What evolution can tell us about protein-DNA interactions

Leonid Mirny

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

1. – Introduction

Protein-DNA interactions are central for the regulation of gene expression in the cell. Much progress has been made since the first DNA-binding protein was isolated [1]. The highest resolution picture of protein-DNA interactions is coming from more than 200 X-ray and NMR solved structures of protein DNA complexes [2]. As this information was accumulated, structures have been thoroughly examined by the authors. Protein-DNA complexes have been studied by chemical modifications (see [3] for review) and site-specific mutagenesis (e.g. [4, 5]); and binding motifs and interactions have been classified [6, 7, 8, 9]. Recently three groups [10, 11, 12] extensively studied representative protein-DNA complexes: chemical and physical properties of the interfaces, their polarity, size, shape and packing.

Although X-ray and NMR structures give us the most detailed picture of protein-DNA interactions, the structures are missing information about the energetics of the interactions and relative importance of different residues and nucleotides in recognition. Hence, by analyzing protein-DNA complexes alone, one can not tell why a protein selects one DNA site to bind instead of the others.

By mutating the protein and the DNA site one can identify the relative importance of different residues and nucleotides in protein-DNA recognition. These experiments are labor-intensive, making it impossible to study all possible mutations of a few residues. An enormous number of such mutations, however, have already been tested in the "natural laboratory" by molecular evolution. Families of homologous proteins tell us about mutations that were tolerated by the protein and those that were not. On the DNA side, sites recognized by the same protein, or by its orthologues in closely related organisms are identified by footprinting assays and bioinformatic techniques. Multiple alignments of both footprinted DNA sites [13] and homologous proteins [14, 15, 16, 17, 18] are publicly available.

In this study we combine and systematically analyze structures of protein-DNA complexes, footprinted DNA sites, and multiple alignments of DNA-binding proteins. The goal is to identify and understand primary determinants of specific DNA recognition by proteins.

In the first section we study how conservation of nucleotides in the DNA site is linked to nucleotides' structural role in the protein-DNA complex. By comparing sites recognized by the same protein we identify base-pairs conserved in evolution. Using structures of protein-DNA complexes we compute the number of interactions every base-pair has with the protein and match this number with the conservation of this base-pair. The first result of this study is that base pairs that have more interactions with the protein are more conserved in the binding sites. As natural as it is, this result has never been reported before.

Next we study the LacI family of homologous proteins and show that certain residues binding DNA exhibit a very special pattern of conservation: they are conserved within orthologues (that have the same binding specificity) and are variable between paralogues (that have different DNA binding specificities). This kind of pattern can serve as a "signature" of specificity determining residues.We develop a method to find such residues in the families of proteins grouped according to their function. This method is somewhat similar to evolutionary trace analysis [19, 20]. Since our method relies on rigorous statistical control, unlike evolutionary trace analysis, it does not require knowledge of protein structure to sort out false-positives. Applied to the LacI family our method identified 12 specificity determinants. When mapped onto the structure, 3 residues are binding DNA and 6 are surrounding the ligand-binding pocket in the ligand-binding domain. Available experimental information supports the critical role of the identified DNA-binding residues in determining the specificity of DNA recognition.

2. – DNA's point of view

2˙1. *Results*. – For our analysis we select all bacterial transcription factors for which sufficient footprinted sites in the DPI database [13] and high resolution X-ray or NMR structures are **both** available. Unfortunately, only five proteins satisfy this criteria: Crp, PurR, TrpR, Ihf and MetJ. For each structure we compute the number of contacts n_i each base pair i has with the protein, i.e. the number of heavy atoms that are at a distance less then $4.0\mathring{A}$ from a protein atom. To focus on sequence-specific interactions of the DNA with the protein, we exclude atoms belonging to the sugar-phosphate DNA backbone because they do not depend on the DNA sequence. We also compute the number of hydrogen bonds (including water-mediated) and the number of hydrophobic interactions each base-pair has with the protein. Hydrogen bonds are computed using

Protein	Number of Sites	Corr.Coeff (r)	Association (γ)
Crp	49	-0.62	**-0.98**
PurR	23	-0.50	**-0.77**
TrpR	4	-0.60	**-0.63**
Ihf	27	-0.72	**-0.64**
MetJ	16	+0.10	**-0.25**
MetJ SELEX holo	75	-0.59	**-0.80**
MetJ SELEX apo	56	-0.50	**-0.80**

Table I. – *Correlation between the number (n) of interactions with the proteins a base pair has and its variability (S) in footprinted sites.*

NUCPLOT/HBPLUS [21]. Two groups are said to have a hydrophobic interaction if both have CHARMM [22] group-charge less then 0.3 and they are separated by less then $3.5\AA$.

For aligned footprinted sites we compute variability (information contents) [23] at each position as

$$S_i = - \sum_{x=A,C,G,T} f_i(x) \log f_i(x) \tag{1}$$

Where $f_i(x)$ is a frequency of nucleotide x in position i at the site. Next we compute correlation between **S** and **n**. We use both the traditional linear correlation coefficient r [24] and a 2×2 association measure γ [25]. The 2×2 measure is used to compute the association between categorical variables. To use it, we classify positions as being variable ($S_i > S_{cut}$) vs conserved ($S_i \leq S_{cut}$) and as strongly involved ($n_i > n_{cut}$) vs slightly involved ($n_i < n_{cut}$) in interactions with the protein. To eliminate ambiguity in setting the cutoff S_{cut} and n_{cut} we use medians of **S** and **n** accordingly. This way we obtain a 2×2 variability-involvement frequency table $\boldsymbol{\rho}$,

$$\begin{aligned}
\rho_{11} &= \text{number of positions with } S_i > S_{cut} \text{ and } n_i > n_{cut} \\
\rho_{12} &= \text{number of positions with } S_i \leq S_{cut} \text{ and } n_i > n_{cut} \\
\rho_{21} &= \text{number of positions with } S_i > S_{cut} \text{ and } n_i \leq n_{cut} \\
\rho_{22} &= \text{number of positions with } S_i \leq S_{cut} \text{ and } n_i \leq n_{cut}
\end{aligned} \tag{2}$$

Then the association between S and n is measured as [25]

$$\gamma = \frac{\rho_{11}\rho_{22} - \rho_{12}\rho_{21}}{\rho_{11}\rho_{22} + \rho_{12}\rho_{21}} \tag{3}$$

Table I summarizes results for all five proteins. Strikingly, for all proteins except MetJ a strong negative correlation is observed. This indicates that base pairs that have more interactions with the protein n are more important for recognition, and hence have lower variability S. This transparent result has never been reported before. Importantly, all types of interactions were counted together. We did not discriminate between hydrogen bonds, hydrophobic or electrostatic interactions. When any single type of interaction is taken into account the correlation is much lower (see [26] for details).

Crp. Figure 1 presents S_i and n_i for the complex of Catabolite gene activator protein (CAP) with its site. CAP is a homodimer. The binding site of each domain can be seen as the region of high n_i and low S_i on the figure. Interestingly, the "right" site is slightly less conserved and correspondingly, has less tight interactions with the protein. Most of the interactions are formed by ARG180, ARG185 and GLU181 in both chains. They form both hydrogen bonds and hydrophobic interactions (by C_β, C_γ atoms interacting with the CH_3 group of T). Neither the hydrogen bonding pattern, nor the hydrophobic pattern alone correlate with observed conservation S. The total number of all types of interactions n, however, exhibits a strong correlation with S (the strongest among the proteins studied here.)

PurR. For purine repressor both S and n are very symmetric (see Fig. 2). However,the perfect symmetry of n is the result of the X-ray structure that was built assuming the two-fold symmetry of the molecule [27]. Correlation between S and n is high with a few exceptions, e.g. base pairs AT in positions 11 and 16 are very conserved, but have no interactions with the protein. Most other positions show a regular trend: S decreases as n increases. On the protein side, residues that have most of the contacts with the bases are THR14, ARG24, LEU52, ALA49 and ALA53. As in the case of Crp, both hydrogen bonding and hydrophobic interactions are involved and only their combination exhibits correlation with S.

TrpR. Only four natural footprinted sites are available for TrpR, leading to a poor profile of S. In spite of this problem, correlation with n is significant (see Fig. 3). Both n and S are symmetric exhibiting the distinct pattern of highly conserved $A_6C_7T_8$ and $A_{17}G_{18}T_{19}$. C_7 and G_{18} are the nucleotides that interact the most with the protein. Both half-sites have lots of hydrophobic interactions with the protein and very few hydrogen bonds. Other conserved base pairs are $G \cdot C_4$ and $C \cdot G_{21}$. Each pair has 7 interactions with the protein and a single hydrogen bond. However, mutations that eliminate this hydrogen bond have a very modest effect on the stability of the complex [28]. Perhaps, other types of interactions are determining specific recognition by Trp.

Comparison with sites obtained by SELEX lead to the correlation $r = -0.43$. However the motif obtained by SELEX is asymmetric and only the half-site is conserved ($GNACTAG$ motif). This inconsistency with the natural sites could result from different modes of binding observed in Trp repressor, which exhibits both dimer and tandem binding [28, 29].

Ihf Integration host factor (IHF) is known to bend DNA 160^o at the binding site. The site consists of two regions: a 5' region with no clear consensus and a 3' region with a significant but very small consensus. In accord with this data,the X-ray structure of the IHF complex shows very few (if any) protein-DNA contacts in the 5' region and tight protein-DNA interactions in the 3' region [30]. Our analysis brings quantitative support to these observations. Figure 4 shows the number of protein-DNA interactions and variability of the base pairs in the IHF site. Our results indicate that conservation in the 3' region can be very well explained by direct protein interactions with the DNA. Two peaks in n correspond to the regions where two proline residues (one from each protein chain) intercalate the DNA. Four arginines, ARG59 and ARG62 (from both A and B chains), are forming almost as many interactions with the bases as intercalating prolines. Most of the other interactions in these regions are formed by LYS65 (chains A&B), ILE72(chain A), ASN63 (chains A&B) and GLY61 (chains A&B). While arginines are involved in direct and water mediated hydrogen bonding, prolines and ilsoleucine are forming hydrophobic interactions with the bases. Two out of three hydrogen bonds with the bases, however, are formed by non-conserved G at position 29 and non-conserved C at position 32. Position 29 is occupied by G only in 15% of the sites and position 32 is occupied by C in 19% indicating that hydrogen bonding of these nucleotides does not lead to strong specificity. Another hydrogen bond and several non-bonded interactions are formed by ARG46(B) with base pairs at positions 20-22. These interactions are also apparently non-specific as base pairs at these positions are not conserved. In summary, a 0.72 correlation is observed in the IHF site, while the hydrogen binding pattern alone can not explain observed conservation. In contrast, hydrophobic interactions seem to correlate with the pattern of conservation.

MetJ is binding to arrays of two to five adjacent copies of an eight base-pair “metbox” sequence. Naturally occurring operators differ from the consensus sequence to a greater extent as the number of metboxes increases. This makes the motif obtained from the individual eight base-pair sites very weak exhibiting no correlation with the number of direct protein-DNA complexes. However,the conservation pattern of SELEX-derived sites does correlate with the number of interactions between the base pairs and the protein.

In summary, we showed that in the five different bacterial transcription factors the number of interactions a base pair has with the protein strongly correlates with conservation of this base pair. The origin of this correlation is clear: some of the direct interactions between the nucleotides and the protein are stabilizing the complex; then mutations of a more interacting base pair are more destabilizing and are eliminated in evolution. For the same reason residues that have more interactions in the protein (buried residues) are more conserved. Although this result for residues has been known for decades, a similar result for base pairs in protein-DNA complexes is reported here for the first time. Another result concerns the role of hydrogen bonds that are widely believed to dominate in determining the specificity and stability of protein-DNA complexes. Our results, on the contrary, indicate that hydrogen bonds alone can not explain the pattern of conservation in the site. Only when hydrogen bonds, hydrophobic and other interactions are taken together, does this number correlate with patterns of conservation.

The nature of protein-DNA interactions is very complex and involves hydrogen bonds, hydrophobic and electrostatic interactions and effects of “indirect readout” related to water extrusion, and local DNA bending and twisting. Surprisingly, such a simple parameter as number of direct interactions (that does not take into account even the different strength of interactions) is able to explain the patterns of conservation in the DNA binding sites. This result makes us believe that more complex models of protein-DNA energetics would be able to predict binding motifs for DNA-binding proteins.

3. – Proteins' point of view

3˙1. *Results.* –

3˙1.1. Conservation of DNA-binding residues. The examination of known protein DNA-complexes reveals several residues binding DNA bases. How conserved are these residues in protein evolution? To address this question we focus on a large LacI family of homologous DNA-binding proteins. All of them are bacterial transcription factors regulating the expression of proteins involved in sugar/nucleotide metabolism.

Figure 5 presents the multiple alignment of the DNA-binding domains in the LacI family. For each residue we computed variability as

$$S_i = -\sum_{x=1}^{20} f_i(x) \log f_i(x) \tag{4}$$

where x is a type of amino acid and $f_i(x)$ is its frequency at position i of the multiple alignment. Variability computed this way shows that some DNA-binding residues are very conserved, while others are not (see Fig 6).

In order to understand the origin of this high variability we split the LacI family into subgroups of orthologous proteins. We start from *E.coli* homologues of LacI. For each of them we find orthologues in a close bacterial genome by the bidirectional-hit method [31]. We discard orthologues, when a bidirectional hit is absent or weak (see [32] for details). By this method we obtain a family of 54 proteins grouped into 15 sub-families of orthologous proteins. All found orthologues are aligned by ClustalW [33] Importantly, most of the residues appearing variable across the *whole* family are conserved *within* every orthologous sub-family. This suggests that such residues can serve as specificity determinants. In fact, orthologous proteins from relatively close genomes are believed to have the same cellular function [31]. Hence, orthologous transcription factors are likely to regulate the same genes and bind similar sites on the DNA. Hence, we expect that residues determining DNA-binding specificity are conserved within a sub-family of orthologous transcription factors (e.g. within PurR proteins from *E.coli, H.influenzae and V.cholerae*). Moreover, specificity determinants must differ among paralogous proteins as they are binding different sites (and regulate different genes) in the same organism (e.g. PurR and GalR in *E.coli*). Some variable DNA-binding residues exhibit exactly this pattern of variation: they are conserved within every single orthologous sub-family and are different in different sub-families.

This pattern of variability can be considered a "signature" of the specificity-determining residues. Based on this idea, we developed a method to search for the specificity-determining residues in protein families.

3·1.2. Specificity determinants of LacI/GalR family. In order to identify residues with the pattern described above, we use *mutual information* as a measure of association between a residue type x and a sub-family index y:

$$I_i = \sum_{x,y} f_i(x,y) \log \frac{f_i(x,y)}{f_i(x) f(y)} \tag{5}$$

where $f_i(x)$ is the frequency of residue type x in position i of the multiple alignment, $f(y)$ is the fraction of proteins belonging to orthologous sub-family y, and $f_i(x,y)$ is the frequency of residue type x in the proteins of sub-family y in position i of the multiple alignment. Summation is over all types of residues $x = 1..20$ and over all sub-families y (for LacI $y = 1..15$). Mutual information has several important properties: (1) it is always positive; (2) it equals zero if and only if x and y are statistically independent; and (3)a large value of I_i indicates a strong association between x and y. The variability and composition of position i in the multiple alignment influence I_i as well. Hence, we can not rely on the value of I_i as an indicator of association; instead we estimate the statistical significance of I_i. We start from the zero-hypothesis of no association between x and y. Next we compute $P(I_i)$ the probability of observing I_i under this zero-hypothesis. Positions in the multiple sequence alignment that exhibit low $P(I_i)$ are likely to be specificity determinants.

We use shuffling to compute $P(I)$. For each position i we take a column of $\mathbf{x}$ and randomly shuffle this vector. Next we compute mutual information for shuffled $\mathbf{x}_{sh}$ and original $\mathbf{y}$: $I_{sh} = I(\mathbf{x}_{sh}, \mathbf{y})$. The procedure is repeated 10^3 times for each i. At each i we get a distribution $f(I_{sh})$, that turns out to be Gaussian even at the tails (data not shown). From this distribution one can compute the mutual information expected under the zero hypothesis as $I_i^0 = E[f(I_{sh})]$. However, I_i^0 is systematically lower than I_i since sequences within an orthologous sub-family are more similar to each other then between sub-families. Importantly, this bias is systematic and is the same for all positions in the protein. To compensate for this bias we make a linear transformation of the mutual informations obtained by shuffling: $I'_{sh} = aI_{sh} + b$. Coefficients a and b are chosen to minimize the squared deviation between I_i and I_i^0, i.e. $\sum_i \left(I_i - I_i^0\right)^2 \to \min$. Then from $f(I'_{sh})$ we compute $P(I_i)$. Details and derivation of the method are published elsewhere [32].

Figure 7 presents the mutual information I, the expected mutual information I^0 and the probability $P(I)$ computed for the LacI family. This plot reveals several important results: (1) The correlation between observed I_i and expected I_i^0 is very high ($\rho = 0.97$); this indicates that (i) the model used to compute expected mutual information is accurate and (ii) the vast majority of positions in the LacI family exhibit no functional association. (2) Very few positions have low $P(I)$ (i.e. I_i much greater than I_i^0) indicating that these positions have a strong association with the function of the protein and are probably specificity determinants.

Positions with strong function association ($P(I) < P_{cutoff} = 10^{-5}$) are : 15, 16, 55, 66, 69, 85, 114, 123, 146, 160, 221, 246. (The numbering is according to the 3D structure of PurR, PDB code 1wet). To understand the role of these residues we map them onto the structure of PurR.

Figure 8 presents the structure of the PurR-DNA complex with specificity-determining residues shown by space-fill. Examination of the structure brings us to the following conclusions. (1) Only three residues THR15, THR16 and LYS55 out of 12 are located in the DNA-binding domain. All three are deeply buried in the DNA grooves forming a net of interactions with the bases. (2) Another 7 specificity-determining residues, SER69, ALA66, CYS123, ASP146, ASP160, PHE221, and GLY246, are located in the ligand-binding pocket or close to it (within $8.5\mathring{A}$ from the ligand; see the insert). Such a structural location indicates that these residues are involved in ligand recognition. Since different orthologues have different ligands these residues change from sub-family to sub-family, but stay the same within sub-families. PHE221 is of special interest as its aromatic rings directly interact with the aromatic ligand of PurR (hypoxanthine or guanine). The other functionally linked residues, ALA66 CYS85 and LYS114, are located far from the DNA and the ligand and are either "false positives" or have some special role in alosteric regulation. For example LYS114 of one PurR chain is located next to LYS114 of the other chain and may be important for correct dimerization. In summary, the structural location of identified residues supports the view that they serve as specificity determinants in the LacI family. This includes the specificity of DNA recognition and ligand binding specificity.

The role of positions 15, 16 and 55 in specific DNA recognition is evident from a series of mutant experiments [34, 35]. When TYR15 and GLN16 of LacI were mutated to residues observed at these positions in the paralogues (malI, rafR, cytR etc) the mutants were preferentially binding corresponding operators of these proteins(malI, rafR, cytR etc.). Similarly, when GalR was mutated to have LacI's residues in positions 15 and 16 it was specifically binding sites of LacI. That is strong experimental evidence that positions 15 and 16 are responsible for determining DNA-binding specificity in proteins of this family. Although residue in 55 is binding DNA in the minor groove, this residue was shown to be critical for DNA recognition by PurR [35].

To the best of our knowledge residues we selected in the ligand-binding domain (except for 146) have not been the subject of mutagenic studies. Although mutations of several other residues were shown to interfere with ligand binding it is not clear how they influence specificity of ligand recognition. Our analysis suggests several possible experiments to test the specificity of ligand binding.

First, one can make single mutations of the putative specificity-determining residues and study the binding affinity and specificity of the mutants. Second, as in experiments with DNA-binding residues, one can "transplant" some or all of the outlined residues from, say, LacI to PurR and measure the selective binding of LacI ligand vs PurR ligand by the mutant. The main question is whether specificity can be re-designed by changing this small set of theoretically predicted residues. Third, since most of these proteins are involved in relatively simple and well-understood transcription regulation processes one can make an *in vivo* study, e.g. PurR protein with LacI DNA specificity and vice versa. If successful this kind of "chimeric" protein can be used to re-design the network of cellular regulation.

In summary, we studied protein-DNA interactions and the evolution of DNA-binding proteins by analyzing their sequences and structures. First, we found that base-pairs that have more interactions with the protein are more conserved in evolution. We also showed that, in contrast to the prevalent view, hydrogen bonds are not the main players in protein-DNA recognition. Only when taken together can hydrogen bonds, hydrophobic and electrostatic interactions explain differing conservation of different base-pairs in the DNA site. In the second part of this work we focused on LacI/PurR family of bacterial transcription regulators. We showed that certain residues responsible for DNA recognition exhibit *no conservation* among homologues. These residues are conserved among orthologous proteins that bind the same site and are different in paralogous proteins that bind different sites. Based on this idea we developed a method to identify such residues in the multiple alignment of proteins grouped according to their specificity. Using this method we found 12 specificity-determining residues in the LacI/PurR family. Structural location and available experimental information strongly support the role of these residues as functional determinants. The method is general and can be applied to any family of proteins grouped according to their function.

The work reported in this lecture was done in collaboration with Dr. Michael Gelfand from the Moscow Institute of Industrial Genetics and was supported by William F. Milton Fund and the Harvard Society of Fellows.

REFERENCES

[1] Gilbert, W. and Muller-Hill, B. *Proc Natl Acad Sci U S A* **58**, 2415–21 (1967).

[2] Berman, H., Zardecki, C., and Westbrook, J. *Acta Crystallogr D Biol Crystallogr* **54**, 1095–104 (1998).

[3] Larson, C. and Verdine, G. *The Chemistry of Protein-DNA Interactions*, 324–346. Bioorganic Chemistry: Nucleic Acids. Oxford University Press (1996).

[4] Fields, D., He, Y., Al-Uzri, A., and Stormo, G. *J Mol Biol* **271**, 178–94 (1997).

[5] Brown, B. and Sauer, R. *Proc Natl Acad Sci U S A* **96**, 1983–8 (1999).

[6] Harrison, S. *Nature* **353**, 715–9 (1991).

[7] Pabo, C. and Sauer, R. *Annu Rev Biochem* **61**, 1053–95 (1992).

[8] Wintjens, R. and Rooman, M. *J Mol Biol* **262**, 294–313 (1996).

[9] Sauer, R. and Harrison, S. *Curr Opin Struct Biol* **6**, 51–2 (1996).

[10] Jones, S., van, H., Berman, H., and Thornton, J. *J Mol Biol* **287**, 877–96 (1999).

[11] Nadassy, K., Wodak, S., and Janin, J. *Biochemistry* **38**, 1999–2017 (1999).

[12] Pabo, C. and Nekludova, L. *J Mol Biol* **301**, 597–624 (2000).

[13] Robison, K., McGuire, A., and Church, G. *J Mol Biol* **284**, 241–54 (1998).

[14] Bateman, A., Birney, E., Durbin, R., Eddy, S., Howe, K., and Sonnhammer, E. *Nucleic Acids Res* **28**, 263–6 (2000).

[15] Corpet, F., Servant, F., Gouzy, J., and Kahn, D. *Nucleic Acids Res* **28**, 267–9 (2000).

[16] Krause, A., Stoye, J., and Vingron, M. *Nucleic Acids Res* **28**, 270–2 (2000).

[17] Dodge, C., Schneider, R., and Sander, C. *Nucleic Acids Res* **26**, 313–5 (1998).

[18] Henikoff, J., Greene, E., Pietrokovski, S., and Henikoff, S. *Nucleic Acids Res* **28**, 228–30 (2000).

[19] Lichtarge, O., Bourne, H., and Cohen, F. *J Mol Biol* **257**, 342–58 (1996).

[20] Lichtarge, O., Yamamoto, K., and Cohen, F. *J Mol Biol* **274**, 325–37 (1997).

[21] Luscombe, N., Laskowski, R., and Thornton, J. *Nucleic Acids Res* **25**, 4940–5 (1997).

[22] chapter CHARMM: The Energy Function and Its Parameterization with an Overview of the Program.

[23] Stormo, G., Schneider, T., and Gold, L. *Nucleic Acids Res* **14**, 6661–79 (1986).

[24] DeGroot, M. *Probability and statistics.* Addison-Wesley Pub. Co, Reading, Mass., (1996).

[25] Goodman, L. and Kruskal, W. *Measures of association for cross classifications.* Springer series in statistics. Springer-Verlag, New York, (1979).

[26] Mirny, L. and Gelfand, M. to be published.

[27] Schumacher, M., Choi, K., Zalkin, H., and Brennan, R. *Science* **266**, 763–70 (1994).

[28] Grillo, A., Brown, M., and Royer, C. *J Mol Biol* **287**, 539–54 (1999).

[29] Lawson, C. and Carey, J. *Nature* **366**, 178–82 (1993).

[30] Rice, P. *Curr Opin Struct Biol* **7**, 86–93 (1997).

[31] Tatusov, R., Galperin, M., Natale, D., and Koonin, E. *Nucleic Acids Res* **28**, 33–6 (2000).

[32] Mirny, L. and Gelfand, M. to be published.

[33] Thompson, J., Higgins, D., and Gibson, T. *Nucleic Acids Res* **22**, 4673–80 (1994).

[34] Lehming, N., Sartorius, J., Kisters-Woike, B., von, W.-B., and Muller-Hill, B. *EMBO J* **9**, 615–21 (1990).

[35] Glasfeld, A., Koehler, A., Schumacher, M., and Brennan, R. *J Mol Biol* **291**, 347–61 (1999).

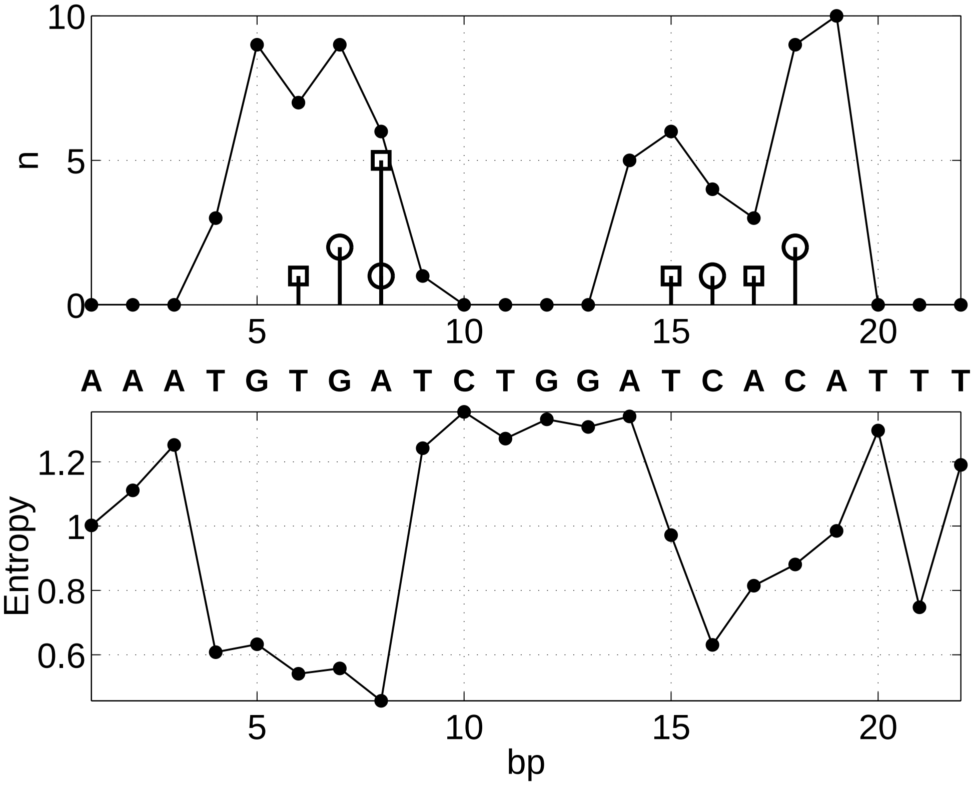

Fig. 1. – Crp site. Top: thin line shows the number of interactions n base-pairs have with the protein. The number of hydrogen bonds formed by the base-pair (including water-mediated bonds) is shown by large circles. The number of hydrophobic interactions between a base-pair and a protein is shown by large squares. Bottom: the variability (entropy) in footprinted DNA sites. The "consensus" (most frequent) nucleotides are shown by letters above the plot

.

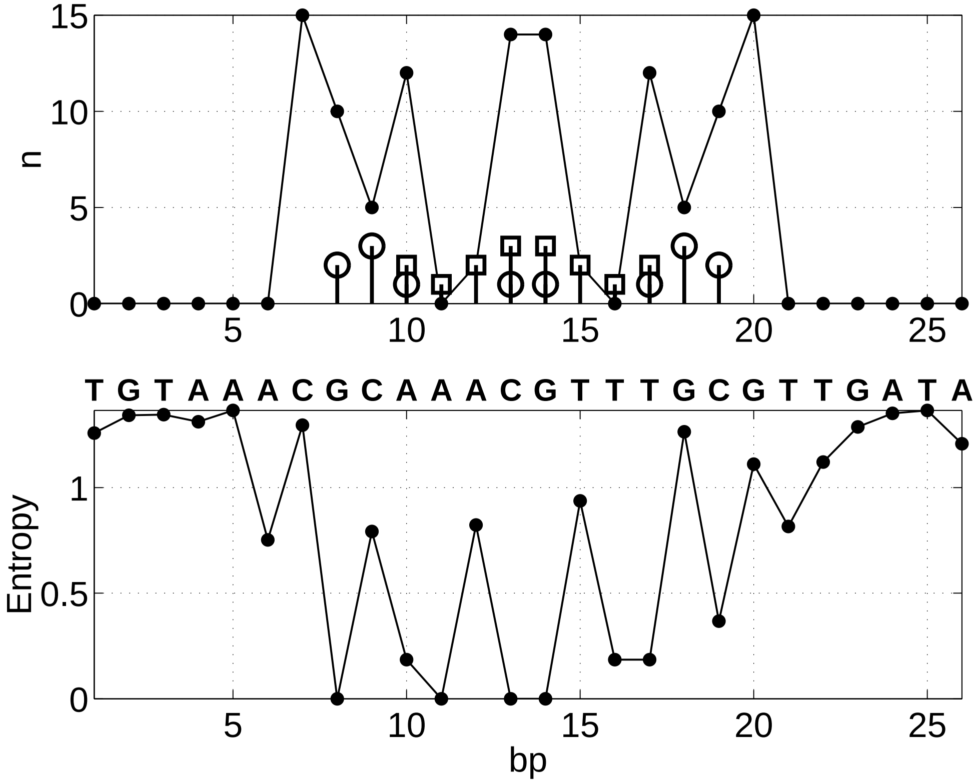

Fig. 2. – PurR site. Notation as on Figure 1

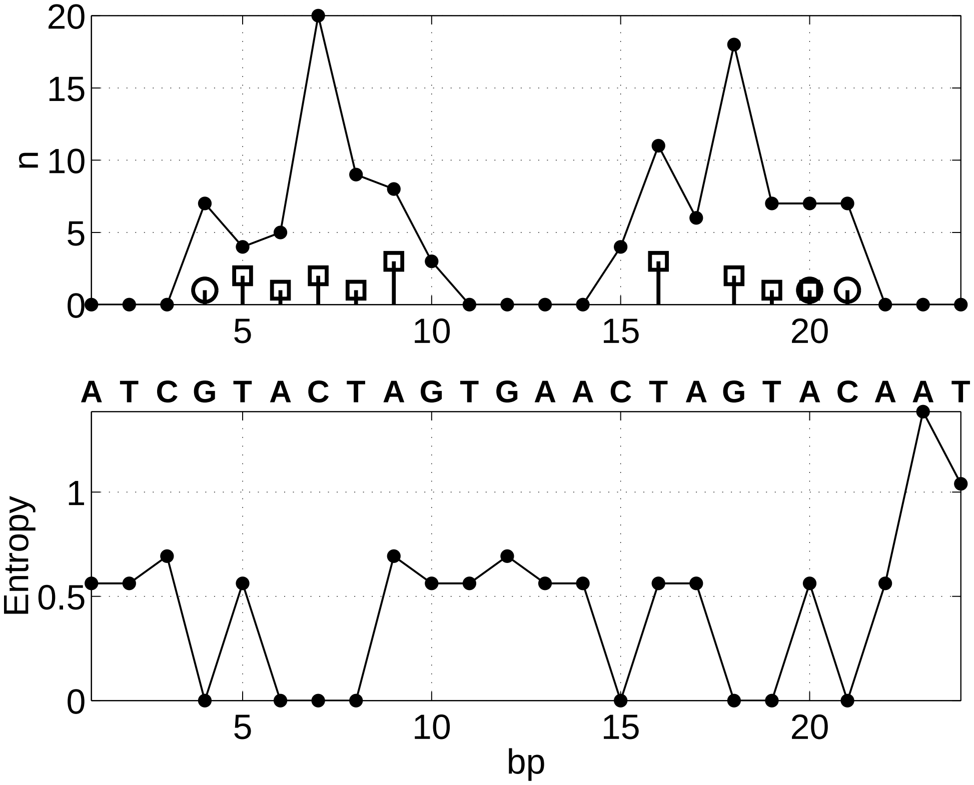

Fig. 3. – Trp site. Notation as on Figure 1

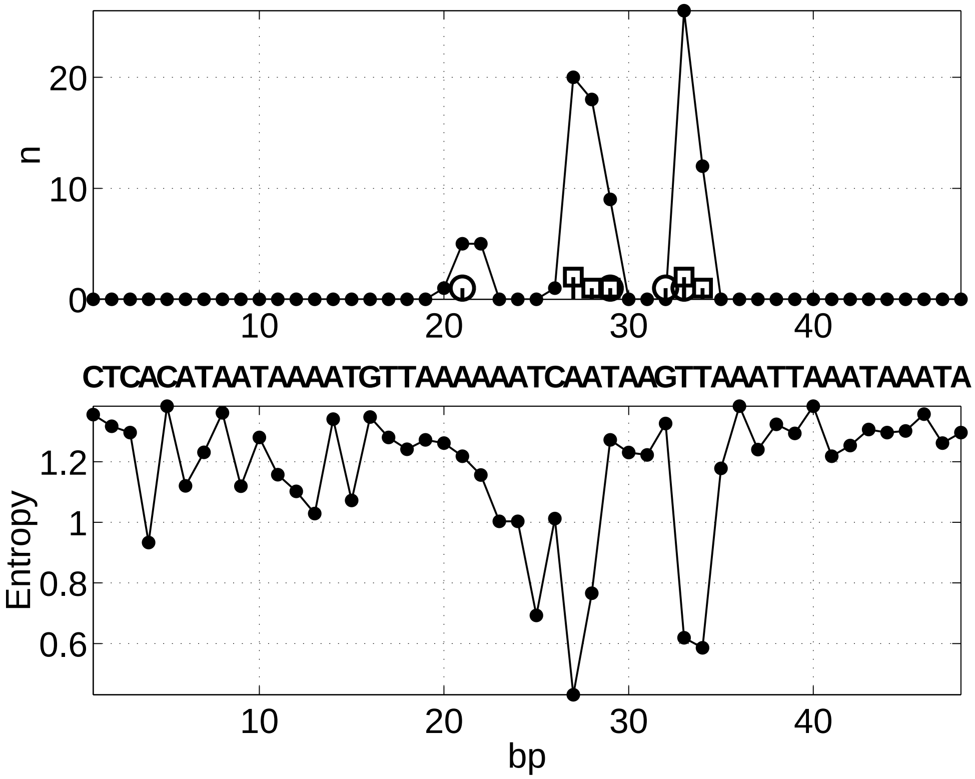

Fig. 4. – Ihf site. Notation as on Figure 1

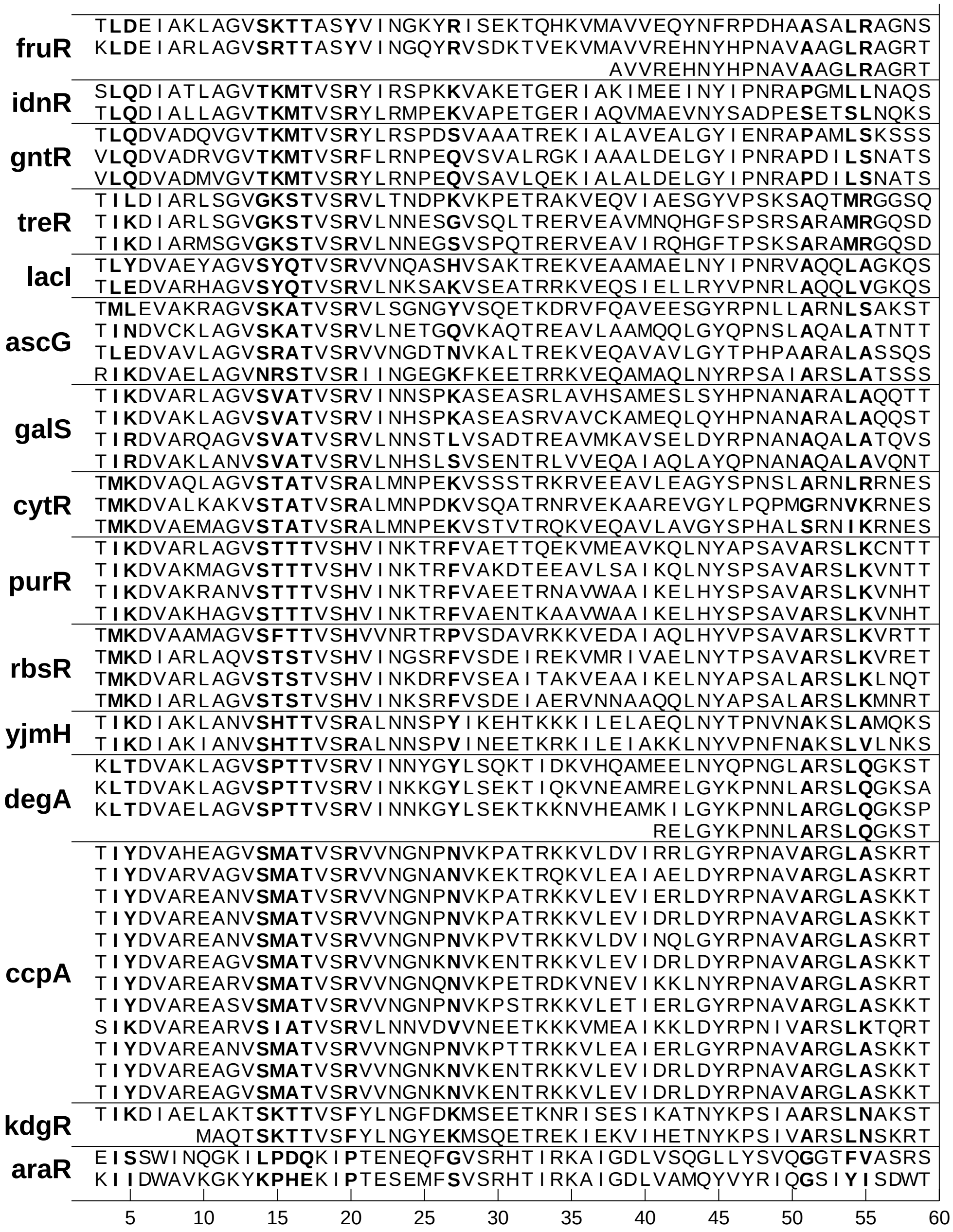
fruR
TLDEIAKLAGVSKTTASYVINGKYRISEKTQHKVMAVVEQYNFRPDHAASALRAGNS
KLDEIARLAGVSRTTASYVINGQYRVSDKTVEKVMAVVREHNYHPNAVAAGLRAGRT
AVVREHNYHPNAVAAGLRAGRT
idnR
SLQDIATLAGVTKMTVSRYIRSPKKVAKETGERIAKIMEEINYIPNRAPGMLLNAQS
TLQDIALLAGVTKMTVSRYLRMPEKVAPETGERIAQVMAEVNYSADPESETSLNQKS
gntR
TLQDVADQVGVTKMTVSRYLRSPDSVAAATREKIALAVEALGYIENRAPAMLSKSSS
VLQDVADRVGVTKMTVSRFLRNPEQVSVALRGKIAAALDELGYIPNRAPDILSNATS
VLQDVADMVGVTKMTVSRYLRNPEQVSAVLQEKIALALDELGYIPNRAPDILSNATS
treR
TILDIARLSGVGKSTVSRVLTNDPKVKPETRAKVEQVIAESGYVPSKSAQTMRGGSQ
TIKDIARLSGVGKSTVSRVLNNESGVSQLTRERVEAVMNQHGFSPSRSARAMRGQSD
TIKDIARMSGVGKSTVSRVLNNEGSVSPQTRERVEAVIRQHGFTPSKSARAMRGQSD
lacI
TLYDVAEYAGVSYQTVSRVVNQASHVSAKTREKVEAAMAELNYIPNRVAQQLAGKQS
TLEDVARHAGVSYQTVSRVLNKSAKVSEATRRKVEQSIELLRYVPNRLAQQLVGKQS
ascG
TMLEVAKRAGVSKATVSRVLSGNGYVSQETKDRVFQAVEESGYRPNLLARNLSAKST
TINDVCKLAGVSKATVSRVLNETGQVKAQTREAVLAAMQQLGYQPNSLAQALATNTT
TLEDVAVLAGVSRATVSRVVNGDTNVKALTREKVEQAVAVLGYTPHPAARALASSQS
RIKDVAELAGVNRSTVSRIINGEGKFKEETRRKVEQAMAQLNYRPSAIARSLATSSS
galS
TIKDVARLAGVSVATVSRVINNSPKASEASRLAVHSAMESLSYHPNANARALAQQTT
TIKDVAKLAGVSVATVSRVINHSPKASEASRVAVCKAMEQLQYHPNANARALAQQST
TIRDVARQAGVSVATVSRVLNNSTLVSADTREAVMKAVSELDYRPNANAQALATQVS
TIRDVAKLANVSVATVSRVLNHSLSVSENTRLVVEQAIAQLAYQPNANAQALAVQNT
cytR
TMKDVAQLAGVSTATVSRALMNPEKVSSSTRKRVEEAVLEAGYSPNSLARNLRRNES
TMKDVALKAKVSTATVSRALMNPDKVSQATRNRVEKAAREVGYLPQPMGRNVKRNES
TMKDVAEMAGVSTATVSRALMNPEKVSTVTRQKVEQAVLAVGYSPHALSRNIKRNES
purR
TIKDVARLAGVSTTTVSHVINKTRFVAETTQEKVMEAVKQLNYAPSAVARSLKCNTT
TIKDVAKMAGVSTTTVSHVINKTRFVAKDTEEAVLSAIKQLNYSPSAVARSLKVNTT
TIKDVAKRANVSTTTVSHVINKTRFVAEETRNAVWAAIKELHYSPSAVARSLKVNHT
TIKDVAKHAGVSTTTVSHVINKTRFVAENTKAAVWAAIKELHYSPSAVARSLKVNHT
rbsR
TMKDVAAMAGVSFTTVSHVVNRTRPVSDAVRKKVEDAIAQLHYVPSAVARSLKVRTT
TMKDIARLAQVSTSTVSHVINGSRFVSDEIREKVMRIVAELNYTPSAVARSLKVRET
TMKDVARLAGVSTSTVSHVINKDRFVSEAITAKVEAAIKELNYAPSALARSLKLNQT
TMKDIARLAGVSTSTVSHVINKSRFVSDEIAERVNNAAQQLNYAPSALARSLKMNRT
yjmH
TIKDIAKLANVSHTTVSRALNNSPYIKEHTKKKILELAEQLNYTPNVNAKSLAMQKS
TIKDIAKIANVSHTTVSRALNNSPVINEETKRKILEIAKKLNYVPNFNAKSLVLNKS
degA
KLTDVAKLAGVSPTTVSRVINNYGYLSQKTIDKVHQAMEELNYQPNGLARSLQGKST
KLTDVAKLAGVSPTTVSRVINKKGYLSEKTIQKVNEAMRELGYKPNNLARSLQGKSA
KLTDVAELAGVSPTTVSRVINNKGYLSEKTKKNVHEAMKILGYKPNNLARGLQGKSP
RELGYKPNNLARSLQGKST
ccpA
TIYDVAHEAGVSMATVSRVVNGNPNVKPATRKKVLDVIRRLGYRPNAVARGLASKRT
TIYDVARVAGVSMATVSRVVNGNANVKEKTRQKVLEAIAELDYRPNAVARGLASKRT
TIYDVAREANVSMATVSRVVNGNPNVKPATRKKVLEVIERLDYRPNAVARGLASKKT
TIYDVAREANVSMATVSRVVNGNPNVKPATRKKVLEVIDRLDYRPNAVARGLASKKT
TIYDVAREANVSMATVSRVVNGNPNVKPVTRKKVLDVINQLGYRPNAVARGLASKRT
TIYDVAREAGVSMATVSRVVNGNKNVKENTRKKVLEVIDRLDYRPNAVARGLASKKT
TIYDVAREARVSMATVSRVVNGNQNVKPETRDKVNEVIKKLNYRPNAVARGLASKRT
TIYDVAREASVSMATVSRVVNGNPNVKPSTRKKVLETIERLGYRPNAVARGLASKKT
SIKDVAREARVSIATVSRVLNNVDVVNEETKKKVMEAIKKLDYRPNIVARSLKTQRT
TIYDVAREANVSMATVSRVVNGNPNVKPTTRKKVLEAIERLGYRPNAVARGLASKKT
TIYDVAREAGVSMATVSRVVNGNKNVKENTRKKVLEVIDRLDYRPNAVARGLASKKT
TIYDVAREAGVSMATVSRVVNGNKNVKENTRKKVLEVIDRLDYRPNAVARGLASKKT
kdgR
TIKDIAELAKTSKTTVSFYLNGFDKMSEETKNRISESIKATNYKPSIAARSLNAKST
MAQTSKTTVSFYLNGYEKMSQETREKIEKVIHETNYKPSIVARSLNSKRT
araR
EISSWINQGKILPDQKIPTENEQFGVSRHTIRKAIGDLVSQGLLYSVQGGTFVASRS
KIIDWAVKGKYKPHEKIPTESEMFSVSRHTIRKAIGDLVAMQYVYRIQGSIYISDWT
5
10
15
20
25
30
35
40
45
50
55
60

Fig. 5. – DNA-binding domain of LacI proteins grouped into orthologous sub-families. Sub-families are separated by lines, the name of each sub-family is shown on the left. Residues binding the nucleotides (non-backbone DNA atoms) are shown in bold. Note that some of them are very conserved (e.g. S14, T17) while some are not. Residues in positions 15, 16 and 55 are specificity determinants as they are conserved within most sub-families and are different between sub-families.

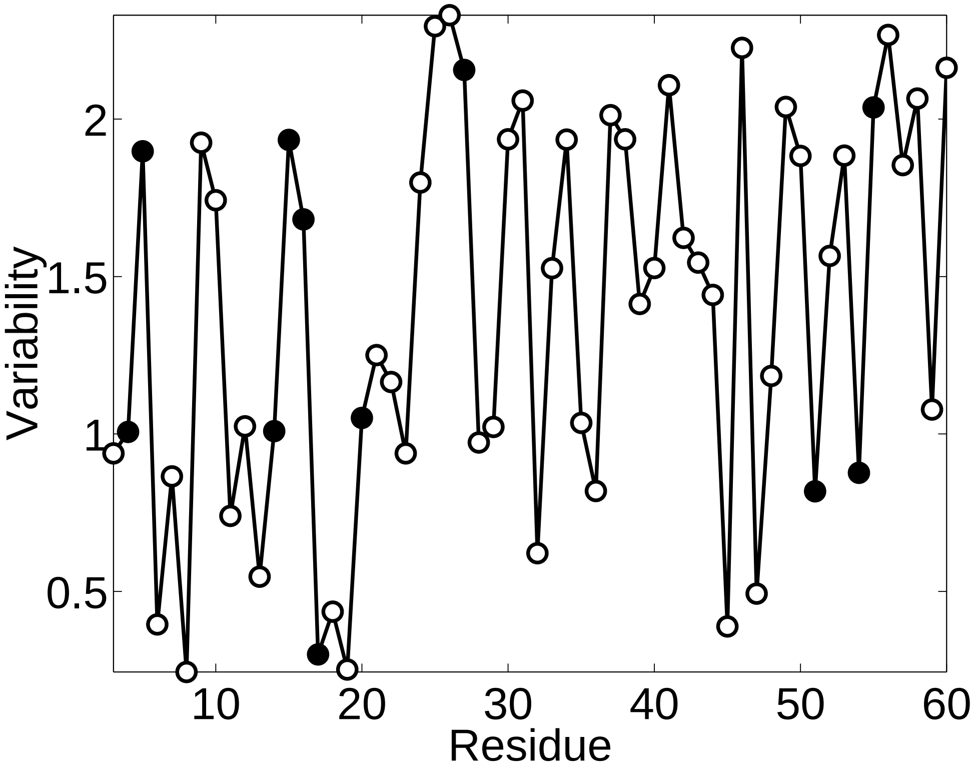

Fig. 6. – Variability (entropy) in the DNA-binding domain of LacI family. Nucleotide-binding residues are shown by filled circles.

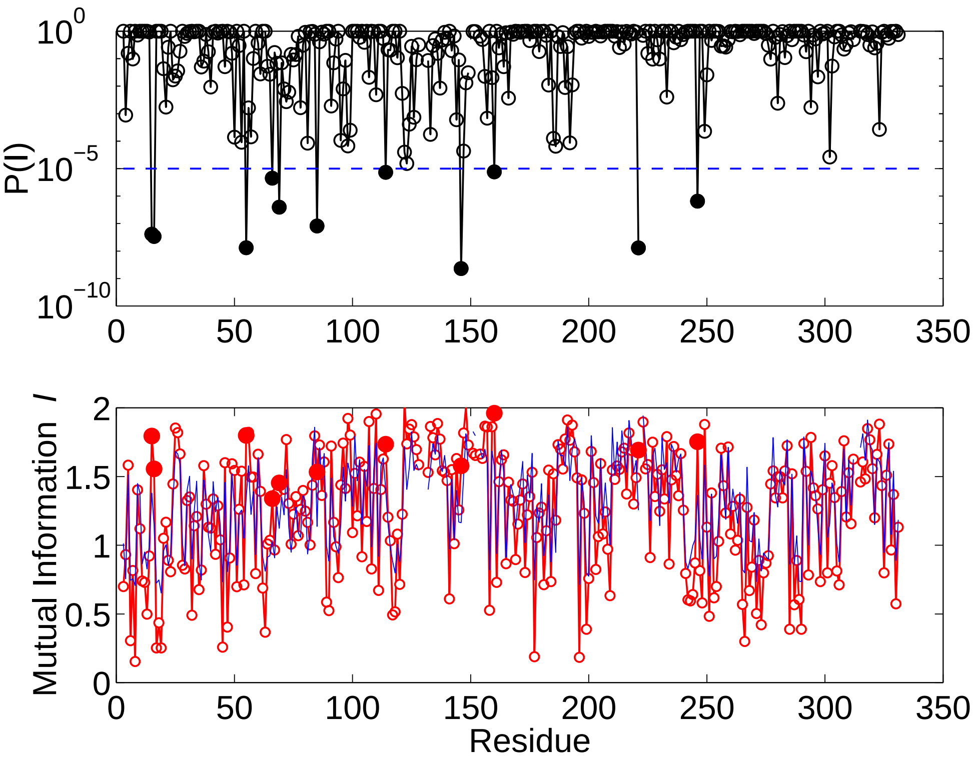

Fig. 7. – Identification of specificity-determinants in LacI family. Top: $P(I)$ probability to observe mutual information I by chance. Positions with $P(I) < 10^{-5}$ (below broken line) are shown by filled circles. These positions are specificity determinants. Bottom: observed (red) and expected (blue) mutual information. Correlation 0.97. Note few specificity determinants (filled red circles) with mutual information much higher then expected

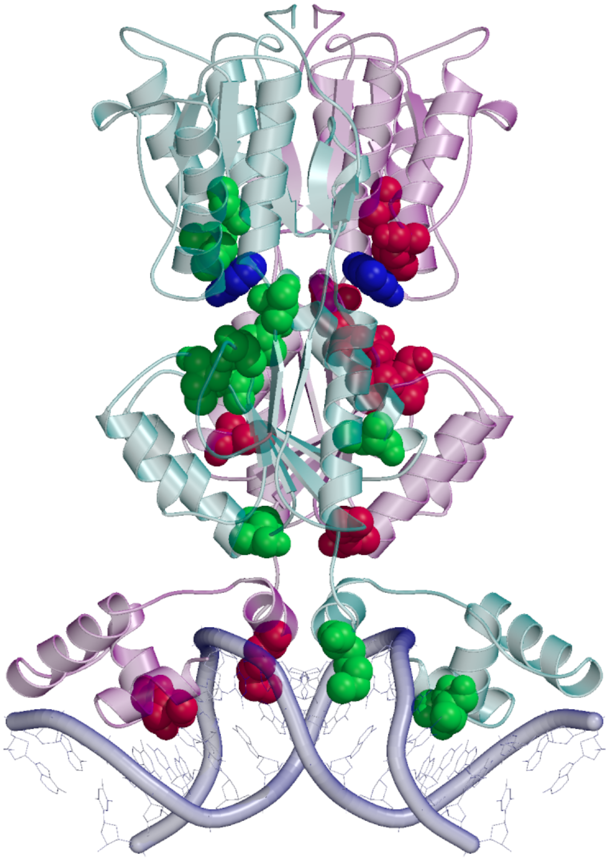

Fig. 8. – Structure of PurR-DNA complex. Two protein chains are shown by semi-transparent ribbons in green and pink. Ligands (guanine) is shown by blue space-fill. Residues identified as specificity determinants are shown by space-fill and colored according to their chain (green and red). Note (1) three residues deeply buried into the DNA, (2) a set of residues in the ligand binding site.