Trees, Graphs and the Evolution of Sequence Families

Dannie Durand

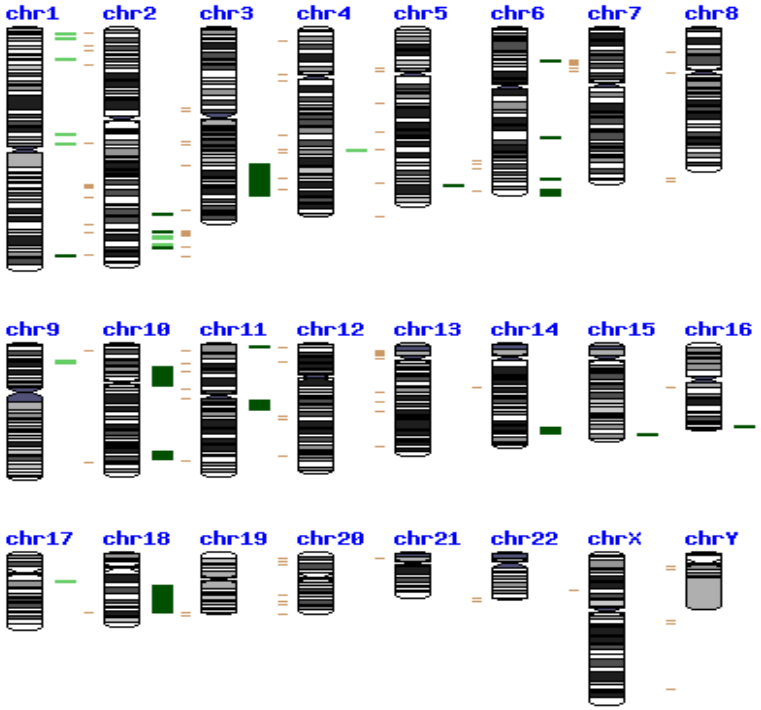
Biological Sciences and Computer Science

Carnegie Mellon University

Outline

- Background: genome evolution
- Graph and trees for sequence families
- Graph theoretical insights into multidomain protein evolution

Foci of genome evolution



Genome sequence

- Coding and coding sequence

Non-coding sequence

- regulation, transposable elements...

All genes

- Intron exon structure, splicing, third positions...

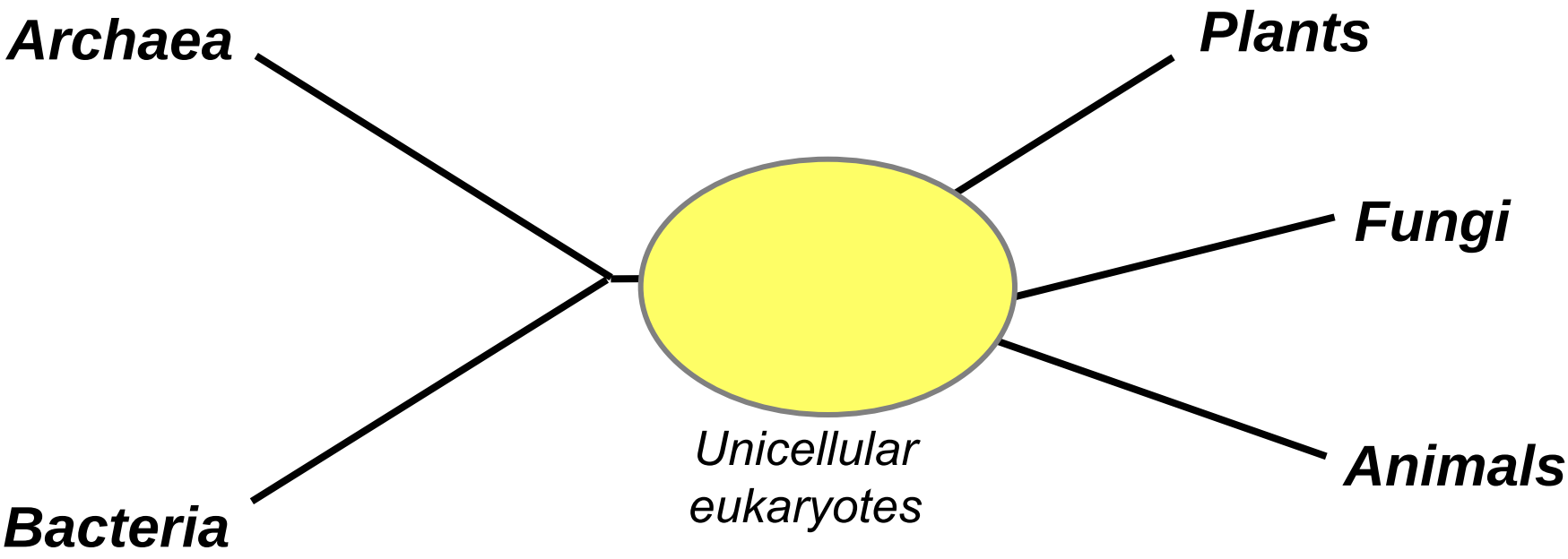
All amino acid sequences

- Gene products

Set of all proteins

- Structures

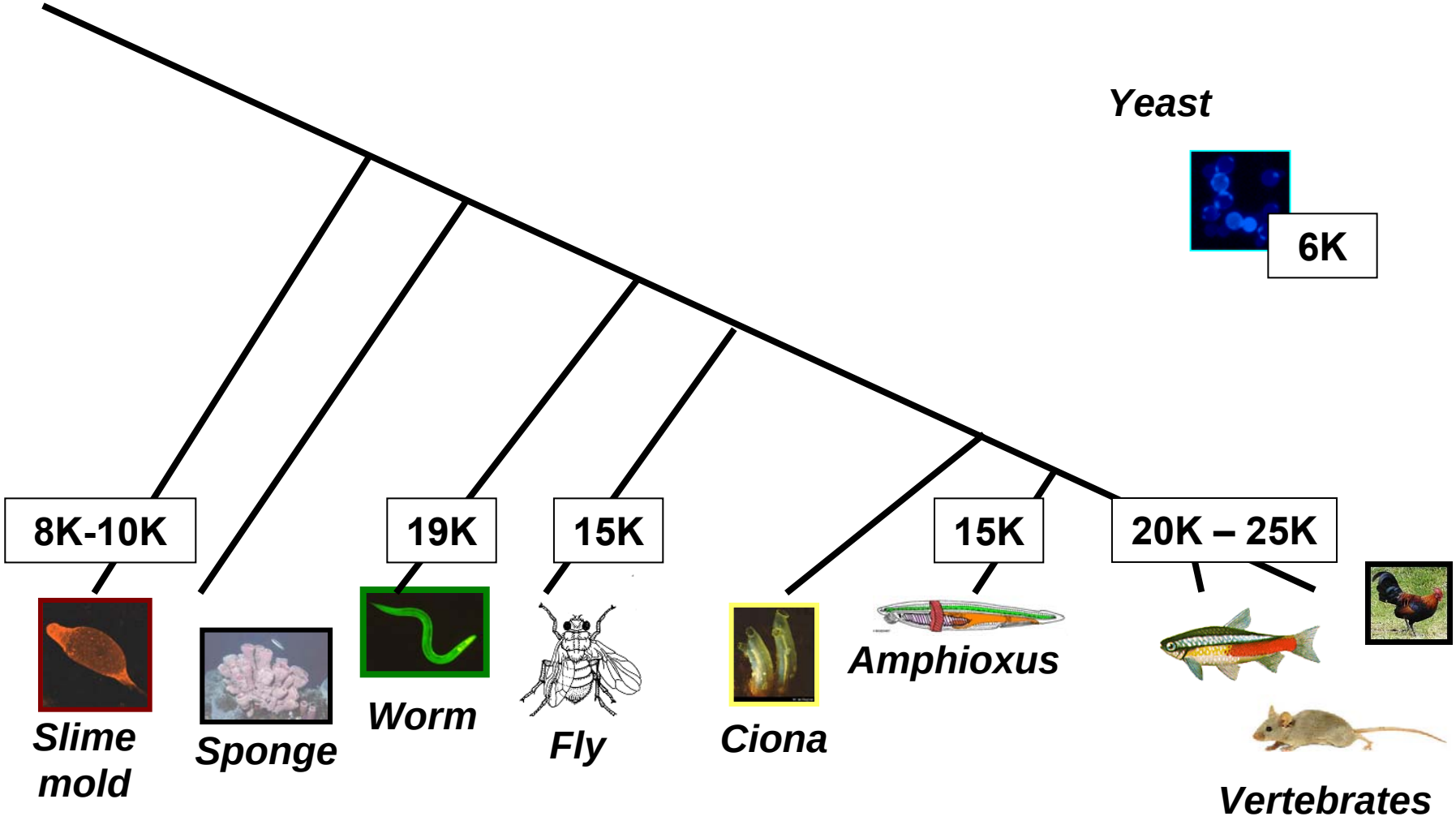
Evolution of the parts list



How does the complement of protein coding sequences change over evolutionary time?

How is this related to the emergence of novel cellular processes? New morphologies? Species evolution?

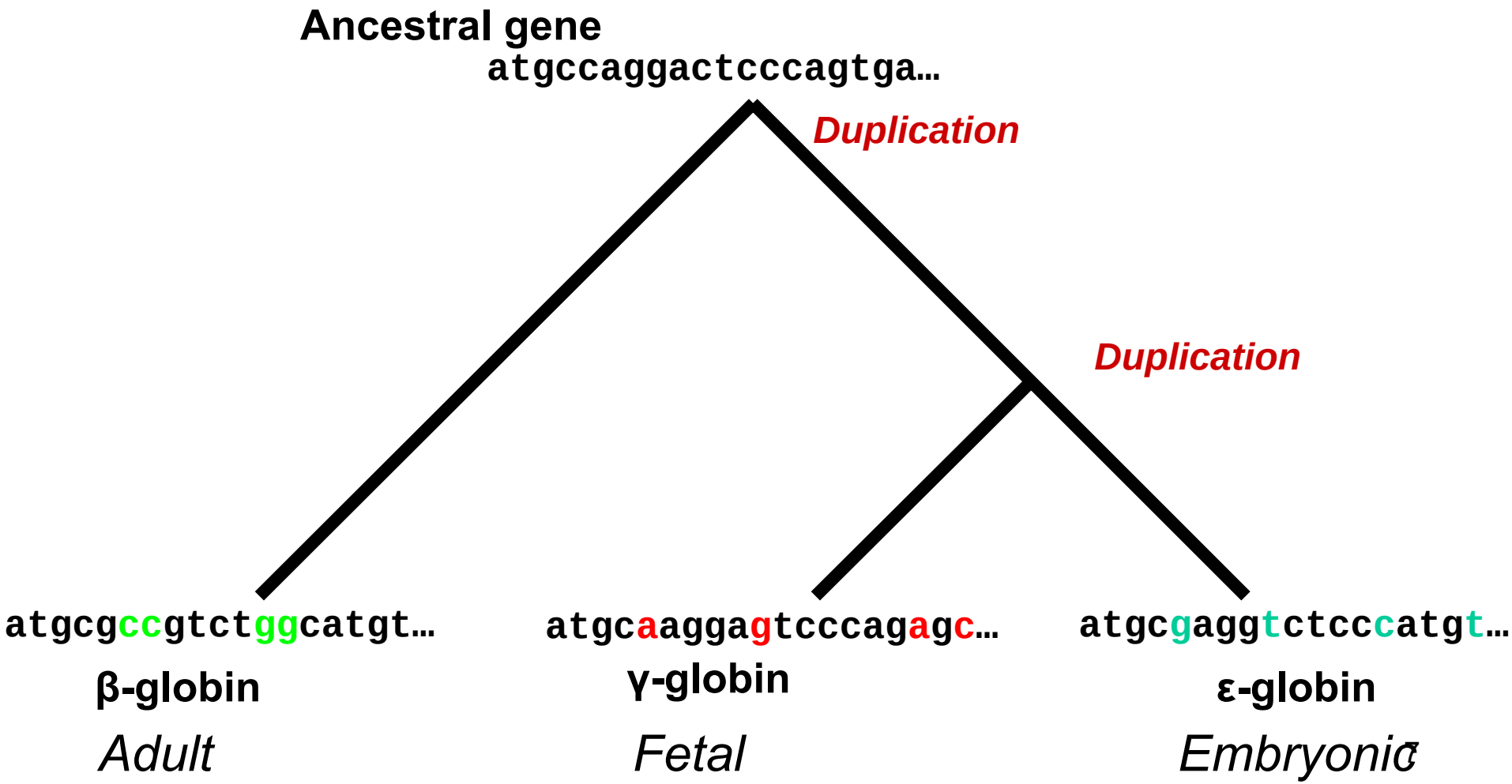
Animal evolution and genome sizes



Where do new genes come from?

- Gene duplication
- Domain shuffling

New genes arise through duplication and modification of existing genes

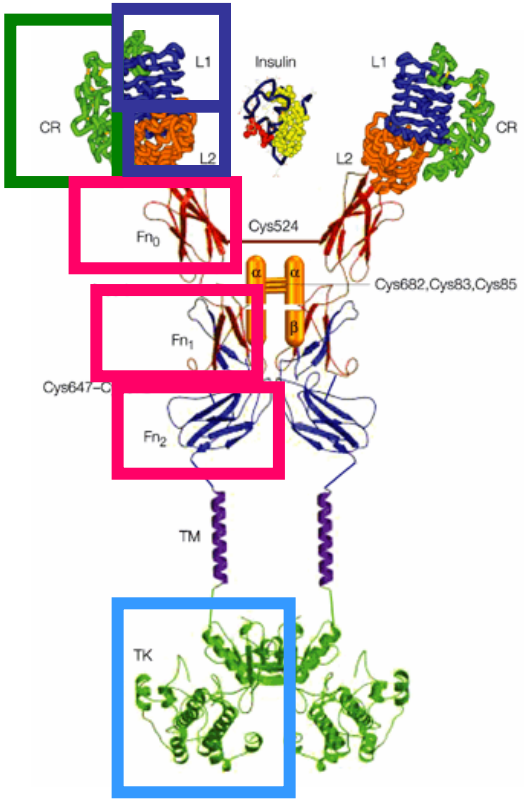


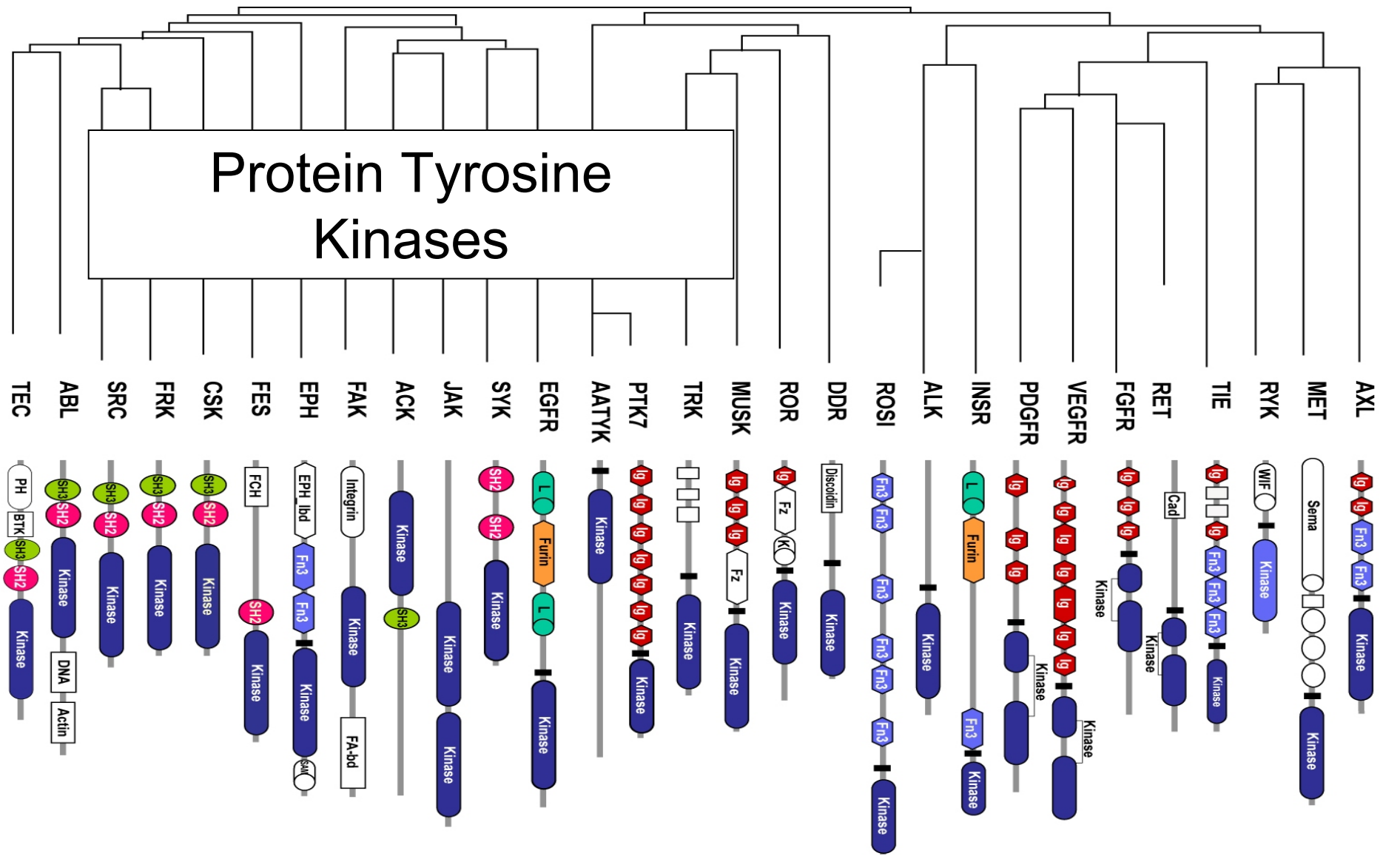
New genes arise through insertion, duplication and rearrangement of domains

Domains:

- Sequence fragments
- Fold independently
- Carry out specific functions
- Found in diverse contexts

Insulin receptor



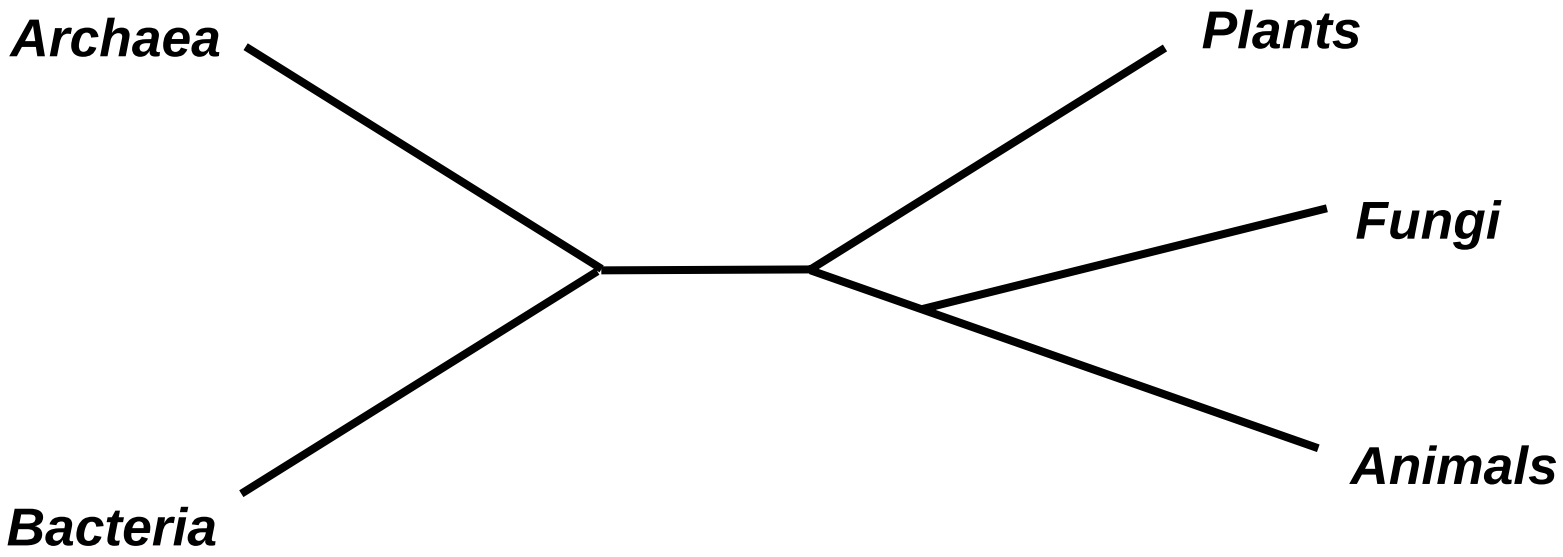


Adapted from Robinson *et al.*, 2000

Multidomain protein statistics

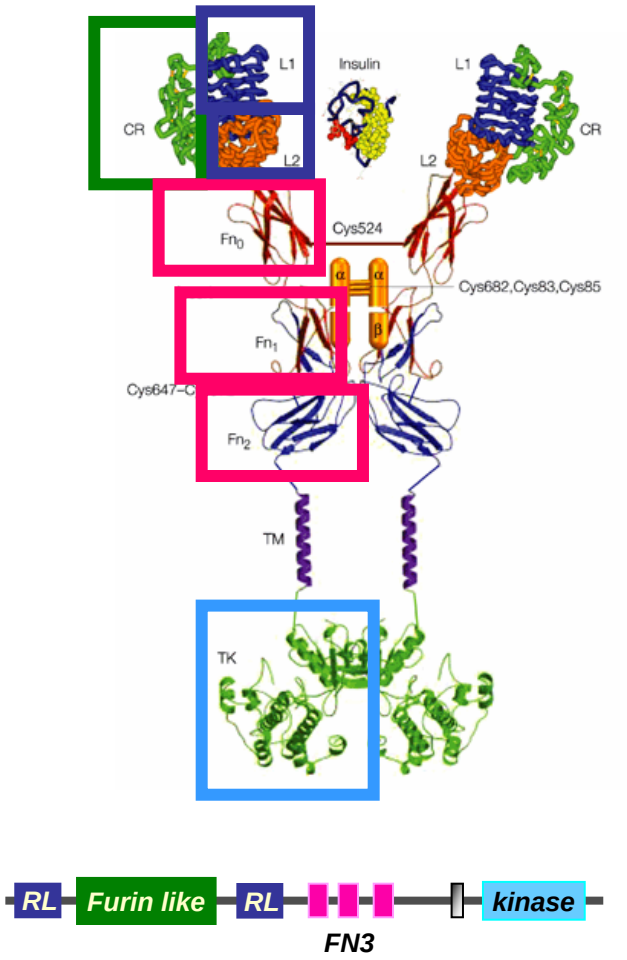
domains	≥2	≥4	≥6	≥11
Bacteria	26.7%	3.9%	1.2%	0.3%
Archaea	23.3%	3.4%	1.1%	0.2%
Plants	35.3%	8.0%	3.8%	1.1%
Fungi	31.6%	6.9%	2.7%	1.2%
Animals	39.3%	11.1%	5.6%	1.7%

Tordai et al., 05



Multidomain Sequences

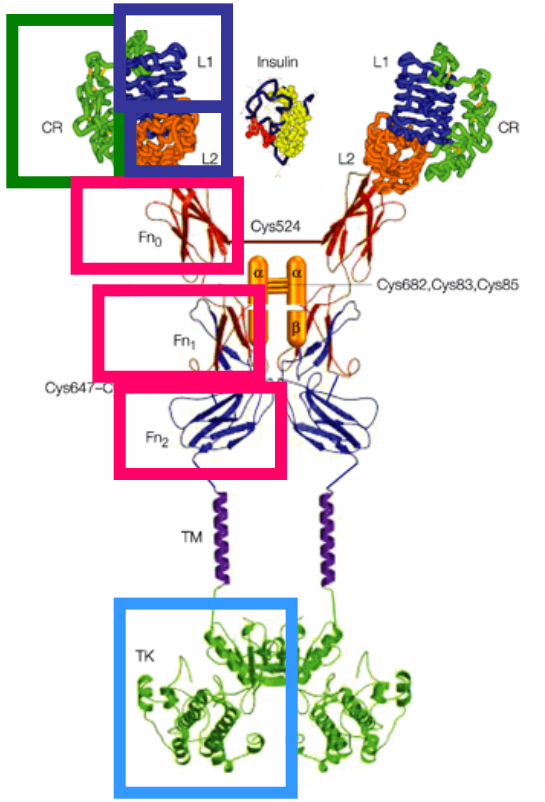
Insulin receptor



- Nature's equivalent of rapid prototyping
- Expanded preferentially in
  - animals
  - vertebrates
- Functional roles:
  - Cell signaling
  - Cell-cell adhesion
  - Tissue repair
  - Cell death
  - Immune response

Questions about Multidomain Evolution

Insulin receptor

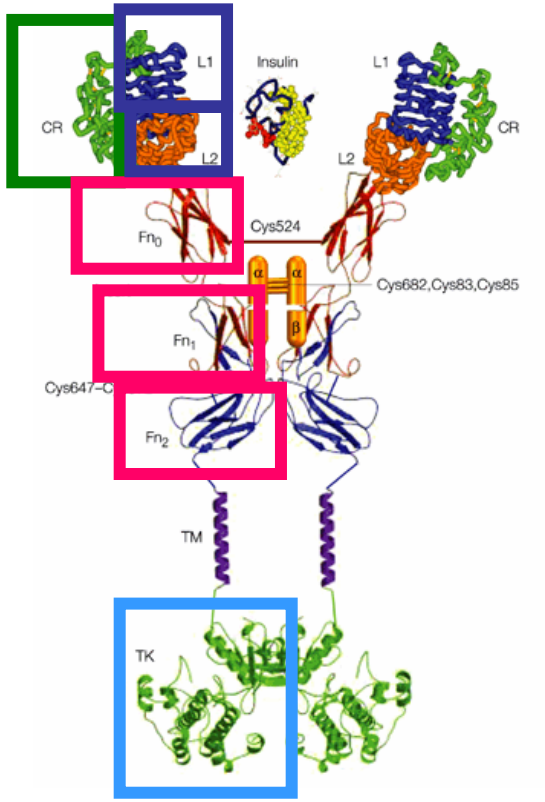


FN3

- Processes of domain acquisition and loss
- Rates of domain acquisition and loss
- Constraints on domain organization
- Evolutionary opportunities offered by domain shuffling

Questions about Multi-Domain Evolution

Insulin receptor



FN3

- Processes of domain acquisition and loss
- Rates of domain acquisition and loss
- Constraints on domain organization
- Evolutionary opportunities offered by domain shuffling
- Are domain insertions rare?
- Do domain architectures persist?

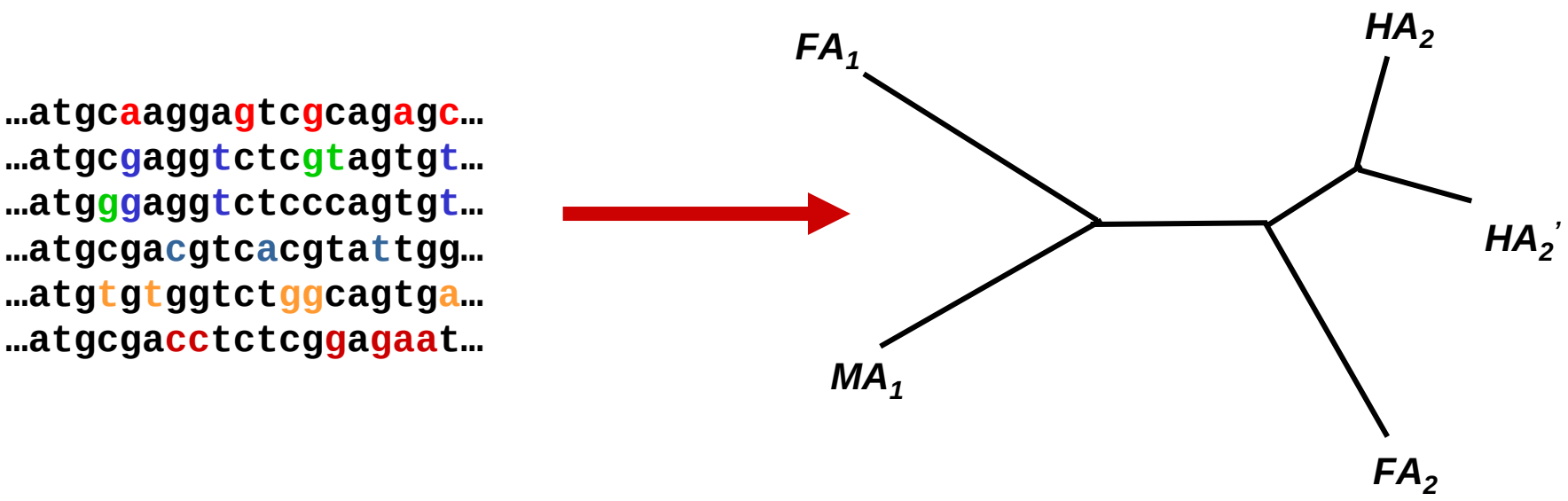
Outline

- Background: genome evolution
- Graph and trees for sequence families
  - Sequence trees
  - Sequence graphs
  - Domain graphs
  - Domain trees
- Graph theoretical insights into multidomain protein evolution

Gene tree reconstruction

Given contemporary sequences, find the tree that

- best explains the data
- under an appropriate sequence evolution model



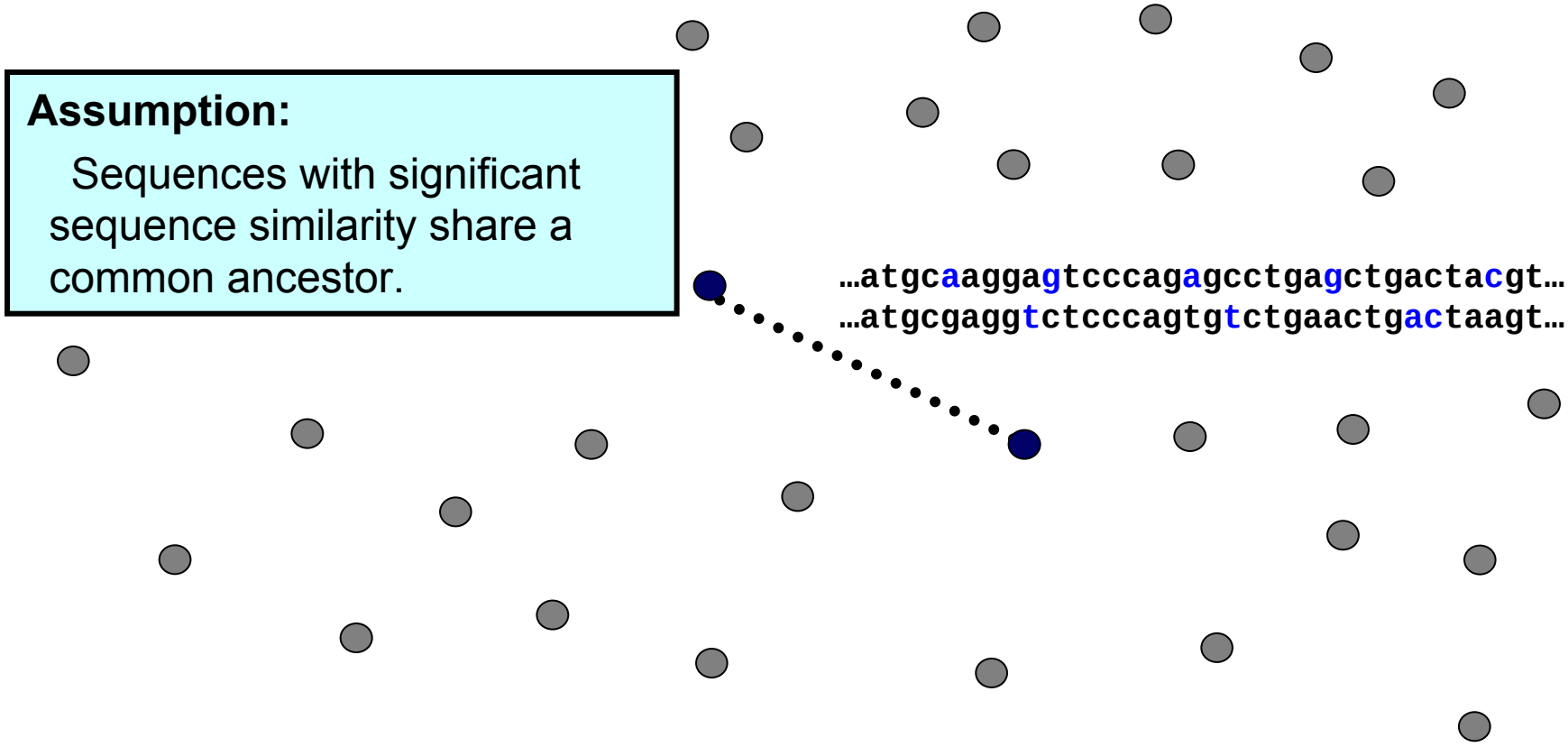
Outline

- Background: genome evolution
- Graph and trees for sequence families
  - Sequence trees
  - Sequence graphs
  - Domain graphs
  - Domain trees
- Graph theoretical insights into multidomain protein evolution

Sequence Similarity Graph

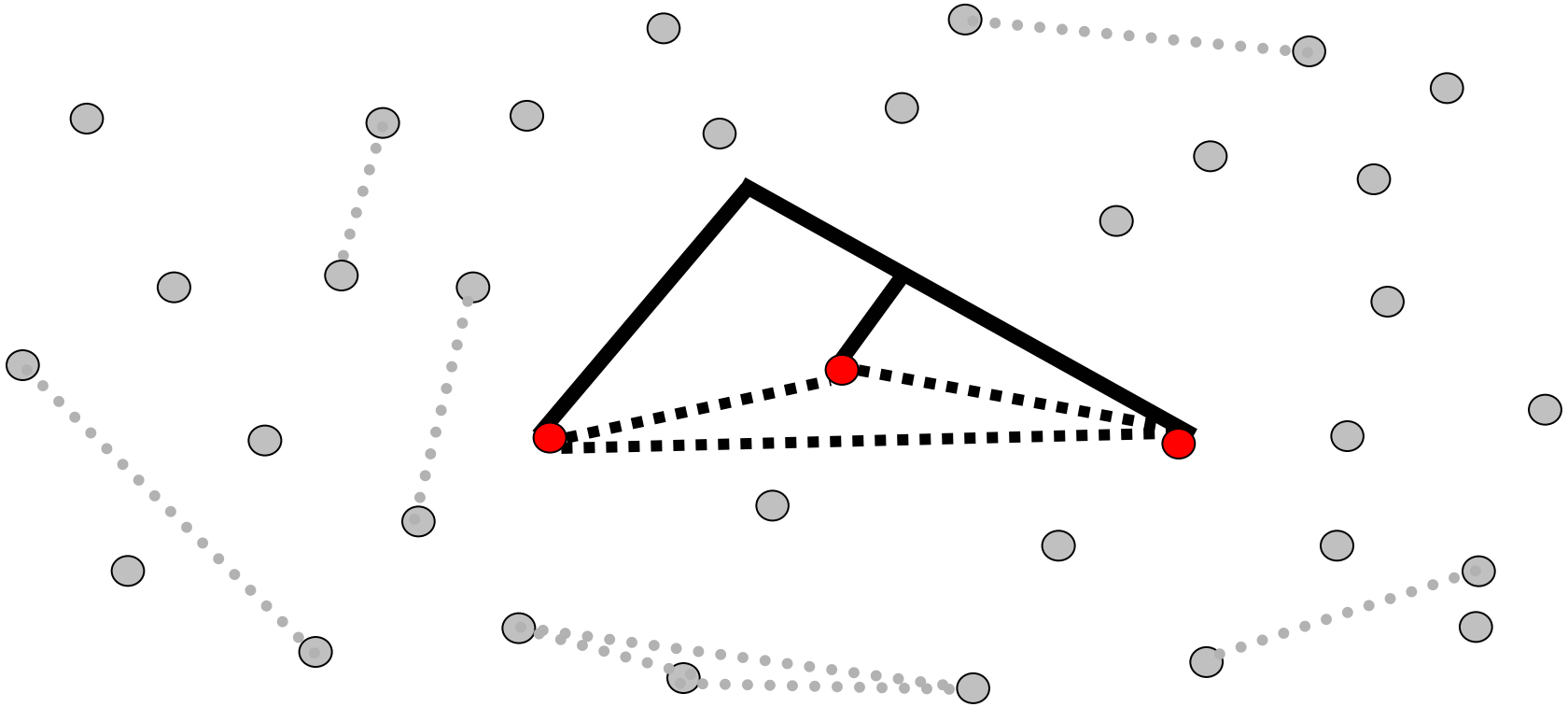
Assumption:

Sequences with significant sequence similarity share a common ancestor.



G = (V,E), V = {all sequences},
E = {(u,v) | u and v share significant similarity}

Sequence Similarity Graph

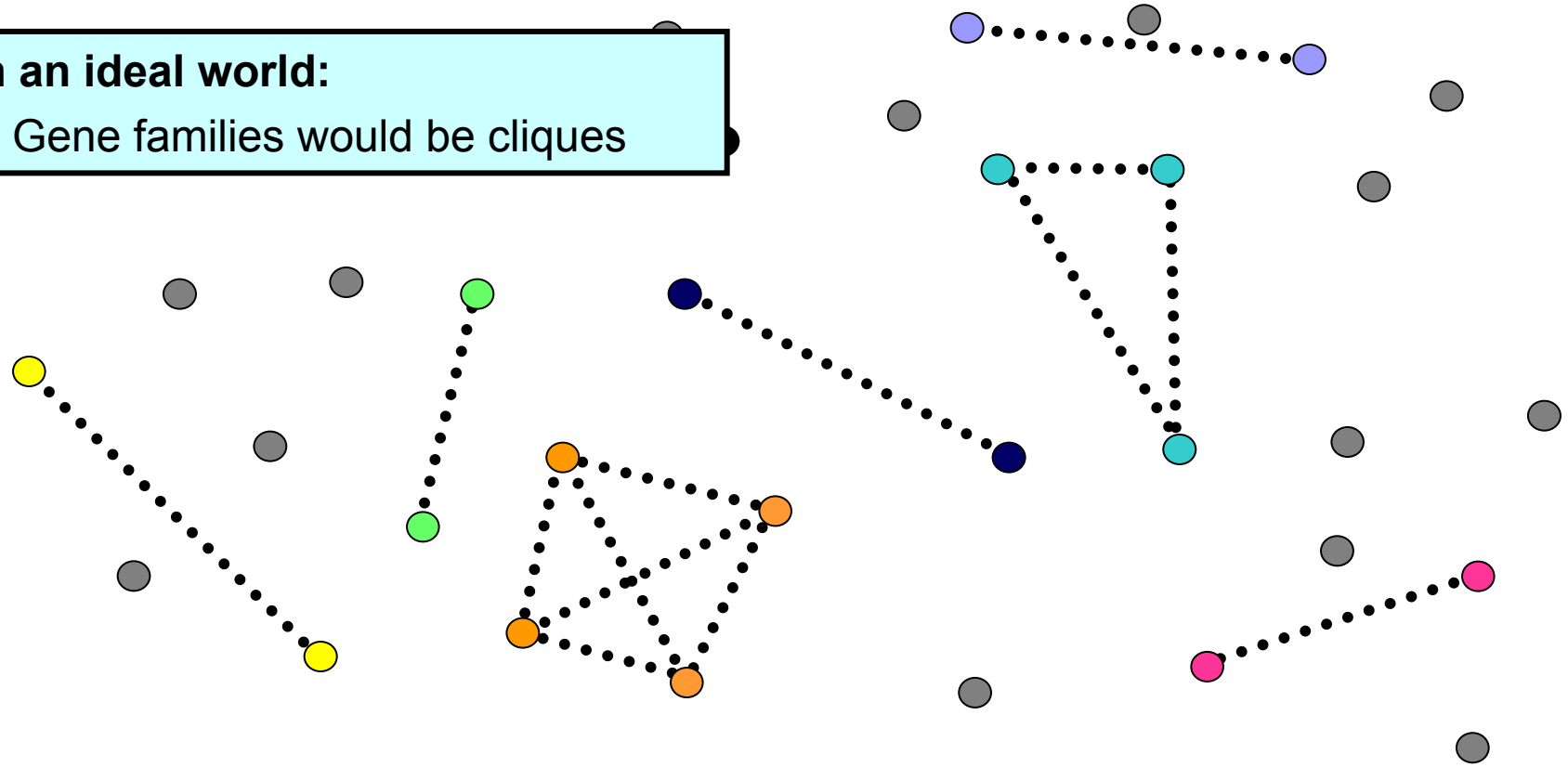


G = (V,E), V = {all sequences},
E = {(u,v) | u and v share significant similarity}

Sequence Similarity Graph

In an ideal world:

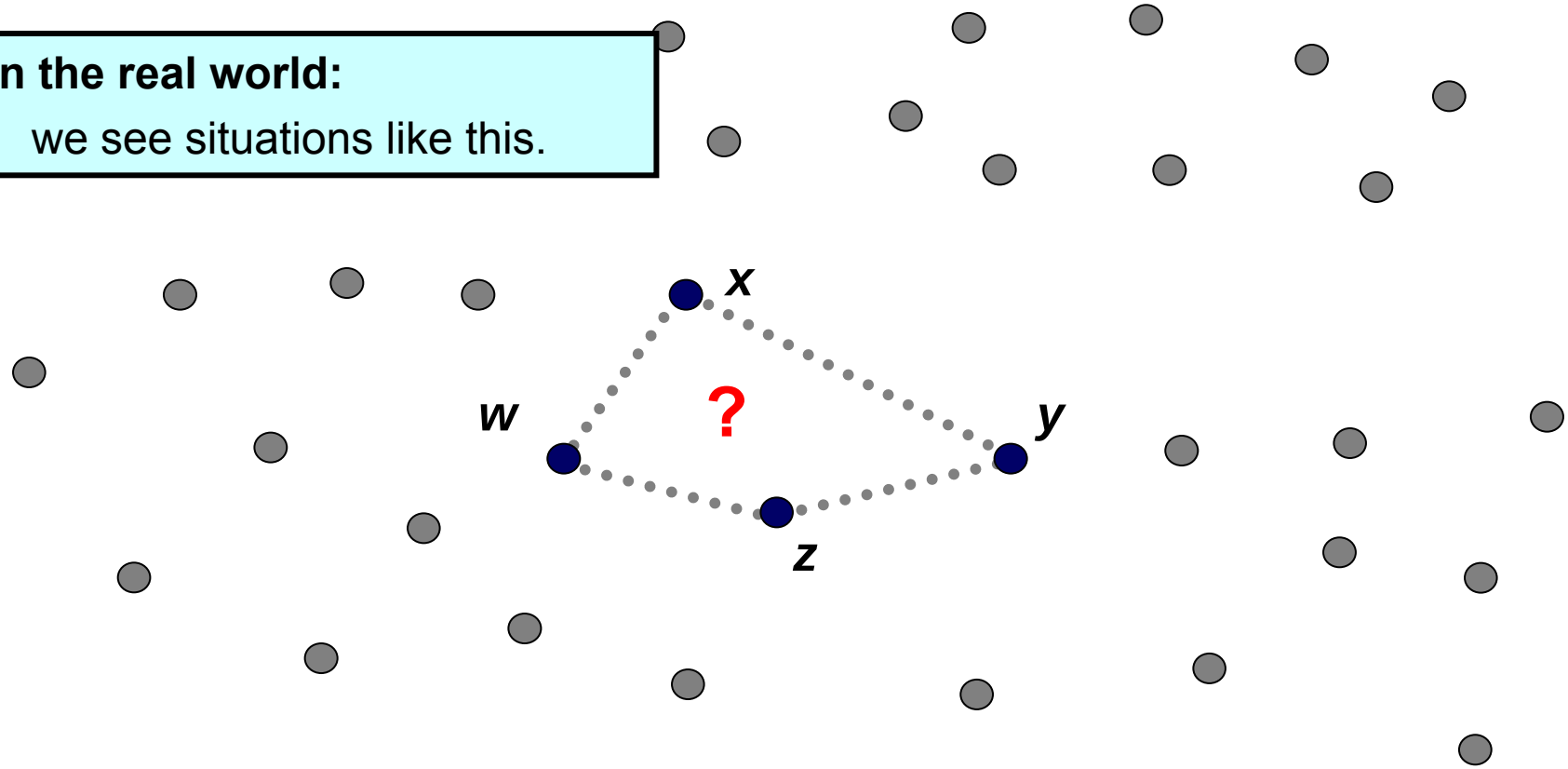
Gene families would be cliques



G = (V,E), V = {all sequences},
E = {(u,v) | u and v share significant similarity}

Sequence Similarity Graph

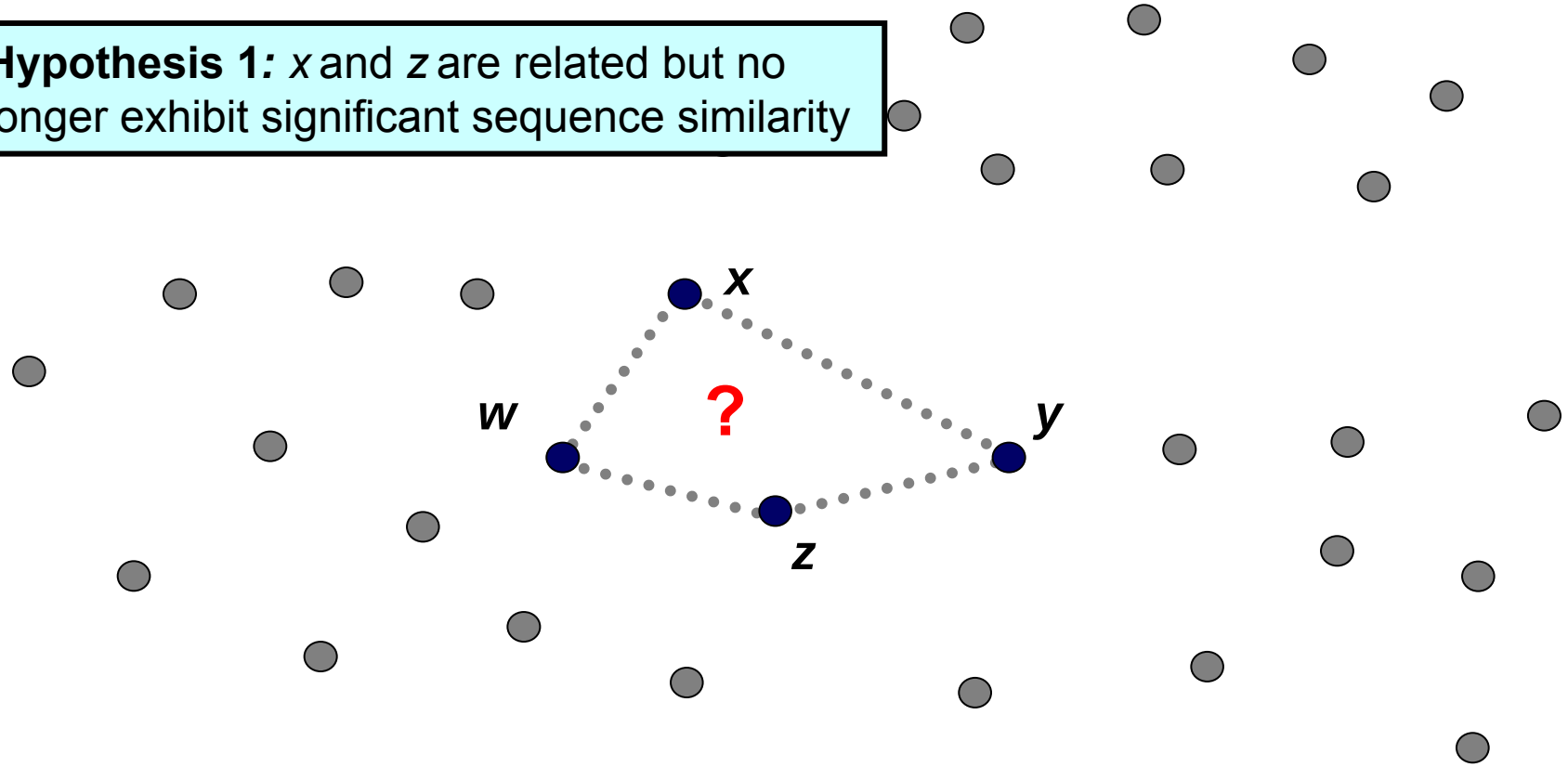
In the real world:
we see situations like this.



G = (V,E), V = {all sequences},
E = {(u,v) | u and v share significant similarity}

Sequence Similarity Graph

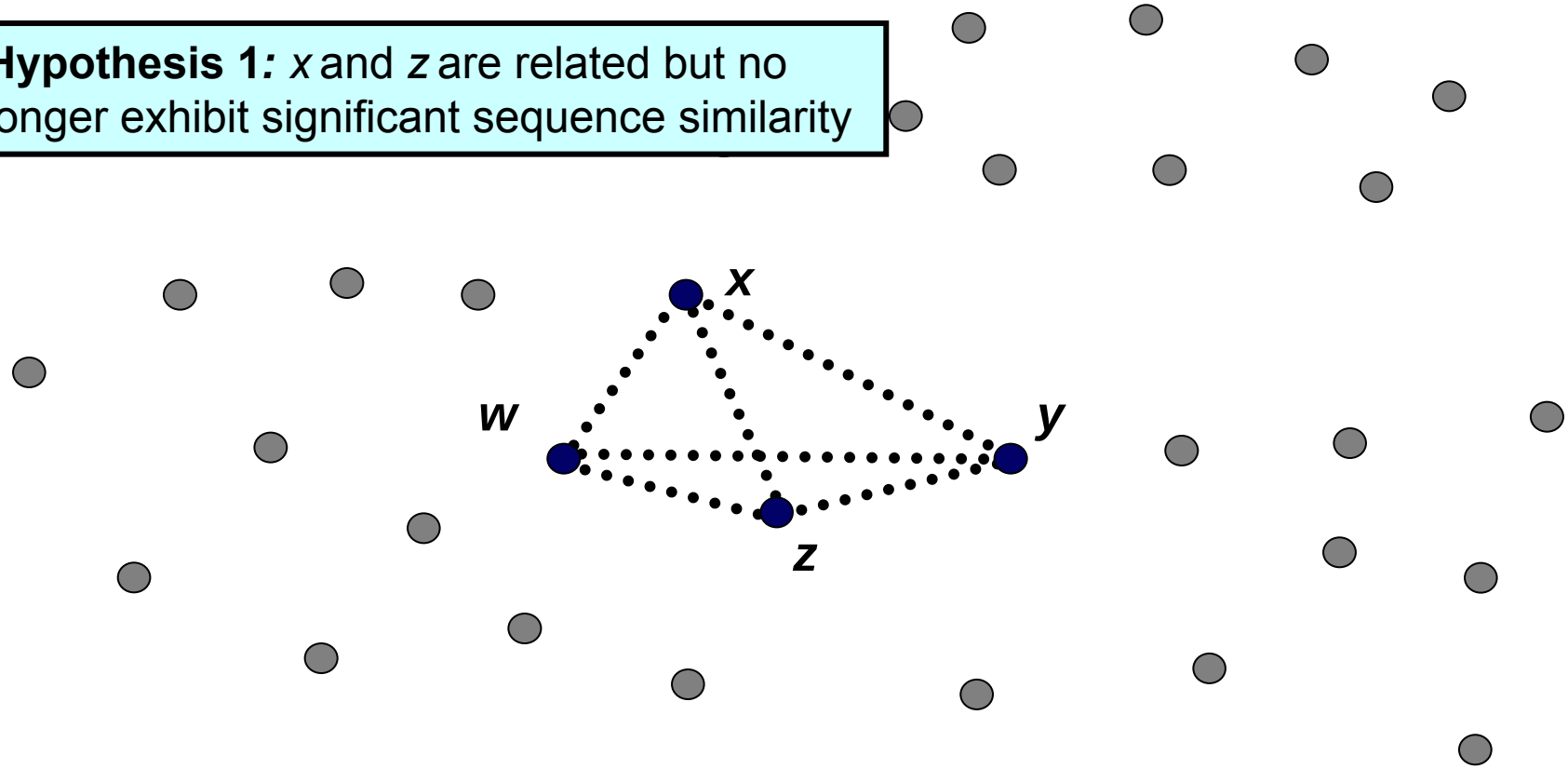
Hypothesis 1: *x* and *z* are related but no longer exhibit significant sequence similarity



G = (V,E), V = {all sequences},
E = {(u,v) | u and v share significant similarity}

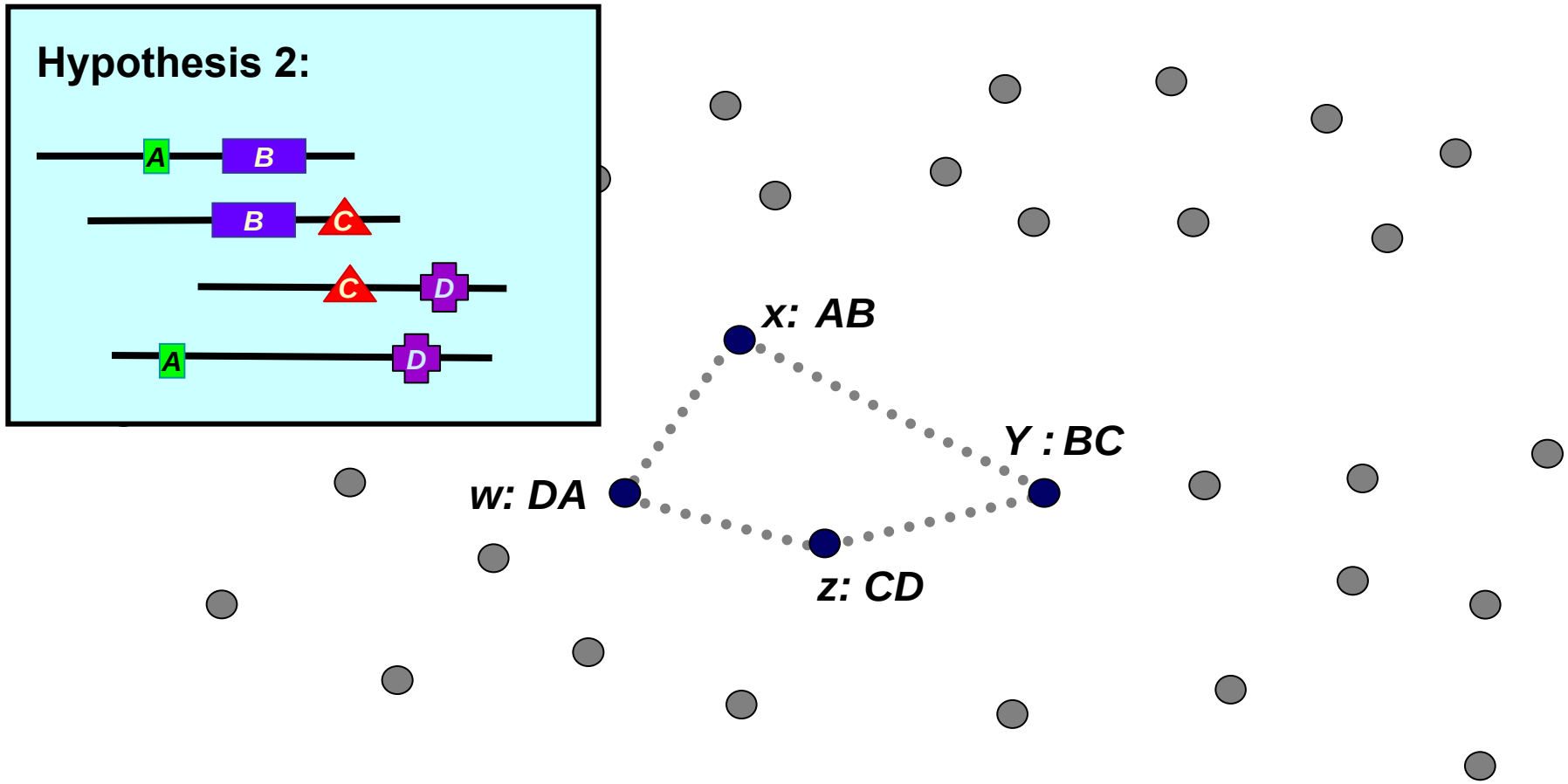
Sequence Similarity Graph

Hypothesis 1: *x* and *z* are related but no longer exhibit significant sequence similarity



G = (V,E), V = {all sequences},
E = {(u,v) | u and v share significant similarity}

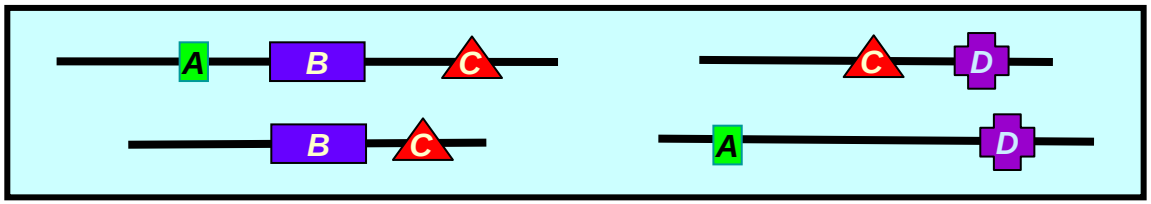
Sequence Similarity Graph



G = (V,E), V = {all sequences},
E = {(u,v) | u and v share significant similarity}

Outline

- Background: genome evolution
- Graph and trees for sequence families
  - Sequence trees
  - Sequence graphs
  - Domain graphs
  - Domain trees
- Graph theoretical insights into multidomain protein evolution

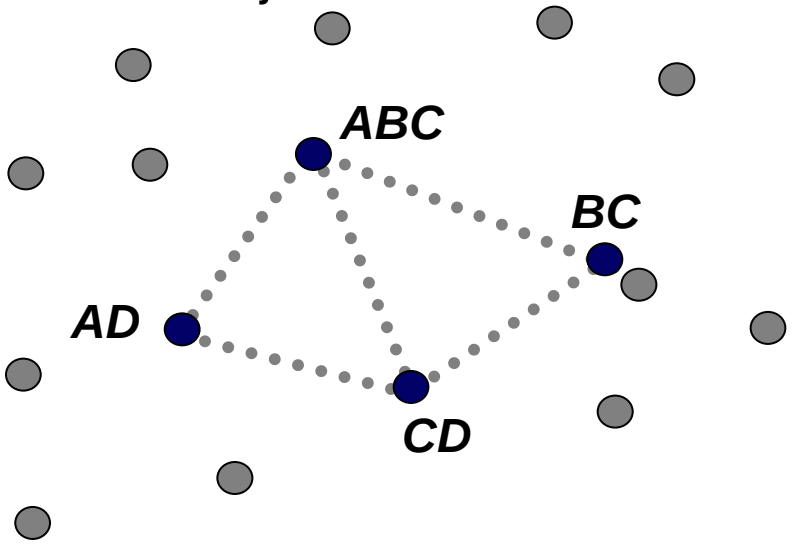


Protein Overlap Graph

G = (V,E),

V = {all domain architectures},

E = {(u.v) | u & v share a domain}

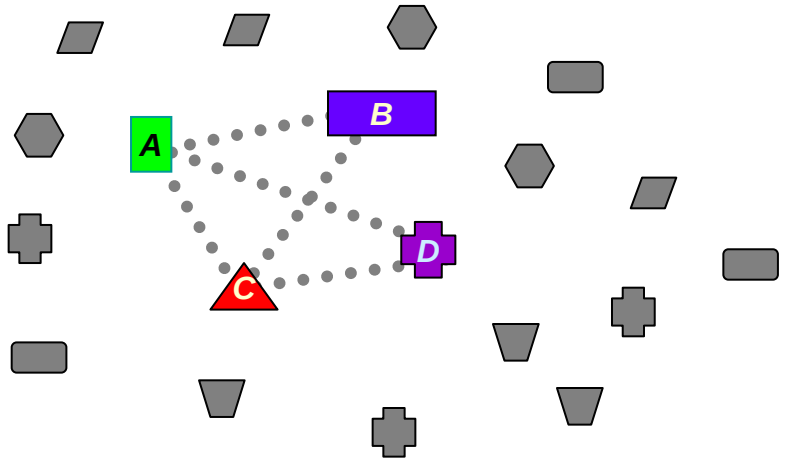


Domain Overlap Graph

G = (V,E),

V = {all domains},

E = {(u.v) | ∃a protein that contains u & v }



Research on the domain graphs

- Graph properties
  - Distribution of node degree, *k*
    - Protein overlap graphs: a^{-k}
    - Domain overlap graphs: $k^{-\gamma}$
  - Clustering coefficient
- Evidence for evolutionary models
  - Preferential attachment
  - Birth, Death, Innovation

Wuchty, 01; Apic, Gough & Teichmann, 01; Wuchty & Almaas, 05; Karev *et al.,* 02*;* Rzhetsky & Gomez, 01; Qian, Lacomb & Gerstein, 01

Outline

- Background: genome evolution
- Graph and trees for sequence families
  - Sequence trees
  - Sequence graphs
  - Domain graphs
  - Domain trees
- Graph theoretical insights into multidomain protein evolution

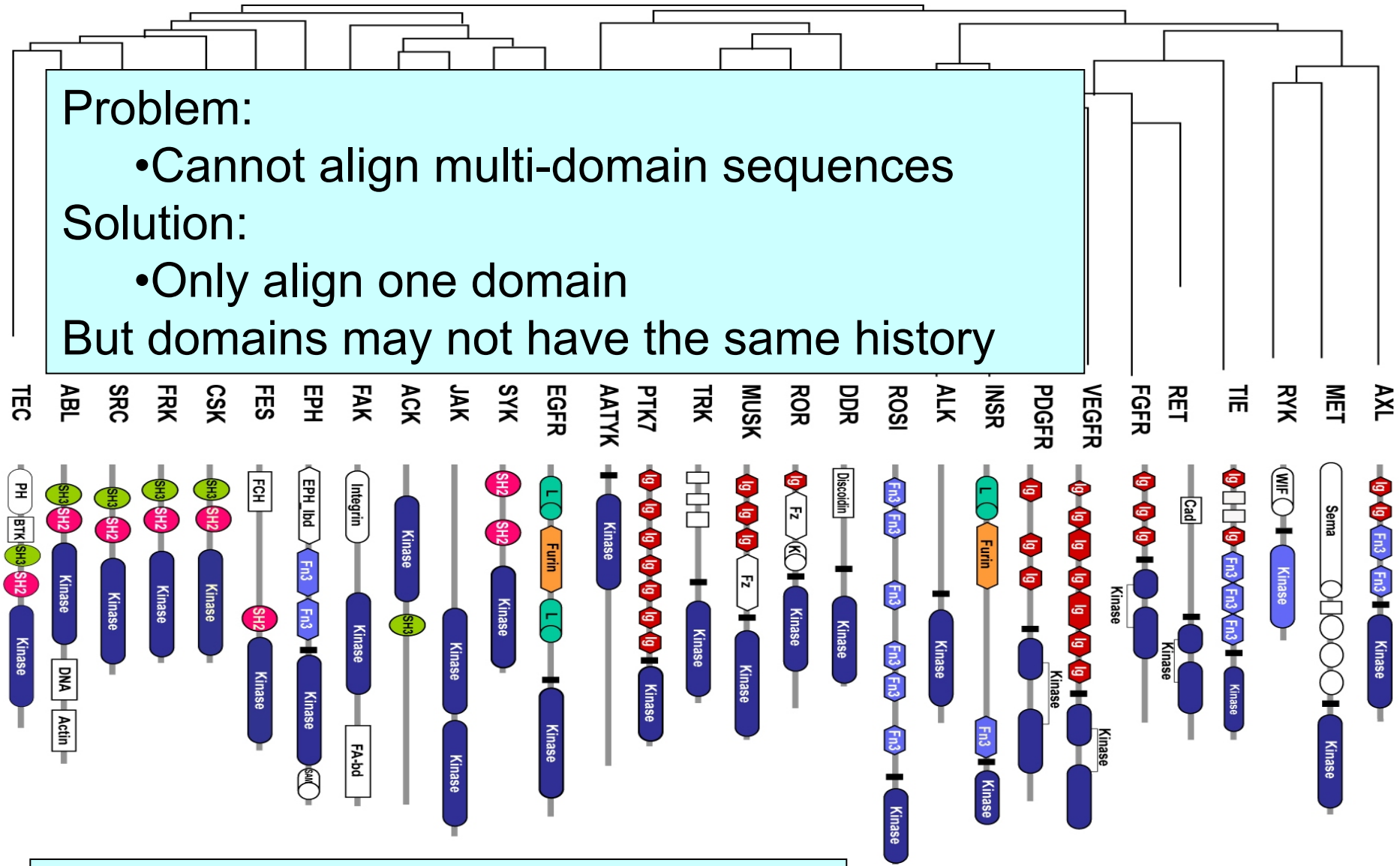
Problem:

- Cannot align multi-domain sequences

Solution:

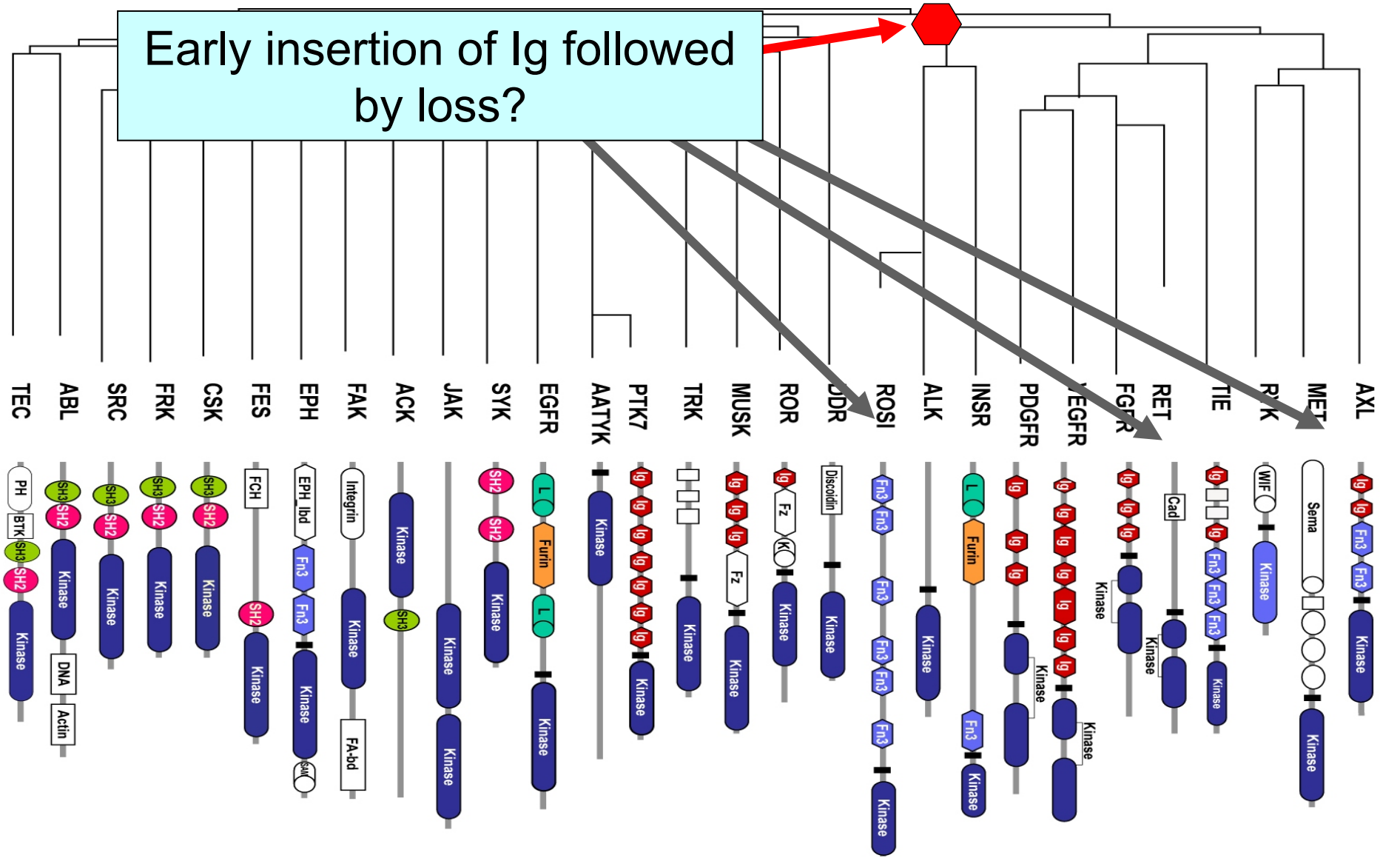
- Only align one domain

But domains may not have the same history

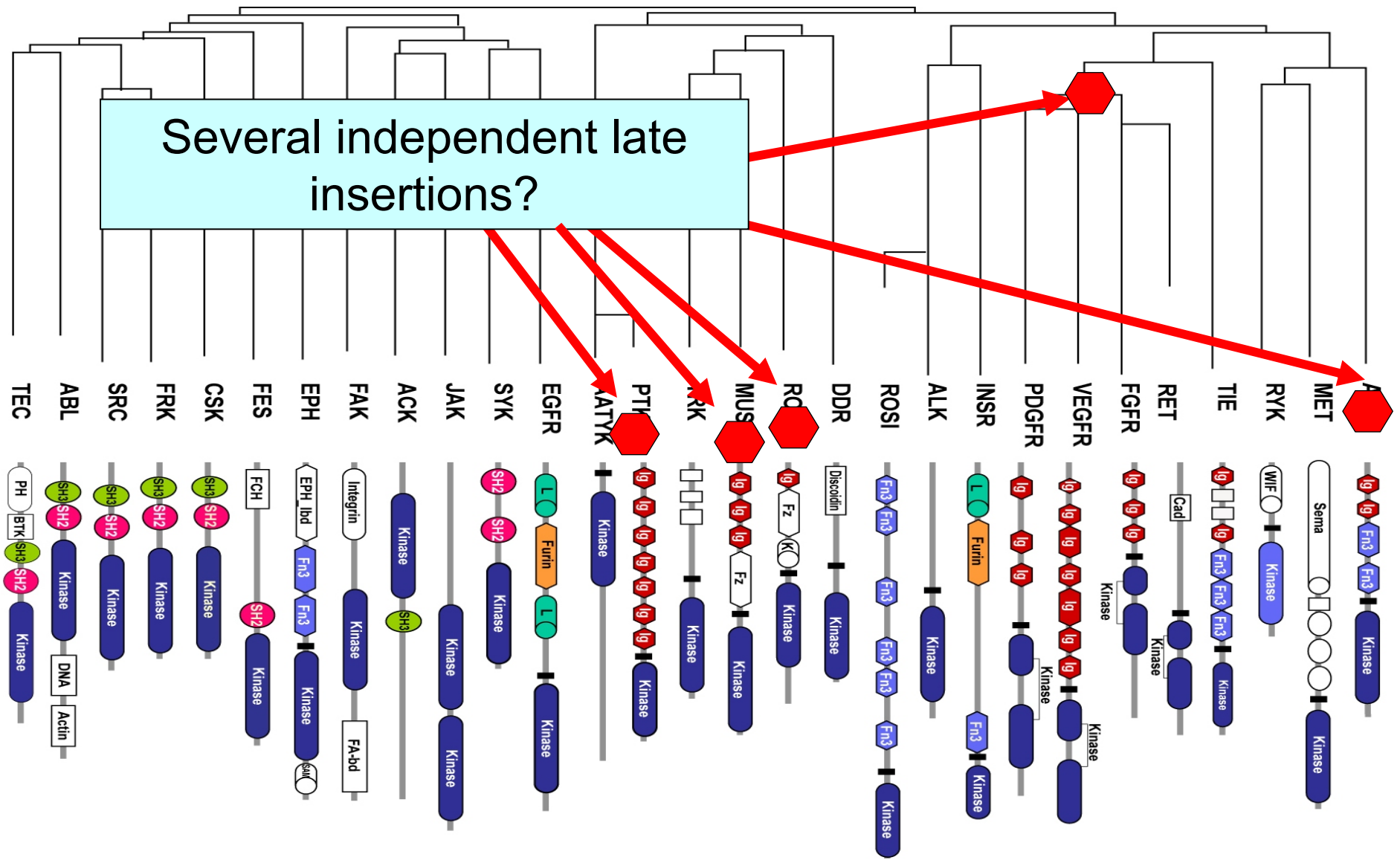


Tree constructed from alignment of kinase domains only

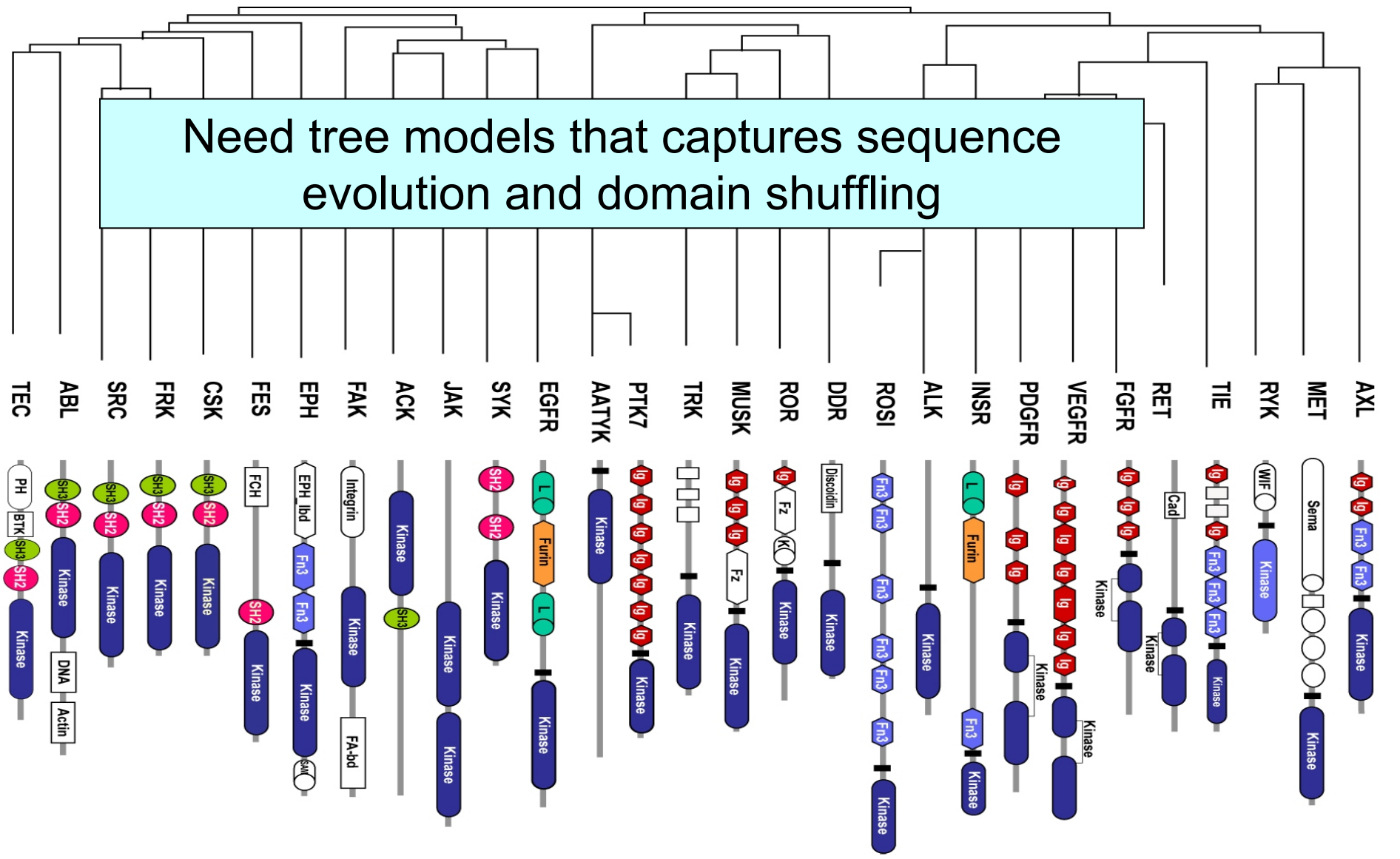
Adapted from Robinson *et al.*, 2000



Adapted from Robinson *et al.*, 2000

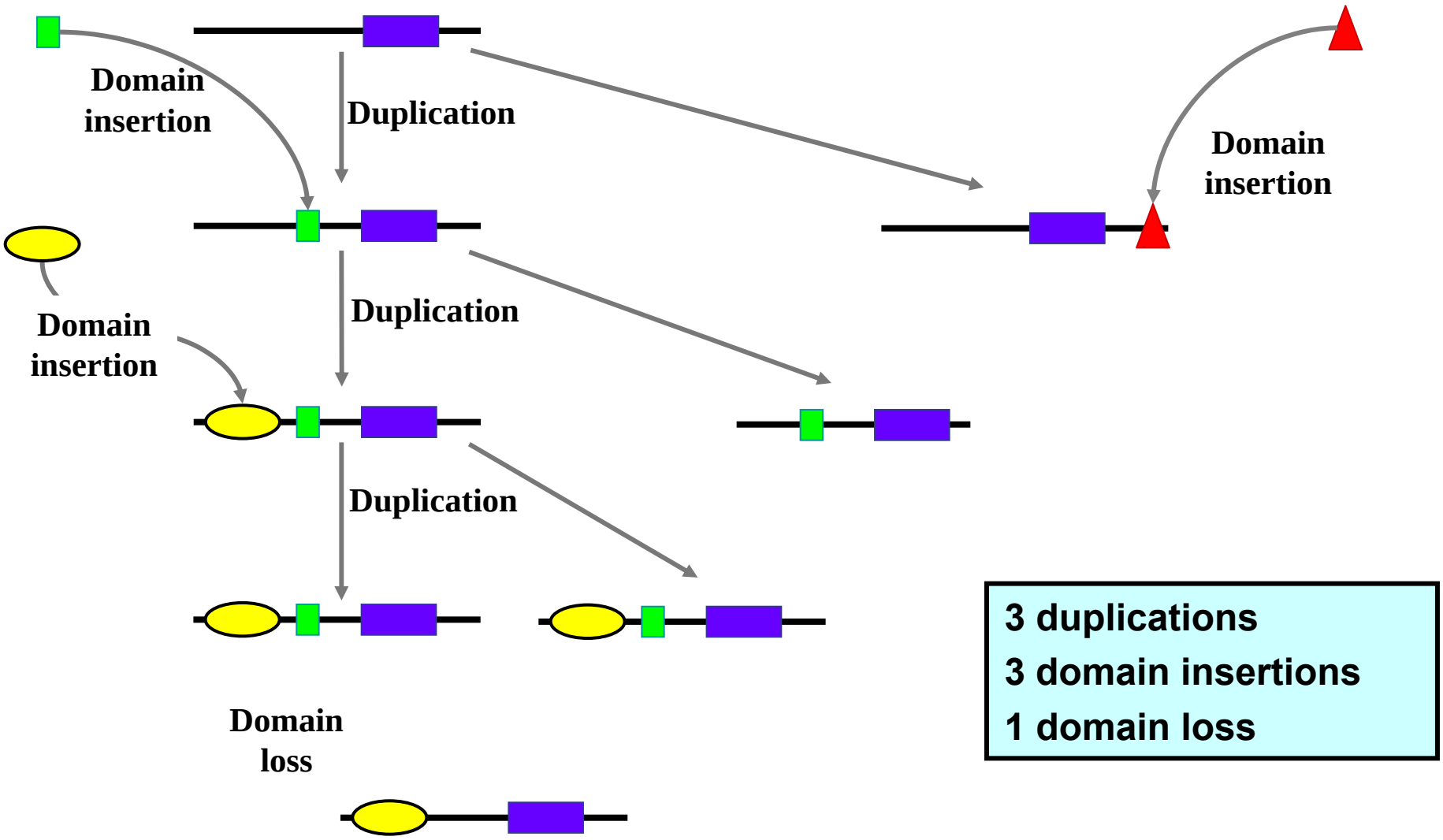


Adapted from Robinson *et al.*, 2000

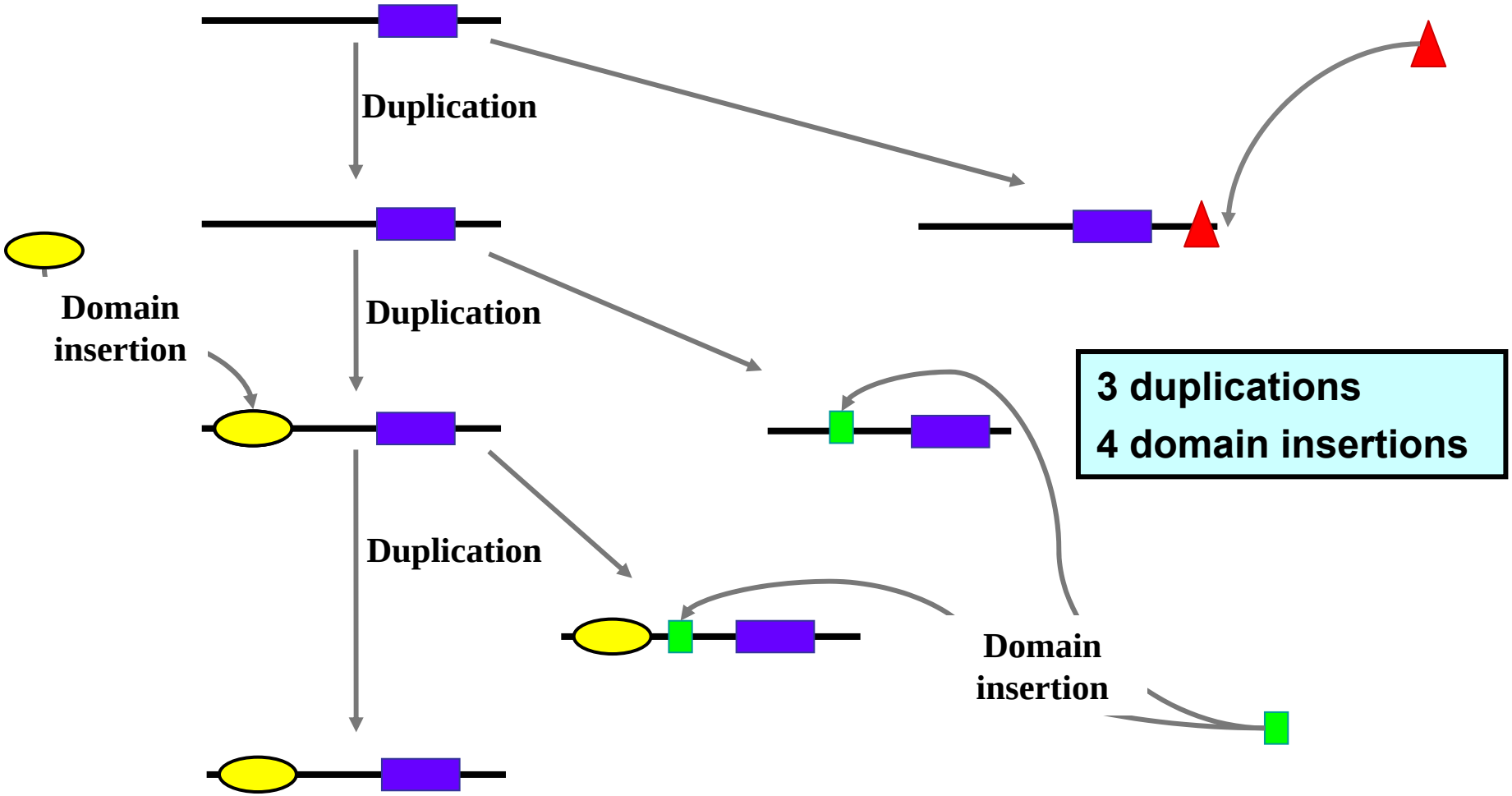


Adapted from Robinson *et al.*, 2000

An example of multidomain family evolution



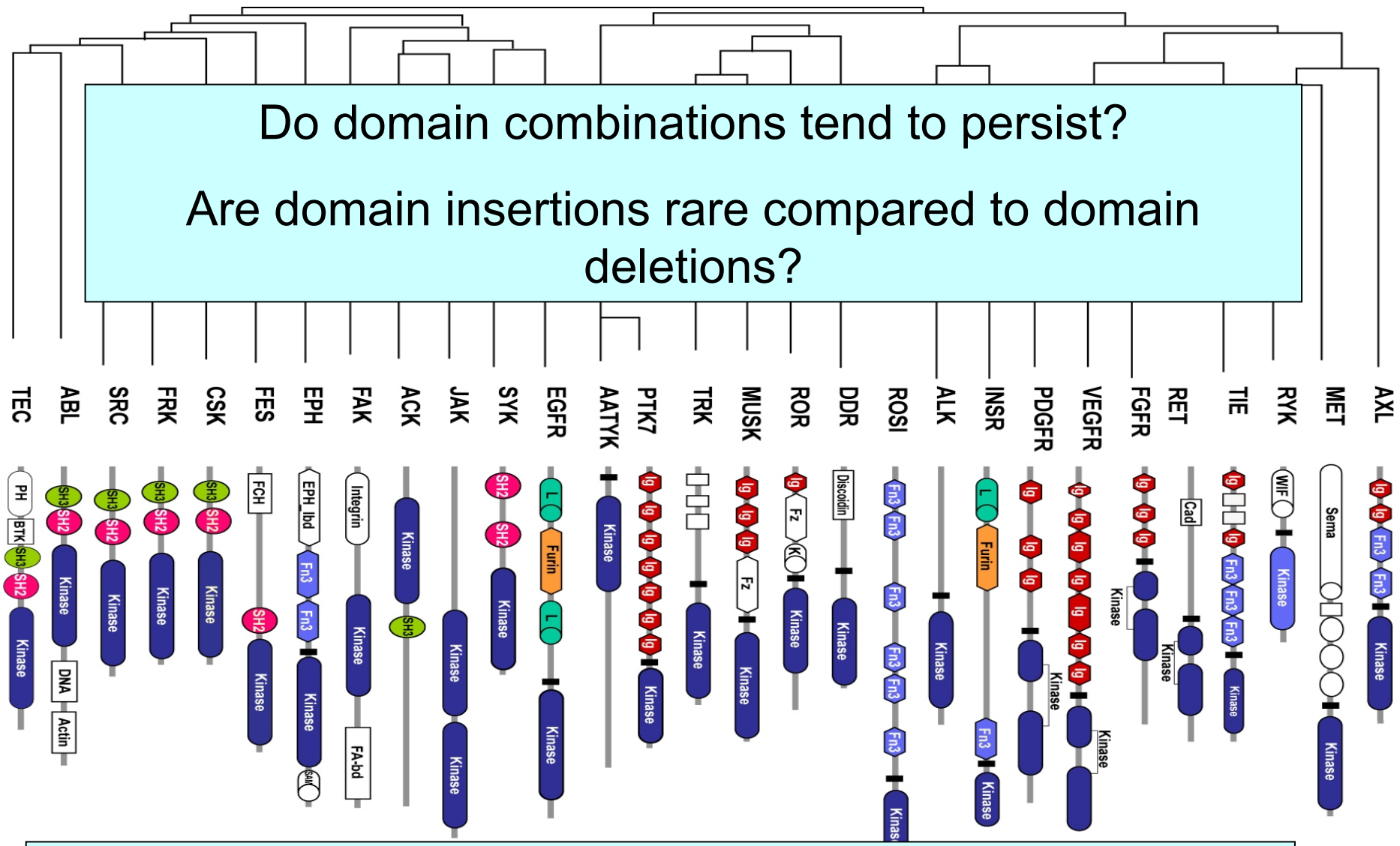
An alternate history



Outline

- Background: genome evolution
- Graph and trees for sequence families
- Graph theoretical insights into multidomain protein evolution
  - Are domain insertions rare?
  - Do domain architectures persist?

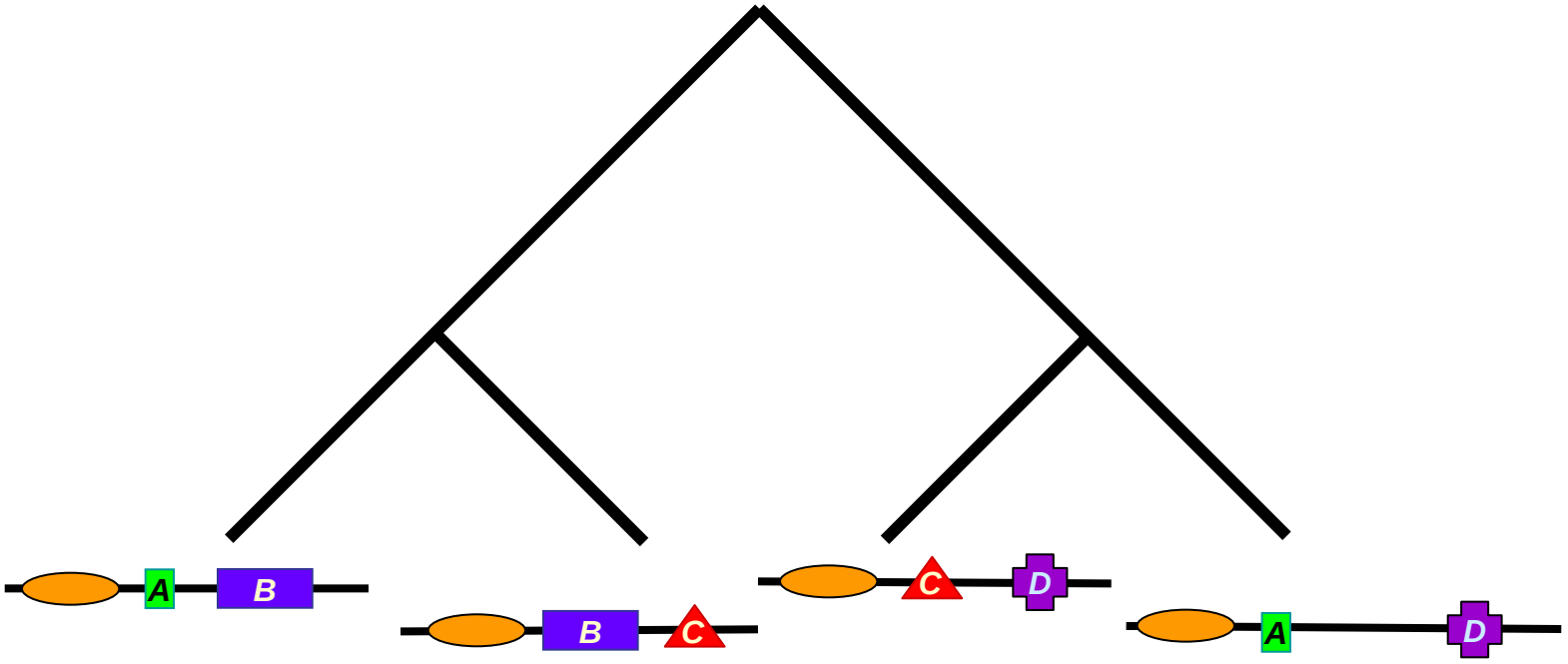
Przytycka, Davis, Song, Durand, JCB, 2006



Note: kinase domain = a reference point for these events

Adapted from Robinson *et al.*, 2000

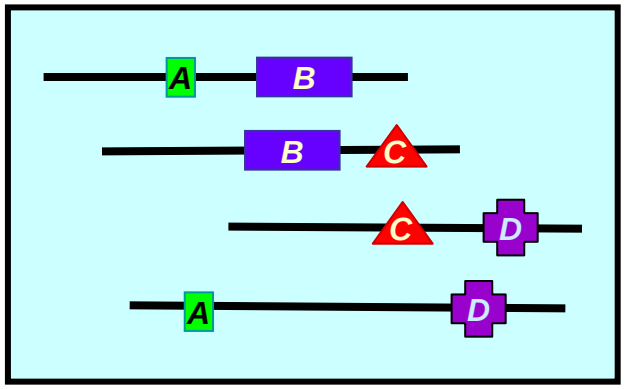
Main Idea

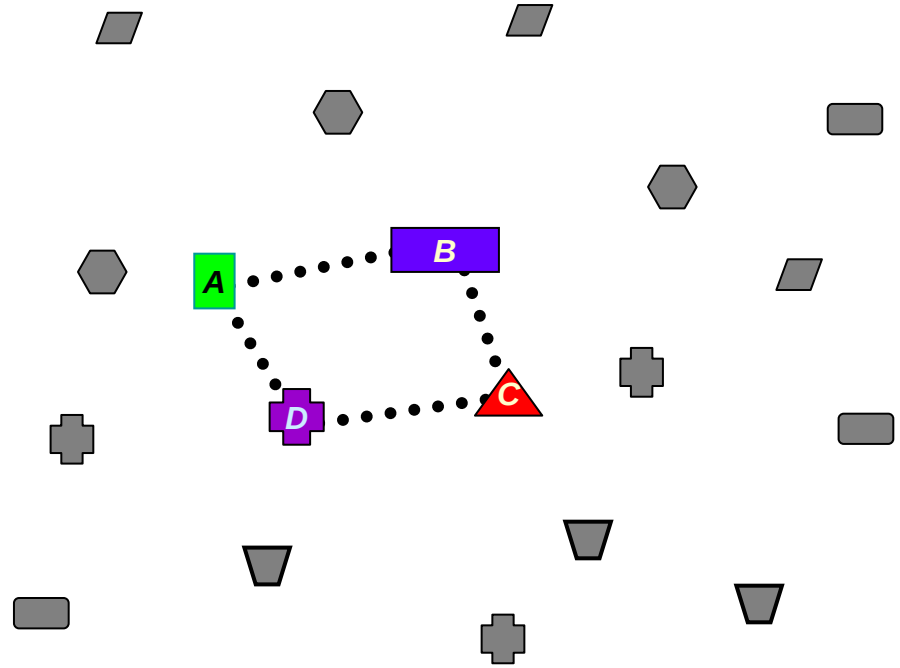


– History of insertions and deletions is encoded in a tree, but trees are hard to reconstruct.

–Test assumptions *domain overlap graph* instead.

Main Idea



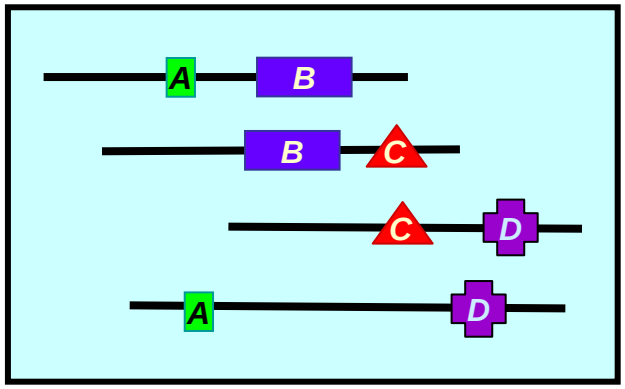


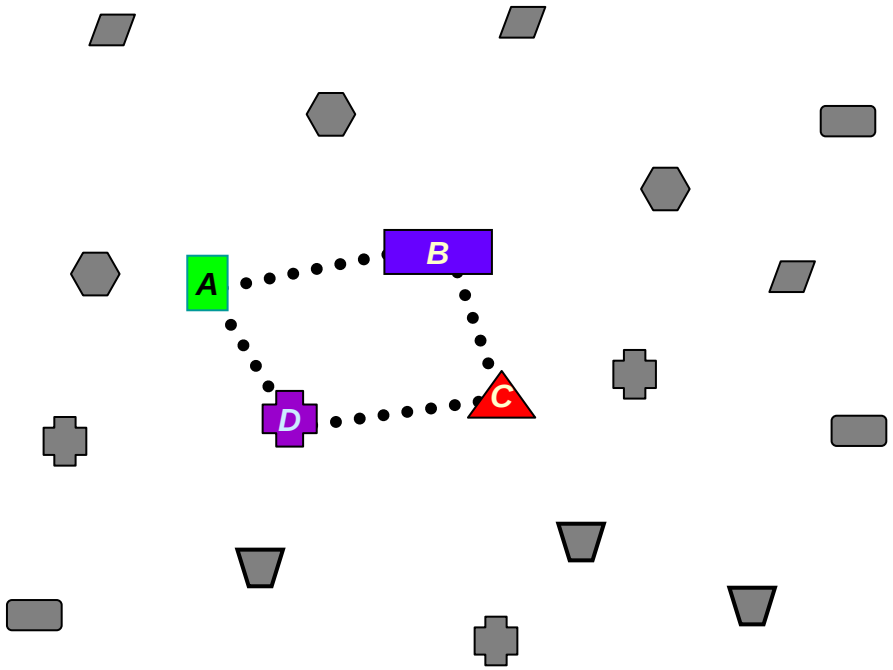
Domain Overlap Graph: $G = (V,E)$

V = {all domains},

$E = \{(u.v) \mid \exists$ a protein that contains u & $v\}$

Main Idea





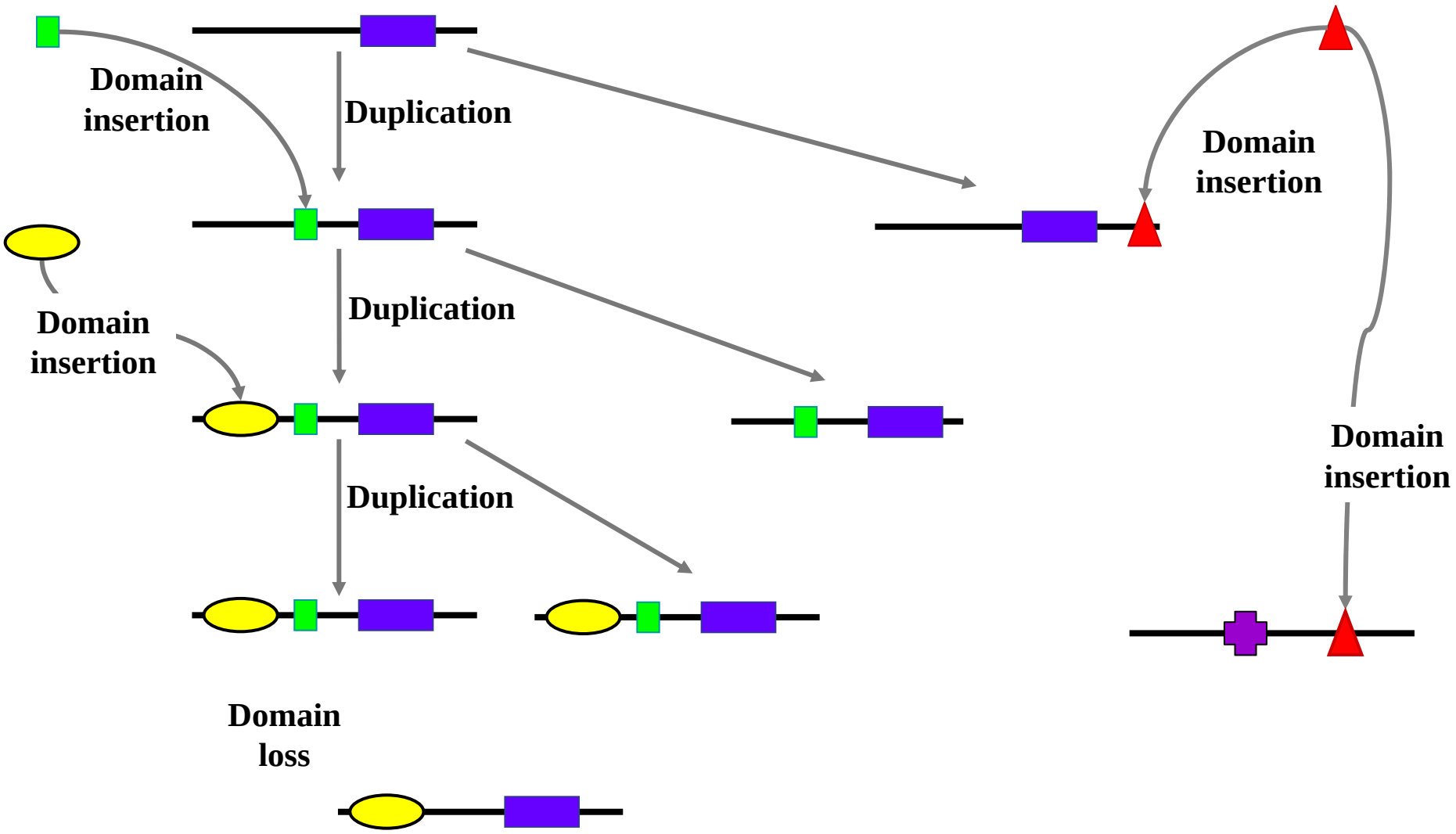
Cycle in the Domain Overlap Graph (DOG)

- Multiple domain insertions must have occurred

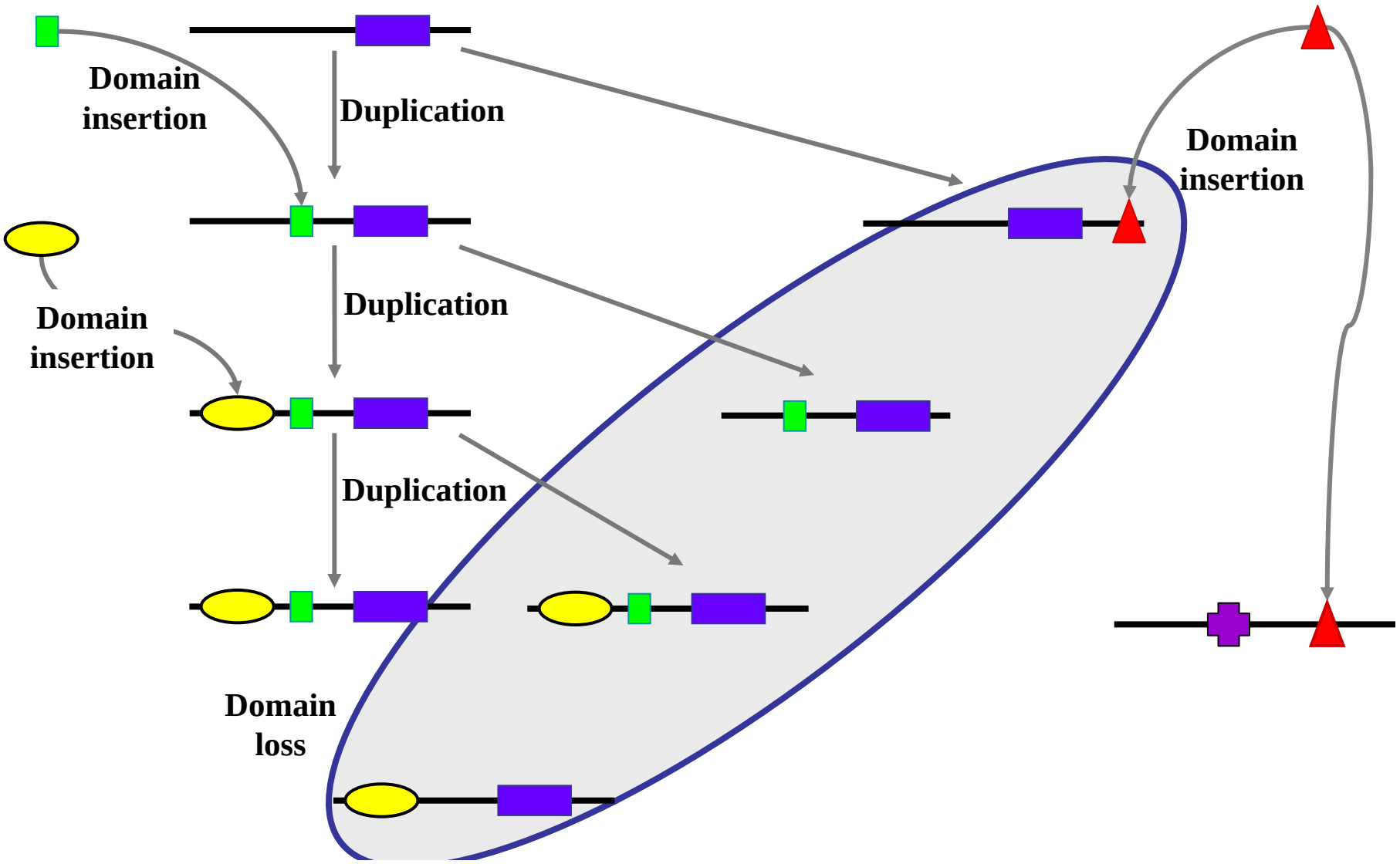
Definitions

- A protein is a set of domains
  - Ignore domain order
  - Ignore copy number
- Events
  - Domain “insertion” (includes unequal crossing over, retrotransposition, read-thru errors, etc.).
  - Domain loss.
  - Gene duplications are free.
- Let *D superfamily* = {all domain architectures containing domain *D*}.

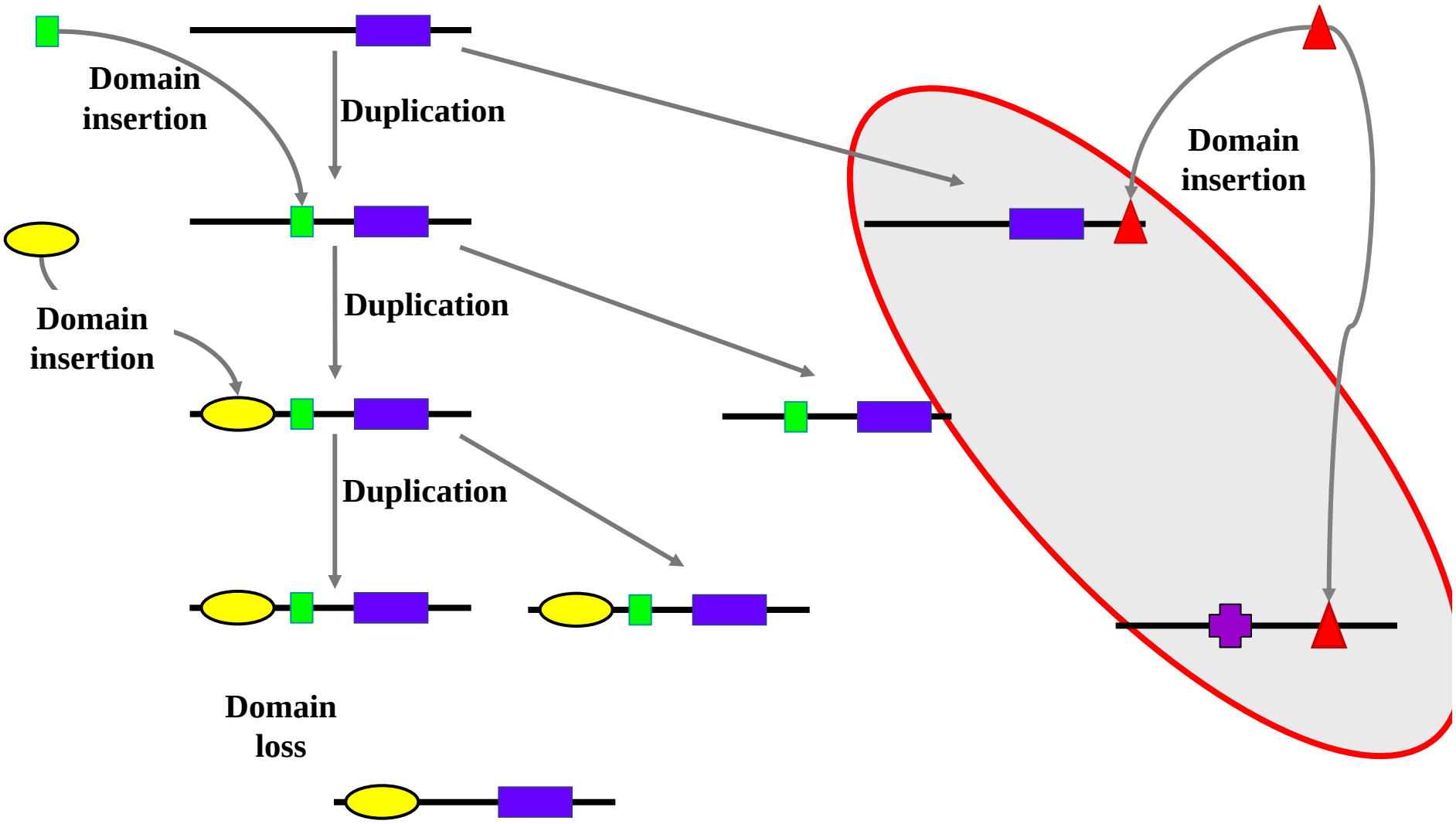
An example of multidomain family evolution



The blue domain superfamily:



The red domain superfamily:



Our Approach

- **Introduce multidomain parsimony models**
  - Conservative Dollo Parsimony (CDP)

    Insertions are rare
  - Static Dollo Parsimony (SDP)

    Domain architectures persist
- Map these models to *Domain Overlap Graph (DOG)*
  - Show CDP ↔ chordality in *DOG*
  - Show SDP ↔ cliques in *DOG* correspond to proteins
  - Adapt fast algorithms for testing these properties to *DOGs.*
- Apply test to all superfamilies in the SwissProt database

Maximum Parsimony

Character data: binary, multistate

Assumptions:

- Mutations are rare events
- The best hypothesis requires the fewest state changes to explain the data
- *Additional constraints may be imposed.*

For a given data set, use model to

- Construct a tree
- Test whether assumptions are violated.

Goal: adapt existing parsimony models to multidomain evolution to test our hypotheses.

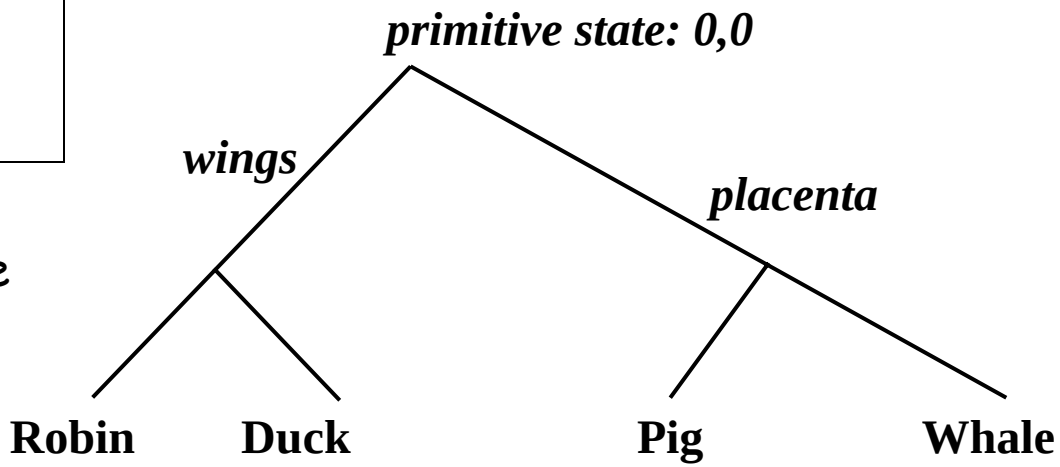
Perfect Phylogeny

Assumptions:

- Mutations are rare events
- Every binary character changes state (0→1) at most once.

Taxa	Characters	
	Placenta	*Wings*
Duck	0	1
Robin	0	1
Whale	1	0
Pig	1	0

This data set satisfies the assumptions



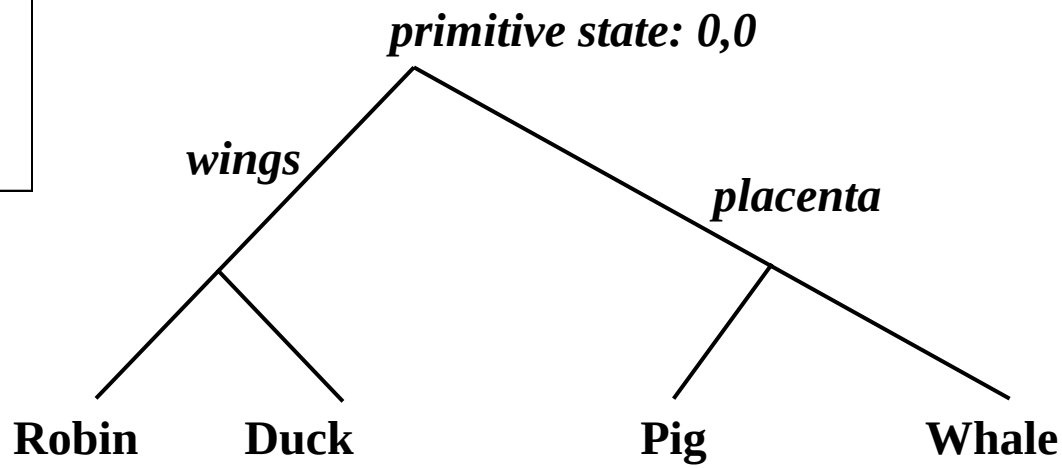
Perfect Phylogeny

Assumptions:

- Mutations are rare events
- Every binary character changes state (0→1) at most once.

Taxa	Character Data	
	Placenta	*Wings*
Duck	0	1
Robin	0	1
Bat	**1**	**1**
Whale	1	0
Pig	1	0

This data set does not admit a perfect phylogeny



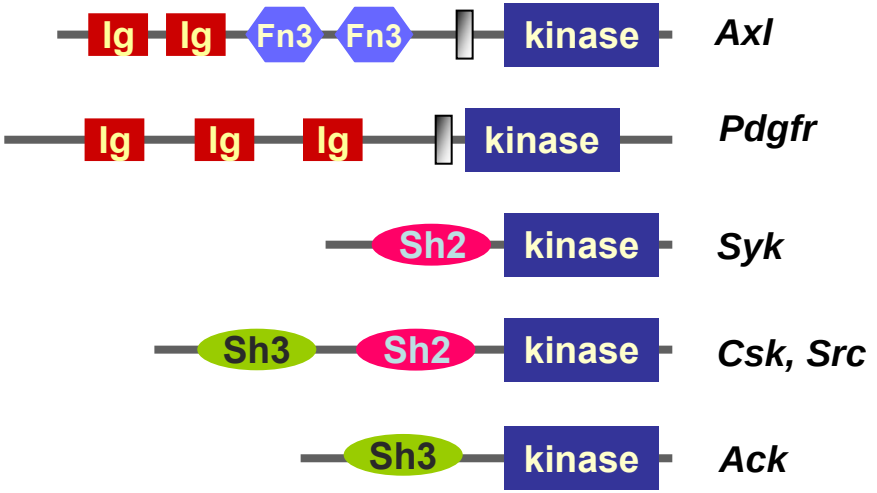
Multidomain Parsimony Model

- A protein is a set of domains
  - Ignore domain order, copy number
- Leaf nodes: proteins
- Binary characters: domains
- If domain *d* is present in protein *p, we* say the state of character *d* in protein *p* is 1.
- Events
  - Domain insertion. *State(d): 0 → 1*
  - Domain loss. *State(d): 1 → 0*
- *Goal:* For each superfamily, *S(D),* does a parsimony tree exist for *S(D)*?

Example: multidomain character data

Some domain architectures from the kinase superfamily

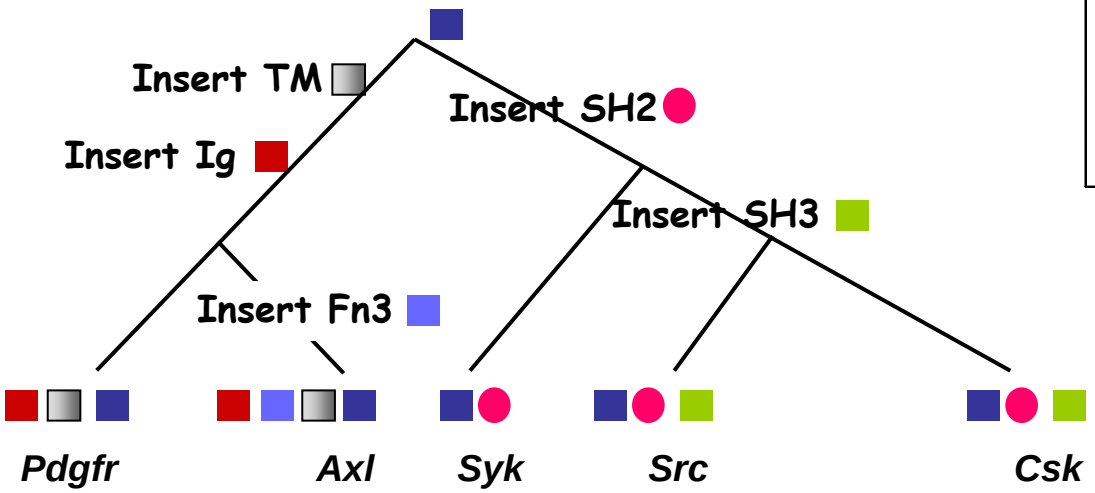
Taxa	Sh2	Sh3	Ig	FN3	TM
Pdgfr	0	0	1	0	1
Axl	0	0	1	1	1
Syk	1	0	0	0	0
Src	1	1	0	0	0
Csk	1	1	0	0	0
Ack	0	1	0	0	0
…	…	…	…		…



- Domains are characters
- Character state: 1 ↔ ‘Domain *d* is present in protein *p.*’

Existing model: Perfect Phylogeny

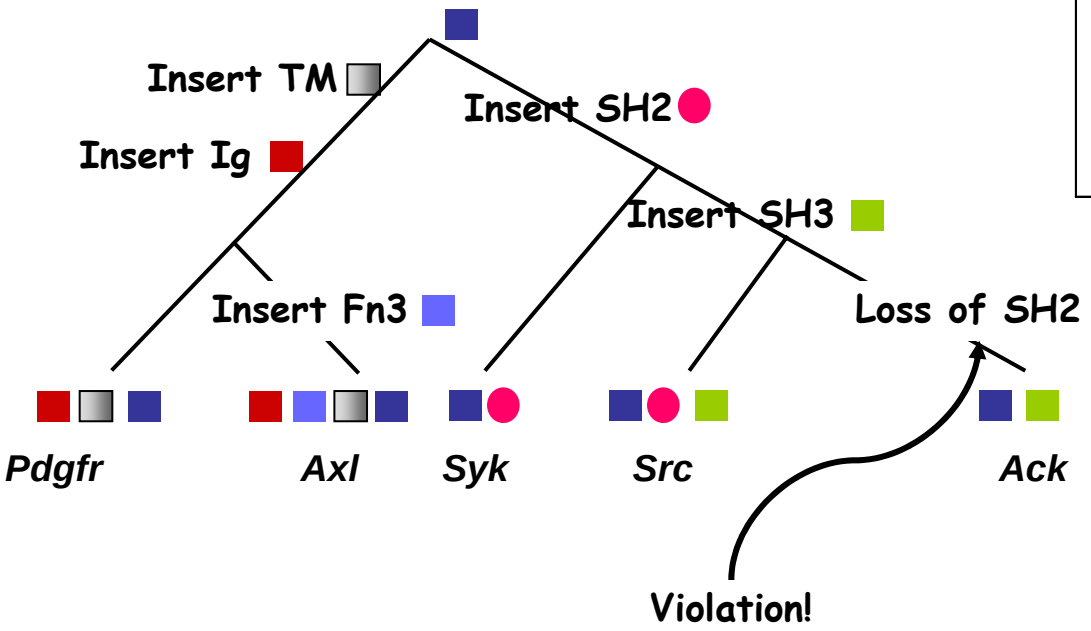
Every binary character changes state (0→1) at most once.



Taxa	Sh2	Sh3	Ig	FN3	TM
Pdgfr	0	0	1	0	1
Axl	0	0	1	1	1
Syk	1	0	0	0	0
Src	1	1	0	0	0
Csk	1	1	0	0	0
Ack	0	1	0	0	0
…	…	…	…		…

Existing model: Perfect Phylogeny

Every binary character changes state (0→1) at most once.



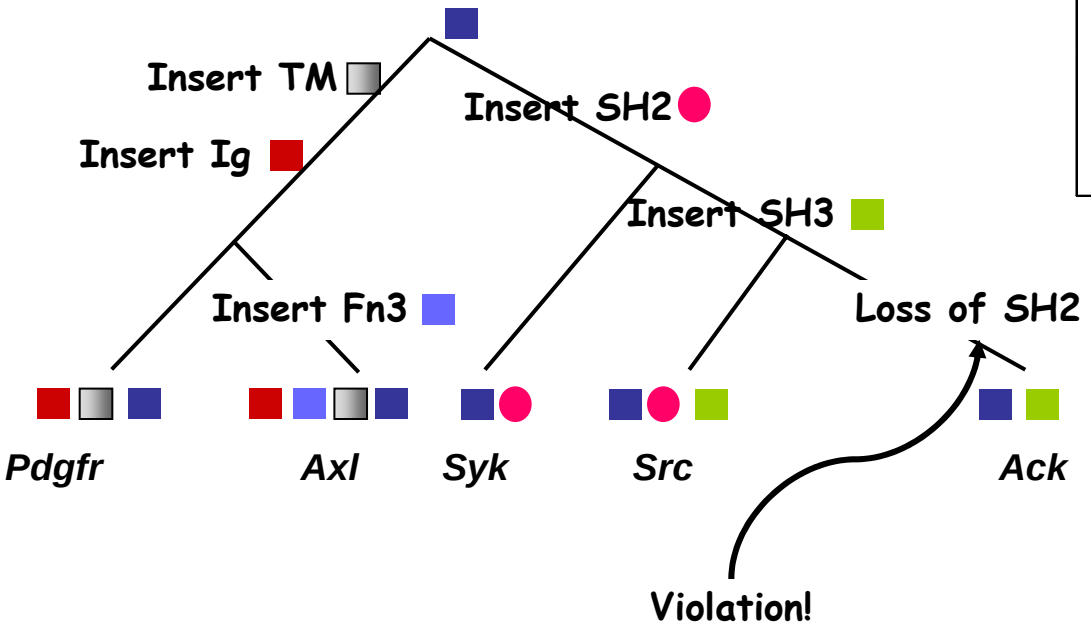
Taxa	Sh2	Sh3	Ig	FN3	TM
Pdgfr	0	0	1	0	1
Axl	0	0	1	1	1
Syk	1	0	0	0	0
Src	1	1	0	0	0
Csk	1	1	0	0	0
Ack	0	1	0	0	0
…	…	…	…		…

Existing model: Perfect Phylogeny

Every binary charac[illegible]ges state at most once.

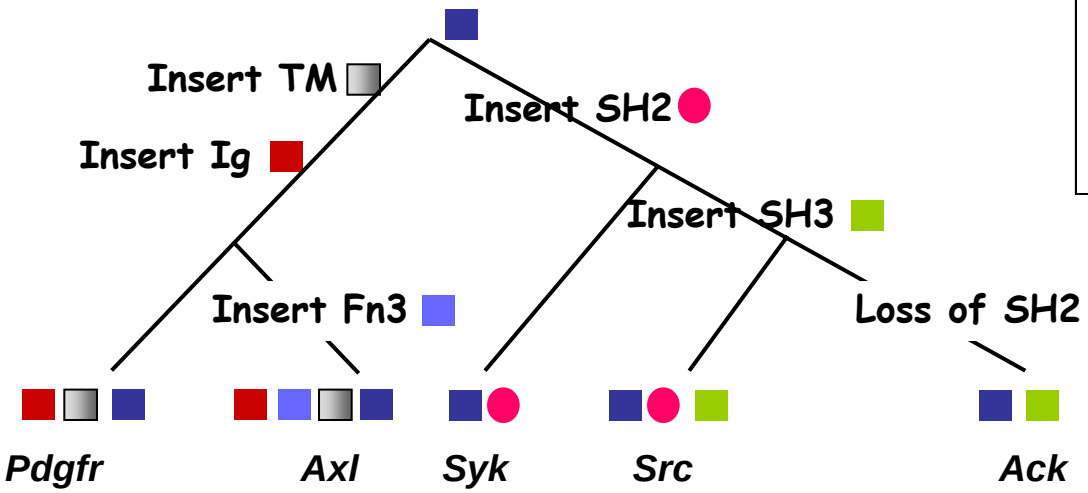
Too restrictive!

Taxa	Sh2	Sh3	Ig	FN3	TM
Pdgfr	0	0	1	0	1
Axl	0	0	1	1	1
Syk	1	0	0	0	0
Src	1	1	0	0	0
Csk	1	1	0	0	0
Ack	0	1	0	0	0
…	…	…	…		…



Existing model: Dollo Parsimony

Every binary character changes state 0 → 1 at most once,
1 → 0 unrestricted

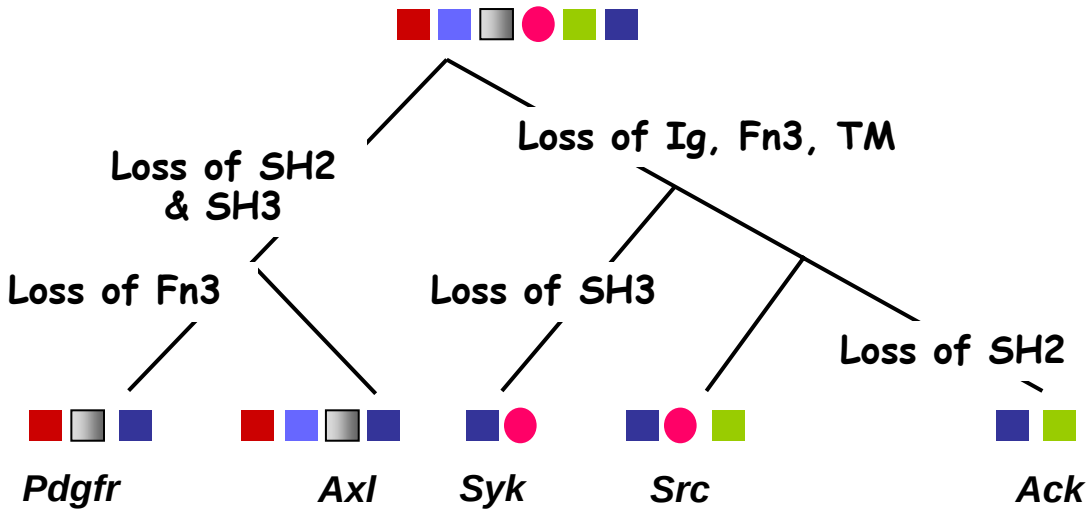


Taxa	Sh2	Sh3	Ig	FN3	TM
Pdgfr	0	0	1	0	1
Axl	0	0	1	1	1
Syk	1	0	0	0	0
Src	1	1	0	0	0
Csk	1	1	0	0	0
Ack	0	1	0	0	0
…	…	…	…		…

This tree satisfies Dollo parsimony

Existing model: Dollo Parsimony

Every binary character changes state 0 → 1 at most once,
1 → 0 unrestricted

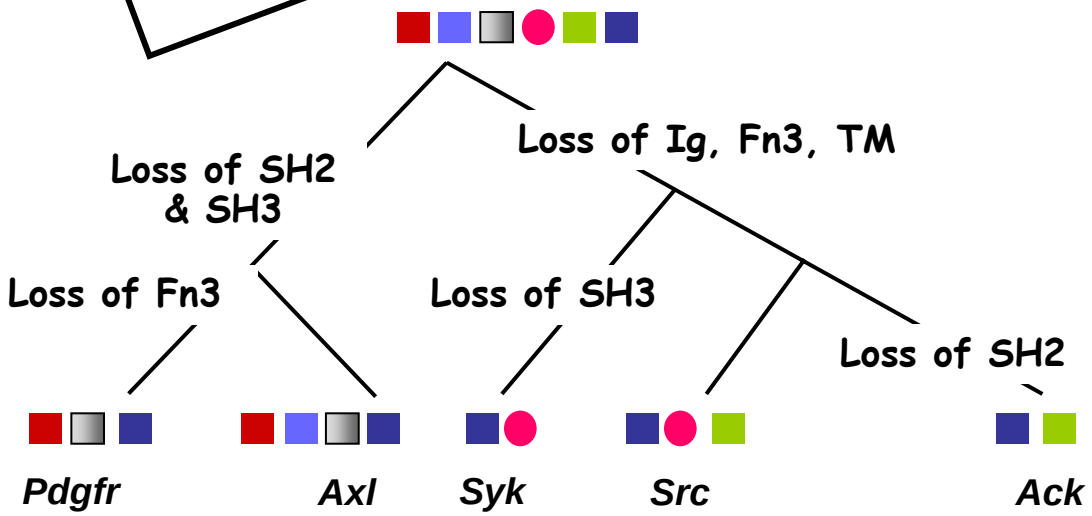


Problem: since losses are unrestricted, we can always construct a Dollo tree by starting with the complete set of domains

Existing model: Dollo Parsimony

Every binary character cha… 0 → 1 at most once,
1 → 0 u…

Not restrictive enough!



Taxa	Sh2	Sh3	Ig	FN3	TM
Pdgfr	0	0	1	0	1
Axl	0	0	1	1	1
Syk	1	0	0	0	0
Src	1	1	0	0	0
Csk	1	1	0	0	0
Ack	0	1	0	0	0
…	…	…	…		…

Problem: since losses are unrestricted, we can always construct a Dollo tree by starting with the complete set of domains

Comparing Existing Parsimony Models

Perfect Phylogeny

- Too restrictive to model multidomain evolution
- Given *n* domains and *m* proteins, find PP in *O(mn)*
- If a PP exists, it is guaranteed to be optimal.

Dollo Parsimony

- Not restrictive enough
- Can always find a DP, but not biologically realistic
- Finding optimal DP is NP-complete.

Goal: Domain parsimony model that is informative and tractable.

Multidomain Parsimony Models

A Dollo phylogeny is ***static (SDP)***

if for every ancestral protein, *p,* the set of domains in *p* is a subset of the domains in some leaf protein.

- **Domains architectures persist**

A Dollo phylogeny is ***conservative (CDP)***

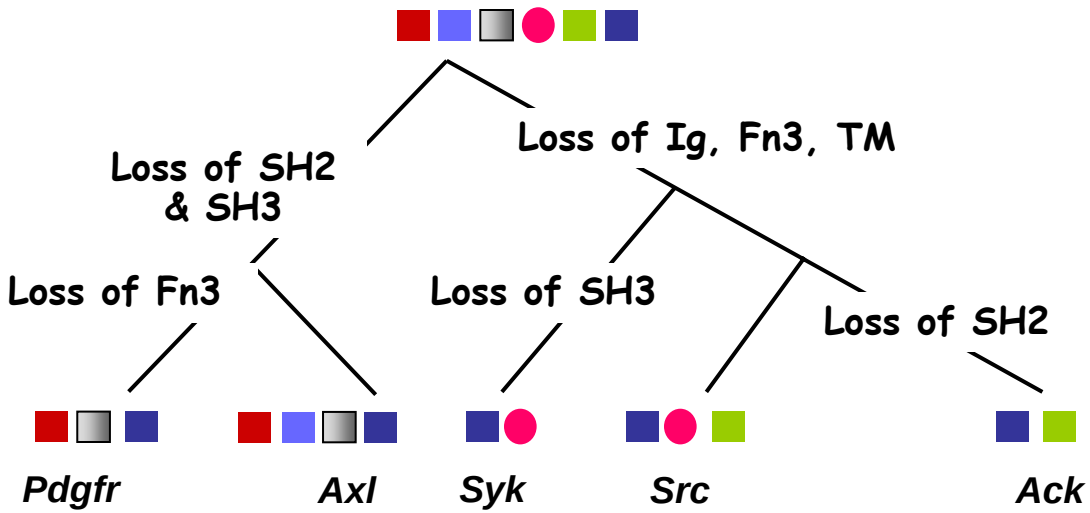
if for every *pair* of domains in an ancestral protein, there exists some leaf protein that also contains this *pair* of domains.

- **Domain insertions are rare**

Every Static Dollo phylogeny is also Conservative.

Static Dollo Parsimony

- Every binary character changes state 0 → 1 at most once.
- Every ancestral architecture is a subset of a leaf architecture



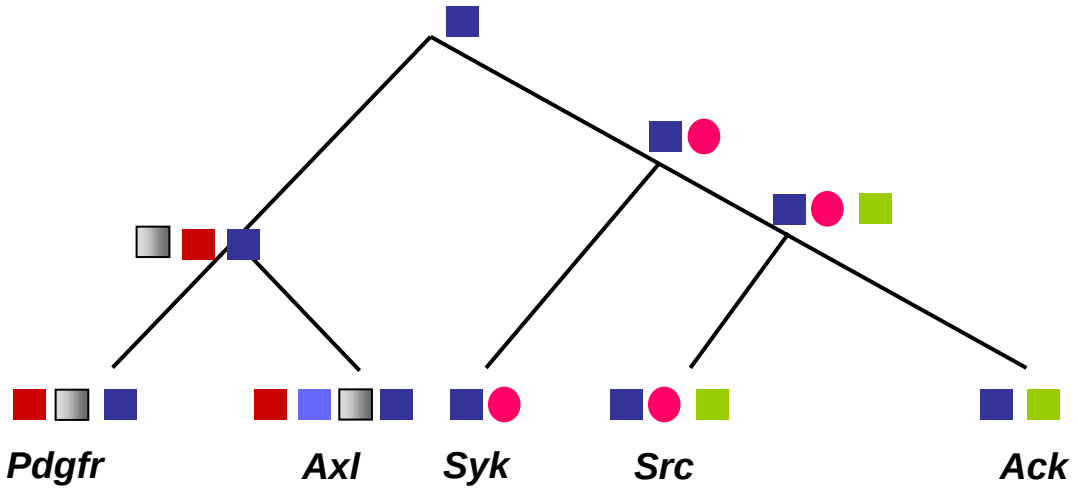
Taxa	Sh2	Sh3	Ig	FN3	TM
Pdgfr	0	0	1	0	1
Axl	0	0	1	1	1
Syk	1	0	0	0	0
Src	1	1	0	0	0
Csk	1	1	0	0	0
Ack	0	1	0	0	0
…	…	…	…		…

This tree is *not* an SDP.

More restrictive than Dollo Parsimony.

Static Dollo Parsimony

- Every binary character changes state 0 → 1 at most once.
- Every ancestral architecture is a subset of a leaf architecture



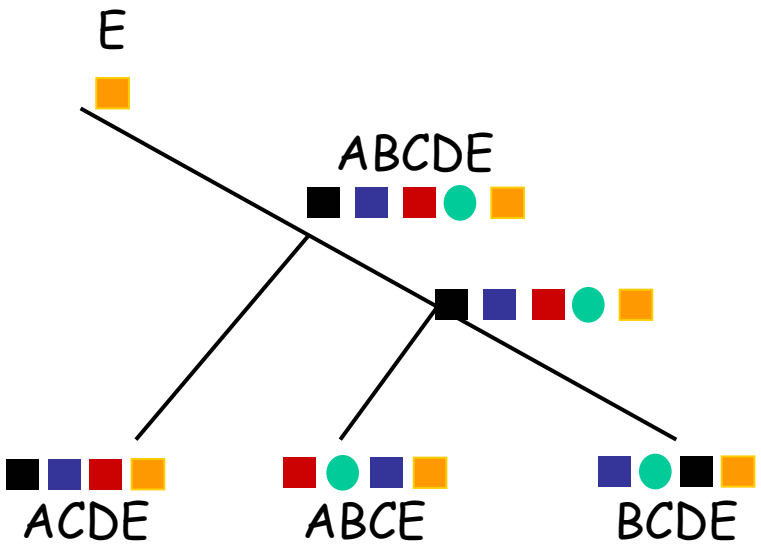
Taxa	Sh2	Sh3	Ig	FN3	TM
Pdgfr	0	0	1	0	1
Axl	0	0	1	1	1
Syk	1	0	0	0	0
Src	1	1	0	0	0
Csk	1	1	0	0	0
Ack	0	1	0	0	0
…	…	…	…		…

This tree is an SDP.

Less restrictive than Perfect Phylogeny

When Does Static Dollo Parsimony Fail?

- Every binary character changes state 0 → 1 at most once.
- Every ancestral architecture is a subset of a leaf architecture



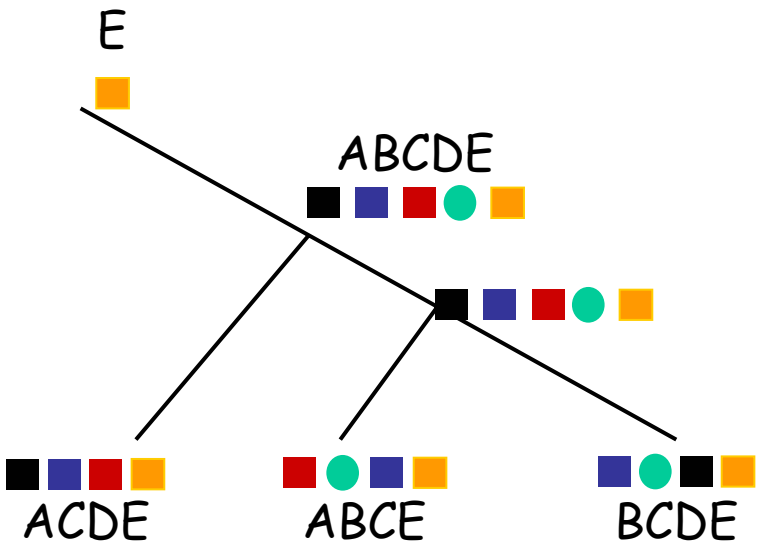
"*E*" superfamily

Taxa	*A*	*B*	*C*	*D*	*E*
Protein 1	0	1	1	1	1
Protein 2	1	0	1	1	1
Protein 3	1	1	1	0	1
…	…	…	…	…	…

This set does not admit an SDP.

Conservative Dollo Parsimony

- Every binary character changes state 0 → 1 at most once.
- Every pair of domains in an ancestral node also appears in a leaf node.



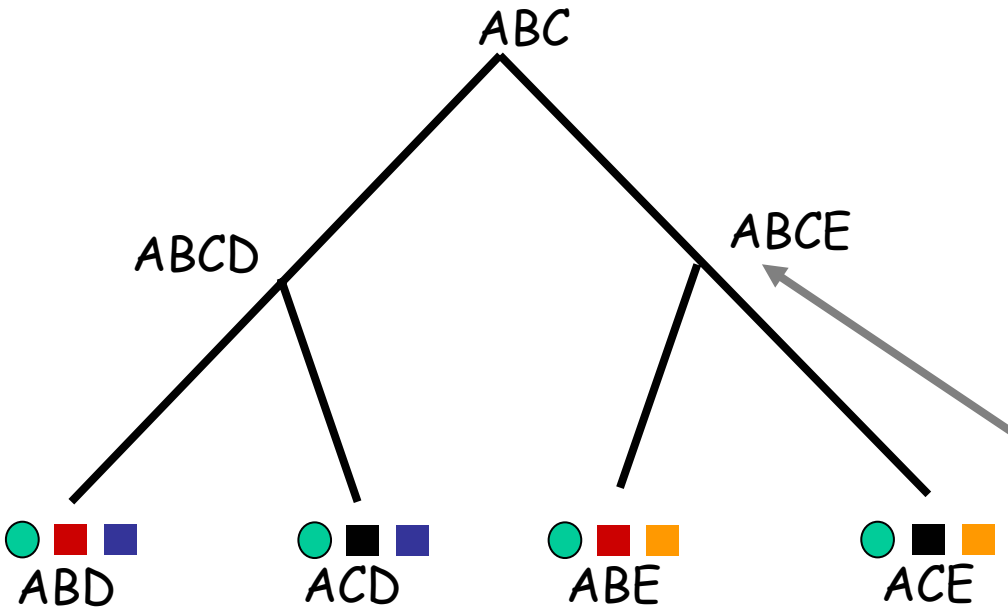
"*E*" superfamily

Taxa	*A*	*B*	*C*	*D*	*E*
Protein 1	0	1	1	1	1
Protein 2	1	0	1	1	1
Protein 3	1	1	1	0	1
…	…	…	…	…	…

This set *does* admit an CDP.

When Does Conservative Dollo Parsimony Fail?

- Every binary character changes state 0 → 1 at most once.
- Every pair of domains in an ancestral node also appears in a leaf node.



"A" superfamily

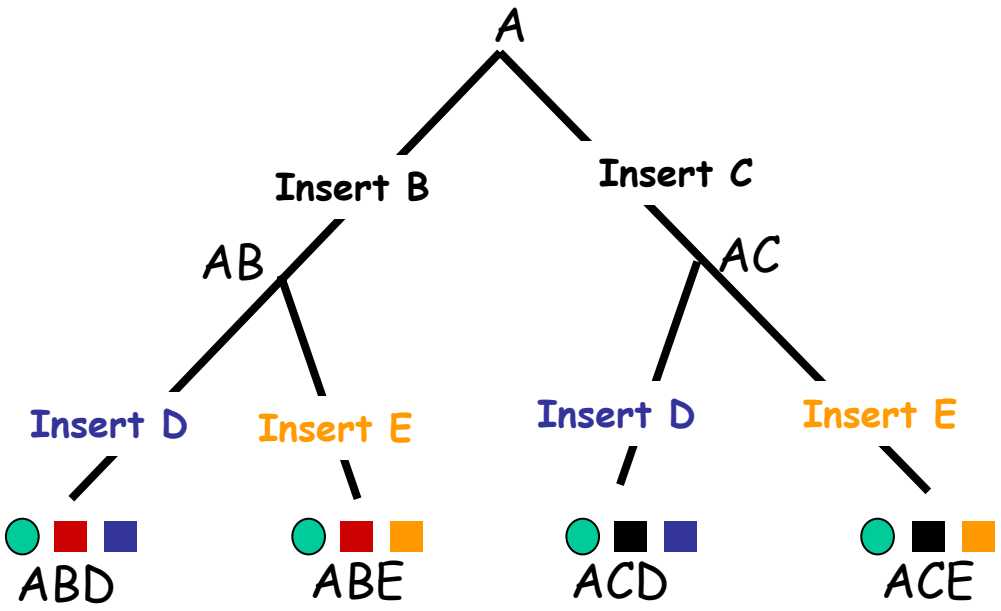
Taxa	B	C	D	E
Protein 1	1	0	1	0
Protein 2	1	0	0	1
Protein 3	0	1	1	0
Protein 3	0	1	0	1
...	...	...		...

This set does *not* admit a CDP.

The pair BC does not appear on any leaf!

When Does Conservative Dollo Parsimony Fail?

- Every binary character changes state 0 → 1 at most once.
- Every pair of domains in an ancestral node also appears in a leaf node.



"*A*" superfamily

Taxa	*B*	C	*D*	*E*
Protein 1	1	0	1	0
Protein 2	1	0	0	1
Protein 3	0	1	1	0
Protein 3	0	1	0	1
…	…	…		…

This set does *not* admit a CDP.

It is *conservative* but requires multiple independent domain insertions.

If a multidomain super family does not admit a *Conservative Dollo Phylogeny* then

- *either* the conservative assumption is violated (i.e., some ancestral protein contained a pair of domains that is not paired in any contemporary protein)
- *or* the Dollo assumption is violated (i.e., multiple independent insertions of the same pair of domains are required to explain the data.)

Our Approach

- **Introduce multi-domain parsimony models**
  - Conservative Dollo Parsimony
  - Static Dollo Parsimony
- ➢**Map these models to *Domain Overlap Graph (DOG)***
  - **Show CDP ↔ chordality in *DOG***
  - Show SDP ↔ cliques in DOG correspond to proteins
  - Adapt fast algorithms for testing these properties to *DOGs.*
- Apply test to all superfamilies in the SwissProt data base

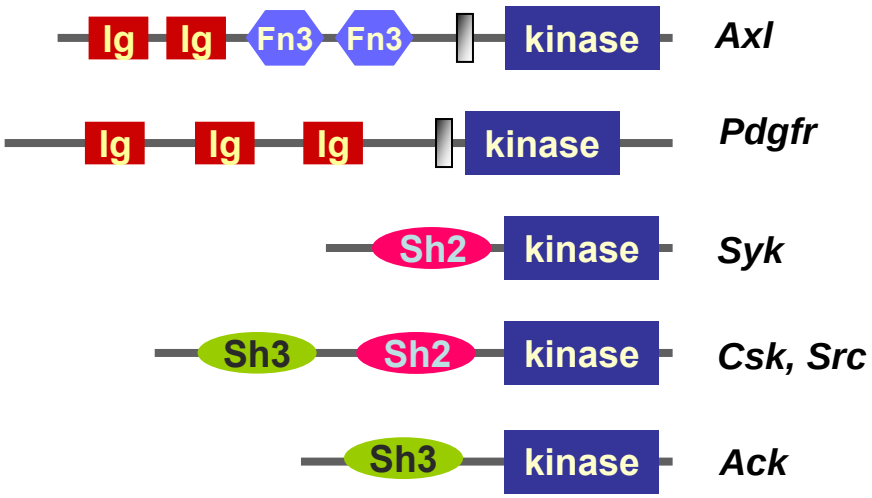
Domain Overlap Graph (DOG)

The DOG of a superfamily, *S(D)*, is the graph *G* such that

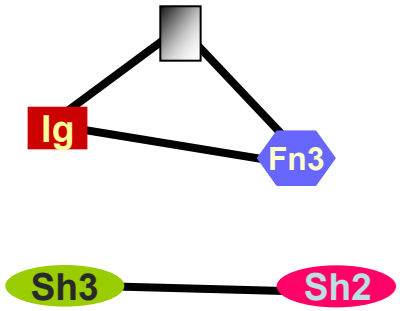
- every domain in *S\D* is a vertex in *G*
- there is an edge between two domains if they appear together in some protein in *S(D).*

EXAMPLE:

Proteins:



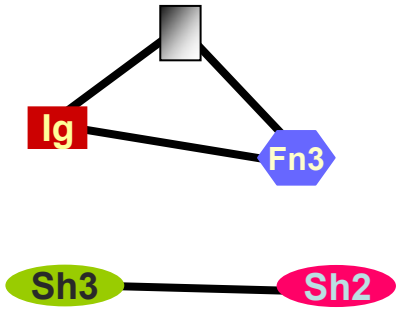
Domain Overlap Graph:



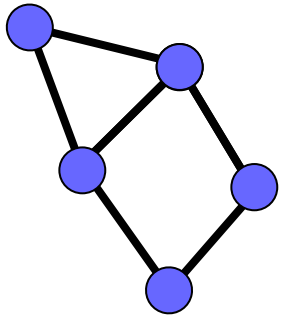
Conservative Dollo Parsimony

Definitions:

1. A *chord* is an edge joining two nonconsecutive vertices of a cycle.
2. A graph is *chordal* if every cycle of length ≥ 4 has a chord.



A chordal graph

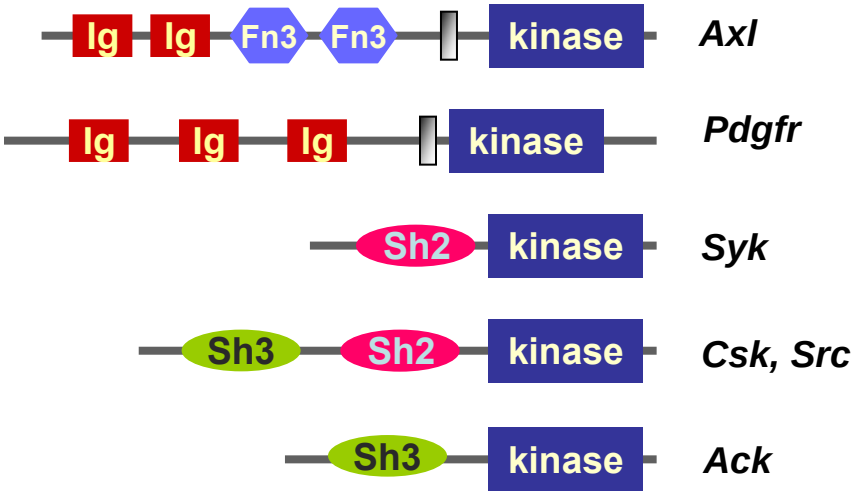


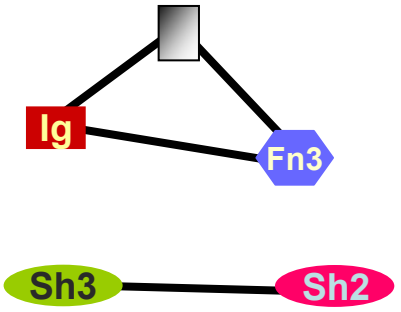
A non-chordal graph

Theorem 1 There exists a *Conservative Dollo Parsimony* tree for a given set of multidomain architectures, if and only if the domain overlap graph for this set is *chordal*.

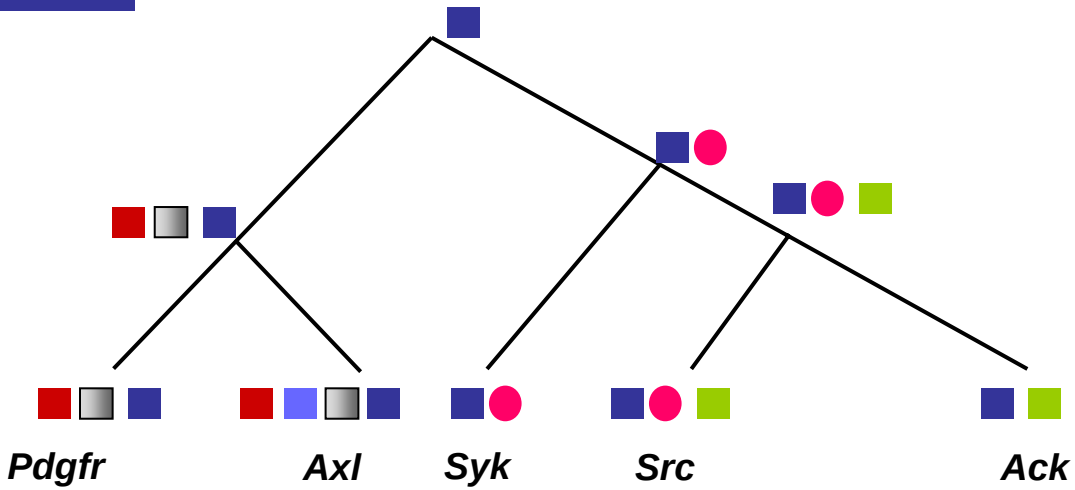
Chordality can be tested in $O(n+e)$ time, where n is the number of vertices and e is the number of edges *(Tarjan and Yannakakis, 1984).*

Examples





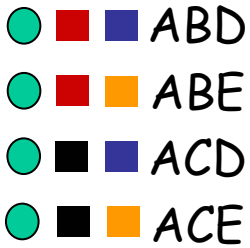
The DOG is chordal

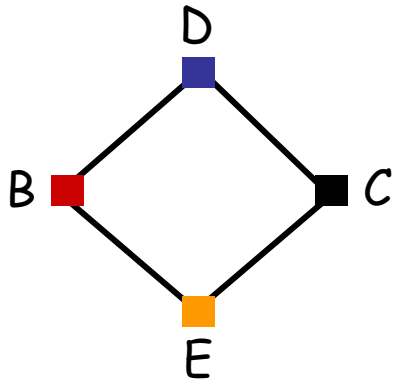


Conservative Dollo Phylogeny

Examples

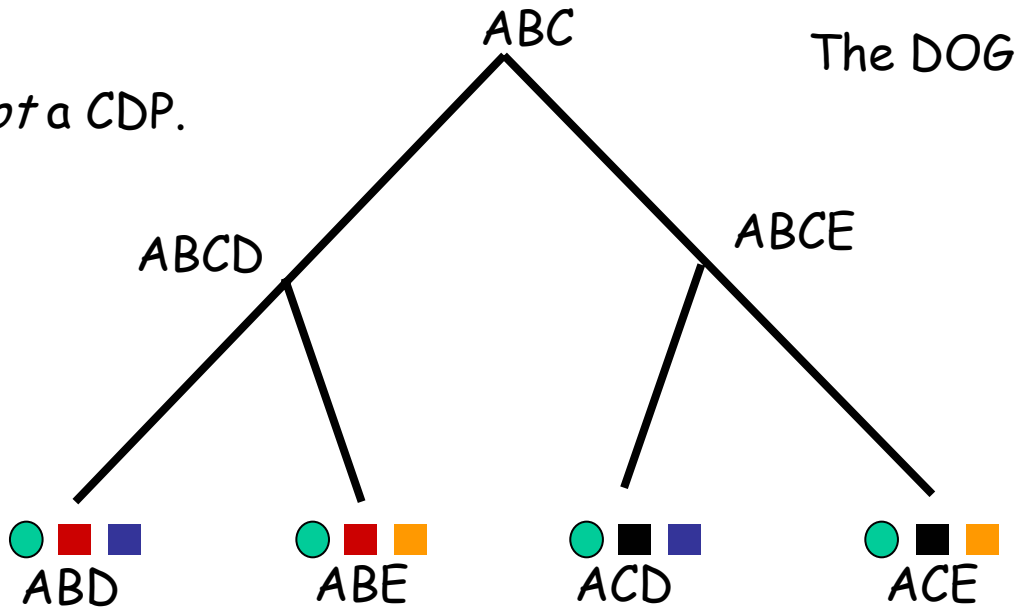
“A” superfamily



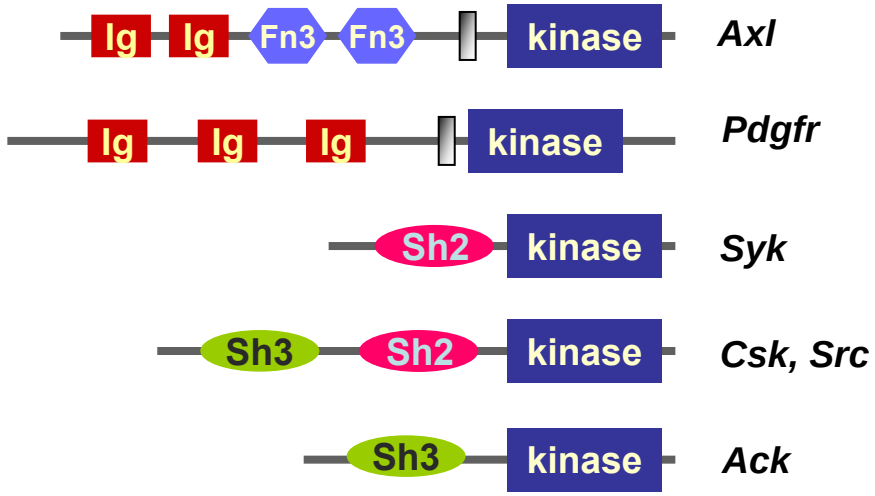


The DOG is *NOT* chordal

This *is not* a CDP.



Conservative Dollo Parsimony

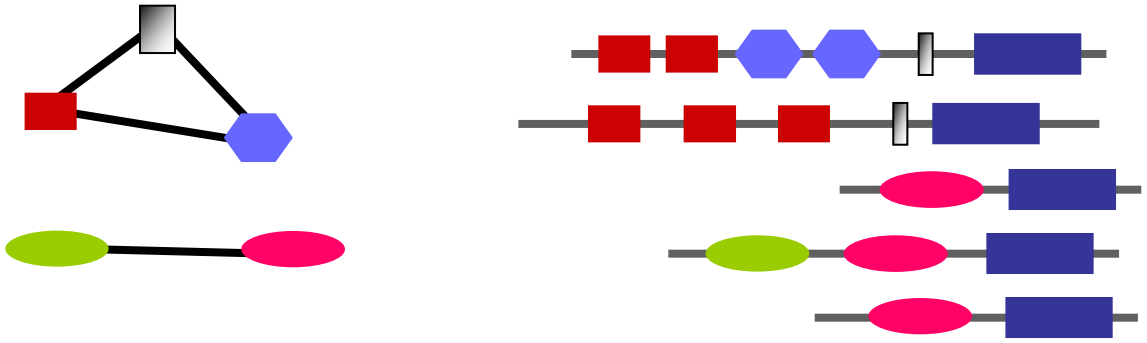


Theorem 2 Finding the optimal *Conservative Dollo Parsimony* tree for a given set of multidomain architectures is NP complete.

Our Approach

- **Introduce multi-domain parsimony models**
  - Conservative Dollo Parsimony
  - Static Dollo Parsimony
- ➢**Map these models to *Domain Overlap Graph (DOG)***
  - Show CDP ↔ chordality in *DOG*
  - **Show SDP ↔ cliques in DOG correspond to proteins**
  - Adapt fast algorithms for testing these properties to *DOGs.*
- Apply test to all superfamilies in the SwissProt data base

Static Dollo Parsimony

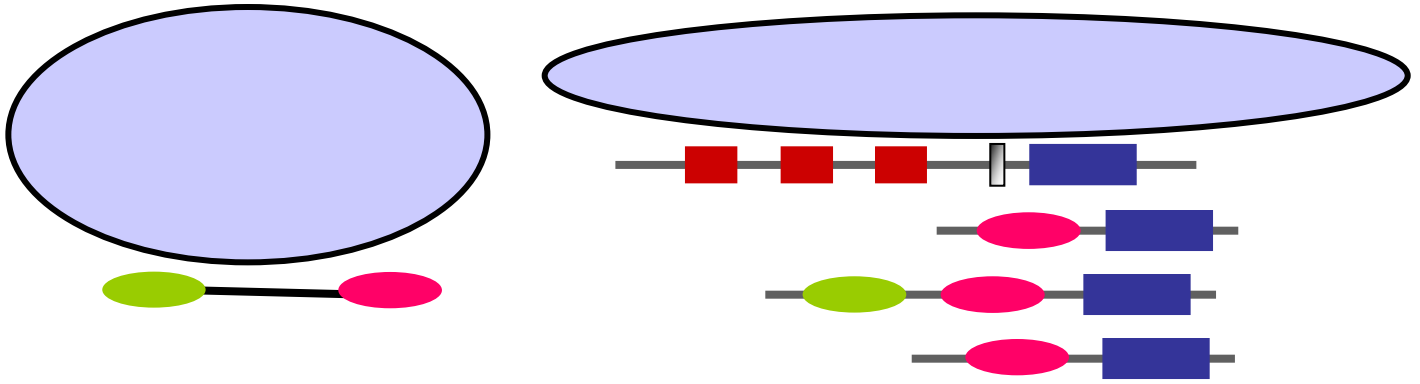


Theorem 3. There exists a *Static Dollo Parsimony* tree for a given set of multi-domain architectures, if and only if

1. the DOG for this set is *chordal*
2. for every clique in the DOG, there exists an architecture that contains all domains in the clique.

All maximal cliques in a chordal graph can be examined in *O(ne)* time *(Tarjan &Yannakakis, 1984).*

Static Dollo Parsimony

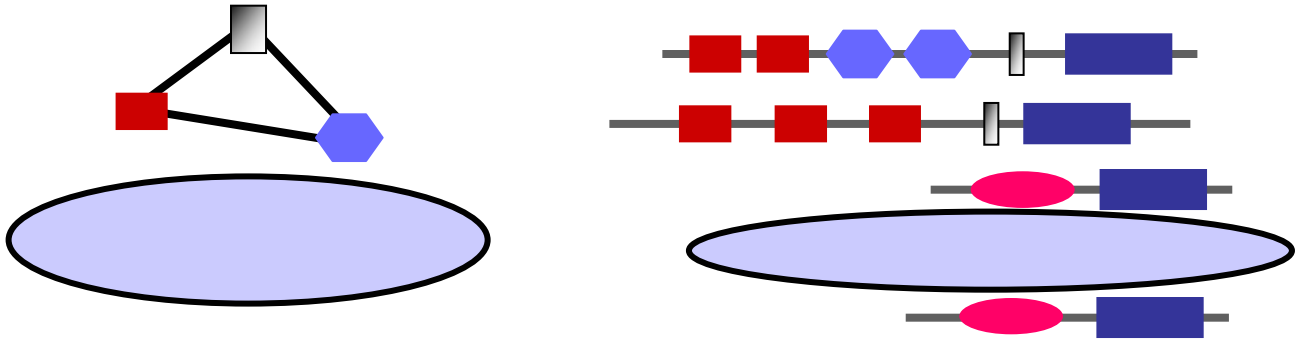


Theorem 2. There exists a *Static Dollo Parsimony* tree for a given set of multi-domain architectures, if and only if

1. the DOG for this set is *chordal*
2. for every clique in the DOG, there exists an architecture that contains all domains in the clique.

All maximal cliques in a chordal graph can be examined in *O(ne)* time *(Tarjan &Yannakakis, 1984).*

Static Dollo Parsimony



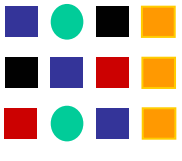
Theorem 2. There exists a *Static Dollo Parsimony* tree for a given set of multi-domain architectures, if and only if

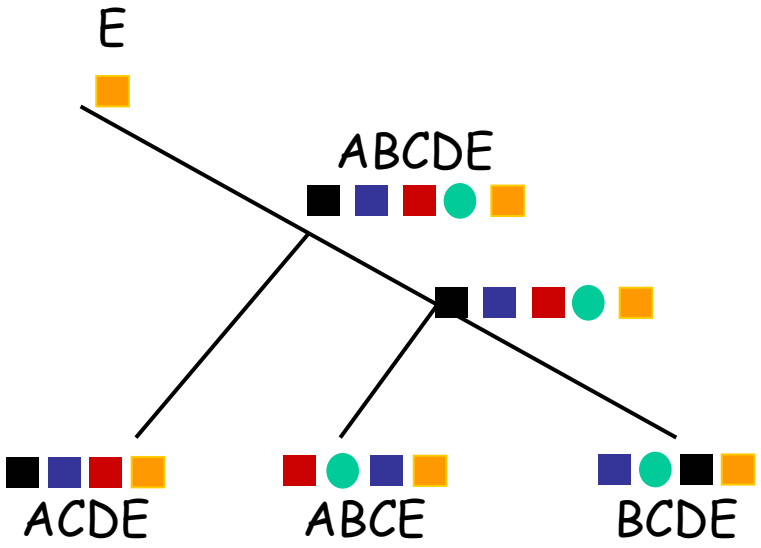
1. the DOG for this set is *chordal*
2. for every clique in the DOG, there exists an architecture that contains all domains in the clique.

All maximal cliques in a chordal graph can be examined in *O(ne)* time *(Tarjan &Yannakakis, 1984).*

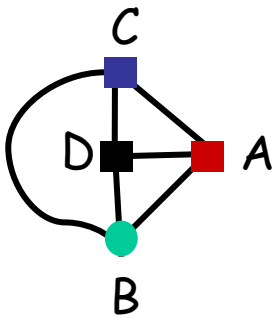
Example

"*E*" superfamily





This set admits a CDP but not an SDP.



The DOG is chordal

It contains the clique ABCD

The domain architecture ABCD does not appear in the protein set

Our Approach

- Introduce multi-domain parsimony models
  - Conservative Dollo Parsimony
  - Static Dollo Parsimony
- Map these models to *Domain Overlap Graph (DOG)*
  - Show CDP ↔ chordality in *DOG*
  - Show SDP ↔ Helly property in *DOG*
  - Adapt fast algorithms for testing these properties to *DOGs.*

➢ **Apply test to all superfamilies in the SwissProt database**

Experiments

- Data:
  - All non redundant (nr90) proteins in SwissProt.
  - Domain composition determined using CDD, based on PSSM domain models.
- For each of the 2896 domain superfamilies in data set:
  - Construct DOG
  - Test existence of PP, SDP, CDP.
- Estimating statistical significance
  - Test existence of PP, SDP, CDP in random graphs with comparable properties.

Estimating Statistical Significance

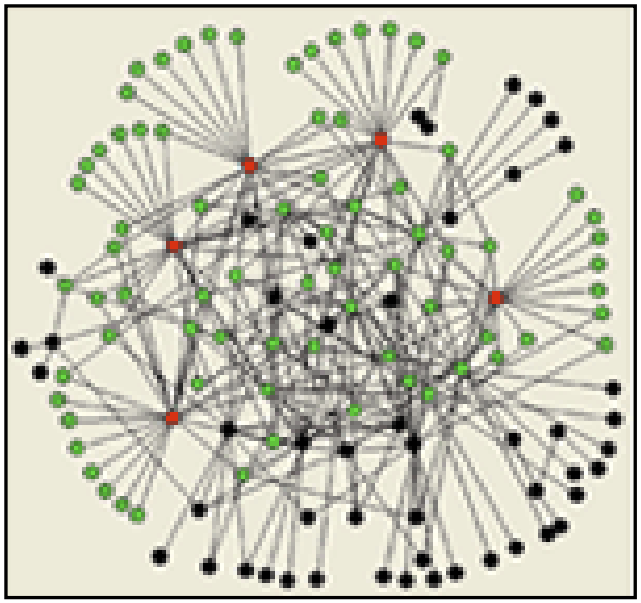
- Eliminate superfamilies that are chordal w/ high probability
  - All graphs with $n \leq 3$ vertices and all acyclic graphs are chordal.
  - Graphs with edge probability $p \leq 1/n$ are acyclic with high probability.
- For remaining 479 DOGs, used simulation to determine probability that a comparable random graph is chordal.

Null Models

Random graph models with comparable density:

1. Uniform random
2. Scale free: $P(\text{degree } k) \sim k^{-\gamma}$

Construction by preferential attachment:
Add nodes, preferentially attaching them to sites that are already well connected.



Results

*n**	# families	%PP	%SDP	%CDP	Random graphs	
					Uniform	Scale-free
4-5	143	57	99	99.5	80	98
6-8	130	37	99	100	31	66
9-10	40	28	100	100	17	25
11-20	104	13	87	99	1.7	1.0
21-30	34	6	53	88	0	0
≥30	28	0	15	50	0	0

34 superfamilies do not safisfy CDP, including TyrKc, Ig, PH, EGF, CUB, SH3, C1, Myosin_Tail

**n* is the number of distinct domains in the superfamily.

Conclusions

- Small families (≤ 20 domains) typically satisfy Static Dollo Parsimony
  - Domain architectures persist (maybe)
- Large families (≥ 30 domains) do not consistently satisfy Conservative Dollo Parsimony
  - Multiple domain insertions needed to explain the data.
- Multidomain superfamilies do not have the same topological structure as corresponding random scale-free graphs
  - Not consistent with evolution by preferential attachment .

Implications for Related Work

Genome Evolution: gene fusion versus gene fission, Snel, Bork & Huynen, TIG, 2000

Relative rates of gene fusion and fission in multi-domain proteins, Kummerfeld & Teichman, TIG, 2005

Supports use of parsimony model to investigate this question for small families. Our results suggest this is not unreasonable.

Implications for Related Work

Scale-free behavior in protein domain networks, Wuchty, MBE, 2001

Multi-domain protein families and domain pairs: Comparison with known structures and a random model of domain combination, Apic, Huber, Teichmann, J. Struct, Func. Genomics, 2003

Multi-domain superfamilies do not have the same topological structure as corresponding random scale-free graphs

- Not consistent with evolution by preferential attachment .

Future work

- Other uses of local graph structure to ask evolutionary and functional questions.
- Models that incorporates sequence mutation and domain insertion and deletion events.

Acknowledgements

Teresa Przytycka, NIH

Aiton Goldman
George Davis
Rose Hoberman
Jacob Joseph
Annette McLeod
Narayanan Raghupathy
Robbie Sedgewick
Nan Song
Maureen Stolzer
Ben Vernot

NHGRI, Packard Foundation