Bioinformatics developments for NGS data analysis at PRABI

Franck Picard, Guy Perrière

Pôle Rhône-Alpes de Bioinformatique, Bât. Gregor Mendel, Université Claude Bernard Lyon 1, Villeurbanne, France

http://www.prabi.fr/

The recent developments performed at PRABI for NGS data analysis are led in three main directions: i) short reads clustering for metagenomic data; ii) Open Reading Frames (ORFs) detection in metagenomes; and iii) statistical detection of peaks applied to the identification of replication origins on the human genome and to chIP-Seq data.

One of the problems frequently encountered with present day metagenomic data is the large amount of reads that have no significant homologs in the repository sequence data banks. In order to see if, at least, those "orphans" share some similarities among themselves, a lot of different clustering strategies have been developed. The strategy we have chosen to explore at PRABI is a distance-based one, as opposed to the model-based ones. More precisely, we have focused on the use of Correspondence Analysis (CA) and derived methods [1]. Due to its simplicity, this method is easy to use, very fast and efficient with large data sets containing hundreds of thousands of reads. On the other hand, its efficiency rapidly decreases when the number of different taxa present in the samples is high.

The approach chosen for ORFs detection is also based on CA. In this case, the analysis is computed on the codon composition of the six possible reading frames of a sequence [2]. The main advantage of this method is that it does not require a training step (like in Glimmer), therefore it can be used on metagenomic data, even if the biodiversity expected in the samples is very high. Tests on simulated metagenomic data sets show that the sensitivity of the program is 59% while specificity is 89%. The low sensitivity is due to fact that the efficiency of the method is highly dependant on the intensity of the codon bias in the coding sequence. Therefore, weakly biased genes (such as lowly expressed genes when there is translational selection in the species considered) are often missed by the method.

Lastly, for the detection of peaks in NGS data, the novelty is to develop a rigorous statistical framework to detect exceptional enrichment of reads using Poisson processes and scan statistics. It is a powerful framework that allows to define a proper P-value and FDR for the peaks, and our project is now to focus on the realistic modeling of the coverage function along the genome in order to adapt the significance of the peaks to a background noise that is highly dependent on the genomic context. As an extension and perspective, we plan to develop a statistical methodology to compare chIP-Seq data between conditions, and to assess the significance of differential peaks. This strategy will be applied also to the detection of differentially expressed small RNAs.

References

1. Perrière, G. and Thioulouse, J. (2002) Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res., **30**, 4548-4555.
2. Fichant, G. and Gautier, C. (1987) Statistical method for predicting protein coding regions in nucleic acid sequences. Comput. Appl. Biosci., **3**, 287-295.

Relevant Web sites

3. http://metasoil.univ-lyon1.fr/