Biophysical Journal, Volume 98

Supporting Material

Tethered-Hopping Model for Protein-DNA Binding and Unbinding Based on Sox2-Oct1-Hoxb1 Ternary Complex Simulations

Peng Lian, Limin Angela Liu, Yongxiang Shi, Yuxiang Bu, and Dongqing Wei

Supporting Materials

for

"Tethered-Hopping Model for Protein-DNA Binding and Unbinding Based on Sox2-Oct1-*Hoxb1* Ternary Complex Simulations"

Peng Lian[1, 2], Limin Angela Liu[2], Yongxiang Shi[1*], Yuxiang Bu[1*], Dongqing Wei[2*]

[1] *School of Life Science and Institute of Theoretical Chemistry, Shandong University, Jinan, P. R. China, 250100*

[2] *Department of Bioinformatics and Biostatistics, College of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, P. R. China, 200240*

[*] Correspondence: shiyx@sdu.edu.cn, byx@sdu.edu.cn, dqwei@sjtu.edu.cn

OVERVIEW OF SOX AND OCT PROTEINS

Sox proteins are an important family of architectural transcription factors that bind to the minor groove of DNA and bend DNA structurally (1-3). They usually recognize a conserved DNA heptamer sequence by their HMG (high-mobity-group) domain (4, 5). The HMG domain consists of about 79 amino acids that form three alpha helices. The DNA bending caused by Sox partially unwinds the DNA duplex to facilitate the binding of additional factors to their respective DNA binding sites. In living cells, Sox proteins usually bind with other transcription factors in forming ternary and higher-order protein-DNA complexes. These complexes exert either synergistic or antagonistic control over cell fate and developmental processes by activating or repressing gene transcriptions, respectively (6, 7).

Sox2 protein, an important member of the Sox family, was found to exert combinatorial gene regulation by binding with different transcription factors in different tissues and cell types during development (7). Sox2 normally recognizes variations of the consensus DNA sequence CTTTGTT (7-9). A recent genomic-scale study (10) suggested that Sox2 shares a large amount of activation sites with Oct3/4, a POU domain protein, and these two transcription factors work in concert in the regulation of gene transcription and maintenance of pluripotency in human embryonic stem cells. Oct1 is another important member of the POU domain family and can also bind with Sox2 to exert combinatorial gene regulation. In contrast to the minor-grooving

binding and DNA-bending ability of Sox proteins, POU domain proteins are a family of major groove binding transcription factors. The POU domain consists of two independent DNA-binding domains, the POU specific domain (POU_S) and the POU homeodomain (POU_H or POU_{HD}), that are connected by a variable linker (11-15). The POU_S domain consists of 75 amino acids that form four alpha helices. The POU_H domain consists of 60 amino acids that form three alpha helices and is similar to other homeodomain proteins in sequence and structure. The POU domain proteins typically recognize the consensus octamer sequence ATGCAAAT (16-19) in which the POU_S domain forms base specific contacts with the 5'-motif ATGC and the POU_H domain recognizes the 3'- motif AAAT. The POU_S and POU_H domains of Oct1 and Oct3/4 proteins share about 58% and 64% sequence identity, respectively. The linker region of Oct3/4 is shorter than that of Oct1 by 7 amino acids. Due to these similarities and differences, Oct1 and Oct3/4 were found to bind to Sox2 in forming many similar ternary complexes with the same DNA control element albeit with often vastly different biological activity of the complex in vivo (20-23).

The discovery of the 3D structures of two Sox2-Oct1-DNA complexes has shed some light on the differences and similarities of protein-DNA recognition and protein-protein interactions. Sox2-Oct1-*FGF4* (24) (PDB entry: 1GT0) and Sox2-Oct1-*Hoxb1* (25) (PDB entry: 1O4X, Fig. S1) are solved using X-ray crystallography and NMR, respectively. The *FGF4* element is nearly identical to the *Hoxb1* element except for one base pair in the binding site of Sox2 and a three base pair insertion between the binding sites of Sox2 and Oct1. The protein-DNA binding interfaces for both the HMG domain and the POU domain are highly similar in both ternary complexes. The HMG domain in each complex forms most of the base specific contacts with DNA using its second alpha helix. The POU_S and POU_H domains both use their third alpha helices for specific DNA binding. Interestingly, due to the extra base pairs in the *FGF4* element, different protein-protein binding interfaces between Sox2 and Oct1 are formed in the two ternary complexes. In the crystal structure of Sox2-Oct1-*FGF4* (24), Sox2 uses its normally disordered C-terminal tail in forming several hydrophobic contacts with the first alpha helix of Oct1's POU_S domain. In the NMR structure of Sox2-Oct1-*Hoxb1* (25), Sox2 uses its third alpha helix in forming a more extensive hydrophobic binding surface with the same alpha helix in Oct1's POU_S domain. In living cells, it is found that only Sox2-Oct1-*Hoxb1* is transcriptionally active (22). We thought that this result may be due to the weak protein-protein interactions that the ternary complex Sox2-Oct1-*FGF4* is unstable and thus transcriptionally inactive in vivo. And we chose the 3D structure of the transcriptionally active ternary complex, Sox2-Oct1-*Hoxb1* (25), as our model system.

We note here that the available 3D structures of Sox2-Oct-DNA ternary complexes were solved using only the DNA binding domains of the Sox2 protein and Oct1 protein. Consequently, these complexes may be more accurately represented as HMG-POU-DNA. As a result, effects on the ternary complex formation due to other parts of the proteins and the interactions of the activation domains on both proteins that are N-terminal and C-terminal to these DNA binding domains as evidenced by the experimental work (20, 21, 23) are both missing in these 3D structures and our simulation.

DETAILED METHODS

Sequences of the protein and DNA chains in the wild type model ternary complex

The sequences of the protein and DNA chains in the wild type model system are shown in Fig. S2. In order to directly compare with experimental work, we follow the same amino acid residue numbering scheme as in the NMR structure (25). The POU_H domain and three DNA base pairs at the 3'- end of the DNA duplex that POU_H binds to in the NMR structure were omitted in the model system. The reason for the removal of POU_H in our study was two-fold. First, the linker region between the POU_S and POU_H domains is unresolved in the NMR structure, and these two domains bind DNA in a relatively independent fashion. Second, the interaction between Sox2 and Oct1 is the focus of the present study. These interactions are confined to the POU_S domain of Oct1 based on the NMR structure.

Mutations in the model ternary complexes

In general practice, mutation of wild type residues to Ala is often done in both experimental and theoretical studies. Occasionally, mutations to other residues, especially those with opposite physical and chemical properties may be of particular interest, as these mutations may lead to more significant findings. In the second model complex ("HMG-**POU_S^M···DNA**"), three amino acids in the POU_S domain that makes base specific contacts with DNA, Gln44, Thr45, and Arg49, were mutated to Glu, Ala, and Ala, respectively. In order to best abolish existent hydrogen bond networks, we mutated Gln44 to Glu44, as Glu44 would maintain the current secondary structure of POU_S, while presenting an unfavorable negative charge for the interaction with DNA. In the third model complex ("**HMG^M···POU_S**-DNA"), we mutated three amino acids in the HMG domain of Sox2 that are important in making hydrophobic contacts with POU_S, Lys59, Arg62, and Met66, to Gly. The reason we didn't mutate the residues to Ala is because Ala is a hydrophobic residue, and would somewhat maintain the hydrophobic

interaction interface, but mutation to Gly would completely abolish the respective favorable hydrophobic interactions. These mutated residues are shown in blue (HMG-**POU_S^M···DNA**) and green (**HMG^M···POU_S**-DNA) in Fig. S2.

Starting HMG-POU_S-DNA structure

The HMG-POU_S-DNA wild type (WT) structure was solvated in TIP3P (26) water molecules in a box of the dimension 70.4×70.9×88.9 $Å^3$. The box size was chosen such that the closest distance between any atoms of the complex from the walls of the box is at least 12 Å. Sodium and chloride ions were added to the system by replacing random water molecules to produce a concentration of 153.8 mM for the two ions. Additional sodium ions were added to obtain neutral charge for the simulation box. The final system consists of 41001 atoms, including 16 DNA base pairs, 75 amino acids of the POU_S domain, 77 amino acids of the Sox2 protein, 12443 TIP3P water molecules, and 67 Na^+ and 50 Cl^- ions.

A 10000-step energy minimization was performed on the water molecules and ions using conjugate gradient and line search algorithm to remove energetically unfavorable contacts. The macromolecules remained fixed during the minimization. The system was then heated to 300 K over 200 ps at constant volume and temperature (300 K) while two harmonic constraints were applied to the complex (20 $kcal·mol^{-1}·Å^{-2}$ on the backbone atoms of HMG and POU_S, 10 $kcal·mol^{-1}·Å^{-2}$ on the backbone atoms of DNA) to allow the water molecules and ions to relax around the macromolecules. Then, a 200 ps equilibration was performed using Nosé-Hoover Langevin piston pressure control (27, 28) and Langevin damping dynamics (29) to keep the system at constant pressure (1 atm) and constant temperature (300 K), respectively. During this equilibration, the harmonic constraints on the backbone atoms of the DNA, HMG domain, and POU_S domain were all set to 10 $kcal·mol^{-1}·Å^{-2}$. The long-range electrostatic forces were evaluated every two time steps using particle mesh Ewald (PME) method (30). The non-bonded cutoff distance was 10 Å. The pair list distance was 12 Å and was updated every four time steps.

Creating a partially dissociated ternary complex structure

In order to compare the dynamics of association for all three model complexes, WT, HMG-**POU_S^M···DNA**, and **HMG^M···POU_S**-DNA, starting from the same geometrical arrangement, we employed steered Molecular Dynamics (SMD) on the solvated and equilibrated HMG-POU_S-DNA system mentioned above to create a partially dissociated HMG-DNA --- POU_S wild type structure. SMD was used because it resembles the natural

dissociation process in yielding a more realistic partially-dissociated complex than simple translation of the POU_S domain away from the complex. Furthermore, SMD often allows macromolecules to return to favorable conformation once the steering force is removed. We applied the constant velocity pulling method (PCV mode) (k = 7 kcal·mol^{-1}·Å^{-2}, v = 0.5 Å·ps^{-1}, t = 200 ps) in the SMD simulation. These parameters were optimized so that the dissociation event could be finished in a relatively short time by computer simulation. The backbone atoms of the DNA and the HMG domain were fixed throughout the SMD simulation. Uniform external steering forces were applied on the backbone atoms of POU_S domain. The side chains of the two proteins and the base pairs of the DNA were allowed to move freely. The external force was applied along the vector that connects the centers of masses of the backbone atoms of the POU_S domain and the atoms of its DNA binding site away from the DNA duplex.

At the end of the SMD simulation, the POU_S domain was pulled to a distance about 3.2 Å away from its bound conformation with DNA. The distance between the mass centers of the POU_S domain and its DNA binding site was at 21.2 Å in this partially-dissociated conformation. In the NMR structure (25), this distance is 18.0 Å. Therefore, the 3.2 Å was obtained from the difference between these two distances. Considering that the typical hydrogen bonds are about 2.5 Å, the SMD procedure has yielded a partially-dissociated complex in which all protein-protein and protein-DNA hydrogen bonds are broken. At this distance, the POU_S domain is in a conformation that is similar to its free-form in water except that the orientation of its out layer atoms is favorable for the reformation of protein-DNA binding with its DNA binding site. Pulling POU_S domain to larger distances led to subsequent association processes that were too slow for molecular dynamics thus were not considered further in this work. This partially-dissociated conformation of the ternary complex became the starting structure for the following association simulation of the WT complex. We also mutated the selected amino acid residues in the HMG domain of the Sox2 protein and the POU_S domain (Fig. S2) to create partially-dissociated mutant complexes for their respective association studies.

To allow the mutant complexes to obtain more favorable conformations due to the amino acid changes, we resolvated the three complexes, WT, HMG-**POU_S^M···DNA** and **HMG^M···POU_S**-DNA, and added sodium and chloride ions to obtain neutral simulation systems and an ionic concentration of NaCl around 150 mM. These systems were then subject to the same minimization, heating, and equilibration molecular dynamics simulation protocol as previously described for the initial preparation of the complex. The final configurations of the equilibrated complexes were then used in the association MD simulations

Starting Point for Dissociation Simulation

A more rigorous way to select the starting point than the method we applied is to obtain well-equilibrated average structures for the three model complexes. For example, the average backbone conformation for a complex can be obtained from the corresponding well-equilibrated association trajectory. The side chains can be re-attached to the average backbone conformation and equilibrated while keeping the backbone fixed. Then a random conformation can be chosen as the starting point of the dissociation simulation. Although our approach outlined in the main text is rather simplistic, later analysis (Table S1) showed that the equilibrated trajectories of the three model complexes are all stable ternary complexes and highly similar in structure. Therefore, the simple approach we used here suffices for the purpose of this study.

Interaction Energy Calculations

The interaction energy of two molecules that form a complex can be described as follows,

$$E_{\text{int}}(\text{A}\cdot\text{B}, t) = P(\text{A-B}, t) - P(\text{A}, t) - P(\text{B}, t) \quad (1)$$

where $E_{\text{int}}(\text{A} \cdot \text{B}, t)$ is the time-dependent interaction energy between molecules A and B, and $P(\text{A-B}, t)$, $P(\text{A}, t)$, and $P(\text{B}, t)$ are the time-dependent potential energies of the bound complex A-B, and unbound molecules A and B, respectively. For molecular dynamics simulation, Eq. (1) implies the need to generate separate simulation trajectories for these three species, A-B, A, and B. In practice, however, a much simpler approach is often taken to obtain this interaction energy(31). If the unbound molecules A and B (free-form) share similar conformational spaces to those in the bound complex A-B (bound-form), one can obtain the time-dependent potential energies for all three molecules from one snapshot of the simulation trajectory of the bound complex A-B. As a result, only the bound complex A-B needs to be simulated.

To test the validity of this simplification, we used the dissociated simulation trajectory (when POU_S is separated from HMG-DNA by >3 Å) to evaluate the conformational differences between the free-form and the bound-form for both POU_S and the HMG-DNA binary complex. Table S2 summarizes the RMSDs for the POU_S backbone atoms and RMSDs for all atoms in the HMG-DNA binary complex, using the NMR structure as a reference. From this table, we can see that the conformational changes for both POU_S and HMG-DNA binary complex were small between their free-form in water and their respective bound-state in the ternary complex. Therefore, the above mentioned simplification is justified for our systems.

When the force field is of the type of pairwise interactions, such as charmm27 used here,

the bonded terms cancel, and the interaction energy becomes simply the non-bonded terms between molecules A and B, as in the following equation,

$$E_{int}(A{\cdot}B, t) = E_{elec}(A\text{-}B, t) + E_{vdw}(A\text{-}B, t) \quad (2)$$

where $E_{elec}(A\text{-}B, t)$ and $E_{vdw}(A\text{-}B, t)$ are the time-dependent electrostatic energy and van der Waals energy between molecules A and B in the bound complex A-B, respectively. When NAMD is used in conjunction with its graphical user interface VMD (32), the NAMDEnergy plugin can be used to generate time series of such interaction energies.

We note here that a limitation of the above method for computing the interaction energy is that the entropic effects are not considered in the energy calculation. Calorimetric work on both minor-groove binding proteins and major-groove binding proteins suggest that the binding free energy of the former is dominated by entropic effects whereas the binding free energy of the latter is dominated by enthalpic effects (33). As POU_S domain is a major-groove binder and the HMG domain remains bound to DNA during all simulations, we chose not to consider the entropic contributions in the present study. One may argue that during the association process, the binding surfaces of the proteins and DNA need to be desolvated first before the protein-DNA binding interface may be formed. Similarly, during the dissociation process, the exposed binding surfaces need to be resolvated. We believe entropy plays a dominant role in these processes, and we choose to consider such entropic effects in a separate study.

ADDITIONAL RESULTS AND DISCUSSION

Structure Comparisons among the Three Model Complexes

Table S2 shows the RMSD values of the backbone atoms in the helical structures of HMG and POU_S with respect to the NMR structure (25) after the DNA backbone atoms in the average structure were superimposed onto those in the NMR structure. The last 1 ns association simulation of the rebound complexes was used for this analysis. Table S2 demonstrates that despite the mutations, the rebound complexes for all three model systems are similar to the NMR structure and to each other, which confirms that stable ternary complexes were reformed after these association simulations. It also confirms that the mutations we chose didn't affect the gross secondary structure of the proteins and maintained the protein-protein and protein-DNA binding conformations of the wild type. This result, together with the distance plot of Fig. 1 of the main text, strongly supports the simple method we used for the selection of starting point for the dissociation simulation.

From Table S2, we also found two interesting trends. First, since the HMG domain binds

DNA using its second alpha helix and POU_S domain binds DNA using its third alpha helix, these helices are found to have the smallest RMSDs (as well as standard deviations of RMSD) compared with other alpha helices. Second, for the POU_S domain, the general order for RMSD values is **HMG^{M}···POU$_S$**-DNA ≥ HMG-**POU$_S^{M}$···DNA** ≥ WT. This trend is weak considering that the RMSD values are large for the three model complexes. Nonetheless, this weak trend can be rationalized as the following. In the HMG-**POU$_S^{M}$···DNA** mutant, the absence of all the base-specific contacts between POU_S and DNA caused the mutant complex structure to be slightly farther away from the wild-type binding position, leading to the somewhat larger RMSD values than the WT complex. In the **HMG^{M}···POU$_S$**-DNA mutant, the perturbation of protein-protein binding by mutations has a stronger effect on the POU_S-DNA binding conformation leading to the largest RMSD values (POU_S being situated farthest away from the wild-type binding position as in Fig. 1 of the main text) among the three model complexes. These results suggest that the protein-protein recognition between HMG and POU_S helps positioning the POU_S domain at its DNA binding surface in the right conformation to facilitate POU_S-DNA binding.

Interaction Energy Plot

Figure S6 clearly demonstrates the magnitude of interaction energy change during the association reaction. Fig. S6A shows that the mutations at the "key amino acids" in the HMG-**POU$_S^{M}$···DNA** mutant greatly increased the electrostatic energy between the POU_S domain and DNA (by about 400 kcal/mol in the rebound complex as in Table 1 column 1 of the main text). Fig. S6D shows that these mutations only caused a slight increase in the van der Waals energy between the POU_S domain and DNA (about 13 kcal/mol in the rebound complex as shown in Table 1 column 2 of the main text). In contrast, Fig. S6E shows that the mutations in the **HMG^{M}···POU$_S$**-DNA mutant mainly increased the van der Waals energy between the POU_S domain and the HMG domain (by about 8 kcal/mol in the rebound complex as in Table 1 column 5 of the main text). We note here that the magnitude of energy change caused by the mutations in the HMG-**POU$_S^{M}$···DNA** complex is rather large, as we have aimed for the abolishment of favorable protein-protein interactions and protein-DNA interactions in the mutant structures. Detailed interaction energy calculations where the contribution from each mutated amino acid was computed show that the Gln44 to Glu44 mutation alone contributes to more than 50% of the loss in binding strength for the HMG-**POU$_S^{M}$···DNA** mutant, indicating that the introduction of the negative charge in Glu caused a severe energy penalty in the mutant. Should more conservative mutations be studied, the corresponding energy change would not

have been so great. Therefore, the behavior in the two mutant model complexes serves as two extreme cases for the energetic study of HMG-POU_S and POU_S-DNA interactions. And it is exactly through such extreme cases that we were able to see significant differences and trends in the kinetic and thermodynamic properties of the model complexes, considering the large standard deviations of these energy terms.

Dissociation Simulations

The difference in the X-axes of the force and energy plots in Figs. 3 and 4 in the main text is explained below. During the 300 ps SMD simulations, forces exerted on the POU_S atoms were evaluated as we pulled POU_S domain away from the HMG-DNA binary complex. From our preliminary work, we found that the dissociation period only had one inflection point or peak for the forces around 100 ps. After that, the force was mainly necessary for moving POU_S domain in solution in one direction. Therefore, during our production runs, we only evaluated forces for the first 150 ps. On the other hand, the energy data points were obtained after the 300 ps SMD simulations were finished, and each trajectory frame was examined for energy evaluation, hence energy data points for the entire 300 ps SMD simulation were obtained.

The distances between POU_S domain and the HMG-DNA binary complex during the beginning phase of the dissociation simulation are shown in Fig. S7 for all three model complexes.

REFERENCES

1. Bewley, C. A., A. M. Gronenborn, and G. M. Clore. 1998. MINOR GROOVE-BINDING ARCHITECTURAL PROTEINS: Structure, Function, and DNA Recognition 1. Annual review of biophysics and biomolecular structure 27:105-131.
2. Grasser, K. D., D. Launholt, and M. Grasser. 2007. High mobility group proteins of the plant HMGB family: Dynamic chromatin modulators. BBA-Gene Structure and Expression 1769:346-357.
3. Weiss, M. A. 2001. Floppy SOX: mutual induced fit in hmg (high-mobility group) box-DNA recognition. Molecular Endocrinology 15:353-362.
4. tros, M., D. Launholt, and K. D. Grasser. 2007. The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. Cellular and Molecular Life Sciences (CMLS) 64:2590-2606.
5. Thomas, J. O., and A. A. Travers. 2001. HMG1 and 2, and related 'architectural'DNA-binding proteins. Trends in Biochemical Sciences 26:167-174.
6. Wilson, M., and P. Koopman. 2002. Matching SOX: partner proteins and co-factors of the SOX family of transcriptional regulators. Current opinion in genetics & development 12:441-446.
7. Kamachi, Y., M. Uchikawa, and H. Kondoh. 2000. Pairing SOX off with partners in the regulation of embryonic development. Trends in Genetics 16:182-187.

8. Pevny, L. H., and R. LovellBadge. 1997. Sox genes find their feet. Current Opinion in Genetics & Development 7:338-344.
9. Wegner, M. 1999. From head to toes: the multiple facets of Sox proteins. Nucleic Acids Research 27:1409-1420.
10. Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. R. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122:947-956.
11. Andersen, B., and M. G. Rosenfeld. 2001. POU Domain Factors in the Neuroendocrine System: Lessons from Developmental Biology Provide Insights into Human Disease 1. Endocrine Reviews 22:2-35.
12. Herr, W., and M. A. Cleary. 1995. The POU domain: versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain. Genes & development 9:1679-1693.
13. Ryan, A. K., and M. G. Rosenfeld. 1997. POU domain family values: flexibility, partnerships, and developmental codes. Genes & development 11:1207-1225.
14. Phillips, K., and B. Luisi. 2000. The virtuoso of versatility: POU proteins that flex to fit. Journal of Molecular Biology 302:1023-1039.
15. Veenstra, G. J. C., P. C. van der Vliet, and O. H. J. Destrée. 1997. POU domain transcription factors in embryonic development. Molecular biology reports 24:139-155.
16. Herr, W., R. A. Sturm, R. G. Clerc, L. M. Corcoran, D. Baltimore, P. A. Sharp, H. A. Ingraham, M. G. Rosenfeld, M. Finney, G. Ruvkun, and H. R. Horvitz. 1988. The Pou Domain - a Large Conserved Region in the Mammalian Pit-1, Oct-1, Oct-2, and Caenorhabditis-Elegans Unc-86 Gene-Products. Genes & Development 2:1513-1516.
17. Verrijzer, C. P., A. J. Kal, and P. C. Vandervliet. 1990. The Oct-1 Homeo Domain Contacts Only Part of the Octamer Sequence and Full Oct-1 DNA-Binding Activity Requires the Pou-Specific Domain. Genes & Development 4:1964-1974.
18. Aurora, R., and W. Herr. 1992. Segments of the Pou Domain Influence One Anothers DNA-Binding Specificity. Molecular and Cellular Biology 12:455-467.
19. Klemm, J. D., and C. O. Pabo. 1996. Oct-1 POU domain DNA interactions: Cooperative binding of isolated subdomains and effects of covalent linkage. Genes & Development 10:27-36.
20. Ambrosetti, D. C., H. R. Scholer, L. Dailey, and C. Basilico. 2000. Modulation of the activity of multiple transcriptional activation domains by the DNA binding domains mediates the synergistic action of Sox2 and Oct-3 on the Fibroblast growth factor-4 enhancer. Journal of Biological Chemistry 275:23387-23397.
21. Nishimoto, M., A. Fukushima, A. Okuda, and M. Muramatsu. 1999. The gene for the embryonic stem cell coactivator UTF1 carries a regulatory element which selectively interacts with a complex composed of Oct-3/4 and Sox-2. Molecular and Cellular Biology 19:5453-5465.
22. Di Rocco, G., A. Gavalas, H. Popperl, R. Krumlauf, F. Mavilio, and V. Zappavigna. 2001. The recruitment of SOX/OCT complexes and the differential activity of HOXA1 and HOXB1 modulate the Hoxb1 auto-regulatory enhancer function. Journal of Biological Chemistry 276:20506-20515.
23. Nishimoto, M., A. Fukushima, A. Okuda, and M. Muramatsu. 2001. The Gene for the Embryonic Stem Cell Coactivator UTF1 Carries a Regulatory Element Which Selectively Interacts with a Complex Composed of Oct-3/4 and Sox-2. Molecular and Cellular Biology 21:978.
24. Remenyi, A., K. Lins, L. J. Nissen, R. Reinbold, H. R. Scholer, and M. Wilmanns. 2003. Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. Genes & Development 17:2048-2059.
25. Williams, D. C., M. L. Cai, and G. M. Clore. 2004. Molecular basis for synergistic transcriptional

activation by Oct1 and Sox2 revealed from the solution structure of the 42-kDa Oct1 center dot Sox2 center dot Hoxb1-DNA ternary transcription factor complex. Journal of Biological Chemistry 279:1449-1457.

26. Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. 1983. Comparison of Simple Potential Functions for Simulating Liquid Water. Journal of Chemical Physics 79:926-935.
27. Nose, S. 1984. A Unified Formulation of the Constant Temperature Molecular-Dynamics Methods. Journal of Chemical Physics 81:511-519.
28. Hoover, W. G. 1985. Canonical Dynamics - Equilibrium Phase-Space Distributions. Physical Review A 31:1695-1697.
29. Brunner, A. T. 1993. X-PLOR, Version 3.1: A system for X-ray crystallography and NMR. Yale University Press, New Haven.
30. Darden, T., D. York, and L. Pedersen. 1993. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. Journal of Chemical Physics 98:10089-10092.
31. Lin, J., N. C. Seeman, and N. Vaidehi. 2008. Molecular-Dynamics Simulations of Insertion of Chemically Modified DNA Nanostructures into a Water-Chloroform Interface. Biophysical Journal 95:1099.
32. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: Visual molecular dynamics. Journal of Molecular Graphics 14:33-38.
33. Privalov, P. L., A. I. Dragan, C. Crane-Robinson, K. J. Breslauer, D. P. Remeta, and C. Minetti. 2007. What Drives Proteins into the Major or Minor Grooves of DNA? Journal of Molecular Biology 365:1-9.
34. Consortium, U. 2008. The universal protein resource (uniprot). Nucleic Acids Res 36:D190-D195.

TABLE S1. Averages and standard deviations of the RMSDs of the POU_S backbone atoms and the RMSDs of all atoms in the HMG-DNA binary complex with respect to the NMR structure in the dissociation simulation trajectory when POU_S is more than 3 Å away from the HMG-DNA binary complex.

Model complex	WT		HMG-**POU_S^M···DNA**		**HMG^M···POU_S**-DNA	
Reactants for association	POU_S	HMG-DNA	POU_S^M	HMG-DNA	POU_S	HMG^M-DNA
RMSDs (Å)	0.89 ± 0.26	0.61 ± 0.09	0.76 ± 0.21	0.66 ± 0.11	0.84 ± 0.25	0.64 ± 0.10

TABLE S2. Average and standard deviation of RMSD values of backbone atoms for the HMG and POU_S domains in the average rebound model complexes with respect to those in the NMR structure. The last 1 ns simulation was used in generating the statistics for each model complex.

RMSD (Å)		WT	HMG-**POU_S^M···DNA**	**HMG^M···POU_S**-DNA
HMG	Helix I	0.42 ± 0.04	0.56 ± 0.04	0.57 ± 0.04
	Helix II	0.38 ± 0.03	0.42 ± 0.03	0.45 ± 0.03
	Helix III	0.67 ± 0.06	0.60 ± 0.06	0.68 ± 0.06
POU_S	Helix I	0.75 ± 0.05	0.82 ± 0.05	0.82 ± 0.06
	Helix II	0.65 ± 0.06	0.70 ± 0.05	0.74 ± 0.04
	Helix III	0.30 ± 0.02	0.34 ± 0.03	0.34 ± 0.03
	Helix IV	0.72 ± 0.07	0.70 ± 0.05	0.78 ± 0.05

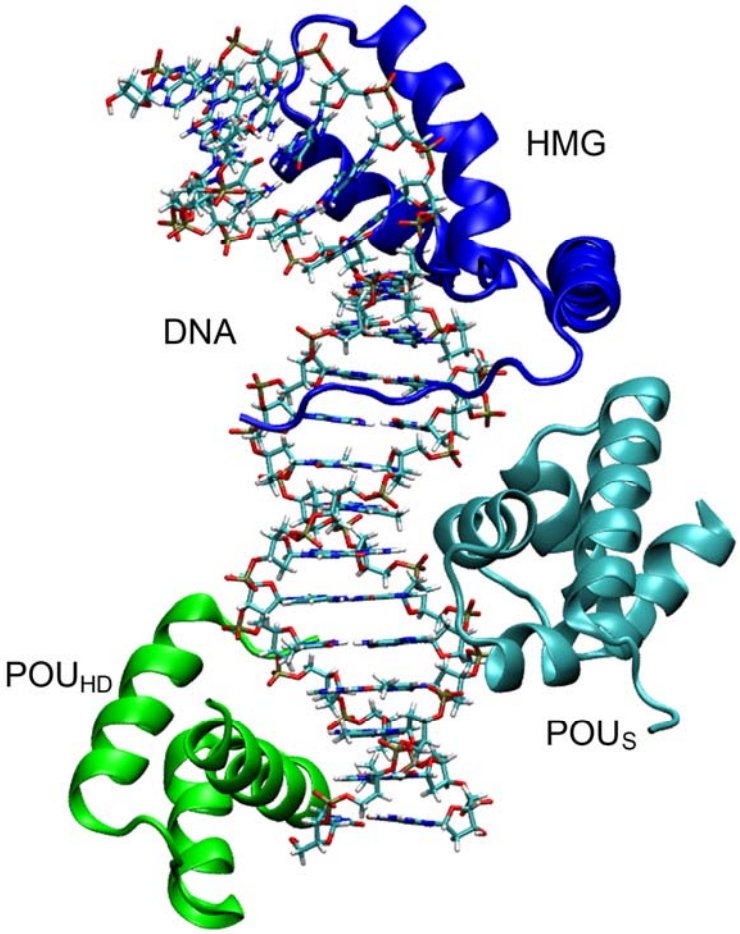

FIGURE S1. NMR Structure of Sox2-Oct1-*Hoxb1* (25). The HMG domain, the POU_S domain, the POU_{HD} domain, and a 19-mer DNA duplex were included in the NMR study. This image rendering was done by Discovery Studio Visualizer 1.7, Accelrys Inc. (http://accelrys.com/products/discovery-studio/visualization/discovery-studio-visualizer.html).

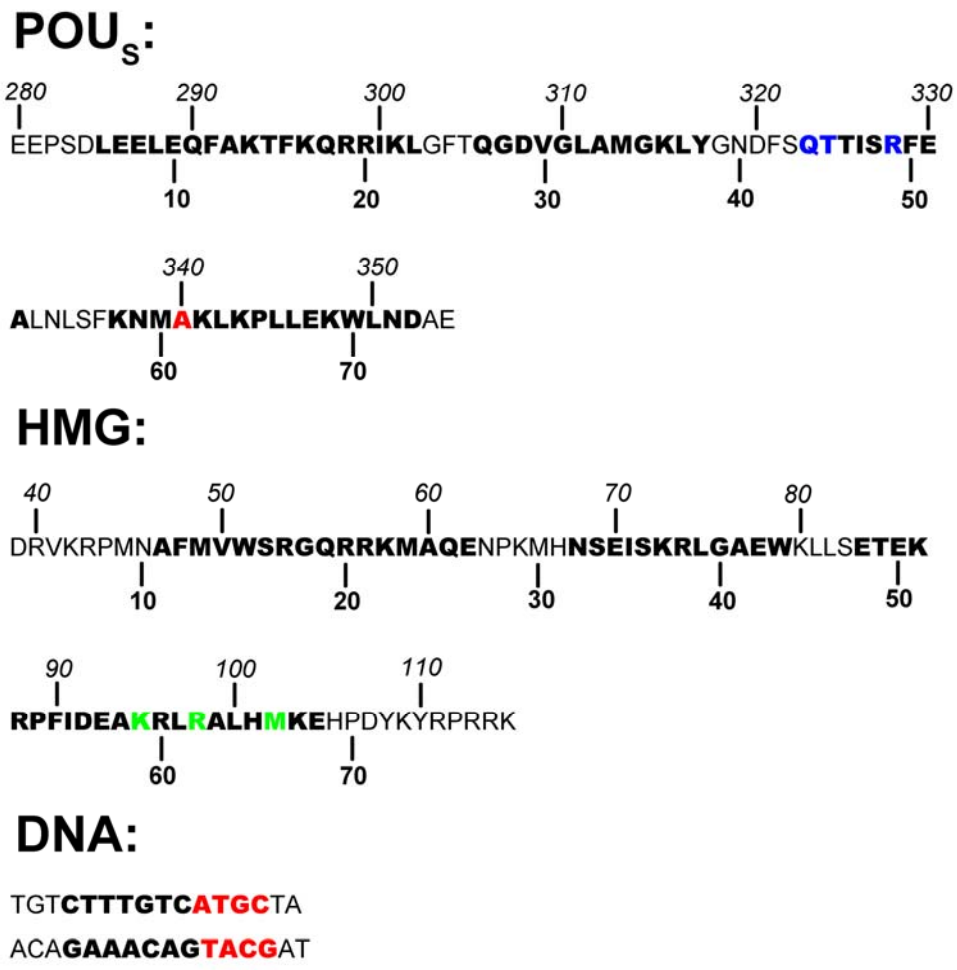

FIGURE S2. The sequences of the protein chains and DNA binding site used in the simulation. The residue numbers in the UniProt database (34) (accession number P14859 for Oct1 protein in human and accession number P48431 for Sox2 protein in human) and in the NMR structure (25) are listed above and below the protein sequences, respectively. Residue number 61 in the POU_S domain (colored red) is a site-mutation from cytosine in the UniProt primary sequence to alanine in the NMR structural study. Bold-faced letters in POU_S and HMG sequences represent the alpha helix structural assignment based on the NMR structure. The bold-faced letters in the DNA sequence represent canonical binding sites for HMG (black) and POU_S (red). The residues in blue in the POU_S domain and those in green in the HMG domain are those mutated in the mutant complex systems for comparison with the wild-type complex.

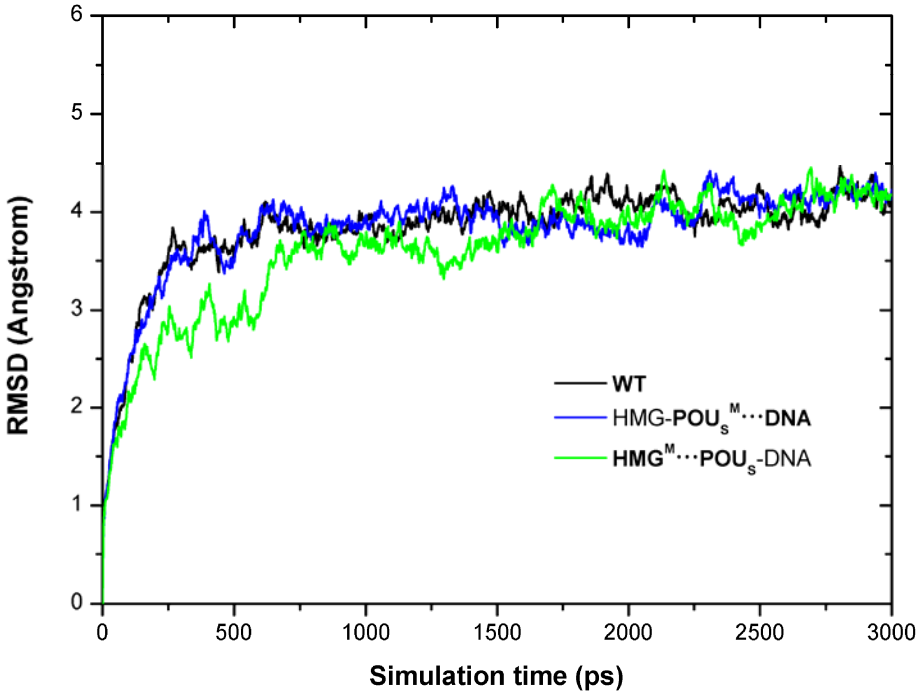

FIGURE S3. RMSDs of three model complexes with respect to the NMR structure during the association reaction. Backbone atoms were used in the RMSD evaluations. The coloring of the figure is the same as Fig. 1 of the main text. The RMSDs for all three complexes increase rapidly and then reach their plateau around 4 Å. The trends of WT and HMG-**POU$_S^M$···DNA** mutant (black and blue) are similar. They both reach equilibration quickly after about 500 ps. The **HMG^M···POU$_S$**-DNA mutant (green) gets equilibrated after about 1500 ps or so. This result suggests that the 3 ns MD simulation of association is adequate for the purpose of our study.

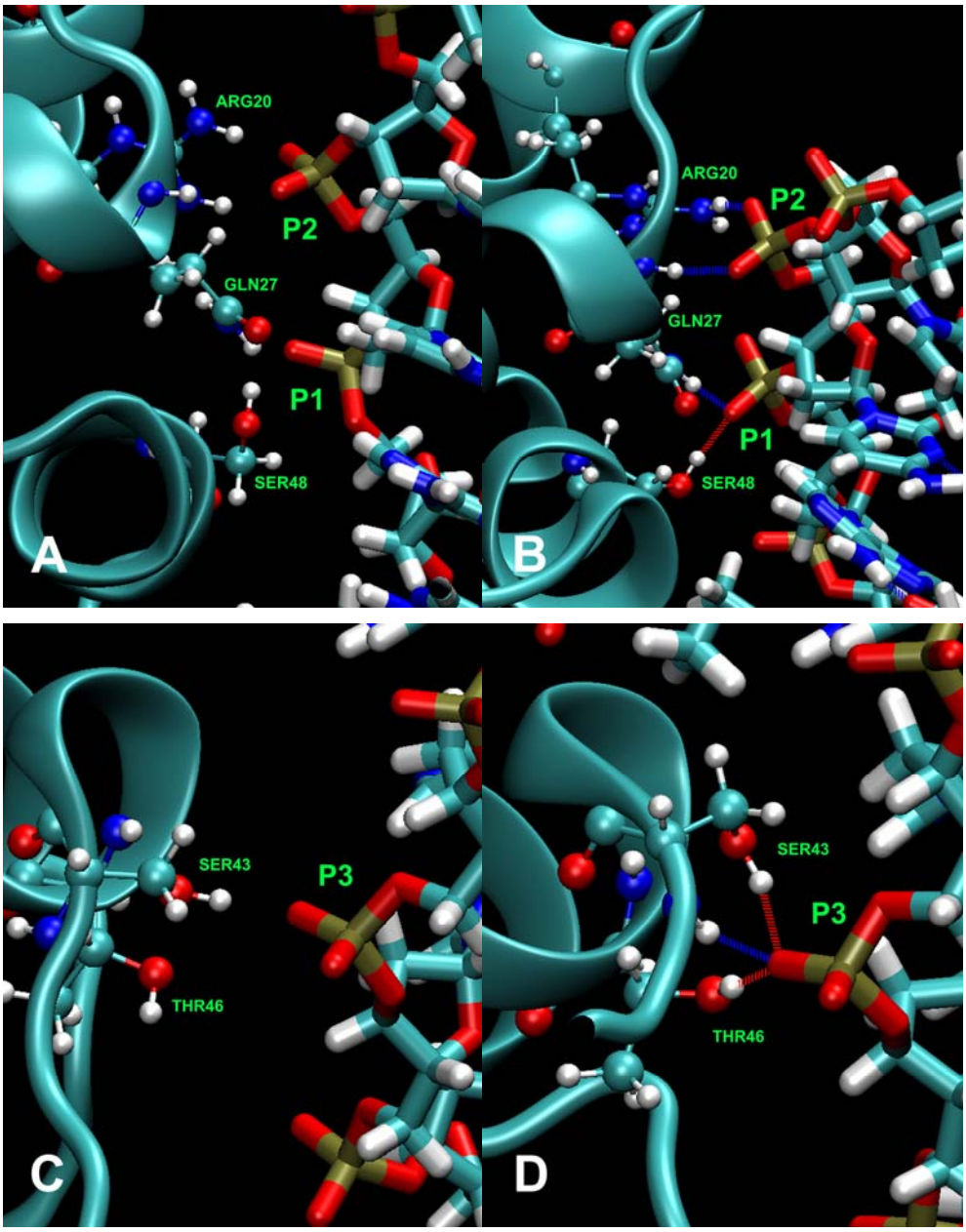

FIGURE S4. Hydrogen-bonding network between the N1 and N2 regions of POU_S domain and the P1, P2, and P3 phosphate groups of the DNA binding site before and after association. Panels A (N1 region) and C (N2 region) show the binding interface between the POU_S domain and the DNA backbone of its binding site before the association simulation. Panels B (N1 region) and D (N2 region) show the same binding interface after the association simulation has reached equilibration. POU_S domain is shown on the left (ribbon representation for backbone atoms and ball-and-stick representation for side chains) whereas the DNA atoms are shown on the right (licorice representation). While no hydrogen bonds exist between the N1 and N2 regions and their corresponding hydrogen bonding partners in DNA (P1 and P2 phosphate groups and P3 group, respectively) before the association, many hydrogen bonds (represented by the dotted lines) have formed at equilibration. The wild-type complex was used for illustration here but similar results were obtained for all three model complexes.

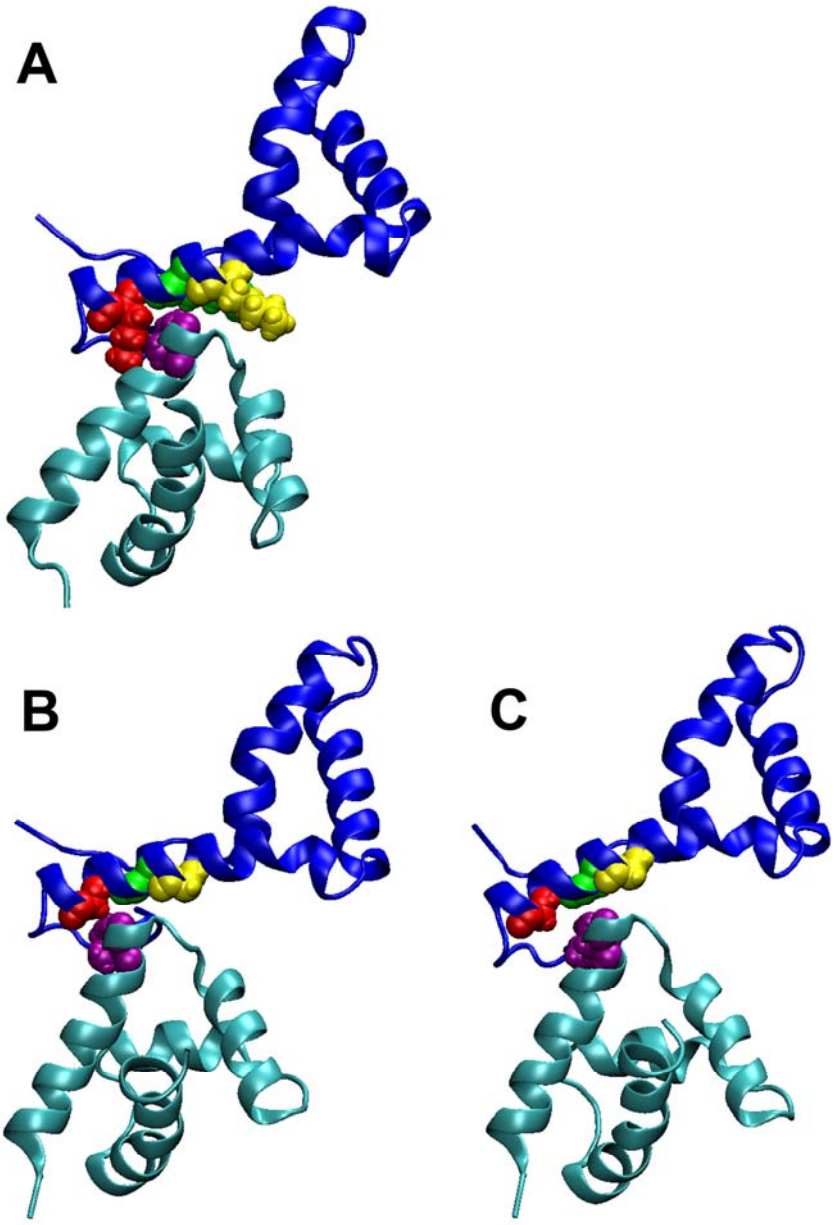

FIGURE S5. Protein-protein interaction between the HMG domain and the POU_S domain. Both proteins are shown in ribbon representation with interfacial amino acids in ball-and-stick representation. The backbones of HMG and POU_S are colored in blue and cyan, respectively. Panel A shows the shape complementarity between these two proteins. In the WT and HMG-**POU_S^M···DNA** complexes, Lys59 (yellow), Arg62 (green) and Met66 (red) on the third alpha helix of HMG form a concave in which Ile21 (purple) of the POU_S domain fits snugly. This interaction forms the core of the hydrophobic interactions observed in the NMR structure. Panels B and C are sample snapshots of the association simulation for the **HMG^M···POU_S**-DNA mutant at 1.2 ns and 1.9 ns. In this mutant, Lys59, Arg62 and Met66 of the HMG domain were replaced by glycine, and the POU_S domain swept along the mutant Sox2 binding surface in a back and forth motion for about 1 ns before finally binding to the HMG-DNA complex.

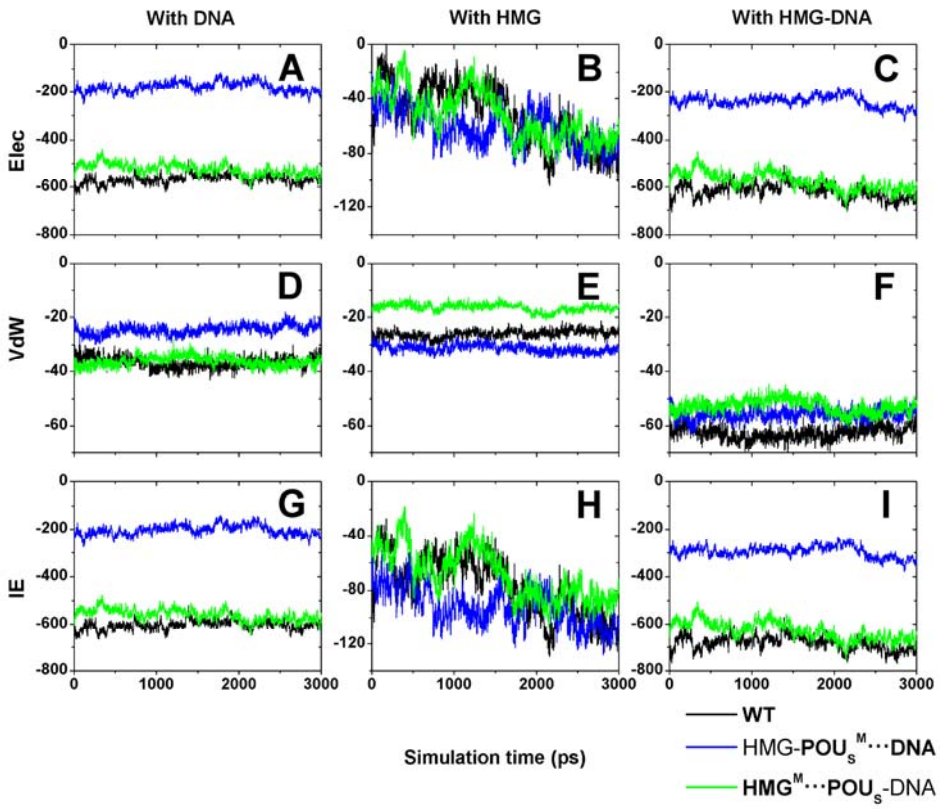

FIGURE S6. Time-dependent interaction energy between POU_S and the HMG-DNA binary complex during the association reaction for all three model complexes. The columns are for energy terms between the POU_S domain with DNA, HMG, and HMG-DNA binary complex, respectively. The rows are for electrostatic energy (Elec), van der Waals energy (VdW), and the interaction energy (IE) which is the sum of these two energy terms, respectively. All energy values are in unit kcal/mol. The X axes in all panels represent the simulation time for the entire duration of the association reaction whereas the Y axes represent the energy values. The curves are colored as in Fig. 1 of the main text. The data shown here demonstrate that the "key amino acids" in the POU_S domain that form the base specific contacts with its DNA binding sites contribute significantly to the total interaction energy of the ternary complex due to the large electrostatic energy between POU_S and DNA. On the other hand, mutations on the HMG domain at the protein-protein binding interface between HMG and POU_S have a relatively small effect on the total interaction energy due to the small van der Waals energy term at the hydrophobic protein-protein interface.

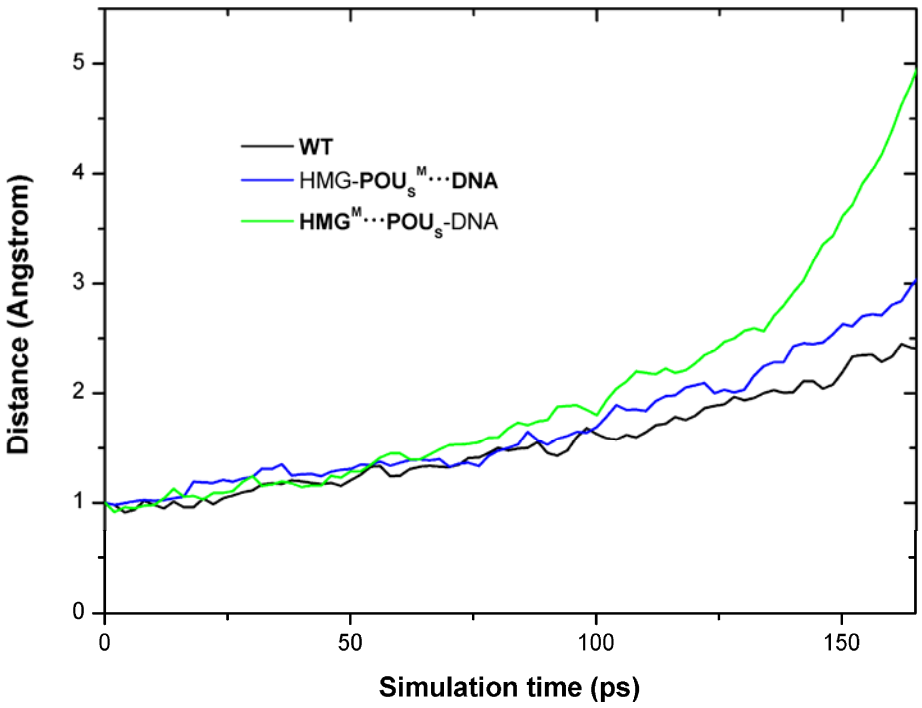

FIGURE S7. Distance of the POU_S domain from its DNA binding position in the NMR structure as a function of simulation time during dissociation for the three model complexes. The distance values were obtained using the same method as in Fig. 1 of the main text. The coloring of the curves is the same as in Fig. 1 of the main text. The dissociation happens fastest in the **HMG^M···POU_S**-DNA mutant, and slowest in the WT complex. This observation is consistent with the force and interaction energy results in Figs. 3 and 4 of the main text.