

How structural biology transformed studies of transcription regulation

Received for publication, March 2, 2021, and in revised form, April 15, 2021. Published, Papers in Press, May 4, 2021, <https://doi.org/10.1016/j.jbc.2021.100741>

Cynthia Wolberger*

From the Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

Edited by Karin Musier-Forsyth

The past 4 decades have seen remarkable advances in our understanding of the structural basis of gene regulation. Technological advances in protein expression, nucleic acid synthesis, and structural biology made it possible to study the proteins that regulate transcription in the context of ever larger complexes containing proteins bound to DNA. This review, written on the occasion of the 50th anniversary of the founding of the Protein Data Bank focuses on the insights gained from structural studies of protein–DNA complexes and the role the PDB has played in driving this research. I cover highlights in the field, beginning with X-ray crystal structures of the first DNA-binding domains to be studied, through recent cryo-EM structures of transcription factor binding to nucleosomal DNA.

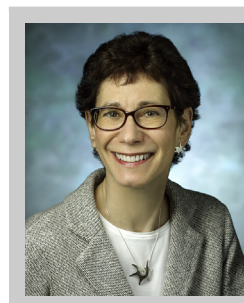
The publication in the early 1980s of the first crystal structures of DNA and of proteins that bind to specific DNA sequences (1–3) marked a turning point in structural biology. X-ray crystallography had already made a profound impact on biology and biochemistry (4), beginning with the first atomic models of hemoglobin (5) and myoglobin (6), to the first structures of enzymes (7), antibodies (8), and tRNA (9, 10). At the same time, the need for large amounts of material to grow crystals of sufficient size and quantity restricted the field to naturally abundant proteins. With the exception of tRNA, it was not possible to obtain the homogeneous samples of RNA or DNA needed for crystallization trials. The advent of molecular cloning and strategies for overexpressing proteins in bacteria, however, dramatically increased the number and types of proteins whose structures could be determined. The publication in the early 1980s of structures of *E. coli* catabolite activator protein (CAP) (2) and of the bacteriophage lambda Cro (1) and *cI* (3) repressor proteins was electrifying and provided the first glimpses of how proteins might bind DNA and regulate transcription. The much broader set of biological problems to which structural methods could now be applied greatly increased the interest in structural biology. At the same time, the development of chemical methods to synthesize DNA oligonucleotides of defined length and sequence made it possible to crystallize and determine structures of DNA (11, 12), as well as of protein–DNA complexes. Indeed, it was the

1981 publication of the crystal structure of a B-DNA dodecamer by Dickerson and colleagues (1BNA) (12) that finally provided experimental proof for the B-DNA model proposed by Watson and Crick in 1953 (13). These combined developments in recombinant DNA technology and chemical synthesis of DNA marked the beginning of a new era in studies of protein–DNA interactions and gene regulation.

The advances in cloning and oligonucleotide synthesis played an additional role in expanding the impact of structural biology beyond simply making it possible to determine structures of protein–DNA complexes. The development of approaches that utilized oligonucleotides to engineer specific amino acid substitutions into proteins (14) meant that one could use structural information to introduce mutations that could be then be used to test mechanistic hypotheses based on crystal structures. An early example was the test of a model for how the helix–turn–helix element (15), which had been identified in early structures of DNA-binding proteins, mediated contacts with DNA base pairs. Site-directed mutagenesis of the bacteriophage 434 repressor validated the proposed model for DNA binding and provided clues as to how side chain contacts determined DNA sequence recognition (16). These new approaches that made it possible to use structural information to drive biochemical and genetic studies further broadened interest in structural biology and helped fuel a dramatic expansion in what had once been a relatively small community of X-ray crystallographers and NMR spectroscopists.

The ability to utilize the new structural information on DNA–protein complexes was, however, limited because many of these new structures were not broadly available. Although the Protein Data Bank (PDB) had been established more than a decade earlier, coordinate deposition was voluntary and many structures of proteins and oligonucleotides were not publicly available (17). Indeed, coordinates for the first DNA-binding proteins mentioned above, CAP (2), lambda cro (1), and lambda *cI* (3), were not deposited in the PDB. Recommendations from the International Union of Crystallography (18) and policy changes at the National Institutes of Health (19) and other funding entities led to mandatory coordinate deposition, making these exciting structures available to all investigators. The number and complexity of protein–nucleic acid complex structures have increased by many orders of magnitude since that time, fueled by technical advances in X-ray

* For correspondence: Cynthia Wolberger, cwolberg@jhmi.edu.



Cynthia Wolberger, Professor of Biophysics and Biophysical Chemistry at the Johns Hopkins University School of Medicine, is a leader in research on transcriptional regulation and ubiquitin signaling.

crystallography, nuclear magnetic resonance (NMR) spectroscopy and, most recently, cryo-electron microscopy (cryo-EM). The availability in the PDB of so many structures of individual transcription factors, enzymes, and nucleosomes has greatly facilitated structure determination of large complexes that contained many of these macromolecules. Most importantly, these structures are easily accessible to all outside the structural biology community and continue to drive new science.

This review focuses on the insights into the regulation of transcription gained from structural studies of protein–DNA complexes and the role the PDB has played in driving this research. I present a historical view of some of the milestones, beginning with structural studies of bacterial and phage repressor proteins bound to DNA, through structures of larger complexes determined by cryo-EM. I have provided the PDB ID in either the text or figure legend for each structure mentioned. Alas, a number of early structures were never deposited in the PDB, so in these cases I also provide a reference to a subsequent structure, along with its corresponding PDB ID. Given its focus on regulation, this review focuses on sequence-specific DNA-binding proteins and does not cover the structural studies of RNA polymerase or of the many transcription factors and chromatin-modifying enzymes required for transcription initiation and elongation. The reader is referred to several recent reviews that cover the remarkable structures of the eukaryotic (20, 21) and bacterial (22) transcription machinery.

Recognition of specific DNA sequences

Regulation of specific genes depends on proteins that can recognize a particular sequence of DNA base pairs in a regulatory region. In bacteria, these proteins either activate or repress transcription by directly interacting with RNA polymerase (23). In eukaryotes, transcriptional regulators have separate domains that may recruit coactivator or corepressor complexes that attach or remove post-translational modifications from histone, reposition nucleosomes, or promote assembly of the transcription preinitiation complex (24). Just a few years after structures of the first isolated DNA-binding domains mentioned above were elucidated (1–3), the first protein–DNA complexes reported in the mid-1980s marked the beginning in our understanding of the molecular basis for recognition of specific DNA sequences. Structures of complexes with the

bacteriophage lambda (1LMB) (25, 26) and 434 repressors (2OR1) (27, 28) and cro (3CRO, 4CRO) (29–31) proteins showed how the second helix in the previously identified helix–turn–helix motif (15) inserted into the major groove of B-DNA (Fig. 1A). Side chains in the recognition helix contacted the edges of the DNA bases directly or *via* water-mediated hydrogen bonds, thereby contributing to sequence specificity, while other regions of the protein formed additional stabilizing contacts with the sugar-phosphate backbone. Although the bacteriophage repressors bound to relatively straight DNA, it turned out that the *E. coli* CAP protein (1CGP) induces a dramatic 90° bend in the helix axis (32) (Fig. 1B). This would be the first of many examples of proteins that induce bends and other distortions in the DNA that modulate the nature of sequence-specific contacts as well as (in most cases) increasing the buried surface area between protein and DNA. The helix–turn–helix motif was soon found in eukaryotic homeodomain proteins such as *Drosophila* engrailed (1HDD) (33) (Fig. 1, C and D) and yeast MAT α 2 (1APL) (34), although the longer recognition helix in homeodomains docked on DNA in a somewhat different manner. In general, all of these structures provided different examples of proteins that form chemically complementary interfaces with the DNA.

An unexpected twist on the nature of DNA sequence recognition emerged with the structure of the bacterial Trp repressor bound to DNA (35). Although Trp repressor also contains a helix–turn–helix, the protein forms no direct contacts with DNA bases. Instead, there are water-mediated contacts between Trp repressor and the base pairs in the major groove (Fig. 1E), with direct contacts formed only with the DNA backbone. The DNA sequence specificity of Trp repressor derives from sequence-dependent variations in DNA structure, a form of recognition termed indirect readout (35). Subsequent analyses have shown that sequence-dependent local variations in DNA structure play a role in a broad array of proteins that bind DNA (36).

As more structures of complexes were determined, the remarkable structural diversity of sequence-specific DNA binding domains and the different modes of interaction with both the major and minor grooves quickly became evident. The early 1990s saw a veritable explosion in the number of novel DNA-binding domains. The structure of the DNA-bound bacterial Met repressor (1CMA) (37) revealed that a pair of beta strands fit in the major groove (Fig. 1F) just as well as an α helix. This validated a prediction, made well before any structures of DNA-binding proteins had been determined, that both α helices and β sheets had the optimal dimensions to fit in the major groove of B-DNA (38). Structures of eukaryotic transcriptional regulators such as the basic region-leucine zipper (bZIP) (39) (Fig. 2A), helix–loop–helix (40) (Fig. 2B), Gal4-type zinc binding domain (41) (Fig. 2C), and the immunoglobulin-like Rel homology domain (42, 43) (Fig. 2D) proteins represented yet other structurally distinct modes of docking on DNA and recognizing specific DNA sequences.

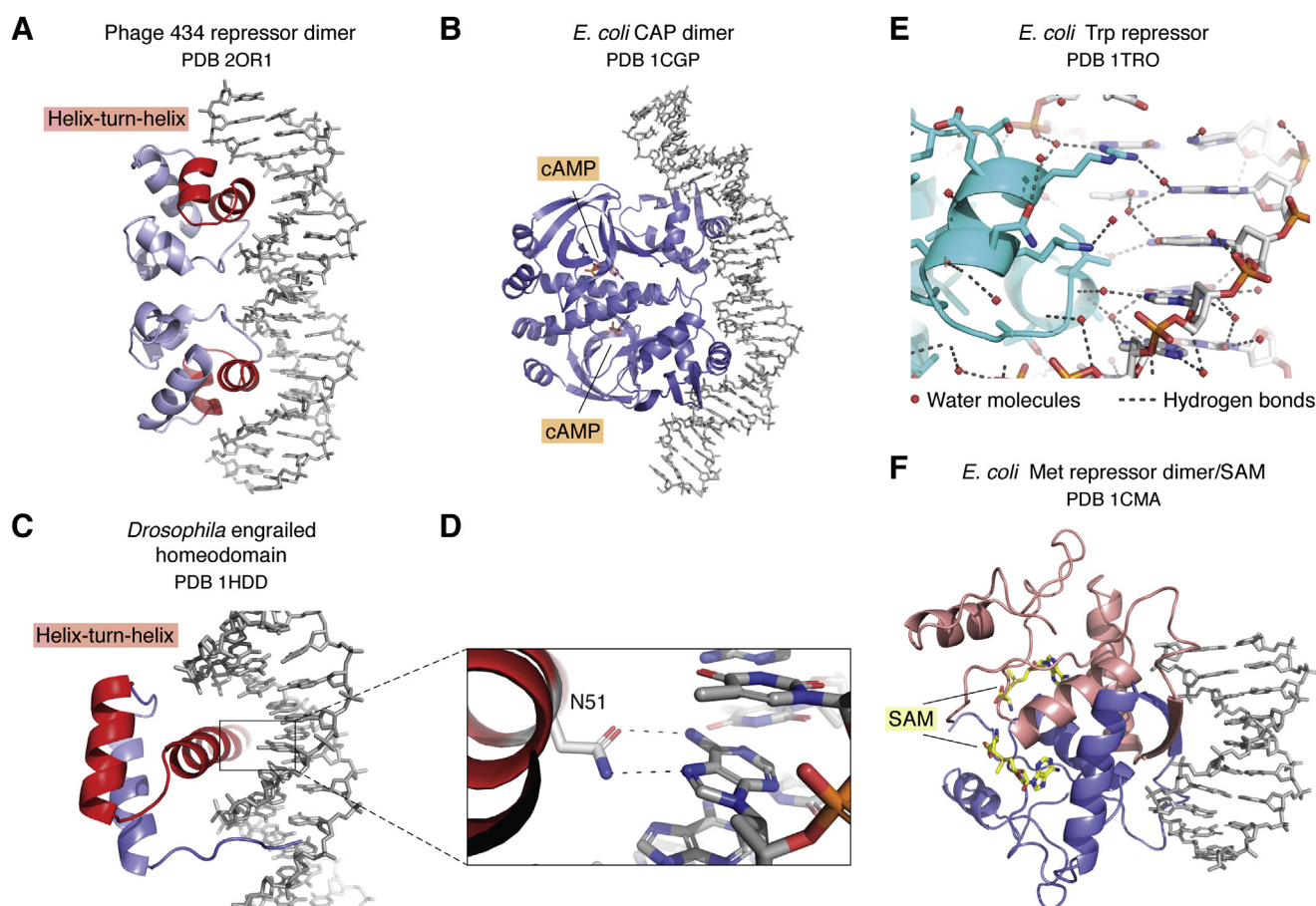


Figure 1. Structures of bacterial and phage repressors and activators bound to DNA. A, phage 434 repressor dimer (2OR1) (28). The helix–turn–helix is highlighted in red. B, *E. coli* CAP dimer with cAMP bound to each monomer (1CGP). C, *Drosophila* engrailed homeodomain with helix–turn–helix is highlighted in red (1HDD). D, hydrogen bonding between Asn and Adenine (1HDD). E, *E. coli* Trp repressor showing waters (*spheres*) and water-mediated hydrogen bonds (1TRO). F, Met repressor dimer with two chains colored differently. There is one molecule of S-adenosine methionine (SAM) bound to each monomer (1CMA).

Perhaps the most unexpected finding from this era was the discovery of the dramatic DNA distortion induced by the eukaryotic TATA-binding protein (TBP), a subunit of the basal transcription factor complex, TFIID, that binds to the TATA box promoter element and helps nucleate assembly of the transcription preinitiation complex (44). In marked contrast to the proteins that insert helices, strands, or loops into the DNA grooves, essentially forming a structurally complementary surface (see Figs. 1 and 2), it is the concave surface of TBP that contacts DNA in the minor groove (1YTB, 1VTL) (45, 46) (Fig. 2E). A severe distortion in the DNA, which contains a nearly 90° bend in the helix axis and is underwound, enables the concave surface of TBP to form sequence-specific contacts with bases in the minor groove.

Zinc finger proteins were distinct from other classes of DNA-binding domains in their modular recognition of DNA sequences, and whose molecular details were first revealed in the structure of the three zinc fingers of Zif268 bound to DNA (1ZAA) (47) (Fig. 3A). Members of this large family of transcriptional regulators contain multiple tandem repeats of the ~33 amino acid domain with a structural zinc coordinated by two histidine and two cysteine side chains (48),

with each zinc finger recognizing 3 to 4 base pairs (47) (Fig. 3A). The modular nature of zinc finger proteins presented an opportunity to engineer proteins with particular DNA-binding specificities (49–52), which could then be used to target nucleases or other domains to specific sites in the genome (53, 54). This marked the first attempt at targeted genome engineering, which was followed a decade later by designed TAL effector proteins (55). Each repeat in these plant DNA-binding proteins recognizes a single base pair (56) (Fig. 3B), which greatly facilitated design of proteins with the desired DNA sequence specificity (57, 58) that could similarly be linked to endonuclease domains for genome engineering (55).

An analysis of PDB depositions as of the year 2000 (59) identified seven broad classes of sequence-specific DNA-binding proteins, with variations within each class. One of the common themes to emerge from all of these studies was the prevalence of DNA sequence recognition *via* contacts in the major groove, where the pattern of nucleobase functional groups is unique to each DNA sequence. Although it had initially been thought by some that there might be a recognition code in the form of a one-to-one correspondence between

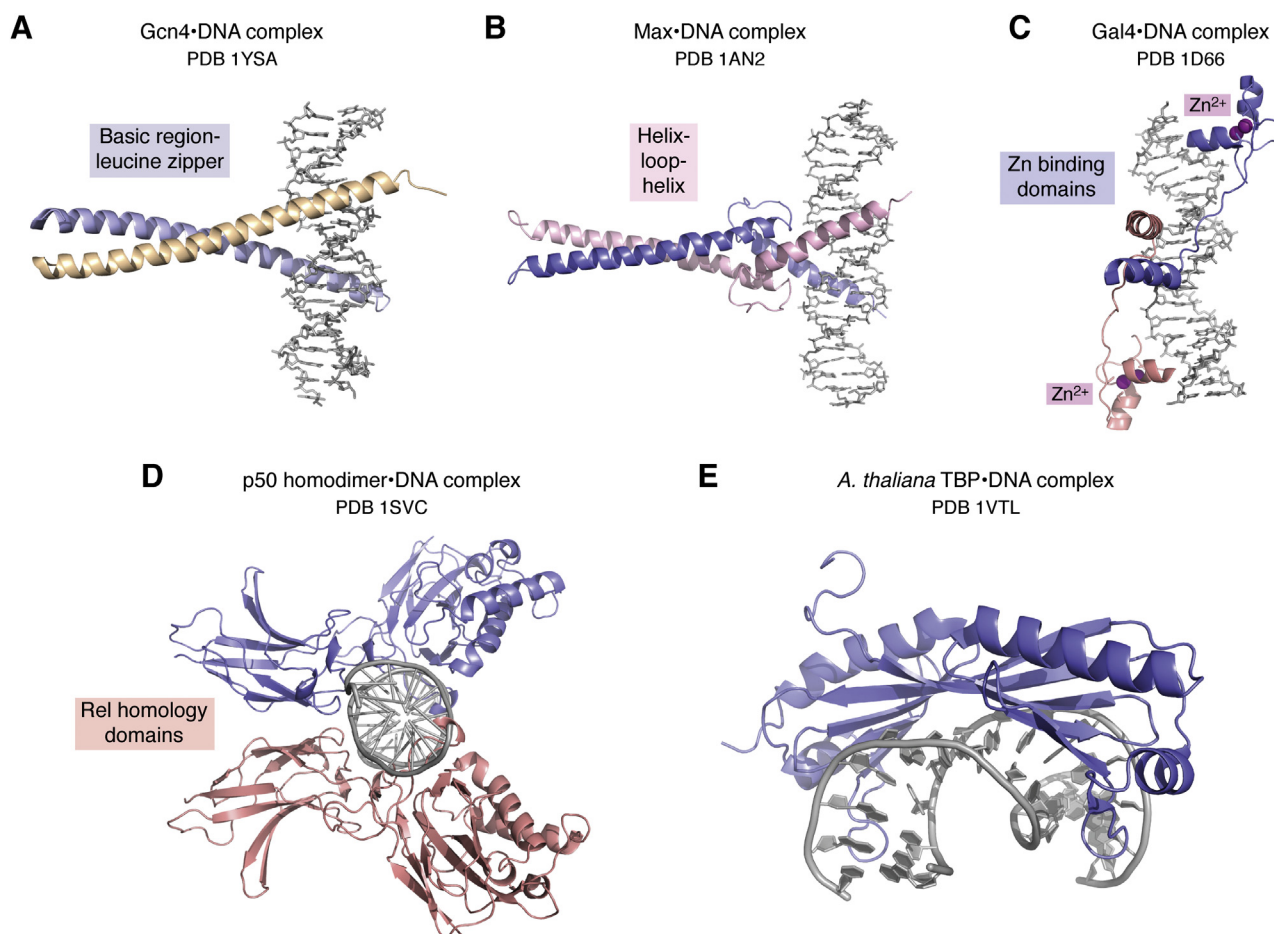


Figure 2. Eukaryotic DNA-binding domains. A, Gcn4, a basic region-leucine zipper (bZIP) protein (1YSA). B, max, a helix-loop-helix protein (1AN2). C, Gal4. Each DNA-binding domain coordinates two Zn^{2+} ions (1D66). D, p50 homodimer, Rel-homology domain (1SVC). E, *A. thaliana* TBP bound to DNA (1VTL).

a particular base and one or more unique side chains, it became apparent early on that there was no such code (60), with the exception of the TAL effector proteins (57, 58). A comprehensive review of the determinants of DNA sequence recognition can be found in (61).

Combinatorial regulation of transcription

Many eukaryotic genes are regulated by multimeric complexes that can regulate transcription in response to multiple inputs. Beginning in the mid-1990s, structural studies of transcriptional regulators advanced to the next level of complexity, with structures determined of multiprotein complexes bound to DNA. One of the first was of the nuclear hormone receptor heterodimer composed of 9-*cis*-retinoic acid receptor (RXR) and thyroid hormone receptor (TR) (2NLL) (62) (Fig. 4A). Members of this family of DNA-binding proteins can form homodimers or heterodimers and contain separate ligand-binding domains, which change conformation upon ligand binding and recruit enzyme complexes that activate (coactivators) or repress (corepressors) transcription (63). Of interest, the RXR and TR DNA-binding domains bind DNA in tandem (62), in contrast with other members of this family, such as glucocorticoid receptor, which bind as symmetric

dimers (64). Structural studies of the homeodomain superfamily revealed an even greater degree of complexity, as selected members of this family can heterodimerize with other homeodomain proteins or with DNA-binding proteins belonging to completely dissimilar structural families. The yeast MAT α 2 homeodomain protein, for example, can heterodimerize with a second homeodomain protein, MAT α 1 (1YRN) (65), or with MCM1 (1MNM) (66), a MADS box DNA-binding protein that is unrelated in structure to homeodomains (Fig. 4, B and C). Structures of *Drosophila* Ubx/Exd (1B8I) (67) and human HoxB1/Pbx1 (1B72) (68) homeodomain heterodimers bound to DNA provided additional insights into how transcription programs are regulated during development.

Through the late 1990s and early 2000s, structures of even larger complexes were determined. With the availability of many structures of smaller protein-DNA complexes in the PDB, it became possible to model regulatory regions and enhancers to which multiple proteins bind. The β -interferon enhancer, for example, contains binding sites within the 55-base pair enhancer sequence for eight proteins that bind cooperatively, together forming an “enhanceosome” (69). By combining structural information from several multiprotein

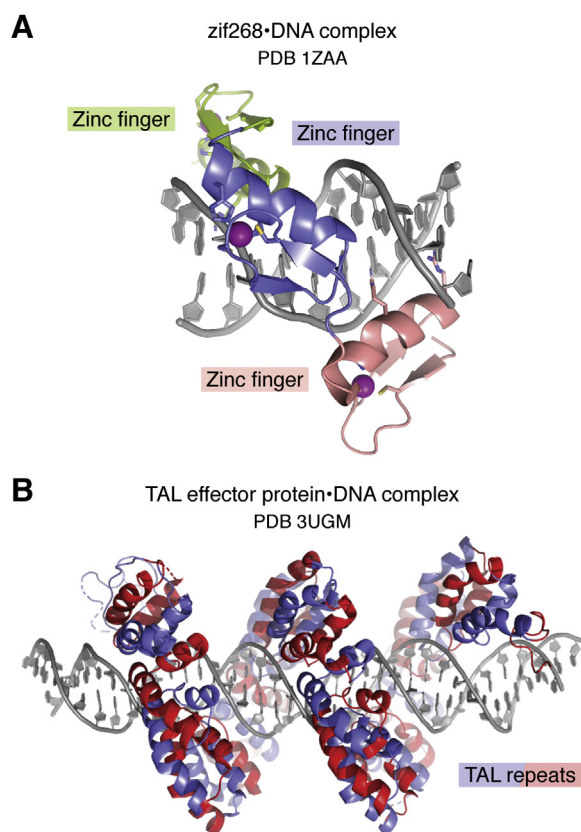


Figure 3. Modular recognition of DNA sequence. *A*, zif268 has three zinc fingers (each colored differently) that bind DNA (1ZAA). Each zinc is coordinated by two His and two Cys side chains (sticks). Each finger contacts bases in the major groove with two side chains. *B*, TAL effector protein (3UGM). Alternate 34 amino acid TAL repeats are colored red and blue.

subcomplexes (2O61, 2O6G, 1T2K), it was possible to assemble a model of the entire enhanceosome containing the bHLH proteins, C-Jun and ATF-2; the Rel homology domain proteins, p50 and RelA; and four IRF proteins, IRF-3A, 3C, 7B, and 7B (Fig. 4D) (70, 71).

Transcription factor binding and chromatin

The packaging of eukaryotic DNA into chromatin impacts all cellular processes requiring access to DNA, including transcription. It is perhaps not surprising, then, that the structure of the fundamental organizational unit of the genome, the nucleosome (1AOI), is the most highly cited structure in the PDB (72). The first high-resolution structure of the nucleosome core particle, determined in 1997 (73), revealed the molecular details of how the 146-base pair DNA duplex wraps twice around the histone octamer core, which contains two copies each of histones H2A, H2B, H3, and H4 (Fig. 5A). One of the important observations to emerge from this and subsequent studies was that the DNA is not smoothly bent but instead contains local kinks that are favored by particular DNA sequences (74). These sequence-dependent differences in the relative energetic penalty for DNA kinking and bending thus play a key role in positioning nucleosomes at particular locations in the genome (74).

Most transcriptional regulators bind to nucleosome-depleted or nucleosome-free regions, where their DNA-binding domains can freely access the DNA. A class of regulatory proteins known as pioneer transcription factors, however, bind directly to nucleosomal DNA in regions of compacted chromatin and reprogram cell fate by altering local chromatin structure (75). Two recent cryo-EM structures have provided the first insights into how pioneer transcription factors, OCT4 and SOX2, bind to DNA that is simultaneously wrapped around the nucleosome (76, 77). The position of the recognition sequence within the nucleosomal DNA, or even whether there is a fixed site, is not clear, so the two studies inserted the recognition sequence in different locations in the DNA based on solution experiments. The structure of both OCT4 and SOX2 bound to DNA near the exit site from the nucleosome (superhelical location -6, or SHL-6) show the DNA peeled away from the histone core (Fig. 5B), suggesting a mechanism by which these pioneer factors could help open chromatin (6T90) (77). In structures of SOX2 (6T7B) and of a homolog, SOX11 (6T7A), bound to an internal DNA site at SHL +2 (Fig. 5C), the DNA is somewhat distorted and bulges away from the histone core (76). Together, these structures constitute an important start in understanding the mechanism by which these and other pioneer transcription factors open chromatin and alter transcription programs.

Visualizing entire transcription initiation complexes

A more complete understanding of how DNA-binding proteins orchestrate transcription will require structures of ever-larger complexes containing all necessary components for transcription initiation. Thanks to the recent “resolution revolution” in cryo-EM (78), one structure after another of huge protein–nucleic acid complexes have provided unprecedented insights into transcription. Since virtually all of these complexes contain proteins whose structures had been determined, usually by X-ray crystallography, the ready availability of coordinates and associated data in the PDB has greatly facilitated map interpretation and model building. The many structures of basal transcription factors bound to DNA, such as TBP (1YTB, 1VTL) (45, 46), TFIIA (79), and TFIIB (80), and RNA polymerase II have provided the foundation on which to interpret structures of transcription initiation and elongation complexes (for a comprehensive recent review, see (21)). The change in the PDB coordinate data format to mmCIF/PDBx (81) was another advance that facilitated working with such large structures. The original PDB format, which was based on the number of characters that could fit on an IBM punch card, could only accommodate structures with up to 62 chains and fewer than 100,000 atoms. The mmCIF/PDBx format has no such limit and also accommodates additional metadata containing information about the macromolecules as well as experimental details. Thus, structures of a 2.7-MDa bacterial expressosome containing RNA polymerase, a ribosome, the NusG transcription factor, duplex DNA, mRNA, and tRNA could be accommodated in a single coordinate file, even though it contains 65 unique chains and more than 175,000

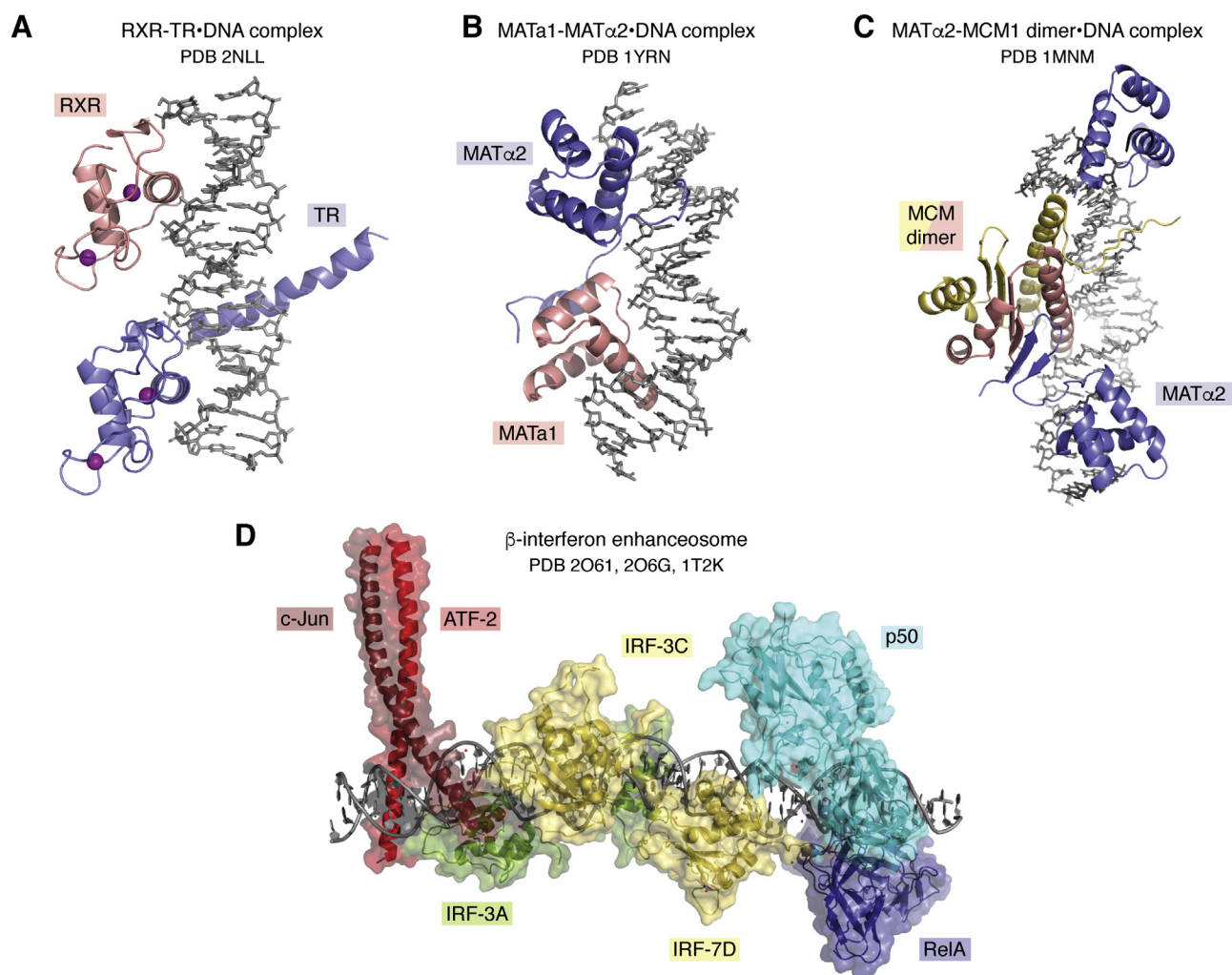


Figure 4. Combinatorial regulation by multiprotein complexes. A, RXR (salmon)-TR (blue) (2NLL). B, Mat a1 (salmon)-Mata2 (blue) (1YRN). C, Mata2 (blue) and an MCM1 dimer (salmon and yellow) (1MNM). D, β -interferon enhanceosome. Model assembled from 2O61, 2O6G, 1T2K.

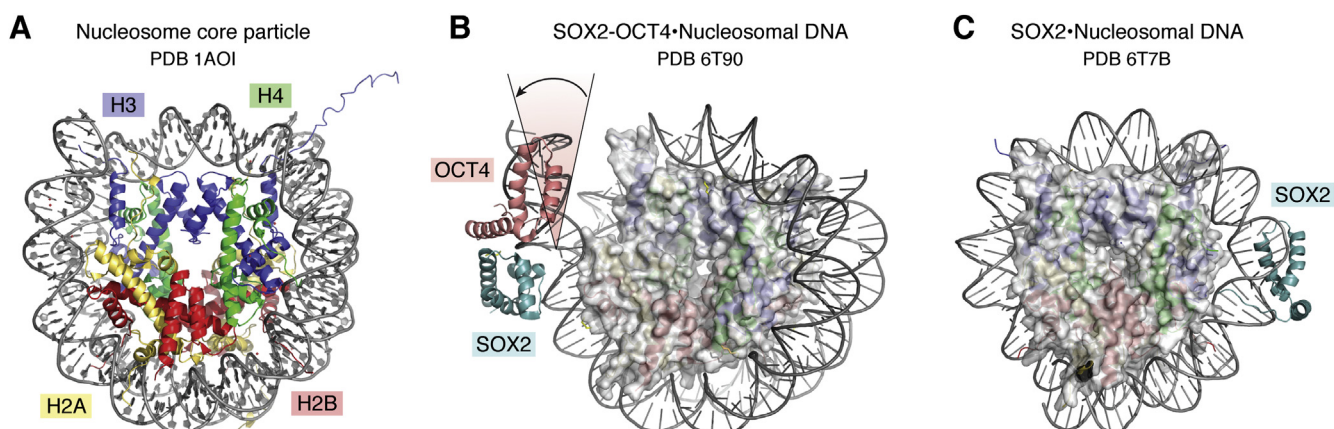


Figure 5. Transcription factor binding to nucleosomal DNA. A, nucleosome core particle. Histones H2A (yellow), H2B (red), H3 (blue), H4 (green) (1AOI). B, Sox2 (blue) and Oct4 (salmon) bound to nucleosomal DNA near the exit site (1T90). Lines and arrow indicate the change in the path of the DNA as compared with the nucleosome in (A). C, Sox2 (blue) bound to an internal site of nucleosomal DNA (6T7B).

non-hydrogen atoms (6ZTJ, 6VU3) (82, 83). The rate at which these huge structures are emerging is so rapid that new structures of the RNA polymerase II preinitiation complex (2.6 MDa) appeared as this article was being revised (84).

The future

What advances in studies of the mechanisms underlying transcription regulation can we anticipate in the near future and how will the PDB continue to support them? As structures have become more complex, it is increasingly common for multiple methods to be used to arrive at the final model. In addition to combining information from X-ray crystallography, NMR, cryo-EM and solution X-ray scattering, data from other biophysical and biochemical methods that provide complementary information are starting to be used to interpret cryo-EM maps of very large complexes with multiple components. This approach, referred to as integrative structural biology (85), incorporates information from methods such as mass spectrometry–cross-linking, hydrogen–deuterium exchange, Förster resonance energy transfer, chromosome capture, and many others. The opportunities presented by integrative structure determination have also presented challenges to the PDB as to how the data should be archived and displayed and how the models should be validated. The PDB has ongoing efforts to address these issues (86).

An important gap in our understanding of how transcription is regulated stems from our inability to capture the dynamics and time-dependent events that link structural snapshots. It is now possible to capture multiple states in a single cryo-EM sample, but linking them temporally involves extrapolation and educated guesswork. Solution and solid-state NMR, along with development of methods for time-resolved structure determination that can be applied to large assemblies, could help fill this gap. There is also the hope that it will eventually be possible to study all these events in their natural, cellular context with further development of cryo-electron tomography (87) or other imaging techniques that have yet to be invented. With so many possibilities ahead, now is as exciting a time in structural biology as when those first few structures of DNA-binding proteins were published 40 years ago.

Acknowledgments—I thank Daniel Panne for providing the coordinates of the β -interferon enhanceosome model. My deepest thanks to all the leaders and staff of the Protein Data Bank, past and present, who have created such a remarkable resource that benefits the entire scientific and education community.

Funding and additional information—Supported by NIGMS, National Institutes of Health grant GM130393 (C. W.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest—The author is a member of the scientific advisory board of Thermo Fisher Scientific.

Abbreviations—The abbreviations used are: CAP, catabolite activator protein; PDB, Protein Data Bank; TBP, TATA-binding protein.

References

- Anderson, W. F., Ohlendorf, D. H., Takeda, Y., and Matthews, B. W. (1981) Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. *Nature* **290**, 754–758
- McKay, D. B., and Steitz, T. A. (1981) Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. *Nature* **290**, 744–749
- Pabo, C. O., and Lewis, M. (1982) The operator-binding domain of lambda repressor: Structure and DNA recognition. *Nature* **298**, 443–447
- Jaskolski, M., Dauter, Z., and Wlodawer, A. (2014) A brief history of macromolecular crystallography, illustrated by a family tree and its nobel fruits. *FEBS J.* **281**, 3985–4009
- Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G., and North, A. C. (1960) Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. *Nature* **185**, 416–422
- Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., and Shore, V. C. (1960) Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å resolution. *Nature* **185**, 422–427
- Blake, C. C., Koenig, D. F., Mair, G. A., North, A. C., Phillips, D. C., and Sarma, V. R. (1965) Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Å resolution. *Nature* **206**, 757–761
- Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P., and Saul, F. (1973) Three-dimensional structure of the Fab' fragment of a human immunoglobulin at 2.8-Å resolution. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3305–3310
- Kim, S. H., Suddath, F. L., Quigley, G. J., McPherson, A., Sussman, J. L., Wang, A. H., Seeman, N. C., and Rich, A. (1974) Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science* **185**, 435–440
- Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F., and Klug, A. (1974) Structure of yeast phenylalanine tRNA at 3 Å resolution. *Nature* **250**, 546–551
- Drew, H., Takano, T., Tanaka, S., Itakura, K., and Dickerson, R. E. (1980) High-salt d(CpGpCpG), a left-handed Z' DNA double helix. *Nature* **286**, 567–573
- Wing, R., Drew, H., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. (1980) Crystal structure analysis of a complete turn of B-DNA. *Nature* **287**, 755–758
- Watson, J. D., and Crick, F. H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737–738
- Hutchison, C. A. 3rd, Phillips, S., Edgell, M. H., Gillam, S., Jahnke, P., and Smith, M. (1978) Mutagenesis at a specific position in a DNA sequence. *J Biol Chem* **253**, 6551–6560
- Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M., and Pabo, C. O. (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature* **298**, 447–451
- Wharton, R. P., Brown, E. L., and Ptashne, M. (1984) Substituting an alpha-helix switches the sequence-specific DNA interactions of a repressor. *Cell* **38**, 361–369
- Barinaga, M. (1989) The missing crystallography data. *Science* **245**, 1179–1181
- Macromolecules, I. U. o. C. C. o. B. (1989) Policy on publication and the deposition of data from crystallographic studies of biological macromolecules. *Acta Cryst.* **A45**, 658
- NIH (1999) *NIH Policy Relating to Deposition of Atomic Coordinates into Structural Databases*, National Institutes of Health, United States
- Greber, B. J., and Nogales, E. (2019) The structures of eukaryotic transcription pre-initiation complexes and their functional implications. *Subcell Biochem.* **93**, 143–192

21. Osman, S., and Cramer, P. (2020) Structural biology of RNA polymerase II transcription: 20 Years on. *Annu. Rev. Cell Dev. Biol.* **36**, 1–34
22. Murakami, K. S. (2015) Structural biology of bacterial RNA polymerase. *Biomolecules* **5**, 848–864
23. Browning, D. F., and Busby, S. J. (2004) The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* **2**, 57–65
24. Weake, V. M., and Workman, J. L. (2010) Inducible gene expression: Diverse regulatory mechanisms. *Nat. Rev. Genet.* **11**, 426–437
25. Jordan, S. R., and Pabo, C. O. (1988) Structure of the lambda complex at 2.5 Å resolution: Details of the repressor-operator interactions. *Science* **242**, 893–899
26. Beamer, L. J., and Pabo, C. O. (1992) Refined 1.8 Å crystal structure of the lambda repressor-operator complex. *J. Mol. Biol.* **227**, 177–196
27. Anderson, J. E., Ptashne, M., and Harrison, S. C. (1987) Structure of the repressor-operator complex of bacteriophage 434. *Nature* **326**, 846–852
28. Aggarwal, A. K., Rodgers, D. W., Drottler, M., Ptashne, M., and Harrison, S. C. (1988) Recognition of a DNA operator by the repressor of phage 434: A view at high resolution. *Science* **242**, 899–907
29. Wolberger, C., Dong, Y. C., Ptashne, M., and Harrison, S. C. (1988) Structure of a phage 434 Cro/DNA complex. *Nature* **335**, 789–795
30. Mondragon, A., and Harrison, S. C. (1991) The phage 434 Cro/ORI complex at 2.5 Å resolution. *J. Mol. Biol.* **219**, 321–334
31. Brennan, R. G., Roderick, S. L., Takeda, Y., and Matthews, B. W. (1990) Protein-DNA conformational changes in the crystal structure of a lambda Cro-operator complex. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 8165–8169
32. Schultz, S. C., Shields, G. C., and Steitz, T. A. (1991) Crystal structure of a CAP-DNA complex: The DNA is bent by 90 degrees. *Science* **253**, 1001–1007
33. Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B., and Pabo, C. O. (1990) Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: A framework for understanding homeodomain-DNA interactions. *Cell* **63**, 579–590
34. Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., and Pabo, C. O. (1991) Crystal structure of a MAT alpha 2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. *Cell* **67**, 517–528
35. Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F., and Sigler, P. B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* **335**, 321–329
36. Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., and Honig, B. (2009) The role of DNA shape in protein-DNA recognition. *Nature* **461**, 1248–1253
37. Somers, W. S., and Phillips, S. E. (1992) Crystal structure of the met repressor-operator complex at 2.8 Å resolution reveals DNA recognition by beta-strands. *Nature* **359**, 387–393
38. Church, G. M., Sussman, J. L., and Kim, S. H. (1977) Secondary structural complementarity between DNA and proteins. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 1458–1462
39. Ellenberger, T. E., Brandl, C. J., Struhl, K., and Harrison, S. C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: Crystal structure of the protein-DNA complex. *Cell* **71**, 1223–1237
40. Ferré-D'Amaré, A. R., Prendergast, G. C., Ziff, E. B., and Burley, S. K. (1993) Recognition by max of its cognate DNA through a dimeric b/HLH/Z domain. *Nature* **363**, 38–45
41. Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S. C. (1992) DNA recognition by GAL4: Structure of a protein-DNA complex. *Nature* **356**, 408–414
42. Ghosh, G., van Duyn, G., Ghosh, S., and Sigler, P. B. (1995) Structure of NF-kappa B p50 homodimer bound to a kappa B site. *Nature* **373**, 303–310
43. Müller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L., and Harrison, S. C. (1995) Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature* **373**, 311–317
44. Sainsbury, S., Bernecky, C., and Cramer, P. (2015) Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 129–143
45. Kim, J. L., Nikolov, D. B., and Burley, S. K. (1993) Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365**, 520–527
46. Kim, Y., Geiger, J. H., Hahn, S., and Sigler, P. B. (1993) Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**, 512–520
47. Pavletich, N. P., and Pabo, C. O. (1991) Zinc finger-DNA recognition: Crystal structure of a zif268-DNA complex at 2.1 Å. *Science* **252**, 809–817
48. Miller, J., McLachlan, A. D., and Klug, A. (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO J.* **4**, 1609–1614
49. Liu, Q., Segal, D. J., Ghiara, J. B., and Barbas, C. F., 3rd (1997) Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 5525–5530
50. Desjarlais, J. R., and Berg, J. M. (1993) Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 2256–2260
51. Rebar, E. J., and Pabo, C. O. (1994) Zinc finger phage: Affinity selection of fingers with new DNA-binding specificities. *Science* **263**, 671–673
52. Greisman, H. A., and Pabo, C. O. (1997) A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **275**, 657–661
53. Kim, Y. G., Cha, J., and Chandrasegaran, S. (1996) Hybrid restriction enzymes: Zinc finger fusions to Fok I cleavage domain. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 1156–1160
54. Porteus, M. H., and Carroll, D. (2005) Gene targeting using zinc finger nucleases. *Nat. Biotechnol.* **23**, 967–973
55. Christian, M., Cermak, T., Doyle, E. L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A. J., and Voytas, D. F. (2010) Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186**, 757–761
56. Mak, A. N., Bradley, P., Cernadas, R. A., Bogdanove, A. J., and Stoddard, B. L. (2012) The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335**, 716–719
57. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509–1512
58. Moscou, M. J., and Bogdanove, A. J. (2009) A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501
59. Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. (2000) An overview of the structures of protein-DNA complexes. *Genome Biol.* **1**, REVIEWS001
60. Matthews, B. W. (1988) Protein-DNA interaction. No code for recognition. *Nature* **335**, 294–295
61. Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., and Mann, R. S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79**, 233–269
62. Rastinejad, F., Perlmann, T., Evans, R. M., and Sigler, P. B. (1995) Structural determinants of nuclear receptor assembly on DNA direct repeats. *Nature* **375**, 203–211
63. Bain, D. L., Heneghan, A. F., Connaghan-Jones, K. D., and Miura, M. T. (2007) Nuclear receptor structure: Implications for function. *Annu. Rev. Physiol.* **69**, 201–220
64. Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., and Sigler, P. B. (1991) Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **352**, 497–505
65. Li, T., Stark, M. R., Johnson, A. D., and Wolberger, C. (1995) Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. *Science* **270**, 262–269
66. Tan, S., and Richmond, T. J. (1998) Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature* **391**, 660–666
67. Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S., and Aggarwal, A. K. (1999) Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* **397**, 714–719
68. Piper, D. E., Batchelor, A. H., Chang, C. P., Cleary, M. L., and Wolberger, C. (1999) Structure of a HoxB1-Pbx1 heterodimer bound to DNA: Role of the hexapeptide and a fourth homeodomain helix in complex formation. *Cell* **96**, 587–597

69. Thanos, D., and Maniatis, T. (1995) Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100
70. Panne, D., Maniatis, T., and Harrison, S. C. (2004) Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. *EMBO J.* **23**, 4384–4393
71. Panne, D., Maniatis, T., and Harrison, S. C. (2007) An atomic model of the interferon-beta enhanceosome. *Cell* **129**, 1111–1123
72. Feng, Z., Verdigué, N., Di Costanzo, L., Goodsell, D. S., Westbrook, J. D., Burley, S. K., and Zardecki, C. (2020) Impact of the Protein Data Bank across scientific disciplines. *Data Sci. J.* **19**, 1–14
73. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., and Richmond, T. J. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260
74. Andrews, A. J., and Luger, K. (2011) Nucleosome structure(s) and stability: Variations on a theme. *Annu. Rev. Biophys.* **40**, 99–117
75. Soufi, A., Garcia, M. F., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K. S. (2015) Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555–568
76. Dodonova, S. O., Zhu, F., Dienemann, C., Taipale, J., and Cramer, P. (2020) Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function. *Nature* **580**, 669–672
77. Michael, A. K., Grand, R. S., Isbel, L., Cavadini, S., Kozicka, Z., Kempf, G., Bunker, R. D., Schenk, A. D., Graff-Meyer, A., Pathare, G. R., Weiss, J., Matsumoto, S., Burger, L., Schübeler, D., and Thomä, N. H. (2020) Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science* **368**, 1460–1465
78. Kuhlbrandt, W. (2014) Biochemistry. The resolution revolution. *Science* **343**, 1443–1444
79. Tan, S., Hunziker, Y., Sargent, D. F., and Richmond, T. J. (1996) Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature* **381**, 127–151
80. Nikolov, D. B., Chen, H., Halay, E. D., Usheva, A. A., Hisatake, K., Lee, D. K., Roeder, R. G., and Burley, S. K. (1995) Crystal structure of a TFIIIB-TBP-TATA-element ternary complex. *Nature* **377**, 119–128
81. Westbrook, J. D., and Bourne, P. E. (2000) STAR/mmCIF: An ontology for macromolecular structure. *Bioinformatics* **16**, 159–168
82. Webster, M. W., Takacs, M., Zhu, C., Vidmar, V., Eduljee, A., Abdelkarim, M., and Weixlbaumer, A. (2020) Structural basis of transcription-translation coupling and collision in bacteria. *Science* **369**, 1355–1359
83. Wang, C., Molodtsov, V., Firlar, E., Kaelber, J. T., Blaha, G., Su, M., and Ebright, R. H. (2020) Structural basis of transcription-translation coupling. *Science* **369**, 1359–1365
84. Chen, X., Qi, Y., Wu, Z., Wang, X., Li, J., Zhao, D., Hou, H., Li, Y., Yu, Z., Liu, W., Wang, M., Ren, Y., Li, Z., Yang, H., and Xu, Y. (2021) Structural insights into preinitiation complex assembly on core promoters. *Science* **372**, eaba8490
85. Rout, M. P., and Sali, A. (2019) Principles for integrative structural biology studies. *Cell* **177**, 1384–1403
86. Sali, A., Berman, H. M., Schwede, T., Trehwella, J., Kleywegt, G., Burley, S. K., Markley, J., Nakamura, H., Adams, P., Bonvin, A. M., Chiu, W., Peraro, M. D., Di Maio, F., Ferrin, T. E., Grunewald, K., *et al.* (2015) Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure* **23**, 1156–1167
87. Turk, M., and Baumeister, W. (2020) The promise and the challenges of cryo-electron tomography. *FEBS Lett.* **594**, 3243–3261