COMPREHENSIVE *IN SILICO* MAPPING OF DNA-BINDING PROTEIN AFFINITY LANDSCAPES

Allocation: GLCPC/310 Knh
PI: Peter Freddolino[1]
Co-PI: Morteza Khabiri[1]
Collaborator: Arttu Jolma[2]

[1]University of Michigan Medical School
[2]University of Toronto

EXECUTIVE SUMMARY

Transcription factors (TFs) and other DNA–binding proteins shape the behavior of all cells, coordinating appropriate gene expression patterns in response to internal or external cues. For any particular transcription factor, maps of the binding affinity for different DNA sequences must be obtained through laborious and expensive experiments. We used Blue Waters as our computing resource to computationally map the TFs' binding free-energy landscapes for several well-studied transcription factors with known crystal structures. Comparing our results with experimental data set on the same systems, we observe generally poor correlations among the computational predictions and experimental data. We used a robust computational protocol for reliable *in silico* determination of TF affinity landscapes; however, there is still some element of uncertainty with nonequilibrium molecular dynamics simulations likely to play a key role in defining the binding free-energy landscape. The remaining bottleneck in high-accuracy prediction of protein–DNA binding free-energy landscapes using these methods thus remains an area of active investigation.

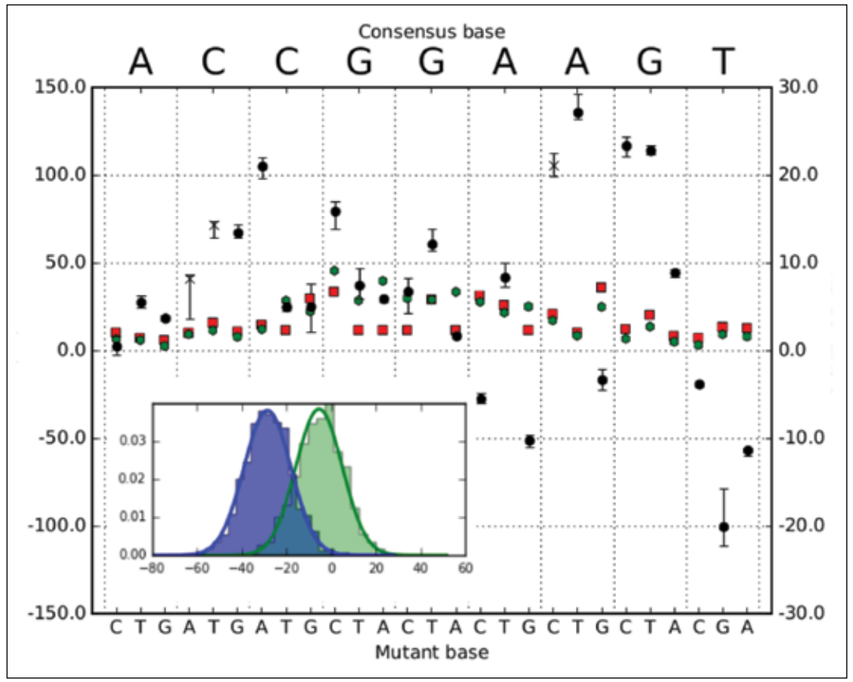

Figure 1: Binding free energy landscape for ELK1. Computational values (black) are compared with human (green) and mouse (red) values from experimental data. The small box in the landscape plot indicates example of work distribution achieved from mutant→wildtype (blue) and wild type→mutant (green) transition. Data from [5].

RESEARCH CHALLENGE

Transcriptional regulation is driven in large part by the action of TFs and other DNA-binding proteins that either recruit or inhibit the recruitment of RNA polymerase. The activity of TFs that interact directly with DNA depends on their potent ability to recognize and bind to specific DNA sequences. In reality, most TFs bind to a range of related sequences; however, the one that provides enough information or has the lowest energy will define a reliable TF binding site. Thus, understanding the binding affinity landscape of each transcription factor to all possible sequences of DNA will help us to predict the behavior of regulatory networks. There have been a number of efforts to predict these binding sites quantitatively based on both experimental methods (e.g., protein-binding microarrays [1] or HT-SELEX experiments [2]). However, most of these experiments carry high costs in terms of time and money. Recent computational efforts by Seeliger and co-workers [4] predicted an accurate DNA–binding affinity landscape based on atomistic molecular dynamics simulations and reproduced experimental results to near-chemical accuracy. This proved to be much more accurate than previous computational methods. By using the massive computing resources provided by Blue Waters, we attempted to apply a similar method for free-energy calculations to that used by Seeliger and coworkers to map the complete binding free-energy landscapes of four TFs with well-characterized binding affinity landscapes.

METHODS & CODES

Building on previous results that showed accurate calculation of protein–DNA binding free energies for a small number of cases [4], we applied the Crooks-Gaussian intersection (CGI) method [6] to calculate the free-energy changes for base pair substitutions in the binding site of the transcription factors of interest. This method requires calculations of very long equilibrium simulations of the protein-DNA complex and the DNA alone for each of two sequences to be compared, followed by many short simulations morphing the system between the two sequences. We performed the free-energy calculations for all possible single nucleotide perturbations of the consensus binding site for the transcription factor of interest. Our results illustrate both very poor overall correlations with the experimental results and, frequently, unphysically large magnitudes of binding free-energy changes [5]. Excluding many sources of possible errors in our simulation setup, we realized that neither extensive control simulations using simplified systems or other free-energy calculation methods, nor careful characterization of the structural features involved in the protein-DNA interface of the simulated complexes, provided an explanation for the poor correlation between calculated and experimental binding free-energy landscapes. We are currently working to resolve this difficulty so that we can realize the promise of computational predictions of DNA-binding affinity landscapes.

WHY BLUE WATERS

The computational work described here requires the capability to efficiently bring huge numbers of nodes together to run dozens of simulations of independent trajectories using GPU-accelerated molecular dynamics software, and then, for each such trajectory, to perform more than 100 short follow-up simulations using CPU-only code for the free-energy calculation. The hybrid architecture of Blue Waters has been absolutely ideal for these applications, providing us with the most efficient possible environment for each portion of our workflow, and allowing us to make progress on huge numbers of mutational calculations simultaneously.

PUBLICATIONS & DATA SETS

Khabiri, M. and P.L. Freddolino, Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein-DNA Binding Free Energy Landscapes. *J. Phys. Chem. B*, 121:20 (2017), pp. 5151–5161.