Week 14, Lecture 27

István Albert
Bioinformatics Consulting Center
Penn State
No homework this week

• Project to be given out next Thursday (Dec 1st)

• Due following Thursday (Dec 8th)

• Turn in all homework you might have missed (partial credit will be given)
Introduction to metagenomics

• genetic material recovered from environmental samples

• nothing lives on its own

• human body is estimated to contain 10 trillion cells

• the human body contains more bacteria than cells (estimates of 10:1) ➔ 100 trillion bacteria!
tract and can perform a number of useful functions for their hosts. The average human body, consisting of about $10^{13}$ (10,000,000,000,000 or about ten trillion) cells, has about ten times that number of microorganisms in the gut. The metabolic activities performed by these microorganisms in the intestines carry about ten times as many microorganisms in the intestines. The metabolic activities performed by these microorganisms are crucial for the human body's health and wellbeing.

http://xkcd.com/978/
I got intrigued!  
So what is the real number?

How many cells are there in the human body?

- Wooley JC, Godzik A, Friedberg I, 2010 A Primer on Metagenomics. PLoS Comput Biol 6(2): e1000667. doi:10.1371/journal.pcbi.1000667 it says to the benefits in agriculture, food industry, and medicine to name a few. We humans have more bacterial cells \(10^{14}\) inhabiting our body than our own cells \(10^{13}\) [2], [3]. It has been stated

The citation are


Let’s go back to 1977!


INTRODUCTION

The adult human organism is said to be composed of approximately $10^{13}$ eukaryotic animal cells (27). That statement is only an expression of a particular point of view. The various body surfaces and the gastrointestinal canals of humans may be colonized by as many as $10^{14}$ indigenous prokaryotic and eukaryotic microbial cells (70).


and down the rabbit hole we go, back to 1971
A man consists of some seven octillion \((7 \times 10^{27})\) atoms, grouped in about ten trillion \((10^{13})\) cells. This agglomeration of cells and atoms has some astounding properties; it is alive, feels joy and suffering, dis-

No citation or source given.

it appears that in 1971 Theodosius Grigorievich Dobzhansky stated that the human body contains ten trillion cells and this has been cited as a fact ever since.
Metagenomics approaches

• metagenome assembly ➔ reconstitute entire genomes from randomly fragmented short reads:
  
  – many (mostly) unknown genomes
  – genomes with various abundances
  – sequences with systematic errors
  – more data worse results?

  How may genomes could there be in the first place?

  Some say: over 1 billion!

  personal note: 1 billion sounds like a really big number
  and we just saw an example of where big numbers come from
16S RNA

• the gene 16S rRNA gene is highly conserved between different species of bacteria and archaea and:
  – parts of these genes are identical in just about all bacteria → universal primers
  – other sections of the 16S rRNA gene contain species-specific signature
  – LSU, SSU → large subunit and small subunit

rNRA → decodes mRNA into aminoacids → essential to life
16S rRNA sequencing

- isolate only the 16S rRNA genes, sequence only these

Pros: require far less coverage, allows, characterize population

Cons: we don’t know what the unknown bacteria are like other then the region that we sequenced
Phylotyping

• Map the reads against all known genetic sequences

• Find exact or partial (homologous) matches, characterize sample by the known genomes

• There are not that many fully sequenced bacteria - thousands? But there is data on lots of partially sequenced ones – millions.

• The tool to use is BLAST (and we will cover it later in this lecture)
OTU based approaches

- OTU – operational taxonomical units, most common approach, no taxonomonomy required:
  - sequences are characterized relative to their similarity to one another
  - groups are formed based on similarity thresholds
Strenghts and weaknesses

• Phylotyping → shifts/biases the view towards known sequences, has to rely on external taxonomy

• OTU → is far more abstract, it cannot tell what is actually there, grouping is subjective

Most often people use a combination of all thetechniques
Complication: competing standards

• Multiple sources of bacterial taxonomical classification
Competing 16s rRNA taxonomical databases

- *silva* – comprehensive ribosomal RNA database → ARB, Germany

- *rdp* → ribosomal database project → Michigan State, USA

- *greengenes* → comprehensive 16S rRNA gene sequence alignment

- also can download from GenBank

Each of these resources creates a slightly different taxonomy → Silva has 9 levels
Some tools mandate the use of the NCBI taxonomy

How to reference the NCBI taxonomy database

The NCBI taxonomy database is not a primary source for taxonomic or phylogenetic information. Furthermore, the database does not follow a single taxonomic treatise but rather attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy experts. Consequently, the NCBI taxonomy database is not a phylogenetic or taxonomic authority and should not be cited as such.
Using BLAST

- It used to be that bioinformatics was synonymous with using BLAST:
  
  Basic **Local** Alignment Search Tool

- Accessible via a web interface or command line

- Aligns sequences in a FASTA query file against a target database

- It can align nucleotides and proteins to other nucleotides and proteins and all combinations + it can also translate a query/target into all reading frames when searching for a match

- there are a fairly large number of tools are included → mastering Blast is a **extremely valuable skill**!
Download the binary for your platform from the NCBI website

- Comes in two versions BLAST and BLAST+, the new version has different program names and slightly different output (and that breaks a lot of other programs)

- NCBI recommends BLAST+

**BLAST+ executables**

BLAST+ is a new suite of BLAST tools that utilizes the NCBI C++ Toolkt. The BLAST+ applications have a number of performance and feature improvements over the legacy BLAST applications. For details, please see the [BLAST+ user manual](http://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) and the article in BMC Bioinformatics (PubMed link).

BLAST programs

old name

new name

blastn
blastx
tblastx
tblastn
blastp
psiblast
rpstblastn
makeblastdb
blastdb_aliaostool

Usage:

blastn [-h] [-help] [-import_search_strategy filename] [-task task_name] [-db datab
-subject subject_input_file] [-subject_loc range] [-query input
-out output_file] [-evalue evalue] [-word_size int_value]
-gapopen open_penalty] [-gapextend extend_penalty]
-perc_identity float_value] [-xdrop_ungap float_value]
Create an index to your target

On cygwin pass a relative path to blast rather than the one with ~ (the windows blast cannot handle unix based paths)

```
$ cd refs

$ makeblastdb -in yeast.fasta -parse_seqids -dbtype nucl

New DB name: yeast.fasta
New DB title: yeast.fasta
Sequence type: Nucleotide
Keep Linkouts: T
Keep MBits: T
Maximum file size: 1073741824B
Adding sequences from FASTA; added 18 sequences in 0.147231 seconds
```

```
$ blastn -query reads.fa -db ../../../refs/yeast.fasta
```
Run BLAST

Take the first 50 bases of chromosome 1 of the yeast genome and put it into the reads.fa

```
$ blastn -query reads.fa -db ../../../refs/yeast.fasta
```

```
Length=51

****** No hits found *****

Lambda    K    H
  1.33    0.621  1.12

Gapped
Lambda    K    H
  1.28    0.460  0.850

Effective search space used: 377054953

Database: yeast.fasta
Posted date: Nov 29, 2011 11:47 AM
Number of letters in database: 12,163,423
```
Turn off low complexity filtering

```
$ blastn -query reads.fa -db ../../refs/yeast.fasta -dust no
```

```
lcl|chr04
-009

>lcl|chr01
Length=230218

  Score = 95.3 bits (51),  Expect = 8e-021
  Identities = 51/51 (100%),  Gaps = 0/51 (0%)
  Strand=Plus/Plus

Query 1  CCACACCACCCACACACCCACACACACACACACACACACACACACACACACACACACACACACAC
        |-----------------------------------------------------
Sbjct 1  CCACACCACCCACACACCCACACACACACACACACACACACACACACACACACACACACACACACAC
        |-----------------------------------------------------
```

--More-- (byte 1155)
next lecture we continue on with BLAST and a metagenomics sample

Please install MEGAN: [http://ab.inf.uni-tuebingen.de/software/megan/](http://ab.inf.uni-tuebingen.de/software/megan/)
the MEtaGenome Analyzer