Phylogeny

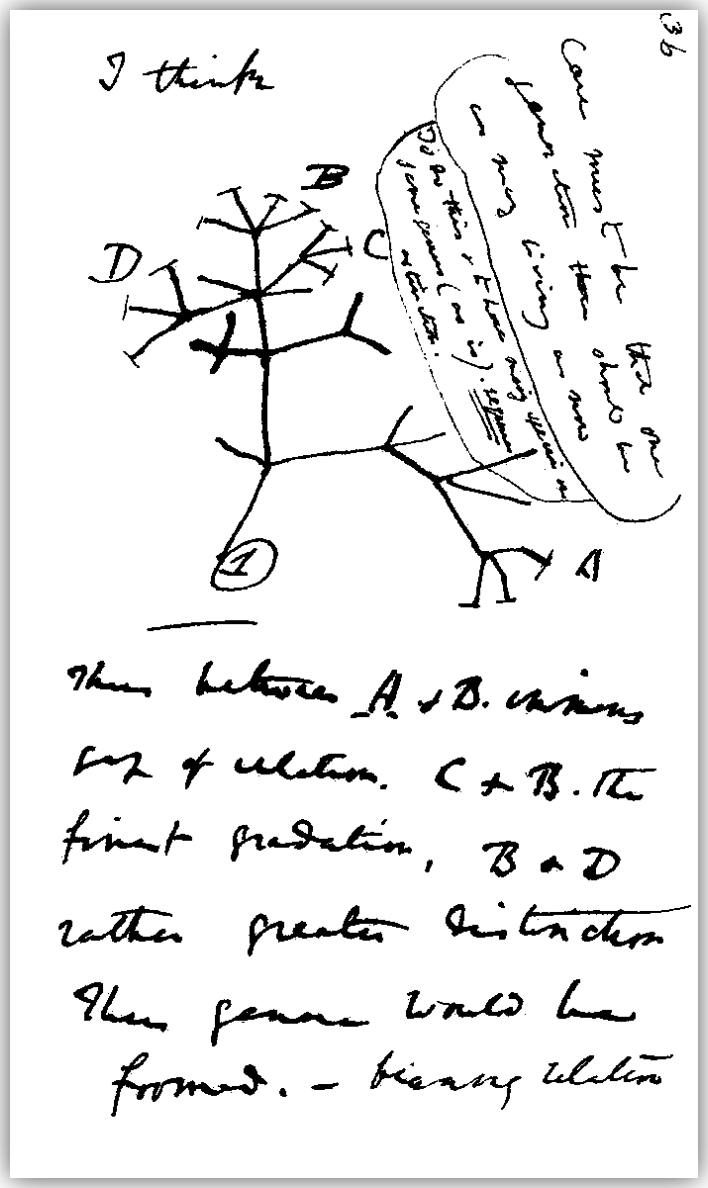
MAISB

ISB101

Roland Krause

Content

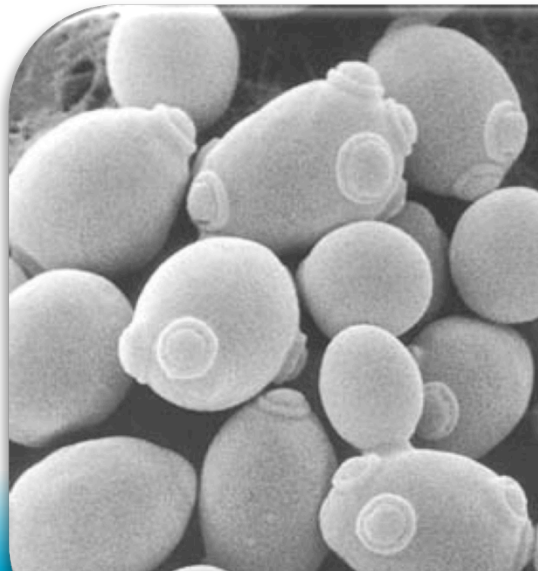
- Orthology and paralogy
  - Refined definitions
  - Practical approaches to orthology
- Building trees
- Tree reconciliation



Definitions for evolutionary genomics

What is “the same gene” in another species?

- Only a small fraction of genes will be characterized experimentally ever
- Model organisms
- Which genes in a given organism perform the same function?
- Transfer of functional information between proteins in different species
  - 10,000s of bacterial genomes
  - 1000 eukaroytic genomes in various stages
  - Vast Metagenomics studies



Pairwise similarity searches

- Similar protein sequences allow the inference of protein function
- Functional assessment and transfer between organism need to be automated

- BLAST based detection
  - FASTA, Smith-Waterman
- Statistics for the similarity of proteins
  - Identity and similarity (percent)
  - Bit-scores
    - Normalized bit-score
    - E-values

50% Identity?

- There is no universal threshold.
- Evolution provides better boundaries for functional transfer
  - Homologs
  - Orthologs
  - Paralogs

Homology

- "...the same organ in different animals under every variety of form and function".
  - Richard Owen, 1843
- Distinction between analogs and homologs
- ("Origin of species", published 1859)
- Homology and common descent are notions introduced by Huxley

Homology and analogy

- **Homology** designates a relationship of common descent between similar entities
  - Bird wings and tetrapod limbs
  - Leghemoglobin and myoglobin
- **Analogy** designates a relationship with **no** common descent
  - Convergent evolution with traits evolving differently
    - Tetrapod and insect limbs
    - Flippers and body shape of dolphins and fish
    - Elements of tertiary structure

Very homologous genes

- Genes (or features) are either homologous or not
- There is no ***70% homology*** (blink)
- The term can also be applied to genomic regions of synteny, exon or even single nucleotides

Elementary events

Microevolution

- Vertical descent (speciation) with modification

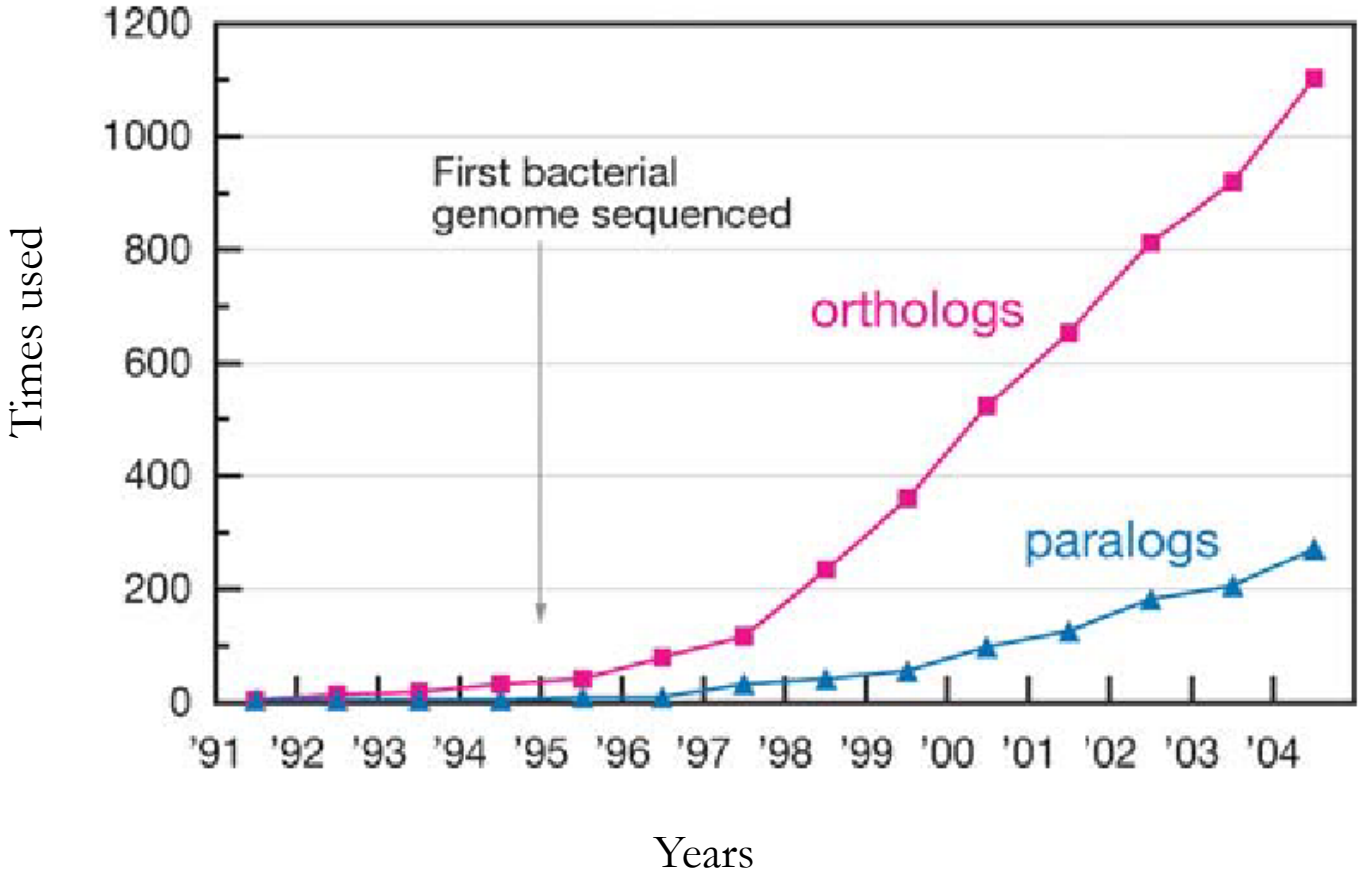
Macroevolution

- Gene duplication
- Gene loss
- Horizontal gene transfer
- Fusion of domains and full-length genes

Gene duplication

- Assessing gene duplications
  - Duplication noted by Fisher in 1928, expanded by Haldane 1932
- Around 1970
  - Ohno: *Evolution by Gene Duplication*
  - Walter Fitch: *Distinguishing homologous from analogous proteins.*
    - The definition of orthology and paralogy as concepts

Usage of the terms



1970 - 1990: 45 mentions in Pubmed
August 2009: 3636 orthologs, 1303 paralogs
August 2011: 4738 orthologs, 1747 paralogs
June 2014: 6566 orthologs, 2581 paralogs

Orthologs

Definition

- Event: Speciation
- Two proteins are considered orthologs if they originated from a single ancestral gene in the most recent common ancestor of their respective genomes

Properties

- Reflexiv
  - If A is o. to B, B is o. to A
- Not transitive
  - If A o. to B and B o. to C, A is not necessarily o. to C
- We cannot show orthology, only infer a likely scenario

Mutations in the Caenorhabditis elegans dystrophin-like gene dys-1 lead to hyperactivity and suggest a link with cholinergic transmission

Catherine Bessou · Jean-Bernard Giugia · Christopher J. Franks · Lindy Holden-Dye
Laurent Ségalat

Received: August 18, 1998 / Accepted: September 15, 1998 / Published online: December 9, 1998

ABSTRACT

Mutations in the human *dystrophin* gene cause Duchenne muscular dystrophy, a common neuromuscular disease leading to a progressive necrosis of muscle cells. The etiology of this necrosis has not been clearly established, and the cellular function of the dystrophin protein is still unknown. We report here the identification of a dystrophin-like gene (named *dys-1*) in the nematode *Caenorhabditis elegans*. Loss-of-function mutations of the *dys-1* gene make animals hyperactive and slightly hypercontracted. Surprisingly, the *dys-1* mutants have apparently normal muscle cells. Based on reporter gene analysis and heterologous promoter expression, the site of action of the *dys-1* gene seems to be in muscles. A chimeric transgene in which the C-terminal end of the protein has been replaced by the human dystrophin sequence is able to partly suppress the phenotype of the *dys-1* mutants, showing that both proteins share some functional similarity. Finally, the *dys-1* mutants are hypersensitive to acetylcholine and to the acetylcholinesterase inhibitor aldicarb, suggesting that *dys-1* mutations affect cholinergic transmission. This study provides the first functional link between the dystrophin family of proteins and cholinergic transmission.

Key words *Caenorhabditis elegans* · Dystrophin · Duchenne muscular dystrophy · Acetylcholine

C. Bessou · J.B. Giugia · L. Ségalat (✉)
IPMC, CNRS-UPR411, 660 route des Lucioles,
F-06560 Sophia Antipolis, France
Tel : +33493957790; Fax: +33493957708
e-mail: segalat@ipmc.cnrs.fr

C.J. Franks · L. Holden-Dye
School of Biological Sciences, University of Southampton,
Bassett Crescent East, Southampton, SO16 7PX, UK

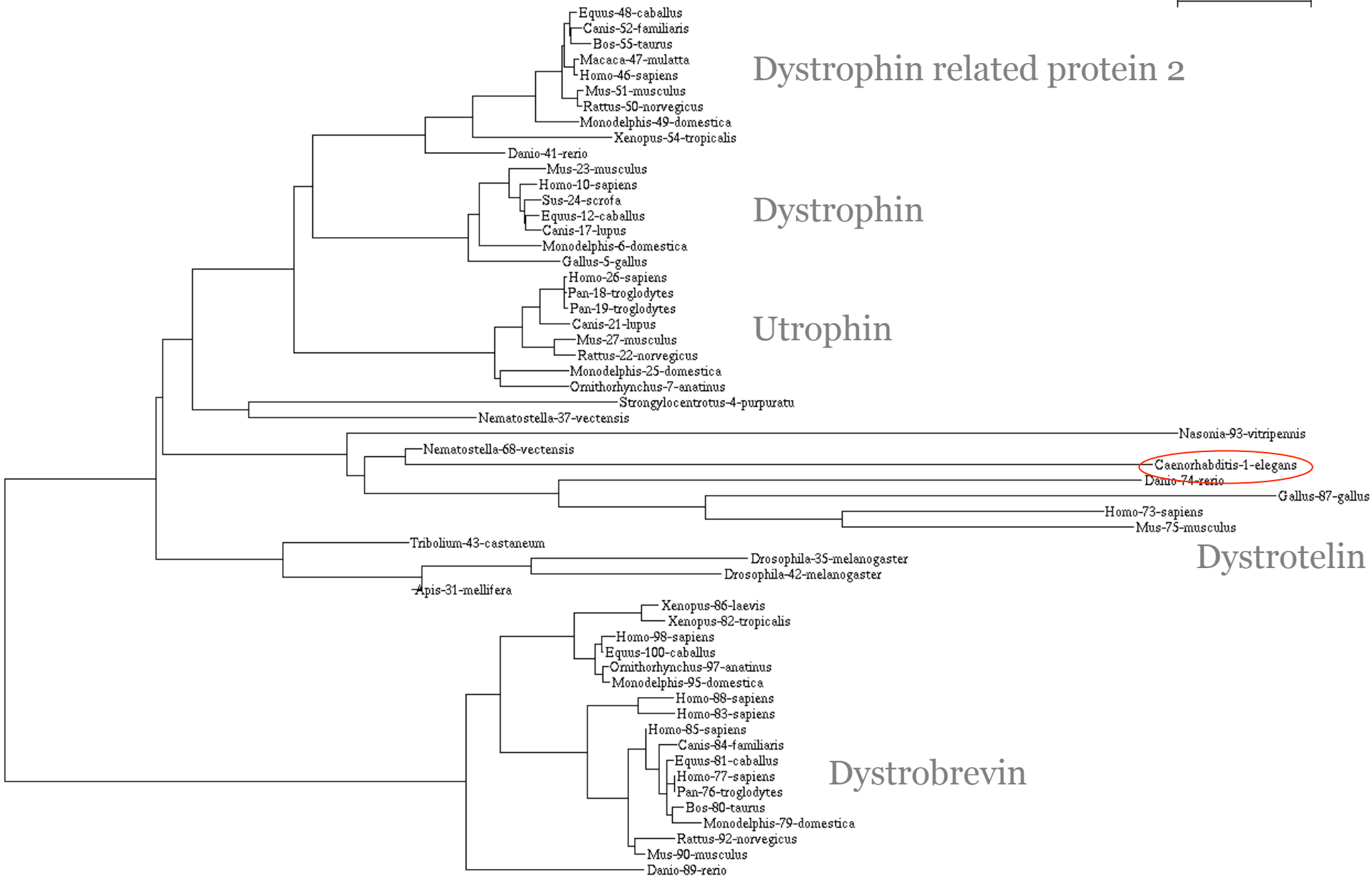
INTRODUCTION

Duchenne muscular dystrophy (DMD) and Becker muscular dystrophy (BMD) are allelic progressive myopathies affecting approximately 1 in 3,500 male births [1]. Dystrophin, the product of the gene mutated in DMD and BMD patients, is a 3,685-amino acid protein found in skeletal and cardiac muscles and in the nervous system. Internal promoters of the dystrophin gene generate shorter forms of the protein expressed in a variety of tissues, including peripheral nerves. Despite extensive cellular and biochemical characterization, the role(s) of dystrophin has(ve) not been clearly demonstrated, nor has the etiology of the disease [2–4]. In muscles, dystrophin is present under the sarcolemma membrane and associates by its N-terminus to the actin cytoskeleton and by its C-terminus to the DGC, a large complex of proteins spanning the membrane. In mammals, utrophin, a protein very similar to dystrophin, and expressed in most cell types, is (at least partly) functionally redundant with dystrophin [5], although its subcellular localization is different in muscles and muscle cells [6, 7]. Mice carrying mutations in either dystrophin or utrophin gene display only a mild neuromuscular defect [8–10], suggesting that these two proteins functionally overlap. A third dystrophin/utrophin-like protein, DRP2, has been identified in mammals. This protein is abundantly expressed in the nervous system [11].

It has been proposed that the function of dystrophin in normal muscles is to increase membrane stability and thus to reduce the risk of membrane shearing during repeated contractions [2, 3]. In non-contractile cell types, however, the role of dystrophin (and of utrophin and DRP2) is probably different and is largely unknown [12]. The dystrophin family of proteins may have a more-general role than increasing membrane stability and may serve as a local membrane organizer, as suggested by the fact that dystroglycan-α, a member of the dystrophin-associated proteins (DAPs), is a putative component of the agrin receptor [13, 14], and that

Orthology:
Who cares?

Found by Martijn Huynen



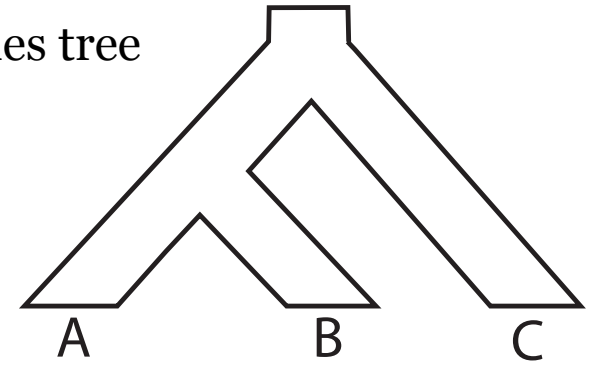
The DYS-1 gene from C.elegans is not orthologous to dystrophin.
No surprise of the knockout on the muscle cells.

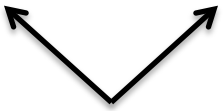
Paralogs

- Two genes are paralogs if they are related by duplication
- Recent paralogs can retain the same function
- Fate in functional divergence
  - Neofunctionalization
    - One copy free of evolutionary constraints evolves a new function or is lost
  - Subfunctionalization
    - Both functions shift into more specific uses
    - Better supported model

Orthologs

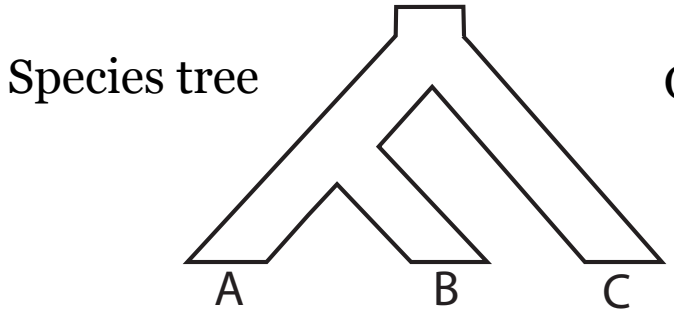
Species tree

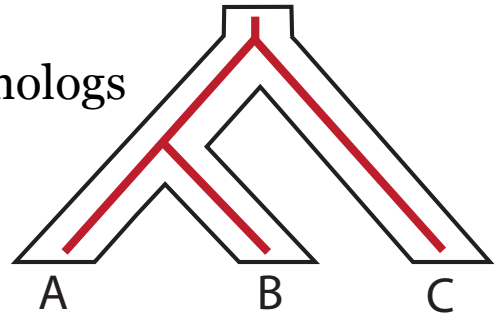




Orthologs

Orthologs and paralogs





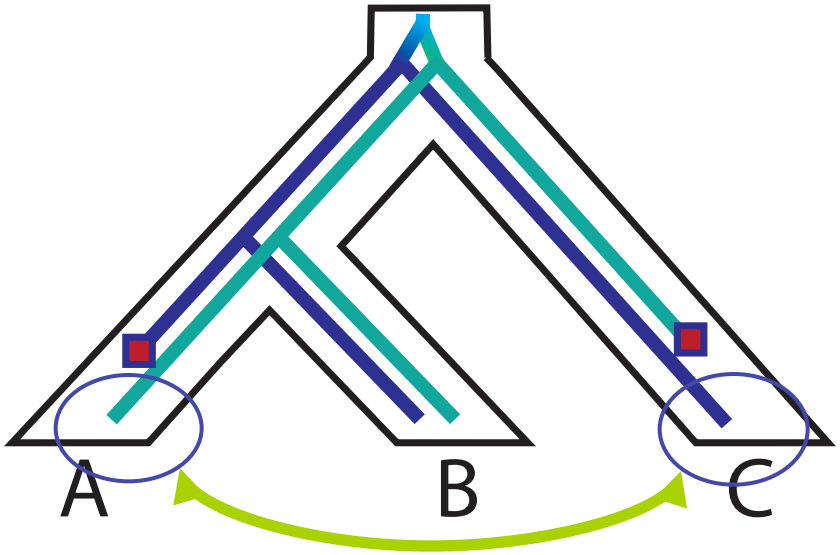
- A, B, C: Species

- Orthologs: Genes related by a speciation event
- Paralogs: Genes related by a duplication event
- In-paralogs: duplication after the relevant speciation
- Out-paralogs: duplication before the relevant speciation
- Co-orthologs: Genes related by speciations that underwent subsequent duplications

Distinction of paralogs

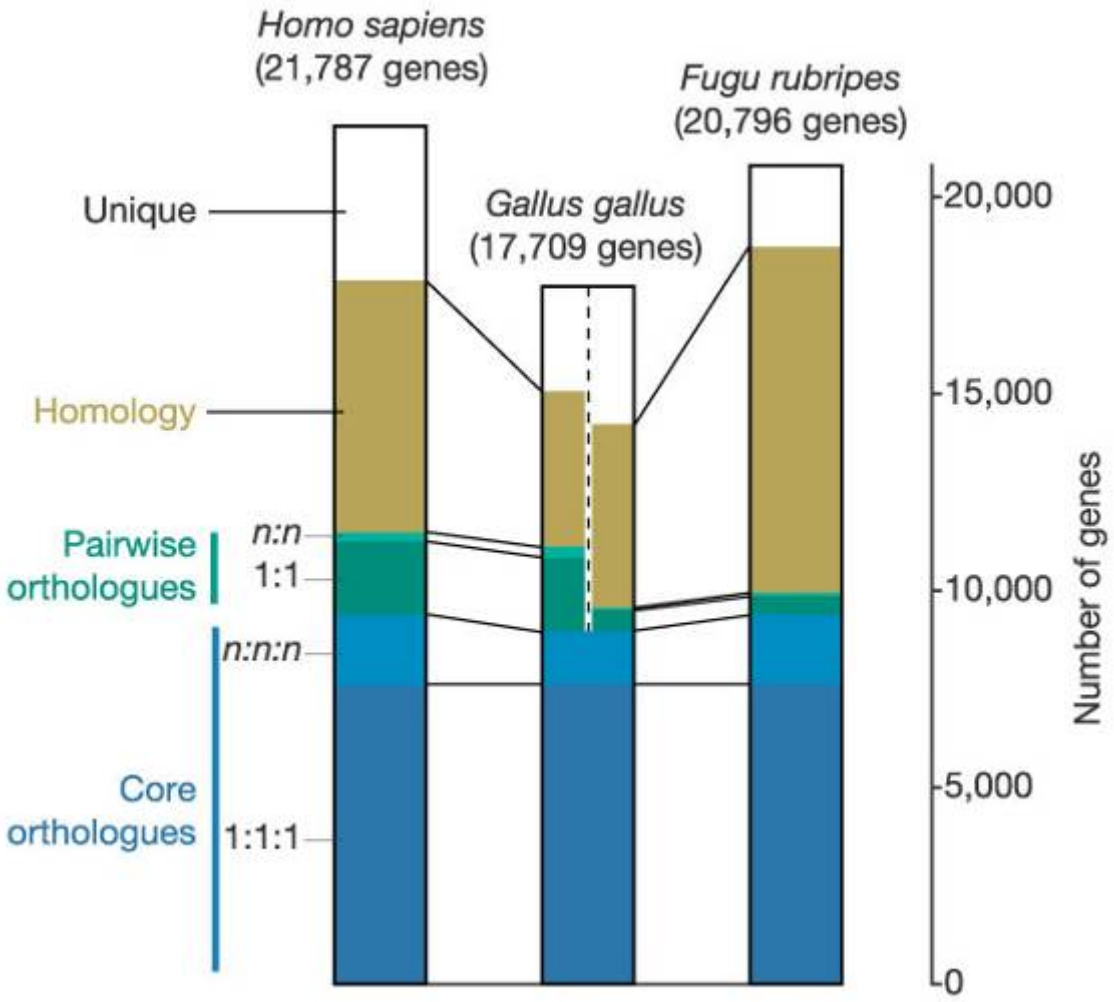
- Time of the speciation event
- In-paralogs (symparalogs, ultra-paralogs)
  - Duplication after species diverged
  - Within a single species
- Out-paralogs (alloparalogs)
  - Ancient duplicates
  - Across species boundaries

The effects of gene loss



Pseudo-
orthologs

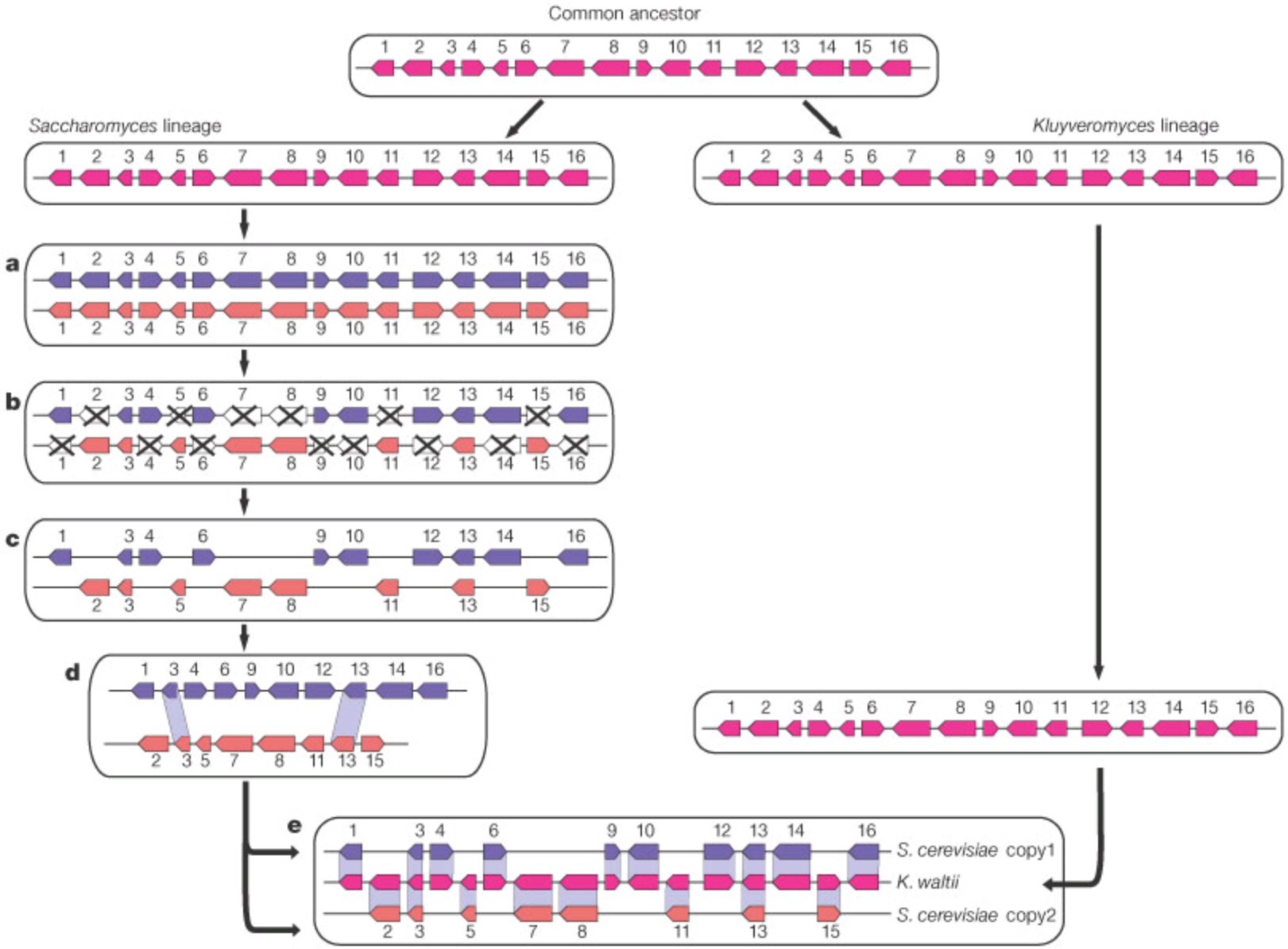
Eukaryotic scenarios



From the Chicken Genome publication

Whole genome duplication

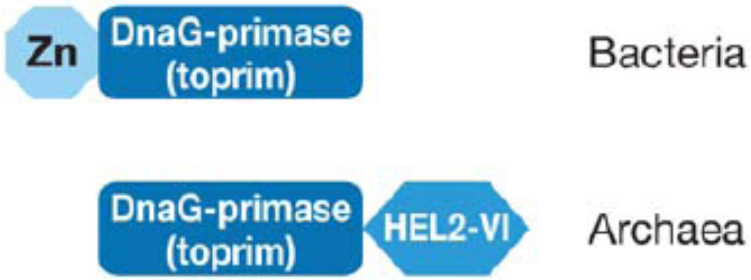
- Genomes are routinely copied in cells
- Replication errors can lead to polyploidy
  - Severe phenotypes in human
  - Very common in plants
- Important whole genome duplications
  - Early metazoan lineage
  - Ray finned fish
  - *Saccharomyces cerevisiae*



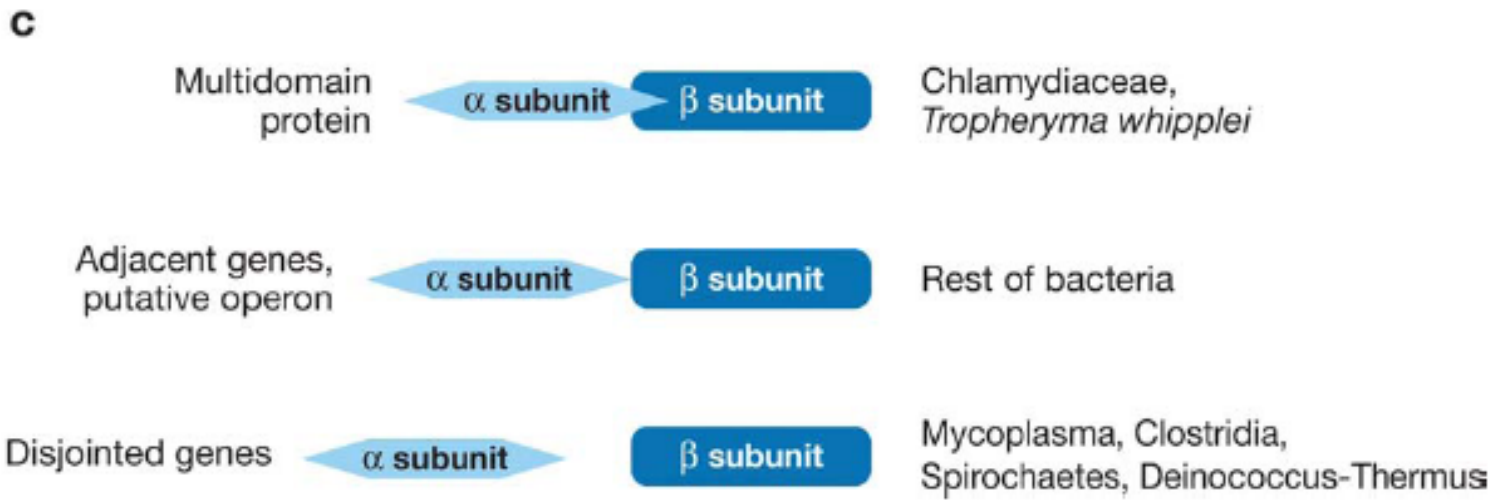
Kellis et al. Nature (2005)

Domain structure

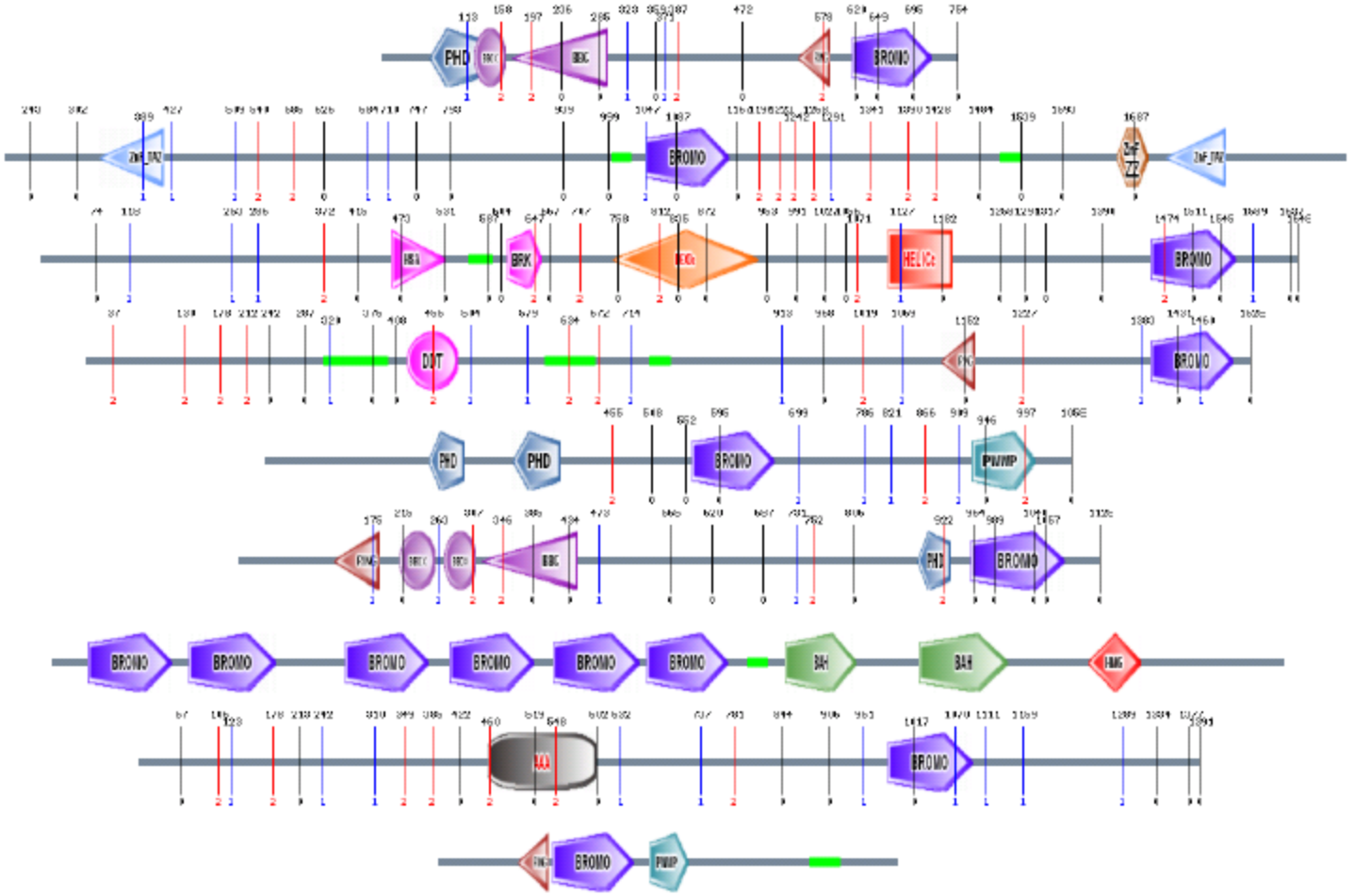
Independent fission events



Prokaryotic scenario



BROMO and friends



Functional transfer

- Koonin et al. inspected 1330 one-to-one orthologs between *E. coli and B. subtilis*
- Few differences in function
  - Transporter specifities/preferences
  - Comprehensive, gene based studies limited
- Use of protein-protein interaction data to judge functional equivalence

Functional orthologs

- Can we prove orthology experimentally?
  - No!
  - Test for functional equivalence
  - Knock-out mutant, replace with cognate copy from other species
    - Developmental genes between fly and worm
    - Metabolic enzymes between Mycobacteria and Enterobacteria

Known limitations

- Differences in genomic structure and life-style
  - Low GC vs high GC genomes
  - Regulatory sequences?
  - Negative results do not disprove orthology (or functional similarity)
  - Paralogs can work as a replacement copy

1-to-1 orthologs

- For complete genomes, genes only separated by speciation events
- Most reliable set, we would typically assume functional eqivalence
- Other names: Superorthologs

Advanced terminology

- In-paralogs
  - Genes duplicated after the last speciation event (orthologs)
- Out-paralogs
  - Genes duplicated before the last speciation event
- Co-orthologs
  - Genes in one lineage that are together ortho
- Xenologs
  - Violation of orthology due to horizontal gene transfer (HGT)
- Pseudo-orthologs
  - Proteins with a common descent due to lineage specific loss of paralogs
- Pseudoparalogs
  - No gene ancestral gene duplication but HGT

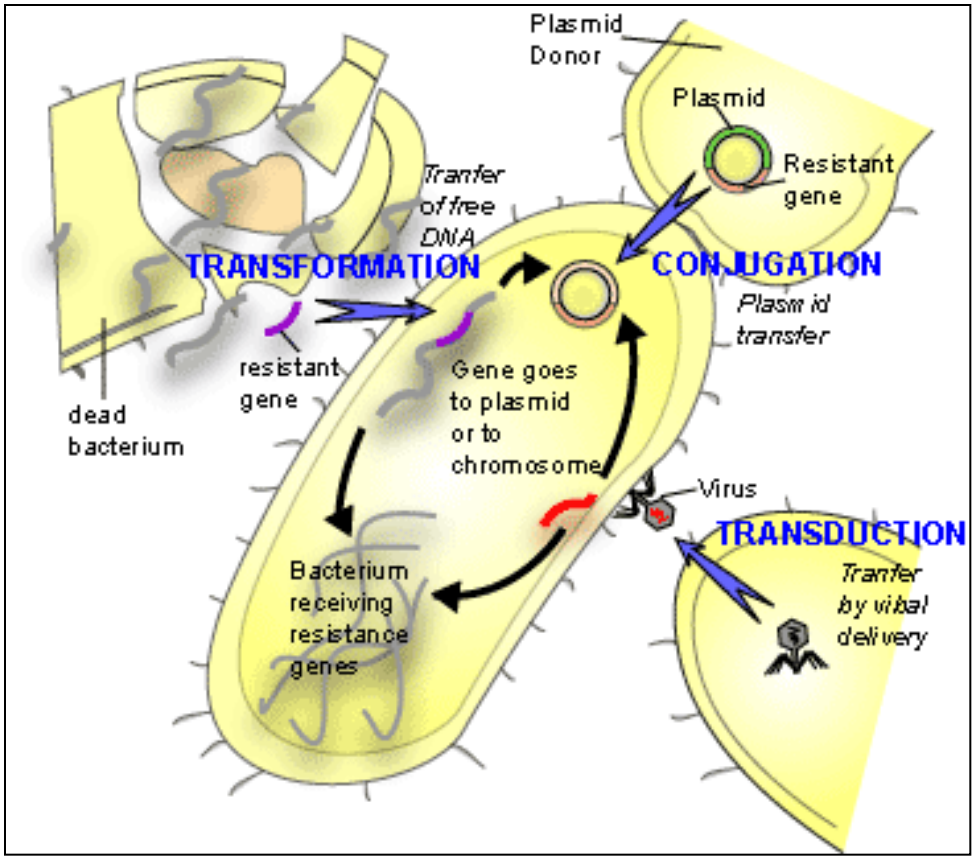
Bonus track

- Ohnologs
  - Gene duplication originating in whole genome duplication (WGD)
- Superorthologs
  - Groups of orthologs that all have a 1-to-1 correspondence

Horizontal gene transfer (HGT)

- Prokayotes exchange genetic material across lineages
  - 5 to 37% of the *E. coli* genome
- Conjugation (Plasmids_
- Transformation (Naked DNA)
- Transduction (Phages)

- Hallmarks of HGT
  - Higher similarity of proteins
  - Unusual GC-content (low)
  - Unusual codon usage



How to build a phylogenetic tree?

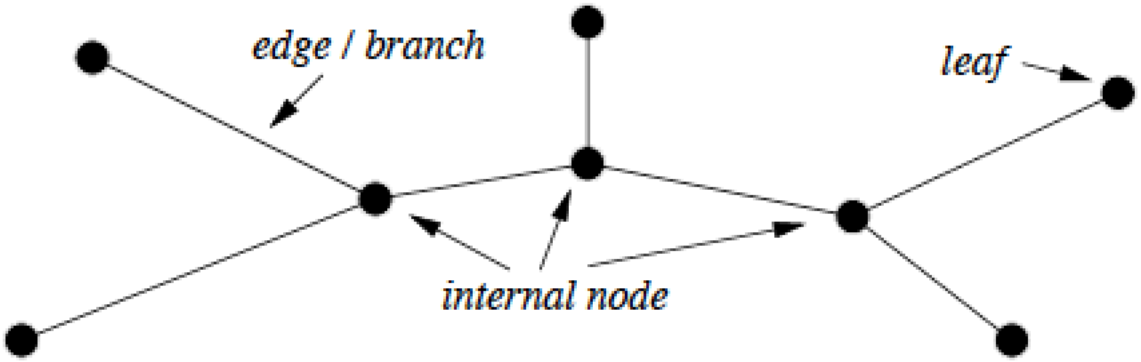
Phylogenetics

Aligned sequence

```
Human   STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRL
Gorilla STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFKL
Horse   SNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRL
Pig     SNADAVMGNPKVKAHGKKVLQSFSDGLKHLDNLKGTFAKLSELHCDQLHVDPENFRL
Cow     STADAVMNNPKVKAHGKKVLDSFSNGMKHLDDLKGTFAALSELHCDKLHVDPENFKL
Deer    SSAGAVMNNPKVKAHGKRVLDAFTQGLKHLDDLKGAFAQLSGLHCNKLHVNPQNFRL
Gull    SSPTAINGNPMVRAHGKKVLTSFGEAVKNLDNIKNTFAQLSELHCDKLHVDPENFRL
```

The tree

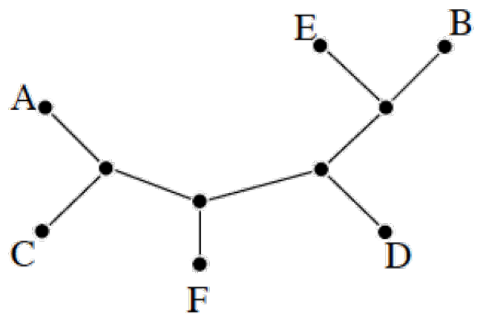
- We cannot observe ancestral sequences
- Find a tree that matches the data best



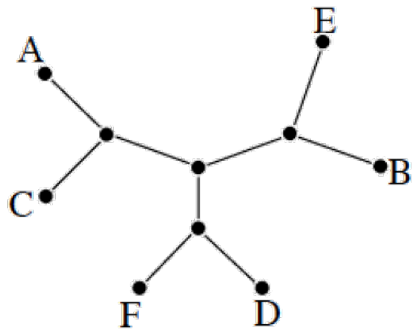
Topology and splits

- Two trees showing the same branching pattern are said to have the same tree topology. Which trees are identical?

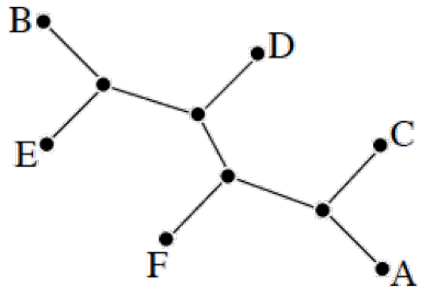
a)



b)



c)



- A split at an edge is phylogenetically informative, if the edge is not connected to a leaf. For the tree in b) the splits (AC|FDEB), (FD|ACEB), (EB|ACFD) are phylogenetically informative.

Maximum parsimony

- Minimal number of substitution events

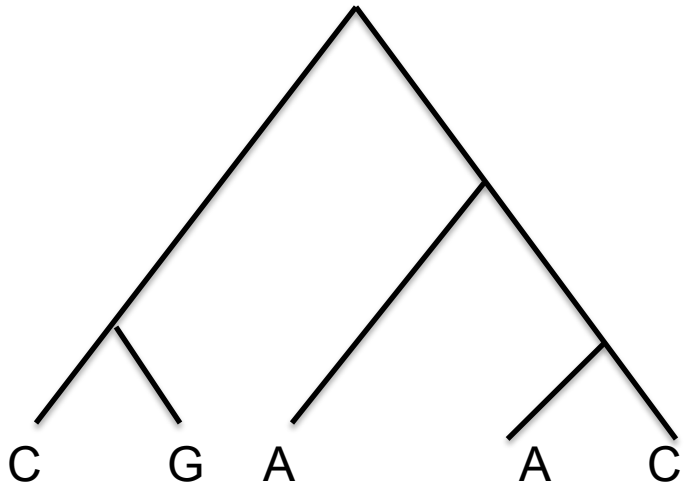
	1	2	3	4	5
S1	G	A	A	T	C
S2	G	C	G	C	T
S3	G	T	G	A	T
S4	G	G	A	C	C

Maximum parsimony definiton

- Tree length *L* is the minimal number of substitutions required to explain a topology.
- The tree with the smallest length is the most parsimonious

MP algorithm (Sankoff)

- For each internal node, record the number of possible substitutions for every possible state (e.g. nucleotide).
- Cost is the minimum of the previous node state plus the cost for transition
- Label the tree from the leaves
- The costs for transitions can be incorporated
- (Example blackboard)

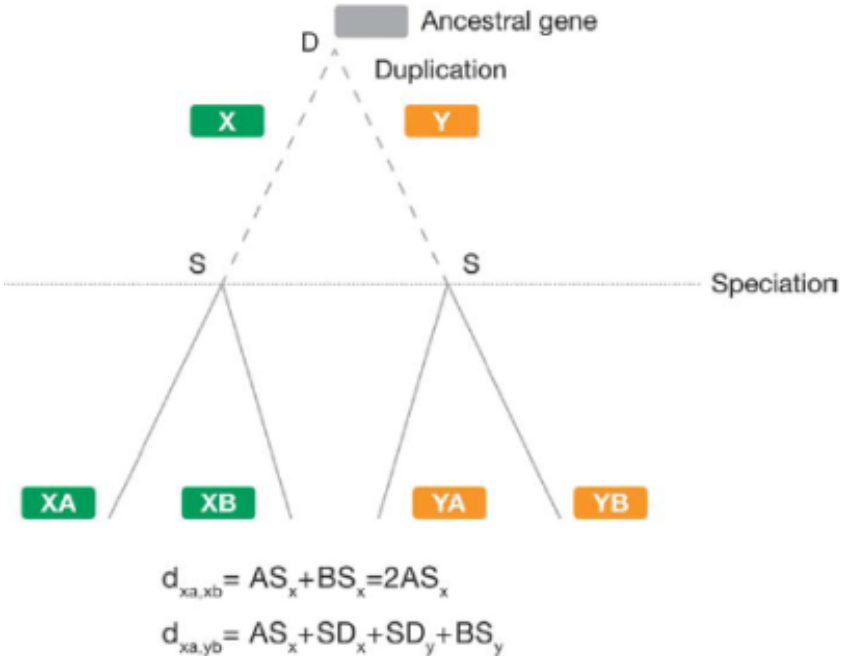


Number of trees

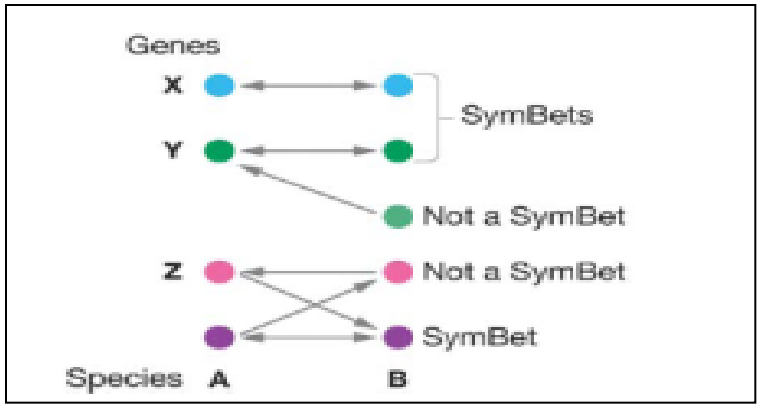
Methods for Orthologous Groups

Using reciprocal best hits

- Orthologs are more similar to each other than any other gene of the genomes considered
- False negatives if one paralogs evolves much faster than the other
- Typically used with BLAST



Lineage specific expansion



- Additional false negatives due to inparalogs
- Typical case for eukaryotic organism

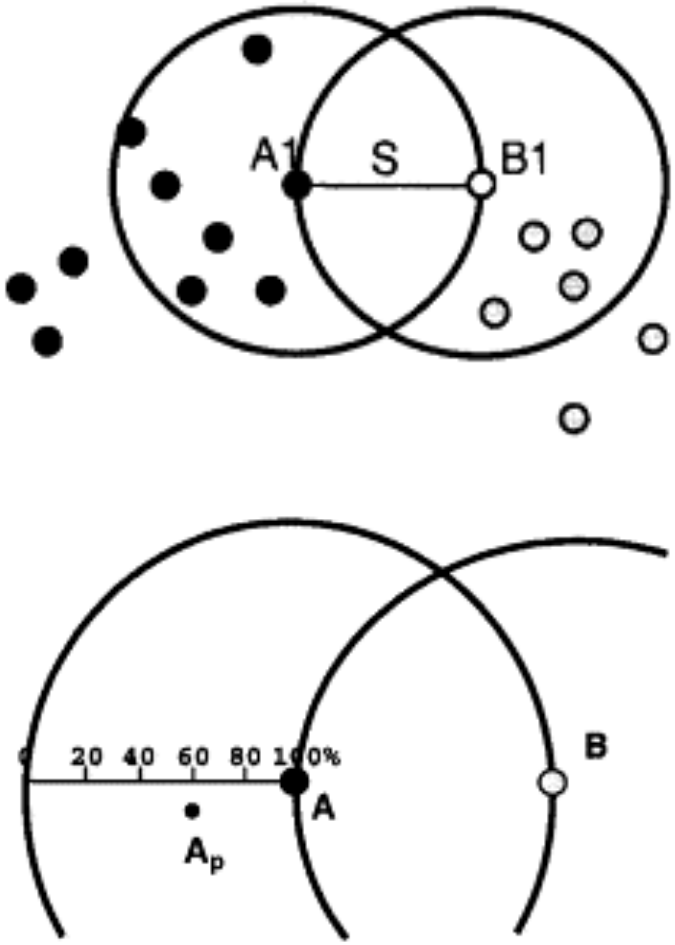
- Only pseudo-orthologs and xenologs will produce false positive orthologs

Orthologous groups

- Define groups of genes orthologous or co-orthologous to each other
  - Uses completely sequenced genomes
- Map protein or sequence fragment to these groups
- Groups of proteins connected by a speciation event
  - Can include paralogs – in- and out!

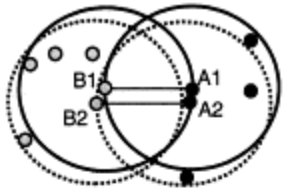
Inparanoid approach

- Main orthologs (mutually best hit) A1 and B1 with similarity score S.
- The main ortholog is more similar to in-paralogs from the same species than to any sequence from other species.
- Sequences outside the circle are classified as out-paralogs.
- In-paralogs from both species A and B are clustered independently.

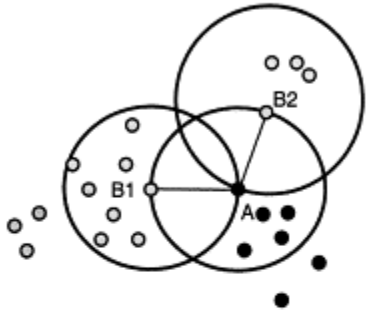


Rules for cluster refinement

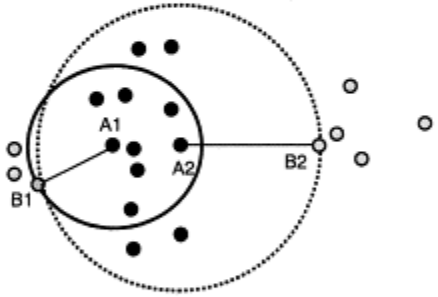
1) MERGE IF BOTH ORTHOLOGS ARE ALREADY CLUSTERED IN THE SAME GROUP



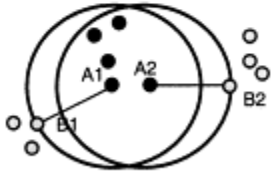
2) MERGE IF TWO EQUALLY GOOD BEST HITS FOUND



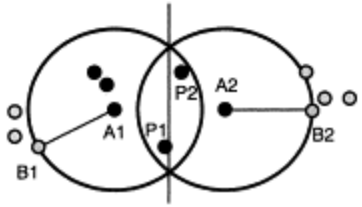
3) DELETE WEAKER GROUP IF (SCORE(A2-B2) - SCORE(A1-B1) > 50 bits)



4) MERGE IF (SCORE(A1-A2) < 0.5 * SCORE(A1-B1))



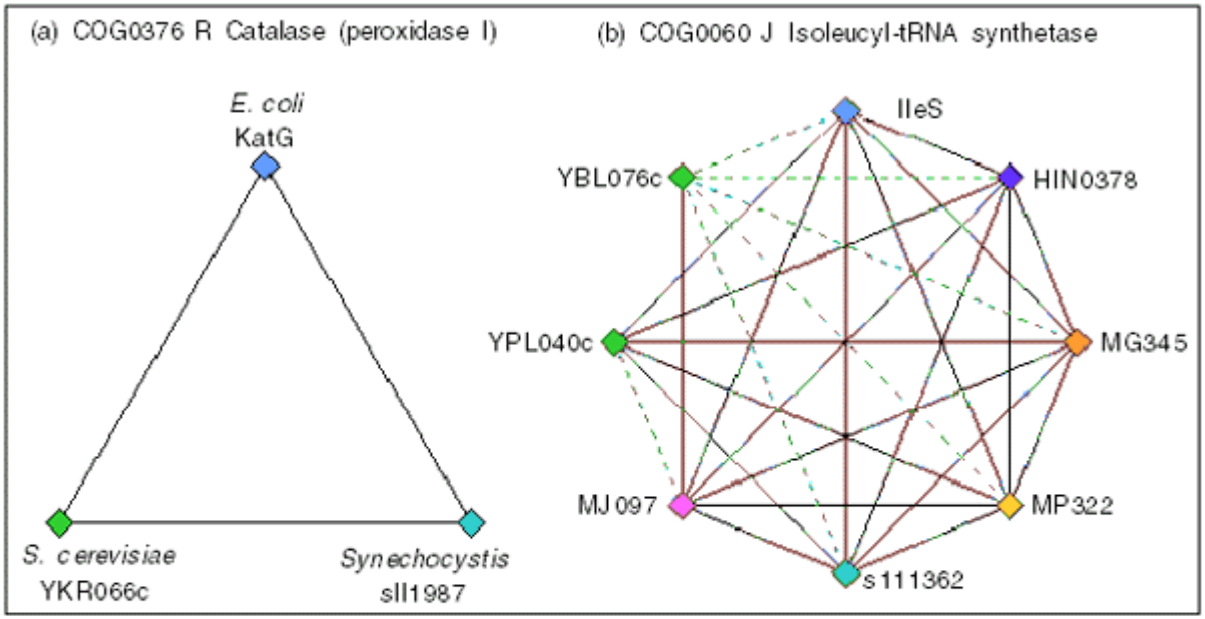
5) DIVIDE IN-PARALOGS IN OVERLAPPING AREAS



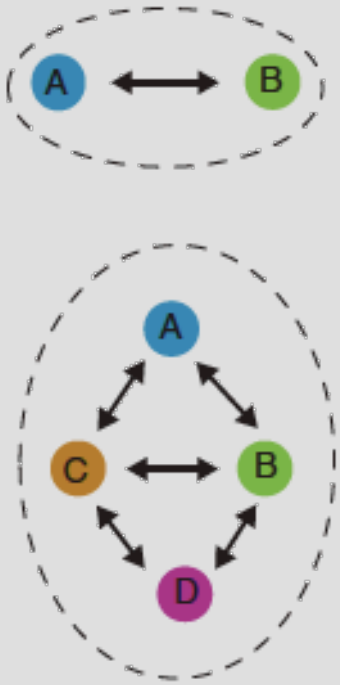
Minimal set of 50% similarity
over 50% of total length

COG database

- Pre-clear inparalogs
- Compute and extend the reciprocal best



Graph-based methods



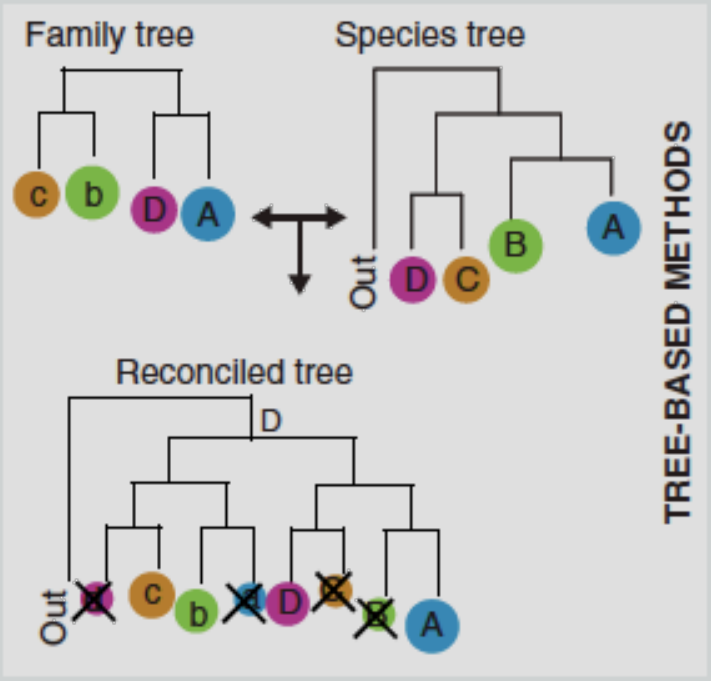
Pairwise species comparison

Multi-species comparison

GRAPH-BASED METHODS

	Phylogenetic distribution	Paralogs	Homology search	Clustering strategy	Hierarchical groups
BRH	ALL*	NO	BLAST	None	-
InParanoid	ALL*	YES	BLAST	None	-
RoundUp	ALL*	NO	Evol. Distance	None	-
COG	ALL	YES	BLAST	Triangles	NO
eggNOG	ALL	YES	BLAST	Triangles	YES
OrthoDB	Eukaryotes	YES	PARALIGN	Triangles	YES
OMA	ALL	YES**	SIMD & Evol. Distance	Maximum weight cliques	YES
OrthoMCL	ALL	YES	BLAST	Markov Clustering	NO

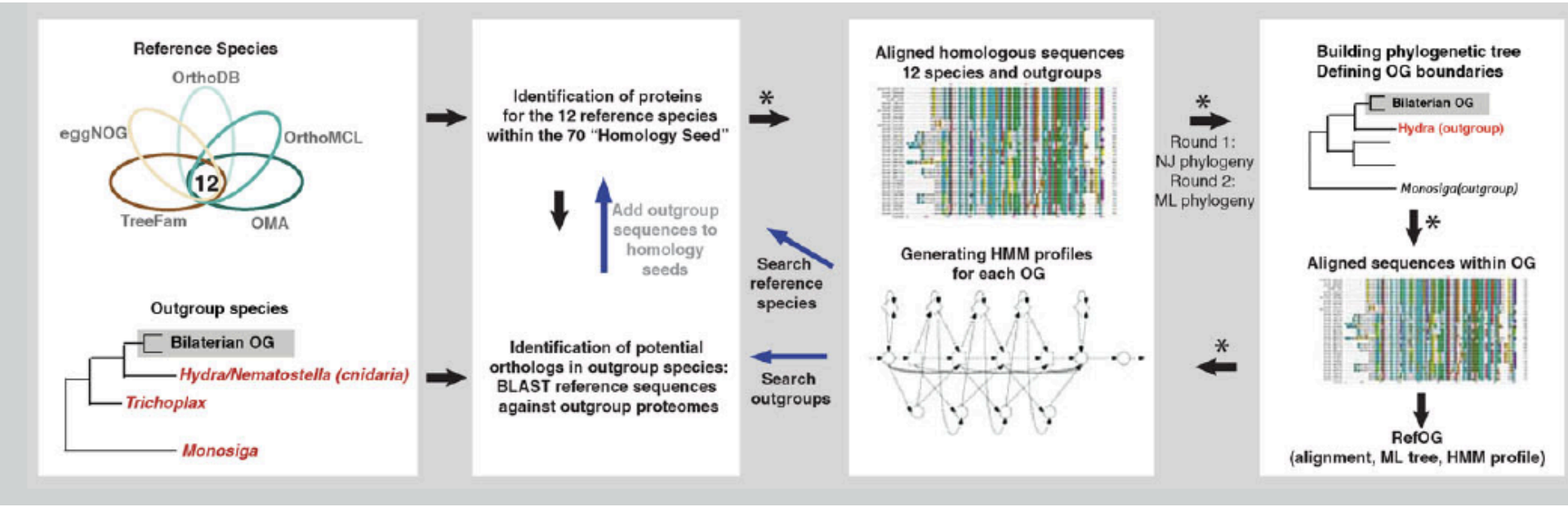
Tree-based methods



TREE-BASED METHODS

TreeFam	Metazoa	YES	BLAST & HMM	Hierarchical clustering	-
Ensembl Compara	Metazoa	YES	BLAST	Hierarchical clustering	-
PhylomeDB	ALL*	YES	BLAST§	None	-

Benchmarking



Trachana et al (Oct. 2011) BioEssays

Tree reconciliation

What is tree reconciliation?

- Bringing the species and the gene tree in congruence
- Mapping duplication and speciation events to a phylogenetic tree
- Several methods exist
- Goodman et al. (1979) described a first algorithm
- Relies on a known species tree and (correct) rooted, binary gene trees

Tree reconciliation (Goldman)

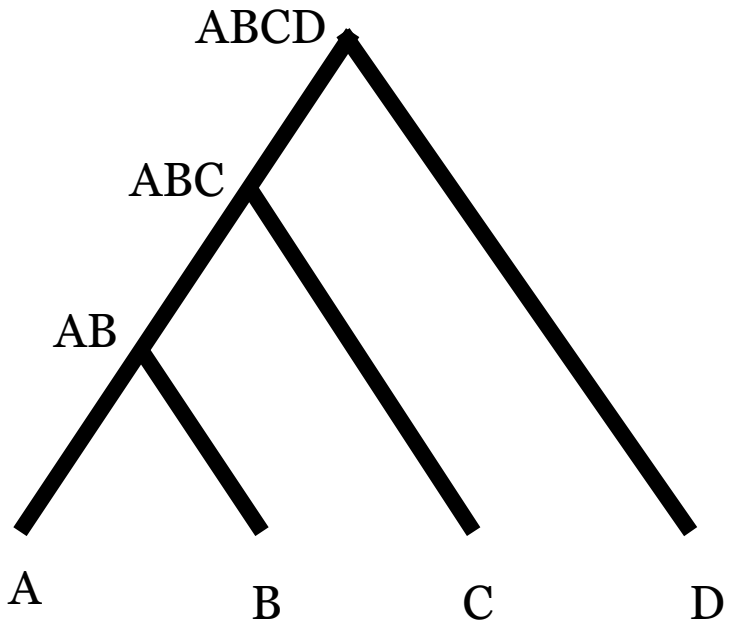
- Label internal nodes of the gene tree
- Label internal nodes of the species trees according to the labels in the gene tree
- Traverse the tree, labeling internal nodes as speciation events of duplication events

Procedure

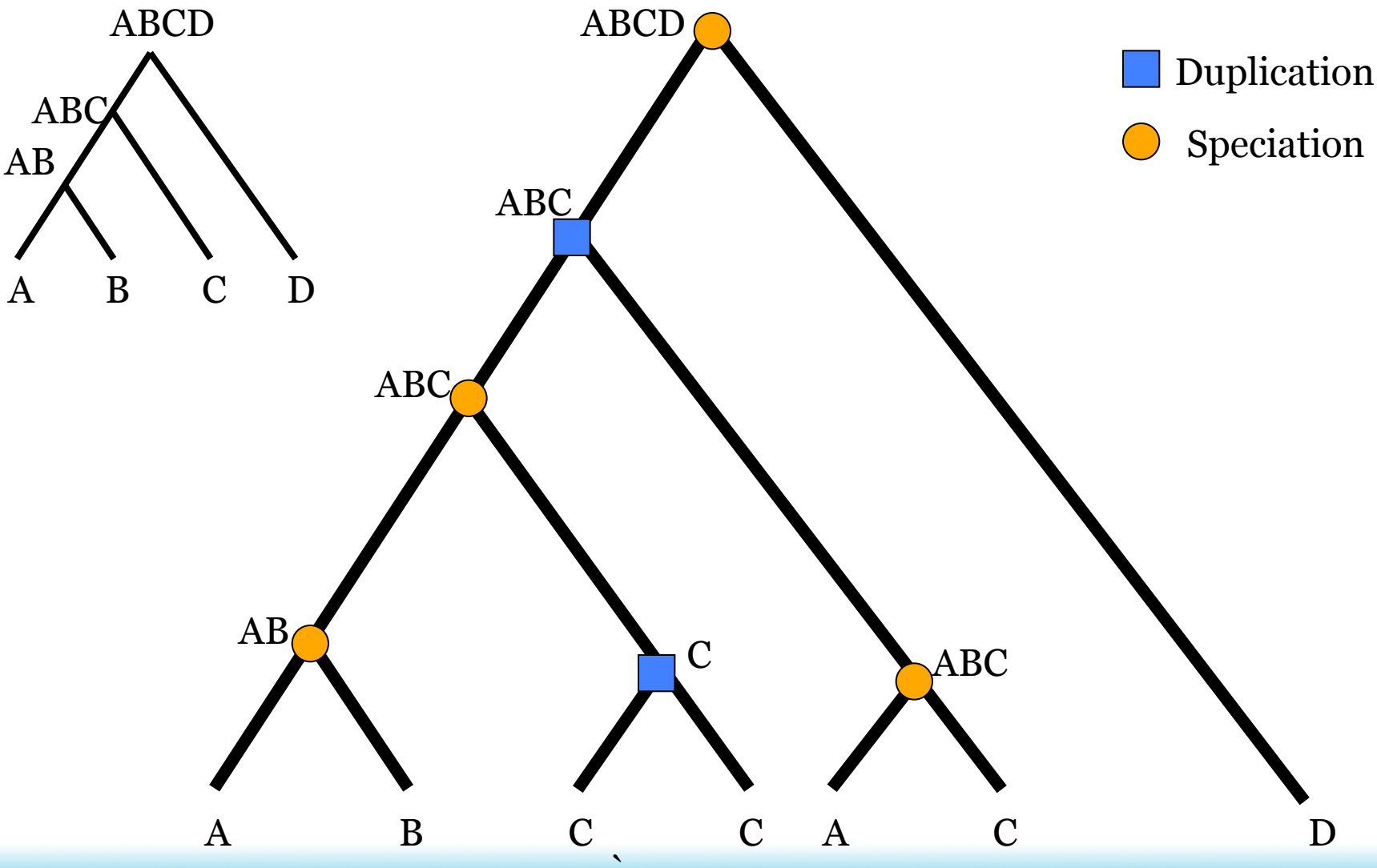
- **Definition Labeling.** Let G be the set of nodes in a rooted binary gene tree and S the set of nodes in a rooted binary species tree. For any node $g \in G$, let $\gamma(g)$ be the set of species in which occur the extant genes descendant from g. For any node $s \in S$, let $\sigma(s)$ be the set of species in the external nodes descendant from s. For any $g \in G$, let $M(g) \in S$ be the smallest (lowest) node in S satisfying $\gamma(g) \in \sigma(M(g))$.

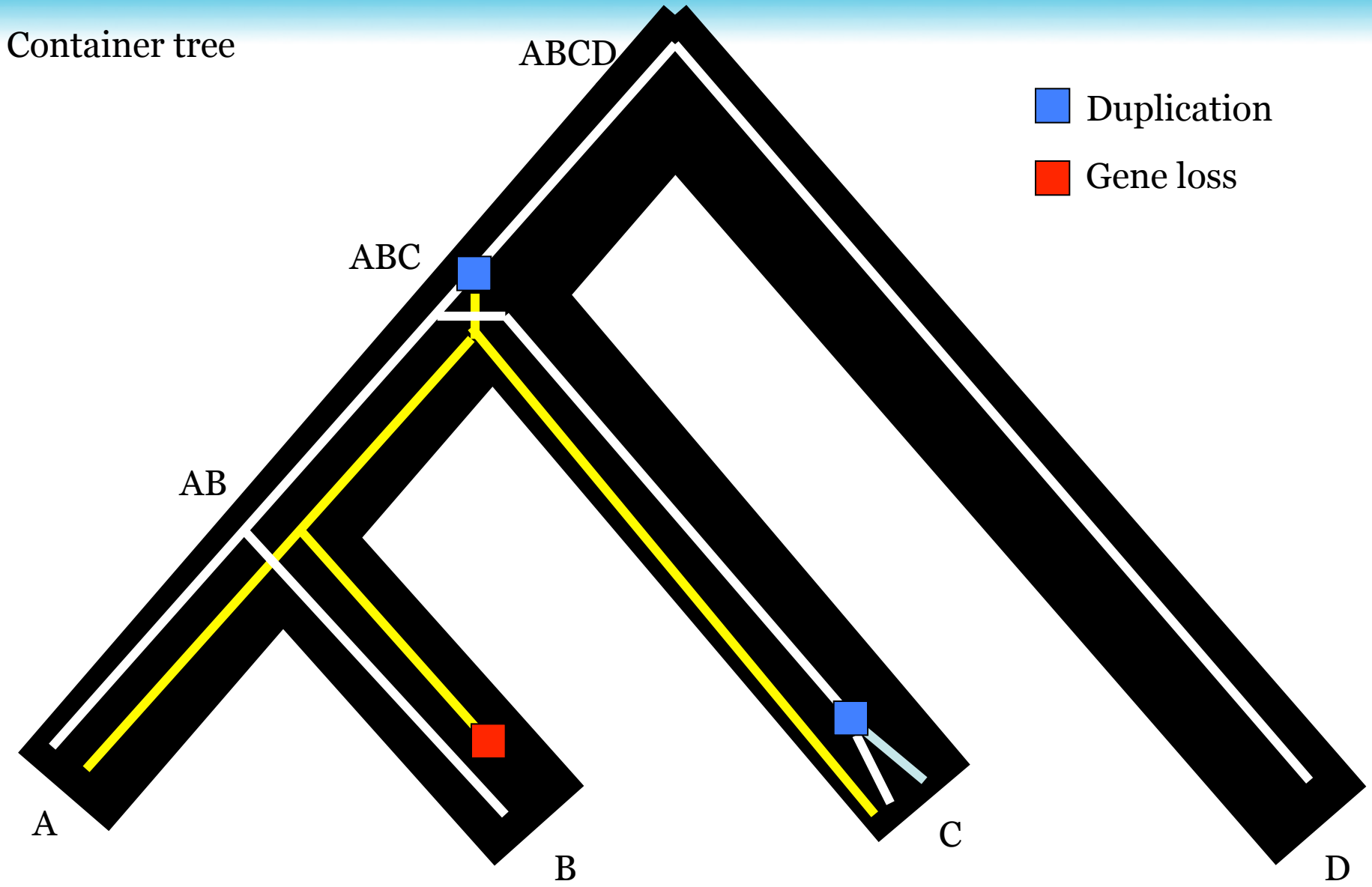
- **Definition Duplication mapping.** Let $g1$ and $g2$ be the two child nodes of an internal node g of a rooted binary gene tree G. Node g is a duplication if and only if $M(g) = M(g1)$ or $M(g) = M(g2)$.

Species tree



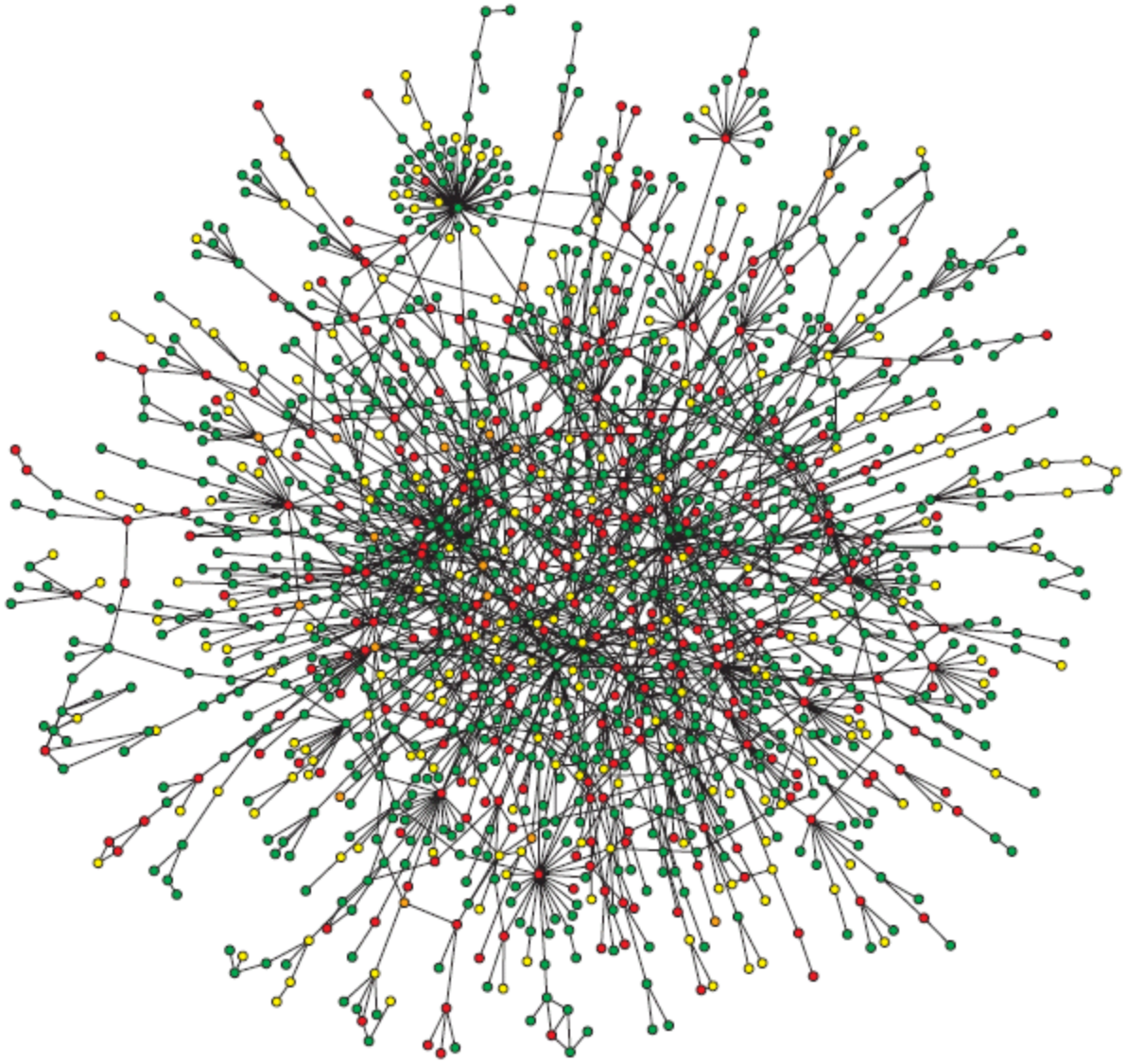
Gene tree

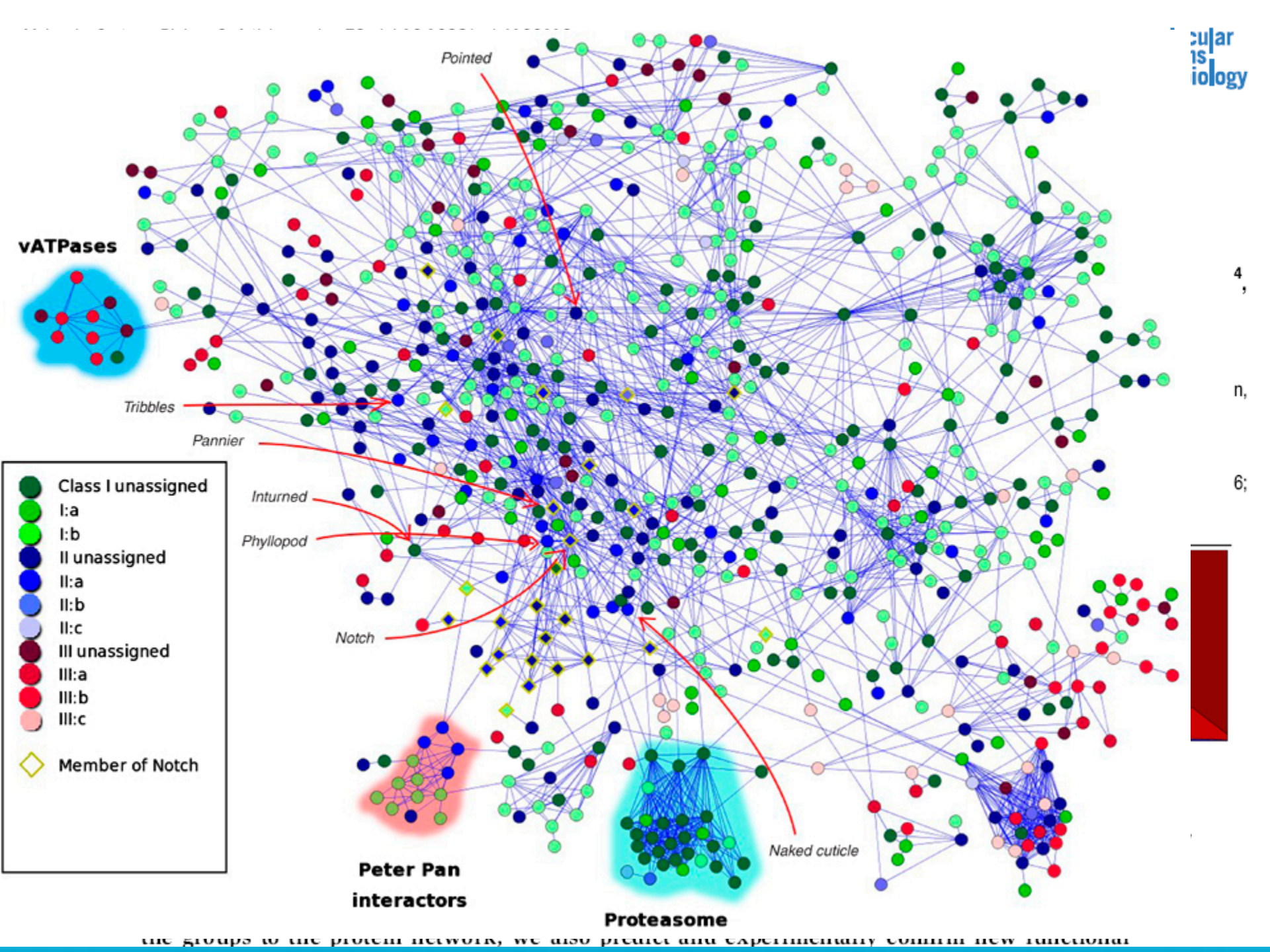

ABCD
ABC
AB
A
B
C
D
ABCD
ABC
ABC
AB
C
ABC
A
B
C
C
A
C
D
Duplication
Speciation


Container tree
ABCD
Duplication
Gene loss
ABC
AB
A
B
C
D

Applied phylogeny in bioinformatics

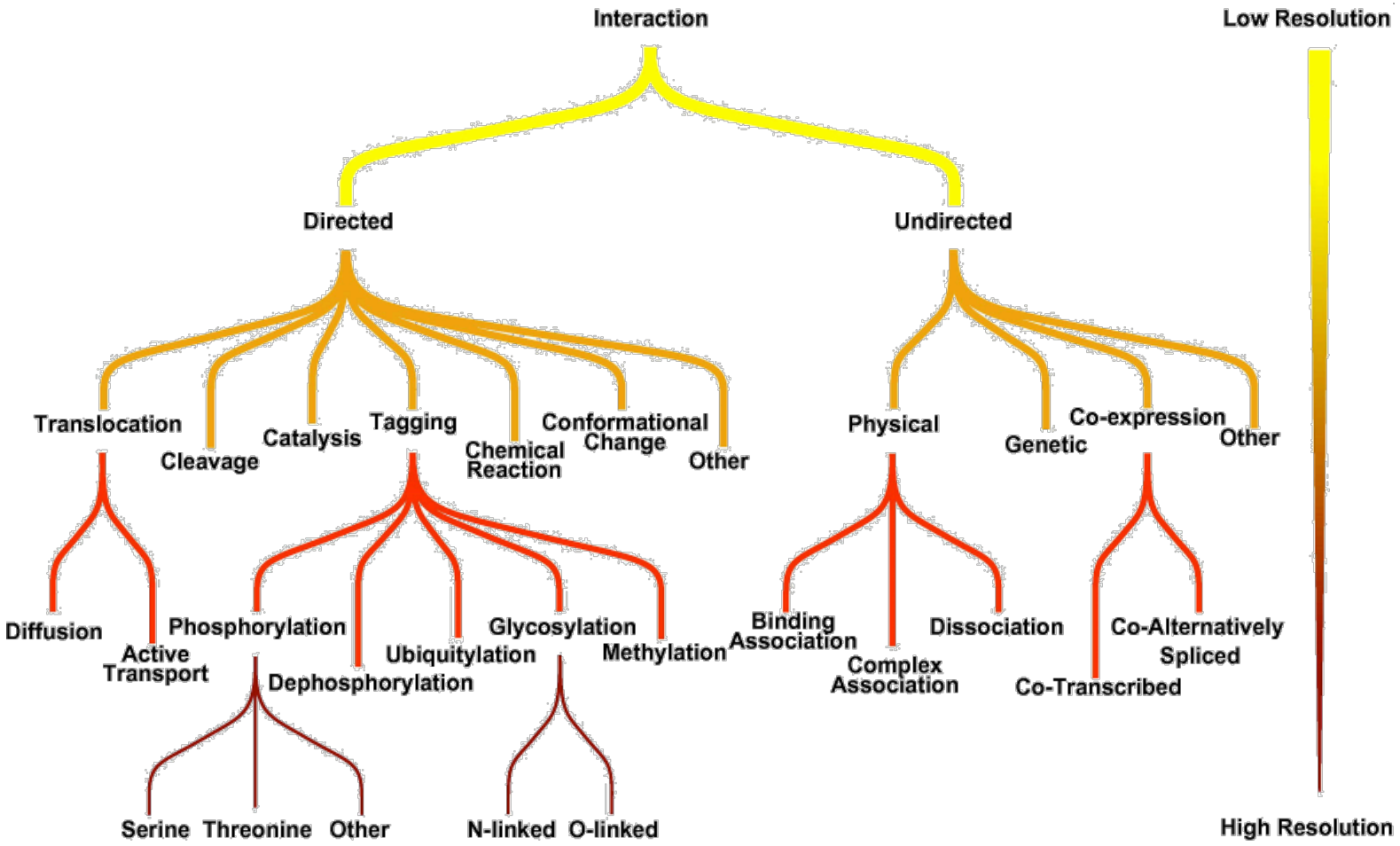
Prediction
of functional interactions





the groups to the protein network, we also predict and experimentally confirm new functional

Biological types of interactions



A proposed ontology for interactions (Lu et al.)

Experimental techniques

High-thoughput methods

- Yeast two-hybrid
- Co-immuno-precipitation (TAP)
- Protein fragment-complemention assay
- Genetic interactions
- Surface plasmon resonance (Biacore)

Bioinformatics predictions

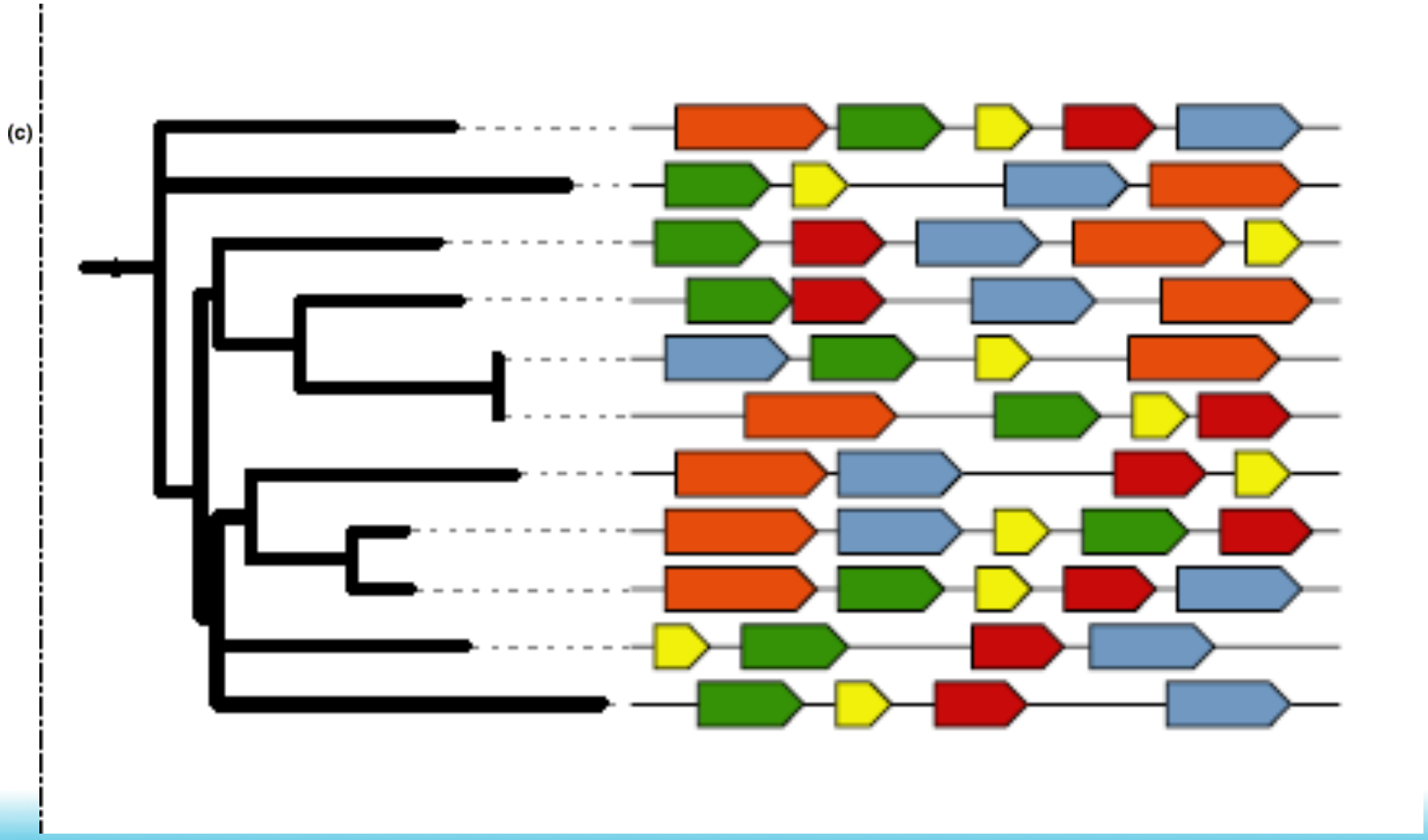
- *Genomic context methods*
- Gene expression
- Computational inference from sequence (machine learning)
- Inference from 3-D structure

Hypothesis generation of protein function

- Homology-based methods
  - BLAST
  - Domain databases
  - Interaction prediction from sequence
  - Typical inference: Enzymatic function
  - Molecular function
- Genome-based methods
  - Protein-protein interaction
  - Operon structures
  - Phylogenetic profiles
  - Protein domain fusion events
  - Typical inference: involved in a metabolic pathway
  - Biological process

Gene neighborhood

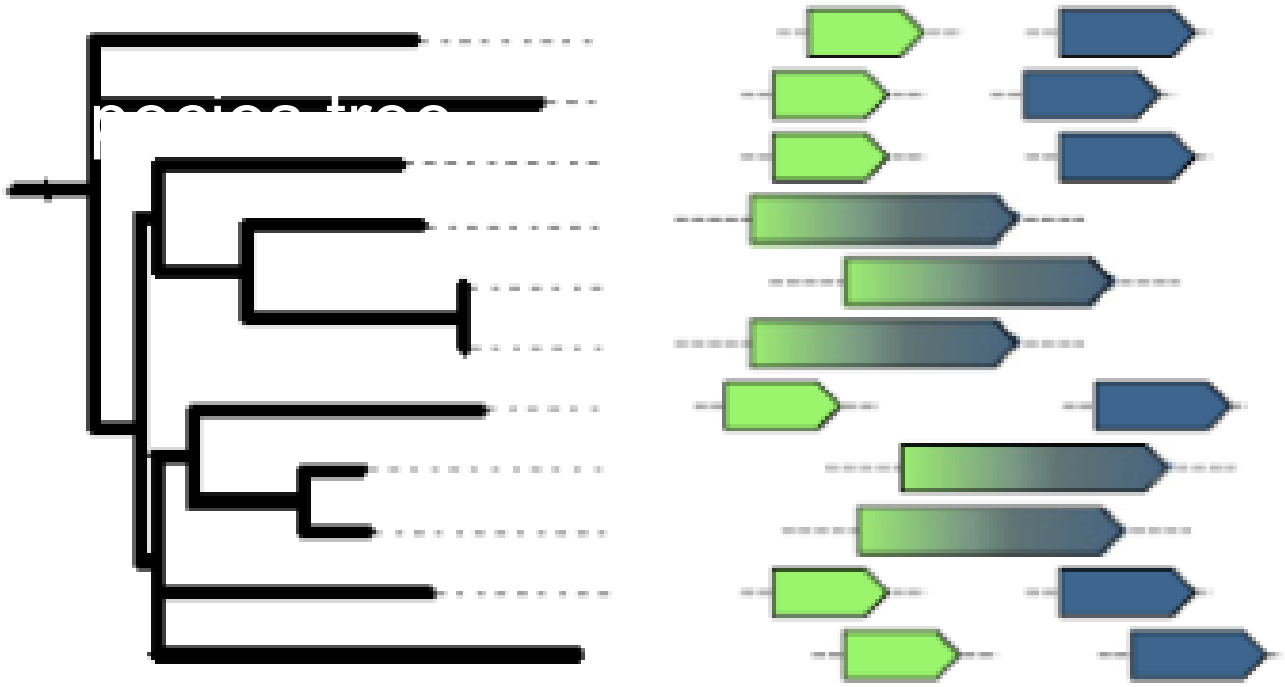
- Operons and Über-operons



Deriving interactions

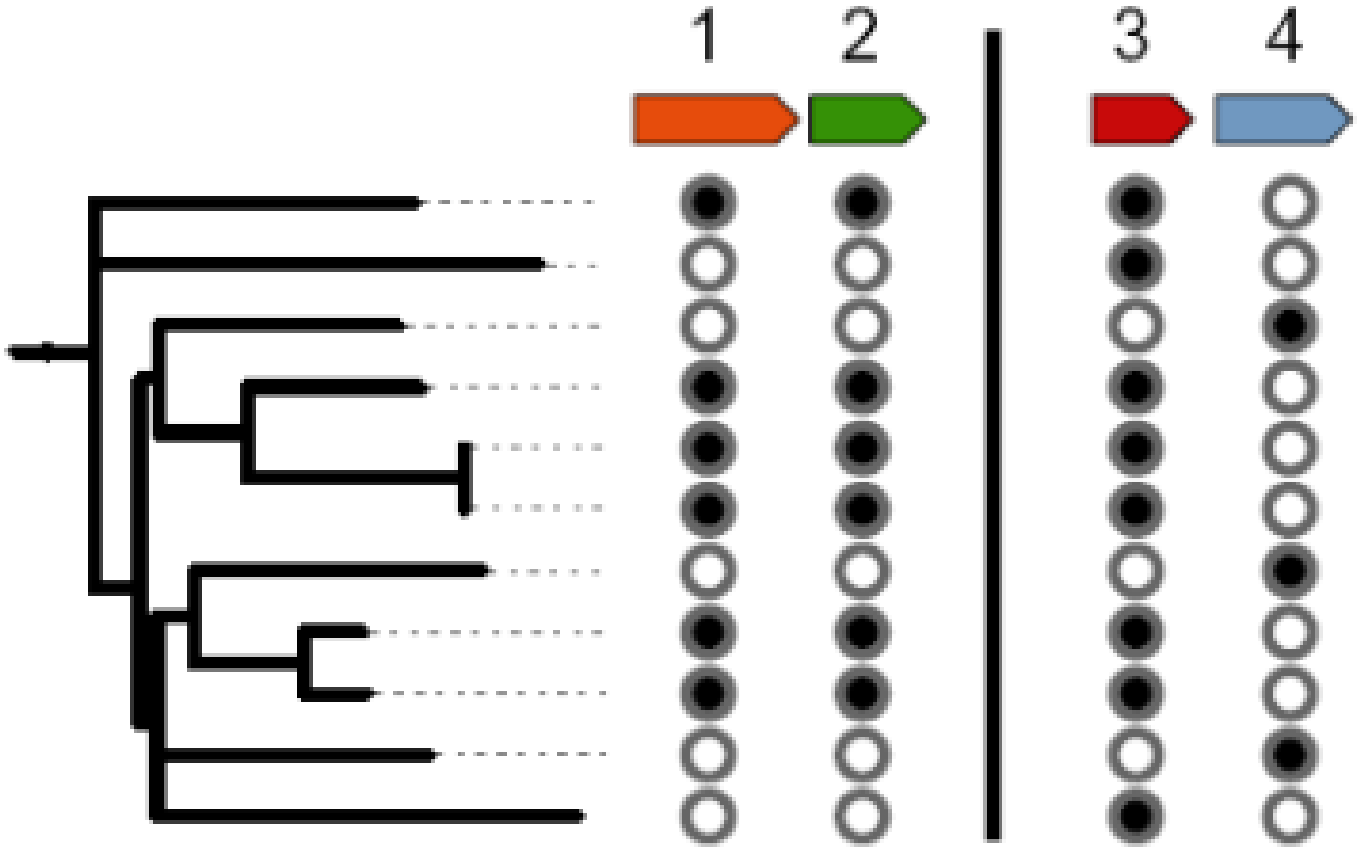
- **Operon prediction**
  - Intergenic distances provide strong signal
  - In *E. coli* 300 nt
  - Additional data
    - *Microarray expression*
- **Gene neighborhood**
  - No explicit operon prediction required
  - Conservation across 500+ genomes provides strong signal
  - Simple to compute

Gene fusion

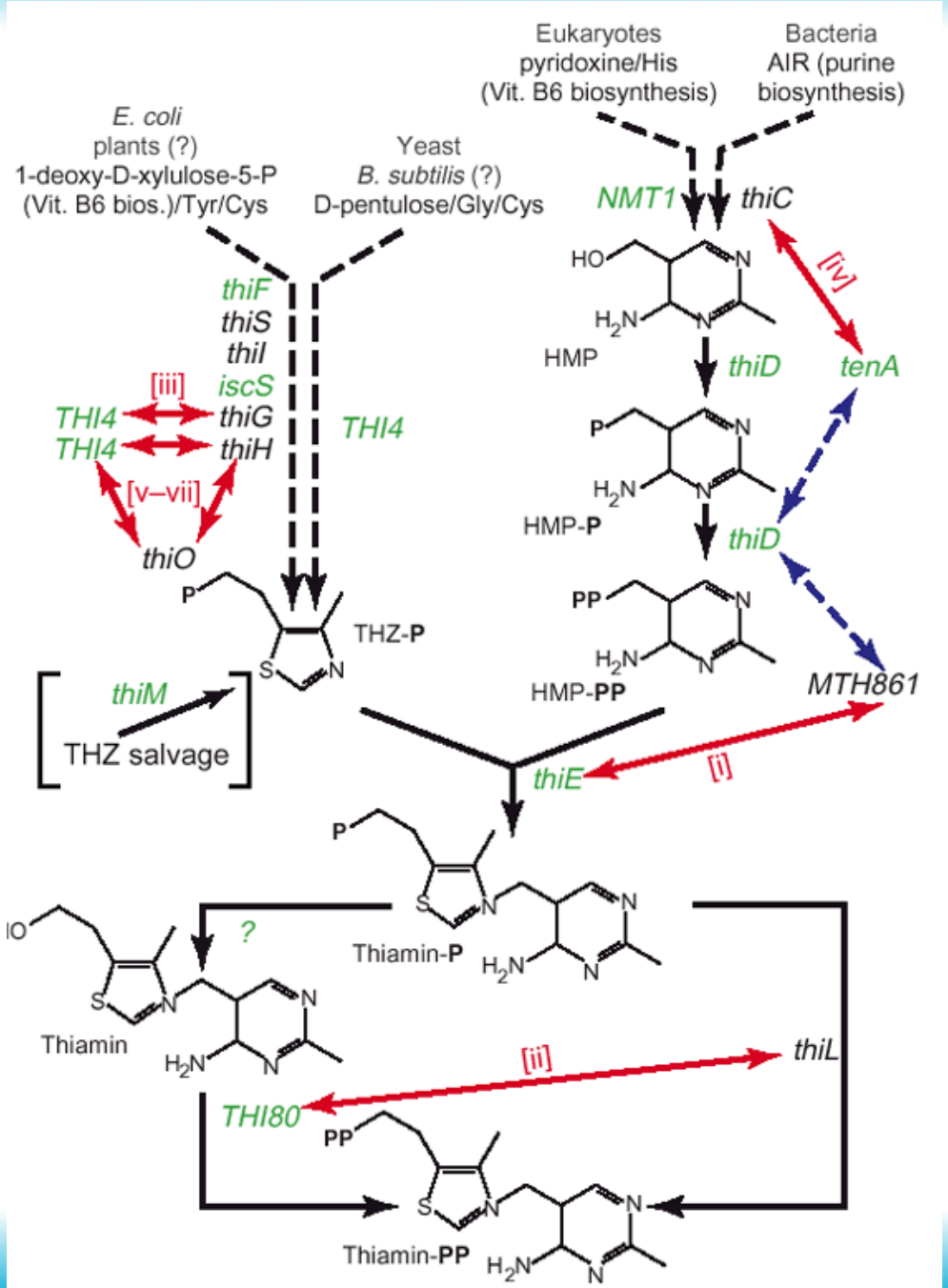


Phylogenetic profiles

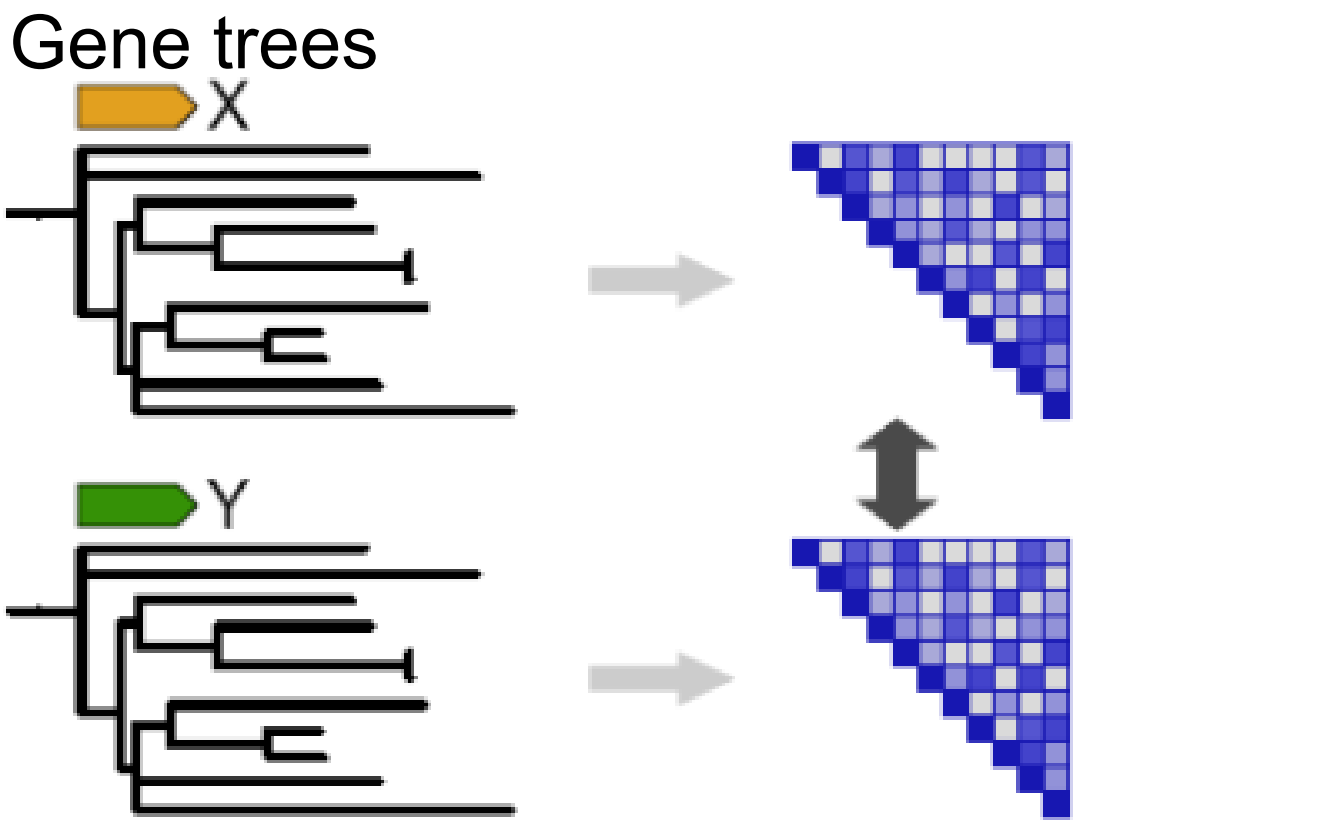
- Pellegrini et al. (1999)



- Thiamine biosynthesis
  - Discovery of an alternative pathway
  - Morett, Korbel Nat. Biot. (2003)

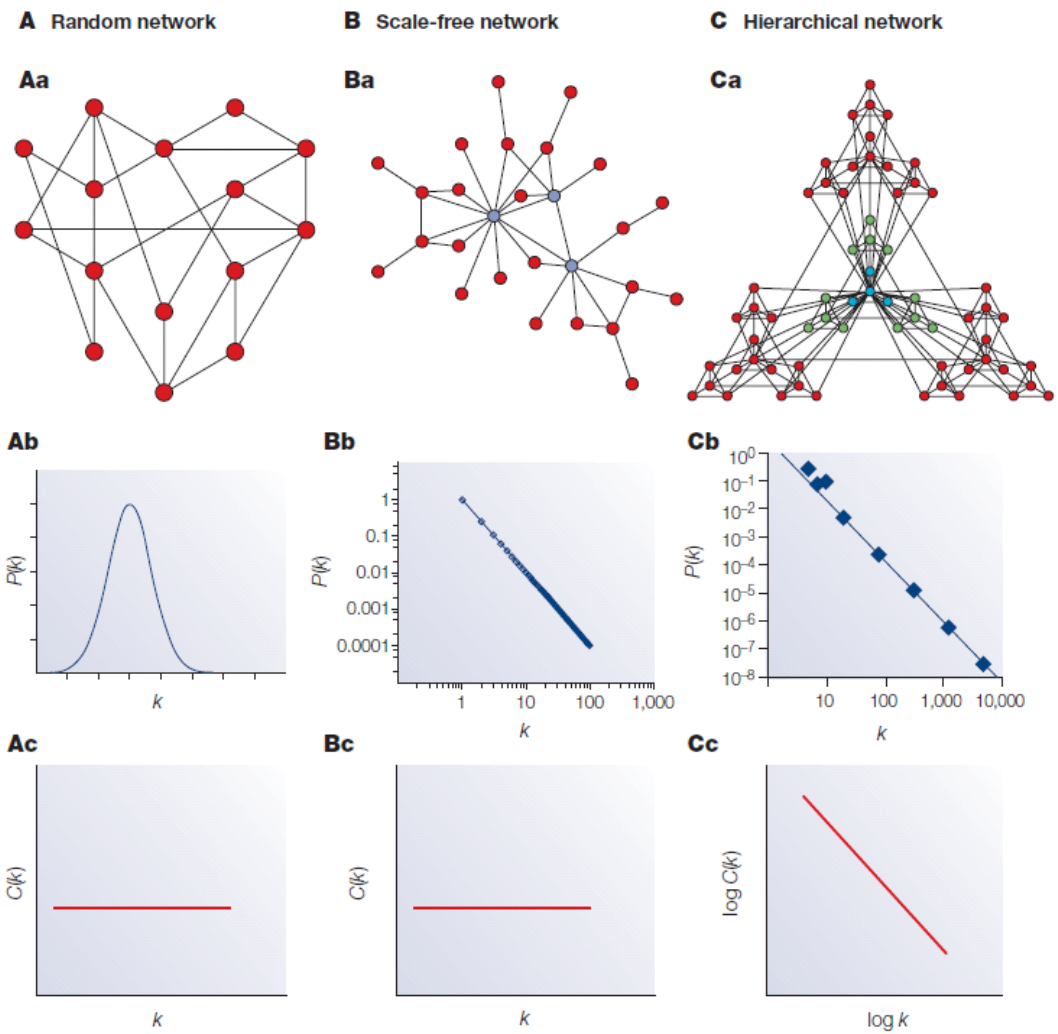


Sequence co-evolution



Different networks

From Barabási (2004), Nature Reviews Genetics



Effect of sampling on topology predictions of protein-protein interaction networks

Jing-Dong J Han[1–3], Denis Dupuy[1,3], Nicolas Bertin[1], Michael E Cusick[1] & Marc Vidal[1]

Currently available protein-protein interaction (PPI) network or 'interactome' maps, obtained with the yeast two-hybrid (Y2H) assay or by co-affinity purification followed by mass spectrometry (co-AP/MS), only cover a fraction of the complete PPI networks. These partial networks display scale-free topologies–most proteins participate in only a few interactions whereas a few proteins have many interaction partners. Here we analyze whether the scale-free topologies of the partial networks obtained from Y2H assays can be used to accurately infer the topology of complete interactomes. We generated four theoretical interaction networks of different topologies (random, exponential, power law, truncated normal). Partial sampling of these networks resulted in sub-networks with topological characteristics that were virtually indistinguishable from those of currently available Y2H-derived partial interactome maps. We conclude that given the current limited coverage levels, the observed scale-free topology of existing interactome maps cannot be confidently extrapolated to complete interactomes.

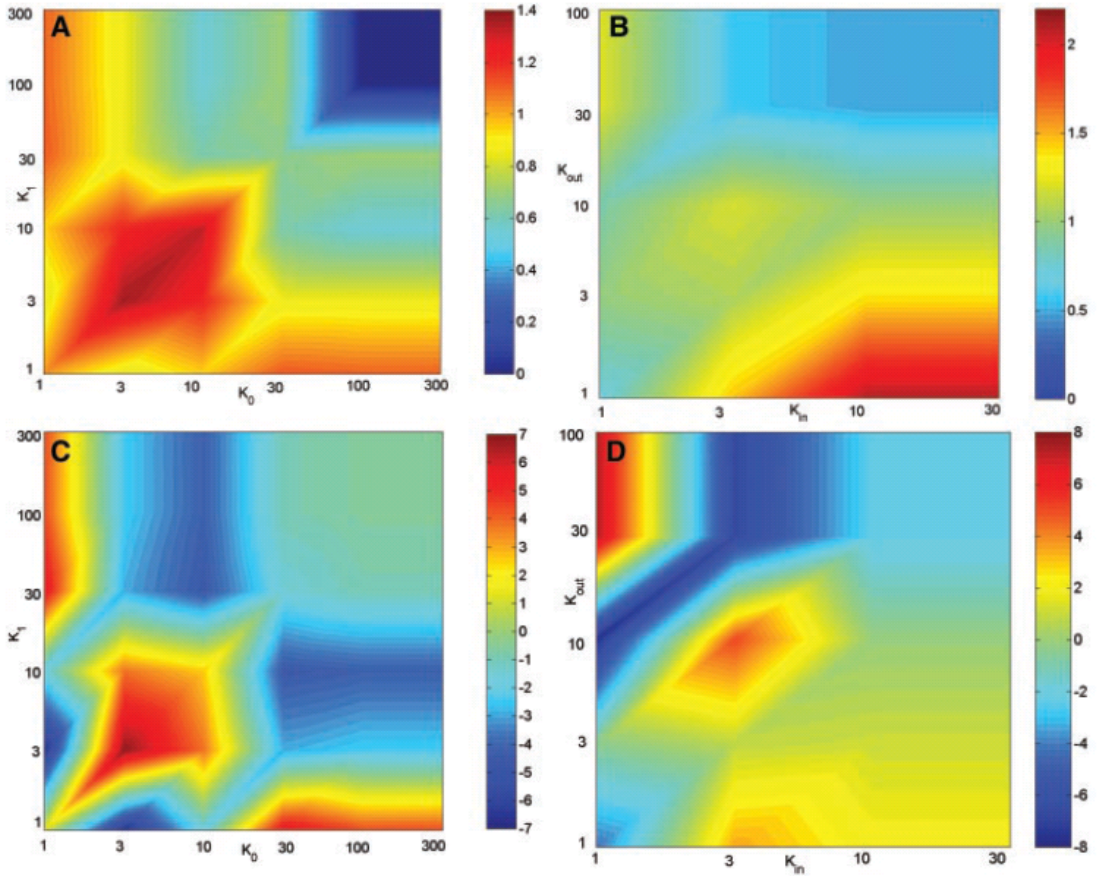
data sets[11,12] contain putative interactions, predicted on the basis of co-membership in a protein complex. Thus, a significant fraction of the PPIs reported in each of these maps may correspond to indirect interactions that would lead to a significant overestimation of their actual coverage. The yeast interactome maps generated by analyzing direct binary interaction assays with the Y2H assay independently cover a mere 3–9% of the complete interactome (948 and 806 defined in the Uetz and Ito core Y2H maps respectively[5,6]) (**Table 1**). The *Caenorhabditis elegans* and *Drosophila melanogaster* Y2H interactome maps show similar limited coverage[8,9]. This low coverage can explain the limited overlap observed between large-scale yeast Y2H data sets[5,6,14,18], between large-scale co-AP/MS data sets[11,12,18] and between *D. melanogaster* Y2H data sets[8,19] (see **Box 2**). To extrapolate the topology of complete interactomes from such incomplete maps requires the assumption that the limited sampling does not affect the overall topological analyses[20]. Recent reports have already noted discrepancies in matching existing interactome networks to a scale-free topology[21,22].

Here, we analyze whether extrapolation of network topologies from

Connections between hubs

Maslov and Sneppen (2002)
Science

Hubs are connected to proteins of low degree, not between each other



Thank you for your attention!