Dissertation Defense

Deep Models for Gene Regulation

Olgert Denas
Emory University

Abstract: The recent increase in the production pace of functional genomics data has created new opportunities in understanding regulation. Advances range from the identification of new regulatory elements, to the prediction of gene expression from genomic and epigenomic features. At the same time, this data-rich environment has raised challenges in retrieving and interpreting information from these data.

Based on recent algorithmic developments, deep artificial neural networks (ANN) have been used to build representations of the input that preserve only the information needed to the task at hand. Prediction models based on these representations have achieved excellent results in machine learning competitions. The deep learning paradigm describes methods for building these representations and training the prediction models in a single learning exercise.

In this work, we propose ANN as tools for modeling gene regulation and a novel technique for interpreting what the model has learned.

We implement software for the design of ANNs and for training practices over functional genomics data. As a proof of concept, use our software to model differential gene expression during cell differentiation. To show the versatility of ANNs, we train a regression model on measurements of protein-DNA interaction to predict gene expression levels.

Typically, input feature extraction from a trained ANN is formulated as an optimization problem whose solution is slow to obtain and not unique. We propose a new efficient technique for classification problems that provides guarantees on the class probability of the features and their norm. We use this technique to identify input features used by the trained model in classification and show how these features agree with previous empirical studies.

Finally, we propose building representations of functional features from protein-DNA interaction measurements using a deep stack of nonlinear transformations. We train the model on a small portion of the input and compute small dimensional representations for the rest of the genome. We show that these reduced representations are informative and can be used to label parts of the gene, regulatory elements, and quiescent regions.

While widely successful, deep ANNs are considered to be hard to use and interpret. We hope that this work will help increase the adoption of such models in the genomics community.

Thursday, March 27, 2014, 4:00 pm
Mathematics and Science Center: W302

Advisor: James Taylor

Mathematics and Computer Science
Emory University