Analysis of pan-genome content and its application in microbial identification - DTU Orbit (11/03/2019)

**Analysis of pan-genome content and its application in microbial identification**

With the rapid development of DNA sequencing technology it is today possible to sequence multiple genomes in a single day at a low cost with a single machine. This has resulted in several large-scale genomic projects, such as Ten Thousand Microbial Genomes (BGI) to explore microbial diversity in China, and understand its influence to the environment and humans; The Human Microbiome project (NIH) to find microorganisms in association with healthy and infected humans; and The 100K Genome Project (University of California, Davis, and FDA), which aims to sequence the genomes of 100,000 infectious microorganisms and eventually speed up the diagnosis of foodborne illnesses. This genomic data can give biologists many possibilities to improve knowledge of organismal evolution and complex genetic systems.

The general interest of this PhD thesis is how to obtain relevant information from growing amounts of genomic data and use this to answer important biological questions. More specifically, comparison of prokaryotic proteomes is used to determine possible sets of functions, essential to sustain microbial life; to extract and interpret similarities and variance in genomic content within different taxonomic groups or genomic structures; and to use the information of a specific proteome to predict which species it might belong to. Two different algorithms, BLAST and profile Hidden Markov Models (HMMs), are used to determine similarity between sequences and to address the questions in this thesis.

The first project, described in Chapter 3, is based on using protein Basic Local Alignment Search Tool (BLAST) comparisons for sequence-based homology searches. Paper I presents comparative genomics of Bifidobacterium, Lactobacillus and related probiotic genera.; and Paper II illustrates the use of in silico analyses for the characterization of two Listeria monocytogenes strains.

Chapter 4 describes the use of profile HMMs for comparative analysis using for sequence-based homology searches. Paper III introduces PanFunPro a new, profile HMM-based method for pan-genome analysis. Paper IV illustrates the application of PanFunPro to a set of more than 2000 genomes; this paper aims to define set of protein families, which are conserved among all the genomes. Papers V demonstrates comparative genomics analysis of proteomes, belonging to Vibrio genus.

In the last project, described in Chapter 5, both BLAST- and profile HMM-based methods are employed to infer taxonomy group-specific gene families, which are used for microbial identification. Paper VI illustrates the use of specific genes for microarray chip design; Paper VII demonstrates the use of the Salmonella enterica core-genome content for epidemiological typing; and Paper VIII represents the application of PanFunPro approach for in silico taxonomy prediction.

In summary, this thesis presents three projects that have contributed to identification and characterization of microbial organisms, and open new possibilities for comparative genomics and epidemiology.

**General information**

State: Published
Organisations: Department of Systems Biology
Contributors: Lukjancenko, O., Ussery, D., Larsen, M. V.
Number of pages: 177
Publication date: 2014

**Publication information**

Publisher: Technical University of Denmark (DTU)
Original language: English
Electronic versions:

PhDThesisOksanaLukjancenko..PDF
Research output: Research › Ph.D. thesis – Annual report year: 2014