REEF and LAP: a computational framework for the identification of chromosomal regions associated to functional features enrichment and differential expression.

ID - 110

Coppe Alessandro[1], Basso Dario[2], Ferrari Francesco[3], Danieli Gian Antonio[1], Bicciato Silvio[2], Bortoluzzi Stefania[1]

[1]Department of Biology, University of Padova, Padova

[2]Department of Chemical Process Engineering, University of Padova, Padova

[3]Department of Biomedical Sciences, University of Modena and Reggio Emilia, Modena

Motivation

Systems biology elevates the study from the single entity level (e.g., genes, proteins) to higher hierarchies, such as entire genomic regions, groups of co-expressed genes, functional modules, and networks of interactions. Since the scientific attention focuses more on critical levels of biological organization and their emerging properties rather than on the single components of the system, the availability of high-throughput gene expression data, coupled with bioinformatics tools for their analysis, represents a scientific breakthrough in the quest for understanding biological mechanisms. The massive accumulation of high-quality structural and functional annotations of genomes imposes the development of computational frameworks able, not only to analyze gene expression profiles per se, but to merge any genomic information. The integration of different types of genomic data (gene sequences, transcriptional levels, functional characteristics) is a fundamental step in the identification of networks of molecular interactions, which will allow turning genomic research into biological hypotheses. We developed two computational procedures integrating functional annotations and genome structural information with transcriptional data: REEF (REgionally Enriched Features in genomes; Coppe et al., 2006) aims at detecting density variations of specific features along the genome sequence while LAP (Locally Adaptive Statistical Procedure; Callegaro et al., 2006) is a methodology for the identification of differentially expressed chromosomal regions.

Methods

REEF is a procedure to detect density variations of specific features, such as a class or group of genes homogeneous for expression and/or functional characteristics, along the genome sequence. For example it can be used to identify genomic regions with significant enrichments of genes which are co-expressed, differentially expressed, or related to particular molecular functions. The algorithm adopts a sliding window approach with the hypergeometric distribution to calculate the statistical significance of local enrichments. False Discovery Rate circumvents the problem of multiple testing when calculating the genome-wide statistical significance. Results of analyses are graphically presented at genome, chromosome and cluster level. A graphical tree structure enables the user to select and view a chromosome or a specific enriched region. REEF exploits UCSC Genome Browser Custom Annotation Tracks facility in order to visualize results as custom tracks together with standard tracks from UCSC Genome Browser. LAP is a method for the identification of differentially expressed chromosomal regions, which incorporates transcriptional data and structural information locally smoothing the expression statistic, along the chromosomal coordinate. The smoothing procedure is approached as a non-parametric regression problem using various methods (local variable bandwidth kernel estimator, spline functions or wavelets). A permutation scheme is used to identify differentially expressed regions, under the assumption that each gene has a unique neighborhood and that the corresponding smoothed statistic is not comparable with any statistic smoothed in other regions of the genome. Specifically, the statistic values are randomly assigned to chromosomal locations through B permutations and then, for each permutation, smoothed over the chromosomal coordinate. The significance of differentially expressed regions (p-value) is computed as the probability that the random null statistic exceeds the observed statistic over B permutations.

Results

REEF is a multiplatform program written in the Python, providing a graphical user interface allowing the interactive display of results. LAP is an R function performing statistical analyses, visualization of results on graphical representations of the genome, and export of the identified regions to genome browsers. The performances of the two algorithms have been accessed and compared first using a simulation approach. Synthetic data mimicking specific modifications or distortions of real gene expression signals have been used to evaluate specificity, sensitivity, and positive predictive values (ROC curves) of the two approaches. Then, REEF and LAP have been applied to the analysis of an integrated dataset of gene expression during myelopoietic differentiation. Results of this study allowed deepen the knowledge on the role of chromatin

remodeling and epigenetic control mechanisms on transcriptional regulation, shedding light on the impact of silencing/induction of specific genes on differential expression, in respect to the contribution of epigenetic mechanisms.

References - Callegaro A, Basso D, Bicciato S. A locally adaptive statistical procedure (LAP) to identify differentially expressed chromosomal regions. Bioinformatics. 2006; 22: 2658-66. - Coppe A, Danieli GA, Bortoluzzi S. REEF: searching REgionally Enriched Features in genomes. BMC Bioinformatics. 2006; 7:453.

Availability: http://telethon.bio.unipd.it/bioinfo/reef/

Email: ale@telethon.bio.unipd.it