KTH Computer Science
and Communication

Data Analysis and Next Generation Sequencing : Applications in Microbiology.

NICOLAS INNOCENTI

Doctoral Thesis
Stockholm, Sweden 2015

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i fysik fredagen den 30 oktober 2015 klockan 14.00 i AlbaNova Universitetscentrum, Kungl Tekniska högskolan, Roslagstullsbacken 21, Stockholm, rum FA32.

TRITA-CSC-A 2015:15 • ISSN-1653-5723 • ISRN-KTH/CSC/A–15/15-SE • ISBN 978-91-7595-699-2

Abstract

Next Generation Sequencing (NGS) is a new technology that has revolutionized the way we study living organisms. Where previously only a few genes could be studied at a time through targeted direct probing, NGS offers the possibility to perform measurements for a whole genome at once. The drawback is that the amount of data generated in the process is large and extracting useful information from it requires new methods to process and analyze it.

The main contribution of this thesis is the development of a novel experimental method coined tagRNA-seq, combining 5'tagRACE, a previously developed technique, with RNA-sequencing technology. Briefly, tagRNA-seq makes it possible to identify the 5' ends of RNAs in bacteria and directly probe for their type, primary or processed, by ligating short RNA sequences, the *tags*, to the beginnings of RNA molecules. We used the method to directly probe for transcription start and processing sites in two bacterial species, *Escherichia coli* and *Enterococcus faecalis.* It was also used to study polyadenylation in *E. coli*, where the ability to identify processed RNA molecules proved to be useful to separate direct and indirect regulatory effects of this mechanism. We also demonstrate how data from tagRNA-seq experiments can be used to increase confidence on the discovery of anti-sense transcripts in bacteria. A detailed analysis of the data revealed subtle artifacts in the coverage signal towards 3'ends of genes, that we were able to explain and quantify based on Kolmogorov's broken stick model. We also discovered evidences for circularization of a few RNA transcripts, both in our own data sets and publicly available data.

Designing the tags used in tagRNA-seq led us to the problem of words absent from a text. We focus on a particular subset of these, the minimal absent words (MAWs), and develop a theory providing a complete description of their size distribution in random text. Genomes from viruses and living organisms have MAWs a large fraction of which are well modeled by the theory, but almost always exhibit a behavior different from random texts in the tail of the distribution. MAWs from this tail are closely related to sequences present in the genome that preferentially appear in regions with important regulatory functions.

Finally, and independently from tagRNA-seq, we propose a new approach to the problem of bacterial community reconstruction in metagenomic, based on techniques from compressed sensing. We provide a novel algorithm competing with state-of-the-art techniques in the field.

TRITA-CSC-A 2015:15 • ISSN-1653-5723 • ISRN-KTH/CSC/A–15/15-SE • ISBN 978-91-7595-699-2