**Week 5, Lecture 10**

**István Albert**

Biochemistry and Molecular Biology and Bioinformatics Consulting Center

Penn State

---

**Sequence alignments**

- Arranging two or more sequences such as to maximize the length of the common regions between the two
- It is a very well developed field – the roots of the bioinformatics started with various alignment software
- We will only cover pair-wise alignments searching a database with a query
- High throughput sequencing poses special constraints: a very large number of very short reads - traditional methods were not feasible

---

**Alignment concepts**

- Perfect match
- One mismatch
- Insertion vs ref.
- Deletion vs ref

NOTE: mismatches or indels can be longer than 1 base!

It gets complicated very quickly

Alignment scoring depends on mismatch scoring (different across bases!), gap open, gap extension penalties
Challenges: repetitive and low complexity regions

Short Read Mappers (aligners)

- Use heuristics to quickly identify locations (hits) where the reads match
- Heuristics → not all hits will be found!
- Tradeoff: resource usage vs speed vs accuracy vs usability
- Each domain of application may have more appropriate tools

A few popular aligners

- BWA, bowtie, SOAP2, Shrimp, BFAST and many others

There is no single best tool, the issues to consider:

1. documentation → can we figure out how it works
2. input features → what type of input can it handle
3. reporting features → will it produce the type of output that we can use
4. performance → is it feasible to run on my resources

Aligner features

- Some cannot handle indels (insertion/deletions) → these tools are typically extremely fast will run on a laptop for even large genomes (bowtie 1) → application Chip-Seq
- May or may not use the quality score during the alignment (usually not, that would alter the alignment score)
- Differences in the they report alignments:
  - unique alignments only (note that this is an ill defined concept!)
  - best alignments above a cutoff
  - all possible alignments (can slow down the process greatly)
BWA (Burrows-Wheeler Aligner) by Heng Li

BWA is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome. It implements two algorithms, bwa-short and BWA-SW. The former works for query sequences shorter than 200bp and the latter for longer sequences up to around 100kb. Both algorithms do gapped alignment. They are usually more accurate and faster on queries with low error rates. Please see the BWA manual page for more information.

http://bio-bwa.sourceforge.net/

Download, unpack, compile with make and link to bin

Download – unpack and make

```
ialbert@porthos ~/src
$ tar jxf bwa-0.6.2.tar.bz2
ialbert@porthos ~/src
$ cd bwa-0.6.2
ialbert@porthos ~/src/bwa-0.6.2
$ make
```

Uses bz2 compression hence the j flag (the z command would expand a gzip file)

Link it the tool into your bin folder

```
ialbert@porthos ~/src
$ cd ~/bin/
ialbert@porthos ~/bin
$ ln -s ~/src/bwa-0.6.2/bwa bwa
ialbert@porthos ~/bin
$ ls
ack blastn fetch makeblastdb blastdbcmd bwa lasts ssearch
ialbert@porthos ~/bin
$`

from: http://www.massgenomics.org/short-read-aligners
Aligning reads with BWA

For BWA is three step process (this is different for each aligner):

1. Index the genome – this only needs to be done once for each genome → bwa index ...
2. Create the alignment → bwa aln ...
3. Report the alignment → bwa samse ...

There is built-in help

Index the genome

For long genomes (over 2GB) you need to index with the –a bwtsw option.

Periodically re-read the manual! BWA has very useful instructions!

Create a FASTQ query sequence

- Select some bases from the beginning of chromosome 1
BWA concepts

- the **index or prefix** (*.fasta): the full path to the fasta reference sequence that has been indexed
- the **query** (*.fq): the file containing the sequences that we want to look up in the reference
- the **alignment** (*.sai): a binary intermediate file that contains the alignment information for each read
- the **sequence alignment map** (*.sam): the result file that contains the alignment cross referenced with the query

Run and report BWA

Homework 10

Create a FASTQ file using the first 60 or so bases from chromosome 1 of the yeast genome. The file should have four sequences:

1. first is perfect match
2. contains one mismatch (past the 10 base)
3. contains a deletion (past the 10 base)
4. contains an insertion (past the 10 base)

Align your file with BWA and print out your SAM file

Tips: Use and editor to make the changes. Remember to add or delete quality scores if the sequence length changes