Statistical approaches for the analysis of RNA-Seq and ChIP-seq data and their integration

Claudia Angelini and Italia De Feis

Istituto per le Applicazioni del Calcolo "Mauro Picone", Naples, Italy

http://www.iac.cnr.it/

The recent introduction of Next-Generation Sequencing (NGS) platforms, able to simultaneously sequence hundreds of millions of DNA fragments, has dramatically changed the landscape of genetics and genomic studies. However, to benefit of this novel sequencing technology, advanced laboratory and molecular biology expertise must be combined with a strong multidisciplinary background in data analysis. In addition, since the output of an experiment consists of a huge amount of data, terabytes of storage and clusters of computers are required to manage the computational bottleneck.

Recently, the Institute of Genetics and Biophysics (IGB) and the Istituto per le Applicazioni del Calcolo (IAC) have started a close collaboration aimed to set up a novel NGS facility in Naples that integrates both the wet laboratory and the bioinformatics core. Therefore, the IGB acquired a SOLiD system (now version 4) and, nowadays it provides all the wet laboratory capabilities and its experience in molecular biology for a wide range of experiments. Our team at IAC provides the experience in the usage and the development of computational methods for their analysis and it is also equipped with a powerful cluster of workstations (http://lilligridbio.na.iac.cnr.it/wordpress/) capable of handling massive computational tasks.

The research activities are directed toward two directions: from one side the effort of our group is devoted to the use of efficient software, the maintenance and development of bioinformatics pipeline for specific applications required by the sequencing facility, on the other hand the scientific interest is also devoted to the development of innovative statistical techniques for the NGS data analysis and to the implementation of novel algorithms using both CPU and GPU systems.

Till now our group has been involved the analysis of a series of independent studies on both RNA-seq and ChIP-seq. The experiments were conducted on the local sequencing facility by dr. Ciccodicola (for the RNA-seq data) and dr. Matarazzo (for the ChIP-seq data) groups at IGB-CNR, which are also members of the SEQAHEAD Cost Action. In this context our ongoing activities are devoted to the implementation of specific pipeline on our local cluster and to the definition of a probabilistic approach to model in terms of "signal plus noise" both transcriptional profiles and chromatin profiles. However, since we believe that integrating ChIP-seq and RNA-seq data is expected to provide much more biological insights for a better understanding of the mechanisms involved in gene expression regulation, rather than using one dataset only, we will focus our attention on the integration of these types of data in a unified statistical framework.

In the light of these considerations our group is aimed to contribute to the goals of the SEQAHEAD project by actively participating to the discussion concerning the development of novel statistical and computational methods for the analysis of RNA-Seq and ChIP-seq data and their integration, and to the development of educational programs on the statistical analysis of NGS data.

References

1. V. Costa, C. Angelini, et al., *Massive-scale RNA-Seq analysis of non ribosomal transcriptome in human trisomy 21*, PLoS ONE 2011.
2. V. Costa, C. Angelini, I. De Feis, A. Ciccodicola. *Uncovering the complexity of transcriptomes with RNA-Seq*. Journal of Biomedicine and Biotechnology vol. 2010, Article ID 853916, 19 pages, (2010).
3. C. Angelini, A. Ciccodicola, V. Costa and I. De Feis. *Analyzing the Whole Transcriptome by RNA-Seq data: the Tip of the Iceberg*, ERCIM NEWS July 2010, Special Theme Computational Biology, pp.16-17. 2010.