The effects of protein jamming on transcription factor target location

Zeba Wunderlich

May 12, 2005

Abstract

The question of how site-specific DNA-binding proteins like transcription factors find their binding sites on the genome is one that has been explored for over 30 years. Currently, a two-step model which incorporates 3D diffusion of the protein in the cytoplasm and 1D diffusion of the protein on the DNA is strongly supported by theoretical arguments and experimental evidence. However, most theoretical studies of this model neglect the presence of other DNA-binding proteins on the DNA. In this study, we attempt to modify the models of transcription factor search to reflect the presence of other proteins and to explore the effect of protein "jamming" on binding site location. We find that, while the presence of other DNA-binding proteins may slow the search of a transcription factor, they may also increase the effective stability of the protein-DNA complex.

Controlling transcription

In multicellular organisms, transcriptional-level control is essential for the differentiation of genetically identical cells into unique cell types. However, Jacob and Monod first described the concept of transcriptional-level regulation in prokaryotes [9]. Though most prokaryotes are unicellular (see [15] for a review of "multicellular" prokaryotes), they also rely heavily on transcriptional-level control. Most prokaryotes have little control over their environment and therefore use different transcriptional programs to adapt to changing environmental conditions. Since eukaryotes possess membrane-bound organelles, most notably the nucleus, which may add a number of complications to the analysis, this study will focus on transcription in prokaryotes.

Transcriptional-level regulation is the most energy-efficient form of regulation, and regulation of transcription initiation is believed to be the most prevalent type of regulation. Since many RNA polymerases are occupied by transcribing mRNAs encoding proteins needed for translation or are bound non-specifically to DNA, RNA polymerase is in short supply in bacteria [8]. Regulating transcription initiation ensures that the limited number of RNA polymerases are prudently distributed among the genome's promoters. Initiation can be controlled using five different components of transcription: promoter sequences, σ factors, small ligands, transcription factors and chromosome condensation [2]. The modulation of transcription by transcription factors (TFs) raises an interesting biophysical question — transcription factors control the transcription of a few specific genes by recognizing specific (*cognate* or *operator*) DNA sequences, so how do transcription factors find their binding sites on DNA?

Transcription factor binding: Fast and specific

Experimental evidence has shown that TF target site location is both fast and specific. Kinetic studies of Lac repressor, one of the first TFs to be isolated [6], show that it can find its site very quickly, with an association rate $k_a = 7 \cdot 10^9$ M^{-1} sec^{-1} [13]. Thermodynamic studies showed that Lac repressor does not bind appreciably to non-specific DNA even when a 300-fold excess is present [12]. One of the surprising aspects of these findings is that Lac repressor finds its binding site much faster than expected if it were searching for its site via a three-dimensional random walk through the volume enclosing the bacterial genome. The expected association rate in that case, which can be estimated using the Debuye-Smoluchowski equation [12, 19], is $k_a = O(10^8)$ M^{-1} sec^{-1}, an order of magnitude smaller than the observed k_a. Subsequent kinetic studies of the Cro repressor from phage λ gave similar results [10].

This discrepancy did not escape the notice of several scientists, who proposed a number of theories to reconcile the expected and observed association rates. On the basis of several pieces of evidence, including studies that showed TFs have an appreciable affinity for non-specific DNA and a study of the effect of salt concentration on association and dissociation rates of Lac repressor to its operator site, Richter and Eigen [11] and Berg, Winter and von Hippel [1] proposed a two-mode search model. In this model, a TF executes a three-dimensional random walk through the medium surrounding the DNA, which is, in the case of prokaryotes, the cytoplasm, and occasionally encounters DNA and non-specifically binds to it. Once bound to DNA, the transcription factor "scans" the DNA in a one-dimensional random walk until it finds its binding site or dissociates from the DNA.

It has been recently noted that the association rates of

most TFs are significantly lower than that of Lac repressor; in fact, most are near the diffusion limit [7]. Nevertheless, the two-mode model is still relevant, since most TFs will initially encounter non-specific sites of the DNA molecule, and, considering that the specific sites are typically a very small portion of the entire DNA molecule, the association rates are expected to be much lower than the diffusion limit.

In the last year, three groups [4, 7, 17] have considered this two-mode model more carefully. The results most relevant to this study are:

1. The total search time, t_s, it takes for a TF to find its binding site can be written $t_s = \sum_{i=1}^{N}(\tau_{1d,i}+\tau_{3d,i}) = M/\bar{n}[\tau_{1d}(\bar{n}) + \bar{\tau}_{3d}]$, where N is the number of search rounds, τ_{1d} and τ_{3d} are the times for 1D and 3D search rounds, M is the DNA strand length and $\bar{n}$ is the average number of bases scanned in a 1D search round. [17]

2. t_s is minimized when $\tau_{1d}(\bar{n}) = \bar{\tau}_{3d}$. [4, 17]

3. The optimal "scan" length, the number of basepairs explored in one round of 1D diffusion is $O(100)$ bps. [7, 17]

In all of these studies, the authors consider the case of "naked" DNA — DNA free of other proteins, a major idealization, since it is well known that a DNA molecule *in vivo* is covered with DNA-binding proteins. This study will attempt to expand on the work of Slutsky and Mirny by considering the effect of protein "jamming," which occurs when a TF encounters other bound proteins on the DNA during its search, on t_s.

The relevance of "jamming"

In eukaryotes, the method of packing long DNA strands into cells with relatively small diameters is well studied. The fundamental step of this process is wrapping DNA into nucleosomes with histone cores. Much less is known about prokaryotic DNA packing, but studies suggest there are a number of proteins (*e.g.* Fis, H-NS, HU and IHF) that resemble histones in their DNA-binding ability, low molecular weight, copy number and electrostatic charge [5]. These proteins control the packing state of bacterial chromosomes, though they do not contribute to the supercoiled loop structure of chromosome [3], and have been implicated in the regulation of gene expression [5]. Considering that just one of these proteins, H-NS, is estimated to be present in $> 20,000$ copies per cell [14], the assumption that DNA is "naked" is questionable.

The effect of "jamming" on search time

In this section, we follow the framework set forth previously [17]. As before, let M be the DNA length in base pairs and $\bar{n}$ be the average number of sites "scanned" in one round of 1D diffusion in the absence of bound proteins. We assume that there are P proteins bound to the DNA so that the average number of basepairs between the proteins $\bar{l} \approx M/P$, assuming each protein binds a negligible length of DNA.

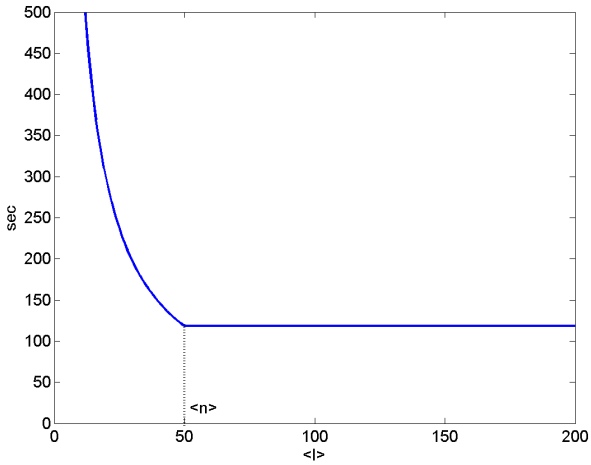

Figure 1: t_s as a function of $\bar{l}$, where $\bar{n} = 50$. When $\bar{l} < \bar{n}$, t_s simply goes as $1/\bar{l}$. When $\bar{l} > \bar{n}$, the proteins are far enough apart that they have a negligible effect on the 1D scans, and t_s is independent of $\bar{l}$.

In this model, a TF executes a 3D random walk in the cell volume, covering on average $l_{3d} \sim 0.1\mu m$ in $\bar{\tau}_{3d} \sim l_{3d}^2/D_{3d}$ seconds, where D_{3d} is the 3D diffusion coefficient of the TF in the cytoplasm. When the TF binds the DNA it executes a 1D random walk. We treat the bound proteins as reflecting barriers. In the "naked" DNA case, τ_{1d}, the time of a round of 1D diffusion, is $\bar{n}^2\pi/(16D_{1d})$, where D_{1d} is the 1D diffusion coefficient of the TF. As shown in Appendix B of [17], $\tau_{1d} \propto e^{\beta E_{ns}}$, where E_{ns} is the non-specific energy of TF-DNA binding. (This implies $\bar{n} \propto e^{\beta E_{ns}/2}$; a fact that is important to keep in mind.) Therefore, τ_{1d} is principally dependent on the non-specific binding energy and should be unchanged by the addition of bound proteins to the DNA. The only effect that bound proteins will have is limiting the number of bases that may be scanned in a round of 1D diffusion. If $\bar{l} < \bar{n}$, the TF can only explore $\bar{l}$ bases instead of $\bar{n}$ bases, which will increase the number of search rounds.

Therefore, in the case of coated DNA, we obtain the following expression for t_s, the search time:

$$t_s = \frac{M}{\min\{\bar{n}, \bar{l}\}}[\tau_{1d}(\bar{n}) + \bar{\tau}_{3d}] \tag{1}$$

The rate of TF association can be expressed as:

$$k_{on} = \frac{1}{t_s[\mathrm{TF}]} \tag{2}$$

Using the following parameter values: $M = 10^6$ bps, $D_{3d} = 10^7 \mathrm{nm}^2/\mathrm{sec}$, $D_{1d} = 10^5 \mathrm{nm}^2/\mathrm{sec}$, $[\mathrm{TF}] = 10^{-9}$ M,

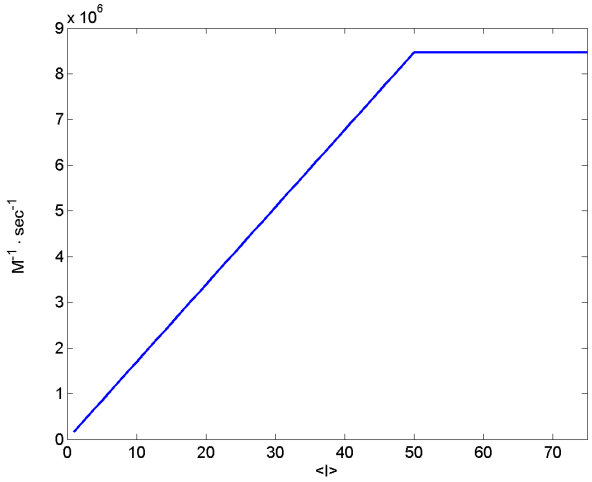

Figure 2: k_{on} as a function of $\bar{l}$, where $\bar{n} = 50$. k_{on} increases linearly with $\bar{l}$ until $\bar{l} = \bar{n}$; after this point, k_{on} is independent of $\bar{l}$.

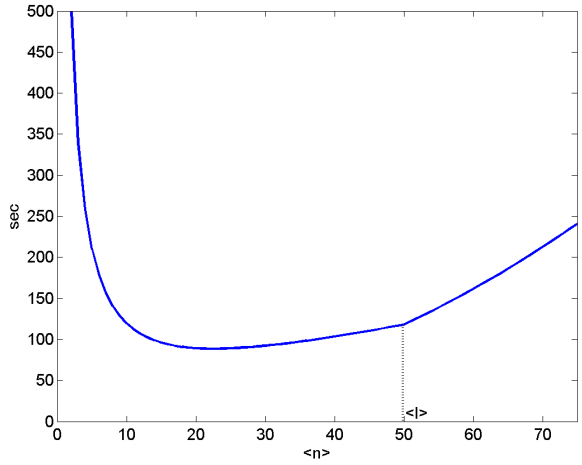

Figure 3: t_s as a function of $\bar{n}$, where $\bar{l} = 50$.
When $\bar{n} < \bar{l}$, the function behaves as it does in the absence of bound proteins — it is large for both small and large $\bar{n}$. When $\bar{n} > \bar{l}$, there is quadratic increase in t_s as a function of $\bar{n}$.

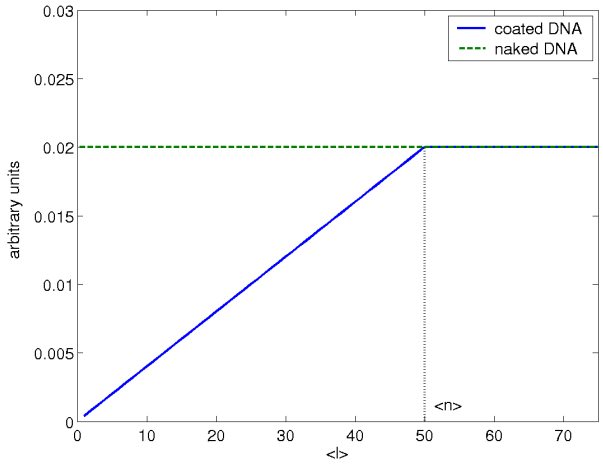

Figure 4: k_{off} as a function of $\bar{l}$, where $\bar{n} = 50$.
For a given $\bar{n}$, when $\bar{l} < \bar{n}$, k_{off} is smaller when proteins are bound than in the absence of protein. The proteins force the TF to repeatedly visit the cognate site.

we obtain Figures 1–3. The results are quite intuitive. When $\bar{n} < \bar{l}$, t_s behaves as it does on "naked" DNA. For small $\bar{n}$ it goes as $\sim 1/\bar{n}$, and for large $\bar{n}$, it goes as $\sim \bar{n}$. It is only when $\bar{l} < \bar{n}$ that we see the effect of the bound proteins, which effectively cap the search speed and k_{on} for a given $\bar{n}$, by making $M/\bar{l}$ rounds of searching necessary, regardless of the length of the 1D search. In this case, t_s is minimized by setting $\bar{n}$ to $\bar{l}$, which prevents unnecessary repetition in the 1D search.

The above observations leads us directly into the discussion of k_{off}. In the "naked" DNA case,

$$k_{off} \propto \frac{1}{\bar{n}e^{-\beta(U_{site})}}$$

where U_{site} is the energy of the TF-cognate site binding event. This expression reflects the two steps necessary for a TF to dissociate from the DNA: first, it must move from the specific site to a non-specific site with a rate proportional to $1/e^{-\beta(U_{site})}$; then, it must dissociate from the non-specific site with a rate proportional to $\bar{n}/e^{\beta(E_{ns})} \propto 1/\bar{n}$.

In the case of DNA with bound proteins, the expression must be modified slightly.

$$k_{off} \propto \frac{\min(\bar{n}, \bar{l})}{\bar{n}^2 e^{-\beta(U_{site})}} \tag{3}$$

So, as before, when $\bar{l} > \bar{n}$, the expression is independent of $\bar{l}$, and when $\bar{n} > \bar{l}$, k_{off} will be smaller than in would in the case of "naked" DNA. In a 1D scan between proteins that surround the target site, the TF will visit the protein $\bar{n}^2/\bar{l}$ times, instead of $\bar{n}$ times, which will decrease k_{off}. (See Figure 4).

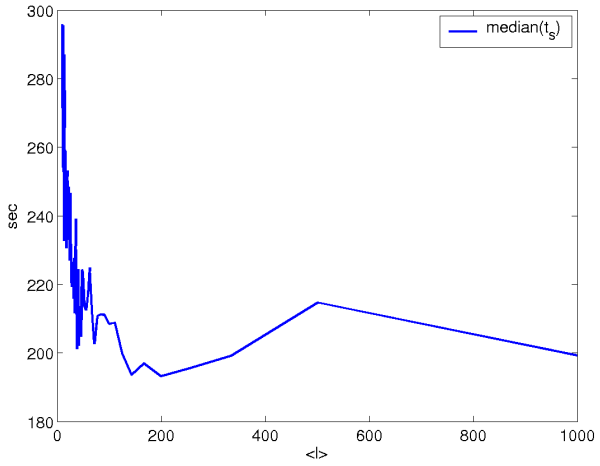

Figure 5: t_s as a function of $\bar{l}$, where $\bar{n} = 100$.
The numerical results agree with the analytical prediction: when $\bar{l} < \bar{n}$, $t_s \propto 1/\bar{l}$. When $\bar{l} > \bar{n}$, $t_s \propto c$.

Search simulations on coated DNA

The above estimates assume that using average values for n and l is an acceptable approximation, since the number of rounds of 3D and 1D searching will be large. This approximation is valid in most cases, as asserted previously [17], but may not be appropriate when $n \approx l$. Presumably, the curves of t_s and k_{on} versus $\bar{l}$ should be smooth in their transition from the case where $\bar{l} < \bar{n}$ to the case where $\bar{l} > \bar{n}$. To obtain these smooth curves, it may be more appropriate to attack this problem using simulation.

Using the same parameters as before, a strand of DNA with one binding site and P randomly distributed proteins is created. Both the cognate site and the bound protein sites are assumed to be 10 bps. Each round of 3D diffusion is assumed to take $\bar{\tau}_{3d}$ seconds, and at the end of the 3D search, the TF associates with the DNA at a random point. Then a 1D random walk is executed with $\bar{n} = 100$ steps, where the bound proteins are treated as reflecting barriers. The simulation was used to determine the dependence of t_s on $\bar{l}$.

The results of the numerical simulation agree closely with the theoretical predictions. Unfortunately, due to the inherent variability of the system, it is diffucult to determine the exact nature of transition from the $\bar{l}$-dependent to the $\bar{l}$-independent regime. In the simulation, the variation in t_s comes from two sources, the variance in the distance between proteins l and the variance in the sliding lengths n. Both processes can be roughly described as negative binomial processes, with the parameters being either the probability of finding a bound protein on the DNA ($p \sim P/M$) or the probability of dissociating from the DNA ($p \sim 1/\bar{n}^2$). However, these processes do not reflect the biological reality well. Variability in l may be quite small *in vivo* if it is the case that prokaryotes, like eukaryotes, use some sort of regular packing for their DNA, which would suggest that bound proteins would be found at fairly regular intervals. Variability in n is determined by the energy landscape of TF-DNA non-specific binding (the σ parameter described in [17]), which is as yet, poorly characterized.

In this simulation, the effects of the 3D conformation of the DNA on the search are excluded. This exclusion affects the spot at which the TF associates with the DNA after a round of a 3D diffusion. In the simulation, the reassociation site is picked randomly, independent of the dissociation site, and time this takes is exactly $\bar{\tau}_{3d}$. In a real search, these are surely correlated [19], and the time of each round may vary considerably if the density of the DNA is heterogeneous. Without knowing the exact conformation of the DNA, however, it is impossible to predict the effect of correlations in the rounds of 3D diffusion.

Conclusions

This study has uncovered two effects of protein "jamming" on transcription factor binding site location:

1. If the density of bound proteins is large enough, bound proteins may interfere with TF sliding and therefore increase TF search times.

2. If the density of bound proteins is large enough, bound proteins may decrease k_{off} and increase the stability of the protein-DNA complex.

The first effect is fairly intuitive and has been anticipated by others previously [18], though it does not seem to have been quantitated . The second effect is slightly subtle, but is quickly understood when one pictures a TF, destined to remain bound to the DNA for an amount of time determined by E_{ns}, bouncing between two proteins. When these proteins are close enough together, it will cause the TF to visit any site between the proteins more frequently than it would in the absence of the proteins.

This study corresponds well to *in vitro* studies of the effect of DNA length on association and dissociation rates using Cro repressor [10]. Kim *et al.* performed a number of studies where the length of the DNA strand containing the TF cognate site is varied and the effect on k_{on} and k_{off} measured. Presumably, protein jamming may have a similar effect to shortening the length of the DNA strand — both DNA ends and bound proteins act as reflecting barriers. They found that both k_{on} and k_{off} increase with increasing DNA length and eventually plateau when the length $\approx$ 500 bps, which is qualitatively similar to the results displayed in figures 2 and 4.

One could imagine a number of refinements may be made to this model. First, the parameter values used in the calculations, particularly the diffusion coefficients, are approximate and seem to result in search times that are too slow. Attempts may be made to find more accurate parameter values. Second, we assume that bound proteins always act as reflecting boundaries, but there

is no experimental evidence supporting this assumption. One could imagine a number of single molecule studies, similar to those used to confirm the presence of 1D sliding [16], to test this assumption. DNA sequences with two binding sites, *e.g.* two Lac sites, at the ends of the strand could be engineered. The DNA could be mixed with an excess of Lac to produce a strand whose ends are fairly stable protein-DNA complexes. The DNA strand could be fixed to a solid surface, and a flow of another TF could be applied perpendicular to the DNA strand. Single molecule tracking could be used to visualize the path of the TF to determine if the fixed Lac molecules act as reflecting boundaries or absorbing boundaries (say, if the Lac molecules promote dissociation of the TF from the DNA). There is also a possibility that the TFs are interacting "moving boundaries," like other TFs in the process of "scanning" the DNA or RNA polymerases that are actively transcribing. The result of these collisions is unknown.

The two results of the study also pose an interesting optimization problem. When a cell wants to up-regulate or down-regulate a gene via transcriptional-level control, there may be two quantities it wishes to optimize, the *speed* of the signal or the *stability* of the protein-DNA complex, which helps determine the longevity of the signal. Given that some parameters in this problem, namely M, the genome length; $\bar{\tau}_{3d}$, which depends largely on the DNA conformation; and the distance between bound proteins $\bar{l}$, which is probably fairly constant during the non-mitotic phases of the cell cycle, are fixed, most of the control comes from varying $\bar{n}$. Setting $\bar{n} > \bar{l}$ will slow the search but increase the signal duration, and setting $\bar{l} > \bar{n}$ will have the opposite effect. Since $\bar{n}$ is determined by E_{ns}, which has been observed to vary considerably from TF to TF, it would be interesting to compare non-specific binding energies to TF function. Perhaps there is evidence that E_{ns} is optimized in some cases to relay a signal rapidly and in other cases to relay a signal for an extended period of time.

It is evident that protein "jamming" may have significant effects on TF-DNA binding. But it is also evident that further experiments that explore the nature and frequency of jamming are needed in order to fully appreciate its significance.

References

[1] Berg OG, RB Winter, and PH von Hippel. 1987. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry.* **20:**6929-6948.

[2] Browning DF and SJW Busby. 2004. The regulation of bacterial transcription initiation. *Nat Rev Microbiol.* **2:**1-9.

[3] Brunetti R, G Prosseda, E Beghetto, B Colonna, and G Micheli. 2001. The looped domain organization of the nucleoid in histone-like defective *Eshcerichia coli* strains. *Biochimie.* **83:**873-882.

[4] Coppey M, O Benichou, R Voituriez, and M Moreau. 2004. Kinetics of target site localization of a protein on DNA: a stochastic approach. *Biophys J.* **87:**1640-49.

[5] Dorman CJ and P Deighan. 2003. Regulation of gene expression by histone-like proteins in bacteria. *Curr Opin Genet Dev.* **13:**179-184.

[6] Gilbert W and B Muller-Hill. 1967. The lac operator is DNA. *Proc Nat Acad Sci.* **58:**2415-21.

[7] Halford SE and JF Marko. 2004. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.* **32:**3040-52.

[8] Ishihama A. 2000. Functional modulation of *Eshcerichia coli* RNA polymerase. *Annu Rev Microbiol.* **54:**499-518.

[9] Jacob F and J Monod. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol.* **3:**318-56.

[10] Kim JG, Y Takeda, BW Matthews, and WF Anderson. 1987. Kinetic studies on Cro repressor-operator DNA interaction. *J Mol Biol.* **196:**149-158.

[11] Richter PH and M Eigen. 1974. Diffusion controlled reaction rates in spheriodal geometry. Application to repressor-operator association and membrane bound enzymes. *Biophys Chem.* **2:**255-263.

[12] Riggs AD, H Suzuki, and S Bourgeois. 1970. *lac* Repressor-operator interaction. I. Equilibrium studies. *J Mol Biol.* **48:**67-83.

[13] Riggs AD, S Bourgeois, and M Cohn. 1970. *lac* Repressor-operator interaction. III. Kinetic studies. *J Mol Biol.* **48:**401-417.

[14] Rimsky S. 2004. Structure of the histone-like protein H-NS and its role in regulation and genome superstructure. *Curr Opin Microbiol.* **7:**109-114.

[15] Shapiro JA. 1998. Thinking about bacterial populations as multicellular organisms. *Annu Rev Microbiol.* **52:**81-104.

[16] Shimamoto N. 1999. One-dimensional diffusion of proteins along DNA. *J Biol Chem.* **274:**15293-15296.

[17] Slutsky M and L Mirny. 2004. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophys J.* **87:**4021-35.

[18] von Hippel PH and OG Berg. 1989. Facilitated target location in biological systems. *J Biol Chem.* **264:**675-678.

[19] Winter RB, OG Berg, and PH von Hippel. 1981. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The*Eshcerichia coli lac* repressor-operator interaction: kinetic measurements and conclusions. *Biochemistry.* **20:**6961-77.