Chemistry & Biology, Vol. 11, July, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.chembiol.2004.07.005

Mapping Key Elements of a Protein Motif

In this issue of *Chemistry & Biology*, Weiss and colleagues use phage display to map residues in the engrailed homeodomain that influence DNA recognition. Their shotgun scanning data provides surprising new insights into the importance of regions outside the recognition helix and N-terminal arm for DNA binding.

How can we define the sequence elements within a protein domain that are critical for its function? With the wealth of genomic information that has become available and tools to mine this data, sequence alignments of orthologs and paralogs from distant species provide a signature (consensus) sequence that gives clues about residues that are critical for function. Supplemented with structural information of the domain, or better yet in complex with a partner of interest (protein, ligand, or nucleic acid), focused hypotheses about the function of residues at various positions can be readily developed. Ultimately though, these hypotheses must be directly tested by mutagenic interrogation of the domain to define the residues that are critical for a particular function.

Dissecting the importance of each residue in domain function has, until recently, been an arduous task of introducing single point mutations throughout a protein domain and then testing these proteins individually for function [1]. Typically, alanine substitutions are introduced to truncate the side chain at the β-carbon, which allows the importance of each side chain in domain function (whether positive or negative) to be probed. Alanine scanning experiments on complexes such as hGH/hGHbp demonstrated the power of this approach for defining regions of a domain that are critical for function [2].

An important shortcut for mapping the critical features of a domain was provided by the advent of combinatorial library approaches that allowed the fitness of a population of mutant proteins to be examined [3, 4]. Instead of making and testing individual mutants, a single library containing a distribution of alanine mutations throughout a region of the protein is tested. The original combinatorial approaches were performed in vivo as some form of genetic selection. However, this cellular requirement creates inherent limitations in the characteristics of the proteins that can be assayed.

In groundbreaking work in 2000, Weiss et al. demonstrated that combinatorial libraries of mutant proteins could be examined as a pool in vitro using phage display (alanine shotgun scanning) [5]. This format provides important advantages over in vivo assays: the ability to define the assay conditions and the nature of the functional readout (protein-protein interaction, protein stability, etc.) and the ability to simultaneously examine large libraries of variants (~10^{10}) in a few rounds of selection. Moreover, if the experiment is performed under equilibrium conditions, the ratio of the wild-type residue to alanine at each position can provide an energetic estimate for the cost of the mutation at a protein-ligand interface [5–7]. Because of the high-throughput nature of phage experiments, multiple libraries that bound neighboring sequences can be used to scan an entire protein domain in blocks; allowing all of the residues that are critical for function to be rapidly mapped in parallel (in weeks) [8, 9].

Weiss and colleagues continue to advance this methodology in this issue of *Chemistry & Biology* [10] by applying two different forms of shotgun scanning (alanine and homolog) to identify elements in the engrailed homeodomain that are critical for sequence-specific DNA recognition. In their alanine scanning experiments, two different libraries, each covering 15 amino acids, mapped residues important for DNA binding around and within helix 1 or helix 2 of the homeodomain (see Figure 1). These two libraries allow half of the residues in this domain to be scanned. The majority of residues in this region of the protein do not directly contact the DNA; they position the recognition elements (helix 3 and the N-terminal arm) for interaction with the major and minor grooves of DNA, respectively [11]. (Some residues that are directly involved in DNA recognition have been defined by previous mutagenesis studies [12, 13].) Consequently, mutations in helices 1 and 2 would be expected to have an indirect affect on DNA binding. These substitutions should reveal residues that help to organize amino acids that directly interact with the DNA (second sphere effects) or that affect DNA binding by altering protein stability.

One of the most exciting aspects of the results from the alanine shotgun scanning experiment is the correlation between the residues most resistant to change and the residues that are most highly conserved among the extended HOX class of homeodomains [14], which includes engrailed. There is a complete correlation between every position at which an amino acid is absolutely conserved within this class of homeodomains and a strong preference for the wild-type residue over alanine in the shotgun scanning data (Figure 1). This result is not necessarily shocking, as the consensus sequence for these proteins would be expected to highlight sequence preferences that are common to their function (i.e., DNA binding), but this correlation demonstrates the power of alanine shotgun scanning to identify key functional elements of a protein domain. The majority of residues that displayed the largest bias against mutation to alanine (F20, L34, L38, L40, I45) pack into the hydrophobic core of the protein. Other key residues that are also resistant to substitution (Y25, R31, K46) interact either directly or via ordered waters with the phosphate backbone [11].

To identify features of each residue that are critical for function, the authors employed homolog shotgun scanning. Like the alanine shotgun experiment, homolog shotgun scanning is performed in a combinatorial manner, where one or a few conservative changes (e.g., F to Y, E to D or Q) are introduced at each position through-

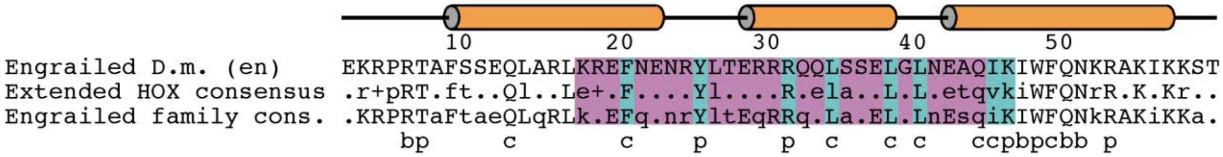
10
20
30
40
50
Engrailed D.m. (en) EKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKST
Extended HOX consensus .r+pRT.ft..Ql..Le+.F....Yl....R.ela..L.L.etqvkiWFQNrR.K.Kr..
Engrailed family cons. .KRPRTaFtaeQLqRLk.EFq.nrYltEqRRq.La.EL.LnEsqiKIWFQNkRAKiKKa.
bp c c p p c c c ccpbpcbb p

Figure 1. Sequence Alignment of the Engrailed Homeodomain with Two Consensus Sequences

Sequence alignment between the engrailed homeodomain from *Drosophila melanogaster* reported in this study [10], the consensus sequence from the human extended HOX homeodomains [14], and the consensus sequence from engrailed orthologs/paralogs in the homeodomain resource database [18] (research.nhgri.nih.gov/homeodomain/). Capital letters and lower case letters in the consensus sequences represent 100% and 60% conservation of a residue, respectively. A plus sign indicates conservation of a positively charged residue. Cylinders above the sequences indicate the location of the three α helices in the homeodomain fold. The region that was mapped by alanine shotgun scanning is denoted in magenta, with the blue bars indicating those positions within this region where the recovered sequences displayed a wild-type to alanine ratio that was greater than 8. Information about the participation of residues in protein-DNA recognition [11] or their burial in the core of the fold [19] is indicated below the sequences: b indicates participation in a base contact, p indicates a phosphate contact (either direct or water mediated), and c indicates residues in the protein core.

out a block of residues in the protein. These more subtle substitutions provide information about the importance of distinct features of each side chain (e.g., the presence of a negative charge). Interestingly, many of the residues that compose the hydrophobic core in engrailed can tolerate, and in some cases modestly prefer, substitution by a residue of similar composition and size. Other positions, such as F20, are intolerant to even a modest change (Y), which is consistent with the absolute conservation of this amino acid within this class of proteins (Figure 1).

The homolog scanning result at position 25 is particularly intriguing. Tyrosine and phenylalanine appear to be completely interchangeable at this position; thus, the hydrogen bond observed between the tyrosine hydroxyl and the phosphate backbone in the protein-DNA complex [11] is not the critical characteristic required for function. Nonetheless, tyrosine is highly to absolutely conserved at this position among human HOX, extended HOX, NK, and paired homeodomains [14]. Moreover, Y25 is absolutely conserved among engrailed orthologs/paralogs from flies to humans, which spans hundreds of millions of years of evolutionary separation (Figure 1). Thus, the presence of the hydroxyl group appears to be tightly linked to domain function. One potential explanation is as follows: the tyrosine hydroxyl serves as a site for phosphorylation that functions as a switch to regulate (directly or indirectly) DNA binding. DNA binding by certain Cys_2His_2 zinc finger proteins has been shown to be regulated by phosphorylation of a conserved linker sequence between the fingers [15, 16]. Consistent with this hypothesis, phosphorylation of tyrosines within the homeodomain of HOX10A was recently shown to modulate its DNA binding activity [17].

Homeodomains are an extremely important class of transcription factors that are present in species from yeast to humans. In higher eukaryotes, these proteins are important regulators of early embryonic patterning and cellular differentiation. As a consequence, missense mutations in homeodomains are associated with many types of diseases. As the authors note, their shotgun scanning data explain how many identified missense mutations outside the primary DNA recognition residues could impact function by affecting residues that scaffold the position of the recognition helix. The results of Weiss and colleagues also identify previously unappreciated features of the domain, such as the plasticity of the hydrophobic core. An additional intriguing result is the identification of substitutions within this core that actually result in improved DNA binding. Why such apparently beneficial substitutions would not be incorporated into gene remains unanswered, but it may result from the fact that gene regulation is typically not the act of an independent protein; instead, gene regulation is usually performed in the context of cooperative interactions with other DNA binding partners. Overall, these experiments provide a deeper understanding of DNA recognition by the homeodomain while presenting many exciting new questions. Finally, the implications of this study for protein design and engineering should not be overlooked. The ability to rapidly define the core functional sequence of a particular domain provides an important foundation for future engineering efforts to alter its specificity, affinity, or function.

Scot A. Wolfe
Program in Gene Function and Expression
Department of Biochemistry and Molecular
Pharmacology
University of Massachusetts Medical School
Worcester, Massachusetts 01605

Selected Reading

1. Wells, J.A. (1991). Methods Enzymol. *202*, 390–411.
2. Cunningham, B.C., and Wells, J.A. (1993). J. Mol. Biol. *234*, 554–563.
3. Gregoret, L.M., and Sauer, R.T. (1993). Proc. Natl. Acad. Sci. USA *90*, 4246–4250.
4. Chatellier, J., Mazza, A., Brousseau, R., and Vernet, T. (1995). Anal. Biochem. *229*, 282–290.
5. Weiss, G.A., Watanabe, C.K., Zhong, A., Goddard, A., and Sidhu, S.S. (2000). Proc. Natl. Acad. Sci. USA *97*, 8950–8954.
6. Vajdos, F.F., Adams, C.W., Breece, T.N., Presta, L.G., de Vos, A.M., and Sidhu, S.S. (2002). J. Mol. Biol. *320*, 415–428.
7. Distefano, M.D., Zhong, A., and Cochran, A.G. (2002). J. Mol. Biol. *322*, 179–188.
8. Morrison, K.L., and Weiss, G.A. (2001). Curr. Opin. Chem. Biol. *5*, 302–307.
9. Sidhu, S.S., Fairbrother, W.J., and Deshayes, K. (2003). Chembiochem *4*, 14–25.
10. Sato, K., Simon, M.D., Levin, A.M., Shokat, K.M., and Weiss, G.A. (2004). Chem. Biol. *11*, this issue, 1017–1023.
11. Fraenkel, E., Rould, M.A., Chambers, K.A., and Pabo, C.O. (1998). J. Mol. Biol. *284*, 351–361.

12. Ades, S.E., and Sauer, R.T. (1995). Biochemistry *34*, 14601–14608.
13. Connolly, J.P., Augustine, J.G., and Francklyn, C. (1999). Nucleic Acids Res. *27*, 1182–1189.
14. Banerjee-Basu, S., and Baxevanis, A.D. (2001). Nucleic Acids Res. *29*, 3258–3269.
15. Dovat, S., Ronni, T., Russell, D., Ferrini, R., Cobb, B.S., and Smale, S.T. (2002). Genes Dev. *16*, 2985–2990.
16. Jantz, D., and Berg, J.M. (2004). Proc. Natl. Acad. Sci. USA *101*, 7589–7593.
17. Eklund, E.A., Goldenberg, I., Lu, Y., Andrejic, J., and Kakar, R. (2002). J. Biol. Chem. *277*, 36878–36888.
18. Banerjee-Basu, S., Moreland, T., Hsu, B.J., Trout, K.L., and Baxevanis, A.D. (2003). Nucleic Acids Res. *31*, 304–306.
19. Marshall, S.A., Morgan, C.S., and Mayo, S.L. (2002). J. Mol. Biol. *316*, 189–199.

Chemistry & Biology, Vol. 11, July, 2004, ©2004 Elsevier Ltd. All rights reserved. DOI 10.1016/j.chembiol.2004.07.004

Cellular Addresses: Step One in Creating a Glycocode

In this issue of *Chemistry & Biology*, a library screening approach reveals at least four types of enzymes that attach galactosamine to build cell surface mucin-type glycoproteins [1]. A better molecular understanding of how these information-carrying oligosaccharides are created sets the stage for designing more selective inhibitors and potential therapeutics.

Cell surfaces are covered in diverse strings and branches of carbohydrate structures that create a kind of three-dimensional address system, or glycocode, which mediates interactions with a variety of biological components such as proteins or other cells [2, 3]. These addresses often change as cells grow and differentiate or become diseased [4]. Pathogens such as viruses and bacteria use such complex sugar structures to adhere to tissue for invasion of their hosts. The inhibition or promotion of these carbohydrate-based interactions serves as a new frontier in therapeutics for the treatment of conditions from cancer to viral and bacterial infections [5–7]. Unfortunately, very little is known about how these complex codes are assembled and regulated at the molecular level. In many cases, the actual changes to the cell surface architecture that occur with age or disease are not even known yet.

The most common protein-associated cell surface carriers of glycocodes in humans are the mucins or mucin-like proteins. *O*-glycosylation of the protein backbone at serine or threonine side chains with *N*-acetylgalactosamine (GalNAc) followed by addition of various other sugars creates densely clustered regions of carbohydrates that often eclipse the protein in size. The original hypothesis that these carbohydrate chains are initiated by only a few polypeptide GalNAc-transferases (ppGalNAcTs) is now being replaced with the realization that the system is far more complex (Figure 1) [8]. The human genome contains 24 putative ppGalNAcTs, and each isoform varies in its spatial and temporal regulation as well as tissue location [9]. A picture is emerging in which the highly glycosylated mucin domains are created by the action of several different ppGalNAcTs on the same protein backbone without simultaneous sugar chain initiation at every amino acid that is ultimately glycosylated [1, 8].

In this issue, the Bertozzi, Gerken, and Tabak groups report the first systematic study of the in vitro peptide or glycopeptide substrate requirements of eight members of the ppGalNAcT family [1]. These groups created a library of compounds based on a 13-amino acid segment of a rat mucin that includes every possible combination of sugar-modified threonine residues with up to four galactosamines. This library was incubated with each of eight glycosyltransferase isoforms and uridine-diphospho-*N*-azidoacetylgalactosamine (UDP-GalNAz) under saturating conditions. The azide-labeled substrate was previously shown to undergo this enzymatic process and provided a convenient method to detect the newly added sugars after Staudinger ligation with a modified phosphine and standard signal amplification [10].

Analysis of the data reveals four basic types of ppGalNacTs. Some prefer peptides with no sugars or only one sugar already attached; glycosylation of nearby amino acids inhibits these so-called early transferases. The intermediate ppGalNAcTs prefer peptides with two, or to a lesser extent, three sugars attached. The late transferases glycosylate peptides that already have three or even four sugars attached nearby. Interestingly, the enzymes in these three categories have some redundant functions, in that two different transferases will glycosylate the same peptide. Therefore, the loss of function of one of these enzymes in vivo can perhaps be rescued in part by other isoforms. In contrast, two of the eight tested ppGalNAcTs form a fourth category, which contains very specialized functions that cannot be taken over by other isoforms. Indeed, the knockout of one of these latter transferases in a fruit fly is shown to be lethal [11, 12]; therefore, the effect of a knockout of the other specialized ppGalNAcT will be of particular interest. Should the other isoforms prove incapable of rescuing the function of this specialized transferase, the real power of the molecular approach to mucin biosynthesis studies reported in this issue will become apparent.

The next question is if the now known differences in substrate acceptance among these isoforms can be