

A Lexicon for Homeodomain-DNA Recognition

Markus Affolter,^{1,*} Matthew Slattery,² and Richard S. Mann^{2,*}

¹Growth and Development, Biozentrum der Universität Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

²Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West 168th Street HHSC 1104, New York, NY 10032, USA

*Correspondence: markus.affolter@unibas.ch (M.A.), rsm10@columbia.edu (R.S.M.)

DOI 10.1016/j.cell.2008.06.008

Decoding the *cis*-regulatory logic of eukaryotic genomes requires knowledge of the DNA-binding specificities of all transcription factors. New work (Berger et al., 2008; Noyes et al., 2008) provides individual specificities for nearly all *Drosophila* and mouse homeodomains, key DNA-binding domains in many transcription factors. The data underscore the complexity of determining target specificities *in vivo*.

The homeodomain is among the most widespread DNA-binding domains in eukaryotes and is a key component of many transcription factors. The homeodomain structure is composed of a bundle of three α helices, loops joining each helix, and an unstructured N-terminal arm. The N-terminal arm sits in the minor groove of DNA, whereas the third α helix makes hydrogen bonds in the major groove (Gehring et al., 1994). With this structural template, the homeodomain superfamily has evolved to carry out an enormous array of biological functions ranging from mating type specification in yeast to the maintenance of pluripotency in human embryonic stem cells. The vast majority of these diverse functions depend on specific DNA binding. Yet herein lies the rub: All analyses of the DNA-binding specificities of homeodomains indicate that even those that carry out very different *in vivo* functions bind to very similar DNA sequences *in vitro*. How then is target specificity achieved?

Two papers in this issue (Berger et al., 2008; Noyes et al., 2008) tackle this question by determining the *in vitro* DNA-binding specificities for most fly and mouse homeodomains, a collection that presumably represents the majority of the homeodomains in all multicellular animals. The two studies use different high-throughput protocols. Noyes et al. (2008) analyzed the binding preferences of 84 homeodomains from the fly *Drosophila melanogaster* using a bacterial one-hybrid system that allows the specificities of any DNA-binding domain to be rapidly characterized in *Escherichia coli* (Noyes et al., 2008). Berger et al. (2008) characterized the binding preferences of 168 mouse home-

odomains using protein binding microarrays, which measure binding to double-stranded oligonucleotides spotted on a microarray platform (see also Warren et al., 2006). Such complete surveys of DNA binding preferences of large families of DNA-binding domains provide an impressive precedent for the analysis of other DNA-binding domains in the future.

Despite their very different methods, the two studies arrive at broadly similar conclusions about homeodomain binding specificities. Homeodomains that fall into distinct classes based on their amino acid sequences typically bind to distinct target sites. Both groups arrive at similar classifications on the basis of binding site preferences (Figure 1). Although this is not surprising, the computer-assisted analyses of the huge data sets from these large-scale approaches provide us with the *in vitro* binding site preferences of nearly all known homeodomains. These data reinforce and extend previous observations suggesting that DNA recognition by homeodomains is guided by a complex set of protein-DNA interactions that dictate distinct binding specificities. Both studies confirm the importance of the N-terminal arm and third α helix in DNA recognition and provide some general rules for relating homeodomain sequences to DNA recognition. There are also some noteworthy differences in the two sets of data. For example, the dominant DNA binding site reported by Noyes et al. for the Iroquois homeodomain class is taACA (uppercase indicates more important positions), whereas Berger et al. report the related symmetric site ACATGT (Figure 1). Similar differences were obtained for the Ladybird group of

homeodomains. These differences may be because protein binding microarrays cannot readily distinguish binding orientation or binding by monomers versus dimers. The design of the bacterial one-hybrid method lends itself to identifying monomeric binding sites and enabled Noyes et al. to make more definitive conclusions about the contributions to specificity of the N-terminal arm versus the third α helix. Other differences in the data sets, such as the site preference of the AbdB homeodomain family (Figure 1), will only be resolved by future studies.

The results also show that residues outside of the protein-DNA binding interface can influence DNA recognition. For example, Noyes et al. suggest that in the Iroquois family of homeodomains, intramolecular interactions influence the position and consequently the recognition properties of the N-terminal arm. Findings such as these illustrate that there is unlikely to be a simple relationship between homeodomain sequence and binding site preference, but nevertheless, they provide an invaluable resource for the eventual understanding of the biophysical basis of homeodomain-DNA recognition. Although the answer to the specificity problem is likely to be complex, these studies also have the practical benefit of allowing researchers to generate new “change of specificity” mutants, some of which may even succeed at swapping specificities *in vivo*.

The two studies also highlight the problem of defining homeodomain specificities. Even though both Noyes et al. and Berger et al. classify most of the homeodomains into distinct groups or subgroups on the basis of DNA-binding specificity, an examination of these

groupings shows that they are actually not very distinct (Figure 1). For example, homeodomains encoded by the Antennapedia (Antp) and Engrailed (En) sub-families (comprising approximately half of the fly homeodomains, and 14 of the 40 Berger et al. subgroups) all show a strong preference for the sequence TAATTA. Yet, despite this common binding site, the range of in vivo functions carried out by this subset of homeodomain proteins is staggering. Both studies point out that if the low-affinity sites identified by these approaches are carefully analyzed, the specificities of even very similar homeodomains can be distinguished. Although true, the results do not provide an explanation for how proteins with similar homeodomains would choose these lower-affinity sites instead of higher-affinity ones in vivo. For example, Berger et al. report that although Lhx2 and Lhx4 both bind to TAATTA with high affinity, they show different preferences for lower-affinity sites—Lhx2 prefers TAACGA, whereas Lhx4 prefers TAATCA (Figure 1). Further, because both of these sequences are likely to exist in multiple copies in most genes, they alone cannot be sufficient to account for specificity.

Although it is remarkable that binding site preferences can be discerned with the reductionist approaches of these two studies, transcriptional activation in vivo is more than a simple binary interaction between a homeodomain and DNA. Indeed, monomeric homeodomain binding sites are not sufficient for in vivo function (for example, see Galant et al. 2002). Moreover, transcriptional regulation requires intricate enhancer-directed assembly of multiprotein complexes that contain many protein-protein interactions in addition to protein-DNA interactions (Levine and Tjian, 2003). There are many examples of cooperative DNA binding by homeodomain proteins and other factors (for example, see Berkes et al., 2004; Mann and Affolter, 1998; Moens and Selleri, 2006). Amino acids flanking the homeodomain can also influence DNA-recognition properties, in part through interactions with cofactors that in turn alter how homeodomains bind to DNA (for example, see Joshi et al., 2007). It is possible that the subtle differences in DNA-binding preferences revealed by the Noyes et al. and Berger et al. studies hint at other mechanisms of












<i>Drosophila</i> groups	Homologous mouse group(s)	
	Dominant motifs	Subgroup motifs
<p>Antp Group (18)</p>  <p>En Group (25)</p> 	<p>Group F (16, HoxA3) nTAATTAn</p> <p>Group A (26, Lhx1) TTAATTAA</p> <p>Group D (2, Nkx6-1) ATTAATTA</p> <p>Group E (7, HoxA7) tTAATTAA</p>	<p>yTAACGAc (9, Hoxa3) cTAATTAC (1, Hoxa2) cTAACGAg (2, Evx2) nnTCATCA (1, Meox1) TAAACGGT (2, Vax1) CGGAaaaAA (1, Gsh2) TAACGAgc (2, Lhx2) ayTAATCA (4, Lhx4) TAATTCGC (12, Phox2a) TAATTAaA (2, Lhx5) CAATAAAA (3, Msx3) cTCAATCA (4, En2) ATAATsrs (2, Nkx6-1) hTAAkrrr (1, Isl2) CTAATTAG (2, Lhx8)</p>
<p>Abd-B Group (5)</p> 	<p>Group J (12, HoxA11) GTCGTAAA</p> <p>Group D (2, Dbx2) atTAATta</p>	<p>gTTGTAAg (2, Cdx2) TCGTTAA (10, Hoxa11) AAATCGAT (2, Dbx2) TAATTraT (1, Hlx1)</p>
<p>Bar Group (6)</p> 	<p>Group C (2, Barh1) yTAATTGg</p> <p>Group H (2, Hmx1) TTAATTGc</p>	<p>cTAATsGG (2, Barh1) cTAATTAg (2, Hmx1) aaTTAATa (1, Tlx2)</p>
<p>NK-1 Group (5)</p> 	<p>Group A (7, Dlx3) aTAATTay</p> <p>Group B (1, Barx2) nTAATTRn</p>	<p>rTAATtgC (5, Dlx3) cTAATgGG (2, Nkx1-1) TAACGAgY (1, Bsx)</p>
<p>Tgif/Exd Group (4)</p> 	<p>Group R (4, Meis1) TGACAGsT</p>	<p>GACAAggy (3, Mrg2) TGACgtaw (1, Tgif1) wtTGATgn (1, Pbx1)</p>
<p>Bcd Group (4)</p> 	<p>Group M (7, Pitx1) tTAATCCc</p>	<p>yTAAGCCh (3, Pitx1) yTAAKcCg (4, Crx)</p>
<p>NK-2 Group (3)</p> 	<p>Group O (8, Nkx2-2) TCAAGTGG</p>	<p>ACTAAGTG (2, Nkx3-1) TTCRAGTG (6, Nkx2-2)</p>
<p>Six Group (3)</p> 	<p>Group P (4, Six1) TGATACCc</p>	<p>none</p>
<p>Iroquois Group (3)</p> 	<p>Group Q (3, Irx2) TACATGTW</p>	<p>none</p>
<p>Ladybird Group (2)</p> 	<p>Group Y (1, Lbx2) yTAATTAr</p>	<p>TAACrAGg (1, Lbx2)</p>

Figure 1. Classification of Homeodomain-DNA-Binding Specificities

Left column: The 11 homeodomain specificity groups in the fruit fly *Drosophila melanogaster*. Middle column: The specificity groups and corresponding high-affinity binding sites identified for those mouse homeodomains that are most closely related to the *Drosophila* proteins in the same row. Right column: The representative low-affinity binding sites for the mouse specificity subgroups. The numbers of proteins in each group and a representative member are shown in parentheses. The figure was generated by a nearest-neighbor homology algorithm (Berger et al. 2008) with data from Noyes et al. (2008) and Berger et al. (2008). Pairings with a nearest neighbor distance greater than 3 were not used. In the one case where a mouse subgroup fell into multiple *Drosophila* groups, priority was given to the pairing with the lowest nearest-neighbor distance. *Drosophila* DNA motif logos are adapted from Noyes et al. (2008). Representative mouse DNA motifs are adapted from motif logos in Berger et al. (2008). Lowercase letters represent positions with low information content, and uppercase letters represent position of higher information content. Grey letters indicate degenerate sequence: n = any base; y = C or T; s = C or G; r = A or G; k = T or G; w = T or A; and h = A or T or C.

DNA recognition that may be amplified or modified by protein-protein interactions in vivo. Such ideas can be tested by assaying combinations of interacting proteins with bacterial one-hybrid experiments and protein-binding microarrays. Further, DNA-binding specificity is only one step in the complex process of transcription regulation. Thus, understanding how this large and important superfamily of DNA-binding homeoproteins ultimately functions in vivo is still a work in progress, but one that now benefits from the first edition of an almost unabridged dictionary of homeodomain-DNA-binding specificities.

REFERENCES

- Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). *Cell*, this issue.
- Berkes, C.A., Bergstrom, D.A., Penn, B.H., Seaver, K.J., Knoepfler, P.S., and Tabscott, S.J. (2004). *Mol. Cell* 14, 465–477.
- Galant, R., Walsh, C.M., and Carroll, S.B. (2002). *Development* 129, 3115–3126.
- Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G., and Wuthrich, K. (1994). *Cell* 78, 211–223.
- Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007). *Cell* 131, 530–543.
- Levine, M., and Tjian, R. (2003). *Nature* 424, 147–151.
- Mann, R.S., and Affolter, M. (1998). *Curr. Opin. Genet. Dev.* 8, 423–429.
- Moens, C.B., and Selleri, L. (2006). *Dev. Biol.* 291, 193–206.
- Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). *Cell*, this issue.
- Warren, C.L., Kratochvil, N.C., Hauschild, K.E., Foister, S., Brezinski, M.L., Dervan, P.B., Phillips, G.N., and Ansari, A.Z. (2006). *Proc. Natl. Acad. Sci. USA* 103, 867–872.

Forging New Ties between *E. coli* Genes

Trey Ideker^{1,*}

¹Department of Bioengineering, University of California San Diego, La Jolla, CA 92093, USA

*Correspondence: trey@bioeng.ucsd.edu

DOI 10.1016/j.cell.2008.06.003

A recent study in *Nature* (Isalan et al., 2008) has examined the effects of systematically adding new transcriptional interactions in the bacterium *Escherichia coli*. Surprisingly, the majority of the engineered connections have no effect on growth, and in some cases the new connections enhance fitness. These findings reveal insights into the robustness and evolvability of gene networks.

Systems biology measures the response of biological networks to systematic perturbations. In many cases, the perturbations involve direct targeting of genes and proteins using knockouts or knockdowns, protein overexpression constructs, or natural variation in the form of single nucleotide polymorphisms (Beyer et al., 2007). Recent work by Isalan et al. (2008) published in *Nature* takes the concept of systematic perturbation to a new level. In this study, the alterations are not to genes or proteins themselves, but to gene and protein interactions. Although the effects of adding or removing interactions have been studied before—for instance using reverse one-hybrid or reverse two-hybrid assays (Vidal et al., 1996)—Isalan et al. alter the transcriptional networks of the bacterium *Escherichia coli* on an unprecedented scale.

Their experimental design is as follows. Any transcriptional regulatory network can be viewed as the superposition of two types of interactions: the set of interactions among transcription factors in which “regulators regulate regulators” (Simon et al., 2001) and the set of interactions connecting transcription factors to downstream responder genes. It is this first network of “regulators regulating regulators” that creates interesting network structures and dynamics such as feed-forward and feed-back loops.

The goal of Isalan et al. was to perturb this network by adding each possible regulatory connection between a pair of transcription factors (Figure 1A). To add a new connection from a given factor A to factor B, the DNA promoter targeted by factor A was placed immediately upstream of the open reading frame (ORF) encoding factor B. All pairwise

combinations (A,B) were considered within a set of 22 *E. coli* transcription factors, which were chosen to represent a range of general and specific regulatory functions. Each rearrangement was introduced into *E. coli* cells on plasmids and, in some cases, also by direct insertion into the genome. Plasmids also encoded a green fluorescent protein (GFP) so that the transcriptional output could be measured in addition to the overall impact on the growth rate of the organism. Note that each transcription factor gene was also left in its original genomic location, such that the net effect was merely to add interactions to the natural network, not to take any away.

What might be the possible consequences of adding a new regulatory input to a given transcription factor? As it turns out, there are at least two. Most simply, it is likely that the expression of the fac-