states up to a chemical potential μ and obeying the hard-core constraint. It is important to distinguish the usage of the word quenched from the phrase "quenched disorder". All our calculations will be done with quenched disorder in the external potential, but the phrase "quenched configuration" refers to the state just described. We would like to compare this configuration with the true $T = 0$ ground state, ρ_G, so the deviation between these two functions, $\Delta = \rho_G - \rho_Q$, will be the main quantity of interest. For larger values of Δ, there will be more metastable configurations, hence more kinetic traps and reduced equilibrium stability. For analytic calculations, we will take the simplest non-trivial case of particles with size $a = 2$ lattice sites, i.e. *dimers*, although when discussing applications the particles will be much larger ($a \sim 20$ for transmembane protein α–helices and $a \sim 150$ for nucleosomes). We don't expect this simplification to make a qualitative difference in our results.

Before calculating ρ_G and ρ_Q, we give a simple argument to show that these two quantities should differ by a finite amount for *all* $\rho_G > 0$. We work on a lattice in the grand canonical ensemble at chemical potential μ and particle size $a = 2$. We assume on-site energies (the

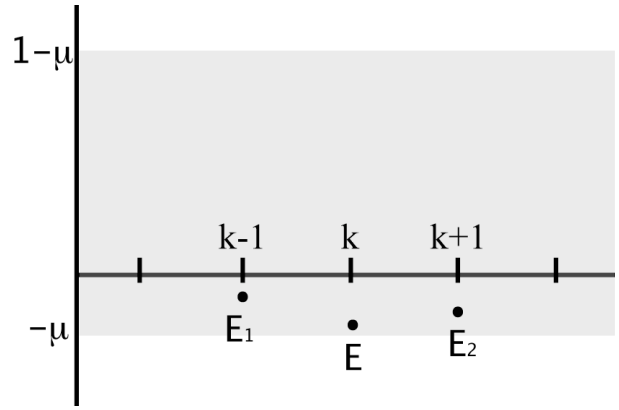

FIG. 1: Schematic of a scenario where the quenched configuration occupies site k, hence blocking sites $k-1$ and $k+1$, but the ground state foregoes occupancy of site k in favor of the flanking sites if $E_1 + E_2 < E$.

energy for *starting* a particle at site i) to be i.i.d. with a uniform distribution, denoted by $E(\epsilon)$, between 0 and 1. The uniform distribution will simplify the analytics but the main results should hold for other, i.e. Gaussian, distributions as well. When μ=0, all sites are repulsive, and when $\mu = 1$, all sites are attractive. Therefore, we will focus on μ between these values. In the following argument, we will absorb μ into the energies which will instead be uniformly distributed between $-\mu$ and $1-\mu$. Let the energy at site k be E such that $E < 0$ and the energies at sites $k-1$ and $k+1$ be E_1 and E_2, respectively. We calculate the probability, P_{switch}, that in the ground state, placement of a particle at k is forfeited in favor of the occupation of sites $k-1$ and $k+1$, despite site k having the lowest energy. The scenario is depicted in Figure 1. This is the simplest way the ground state and quenched configurations could differ because $\rho_G(k) = 0$ but $\rho_Q(k) = 1$. Averaging over the value of E,

$$P_{\text{switch}} = -\frac{\int_{-\mu}^{0} P(E_1 + E_2 < E)\mathrm{d}E}{\int_{-\mu}^{0} \mathrm{d}E} \tag{1}$$

Since the sites $k-1$ and $k+1$ must also be attractive, $P(E_1+E_2 < E) = \int_E^0 dE_1 \int_E^0 dE_2 \theta\left(E - (E_1+E_2)\right)$ which equals $E^2/2$, and we can see that $P_{switch} = \mu^2/6$. Multiplying this by the average number of sites with $E < 0$, i.e. μL, gives the average number of these switches in a finite system of size L. Setting this equal to 1 gives the chemical potential at which we expect the first switch: $\mu_c \sim L^{-1/3}$. Clearly, as $L \to \infty$, $\mu_c \to 0$. Also, since each such switch increases the ground state density relative to the quenched, we also have that $\Delta = \rho_G - \rho_Q \sim \mu^3/6$ for small μ, so $\Delta = 0$ only at $\mu = 0$. The above argument neglects the contributions of sites $k-2$ and $k+2$ etc. but these effects are higher order in μ and hence can be neglected for $\mu \ll 1$. We will see this behavior of Δ reproduced precisely from the exact result.

We now derive the disorder-averaged ground state density ρ_G for a $1D$ lattice of dimers at chemical potential μ and with on-site energies uniformly distributed between 0 and 1. The analysis follows that of Fonk and Hillhurst [7] who considered the problem with a different energy distribution not easily generalizable to a non-zero chemical potential. Define $E_k^1(E_k^0)$ to be the minimum energy of the first k sites subject to the constraint that a particle is present (absent) at site k. These quantities can be seen to obey the recursion relations,

$$E_k^1 = -\epsilon_k + E_{k-1}^0 \tag{2}$$

$$E_k^0 = \min\left(E_{k-1}^0, E_{k-1}^1\right) \tag{3}$$

where ϵ_k is the energy to start a particle at site k. Defining difference variables $\xi_k = E_k^0 - E_k^1$ and subtracting the above equations gives a recursion relation for ξ_k,

$$\xi_k = \epsilon_k + \min\left(0, -\xi_{k-1}\right) \tag{4}$$

which, by averaging over the on-site energy distribution $E(\epsilon)$, is readily transformed into an integral recursion relation for $P(\xi)$, the distribution function of ξ,

$$P_k(\xi) = E(\xi)\int_{-\infty}^{0} \mathrm{d}\xi' P_{k-1}(\xi') \tag{5}$$

$$+ \int_0^{\infty} \mathrm{d}\xi' E(\xi+\xi') P_{k-1}(\xi') \tag{6}$$

The fixed point distribution of $P(\xi)$ contains the required information for bulk quantities, so we can drop the subscripts on $P(\xi)$. With $E(\epsilon) = \theta(\epsilon+\mu)\theta(1-\mu-\epsilon)$, we see that $P(\xi) = 0$ for $\xi > 1-\mu$ and hence also for $\xi < -1$. Then there are three regions to consider:

Region 1: $-1 < \xi < -\mu$

$$P(\xi) = \int_{-\xi-\mu}^{1-\mu} P(\xi')\mathrm{d}\xi' \tag{7}$$

Region 2: $-\mu < \xi < 0$

$$P(\xi) = \int_{-1}^{0} P(\xi')\mathrm{d}\xi' + \int_0^{1-\mu} P(\xi')\mathrm{d}\xi' = 1 \tag{8}$$

Region 3: $0 < \xi < 1-\mu$

$$P(\xi) = \int_{-1}^{0} P(\xi')\mathrm{d}\xi' + \int_0^{1-\mu-\xi} P(\xi')\mathrm{d}\xi' \tag{9}$$

If we know the solution in region 3, we can integrate to find the solution in region 1. Region 2 has a flat value of 1 (since $P(\xi)$ is a normalized probability distribution). We convert (9) to the differential equation

$$\frac{dP(\xi)}{d\xi} = -P(1-\mu-\xi) \tag{10}$$

which can be reduced to two coupled ODEs by the replacement $Q(\xi) = P(1-\mu-\xi)$, resulting in the solution for region 3:

$$P(\xi) = \cos\xi + \frac{\sin\left(\frac{1-\mu}{2}\right) - \cos\left(\frac{1-\mu}{2}\right)}{\sin\left(\frac{1-\mu}{2}\right) + \cos\left(\frac{1-\mu}{2}\right)} \sin\xi \tag{11}$$

There is an undetermined multiplicative constant fixed by requiring that $P(\xi)$ integrates to 1. From the form of the equation in region 3, we see that this is equivalent to requiring $P(1-\mu) + \int_0^{1-\mu} P(\xi')\mathrm{d}\xi' = 1$. The necessary constant turns out to be unity. One can then directly integrate to find the solution in region 1:

$$P(\xi) = \frac{2\sin\left(\frac{1+\xi}{2}\right)\left[\sin\left(\frac{\xi+\mu}{2}\right) + \cos\left(\frac{\xi+\mu}{2}\right)\right]}{\sin\left(\frac{1-\mu}{2}\right) + \cos\left(\frac{1-\mu}{2}\right)} \tag{12}$$

Eqs. (11) and (12), along with $P(\xi) = 1$ in region 2, comprise the required solution of the integral equation.

We now use the form of $P(\xi)$ to solve for the disorder-averaged ground state density. This can be found, again following [7], by defining $E^0(E^1)$ to be the minimum energy of the *entire* system (not just the left half), subject to the constraint that a particle is absent (present) at some site k deep in the bulk. The density will be given by $1 - P(E^0 < E^1)$. A similar recursive calculation, this time including sites to the left and right, leads to

$$P(E^0 < E^1) = \int_{-1}^{1-\mu} \mathrm{d}\xi_1 P(\xi_1) \int_{-1}^{1-\mu} \mathrm{d}\xi_2 P(\xi_2) \times \tag{13}$$

$$\int_{-\mu}^{1-\mu} \mathrm{d}\epsilon\theta\big(-\epsilon - \min[0,-\xi_1] - \min[0,-\xi_2]\big) \tag{14}$$

The theta function can be split up into four cases

$$\int_{-\mu}^{1-\mu} \mathrm{d}\epsilon\theta\big(-\epsilon - \min[0,-\xi_1] - \min[0,-\xi_2]\big) = \tag{15}$$
$$\big[\theta(\xi_1)\theta(-\xi_2) + \theta(-\xi_1)\theta(\xi_2)\big](\xi_1+\mu) +$$
$$\theta(\xi_1)\theta(\xi_2)\min(1, \xi_1+\xi_2+\mu) + \theta(-\xi_1)\theta(-\xi_2)(\mu)$$

and each term integrated with the form of $P(\xi)$ derived above. After cleaning up all the mess and using $\rho_G = 1 - P(E^0 < E^1)$, we find the remarkably simple result

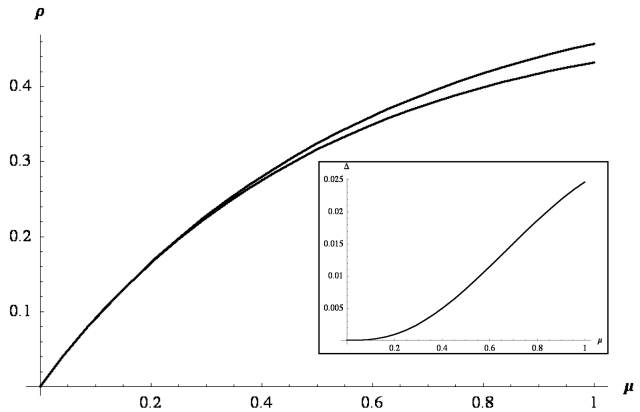

FIG. 2: Plot of the analytic forms of the ground and quenched state densities as derived in the text. Both rise linearly from zero but the quenched density peels off at higher μ. ρ_G is bounded below by ρ_Q and the two functions only intersect at zero. Inset: plot of $\Delta = \rho_G - \rho_Q$ vs. μ. It rises as $\mu^3/6$ for small μ, exactly as predicted earlier.

$$\rho_G = \frac{1}{1+\csc\mu} \tag{16}$$

It is important to remember that this result is valid in the regime $0 \le \mu \le 1$.

We now turn to the calculation of the quenched state density. To do this, we will use the formalism of random sequential adsorption (RSA). The problem of RSA was first studied by Flory [8] and later by many others, see [9], in the continuum as well as for different particle sizes on a lattice. In particular, we will use the dynamic formulation of RSA with a random distribution of binary adsorption rates [10] in the determination of the quenched density. The reason is that, since the lattice potential is random and uncorrelated, the process of sequentially filling the deepest minima is identical to RSA [7]. However, with $\mu < 1$, some sites are repulsive and have an "on rate" of zero. Therefore, we need to consider RSA with two adsorption rates, α and β, take $\beta \to 0$ (while α remains arbitrary) and look for the $t \to \infty$ density.

Let the adsorption rate of site n be denoted by α_n and let the probability that site n is occupied by a particle at time t be $\rho_n(t)$. This probability should be thought of as an average over different realizations of the adsorption process for a fixed choice of the α_n. Using the results of [10], ρ_n varies in time as

$$\frac{d\rho_n(t)}{dt} = \alpha_n \exp(-\alpha_n t) Q^-_{n+1} Q^+_{n-1} \tag{17}$$

where the Q's are time-dependent and obey

$$\frac{dQ^-_n}{dt} = -\alpha_n \exp(-\alpha_n t) Q^-_{n+1} \tag{18}$$

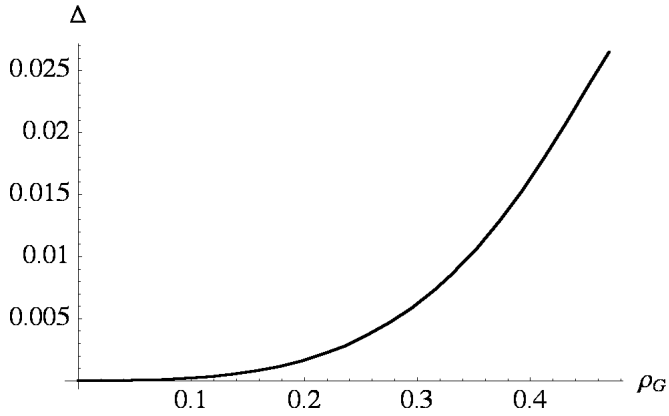

FIG. 3: Parametric plot of $\Delta = \rho_G - \rho_Q$ vs. ρ_G. Surprisingly, Δ remains relatively small until the lattice is nearly half full.

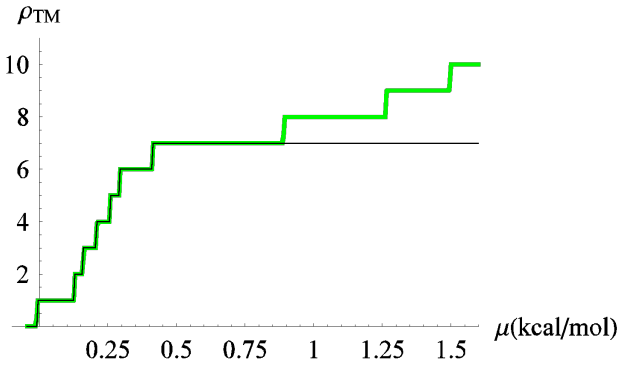

FIG. 4: Ground (green) and quenched (black) states for Bacteriarhodopsin, known to have 7 transmembrane helices. Both configurations agree through the insertion of the 7th helix, above which they diverge.

with a similar equation for Q_n^+ except with the replacement $n+1 \to n-1$. To make use of these equations, we choose the arbitrary α_n's to be random variables equal to α with probability μ and equal to 0 with probability $1-\mu$. Since in our calculation of the ground state density, we have chosen the on-site energies to be uniformly distributed between 0 and 1, the chemical potential μ gives the fraction of attractive sites which therefore have adsorption rate $\alpha \neq 0$. The rest of the sites are repulsive and hence have an adsorption rate of 0. The sequential filling of lowest energy minima will then be mimicked by the dynamic RSA process.

To find the average density of the quenched configuration, ρ_Q, we need to average over the α_n's and take the $t \to \infty$ limit. Since Q_{n+1}^- only depends on sites $m \geq n+1$ (and Q_{n-1}^+ on $m \leq n-1$), the average over α_n, denoted by $\langle . \rangle$, simply factorizes [10]:

$$\frac{d\langle \rho_Q(t) \rangle}{dt} = \alpha \exp(-\alpha t) \langle Q \rangle^2 \tag{19}$$

$$\frac{d\langle Q \rangle}{dt} = -\mu\alpha \exp(-\alpha t) \langle Q \rangle \tag{20}$$

where we have used $\langle \alpha_n \exp(-\alpha_n t) \rangle = \mu\alpha \exp(-\alpha t)$ and the Q's become independent of position after the averaging. Clearly, we then have $\langle Q \rangle = \exp[\mu(e^{-\alpha t} - 1)]$ and upon integration of (19) for $t \to \infty$, we find that the quenched density is given by

$$\rho_Q = \frac{1}{2}(1 - e^{-2\mu}) \tag{21}$$

ρ_Q rises linearly from zero at $\mu = 0$ and saturates at $\mu = 1$ to the "jamming" density of dimers $\simeq .432$.

The two densities (16) and (21) are plotted in Figure 2, and their difference Δ is plotted in the inset. ρ_Q sets a lower limit for ρ_G and both track each other, rising linearly from zero for small values of μ. For larger μ, ρ_Q peels off due to the jamming caused by irreversible adsorption. When we expand $\rho_G - \rho_Q$ for small μ, the first non-zero term is $\mu^3/6$ (see inset to Fig. 2), which is identical to what we predicted earlier in our heuristic argument. We also plot Δ vs. ρ_G in Figure 3 to show that Δ remains quite small until the lattice is nearly half full ($\rho_G = .25$).

The picture that emerges from our exact solution leads to a few general conclusions. First, the quenched and ground states of hard core particles in disordered landscapes *always* differ in an extensive fashion, except at $\rho_G = 0$. Despite this, there are two qualitatively different regimes where Δ can be either small or large. In the small Δ regime, the ground state is kinetically accessible and nearly degenerate metastable states are rare, separated by stretches of unique, well-defined positions. In contrast, the large Δ regime is characterized by ubiquitous metastability and positioning ambiguity.

This picture is very useful for characterizing and contrasting the behaviors of more complicated models that arise in biological systems. The two ends of the spectrum are beautifully illustrated by transmembrane protein α–helix prediction and nucleosome positioning.

briefly review TM helix prediction

The number of transmembrane helices present in the ground and quenched states for bacteriarhodopsin (Br) was computed in [11] and is plotted in Figure 4 as a function of μ, which here controls the zero of the hydrophobicity scale. (There is an important distinction between the chemical potential per amino acid residue and the chemical potential per α–helix because the helices may tilt in the membrane and thus have a somewhat variable length.) Br is known to have 7 transmembrane helices *in vivo* and we see that both densities agree precisely through the 7th step, beyond which the ground state continues to add particles while the quenched state is

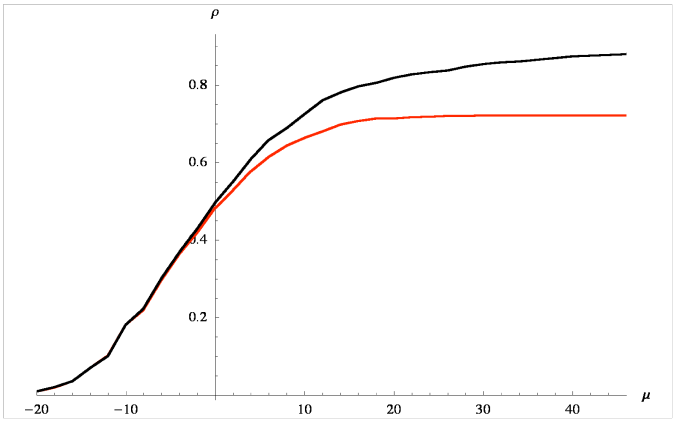

FIG. 5: Ground (black) and quenched (red) states for a 100,000 base pair stretch of DNA from chromosome II of the *S. cerevisiae* genome. The two configurations agree quite well for moderately low densities but the quenched state jams at around 70% coverage. Note that the *in vivo* occupancy of nucleosomes is between 75-90%.

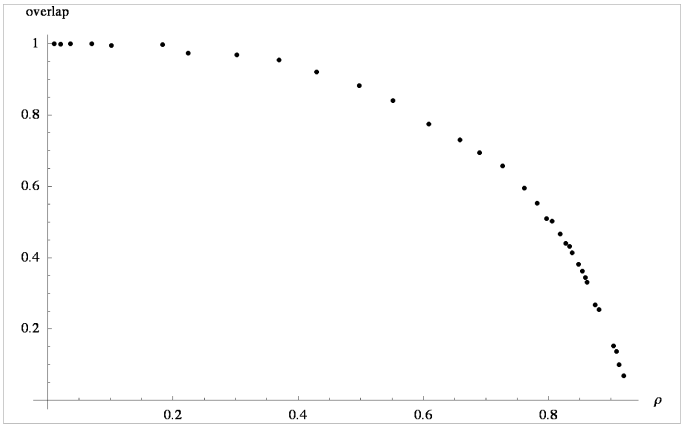

FIG. 6: Overlap of quenched and ground states, defined in the text, as a function of the average ground state density.

jammed. As discussed earlier, a finite system will generically have both configurations identical up to a certain μ_c (in the case of dimers $\mu_c \sim L^{-1/3}$), so it is important to compare to a control sequence. When the sequence of Br is randomized, it is found that the two configurations generically diverge before the insertion of the 7th helix, indicating that their agreement for the true sequence is a result of *design* and not of chance. This turns out to be a general result [11]. The fact that transmembrane proteins design their ground and quenched states to coincide is an important statement about their foldability. Since there are fewer metastable configurations competing for occupancy, the ground state is easier to find kinetically as well as more stable against fluctuations in chemical potential or temperature. See [11] for a much more thorough discussion of transmembrane proteins in this context.

We next turn to nucleosome positioning. **briefly review nucleosome positioning**. A plot of the ground and quenched states for a 100,000 base pair stretch of chromosome II of *S. cerevisiae* is shown in Figure 5. For low to moderate densities, the two curves agree very well, whereas for larger μ the quenched density jams at around 70% coverage while the ground state continues to pack more particles. This is striking because nucleosomes *in vivo* cover between 75 and 90% of the DNA, so in their native environment, nucleosomes must exist in a world of metastability and kinetic traps. How they manage to find their correct positions despite these difficulties is still an open question.

Since the densities in Figure 5 are an average over a 100,000 base pair stretch of DNA, information about the differences in precise positioning of the nucleosomes is lost. Since the exact location of nucleosomes is important for gene regulatory purposes, [4, 5], it is instructive to study the overlap between the ground and quenched density profiles, defined as

$$\frac{\sum_{n=1}^{L} \rho_G(n)\rho_Q(n)}{\sum_{n=1}^{L} \rho_G(n)} \tag{22}$$

This quantity, plotted in Figure 6 as a function of the average ground state density, gives the fraction of particles that are correctly positioned in the quenched configuration. We see that in the biologically relevant regime of 75-90% occupancy, the overlap is a rapidly decreasing function varying from 0.6 down to 0.1. This implies that that density changes in this regime are characterized not by the addition of autonomous single particles, but rather by positional "switching" wherein a local configuration with m particles is overtaken by another with $m+1$.

The ambiguity in positioning nucleosomes was first recognized in [4] They speculated that such positioning ambiguity may play a *functional* role for gene regulation. The idea is that, since nucleosomes occlude regulatory proteins from binding to the DNA on which they sit, (meta)stable configurations could serve as switches. When nucleosomes are organized locally in their ground state, the gene is, e.g., "on" while in a metastable configuration, the gene is "off". We see that the fact that nucleosomes sit on the large Δ side of the spectrum, may actually be useful.

possibly include discussion of differences between dimers and larger particles. e.g. dimers can only switch once through a chemical potential sweep, whereas larger particles can switch many times. this changes the *hysteresis* effects if i study the T=0 dynamics (similar to random field ising model [dahmen, sethna] in external field which is here represented by μ). maybe worth discussing

since hysteresis is a symptom of metastability.

[1] Lewi Tonks. The complete equation of state of one, two and three-dimensional gases of hard elastic spheres. *Phys. Rev.*, 50(10):955–963, Nov 1936.

[2] Jerome K. Percus. Equilibrium state of a classical fluid of hard rods in an external field. *Journal of Statistical Physics*, 15(6):505–511, Dec 1976.

[3] T. K. Vanderlick, L. E. Scriven, and H. T. Davis. Solution of percus's equation for the density of hard rods in an external field. *Phys. Rev. A*, 34(6):5130–5131, Dec 1986.

[4] David J. Schwab, Robijn F. Bruinsma, Joseph Rudnick, and Jonathan Widom. Nucleosome switches. *submitted to PRL*, 2008.

[5] Eran Segal, ..., and Jonathan Widom. A genomic code for nucleosome positioning. *Nature*, 2006.

[6] Blast. ?

[7] Y. Fonk and H. J. Hilhorst. Ground-state and quenched-state properties of a one-dimensional interacting lattice gas in a random potential. *Journal of Statistical Physics*, 49(5):1235–1254, 1987.

[8] Paul J. Flory. Intramolecular reaction between neighboring substituents of vinyl polymers. *Journal of the American Chemical Society*, 61(6):1518–1521, 1939.

[9] J. W. Evans. Random and cooperative sequential adsorption. *Rev. Mod. Phys.*, 65(4):1281–1329, Oct 1993.

[10] Stacchiola D.J., Eggarter T.P., and Zgrablich G. Exact solution of a class of cooperative sequential adsorption problems. *Journal of Physics A: Mathematical and General*, 31:185–194(10), 1998.

[11] unpublished. 2008.