DNA Shape Readout Specificities of Different TF Families

Yaron Orenstein[1], Lin Yang[2], Arttu Jolma[3], Jussi Taipale[3], Remo Rohs[2], Ron Shamir[1]
[1]*Tel-Aviv University.* [2]*University of Southern California.* [3]*University of Helsinki.*

ABSTRACT

Protein-DNA binding is key in regulation of gene transcription. To date, PWM models have been the prevailing tool for predicting protein-DNA binding. New high-throughput data have shown that the positional-independence assumption underlying the PWM model is inaccurate. DNA shape models encode these dependencies via a biophysical interpretation of protein DNA shape readout. They have been used successfully to improve binding prediction and to understand its mechanism, but for only a few proteins. Recently, HT-SELEX data have been published for more than 400 mouse and human proteins, representing 40 different protein families (Jolma *et al*., Cell 2013). This provided the first opportunity to explore DNA shape contributions extensively.

Using HT-SELEX data, we analyzed DNA shape models for 106 proteins from 25 families. We used 10-fold more sequencing data than reported in Jolma *et al*. in order to allow for better inference of k-mer binding scores. For each experiment, we scored k-mers containing the protein core motif, and used machine learning methods to construct binding models. The shape-augmented PWM models performed better than PWM models by a margin of more than 10% in predicting binding intensity (Fig. 1A). Moreover, we used feature selection to pinpoint which positions along the binding sites are more likely to play a part in the shape readout, and summarized this information in 'DNA shape logo' (Fig. 1B).

Our results show that TF shape readout is important for most of protein families, and that DNA shape models improve binding prediction. High-resolution positional shape preference profiles can be derived from high-throughput quantitative binding data. As a byproduct, we provide a pipeline to derive accurate k-mer scores from HT-SELEX data. These scores give a much richer description of the binding landscape than PWMs, and can be used for other applications.

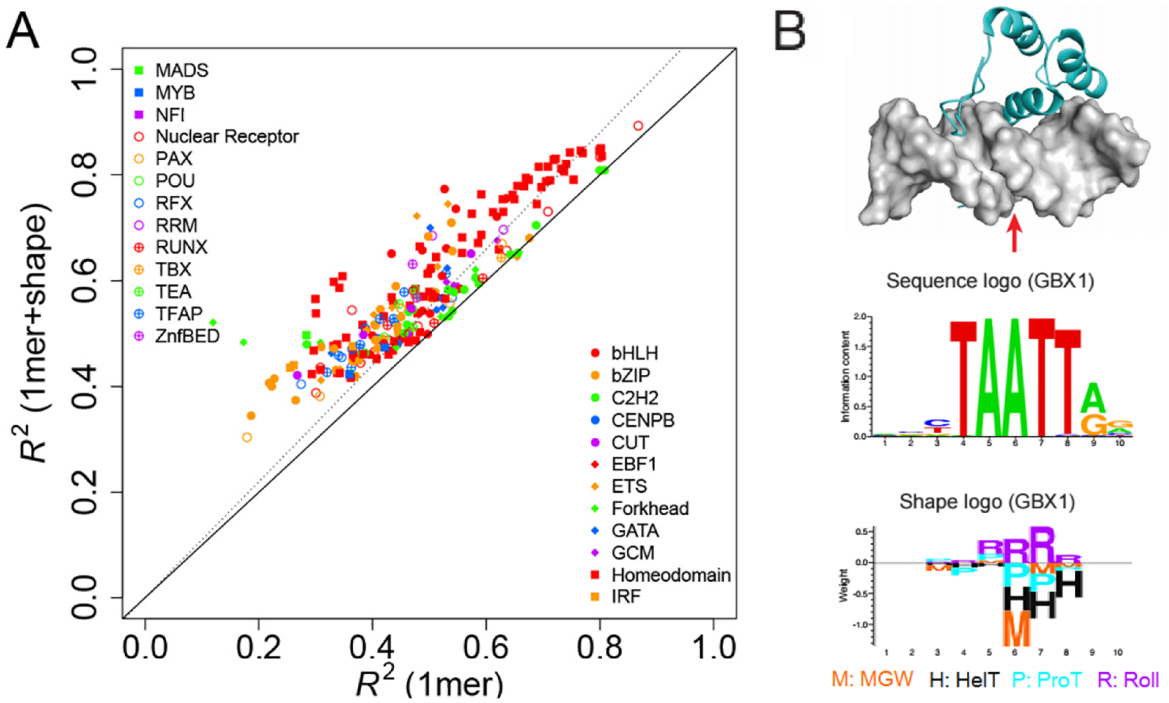

Figure 1. Binding model that utilize DNA shape improve binding prediction and understanding of the binding mechanism for diverse TF families. A) For all tested TF families, models using both PWM and shape features have improved binding prediction compared to PWM models. B) Sequence and shape logos for GBX1. The shape logo represents the positional preference for different shape features, and pinpoints positions that are more likely to play a part in shape readout. MGW: minor groove width, H: helix twist, ProT: propeller twist, R: roll.