A global feature selection algorithm for the model selection step in the identification of cell signaling networks

Gustavo Estrela de Matos[1], Lulu Wu[2], Vincent Noel[3], Marco Dimas Gubitoso[4], Carlos Eduardo Ferreira[4], Junior Barrera[4], Hugo A. Armelin[3], Marcelo S. Reis[3]

1 CENTER OF TOXINS, IMMUNE-RESPONSE AND CELL SIGNALING, INSTITUTO BUTANTAN, INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
2 CENTER OF TOXINS, IMMUNE-RESPONSE AND CELL SIGNALING, INSTITUTO BUTANTAN
3 INSTITUTO BUTANTAN
4 INSTITUTO DE MATEMÁTICA E ESTATÍSTICA, USP

Abstract

In the context of cell signaling network identification, model selection is the choice of a dynamic model from a set of possibilities; the chosen model should be the most suitable one according to a given cost function (e.g., curve-fitting optimization). If these possibilities are defined by differences in the chemical species and/or reactions that compose each of them, then a feature selection procedure could be carried out to accomplish the model selection. Recently, it was proposed a method to carry out model selection through examination of interactome databases. However, such databases typically yield huge search spaces during the feature selection procedure; hence, only a greedy sequential approach could be explored so far. Therefore, there is a need for development of efficient global feature selection methods to tackle this hard combinatorial optimization problem. In this work, we introduce a new global feature selection method, which may be used to assist the model selection step during the identification of cell signaling networks. This method, called Parallelized U-Curve Search (PUCS), relies on the fact that the chain costs of the Boolean lattice induced by the search space are decomposable in U-shaped curves; this latter phenomenon is due the curse of dimensionality, that is, the impact the lack of samples brings to the cost function as the number of considered features increases. To implement and evaluate the PUCS algorithm, we used featsel, a framework for benchmarking of feature selection algorithms and cost functions. To compute the cost function (i.e., the fitness of a candidate model), we are employing the Signaling Network Simulator (SigNetSim), a tool for building, fitting, and analyzing mathematical models of molecular signaling networks. Initial results with synthetic data showed that PUCS outperforms golden standard algorithms in feature selection such as the Sequential Forward Floating Search (SFFS). Currently, we are applying PUCS into the model selection of real-data signaling networks extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) interactome database. We expect that this new feature selection method will become a critical asset to identify fully predictive dynamic models, which in turn will help researchers to unveil intricate molecular mechanisms underlying cell phenotype changes due extracellular stimuli.

Funding: CAPES, CNPq and grants #2013/07467-1 and #2016/25959-7, São Paulo Research Foundation (FAPESP)