Novel bioinformatics tools to assess microbial diversity in life support systems

Mohamed Mysara, Natalie Leys, Pieter Monsieurs

mahemd@sckcen.be; pmonsieu@sckcen.be
Bacterial communities

• Bacterial species live in communities, rather than individual species
• They interact, depend and talk to each other
• These dynamic communities referred to as microbiome
Microbiome studies

However, culture based approach exhibit various disadvantages:

- Time and labor consuming
- Most of the species can not be cultured in lab conditions
From sample to data (16S rRNA gene sequencing)

Extract DNA → Amplify one common gene “PCR” → Determine DNA sequence

Illumina MiSeq

Pseudomonas = atcccgtagtaccg
Cupriavidus = atcccgtagtaccg
Acidovorax = atcccgtagtaccg
Arthrosipra = atcccgtagtaccg
Gracilibacter = atcccgtagtaccg

Each sequence is mapped to a species

Thousands – millions of sequences
Microbiome & 16S rRNA metagenomics

17 bacterial species

Extract DNA

PCR of 16S rRNA gene

Amplify one

Hundred more sequences

atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg

atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg

atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg
atcccgtagtaccg

Thousands of sequences
16S rRNA Metagenomics Analysis Pipeline

Illumina (MiSeq)
454 Pyrosequencing

Ion Torrent (PGM)
PacBio (SMRT)
Nanopore (MinION)

Demultiplexing and primer clipping
Assembly (MiSeq)
Quality filtering
Denoising

IPED: Denoising algorithm for PE Illumina MiSeq data
Chapter 2 & 3: Training artificial intelligent model to handle this problem

I) Selection of datasets where the ground truth is known (Mock dataset)

<table>
<thead>
<tr>
<th>What does the machine report</th>
<th>A</th>
<th>G</th>
<th>-</th>
<th>A</th>
<th>G</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>What should the machine report</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>-</td>
<td>A</td>
<td>T</td>
</tr>
<tr>
<td>Status</td>
<td>T</td>
<td>T</td>
<td>D</td>
<td>T</td>
<td>T</td>
<td>S</td>
<td>T</td>
<td>I</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

II) Identification of features contributing to the sequencing errors
    e.g. Position in the read

For this purpose, we have developed two artificial intelligence tools:
A) NoDe for 454
B) IPED for MiSeq

IV) Selection of the best performing model
Denoising algorithms concept

ATCCC–TACTACCGA–CGCGTACTACC–G
ATCCC–TACTACCGA–CCCGTACTACC–G
ATCCC–TACTACCGA–CCCGTACTCC–G
ATCCC–TACTACCGA–CCCGTACTACC–G
ATCCC–TACTACCGA–CCCGTACTACC–G

The Classifier
(i) Extracting quality-features for each position (Perl)
(ii) Running a pre-trained classifier (WEKA using JAVA)
(iii) Marking the potentially erroneous positions (Perl)

ATCCC–TACTACCGA–CCCGTACTACC–G
ATCCC–TACTACCGA–CXCGTACTACC–G
ATCCC–TACTACCGA–CCCGTACTXCC–G
ATCCC–TACTACCGA–CCCGTACTACCXG

X = correctly marked erroneous positions

Modified Pre-cluster (mothur using C++)

ATCCC–TACTACCGA–CCCGTACTACC–G

Representative Read (Count = 110)
# Chapter 3: IPED comparative analysis

## Table: Error and CPU cost comparison

<table>
<thead>
<tr>
<th></th>
<th>Error</th>
<th>CPU Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNoise</td>
<td>0.18%</td>
<td>14 sec</td>
</tr>
<tr>
<td>Pre-cluster</td>
<td>0.18%</td>
<td>12 sec</td>
</tr>
<tr>
<td>IPED</td>
<td>0.10%</td>
<td>70 sec</td>
</tr>
</tbody>
</table>
16S rRNA Metagenomics Analysis Pipeline

**Illumina (MiSeq)**
454 Pyrosequencing

**Ion Torrent (PGM)**
**PacBio (SMRT)**
**Nanopore (MinION)**

Demultiplexing and primer clipping

Assembly (MiSeq)

Quality filtering

Denoising

Chimera Detection

**IPED: Denoising algorithm for PE Illumina MiSeq data**

**CATCh: Chimera Detection**
Chimeric problem

16S rRNA Amplicon Sequencing

PCR Amplification

Sequencing 454, MiSeq, SMRT, PMG, MinION

PCR cycle 1

PCR cycle 2

PCR cycle 3

Species A

Species B
This hybrid sequence is not real, it could be mistaken as a false NOVEL species. Chimeric rate in Next generation sequencing run can reach up to 45% of the reads.
CATCh Training/Running

Extract the scores and features of these tools

- Uchime
- ChimeraSlayer
- Pintail
- Decipher
- Perseus
Chapter 4: Chimera Detection Challenges

Chimeric Range
Length added by the smaller parent

Divergence
A measure of the differences between parents

Number of parents
Number of parent read forming the chimeras
CATCh Comparative Analysis

### How sensitive is the tool in detecting all chimera

#### Sensitivity

<table>
<thead>
<tr>
<th>Tool</th>
<th>UCHIME</th>
<th>ChimeraSlayer</th>
<th>Pintail</th>
<th>DECIPHER</th>
<th>CATCh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>78</td>
<td>67</td>
<td>29</td>
<td>57</td>
<td>85</td>
</tr>
</tbody>
</table>

### How specific is the tool not to falsely detect non-chimeric reads

#### Specificity

<table>
<thead>
<tr>
<th>Tool</th>
<th>UCHIME</th>
<th>ChimeraSlayer</th>
<th>Pintail</th>
<th>DECIPHER</th>
<th>CATCh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity</td>
<td>97</td>
<td>98</td>
<td>75</td>
<td>97</td>
<td>96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tool</th>
<th>CATCh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specificity</td>
<td>95</td>
</tr>
</tbody>
</table>

**Note:** The table compares the performance of different tools in terms of sensitivity and specificity for chimera detection.
16S rRNA Metagenomics Analysis Pipeline

NoDe & IPED: Denoising Algorithms for 454 Pyrosequencing and Illumina MiSeq
CATCh: Chimera Detection
DynamiC: Over-merging

Illumina (MiSeq) 454 Pyrosequencing
Ion Torrent (PGM)
PacBio (SMRT)
Nanopore (MinION)

Reads
Demultiplexing and primer clipping
Assembly (MiSeq)
Quality filtering
Denoising
Chimera Detection
OTU Clustering
16S rRNA Metagenomics Analysis Pipeline

16S rRNA amplicon sequencing
- Illumina (MiSeq)
- Roche (454 Pyrosequencing)
- Ion Torrent (PGM)
- PacBio (SMRT)
- Nanopore (MinION)

Pre-assembly demultiplexing and primer clipping
Assembly (MiSeq)
Quality filtering
Denoising
Chimera detection
OTU clustering

Optimized CATCh, mothur, IPED, UPAARSE and SPAdes
(OCToPUS)
Benchmark - accuracy

MOCK1 (V3-V4)
- USEARCH
- Mothur
- OCToPUS
- QIIME
- LotuS

MOCK1 (V4)
- USEARCH
- Mothur
- OCToPUS
- QIIME
- LotuS

MOCK2 (V4)
- USEARCH
- Mothur
- OCToPUS
- QIIME

MOCK2 (V4-V5)
- USEARCH
- Mothur
- OCToPUS
- QIIME
- LotuS

MOCK3 (V3-V4)
- USEARCH
- Mothur
- OCToPUS
- QIIME
- LotuS

Copyright © 2014
SCK•CEN
Benchmark – data retrieval
Benchmark – Computational cost
Wide range of applications

Introduction

Palliative radiotherapy is a commonly used treatment to treat specific types of cancer (e.g., colon cancer). After exposure to ionizing radiation, the intestine is always affected. The intestinal epithelium is very sensitive to the increased oxidative stress after exposure to ionizing radiation (Riley, 1994), resulting in a loss

Conclusion

In future experiments, first the dose of the irradiation has to be optimized for our mouse model in order to exhibit relevant clinical symptoms. Second, a more specific examination of apoptosis and examination of oxidative stress will be performed. And third, a larger subset of inflammation markers will also be used. In addition, the formulation and dose of Arthrospira sp. as food supplement will also be further optimized.
Summary

NoDe & IPED: Denoising Algorithms for 454 Pyrosequencing and Illumina MiSeq

CATCh: Chimera Detection

DynamIC: Over-merging

Pre-processing

Demultiplexing and primer clipping

Assembly (MiSeq)

Quality filtering

Denoising

Chimera Detection

OTU Clustering

Biodiversity analysis

Post-processing

Clipped Reads

Forward

Reverse

Low quality reads

Sequencing errors

Chimeric reads

OTU1

OTU2

OTU3

Reads

Illumina (MiSeq)

454 Pyrosequencing

Ion Torrent (PGM)

PacBio (SMRT)

Nanopore (MiniION)

Publications

CATCh: Mysara et al., AEM (2015)
NoDe: Mysara et al. BMC Bioinformatics (2015)
iPED: Mysara et al. BMC Bioinformatics (2016)
OCToPUS: Mysara et al. FEMS Ecology (2017)

Future work – Extend expertise

• New long-read sequencing technologies (NanoPore, PacBio)

• Complex mock communities (n > 200)

• Shotgun metagenomics

https://github.com/M-Mysara/OCToPUS
Future work

- New long-read sequencing technologies
  - NanoPore
  - PacBio

- Challenging 16S rRNA amplicon sequencing pipelines
  - Complex mock communities (n > 200 strains)
  - Strain / subspecies variation detection

- Shotgun metagenomics
  - Taxonomy versus metabolic potential
Acknowledgements

Dr. Mohamed Mysara
Ruben Props
Dr. Natalie Leys

Prof. Dr. Peter Vandamme
Prof. Dr. Yvan Saeys

Prof. Dr. Jeroen Raes

Prof. Dr. Daniel Charlier
Microarrays

- **Types**
  - Two-color arrays: *Rhodospirillum rubrum, Cupriavidus metallidurans*
  - Affymetrix: *Pseudomonas aeruginosa*
  - Nimblegen: *Arthrospira sp. PCC8005*
  - Agilent: *Cupriavidus metallidurans*

- **In-house facilities**
  - Two-color microarrays
  - Affymetrix arrays

- **Analysis**
  - MIC: BioConductor
  - RDB: Partek

Transcriptomics via RNA-seq

- **Application area**
  - MIC
    - *Arthrospira* sp. PCC8005
    - *Clostridium bytiricum*
    - *Pseudomonas aeruginosa*
  - BIS
    - To be started

- **Analysis**
  - BWA / Bowtie
  - EdgeR
**Genome assembly**

- **De novo genome assembly**
  - Based on 454 pyrosequencing data
  - Based on Illumina HiSeq data
  - Annotation ➔ GenoScope - MaGe
    - CmetScope
    - ArthroScope

- **Applications**
  - MIC: Metal resistant bacteria
  - BIS: *Lemna minor* (~ 400 Mbp genome)

---

<table>
<thead>
<tr>
<th>Platform</th>
<th>Insert Size</th>
<th>Read Length</th>
<th>No. Reads</th>
<th>No. Nucleotides</th>
<th>Genome Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina HiSeq 2000</td>
<td>200 bp</td>
<td>2*100 bp</td>
<td>207,985,822</td>
<td>40,730,561,447</td>
<td>100X</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Statistics</th>
<th>SOAPdenovo2(^2)</th>
<th>CLC Bio</th>
<th>CLC Bio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>preprocessed data</td>
<td>preprocessed data</td>
<td>processed reads + flash data</td>
</tr>
<tr>
<td>K-mer size</td>
<td>63</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>No. Scaffolds</td>
<td>108,607</td>
<td>116,254</td>
<td>117,403</td>
</tr>
<tr>
<td>Max scaffold length</td>
<td>1,101,160</td>
<td>1,227,158</td>
<td>1,299,833</td>
</tr>
<tr>
<td>genome length</td>
<td>401</td>
<td>388</td>
<td>410</td>
</tr>
<tr>
<td>N50</td>
<td>10,194</td>
<td>8,059</td>
<td>8,423</td>
</tr>
</tbody>
</table>

---

Janssen et al., 2010, Monsieurs et al., 2013, Monsieurs et al., 2014a, Monsieurs et al., 2014b

---

Arne Van Hoeck
Bacterial resequencing - workflow

Read in Data

Quality Control

BWA aligner

Quality Control

SNP detection & validation

SNP’s and small indels

Incorrect mapped reads

Large insertions

Large Deletions

Insert size plotting

BWA / Samtools / GATK genomic suite
Bacterial re-sequencing

**MUT2**
- ISRme3
- agrS

**MUT3**
- ISRme5
- IS1086

**Normal Insertion**
- agrS

Reference vs query with 300 nt, 1000 nt, and 300 nt differences indicated.
Facilities

- Bio-informatics server – calculations
  - 24 processors – 3.2 GHz
  - 96 Gb memory
  - 6.0 Tb hard drive (RAID5 configuration)

- Data server - storage
  - 1 Tb hard drive in RAID10 configuration
  - Automatic backup to EqualLogic tapes
  - Storage of all raw microarray and sequencing data

- FERMI cluster
  - Details to be added
To be added?: Slides Roel on Affymetrix and Partek
16S rRNA metagenomics algorithms

From reads to operational taxonomic units: an ensemble processing pipeline for MiSeq amplicon sequencing data

Mohamed Mysara; Mercy Njima; Natalie Leys; Jeroen Raes; Pieter Monsieurs

Gigascience gv017. DOI: https://doi.org/10.1093/gigascience/gjw017
Published: 18 January 2017

Abstract

Introduction: The development of high-throughput sequencing technologies has provided microbial ecologists with an efficient approach to assess bacterial diversity at an unseen depth, particularly with the recent advances in the Illumina MiSeq sequencing platform. However, analysing such high-throughput data is posing important computational challenges, requiring specialized bioinformatic solutions at different stages during the processing pipeline, such as assembly, quality control, alignment, and classification of taxonomic units.

Background: The development of high-throughput sequencing technologies has revolutionized the field of microbial ecology via the sequencing of phylogenetic marker genes (e.g. 16S rRNA gene amplicon sequencing). Denoising, the removal of sequencing errors, is an important step in preprocessing amplicon sequencing data. The increasing popularity of the Illumina MiSeq platform for these applications requires the development of appropriate denoising algorithms.
Proteogenomics pipeline

- Proteomics data
  - Mascot
  - Filter peptides mapping to genomic DNA

- Genomic DNA sequence
  - EMBOSS Transeq
  - Translation to 6 reading frames

- Blast
  - Get positions of peptide in genomic DNA
  - Compare with structural annotation
Rhodospirillum rubrum ATCC 11170 - chromosome Rru_A NC_007643
Rhodospirillum rubrum ATCC 11170 - chromosome Rru_A NC_007643
2357930 -- 2362930

(sequence length: 4352825 bases)
Phylogenetic footprinting

Cluster motif models:
Markov clustering (MCL) of output motif models based on the Pearson Correlation Coefficient
Other pipelines

Phylogenetic footprinting

AgrR

gene A

Protein A

Compare between ≠ species

Motif Sampler
AligACE
MEME
Weeder
GLAM

List of motif models
List of motif models
List of motif models
List of motif models
List of motif models

Cluster motif models:
Markov clustering (MCL) of output motif models based on the Pearson Correlation Coefficient
Phenotypic Biomarkers

Multivariate data

Red Fluorescence

Green Fluorescence

2 parameters

# cells
Introduction: Amplicon sequencing and data analysis

- Avoids the time/labour consuming culture base approaches.
- Allows the detection of unculturable bacteria.
- Can target a wide range of bacterial species.

16S rRNA Amplicon Sequencing

Extract DNA → PCR Amplification → Sequencing

Mock Samples

Chimera

Known microbial composition

Sequencing errors

Over-merging
Flow Cytometry

A Flow cytometric analysis

Data extraction

Create Fingerprint model

Create Fingerprints
However, culture based approach exhibit various disadvantages:

- Time and labor consuming
- Most of the species can not be cultured in lab conditions
- Do not provide a real time assessment of microbial community
Flow cytometry – test case