

Notes on representation of genomic coordinates: Interbase versus One-based

Interbase:

Bed files and output from the cDNA-to-genome alignment program `blat` use the interbase coordinate system, which numbers between bases, as shown in the diagram below. Note this is identical to how taking slices of strings works in python.

In interbase coordinates, the first base of a sequence string begins at position zero, the next base at position one, the third base at position two, etc.

Interbase (and bed files) define blocks, or ranges, of genomic sequence using start and end coordinate pairs (s,e) where s (start) indicates the index of the first base and e (end) indicates the index of the first base not included in the range. In addition, $e \geq s$ and the length of a range (the number of bases it covers) is always $e - s$.

For example, R1 in the diagram below has

$(s,e) = (2,5)$ with length $5-2 = 3$ bases.

```

    _____ R1 (start,end) = 2,5
   a t t t a a a a
  1 2 3 4 5 6 7 8 9
```

Two ranges R1 and R2 overlap when they cover some of the same sequence bases.

To identify the start and end positions of an overlap R3 between ranges R1 and R2, note that R3 can never include any bases to the left of the largest start position. Nor can it contain any bases to the right of the smallest end position. The start and end positions of R3 (s3,e3) are the maximum of (s1,s2) and the minimum of (e1,e2). R1 overlaps R2 whenever the length of R3 > 0 .

Consider the following scenarios:

R1: (2,5) R2: (5,8) R3: $(\max(2,5), \min(5,8)) = (5,5)$

length(R3) = $5-5 = 0$ NO OVERLAP

```

   _____ _____
  a t t t a a a a
  1 2 3 4 5 6 7 8 9
```

R1: (2,6) R2: (4,7) R3: $(\max(2,4), \min(6,7)) = (4,6)$

length(R3) = $6-4 = 2$ OVERLAP

```

   _____
  a t t t a a a a
  1 2 3 4 5 6 7 8 9
```

One-based:

This is the coordinate scheme used by most bioinformatics programs, e.g., blast. It is also the same scheme used in GFF (general gene format).

In this scheme, we number the bases, not the gaps between them:

R: start = 3, end = 5, length = end-start+1 = 5 - 3 + 1 = 3

```
      _____ R
a t t t a a a a
1 2 3 4 5 6 7 8
```

Comparing interbase and one-based schemes:

The one-based scheme breaks down when we need to designate gaps between bases or boundaries between adjacent segments. If you wanted to use one-based coordinates to designate a boundary between two segments of sequence, you would have to come up with some perhaps non-intuitive scheme. In interbase, you could indicate a boundary like so:

B = (s, s)

```
      | B = (2,2)
a t t t a a a a
1 2 3 4 5 6 7 8
```