Whole genome alignments

http://faculty.washington.edu/jht/GS559_2012/

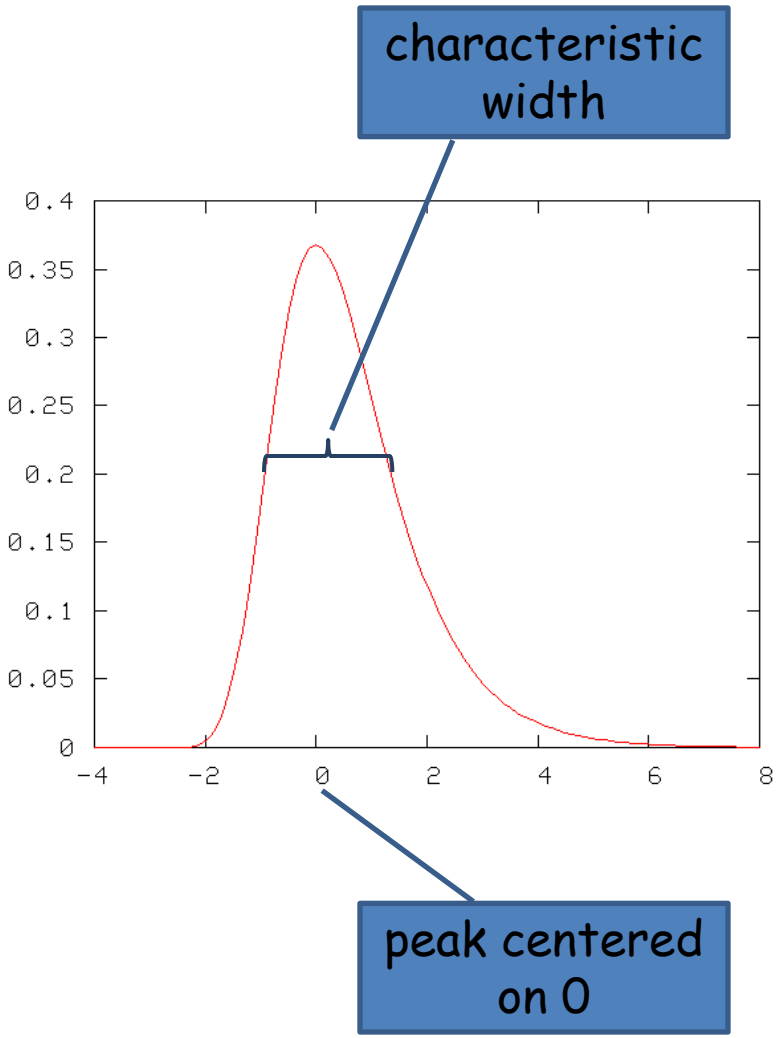
Genome 559: Introduction to Statistical and Computational Genomics

Prof. James H. Thomas

Problem set 3 will be posted by tonight.

The Python part may take you a LONG time!

Unscaled EVD equation

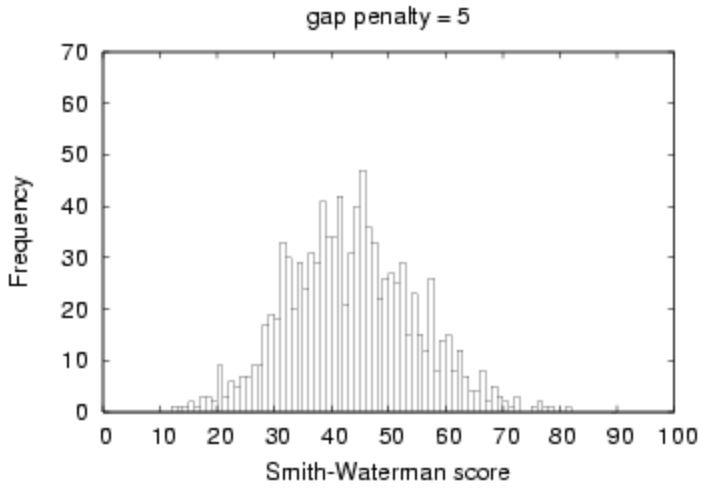


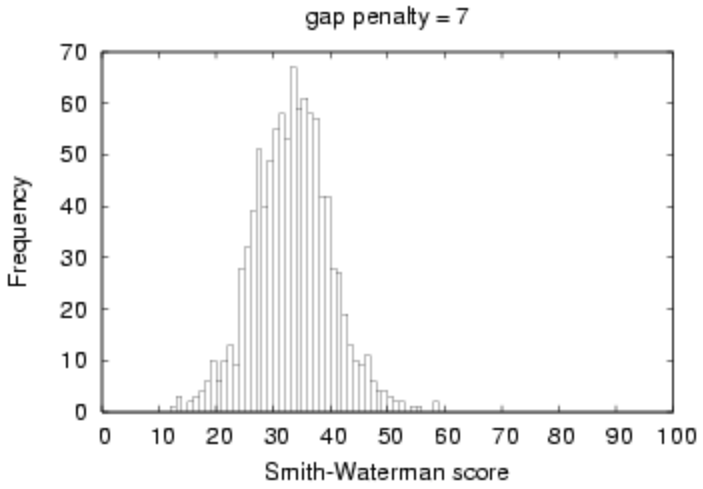
$$P\ \ S \geq x\ \ = 1 - e^{(-e^{-x})}$$

S is data score, x is test score

(FYI this is 1 minus the cumulative density function or CDF)

Scaling the EVD





- An EVD derived from, e.g., the Smith-Waterman algorithm with BLOSUM62 matrix and a given gap penalty has a characteristic mode parameter μ and scale parameter λ.

$$P\ \ S \geq x\ \ = 1 - e^{(-e^{-x})}$$ scaled: $$P\ \ S \geq x\ \ = 1 - e^{(-e^{-\lambda(x-\mu)})}$$

λ and μ depend on the size of the query, the size of the target database, the substitution matrix and the gap penalties.

Similar to scaling the standard normal

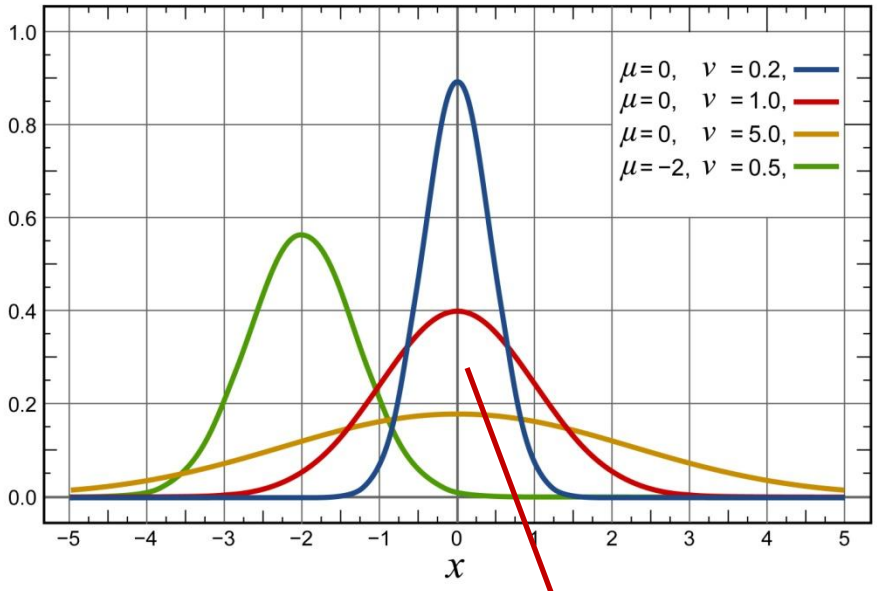
$$PDF_{stdNormal} = Ce^{-x^2/2}$$

where $C = 1/\sqrt{2\pi}$

$$PDF_{scaledNormal} = Ce^{-(x-\mu)^2/2v}$$

where $C = 1/\sqrt{2\pi v}$

v is variance, μ is mean



(μ moves peak and v adjusts peak width)

Summary score significance

- A distribution plots the frequencies of types of observation.
- The area under the distribution curve is 1.
- Most statistical tests compare observed data to the expected result according to a null hypothesis.
- Sequence similarity scores follow an extreme value distribution, which is characterized by a long tail.
- The p-value associated with a score is the area under the curve to the right of that score.
- Selecting a significance threshold requires evaluating the cost of making a mistake.
- Bonferroni correction: Divide the desired p-value threshold by the number of statistical tests performed.
- The E-value is the expected number of times that a given score would appear in a randomized database.

Whole genome alignments

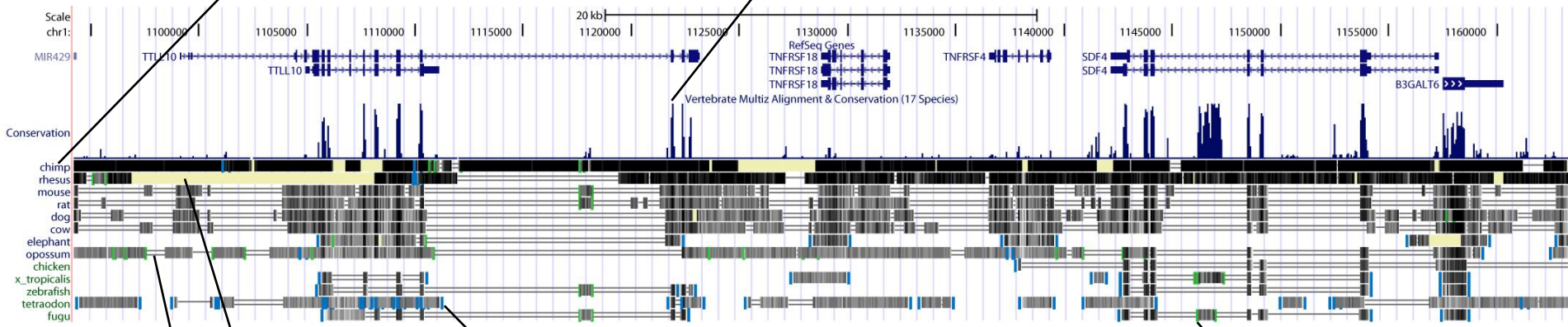
Why?

- genome-wide alignment data (efficient)
- inference of shared (orthologous) genes across species
- genome evolution

UCSC Browser track

individual genome alignments, darker = higher scoring

averaged conservation for 17 genomes

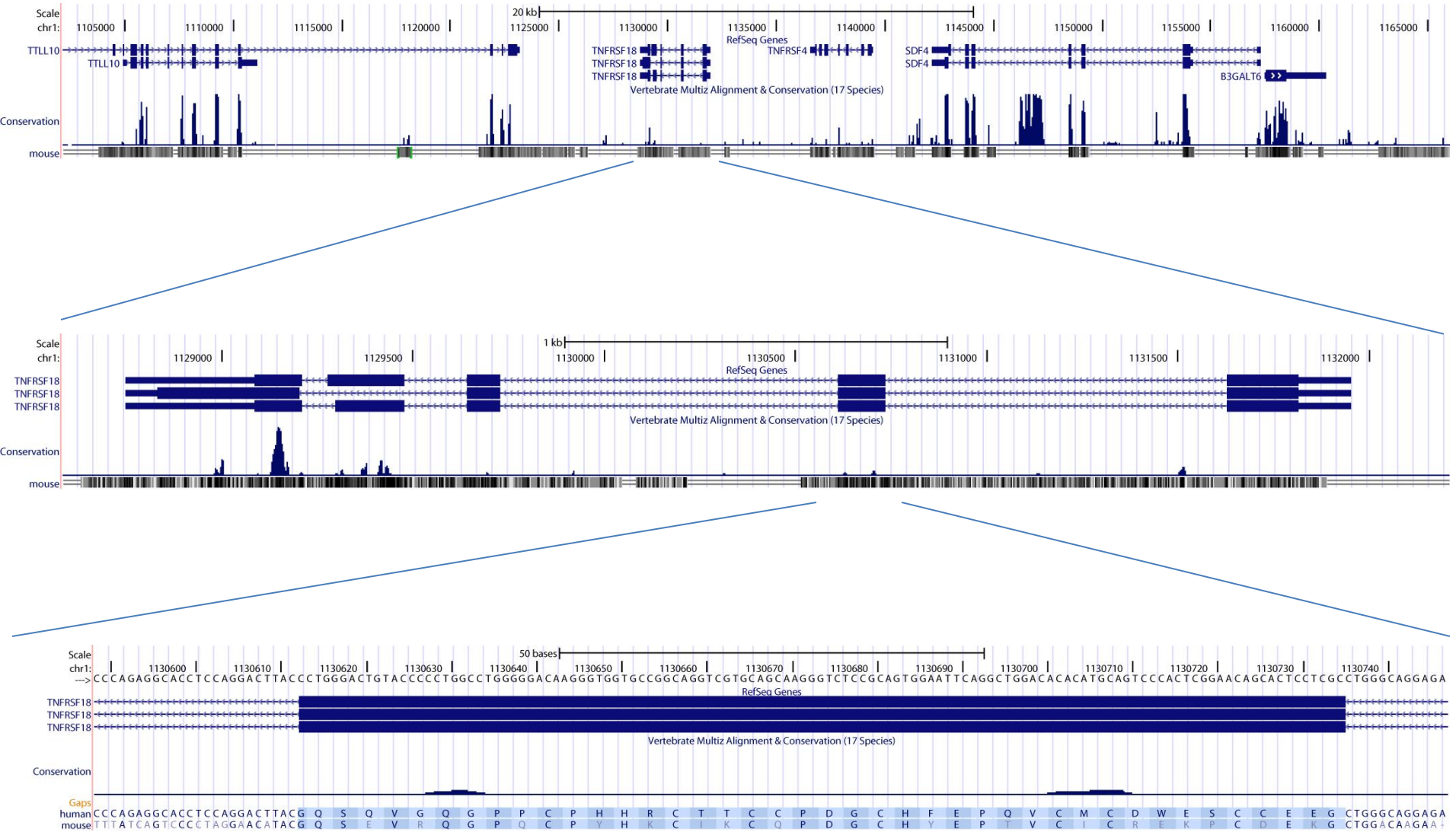


known gap in assembly

alignment discontinuity (e.g. translocation break point)

questionable alignment segment

= sequence present but unalignable



GQSQVGQGPPCPHHRCTTCCPDGCHFEPQVCMCDWESCCEEG

GQSEVRQGPQCPYHKCIKCQPDGCHYEPTVCICREKPCDEKG

How are genome-wide alignments made?

- mouse and human genomes are each about 3×10^9 nucleotides.

- how many calculations would a dynamic programming alignment have to make?

- at a minimum - 3 integer additions and 3 inequality tests for each DP matrix position

- DP matrix size is 3×10^9 by 3×10^9

- about $6 \times (3 \times 3 \times 10^{18}) = 5.4 \times 10^{19}$ calculations!

  Age of the universe is about 4.3×10^{17} seconds

(by the way, there are other problems too, including assuming colinearity)

Making large searches faster

- Most common method is the BLAST search (Basic Local Alignment Search Tool). Only the initial step is different from dynamic programming alignment.

- Search sequence is broken into small **words** (usually 3 residues long for proteins). 20 * 20 * 20 = 8,000 protein words. These act as **seeds** for searches.

- The target dataset is pre-indexed for all the positions in the database sequences that match each search word above some score threshold (using a score matrix such as BLOSUM62).

BLAST searches (cont.)

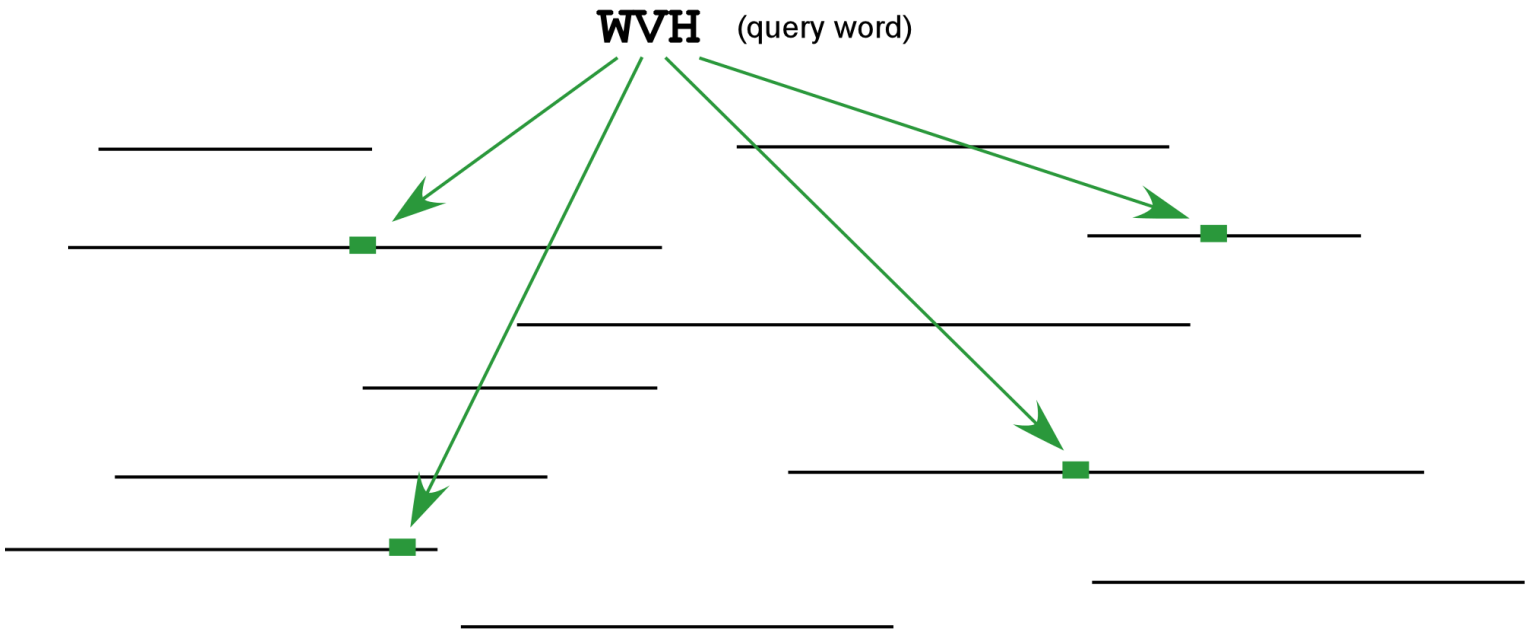
- For example, the search sequence word "WVH" might score above threshold with these indexed sequences:

Indexed word	Score
WVH	**23**
WIH	**22**
WVY	**17**
WIY	**16**

- Target sequences around each indexed word hit are retrieved and the initial match is extended in both directions:

```
...VFEWVHLLP...      your sequence
   ⟵ WIY ⟶           database (many sites)
```

Schematic of indexed matches



Result – instead of aligning these 3 amino acids to everything, they are aligned only with the tiny fraction of sequence regions that are good candidates for a valid alignment.

(note- blast actually looks for two such matches close to each other)

Extension and scoring

	Match Score:	Total Score:
...QSVFEWVHLLPGA... ..WIY..	16	16
...QSVFEWVHLLPGA... ..WIY**Q**..	-3	13
...QSVFEWVHLLPGA... ..WIYQ**K**..	-2	11
...QSVFEWVHLLPGA... ..WIYQK**A**..	-1	10

[mention gap variant]

Extension termination

- Extension is continued until the cumulative score drops below some threshold (usually 0).

- This permits the match to cross a region of marginal similarity or frank mismatching (e.g. a small intron in tblastn) if it flanks a region of high similarity.

- Extensions whose maximal cumulative score is above some threshold are kept for reporting to user.

- For web interfaces, various formatting, links, and overviews are added.

- It is also easy to set up blast on your local computer; useful for custom databases and automation.

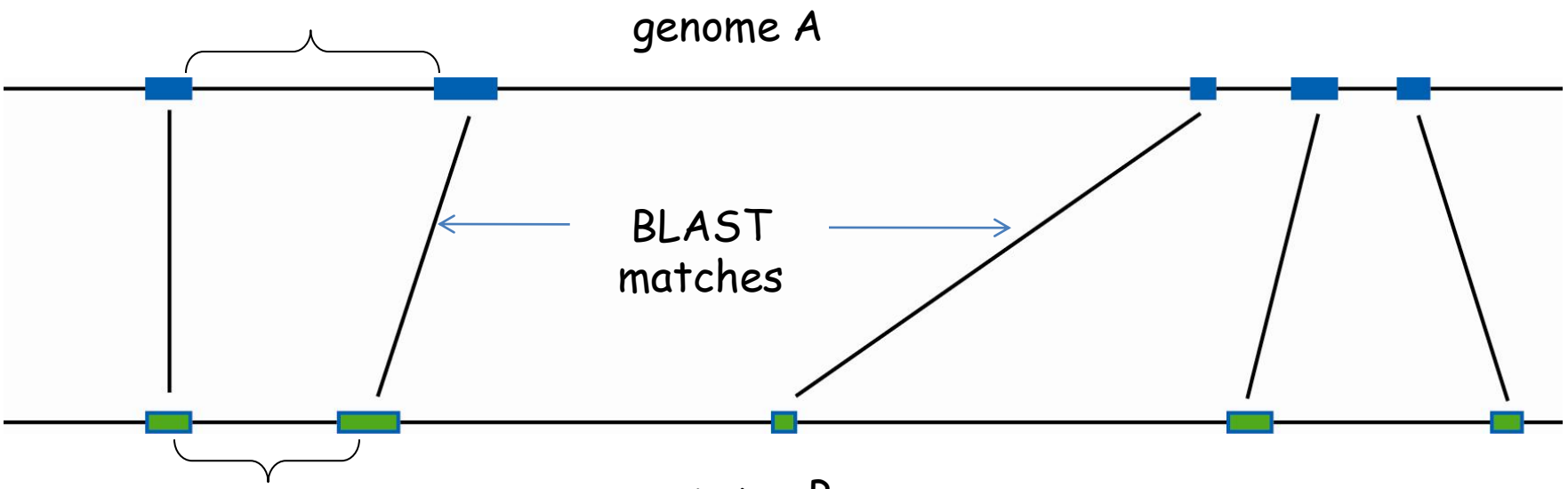
Key to speed: word matching and prior indexing

- Though gapped blast local alignment is slow, only a very small part of total search space is analyzed.

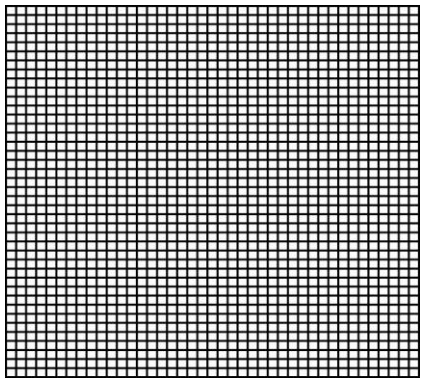
- Because the positions of all database word matches are indexed prior to the search, the relevant parts of search space are reached quickly.

- Tradeoff is in sensitivity – occasionally matches will be missed (when they are distant enough and dispersed enough that no local word pairs match well enough).

Dynamic programming after BLAST matching



DP alignment region



M x N manageable

Anchored DP alignment: if two blast matches are nearby and in the same orientation, DP align everything between them.