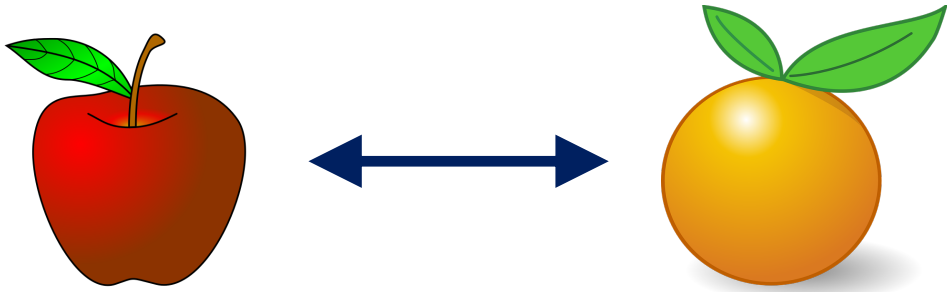
Homology & BLASTology



Comparing apples to oranges tells you nothing about apples or oranges

BIOL 435/535: Bioinformatics

January 20, 2022

How homologous are the two amino acid sequences?

```
Seq1 – PLSQMFFWAF
Seq2 – PLSQVFFWTF
       ***** **** *
```

* = Identical amino acid

Trick question –

Homology is an inference of shared common ancestry
Similarity is the criterion used to make that inference

```
Seq1 – PLSQMFFWAF
Seq2 – PLSQVFFWTF
       ***** **** *
```

Homology is binary:
Either two sequences share a common ancestor, or they don't

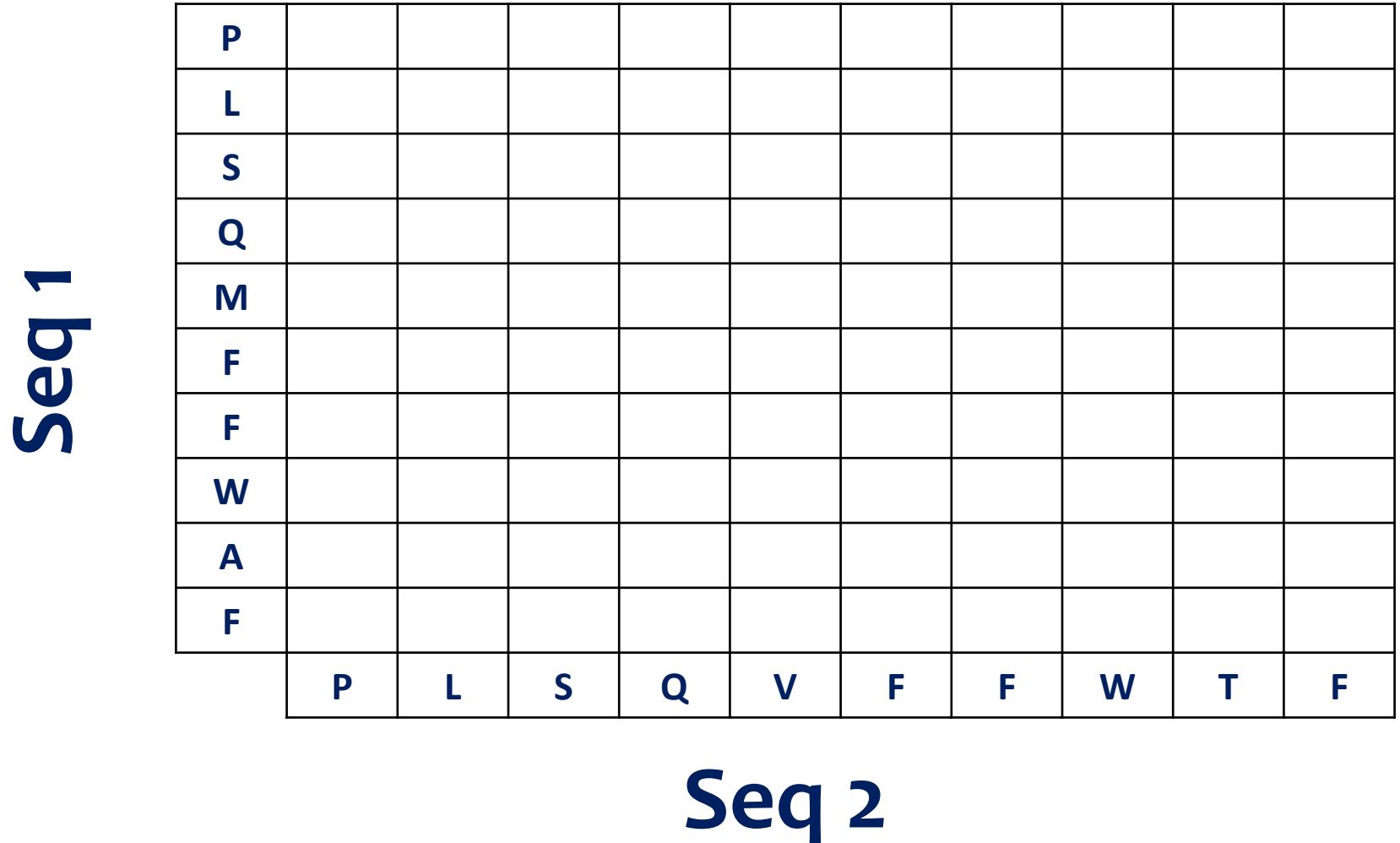
How similar is similar enough?

-Eyeball it

```
Seq1 – PLSQMFFWAF
Seq2 – PLSQVFFWTF
       ***** **** *
```

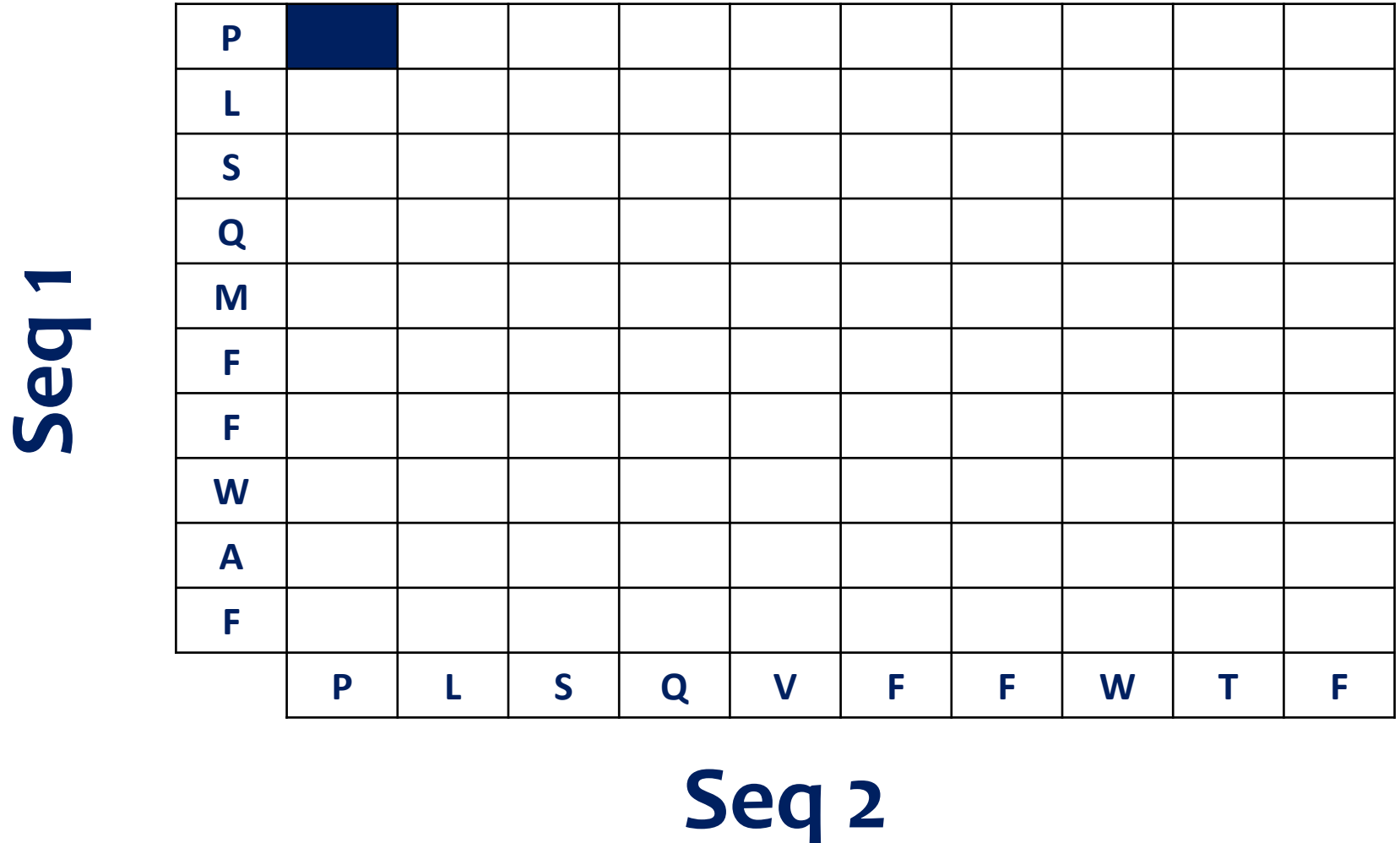
How similar is similar enough?

-Dot plots



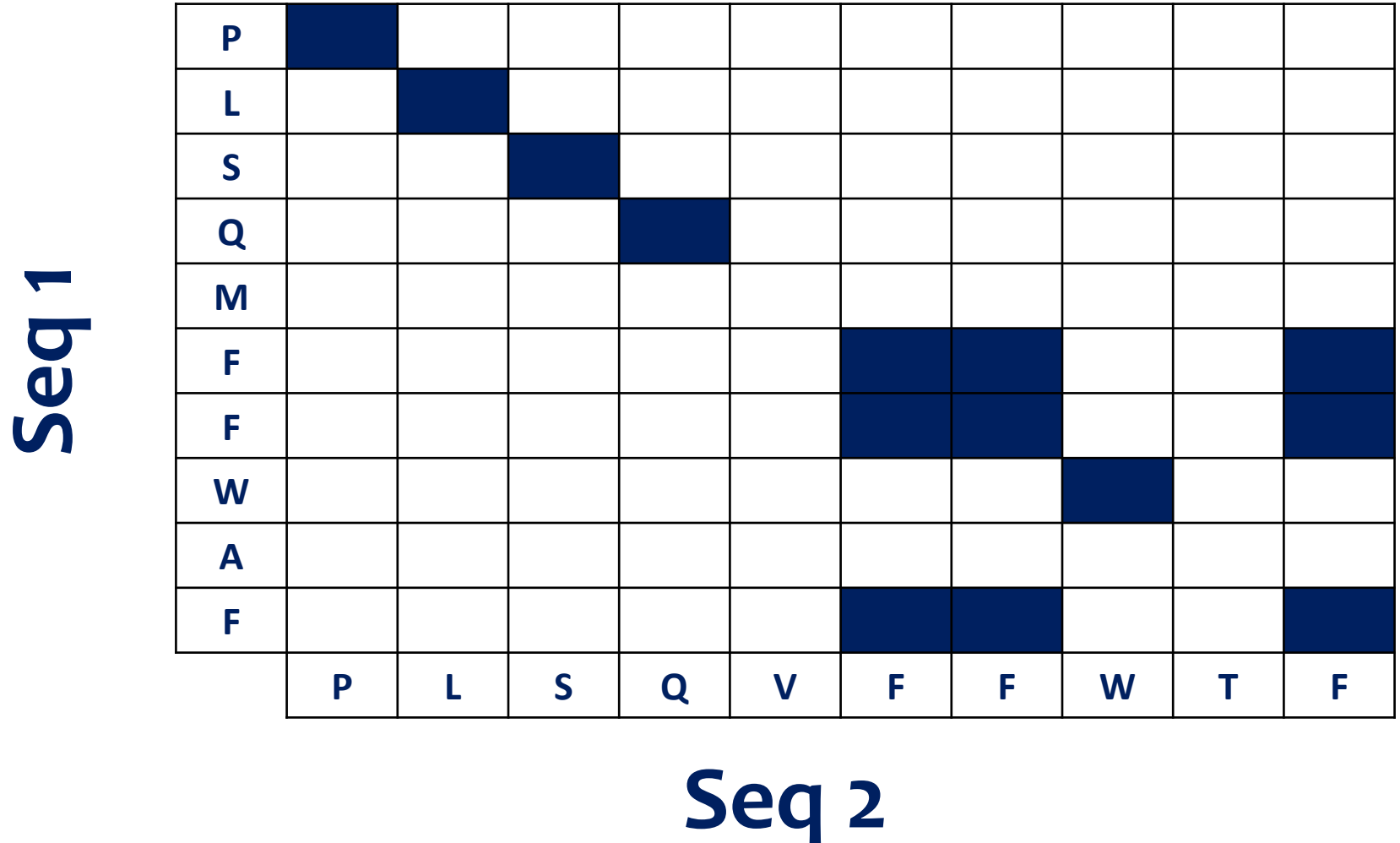
How similar is similar enough?

-Dot plots



How similar is similar enough?

-Dot plots



How similar is similar enough?

-Expect (E)-value

E-values represent the number of hits with a similar score **expected by random chance**

- Size of the database
- Length of the sequence

- E-values ≤ 10^{-5} are commonly considered "significant" hits

Similarity arising by random chance – nucleotide sequence

4 bases (A,C,G,T) –> probability of a random similarity at an individual site is **1/4 = 0.25**

Longer sequences are less likely to be similar due to random chance

Similarity arising by random chance – nucleotide sequence

4 bases (A,C,G,T) –> probability of a random similarity at an individual site is **1/4 = 0.25**

$P_{identity} = 0.25^{L}$

L = alignment length

L=12, $P_{identity} = 5.9 \times 10^{-8}$

L=100, $P_{identity} = 6.22 \times 10^{-61}$

Longer sequences are less likely to be similar due to random chance

Similarity arising by random chance – protein sequence

20 amino acids –> probability of a random similarity at an individual site is **1/20 = 0.05**

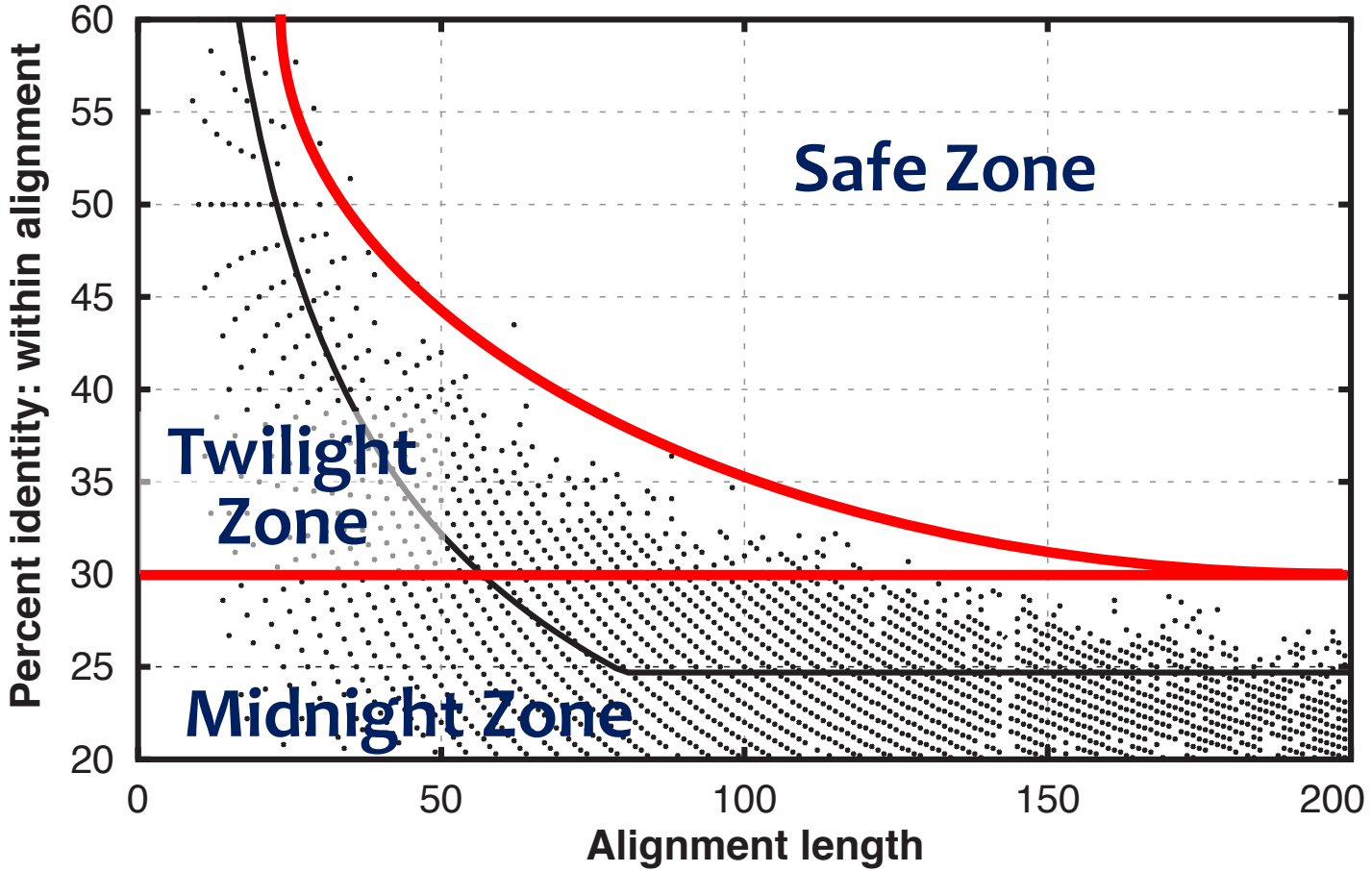
$P_{identity} = 0.05^{L}$

L = alignment length

L=10, $P_{identity} = 9.8 \times 10^{-14}$

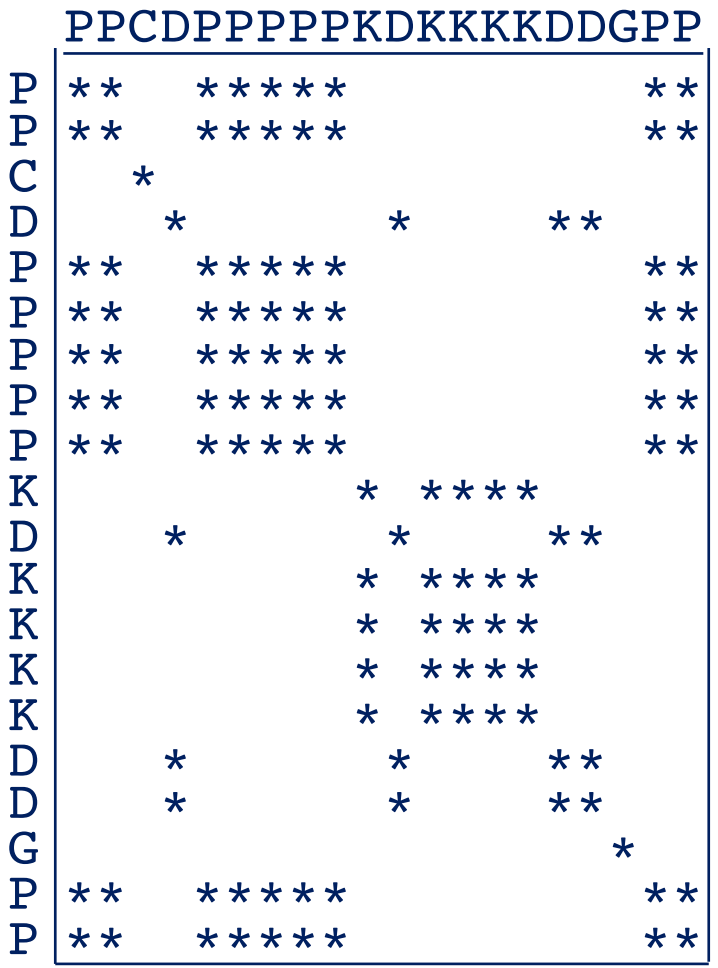
Protein sequences can detect lower degrees of similarity than nucleotide sequences

Similarity arising by random chance – protein sequence



Brenner et al., 1998 PNAS
Rost 1999 Prot. Eng.

Similarity arising by random chance – sequence complexity



Similarity arising by random chance – sequence complexity

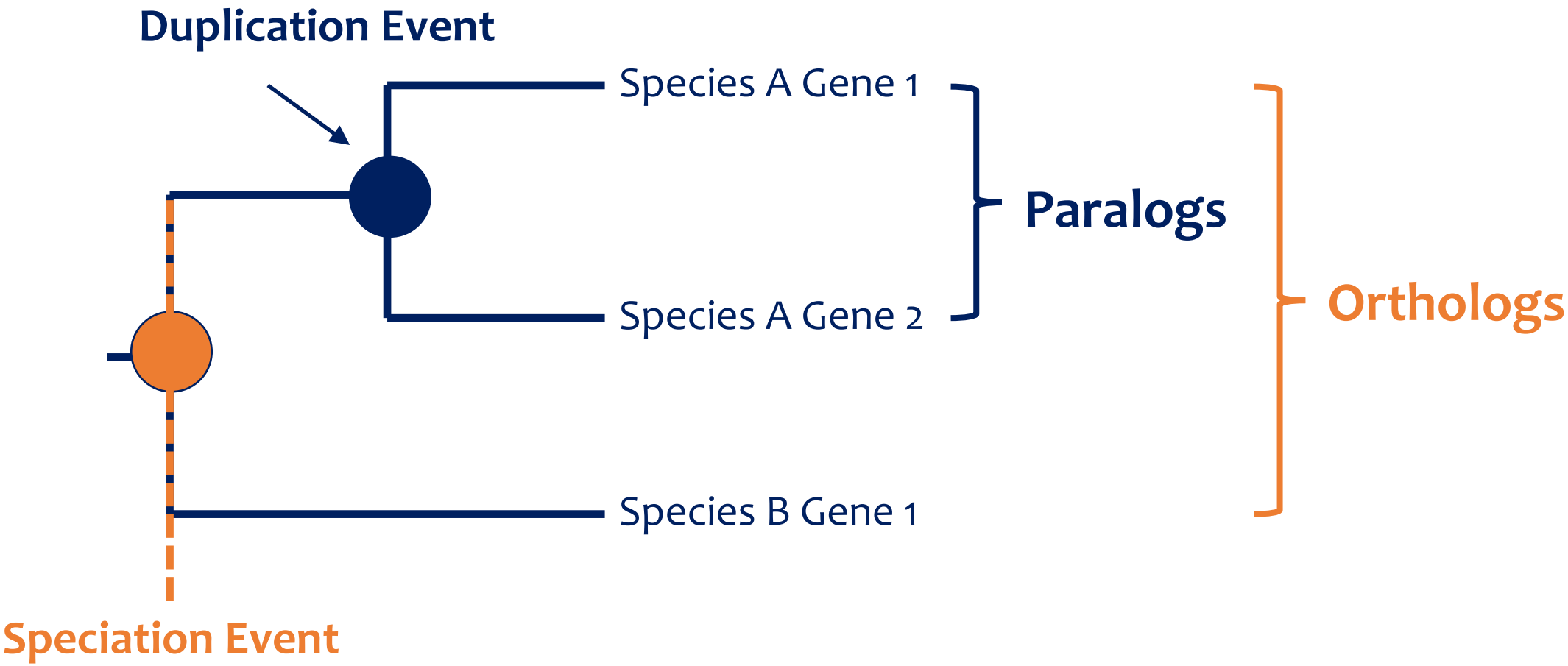
- **Low complexity sequences can give artefactually high similarity scores**
  - Mask low complexity sequences from search
  - Increase similarity threshold

Homologs: Orthologous or paralogous?

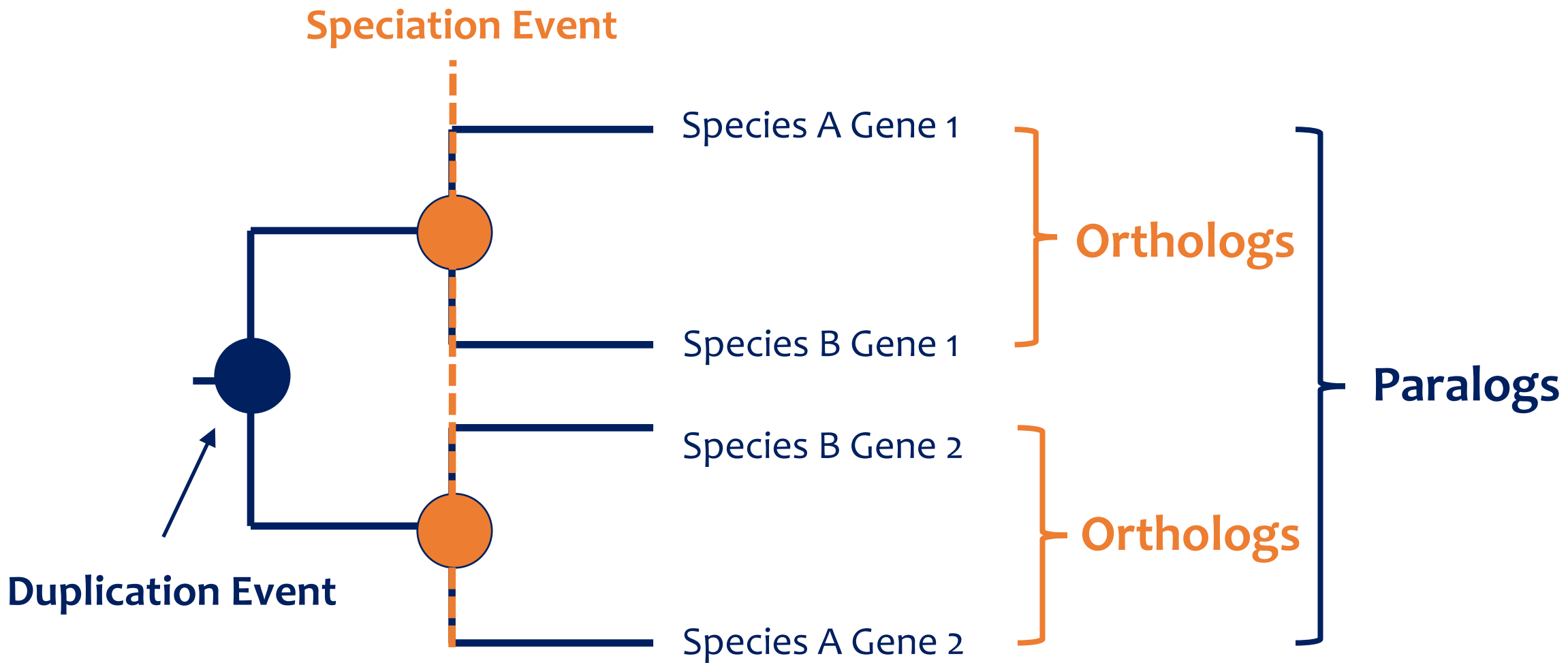
- Orthologs are homologous sequences that are descended from a speciation event

- Paralogs are homologous sequences that are descended from a duplication event

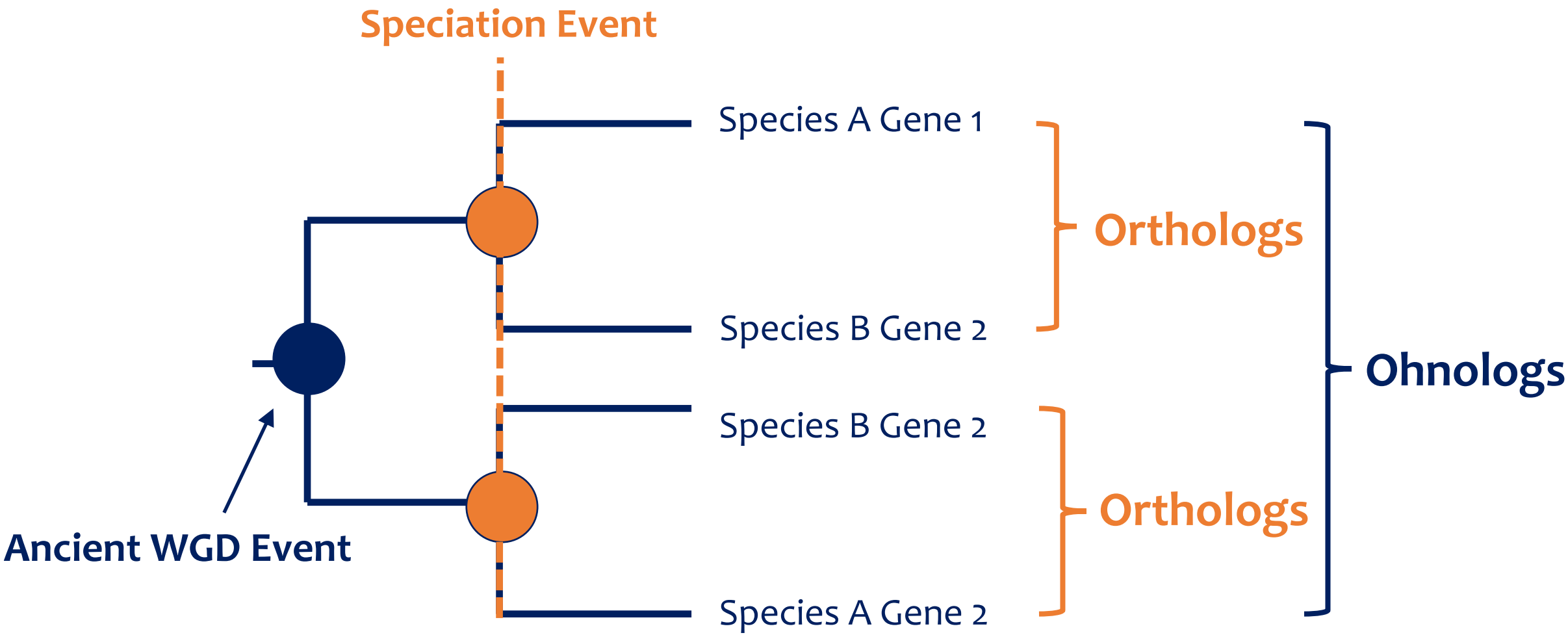
Homologs: Orthologous or paralogous?



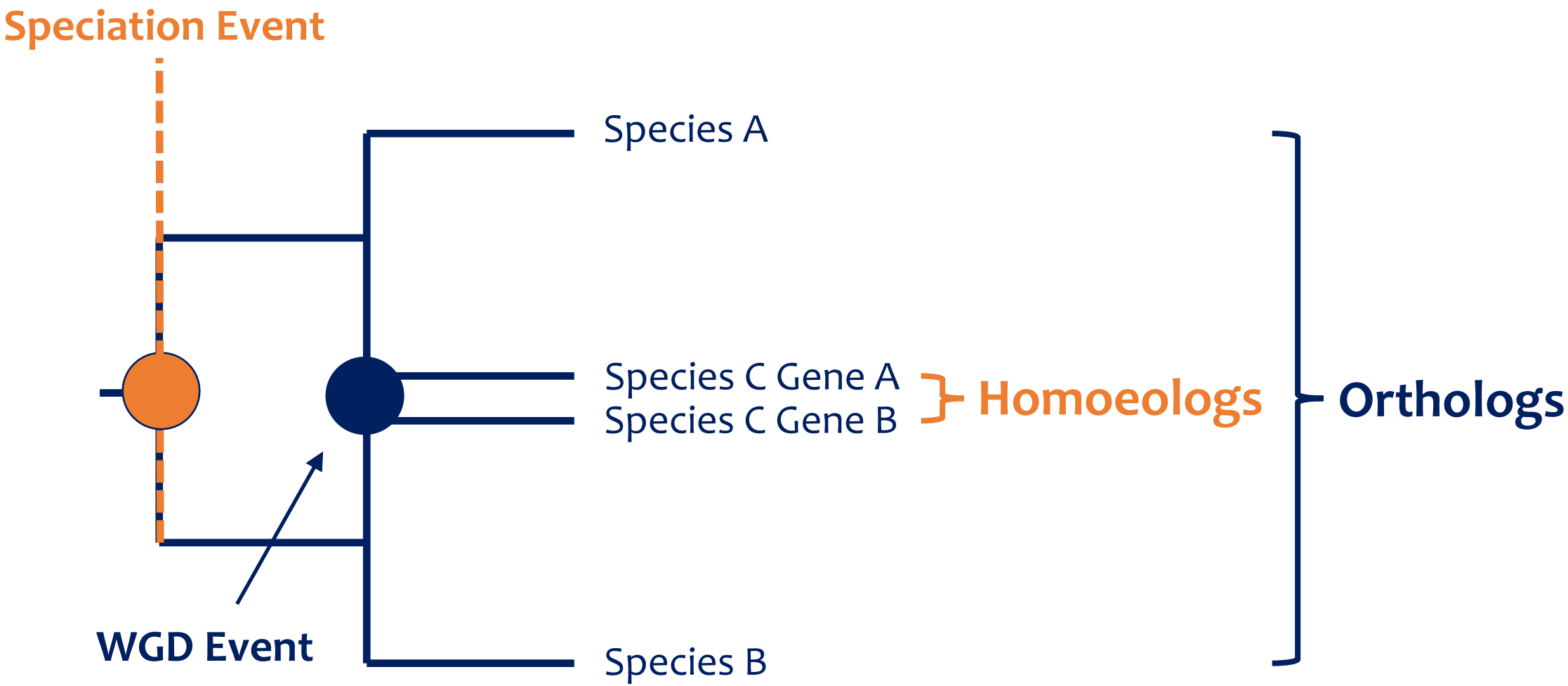
Homologs: Orthologous or paralogous?



Whole Genome Duplications



Hybrid Whole Genome Duplications



Homology can be assessed at different levels of organization

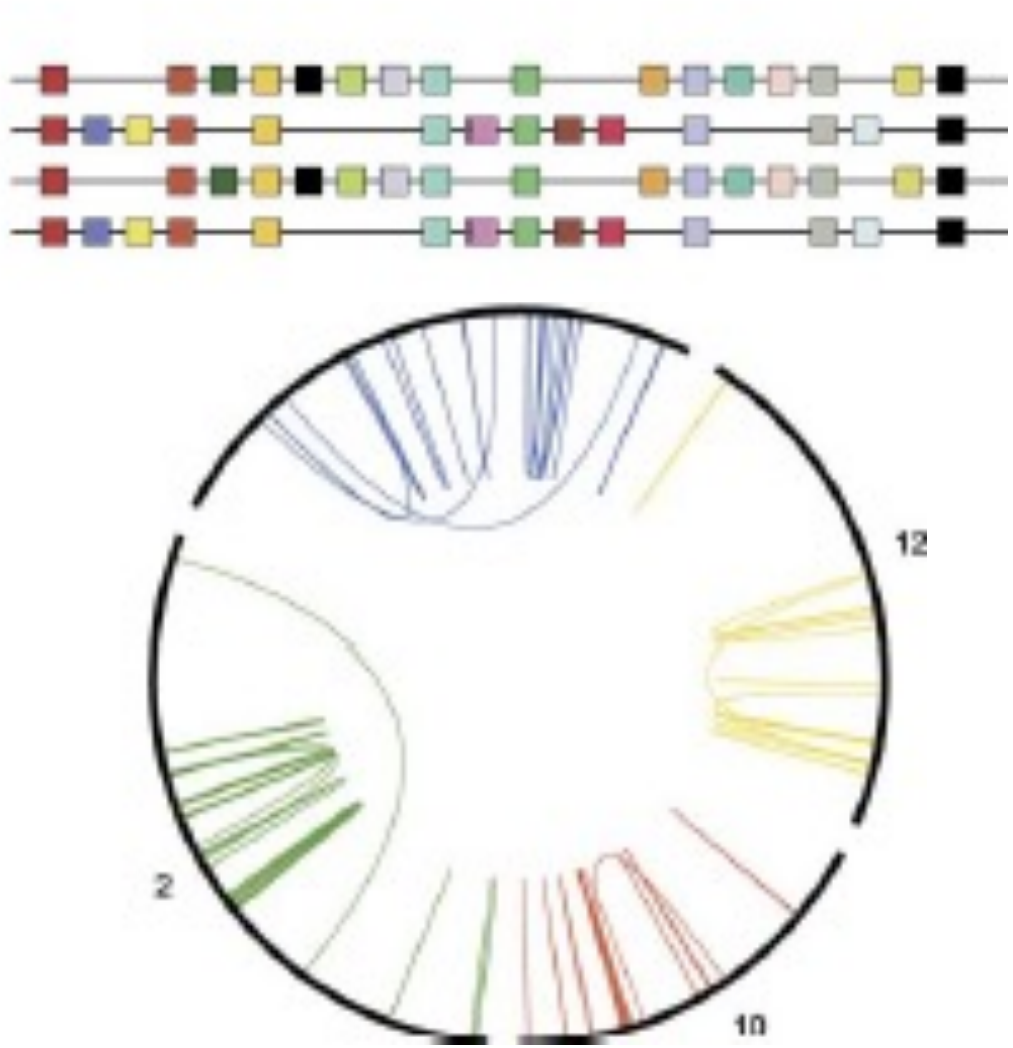
- Nucleotide/protein sequence
- Gene order (synteny)
- Phenotypes (e.g., morphology)

Understand models/mechanisms of evolution before assessing homology!

Syntenic homology

Gene order can evolve via:

- Structural rearrangements
- Transposition
- Duplication
- Deletion

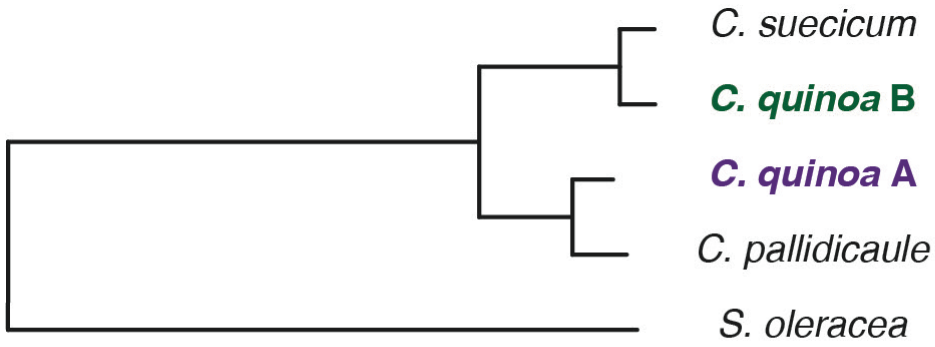


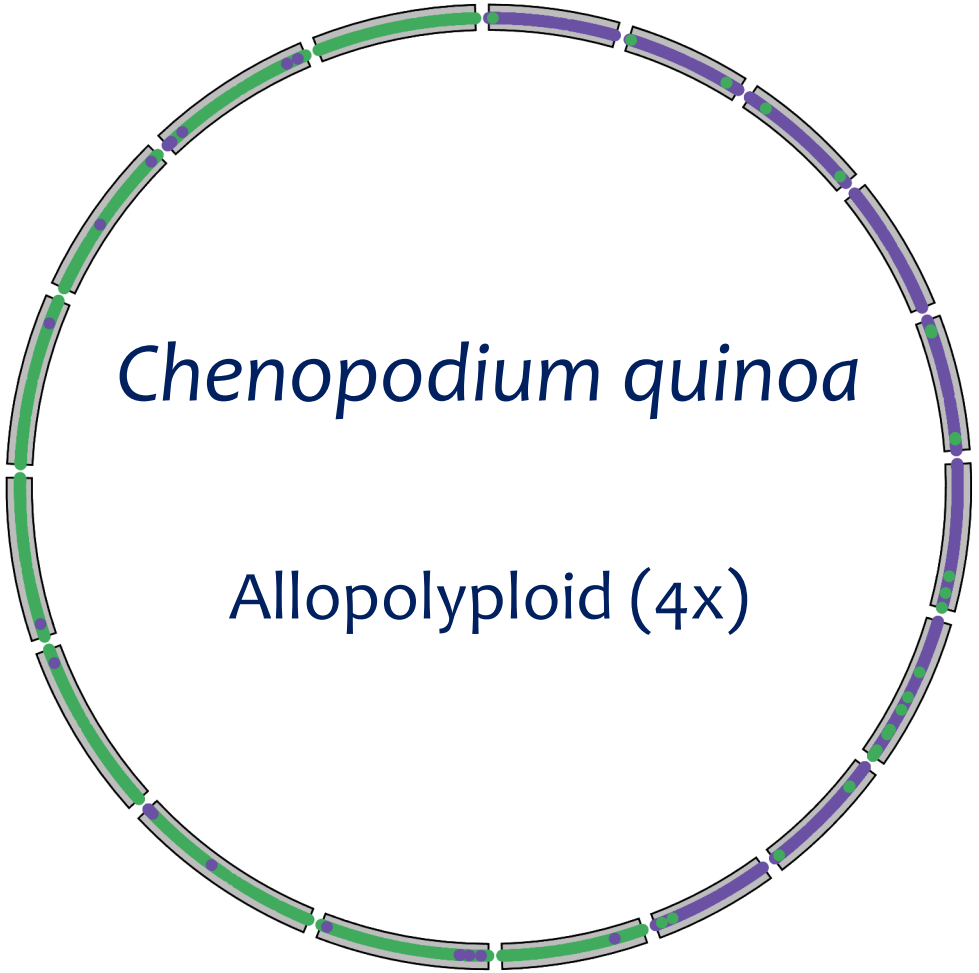
Dehar and Boore 2005 PLoS Biol

Syntenic homology

Gene order can evolve via:

- Structural rearrangements
- Transposition
- Duplication
- Deletion

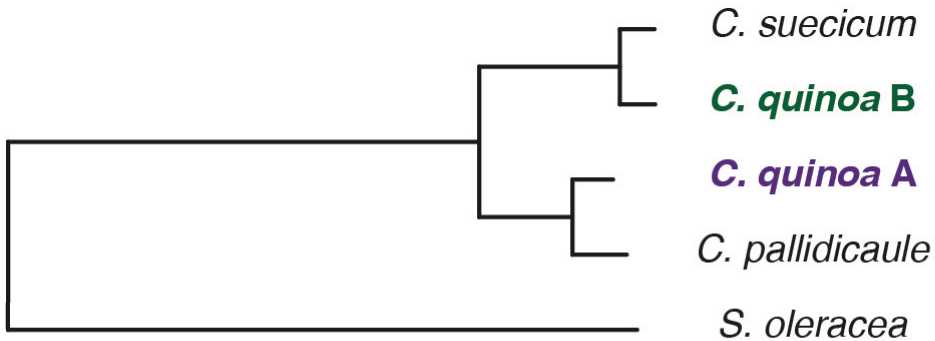


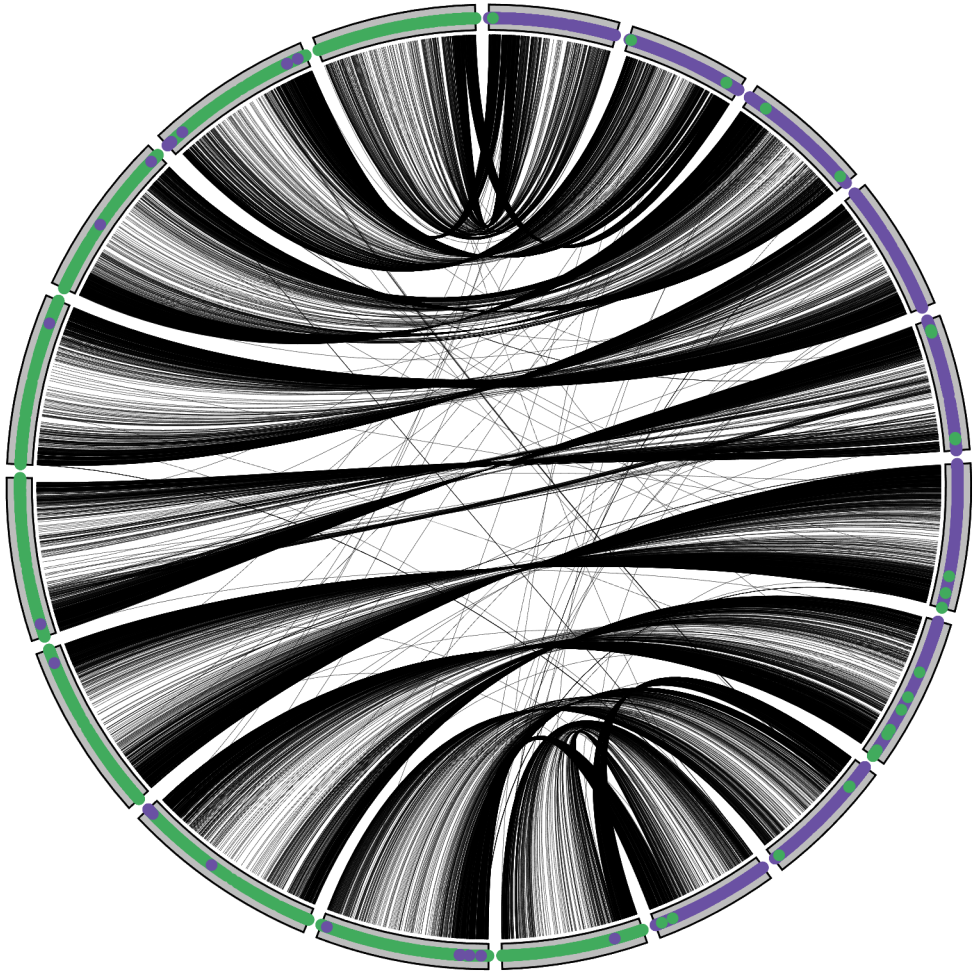


Syntenic homology

Gene order can evolve via:

- Structural rearrangements
- Transposition
- Duplication
- Deletion

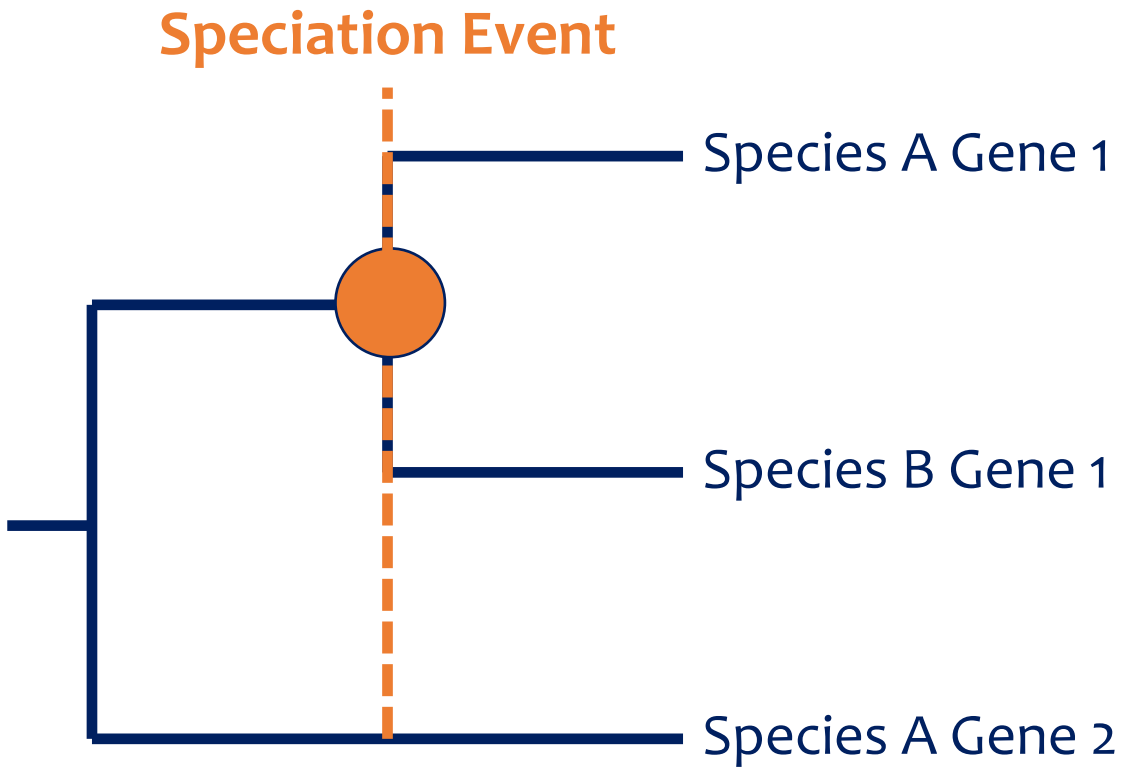




Chenopodium quinoa

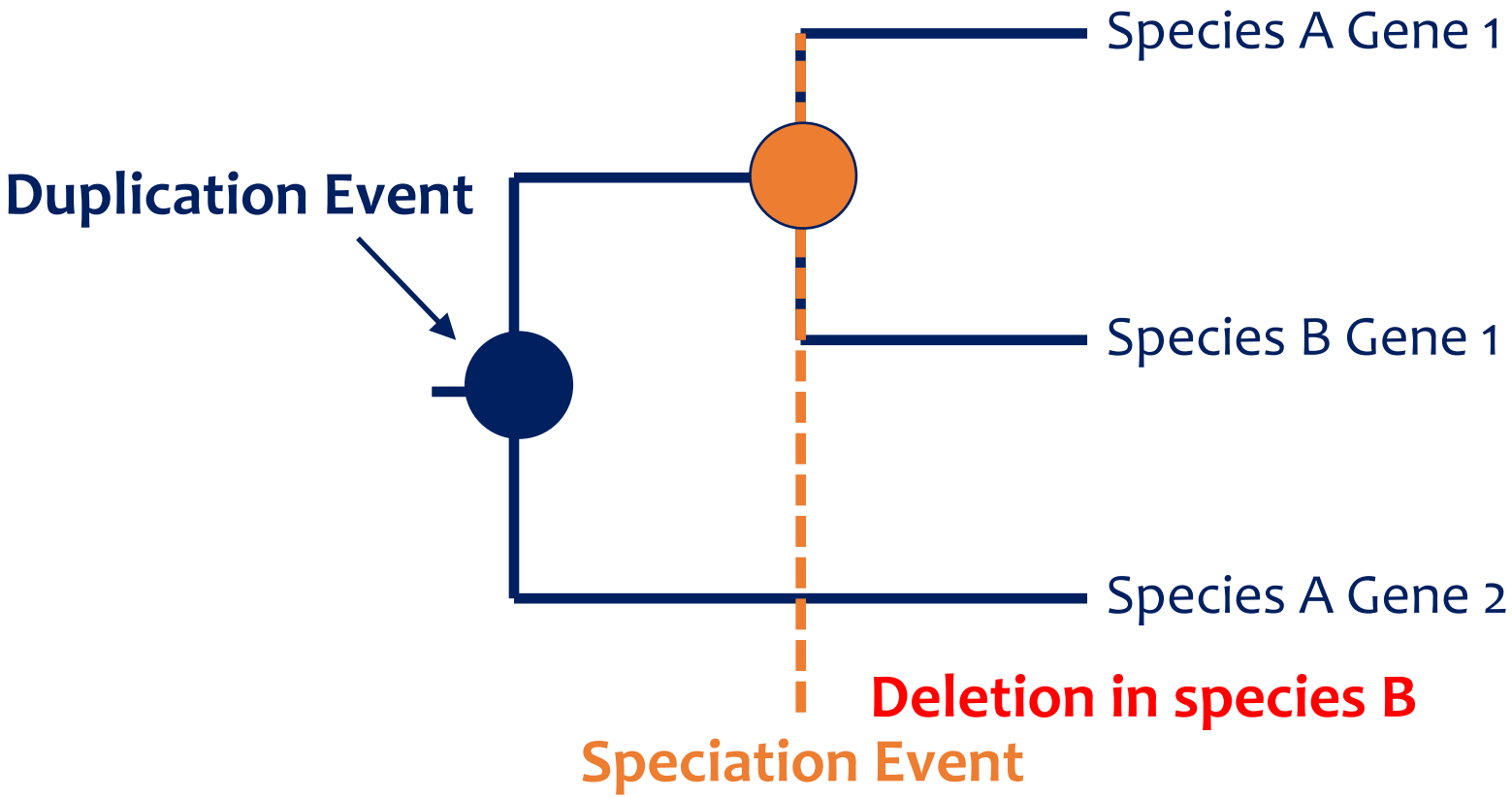
Problems in inferring homology

- Duplication and subsequent deletion



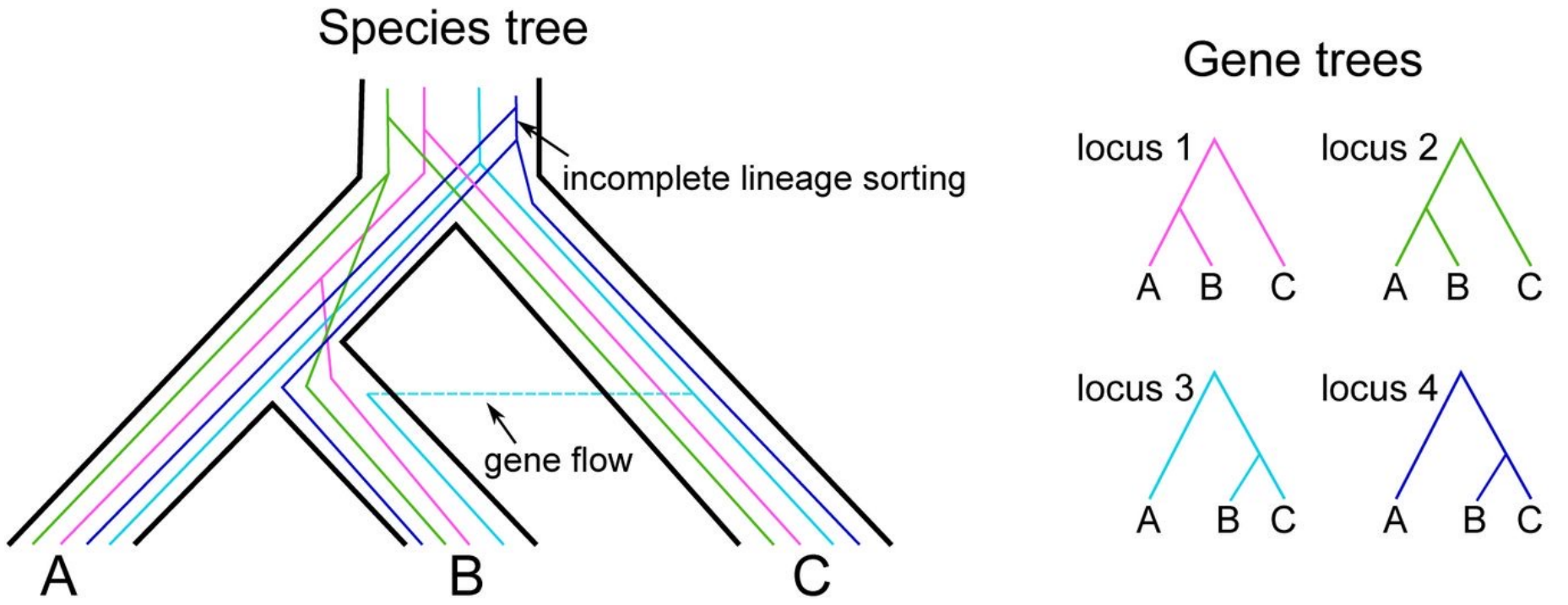
Problems in inferring homology

- Duplication and subsequent deletion



Problems in inferring homology

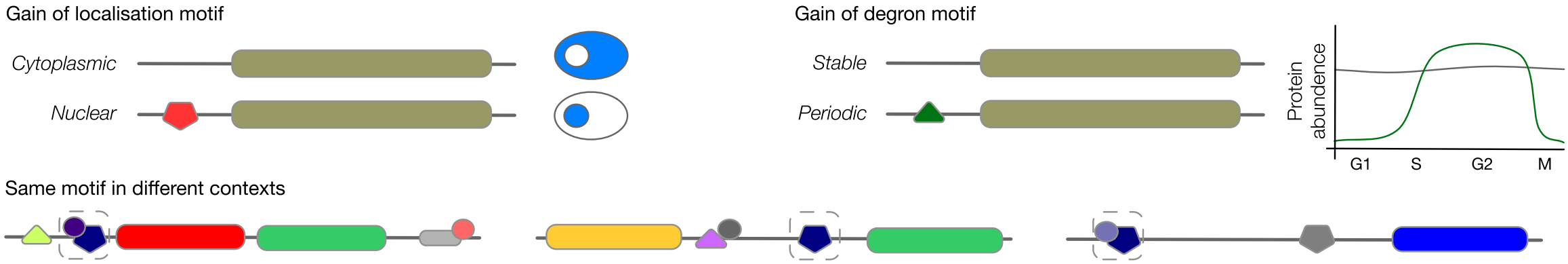
- **Introgression** and **incomplete lineage sorting** will result in different tree topologies



Marin et al 2020 JEB

Problems in inferring homology

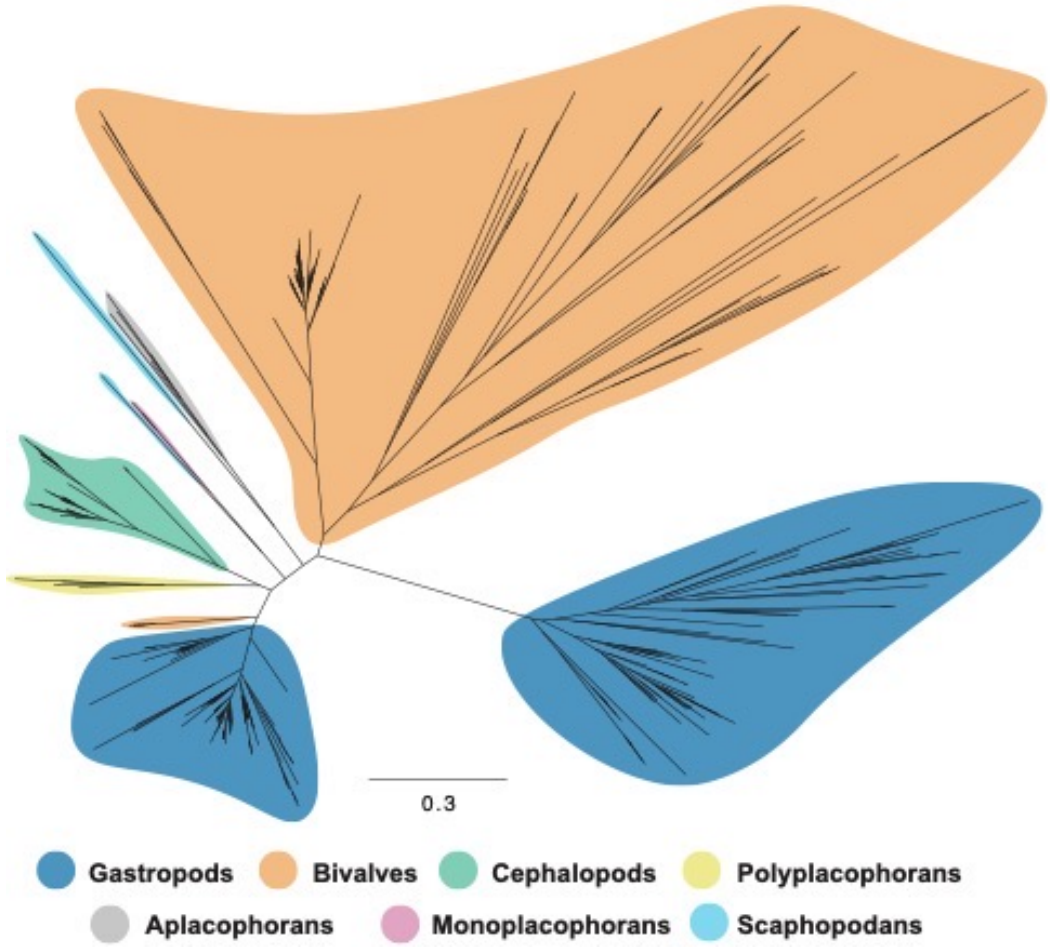
- Motif gain/loss



Roey and Davey 2015 Cell Comm

Problems in inferring homology

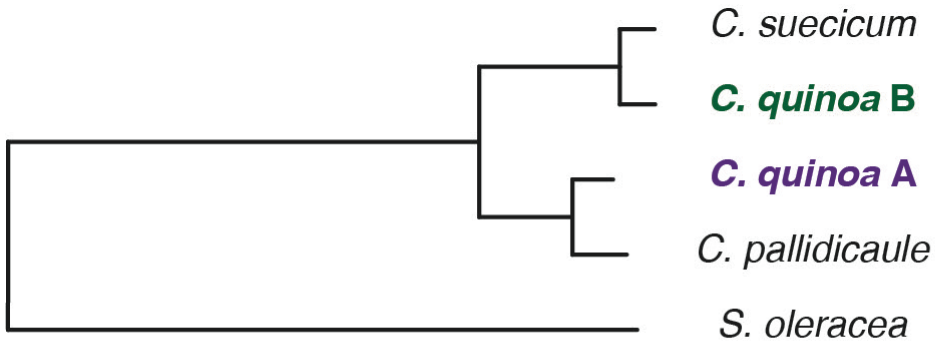
- Long branch attraction
  - Erroneous grouping of two or more long branches as sister groups due to methodological artifacts

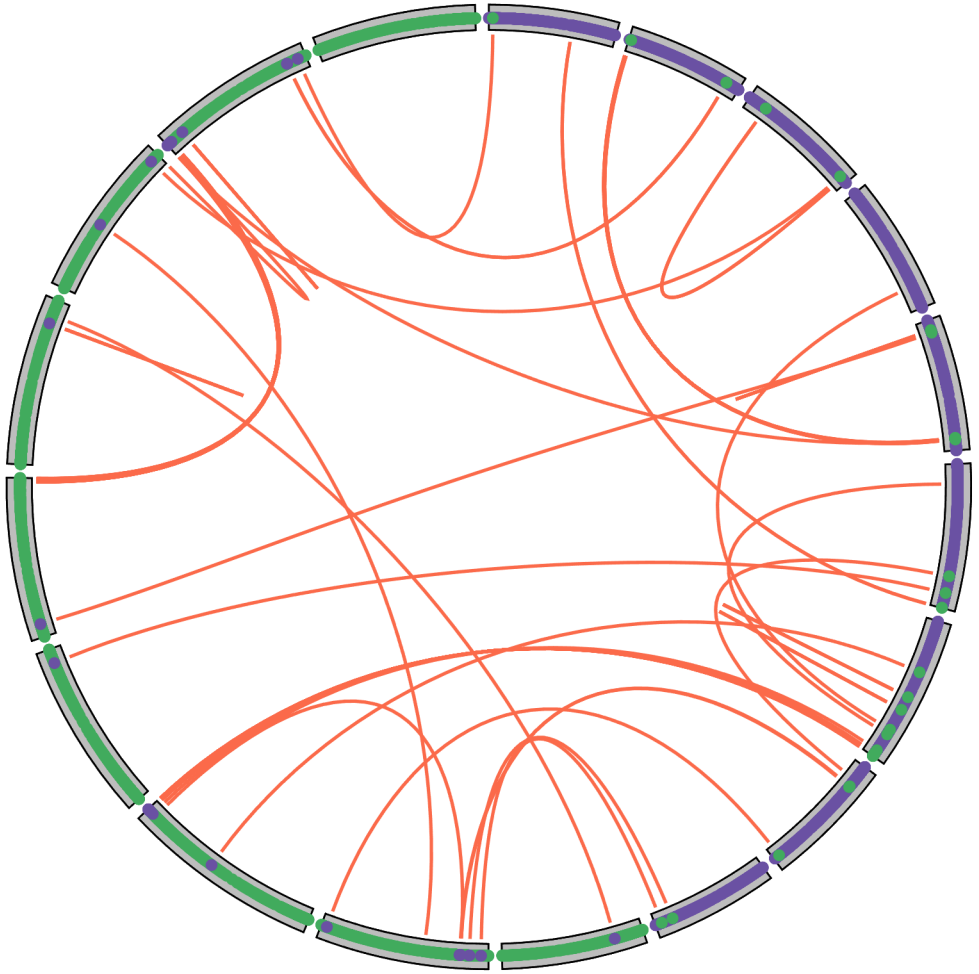


Felsenstein 1978 Syst Biol
Bergsten 2005 Cladistics

Problems in inferring homology

- Illegitimate recombination
  - Gene conversion
  - Structural rearrangements
  - Etc.





Chenopodium quinoa

How homologous are the two amino acid sequences?

```
Seq1 – PLSQMFFWAF
Seq2 – PLSQVFFWTF
```

How ~~homologous~~ similar are the two amino acid sequences? Are they homologous?

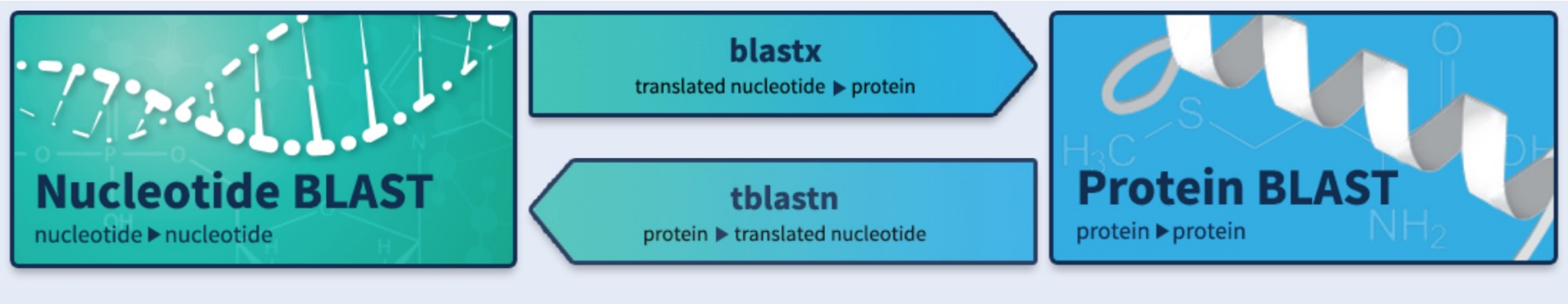
```
Seq1 – PLSQMFFWAF
Seq2 – PLSQVFFWTF
```

Make a dot plot of these two sequences.

What % similarity are they?

Based on that, are they are they homologous?

BLAST



Most important card in the Bioinformatician's deck

Basic Local Alignment Search Tool

1. Compiling list of high-scoring **words**
2. Scanning database for hits
3. **Word** extension

Words are *w*-mers of sequence

PLSQMFFWAF

***w* = 5**

Length = 10

PLSQM

***w*-mer**

 # words = Length – *w* + 1

Words are *w*-mers of sequence

PLSQMFFWAF

w = 5

Length = 10

PLSQM LSQMF

w-mers

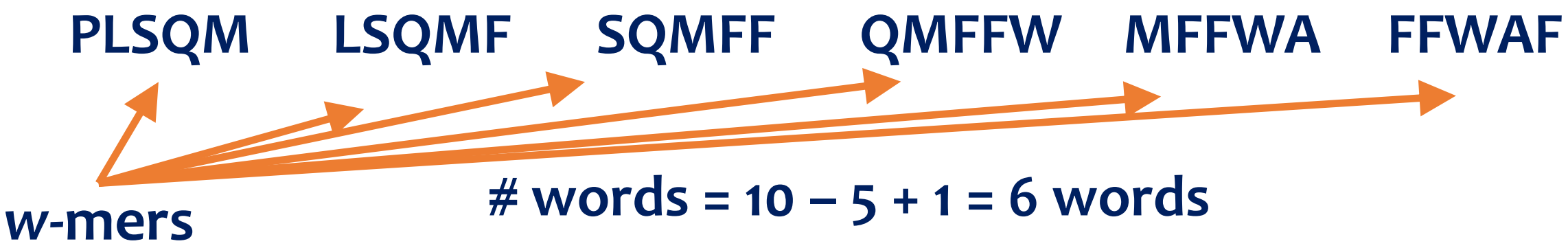
words = Length – *w* + 1

Words are *w*-mers of sequence

PLSQMFFWAF

***w* = 5**

Length = 10



Words are *w*-mers of sequence

w-mer = *k*-mer

Compiling list of high-scoring words

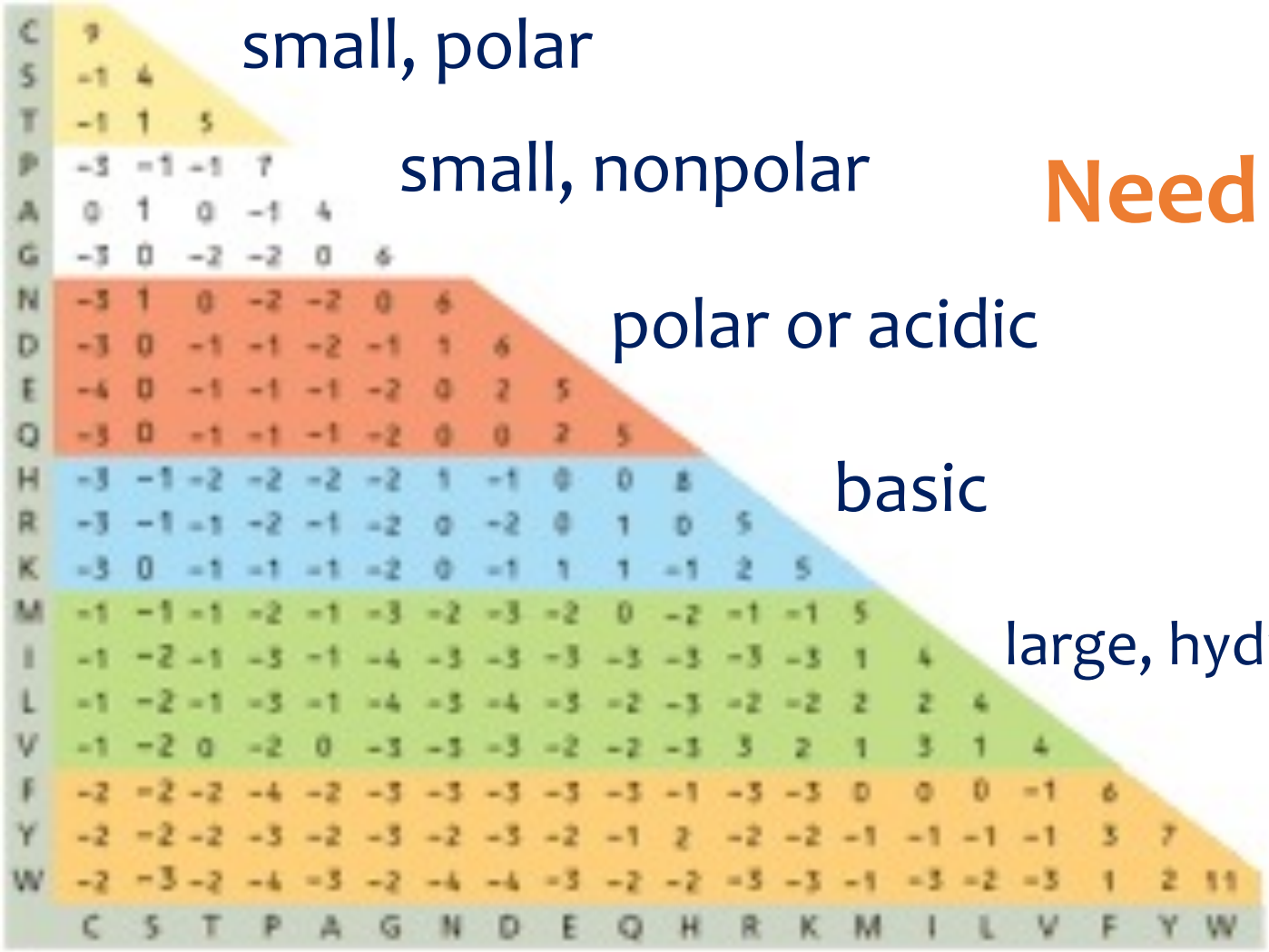
Scan for words heuristically

Calculate scores

High scoring pairs with scores > threshold *T* *kept*

Compiling list of high-scoring words

PAM-120



small, polar

small, nonpolar

polar or acidic

basic

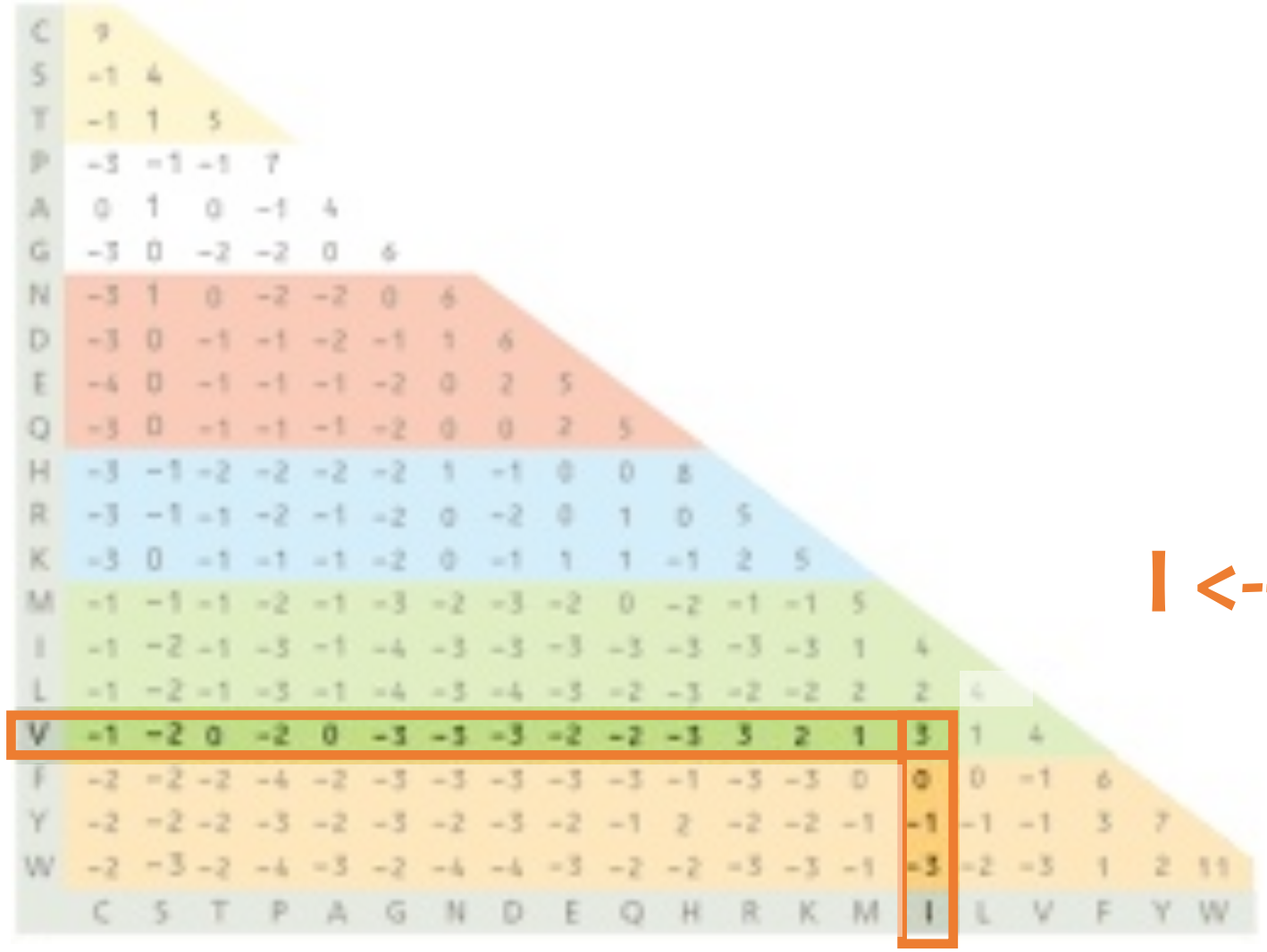
large, hydrophobic

aromatic

Need a scoring system

Compiling list of high-scoring words

PAM-120



I <--> V = 3

Scan database for hits

Computationally intensive part

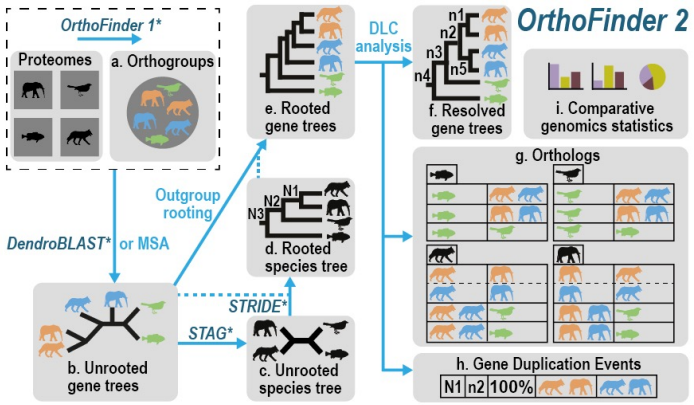
Uses only HSPs from previous step
to reduce the search space

Word extension

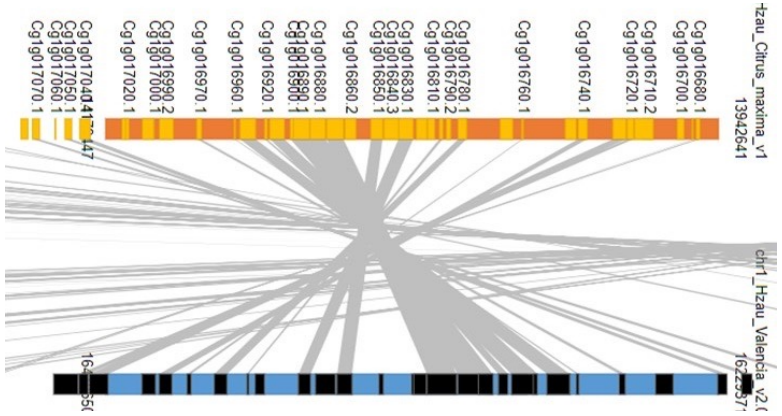
Extend 1 character a time in both directions, re-compute score

If score is lower than the minimal score computed so far, extension in that direction stops

Tools that rely on BLAST



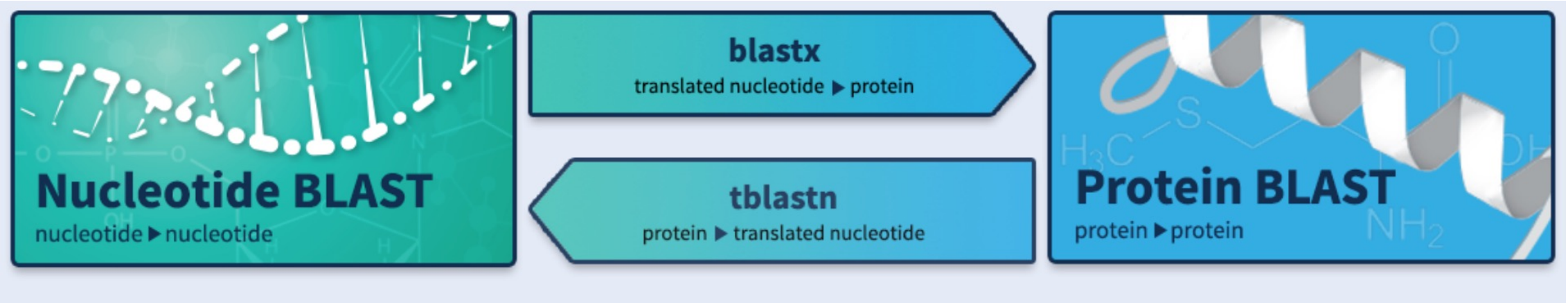
Orthofinder



MCScanX

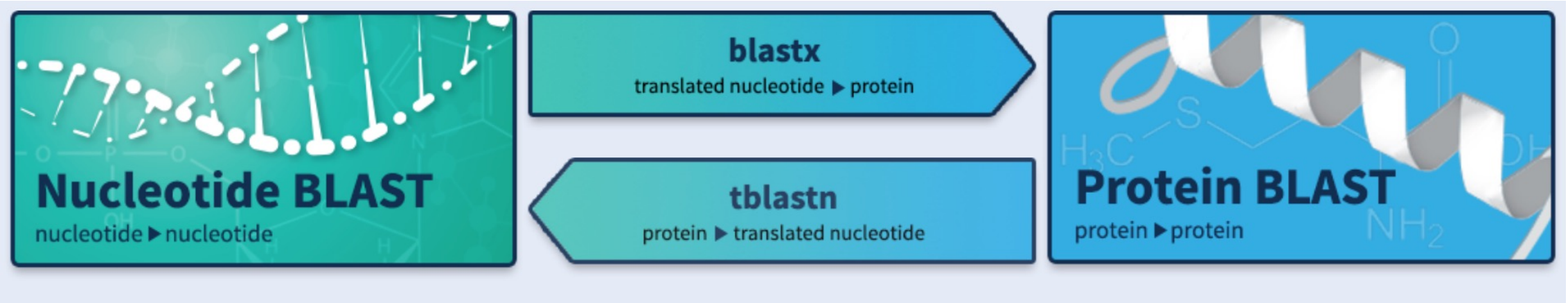
Lots of others, and the BLAST manual can help you implement BLAST in your program

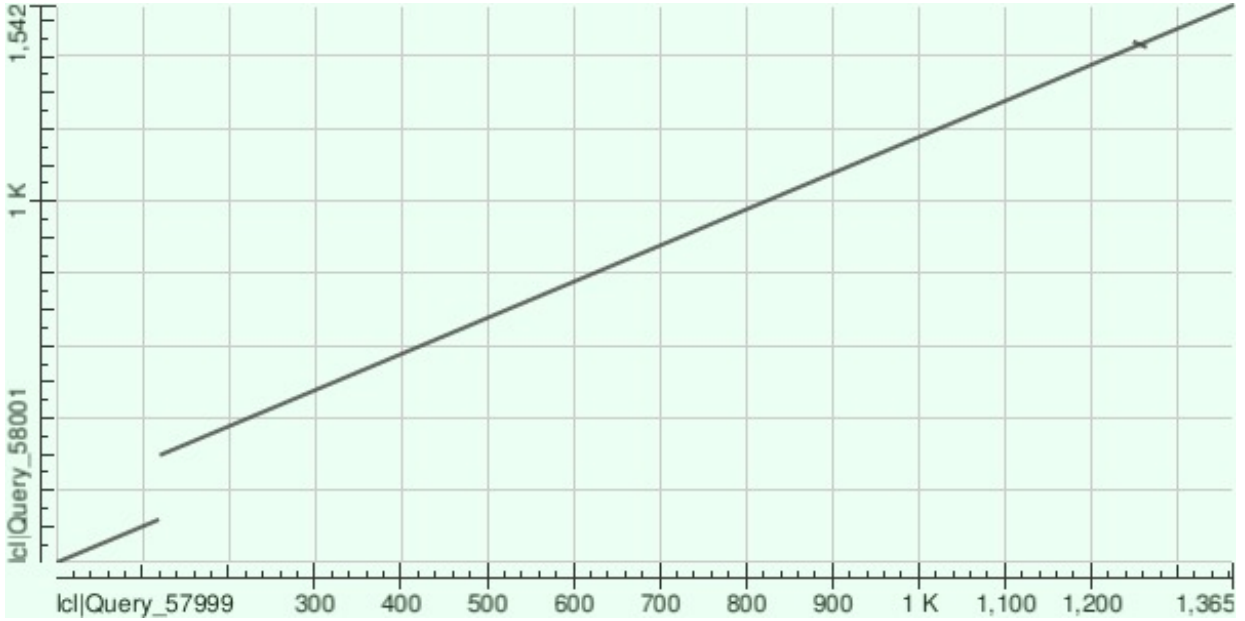
GUI vs. Command line




```
blastn [-h] [-help] [-import_search_strategy filename]
  [-export_search_strategy filename] [-task task_name] [-db database_name]
  [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
  [-negative_gilist filename] [-negative_seqidlist filename]
  [-entrez_query entrez_query] [-db_soft_mask filtering_algorithm]
  [-db_hard_mask filtering_algorithm] [-subject subject_input_file]
  [-subject_loc range] [-query input_file] [-out output_file]
  [-evalue evalue] [-word_size int_value] [-gapopen open_penalty]
  [-gapextend extend_penalty] [-perc_identity float_value]
  [-qcov_hsp_perc float_value] [-max_hsps int_value]
  [-xdrop_ungap float_value] [-xdrop_gap float_value]
  [-xdrop_gap_final float_value] [-searchsp int_value]
  [-sum_stats bool_value] [-penalty penalty] [-reward reward] [-no_greedy]
  [-min_raw_gapped_score int_value] [-template_type type]
  [-template_length int_value] [-dust DUST_options]
```

GUI vs. Command line





GUI vs. Command line

```
  blastn [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-negative_seqidlist filename]
    [-entrez_query entrez_query] [-db_soft_mask filtering_algorithm]
    [-db_hard_mask filtering_algorithm] [-subject subject_input_file]
    [-subject_loc range] [-query input_file] [-out output_file]
    [-evalue evalue] [-word_size int_value] [-gapopen open_penalty]
    [-gapextend extend_penalty] [-perc_identity float_value]
    [-qcov_hsp_perc float_value] [-max_hsps int_value]
    [-xdrop_ungap float_value] [-xdrop_gap float_value]
    [-xdrop_gap_final float_value] [-searchsp int_value]
    [-sum_stats bool_value] [-penalty penalty] [-reward reward] [-no_greedy]
    [-min_raw_gapped_score int_value] [-template_type type]
    [-template_length int_value] [-dust DUST_options]
    [-filtering_db filtering_database]
    [-window_masker_taxid window_masker_taxid]
    [-window_masker_db window_masker_db] [-soft_masking soft_masking]
    [-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
    [-best_hit_score_edge float_value] [-window_size int_value]
    [-off_diagonal_range int_value] [-use_index boolean] [-index_name string]
    [-lcase_masking] [-query_loc range] [-strand strand] [-parse_deflines]
    [-outfmt format] [-show_gis] [-num_descriptions int_value]
    [-num_alignments int_value] [-line_length line_length] [-html]
    [-max_target_seqs num_sequences] [-num_threads int_value] [-remote]
    [-version]

DESCRIPTION
   Nucleotide-Nucleotide BLAST 2.7.1+
```

Next up: Pairwise Sequence Alignment

Homework #1 (BLASTology) Posted on Canvas

Please Read Altschul et al., 1990