PERSPECTIVE

nature structural & molecular biology

© 2004 Nature Publishing Group http://www.nature.com/natstructmolbiol

Combinatorial control of gene expression

Attila Reményi[1,2,4], Hans R Schöler[1,3] & Matthias Wilmanns[2]

Revealing the molecular principles of eukaryotic transcription factor assembly on specific DNA sites is pivotal to understanding how genes are differentially expressed. By analyzing structures of transcription factor complexes bound to specific DNA elements we demonstrate how protein and DNA regulators manage gene expression in a combinatorial fashion.

Unlike prokaryotes that often use single proteins for transcriptional regulation of a gene, eukaryotic gene expression regulation involves the coordination of multiple proteins and is therefore combinatorial. This mechanism effectively integrates many different signaling pathways to provide a more complex network to meet the higher regulatory demand of a higher organism. Gene expression is regulated by the binding of transcription factors to DNA elements located in promoter or enhancer regions. Notably, the expression of all genes (>30,000) in a complex spatio-temporal pattern is achieved by a relatively low number of protein regulators (2,000–3,000). How is this limited set of protein regulators capable of building up a complex gene expression network? What are the molecular principles in the combinatorial assembly of transcriptional regulators?

In this work, we review four prototypic transcription factor families: nonsteroid nuclear receptors[1], MADS box–containing proteins[2], SOX proteins[3] and POU factors[4]. Members of these transcription families interact with other members of the same or unrelated transcription factor families. They control gene expression from various DNA enhancers in a combinatorial fashion. To understand the principles behind regulation of combinatorial gene expression at a molecular level, structures of protein–DNA complexes have to be determined for the same transcription factor in complex with other interacting partners bound to the same *cis*-acting DNA element and/or to other DNA enhancers. Now that a handful of structures of protein–DNA complexes are available that fulfill the above requirements, we describe the common principles that emerge at this stage.

Protein-DNA and protein-protein interactions

Transcription factors are composed of two regions. A DNA-binding domain recognizes and binds specific short stretches of DNA (~5–15 base pairs long). Activating regions interact either directly or indirectly via coregulators with different components of the basal transcription machinery, leading to or facilitating transcription. Expression of eukaryotic genes could thus be activated by 'regulated recruitment' of the RNA polymerase to promoter-enhancer regions[5]. These DNA regions—often referred to as *cis*-acting DNA elements—of eukaryotic genes, however, mediate the assembly of stereo-specific protein–DNA complexes[6]. The individual proteins often adapt their conformation and function according to the specific interacting partners and the nature of the DNA enhancer site[7]. Elucidation of the function of the complex as a whole necessitates careful inspection of the interplay of individual components comprising the regulatory protein–DNA complexes.

At the structural level, DNA binding may lead to various alterations in protein structure, including the formation of additional secondary structural elements, reorientation of loops, rearrangements of hydrophobic cores, and changes in their quaternary structure. Eukaryotic transcription factors often homo- or heterodimerize upon binding to *cis*-acting DNA elements, also contributing to structural complexity. Therefore, protein-protein and protein-DNA interactions from these relatively simple macromolecular complexes may help us understand their role in generating diverse patterns of eukaryotic gene expression.

The importance of DNA site spacing

Members of the nonsteroid nuclear receptor superfamily of transcription factors contain a zinc finger DNA-binding domain and operate by binding to hormone response elements (HREs). HREs consist of two minimal core hexad sequences, AGGTCA, which can be configured into various functional motifs. The orientation and spacing between these two hexamer sequences dictate the identity and the mode (monomer, heterodimer or homodimer) of nuclear receptor binding. The HREs (→) can form direct (→→), inverted (→←) and everted (←→) repeats based on their relative orientation. On inverted and everted repeats, the receptors dimerize via a small molecule-induced surface patch within the C-terminal ligand-binding domain. On direct repeats, however, the receptors form an additional interface between the conserved zinc finger DNA-binding domains (**Fig. 1a**).

In contrast to steroid receptors that bind to two half-sites of a DNA palindrome as homodimers, thyroid hormone receptors (TRs), retinoic acid receptors (RARs), vitamin D receptors (VDRs), peroxisome proliferator-activated receptors (PPARs) and several orphan receptors bind to direct repeats (DRs) as heterodimers with retinoid X receptors (RXRs). The binding specificity of these receptors is determined by the spacing between the half-sites, as PPARs, VDRs, TRs and RARs bind preferentially to direct repeats spaced by 1, 3, 4 or 5

[1]Gene Expression Program, European Molecular Biology Laboratory, 69117 Heidelberg, Germany. [2]European Molecular Biology Laboratory, Hamburg Outstation, Notkestrasse 85, D-22603 Hamburg, Germany. [3]Max-Planck-Institute for Molecular Biomedicine, Department of Cell and Developmental Biology, Mendelstrasse 7, 48149 Münster, Germany. [4]Present address: University of California, San Francisco, Department of Cellular and Molecular Pharmacology, 600 16th Street, San Francisco, California 94143-2240, USA. Correspondence should be addressed to A.R. (remenyi@itsa.ucsf.edu).

Published online 26 August 2004; doi:10.1038/nsmb820

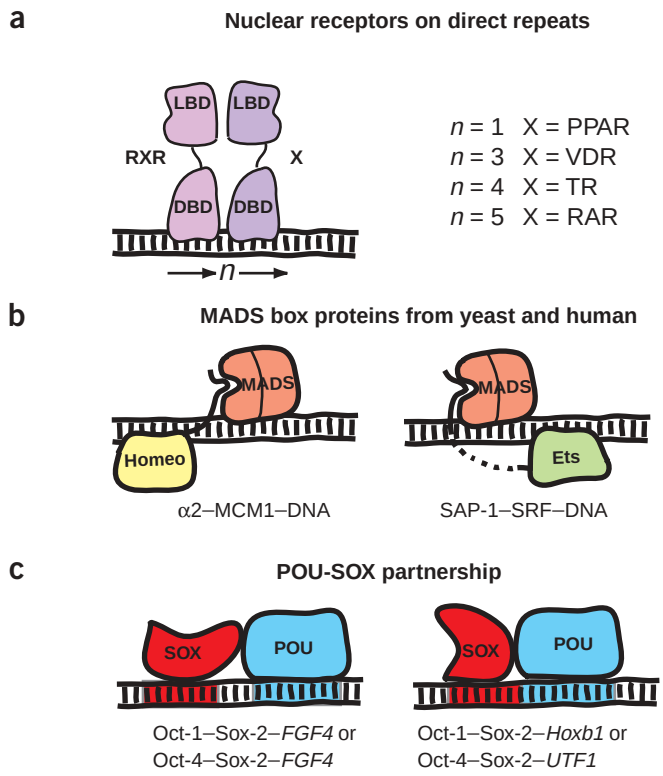

Figure 1 Examples of transcription factor–DNA complexes involved in combinatorial control of gene expression. (**a**) Heterodimerization of nuclear receptors on direct repeats. The number of base pairs between the direct repeat elements (AGGTCA, arrow) in the different hormone response elements (HREs) is denoted by *n*. X designates the name of the interacting partner of the RXR protein on different HREs. Nonsteroid nuclear receptors on direct repeats also interact via their DNA-binding domain (DBD) in addition to their association mediated by the ligand-binding domain (LBD). (**b**) Schematic representation of the crystal structures of the MATα2–MCM1–DNA[12] and SAP-1–SRF–DNA[11] ternary complexes. Dotted line, flexible linker from SAP-1 connecting the Ets domain with the MADS interaction domain, the so-called B box. (**c**) The enhancer regions of the *FGF4* and *Hoxb1* or *UTF1* genes contain POU- and SOX-binding sites that are differently spaced relative to each other (3 or 0, respectively).

nucleotides, respectively[1]. Thus, nuclear receptor heterodimerization on direct repeats exemplifies how a large family of related proteins may be involved in recognizing a wide repertoire of DNA motifs while employing one common protein interacting partner (RXR) to regulate transcription. A change in one nucleotide in the spacer region requires the RXR partner to be rotated by ~36° around and translated by 3.4 Å along the double helix in order for binding to occur. RXR, therefore, cannot use a single interface to bind different receptor partners, but requires a series of interaction surfaces related to the rotational and distance changes of its partners[8,9]. The interaction surface patch on RXR demonstrates the flexibility of an interaction module to adapt to various protein partners. In contrast, the specific receptor partner (such as TR) typically contains only one single dimerization interface that projects toward RXR.

An interface for divergent protein partners

Transcription factors belonging to the MADS box family contain a highly conserved DNA-binding domain that was originally identified by comparison of four transcription factors: the yeast MCM1, AG and DEFA from plants, and the human serum response factor (SRF). SRF is a ubiquitous protein important for cell proliferation and differentiation. It also contributes to the activation of a gene expression program called immediate-early gene response triggered by an extracellular mitogenic stimulus[10]. SRF operates by binding to the serum response element (SRE) found in numerous immediate-early gene promoters. The SRE DNA sequence, for example, mediates rapid transcriptional induction of the human *c-fos* proto-oncogene in response to growth factors. Full activation of the SRE requires the binding of SRF, and another protein, SRF-associated protein 1 (SAP-1). SAP-1 has an Ets DNA-binding domain and the structure of the SAP-1–SRF–SRE complex reveals that SAP-1 interacts with the SRF MADS domain by a module called the B box. This interaction module is connected to the Ets domain via a flexible linker region[11].

A homolog of SRF, a protein called MCM1, plays a prominent role in mating type determination in *Saccharomyces cerevisiae*. In the haploid α-cell MCM1 interacts with the homeodomain-containing protein MATα2 to repress genes specifying the **a** mating type. The crystal structure of the MATα2–MCM1–DNA ternary complex showed that an N-terminal extension to the MATα2 homeodomain interacts with the MADS box of MCM1 (ref. 12). Comparison of the crystal structures of the SAP-1–SRF–SRE and MATα2–MCM1–DNA complexes reveals a marked similarity between the interaction of a MADS domain protein (SRF or MCM1) with unrelated DNA-binding partners: the same hydrophobic groove on the MADS box domain surface is used to form an intermolecular β-sheet with the interacting partner[11]. This example demonstrates the versatile nature of an interaction surface patch for establishing functional partnerships with a diverse set of other protein regulators of transcription (**Fig. 1b**).

POU and SOX protein partnership

POU proteins are involved in a broad range of biological processes ranging from housekeeping gene functions (Oct-1) to those required to maintain the pluripotency of embryonic stem cells (Oct-4), the development of the immune system (Oct-1 and Oct-2) and cell type specification during organ development (Pit-1)[13]. This wide spectrum of activity is at odds with the small number of family members (15 in human, 5 in the fly (*Drosophila melanogaster*) and 4 in the worm (*Caenorhabditis elegans*). To function in such diverse processes, POU factors rely on control mechanisms, which in part are mediated by interactions with other family members or with unrelated regulatory proteins. SOX proteins, for example, are known to interact with various POU factors, and their genes are critically involved in the determination of cell fate[3]. The highly conserved DNA-binding segments of SOX factors are limited in their ability to bind to specific target sites, but assembly with their partner regulator proteins, such as POU or PAX, provides a plausible explanation for how they can distinguish their targets as well as act in a cell type–specific fashion[14,15].

Heterodimerization between the embryonic stem cell–specific POU factor Oct-4 and the Sox-2 protein provides the best-characterized example of a POU and SOX partnership[16]. The Oct-4–Sox-2 interaction on DNA is considered to be a combinatorial code in early embryonic development, as the expression level of these transcription factors directs the establishment of the first three lineages in the mammalian embryo[17]. Oct-4 and Sox-2 can interact on two distinct enhancer elements, *FGF4* and *UTF1*, and exert different degrees of cooperativity[18]. This difference may result in different transcriptional readouts for downstream genes, like *FGF4* and *UTF1*, based on the amount of Oct-4 and Sox-2 proteins present in the cell. The FGF4 enhancer contains three base pairs between the POU- and SOX-binding sites, whereas the UTF1 enhancer contains no such spacer (**Fig. 1c**).

© 2004 Nature Publishing Group http://www.nature.com/natstructmolbiol

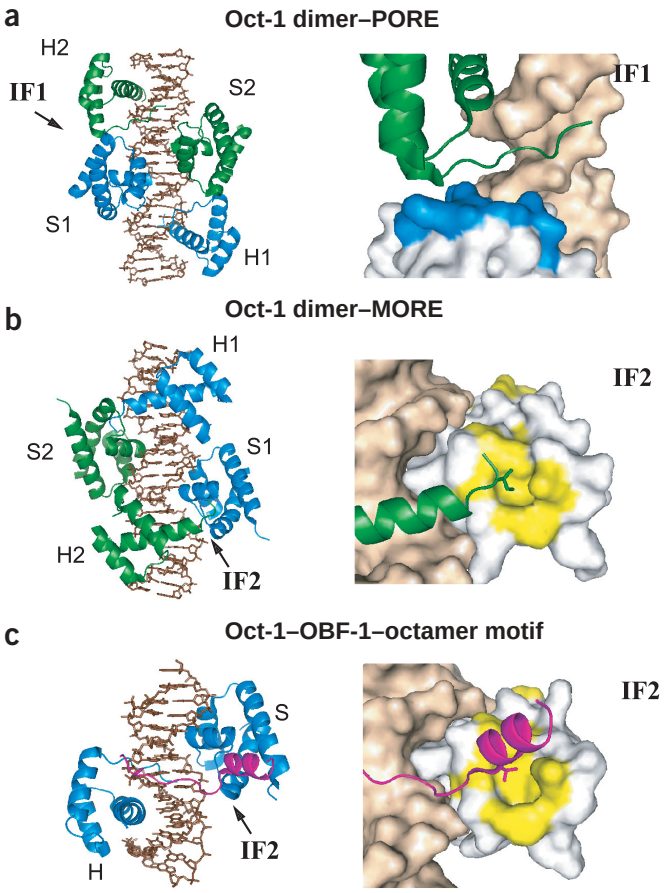

Figure 2 Importance of DNA site architecture in POU factor–mediated gene expression regulation. POU factors are composed of two autonomous DNA-binding domains connected by a flexible linker region (POU-specific domain, S; POU homeodomain, H). The flexible linker, invisible in the crystal structures, allows various arrangements of the two DNA-binding domains. (**a**–**c**) The structural basis for the differential interaction of the POU dimer formed on the PORE or on the MORE with the OBF-1 coactivator protein is revealed by comparing the crystal structures of the Oct-1 dimer–PORE (**a**) and the Oct-1 dimer–MORE (**b**) to the structure of the ternary complex of the Oct-1–OBF-1–octamer motif [24,25] (**c**). Oct-1 has two protein-protein interaction surface patches for homodimerization (IF1 and IF2, right-hand panels). POU domains of the two Oct-1 molecules are green and blue; DNA is brown. The PORE-like interface (IF1) is blue whereas the hydrophobic residues aligning the MORE interface (IF2) on the surface of the POU-specific domain are yellow. The interacting region of the OBF-1 protein is magenta.

The Hoxb1 DNA regulatory element also contains contiguous POU- and SOX-binding sites, with the same relative spacing as in the *UTF1* element[19]. This regulatory element functions to selectively promote transcription of the *Hoxb1* gene in a specific region of the hindbrain during embryogenesis. A solution structure of the Oct-1–Sox-2–*Hoxb1* ternary complex has revealed the molecular basis for the POU-SOX interaction on this DNA site[20] and allowed the generation of a reliable model of the POU–SOX–*UTF1* ternary complex. In this model, the protein-protein interface is formed by a different SOX surface patch compared with the one observed in the POU–SOX–*FGF4* complex, whereas the interaction site on the POU domain is conserved. Moreover, further studies have shown that Sox-2 uses one of these two interaction surfaces to bind yet another transcription factor of unrelated structure and function, Pax-6. This interaction is operative in later stages of embryogenesis compared with Oct-4–Sox-2 and serves as a developmental code in eye development[21].

POU factor dimerization

POU factors were originally identified to function as monomeric transcriptional regulators[22]. However, recent data have demonstrated that POU factors can also homo- and heterodimerize on specific DNA response motifs. This dimerization is likely to provide a higher level of functional diversity via different oligomeric arrangements. For instance, the same POU dimer (Oct-1 or Oct-2) binds to two DNA response elements, PORE and MORE, present in B-cell-specific genes. Although binding to PORE can lead to the recruitment of the B-cell-specific transcriptional coactivator OBF-1, the POU protein and its coactivator do not interact in the presence of MORE, suggesting two distinct DNA-mediated dimer configurations[23]. Crystal structures of the POU domain of Oct-1 in complex with the MORE or with the PORE indeed revealed the existence of two nonoverlapping surface patches for dimerization (that is, PORE-like and MORE-like patches). In the MORE-mediated POU dimer arrangement, the coactivator-binding site is blocked. In contrast, the PORE-mediated POU dimer quaternary structure is conducive to OBF-1 interaction[24,25] (**Fig. 2**).

Notably, the PORE-like and MORE-like interfaces of POU factors seem to be versatile in mediating protein-protein interactions with various other partners as well. The PORE-like interface on Oct-4, for example, is used to bind Sox-2 in the Oct-4–Sox-2 complexes formed on *FGF4* and *UTF1* (ref. 18). Furthermore, the PORE-like interaction module of Oct-1 also associates with Sox-2 in the Oct-1–Sox-2–*Hoxb1* ternary complex[20]. The MORE-like interface, on the other hand, apart from interacting with OBF-1, has also been reported to be operational in binding to the STAT5 protein on the *cyclin D1* promoter sequence[26].

The POU domain is composed of two regions: the POU-specific and POU-homeodomain regions, which are structurally and functionally autonomous in DNA binding. A flexible linker region between

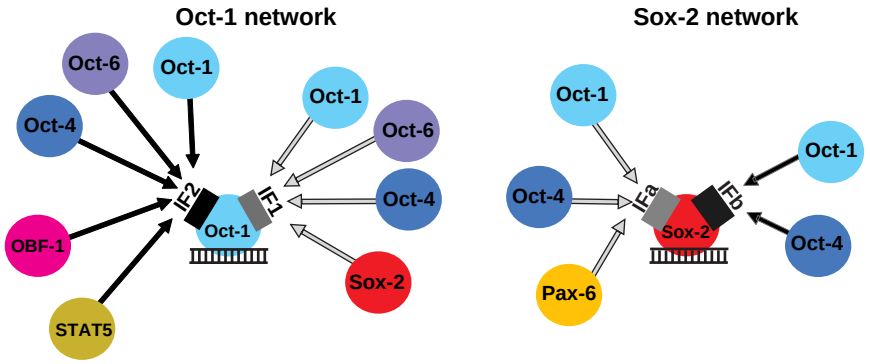

Figure 3 Interaction diagram of Oct-1 and Sox-2. Transcription factors are depicted as protein molecules with surface patches that can interact with a whole array of different partners provided that the protein is bound to a specific DNA element. DNA-bound Oct-1 and Sox-2 are depicted schematically with protein-protein interaction surface patches that are instrumental in binding to other partners. IF1 and IF2 on the Oct-1–DNA complex denote two interfaces of Oct-1 that are accessible and used for interaction on various DNA. Similarly, IFa and IFb designate interfaces of Sox-2 that are used for interaction on different DNA sites.

© 2004 Nature Publishing Group http://www.nature.com/natstructmolbiol

these two subdomains allows various domain arrangements on DNA. Owing to the flexible nature of the linker region, differences within the same POU dimer configuration could also lead to the selective recruitment of other transcriptional regulators. This has been demonstrated by the pituitary-specific POU factor Pit-1 (ref. 27). Two structures of Pit-1 bound to two related DNA response elements within the prolactin (*Prl*) and growth hormone (*GH*) promoters, differing only by a TT insertion in the *GH* promoter, showed how the MORE-like dimer arrangement can accommodate different spacings between the POU subdomain–binding sites. The extended quaternary structure adopted by Pit-1 on *GH*, but not on *Prl*, enables the recruitment of a corepressor complex containing N-CoR. The recruitment of such a complex on *GH* could explain the repression of growth hormone gene expression in lactotrope pituitary tissues.

POU factors have considerable sequence similarity, particularly within the segments involved in DNA binding and POU-POU interactions. It has been shown that the MORE and PORE *cis*-acting DNA elements both mediate versatile homo- and heterodimerization among all four Oct factors tested (Oct-1, Oct-2, Oct-4 and Oct-6)[23]. Molecular modeling combined with biochemical analysis suggested that the interaction surfaces for homo- and heterodimerization are the same. Despite their high degree of sequence similarity in their POU domain, Oct factors have divergent activation domains. These, in turn, are very likely to mediate interactions with divergent sets of other transcription factors or coregulators. Because these four Oct factors have overlapping spatio-temporal expression patterns during embryogenesis and in adult tissues, it is tempting to speculate that by heterodimerization, differential transcriptional activities could be simply acquired from the same *cis*-acting DNA element. What determines which heterodimer pairs are possible on a certain DNA regulatory element, however, is poorly understood. The assembly of POU dimers might be regulated by upstream signals via post-translational modification, as both the MORE- and PORE-like interfaces contain potential target sites for phosphorylation[24].

Concluding remarks and outlook

In summary, these examples demonstrate that transcription factors can bind to multiple partner proteins in a similar manner or use a different set of interaction surfaces. Dimerization surface patches seem to be adept at mediating interactions with different interacting protein partners, but their usage is strictly dependent on the *cis*-acting DNA element that coordinates the assembly of the protomers. It is the architecture, or, more simply, the sequence of the DNA, that determines whether an interaction module is properly aligned in the stereospecific complex for a particular protein partner or whether it is accessible to coactivators or corepressors (**Figure 3**).

Although this feature of protein regulators of transcription is certainly not the only one governing 'combinatorial control,' it represents one important aspect of how multiple transcription factors come together to exert specific functions. With the increased success of identifying genomic locations for transcription factor–binding sites in higher eukaryotes experimentally, as well as *in silico* (reviewed in refs. 28 and 29), we anticipate the discovery of more *cis*-regulatory DNA elements that are involved in combinatorial control of genes. Identification of the regulatory components (proteins and DNA elements alike) that govern these processes should be followed by the establishment of their interaction maps. A better understanding of gene expression control at the structural level could then set the stage for systematic mapping of the protein-protein interaction surfaces in protein–DNA complexes. Conducting *in vivo* structure-function relationship analysis in model organisms, in turn, could then provide valuable data for testing and exploring gene network organization. Ultimately, this information may be instrumental in modulating cell fate, pluripotency or embryonic stem cell plasticity.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Received 5 April; accepted 30 June 2004
Published online at http://www.nature.com/nsmb/

1. Glass, C.K. Differential recognition of target genes by nuclear receptor monomers, dimers, and heterodimers. *Endocr. Rev.* **15**, 391–407 (1994).
2. Shore, P. & Sharrocks, A.D. The MADS-box family of transcription factors. *Eur. J. Biochem.* **229**, 1–13 (1995).
3. Wegner, M. From head to toes: the multiple facets of Sox proteins. *Nucleic Acids Res.* **27**, 1409–1420 (1999).
4. Herr, W. & Cleary, M.A. The POU domain: versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain. *Genes Dev.* **9**, 1679–1693 (1995).
5. Ptashne, M. & Gann, A. *Genes & Signals* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2002).
6. Tjian, R. & Maniatis, T. Transcriptional activation: a complex puzzle with few easy pieces. *Cell* **77**, 5–8 (1994).
7. Lefstin, J.A. & Yamamoto, K.R. Allosteric effects of DNA on transcriptional regulators. *Nature* **392**, 885–888 (1998).
8. Rastinejad, F., Perlmann, T., Evans, R.M. & Sigler, P.B. Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature* **375**, 203–211 (1995).
9. Mangelsdorf, D.J. & Evans, R.M. The RXR heterodimers and orphan receptors. *Cell* **83**, 841–850 (1995).
10. Pellegrini, L., Tan, S. & Richmond, T.J. Structure of serum response factor core bound to DNA. *Nature* **376**, 490–498 (1995).
11. Hassler, M. & Richmond, T.J. The B-box dominates SAP-1-SRF interactions in the structure of the ternary complex. *EMBO J.* **20**, 3018–3028 (2001).
12. Tan, S. & Richmond, T.J. Crystal structure of the yeast MATα2/MCM1/DNA ternary complex. *Nature* **391**, 660–666 (1998).
13. Ryan, A.K. & Rosenfeld, M.G. POU domain family values: flexibility, partnerships, and developmental codes. *Genes Dev.* **11**, 1207–1225 (1997).
14. Kamachi, Y., Uchikawa, M. & Kondoh, H. Pairing SOX off: with partners in the regulation of embryonic development. *Trends Genet.* **16**, 182–187 (2000).
15. Dailey, L. & Basilico, C. Coevolution of HMG domains and homeodomains and the generation of transcriptional regulation by Sox/POU complexes. *J. Cell Physiol.* **186**, 315–328 (2001).
16. Ambrosetti, D.C., Basilico, C. & Dailey, L. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Mol. Cell. Biol.* **17**, 6321–6329 (1997).
17. Avilion, A.A. *et al.* Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.* **17**, 126–140 (2003).
18. Remenyi, A. *et al.* Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev.* **17**, 2048–2059 (2003).
19. Di Rocco, G. *et al.* The recruitment of SOX/OCT complexes and the differential activity of HOXA1 and HOXB1 modulate the Hoxb1 auto-regulatory enhancer function. *J. Biol. Chem.* **276**, 20506–20515 (2001).
20. Williams, D.C. Jr., Cai, M. & Clore, G.M. Molecular basis for synergistic transcriptional activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1.Sox2.Hoxb1-DNA ternary transcription factor complex. *J. Biol. Chem.* **279**, 1449–1457 (2004).
21. Kamachi, Y., Uchikawa, M., Tanouchi, A., Sekido, R. & Kondoh, H. Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. *Genes Dev.* **15**, 1272–1286 (2001).
22. Klemm, J.D., Rould, M.A., Aurora, R., Herr, W. & Pabo, C.O. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* **77**, 21–32 (1994).
23. Tomilin, A. *et al.* Synergism with the coactivator OBF-1 (OCA-B, BOB-1) is mediated by a specific POU dimer configuration. *Cell* **103**, 853–864 (2000).
24. Remenyi, A. *et al.* Differential dimer activities of the transcription factor Oct-1 by DNA-induced interface swapping. *Mol. Cell* **8**, 569–580 (2001).
25. Chasman, D., Cepek, K., Sharp, P.A. & Pabo, C.O. Crystal structure of an OCAB peptide bound to an Oct-1 POU domain/octamer DNA complex: specific recognition of a protein-DNA interface. *Genes Dev.* **13**, 2650–2657 (1999).
26. Magne, S., Caron, S., Charon, M., Rouyez, M.C. & Dusanter-Fourt, I. STAT5 and Oct-1 form a stable complex that modulates cyclin D1 expression. *Mol. Cell. Biol.* **23**, 8934–8945 (2003).
27. Scully, K.M. *et al.* Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science* **290**, 1127–1131 (2000).
28. Bulyk, M.L. Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5**, 201 (2003).
29. Taverner, N.V., Smith, J.C. & Wardle, F.C. Identifying transcriptional targets. *Genome Biol.* **5**, 210 (2004).

© 2004 Nature Publishing Group http://www.nature.com/natstructmolbiol