A SURROGATE MODELING FRAMEWORK FOR INTERPRETING DEEP NEURAL NETWORKS IN FUNCTIONAL GENOMICS

Evan Seitz, Peter Koo, Justin Kinney

Cold Spring Harbor Laboratory, Simons Center for Quantitative Biology, Cold Spring Harbor, NY

Understanding the cis-regulatory grammars that coordinate how proteins interact to regulate transcription is a major goal in genomics. Deep neural networks (DNNs) applied to this task have greatly enhanced our ability to accurately predict experiments in regulatory genomics. Despite their impressive performance compared to traditional methods in computational genomics, it remains difficult to determine how these networks form their decisions. To address this gap, attribution methods are being increasingly used to gain mechanistic insights underlying DNN predictions. Attribution methods probe the trained DNN to assess the importance of each nucleotide in a sequence to produce an attribution map, which has been shown to visualize known functional motifs and their locations. However, current attribution methods are sensitive to the local function properties learned by the DNN, making identification of functional motifs difficult. Due to the high expressivity of DNNs, there is no guarantee this issue can be resolved by altering the DNN to learn smoother functions amenable to attribution maps.

Instead, we surmise that attribution-based explanations can be made more robust by approximating a larger region of function space anchored at a given sequence of interest with an interpretable surrogate model, for which the parameters provide direct interpretations of variant importance similar to attribution maps. Here we introduce this new surrogate modeling approach into genomics, where it has not yet been explored. Our approach, called SQUID for Surrogate QUantitative Interpretability of Deepnets, is a general framework that leverages interpretable surrogates to quantitatively model the sequence-function relationship learned by any black-box genomic model. We demonstrate our framework across several existing DNNs designed to perform a variety of regulatory genomics prediction tasks. For each of these DNNs, we show that SQUID outperforms existing attribution methods in studies spanning ensembles of high-functioning motifs and genomic sequences. From this comparison, we find that SQUID is able to more robustly characterize the direct effect of motifs and their higher-order interactions on predictions, consistently model larger sequence contexts, identify weaker binding sites that enable opportunities for better annotation, and provide better approximations to variant effect predictions. SQUID provides a leap forward in our ability to decipher the quantitative effects of cis-regulatory elements throughout the genome.