Quarterly Reviews of Biophysics **38**, 4 (2005), pp. 339–344. © 2006 Cambridge University Press
doi:10.1017/S0033583505004105 Printed in the United Kingdom
First published online 6 March 2006

ESSAY

Recognizing DNA

Richard Lavery

Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, Paris, France

Abstract. It has become clear that there is no simple 'code' for protein–DNA recognition and that selecting an optimal binding sequence along the DNA double helix corresponds to more than simply forming a set of specific hydrogen bonds or steric interactions. However, it has been difficult to characterize the so-called indirect components of recognition. While DNA deformation certainly underlies indirect recognition, it is not easy to determine how local fine structure and deformability depend on base sequence or exactly what percentage of recognition should be attributed to such factors. Molecular modelling can help to develop these ideas into a quantitative model, provided the calculations can be carried out fast enough to enable a comprehensive survey of base-sequence effects. I present here some recent results from our group and their consequences for improving our understanding of protein–DNA binding, and their potential for predicting, and eventually modulating, protein–DNA binding.

Why is recognizing DNA a problem? As early analyses pointed out, at least the major groove of the double helix is sterically accessible, and, within this groove, the various Watson–Crick base pairs can be distinguished by their distinctive pattern of hydrogen-bond donor and acceptor groups. If amino acid (aa) side-chains could be correctly organized on the surface of a binding protein, it should be relatively simple to recognize a series of consecutive base pairs by constructing a series of specific aa-base interactions. Specificity could be further improved by using side-chains which could make two simultaneous hydrogen bonds to a given base (e.g. Gln-A, Arg-G). This is the so-called direct view of DNA recognition. It can also be seen as a digital model, where recognition results from adding up many bits of information represented by the individual interactions.

As more crystal structures of protein–DNA complexes were obtained, it became clear that this digital model was only partially correct. First, although certain amino acids effectively showed a bias towards binding to certain bases, there was clearly no one-to-one recognition code. Second, several complexes had too few specific aa-base contacts to explain the length of the base sequence actually recognized by the protein. Something else was needed. The missing feature was a common element of most biological interactions, the structural adaptability of the interacting partners. In this case, it was the local structure and/or the deformability of DNA as a function of its base sequence that needed to be taken into account. Since it was already known that the base sequence leads to static and mechanical heterogeneity within the DNA double helix, it was not

Address for correspondence: R. Lavery, Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, Paris 75005, France.
Tel.: 33-1-58-41-50-16; Fax: 33-1-58-41-50-26; E-mail: rlavery@ibpc.fr

https://doi.org/10.1017/S0033583505004105 Published online by Cambridge University Press

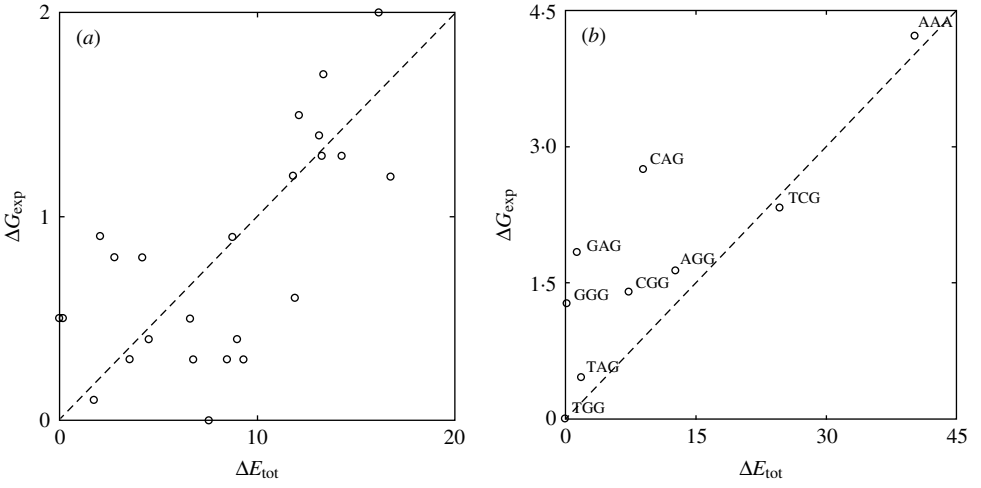


Fig. 1. Comparison of experimental binding free energies and calculated energies for mutants in the flanking regions of a *Bam*HI–DNA complex and in the centre of a Zif268–DNA complex (energies in kcal mol^{-1}).

unreasonable to assume that a binding protein could exploit this heterogeneity as part of the recognition process. This is the so-called indirect view of DNA recognition. It can also be seen as an analogue model, where recognition results from a collective property of the recognition site.

If we want to really understand protein–DNA binding and especially if we want to be able to rationally design proteins that will bind to specific DNA sites, it is necessary to turn the general direct/indirect classification into something much more quantitative. This is not an easy task, since neither structures nor binding free energies can be easily deconvoluted into direct and indirect contributions. The same difficulty applies to point mutation studies. In principle, modelling approaches could help, since deconvoluting energetic factors is one of their strong points. The problem is that calculating the free energy of binding for a protein–DNA complex is already a difficult computational task, even for a single DNA sequence. Resolving recognition mechanisms can easily require comparing millions of sequences.

It is, therefore, necessary to introduce some approximations. Amongst the main factors that contribute to protein–DNA recognition we can list: protein–DNA interaction energy, DNA deformation energy, DNA and protein dehydration, ion release, and conformational and translational entropy loss. This list has been roughly ordered in terms of the decreasing sensitivity of each contribution to the DNA base sequence. Although there may be some arguments about ordering at the end of the list, it seems likely that the protein–DNA interaction energy (ΔE_{int}) and the DNA deformation energy (ΔE_{def}) are the prime candidates for sequence sensitivity (whether or not they dominate the total interaction free energy). Happily, these two terms can also be linked respectively to the notions of direct and indirect recognition. As we compare different DNA sequences, all direct recognition must show up in variations of the protein–DNA interaction energy, while the majority of indirect recognition will be reflected by variations in the DNA deformation energy. Figure 1 shows that these two terms ($\Delta E_{tot} = \Delta E_{int} + \Delta E_{def}$), calculated with an all-atom model and an appropriate force field, do surprisingly well in reproducing the effect of point mutations in two typical protein–DNA complexes.

Although calculating molecular mechanics energies is much faster than calculating free energies, it is still too slow for studying sequence specificity. As the length N of a recognition site

https://doi.org/10.1017/S0033583505004105 Published online by Cambridge University Press

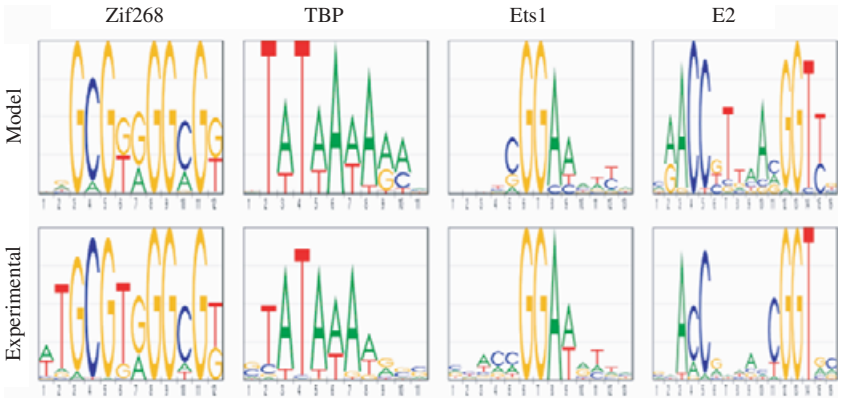


Fig. 2. Experimental and calculated consensus sequence logos for a variety of DNA-binding proteins (experimental results derived from the TRANSFAC database).

increases, the number of possible base combinations goes up as 4^N. For 6 base pairs (bp) (a typical site for a nuclease) there are already 4096 possibilities, for 10 bp this number becomes one million, and for 15 bp it exceeds 10^9. The latter number roughly corresponds to the size of the human genome, and consequently a given 15-bp sequence would only be expected to occur once by chance. This explains why highly specific eukaryotic proteins need to recognize sites which are roughly 15-bp long.

In order to speed up our calculations we used a computational trick. Instead of calculating the energy of a protein–DNA complex (and of the corresponding isolated DNA fragment) for a given sequence, we calculated the energy components corresponding to every possible base pair at every position within the DNA fragment. These components are stored in a matrix which contains the information on protein–DNA interaction energies and internal DNA energies, base pair by base pair. Using this matrix, it is possible to calculate the total energies ΔE_{int} and ΔE_{def} by simply adding up the appropriate matrix elements. This is much faster than the original calculations and is easily applicable to studies of millions of sequences.

To analyse the results of such a study, we order the sum of ΔE_{int} and ΔE_{def} (this is our estimate of the total complexation energy, ΔE_{tot}) for all sequences and then select those sequences which fall within a given energy cut-off (5 kcal mol^{-1} with respect to the best binding sequence). This is the computational equivalent of carrying out millions of quantitative binding assays. The outcome can be converted into an easily readable 'sequence logo', where the height of each letter at each position is proportional to the probability of its appearance in the selected set of strongly binding sequences. Figure 2 shows that these results compare well with experimental sequence logos for a variety of DNA-binding proteins.

We can now get back to the original question of how direct and indirect effects contribute to recognition. Answering this question involves finding out how much ΔE_{int} and ΔE_{def} contribute to defining the set of best binding sequences. To do this, we take this set of sequences and see how it is distributed within the total list of sequences, ordered in terms of ΔE_{int} or ΔE_{def}. If the order is similar to that of ΔE_{tot}, the corresponding energy term makes an important contribution. If the order is different, and the best binding sequences are mixed in with a large set of other sequences, the corresponding energy term in not important. Sequence logos can again be used to analyse the results. However, it is also possible to get a more quantitative result by noting that recognition can be expressed in units of base pairs. Consider a DNA fragment containing

https://doi.org/10.1017/S0033583505004105 Published online by Cambridge University Press

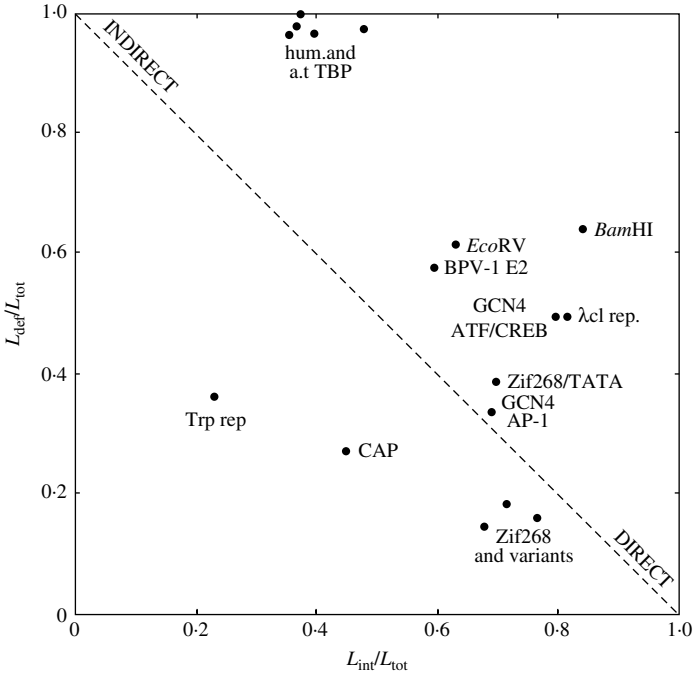


Fig. 3. Indirect (L_{def}/L_{tot}) *versus* direct (L_{int}/L_{tot}) contributions to DNA recognition for various DNA-binding proteins. (Adapted from Paillard & Lavery, 2004.)

N base pairs. This fragment can be built with 4^N possible base sequences. If our analysis of protein binding leads to a set of M best binding sequences, this implies that the protein tolerates a sequence variability corresponding to S base pairs, where $M=4^S$. Conversely, this also implies that protein binding has defined a site with an effective length of L_{tot} base pairs, where $L_{tot}=N-S=N-\log(M/4)$. We can apply the same reasoning to protein–DNA interaction and DNA deformation by calculating the total number of sequences encountered before reaching the least favourable of the set of M best binding sequences defined with ΔE_{tot}. Using the same equation as above, we obtain L_{int} and L_{def}, the binding-site lengths corresponding respectively to direct and indirect recognition.

It turns out that L_{int} and L_{def} vary quite dramatically amongst different protein–DNA complexes. If we normalize the total binding-site length to unity, by dividing L_{int} and L_{def} by L_{tot}, we can plot different complexes on the same graph (see Fig. 3). In this figure, the bottom right-hand corner corresponds to pure direct (digital) recognition, while the top left-hand corner corresponds to pure indirect (analogue) recognition. In fact, there are proteins which are not far from belonging to these two extreme cases. Zif268 (direct recognition) and the TATA-box binding protein (TBP; indirect recognition) are good examples. However, most proteins seem to find their target sites using hybrid mechanisms, where both direct and indirect terms contribute. It was a surprise to discover that indirect terms can play an important role even for complexes where the bound state shows relatively little DNA deformation. We will return to this point shortly.

Note that, in general, $L_{int}+L_{def}\neq L_{tot}$. If this were the case, all the points in Fig. 3 would lie along the leading diagonal. In fact, the positions of the points with respect to this diagonal tell us

https://doi.org/10.1017/S0033583505004105 Published online by Cambridge University Press

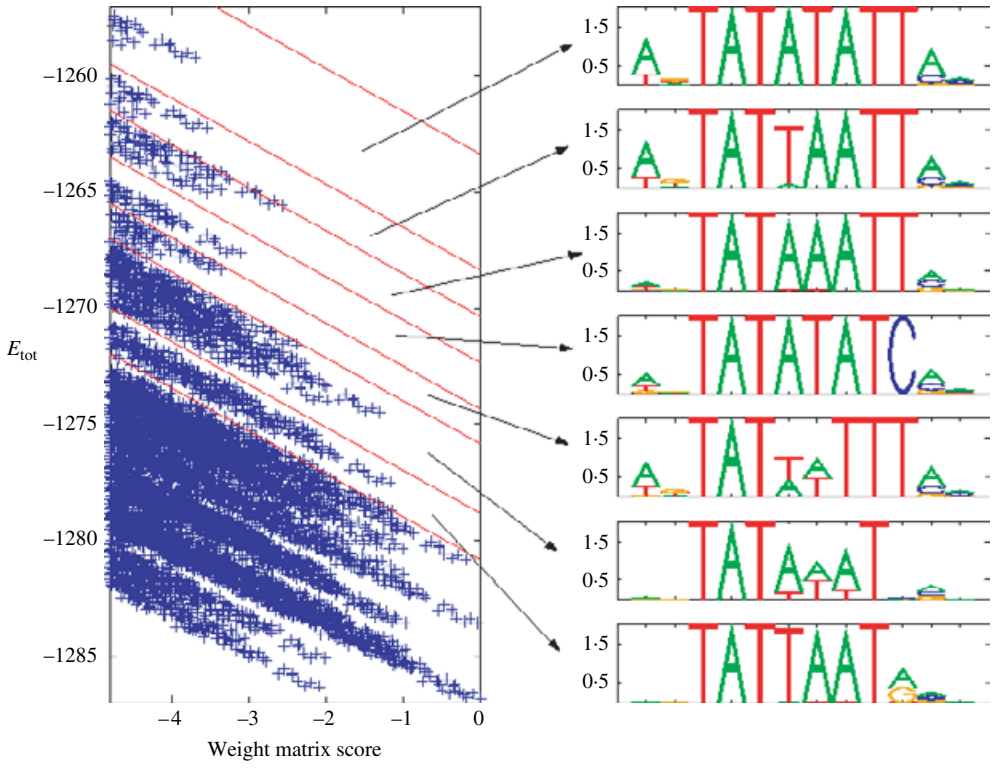


Fig. 4. Theoretical binding energy for TBP for all possible sequences plotted against the score of a mononucleotide weight matrix built on the basis of the sequences falling in the best 5 kcal mol^{-1} of the distribution. (Adapted from O'Flanagan *et al.* 2005.)

another fact about recognition mechanisms. For points above the diagonal ($L_{int} + L_{def} > L_{tot}$), direct and indirect mechanisms are working together to select the same optimal binding sequences. For points below the diagonal ($L_{int} + L_{def} < L_{tot}$), the final binding sequences are selected by a compromise between divergent sets of sequences selected by the direct and indirect mechanisms. We have termed these behaviours concordant and discordant recognition respectively. While most of the proteins we have studied so far belong to the first category, several are clearly in the second group (e.g. Trp repressor, CAP).

Our analysis of recognition lets us look at one other factor which is again difficult to analyse experimentally. This factor is correlation between the binding sequence preferences at different base-pair positions along a binding site. The direct view of recognition supposes that each base pair is recognized separately by one or more aa-base interactions. Correlations should, therefore, be absent. This view is reflected in the sequence logos used to illustrate binding preferences and also in the weight matrices used to search for sites, where the sequence score is a simple sum reflecting the sequence preferences of each successive base pair. On the other hand, the indirect view of recognition supposes that collective DNA deformation plays a significant role. In this case, correlations could easily occur. An obvious cause is that deformation would probably involve perturbations of base stacking and this, in turn, depends on the two adjacent base pairs which make up each step along the binding site. Although, experimental binding assays have pointed to the existence of correlation in a limited number of cases, the number of assays which

https://doi.org/10.1017/S0033583505004105 Published online by Cambridge University Press

must be carried out makes such studies difficult. This problem can, however, be addressed with our modelling approach.

Correlation can be detected by checking whether the sequence probabilities for multi-base-pair fragments along a binding site are equal to the sum of the probabilities of the individual base-pair positions. Figure 4 shows that this is not the case by plotting the scores generated using a conventional weight matrix (consequently ignoring correlation) against the calculated binding energies, ΔE_{tot}. Without correlation, one would expect the points to fall on a single diagonal. This is far from being the case and the points are seen to group into a series of diagonal distributions, each distribution corresponding to a different selection of bases in the centre of the binding site. With a detailed analysis (carried out in collaboration with R. A. O'Flanagan and A. M. Sengupta, BioMaPS Institute, Rutgers University), we have been able to show that most of the correlation is effectively due to base-pair stacking and, consequently, the problem can be almost completely solved by using weight matrices based on overlapping dinucleotide sequence probabilities, rather than on mononucleotide probabilities. A useful observation from this work is that correlation, when it occurs, is generally only important at a limited number of positions along the binding site. This implies that good results can be obtained with weight matrices which only include correlation at these positions (introducing only some dinucleotide terms), consequently reducing the number of experimental binding assays which would be necessary to obtain the necessary parameters.

As always, answering one set of questions, leads to new questions. One of the surprising aspects of these calculations was that even small DNA deformations can have a strong effect of sequence specificity. Thus far, it has not been possible to understand the molecular mechanism of this effect. Another problem yet to be solved is the role of water at the protein–DNA interface. This is a common feature of specific complexes and is an even more important problem for the low-specificity binding involved in 1D protein searches along the DNA double helix and also in nucleosome positioning. Protein adaptability also needs to be taken into account. The calculations discussed here do not allow for local (side-chain or loop rearrangement) or global repositioning of the protein with respect to DNA. This is a reasonable approximation when we have a good starting structure and we only modify DNA sequence. Mutating the protein can, however, lead to changes that affect not only the optimal DNA-binding sequence, but also the mechanism of recognition, as recent studies of a zinc-finger mutant have shown. Lastly, biology is not just based on binary complexes. An exciting challenge for the future is to understand how a group of proteins can cooperatively act on DNA to control processes as important for the cell as replication or transcription.

Further Reading

Lafontaine, I. & Lavery, R. (2000). Direct optimization of nucleic acid sequences. *Biophysical Journal* **79**, 680–685.

Lafontaine, I. & Lavery, R. (2001). ADAPT: a molecular mechanics approach for studying the structural properties of long DNA sequences. *Biopolymers* **56**, 293–310.

O'Flanagan, R. A., Paillard, G., Lavery, R. & Sengupta, A. M. (2005). Non-additivity in protein–DNA binding. *Bioinformatics* **21**, 2254–2263.

Paillard, G., Deremble, C. & Lavery, R. (2004). Looking into DNA recognition: zinc-finger binding. *Nucleic Acids Research* **32**, 6673–6682.

Paillard, G. & Lavery, R. (2004). Analyzing protein–DNA recognition mechanisms. *Structure* **12**, 113–122.

https://doi.org/10.1017/S0033583505004105 Published online by Cambridge University Press