Protein-DNA Recognition: New Perspectives and Underlying Themes

Peter H. von Hippel

In a Research Article in this issue of *Science*, Spolar and Record (*1*) present a detailed and quantitative explanation of the stability of specific regulatory complexes of proteins and DNA. This work comprises the first major effort to define the thermodynamics of protein-DNA recognition in structural terms. Its publication provides an appropriate occasion to take stock of the avenues that have led to this point in our understanding of these central regulatory interactions, to assess how these approaches fit together, and to ask how far we still have to go to reach a total molecular understanding.

The genetic makeup of every organism is encoded in its DNA, and the central problem that each faces is to express its genes as proteins, in correct amounts and with correct timing, relative to cellular and developmental cycles. The first level of this expression is transcription, and most of the protein-DNA complexes that Spolar and Record analyze are transcriptional activators of specific genes or classes of genes. Since the formulation of the operon hypothesis (*2*) and the establishment that these genetic regulators are proteins (*3*, *4*), attempts to gain a molecular understanding of the interactions of these proteins with the promoters, operators, and enhancers that comprise their regulatory targets has been central to molecular biology.

This problem has thermodynamic and kinetic components, which are now moving toward discrete structural and energetic explanations. However, early thinking focused more on specificity—asking how, in principle, a protein might find, recognize, and discriminate a particular sequence of DNA base pairs within a huge linear genome—than on the stability of the complexes themselves. Lacking a structure for any DNA-binding protein, the stability component was treated as a "black box" carrying a matrix of defined binding elements, and recognition was assumed to result from the docking of this matrix against complementary elements located on the DNA. In the absence of information to the contrary, the conformation of both partners was viewed as unchanged by the interaction, although this was recognized as an assumption that could be relaxed as further information was accumulated. The problem then developed along the following lines.

The author is in the Institute of Molecular Biology and Department of Chemistry, University of Oregon, Eugene, OR 97403, USA.

Information Content

It was early apparent that particular sequences of the four canonical Watson-Crick base pairs, assisted perhaps by unspecified flanking sequences, must be the coding elements that in combination specify functional binding sites for DNA regulatory proteins (*5*). One could then ask, using conditional probability or information theory approaches (*6*), how long a sequence of base pairs must be to form a site that reoccurs at random less than once per genome. For *Escherichia coli*, a defined sequence of at least 12 base pairs is required. Mutation and deletion analyses are consistent with such site sizes and have been used to probe the relative importance of different base pair loci within the target sequences (*7*). These data can then be further analyzed in attempts to link the information content, the evolution, and the biological activity of these regulatory sites (*6*).

Recognition

How can a protein recognize such a sequence of base pairs? Just as in specific DNA-DNA interactions, where Watson-Crick base-pairing provides the primary recognition motif, it was early appreciated that the central elements in the recognition of a particular DNA sequence by a protein must again be the hydrogen bond donors and acceptors of the base pairs, although here it is those projecting into the grooves of the double helix that must interact with the complementary recognition matrix within the protein binding site (*8*). Subsequent work showed that readout of these hydrogen bond–based recognition interactions could be indirect [involving specific water molecules as intermediates (*9*)], as well as direct, and could be facilitated by sequence-specific distortion of the DNA, the protein, or both to bring appropriate charges into register and generally to improve the physical (and thermodynamic) complementarity of the interacting protein and nucleic acid surfaces (*9*, *10*).

DNA Target Location and Discrimination

Of course it is not enough just to recognize the correct DNA site; the protein must also find it rapidly and bind to it sufficiently tightly to discriminate it from the millions of competing and overlapping nonspecific sites that are explored in the course of specific target location. A largely electrostatic nonspecific binding affinity, based on the displacement of condensed counterions from the DNA (*11*), permits rapid exploration of the DNA by protein sliding and intersegment transfer processes (*12*), while rapidly reversible conformational changes may permit switching between nonspecific and specific (and perhaps pseudospecific) protein binding modes during the exploration process (*9*).

Structure

Most of these ideas were developed long before anyone had ever "seen" a DNA-binding protein. Now a virtual explosion of elegant structural solutions has put molecular flesh on the above thermodynamic and kinetic bones. Modern x-ray diffraction and nuclear magnetic resonance (NMR) techniques, coupled with advances in protein overexpression and chemical synthesis of specific DNA sequences, have provided detailed molecular structures for dozens of important DNA regulatory proteins and protein-DNA complexes (*13*). These studies have defined a number of different types of protein recognition domains (helix-turn-helix, helix-loop-helix, zinc fingers, and so forth) that form the actual molecular platforms on which the protein components of the complementary recognition surfaces are positioned in space. This work has also revealed specific protein-protein interaction domains that can form homo- and heterodimers of protein subunits that are both varied enough and extensive enough to recognize a spectrum of specific DNA target sites.

Stability, Structure, and Conformational Change

This progress now permits us to ask again, but now within the constraints and opportunities of this new structural richness and diversity, for a detailed molecular understanding of the interactions that assemble and stabilize specific protein-DNA complexes. Clearly many aspects of earlier views were oversimplified and Spolar and Record (*1*) highlight situations where the quantitative approaches now available can lead to significant new insights. One striking feature revealed by recent structural studies of protein-DNA complexes is the extent to which the conformations of both partners may differ from those of the unbound species. These studies raise the question of the role of conformational change in protein-DNA recognition (*14*).

On the basis of the wealth of structural and binding data now available, Spolar and

Record (*1*) ask whether processes that couple conformational change to binding have a thermodynamic signature that corresponds to structural predictions. One early insight provided by Record and coworkers is that the formation of specific protein-DNA complexes, like other processes involving protein folding and assembly (*15*), is characterized by a large negative heat capacity change (*16*). Thus, thermodynamically, the problem of forming a specific protein-DNA complex is clearly related to that of identifying the forces and interactions that underlie and direct the folding and assembly of polypeptide chains into functional protein molecules and complexes.

Spolar and Record use the relations developed with proteins, in conjunction with thermodynamic and structural data and the quantitative "liquid hydrocarbon" model for the hydrophobic effect (*17*), to predict thermodynamic parameters for protein-DNA complexes. Their approach seems to work, and (subject to the key assumption that the more polar and highly charged nature of the nucleic acid surface does not severely perturb the outcome) their general conclusion that binding is coupled to (and drives) processes that bury additional nonpolar surfaces and are entropically costly appears to be well founded and relatively independent of the quantitative details of the argument.

Specificity and Stability

This brings us back to the continuing interplay, at progressively increasing levels of sophistication, of the concepts of specificity and stability. The Spolar and Record (*1*) analysis deals with the stability of specific protein-DNA complexes. Understanding the specificity with which regulatory proteins recognize their target sequences within the DNA genome requires that we also understand the structure and stability of nonspecific protein-DNA complexes, since it is the competition between these two types of complexes for the available protein that defines the specificity parameter. In thinking about what still needs to be learned, it is useful to contrast our now well-defined structural (and developing thermodynamic) view of specific protein-DNA complexes with what we know of nonspecific complexes, which are thought to be largely electrostatically stabilized, to retain most of the hydration properties of the individual partners and to be held together primarily by the displacement of condensed monovalent counterions from the DNA by the protein ligand (*11*). This description is consistent with the apparent lack of an anomalous heat capacity change associated with the formation of these complexes (*18*), but may require modification as more nonspecific [and pseudospecific (*10*)] protein-DNA complexes are investigated in the future.

Perspectives and Problems

Despite significant progress, many questions remain before we can consider our understanding of protein-DNA recognition to be complete. How, for example, can one really define the hydrophobic content of a nucleic acid surface and, for that matter, what is the molecular basis of defining the hydrophobicity of such surfaces in proteins? Can one really think of the protein-DNA interface of a nonspecific complex as retaining full hydration, and what role does the expulsion of water bound to polar groups at the interface play in stabilizing the specific complexes that form when the DNA target site is reached? Do conformational changes induced by DNA binding occur in nonspecific complexes as well and, if so and if nonspecific complex formation lies directly on the pathway to specific complex formation, how does one distribute the thermodynamic consequences of these effects in calculating binding specificity? And finally, of course, to what extent do these ideas actually apply within the cell?

These are problems for the future, and clearly those who work on them will not soon run out of things to do. The day is still far off when we can put the structure of a regulatory protein or protein complex into a computer together with that of a DNA sequence, massage the partners together, and ask for the relative free energy costs of forming specific and nonspecific complexes of every conceivable conformation. However, the Spolar and Record work gives us the beginnings of a data set to explore and calibrate such an approach, and thus brings a defined molecular understanding of protein-DNA recognition one step closer.

References and Notes

1. R. S. Spolar and M. T. Record Jr., *Science* **263**, 777 (1994).
2. F. Jacob and J. Monod, *J. Mol. Biol.* **3**, 318 (1961).
3. W. Gilbert and B. Muller-Hill, *Proc. Natl. Acad. Sci. U.S.A.* **56**, 1891 (1966).
4. M. Ptashne, *Nature* **214**, 232 (1967).
5. The use of the word "coding" in this context does not imply a 1:1 recognition of specific base pairs by specific amino acid residues or residue combinations; general codes of this type probably do not exist.
6. P. H. von Hippel, in *Biological Regulation and Development*, R. F. Goldberger, Ed. (Plenum, New York, 1979), vol. 1, pp. 279–347; T. D. Schneider, G. D. Stormo, L. Gold, A. Ehrenfeucht, *J. Mol. Biol.* **188**, 415 (1986); O. G. Berg and P. H. von Hippel, *ibid.* **193**, 723 (1987).
7. D. H. Ohlendorf *et al.*, *Nature* **298**, 718 (1982); D. K. Hawley and W. R. McClure, *Nucleic Acids Res.* **11**, 2237 (1983).
8. M. Yarus, *Annu. Rev. Biochem.* **38**, 841 (1969); P. H. von Hippel and J. D. McGhee, *ibid.* **41**, 231 (1972); N. C. Seeman, J.M. Rosenberg, A. Rich, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 804 (1976).
9. Z. Otwinowski *et al.*, *Nature* **335**, 321 (1988).
10. P. H. von Hippel and O. G. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1608 (1986); M. C. Mossing and M. T. Record Jr, *J. Mol. Biol.* **186**, 295 (1985).
11. M. T. Record Jr., T. M. Lohman, P. L. deHaseth, *J. Mol. Biol.* **107**, 145 (1976); P. L. deHaseth, T. M. Lohman, M. T. Record Jr., *Biochemistry* **16**, 4783 (1977); A. Revzin and P. H. von Hippel, *ibid.* **16**, 4769 (1977).
12. R. B. Winter, O. G. Berg, P. H. von Hippel, *Biochemistry* **20**, 6961 (1981); P. H. von Hippel and O.G. Berg, *J. Biol. Chem.* **264**, 675 (1989); H. Kabata *et al.*, *Science* **262**, 1561 (1993).
13. S. C. Harrison and A. K. Agarwal, *Annu. Rev. Biochem.* **59**, 933 (1990); C. O. Pabo and R. T. Sauer, *ibid.* **61**, 1053 (1992).
14. A. D. Frankel and P. S. Kim, *Cell* **65**, 717 (1991); T. Alber, *Curr. Biol.* **3**, 182 (1993).
15. J. M. Sturtevant, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 2236 (1977).
16. R. S. Spolar, J. H. Ha, M. T. Record Jr., *ibid.* **86**, 8382 (1989); J. H. Ha, R. S. Spolar, M. T. Record Jr., *J. Mol. Biol.* **209**, 801 (1989).
17. R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 2236 (1986).
18. Y. Takeda, P. D. Ross, C. P. Mudd, *ibid.* **89**, 8180 (1992).

19. I thank my many colleagues, both at the University of Oregon and elsewhere, for helpful discussions on these issues over the years. Preparation of this article was supported by USPHS research grants GM-15792 and GM-29158. P.H.v.H. is an American Cancer Society Research Professor of Chemistry.