Using mRNA secondary structure predictions improves recognition of known yeast functional uORFs

Selpi, S, Bryant, CH and Kemp, GJL

Title	Using mRNA secondary structure predictions improves recognition of known yeast functional uORFs
Authors	Selpi, S, Bryant, CH and Kemp, GJL
Type	Book Section
URL	This version is available at: http://usir.salford.ac.uk/1752/
Published Date	2008

Using mRNA Secondary Structure Predictions Improves Recognition of Known Yeast Functional uORFs

Selpi[1*], Christopher H. Bryant[2], and Graham J.L. Kemp[3]

[1] School of Computing, The Robert Gordon University,
St. Andrew Street, Aberdeen, AB25 1HG, UK
[2] School of Computing, Science and Engineering, Newton Building,
University of Salford, Salford, Greater Manchester, M5 4WT, UK
[3] Department of Computer Science and Engineering,
Chalmers University of Technology, SE-412 96 Göteborg, Sweden

Abstract. We are interested in using inductive logic programming (ILP) to generate rules for recognising functional upstream open reading frames (uORFs) in the yeast *Saccharomyces cerevisiae*. This paper empirically investigates whether providing an ILP system with predicted mRNA secondary structure can increase the performance of the resulting rules. Two sets of experiments, with and without mRNA secondary structure predictions as part of the background knowledge, were run. For each set, stratified 10-fold cross-validation experiments were run 100 times, each time randomly permuting the order of the positive training examples, and the performance of the resulting hypotheses were measured. Our results demonstrate that the performance of an ILP system in recognising known functional uORFs in the yeast *S. cerevisiae* significantly increases when mRNA secondary structure predictions are added to the background knowledge and suggest that mRNA secondary structure can affect the ability of uORFs to regulate gene expression.

1 Introduction

Uncovering the mechanisms that regulate gene expression at a system-level is an important task in systems biology. Understanding the roles of post-transcriptional regulatory elements in gene expression is one aspect of this. Upstream open reading frames (uORFs) are among the regulatory elements that can be present in the 5′ untranslated region (UTR) of messenger RNA (mRNA). In the yeast *Saccharomyces cerevisiae*, some uORFs have been well studied and it has been verified that some of these regulate gene expression (i.e. they are functional) [1–5], while a few others do not (i.e. they are non-functional) [6, 7]. The mechanism by which uORFs regulate genes is still only partially understood. This is mainly because wet-lab experiments to test whether a gene contains functional uORFs are costly and time-consuming.

[*] To whom correspondence should be addressed.

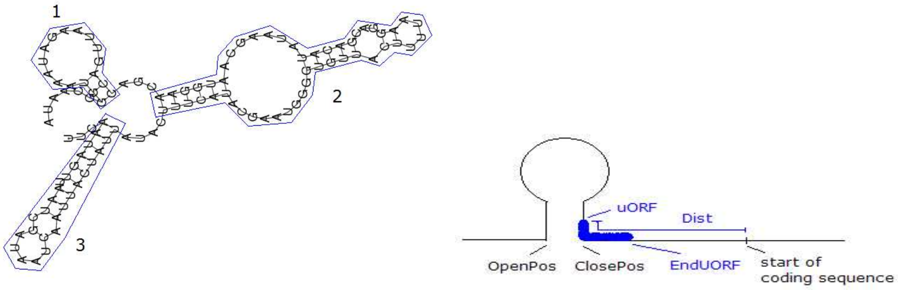

Fig. 1. Left: A predicted secondary structure of the 5′ UTR sequence and ten nucleotides of gene *YAP2* (YDR423C), made by RNAfold. The boxes have been added to show how we view the structure as three stem-loop structures. Right: Illustration of a uORF intersects with an mRNA secondary structure on the uORF's left (upstream) part.

It has been shown that inductive logic programming (ILP) can automatically generate a set of hypotheses which makes searching for novel functional uORFs (i.e. uORFs which can regulate gene expression) in the yeast *S. cerevisiae* more efficient than random sampling [8]. Those hypotheses were simple and easy to understand, but appeared to be too general. This is due not only to the limited number of positive examples and the high degree of noise in the data, two problems which cannot be easily rectified, but also due to the limited background knowledge.

In this paper, we investigate whether incorporating predicted mRNA secondary structure as background knowledge can increase the performance of the resulting hypotheses in recognising functional uORFs in the yeast *S. cerevisiae*. The type of mRNA secondary structure we consider is the stem-loop. A stem-loop is a simple RNA secondary structure motif that can occur when the transcribed sequence contains an inverted repeat sequence (see Fig. 1). Based on their study on a maize gene, Wang and Wessler [9] concluded that uORF and mRNA secondary structure regulate gene expression independently. However, the results from [10], based on studies on human genes, are rather different; the presence of secondary structure seems to affect uORFs' ability to regulate gene expression. The difference between the conclusions of [9] and those of [10] leave open the question whether mRNA secondary structure influences uORFs' ability to regulate gene expression. This motivates our study; to test whether mRNA secondary structure predictions could help in recognising known functional uORFs in the yeast *S. cerevisiae*.

The rest of this paper is organised as follows. Section 2 describes the dataset and the learning system used in this work. The experimental method, including how we incorporate mRNA secondary structure predictions as background knowledge, is detailed in Section 3. Our results are presented in Section 4. Finally, in Section 5, we discuss our main results and suggest directions for future work.

2 The Dataset and the Learning System

The same dataset that was used for training and testing in [8] was used here for training and testing. For the task of learning which uORFs regulate gene expression, positive examples are verified functional uORFs, and negative examples are verified non-functional uORFs. Since confirmed negative examples are scarce (there are only two compared to 20 positive examples) and given that there are 380 random examples (unlabelled uORFs, most of which are probably negatives), we use the positive-only setting [11] of CProgol [12] version 4.4 [13]; the same was used in [8]. CProgol is an established ILP system which uses a covering approach for hypotheses construction. CProgol has been successfully applied to many different problems, including some in bioinformatics. The positive-only setting of CProgol4.4 learns from both positive and random examples; the random examples can either be provided by the user or generated automatically by CProgol. The random examples used for our experiments here are the 380 unlabelled uORFs. We did not use the system-generated random examples because these could be less informative than unlabelled uORFs and might not represent true examples (i.e., true uORFs).

3 Methods

To enable us to test whether incorporating mRNA secondary structure predictions as background knowledge increases ILP performance when learning which uORFs in yeast are functional, we run two sets of experiments, with and without mRNA secondary structure predictions as part of the background knowledge. For each set, stratified 10-fold cross-validation experiments were run 100 times, each time with a random permutation of the order in which positive training examples are presented to the ILP system; this was done because CProgol4.4 may generate different hypotheses when given different orderings of positive training examples. The same 100 random orderings were used for both sets of experiments. Stratified 10-fold cross-validation means that the set of positive examples is divided into ten roughly equal partitions and the same is done to the set of random examples; each of these positive and random partitions are in turn used as a test set while the rest of the partitions are used as training set. Table 1 summarises our experimental procedure.

The ILP learner was instructed to learn a predicate `has_functional_role/1` from a set of training examples. Positive examples were represented as instances of the predicate `has_functional_role(X)`, where `X` is a uORF ID. A uORF ID is a composite of the systematic name of the gene to which the uORF belongs (for example, YDR423C is the systematic name of gene *YAP2*) and a uORF identifier (e.g., uORF1, uORF2, *etc.*). The definition of the hypotheses space for

Table 1. The Experimental Procedure

```
For i=1 to 100
   Randomly permute the order of examples
   Divide dataset into stratified 10 folds
      Divide set of positives into 10 equal partitions
      Divide set of randoms into 10 roughly equal partitions
      For j=1 to 10
         Concatenate partition j of positives and partition j of randoms
         to create fold j
   For each set of background knowledge
      For j=1 to 10
         Use fold j as test set
         Construct hypotheses using the other nine folds
         Use the resulting hypotheses to classify the test set
      Get the performance of stratified 10-fold cross-validation
      experiments
```

Table 2. Representation of a predicted structure shown in Figure 1. `has_stemloop(X,Y)` represents the relationship between UTR `X` and stem-loop `Y`. `stemloop(W,X,Y,Z)` states that stem-loop `W` has its opening and closing positions in `X` and `Y` bases to the coding sequence; and there are, in total, `Z` base pairs within `W`.

```
has_stemloop(YDR423C, YDR423C_sl3).    stemloop(YDR423C_sl3, 98, 71, 10).
has_stemloop(YDR423C, YDR423C_sl2).    stemloop(YDR423C_sl2, 66, 17, 13).
has_stemloop(YDR423C, YDR423C_sl1).    stemloop(YDR423C_sl1, 13, -3,  3).
```

the experiments without mRNA secondary structure predictions were the same as in Table 5 of [8].

RNAfold [14][4] was used, with its default settings, to generate mRNA secondary structure predictions from sequence data. For each of the 17 well-studied genes, the 5′ UTR sequence and the first ten nucleotides of the coding sequence was used as an input for RNAfold. The length of 5′ UTRs were taken from the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database[5], where available, or, failing that, [1]. The output from RNAfold was transformed into Prolog predicates representing predicted mRNA secondary structure as extensional background knowledge. In this work, we view the predicted mRNA secondary structure from the highest level. This means that we do not consider a nested stem-loop as an independent stem-loop. For example, we only consider *YAP2* to have the three stem-loop structures shown in the left part of Fig. 1 and Table 2.

[4] ViennaRNA-1.6.1 was downloaded from `http://www.tbi.univie.ac.at/~ivo/RNA/`

[5] `ftp://ftp.ebi.ac.uk/pub/databases/UTR/data/5UTR.Fun_nr.dat.gz` version 16 June 2006

Table 3. Additional mode declarations used in experiments with mRNA secondary structure predictions[a].

```
:- modeb(1,is_inside_stemloop(+uORF))?
:- modeb(1,intersectleft_with_stemloop(+uORF))?
:- modeb(1,intersectright_with_stemloop(+uORF))?
:- modeb(*,has_stemloop(+uORF,-stemloop))?
:- modeb(1,stemloop(+stemloop,-pospair1,-pospair2,-numberofpairs))?
:- modeb(1,+numberofpairs=< #int)?
:- modeb(1,+numberofpairs>= #int)?
:- modeb(1,+numberofpairs= #int)?
:- modeb(1,+pospair1=< #int)?          :- modeb(1,+pospair2=< #int)?
:- modeb(1,+pospair1>= #int)?          :- modeb(1,+pospair2>= #int)?
:- modeb(1,+pospair1= #int)?           :- modeb(1,+pospair2= #int)?
```

[a] `modeb` describes the predicates to be used in a hypothesis and has the format: `modeb(RecallNumber,Template)`. `RecallNumber` specifies how many times the `Template` can be called successfully; `*` means the `Template` can be called successfully up to 100 times. `Template` is n-ary predicates, with $n \geq 1$ and each of the arguments is a *variable type* preceded by either a '+' (indicates that the argument should be an input), '-' (indicates that the argument should be an output), or '#' (indicates that the argument should be a constant). The types `uORF` and `stemloop` were declared by defining a set of instances of the predicates `uORF(X)` and `stemloop(Y)` respectively, where `X` is a uORF ID and `Y` is a stem-loop ID. The types of `pospair1`, `pospair2`, and `numberofpairs` were all defined as integer. `pospair1` and `pospair2` represent the opening and closing positions of the stemloop. `numberofpairs` represents the length of stem.

[15] and [16] suggested that the stability of a secondary structure and its distance from the coding sequence influence its ability to inhibit the translation of the coding sequence. Therefore, the predicate `stemloop/4` (see Table 3) was designed to capture both the distance (the opening and the closing positions in the right part of Fig. 1) of a predicted stem-loop structure to the coding sequence and the stability. Here, the stability was represented by the number of base pairs (the length of the stem); the longer the stem the more stable the secondary structure and the more energy is needed to unwind it. We do not use the predicted minimum free energy because of the way we view the predicted mRNA secondary structure. For example, we consider three stem-loop structures while there was only one predicted minimum free energy for the overall predicted structure shown in the left part of Fig. 1.

With the biological knowledge gained from literature, we defined several declarative rules that identify if a uORF intersects with any predicted secondary structure on the uORF's left (upstream) part (see an illustration in Fig. 1), on the uORF's right (downstream) part, or is inside any predicted secondary structure. To instruct CProgol to include mRNA secondary structure predictions in its hypothesis space, we defined additional mode declarations (Table 3). Some adjustments were made to the parameter settings used in [8] to allow CProgol to consider a larger hypotheses space. The parameter **c** (the maximum number of atoms in the body of the rules constructed) was increased from 6 to 10; **nodes** (the maximum number of nodes explored during clause searching) was increased from 7,000 to 50,000; and **h** (the maximum depth of resolutions allowed when proving) was increased from 30 (default value) to 100.

4 Results

To statistically evaluate the impact of incorporating mRNA secondary structure predictions as part of the background knowledge on the task of recognising yeast functional uORFs, we compared the relative advantage (RA) values [17, Appendix A] from 100 experiments with and without mRNA secondary structure predictions. RA was used as a performance measure in [8]. The characteristics of the data used here matched with the characteristics for which RA is claimed to be useful. The idea of using RA is to predict the cost reduction in finding functional uORFs using a recognition model compared to using random sampling. In this application domain, RA is defined as

$$RA = \frac{A}{B}$$

where

- A is the expected cost of finding one functional uORF by repeated independent random sampling from the set of possible uORFs and performing a lab analysis of each uORF;
- B is the expected cost of finding one functional uORF by repeated independent random sampling from the set of possible uORFs and analysing only those uORFs which are predicted by the learned model as functional uORFs.

In 87 experiments out of 100, the mean RA values from the experiments with mRNA secondary structure predictions are better than the mean RA values from the corresponding experiments without mRNA secondary structure predictions (see Fig. 2). The result from a *Wilcoxon Signed Rank test* shows that there was a statistically significant increase from the mean RA values from the experiments without mRNA secondary structure predictions to those from the corresponding experiments with mRNA secondary structure predictions (mean RA values: mean without=34.05, mean with=61.53, $p < 0.0005$).

The analysis made so far is based on the mean RA values from our experiments. However, RA is less well known than other performance measures such as precision, recall (also known as sensitivity), specificity, and F_1 score. Therefore, to support our analysis, we also measured the precision, recall, specificity[6], and F_1 score. We found that there were statistically significant increases in the values of precision, recall, specificity, and F_1 score from the experiments

[6] In this case, specificity measures the fraction of randoms which are predicted as randoms.

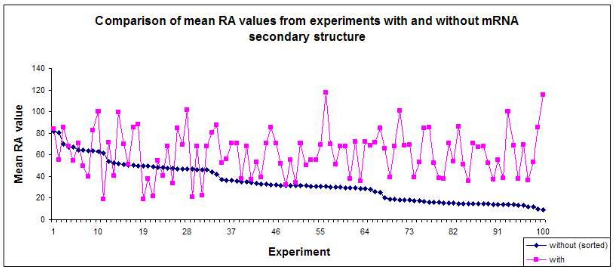

Fig. 2. Comparison of mean RA values from 100 experiments with and without mRNA secondary structure predictions. Experiments are sorted with respect to mean RA values from the experiments without mRNA secondary structure predictions. In 87 experiments, the mean RA values from experiments with mRNA secondary structure predictions are better than those from experiments without mRNA secondary structure predictions.

Table 4. Spearman's rank correlation between mean RA and other performance measures from 100 experiments with and without mRNA secondary structure predictions.

Experiment		Precision	Recall	Specificity	F_1 score
with	Mean RA	0.94	-0.02	0.73	0.74
without	Mean RA	0.91	-0.05	0.72	0.70

Note: There is no significant correlation between mean RA and recall. All other correlations are significant with $p < 0.0005$.

without mRNA secondary structure predictions to those from the corresponding experiments with mRNA secondary structure predictions (precision: mean without=0.45, mean with=0.63; recall: mean without=0.77, mean with=0.87; specificity: mean without=0.94, mean with=0.96; F_1 score: mean without=0.54, mean with=0.70; all were based on Wilcoxon Signed Ranks test with $p < 0.0005$).

Spearman's rank correlation was used to find out whether there are relationships between RA and the other measures (Table 4). We conclude that mean RA has a strong positive correlation with precision and specificity. Spearman's correlation also shows that there was a strong positive correlation between mean RA and F_1 score. This is due to the strong positive correlation between mean RA and precision, since there was no significant correlation between mean RA and recall; precision and recall are the two components used for calculating F_1 score.

The content of the hypotheses were also analysed. The hypotheses from the 10 experiments that give the 10 highest average cross-validation performances (mean RA) suggest that mRNA secondary structure influences uORFs' ability to regulate gene expression in the yeast *S. cerevisiae.* The rules also suggest that a functional uORF is likely to lie inside a stem-loop structure, or to intersect with a stem-loop structure on the uORF's left part. In our data, 17 of the 20 functional uORFs (positive examples) lie inside stem-loop structures predicted in the associated UTRs. For 3 of the 20 uORFs, their left part intersect with stem-loop structures predicted in the associated UTRs; 2 of these 3 uORFs do not lie inside stem-loop structures predicted in the associated UTRs.

5 Discussion and Future Work

Our empirical results show that the performance of an ILP system, CProgol 4.4, in recognising known functional uORFs in the yeast *S. cerevisiae* significantly increases when mRNA secondary structure predictions are added to the background knowledge (mean RA values: mean without=34.05, mean with=61.53, $p < 0.0005$). This conclusion still holds when performance is measured using precision, recall, specificity, and F_1 score, which are very well known in both machine learning and bioinformatics domains.

In this work, the background knowledge regarding mRNA secondary structure was derived from predictions made by RNAfold on the given *S. cerevisiae* sequences. However, the reliability of predictions made by RNAfold, and other similar software based on thermodynamic energy minimisation, is often questioned because each prediction is made based on a single sequence. Therefore, for future work, one could consider deriving the background knowledge from mRNA secondary structures that are predicted to be conserved among yeast species.

Here, we view the predicted mRNA secondary structure from the highest level, and do not consider a nested stem-loop as an independent stem-loop. Thus, we limited the type of background knowledge that was derived from the predicted mRNA secondary structure. It would be interesting to investigate the effect of including more detailed background knowledge of the mRNA secondary structure predictions on the ILP system's performance in recognising functional uORFs.

Acknowledgments. We thank Alex Wilson and Dr John McCall for useful discussions on statistics and experimental design.

References

1. Vilela, C., McCarthy, J.E.G.: Regulation of fungal gene expression via short open reading frames in the mRNA 5′ untranslated region. Molecular Microbiology **49**(4) (2003) 859–867
2. Vilela, C., Ramirez, C.V., Linz, B., Rodrigues-Pousada, C., McCarthy, J.E.G.: Post-termination ribosome interactions with the 5′ UTR modulate yeast mRNA stability. The EMBO Journal **18**(11) (1999) 3139–3152
3. Hinnebusch, A.G.: Translational Regulation of Yeast *GCN4*. A Window on Factors that Control Initiator-tRNA Binding to the Ribosome. J. Biol. Chem. **272**(35) (1997) 21661–21664
4. Fiaschi, T., Marzocchini, R., Raugei, G., Veggi, D., Chiarugi, P., Ramponi, G.: The 5′-untranslated region of the human muscle acylphosphatase mRNA has an inhibitory effect on protein expression. FEBS Letters **417**(1) (1997) 130–134

5. Iacono, M., Mignone, F., Pesole, G.: uAUG and uORFs in human and rodent $5'$untranslated mRNAs. Gene **349** (2005) 97–105
6. Morris, D.R., Geballe, A.P.: Upstream Open Reading Frames as Regulators of mRNA Translation. Molecular and Cellular Biology **20**(23) (2000) 8635–8642
7. Krummeck, G., Gottenöf, T., Rödel, G.: AUG codons in the RNA leader sequences of the yeast *PET* genes *CBS1* and *SCO1* have no influence on translation efficiency. Current Genetics **20** (1991) 465–469
8. Selpi, Bryant, C.H., Kemp, G.J.L., Cvijovic, M.: A First Step towards Learning which uORFs Regulate Gene Expression. Journal of Integrative Bioinformatics **3**(2) (2006) 31
9. Wang, L., Wessler, S.R.: Role of mRNA Secondary Structure in Translational Repression of the Maize Transcriptional Activator *Lc*. Plant Physiology **125**(3) (2001) 1380–1387
10. Kwon, H.S., Lee, D.K., Lee, J.J., Edenberg, H.J., ho Ahn, Y., Hur, M.W.: Post-transcriptional Regulation of Human *ADH5/FDH* and *Myf6* Gene Expression by Upstream AUG Codons. Archives of Biochemistry and Biophysics **386**(2) (2001) 163–171
11. Muggleton, S.: Learning from Positive Data. In Muggleton, S., ed.: Inductive Logic Programming Workshop. Volume 1314 of Lecture Notes in Computer Science., Springer (1996) 358–376
12. Muggleton, S.: Inverse Entailment and Progol. New Generation Computing **13**(3&4) (1995) 245–286
13. Muggleton, S., Firth, J.: CProgol4.4: a tutorial introduction. In Džeroski, S., Lavrač, N., eds.: Relational Data Mining. Springer-Verlag (2001) 160–188
14. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P.: Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package). Monatsh. Chem. **125**(2) (1994) 167–188
15. Baim, S.B., Sherman, F.: mRNA Structures Influencing Translation in the Yeast *Saccharomyces cerevisiae*. Molecular and Cellular Biology **8**(4) (1988) 1591–1601
16. Laso, M.R.V., Zhu, D., Sagliocco, F., Brown, A.J.P., Tuite, M.F., McCarthy, J.E.G.: Inhibition of Translation Initiation in the Yeast *Saccharomyces Cerevisiae* as a Function of the Stability and Position of Hairpin Structures in the mRNA Leader. The Journal of Biological Chemistry **268**(9) (1993) 6453–6462
17. Muggleton, S.H., Bryant, C.H., Srinivasan, A., Whittaker, A., Topp, S., Rawlings, C.: Are Grammatical Representations Useful for Learning from Biological Sequence Data?-A Case Study. Journal of Computational Biology **8**(5) (2001) 493–521