Module Inducer – a tool to automatically extract knowledge from biological sequences

Oksana Korol and Marcel Turcotte
School of Information Technology and Engineering, University of Ottawa
okoro103@uottawa.ca, turcotte@site.uottawa.ca

In the past decade fast advancement have been made in the sequencing, digitalization and collection of the biological data. However the bottleneck remains at the point of analysis and extraction of patterns from the data. We have developed an application that is aimed at widening this bottleneck by automating the knowledge extraction from the biological data. Our approach is aimed at discovering patterns in the set of sequences based on the location of transcription factor binding sites or any other biological markers with the emphasis of discovering relational dependencies. The core of our approach lies in an inductive logic programming engine (using Aleph[1]), which, based on the positive and negative examples as well as designed set of background knowledge rules, is able to induce a descriptive, human-readable theory, describing the data. An application provides an end-to-end analysis of the set of DNA sequences. A simple to use Web interface (induce.site.uottawa.ca) accepts a set of co-regulated sequences to be analyzed, set of negative example sequences to contrast the main set (optional), and a set of possible genetic markers as position weight matrices. A Java-based backend formats the sequences and determines the location of the genetic markers inside them by invoking Patser v3e.1[2]. The information is then passed on to the ILP engine, which induces the theory.

The application has been tested on synthetic as well as real biological data and has shown to produce consistent theory with good sensitivity (over 0.513) and high specificity (over 0.975). Currently the tool is being applied to further analyze ChIP-Seq data from a recent study investigating TAL1 binding in normal erythroid and leukaemic T cells [3]. Up-to-date results will be presented at the conference. We believe that our application will be a valuable aid at analyzing the results of high-throughput biological experiments as well as processing currently available data.

References

[1] Srinivasan A. The Aleph Manual. 2007. http://web.comlab.ox.ac.uk/oucl/research/areas/-machlearn/Aleph/.

[2] Hertz G. Z., Stormo G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*. **15**:563–577, 1999.

[3] Palii C. G., Perez-Iratxeta C., Yao Z., Cao Y., Dai F., Davison J., Atkins H., Allan D., Dilworth F. J., Gentleman R., Tapscott S. J., Brand M. Differential genomic targeting of the transcription factor TAL1 in alternate haematopoietic lineages. *The EMBO Journal*. **30**:494–509; doi:10.1038/emboj.2010.342, 2011.