Leveraging contextual information to explain transcriptional activation domain activity

Charles Hodgens

North Carolina State University

Certain intrinsically disordered regions (IDRs) in transcription factors serve a functional role in gene regulation by recruiting the Mediator complex. These regions, known as transcriptional activation domains (ADs), have undergone decades of research, yet we lack a complete understanding of the rules governing their function. Activation domains are often explained using the acid-exposure model, where negatively charged acidic residues expose buried hydrophobic and aromatic residues which serve as contact points for other proteins. Through analysis of large datasets of *Arabidopsis* transcription factor-derived sequences and their activity as activation domains, our group has shown that the acid-exposure model is insufficient to explain all activation domains. Clearly, the rules governing activation domain function are more complex than previously thought.

We hypothesize that cellular context is an important variable for activation domain function. Activation domain sequences may be functional in only certain cell types, due to differences in cytoplasmic or nuclear contents or due to biochemical state; if true, contextual information is a required component for any model attempting to explain activation domain function. We have published a large dataset describing *Arabidopsis* protein fragments and their *in vivo* function as activation domains. We plan to supplement this data with cell type specific transcriptional data derived from *Arabidopsis* root tissue atlases.

Our final dataset—consisting of protein fragments linked with independently acquired context-dependent measures of gene expression—will be analyzed using a conditional variational autoencoder (CVAE) model. Variational autoencoders (VAEs) are a class of neural networks capable of learning summarized representations (a latent space) of input data and non-deterministically generating novel outputs. Conditional autoencoders (CAEs) supplement the latent space with contextual information, freeing the training process to focus on features independent of the provided context. Critically, CAE models permit the user to generate completely novel output by modifying the context provided to the latent space. A model combining features of both archetypes should permit the tailored generation of novel protein fragments with desired biochemical function. Development of this model, and interrogation of the input features it is attentive to, will improve our understanding of the biochemical features controlling activation domain activity and provide tools for designing novel protein fragments to adjust cell type specificity or activation ability.