Supporting Material

Positive and negative design for nonconsensus protein-DNA binding affinity in the vicinity of functional binding sites

Ariel Afek and David B. Lukatsky*
Department of Chemistry, Ben-Gurion University of the Negev, Beer-Sheva 84015, Israel

**Corresponding author:*
Email: lukatsky@bgu.ac.il

Supporting Figure Legends

Figure S1. This figure demonstrates the robustness of the computed free energy of nonconsensus protein-DNA binding around the Reb1 consensus motif, with respect to the transcription factor size, M, and with respect to the width of the sliding window, L. The computed normalized, average free energy per bp, $\langle \delta f \rangle = \left\langle \langle \delta F \rangle_{TF} \right\rangle_{seq} / M$, in the interval (-200,200) around specific, experimentally bound Reb1 motifs, as compared with the corresponding $\langle \delta f \rangle$ computed around the control set of unbound motifs. Here, $\delta F = F - F_{rand}$, and F_{rand} is the free energy computed for a randomized sequence (in the same sliding window as F), and averaged over 20 random realizations. Top panel (from left to right): Different values of M were used: $M = 4$, $M = 6$, and $M = 8$. Bottom panel (from left to right): Different values of L were used: $L = 30$, $L = 50$, and $L = 70$. See Materials and Methods for the calculation of the *p*-values.

Figure S2. This figure demonstrates the robustness of the computed free energy of nonconsensus protein-DNA binding around the Reb1 consensus motif, with respect to the global variability of the nucleotide content along the genome. Solid curves: The computed average free energy per bp, $\langle f \rangle = \left\langle \langle F \rangle_{\mathrm{TF}} \right\rangle_{seq} / M$, in the interval (-200,200) around the specific, experimentally bound Reb1 motif, as compared with the corresponding *normalized* $\langle \delta f \rangle = \left\langle \langle \delta F \rangle_{\mathrm{TF}} \right\rangle_{seq} / M$, where $\delta F = F - F_{rand}$. For a given TF, F is computed as described in the main text, and F_{rand} is the free energy computed for a randomized sequence (in the same sliding window as F), and averaged over 20 random realizations. Dotted curves: Similar calculations are performed for the non-functional (unbound) sequences (see the main text). We used $M = 8$ and $L = 50$ in our calculations. The described procedure removes a possible bias in the free energy, stemming from the global variability of the nucleotide content.

Figure S3. Specific, functional Reb1 binding sites are surrounded by the genomic background with enhanced nonconsensus protein-DNA binding free energy. We searched for Reb1 binding motifs exclusively within the interval (-400,400) around the annotated TSSs. The computed average free energy per bp, $\langle f \rangle = \left\langle \langle F \rangle_{\mathrm{TF}} \right\rangle_{seq} / M$, in the interval (-200,200) around 415 bound specific Reb1 motifs (black), as compared with the corresponding $\langle f \rangle$ for 271 unbound specific Reb1 motifs (grey), as measured in (1). The second averaging is performed over the sequences aligned with respect to the center of the specific Reb1 binding motif (TTACCCG/T); and M is the motif length. We used $M = 8$ and $L = 50$ in our calculations. The computed *p*-value is highly significant. The *p*-value is computed analogously to **Figure 2**.

Figure S4. This figure demonstrates the robustness of the computed free energy of nonconsensus protein-DNA binding with respect to the global variability of the nucleotide content along the genome. The computed average free energy per bp, $\langle f \rangle = \left\langle \langle F \rangle_{\mathrm{TF}} \right\rangle_{seq} / M$, in the interval (-400,400) around specific, experimentally bound CTCF motifs, as compared with the corresponding *normalized* $\langle \delta f \rangle = \left\langle \langle \delta F \rangle_{\mathrm{TF}} \right\rangle_{seq} / M$, where $\delta F = F - F_{rand}$. For a given TF, F is computed as described in the main text, and F_{rand} is the free energy computed for a randomized sequence (in the same sliding window as F), and averaged over 25 random realizations. The described procedure removes a possible bias in the free energy, stemming

from the global variability of the nucleotide content. We used $M = 8$ and $L = 50$ in our calculations. **A**. The calculation is performed for Chromosome 2. **B**. The calculation is performed for the 10% highest CTCF occupancy binding sites. **C**. The calculation is performed for the 10% intermediate CTCF occupancy binding sites.

Figure S5. This figure demonstrates the robustness of the computed free energy of nonconsensus protein-DNA binding with respect to the transcription factor size, M, and with respect to the width of the sliding window, L. The computed normalized, average free energy per bp, $\langle \delta f \rangle$, in the interval (-400,400) around specific, experimentally bound CTCF motifs, as compared with the corresponding $\langle \delta f \rangle$ computed around the control set of unbound motifs. Top panel (from left to right): Different values of M were used: $M = 4$, $M = 6$, and $M = 8$. Bottom panel (from left to right): Different values of L were used: $L = 30$, $L = 50$, and $L = 70$. See Materials and Methods for the description of the control set of unbound motifs and for the calculation of the *p*-values. The calculations are performed for Chromosome 2.

Supporting References

1. Rhee HS & Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147(6):1408-1419.

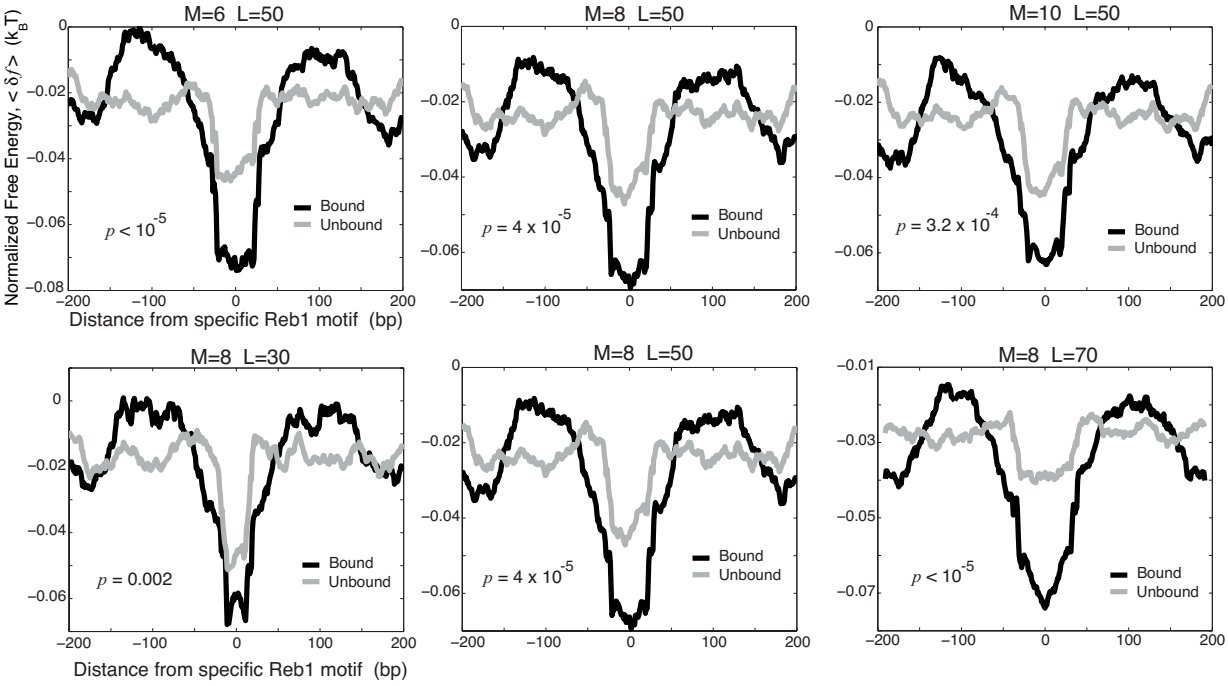

Figure S1

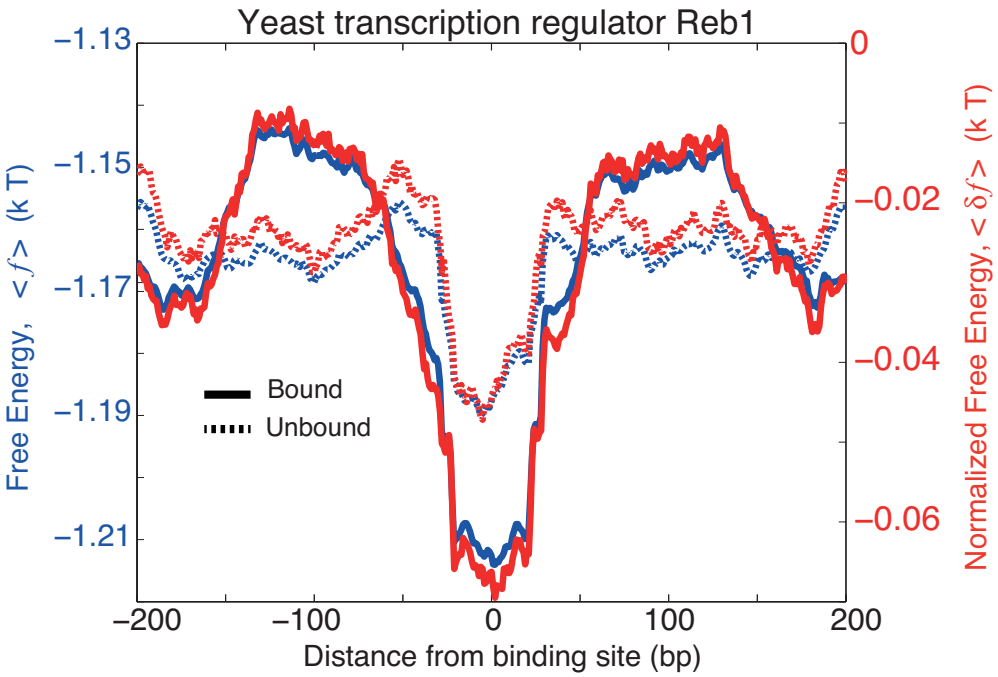

Figure S2

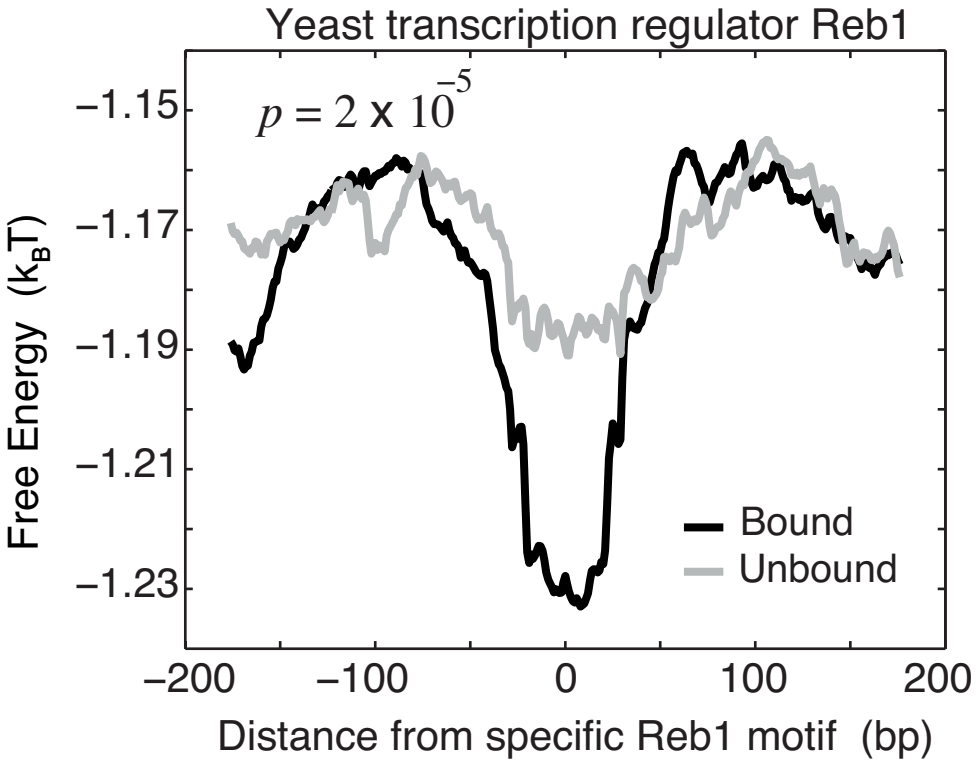

Figure S3

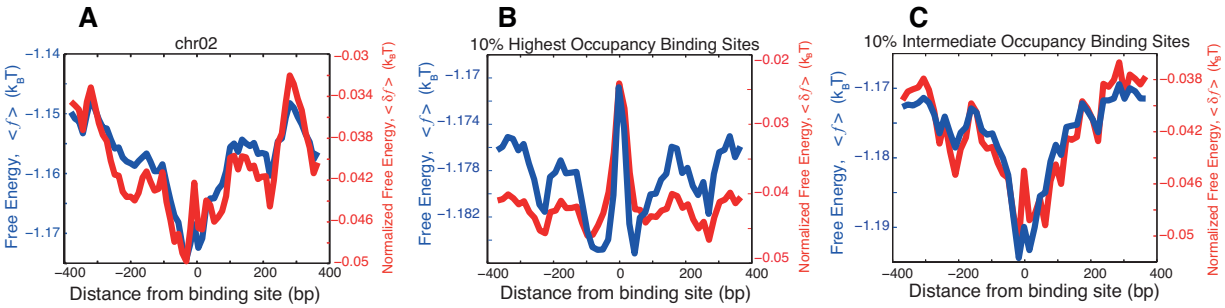

Figure S4

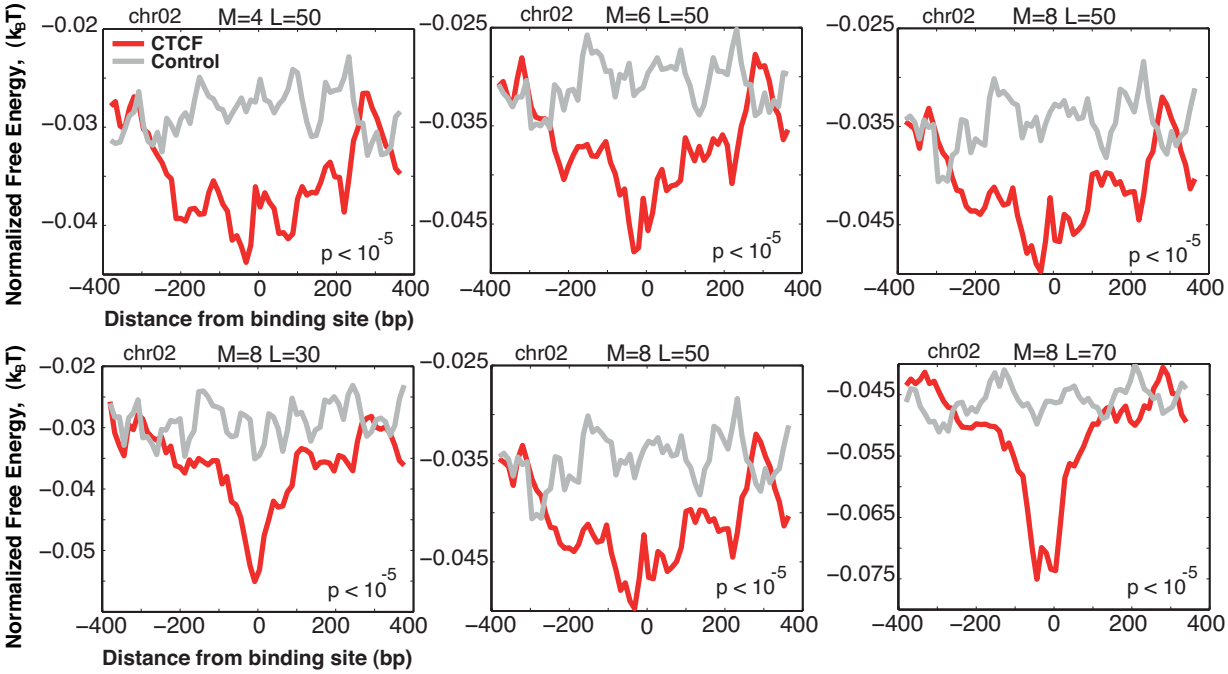

Figure S5