Chapter 5

Understanding the factors that dictate the protein-DNA binding specificity
Abstract

It is generally believed, that the transcription factors, through a combination of short-range slides (1D) and hops (3D) with long-range jumps (3D) find their target site on the extremely large DNA molecule. But how the various transcription factors are distributed over the entire bacterial genome when they are not at their target sites, or whether their location on the genome plays a role in finding their target site is not clearly understood. We chose the Gal repressor-operator system of E. coli to study the binding interactions of each of the possible single base operator mutants with the repressor protein in vitro using fluorescence anisotropy measurements. Single basepair operator mutations only at a few positions seem to keep the binding energy largely unchanged from the wild type operator while most positions of the operator are important for the interaction. From the energy values thus obtained, the total binding energy difference between any arbitrary sequence and the specific operator sequence could be estimated provided there is additivity. Using these binding energy results a program was used to find the propensity of this repressor for occupancy of all the 16-basepair sites on the genome. We conclude that very tight non-specific interaction of the protein with the bacterial genome is not favored energetically. Neither are their clusters of preferred energy sites in the genome where the protein can be found to be concentrated.
Introduction

Recognition of specific DNA sequences by proteins plays a central role in the regulation of gene expression. Increasing amounts of data on the structures of protein-DNA complexes provide us with information about these specific interactions. However, analysis of these structures reveals no clear one-to-one correspondence between amino acids and bases. Takeda and Sarai (Sarai, A. et al., 1989) probed how these specific contacts are used to recognize the specific operator sequences, and whether some other factors like phosphate interactions or changes in DNA conformation also play major role in sequence recognition. Systematic base substitutions in λ operator were made and filter binding assays were done to precisely locate sequence specific interactions within the binding site. No other systematic base substitution experiments have been carried out.

In this work we have studied the interaction of gal repressor with the gal operator DNA. To carry out detailed analysis of sequence specific interactions we wanted to understand at which positions in the 16 base pair operator sequence, the specific interaction points are located. For that we created a series of single base operator mutants with systematic base substitutions at each of the 16 positions. For each of these operator mutants the binding energy of the gal repressor protein was measured quantitatively. The energy changes thus obtained suggest how each position and more importantly how the specific base at each position contributes specifically to the binding energy.

Position weight matrix is one of the most commonly used representations of transcription factor-DNA binding specificity. It is a probability matrix with four rows corresponding to the four bases and n columns corresponding to each position in the binding site. The log-likelihood value is obtained by dividing the base probabilities at a position by the
background probability of the base and taking the logarithm of that value. This value is entered in the weight matrix. The entry corresponding to the $i^{th}$ column (position) and $j^{th}$ row (base) is $f_{ij}$. Using statistical mechanics theory it was shown that the logarithms of the observed to the expected base frequencies should be proportional to the binding energy contribution of the base (Heumann, J.M. et al., 1994). We used the binding energy difference of each mutant due to the single base change for creating the position weight matrix. Using this weight matrix we calculated the energy of all possible binding sites in the entire *E. coli* genome.

A transcription factor undertakes a one-dimensional search by hopping to neighboring sites by normal diffusion, where the diffusion coefficient varies exponentially with the roughness of the binding energy landscape. An optimal one-dimensional/threedimensional combination of movement leads to a significant speed-up of the search process. A protein cannot find its target site in biologically relevant time-scale if the roughness of the specific binding landscape is greater than $\sim 2 \text{ K}_B T$. Thus, rapid search requires a fairly smooth landscape, whereas stability of the protein-DNA complex at the target site requires a low energy of the target site, possibly less than $15 \text{ K}_B T$. Mirny and coworkers suggest that there are two modes of protein-DNA binding; the *search* mode and the *recognition* mode. The transition between the modes may happen when the protein is trapped at a low-energy site of the search landscape (Slutsky, M. et al., 2004). Our goal was to find out whether the *E. coli* genome has several such low energy sites which can trap the repressor protein other than the unique operator site.

Takeda and coworkers have shown that the binding energy changes are additive for lambda repressor and Cro and thus can be summed over the total sequence (Takeda, Y.
et al., 1989). From the binding energy obtained for each mutant, using the additive rule, we could create a 4 (for positions A T G C) by 16 (number of bases of the operator) matrix assigning the energy values for each base at each position as obtained from each mutant. Using a FORTRAN program we could thus measure the energy values for binding to all 16 base pair sequences spanning over the entire E. coli genome calculated both in the forward and the reverse complementary directions.

The gal repressor (GalR) is a close homologue of the lac repressor (LacI) and the purine repressor (PurR). The GalR forms a tetramer only as a nucleoprotein complex bound to operator, in contrast to the Lac-repressor which adopts a tetramer conformation in vivo when unbound to DNA (Aki, T. et al., 1997; Geanacopoulos, M. et al., 2001). GalR and LacI, were shown to have homology in their amino acid sequences extending beyond the DNA binding domains (von Wilcken-Bergmann, B. et al., 1982). LacI / GalR monomers contain common structural features that include a DNA binding domain with a helix-turn-helix motif, a linker between the two major domains, and a regulatory domain that encompasses regions for oligomerization and for effector binding. In the operator-bound lac repressor bound to operator structure; the repressor head-piece domain (HTH motif) binds to the major groove of the operator; the hinge-helix binds to the centre of the operator in the minor groove introducing a 40° bend. Looking at the Lac repressor / O1 complex crystal structure it is evident that the ~35-40% of the bases of the operator sequence are in direct contact with the repressor protein. It would be expected that the gal repressor given its homology with the lac repressor would be similar in operator binding. With no crystal structure of the gal repressor available, from our binding experiments
with different operator mutants we tried to assess the regions of operator that confer the specificity of binding to the repressor.
Results

Expression and Purification of GalR

The plasmid pSEM1026 bearing the gene for gal repressor under pBAD promoter was expressed in DH5α cells by induction with arabinose and purified through Ni-NTA agarose affinity column. The details of this purification are discussed in Materials and Methods section. The protein was purified to greater than 95% purity. The purification profile of GalR is shown in Figure 5.1.

![Figure 5.1: Purification profile of GalR.](image)

The LFT fraction denotes the loading flow through, WFT denotes the washing flow through, and 1-12 which denote the elution fractions were run in a 10% SDS PAGE.
Binding studies of GalR with the wild type $O_E$ operator and the mutant $O_R$ operators

Fluorescein-labeled wild type operator ($O_E$) and several single base mutant operators were used to study the binding with GalR by measuring the increase in anisotropy values. The wild type repressor-operator has an apparent dissociation constant of $4 \times 10^{-9}$ M which has been reported from earlier work in our laboratory (Chatterjee, S., et al., 1997). This result was reproduced and $K_d$ values for the several mutant operators were generated on the basis of the fluorometric titration with these operators. Least square fitting of the anisotropy data to a single site binding equation revealed the dissociation constant ($K_d$) values for each of the mutants.

Figure 5.2 shows the gal operator. Figure 5.3 shows the binding isotherm of the wild type operator with gal repressor and Figure 5.4 (A to P) shows the binding isotherms of the 46 mutant operators with the repressor. Two mutant operators showed no binding till 1 $\mu$M protein concentration and thus were considered to have very high dissociation constants. The data for these two are not shown.

A control titration was done to estimate the dissociation constant of the gal repressor with a non-specific DNA. A fluorescein labeled operator DNA duplex, $O_R1$, from the bacteriophage $\lambda$, 21 bases in length, was used. Using same parameters and buffer, as for the rest of the titrations, the dissociation constant of the non-specific DNA with GalR was found to be 10.9 $\mu$M. Figure 5.5 shows the control titration. The apparent dissociation constants obtained from these titrations were used to calculate the binding free energy change for each of the mutants. Table 5.1 lists the apparent dissociation constants and the binding free energy change thus calculated for each of the mutants.
Figure 5.2: The Galactose operator sequence. The numbers show the 16 base positions.

```
5'-GC GTGTAAACGATTCCAC
3'-CG CACATTTGCTAAGGTG
```

Figure 5.3 Binding isotherm of wild type O_E with Gal repressor. 2 nM of fluorescein labeled operator was incubated with increasing concentrations of purified GalR and anisotropy values were measured at each point. The excitation and emission wavelengths used for titration were 490 nm and 525 nm respectively, with a band pass of 5 nm in each channel. The titration was carried out in 12.5 mM Tris-HCl buffer pH 8.0, containing 300 mM KCl and 0.5 mM EDTA. The temperature was kept at 25°C. The plot is an average of 5 titrations.
Figure 5.4A

$O_{1}^{G-T}$

$[\text{GalR}]$ M

Figure 5.4A1

$O_{1}^{G-A}$

$[\text{GalR}]$ M

Figure 5.4A2

$O_{1}^{G-C}$

$[\text{GalR}]$ M
Figure 5.4C1

Figure 5.4C2

Figure 5.4C3

Figure 5.4C4
Figure 5.4D

Anisotropy vs. [GaIR] M for different base pairs:

- $O_{e4} (T-A)$
- $O_{e4} (T-C)$
- $O_{e4} (T-G)$

Figure 5.4D1

Anisotropy vs. [GaIR] M for different base pairs:

- $O_{e4} (T-A)$
- $O_{e4} (T-C)$
- $O_{e4} (T-G)$

Figure 5.4D2
Anisotropy

$O_{e5}$ (A–C)

$O_{e5}$ (A–T)

$O_{e6}$ (A–G)

Figure 5.4E

Figure 5.4E1

Figure 5.4F
Figure 5.4F1

$O_{66} (A\rightarrow C)$

Figure 5.4F2

$O_{66} (A\rightarrow T)$

Figure 5.4G

$O_{67} (A\rightarrow T)$

140
Figure 5.4H1

Figure 5.4I

Figure 5.4I1
Figure 5.412

Figure 5.4J

Figure 5.4J1
Figure 5.4J2

Figure 5.4K

Figure 5.4K1
Figure 5.4K2

Figure 5.4L

Figure 5.4L1
Anisotropy

Figure 5.4M2

Figure 5.4N

Figure 5.4N1
Figure 5.4O

Figure 5.4O1
Figure 5.4O2

Figure 5.4P

Figure 5.4P1
Figure 5.4: The Binding Isotherms of the various O_e Mutants with the Repressor GalR. The numbers 1 to 16 following O_e in the title of each of these plots denote the position of base in the operator where the mutation has been incorporated. The alphabets in parentheses denote the original nucleotide followed by the nucleotide it has been mutated to. Each titration was carried out in the same buffer which was used for the wild type titration. All other conditions were also similar to the wild type titration. All titrations were done in triplicate. The line shown for each plot is the best fit to a single site binding equation. The plots are numbered as 5.3A, 5.3A1, 5.3A2 for the three mutants at position 1; 5.3B, 5.3B1, 5.3B2 for the three mutants at position 2 and so on till the 16th position.
Figure 5.5: Binding isotherm of fluorescein labeled non-specific $\lambda$-OR1 duplex DNA with GalR. The control titration was done in the same buffer and under the same conditions as the rest of the titrations.
<table>
<thead>
<tr>
<th>Base position</th>
<th>I Kd (nM)</th>
<th>ΔΔG (Kcal/mole)</th>
<th>II Kd (nM)</th>
<th>ΔΔG (Kcal/mole)</th>
<th>III Kd (nM)</th>
<th>ΔΔG (Kcal/mole)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>G-T</td>
<td>26.7</td>
<td>1.10</td>
<td>G-A</td>
<td>92.5</td>
<td>1.83</td>
</tr>
<tr>
<td>2</td>
<td>T-G</td>
<td>529</td>
<td>2.86</td>
<td>T-A</td>
<td>453</td>
<td>2.77</td>
</tr>
<tr>
<td>3</td>
<td>G-C</td>
<td>637</td>
<td>2.97</td>
<td>G-A</td>
<td>416</td>
<td>2.72</td>
</tr>
<tr>
<td>4</td>
<td>T-A</td>
<td>308</td>
<td>2.54</td>
<td>T-C</td>
<td>361</td>
<td>2.64</td>
</tr>
<tr>
<td>5</td>
<td>A-C</td>
<td>13.5</td>
<td>0.7</td>
<td>A-G</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>A-G</td>
<td>361</td>
<td>2.64</td>
<td>A-C</td>
<td>160</td>
<td>2.16</td>
</tr>
<tr>
<td>7</td>
<td>A-T</td>
<td>9.27</td>
<td>0.48</td>
<td>A-G</td>
<td>40.4</td>
<td>1.35</td>
</tr>
<tr>
<td>8</td>
<td>C-A</td>
<td>425</td>
<td>2.73</td>
<td>C-T</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>9</td>
<td>G-A</td>
<td>451</td>
<td>2.77</td>
<td>G-C</td>
<td>369</td>
<td>2.65</td>
</tr>
<tr>
<td>10</td>
<td>A-T</td>
<td>232</td>
<td>2.38</td>
<td>A-G</td>
<td>34.3</td>
<td>1.25</td>
</tr>
<tr>
<td>11</td>
<td>T-G</td>
<td>679</td>
<td>3.01</td>
<td>T-A</td>
<td>1040</td>
<td>3.26</td>
</tr>
<tr>
<td>12</td>
<td>T-C</td>
<td>989</td>
<td>3.23</td>
<td>T-A</td>
<td>44.1</td>
<td>1.4</td>
</tr>
<tr>
<td>13</td>
<td>C-G</td>
<td>97.6</td>
<td>1.87</td>
<td>C-T</td>
<td>103</td>
<td>1.9</td>
</tr>
<tr>
<td>14</td>
<td>C-A</td>
<td>259</td>
<td>2.44</td>
<td>C-T</td>
<td>330</td>
<td>2.58</td>
</tr>
<tr>
<td>15</td>
<td>A-C</td>
<td>539</td>
<td>2.87</td>
<td>A-T</td>
<td>166</td>
<td>2.18</td>
</tr>
<tr>
<td>16</td>
<td>C-G</td>
<td>188</td>
<td>2.25</td>
<td>C-T</td>
<td>56</td>
<td>1.54</td>
</tr>
</tbody>
</table>

**Table 5.1**: This table lists the dissociation constants and the free energy changes of the mutant operators relative to the wild type operator. The columns I, II and III show the three sets of single base mutants at each of the 16 positions of the operator O_T. The ΔΔG values < 1.5 Kcal/mole are shaded.
Analysis of distribution of the repressor in the *E. coli* genome

From the binding energy values thus obtained, we wanted to see, what is the distribution of binding sites for the gal repressor protein over the entire *E. coli* genome, other than the unique operator site. The output of a FORTRAN based program was so designed that it gave us the position of each binding site (start site for a 16 base sequence) within the *E. coli* genome and the corresponding binding energy value of the repressor to that site. The values were grouped into different sets of <1, <2, <3, and so on, up to <6 Kcal/mole. Binding energy sites thus obtained can be viewed over the genome by a plot of genome positions versus energy values grouped from <1 to <6 Kcal/mole. The Figure 5.6 shows such a plot. This plot gives us a picture of distribution of repressor binding sites with energy difference of < 6Kcal/mole from the unique site energy distributed over the *E. coli* genome. The sites which have larger difference in energy values from the operator site seem to be of little importance in serving as binding sites for the repressor.
Figure 5.6: This polar plot shows the distribution of binding energy sites for the gal repressor with $\Delta\Delta G$ values from 0 to 6 Kcal/mole in the *E. coli* genome.
**Discussion**

The structure of the Lac repressor bound to the natural O1 operator shows that the repressor interacts more strongly with the left-half site of the operator than it does with the right half-site (Sadler, J.R. et al., 1983). Due to low resolution and thermal motion of the DNA in the crystals of the complex, detailed interactions of the side chains of the repressor with the DNA bases could not be interpreted (Bell, C.E. et al., 2001). But it appears from the structure that nucleotides at positions 6, 7, 8, 12, 13, 14 and 15 out of the 22 bases of the operator make direct contact with the repressor. It may be said that other bases which do not make direct contact with the protein may also have an influence on the overall interaction.

From our results, we conclude that it is not possible to define any particular pattern in which the bases of the operator DNA may seem to interact with the repressor. Only three out of the 48 mutations seem to have negligible effect on the binding energy, where the ΔΔG values are less than 1 Kcal/mole; the adenine to cytosine mutations at base positions 5 and 10, and the adenine to thymine mutation at base position 7. Six other mutations, randomly spread over the operator sequence, have moderately low effect on the interaction, with ΔΔG values less than 1.5 Kcal/mole. But all other mutations seem to have a considerably large effect on the interaction between the operator and repressor increasing the binding energy by values between 1.5 and 3.5 Kcal/mole or even greater for the adenine to guanine mutation at base position 5 and the cytosine to thymine mutation at base position 8 (values not shown).

The plot of distribution of repressor binding sites over the entire *E. coli* genome shows only a handful of sites that exist, with energy values within 6 Kcal/mole. Out of those
sites, only two are within energy values of 2.63 Kcal/mole and 3.07 Kcal/mole, the rest being above 4 Kcal/mole. It is known that the transcription factors find their target site by a combination of one-dimensional and three-dimensional diffusion. It has been established theoretically that more the roughness of the binding energy landscape the more difficult it gets for the protein to diffuse and thus it takes a longer search time to find its target (Slutsky, M. et al., 2004). From our results we conclude that the E. coli genome appears to offer a smooth binding energy landscape to the repressor protein with little or no quasistable sites for the protein other than the unique site. This perhaps also explains the efficient regulation of transcription and very low dissociation constant of the gal repressor-operator binding.