Learning Yeast Functional Upstream Open Reading Frames from Heterogeneous Data Sources

Selpi[1], Christopher H. Bryant[1], Graham J.L. Kemp[2], Alexandra Jauhiainen[3], Janeli Sarv[3], Erik Kristiansson[3], Marija Cvijovic[4], Per Sunnerhagen[5], Olle Nerman[3]

[1]School of Computing, The Robert Gordon University, United Kingdom
[2]Department of Computer Science and Engineering, Chalmers University of Technology, Sweden
[3]Department of Mathematical Statistics, Chalmers University of Technology, Sweden
[4]Max Planck Institute for Molecular Genetics, Germany
[5]Department of Cell and Molecular Biology, Göteborg University, Sweden

Regulation of gene expression is an important process, yet it is not fully understood. We believe this is mainly because very little is known about regulatory elements. Several upstream open reading frames (uORFs) have been shown to play important roles as post-transcriptional regulators in protein expression. However, the mechanism of how these uORFs regulate protein expression is still unclear. To be able to draw a whole understanding of the mechanism, we would expect that a large number of functional uORFs would be needed. Unfortunately, lab-based experiments to identify functional uORFs are extremely expensive and time-consuming. Therefore, an in silico prediction method, which can help in selecting sets of candidate functional uORFs for experimental studies, is essential.

Here, we present a new approach to predicting functional uORFs in the yeast Saccharomyces cerevisiae. Our method uses a machine learning technique, called inductive logic programming (ILP). As negative training data (verified non-functional uORFs) is even more difficult to get and thus scarce, positive-only learning was explored. Unlike many learning techniques which only learn from examples, ILP can make use background/domain knowledge to construct hypotheses (a set of rules). In this study, knowledge derived from biological sequences of several different yeast species, analysis of several publicly available expression data sets measuring translational activity under different stress conditions, and gene ontology annotations were used to form the background knowledge. The input (examples and background knowledge) and the output (hypotheses) of ILP are all represented in predicate logic. This representation is easily translated into English and thus can be read and understood easily by scientists.

The performance of the hypotheses is assessed by leave-one-out cross-validation. This means that each example is in turn used as a test set, while all the remaining examples are used as a training set. This assessment shows that the hypotheses can correctly recognise 81% of the total examples (i.e., sensitivity = 81%). The hypotheses are simple and suggesting that features like conservation in at least two other yeast species, involvement of the main gene product in regulation of biological process, cellular metabolic process, and nucleic acid binding may be important in recognising functional uORFs.

The whole set of examples is then used to construct a set of rules, which is then used to predict novel functional uORFs from a set of unlabelled uORFs within the genome of S.cerevisiae. Our method predicts around 300 further genes to have novel functional uORFs.