Comparative Genomics

Some Definitions

Anne Lopes

November - 2020

Majors events in genome evolution

macro

scales

micro

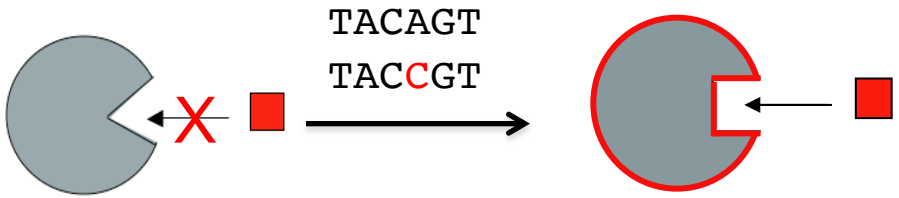
mechanisms

species:
- speciation
- extinction

DNA fragment:
- duplication
- translocation
- fusion/fission

gene:
- duplication
- fusion/fission
- HGT
- loss
- de novo emergence

nucleotides:
- substitution/insertion/deletions



Majors events in genome evolution

macro

scales

micro

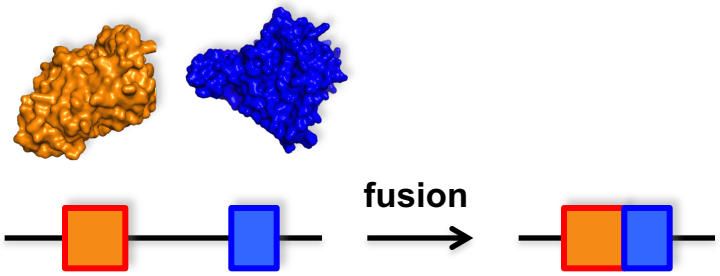
mechanisms

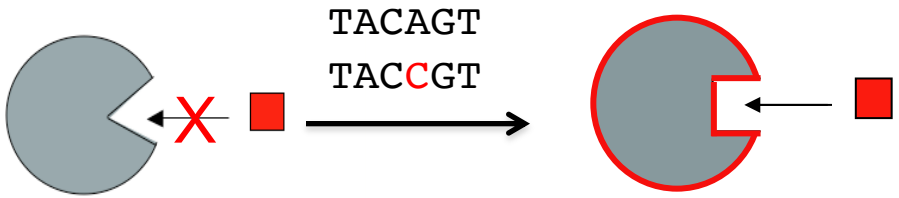
species:
- speciation
- extinction

DNA fragment:
- duplication
- translocation
- fusion/fission

gene:
- duplication
- fusion/fission
- HGT
- loss
- de novo emergence

nucleotides:
- substitution/insertion/deletions





Majors events in genome evolution

macro

scales

micro

mechanisms

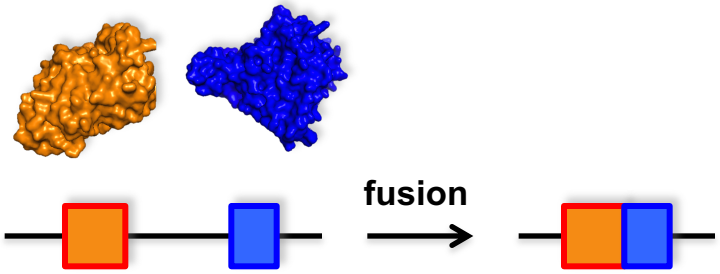
species:
- speciation
- extinction

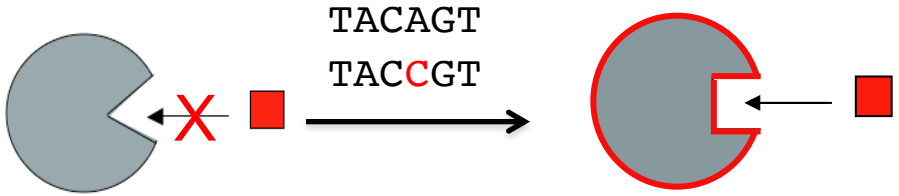
DNA fragment:
- duplication
- translocation
- fusion/fission

gene:
- duplication
- fusion/fission
- HGT
- loss
- de novo emergence

nucleotides:
- substitution/insertion/deletions

WANTED





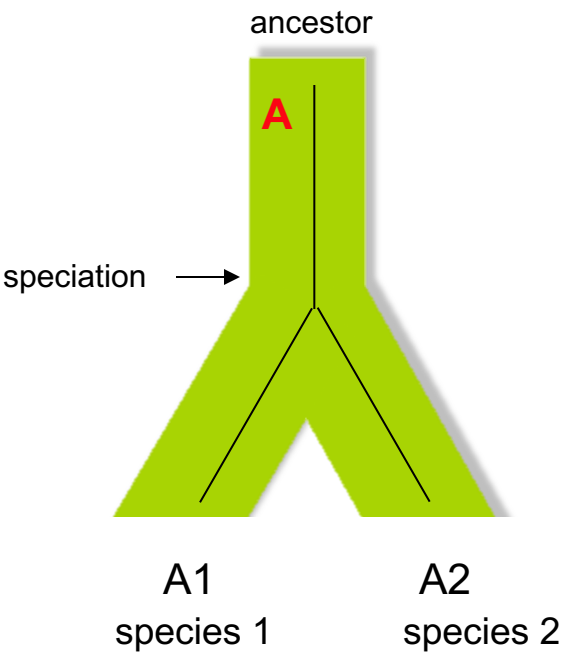
Orthologs/paralogs
D e f i n i t i o n s

Homologs: genes resulting from a common ancestor (including orthologs and paralogs)

Orthologs: genes resulting from a speciation event (i.e. genes in different species evolved from a common ancestral gene). Usually expected to share the same function, though it not always the case.

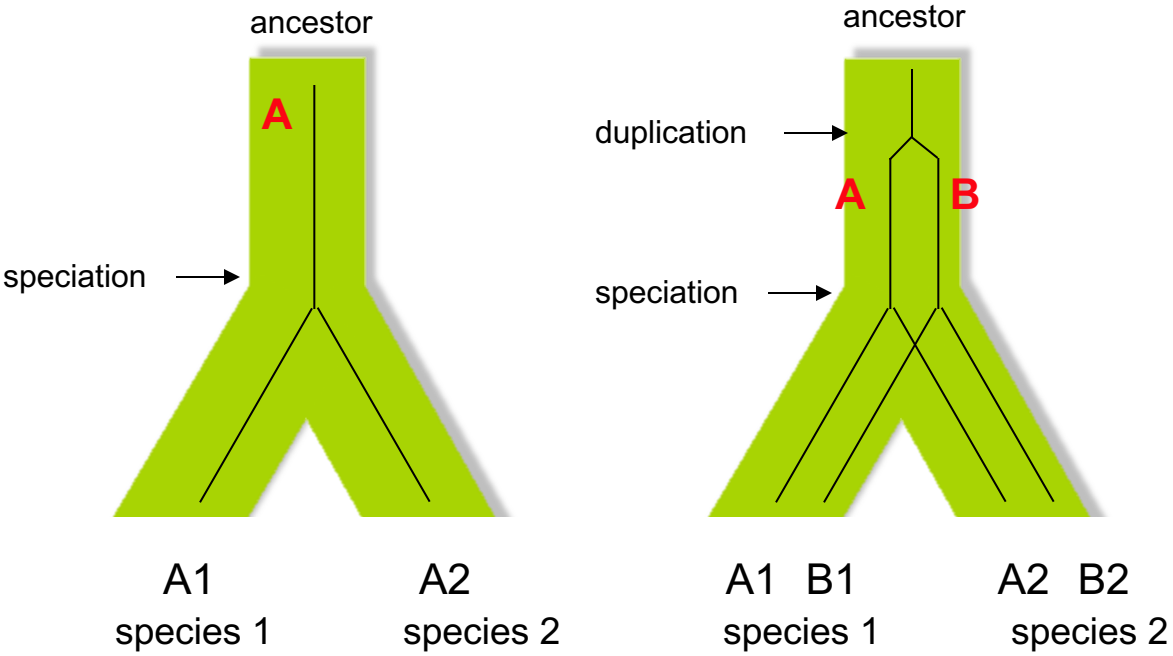
Paralogs: genes resulting from a duplication event (i.e. copies in the same genome or between two genomes). Usually, paralogous genes are associated with different (though generally related) functions since the selective pressure operates on only one copy, the other is therefore free to sample new sequence spaces.

Orthologs and Paralogs



orthologs
(A1-A2)

Orthologs and Paralogs

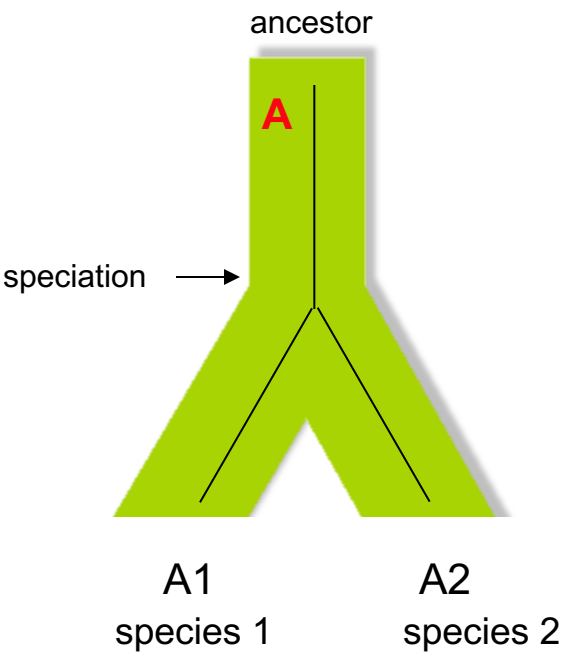


paralogs
(A1-B1), (A1-B2),(A2-B2)
and (B1-A2)

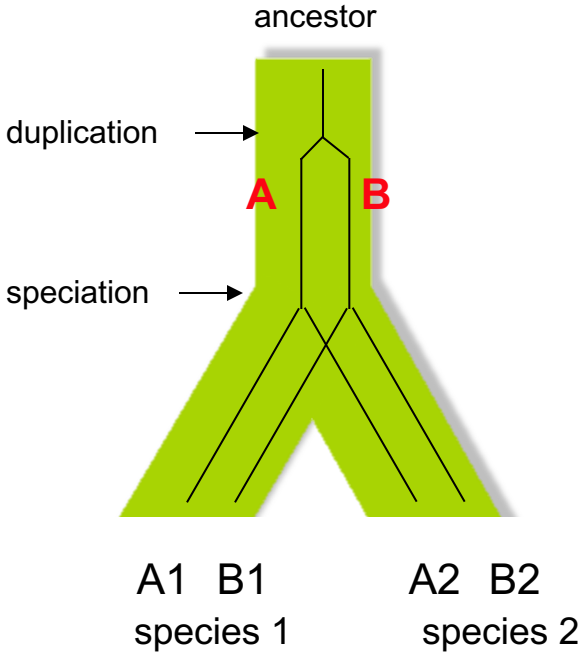
orthologs
(A1-A2)

orthologs
(A1-A2) and (B1-B2)

Orthologs and Paralogs

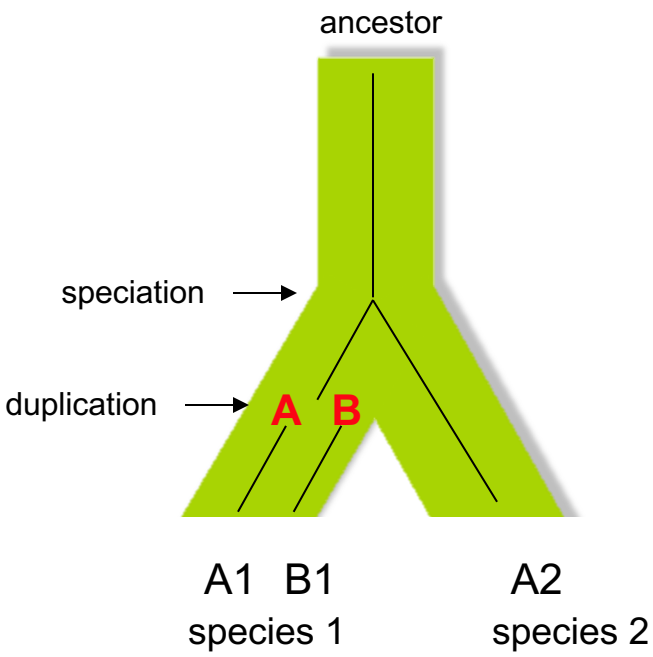


orthologs
(A1-A2)



paralogs
(A1-B1), (A1-B2),(A2-B2)
and (B1-A2)

orthologs
(A1-A2) and (B1-B2)



in-paralogs
(A1-B1)

orthologs
(A1/B1-A2)

Identification

Many methods!

Review

Cell PRESS

The quest for orthologs: finding the corresponding gene across genomes

Arnold Kuzniar[1], Roeland C.H.J. van Ham[1], Sándor Pongor[2] and Jack A.M. Leunissen[1]

[1] Laboratory of Bioinformatics, Wageningen University, Dreijenlaan 3, 6703 HA Wageningen, the Netherlands
[2] Protein Structure and Bioinformatics, International Centre for Genetic Engineering and Biotechnology, AREA Science Park, Padriciano 99, 34012 Trieste, Italy

Orthology is a key evolutionary concept in many areas of genomic research. It provides a framework for subjects as diverse as the evolution of genomes, gene functions, cellular networks and functional genome annotation. Although orthologous proteins usually perform equivalent functions in different species, establishing true orthologous relationships requires a phylogenetic approach, which combines both trees and graphs (networks) using reliable species phylogeny and available genomic data from more than two species, and an insight into the processes of molecular evolution. Here, we evaluate the available bioinformatics tools and provide a set of guidelines to aid researchers in choosing the most appropriate tool for any situation.

Homology: refers to a testable hypothesis that characters in different species sharing significant sequence similarity (at least 30–35% as a rule of thumb for protein sequences) descend from a single common ancestral character. Sequences that are evolutionarily related to each other in this way are known as homologs. Note that homology is independent of the size and molecular nature of a biological sequence.

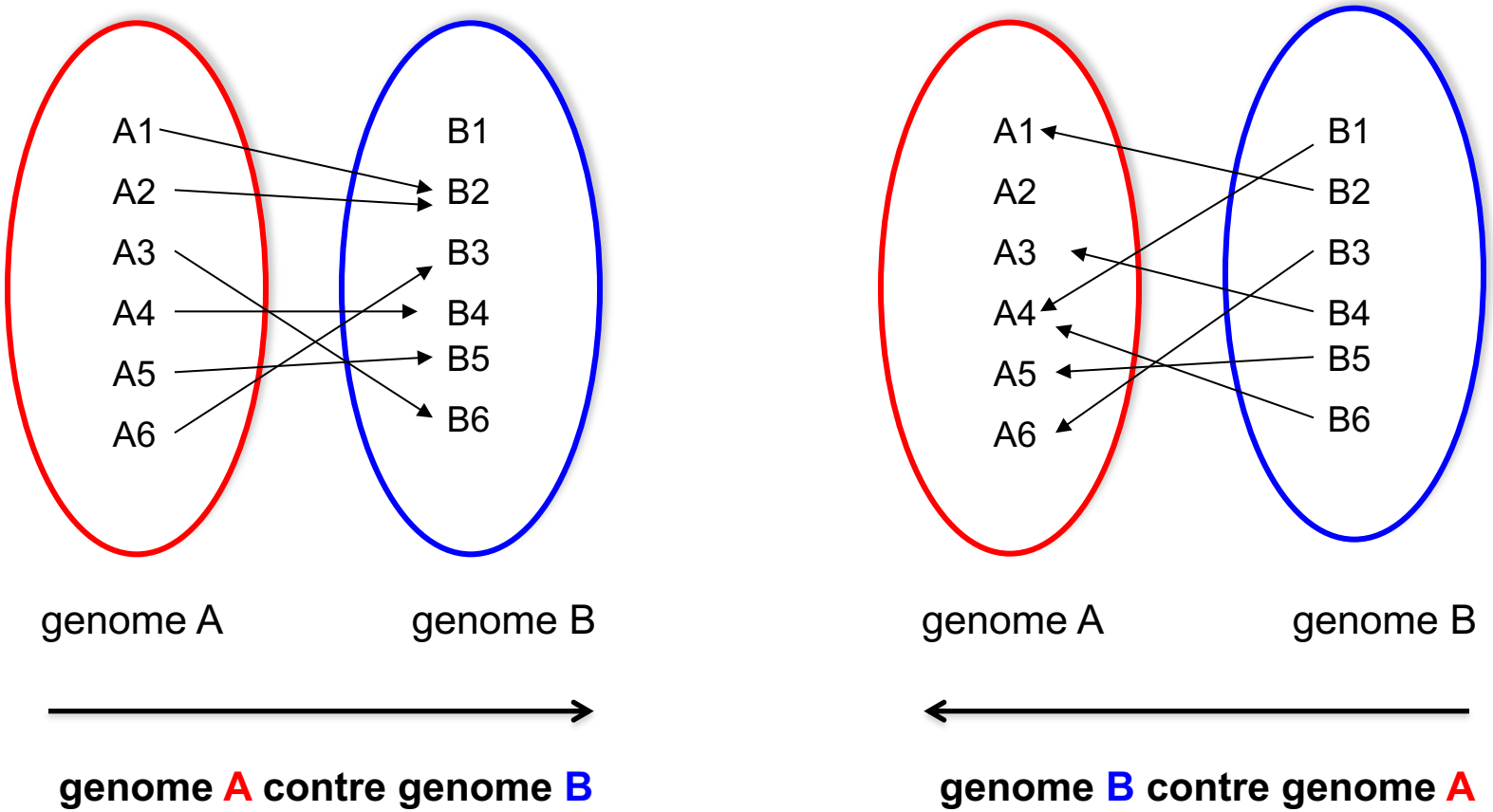
Horizontal gene transfer (HGT): an evolutionary process that involves transfer of genetic material between species but does not follow the vertical descent from a parental lineage to its offspring. HGT is an important phenomenon in the evolution of prokaryotes and eukaryotes [66–68].

In-paralogs: paralogs that result from a lineage-specific duplication(s) subsequent to a given speciation event (sometimes termed 'recent' paralogs). They are likely to have retained similar functions within a species.

Graph-based methods

BBH (Bidirectional Best Hits)

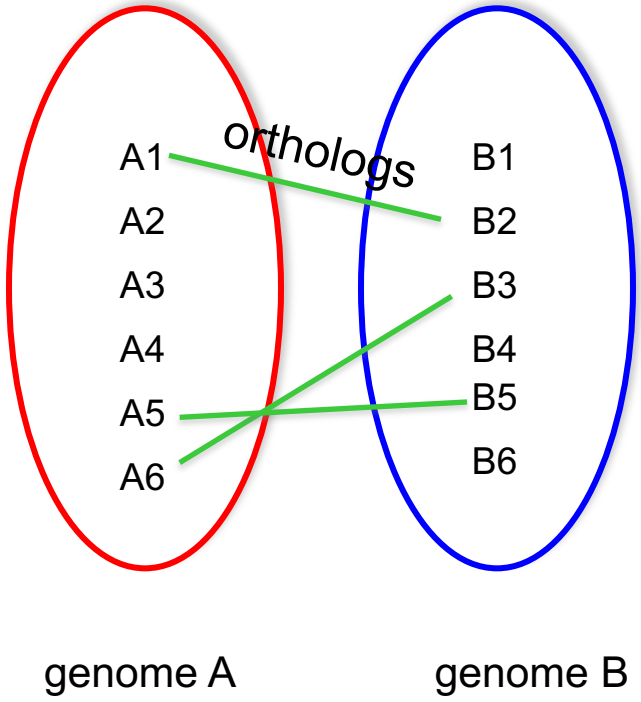
idea: search for homologs + « clustering »



Graph-based methods

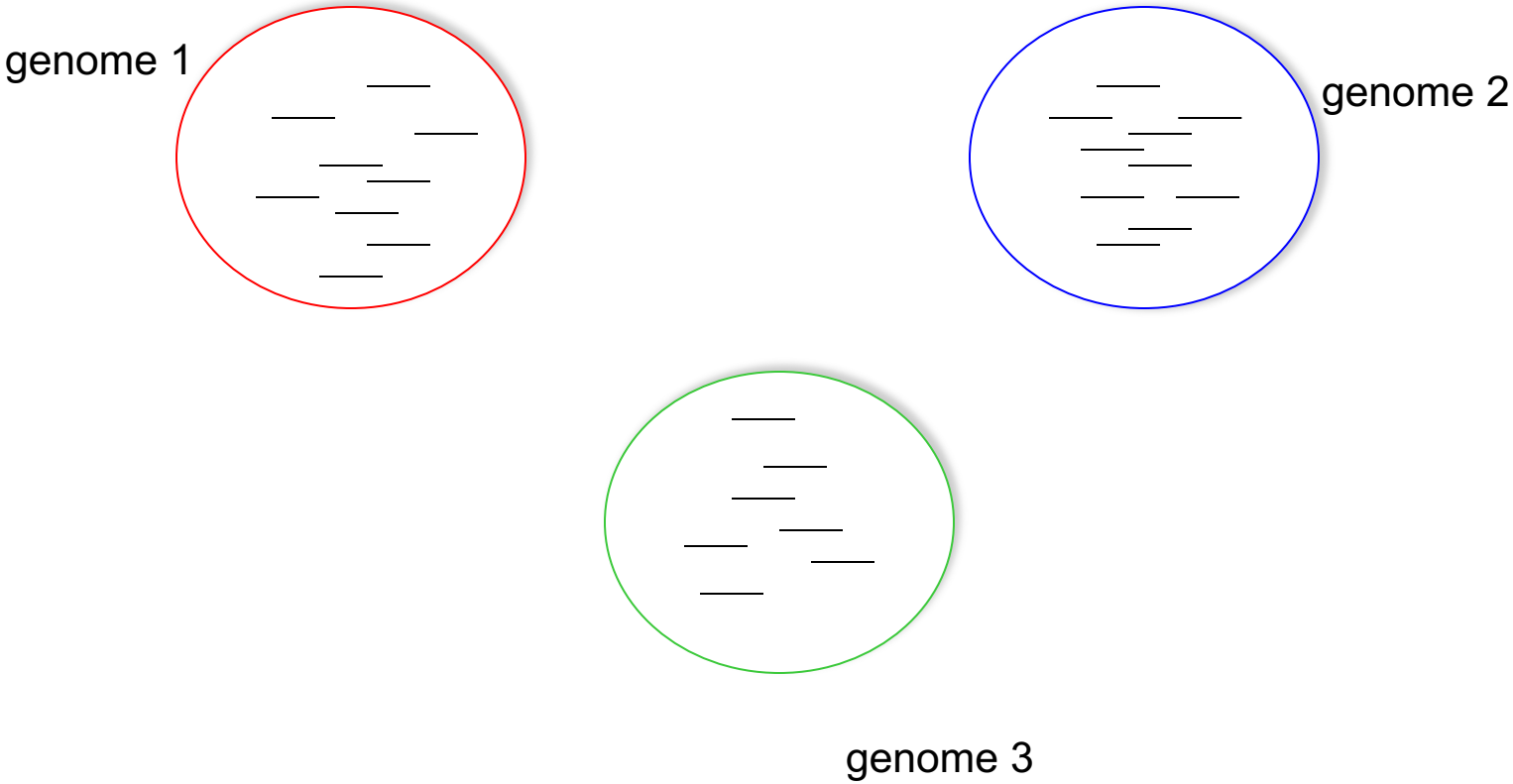
BBH (Bidirectional Best Hits)

idea: search for homologs + « clustering »



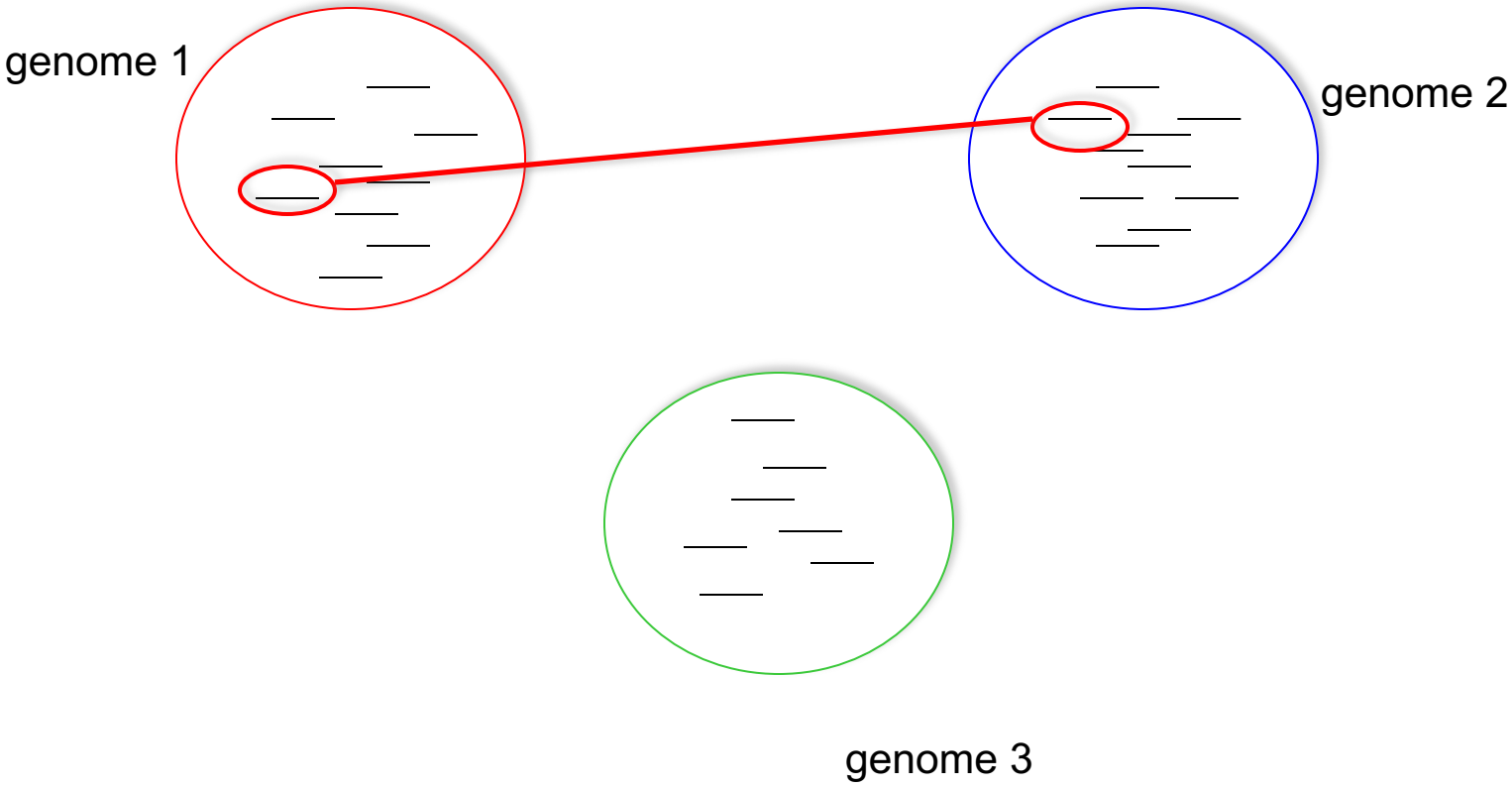
Graph-based methods

BBH (Bidirectional Best Hits)



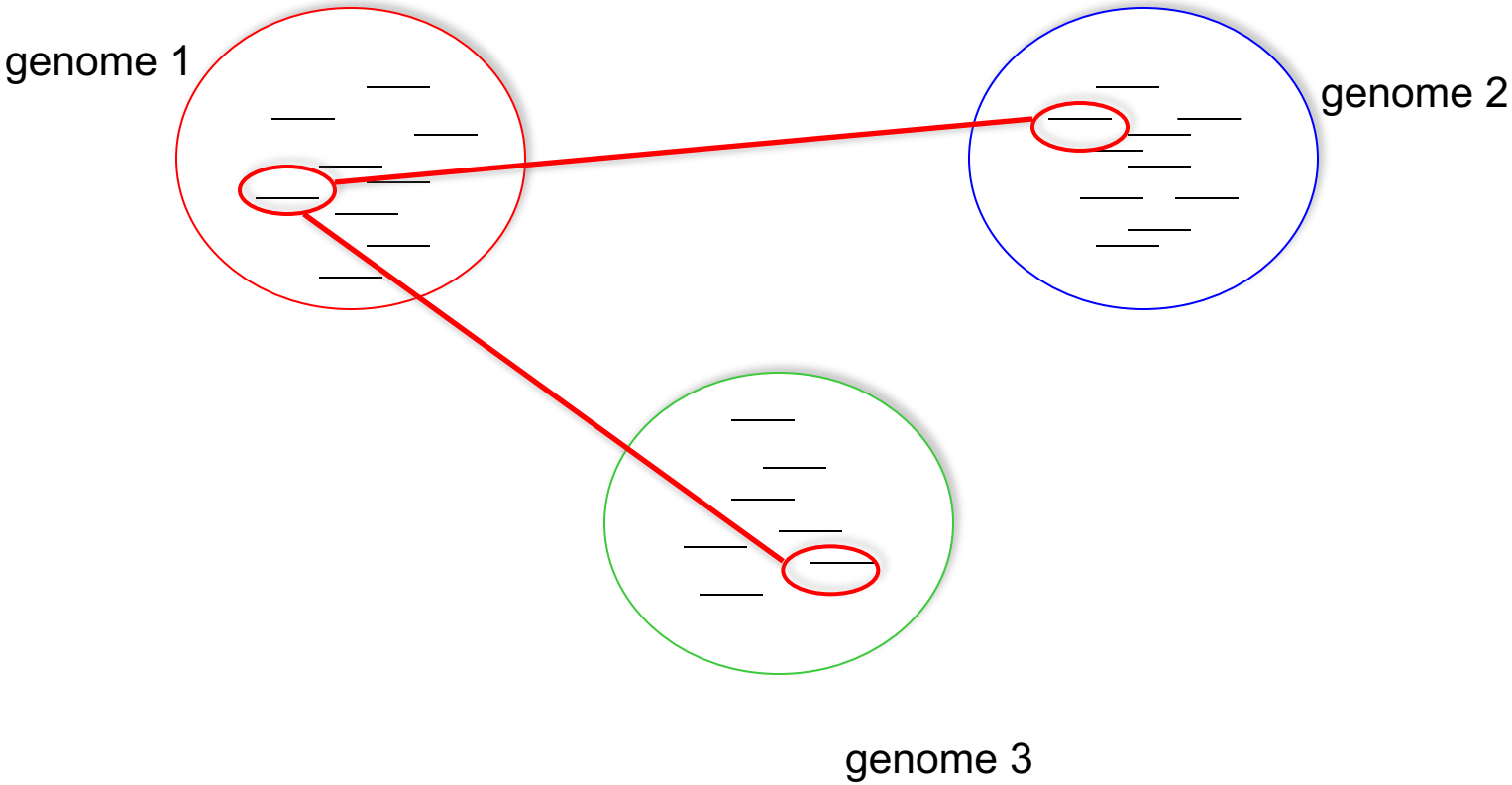
Graph-based methods

BBH (Bidirectional Best Hits)



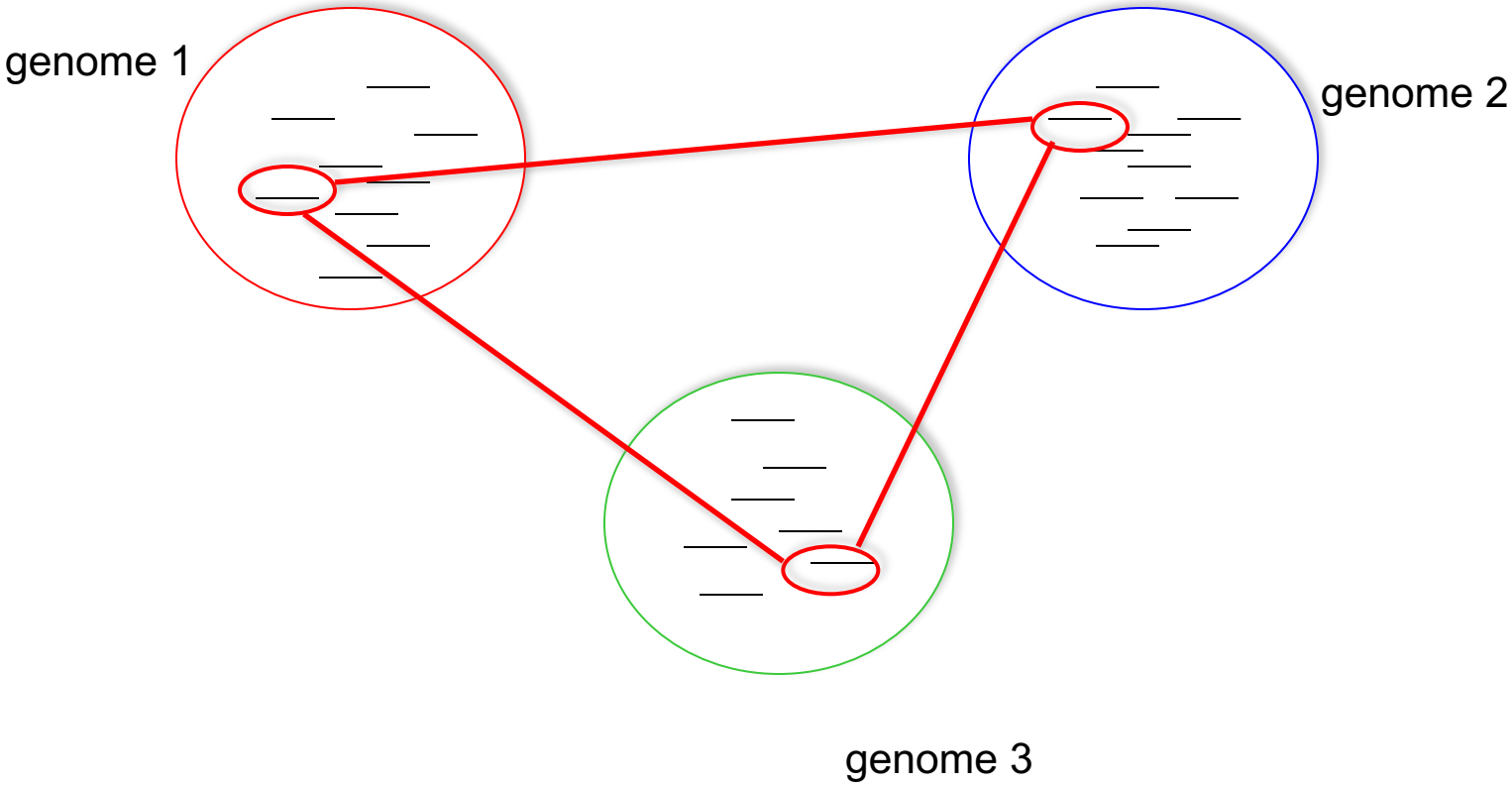
Graph-based methods

BBH (Bidirectional Best Hits)



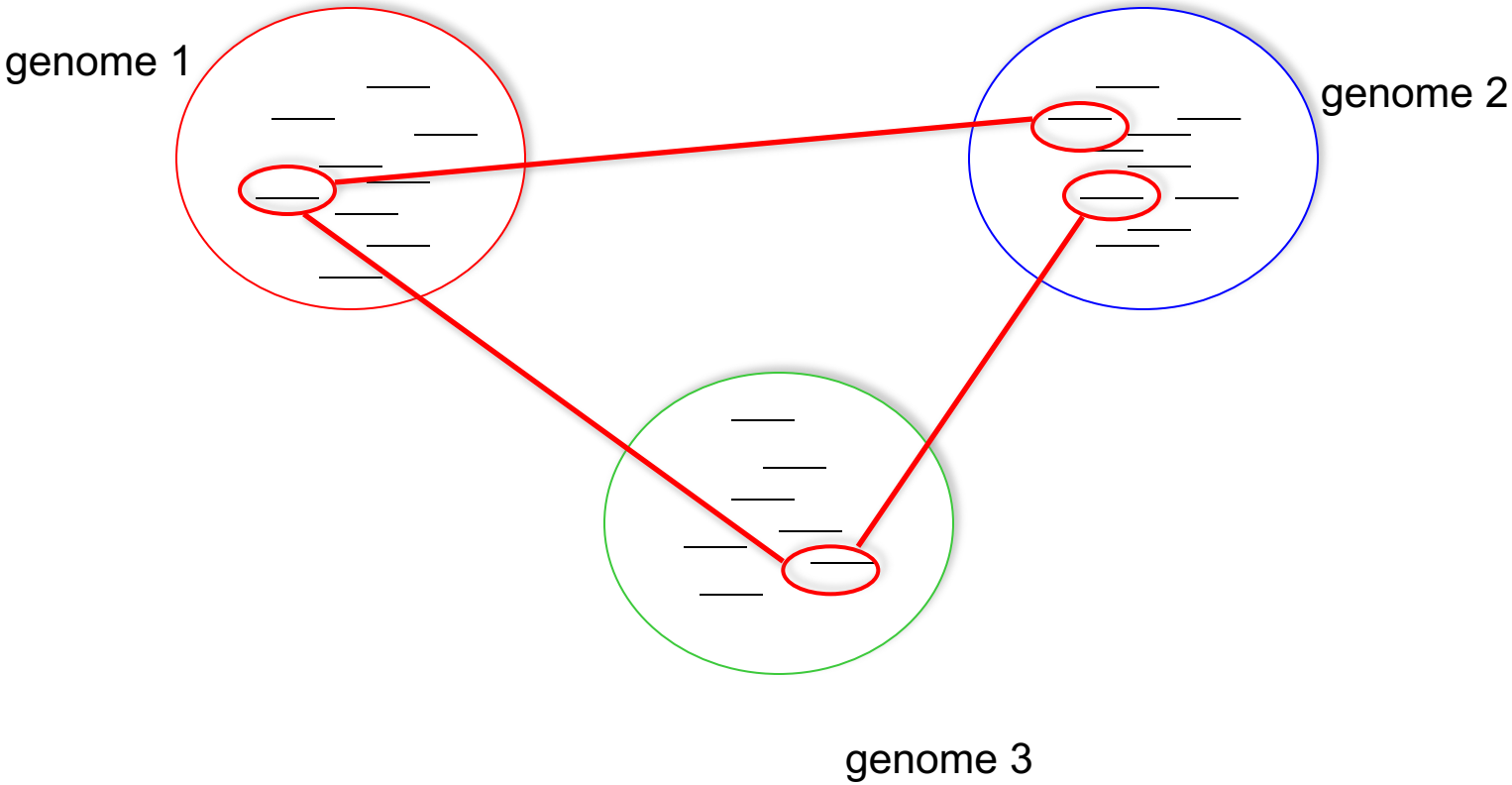
Graph-based methods

BBH (Bidirectional Best Hits)



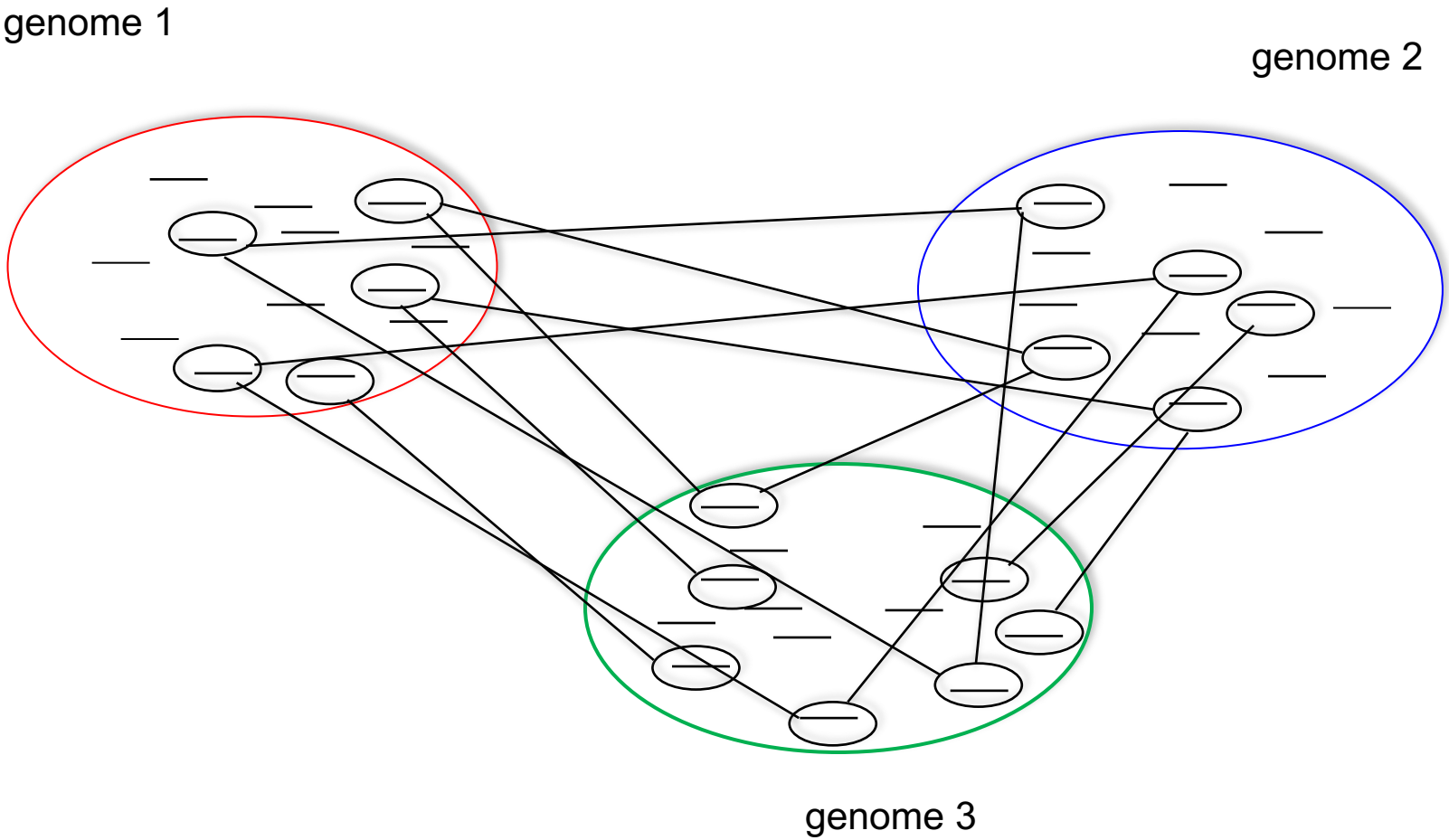
Graph-based methods

BBH (Bidirectional Best Hits)



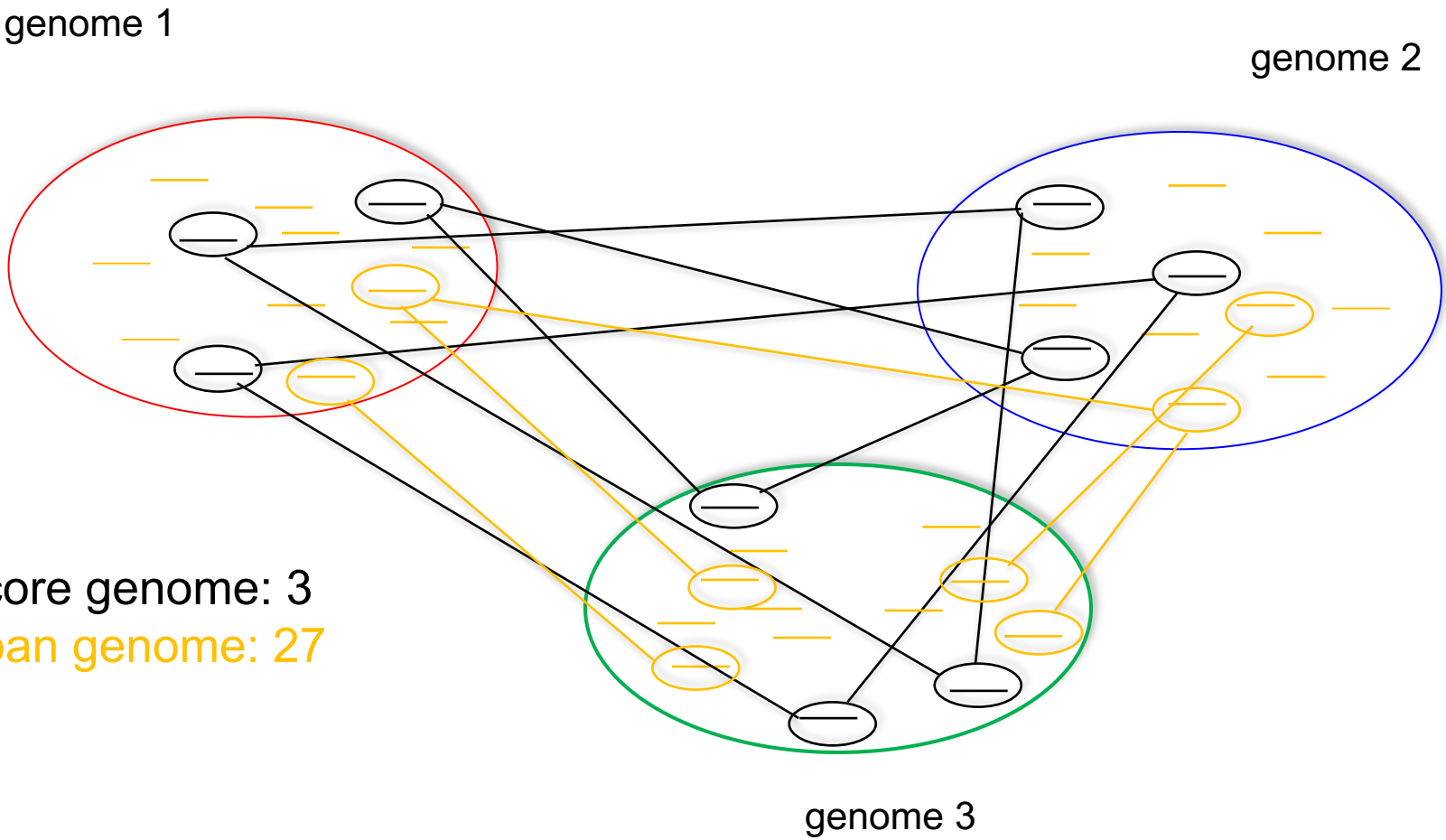
Graph-based methods

BBH (Bidirectional Best Hits)



Graph-based methods

BBH (Bidirectional Best Hits)



Graph-based methods
BBH (Bidirectional Best Hits)

- main differences:
  - definition of BBH (thresholds, evalue, coverage, seqid?)
  - clustering (single linkage (transitive), complete linkage (cliques), …)?

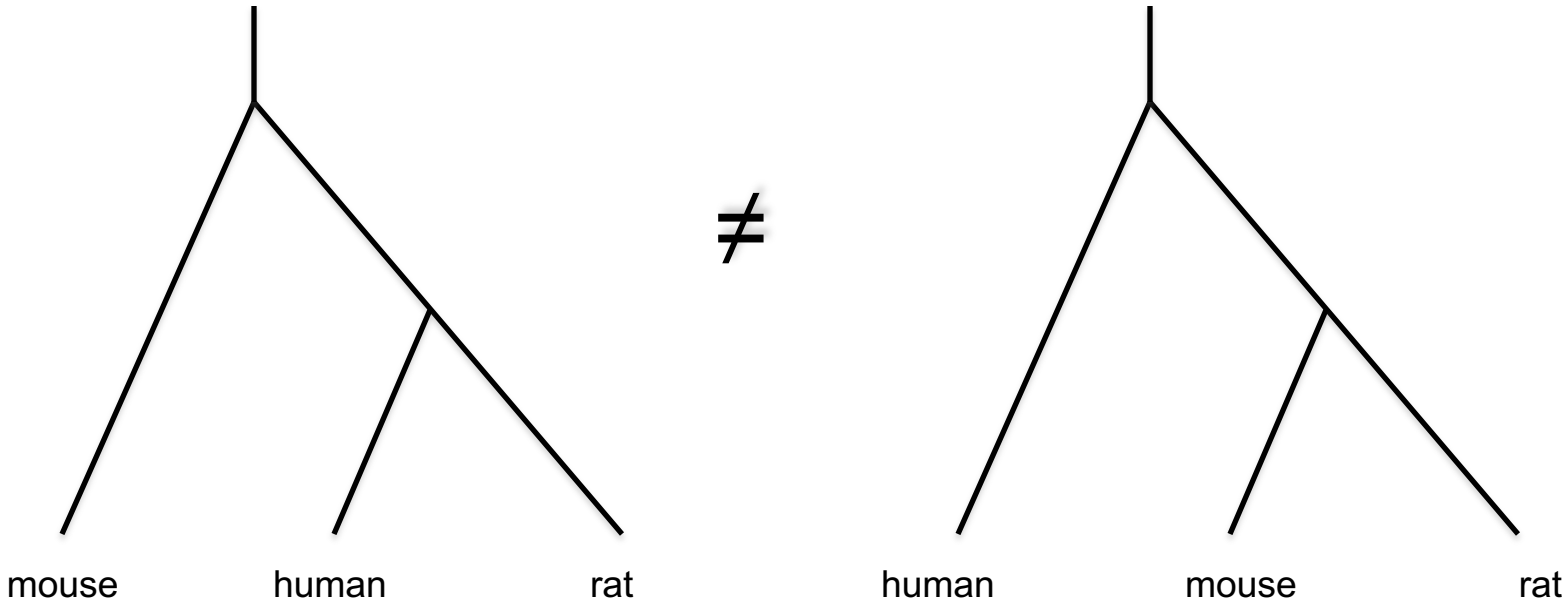
- main difficulties:
  - highly divergent sequences (fast evolving genes or remote species), twilight zone (convergence, divergence?)

  - domain shuffling (repeated domains, different domain organizations/orders…).
    Partial similarity between proteins via different domains.
    common domains ≠ identical functions
- advantages :
  - sensitive
  - fast
  - phylogeny of species is not neecssary

- disadvantages :
  - only 1:1 relationships (BBH)
  - does not consider the evolutionary history of species

- examples : OMA, COG/KOG, InParanoid

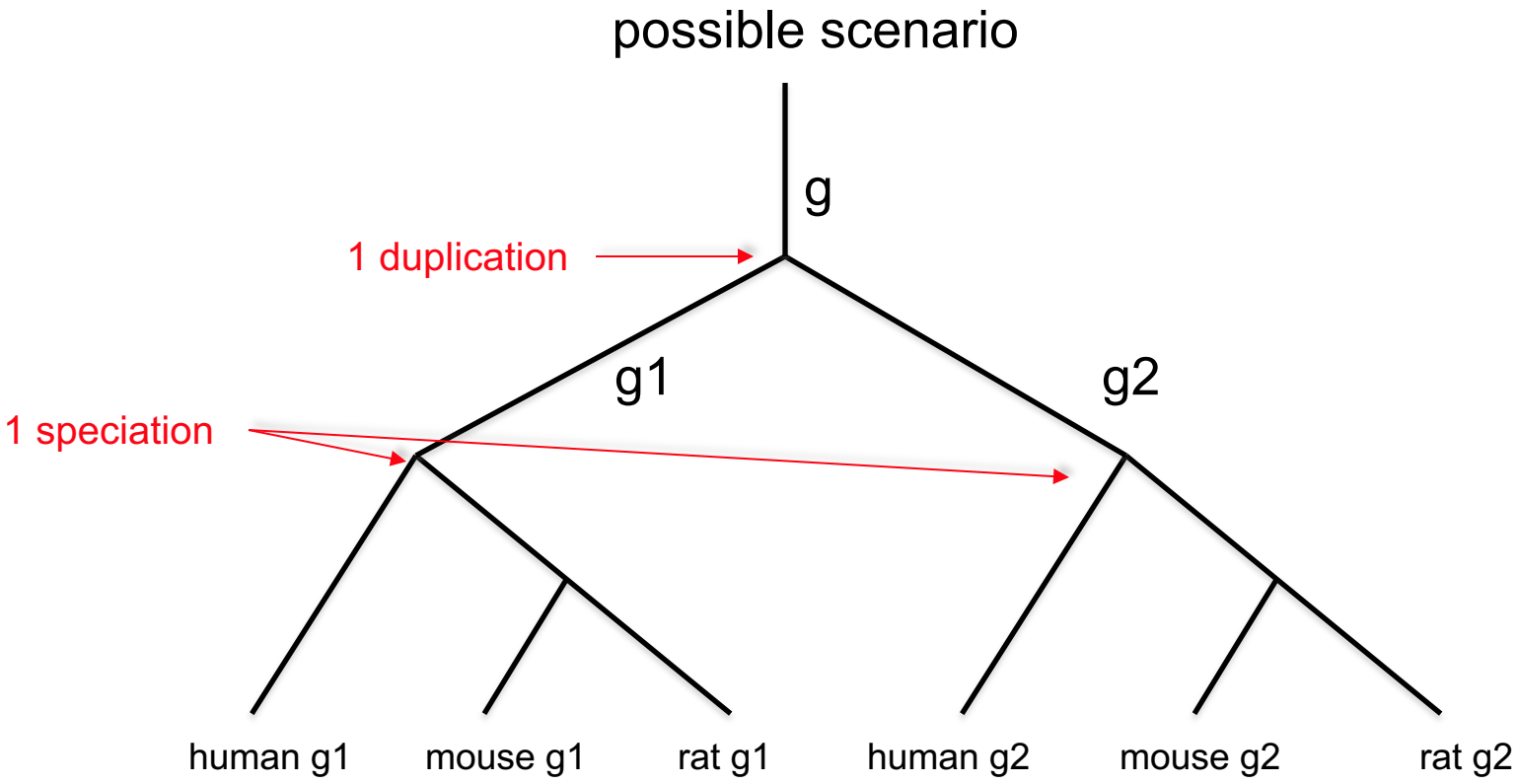
Phylogeny-based methods



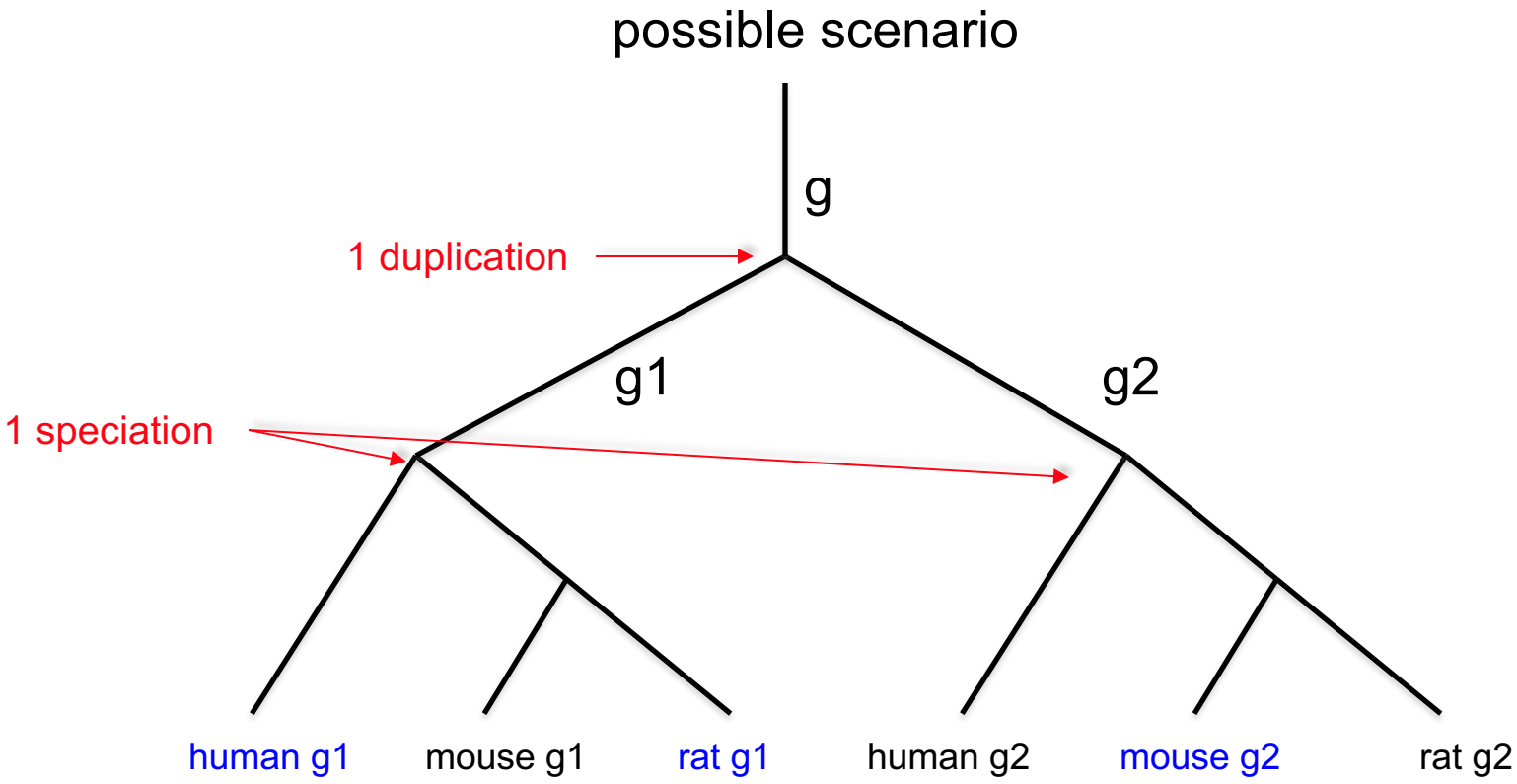
gene tree

species tree

Phylogeny-based methods



Phylogeny-based methods



The gene tree and the species tree are reconciled now!

Phylogeny-based methods

- advantages :
    - more specific since rely on and build evolutionary events (consider evolutionary history)

- disadvantages :
    - high computational time cost
    - struggle to cope with large protein families (computational time + reliability of results)
    - strongly dependent on the quality of the species tree
        - especially if the phylogeny of the species is unknown (bacteria? Viruses?)
        - if species are highly divergent (poor quality of the resulting tree)

- exemples : Orthostrapper, COCO-CL, Ensembl

Duplication

Major mechanism in the evolution of genomes.

Some numbers: 15% and 26% of duplicated genes in Human and yeast respectively

- duplication of:
  - a gene
  - a DNA fragment
  - a whole genome (WGD)

Gene Duplication

Consequences:

- pseudogenization : loss of function and/or transcription/translation for one copy
  (apparition of STOP codons, deletion of a TATA box, frameshift event…)
- increase in protein concentration: selective pressure operates on all copies
- neofunctionalization : one copy accumulates mutations -> new function
  ex: olfactive receptors
- subfunctionalization : initial gene had two functions => each copy ensures one of the two initial functions
  - the organism still ensures both functions
  ex : *exonuclease* + *recombinase in phages*

Gene Duplication

Consequences:

- pseudogenization : loss of function and/or transcription/translation for one copy
  (apparition of STOP codons, deletion of a TATA box, frameshift event…)
- increase in protein concentration: selective pressure operates on all copies
- neofunctionalization : one copy accumulates mutations -> new function
  ex: olfactive receptors
- subfunctionalization : initial gene had two functions => each copy ensures one of the two initial functions
  - the organism still ensures both functions
  ex : *exonuclease* + *recombinase in phages*

One of the major evolutionary event but not the only one :)