Genome-wide prediction of minor groove electrostatics enables biophysical modeling of protein-DNA binding

Tsu-Pei Chiu and Remo Rohs\*

Molecular and Computational Biology Program, Departments of Biological Sciences, Chemistry, Physics, and Computer Science, University of Southern California, Los Angeles, CA 90089, USA

\* To whom correspondence may be addressed. Email: rohs@usc.edu

Abstract: Protein-DNA binding is at the core of gene regulatory and other cellular processes. However, the underlying mechanisms of how proteins recognize their genomic target sites are still not completely understood. Protein-DNA recognition involves chemical (base readout) and structural (shape readout) properties of DNA. Using structural analysis, we previously concluded that minor groove geometry was as an indirect representation of electrostatic interactions between positively charged amino acids and the minor groove. Here, we confirm this readout mechanism by directly using electrostatic potential in genome-wide studies of protein-DNA binding based on massive sequencing data. For this study, we required a method for predicting electrostatic potential in the minor groove in a computationally very efficient way without compromising accuracy. The methodology we introduce uses a sliding-window approach to mine results based on nonlinear Poisson–Boltzmann (NLPB) calculations on DNA structures obtained from Monte Carlo simulations. This approach only requires nucleotide sequence as input and instantly predicts electrostatic potential. We validated this method based on direct comparison with NLPB calculations for available crystal structures. Using machine-learning approaches, we showed that adding electrostatic potential as biophysical feature can improve the predictive power of existing binding specificity models across 25 different transcription factor families. High-throughput prediction of electrostatic potential offers a novel way to integrate biophysical and genomic studies of DNA-binding proteins and their target sites.