Week 11, Lecture 21

István Albert

Biochemistry and Molecular Biology
and Bioinformatics Consulting Center

Penn State
Genome representation concepts

- At the simplest level of abstraction the genome is represented by a one dimensional “space” (lines)
- Genome is two stranded → a line corresponds to each strand
- Each strand has a polarity → each line has a direction
- Strands (lines) are paired
- The smallest unit is one base → one integer on the number line
- Annotations (features) are segments (coordinates) on each line
Genomic coordinates – brief overview

DNA two stranded and directional
But there is only one coordinate system

The upstream region – before the 5’ end relative to the direction of transcription

Standard formats use start < end even for the reverse strand
Coordinate systems

• 0 based $\rightarrow$ 0, 1, 2, ... 9
• 1 based $\rightarrow$ 1, 2, 3, ... 10

Typically

• 0 based are non-inclusive $10:20 \rightarrow [10, 20)$
• 1 based include both ends $10:20 \rightarrow [10, 20]$
## Comparing coordinate systems

<table>
<thead>
<tr>
<th></th>
<th>1 based indexing</th>
<th>0 based indexing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third element</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>First ten</td>
<td>1, 10</td>
<td>0, 10</td>
</tr>
<tr>
<td>Second ten</td>
<td>11, 20</td>
<td>10, 20</td>
</tr>
<tr>
<td></td>
<td>21, 30</td>
<td>20, 30</td>
</tr>
<tr>
<td>One base long starting at 10</td>
<td>10, 10</td>
<td>10, 11</td>
</tr>
<tr>
<td>Length of interval</td>
<td>end - start + 1</td>
<td>end - start</td>
</tr>
<tr>
<td>Five elements starting at 1000</td>
<td>1000, 1004</td>
<td>1000, 1005</td>
</tr>
<tr>
<td>Empty interval</td>
<td>?</td>
<td>start, start (0,0)</td>
</tr>
</tbody>
</table>

Vote for what you think is better
• **SAM/BAM** – Sequence Alignment Map

• **BED/GFF** → Gene Annotation representation

• **VCF/BCF** → for variant calls
Pick one coordinate system and stick with it!

1. Pick the standard your group is using and convert every new data to this standard!
2. If you have a choice of what to use pick the one based system! (GFF).
What is a genomic feature?

- Feature: a genomic region (interval) associated with a certain annotation (description).

Typical attributes to describe a feature:

1. chromosome
2. start
3. end
4. strand
5. name
Values on intervals

• A single value characterizes an entire interval → score (value) for the interval

• Continuous values → different value for each base of the interval → analogous to a series of 1bp long intervals

Different data representation formats
http://genome.ucsc.edu/FAQ/FAQformat.html
Two commonly used formats

• **BED** – UCSC genome browser → 0 based non inclusive → also used to display tracks in the genome browser (US “standard”) (variants: **bigBed**, **bedgraph**)

• **GFF** – Sanger institute in Great Britain → 1 based inclusive indexing system (“European standard”), (variants: **GTF**, **GFF 2.0**)
**BED format**

Search for BED format

Tab separated 3 required and 9 optional columns. Lower numbered filed must be filled.

1. `chrom` (name of the chromosome, sequence id)
2. `chromStart` (starting position on the chromosome)
3. `chromEnd` (end position of the chromosome, **note** this base is not included!)
4. `name` (feature name)
5. `score` (between 0 and 1000)
6. `strand` (+ or -)
7. `thickStart` (the starting position at which the feature is drawn thickly)
8. `thickEnd` (the ending position at which the feature is drawn thickly)
9. `itemRGB` (RGB color → 255, 0, 0 display color of the data contained)
10. `blockCount` (the number of blocks (exons) in the BED line.)
11. `blockSizes` (a comma-separated list of the block sizes)
12. `blockStarts` (a comma-separated list of the block starts)
GFF format

Search for GFF3 ⇒ http://www.sequenceontology.org/gff3.shtml

Tab separated with 9 columns. Missing attributes may be replaced with a dot ⇒ .

1. **Seqid** (usually chromosome)
2. **Source** (where is the data coming from)
3. **Type** (usually a term from the sequence ontology)
4. **Start** (interval start relative to the seqid)
5. **End** (interval end relative to the seqid)
6. **Score** (the score of the feature, a floating point number)
7. **Strand** (+ or −)
8. **Phase** (used to indicate reading frame for coding sequences)
9. **Attributes** (semicolon separated attributes ⇒ Name=ABC;ID=1)

people like to stuff a lot of information here
Wiggle format

- two versions → fixed step and variable step each trying to optimize the amount of data storage

```plaintext
fixedStep chrom=chr1 start=100 step=1
10
15
11
22
...

variableStep chrom=chr1
100 10
101 15
102 11
103 22
variableStep chrom=chr2
2000 23
2005 40
...
```

Wiggle is a nasty format – it looks simpler than it is – please avoid
We may have data in different coordinate systems!

Being “one off” is one of the most common errors in bioinformatics.

Conversion from GFF to BED

\[(\text{start, end}) \rightarrow (\text{start} - 1, \text{end})\]

Conversion from BED to GFF

\[(\text{start, end}) \rightarrow (\text{start} + 1, \text{end})\]

Not that there will be differences when selecting positions that depend on the END coordinate!
Handling coordinates relative to intervals

What are the coordinate of the base preceding and following the interval

GFF [start, end] → preceding base is at start - 1
BED [start, end) → preceding base is at start - 1

GFF [start, end] → next base is at end + 1
BED [start, end) → next base is at end
Representing interval relationships

- We have a gene with three splicing variants

Note: each exon is 1kb separated by multiples of 1kb

How to represent this in data analysis?
Data representation

- Both BED and GFF files can represent them.

- Two common versions of GFF → **GTF2** and **GFF3**
  (note: tool documentation can often wrong and shows a weird combination of these two formats)

- In GFF the content of the ATTRIBUTE (9\text{th}) column specifies the relationship between features.
GTF/GFF formats

GTF attributes:

- `gene_id` value;  
  a globally unique identifier for the genomic source of the transcript
- `transcript_id` value  
  a globally unique identifier for the predicted transcript.

`gene_id “G1” transcript_id “T1”`

GFF attributes:

`ID=exon1; Parent=T1`

See the GFF3 site for exact specification of these terms.
Important: More than one parent may be listed!
A distinct line is entered for each exon, repeated for each transcript.
Example interval as GFF 3

The same exon may be part of different transcripts (parents)

<table>
<thead>
<tr>
<th></th>
<th>gff-version</th>
<th>chrom</th>
<th>source</th>
<th>start</th>
<th>end</th>
<th>strand</th>
<th>Parent</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>chrI</td>
<td>demo</td>
<td>1000</td>
<td>2000</td>
<td>+</td>
<td>T1,T2,T3;</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>chrI</td>
<td>demo</td>
<td>3000</td>
<td>4000</td>
<td>+</td>
<td>T1,T2;</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>chrI</td>
<td>demo</td>
<td>5000</td>
<td>6000</td>
<td>+</td>
<td>T1,T3;</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>chrI</td>
<td>demo</td>
<td>7000</td>
<td>8000</td>
<td>+</td>
<td>T1,T2,T3;</td>
</tr>
</tbody>
</table>
Example interval in BED

From the BED format specification

6. strand - Defines the strand - either '+' or '-'.
7. thickStart - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).
8. thickEnd - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. itemRgb - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line itemRgb attribute is set to "On", this RGB value will determine the display color of the data contained in this BED line. NOTE: It is recommended that a simple color scheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. blockCount - The number of blocks (exons) in the BED line.
11. blockSizes - A comma-separated list of the block sizes. The number of items in this list should correspond to blockCount.
12. blockStarts - A comma-separated list of block starts. All of the blockStart positions should be calculated relative to chromStart. The number of items in this list should correspond to blockCount.
Visualizing in IGV
Homework 21

• Create and visualize in IGV an interval file that contains three splice variants of a 1kb long gene with 5 exons.

• Show the file and a screenshot