Three-dimensional intricacies in protein-DNA recognition and transcriptional control

Stephen C Harrison

Transcription factors 'recognize' relatively short DNA consensus sequences; their full specificity must depend on a broader set of protein-protein and protein-DNA interactions. Joshi *et al.* show that, in addition to forming base pair–specific hydrogen bonds in the DNA major groove, certain Hox proteins detect DNA shape in the minor groove.

Patterns of gene expression that determine the spatial organization of multicellular organisms depend on combinations of a remarkably modest collection of transcriptional regulators. For example, transcription factors in the Hox family of homeodomain-containing proteins control developmental morphology along the anterior-posterior axis in both vertebrates and invertebrates. Yet *Drosophila melanogaster* has only eight Hox paralogs, each of which controls, with confusingly overlapping specificities, a large number of target genes[1]. The analysis in a recent paper by Joshi *et al.*[2] elegantly sorts out one such set of ambiguities, by combining X-ray crystallography and genetics to determine how the DNA base sequence between the site for a particular Hox paralog and the neighboring site for a DNA-binding cofactor influences developmental outcome. The resolution depends on a previously detected 'invasion' of the cofactor, Extradenticle (Exd), by an N-terminal extension of the Hox protein. Joshi *et al.*[2] show that one particular sequence of 3–4 nucleotides between the two sites allows the invading linker to settle comfortably into the DNA minor groove and to establish stabilizing contacts. This intervening DNA sequence is present in a site upstream of the *fork head* gene (*fkh*), but not in a consensus Hox-Exd target sequence.

The homeodomain is a simple, three-helix structure with an N-terminal extension ('arm'). The third ('recognition') helix lies in the DNA major groove, facing a sequence that in many sites conforms to a consensus TAAT (or ATTA, depending on which strand is read). Hydrogen bonds and van der Waals contacts with residues at three positions in the recognition helix can specify bases xxATNN, and a minor-groove insertion of the N-terminal arm allows an arginine side chain to specify T or A at the initial position in the TAAT[3–5]. But how can so slight an array of interactions effect the multitude of site distinctions required for the ensemble of homeodomains in any organism? One part of the answer is that most homeodomain-containing proteins interact with other DNA-binding proteins, which in turn have specific, adjacent sites. For example, Mat-α2 and Mat-a1 of budding yeast bind cooperatively to sites on DNA spaced about a turn apart in such a way that a C-terminal extension of Mat-α2 folds against the homeodomain of Mat-a1, forcing a bend in the DNA[6]; the *Drosophila* Hox protein, Ubx, and the somewhat atypical homeodomain of Exd interact through a linker N-terminal to the Ubx homeodomain[7].

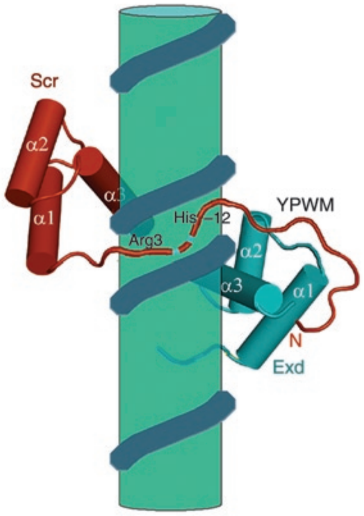

Figure 1 Cooperative binding of the Scr and Exd homeodomains to a site that regulates *fkh*. A YPWM motif in the N-terminal extension of Scr covers a hydrophobic patch on the surface of Exd. His –12 and Arg3, in the segment between YPWM and the Scr homeodomain proper, insert into the unusually narrow minor groove of the authentic *fkh*^{250} site, but fail to insert into the somewhat wider minor groove of a consensus site. The three α-helices of the homeodomains are labeled; α3, the 'recognition helix', lies in the major groove. Modified from Joshi *et al.*[2].

The Sex combs reduced gene product, Scr, is the only Hox paralog in *Drosophila* that can initiate salivary-gland development, although several paralogs can repress a critical antenna-specifying gene (see references in Joshi *et al.*[2]). In the former case, Scr binds together with Exd; in the latter, no cofactor seems to be required. Previously determined structures in which a Hox protein and Exd (or Pbx, the mammalian Exd ortholog) are both present in a DNA complex have not, however, shown a clear reason why the Exd interaction should be specific for Scr[7,8]. That is, the Hox-Exd interaction involves residues common to most Hox proteins (including

*Stephen C. Harrison is in the Departments of Biological Chemistry and Molecular Pharmacology and of Pediatrics, Harvard Medical School, and is an investigator of the Howard Hughes Medical Institute, Boston, Massachusetts 02115, USA.
e-mail: harrison@crystal.harvard.edu*

Ubx, as mentioned in the preceding paragraph), and the Hox site itself is too short to distinguish among paralogs. Joshi *et al.*[2] have now compared structures of Scr and Exd bound with two closely related DNA fragments—one (designated *fkh*^{250}) having the base sequence of an authentic site upstream of the *fkh* gene and the other having a 'consensus' Hox-Exd sequence. In both complexes, the so-called 'linker' segment of Scr, a region N-terminal to the 'arm' described above, extends into an intimate contact with Exd (**Fig. 1**). A YPWM motif (residues –19 to –16, where position 1 is the conventional first residue in the homeodomain) covers an exposed hydrophobic patch on the Exd homeodomain. In the specific complex, however, more residues in the arm and in the linker are ordered than in the consensus complex. These include an arginine in the arm at position 3 in the standard homeodomain numbering and a histidine in the linker at position –12 (that is, 12 positions N-terminal to the first residue of the homeodomain as conventionally defined); most of the intervening residues are still disordered (could they attract yet another protein *in vivo*?). Arg3 and His –12 dip into the DNA minor groove, between the Scr and Exd contacts, over a stretch in which the minor groove is exceptionally narrow—narrower than the corresponding groove in the complexes on consensus sites. Neither side chain has direct hydrogen-bonding interactions with bases, but Joshi *et al.*[2] propose that they are, in effect, recognizing the DNA conformation.

To show that His –12 and Arg3 are indeed relevant, both to DNA binding *in vitro* and to gene regulation *in vivo*, Joshi *et al.*[2] mutated the two residues singly and in combination. These changes reduced binding to the specific site and produced effects on the cuticle pattern in transgenic *Drosophila* misexpressing wild-type and mutant Scr that were consonant with expectations from the proposed mechanism. Moreover, activation of LacZ synthesis under control of the specific *fkh* regulatory site was compromised in the double mutant. Thus, the insertion of His –12 and Arg3 into a narrow minor groove seems to enhance specificity, as predicted. The electrostatic potential of a narrowed groove is particularly negative, and the arginine-histidine pair provides a robustly anchored, positively charged probe. The two side chains appear to be linked by a hydrogen bond, which will have the effect of buttressing the arginine without neutralizing it. Indeed, mutation of the (presumably unprotonated) histidine has a less marked effect on binding and phenotype than does mutation of the arginine.

At least one of the collaborators should not have been surprised by the way complementarity to minor-groove conformation modulates Scr specificity. Aggarwal *et al.*[9] showed nearly 20 years ago how an essentially similar mechanism governs the genetic switch in bacteriophage 434. In that case, contacts between dyad-related helix-turn-helix modules enforce a bend that narrows the minor groove, into which inserts an arginine side chain from a loop just C-terminal to the recognition helix. As in the present example of two homeodomains, inspection of the protein-DNA structures alone does not distinguish between a preformed narrow groove and a groove that can be narrowed more readily—that is, between recognition of DNA conformation or of DNA conformability. Computational modeling carried out by Joshi *et al.*[2] suggests that the minor groove of the *fkh*^{250} site may be stably narrow. Moreover, ApA, TpT and ApT DNA base steps generally favor narrow minor grooves, because strong negative propeller twists stabilize interbase hydrogen bonds in the major groove, whereas this does not happen with TpA. The specific *fkh*^{250} sequence (5′-AGATTAATCG) contains a TpA step between the position at which Arg5 inserts (an almost universal homeodomain interaction) and the position where Arg3 and His –12 project; the consensus fkh sequence (5′-TGATTTATGG) contains a TpA just opposite the (disordered) Arg3/His –12 pair.

There was once a view that base sequence–specific DNA-binding proteins might have relatively simple 'read-out' properties, and much speculation went into considering potential transcription-factor 'codes'. Early studies of phage repressors more or less dispelled such notions (although they have popped up from time to time since), and true 'altered-specificity' mutants have proved exceptionally hard to isolate or design. Various recently determined examples of multiple proteins bound to DNA—for example, the interferon-β enhanceosome[10]—should help dispel the further illusion that the logic of transcriptional regulation can be extracted from genomic sequences by matching consensus transcription-factor sites. The actual encoding of regulatory information is far more intricate, and contingent specificities such as the one analyzed by Joshi *et al.*[2] are probably common features. The conformational characteristics of DNA and of its protein partners do far more than merely create a framework within which projecting side chains interrogate base-pair functional groups.

1. Pearson, J.C., Lemons, D. & McGinnis, W. *Nat. Rev. Genet.* **6**, 893–904 (2005).
2. Joshi, R. *et al. Cell* **131**, 530–543 (2007).
3. Qian, Y.Q. *et al. Cell* **59**, 573–580 (1989).
4. Kissinger, C.R., Liu, B.S., Martin-Blanco, E., Kornberg, T.B. & Pabo, C.O. *Cell* **63**, 579–590 (1990).
5. Gehring, W.J. *et al. Cell* **78**, 211–223 (1994).
6. Li, T., Stark, M.R., Johnson, A.D. & Wolberger, C. *Science* **270**, 262–269 (1995).
7. Passner, J.M., Ryoo, H.D., Shen, L., Mann, R.S. & Aggarwal, A.K. *Nature* **397**, 714–719 (1999).
8. Piper, D.E., Batchelor, A.H., Chang, C.P., Cleary, M.L. & Wolberger, C. *Cell* **96**, 587–597 (1999).
9. Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M. & Harrison, S.C. *Science* **242**, 899–907 (1988).
10. Panne, D., Maniatis, T. & Harrison, S.C. *Cell* **129**, 1111–1123 (2007).