Using Inductive Logic Programming to Predict Functional Upstream Open Reading Frames in Yeast

Selpi\textsuperscript{1}, Christopher H. Bryant\textsuperscript{1}, Graham J.L. Kemp\textsuperscript{2}, Marija Cvijovic\textsuperscript{3}

\textsuperscript{1}School of Computing, The Robert Gordon University, St. Andrew Street, Aberdeen, AB25 1HG, United Kingdom
\textsuperscript{2}Department of Computer Science and Engineering, Chalmers University of Technology, SE-412 96, Göteborg, Sweden
\textsuperscript{3}Max Planck Institute for Molecular Genetics, Department Lehrach, Kinetic Modelling Group, Ihnestrasse 63-73, 14195, Germany

Upstream open reading frames (uORFs), that can be present in the 5' untranslated region (UTR) of mRNA, can play important roles as post-transcriptional regulators in gene expression. Laboratory experiments to test whether a gene contains functional uORFs are costly and time-consuming, and so far very few functional uORFs have been verified experimentally. This is an obstacle in understanding uORFs’ roles.

Predicting functional uORFs can help in understanding uORFs’ roles in gene expression. We have developed a novel, inductive logic programming (ILP), approach to predicting functional uORFs in the yeast Saccharomyces cerevisiae. The input to the ILP system is a set of positive examples and background knowledge. In this case, the positive examples are a set of known functional uORFs, and the background knowledge is a set of features and rules that describe uORFs, UTRs and the relationships between these. The ILP system uses the positive examples and background knowledge to generate a hypothesis (a set of rules, also called the model). These rules can be used to predict whether unseen uORFs are functional. The examples, background knowledge and the hypotheses are all represented in predicate logic. This representation is easily translated into English and thus the hypotheses generated by the system can be read and understood easily by scientists. The method makes searching for novel functional uORFs more efficient than random sampling. The model is simple and may be biologically meaningful.