

Modeling of Signal-Response Cascades using Decision Tree Analysis

Sampsa Hautaniemi¹, Sourabh Kharait², Akihiro Iwabu², Alan Wells²,
Douglas A. Lauffenburger¹

Keywords: Systems biology, signaling transduction, migration, simulation, decision trees

1 Introduction.

Signal transduction cascades governing cell functional responses to stimulatory cues play crucial roles in cell regulatory systems and represent promising therapeutic targets for complex human diseases. However, mathematical analysis of how cell responses are governed by signaling activities is challenging due to their multivariate and nonlinear nature. Moreover, inherent noise in the measurements and small sample sizes further complicate the analysis.

We have two objectives in our analysis of signal transduction cascades. The first is to build a model from which the most relevant signaling proteins in regard to response can be identified. The second is to assess the prediction accuracy of the model. The algorithmic approach we propose to employ to meet these goals is decision tree modeling.

2 Methods.

We present a framework of computational techniques for elucidating useful models of the relationships between protein signals and cell functional responses to extracellular cues in a manner driven by quantitative data across diverse conditions. The discussion is divided into three parts: preprocessing, simulation and classification.

Often the experimental data are noisy and the amount of observations is inadequate for dependency modeling or prediction. Therefore, before analyzing the data with decision trees, the data should be preprocessed. Here we present and apply an analysis of variance (ANOVA) based quality control approach for replicate measurements.

Despite the development of high-throughput measurement technologies, in a typical biological experiment the amount of the data is insufficient for robust analysis. Here we suggest and apply an interpolative simulation of additional, internally-consistent data points. The simulation is based on polynomial fitting, where the degree of the polynomial is determined with the normalized maximum likelihood (NML), which in turn is based on the minimum description length (MDL) principle.

We perform the classification using decision trees, which have many appealing properties in biomedical research. For example, decision trees are robust against outliers and can be effectively applied to any data structure, such as discrete, continuous or mixed data. Moreover, prediction rules are easy to interpret and tree visualization provides information on the predictive structure of the data.

We demonstrate the presented methodology with a specific example case study of five intracellular signals (EGFR, ERK, MLC, PKC δ , PLC γ) influencing fibroblast migration under eight conditions: Four substratum fibronectin levels and presence vs. absence of

¹Biological Engineering Division, MIT, E-mails: sampsa@mit.edu, lauffen@mit.edu

²Department of Pathology, University of Pittsburgh. E-mails: sok4+pitt.edu, akihiro.iwabu@nifty.com, wellsa@upmc.edu

epidermal growth factor. Our experimental measurements are accomplished by quantitative immunoblotting procedure after either 5 minutes or 1 hour after introducing epidermal growth factor.

3 Results and conclusions.

For this specific case study, our approach achieves 70% and 76% overall classification accuracies for 5 minutes and 1h data, respectively. An example of a decision tree for 5min data set is given in Figure 1. The decision tree models reveal insights concerning the combined roles of the various signaling activities in governing cell migration speed and provide directions to hypothesize the response of signaling proteins to extracellular cues. The methodology above and the results for the 5 minutes data set are published in [1].

These results highlight the central idea of systems biology, i.e. complex biological processes cannot be analyzed by perturbing only one component at a time but there is a need to study several components simultaneously. We conclude that a decision tree methodology is likely to be useful in a wide spectrum of biomedical research by facilitating elucidation of signal-response cascade relationships and offering directions for future experiments which could test predictions arising from these relationships.

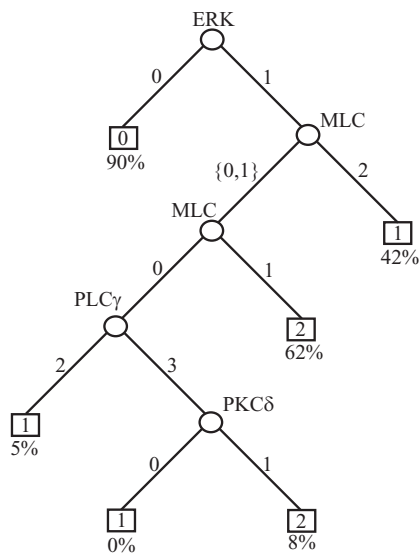


Figure 1: The decision tree for the 5min data set. Round nodes denote proteins, while square nodes represent estimated migration speed categories. Integers attached to the arcs correspond to the splits of the parent node. Under each migration speed category the fraction of cases explained by that classification rule is given. For example, if $ERK = 0$, the migration speed category is 0, and 90% of the observations (in the training set) for migration speed category 0 can be explained with this rule.

References

- [1] Hautaniemi, S., Kharait, S., Iwabu, A., Wells, A. and Lauffenburger, D. A. 2005. Modeling of signal-response cascades using decision tree analysis. In print *Bioinformatics*.