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

Allocation: GLCPC/383 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. Using the massive computing resources available through Blue Waters, we are pursuing a strategy to computationally map the DNA-binding affinity landscapes of several human transcription factors. Through comparison with experimental results on the same systems, we will validate and refine computational protocols for allowing reliable *in silico* determination of TF affinity landscapes, obtain completely novel insight into the structural basis for these affinity landscapes, and catalog the effects of the binding of different transcription factors on DNA structure, which appears likely to play a key role in the interplay between different transcription factors regulating the same gene *in vivo*.

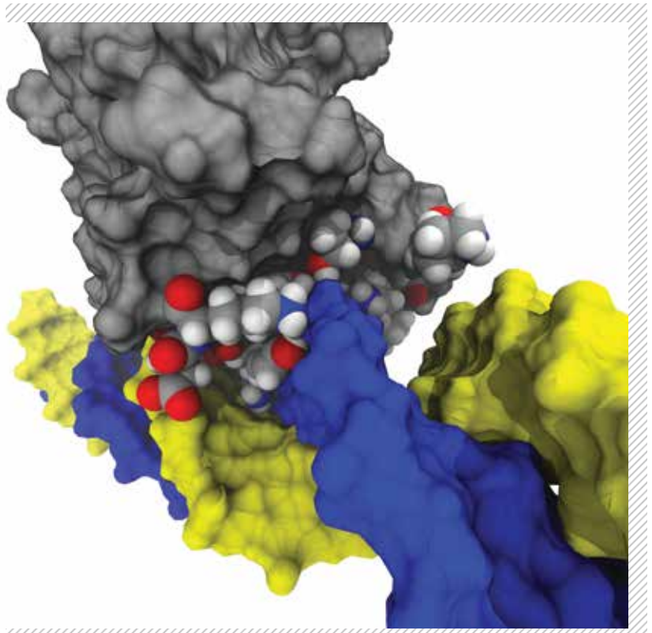

FIGURE 1: Structure of the transcription factor ELK1 (grey) bound to DNA (yellow, blue); the key residues on the protein involved in binding are shown outside of the surface.

INTRODUCTION

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; thus, to understand and predict the behavior of transcriptional regulatory networks, it is necessary to know the landscape of binding affinities of each transcription factor to all possible sequences of DNA. While experimental methods have been developed to measure these landscapes (e.g., protein binding microarrays (PBMs) [1] or HT-SELEX experiments [2]), these experiments remain expensive, labor intensive, and provide no structural insight into the nature of the protein-DNA complex or the specific interactions governing affinity landscapes.

The need for a **more efficient method** to obtain sequence affinity landscapes, insight into the structural basis of these landscapes, and information on the structural effects of different transcription factors on DNA all argue for the large-scale application of molecular modeling to study the affinity landscapes of DNA-binding proteins. Prior efforts to predict TF binding affinity landscapes using approximate methods performed well in identifying the native binding site for a particular protein, but poorly in efforts to more broadly map the binding affinity of a given protein for a variety of sequences (reviewed in [3]). Extremely promising preliminary work using all-atom simulations with explicit solvent (and thus requiring far fewer approximations) has shown near chemical accuracy (average absolute error of 0.31 kcal/mol to experimental values) for all possible single base pair perturbations of the native target site for the zinc-finger transcription factor Zif268 [4].

Using the massive computing resources provided by Blue Waters, we are applying similar atomistic free energy simulations to map the complete sequence affinity landscapes for a set of four carefully chosen human transcription factors for which direct comparisons with experimental results are possible. Our computational protocol requires long equilibrium simulations of the system with each of the DNA sequences being considered. Thus, we will obtain information on the binding landscape of the protein and gain insight into the structural basis for specificity of different transcription factors by allowing detailed analysis of the structural changes associated with mutations in the target sequence. We will use the availability of direct experimental comparisons to **benchmark** the accuracy of our methods, and test force field modifications to enhance performance.

METHODS & RESULTS

Building on previous results that showed accurate calculations of protein-DNA binding free energies for a small number of cases [4], we are applying the Crooks-Gaussian intersection (CGI) method [5] to calculate the free energy changes for base pair substitutions in the binding site of the transcription factors of interest. The 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 will perform the free energy calculations for all possible single nucleotide perturbations of the consensus binding site for the transcription factor of interest, and subsequently pursue additional mutations through a tree-like approach (suggested by [4]) in which at each layer of mutations, only those which have not been shown to strongly inhibit binding will be considered in the next round.

Our results to date illustrate that the free energy changes from *in silico* calculations on the protein-DNA complex can indeed reproduce the consensus binding sequence of transcription factors obtained from HT-SELEX data, and additionally show that the ensemble of thermally accessible protein-DNA interaction conformations in the bound state is far broader than what might be inferred from crystallographic structures. Our findings on the latter point may be particularly relevant to future efforts in developing streamlined prediction or design of protein-DNA interfaces.

WHY BLUE WATERS

The computational work described here requires the capability to bring huge numbers of nodes efficiently together to run dozens of simulations of independent trajectories using graphics processing unit-accelerated molecular dynamics software, and then for each such trajectory to perform more than 100 short follow-up simulations using central processing unit-only code for the free energy calculation. The hybrid architecture of Blue Waters has been 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.

NEXT GENERATION WORK

Recent findings by our collaborators and others have demonstrated that when two or more transcription factors bind to nearby sites on DNA, they can alter each other's sequence affinity landscapes, both through direct protein-protein interactions and the effects of binding on DNA structure. We will use the Track 1 system to simulate two cases each of protein-mediated and DNA-mediated binding site modulation, with an aim toward identifying the biophysical basis for the observed changes in sequence specificity.