Metagenomic sample clustering enhanced by SRA metadata and NLP

Rikke Møller Larsen

201607214@post.au.dk

Supervised by Associate Professor Thomas Bataillon
&
Co-supervised by Patrick Ettenhuber and Fernando Tavares at QIAGEN

Bioinformatics Research Centre,
Universitetsbyen 81
8000 Aarhus C

Aarhus University

June 1, 2022

Preface

The work within this Master's thesis was carried out under the supervision of Associate Professor Thomas Bataillon at the Bioinformatics Research Centre at Aarhus University. The project was in collaboration with QIAGEN Aarhus, with Patrick Ettenhuber and Fernando Tavares, (Senior Product Owner and Project Manager/Senior Software Developer, respectively) from QIAGEN as co-supervisors.

Thank you to my supervisor Thomas Bataillon for trusting me to work independently, while also offering helpful counsel and direction, whenever I needed it.

A heartfelt thank you to QIAGEN and its employees for their hospitality through these four months, in particular to Patrick Ettenhuber and Fernando Tavares, and the rest of the Microbial Team. Thank you for your guidance and advice, and for believing in me enough to invite me to stay for the coming year.

I also thank the professors, staff and my fellow Master's students at BiRC for making it two wonderful years despite an ongoing pandemic.

I would like to thank Rikke Myrhøj Jensen, Stine Katrine Rye Østergaard, Janne Auning Engestoft and Benjamin Nichum for their continued support, both moral and academic, and most important of all, their friendship.

Abstract

The data available from metagenomic sequencing experiments are exponentially increasing, in accordance with an increased focus on microbiomes and their effect on health and disease. Databases, such as NCBI's Sequence Read Archive (SRA) seek to make this data publicly and easily accessible with the use of standardized metadata. This metadata can contain important information for the description and distinction of experiments, but given its text-based form, it is not directly available for computer-based methods.

Utilizing biomedical natural language processing (NLP) to harness this data, has the potential to improve the ability of automatically classifying new experiments relative to using only analysis data, which ultimately can help facilitate novel research and discoveries.

While the metadata may contain valuable information, it is also prone to be faulty or incomplete, so there is a need for the development of unsupervised corrective methods, in order to obtain training data of high enough quality.

Abbreviations

CBOW - Continuous Bag-of-words

CRF - conditional random field

LCA - Lowest common ancestor

LSTM - Long short term memory

MeSH - Medical subject headings

MIMIC-III - Medical Information Mart for Intensive Care

NCBI - National Center for Biotechnology Information

NGS - Next-generation sequencing

NLP - Natural language processing

OOV - Out-of-vocabulary

OTU - Operational taxonomic unit

QMI-PTDB - QIAGEN Microbial Insights - Prokaryotic Taxonomy Database

SRA - Sequence read archive

S2V - sent2vec

UMAP - Uniform Manifold Approximation and Projection

WGS - Whole genome sequencing

W2V - word2vec

ΔWB - Delta within-between

Contents