# Discovering Relational Association Rules for the Characterization of UTR cis-regulatory modules

Logisci C[(1)], Salvemini E[(1)], Turi A[(2)], Grillo G[(2)], Malerba D [(1)],D'Elia D[(2)]

[(1)] Department of Computer Science, University of Bari,
Via Orabona 4, 70125 Bari, Italy
[(2)] Institute for Biomedical Technologies, CNR,
Via Amendola 122/D, 70126 Bari, Italy

## Motivation

Understanding the nature and distribution of regulatory motifs in mRNA untranslated regions (UTRs) is an important step towards the comprehension of post-transcriptional mechanisms regulating gene expression. We have been working on this important task, experimenting the use of a two-stepped data mining approach to discover Frequent Sequential Pattern (FSP) of UTR regulatory motifs. Characterization is based on the discovery of frequent patterns of spaced motifs. The results of this approach are currently collected in the UTRminer web site (http://beagle10.ba.itb.cnr.it/UTRminer/) which provides FSPs detected in UTR sequences of mRNA targeting mitochondria in metazoan species. Here we present an extension of our previous work, in which the algorithm GSP (Generalized Sequential Pattern) was considered, making a comparative analysis with another mining software that could better fit the biological issue under investigation. The extension aims to:
1) find association rules, which convey additional information on inferential confidence;
2) possibly increase the number of motifs involved in discovered patterns;
3) reduce possible effects of pre-discretization of spacers on pattern support;
4) prevent the generation of uninteresting frequent patterns and association rules.

## Methods

The generation of the FSPs proceeds in two steps. In the first step co-occurring sets of motifs are found without taking into account their spatial displacement. The search is based on the levelwise method by Mannila and Toivonen (1997), that is, a breadth-first search is performed in the lattice of sets of motifs. At the end of this first mining step, sets of motifs which frequently co-occur in UTRs are returned. The length of spacers between motifs, which is expressed as the number of nucleotides between the starting position of two strictly subsequent motifs, is discretized. A spaced motif is represented as a sequence where each "Mi" denotes a motif while each "Si" corresponds to an interval of spacers returned by the discretization procedure. A deductive database is generated such that sequences of spaced motifs are stored in the extensional part of the deductive database, while rules for the generalization over intervals are stored in the intensional part. The algorithm SPADA is used to mine association rules in the generated deductive database. Association rules take the form: P => Q (s%, c%), where the antecedent P is a conjunction of atomic formulas (or atoms) which define the conditions, while the consequent Q is a conjunction of atoms which define some entailed properties. The parameter s% represents the support of the conjunction PQ , while c% represents the confidence of the implication P => Q. When the confidence is high, e.g., greater than 95%, an association rule can be used to infer Q when P is true. In this work, P Q represents a sequence of annotated motifs augmented with information on distance range between motifs. Association rules with high support suggest sequences of motifs strongly conserved and might have some functional property, while association rules with high confidence can be used to predict part of a sequence (Q) given another part (P). The sets of association rules discovered by SPADA are grouped by pattern length and can be ordered by support, confidence, antecedent length and consequent length, in order to support their analysis by biologists.

## Results

Experimental results prove the effectiveness of the approach. Mined association rules have generally high confidence, which means that it is possible to use association rules for prediction purposes as well. By generalizing over spacing intervals between two subsequent motifs, SPADA generates non-empty sets of patterns for minimum support thresholds higher than that used in experiments with GSP. Increased support allows for discovering more complex patterns, i.e. with a larger number of motifs.

**Contact :** domenica.delia@ba.itb.cnr.it