Cis-cop: Multiobjective identification of cis-regulatory modules based on constraints

Rocío Romero-Zaliz[1] **M. Martinez Ballesteros**[1] **Igor Zwir**[1,2] **Coral del Val**[1]

[1] DECSAI, UGR, Granada, Spain, {rocio,delval,igor}@decsai.ugr.es
[2] Howard Hughes Medical Institute, St. Louis, Missouri, USA

Abstract

Gene expression regulation is an intricate, dynamic phenomenon essential for all biological functions. The necessary instructions for gen expression are encoded in cis-regulatory elements that work together and interact with the RNA polymerase to confer specific spatial and temporal patterns of transcription. Therefore, the identification of these elements is currently an active area of research in computational analysis of regulatory sequences. However, the problem is difficult since the combinatorial interactions between the regulating factors can be very complex. Here we present a web server, Cis-cop, that identifies cis-regulatory modules given a set of transcription factor binding sites and, additionally, also RNA pol sites for a group of genes.

1 INTRODUCTION

Gene expression regulation is an intricate, dynamic phenomenon essential for all biological functions. The necessary instructions for gen expression are encoded in cis-regulatory elements, or modules (CRMs). A CRM consist of a set of transcriptional factors (TFs) that work together [6] and interact with the RNA polymerase (RNA pol) to confer specific spatial and temporal patterns of transcription. Therefore, the identification of CRMs is currently an active area of research in computational analysis of regulatory sequences [2, 8]. However the problem is difficult since the combinatorial interactions between the regulating factors can be very complex.

Here we present a web server, Cis-cop, that identifies CRMs given a set of TFs binding sites and, additionally, also RNA pol sites for a group of genes. The server takes into account spatial constraints on the arrangement of cis-elements such as relative position, distance and strand orientation. Cis-cop is also able to deal with more than one binding site for a certain TF in the same promoter region. The web server infers these modules using multi-objective and multimodal techniques [4]. Our algorithm is able to identify not only the most obvious solutions (i.e., the most frequent), but also those that are not locally dominated and that represent less frequent but relevant ones.

There is a large number of DNA motif finding algorithms and a lack of standards to measure their correctness. Most of the algorithms perform better in lower organisms, including yeast, as compared to higher organism [9]. For this reason, we leave to the user the selection of one or multiple DNA motif finding algorithms or databases available. These algorithms may use combinatorial enumeration, probabilistic modeling (Stormo, Gibbs, AlineACE, ANNSpec, MEME), mathematical programming, neural networks and/or genetic algorithms (EC, GAME), while databases can be TRANSFAC, REGULON DB, JASPAR, etc.

2 METHODOLOGY

The proposed methodology is depicted in Figure 1. The first step is the generation of Association Rules [1]. That is, discover the sets of features that appear frequently together in the set of cis-features in the input csv file. To do this, we use a modification of the Apriori algorithm [1] on Borgelt's eclat implementation [3] where candidate itemsets are counted in a pass and not generated on-the-fly. Borgelt's implementation is based on the idea to organize the counters for the itemsets in a special kind of prefix tree, which not only allows us to store them efficiently using little memory, but also supports processing the transactions as well as generating the rules. A further modification has been designed in order to deal with multi-objective prob-

lems. Apriori's method can be interpreted as a search problem with one objective function: the support of the itemset. In our case, we introduce a new objective to be maximized, complexity. Support is given by the number of genes that have the same cis-elements as the itemset, while complexity is calculated by the number of different cis-elements of the itemset. We extract itemsets from a database of transactions that maximize both support and complexity functions by post-processing the information generated by the Apriori algorithm (Figure 1).

As the number of features and genes (transactions) grows, the number of possible solutions increases exponentially. At a certain level of features and gene combination the eclat implementation is unable to produce results halting by a memory problem. For such cases we have developed a heuristic method. It uses an evolutionary approach to the problem of finding the best itemsets in a database of transactions. We propose a genetic algorithm (GA) based on the NSGA-II multi-objective approach.

The NSGA-II algorithm has been demonstrated as one of the most efficient algorithms for multi-objective optimization on a number of benchmark problems [5]. NSGA-II uses non-dominated sorting for fitness assignments. All individuals not dominated by any other individuals, are assigned front number 1. All individuals only dominated by individuals in front number 1 are assigned front number 2, and so on. Selection is made, using tournament between two individuals. The individual with the lowest front number is selected if the two individuals are from different fronts. The individual with the highest crowding distance is selected if they are from the same front. i.e., a higher fitness is assigned to individuals located on a sparsely populated part of the front. There are N parents and in every iteration N new individuals (offspring) are generated. Both parents and offspring compete with each other for inclusion in the next iteration.

The proposed GA instantiates the NSGA-II algorithm using a simple chromosome composed by a list of module structures (e.g., <(AP2A,0.3,D,0), (SP1,0.6,D,1), (SP1,0.9,D,110), (SP1,0.5,D,25), (SP1,0.7,D,32)>). A module structure consists of five features: name, score, orientation and distance (e.g., name: CAAT, score: 0.7, orientation: D, distance: 30: (CAAT,0.76,D,30)). Some of the features of the chromosome may be not used in some cases, for instance, when strand is ignored, then the orientation feature will be ignored. One-point crossover is used, that is, the chromosomes of the parents are cut at some randomly chosen common point (between module structures) and the resulting sub-chromosomes are swapped. Different mutation operators are used: add, erase and modify. An add mutation operator simply extends the chromosome by adding a new module structure at the beginning or the end of the chromosome. Erase mutation operator selects any module structure from the chromosome and deletes it. Finally, a modification mutation operator selects one of the features of the module structure and changes it. If the name is chosen, then a random name from the set of valid names is chosen and replaced in the module structure; if the orientation is chosen, then strand changes from D to R and vice versa. Score feature is not used in the evaluation process; therefore it is not used in any mutation operator. Finally, if distance is chosen, a small integer is added or subtracted from the one in the selected module structure from the chromosome. Again, strand and distance features are used only if the user selected them in the original parameters.

3 WEB SERVER

3.1 INPUT

The input form requires a file, provided by the user in gff format. Each entry of the csv input file contains the information about which TFBs are present in which promoters. The csv file contains the following fields: sequence name, source, feature, start, end, score, and strand and DNA pattern. As input parameters, the user can select which restrictions to apply to the search process: strand, same ordering of TFs in the module, same relative distance among the TFs, which can be crisp or fuzzy. There are two Cis-cop versions: exhaustive, for small input files, and heuristic for larger datasets. Once the parameters are selected, the query starts by clicking the 'submit' button.

3.2 OUTPUT

Results are provided as HTML for visual inspection and can also be received by e-mail. In case of error, a human-readable message is displayed. The output is a graphical panel containing a detailed description of all cis-features found in the CRM (Figure 2). The user can explore individually each module. All views can be saved as graph (e.g., png, pdf, jpg) or text (e.g., txt, csv, html).

4 DISCUSSION

The server is available at `http://gps-tools2.wustl.edu/modulos/modules.html` and it is open to all users and no login is required. There are two versions available, exhaustive and heuristic, both of them as required have an available tutorial along with example test files. The tu-

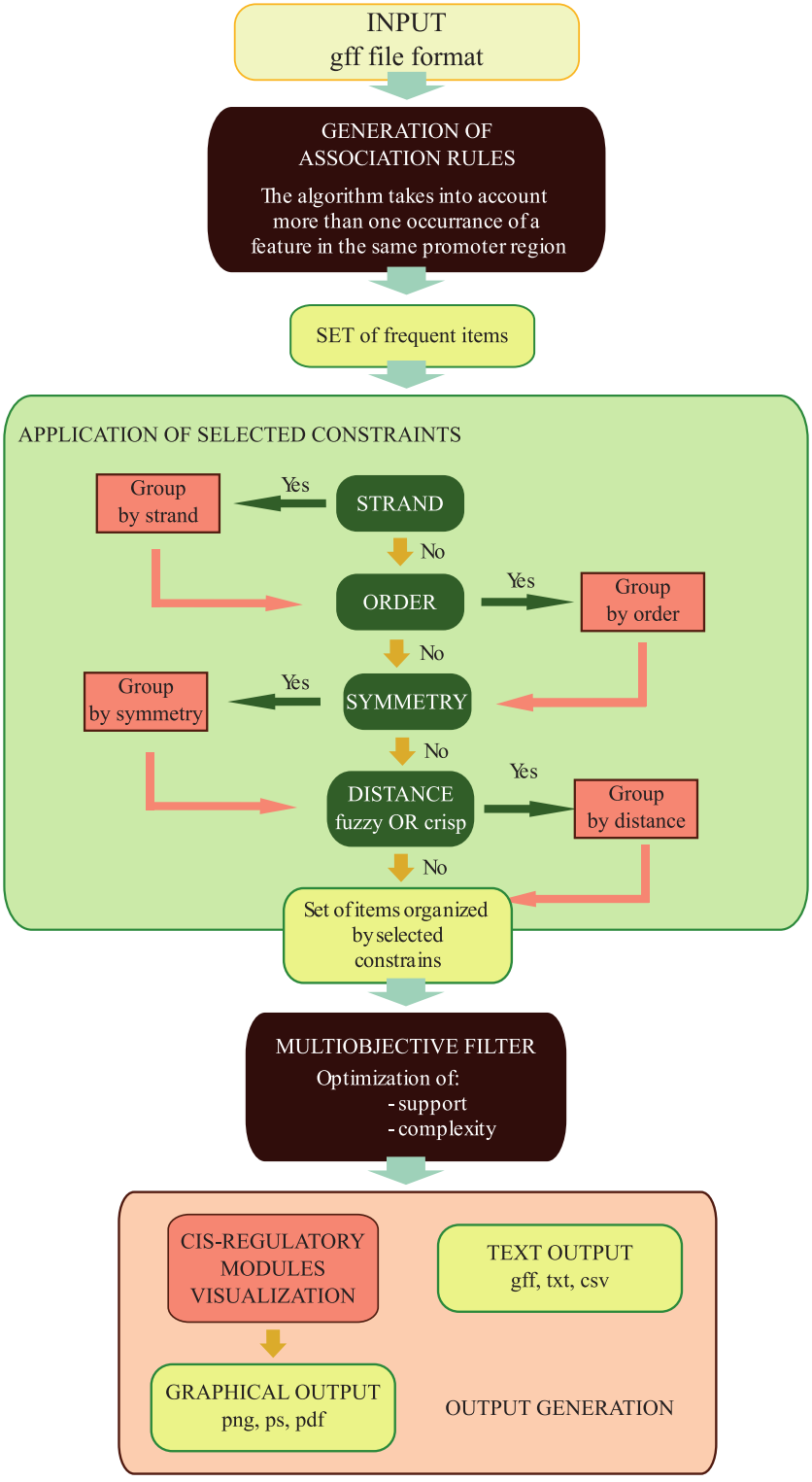

Figure 1: Pipeline

Figure 2: Modules image

torials explain which parameters can be modified and cover the following help topics: input, maximal number of features to be considered in a combinatorial module, order and distance, orientation, e-mail results, and output results. The development of the server presented here has been user-driven from the beginning. The here described approach has been already successfully used to identify PhoP regulated promoters harboring more than one binding site for the TF PhoP and sharing an atypical orientation and distance of the PhoP box/boxs to the RNA polymerase site [10] leading the results to the inference of PhoP transcription control over acid resistance genes in Salmonella typhimurium [10, 7]. The server has been active for the last year and the number of sequences used by de users varied very much from 10-500.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 1993.

[2] B.P. Berman, Nibu Y., Pfeiffer B.D., Tomancak P., Celniker S.E., Levine M., Rubin G.M., and Eisen M.B. Expliting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *drosophila* genome. *PNAS*, 99:757–762, 2002.

[3] Borgelt C. Efficient implementations of apriori and eclat. In *Workshop of Frequent Item Set Mining Implementations*, 2003.

[4] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Inc., 2001.

[5] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6:182–197, 2002.

[6] Davidson E.H. Genomic regulatory systems. *Nat. Genet.*, 29:153159, 2001.

[7] Zwir I., Harari O., and Groisman E.A. Gene promoter scan methodology for identifying and classifying coregulated promoters. *Methods Enzymol.*, 422:361–385, 2007.

[8] Markstein M. and Levine M. Decoding cis-regulatory dnas in the *drosophila* genome. *Curr Opin Genet Dev.*, 12(5):601–606, 2002.

[9] Das M.K. and Dai H.K. A survey of dna motif finding algorithms. *BMC Bioinformatics*, 8(Suppl. 7):S21, 2007.

[10] I. Zwir, H. Huang, and E.A. Groisman. Analysis of differentially-regulated genes within a regulatory network by gps genome navigation. *Bioinformatics*, 21:4073–4083, 2005.