Next Generation Sequencing

Includes slides from: Walter Pirovano

Sanger sequencing

Overview

Applications
- Reference genome
- Methylation
- Deep sequencing (RNA)
- ChIP-Seq

Algorithms
- Alignment
- Hashes
- Assembly
- De Bruijn Graphs

BaseClear

Sanger sequencing

- Introduced by Fred Sanger in 1977
- First sequenced genome was PhiX-174 virus
  - 5K-bases - 2.9 million K-bases human genome
- The bases are detected using fluorescence
  - 4 different labels
- Nobel prize in 1980
Chain termination sequencing cannot deal with sequences longer than 2 k-BP.

The de Bruijn graph has a node for every k-mer observed in the sequence set and an edge between nodes if these two k-mers are observed adjacent in a read. Such edges are therefore associated with the single-step base difference.
Linear stretches:
continuous sequence

3. Simplification of linear stretches

Tip & Bubble removal

Error (tip and bubble) removal

Full de Bruijn Graph

A full de Bruijn graph of two related plasmids that have a locus in common. The de Bruijn graph was created with 30-bp k-mers. The open loops are regions that differ between the two plasmids, whereas the heavier lines indicate common regions.

Sanger sequencing

- Fragment DNA (shotgun)
- Amplify DNA - generate high numbers
- Extension & chain termination
  - Low number of fluorescent ddNTP's, that block the reaction
- Electrophoresis of single strand to sort on length

Human genome project (1990 – 2003)

- ‘a milestone for humanity’
- performed using traditional methods

Human genome project - in numbers

- 23 chromosome pairs
- 20,000 genes
- 2.9 billion base pairs
Sanger sequencing

- Still very much used:
  - Accurate
  - Long reads (up to 2000 bp)
  - $0.5 per K-base

- But:
  - Slow (96 - 384 simultaneous channels)
  - Need specific primer design for each sequencing reaction

Next Generation Sequencing

- Fragmentation
- Same ‘primer’ (for all fragments
- Create clusters of same sequence fragments or polonies
- Massively parallel sequencing on array

Next Generation Sequencing

- Massive parallel sequencing of millions of short fragments
- Very fast
- Huge amounts of data
- Reduced costs
  (e.g. 1 human genome ~10.000 $ vs. HGP 3 billion $)

Next-generation sequencing

- Platforms are based on +/- same principle
- Differences reside mainly in chemical usage and the way fragments are stuck to the surface
  - 454 Pyro sequencing (Roche)
    - long reads (up to 750 bp)
    - expensive, smaller amount of data
  - Solexa (Illumina) / SOLiD (Applied Biosystems)
    - huge amounts of data, relatively cheap
    - Shorter reads (up to 150 bp)

Basic strategy sequencing

Add same adapter sequence to each fragment → need only one unique primer!

Prepare library

Illumina: attaching adapters
Create Polonies

Pyro sequencing, SOLiD

Cyclic Array Sequencing

Illumina sequencing

$1 per M-Base

Illumina sequencing

What kind of data can we sequence?

• DNA
• RNA (which will be copied into cDNA)
  – mRNA
  – tRNA
  – rRNA
  – smallRNA
• ChIP-Seq (DNA with bound proteins)

NGS OUTPUT

What to do next? BIOINFORMATICS

Three examples:

– Reference genome alignment
– Transcriptome analysis
– De novo assembly of a genome

Reference alignment

• INPUT
  – Millions of sequenced fragments
  – A reference genome sequence
• PROCESS
  – String matching of the sequence reads against the reference (allow for small variations)
  – Reference and sequenced organism need to be closely related (at least the same species)
• OUTPUT
  – Alignment and sequence of new strain
Reference alignment
Match sequence to a given genome

Reference alignment
Match sequence to a ‘similar’ genome

Why do we need reference alignment techniques?

- What kind of sequencing experiments would create the problem of matching short sequence reads with an existing similar genome?

Alignments - Hash tables

Hash is like a dictionary in python

Alignments

- Hash Table:
  - Create hash from genome and map read
  - Create hash from reads and map genome

- Once a match between genome and read has been found, gapped and ungapped Smith-Waterman (dynamic programming) may be used for alignment

Mismatches

- In the previous example what may be cause of any mismatches alignment between the sequence read and the reference genome?

- How may this be different from a protein sequence alignment?
Programs for reference alignment

- Bowtie
- Maq
- SOAP
- CLCbio (commercial package)

Reference alignment application

- Mainly used for SNP detection
- More complicated: larger structural variance (INDELS, rearrangements)

Transcriptome alignment + analysis

- Measure transcription levels (mRNA, small RNAs) between different samples/tissues
- Find alternative splicing sites
- Will replace microarray platforms

Combining all data

- Sequence DNA and assembly
- Annotate your genome
- Sequence the RNA and measure transcription levels
- Study gene networks
- Study metabolic pathways
  - from genomics
  - to transcriptomics
  - to systems biology

De Bruijn graphs
**De novo assembly**

- **INPUT**
  - Millions of sequenced fragments

- **PROCESS**
  - Cut reads in k-mers and determine overlap through string-matching (allow for small variations)
  - No reference needed

- **OUTPUT**
  - Alignment and sequence of new strain

---

**Programs**

- Velvet (EBI)
- Abyss (Michael Smith Genomic Sciences Centre, Canada)
- Curtain (mix of Velvet/Abyss)
- SOAPdenovo (BGI China)
- CLCbio (commercial package)

---

**Assembly: Puzzling reads into a genome**

```
ACAG GAGGT read1
GAGGTCCAGA read2
CCAGATGATGATGATA read3
ACAGGAGGTCCAGAATGATGATGATA contig
```

---

**De novo assembly**

- The goal is to reconstruct the genome into a single sequence
- But in practice this is not possible
  - Sequencing errors
  - Lack of coverage
- Result: many sequences (contigs) of unknown order

---

**Sequencing errors**

- Incorrect information within the reads because of errors of sequencing machines
- If many sequencing errors: the assembler has problems with contig formation.
  Example;
  - ACAGGAGGT read1
  - GAGGTCCAGA read2
  - GAAGTCCAGA read3
  - GAGGTCCAGA read4

---

**Lack of coverage**

- Lack of coverage of the reads on the original genome causes multiple contigs
  - Due to randomness of shearing process there is a chance that some regions of the genome are unsequenced

---
Repetitive elements - 1

- Repeats can cause two major problems to the assembler;
  - 1) May be collapsed in two separate contigs

Repetitive elements - 2

- 2) Cause misjoin of non-adjacent reads and produces incorrect contigs by joining two distant regions of the genome together

De novo sequencing

- Reconstruction of novel genomes through a combination of different sequencing strategies
  - Single-read
  - Paired-end
  - Matepair

Single-read strategy

- Sequence from one end of the fragment
- Results is a number of contig sequences
- Problem: impossible to determine their order and orientation
- Need long range positional information to solve e.g. repeats

Paired-end strategy

- Sequence from both ends of the fragment of known size
- Using interspacing size to determine distance between contigs (and solve e.g. repeat regions)
- Problem: fragment size is limited to few hundreds of basepairs -> fill up minor gaps

Matepair strategy

- Illumina technique developed to solve larger repeats
- Allows for larger fragment size (up to 10kb)
- Consequently allows for larger distance between the two reads (sequenced at each fragment end)
Scaffolding contigs using paired-end and/or matepair data

- Find the links between the contigs
- Merge different contigs into scaffolds (supercontigs)
- In the end we shed (some) light on:
  - The contig order
  - The contig orientation
  - The distance between contigs

In the ideal scenario ...

Output of BaseClear scaffold

SSPACE

- FASTA file with scaffolds (supercontigs)
- Graphical display

A note on de novo assembly quality

- No official measure, only the true genome sequence can provide this!
- N50 value (sort all contigs on size, and determine in which contig 50% of all bases are assembled, the larger the better)
- Number of contigs
- Average contig size
- Total genome size (does this meet expectations?)
- Same holds for % GC content

Overview

- Sanger sequencing
  - Chain termination
  - Shotgun
  - Assembly...
- NGS
  - Pyro 545
  - Illumina (Solexa)
  - Others...
- Outlook - single molecule
- Applications
  - Reference genome
  - Methylation
  - Deep sequencing (RNA)
  - ChIP-Seq
- Algorithms
  - Alignment
  - Hashes
  - Assembly
  - De Bruijn Graphs
Our home-brewed scaffolding tool

- **SSPACE** is a stand-alone program for scaffolding pre-assembled contigs using paired Illumina data
- It is based on the SSAKE scaffolding protocol (Warren et al. 2007)
- Main features are:
  - Easy to use
  - Reduced runtime and memory usage
  - Multiple library input of paired-end and/or matepair datasets
  - Optional: contig extension using unmapped reads

SSPACE algorithm

1. Read in pre-assembled contigs (fasta)

2. Read in (multiple) Illumina paired-read sequences; Filter on unique paired-reads

3. Align the paired reads to the contigs using Bowtie

4. Optional: Use unaligned reads for contig-extension

5. Combine contigs into scaffolds if linked by a specified number of reads (within limit of library insert size)

6. Use insert size to estimate gap between the contigs

SSPACE output

- FASTA file with scaffolds (supercontigs)

- Graphical display

Our home-brewed scaffolding tool

- **SSPACE** is a stand-alone program for scaffolding pre-assembled contigs using paired Illumina data
- It is based on the SSAKE scaffolding protocol (Warren et al. 2007)
- Main features are:
  - Easy to use
  - Reduced runtime and memory usage
  - Multiple library input of paired-end and/or matepair datasets
  - Optional: contig extension using unmapped reads