An improved procedure for clustering and assembly of large transcriptome data

E. Picardi[1] ✉, V. Bevilacqua[2], F. Stoppa[2], G. Pesole[1]

[1]Istituto di Biomembrane e Bioenergetica del Consiglio Nazionale delle Ricerche, Bari, Italy
[2]Dipartimento di Elettrotecnica ed Elettronica, Politecnico di Bari, Bari, Italy

Motivations

Expressed sequence tags and full-length cDNAs represent an invaluable source of evidence for inferring reliable gene structures and discovering potential alternative splicing events [1]. However, to fully exploit their biological potential, correct and reliable EST clusters are required. To fill this gap we developed the program EasyCluster that resulted the most accurate when compared to software at the state of the art in this field [2]. Recent technological advances are dramatically increasing the number of available transcriptome reads. EST-like sequences can now be generated by pyrosequencing using Roche 454 platform which generates approximately one million reads per run. Handling such huge amount of EST-like data is basic to detect alternative splicing events, improve gene annotations or simply create gene-oriented clusters for expression studies. Sometimes EST-like data provide a fragmented overview of their genomic loci of origin and, thus, transcript assembly may be an optimal solution to annotate user-produced sequences. For these reasons we propose here a new implementation of EasyCluster able to manage genome scale transcriptome data and generate reliable gene-oriented clusters from 454 reads. The new version of EasyCluster software can facilitate downstream analyses because it enables the assembly of full-length transcripts per cluster, improves the clustering procedure using available annotations and embeds a graphical browser to provide an overview of results at genome level.

Methods

EasyCluster is based on the well-known EST-to-genome mapping program GMAP [3] since it can perform a very quick mapping of whatever expressed sequence onto a genomic sequence and can detect splicing sites according to a so defined "sandwich" dynamic programming that is organism independent. Providing EST-like data from Roche 454 sequencer, EasyCluster initially runs GMAP program and parses results in order to create an initial collection of pseudo-clusters by grouping EST-like reads according to the overlap of their genomic coordinates on the same strand. Then EasyCluster refines the EST grouping by including in each cluster only expressed sequences sharing at least one splice site. An ad hoc procedure is used to correct potential GMAP errors near splice sites and unspliced ESTs are added to each refined cluster. Finally, full-length transcripts are assembled for each cluster in order to valuate the alternative splicing extent and provide gene expression levels according to user supplied annotations.

Results

The new implementation of EasyCluster is written in Java programming language and provides improved clusters of EST sequences. It has been conceived to handle huge amount of EST-like reads produced by Roche 454 machines and supply a unique tool to cluster and assembly such transcriptome reads. Moreover, EasyCluster can now include unspliced reads and take benefit from available annotations. Alternative splicing is also inferred from each cluster after a refining procedure near exon-intron boundaries to reduce mapping errors due to GMAP. Accuracy and performances have been tested on simulated 454 reads by MetaSim software. Preliminary results indicate that the new EasyCluster implementation is highly efficient to manage and analyze deep transcriptome data from Roche 454 technology.

References

1. Nagaraj, S.H., Gasser, R.B. and Ranganathan, S. (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Briefings in bioinformatics, 8, 6-21.
2. Picardi, E., Mignone, F. and Pesole, G. (2009) EasyCluster: a fast and efficient gene-oriented clustering tool for large-scale transcriptome data. BMC bioinformatics, 10 Suppl 6, S10.
3. Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics (Oxford, England), 21, 1859-1875.