Universidade de São Paulo

Tópicos em Bioinformática

Abdalla Almeida

Proposta de resenha: Análise Computacional para predição de genes em Metagenomas. A Metagenomica é uma abordagem que caracteriza um conjunto de genomas de uma determinada comunidade. A predição de genes de interesse requer um processo computacional utilizando softwares. A proposta é avaliar a eficácia e metodologia pra predizer genes utilizando a partir da metagenômica, já que esta é uma abordagem que trabalha com diversos genomas em um mesmo momento, seguido de análise computacional das sequência.

1-Katharina J Hoff. The effect of sequencing errors on metagenomic gene prediction. BMC Genomics 2009, 10:520

Abstract: Background: Gene prediction is an essential step in the annotation of metagenomic sequencing reads. Since most metagenomic reads cannot be assembled into long contigs, specialized statisticalgene prediction tools have been developed for short and anonymous DNA fragments, e.g. MetaGeneAnnotator and Orphelia. While conventional gene prediction methods have been subject to a benchmark study on real sequencing reads with typical errors, such a comparison has not been conducted for specialized tools, yet. Their gene prediction accuracy was mostly measured on error free DNA fragments. Results: In this study, Sanger and pyrosequencing reads were simulated on the basis of models that take all types of sequencing errors into account. All metagenomic gene prediction tools showed decreasing accuracy with increasing sequencing error rates. Performance results on an established metagenomic benchmark dataset are also reported. In addition, we demonstrate that ESTScan, a tool for sequencing error compensation in eukaryotic expressed sequence tags, outperforms some metagenomic gene prediction tools on reads with high error rates although it was not designed for the task at hand. Conclusion: This study fills an important gap in metagenomic gene prediction research. Specialized methods are evaluated and compared with respect to sequencing error robustness. Results indicate that the integration of error-compensating methods into metagenomic gene prediction tools would be beneficial to improve metagenome annotation quality.

2- Mina Rho, Haixu Tang and Yuzhen Ye. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research, 2010, 1–12.

Abstract: The advances of next-generation sequencing technology have facilitated metagenomics research that attempts to determine directly the whole collection of genetic material within an environmental sample (i.e. the metagenome). Identification of genes directly from short reads has become an important yet challenging problem in annotating metagenomes, since the assembly of metagenomes is often not available. Gene predictors developed for whole genomes (e.g. Glimmer) and recently developed for metagenomic sequences (e.g. MetaGene) show a significant decrease in performance as the sequencing error rates increase, or as reads get shorter. We have developed a novel gene prediction method FragGeneScan, which combines sequencing error models and codon usages in a hidden Markov model to improve the prediction of protein-coding region in short reads. The performance of FragGeneScan was comparable to Glimmer and MetaGene for complete genomes. But for short reads, FragGeneScan consistently outperformed MetaGene (accuracy improved 62% for reads of 400 bases with 1% sequencing errors, and 18% for short reads of 100 bases that are error free). When applied to metagenomes, FragGeneScan recovered substantially more genes than MetaGene predicted (>90% of the genes identified by homology search), and many novel genes with no homologs in current protein sequence database.

3- Fengfeng Zhou and Ying Xu. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. 2010. Bioinformatics Vol. 26 no. 16 2010, pages 2051–2052.

Abstract: Huge amount of metagenomic sequence data have been produced as a result of the rapidly increasing efforts worldwide in studying microbial communities as a whole. Most, if not all, sequenced metagenomes are complex mixtures of chromosomal and plasmid sequence fragments from multiple organisms, possibly from different kingdoms. Computational methods for prediction of genomic elements such as genes are significantly different for chromosomes and plasmids, hence raising the need for separation of chromosomal from plasmid sequences in a metagenome. We present a program for classification of a metagenome set into chromosomal and plasmid sequences, based on their distinguishing pentamer frequencies. On a large training set consisting of all the sequenced prokaryotic chromosomes and plasmids, the program achieves ∼92% in classification accuracy. On a large set of simulated metagenomes with sequence lengths ranging from 300 bp to 100 kbp, the program has classification accuracy from 64.45% to 88.75%. On a large independent test set, the program achieves 88.29% classification accuracy.