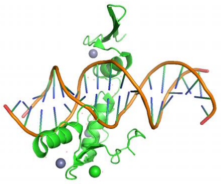
Predicting protein-DNA binding affinity from structure and sequence

Alex Tseng

Background



Some proteins bound to DNA have been captured in detail using X-ray crystallography.

Using ChIP-seq, we can greatly expand the set of DNA sequences a protein binds to and does not bind

Goals

Learn a model that predicts whether or not a protein will bind to a DNA strand.

Given an example of a protein bound to any sequence, use that structure to determine binding affinity to any new sequence.

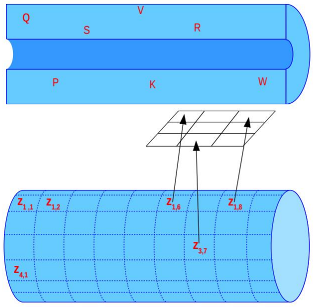
Be able to mutate the protein or a DNA strand to determine how it affects binding affinity.

Dataset

500 crystalline structures of different proteins bound to a single DNA sequence.

With ChIP-seq data, this gives 50000 total protein-to-DNA pairs (half binding, half non-binding).

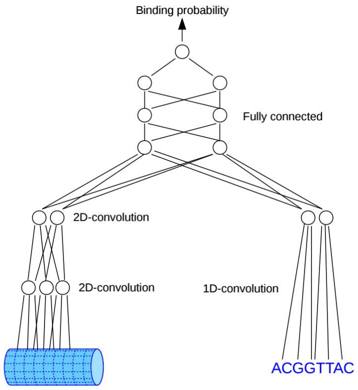
Featurization



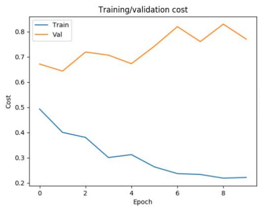
Each entry is a vector encoding location/direction, physical features, and some identity features.

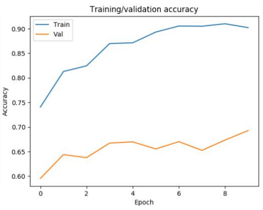
For DNA, 1-hot encode the sequence, and add shape features: Roll, MGW, ProT, HelT.

Architecture



Results





But similar results are achieved by zeroing out the protein input, which means that structure is being ignored.

Also see similar results even if the dataset is limited to only the canonical sequences seen in the crystalline structures themselves.

Discussion

Structure is not being utilized, and learning is based on identifying binding based on DNA sequence alone.

The same results come from only the canonical sequences, even though it is possible to learn from these canonical sequences alone (as demonstrated by the derivation of working statistical potentials).

The issue is likely that the selected features for the protein-DNA binding interface structure are inappropriate/insufficient.

Next steps are to consider new features, perhaps closer to an atomistic view but sufficiently high-level to allow for simplicity when mutating proteins/DNA.

Acknowledgements: Joe Paggi, David Eng, Ron Dror