DOI 10.18699/SBB-2022-675

Argo_CEL: GPU based composite elements discovery

Vishnevsky O.[1, 2]*, Bocharnikov A.[2], Ignatieva E.[1, 2], Kolchanov N.[1, 2]
[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[2] *Novosibirsk State University, Novosibirsk, Russia*
* *oleg@bionet.nsc.ru*

Key words: motif discovery, composite elements, transcription regulation, ChIP-Seq

Motivation and Aim: Composite elements play an important role in the regulation of transcription and provide a way for crosstalk between different regulatory pathways. Existing methods for the revealing of potential composite elements (PCE) are usually based on assessment of the significance of the mutual presence of the predicted transcription factor binding sites (TFBS) or ChIP-Seq pikes. Thus, such methods essentially depend on the completeness of training samples and experimental information about localization of TFBs. Methods for *de novo* discovery of significant oligonucleotide motifs are rather fast and require neither multiple alignments of query sequences nor the information about localization of transcription factor binding sites. However, existing heuristic approaches do not guarantee discovery of the most represented and significant motifs in huge experimental data sets. We developed new GPU based exhaustive approach ARGO_CEL for potential composite elements discovery in large DNA datasets.
Methods and Algorithms: Total scheme of ARGO_CEL approach is:
1. *De novo* discovery of the significant oligonucleotide IUPAC motifs specific for the set of regulatory sequences. The degenerate oligonucleotide motif, which was obtained as a result of this procedure, is considered significant if it meets the following criteria:
$$\begin{cases} F > f_0 \\ Q > q_0 \\ P(n,N) > p_0 \end{cases}$$
F is the proportion of query sequences containing the motif; Q denotes the enrichment of the motif expected by chance; P(n,N) is the binomial probability of observing by chance the motif in n out of N analyzed sequences; f_0, q_0, p_0 are the threshold levels.
2. Clustering of the revealed motifs. Clusters include the motifs significantly mutually appearing by chance in a set of random sequences. Estimation of correlation of all revealed motifs to each other on the set of random sequences was performed using phi-coefficient of Pearson for binary variables.
3. Revealing in the analyzed set of regulatory sequences significantly correlated clusters of motifs not correlated by chance.
4. Classification of the clusters with Tomtom. The assessment of similarity with the known TFBS was carried out in the JASPAR and HOMOCOMO databases. When annotating, a significant similarity with the weight matrices of several TF binding sites was normally observed for each oligonucleotide motif revealed. In this case, if JASPAR or HOMOCOMO annotations showed similarity with the weight matrix of the major TF, for which the experiment was carried out, then we considered that this oligonucleotide motif corresponds to the binding site of this TF. Otherwise, we identified PWMs that

had a significant similarity to the annotated motif in both JASPAR and HOMOCOMO databases.

Results: In order to carry out a comparative assessment of the efficiency of the developed motif discovery system, test sets of randomly generated sequences were created. In these sets, random IUPAC motifs of a given level of degeneracy were placed at random positions and in a given part of the sequences. Average information content (AIC) was used to measure the level of degeneracy of the motifs. Well-known motif discovery programs MEME-ChIP and CisFinder have been used to reveal embedded motifs. It was demonstrates that as soon as the abundance, F, of motifs falls below 30 %, the degeneracy of motifs increases, and AIC decreases below 1.5, the number of sequences N decreases and pr (probability, of observing by chance motifs embedded in the set) becomes higher than 0.05, the efficiency of all programs decreases abruptly. Nevertheless, as pr falls below 0.05, Argo_CEL still accurately identifies motifs, while all the other methods mentioned do not, especially at AIC = 1.25, – their outputs are significantly different from the original motifs embedded in the sets. It is most clearly observed at F < 30 %. Thus, it becomes necessary to use Argo_CEL to identify poorly abundant and highly degenerate context signals. In [1], we showed that ~70 % of weight matrices from TRANSFAC and Jaspar have AIC ≤ 1.25. Thus, we have proven that Argo_CEL can be more efficient than some other approaches in *de novo* discovery of about half of the TFBS, especially if they are present in less than 30 % of the sequences of the DNA set. We applied Argo_CEL system, to detect *de novo* conserved motifs in the sets of DNA sequences obtained during ChIP-Seq experiments with three transcription factors: FoxA2, Sp1, and GATA1 downloaded from the Cistrome DB [2]. For each of the ChIP-Seq experiments, 50 most significant motifs were annotated using Tomtom. In each of the considered ChIP-Seq experiments, for the annotations of the 50 most significant motifs obtained in this way, a search was made in scientific publications. For the vast majority of the considered motifs, experimental confirmation of the synergistic interaction of predicted TFs with target TFs was found.

Conclusion: We developed a new method ARGO_CEL for revealing of correlated binding sites in the ChIP-seq data, based on *de novo* motif discovery approach. This is an exhaustive method based on the use of high-performance GPU technologies. In comparison with existing methods of *de novo* motif discovery, it demonstrates higher efficiency, especially for strongly degenerate and underrepresented motifs. We have proven that Argo_CEL can be more efficient than some other approaches for *de novo* discovery of about half of the TFBS, especially if they are present in less than 30 % of the sequences of the DNA set. These correlated binding sites may correspond to sites – partners of transcription factors that take part in coordinated regulation of expression. The use of ARGO_CEL for the analysis of ChIP-Seq data made it possible both to identify potential binding sites for previously experimentally confirmed transcription factors – partners, and to propose motifs – targets for further experimental analysis.

Acknowledgements: The study is supported by the budget project of Institute of Cytology and Genetics, SB RAS (FWNR-2022-0020).

References

1. Vishnevsky O.V., Bocharnikov A.V., Kolchanov N.A. ARGO_CUDA: Exhaustive GPU based approach for motif discovery in large DNA datasets. *J Bioinformatics Computation Biol.* 2018;16(1):1740012.
2. Mei S., Qin Q., Wu Q. et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.* 2017;45(D1):D658-D662.