# New Approaches to Genome Analysis Based on the Integration of DNA Sequence and Shape

Iris Dror[1,2], Tianyin Zhou[1], Lin Yang[1], Yael Mandel-Gutfreund[2] & Remo Rohs[1,*]

1 Molecular and Computational Biology Program, Departments of Biological Sciences and Chemistry, University of Southern California, Los Angeles, United States.
*rohs@usc.edu
2 Faculty of Biology, Technion – Israel Institute of Technology, Haifa, Israel.

High-throughput sequencing technologies continue to produce large amounts of DNA sequence information. Whereas analyzing the genome as a linear one-dimensional string of letters provides answers to many biological questions, proteins recognize DNA as a three-dimensional object. Considering DNA as a double helix with sequence-dependent shape enables the biophysical characterization of protein-DNA readout. We developed a novel high-throughput approach to compute DNA shape on a genomic scale and used this method to analyze the DNA shape of thousands of sequences from SELEX-seq and PBM experiments for Hox binding sites in Drosophila and mouse. We have found that anterior and posterior Hox proteins prefer DNA sequences with distinct minor groove topographies. More so, we suggest that DNA shape indicates how Hox genes have differentiated in evolution. We also used our new DNA shape prediction to develop a novel approach for the de-novo discovery of TF binding motifs that incorporate sequence and shape features. We have tested this method based on Hoxa2 ChIP-seq data and ChIP-chip data for 81 yeast TFs and were able to show that an integration of sequence and structural information can help not only in adding shape information to already known binding motifs, but also in finding novel binding motifs which were not found using sequence alone.