Sequence alignment

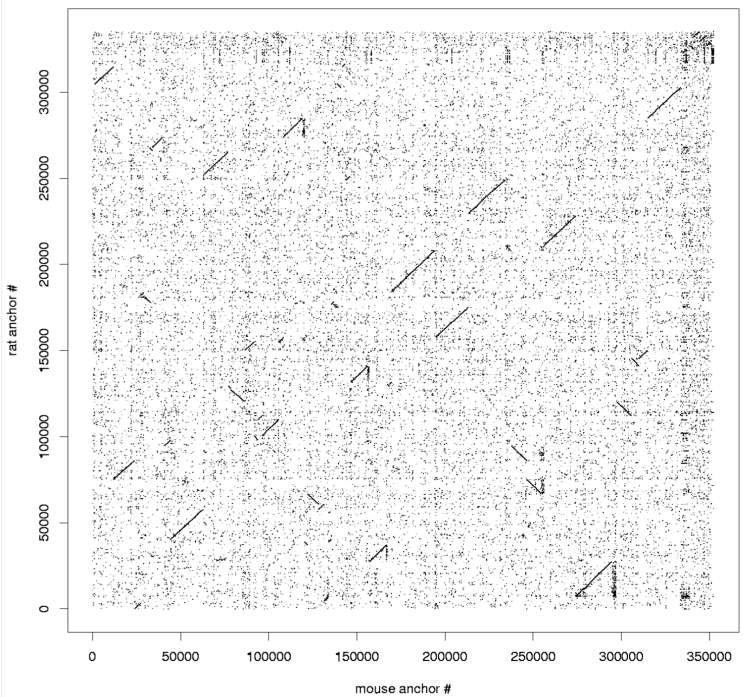
The pairwise alignment task

Outline

- How do we cast sequence alignment as a computational problem?
- Classes of pairwise alignment
- Task definition
- Scoring an alignment

Dot plots

- Not technically an "alignment"
- But gives picture of correspondence between pairs of sequences
- Dot represents similarity between segments of the two sequences



Issues in Sequence Alignment

- the sequences we're comparing probably differ in length
- there may be only a relatively small region in the sequences that matches
- we want to allow partial matches (i.e. some amino acid pairs are more substitutable than others)
- variable length regions may have been inserted/deleted from the common ancestral sequence

Two main classes of pairwise alignment

- Global: All positions are aligned

```
CA--GAGTCGAAT
CGCCGA-TC-A--
```

- Local: A (contiguous) subset of positions are aligned

```
  ..GAGTC....
....GA-TC.
```

Pairwise Alignment: Task Definition

- Given
  - a pair of sequences (DNA or protein)
  - a method for scoring a candidate alignment
- Do
  - find an alignment for which the score is maximized

Scoring An Alignment: What Is Needed?

- substitution matrix
  - *S(a,b)* indicates score of aligning character *a* with character *b*
- gap penalty function
  - *w(k)* indicates cost of a gap of length *k*

Linear Gap Penalty Function

- different gap penalty functions require somewhat different algorithms
- the simplest case is when a linear gap function is used

$$w(k) = s \times k$$

where s is a constant

- we'll start by considering this case

Scoring an Alignment

- the score of an alignment is the sum of the scores for pairs of aligned characters plus the scores for gaps
- example: given the following alignment

```
VAHV---D--DMPNALSALSDLHAHKL
AIQLQVTGVVVTDATLKNLGSVHVSKG
```

- we would score it by

  S(V,A) + S(A,I) + S(H,Q) + S(V,L) + 3*s* + S(D,G) + 2s …

Summary

- Global and local pairwise alignment
- Scoring an alignment