Mining spatial association rules of multiple co-occurring motifs to discover cis-regulatory modules

Ceci M[1], Loglisci C[1], Salvemini E[1], Grillo G[2], D'Elia D[2], Malerba D[1]

Motivation

Biological activities are typically co-regulated by several factors and this feature is properly reflected by higher-order structures called cis-regulatory modules (CRM) and represented by non-random clusters of regulatory motifs. Several methods have been proposed for the de novo discovery of modules. We propose an alternative approach based on the discovery of rules which define strong spatial associations between single motifs and suggest the structure of a module. Rules are expressed in a first-order logic formalism and are mined by means of an inductive logic programming (ILP) system. We also propose computational solutions to two issues: the hard discretization of numerical inter-motif distances and the choice of a minimum support threshold. All methods have been implemented and integrated in a prototypal tool designed to support biologists in the discovery and characterization of *cis-regulatory modules*.

Methods

The method we propose exploits an association rule mining method[1] which takes advantage of the logic representation and inference techniques. Association rules are a class of patterns which describe regularities or co-occurrence relationships in a set T of homogeneous data structures (e.g. sets, sequences, and so on). Formally, an association rule R is expressed in the form of A => C, where A (the antecedent) and C (the consequent) are disjoint conditions on properties of data structures (e.g. the presence of an item in a set). The meaning of an association rule is quite intuitive: if a data structure satisfies A, then it is likely to satisfy C. To quantify this likelihood, two statistical parameters are usually computed, namely support and confidence. The former, denoted as sup(R,T), estimates the probability P(A/\C) by means of the percentage of data structures in T satisfying both A and C. The latter, denoted as conf(R,T), estimates the probability P(C|A) by means of the percentage of data structures which satisfy condition C, out of those which satisfy condition A. The task of association rule mining consists in discovering all rules whose support and confidence values exceed two respective minimum thresholds. Frequency of sets of co-occurring motifs in a set of functional related sequences and inter-motifs distance are used as indication for the prediction of CRM.

[1] Dipartimento di Informatica, Università degli Studi di Bari, Bari [2] Istituto di Tecnologie Biomediche (ITB), CNR, Bari

Results

This work represents an extension of a previous work, carried out on regulatory motifs located in the UTRs of mRNAs [2], exploiting SPADA [3] in place of a traditional SM algorithm, and it is realized through an integrated system interacting with the user though a user-friendly GUI. This extension aims to: find association rules which convey additional information with respect to frequent sequential patterns; discover more significant inter-motif distances by means of a new discretization algorithm which does not require a-priori specifications; automatically select the best minimum support threshold; filter out redundant rules and return those interesting for the end-user. By using the GUI the user can set of the CRM mining process with annotations of regulatory motifs inside a set of sequences. Annotated sequences, which support specific frequent sets of motifs, are abstracted into sequences of spaced motifs, which are defined as ordered collections of motifs interleaved with gaps, i.e. inter-motif distances. Spatial association rules are mined from these abstractions. In order to deal with numerical information on the inter-motif distance, a density-based unsupervised discretization algorithm is applied. The algorithm takes into account the distribution of the distances in the set of sample sequences and does not significantly depend on input parameters as in the case of classical equal width or equal frequency discretization algorithms. The GUI hides the system complexity from the biologist by allowing to set parameters (minimum support and confidence) and by showing spatial association rules in a tabular form. The output generator module transforms the mined association rules in a tabular format in order to make them human readable. SPADA discovers many similar spatial association rules which basically differ in the size of the interval for some inter-motif distances. A very huge number of association rules makes interpretation of results cumbersome for the biologist. For this reason, the system includes a filtering module which identified rules of interest according to three different criteria. A case study is reported in order to show the potentialities of the tool.

Contact e-mail

domenica.delia@ba.itb.cnr.it

Supplementary information

References

[1] S.H. Nienhuys-Cheng and R. De Wolf. Foundations of Inductive Logic Programming, volume 1228 of LNAI. Springer-V., 1997 [2] A. Turi, C. Loglisci, E. Salvemini, G. Grillo, D. Malerba, and D. D'Elia. Computational annotation of UTR cis-regulatory modules through frequent pattern mining. BMC Bioinformatics, 10(S-6), 2009 [3] F.A. Lisi and D. Malerba. Inducing multi-level association rules from multiple relations. Machine Learning, 55 (2) :175–210, 2004