Phylogenetic Transfer of Knowledge

Bernard M.E. Moret and Xiuwei Zhang

Laboratory for Computational Biology and Bioinformatics

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

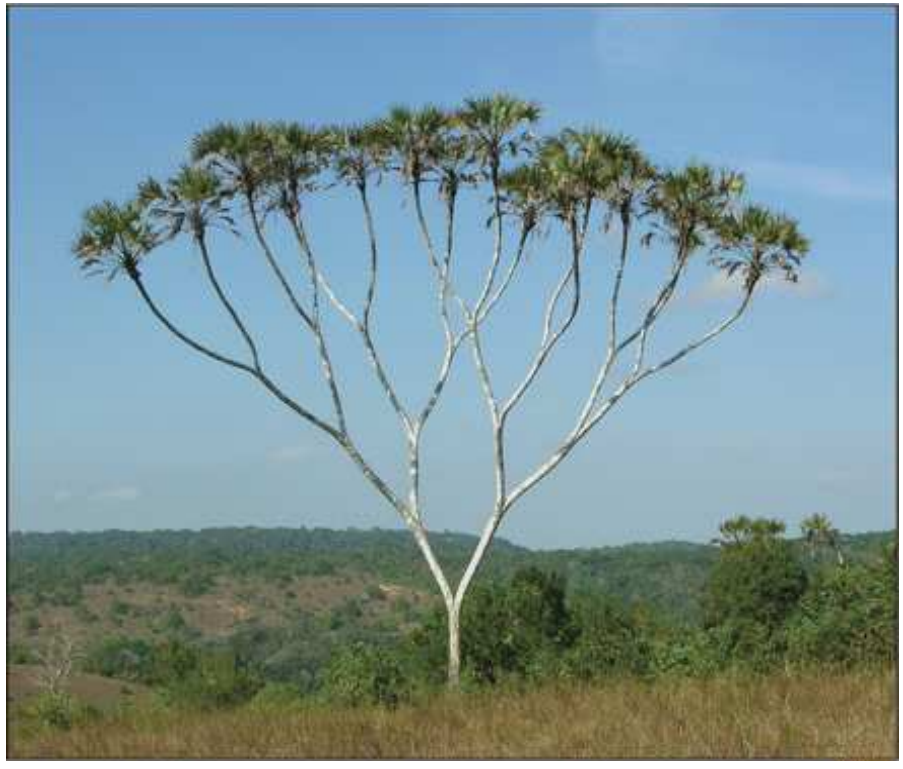
What is Phylogenetic Inference?

Phylogenetic inference attempts to reconstruct (a cartoon of) the evolutionary history of a collection of taxa (e.g., species).

What is Phylogenetic Inference?

Phylogenetic inference attempts to reconstruct (a cartoon of) the evolutionary history of a collection of taxa (e.g., species).

This history typically takes the form of a tree.

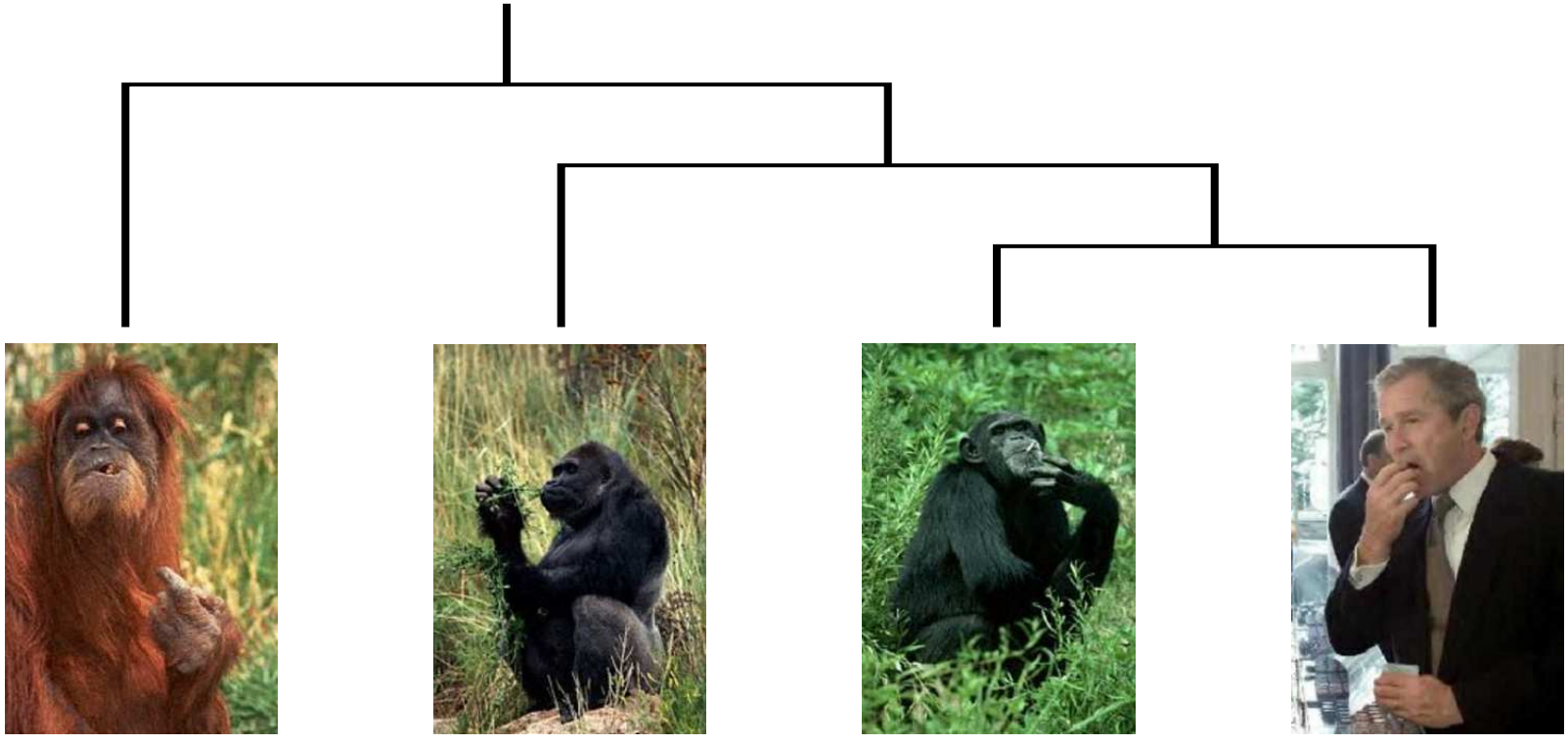


The Doum palm Hyphaene compressa *in Kenya (Photo: Charles Godfray)*

What is Phylogenetic Inference?

Phylogenetic inference attempts to reconstruct (a cartoon of) the evolutionary history of a collection of taxa (e.g., species).

A real phylogeny.



Uses of Phylogenetic Inference

As famously stated by Th. Dobzhansky (1973):

nothing makes sense in biology except in the light of evolution.

Phylogenies embody and display evolution.

Phylogenetic inference has become a mainstay of computational biology, with over 10,000 citations per year to inference packages.

Phylogenies can be used with

any system that evolved from a common ancestor .

Taxa can be biological species, but also genes, protein folds, biological networks, pathogens and their hosts, pattern of epidemics, etc.

Beyond these, taxa can also be languages, ethnic customs, craft products (from flint arrowhead to computer worms), artistic styles, fashions, etc.

How Do We Infer a Phylogeny?

We need:

Comparable data (homologous characters) on contemporary taxa

DNA sequences, protein structures, contact networks, regulatory networks, morphological characters, brush strokes, etc.

A model of evolution for these data

character substitution matrices, gains and losses of morphologic characters, tandem and segmental duplication of genomic regions, genomic rearrangements, etc.

An inference algorithm

simple heuristics such as neighbor-joining as well as search and estimation procedures for NP-hard optimization criteria such as maximum parsimony and maximum likelihood

Uses of Phylogenies

Phylogenies are (almost) everywhere:

Fundamental research in evolutionary biology, systems biology, biomedicine, clinical medicine, etc.

Public health (host-pathogen co-evolution, vaccine design, spread of disease)

Drug design

Agriculture

Conservation biology

Linguistics

Anthropology (e.g., migration patterns, dispersion of memes)

Sociology (e.g., evolution of social networks)

Art history (e.g., evolution of styles and techniques, forgery detection)

Security (ditto)

Uses of Phylogenies

Phylogenies are (almost) everywhere:

Fundamental research in evolutionary biology, systems biology, biomedicine, clinical medicine, etc.

Public health (host-pathogen co-evolution, vaccine design, spread of disease)

Drug design

Agriculture

Conservation biology

Linguistics

Anthropology (e.g., migration patterns, dispersion of memes)

Sociology (e.g., evolution of social networks)

Art history (e.g., evolution of styles and techniques, forgery detection)

Security (ditto)

But they are not used enough!

Transfer of Knowledge

Comparative methods

The workhorse of computational biology, also known as "guilt by association."

Knowledge gained in well studied systems is transferred to a system under study using pairwise comparisons.

Transfer of Knowledge

Comparative methods

The workhorse of computational biology, also known as "guilt by association."

Knowledge gained in well studied systems is transferred to a system under study using pairwise comparisons.

Transfer learning / Inductive transfer

In machine learning, an approach to abstracting knowledge gained on one or more problems in order to apply it to another problem, often using graphical models.

Pairwise Comparisons in CompBio

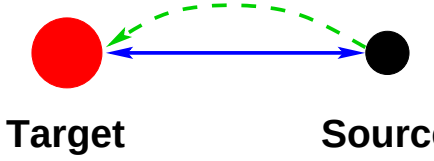
The basis for most homology and orthology assignments.
The foundation of all homology-based inference methods (gene hunting, structure prediction, functional prediction, etc.)

Works well for closely related systems, but degrades rapidly with increased evolutionary distance.
Fortunately, there remains a lot of low-hanging fruit.

To handle more distantly related systems, biologists have used multiple pairwise comparisons—between the target system and several known systems).
However, reconciling conflicting predictions becomes a difficult problem.

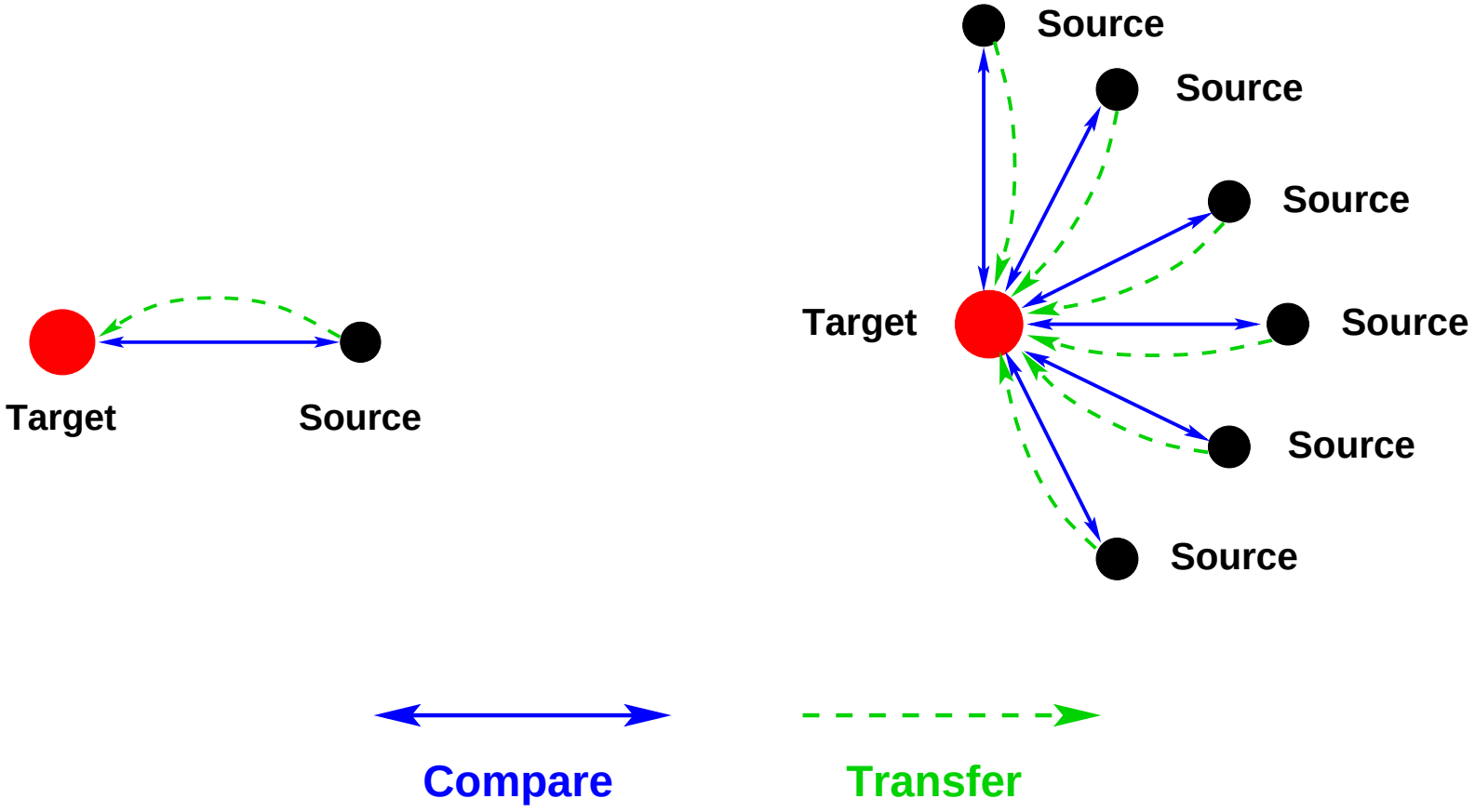
What we need is a model for integration: an evolutionary context.

Schemata for Transfer of Knowledge



single prediction

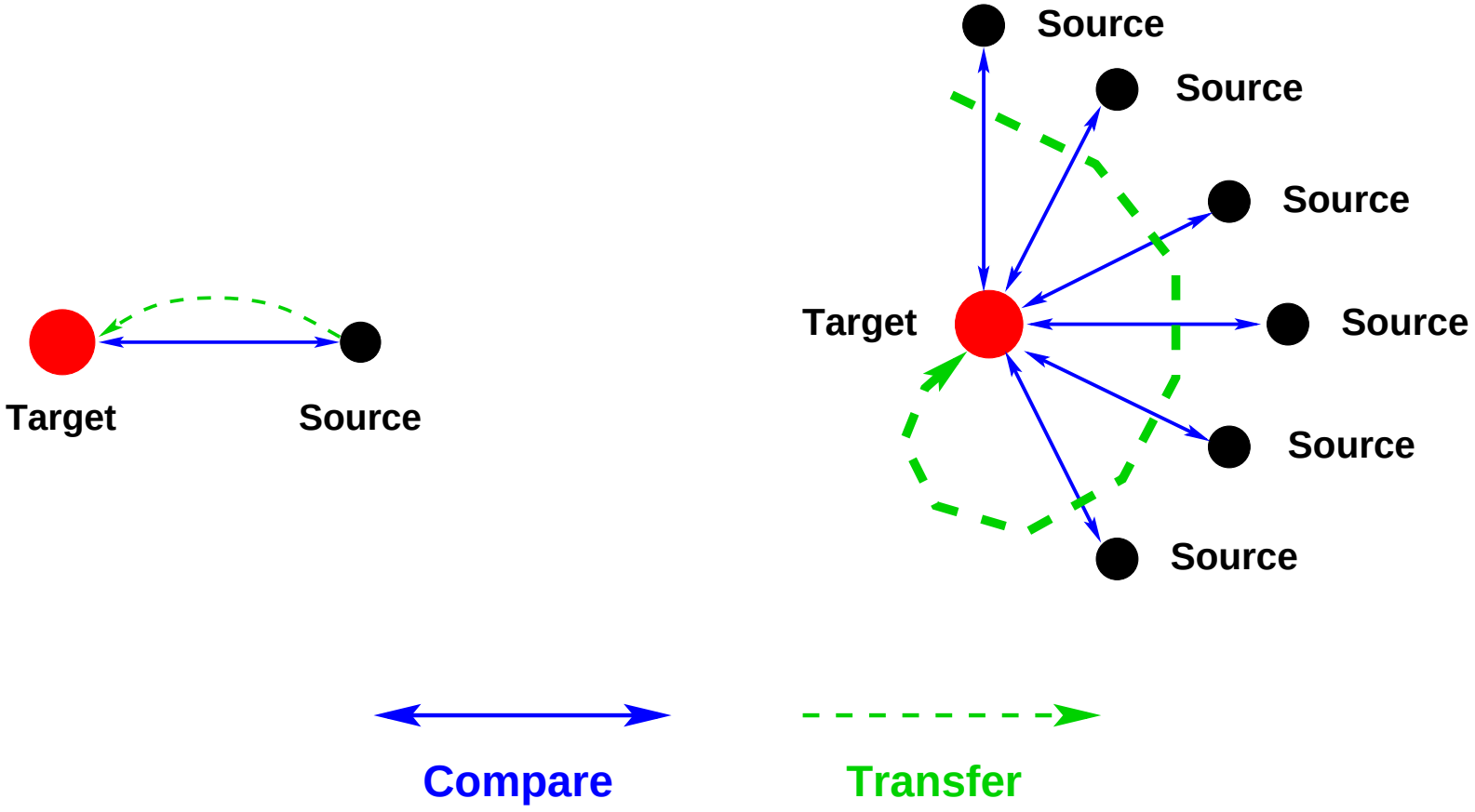
Schemata for Transfer of Knowledge



single prediction

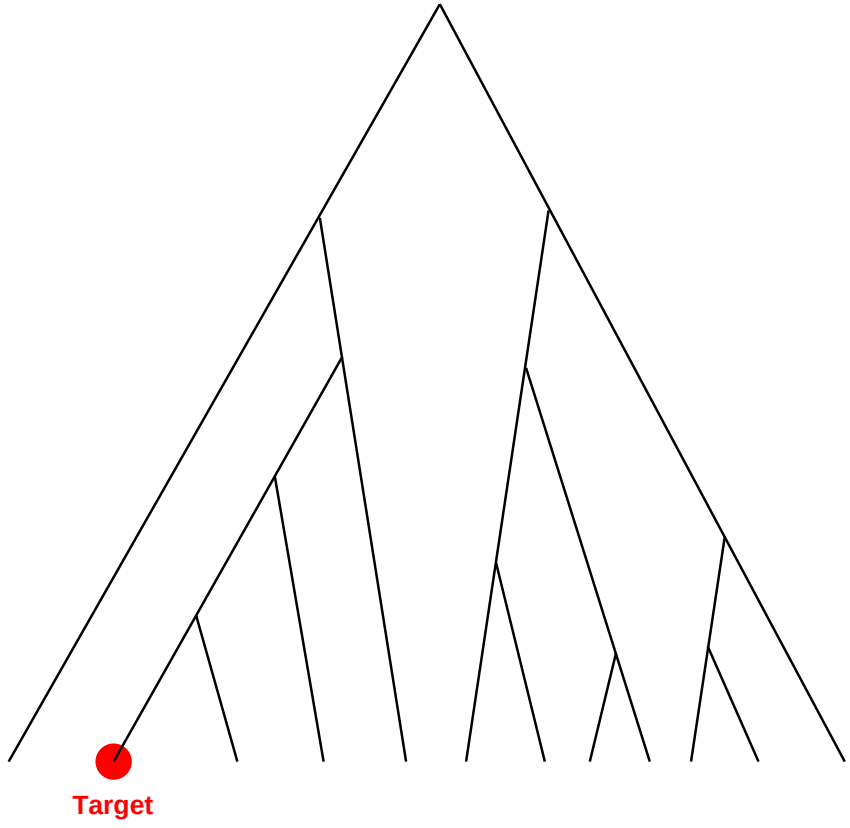
many independent predictions

Schemata for Transfer of Knowledge



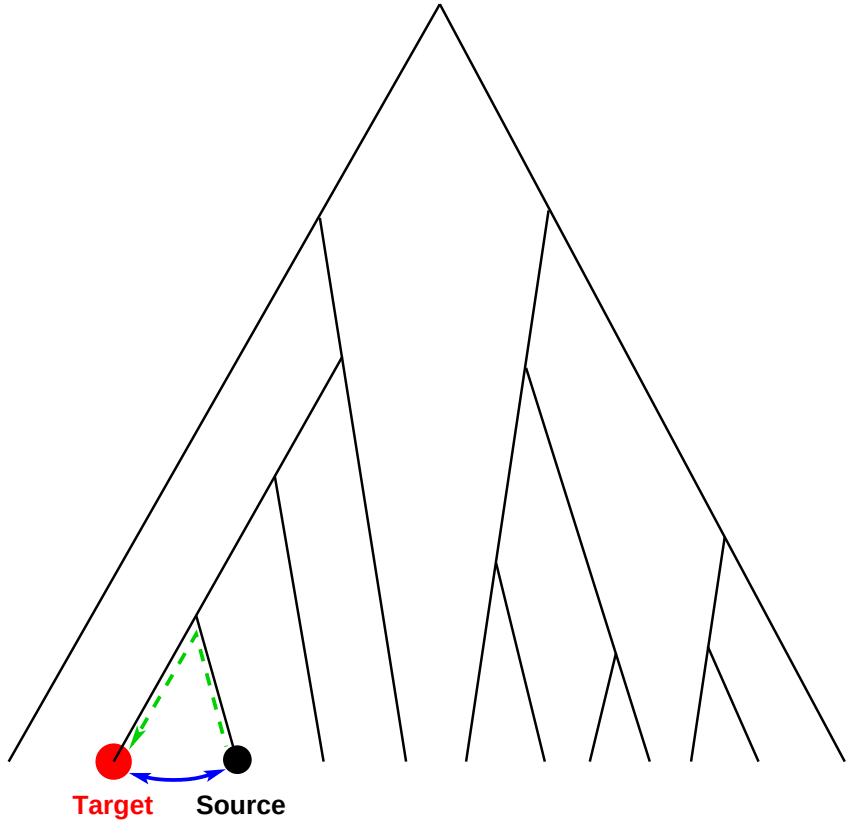
how are the multiple predictions integrated?

The Same Schemata in Context



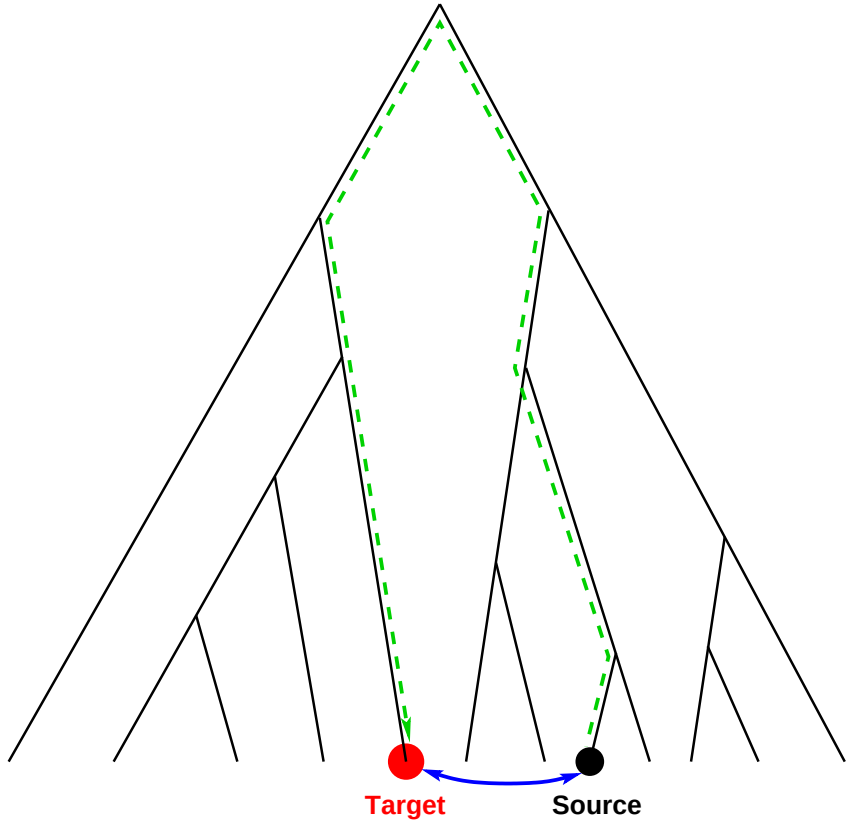
the evolutionary context: a phylogeny

The Same Schemata in Context



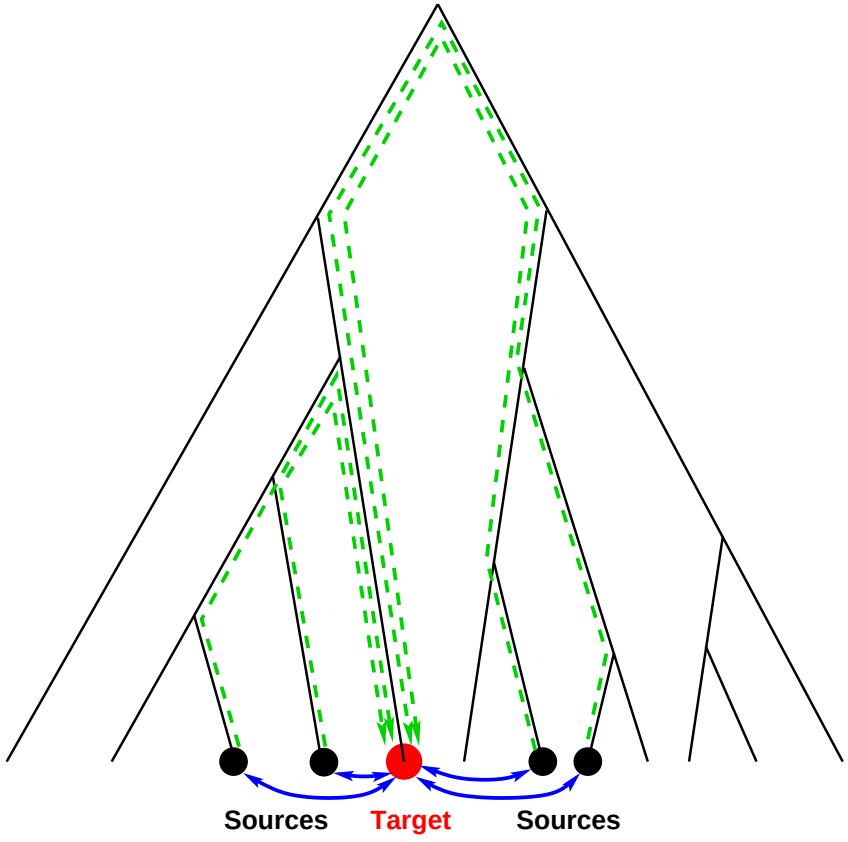
pairwise comparison: OK for closely related objects

The Same Schemata in Context



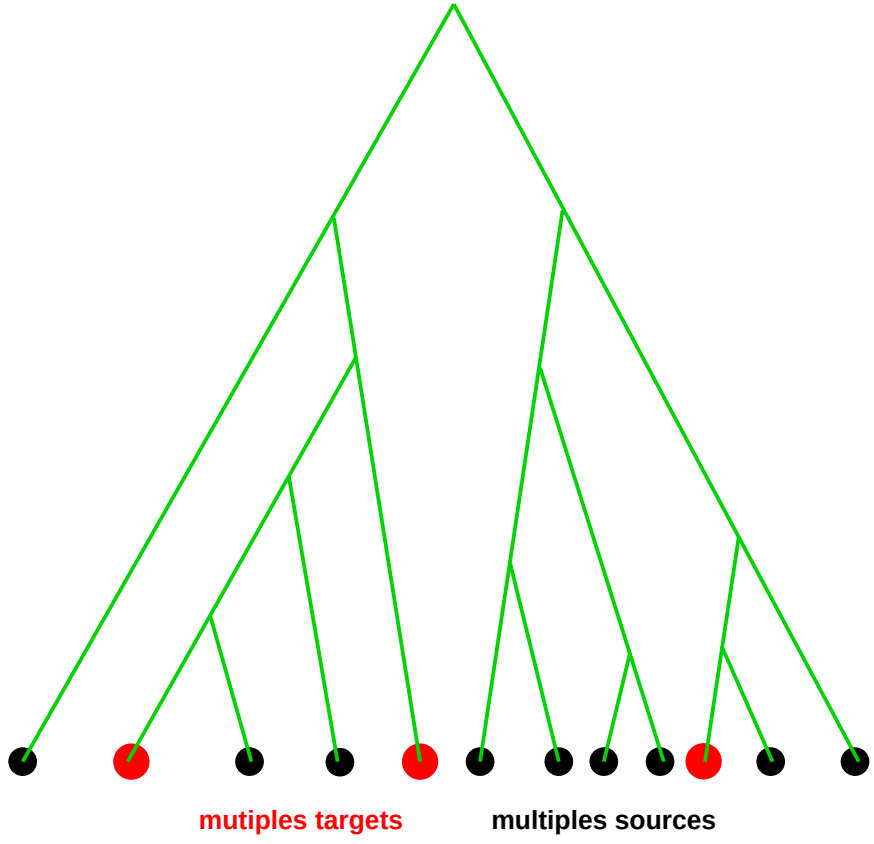
pairwise comparison: problematic for distantly related ones

The Same Schemata in Context



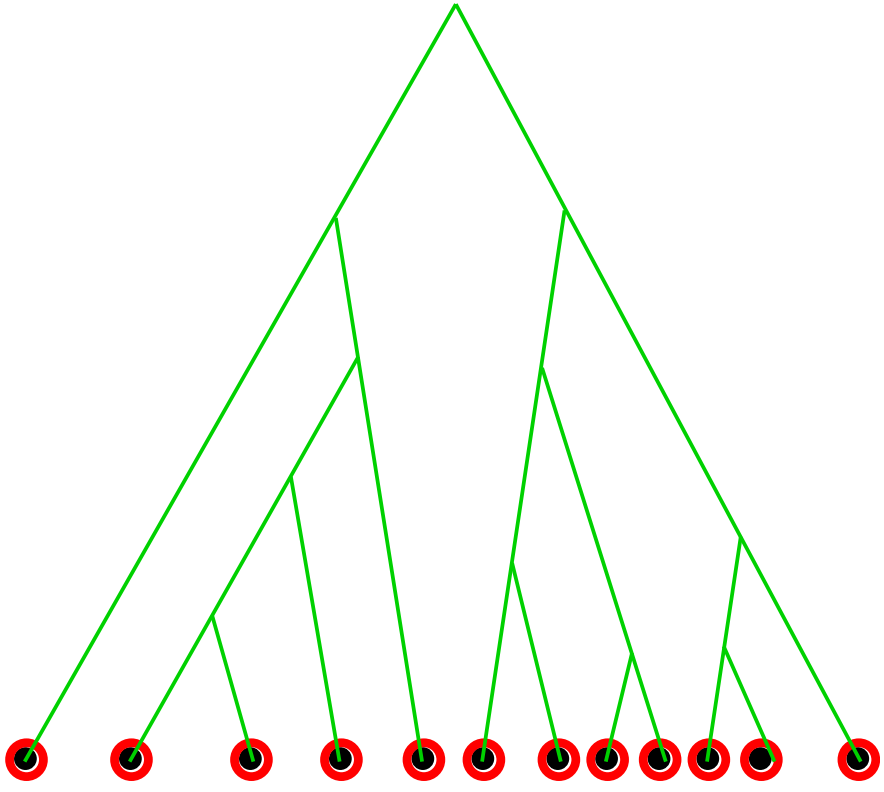
multiple pairwise comparisons: all sorts of problems here!

Use the Phylogeny for Integration



phylogenetic transfer: the full context is in play

Integration Offers Additional Benefits



every object is both source and target

target or source? it's just a matter of confidence

Phylogenetic Transfer of Knowledge (PTK)

$\Rightarrow$ USE EXISTING PHYLOGENETIC KNOWLEDGE TO IMPROVE INFORMATIONAL TRANSFER $\Leftarrow$

- *Use multiple sources of knowledge (well studied taxa) and multiple transfer targets.*
- *Design a graphical inference model that incorporates the known phylogenetic relationships among taxa.*
- *Adapt or enhance standard ML techniques to carry out the inference on the graphical model.*

PTK Attributes

transfer learning

multiple sources of knowledge and multiple transfer targets

refinement

an object can be both source of knowledge and transfer target

integration

very different sources of knowledge in a single model

formal inference model

interplay of inferences defined by the tree,

not by ad hoc consensus or majority

accuracy

large improvement for transfer targets

PTK Issues

- *designing good graphical models is a difficult art*
- *tree must be known and be fairly accurate*
- *may need to reconcile species tree and gene trees*
- *need to infer ancestral data*
- *inference can become very complex*

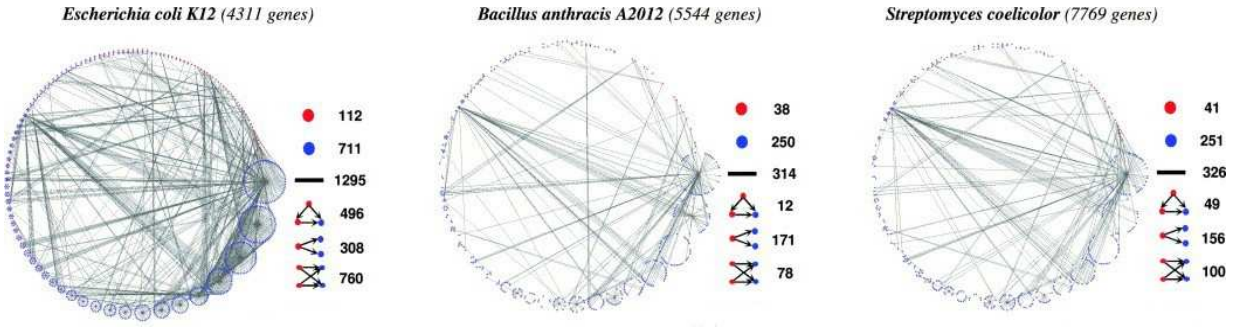
Application to Biological Networks

- *Direct determination (bench work) is slow and expensive.*
- *Computational methods use high-throughput data (microarrays, RNASeq, ChIPSeq, etc.) or pairwise transfer of knowledge to infer the networks.*
- *Error rates are high, esp. false positives.*
- *All (but one) studies up to 2006 focussed on just one network.*

We set out to demonstrate that PTK would significantly improve the accuracy of networks.

Regulatory Networks

Transcriptional regulatory networks represent regulatory connections between genes, gene products, etc.
The simplest are given as directed graphs: an arc from A to B indicates that gene A influences the rate of transcription of gene B.

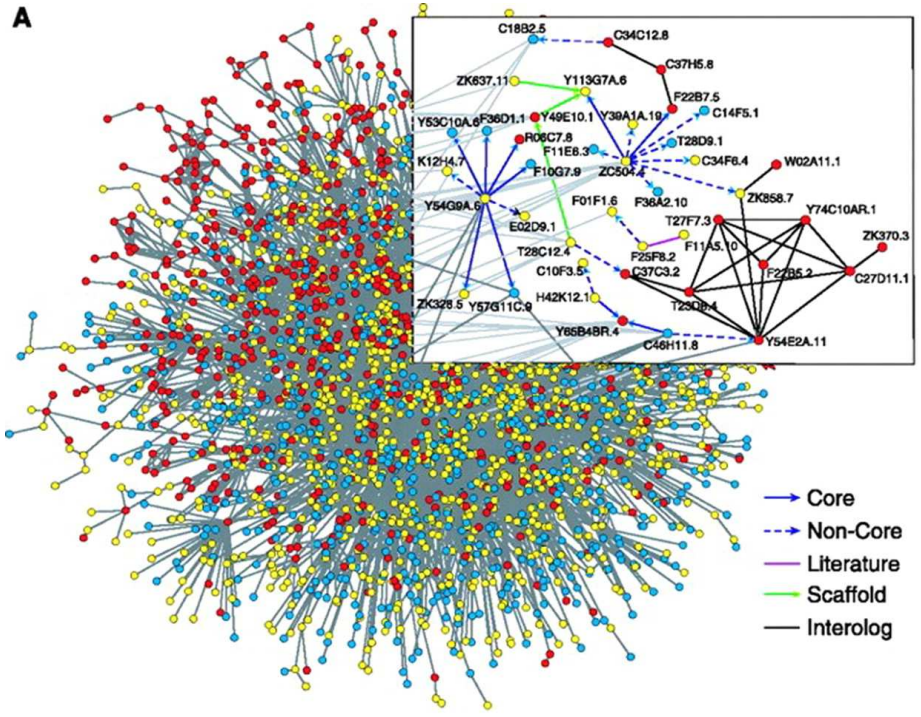


Inference of such networks (from, e.g., microarray or RNAseq data) is notoriously difficult:

- *including all putative regulatory connections gives poor specificity*
- *including only experimentally verified connections gives very poor sensitivity*

Protein-Protein Interaction Networks

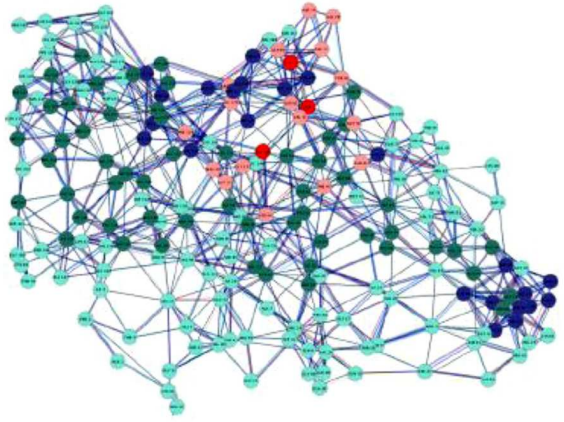
PPI networks are usually undirected graphs, where vertices correspond to proteins and edges to interactions between two proteins.



Most PPI networks published in the literature include every connection that has been observed, along with many that have been inferred, including some inferred from weak evidence, such as frequent co-occurrence with other terms.

Residue Contact Networks

Contact networks abstract the structure of a protein by representing each aminoacid by a node and connecting nodes that correspond to aminoacids close enough to each other to exert significant force upon each other.



The graph model highlights fundamental structural elements and can be used for prediction and comparison.

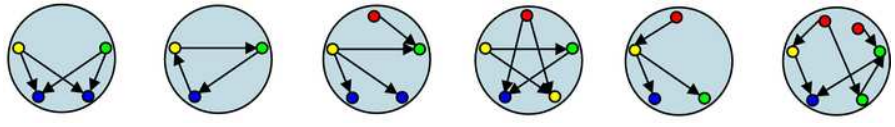
Establishing the contact structure experimentally is expensive and time-consuming, especially for protein complexes. We can use a computational approach based on the evolution of the contact networks.

A Probabilistic Phylogenetic Model

We designed ProPhyC*, a probabilistic phylogenetic model with confidence values, to take advantage of phylogenetic relationships.*

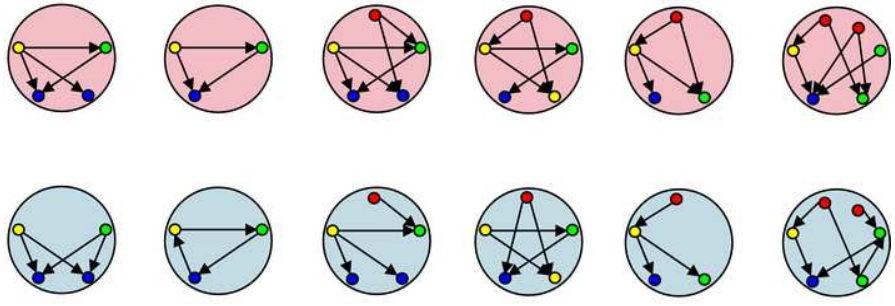
A Probabilistic Phylogenetic Model

We designed ProPhyC*, a probabilistic phylogenetic model with confidence values, to take advantage of phylogenetic relationships.*



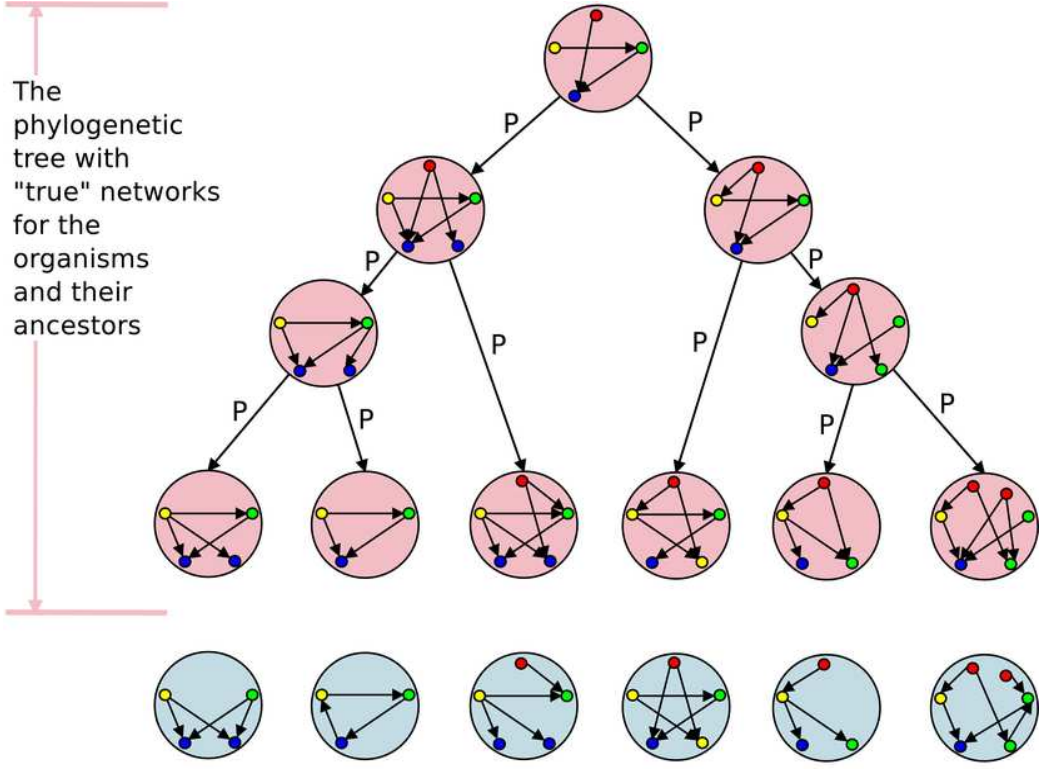
A Probabilistic Phylogenetic Model

We designed ProPhyC*, a probabilistic phylogenetic model with confidence values, to take advantage of phylogenetic relationships.*



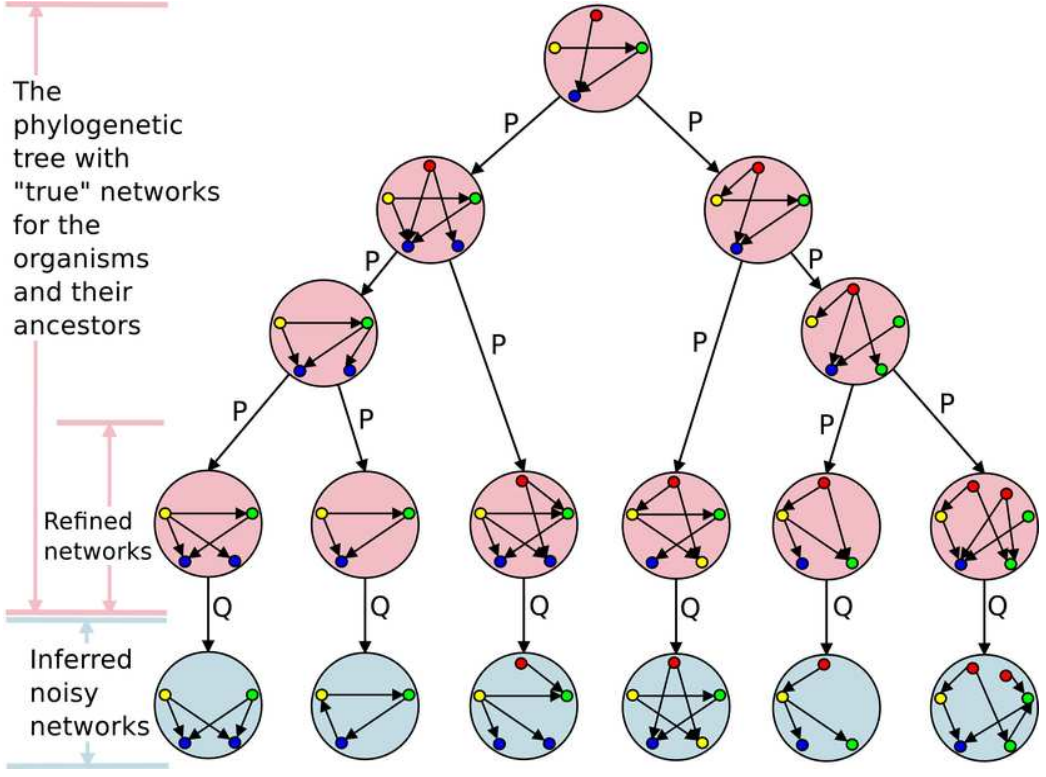
A Probabilistic Phylogenetic Model

We designed ProPhyC*, a probabilistic phylogenetic model with confidence values, to take advantage of phylogenetic relationships.*



A Probabilistic Phylogenetic Model

We designed ProPhyC*, a probabilistic phylogenetic model with confidence values, to take advantage of phylogenetic relationships.*



ProPhyC Attributes

- *The probabilities P reflect evolutionary events, while the probabilities Q reflect confidence in the data and/or probabilities of error in the original inferences.*
- *By assuming independence among the variables, we can use dynamic programming to compute a maximum-likelihood assignment of values to every network in the tree.*
- *Limited dependencies can be modelled with extra variables.*
- *Computations can be biased to improve sensitivity or specificity.*
- *More complex inferences can use EM techniques.*

Inference with ProPhyC

Dynamic programming runs in time linear in the size of the tree, but requires

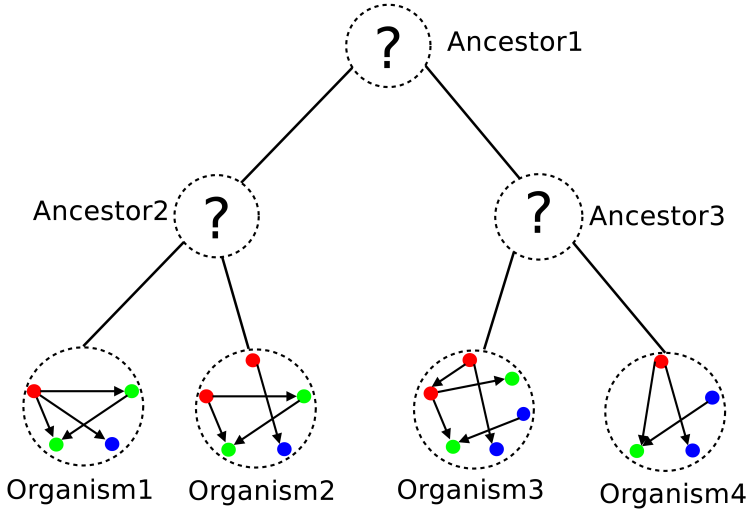
- *independence among events*
- *unified network representation*

Steps:

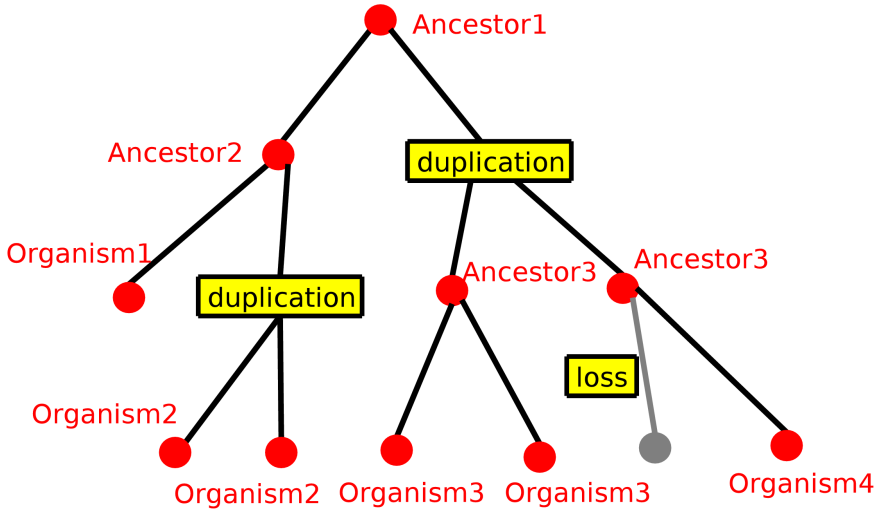
1. *Infer gene duplication and loss for each gene family.*
2. *Devise a unified representation for all networks so as to reduce the problem to single-site inference.*

Step 1: Duploss History

For each family, build the gene tree and reconcile it to the phylogeny.



phylogeny and networks



the phylogeny of the “red” gene family

Step 2: Unified Representation

Use a special character to represent absence of a node (gene).
For simple graph representations, this gives alphabet $\{0, 1, x\}$.

Use probability parameters

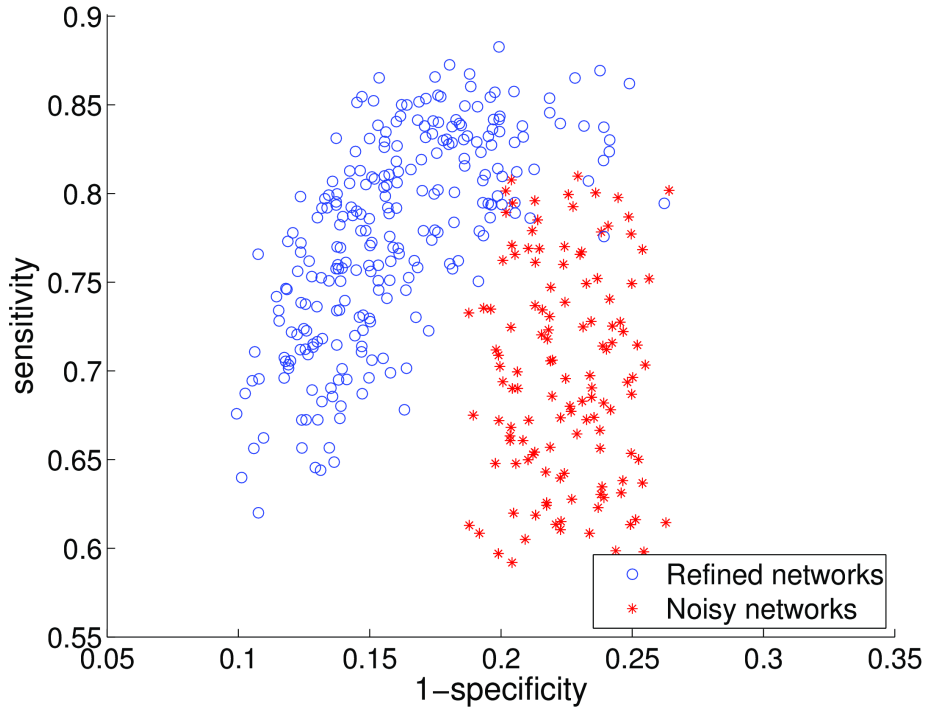
- p_d *and* p_l*: gene getting duplicated or lost*
- p_{00}, p_{01}, p_{10}, *and* p_{11}*: edge gain or loss (or no change)*
- π_0 *and* π_1*: ground probabilities*

Now set

$$P' = \begin{pmatrix} p'_{00} & p'_{01} & p'_{0x} \\ p'_{10} & p'_{11} & p'_{1x} \\ p'_{x0} & p'_{x1} & p'_{xx} \end{pmatrix} = \begin{pmatrix} (1-p_l)\cdot p_{00} & (1-p_l)\cdot p_{01} & p_l \\ (1-p_l)\cdot p_{10} & (1-p_l)\cdot p_{11} & p_l \\ p_d \cdot \pi_0 & p_d \cdot \pi_1 & 1-p_d \end{pmatrix}$$

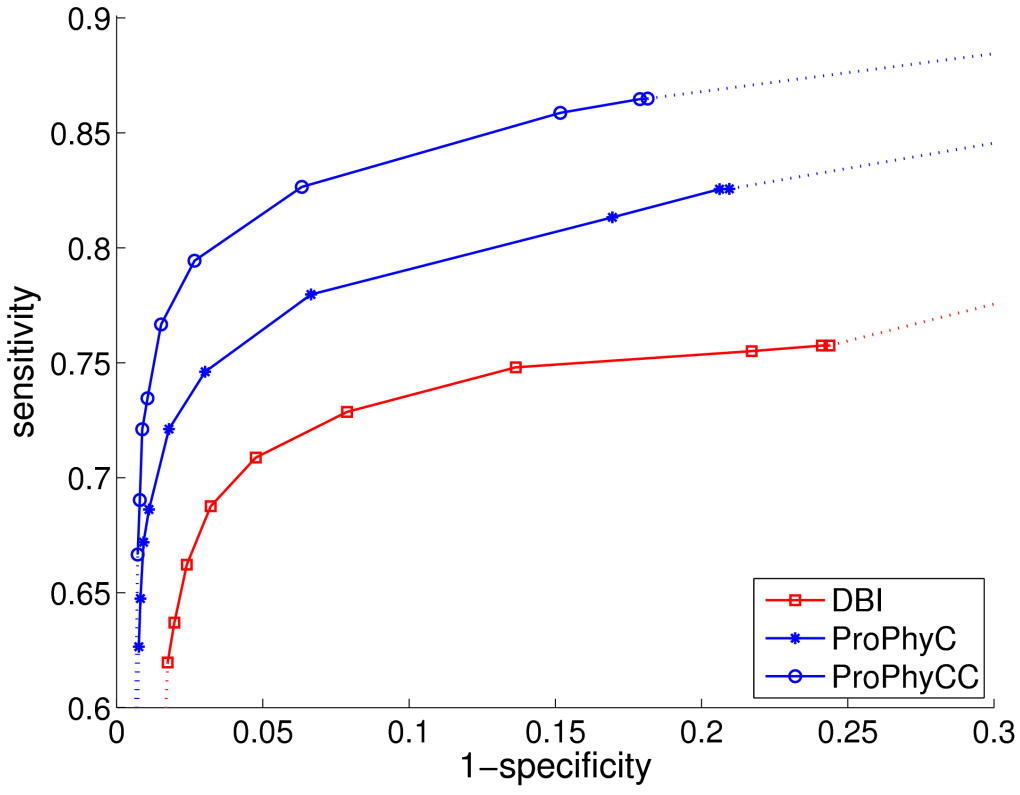
Regulatory Networks: Drosophila

Regulatory modules for 12 species of Drosophila*:*



	sensitivity	*specificity*
improve both	*59.9% → 66.3%*	*80.0% → 86.5%*
focus on sensitivity	*59.5% → 69.2%*	*69.3% → 72.7%*
focus on specificity	*57.7% → 58.5%*	*70.1% → 80.0%*

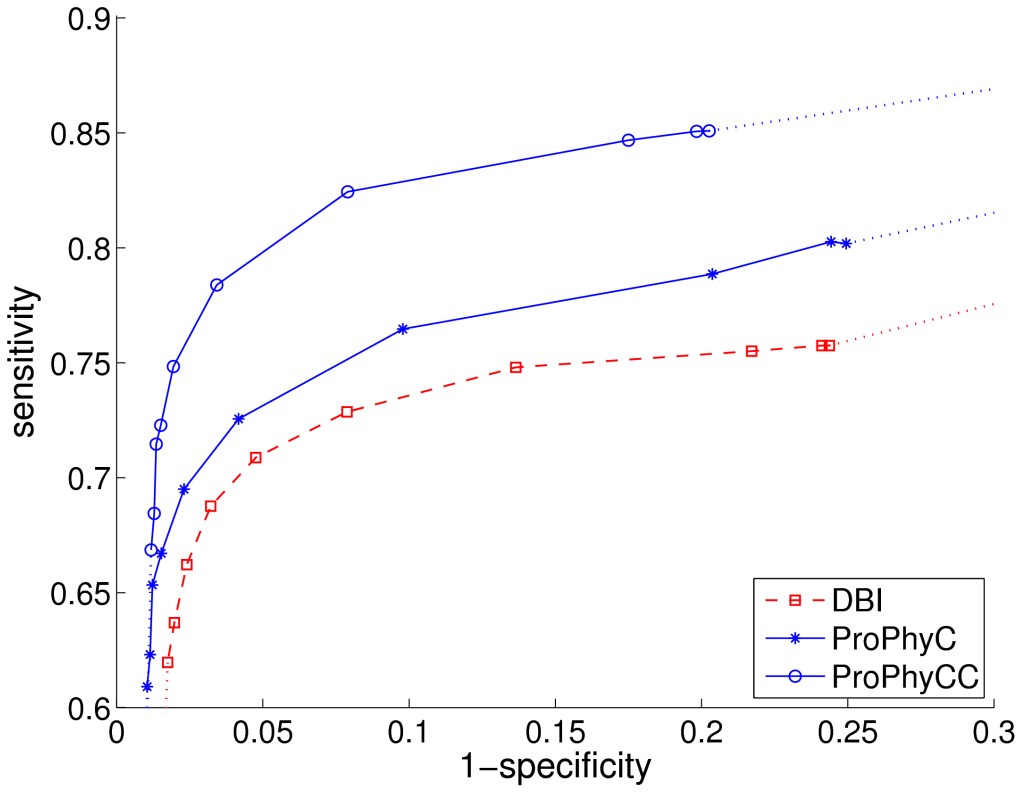
Regulatory Networks: Simulations



True duplication-loss history.

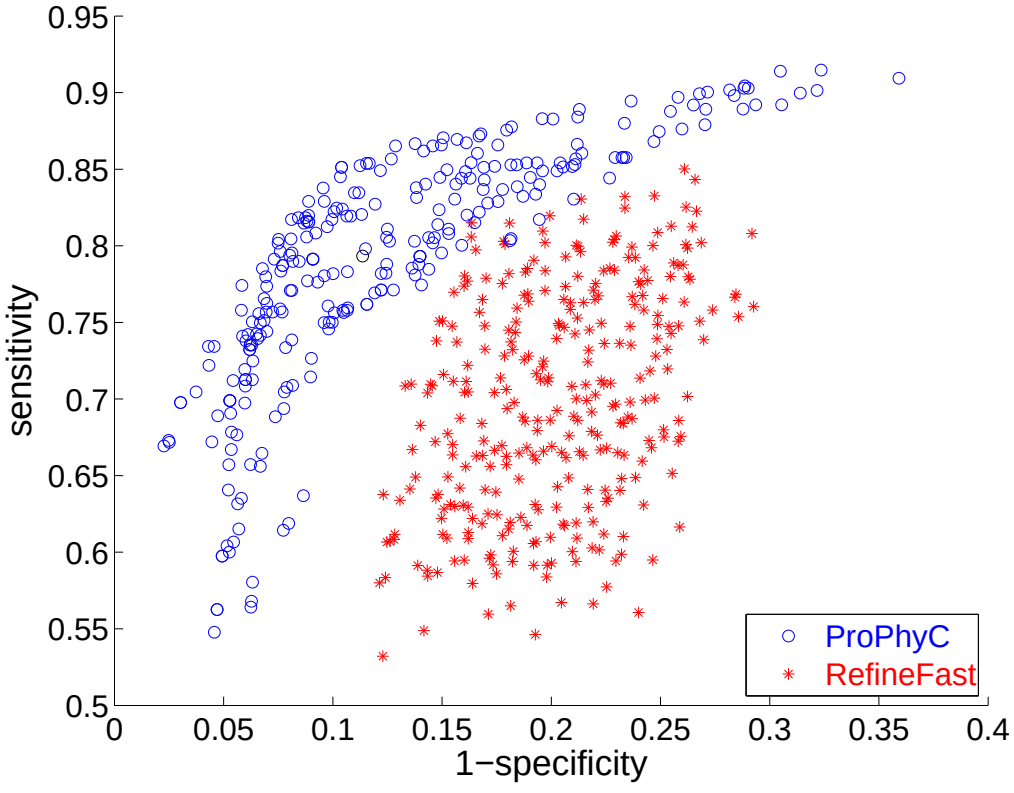
DBI is a standard Bayesian tool for network inference and matches the simulation model. ProPhyC uses a uniform noise model (middle curve) while ProPhyCC uses confidence values derived from the conditional probability tables (top curve).

Regulatory Networks: Simulations



Duplication-loss history of genes is inferred by Notung and so not very accurate.
DBI is a standard Bayesian tool for network inference and matches the simulation model.
ProPhyC uses a uniform noise model (middle curve) while ProPhyCC uses confidence values derived from the conditional probability tables (top curve).

Testing the Graphical Model



The two clouds of points represent results on the same dataset from ProPhyC and from an earlier approach that does not use the noisy observation layer, under a large variety of parameters. ProPhyC clearly dominates.

Confounding Factors

Averaging accounts for some of the improvement. Because it is nearly independent of the labelling of the leaves, we can assess it by randomizing the assignment of networks to leaves.

Averaging accounted for under one-half of the improvement.

Duploss history is hard to infer accurately. Its effect can be evaluated under simulation, comparing to results that used the true history.

Inferring duploss history caused a 5–10% loss in accuracy.

Related Work

- *Bourque and Sankoff (2004) proposed a tree-guided inference procedure to recover regulatory networks from gene-expression levels.*
- *Pinney* et al. *(2007) and Dutkowski and Tiuryn (2009) applied variants of PTK to PPI networks.*
- *Gaudet* et al. *(2011) used a form of PTK to propagate annotations for genes and proteins.*
- *Roy* et al. *(2013) developed the Arboretum software, which uses phylogenetic information to infer regulatory modules from gene-expression data.*
- *Patro and Kingsford (2013) used a form of PTK to infer PPI networks by inferring their evolutionary history.*

Pure Transfer of Knowledge: Simulations

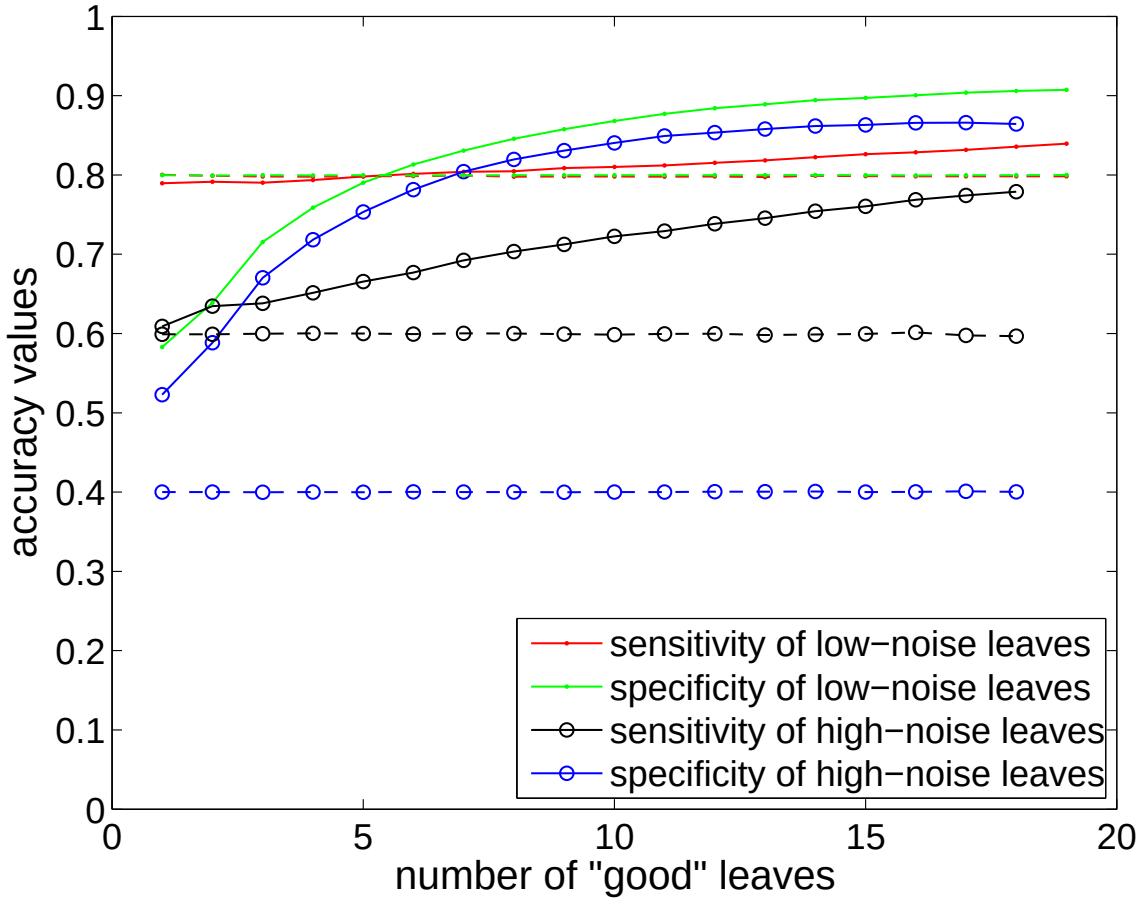
What is a minimum proportion of “good” leaves to “poor” leaves?

We ran extensive simulations on a variety of smaller trees (up to 20 leaves), with various evolutionary models, letting the number of good leaves (high specificity and sensitivity) vary from almost none to almost all, using a matching model for PTK.

PTK sharply increased specificity and sensitivity in the bad leaves for a relatively small decrease of the same measures in the good leaves.

Pure Transfer of Knowledge: Results

Using varying numbers of "good" leaves (80% specificity and sensitivity) and "bad" leaves (40% specificity and 60% sensitivity) on a tree of 19 leaves:



Conclusions

- *Phylogenetic inference—or simply an evolutionary approach—is used for an increasingly diverse collection of problems, but this collection remains a small fraction of what can be addressed productively.*
- *PTK, an "evolutionary" extension of transfer learning, effectively leverages information about evolutionary relationships to structure and balance the transfer of knowledge among extant systems.*
- *Many hard problems remain (such as dealing with interdependencies), requiring advances in theory.*

Details on PTK

Xiuwei Zhang
EBI, Hinxton, and Cambridge U., UK

Zhang, X., Ye, M., and Moret, B.M.E., "Phylogenetic transfer of knowledge for biological networks," PeerJ Preprints, 2, e401v1 (2014).

Zhang, X., and Moret, B.M.E., "Refining regulatory networks through phylogenetic transfer of information," IEEE/ACM Trans. on Comput. Bio. & Bioinf. *9, 4 (2012), 1032–1045.*

Zhang, X., and Moret, B.M.E., "Refining transcriptional regulatory networks using network evolutionary models and gene histories," BMC Algs. in Mol. Bio. *5(1):1 (2010).*

Zhang, X., and Moret, B.M.E., "Boosting the performance of inference algorithms for transcriptional regulatory networks using a phylogenetic approach," Proc. 8th Workshop on Algs. in Bioinf. WABI'08*, in* LNCS **5251***, 245–258 (2008).*