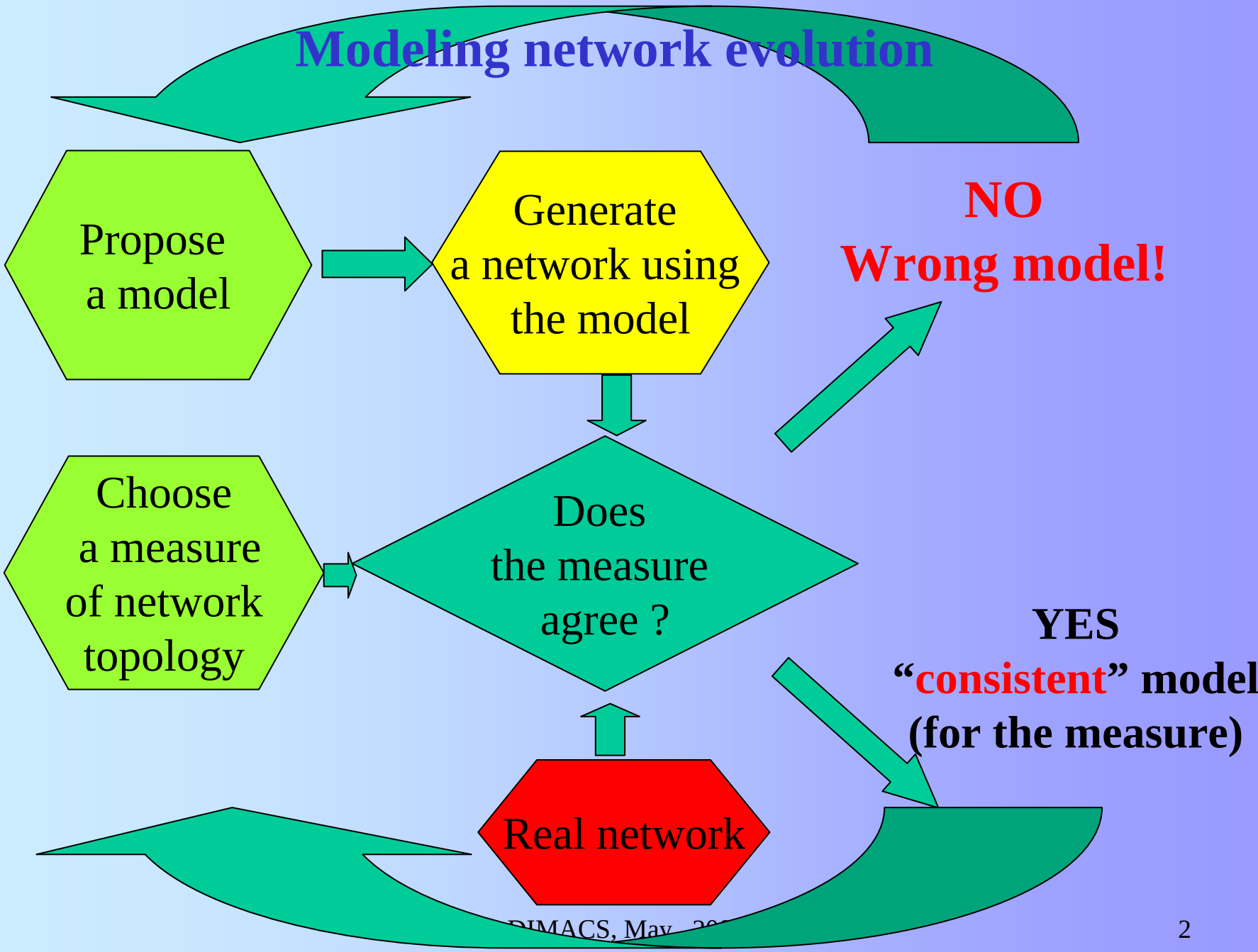
Network topology and evolution of hard to gain and hard to loose attributes

Teresa Przytycka

NIH / NLM / NCBI

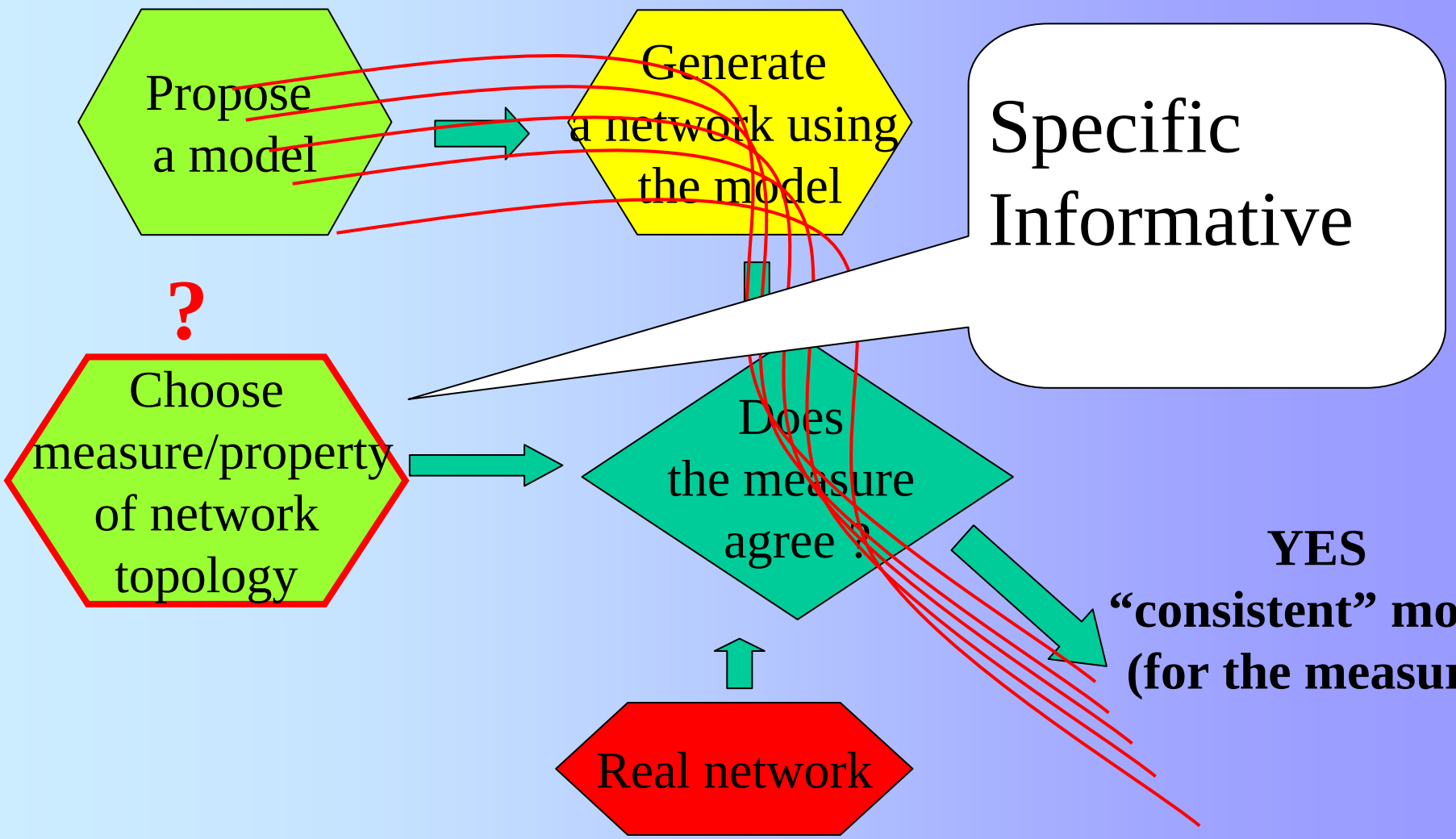
Modeling network evolution



Network Measurements

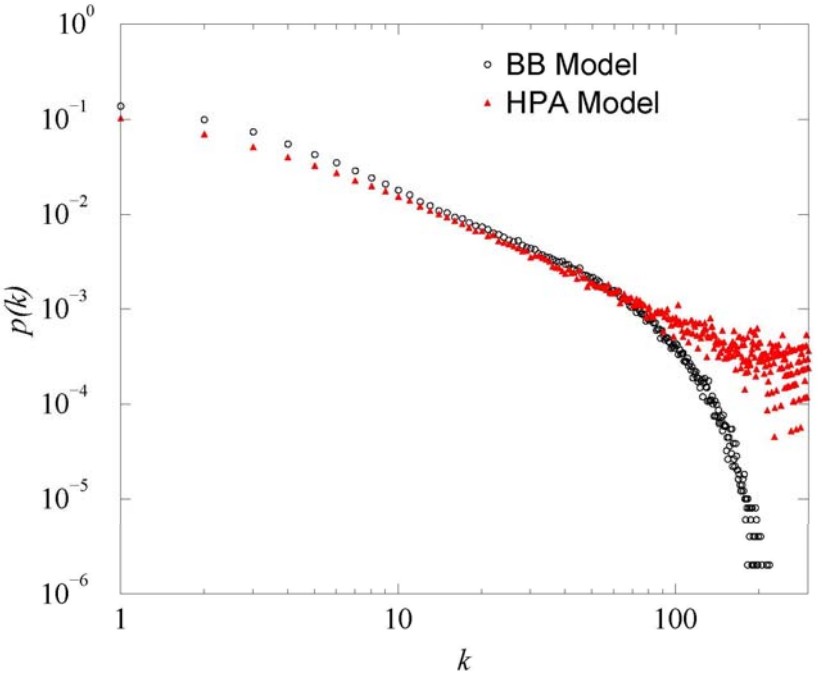
- Degree distribution
- Diameter
- Clustering coefficient
- Distribution of connected / bi-connected components
- Distribution of networks motifs

Choosing network measurement



Degree distribution is not specific

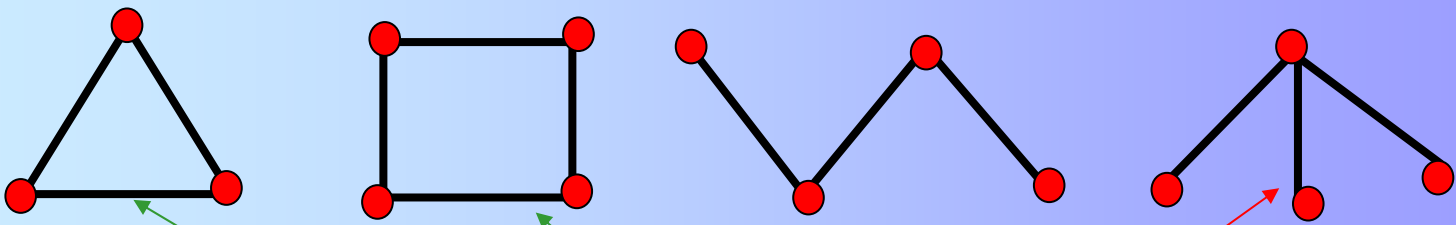
T. Przytycka, Yi-Kou Yu, short paper ISMB 2004
J. Comp. Biol. and Chem. 2004



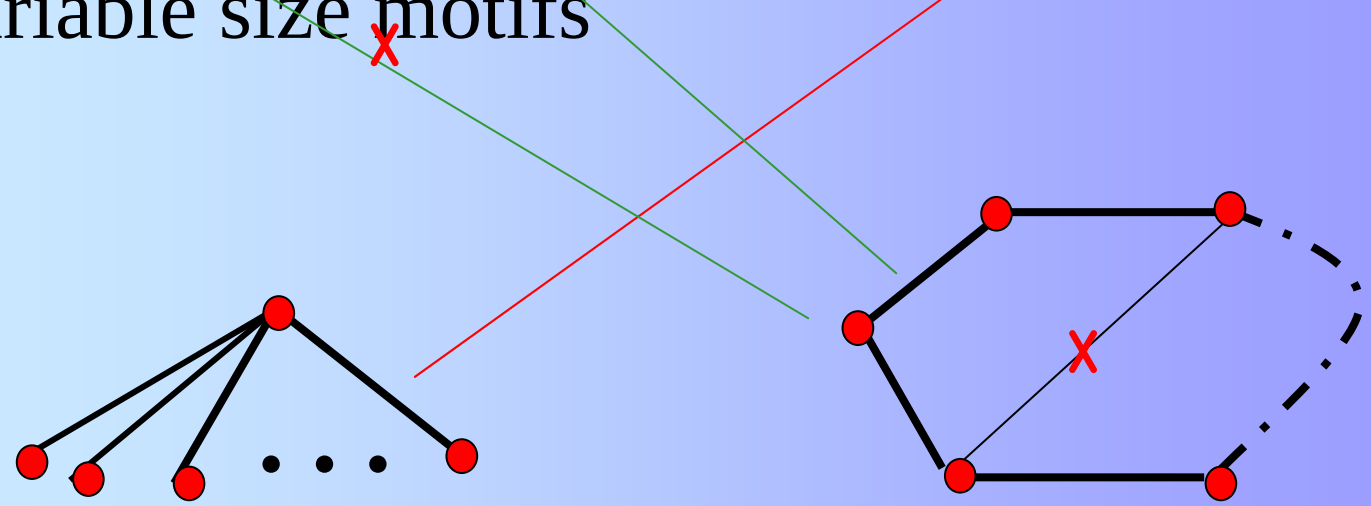
Both models agree not only with other on a large interval but also with the data (data points not shown). The real data is in the interval (1,80)

Network motifs

- Fixed size motif



- Variable size motifs



Working with small fixed size motifs

Enumerate all small size motifs in a network and observe which are under and over represented and try to understand the reasons.

[Alon, Kashtan, Milo, Pinter, ….]

Machine learning approach - use small size network motif to train a machine learning program to recognize networks generated by a given model.

[Middendorf, Ziv, Wiggins, PANS 2005]

Variable size motifs

- Have to focus on particular families (cannot handle all possibilities)
- Even within one family motifs can be enumerated efficiently (eg. fans) some are known to be hard – holes, cliques
- Which motives are may be interesting to examine?
- THIS TALK: holes and their role in identifying networks that corresponds to hard to gain and hard to loose characters.

Character Overlap Graph

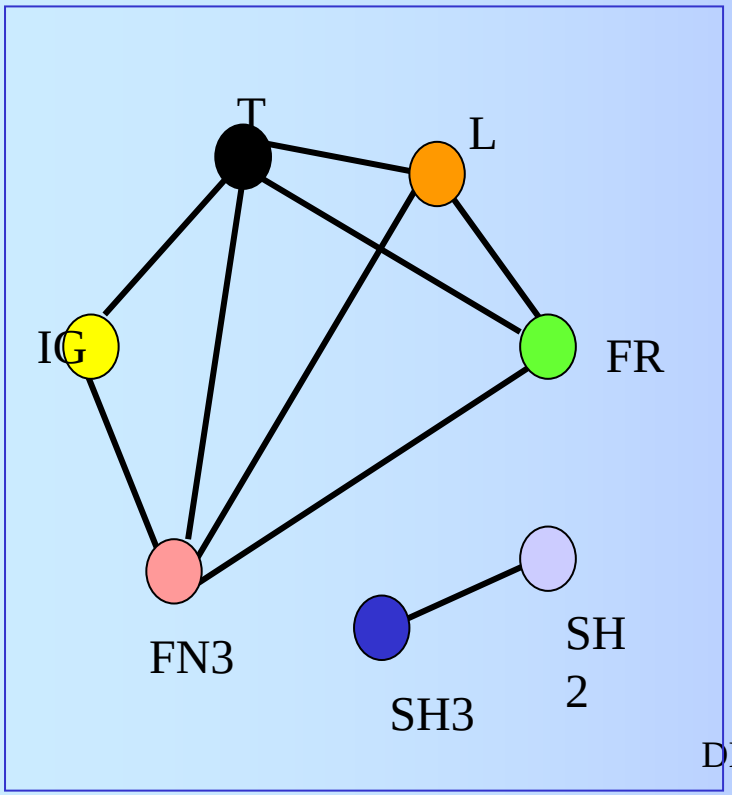
- Given:
  - set of biological units,
  - each described by a set of characters
  - the units evolve by loosing and gaining characters
- Examples

biological units	characters
multidomain proteins	domains
genes	introns
genomes	genes

continued

Character overlap graph

- Characters = nodes
- Two nodes are connected by an edge if there is a unit which contains both characters



Domain (overlap) graph

-Wuchty 2001

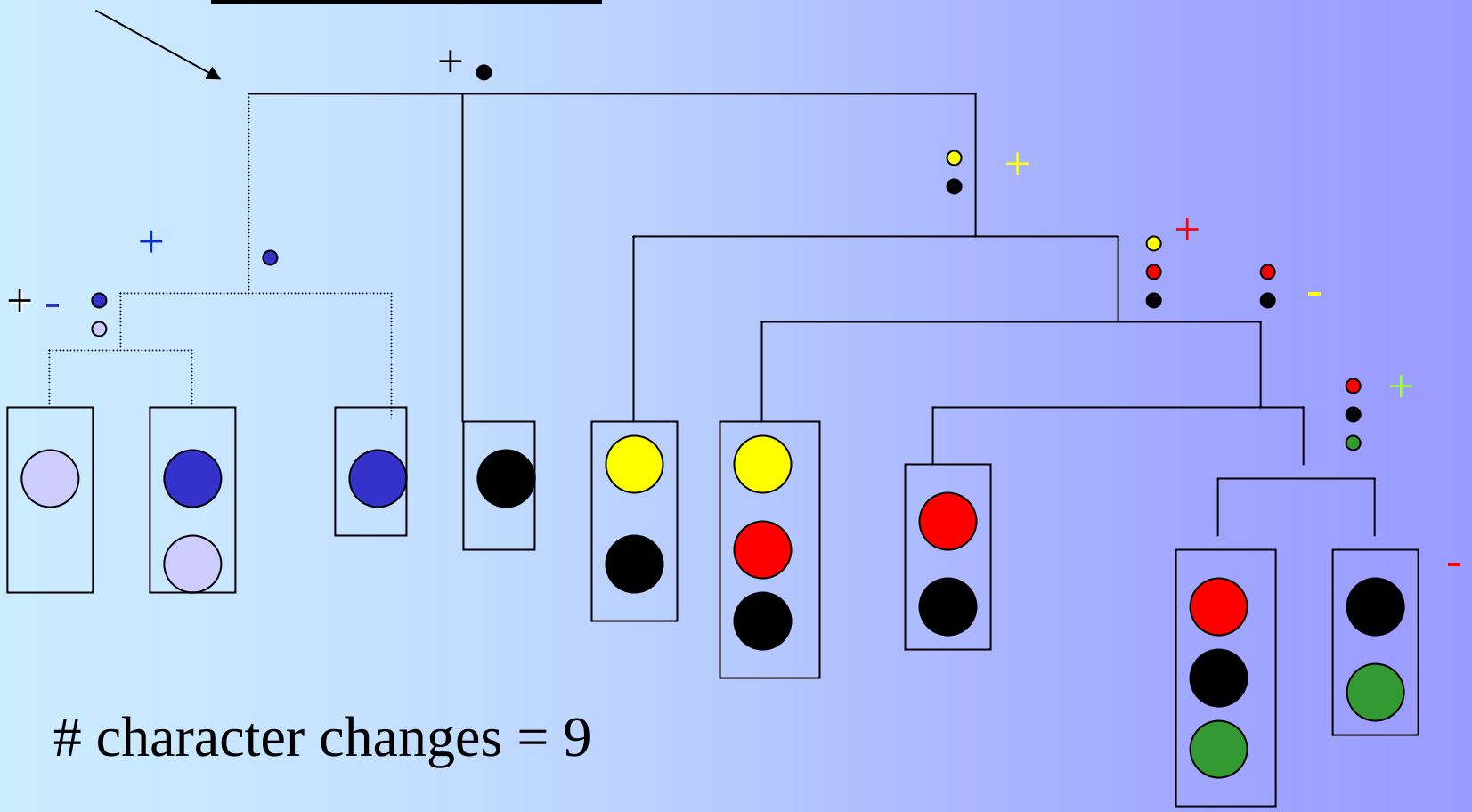
-Apic, Huber, Teichmann, 2oo3

-Przytycka, Davis, Song, Durand
RECOM 2005

Characters hard to gain and Dollo parsimony

- Maximum Parsimony: Build a tree with taxa in the leaves and where internal nodes are labeled with inferred character state such that the total number of character insertions and deletion along edges is minimized.
- Dollo parsimony – only one insertion per character is allowed

Example:



character changes = 9

- **If it is known that characters are hard to gain then use of Dollo parsimony is justified**

- **but if "hard to get" assumption is incorrect Dollo tree can still be constructed.**

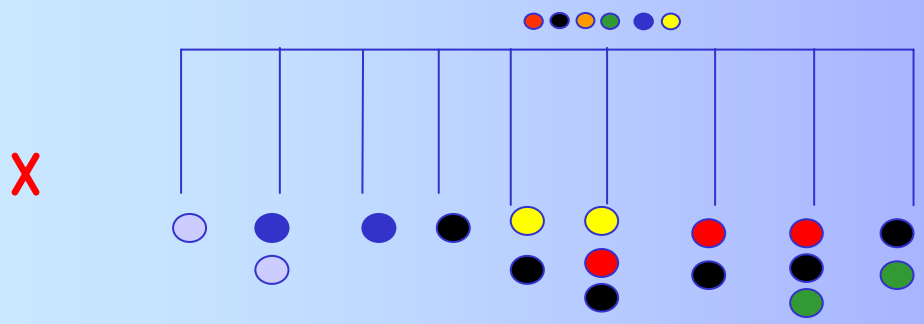
- **Is there a topological signal that would indicate that the assumption is wrong?**

Conservative Dollo parsimony

Przytycka, Davis, Song, Durand *RECOM 2005*

Every pair of domains that is inferred to belong to the same ancestral architecture (internal node) is observed in some existing protein (leaf)

Motivation: Domains typically correspond to functional unit and multidomain proteins bring these units together for greater efficiency



Theorem: Przytycka, Davis, Song, Durand *RECOM 2005*

There exits a conservative Dollo parsimony if and only if character overlap graphs does not contain holes.

(A graph without holes is also called **chordal** or **triangulated**)

Comment 1: There exist a fast algorithm for testing chordality (Tarjan, Yannakakis, 1984)

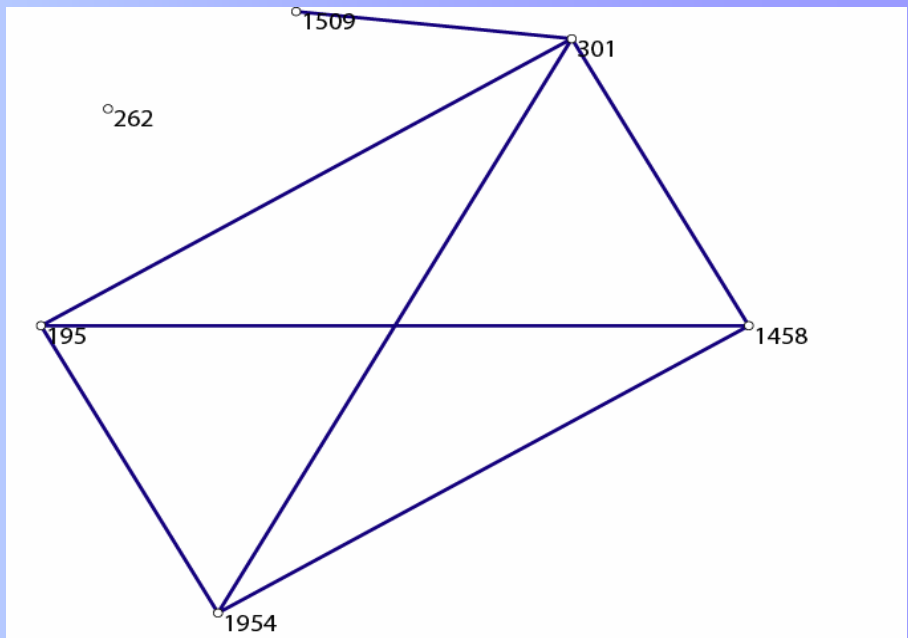
Comment 2: Computing Dollo tree is NP-complete (Day, Johnson, Sankoff, 186)

Holes and intron overlap graphs

- Intron data:

684 KOGs (groups of orthologous genes form 8 genomes) [Rogozin, Wolf, Sorokin, Mirkin, Koonin, 2003]

- Only two graphs had holes.
- Possible explanations:
  - Most of the graphs are small and it is just by chance?
  - Something else?



KOG0010

	195	262	301	1458	1509	1954
Sc	0	0	0	0	0	0
Dr	1	0	0	0	0	1
Ar	0	0	1	0	1	0
Ca	1	0	1	1	0	1
Ho	1	0	0	1	0	0
Sa	0	0	0	0	0	0
An	1	0	0	0	0	0
Pl	0	1	0	0	0	0

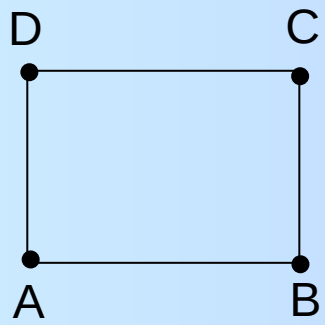
Assume that characters are hard to gain, if additionally they are hard to loose, what would the character overlap graph be like?

Assume parsimony model where each character is gained once and lost at most once.

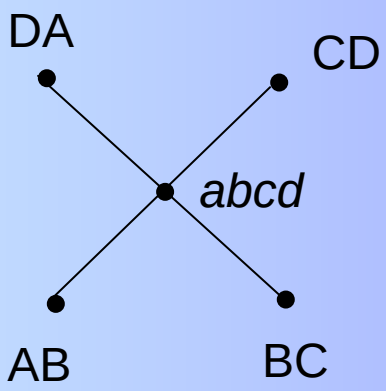
- **Theorem: If each character gained once and is lost at most once than the character overlap graph is chordal (that is has no holes).**

- **Note no "if and only if"**

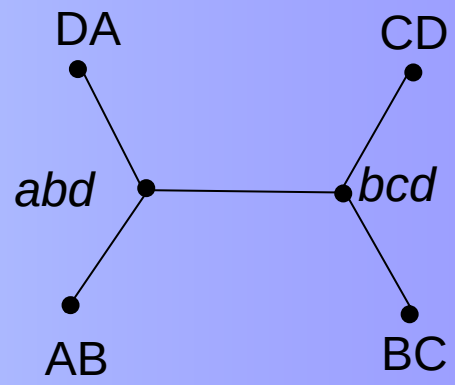
Informal justification



a)



b)



c)

Applying the theorem to intron overlap graph and domains overlap graph

Before we go further …

- **We have a hammer that is too big for the KOG intron data.**
- **Why**
  - There are only 8 genomes – NP completeness of computing optimal Dollo tree is not a an issue here
  - We know the taxonomy tree for these 8 species thus all one needs to do is to find the optimal (Dollo) labeling of such three which is computationally easy problem.
  - Such fixed taxonomy tree analysis has been already done (by the group whose data we are using)
- **But we hope to gain an insight into a general principles not particular application**

Intron data

- Concatenated intron data has been used (total 7236 introns; 1790 after removing singletons)
- But now the question if the graph is chordal has an obvious no answer (we seen already two examples while analyzing KOGs separately)
- Counting all holes is not an option
- We count holes of size four (squares)

Domain Data

Przytycka, Davis, Song, Durand *RECOM 2005*

Superfamily = all multidomain architectures that contain a fixed domain.

We extracted 1140 of such superfamilies and constructed domain overlap graph for superfamily separately (considered graphs that have at least 4 nodes only).

Null model

Have the same number of biological units as the real model

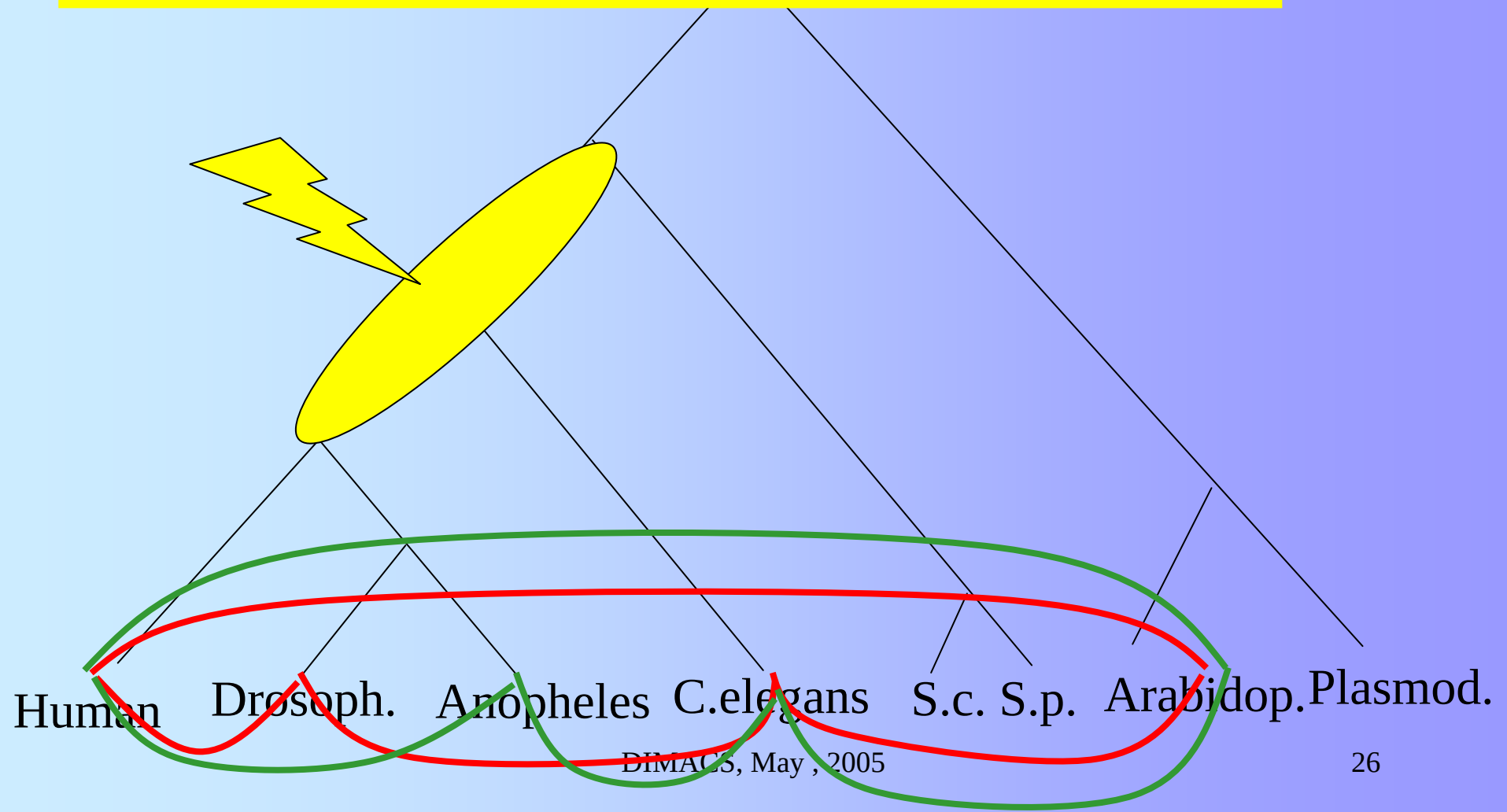
Each unit from the null model corresponds to one real unit and has the same number of characters as the real unit but randomly selected.

Results

Type of character overlap graph	Number of squares in real data	Number of squares in null model
domains	251	55,983
introns	145,555	84,258

Where are these intron holes ?

Multiple independent (?) deletions more than in null model



Compare to the following picture

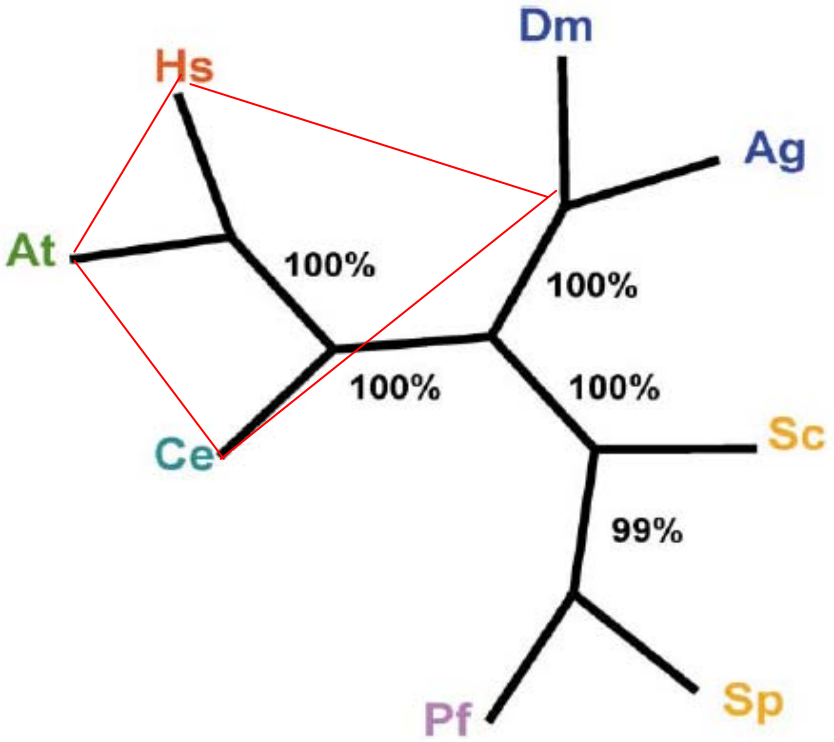


Figure 2. A Maximum Parsimony Tree Based on the Concatenated Intron Absence/Presence Data

Only the data for conserved alignment regions were analyzed. The unrooted tree was constructed by using Dollo parsimony. Only one most parsimonious tree was obtained; the numbers at the interior branches are bootstrap values with 1000 replicates. The species abbreviations are as in Figure 1.

Summary & Conclusions

- We identified network motifs that are very informative in establishing whether characters are hard to gain and hard to loose.
- We identified them following a graph theoretical reasoning rather than discovering difference between real and null model and then proposing an explanation.

Acknowledgments

- Conservative Dollo tree theorem is part of a joint work on evolution of multidomain architecture done in collaboration with Dannie Durand and her lab (RECOMB 2005)
- Lab members
  - Raja Jothi
  - Elena Zotenko
- And NCBI journal club discussion group