bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima

Ekaterina Morgunova[1], Yimeng Yin[1], Arttu Jolma[1], Alexander Popov[3], You Xu[2], Lennart Nilsson[2] and Jussi Taipale[1, 4, 5]

[1]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE 171 77 Stockholm, Sweden

[2]Department of Bioscience and Nutrition, Karolinska Institutet, SE 141 83 Huddinge, Sweden

[3]European Synchrotron Radiation Facility, Grenoble, France

[4]Genome-Scale Biology Research Program, P.O. Box 63, FI-00014 University of Helsinki, Finland

[5]Department of Biochemistry, University of Cambridge, United Kingdom

address correspondence to:

ajt208@cam.ac.uk

ekaterina.morgunova@ki.se

Lead Contact: Dr. Jussi Taipale, ajt208@cam.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**SUMMARY**

**Most transcription factors (TFs) are thought to recognize a single optimal sequence, with mutations to this sequence decreasing binding in a multiplicative manner. However, some TFs display epistasis, with multiple substitutions to their optimal site alleviating each other's effects. These TFs can bind to two distinct sequences that represent two local optima in the Gibbs free energy of binding (ΔG). To determine the molecular mechanism behind this effect, we solved the structures of human HOXB13 and CDX2 proteins bound to their two optimal DNA sequences, CAATAAA and TCGTAAA. Striking differences were observed in the recognition of the distinct part of the sequence. Whereas the CAA trinucleotide was recognized by a network of direct and water-mediated hydrogen bonds, no such interactions were visible in the TCG structures of both TFs, suggesting that the solvent molecules at the interface were disordered. Thermodynamic analyses by isothermal titration calorimetry (ITC) revealed that both sites were bound with similar ΔG. However, the interaction with the CAA sequence was driven by change in enthalpy (ΔH), whereas the TCG site was bound with similar affinity due to a smaller loss of entropy (ΔS). Additional analysis of BARHL2 and MYF5 using ITC confirmed the presence of entropic and enthalpic optima also for these TFs that can recognize two distinct sequences. The common presence of at least two local optima is general to all macromolecular interactions, as ΔG depends on two partially independent**

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**variables ΔH and ΔS according to the central equation of thermodynamics, ΔG = ΔH - TΔS.**

**Keywords:** transcription, transcription factor, DNA recognition, HOXB13, CDX2, thermodynamics, enthalpy, entropy, epistasis.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## INTRODUCTION

The binding of transcription factors (TFs) to their specific sites on genomic DNA is a key event regulating cellular processes. Analysis of structures of known TFs bound to DNA has revealed three different mechanisms of recognition of the specifically bound sequences: **1)** the "direct readout" mechanism involving the formation of specific hydrogen bonds and hydrophobic interactions between DNA bases and protein amino acids (*1-3*); **2)** "indirect readout" of the DNA shape and electrostatic potential (*4-8*) by protein contacts to the DNA backbone or the minor groove, and **3)** water mediated interactions between bases and amino-acids (*9-15*). Each of these mechanisms contributes to binding specificity of most TFs, with their relative importance varying depending on the TF and the recognized sequence.

The modes of DNA recognition differ from each other also in their thermodynamic characteristics. For example, direct hydrogen bonds can contribute strongly to enthalpy of binding, whereas indirect hydrogen bonds mediated by water are weaker, due to the loss of entropy caused by immobilization of the bridging water molecule. In many cases, the contributions of the loss of entropy and the gain in enthalpy are similar in magnitude, leading to a phenomenon called "enthalpy-entropy compensation" (*13, 16-18*). This leads to binding promiscuity, allowing a TF to bind to several different but closely related sequences with a biologically relevant affinity.

Many transcription factors appear to only recognize sequences closely related to a single optimal site. Their binding to DNA can be approximated by a position weight

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

matrix (PWM) model, which describes a single optimal site, and assumes that individual substitutions affect binding independently of each other. Thus, the combined effect of multiple mutations is predictable from the individual effects. However, some TFs have been shown to bind with high affinity to multiple different sequences, and populations of sequences that are closely related to these optimal sites (*19-23*). In such cases, the effect of substitution mutations is not independent, and instead the mutations display strongly epistatic behavior (*24, 25*), where the combined effect of two mutations can be less severe than what is predicted from the individual effects. Many cases of such multiple specificity can be explained by different spacing of homodimeric TFs (*2*), but in some cases a single monomeric TF appears to be able to bind to two distinct sequences with similar affinities.

The molecular mechanism behind the phenomenon has not been understood. To elucidate the mechanism, we performed structural analysis of two homeodomain proteins, the posterior homeodomain protein HOXB13, and the parahox protein CDX2, each bound to two distinct high-affinity sequences. The optimal sequences for HOXB13 are CCAATAAA and CTCGTAAA that differ from each other by the three underlined base pairs, whereas CDX2 binds with high affinity to similar two sequences that begin with a G instead of a C.

This analysis, together with thermodynamic measurements of HOXB13, CDX2 and two other transcription factors, BARHL2 and MYF5 that also display multiple specificity revealed that in each case, one of the optimal sequences is bound primarily due to an optimal enthalpic contribution, whereas the other is bound due to an optimum

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of entropy. This result is likely to be general to most macromolecular interactions, as they commonly involve interaction of the macromolecules with a network of interconnected water-molecules, whose formation involves a trade-off between enthalpy and entropy.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## RESULTS AND DISCUSSION

### Modeling the binding of many TFs requires more than one PWM model

Many TFs have been reported to display multiple specificity, including many biologically important transcription factors such as the MYF family of basic helix-loop helix factors (*26*), the nuclear receptor HNF4A (*19*), and the homeodomain proteins BARHL2, CDX2 and HOXB13 (*23, 27*). Analysis of enrichment of subsequences by MYF6, BARHL2, CDX2 and HOXB13 in SELEX reveals that a single PWM model cannot describe the binding affinity of these factors to DNA (**Figure 1A-D**). Each of these factors has more than one locally optimal sequence. All sequences between these optima have lower affinity and enrich less in SELEX than the optimal sequences. Therefore, more than one positionally independent position weight matrix (PWM) model is required for describing their affinity towards DNA (**Figure 1**).

Combinations of mutations affecting the optimal sites of these TFs display extremely strong epistatic effects. For example, the effect of mutating three first bases of the optimal HOXB13 motif TCGTAAAA is more than 400-fold smaller than what is expected from the individual single mutants (**Figure 1E**, **F**), and the generated CAATAAAA site binds to HOXB13 with almost the same affinity as the initial unmutated sequence.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Structural analysis of HOXB13 and CDX2 bound to DNA$^{TCG}$ and DNA$^{CAA}$**

To understand the molecular basis of the epistatic effect, we decided to solve the structure of HOXB13 and CDX2 bound to their two optimal sequences. These proteins are related, but diverged significantly in primary sequence, showing 43% identity at amino-acid level (**Figure S1 and S2**). For structural analysis, the DNA-binding domains (DBD) of HOXB13 (the 75 amino-acids Asp-209 to Pro-283) and CDX2 (the residues Arg-154 – Gln256) were expressed in *E.coli*, purified and crystallized bound to synthetic 19 or 18 bp double stranded DNA fragments containing the CTCGTAAA/GTCGTAAA (DNA$^{TCG}$) or CCAATAAA/GTCGTAAA (DNA$^{CAA}$) motifs, respectively. These core sequences were obtained by PBM (*28*) and HT-SELEX (*29*), and validated by ChIP-seq experiments (*26*), and represent the two distinct binding sites of HOXB13 and CDX2 (**Figure 2**). The structures were solved using molecular replacement at resolutions 3.2 and 2.2 Å for HOXB13, and 2.57 and 2.95 Å for CDX2, respectively.

All complexes displayed a high overall similarity to HOXB13 bound to methylated DNA (*26*), and to the previously known DNA-bound HOX protein structures (*7, 30-34*) (**Figure 2**; **Figure S1**). Two parts of both HOXB13 and CDX2 DBDs interact with DNA: the recognition helix α3, which tightly packs into the major groove, and the N-terminal tail interacting with the minor groove (**Figure 2A**, **C**). The residue Gly-84 that is affected by a coding variant that is strongly implicated in prostate cancer was not included in our construct; two other residues mutated in single prostate cancer families (*35*) were predicted to destabilize the protein, or its interaction with DNA (**Figure S2**).

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The core interactions between both HOXB13 and CDX2 DBDs and DNA are similar to those known from earlier structures (*7, 30-34*). The TAAA sequence characteristic of the posterior homeodomains is recognized by a combination of a direct hydrogen bond to the $A_{10}$ base opposite of the T, $A_{10}$ base on the other stand and an insertion of the N-terminal basic amino-acids to the narrow minor groove induced by the stretch of four As. The overall protein structure in the four complexes is highly similar, showing only minor differences in the conformation of the N-termini (**Figure 2A**, **C**). The most remarkable difference between the complexes is in the conformation of DNA of the HOXB13-$DNA^{TCG}$ complex at the position of the divergent bases (**Figure 2A**, **C**; **Figure S3**). To quantitate the shape of the DNA in the protein binding region we determined the helicoidal parameters using the program Curves+ (*36*), and found the most prominent differences between the two complexes were in twist, shift, slide, X- and Y-displacement, minor groove width, and major groove depth at the positions of the divergent CAA and TCG sequences (**Figure S3B**). The $DNA^{TCG}$ backbone is bent towards the major groove, facilitating contact with Arg-258 of the recognition helix with the DNA backbone. The corresponding contact (Arg-228 to backbone) is also observed in both CDX2 structures. In contrast, the $DNA^{CAA}$ backbone is bent towards the minor groove, leading to a contact with N-terminal Arg-217 and Lys-218 (**Figure 2B**; **Figure 3**). Instead of contacting the DNA backbone, Arg-258 assumes an alternative conformation in which it turns inside of the major groove, forming a water-mediated contact with Gln-265. The Gln-265, in turn, recognizes $C_{6'}$ via a direct hydrogen bond. In addition, the CAA sequence is recognized by a hydrophobic interaction between Ile-262

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and the $T_{11}$ methyl group. In CDX2 complexes the DNA bend in CDX2:DNA$^{TCG}$ is slightly smaller due to the replacement of Thr-261 with Lys-231 which does not allow the alternative conformation of Arg-228. The other contacts in CDX2:DNA complexes are very similar to those listed for HOXB13:DNAs.

In order to understand the role of individual residues in binding of specific DNA we created 48 different single and combined mutations in DBD of HOXB13. The resulting data are presented in **Figures 2E** and **S4**. The replacement of Thr-261 either as a single mutation or in combination with any other amino-acids resulted in changing of HOXB13 specificity from Ctcg/Ccaa towards the sequence recognized by CDX2 (Gtcg/Gcaa; **Figure 2E**, left panel). No substitutions were identified that would lead to a specific loss of binding to the TCG or CAA sequences. However, several mutations affecting backbone contacts between HOXB13 amino-acids and DNA 5' of the divergent trinucleotide moderately increased the relative affinity towards the CAA sequence (**Figure 2E**, right panel).

Analysis of the mutation data together with the structures revealed that amino-acids involved in the protein-DNA interface formation cannot fully explain the specificity preferences of HOXB13 and CDX2. The lack of direct interactions between protein and DNA in this region instead suggests that the specificity would be conferred in part by bridging water-molecules located at the protein-DNA interface.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Role of water molecules in the protein-DNA interface**

The main difference between the complexes with $DNA^{CAA}$ and $DNA^{TCG}$ is revealed by analysis of the bridging water molecules. The HOXB13-$DNA^{CAA}$ structure (2.2 Å) contains chains of water molecules that interact with both HOXB13 amino-acids and each of the DNA bases in the CAA sequence (**Figure 4A-C**). In contrast, no water molecules are visible in the HOXB13-$DNA^{TCG}$ structure, despite the 3.2 Å resolution that should allow identification of strongly bound water molecules as well as much fewer water molecules are found in the complex CDX2:$DNA^{TCG}$. A relatively large solvent channel (6.4 Å in smallest diameter) exists between the α3 helix of HOXB13 and DNA (**Figure 4D**). The electron density in this region is low ($\sigma < 0.5$), similar to that found in the surrounding solvent, indicating that the water-molecules in this region are highly mobile. Thus, the optimal binding of HOXB13 to the CAA sequence can be rationalized by the visible interactions that contribute to the enthalpy of binding (ΔH). In contrast, no such interactions can be identified that could explain the preference of HOXB13 to the TCG trinucleotide. The absence of ordered solvent molecules, and the lower resolution of the HOXB13-$DNA^{TCG}$ structure is consistent with the possibility that the TCG sequence is preferred because it represents a relatively disordered, high entropy state. In complex of CDX2:$DNA^{TCG}$ with high resolution (2.57 Å) the water molecules were well visible but they did not form the corresponding water-chains (**Figure 4E**, **F**) supporting the idea of entropically driven binding.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Thermodynamic features of the protein-DNA interactions**

We next performed molecular dynamics simulations and free energy perturbation calculations to probe the behavior of water molecules in the protein-DNA interface for the two optimal sequences for HOXB13. The relative free energy (*37*) estimates for the affinities of HOXB13 for the two DNA sequences obtained from the simulations indicate that both sequences are bound with similar affinities ($\Delta\Delta G$ = -0.1 kcal/mol). Analysis of the mobility of water molecules at the protein-DNA interface revealed that, while there is a similar number of water molecules in both systems, the waters at the HOXB13-DNA$^{TCG}$ interface are more mobile (**Figure S5**), consistent with a model where this complex has higher entropy than the HOXB13-DNA$^{CAA}$ complex.

To more directly test if the two states are driven by enthalpy and entropy, we measured these thermodynamic parameters using isothermal titration calorimetry (ITC). ITC directly measures the heat of binding ($\Delta H$) and $K_d$ of the binding reaction. Conversion of the $K_d$ to $\Delta G$ then allows the inference of the entropy of binding ($\Delta S$) from the data. The measured thermodynamic parameters for the TCG site were very similar to those we reported previously (Yin et al., 2017). Comparison of the parameters for the TCG and CAA sites revealed that consistent with SELEX (*29*) and molecular modeling data, the $\Delta G$ values for both sequences were similar. However, as predicted, the CAA site displayed much higher change in enthalpy, and larger loss of entropy compared to those of the TCG site (**Figure 5A**, **B**). These results indicate that HOXB13 binding to one optimal site, CAA, is driven by enthalpy, whereas strong binding to the other, TCG, is due to a lower loss of entropy.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To test if the identified mechanism is general to other cases of multiple specificity, we used ITC to determine the thermodynamic parameters for CDX2 and two other TFs, the MYF family TF MYF5 and the homeodomain protein BARHL2, both of which can optimally bind to two distinct sequence populations. Analysis of the data confirmed that in both cases, the ΔG values for the two optimally bound sequences were similar, whereas the relative contributions of entropy and enthalpy to the binding were strikingly different (**Figure 5C** to **H**). These results suggest that the ability of some TFs to bind to two distinct sequences with high affinity can be caused by the presence of both an enthalpic and an entropic optima.

## Conclusions

In drug development, multiple optimal compounds can often be found that bind to a particular target molecule (*18*). However, biological macromolecules are composed of a small set of relatively large monomers, and thus populate the shape-space more sparsely than synthetic small molecules, which can be modified at the level of single atoms. Therefore, the finding that TFs can bind to two distinct DNA sequences with equal affinity was unexpected (*28, 29*), and has been controversial in the field. Our initial hypothesis was that the two optimal states could be due to an ability of the TF to adopt two distinct conformational states, or due to a similarity of the shapes of the two distinct DNA sequences. To address these hypotheses, we solved the structure of HOXB13, a central transcription factor involved in both development (*38-40*) and tumorigenesis (*35, 41, 42*) bound to its two optimal DNA sequences. Surprisingly, the conformational

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

differences between the HOXB13 proteins in the two structures were minor, and the same effect we observed in two complexes of CDX2. In addition, the shape and charge-distribution of the optimally bound DNA sequences were not similar to each other. Thus, the structural analysis failed to support either the dual protein conformation or the DNA shape similarity models. Instead, thermodynamic analyses of HOXB13, CDX2, BARHL2 and MYF5 revealed that the two optimal states were bound because of their distinct effects on enthalpy and entropy, principally caused by differential stability of the water network at the protein-DNA interface.

The mechanism by which TFs bind to two optimal DNA sequences is fundamental, and applies to all macromolecular interactions. In principle, enthalpy and entropy of binding vary partially independently as a function of the shape and charge distribution of the interacting molecules. Thus, different sequences are likely to be optimal with respect to enthalpy and entropy, in such a way that one optimal sequence is close to the optimal enthalpy state, and another is close to the optimal entropy state (**Figure 6**). We are not aware of work that has previously described such a situation. The observed effect that commonly results in at least two distinct ΔG minima may have been missed before because it is generally strongest when there are solvent molecules at the interacting surface. Interacting through solvent molecules can increase enthalpy of binding, but also causes a large loss of entropy due to fixing of the solvent molecule (s). However, in macromolecular interactions that are driven by direct contact between residues, entropy generally has lower impact on binding than enthalpy, and thus one of the optima is at a higher ΔG than the other. Another reason for overlooking this

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

mechanism could have been the fact that multiple local optima can also exist via other mechanisms (see for example (*18*)), and measurements that allow inference of entropy and enthalpy separately are not commonly performed in studies of macromolecular interactions. In addition, simple additive binding models such as position weight matrices (PWMs) can hide the effect, as they can only describe a single optimal state.

The cases we studied here represent some of the strongest deviations from the PWM model, and also present two optima of very similar ΔG that are located relatively far from each other in sequence space. It is likely that many other biologically relevant examples will be identified where the sequences representing entropic and enthalpic optima are more closely related to each other. This would manifest as a "flat bottom" in the affinity landscape, where many sequences would bind with similar affinity. In addition, situations may be found where one of the local optima is located at lower affinity than the other. This would manifest as a minor peak or a shoulder in the affinity landscape farther from the optimal sequence. In each case, the measured affinities would deviate from those predicted from a single PWM model. Our results and the underlying theory suggest that the ability of TFs to bind to distinct sequences could be widespread, and that the importance of the optimal states in determining TF-DNA binding preferences should be reinvestigated. In addition, models for TF binding that are used to identify TF sites should also be adjusted to include features that allow two or more optima. In a broader sense, our results are relevant to all macromolecular interactions, particularly in the presence of a polar solvent such as water that can contribute to bridging interactions, whose contributions to the enthalpy and entropy of binding are in the same order of

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

magnitude. Therefore, in addition to explaining the observed epistasis in protein-DNA interactions, the presence of two optima is likely to also explain the molecular mechanisms behind other types of genetic epistasis.

## Author Contributions

E.M. and J.T. designed the project; E.M. performed crystallization, solved and refined crystal structures, performed ITC experiments, analyzed the results, A.P. collected data in ESRF, Y.Y. and A.J. performed HT-SELEX experiments and data analysis, L.N and Y.H. performed and analyzed molecular dynamics simulations, E.M. and J.T. prepared the figures and wrote the manuscript

## Acknowledgments

The authors thank Drs. Minna Taipale, Inderpreet Sur, Bernhard Schmierer, Eevi Kaasinen, Sten Linnarsson and Johan Elf for the critical review of the manuscript, Karolinska Institutet Protein Science Facility and Sandra Augsten for protein production, as well as Lijuan Hu and Anna Zetterlund for technical assistance. This work was supported by the Center for Innovative Medicine at Karolinska Institutet, Cancerfonden, the Knut and Alice Wallenberg Foundation and the Swedish Research Council.

## Accession codes

The atomic coordinates and diffraction data have been deposited to Protein Data Bank with the accession codes 5EDN and 5EEA, for HOXB13-DNA$^{TCG}$ and HOXB13-

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$DNA^{CAA}$, respectively and 6ES3 and 6ES2 for CDX2- $DNA^{TCG}$ and CDX2- $DNA^{CAA}$, respectively.

## Competing Financial Interests

The authors declare no competing financial interests.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## REFERENCES


1. J. E. Anderson, M. Ptashne, S. C. Harrison, Structure of the repressor-operator complex of bacteriophage 434. *Nature* **326**, 846-852 (1987).
2. A. K. Aggarwal, D. W. Rodgers, M. Drottar, M. Ptashne, S. C. Harrison, Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242**, 899-907 (1988).
3. C. Wolberger, Y. C. Dong, M. Ptashne, S. C. Harrison, Structure of a phage 434 Cro/DNA complex. *Nature* **335**, 789-795 (1988).
4. J. Hizver, H. Rozenberg, F. Frolow, D. Rabinovich, Z. Shakked, DNA bending by an adenine--thymine tract and its role in gene regulation. *Proc Natl Acad Sci U S A* **98**, 8490-8495 (2001).
5. R. Lavery, Recognizing DNA. *Q Rev Biophys* **38**, 339-344 (2005).
6. R. Rohs, H. Sklenar, Z. Shakked, Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure* **13**, 1499-1509 (2005).
7. R. Joshi *et al.*, Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131**, 530-543 (2007).
8. I. Dror, T. Zhou, Y. Mandel-Gutfreund, R. Rohs, Covariation between homeodomain transcription factors and the shape of their DNA binding sites. *Nucleic Acids Res* **42**, 430-441 (2014).
9. M. Bastidas, S. A. Showalter, Thermodynamic and structural determinants of differential Pdx1 binding to elements from the insulin and IAPP promoters. *J Mol Biol* **425**, 3360-3377 (2013).
10. M. M. Garner, D. C. Rau, Water release associated with specific binding of gal repressor. *Embo J* **14**, 1257-1263 (1995).
11. J. E. Ladbury, J. G. Wright, J. M. Sturtevant, P. B. Sigler, A thermodynamic study of the trp repressor-operator interaction. *J Mol Biol* **238**, 669-681 (1994).
12. C. J. Morton, J. E. Ladbury, Water-mediated protein-DNA interactions: the relationship of thermodynamics to structural detail. *Protein Sci* **5**, 2115-2118 (1996).
13. G. Patikoglou, S. K. Burley, Eukaryotic transcription factor-DNA complexes. *Annu Rev Biophys Biomol Struct* **26**, 289-325 (1997).
14. G. M. Poon, Sequence discrimination by DNA-binding domain of ETS family transcription factor PU.1 is linked to specific hydration of protein-DNA interface. *J Biol Chem* **287**, 18297-18307 (2012).
15. R. S. Spolar, M. T. Record, Jr., Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263**, 777-784 (1994).
16. L. Jen-Jacobson, L. E. Engler, L. A. Jacobson, Structural and thermodynamic strategies for site-specific DNA binding proteins. *Structure* **8**, 1015-1023 (2000).
17. J. D. Chodera, D. L. Mobley, Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu Rev Biophys* **42**, 121-142 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

18. G. Klebe, Applying thermodynamic profiling in lead finding and optimization. *Nat Rev Drug Discov* **14**, 95-110 (2015).
19. G. Badis *et al.*, Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720-1723 (2009).
20. D. S. Johnson, A. Mortazavi, R. M. Myers, B. Wold, Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502 (2007).
21. Y. Zhao, G. D. Stormo, Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol* **29**, 480-483 (2011).
22. Q. Morris, M. L. Bulyk, T. R. Hughes, Jury remains out on simple models of transcription factor specificity. *Nat Biotechnol* **29**, 483-484 (2011).
23. A. Jolma *et al.*, DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384-388 (2015).
24. B. Lehner, Molecular mechanisms of epistasis within and between genes. *Trends Genet* **27**, 323-331 (2011).
25. D. W. Anderson, A. N. McKeown, J. W. Thornton, Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* **4**, e07864 (2015).
26. Y. Yin *et al.*, Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, (2017).
27. K. R. Nitta *et al.*, Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **4**, (2015).
28. M. F. Berger *et al.*, Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266-1276 (2008).
29. A. Jolma *et al.*, DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327-339 (2013).
30. J. M. Passner, H. D. Ryoo, L. Shen, R. S. Mann, A. K. Aggarwal, Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. *Nature* **397**, 714-719 (1999).
31. D. E. Piper, A. H. Batchelor, C. P. Chang, M. L. Cleary, C. Wolberger, 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 (1999).
32. S. Hovde, C. Abate-Shen, J. H. Geiger, Crystal structure of the Msx-1 homeodomain/DNA complex. *Biochemistry* **40**, 12013-12021 (2001).
33. N. A. LaRonde-LeBlanc, C. Wolberger, Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. *Genes Dev* **17**, 2060-2072 (2003).
34. Y. Zhang, C. A. Larsen, H. S. Stadler, J. B. Ames, Structural basis for sequence specific DNA binding and protein dimerization of HOXA13. *Plos One* **6**, e23069 (2011).
35. C. M. Ewing *et al.*, Germline Mutations in HOXB13 and Prostate-Cancer Risk. *New Engl J Med* **366**, 141-149 (2012).
36. R. Lavery, M. Moakher, J. H. Maddocks, D. Petkeviciute, K. Zakrzewska, Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res* **37**, 5917-5929 (2009).
37. T. Hansson, J. Marelius, J. Aqvist, Ligand binding affinity prediction by linear interaction energy methods. *J Comput Aided Mol Des* **12**, 27-35 (1998).

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.


38. C. A. Nolte, Tara B; and Krumlauf, Robb, Mammalian Enbryo: Hox Genes. *eLS*. 2015 (10.1002/9780470015902.a0000740.pu3).
39. R. Krumlauf, Hox Genes in Vertebrate Development. *Cell* **78**, 191-201 (1994).
40. K. D. Economides, M. R. Capecchi, Hoxb13 is required for normal differentiation and secretory function of the ventral prostate. *Development* **130**, 2061-2069 (2003).
41. H. Huang, B. Cai, G84E mutation in HOXB13 is firmly associated with prostate cancer risk: a meta-analysis. *Tumor Biol* **35**, 1177-1182 (2014).
42. M. M. Pomerantz *et al.*, The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat Genet* **47**, 1346-1351 (2015).
43. P. Savitsky *et al.*, High-throughput production of human proteins for crystallization: the SGC experience. *J Struct Biol* **172**, 3-13 (2010).
44. G. P. Bourenkov, A. N. Popov, A quantitative approach to data-collection strategies. *Acta Crystallogr D Biol Crystallogr* **62**, 58-64 (2006).
45. W. Kabsch, Xds. *Acta Crystallogr D Biol Crystallogr* **66**, 125-132 (2010).
46. M. D. Winn *et al.*, Overview of the CCP4 suite and current developments. *Acta Crystallogr D Biol Crystallogr* **67**, 235-242 (2011).
47. A. J. McCoy *et al.*, Phaser crystallographic software. *J Appl Crystallogr* **40**, 658-674 (2007).
48. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).
49. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).
50. P. V. Afonine *et al.*, Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* **68**, 352-367 (2012).
51. A. Jolma *et al.*, Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* **20**, 861-873 (2010).
52. R. B. Best *et al.*, Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput* **8**, 3257-3273 (2012).
53. A. D. MacKerell *et al.*, All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**, 3586-3616 (1998).
54. K. Hart *et al.*, Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *J Chem Theory Comput* **8**, 348-362 (2012).
55. N. Foloppe, A. D. MacKerell, All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J Comput Chem* **21**, 86-104 (2000).
56. A. D. MacKerell, Jr., M. Feig, C. L. Brooks, 3rd, Improved treatment of the protein backbone in empirical force fields. *J Am Chem Soc* **126**, 698-699 (2004).
57. B. R. Brooks *et al.*, CHARMM: the biomolecular simulation program. *J Comput Chem* **30**, 1545-1614 (2009).
58. M. S. Friedrichs *et al.*, Accelerating molecular dynamic simulation on graphics processing units. *J Comput Chem* **30**, 864-872 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

59. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of Simple Potential Functions for Simulating Liquid Water. *Journal of Chemical Physics* **79**, 926-935 (1983).
60. J. P. Ryckaert, G. Ciccotti, G., Berendsen H.J.C., Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **23**, 15 (1977).
61. R. W. Zwanzig, High-Temperature Equation of State by a Perturbation Method .1. Nonpolar Gases. *Journal of Chemical Physics* **22**, 1420-1426 (1954).
62. C. H. Bennett, Efficient Estimation of Free-Energy Differences from Monte-Carlo Data. *J Comput Phys* **22**, 245-268 (1976).

## Figure legends

### Figure 1.

**Multiple TFs prefer to bind to two optimal sequences.** (**A)** MYF6; (**B)** BARHL2; (**C**) HOXB13; (**D)** CDX2. Note that single PWM models (top) fail to describe sequence specificity towards different sequences shown in the bar graphs (middle). For example, a single PWM model for HOXB13 (panel C, top) predicts near-equal affinities towards sequences TCGT and TCAT at the position of the bracket, and lower affinity towards CAAT. Analysis of the counts of the subsequences (middle), instead, reveals that the TCAT sequence is bound more weakly than the two most preferred sequences TCGT and CAAT. Counts for maxima (dark blue) and related sequences that differ from the maxima by one or more base substitutions are also shown (light blue). The bars between the maxima represent sequences that can be obtained from both maximal sequences and have the highest count between the maxima. Bottom of each figure: Two distinct models that can represent the binding specificity of the TFs, the divergent bases are indicated by shading. For clarity, the PWM for the MYF6 optima that contains both AA and AC

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dinucleotide flanks (middle bar in a) is not shown. (**E**) Sequences representing the highest (blue line) and lowest (red line) affinity sequences between the two optimal HOXB13 sequences. y-axis: counts for 8-mer sequences containing the indicated trinucleotide followed by TAAA. (**F**) Epistasis in HOXB13-DNA binding. The effect of individual mutations (single mutants) to the optimal sequence TCGTAAAA (top) are relatively severe, with binding decreasing by more than 70% in all cases (observed binding). However, combinations of the mutations (double mutants) do not decrease HOXB13 binding in a multiplicative manner (compare predicted and observed binding). A multiplicative model predicts that combining all three substitutions would abolish binding, but instead the CAA site is bound more strongly than any other mutant (triple mutant).

**Figure 2.**

**Comparison of HOXB13-DNA complexes. (A)** The view of superposition of HOXB13 (wheat) bound to $DNA^{TCG}$ and HOXB13 (red) bound to $DNA^{CAA}$ (rmsd = 0.813 Å on 57 residues). The respective DNAs are in blue and green. The dissimilar base pairs are presented as ball-and-stick models and colored as the proteins, $DNA^{TCG}$ is wheat and $DNA^{CAA}$ is red. Note the different bending of the DNA backbone at these positions (orange). **(B)** Schematic representation of interactions formed between HOXB13 DBD and the two different DNAs: left panel shows the interactions between HOXB13 and the primary binding site ($DNA^{TCG}$) and right panel represents the interactions of HOXB13

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with the secondary site ($DNA^{CAA}$), respectively. Dashed lines represent interaction with backbone phosphates and deoxyribose and solid lines interactions with the bases. The protein residues belonging to the HOXB13-$DNA^{TCG}$ and HOXB13-$DNA^{CAA}$ structures are colored wheat and red, respectively. The divergent parts of the DNA sequences are highlighted by a light green box. Note that the TCG site lacks direct contacts to the DNA bases, whereas the CAA site is recognized by direct contacts by Gln-265 and Ile-262. Most other contacts are well conserved in both structures. The four As of the TAAAA sequence are recognized by N-terminal amino-acids interacting with the DNA backbone via the minor groove, whereas the T is recognized by a bidentate interaction formed between its complementary adenine $A_{10}$ and side chain of asparagine Asn-266. Two hydrogen bonds are formed between nitrogen atoms $N^6$ and $N^7$ from adenine base and oxygen and nitrogen atoms of the Asn-266 side chain. This adenine-specific asparagine is totally conserved in the HOX family. **(C)** Superposition of CDX2 (cyan) bound to $DNA^{TCG}$ and CDX2 (magenta) bound to $DNA^{CAA}$ (rmsd = 0.270 Å on 64 residues). The respective DNAs are in blue and green. The dissimilar base pairs are presented as ball-and-stick models and colored as the proteins, $DNA^{TCG}$ is green and $DNA^{CAA}$ is blue. Note the different bending of the DNA backbone at these positions (orange). **(D)** Schematic representation of interactions formed between CDX2 DBD and the two different DNAs. **(F)** Structural interpretation of mutations that change the specificity of HOXB13: the mutations changing Ccaa/Ctcg to Gcaa/Gtcg are shown in a small box and, as a close view, on the left panel, and mutations, which switch the preferences of HOXB13 from CTCG to CCAA, are shown in big box and, as a close view, on the right

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

panel. The mutations are presented in structural alignment of HOXB13 (red), HOXA9 (blue, PDB entry 1PUF) and CDX2 (pink) bound to DNA. Note the unique mutation of Lys (small box), which is conserved in all known HOXes, to Thr in HOXB13 allows HOXB13 to accept any base pair in the position before TCG/CAA. The left panel is representing the close view to the interactions formed by Lys in HOXA9 and CDX2. Long aliphatic chain of Lys increases the hydrophobicity of this part of protein-DNA interface, pushing out the water molecules. Dashed line indicates water-mediated interaction between the ε-Amino group of Lys and the N7 and O6 of the guanine base at the Gtcg sequence. The right panel is representing the close view of triple mutation in the loop connecting helix 1 and helix 2: Lys-239/Met, Phe-240/Tyr and Ile-241/Leu; and single mutation of Lys-272/Arg. Those mutations change the hydrogen bond network inside the protein and between protein and DNA and lead to a preference towards the more rigid, more B-shaped $DNA^{CAA}$.

**Figure 3.**

**Close view of the protein-DNA interactions.** (**A**) HOXB13-$DNA^{TCG}$ and (**B**) HOXB13-$DNA^{CAA}$ complexes. The 2mFo-Fc maps contoured with 1.5σ are shown around the key residues. The residues and base pairs involved in protein to DNA contacts are also labeled. (**C, D**) Surface representation of the major groove in HOXB13-$DNA^{TCG}$ and HOXB13-$DNA^{CAA}$ complexes, respectively. The divergent bases are colored to indicate electrostatic charges of the atoms: neutral carbon atoms are green, oxygen atoms (negative) are red and nitrogen atoms (positive) are blue. Note the larger solvent-

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

accessible space between amino-acids and bases in the TCG structure (C) and the difference in distribution of the positively and negatively charged spots on the surface that can contribute to differences in distribution of water molecules on the surface. (**E**) CDX2-DNA$^{TCG}$ and (**F**) CDX2-DNA$^{CAA}$ complexes. The 2mFo-Fc maps contoured with 1.5σ are shown around the key residues. The residues and base pairs involved in protein to DNA contacts are also labeled.

**Figure 4.**

**Close view of the role of water molecules in HOXB13-DNA interaction. (A)** Schematic representation of water-mediated interactions between amino-acids (red typeface) of HOXB13 and DNA bases in the HOXB13-DNA$^{CAA}$ structure. Different water chains are indicated with different shades of blue. Thick dashed lines represent interactions formed between water molecules and bases or amino acids; thin dashed lines represent contacts formed between water molecules, and solid blue line indicates the direct interaction between $A_{10}$ and Asn-266. Note that all of the base positions in the CAA sequence (boxes) are recognized via direct or water-mediated hydrogen bonds. **(B)** Structural representation of the network of interactions schematically presented in (A). Note the three water chains colored by slightly varied blue color. The amino acids and bases involved in interactions are presented as stick models. **(C)** Close view to the different conformations of amino-acids observed in HOXB13-DNA$^{TCG}$ and HOXB13-

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$DNA^{CAA}$ structures. Note that the conformations of the key amino-acids Gln-265 and Arg-258 that interact with the water network in HOXB13-$DNA^{CAA}$ (amino-acids in red, DNA carbons in green) are not suitable for interacting with the network in HOXB13-$DNA^{TCG}$ (amino-acids and DNA carbons in wheat). **(D)** Surface representation of protein-DNA interface of HOXB13-$DNA^{TCG}$ complex. Relatively large channel between the protein and DNA that goes through the protein-DNA interface (white) lend support to the presence of mobile water molecules in this region. TCG-bases are colored by atoms: carbon atoms are yellow; oxygen atoms are red and nitrogen atoms are blue. **(E)** Schematic representation of water-mediated interactions between amino-acids (cyan typeface) of CDX2 and DNA bases in the CDX2-$DNA^{TCG}$ structure. Different water chains are indicated with different shades of red. Thick dashed lines represent interactions formed between water molecules and bases or amino acids; thin dashed lines represent contacts formed between water molecules, and solid red line indicates the direct interaction between $A_{10}$ and Asn-236. Note that only the position of the GC pair is recognized (boxes) via water-mediated hydrogen bonds. **(F)** Structural representation of the network of interactions schematically presented in **(E)**. Note the three water chains colored by varied red-pink color. The amino acids and bases involved in interactions are presented as stick models.

**Figure 5.**

**Calorimetric titration data reveals that two optimal DNA sequences recognized by HOXB13 (A, B), CDX2 (C, D), BARHL2 (E, F) and MYF5 (G, H) represent**

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**enthalpy and entropy optima.** The optimal sequences with higher enthalpic contribution to binding are presented on the left side (A, C, E, G) and the reactions with higher entropic contribution are presented on the right side (B, D, F, H). Note that for each protein both DNAs are bound with similar ΔG. The top panels of the ITC figures represent raw data; the bottom panels show the integrated heat of the binding reaction. The red line represents the best fit to the data, according to the model that assumes a single set of identical sites. The determined changes of enthalpy and calculated losses of entropy are shown on the bottom panel. The changes of Gibbs free energy, $\Delta G=\Delta H-T\Delta S$, are also calculated and presented on the bottom panel of each isotherm.

**Figure 6.**

**The two optimal sites bound by HOXB13 represent enthalpy and entropy-driven optima**. (**A-B)** Schematic illustrations of the binding mechanism driven by the low enthalpy (A) and by high entropy (B) are presented in the left panels. The DNA bases are presented as pyrimidine and purine rings, protein is represented as ellipsoid, N-terminus is shown bound to the minor groove created by A-stretch, and water molecules are shown schematically and colored blue. The dashed lines represent hydrogen bonds observed in the low enthalpy state; solid line represents direct interactions between amino acids and bases. The blurred water molecules indicate the high entropy state. Hydrogen bonds that are common to both complexes are omitted for clarity. Graphs on the right show schematic illustration of the variance of enthalpy (ΔH, top), entropy (-TΔS, middle) and Gibbs free energy (ΔG) (bottom) as a function of an idealized one-dimensional

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

continuous variable representing the high-dimensional variables of shape, electrostatic charge and vibration of DNA that vary as a function of the DNA sequence. As DNA is composed of only four bases, only discrete positions in this axis are possible (indicated by dots). Example models of shape and charge distribution of different DNA sequences (from Figure 1C) are shown as surface representation above the scheme. The surfaces are colored according to the charge distribution: positively charged atoms are in blue, negatively charged are in red and neutral atoms are in green. Note that enthalpy and entropy are partially negatively correlated, leading to binding promiscuity (wider optima in $\Delta G$ compared to $\Delta H$ and $\Delta S$). The remaining uncorrelated component leads to the presence of two optima for $\Delta G$ (bottom). Shaded boxes on the right show simplified dinucleotide binding models that illustrate how this leads to two distinct locally optimal sequences.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Table 1. Data collection and refinement statistics.**

| | HOXB13-DNA$^{TCG}$ | HOXB13-DNA$^{CAA}$ | CDX2-DNA$^{TCG}$ | CDX2-DNA$^{CAA}$ |
|---|---|---|---|---|
| **Data collection** | | | | |
| **Wavelength (Å)** | 0.9724 | 0.9724 | 0.9724 | 0.9724 |
| **Resolution range (Å)** | 45.95 - 2.19 (2.27 - 2.19) | 46.29 - 3.2 (2.97 - 2.87) | 43.23 - 2.57 (2.66 - 2.57) | 55.96– 2.95 (3.13 - 2.95) |
| **Space group** | P 1 2 1 | P 2 2 $2_1$ | C 1 2 1 | I 1 2 1 |
| **Unit cell (Å, °)** | 77.35 57.92 101.28; 90 101.57 90 | 52.62 52.52 389.33; 90 90 90 | 127.95 46.49 68.89; 90 113.27 90 | 70.25 46.69 128.63; 90 101.40 90 |
| **Total reflections** | 241614 (21747) | 86877 (3476) | 19575 (1958) | 27018 (4003) |
| **Unique reflections** | 44125 (3912) | 20590 (1049) | 12095 (1197) | 8802 (1264) |
| **Multiplicity** | 5.5 (5.6) | 4.2 (3.3) | | 3.2 (3.2) |
| **Completeness (%)** | 97.42 (87.37) | 97.5 (90.4) | 99.5 (100) | 96.6 (90.5) |
| **Mean I/sigma(I)** | 8.11 (1.10) | 7.91 (0.10) | 8.47 (2.77) | 7.5 (1.1) |
| **R-merge** | 0.12 (1.21) | 0.085 (4.59) | 0.13 (5.49) | 0.071 (7.24) |
| **R-meas** | 0.13 | 0.09 | 0.08 | 0.09 |
| **CC1/2** | 0.99 (0.71) | 0.99 (0.72) | 0.99 ( 0.80) | 0.99 (0.61) |
| **Refinement** | | | | |
| **R-work** | 0.25 (0.37) | 0.21 | 0.23 | 0.19 |
| **R-free** | 0.29 (0.35) | 0.28 | 0.29 | 0.25 |

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

| | | | | |
|---|---|---|---|---|
| **Number of non-hydrogen atoms** | 5591 | 5197 | 2841 | 2783 |
| **macromolecule** | 5072 | 5172 | 2748 | 2717 |
| **water** | 519 | 17 | 93 | 66 |
| **Protein residues** | 274 | 242 | 144 | 141 |
| **RMS (bonds)** | 0.011 | 0.025 | 0.018 | 0.012 |
| **RMS (angles)** | 1.26 | 2.03 | 2.11 | 1.83 |
| **Ramachandran favored (%)** | 97 | 92 | 97.8 | 97.1 |
| **Ramachandran outliers (%)** | 0.41 | 1.7 | 1.43 | 0.73 |
| **Clashscore** | 5.31 | 10.51 | 4.42 | 6.43 |
| **Average B-factor** | 41.70 | 124.40 | 30.54 | 74.75 |
| **macromolecule** | 42.10 | 124.70 | 29.30 | 74.41 |

Statistics for the highest-resolution shell are shown in parentheses.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## MATERIAL AND METHODS

### Protein expression, purification and crystallization

Expression and purification of the DNA-binding domain fragment of human HOXB13 (residues 209-283) were performed as described in Refs. (*43*) and (*26*). The DNA fragments used in crystallization were obtained as single strand oligos (Eurofins), and annealed in 20 mM HEPES (pH 7.5) containing 150 mM NaCl and 0.5 mM *Tris* (2-carboxyethyl) phosphine (TCEP) and 5% glycerol. For each complex, the purified and concentrated protein was first mixed with a solution of annealed DNA duplex at a molar ration 1:1.2 and after one hour on ice subjected to the crystallization trials. The crystallization conditions for both complexes were optimized using an in house developed crystal screening kit of different PEGs. Complexes were crystalized in sitting drops by vapor diffusion technique from solution containing 50 mM Tris buffer (pH 8.0), 100 mM $MgCl_2$, 150 mM KCl and different concentrations of various PEGs and additives. PEG (3350) was used at 14 % for HOXB13-$DNA^{TCG}$ and 21 % (w/v) of polyethylene glycol monomethyl ether (PEGmme (5000)) was used in crystallizations of HOXB13 with $DNA^{CAA}$. Both complexes of HOXB13 with $DNA^{TCG}$ and $DNA^{CAA}$ were crystallized with addition of 8% of PEG (400). The data sets were collected at ESRF from a single crystal on beam-lines ID29 (HOXB13:$DNA^{TCG}$) and ID23-1 (HOXB13:$DNA^{CAA}$), at 100 K using the reservoir solution as cryo-protectant. The data collection strategy was optimized with the program BEST (*44*). Data were integrated with the program XDS (*45*) and scaled with SCALA (*46*). Statistics of data collection are presented in **Table 1**.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

**Structure determination and refinement**

Both structures were solved by molecular replacement using program Phaser (*47*) as implemented in Phenix (*48*) with the structure of HOXA9 (pdb entry 1PUF) as a search model. After the positioning of protein, the density of DNA was clear and the molecule was built manually using COOT (*49*). The rigid body refinement with REFMAC5 was followed by restrain refinement with REFMAC5, as implemented in CCP4 (*46*) and Phenix.refine (*50*). The manual rebuilding of the model was done using COOT. The refinement statistics are presented in **Table 1**. The first 7 amino acid from N-termini and the last 7 from C-termini were found disordered and were not built in the maps. The end base pairs of the DNA in HOXB13-DNA$^{TCG}$ structure were also found slightly disordered but it was possible to build them to the maps.

**HT-SELEX and motif analysis**

MYF6 and BARHL2 HT-SELEX experiments were performed essentially as described in Yin et al. (Ref. (*26*)). The PWM models were generated from cycles 4, 3 and 3 of HOXB13 (from Ref. (*26*)), MYF6 and BARHL2 HT-SELEX reads, respectively, using the multinomial (setting=1) method (*51*) with the following seeds: HOXB13 single PWM: NCYMRTAAAAN, TCG: NCTCGTAAAAN, CAA: NCCAATAAAAN; MYF6 single PWM: NRWCAGCTGWYN, AA...TT flank: NAACAGCTGTTN, GT...AC flank: NGTCAGCTGACN; BARHL2 single PWM: NSYTAAWYGNYN, TT: NSYTAATTGNYN, AC: NSYTAAACGKYN.

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Molecular Dynamics

Molecular dynamics simulations were performed for HOX13B complexed with either $DNA^{TCG}$ or $DNA^{CAA}$; the DNA sequence used in the simulations contained nucleotides G5 – C18 from the crystal structure. The CHARMM 36 forcefield (*52-56*) and CHARMM program (*57*), with the CHARMM interface to OpenMM(*58*) to allow the use of NVIDIA graphical processing units (GPUs), were used for all simulations. The starting structure was placed in a cubic solvent box with 8 nm side length containing water (*59*) and 0.15M NaCl; $Na^+$ ions were then added to neutralize the system. After energy minimization to relax initial strain the systems were heated from 100K to 300K over 0.1 ns followed by 0.3 ns simulation at constant pressure (1 bar) and constant temperature (300K), with soft harmonic positional restraints on the protein and DNA atoms. For each complex 3 x 0.8 μs production runs were performed using the GPU, with the pressure and temperature maintained at 1 bar and 300K, respectively, and without the positional restraints. Particle mesh Ewald summation was used to treat the long range electrostatic interactions, using a $6^{th}$ order cubic spline interpolation for the charge distribution on the 0.1nm spaced grid points, kappa=0.34. The same 0.9 nm cutoff was used for both the direct space part of the PME and for the van der Waals interactions, which were switched to zero from 0.8 nm to 0.9 nm, and the non-bond list was generated with a 1.1 nm cutoff. SHAKE (*60*) was used to keep the lengths of all covalent X-H bonds fixed, allowing a time-step of 2 fs.

In the free energy perturbation calculations (*61*) we changed the three base pairs in the TCG sequence into those of the CAA sequence using a total of 43 intermediate states, where the order of change was: turn off charges, change Lennard-Jones parameters, turn on charges. In each state, a 10 ns

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

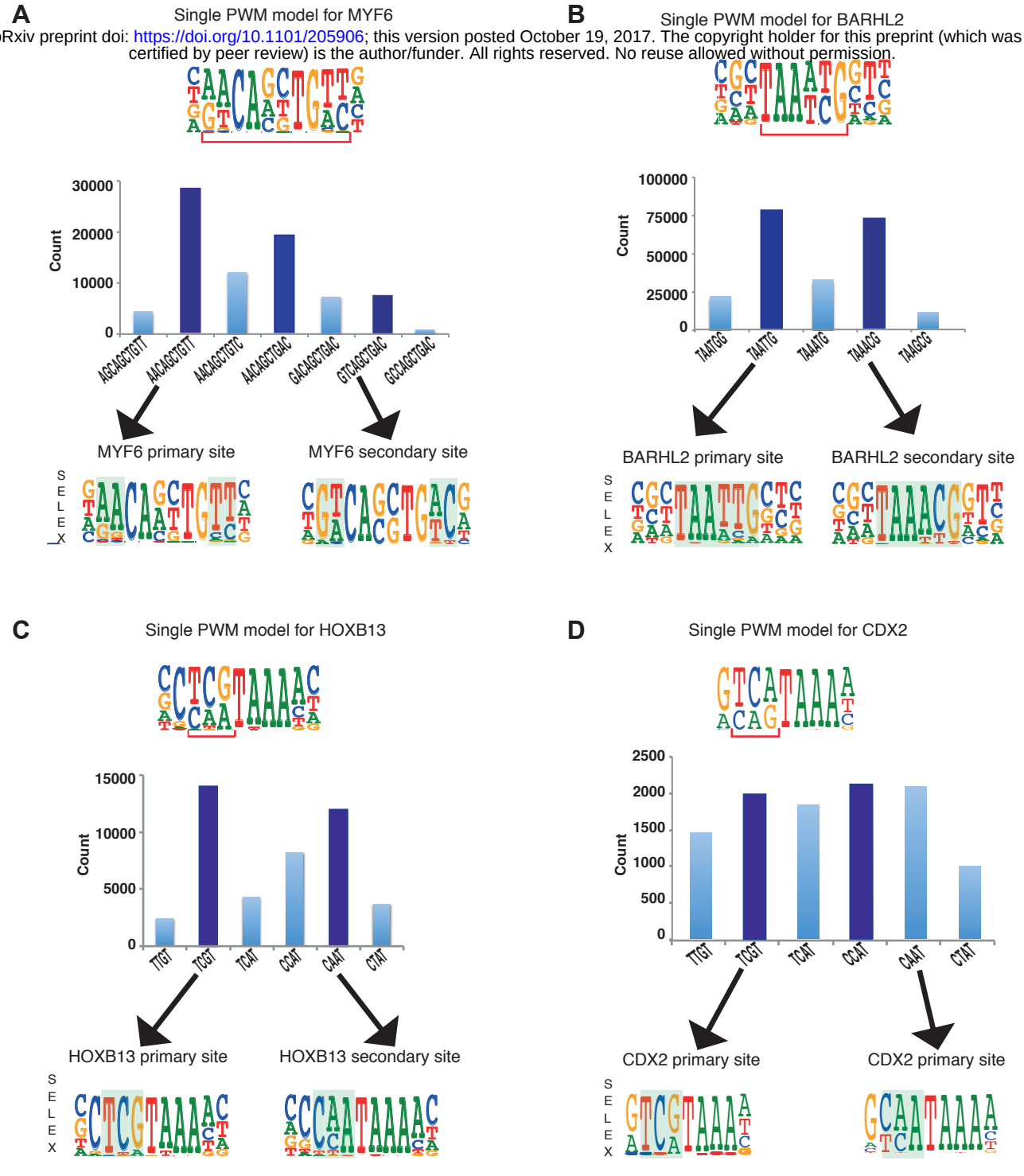
equilibration was followed by 10 ns production. The free energies were calculated using the Bennett Acceptance Ratio method (*62*).

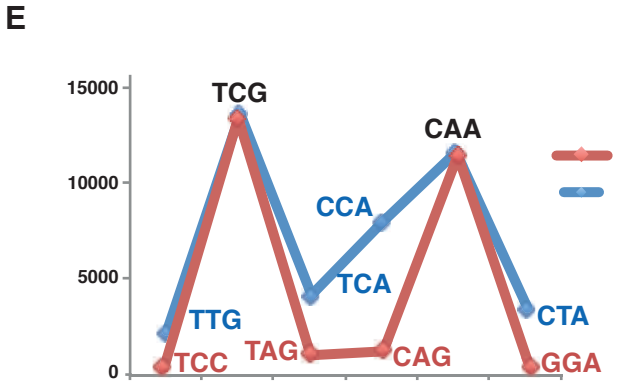
**Isothermal titration calorimetry**.

The ITC experiments were carried similarly to described in Ref. (*26*). Briefly, an ITC200 microcalorimeter (MicroCal Inc., Northampton, Massachusetts, USA) in PSF (Protein Science Facility at Karolinska Institute, Sweden) was used to measure binding isotherms of DNAs by direct titration of protein to the cell containing DNA. The measurements were taken at 25 Cº. Both protein and DNA were prepared in a buffer containing 20 mM HEPES pH 7.5, 300 mM NaCl, 10% glycerol and 2 mM Tris (2-carboxyethyl)phosphine (TCEP). To measure binding affinity, a solution of 0.15mM protein was titrated to 0.012–0.016 mM solution of DNA. A total of 23 injections were made with 240 s between injections. Each experiment was repeated three times for the reliability of the results. All data were evaluated using the OriginPro 7.0 software package (Microcal) supplied with the calorimeter. The apparent binding constant $K_b$, binding enthalpy $\Delta H$ and stoichiometry n, together with their corresponding standard deviation (s.d.), were determined by a nonlinear least square fit of the data to standard equations for the binding using a model for one set of independent and identical binding sites as implemented in the package. The entropy and free energy of binding were obtained from the relation $\Delta G = -RT\ln K_d = \Delta H - T\Delta S$.

Morgunova et al., 2017 Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

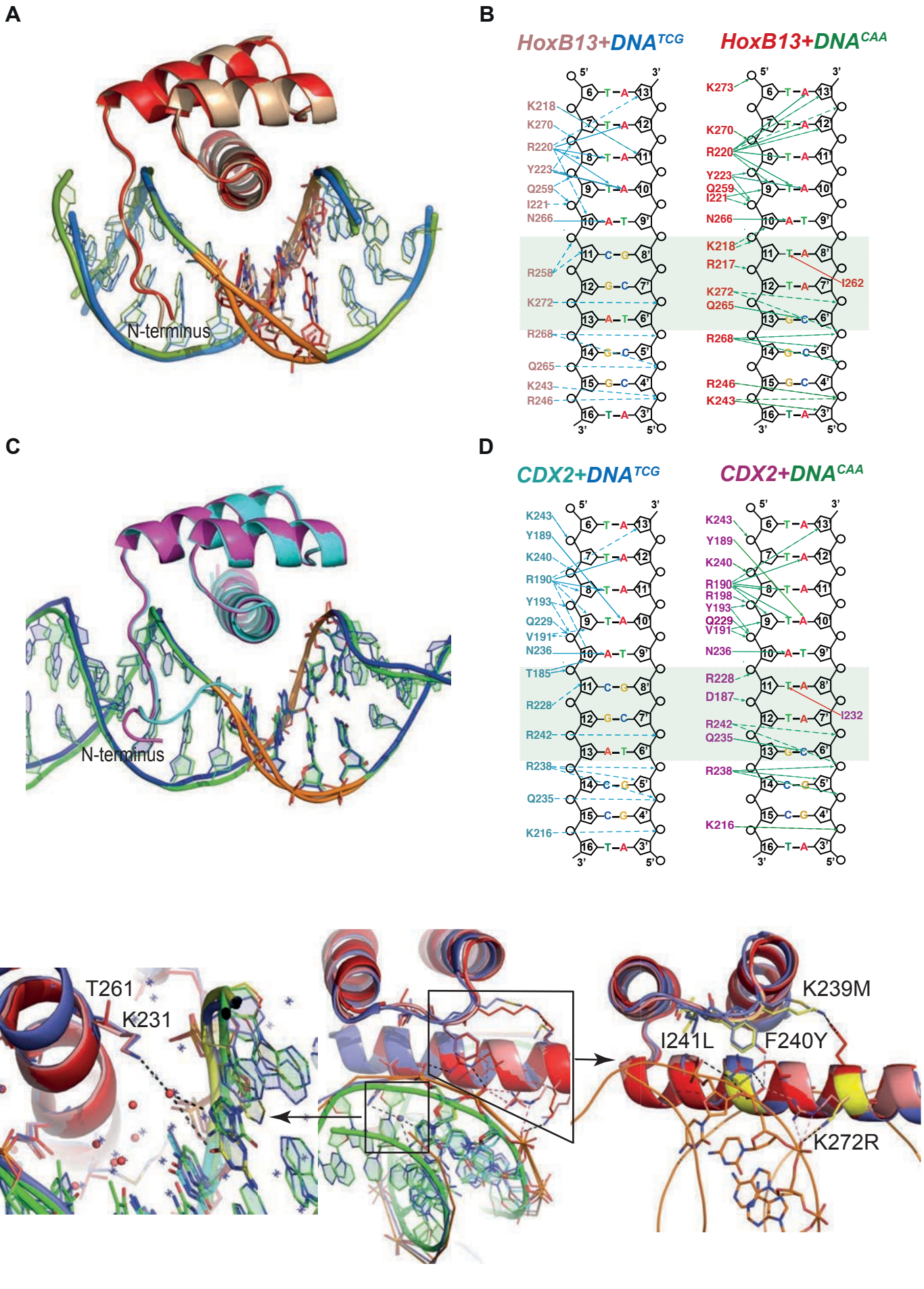


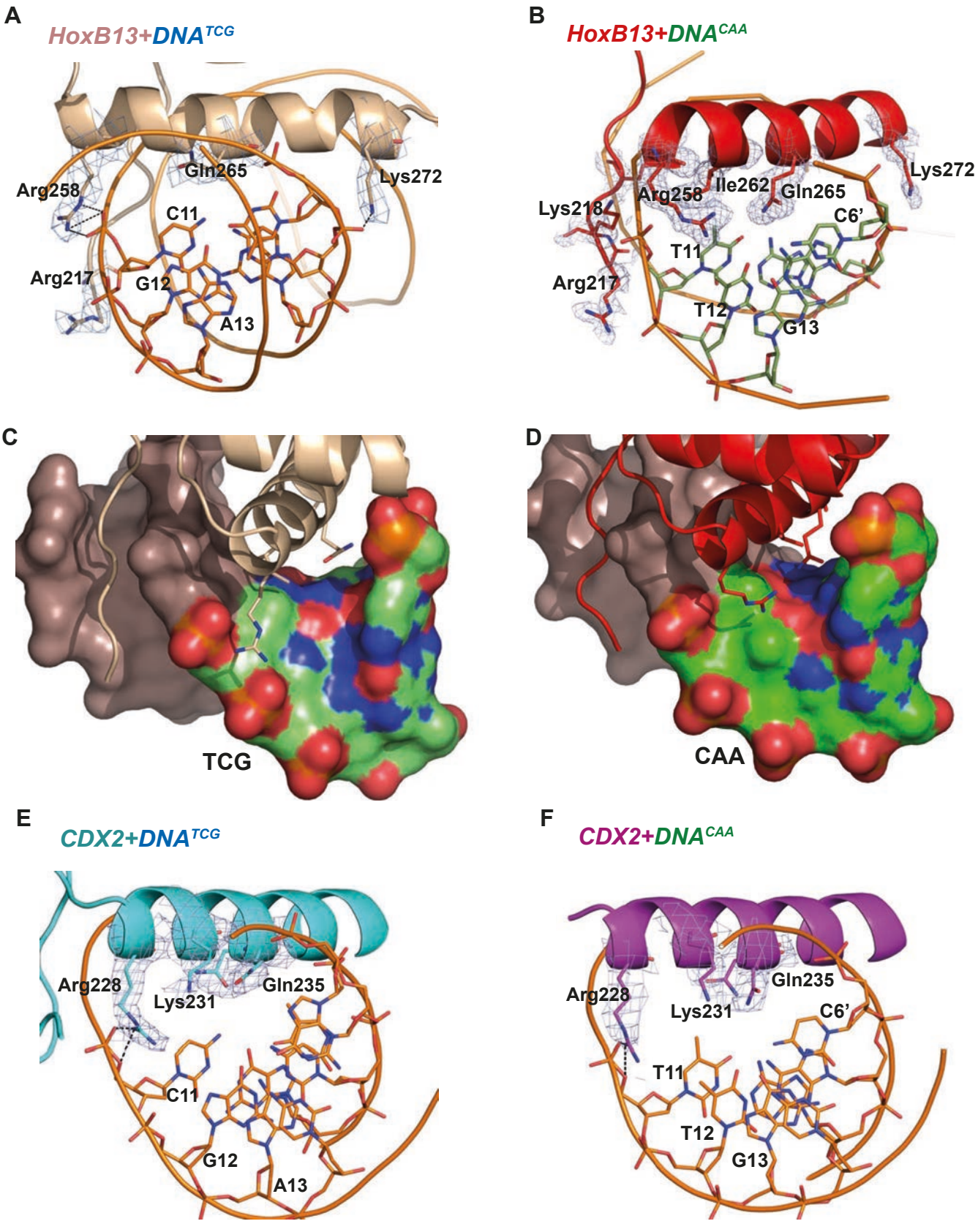

**F**

| **EPISTASIS** | | **TCG**(TAAAA) 100% | |
|---|---|---|---|
| **Single mutants** | CCG | TAG | TCA |
| observed binding | 12.6% | 5.3% | 29.6% |
| **Double mutants** | CAG | CCA | TAA |
| observed binding | 7.1% | 58.2% | 18.4% |
| predicted binding | 0.7% | 3.7% | 1.6% |
| fold difference | 10.5 | 15.7 | 11.6 |
| **Triple mutant** | | CAA | |
| observed binding | | 85.3% | |
| predicted binding | | 0.2% | |
| fold difference | | 429.2 | |

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

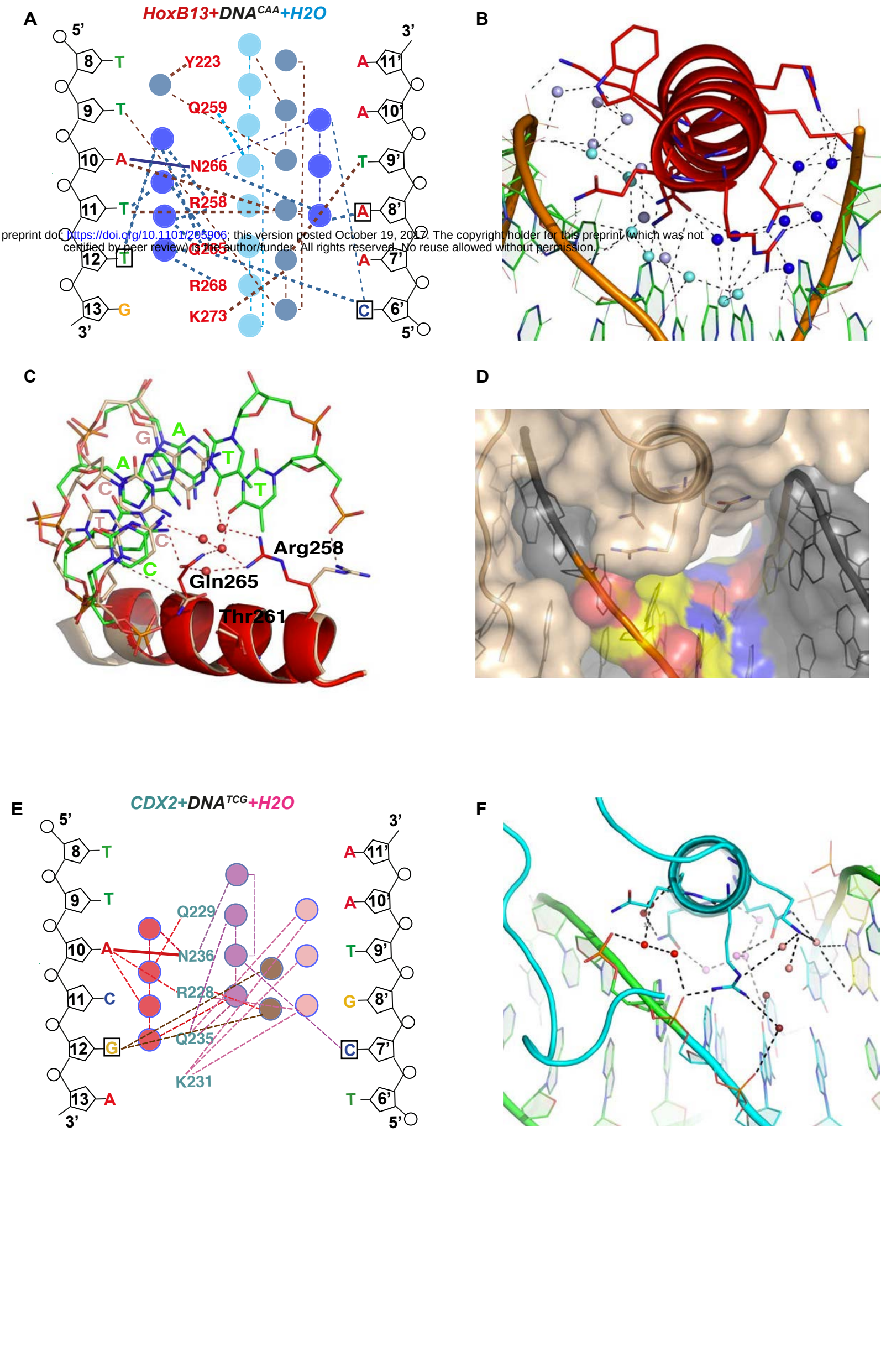


bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



Morgunova et al., 2017 Figure 4

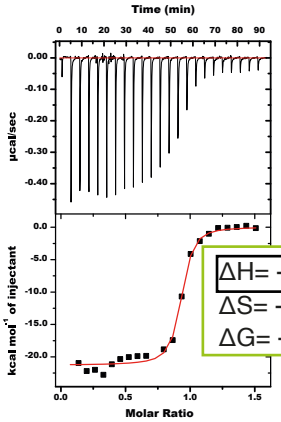
bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A

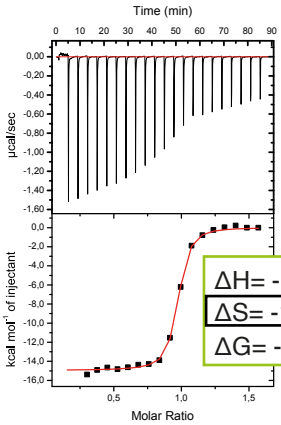
**HOXB13:DNA$^{CAA}$**




ΔH= -21250 cal/mol
ΔS= -37.5 cal/mol/deg
ΔG= -10075 cal/mol

B

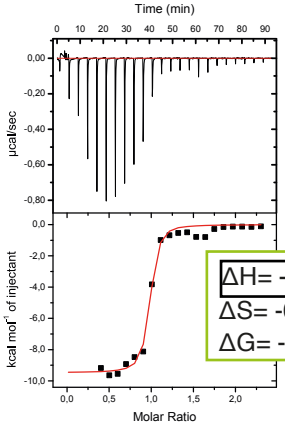
**HOXB13:DNA$^{TCG}$**




ΔH= -14950 cal/mol
ΔS= -16.8 cal/mol/deg
ΔG= - 9943,6 cal/mol

C

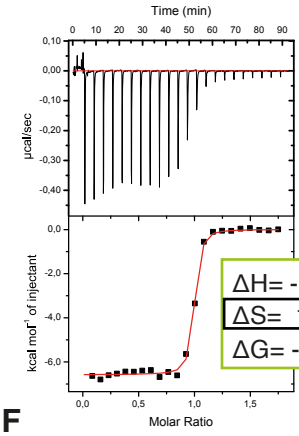
**CDX2:DNA$^{CAA}$**




ΔH= -9809.6 cal/mol
ΔS= -0.7 cal/mol/deg
ΔG= -9601cal/mol

D

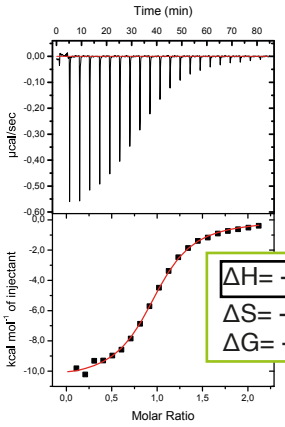
**CDX2:DNA$^{TCG}$**




ΔH= - 6577 cal/mol
ΔS= 12.3 cal/mol/deg
ΔG= - 10242.4 cal/mol

E

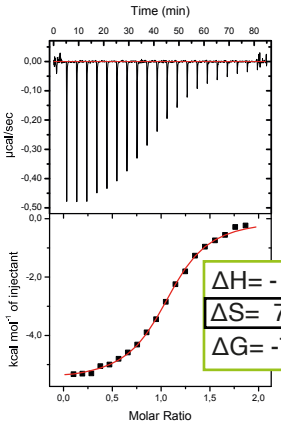
**BARHL2:DNA$^{AC}$**




ΔH= -10570 cal/mol
ΔS= -8.84 cal/mol/deg
ΔG= -7935 cal/mol

F

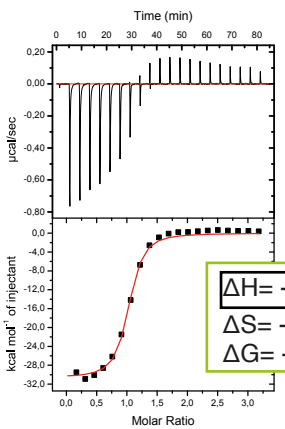
**BARHL2:DNA$^{TT}$**




ΔH= - 5563 cal/mol
ΔS= 7.55 cal/mol/deg
ΔG= -7813 cal/mol

G

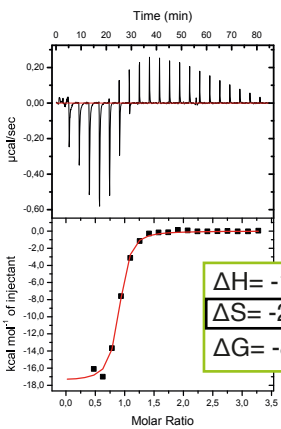
**MYF5:DNA$^{AACAGCTGTT}$**



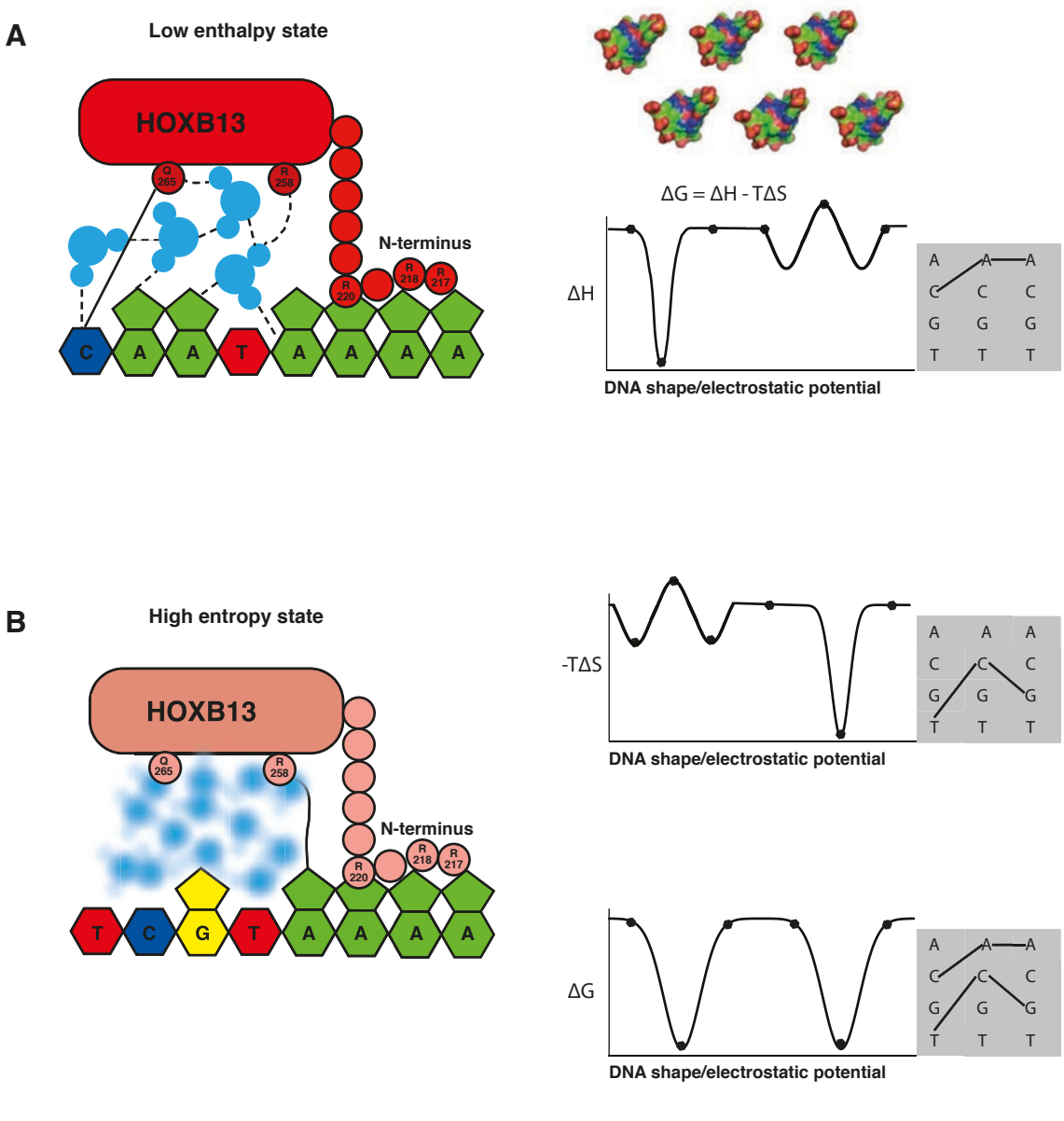

ΔH= -30690 cal/mol
ΔS= -73.0 cal/mol/deg
ΔG= -8936 cal/mol

H

**MYF5:DNA$^{GTCAGCTGAC}$**




ΔH= -17470 cal/mol
ΔS= -28.9 cal/mol/deg
ΔG= -8857.8 cal/mol

Morgunova et al., 2017 Figure 6

bioRxiv preprint doi: https://doi.org/10.1101/205906; this version posted October 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.