Week 14, Lecture 27

István Albert

Biochemistry and Molecular Biology
and Bioinformatics Consulting Center

Penn State
Pet peeve: there is no PROTOCOL!

Data analysis ≠ PROTCOL

Beware of countless false prophets that claim that following a given set of instructions will lead to discovery.

Data analysis -> know what is possible, know what tools do, pick the right one, apply it correctly, evaluate the results, decide what needs be done next.
Chip-Seq Frameworks

Often these come as entire packages of many related tools (most of which are interval operation)

Chip-Seq studies are a series of independent operations:

1. Align the reads
2. Identify intervals that show enrichment
3. Place these intervals into genomic context → distance to markers, sequences that are covered by the intervals,

Step 3 is usually a very complex procedure that is independent of others
Some of the numerous Chip-Seq frameworks

- **ChipSeeqer**: A comprehensive framework for the analysis of ChIP-seq data
- **USeq**: a collection of software tools for ChIP-Seq and RNA-Seq studies
- **HOMER**: software for motif discovery and ChIP-Seq sequencing analysis
- **MochiView**: genome browser with a ChIP-Seq focus

NOTE: many of the ChIP-Seq techniques also apply to RNA-Seq – both approaches produce counts over intervals
Chip-Seq peak finders

• Very large selection of tools and techniques:

  ERANGE, FindPeaks, MACS, QuEST, CisGenome, SISSRS, PeakSeq, SPP, ChIPSeqR, GLITR, ChIPDiff, T-PIC, BayesPeak, MOSAiCS, CCAT, CSAR,

• I also developed one – **GeneTrack**: a genomic data processing and visualization framework Bioinformatics, **2008**
I lost interest around 2009 and moved on with different projects.

It has been only much later when I realized that people still use it.

Lesson: **have more faith...**

I still consider it be the most accurate and easiest to comprehend approach to peak predictions.
MACS

- MACS – Model Based Analysis for Chip-Seq (one of the most commonly used peak predictors)

- SSIRS: Genome-Wide Identification of in Vivo Protein-DNA Binding Sites From ChIP-Seq Data
Example predictions with macs and sisrs.pl

Neither of which looks quite right
The nomenclature is a bit hazy

- The term “peaks”, “signal”, and “enriched regions” are used interchangeably → leads to a lot of confusion

Proper definition would be

- Peaks → regions with a well defined shape, a midpoint and a spread that describe one or more measurements

- Enrichment → selecting those peaks that show statistically significant properties when compared to background/control
More on how GeneTrack works

• In general when using any peak predictor make a concerted effort to understand the mechanism by which the tool operate

• Create test inputs and verify what the tools does

• We use the following slides as a demonstration on what issues every peak predictor has to address – we’ll show how we addressed them in GeneTrack
What is the Error Model

• Measurements are affected by error
• Replace every read with the probability of the read falling onto the observed location.

Each single coordinate will be replaced with a “smudge” of values (hundreds of them).
Central Limit Theorem guarantees another normal function
Summing up multiple errors
GeneTrack: revisit the reads
GeneTrack: with smoothing
GeneTrack also does peak prediction
Command line GeneTrack

• Author Pindi Albert (written as a rotation freshman)

• The functionality of GeneTrack website ported to a command line tool
Clone the **chipexo** repository from **github**

```bash
ialbert@porthos ~/src
$ git clone https://github.com/ialbert/chipexo.git
Cloning into chipexo...
remote: Counting objects: 350, done.
remote: Compressing objects: 100% (136/136), done.
remote: Total 350 (delta 206), reused 349 (delta 205)
Receiving objects: 100% (350/350), 885.81 KiB, done.
Resolving deltas: 100% (206/206), done.

ialbert@porthos ~/src
$ python ~/src/chipexo/genetrack/genetrack.py
Usage: genetrack.py [options] input_paths

input_paths may be:
- a file to run on
- "-" to run on standard input

element usages:
python genetrack.py -s 10 /path/to/a/file.txt
python genetrack.py -s 5 -e 50 -

Options:
- h, --help show this help message and exit
- s SIGMA Sigma to use when smoothing reads to call peaks. Default 5
- e EXCLUSION Exclusion zone around each peak that prevents others from
  being called. Default 20.
```
GeneTrack output
Homework 27

Generate peak calls with a peak predictor for the long and short dataset from lecture 26.

Visualize the results and discuss what you see