Netform: From multi-species raw data to causality

Our goal is to use data about multiple

species to infer causal relationships among the genes of individual species.

To achieve this, we start with raw experimental data from several species, process them through a user-chosen variety of cleaning and normalization steps (aim 1) (JI). We then apply a set of analytical tools to infer likely causal relationships; the choice of tool and the hyperparameter settings will be determined automatically (aim 2)(Dennis).

Finally we offer visualization tools to understand both the causal links and .... [whatever else] (Aim3 - Manny)

The main contributions of NetForm are:

1. Creating a multispecies raw data-to-causality pipeline.

2. The ability to parallelize processing to achieve scalability.

3. The capability to handle non-linear as well as linear relationships

4. Applying ideas from automatic machine learning to the entire raw data-to-causality pipeline.

Recent work [hit-and-run paper and friends \cite{Gitter:2009}] by our lab and others has shown that binding may or may not be closely related to transcriptional influence. Other work [e.g. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0153295>

] has shown that the relationship between transcription factors and each target may be non-linear (i.e. more than a weighted sum of individual transcriptional effects). Still other work [need refs supporting multi-networks] has shown that transcriptional relationships from one species may carry over to homologous genes in other species. Finally, the effectiveness of automatic machine learning suggests that parameter and hyperparameter settings that influence both data preparation (e.g. fold change cutoffs) and analytical processing (e.g. depth of trees in a random forest) can be learned.

NetForm is a set of tools that can be organized into a pipeline by the user and can be tuned using machine learning techniques that (i) can apply a variety of normalization and data cleaning techniques for data preparation, (ii) can infer causal (or at least plausibly causal) relationships between transcription factors and targets within a species s using non-linear machine learning techniques on data from s and from other species s1, s2, … through homology, (iii) can run in parallel across the target genes of species s, (iv) preserves the provenance of each output conclusion for purposes of reproducibility, (v) provides visualization tools to ….

[Ji does the variety of normalization and data cleaning]

[Manny does the visualization]

Overview of the Analytical Part

The heart of our analytical pipeline is an analytical method called OutPredict [under revision at …] that not only constructs non-linear functions from the expression of transcription factors to the genes but also incorporates into those functions hyperparameters such as (i) the number and depth of trees in the forest, (ii) the extent to which binding and methylation assays should influence the selection of transcription factors to be branch points in the individual decision trees, and (iii) in the case of time series data the way to model the time difference between consecutive measurements.

OutPredict is based on random forests [Breiman] like other state-of-the-art methods [Genie 3 reference], because random forests express non-linear functions and require less data than neural networks. Unlike other random forest models, OutPredict can incorporate a variety of data sources such as the hyperparameters mentioned above but also, potentially, data from homologous genes of other species. Further, because OutPredict constructs a model for each gene independently, it can run in parallel thus achieving scalability.

Our experiments have shown that OutPredict is better at predicting out-of-sample expression values in time series than other state-of-the-art methods. Because transcription factors high in the decision trees of the random forest for each gene g that OutPredict produces have greater influence on g than other transcription factors, OutPredict naturally produces candidate causal transcription factors for gene g.

of functional genomics data types.