UMEÅ UNIVERSITET

Hidden patterns that matter

Statistical methods for analysis of DNA and RNA data

Therese Kellgren

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Hörsal B, Lindellhallen,
fredagen den 16 oktober, kl. 09:00.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor, Rebecka Jörnsten,
Institutionen för Matematiska Vetenskaper, Chalmers tekniska högskola och Göteborgs universitet, Göteborg, Sverige.

Department of mathematics and mathematical statistics

Organization
Umeå University
Department of mathematics
and mathematical statistics

Document type
Doctoral thesis

Date of publication
25 September 2020

Author
Therese Kellgren

Title
Hidden patterns that matter
Statistical methods for analysis of DNA and RNA data

Abstract

Understanding how the genetic variations can affect characteristics and function of organisms can help researchers and medical doctors to detect genetic alterations that cause disease and reveal genes that causes antibiotic resistance. The opportunities and progress associated with such data come however with challenges related to statistical analysis. It is only by using properly designed and employed tools, that we can extract the information about hidden patterns. In this thesis we present three types of such analysis.

First, the genetic variant in the gene COL17A1 that causes corneal dystrophy with recurrent erosions is reveled. By studying Next-generation sequencing data, the order of the nucleotides in the DNA-sequence was be obtained, which enabled us to detect interesting variants in the genome. Further, we present results of an experimental design study with the aim to make the best selection from a family that is affected by an inherited disease.

In second part of the work, we analyzed a novel antibiotic resistance Staphylococcus epidermidis clone that is only found in northern Europe. By investigating its genetic data, we revealed similarities to a world known antibiotic resistance clone. As a result, the antibiotic resistance profile is established from the DNA sequences.

Finally, we also focus on the challenges related to the abundance of genetic data from different sources. The increasing number of public gene expression datasets gives us opportunity to increase our understanding by using information from multiple sources simultaneously. Naturally, this requires merging independent datasets together. However, when doing so, the technical and biological variation in the joined data increases. We present a pre-processing method to construct gene co-expression networks from a large diverse gene-expression dataset.

Keywords

Genome, Next-generation sequence, statistics, microarrays, bacteria, antibiotic resistance, inherited diseases, Co-expression networks, centralization within subgroups

Language
English

ISBN
print: 978-91-7855-240-5
PDF: 978-91-7855-241-2

ISSN
1653-0829

Number of pages
26 + 4 papers