Supplementary materials

Primer sequences used for amplifying and barcoding purified dsDNA libraries:

PE1.0-genetics:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

GATAGTCTCATTTTCACC

Indexed PE2.0 primer:

CAAGCAGAAGACGGCATACGAGAT-index-GTGACTGGAGTTCAGACGTGTGCTCTTCCGA

Primer sequences used for making randomized dsDNA library:

LacO-irand1:

5’-GATAGTCTCATTTTCACC NNNNGTGAGCGGATAACAATT AGATCGGAAGAGCACACG-3’

LacO-irand2:

5’-GATAGTCTCATTTTCACC AATNNNNAGCGGATAACAATT AGATCGGAAGAGCACACG-3’

LacO-irand3:

5’-GATAGTCTCATTTTCACC AATTGTNNNNGGATAACAATT AGATCGGAAGAGCACACG-3’

LacO-irand4:

5’-GATAGTCTCATTTTCACC AATTGTGAGNNNNTAACAATT AGATCGGAAGAGCACACG-3’

LacO-irand5:

5’-GATAGTCTCATTTTCACC AATTGTGAGCGGNNNNCAATT AGATCGGAAGAGCACACG-3’

LacO-irand6:

5’-GATAGTCTCATTTTCACC AATTGTGAGCGGATANNNNTT AGATCGGAAGAGCACACG-3’

LacO-irand7:

5’-GATAGTCTCATTTTCACC AATTGTGAGCGGATAACNNNN AGATCGGAAGAGCACACG-3’

LacO-R2:

5’-GATAGTCTCATTTTCACC AATTGTNANCGNTNACAATT AGATCGGAAGAGCACACG-3’

LacO-R3:

5’-GATAGTCTCATTTTCACC AATTGTNANCNGNTNACAATT AGATCGGAAGAGCACACG-3’

LacO-R4:

5’-GATAGTCTCATTTTCACC AATTGTNANCCGGNTNACAATT AGATCGGAAGAGCACACG-3’

PurR-PR2:

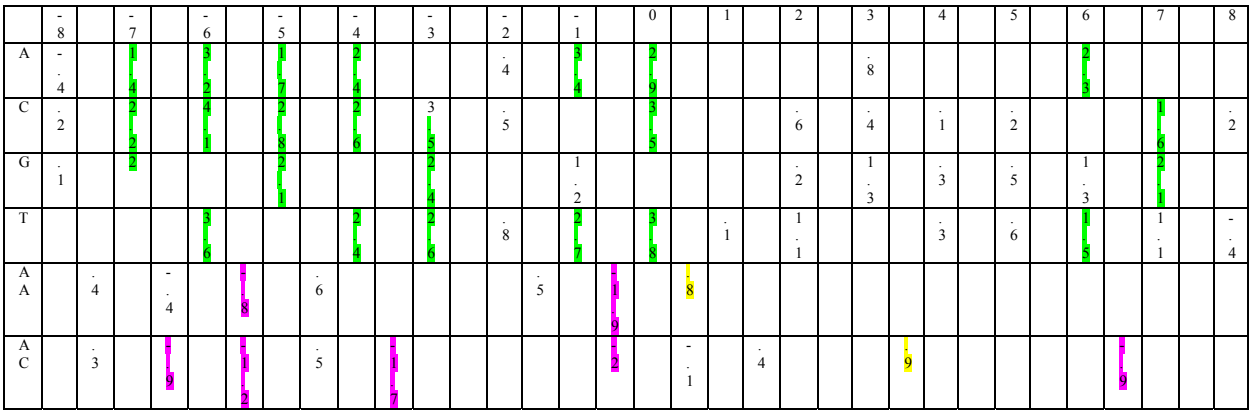
5’-GATAGTCTCATTTTCACC ACGCAAA CG NNNNCGT AGATCGGAAGAGCACACG-3’

PurR-PR3:

5’-GATAGTCTCATTTTCACC ACGCAAA CGG NNNNCGT AGATCGGAAGAGCACACG-3’

PurR-PR3-P4:

5’-GATAGTCTCATTTTCACC ACGCGAA CNG NNNNCGT AGATCGGAAGAGCACACG-3’



AG		.7				-1				-.6				1.5						-.3				.7						-1.9				
AT				-1.7						-1.1				1.1			-2.8			-.4		1.4				-.1						.1			
CA		.2		-1.9		-1.2		-.4				.2		-.8						.4				-.3						-.5				
CC		-1.4		-1.8		-1.9		-.5		-1.8		.4								-1.2		-1.8		.5		.7		.2						.1
CG						-1.4				.-5				.2								-.4		-.1		.1		-.5		.9				.1
CT				-1.3				-.5		-.7		.3		.6						-1.7		-.5				-.2		-.1		.3				.2
GA		.7		-1.1				-.4				1.2				.4								.1						.9				
GC		-.1		-1.6				-1.4				.7												.5		.6		.5				.5		.1
GG		.6																						.2		.6		-2.2		-1.9		.2		-.1
GT				-1.5				-.5				.7				-.5										-.3		-.8		-1.4		-1.1		.3
TA						-1.3						1.2		.4		-1.3		.5						-.6						-.4				
TC						-1.7				-1.2		.9				-1.5		.3		-.3		-.1						-1.2				.5		.5
TG						-1.3				-.5				2						-.5		-1.1						-2.		-.2				.5
TT										-1.1		.9		.6		-2		.3		-1.3								-2.		-1.1		.9		.1

Table S1

Energy deviation values in Figure 2B were classified based on di-nucleotide compositions and operator positions. Pink labels indicate compensatory additivity lower than -0.6kT, whereas yellow labels indicate anti-compensatory additivity beyond 0.6kT.

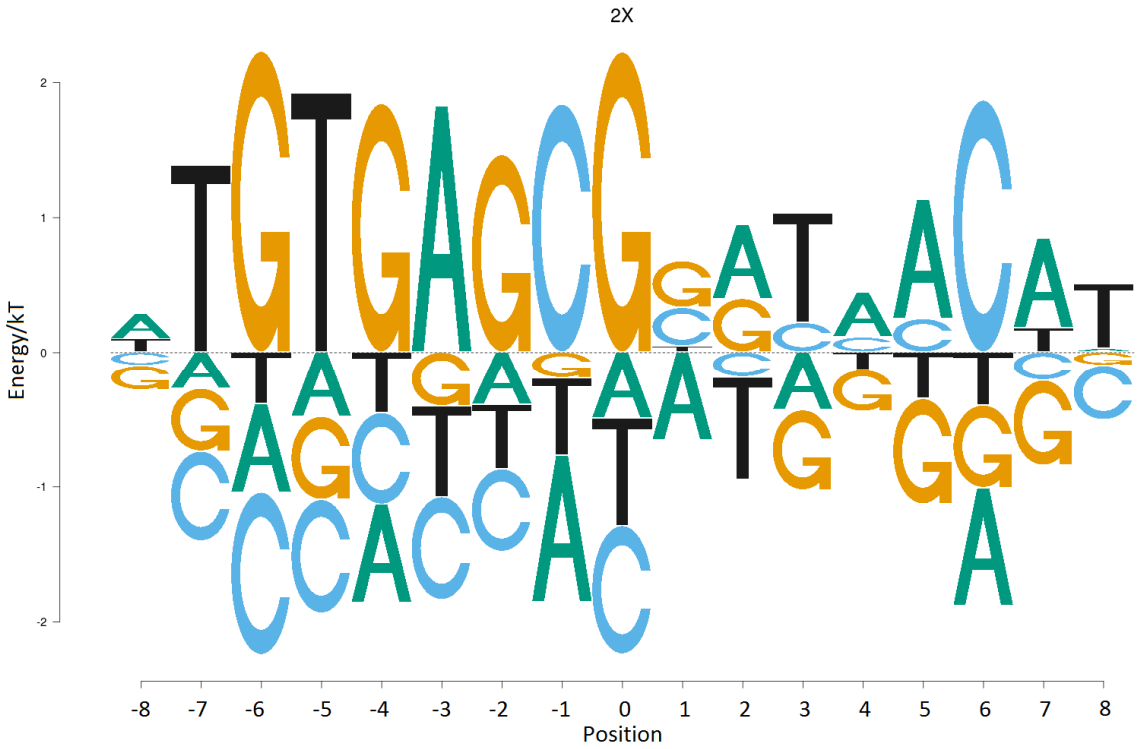


Figure S1

Lac repressor specificity in 2X NEB buffer4, as compared to Figure 2A.

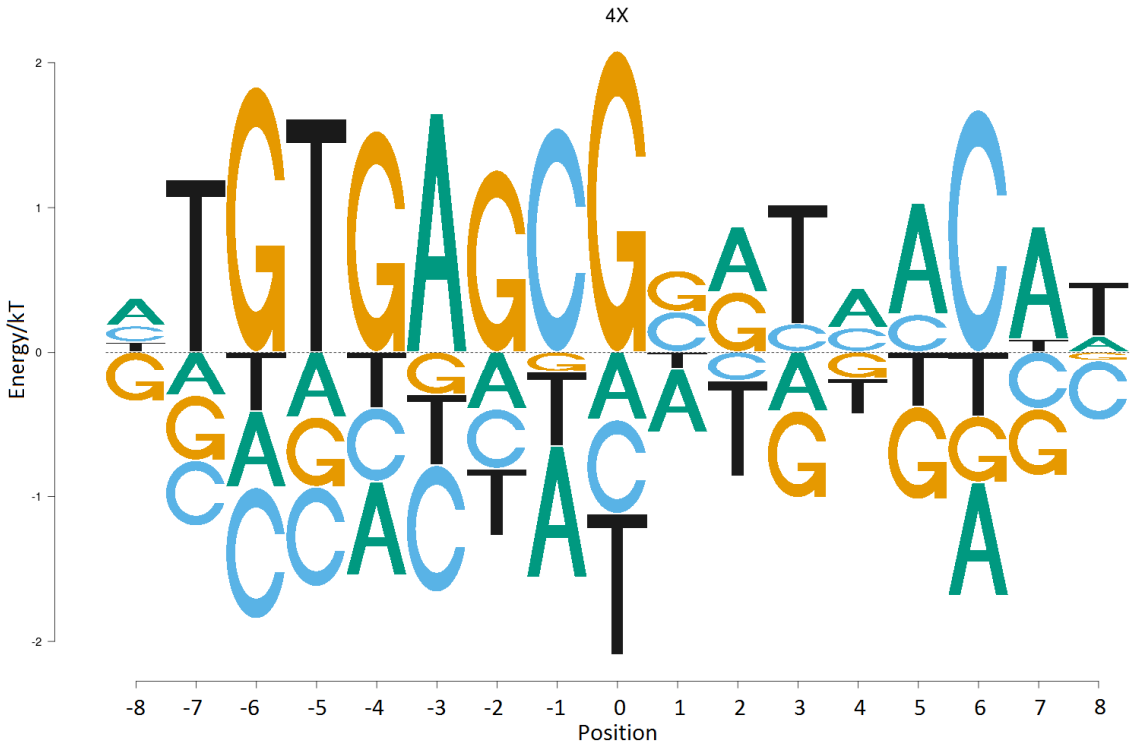


Figure S2

Lac repressor specificity in 4X NEB buffer4, as compared to Figure 2A.

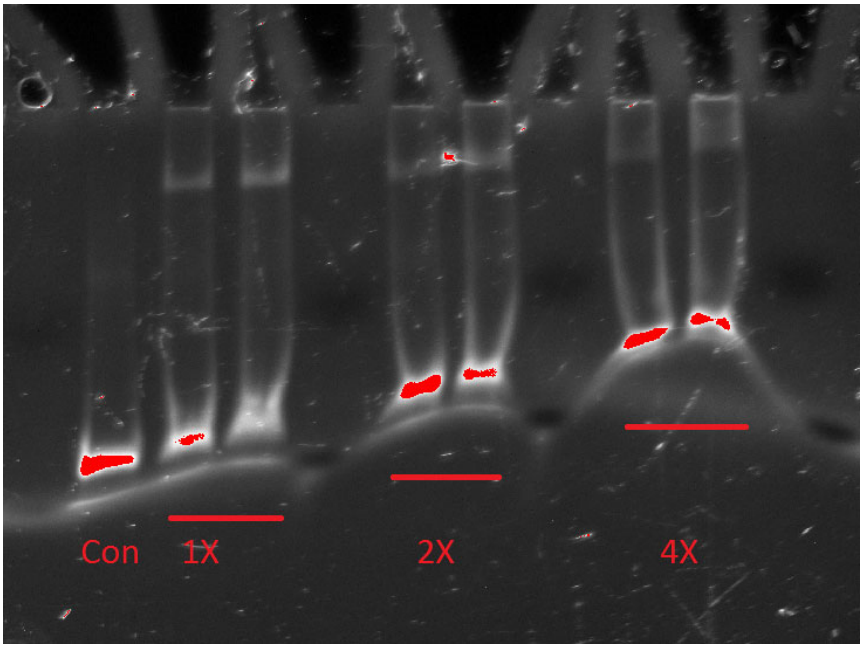


Figure S3

EMSA gel picture for the specificity profiling of the whole lac operator under 1X, 2X, and 4X NEB buffer 4 conditions.

Con: control without protein;

Two protein concentrations (100ng, 200ng/15uL) were used for each ionic conditions; the lane with higher protein concentration was used for Spec-seq sequencing.

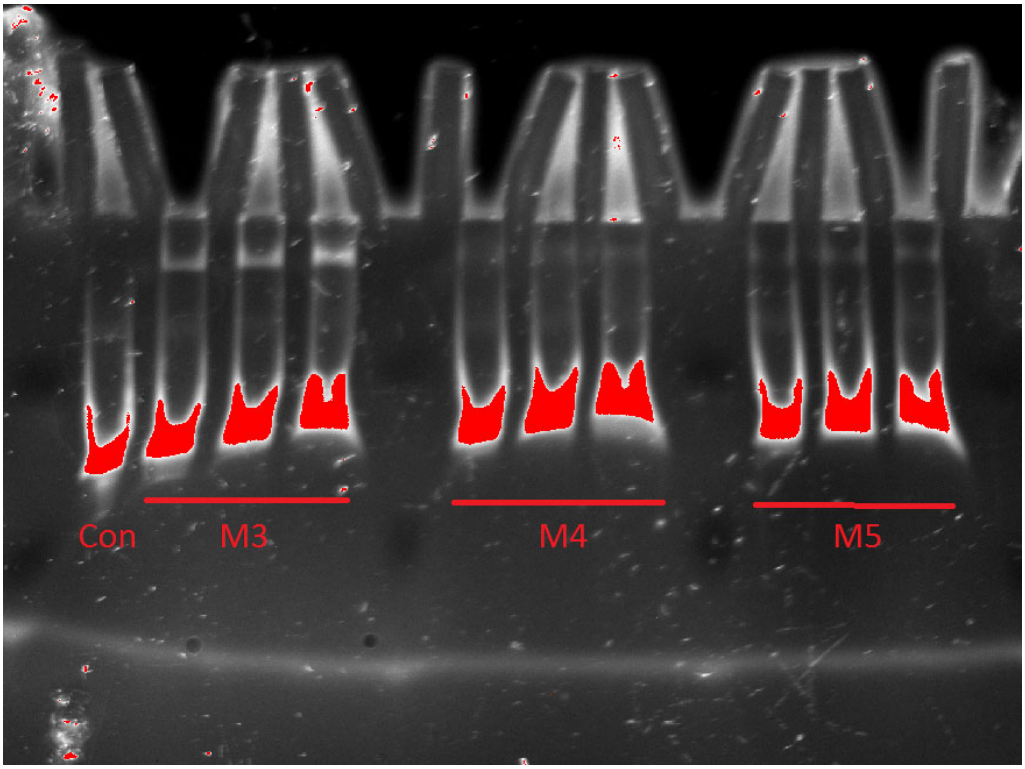


Figure S4 EMSA gel for lacI mutants m4, m4, and m5

Protein concentration increases 2 fold per lane from 50ng/15uL to 200ng/15uL.

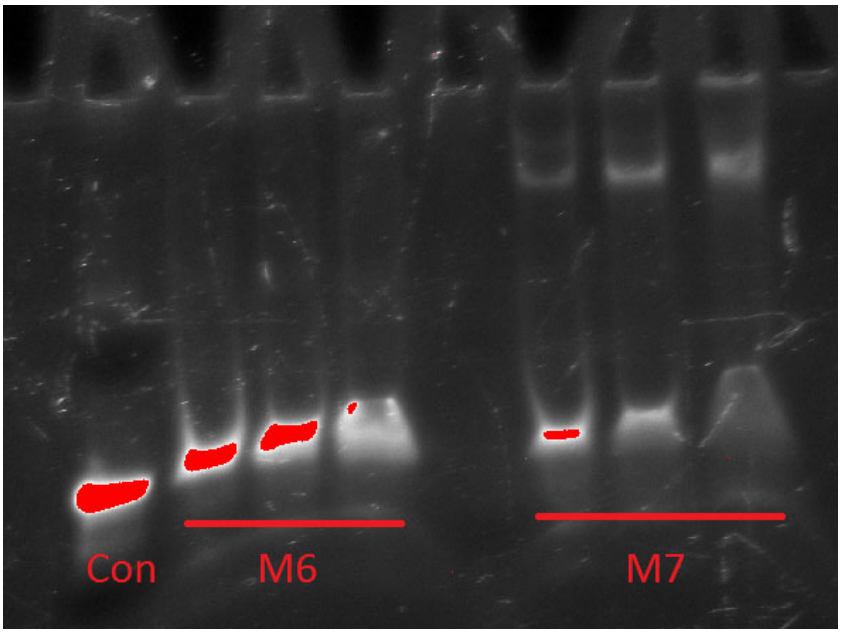


Figure S5 EMSA gel for lacI mutants m6, m7.

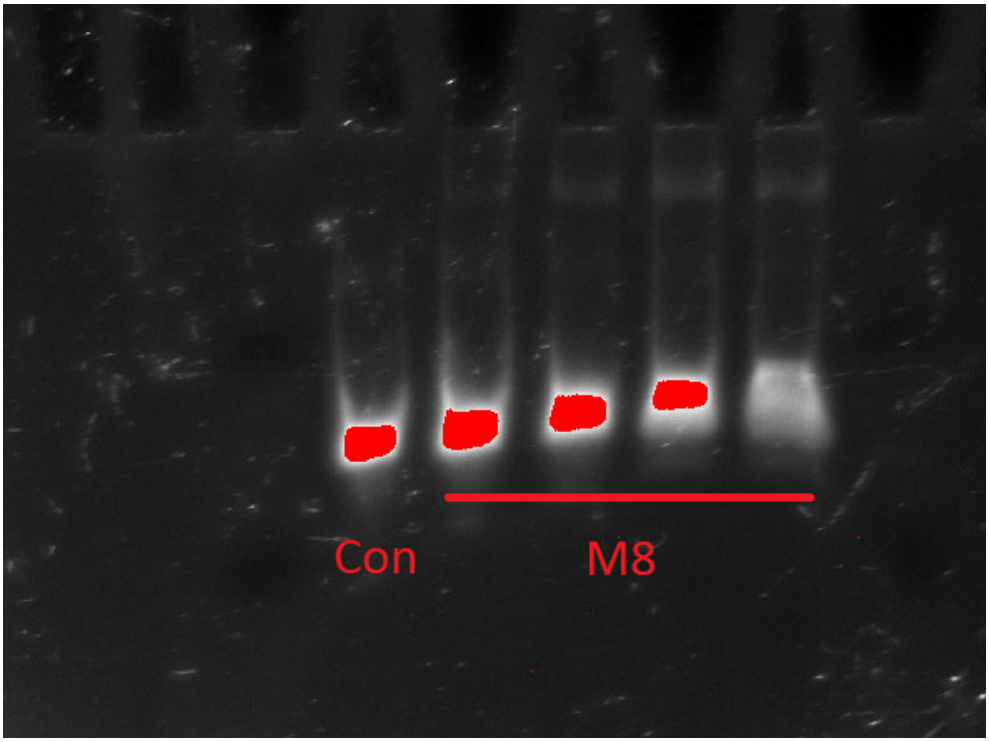


Figure S5 EMSA gel for lacI mutant m8.

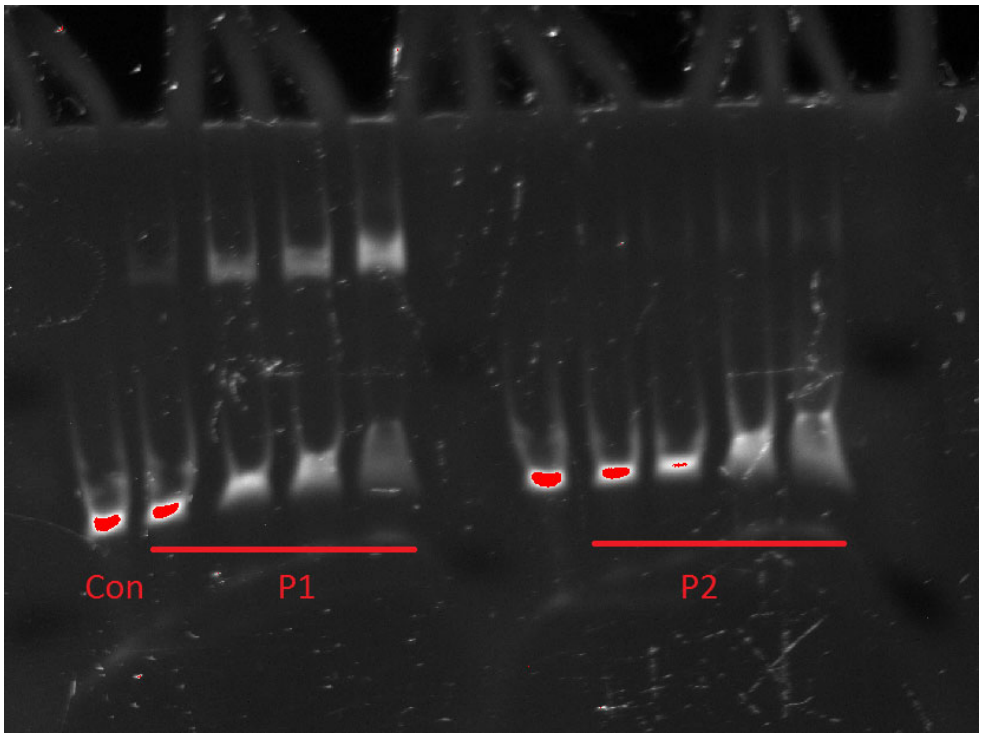


Figure S6 EMSA gel for purR mutants p1 and p2.

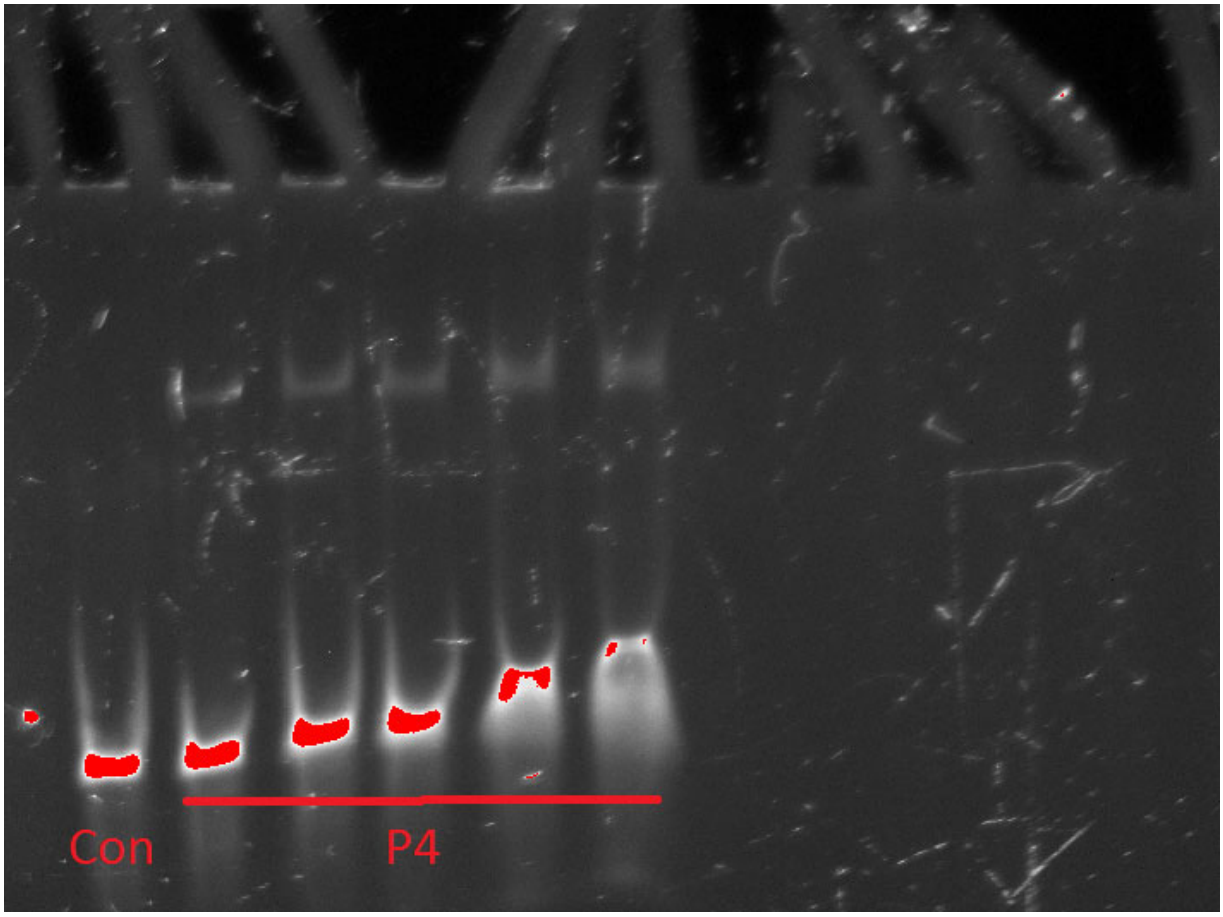


Figure S7 EMSA gel for purR mutant p4.

Positioning onto lac operator: A simplified model

For lac repressor, now we have its systematic specificity data for the first time so it becomes possible to build a positioning model in a theoretical rigorous way.

Positioning without looping

For a primitive positioning model, we assume there is no DNA looping or cooperative interaction between dimeric lac repressors. Thus the binding probability for each individual locus on the genome should follow Fermi-Dirac distribution as in Equation S.1.

$$P_i = \frac{1}{e^{E_i - \mu/kT} + 1} \quad \text{(S.1)}$$

E_i is the binding energy level for the underlying locus (16, 17, or 18bp long), and here we choose the wild-type O1 operator site as the reference site with zero energy level,. μ is the chemical potential determined by protein concentration inside living cells. In reality the genomic DNA concentration is so high (~10mM) that more than 90 percent of time lac repressor is always bound to the genome (Kao-Huang, Revzin et al . 1977). Thus we can have the following constrain:

$$\sum_{i=1}^{Genome\ size} P_i = \textbf{copy number} \quad \text{(S.2)}$$

For lac repressor, it was estimated that there are totally no more than 10 copies per cell(Müller-Hill 1996), which also depends on the cell replication states.

We measured the relative binding energy levels for all the single and adjacent double variants for wild-type O1 operator. Thus it is possible to build the energy weight matrix to predict the relative binding energies for all other 17bp long variants as in Figure 2. Also, we know lac repressor can take two other spacer formats and recognize 16bp or 18bp long binding sites. Since the motif in either case is symmetric around the central CG dinucleotide and resemble the left half and right half for the 17bp case respectively, the energy matrix in Figure 2 is truncated into halves to estimate the 16bp and 18bp binding sites energy levels.

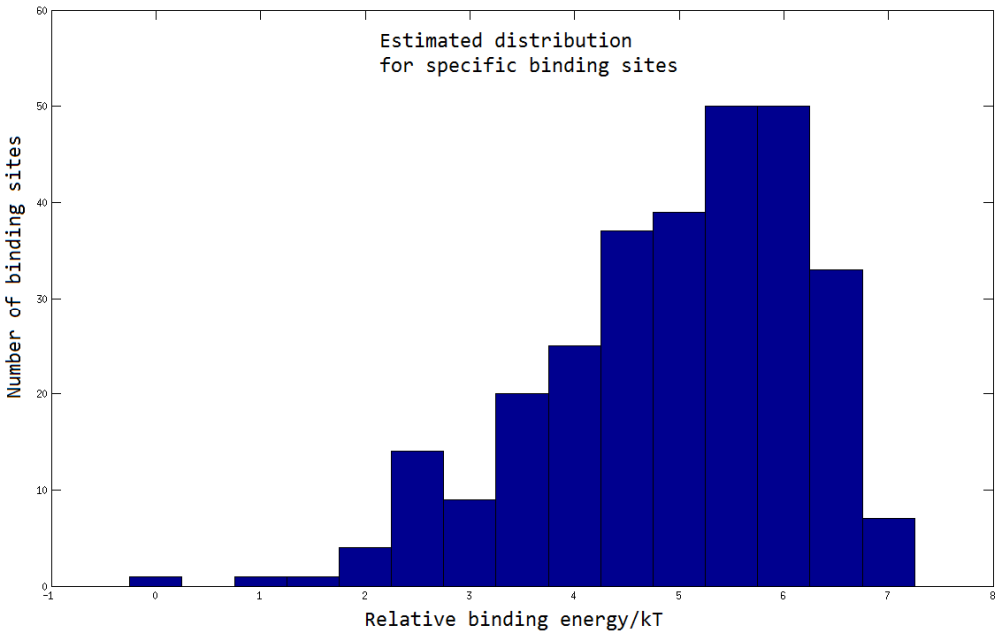


Figure S8

Distribution of all specific binding sites across E. coli genome.

Figure S8 is the estimated binding energy distribution for all the potential specific binding sites across the whole *E. coli* genome. Here we have some additional assumptions for simplification, i.e., all those specific binding sites can neither have more than 4 mismatches to consensus sites nor have binding energy higher than 7kT. Under these two assumptions, it is estimated there are no more than 291 binding sites that can be bound specifically by lac repressor in *E. coli* genome.

Besides lac repressor specificity and its copy number inside cells, there is another important parameter to be considered in our model, i.e., the non-specific energy level $\boldsymbol{E_{NS}}$ compared to reference O1 operator. In order to estimate how high the $\boldsymbol{E_{NS}}$ can reach, one additional negative control sequence was introduced into the library which has no resemblance to any operator sequence in our Spec-seq runs. We observed 1000 fold enrichment of consensus site to this negative control site. Since the whole dsDNA fragments used in our experiments are ~50bp long, the observed binding affinity should be summed up from these 50 individual sites. Therefore in our case we can estimate the $\boldsymbol{E_{NS}}$ level at $\boldsymbol{\log(50 \times 1000) \approx 11kT}$. Alternatively, using *in vivo* fluorescent imaging, Elf, Li et al. (2007) showed that the average residence time for non-specific site is no more than 5ms, whereas for lac O1

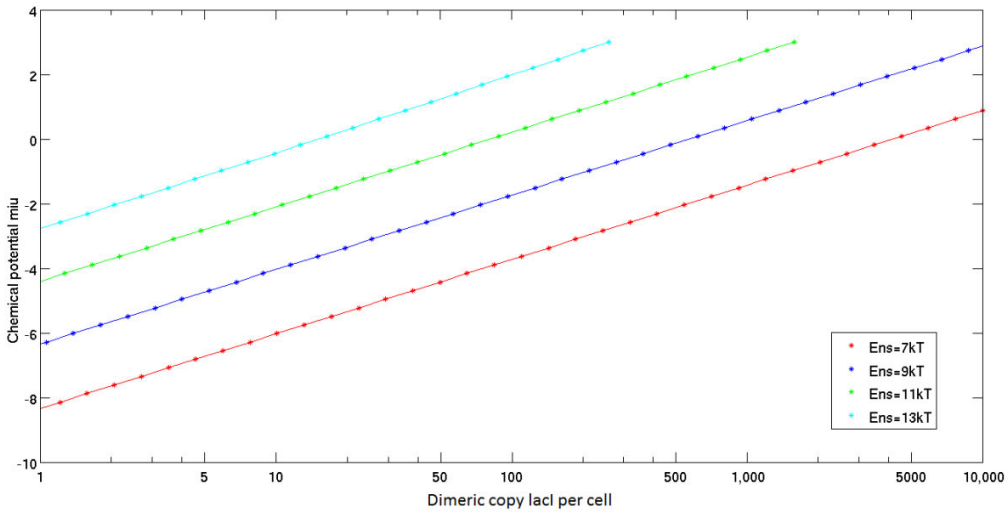
operator site it is ~5mins(Hammar, Wallden et al. 2014). This number is consistent with our estimation, even though $\boldsymbol{E_{NS}}$ might be highly sensitive to salts concentration, PH, and temperatures.

If we know the energy levels for specific and non-specific binding sites both, it is possible to rewrite Equation S.2 as follows:

$$\sum_{i=1}^{Specific\ sites} \frac{1}{e^{E_i-\mu/kT}+1} + \sum_{j}^{Non-specific\ sites} \frac{1}{e^{E_{NS}-\mu/kT}+1} = \boldsymbol{copy\ number} \qquad \text{(S.3)}$$

Given Equation S.3 as constrain for $\boldsymbol{\mu}$, we can then calculate the chemical potential level. Figure S9A plotted the calculated $\boldsymbol{\mu}$ values changing with lacI copy number/protein concentration at different $\boldsymbol{E_{NS}}$ levels. Furthermore, the binding probability of O1 operator can be written down as:

$$\boldsymbol{P_{O1}} = \frac{1}{e^{-\mu/kT}+1} \qquad \text{(S.4)}$$



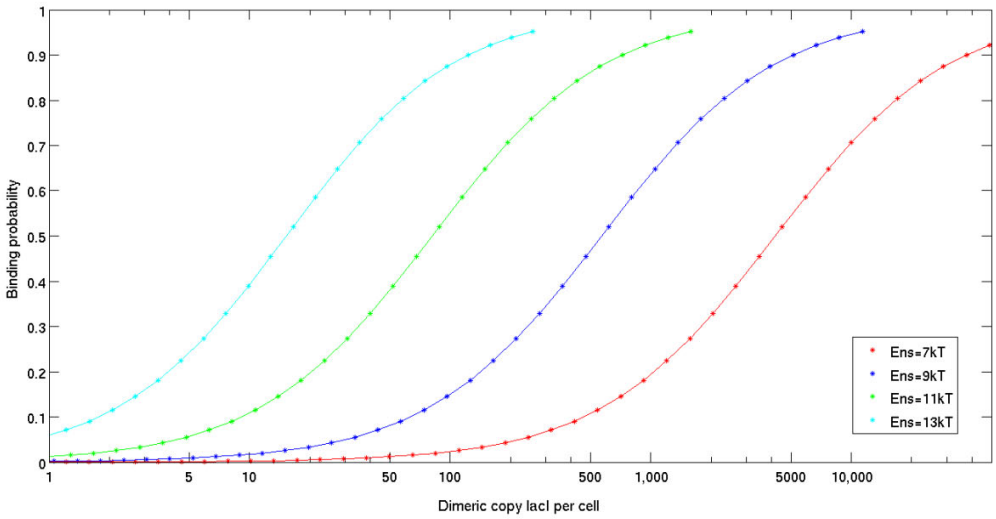


Figure S9

Positioning without looping

Figure S9B plotted the predicted occupancy level of O1 operator at different repressor copy number and $\boldsymbol{E_{NS}}$ levels. Clearly, if we assume lacI at 10 copies/cell and $\boldsymbol{E_{NS}}$ as 11kT, then the O1 occupancy level won't go beyond 20 percent. Some previous work supports this prediction(Elf, Li et al. 2007), even though it also contradicts with other quantification done by lacZ activity assay(Oehler, Eismann et al. 1990). But as one simplified positioning model, the most important message is, without protein-protein interaction between lacI dimers and DNA looping, it is impossible to achieve desired high occupancy level and repression ratio (RR).

A looped case

Ever since the auxiliary lac operators O2 and O3 were discovered, there have been many efforts to decipher the DNA looping mechanism experimentally and theoretically. Oehler, Eismann et al. (1990) proved that all three operators cooperate together to enhance repression ratio. Royer, Chakerian et al. (1990) measured the association energy for tetramerization. Han, Garcia et al. (2009) used *in vitro* single molecule assay to estimate the Looping factor J.

In living E. coli cells, the dimeric lacI concentration is estimated to be 20nM, i.e., 10copies/cell, which is a lot higher than the dissociation constant for tetramerization (4.5nM). So it's easy to imagine that lac repressor stays as tetramer most of time. Also, if the protein is predominantly bound to genomic DNA as discussed in previous section, we can simply assume the looped DNA bound by tetramer lac repressor is the normal state by default. From a modeling perspective, we can divide the free energy for the whole looped complex into four parts:

$$\boldsymbol{G_{looped\ complex} = G_{DNA,i} + G_{DNA,j} + G_{tetramer} + G_{looping\ cost\ i,j}} \qquad \text{(S.5)}$$

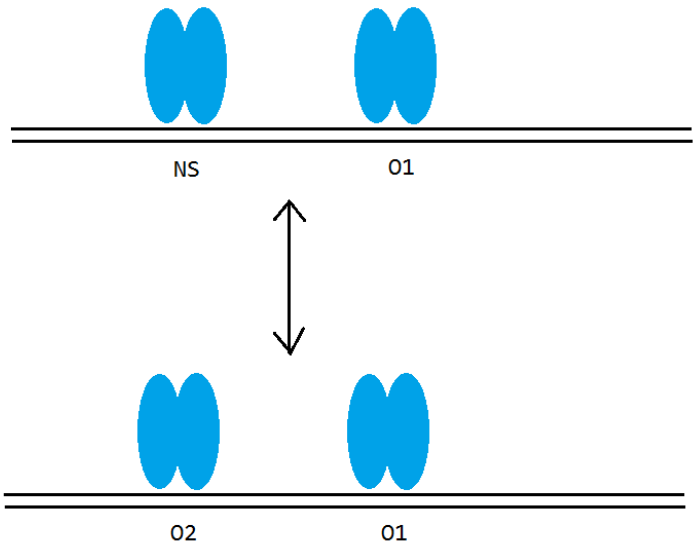
where $\boldsymbol{G_{DNA,i}}$ and $\boldsymbol{G_{DNA,j}}$ are the binding energy of dimeric lac repressor to genomic locus i and j respectively; $\boldsymbol{G_{tetramer}}$ is the association energy between two lac repressor dimers; $\boldsymbol{G_{looping\ cost\ i,j}}$ is the extra energy required to form a DNA loop, including the energy to bend DNA and the entropy cost for a rigid structure, which depends the relative distance between locus i and j(Law, Bellomy et al. 1993).

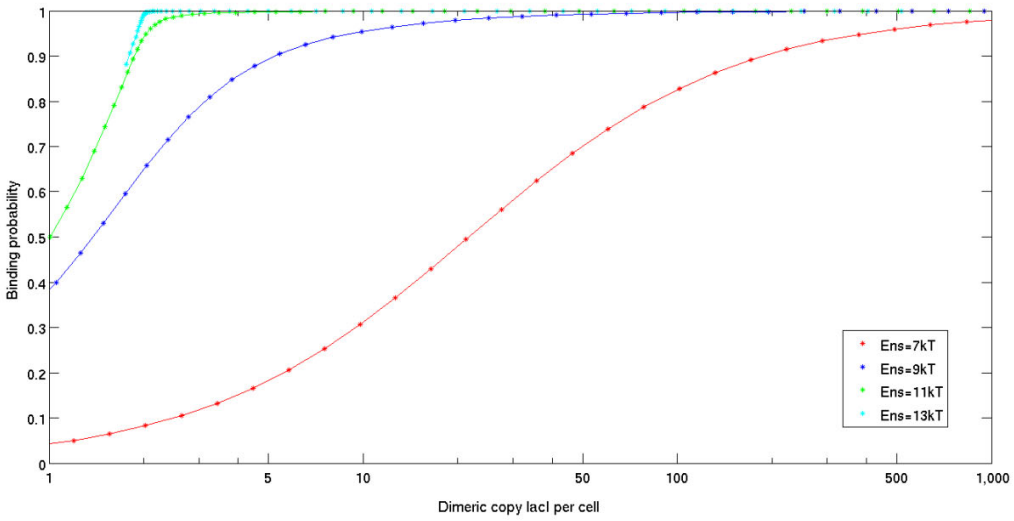
If only single lac operator O1 existed in *E. coli* genome with no auxiliary O2 and O3, we can imagine there would be no difference in the predicted O1 occupancy level between looped and no-looped case. Because even the lac repressor binds to the O1 site and form a loop, the other bound site can only be some non-specific site, i.e., a looped O1-NS structure. This looped O1-NS structure brings no extra energy benefit compared to any other looped NS-NS structure. Thus we have

$$\mathbf{P}(\boldsymbol{O1, NS|tetramer}) = \mathbf{P}(\boldsymbol{O1|dimer}) = \frac{\mathbf{1}}{\boldsymbol{e^{-\mu/kT}+1}} \text{ (S.6)}$$

Now if we put back O2 into the system, we essentially replaced the old O1-NS site with an enhanced tetramer binding sites O1-O2. Since there is no other known good looping site in *E. coli*, we can have

$$\mathbf{P}(\boldsymbol{O1, O2|tetramer}) = \frac{\mathbf{1}}{\boldsymbol{e^{(E_{O2}-E_{NS})-\mu/kT}+1}} \qquad \text{(S.7)}$$


NS
O1
O2
O1


Binding probability
Dimeric copy lacI per cell
Ens=7kT
Ens=9kT
Ens=11kT
Ens=13kT
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
5
10
50
100
500
1,000

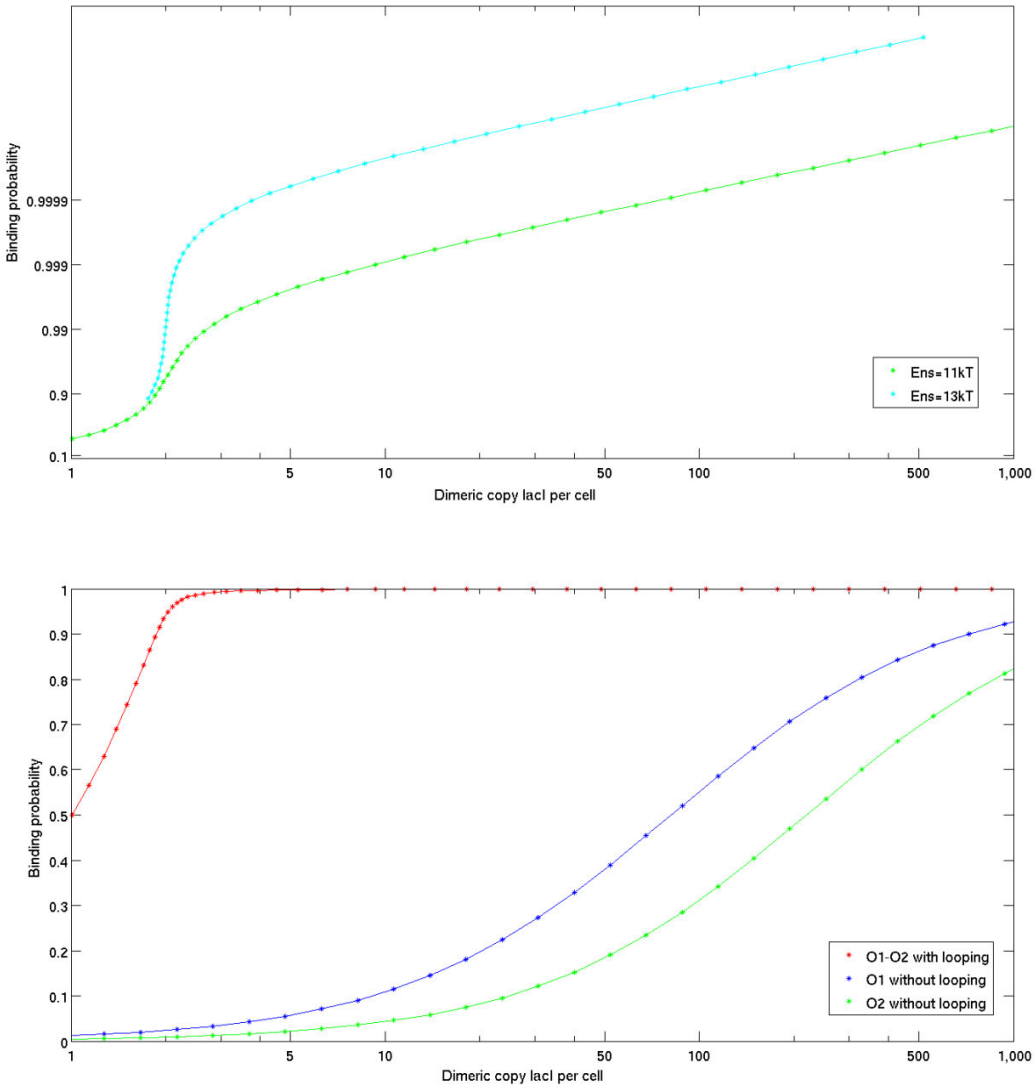


Figure S10 Positioning with looping

Figure S10B and S10C plotted copy number/cell vs. the predicted occupancy level under looped O1-O2 case. The chemical potential μ is recalculated assuming that O1-O2 state has new energy level $\boldsymbol{E_{O2} - E_{NS}}$, while other sites keep their original levels. If we choose $\boldsymbol{E_{NS} = 11kT}$, the binding probability gets significantly increased to 0.999, which corresponds to 1000 fold repression. Figure S10D further compared the predicted occupancy level under different operator configurations, i.e., looped O1-O2, O1 alone, and O2 alone. .

One intuitive way to understand this model is the tetramer lac repressor search the genome like a "monkey bar". The tetramer molecule diffuse or slide along the genome until it finds the first lac operator, either O1 or O2. Thus it switches from NS-NS state to semistable O1-NS or NS-O2 state. Afterwards, it may continue search for the other operator to finally fall into the stable O1-O2 state.

To theoretically treat this "Monkey bar" model more vigorously, we not only need to know the binding specificity for each individual dimer protein, but also have some ideas on how the $\boldsymbol{G_{looping\ cost\ i,j}}$ term changes according to distance between two dimer binding sites. After taking all possible tetramer conformations into account, we would have a more accurate characterization for this binding probability in looped case.

REFERENCES

1. Elf, J., Li, G. W. and Xie, X. S. (2007) Probing transcription factor dynamics at the single-molecule level in a living cell. Science, 316, 1191–1194doi:10.1126/science.1141967. PMID:17525339

2. Hammar, P., Walldén, M., Fange, D., Persson, F., Baltekin, O., Ullman, G., Leroy, P. and Elf, J. (2014) Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. Nat. Genet., 46, 405–408doi:10.1038/ng.2905. PMID:24562187

3. Han, L., Garcia, H. G., Blumberg, S., Towles, K. B., Beausang, J. F., Nelson, P. C. and Phillips, R. (2009) Concentration and length dependence of DNA looping in transcriptional regulation. PLoS ONE, 4, e5621doi:10.1371/journal.pone.0005621. PMID:19479049

4. Kao-Huang, Y., Revzin, A., Butler, A. P., O'Conner, P., Noble, D. W. and von Hippel, P. H. (1977) Nonspecific DNA binding of genome-regulating proteins as a biological control mechanism: measurement of DNA-bound Escherichia coli lac repressor in vivo. Proc. Natl. Acad. Sci. U.S.A., 74, 4228–4232doi:10.1073/pnas.74.10.4228. PMID:412185

5. Law, S. M., Bellomy, G. R., Schlax, P. J. and Record, M. T. Jr. (1993) In vivo thermodynamic analysis of repression with and without looping in lac constructs. Estimates of free and local lac repressor concentrations and of physical properties of a region of supercoiled plasmid DNA in vivo. J. Mol. Biol., 230, 161–173doi:10.1006/jmbi.1993.1133. PMID:8450533

6. Muller-Hill, B. (1996). The lac operon: A short history of a genetic paradigm. 1996, Berlin, Walter de Gruyter,.

7. Oehler, S., Eismann, E. R., Krämer, H. and Müller-Hill, B. (1990) The three operators of the lac operon cooperate in repression. EMBO J., 9, 973–979 PMID:2182324.

8. Royer, C. A., Chakerian, A. E. and Matthews, K. S. (1990) Macromolecular binding equilibria in the lac repressor system: studies using high-pressure fluorescence spectroscopy. Biochemistry, 29, 4959–4966doi:10.1021/bi00472a028. PMID:2194564