Stochastic Models and Algorithms for Large-scale Comparative Genomics under Complex Evolutionary Scenarios

Kevin J. Liu

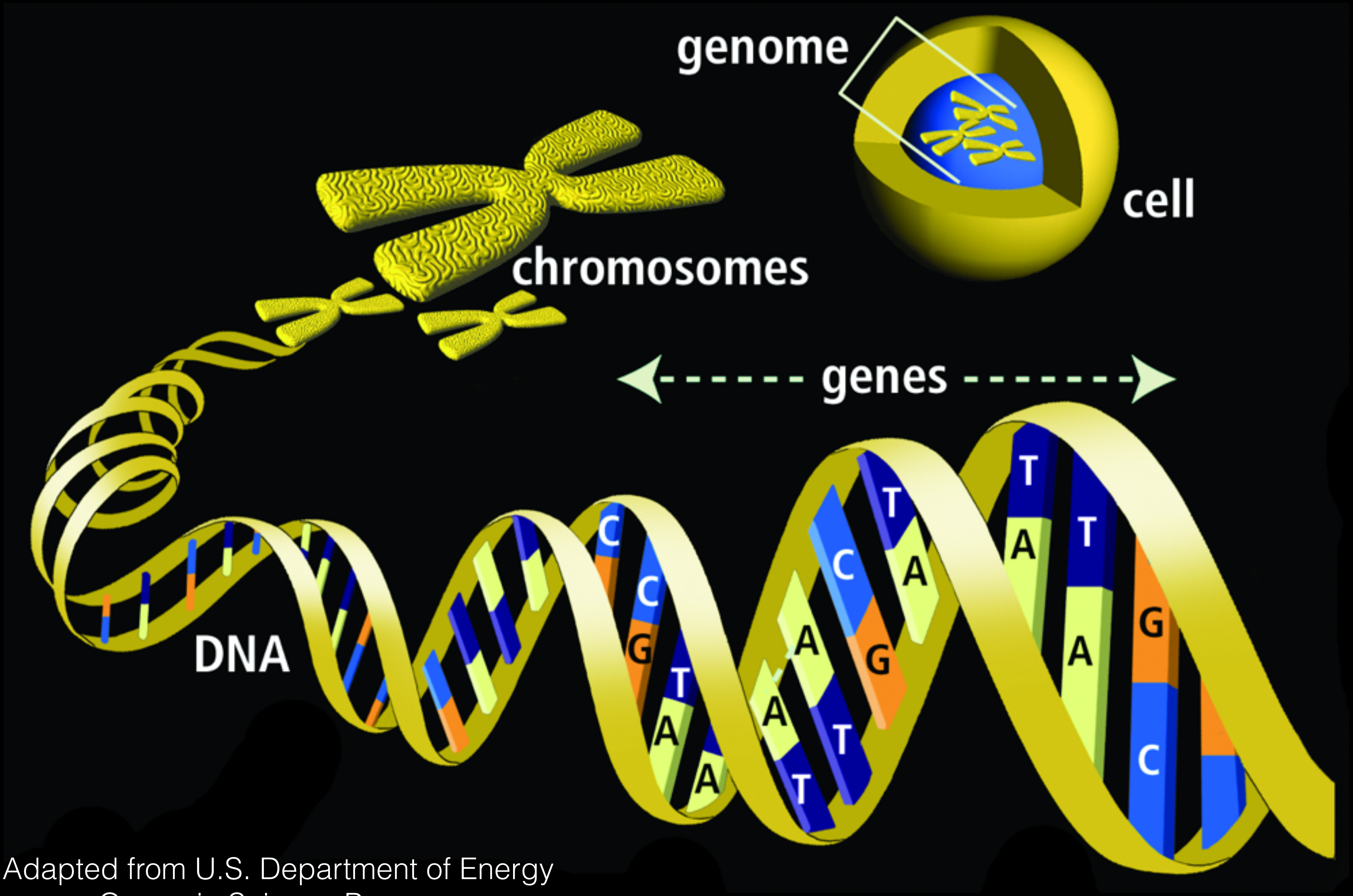
Department of Computer Science

Rice University

Outline

- Comparative genomics: Promises and challenges
- Part I: Fast and accurate alignment and tree estimation on large-scale data sets
- Part II: Modeling and inference under more complex evolutionary scenarios
- Directions for future research and summary

Comparative Genomics

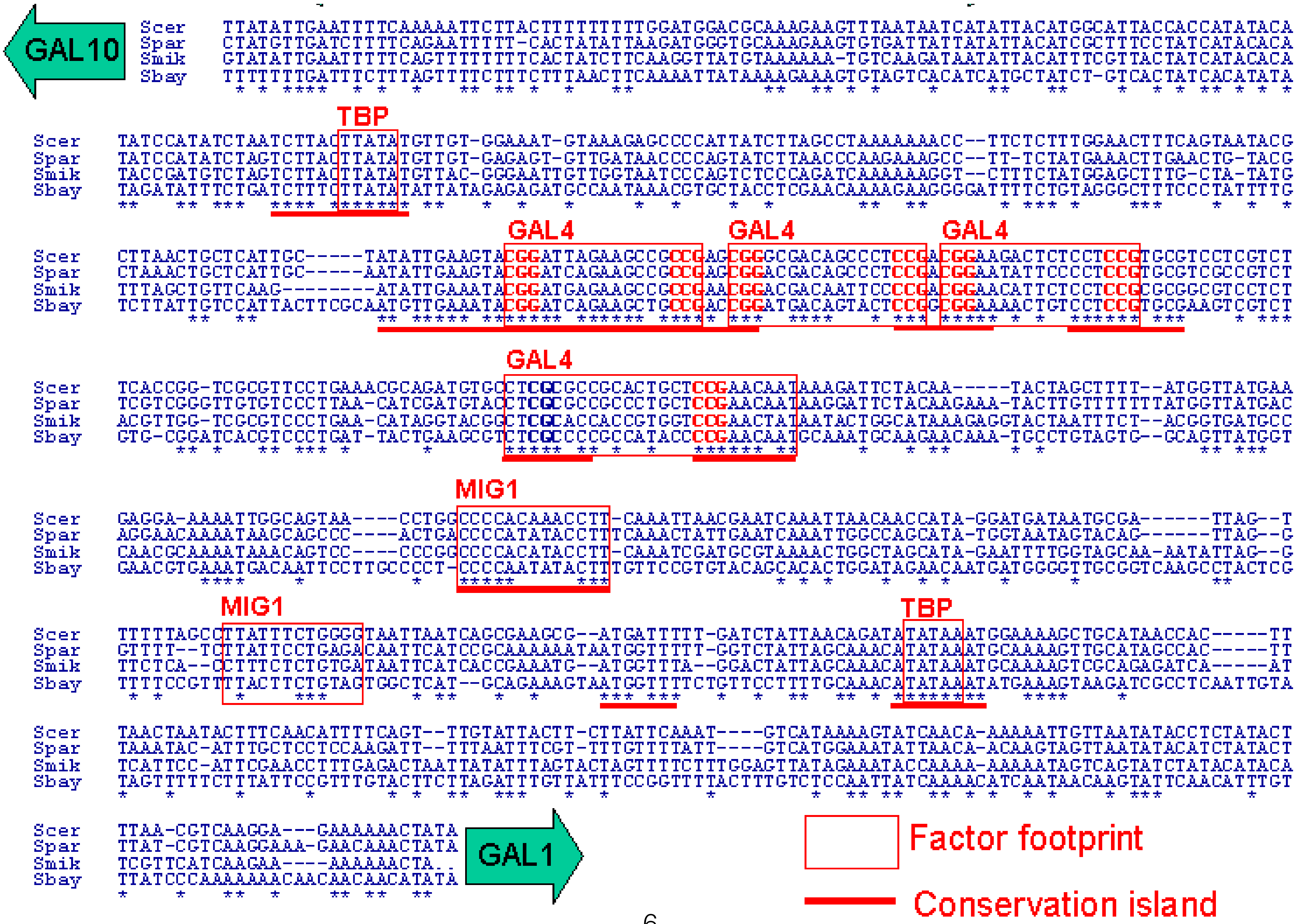


Adapted from U.S. Department of Energy
Genomic Science Program
(genomicscience.energy.gov)

- Advanced genome sequencing technologies are generating data at an unprecedented rate.
- How do we make sense of all of this data?
- One answer: “Nothing in biology makes sense except in the light of evolution.” T. Dobzhansky

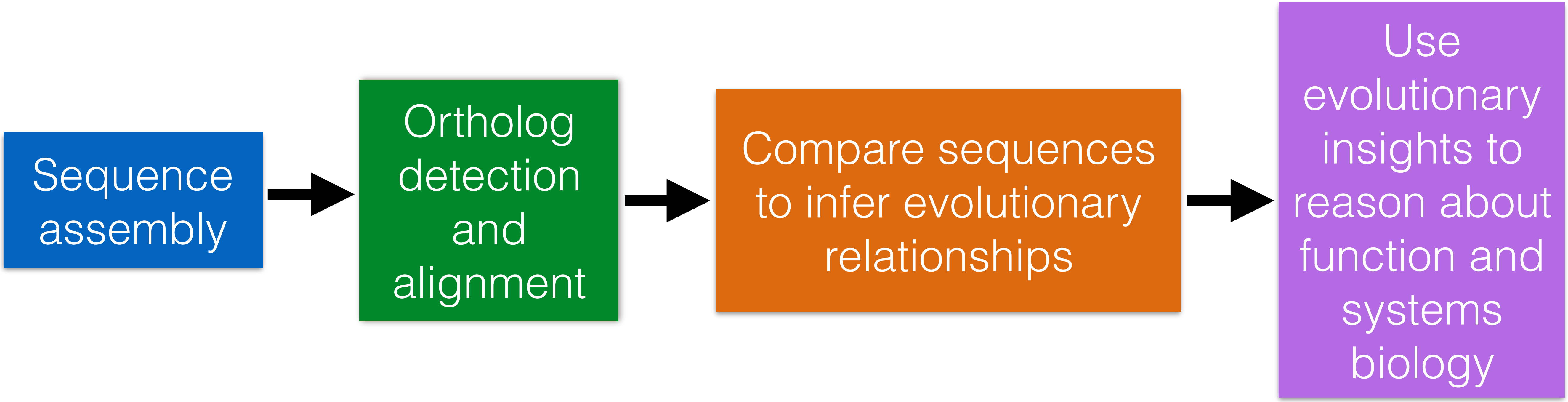
Input and Output of An Example Comparative Genomic Study

(Nature 423 2003)



```
Scer    TTATATTGAATTTTCAAAAATTCTTACTTTTTTTTTGGATGGACGCAAAGAAGTTTAATAATCATATTACATGGCATTACCACCATATACA
Spar    CTATGTTGATCTTTTCAGAATTTTT-CACTATATTAAGATGGTGCAAAGAAGTGTGATTATTATATTACATCGCTTTCCTATCATACACA
Smik    GTATATTGAATTTTTCAGTTTTTTTCACTATCTTCAAGGTTATGTAAAAAA-TGTCAAGATAATATTACATTTCGTTACTATCATACACA
Sbay    TTTTTTTGATTTCTTTAGTTTTCTTTCTTTAACTTCAAAATTATAAAAGAAAGTGTAGTCACATCATGCTATCT-GTCACTATCACATATA
         * * ****  * *  *   ** ** *  *   **           **  ** * *    *    **   **    *  * * ** * * *

Scer    TATCCATATCTAATCTTACTTATATGTTGT-GGAAAT-GTAAAGAGCCCCATTATCTTAGCCTAAAAAAACC--TTCTCTTTGGAACTTTCAGTAATACG
Spar    TATCCATATCTAGTCTTACTTATATGTTGT-GAGAGT-GTTGATAACCCCAGTATCTTAACCCAAGAAAGCC--TT-TCTATGAAACTTGAACTG-TACG
Smik    TACCGATGTCTAGTCTTACTTATATGTTAC-GGGAATTGTTGGTAATCCCAGTCTCCCAGATCAAAAAAGGT--CTTTCTATGGAGCTTTG-CTA-TATG
Sbay    TAGATATTTCTGATCTTTCTTATATATTATAGAGAGATGCCAATAAACGTGCTACCTCGAACAAAAGAAGGGATTTTCTGTAGGGCTTTCCCTATTTTG
        **   ** ***  **** ******* **   *  *   *     *  *    *  *       **  **      * *** *    ***    *  *  *

Scer    CTTAACTGCTCATTGC-----TATATTGAAGTACGGATTAGAAGCCGCCGAGCGGGCGACAGCCCTCCGACGGAAGACTCTCCTCCGTGCGTCCTCGTCT
Spar    CTAAACTGCTCATTGC-----AATATTGAAGTACGGATCAGAAGCCGCCGAGCGGACGACAGCCCTCCGACGGAATATTCCCCTCCGTGCGTCGCCGTCT
Smik    TTTAGCTGTTCAAG--------ATATTGAAATACGGATGAGAAGCCGCCGAACGGACGACAATTCCCCGACGGAACATTCTCCTCCGCGCGGCGTCCTCT
Sbay    TCTTATTGTCCATTACTTCGCAATGTTGAAATACGGATCAGAAGCTGCCGACCGGATGACAGTACTCCGGCGGAAAACTGTCCTCCGTGCGAAGTCGTCT
              **  **          ** ***** ******* ****** ***** ***  ****   * *** ***** * *  ****** ***    * ***

Scer    TCACCGG-TCGCGTTCCTGAAACGCAGATGTGCCTCGCGCCGCACTGCTCCGAACAATAAAGATTCTACAA-----TACTAGCTTTT--ATGGTTATGAA
Spar    TCGTCGGGTTGTGTCCCTTAA-CATCGATGTACCTCGCGCCGCCCTGCTCCGAACAATAAGGATTCTACAAGAAA-TACTTGTTTTTTTATGGTTATGAC
Smik    ACGTTGG-TCGCGTCCCTGAA-CATAGGTACGGCTCGCACCACCGTGGTCCGAACTATAATACTGGCATAAAGAGGTACTAATTTCT--ACGGTGATGCC
Sbay    GTG-CGGATCACGTCCCTGAT-TACTGAAGCGTCTCGCCCCGCCATACCCCGAACAATGCAAATGCAAGAACAAA-TGCCTGTAGTG--GCAGTTATGGT
             ** *   ** *** *      *      ***** ** *  *   ****** **     *   * **     * *             ** ***

Scer    GAGGA-AAAATTGGCAGTAA----CCTGGCCCCACAAACCTT-CAAATTAACGAATCAAATTAACAACCATA-GGATGATAATGCGA------TTAG--T
Spar    AGGAACAAAATAAGCAGCCC----ACTGACCCCATATACCTTTCAAACTATTGAATCAAATTGGCCAGCATA-TGGTAATAGTACAG------TTAG--G
Smik    CAACGCAAAATAAACAGTCC----CCCGGCCCCACATACCTT-CAAATCGATGCGTAAAACTGGCTAGCATA-GAATTTTGGTAGCAA-AATATTAG--G
Sbay    GAACGTGAAATGACAATTCCTTGCCCCT-CCCCAATATACTTTGTTCCGTGTACAGCACACTGGATAGAACAATGATGGGGTTGCGGTCAAGCCTACTCG
               ****    *         *   *****     ***              * * *    *  * *    *     *           **

Scer    TTTTTAGCCTTATTTCTGGGGTAATTAATCAGCGAAGCG--ATGATTTTT-GATCTATTAACAGATATATAAATGGAAAAGCTGCATAACCAC-----TT
Spar    GTTTT--TCTTATTCCTGAGACAATTCATCCGCAAAAAAATAATGGTTTTT-GGTCTATTAGCAAACATATAAATGCAAAAGTTGCATAGCCAC-----TT
Smik    TTCTCA--CCTTTCTCTGTGATAATTCATCACCGAAATG--ATGGTTTA--GGACTATTAGCAAACATATAAATGCAAAAGTCGCAGAGATCA-----AT
Sbay    TTTTCCGTTTTACTTCTGTAGTGGCTCAT--GCAGAAAGTAATGGTTTTCTGTTCCTTTTGCAAACATATAAATATGAAAGTAAGATCGCCTCAATTGTA
         * *      *    ***       * **   *  *     *** ***   *  *  **  ** * ********   ****    *

Scer    TAACTAATACTTTCAACATTTTCAGT--TTGTATTACTT-CTTATTCAAAT----GTCATAAAAGTATCAACA-AAAAATTGTTAATATACCTCTATACT
Spar    TAAATAC-ATTTGCTCCTCCAAGATT--TTTAATTTCGT-TTTGTTTTATT----GTCATGGAAATATTAACA-ACAAGTAGTTAATATACATCTATACT
Smik    TCATTCC-ATTCGAACCTTTGAGACTAATTATATTTAGTACTAGTTTTCTTTGGAGTTATAGAAATACCAAAA-AAAAATAGTCAGTATCTATACATACA
Sbay    TAGTTTTTCTTTATTCCGTTTGTACTTCTTAGATTTGTTATTTCCGGTTTTACTTTGTCTCCAATTATCAAAACATCAATAACAAGTATTCAACATTTGT
        *   *     *     *      * *  **  ***   *  *        *        *  ** **  ** * *  * *    * ***       *

Scer    TTAA-CGTCAAGGA---GAAAAAACTATA
Spar    TTAT-CGTCAAGGAAA-GAACAAACTATA
Smik    TCGTTCATCAAGAA----AAAAAACTA..
Sbay    TTATCCCAAAAAAACAACAACAACATATA
        *    *   **  *    ** **  **
```

A Comparative Genomics Pipeline

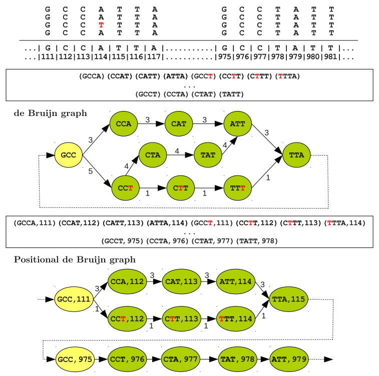


Sequence assembly

Ortholog detection and alignment

Compare sequences to infer evolutionary relationships

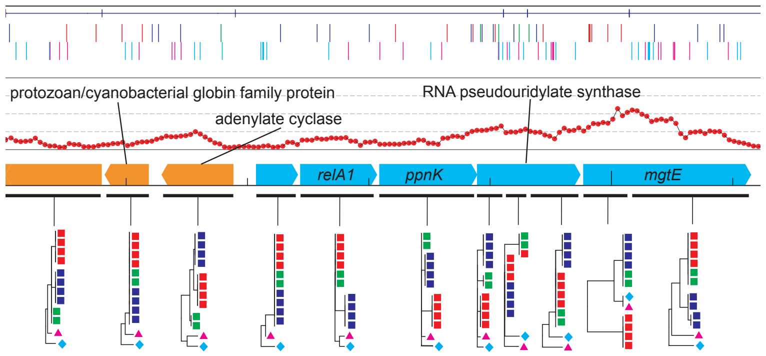
Use evolutionary insights to reason about function and systems biology



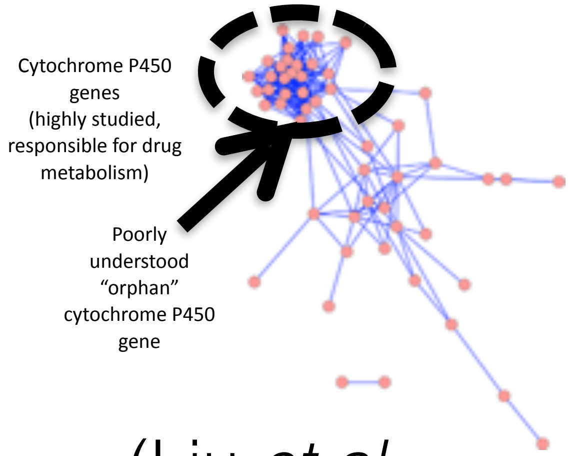
(Bioinformatics 28, 2012)

Ot QDIKLSDYRG--KYVVLFFYPL
Se QTIKLSNYRG--KYVVLFFYPL
At IKVKLSDYNGK-KYVILFFYPL
Hs KEVKLSDYKG--KYVVLFFYPL
Mm KEIKLSDYRG--KYVVLFFYPL
Ce VDVSLSDYKG--KYVVLFFYPL
Se DEVSLDKYKG--KYVVLAFIPL
Dm KDIKLSDYKG--KYLVLFFYPL
Nc -PIDFHEFIGD-NWVILFSHPE
Has TRLGLTDALADNRAVVLFFYPF

(Nature 485, 2012)



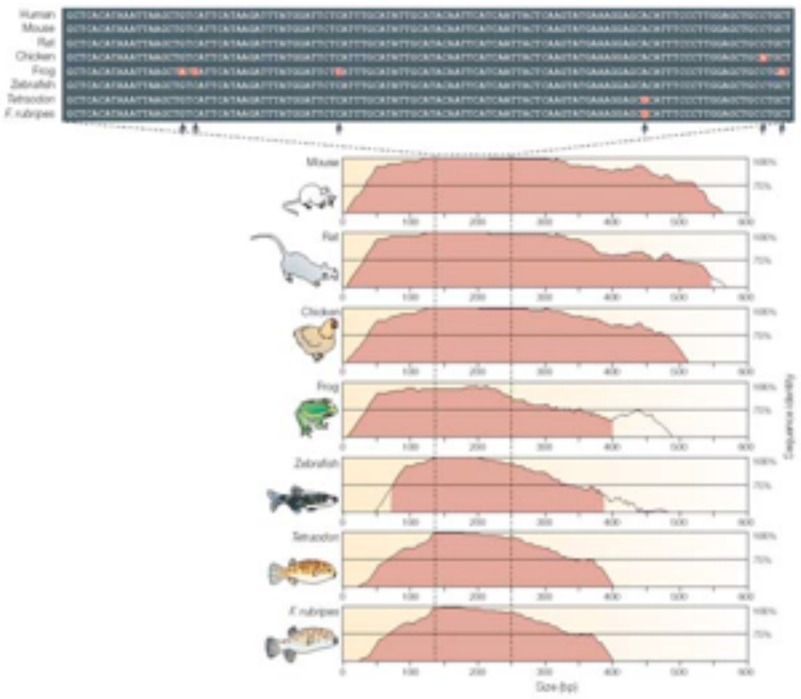
(MBE 29, 2013)


Cytochrome P450 genes (highly studied, responsible for drug metabolism)
Poorly understood “orphan” cytochrome P450 gene

(Liu *et al.*, submitted.)

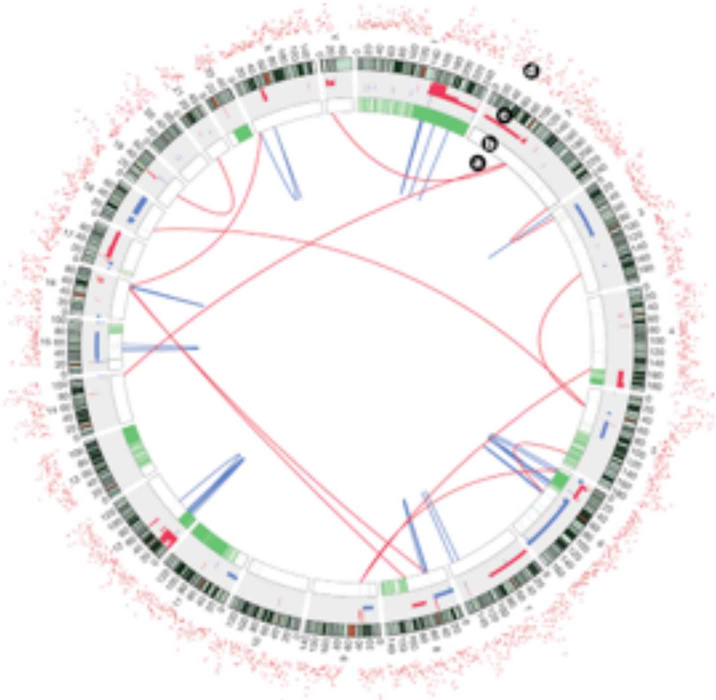
Applications

Detecting regulatory elements



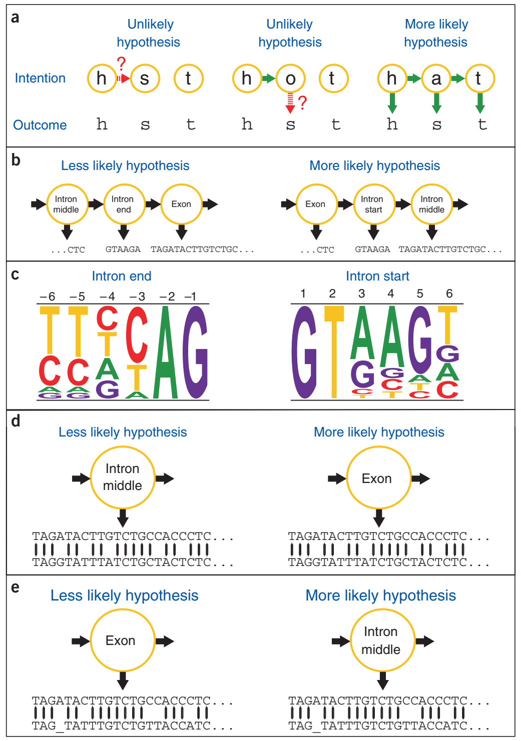
(Nature Reviews Genetics 5, 2004)

Detecting cancer mutations



(Nature 465, 2010)

Gene finding



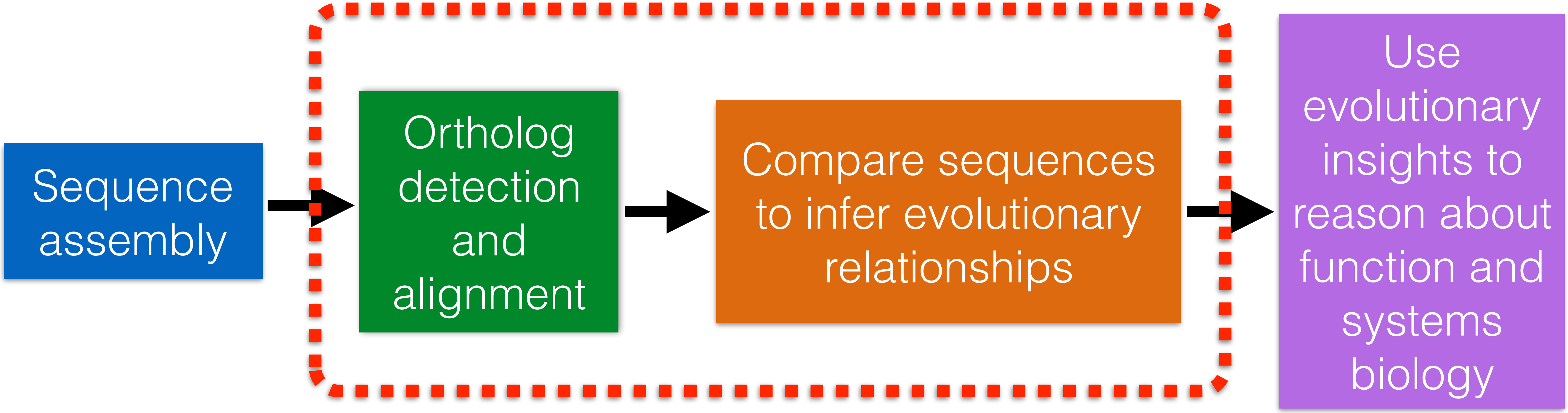
(Nature Biotechnology 25, 2007)

And many, many more …

Three Major Challenges

- **Computational challenge:** accurate and scalable algorithms and tools for large-scale analyses
- **Statistical challenge:** realistic yet tractable models of genome evolution
- **Biological challenge:** co-occurrence of multiple complex evolutionary events

My Contributions


Sequence assembly
Ortholog detection and alignment
Compare sequences to infer evolutionary relationships
Use evolutionary insights to reason about function and systems biology

Graduate work:

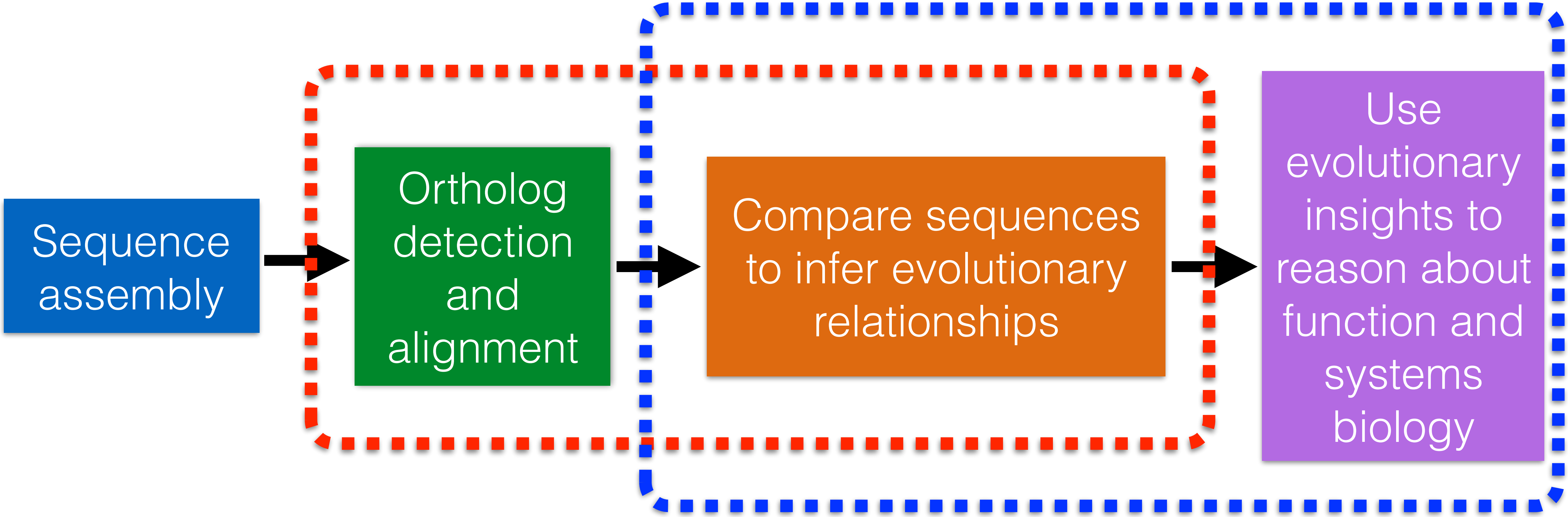
SATé,

SATé-II,

DACTAL,

etc.

My Contributions


Sequence assembly
Ortholog detection and alignment
Compare sequences to infer evolutionary relationships
Use evolutionary insights to reason about function and systems biology

Graduate work:

SATé,

SATé-II,

DACTAL,

etc.

Postgraduate work:

PhyloNet-HMM,

etc.

Part I: Fast and Accurate Alignment and Tree Estimation on Large-Scale Datasets

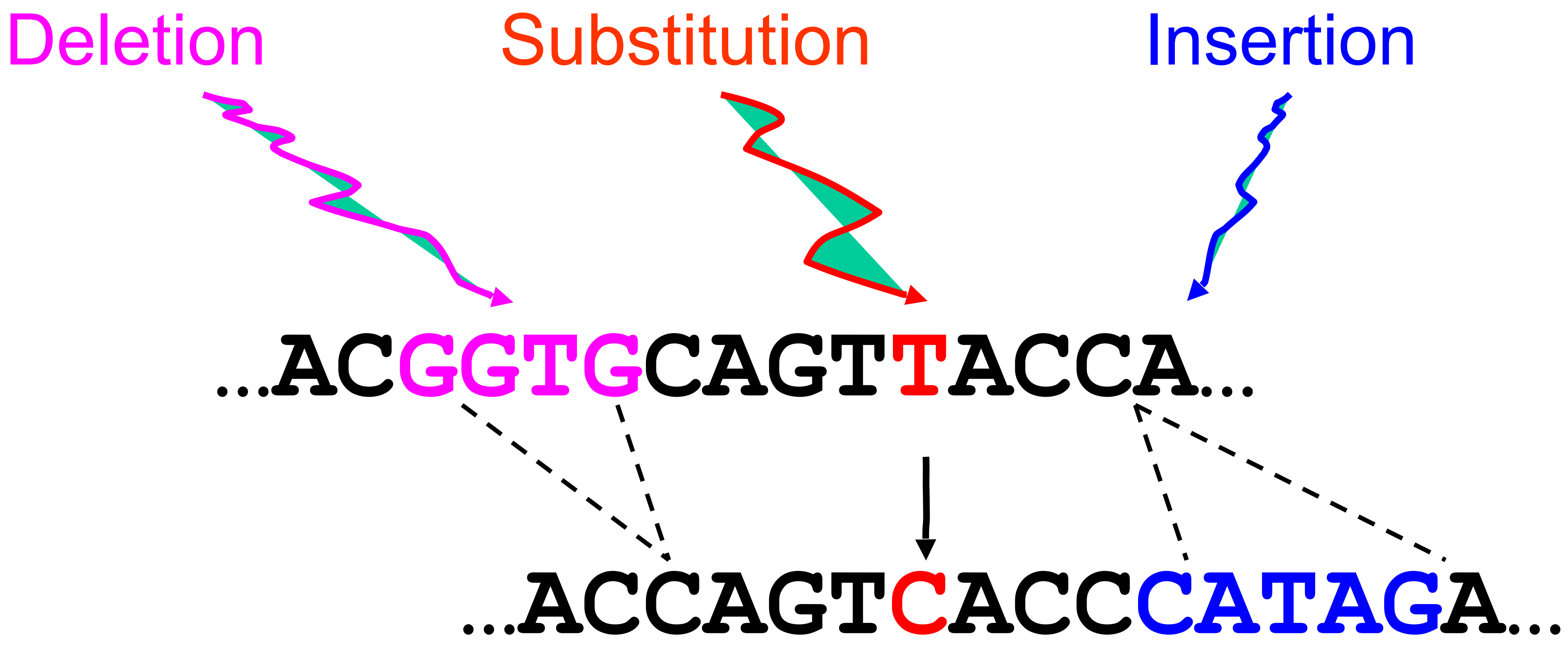
SATé: Simultaneous Alignment and Tree estimation (Liu *et al.* Science 2009)

- Standard methods for alignment and tree estimation have unacceptably high error and/or cannot analyze large datasets

- SATé has equal or typically better accuracy than all existing methods on datasets with up to thousands of sequences

- 24 hour analyses using standard desktop computer

- SATé-II (Liu *et al.* Systematic Biology 2012) is more accurate and faster than SATé on datasets with up to tens of thousands of taxa



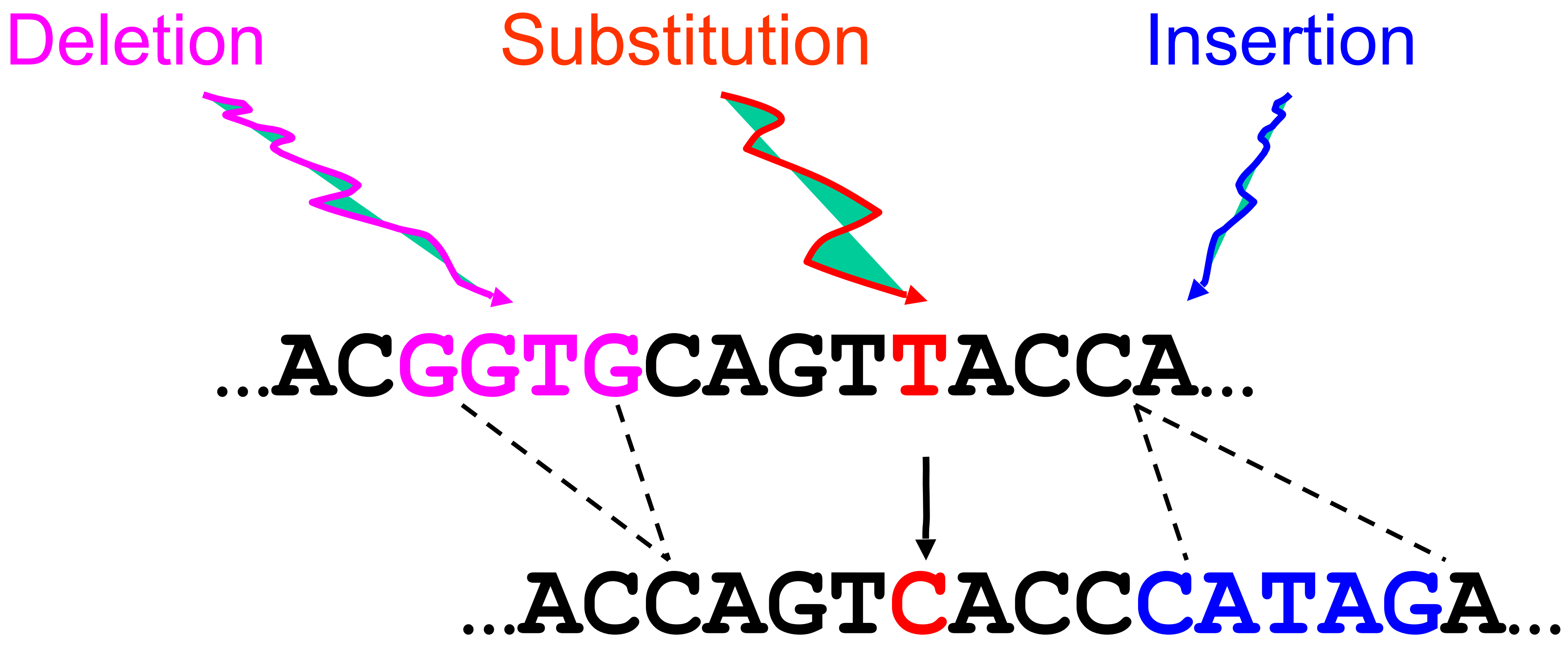
Deletion

Substitution

Insertion

…ACGGTGCAGTTACCA…

…ACCAGTCACCCATAGA…



Deletion

Substitution

Insertion

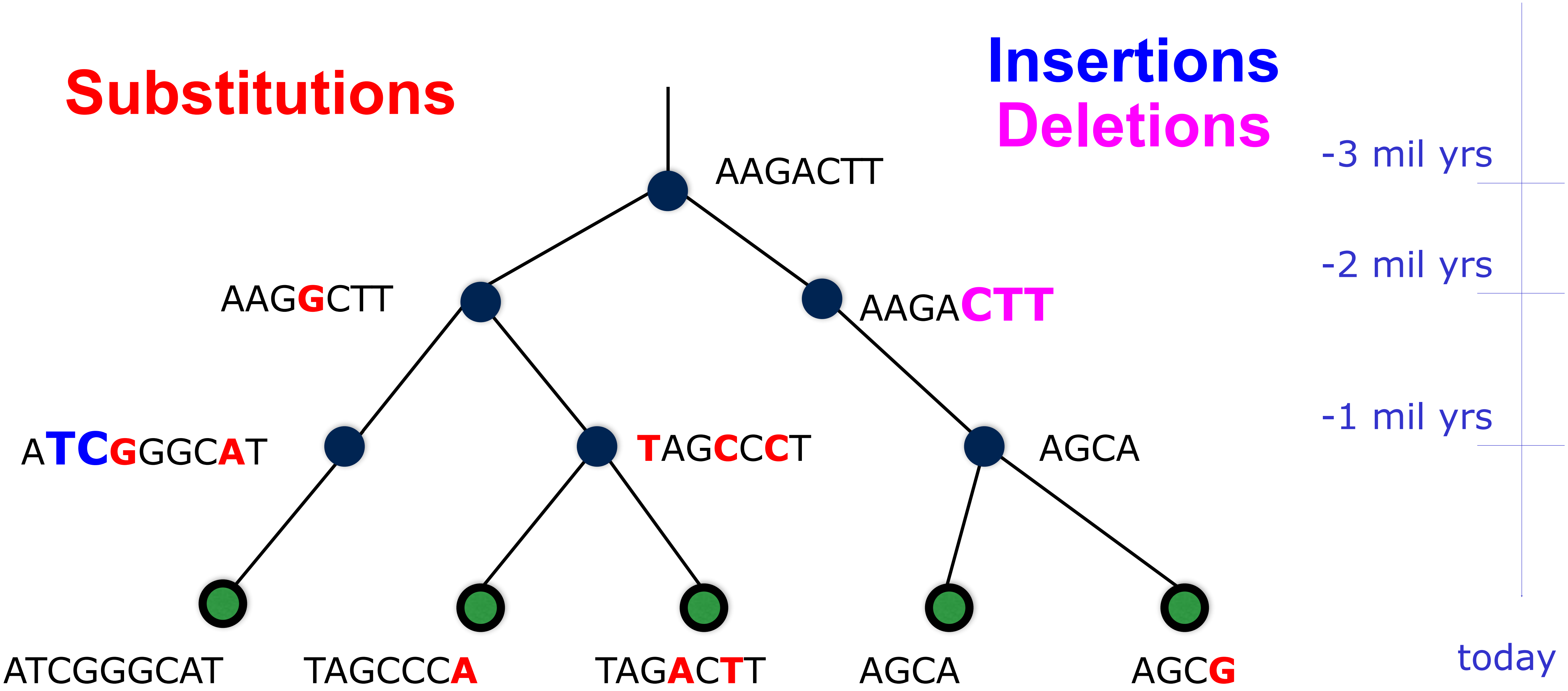
...ACGGTGCAGTTACCA...

...ACCAGTCACCCATAGA...

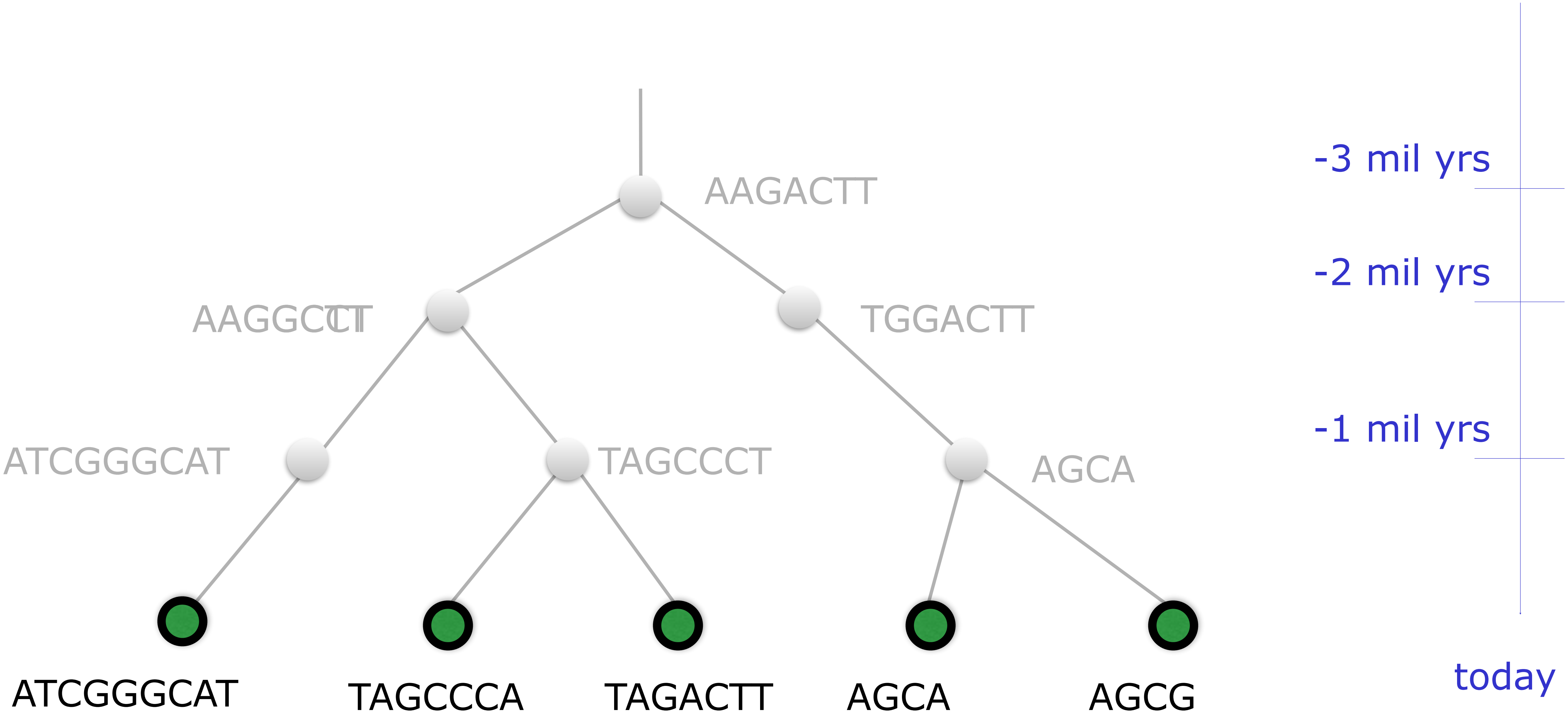
The **true alignment** is:

```
...ACGGTGCAGTTACC-----A...
...AC----CAGTCACCCATAGA...
```

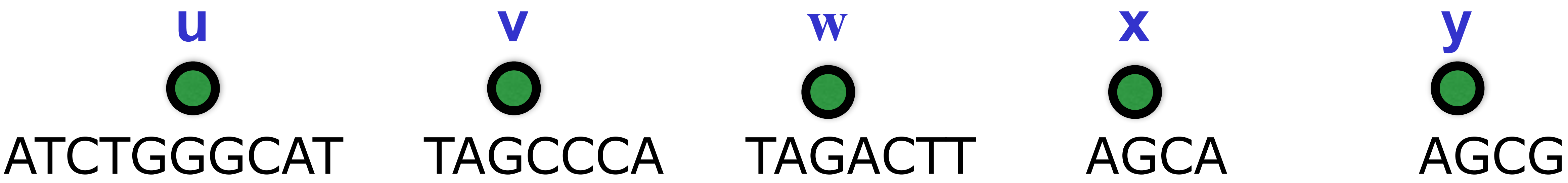
DNA Sequence Evolution (Example)

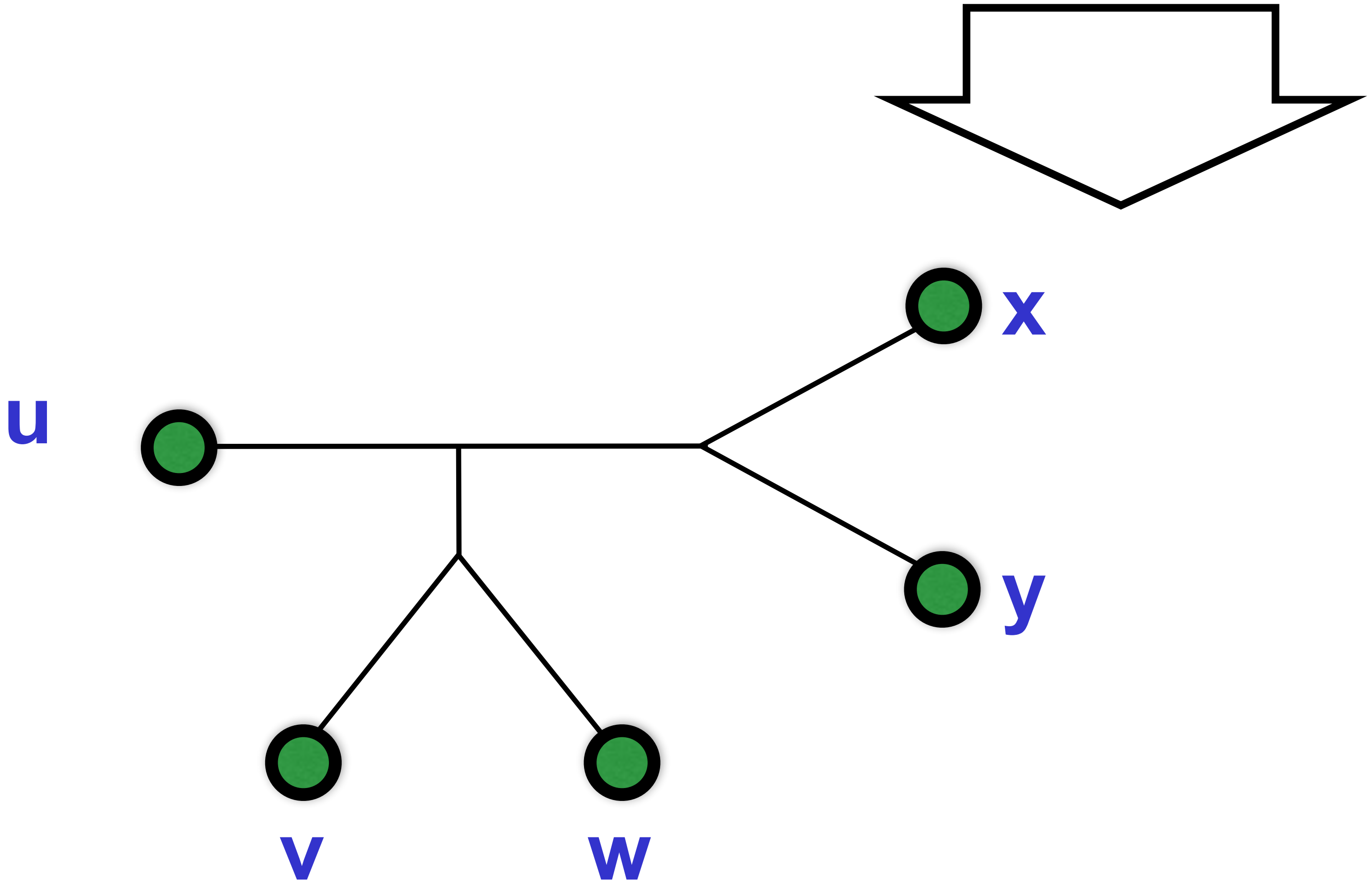


DNA Sequence Evolution (Example)



Tree and Alignment Estimation Problem (Example)





```
u = ATCTGGCAT
v = T--AGCCCA
w = T--AGACTT
x = AGCA-----
y = AGCG-----
```

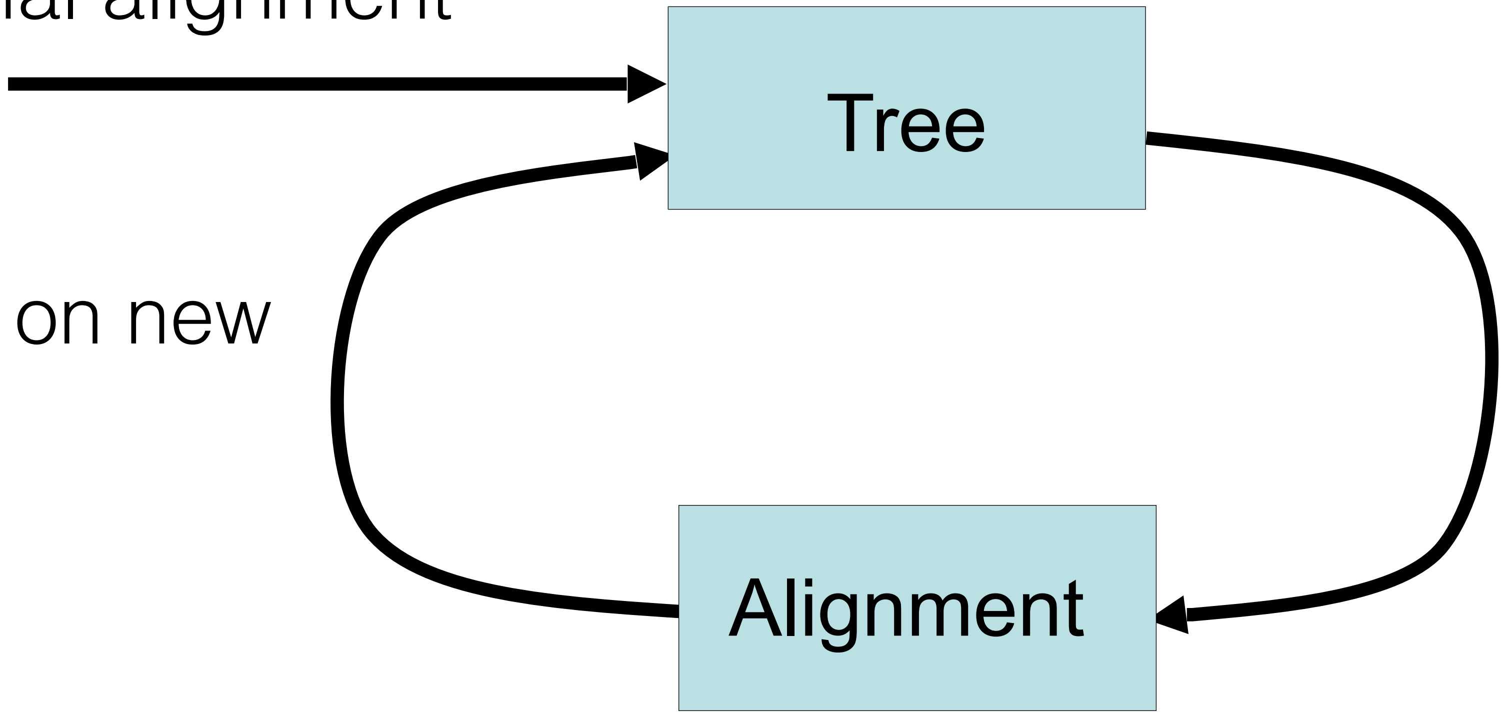
Many Trees and Many Alignments

- Number of trees |*T*| grows exponentially in n, the number of leaves:

$$|T| = (2n-5)!!$$

- The number of alignments |*A*| also grows exponentially in *n* and the length of the longest unaligned sequence.
- All of the common and useful optimization problems are NP-hard.

SATé Algorithm



Obtain initial alignment and tree

Tree

Estimate tree on new alignment

Alignment

Insight:

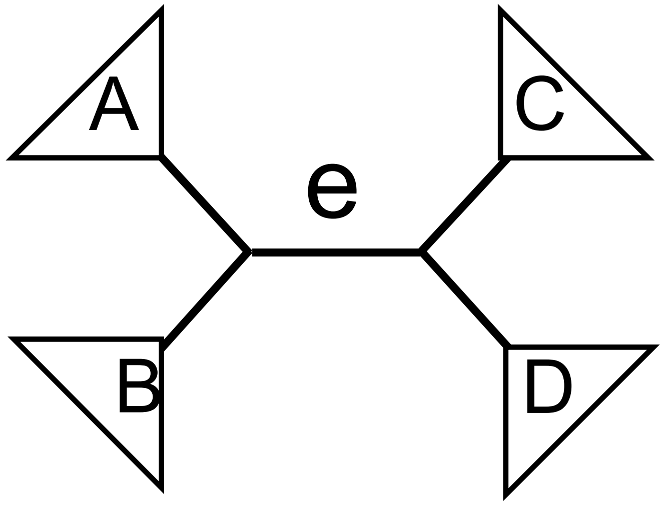
Use tree to perform divide-and-conquer alignment

Insight: iterate - use a moderately accurate tree to obtain a more accurate tree

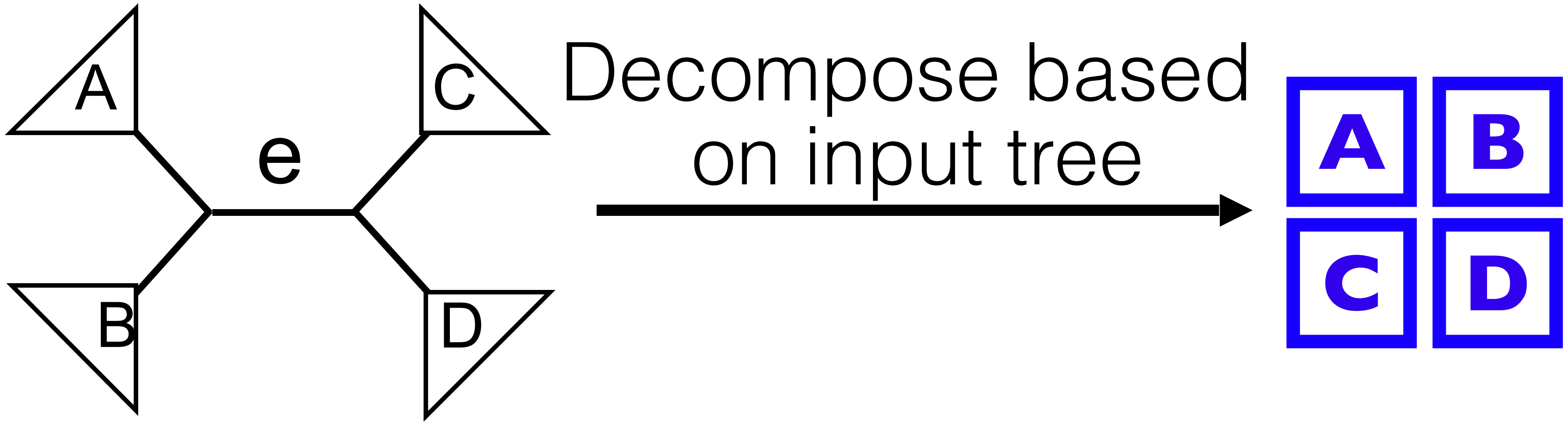
If new alignment/tree pair has worse likelihood, realign using a different decomposition

Repeat until convergence under the maximum likelihood optimization criterion

SATé iteration
(Actual decomposition size is configurable)

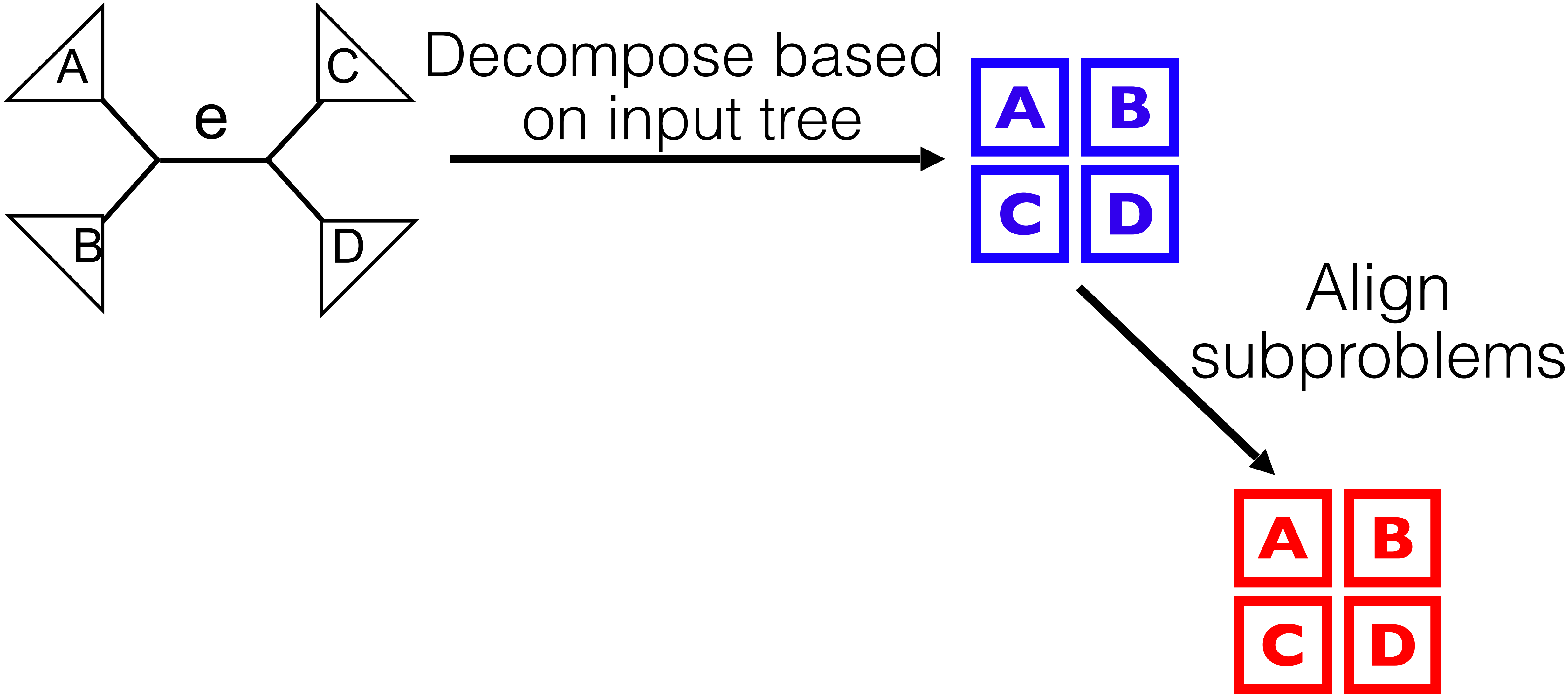


SATé iteration
(Actual decomposition size is configurable)



Decompose based on input tree

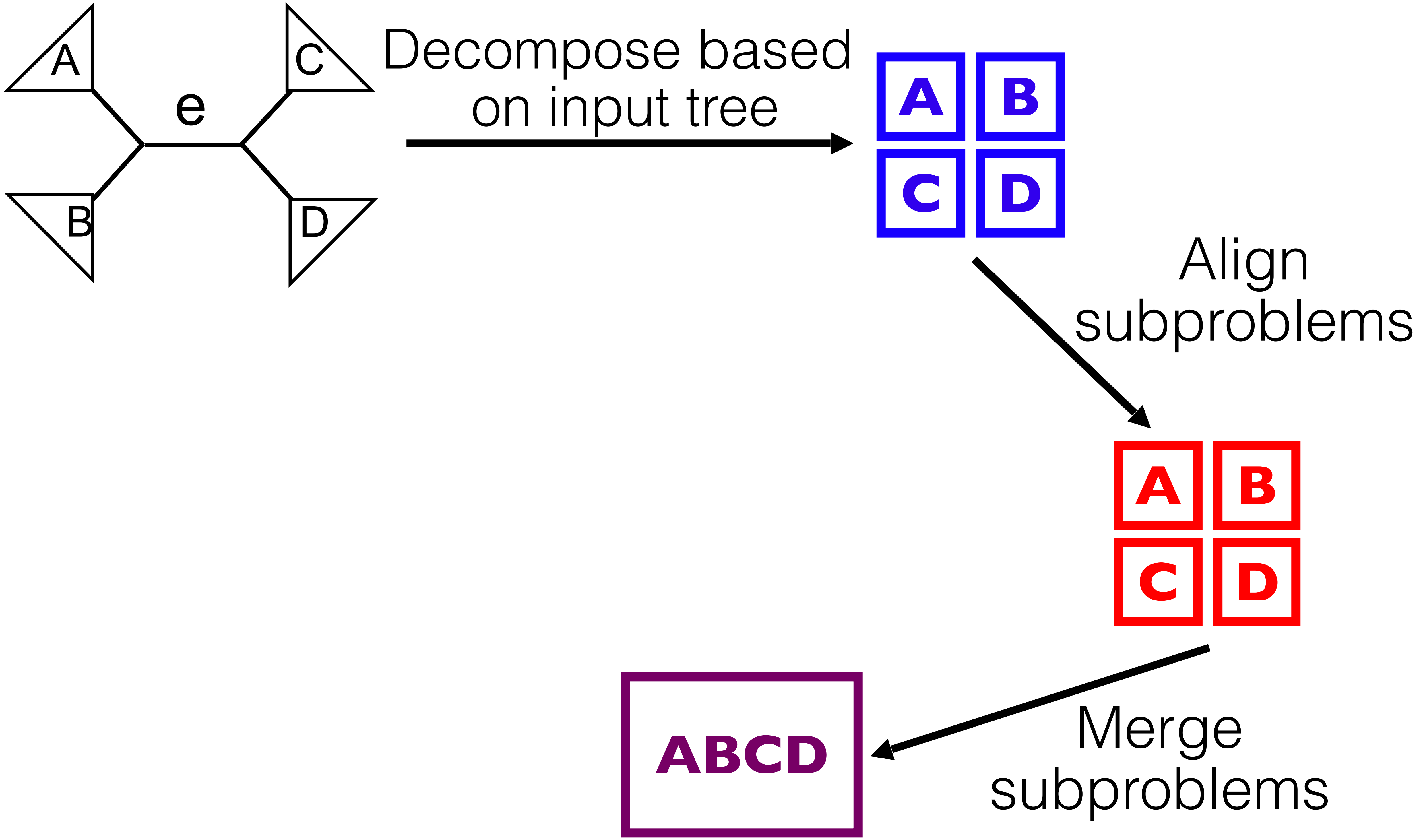
SATé iteration
(Actual decomposition size is configurable)



Decompose based on input tree

Align subproblems

SATé iteration
(Actual decomposition size is configurable)

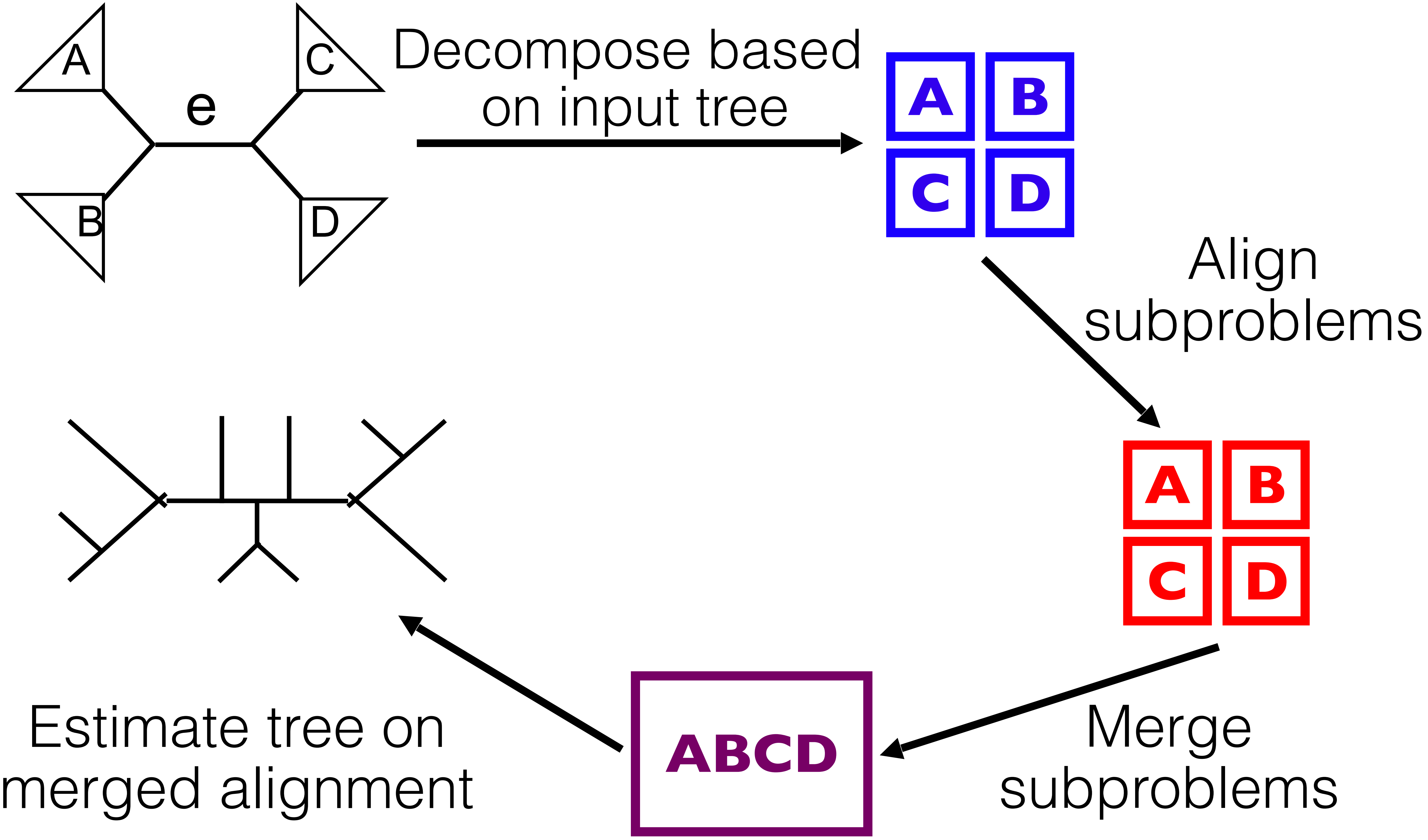


Decompose based on input tree

Align subproblems

Merge subproblems

SATé iteration
(Actual decomposition size is configurable)



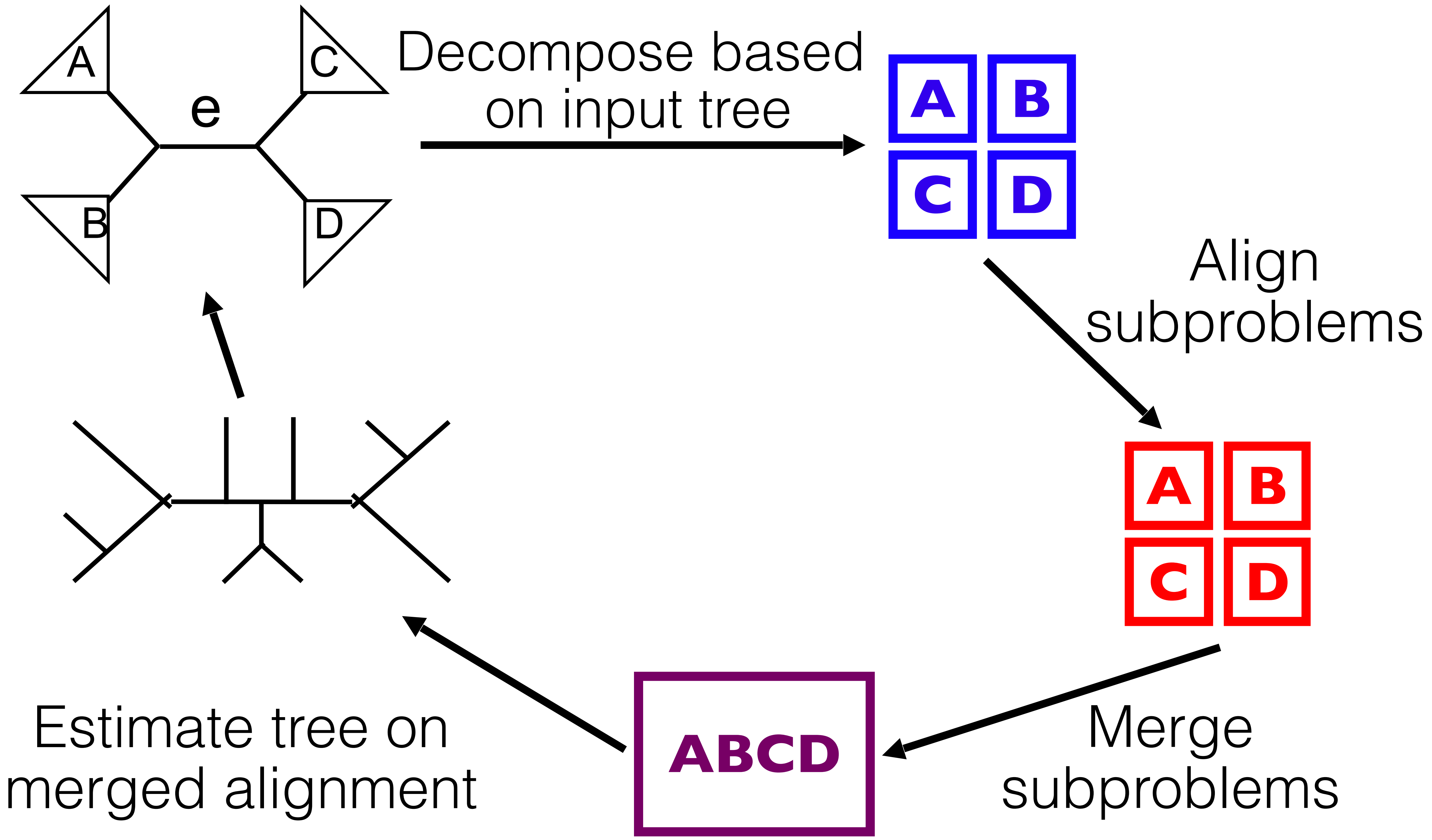
Decompose based on input tree

Align subproblems

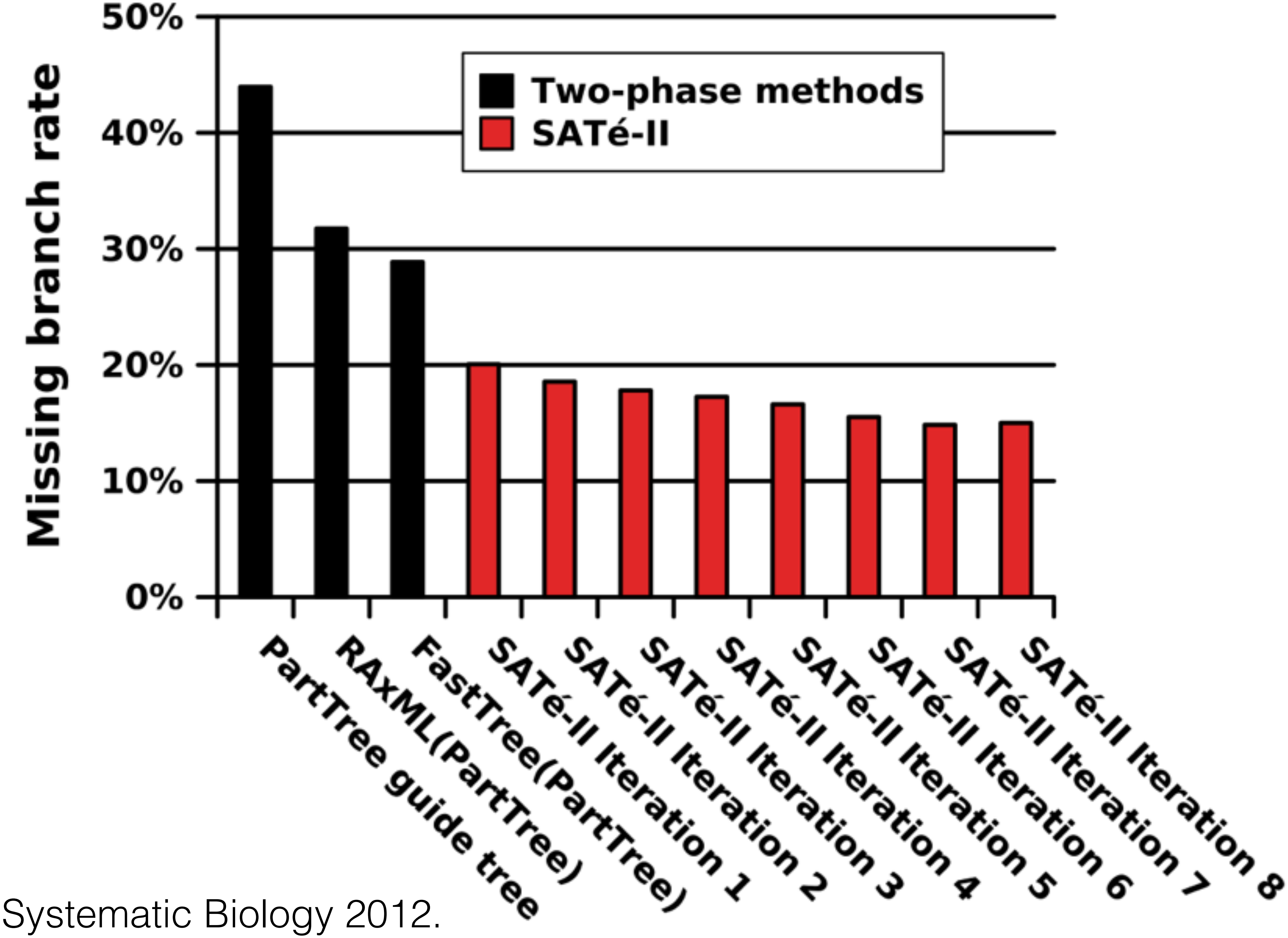
Merge subproblems

Estimate tree on merged alignment

SATé iteration
(Actual decomposition size is configurable)


A
C
e
B
D
Decompose based on input tree
A
B
C
D
Align subproblems
A
B
C
D
Merge subproblems
ABCD
Estimate tree on merged alignment

Results on a Dataset with 27,000 Sequences



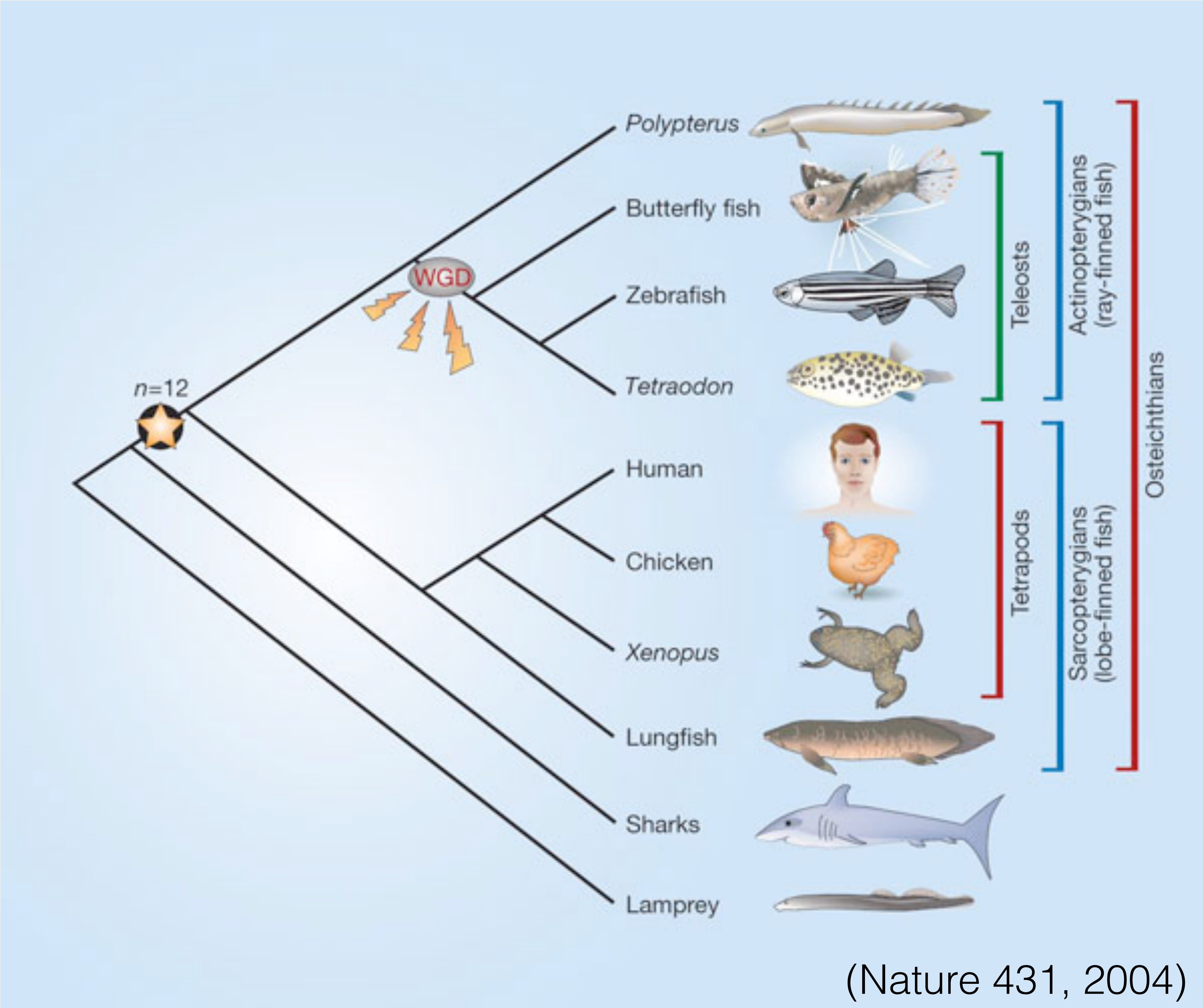
Liu *et al.* Systematic Biology 2012.

Summary of Part I

- Created novel tree-based divide-and-conquer techniques for simultaneous alignment and tree estimation, enabling:
    - Scalability to thousands of sequences or more
    - High accuracy
- Family of algorithms included:
    - SATé (Liu *et al.* Science 2009)
    - SATé-II (Liu *et al.* Systematic Biology 2012)
    - and others

Part II: Beyond Trees

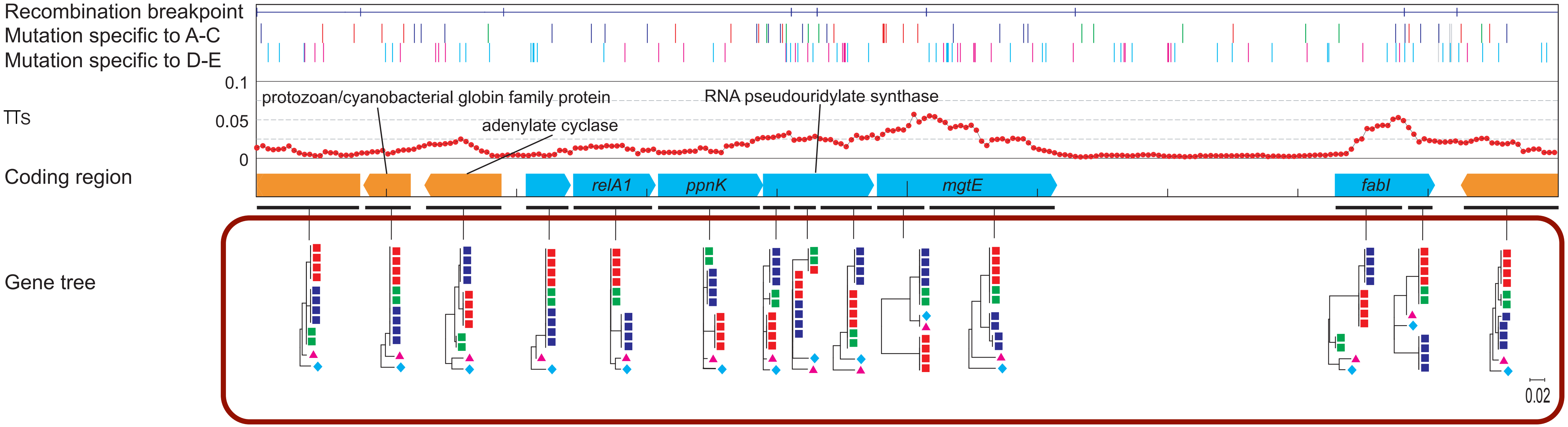
Almost all comparative genomic approaches assume that genomes have evolved down a tree.



(Nature 431, 2004)

- However, it has been shown that:
  - different genomic regions might evolve down different trees, and
  - the set of species might not have evolved in a strictly diverging manner.

(MBE 29, 2013)

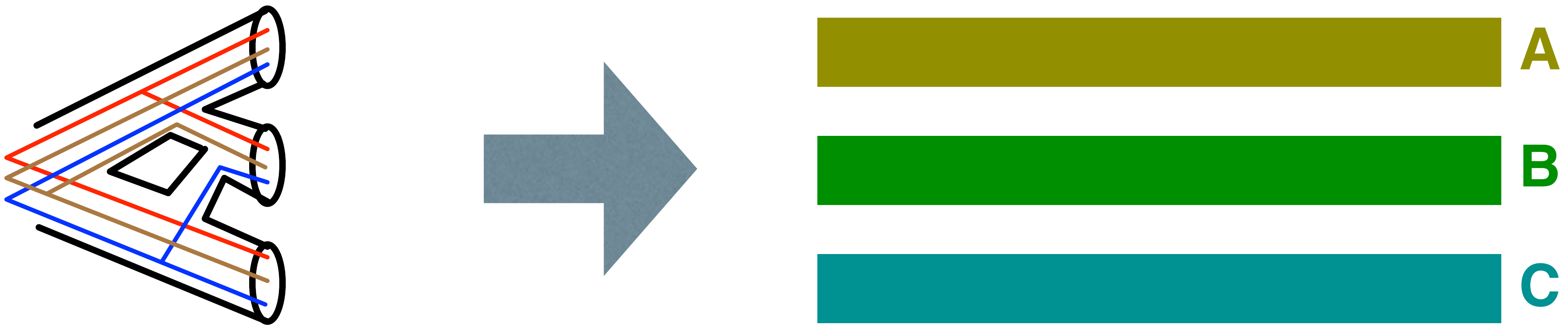


different gene trees for different regions in the Staph aureus genomes, due to horizontal gene transfer!

A Machine Learning View of Comparative Genomics

Species network
(DAG)
+
gene trees

Genomes



Stochastic
Generative
Model

→

Observed Data
(Genomic sequences)

Overarching Goal

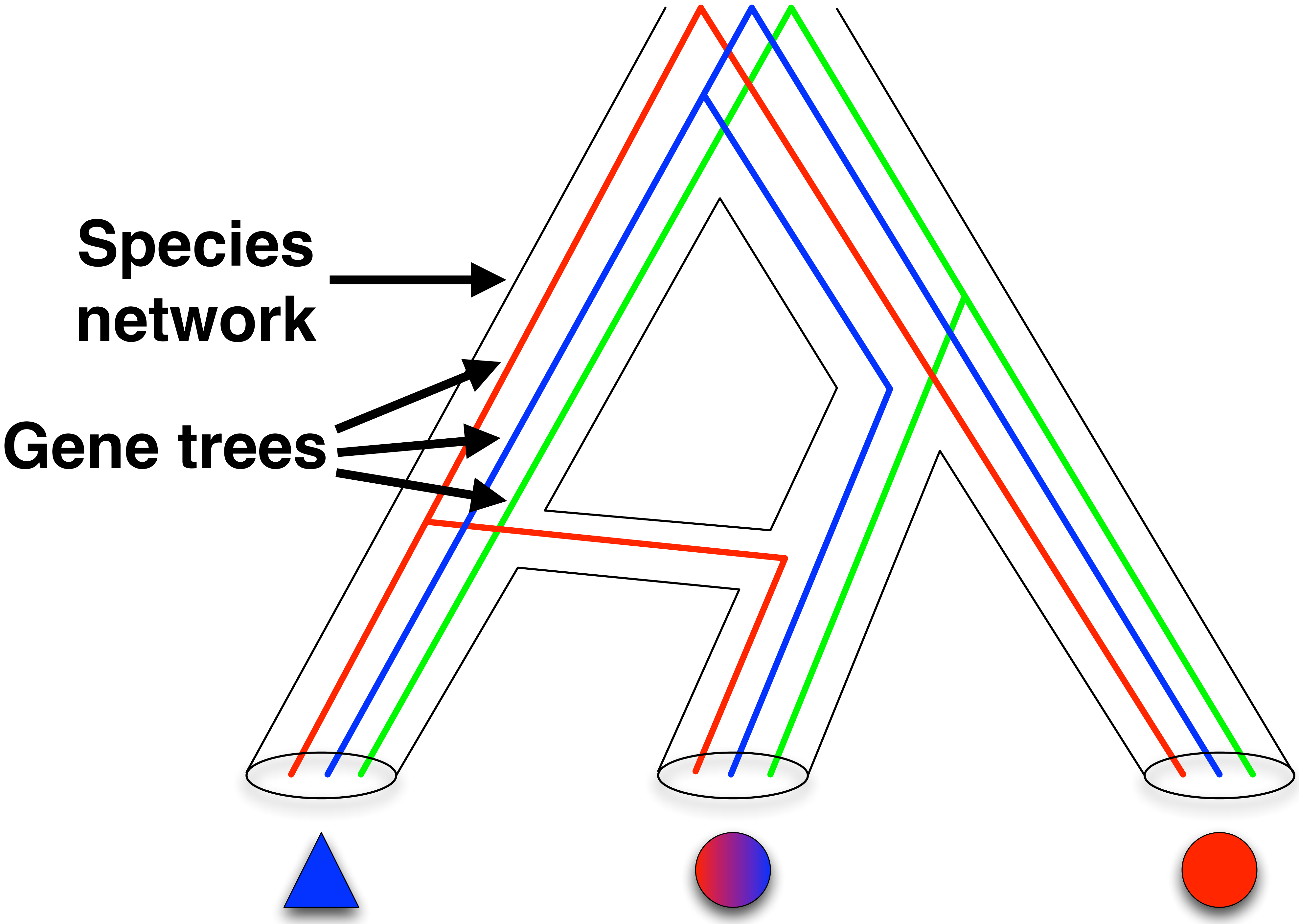
- For every site in the genome, learn:
  - the local gene tree along which the site evolved, and
  - the evolutionary trajectory that the local gene tree took within the species network.
- We also want a confidence measure for the inference.

My Approach

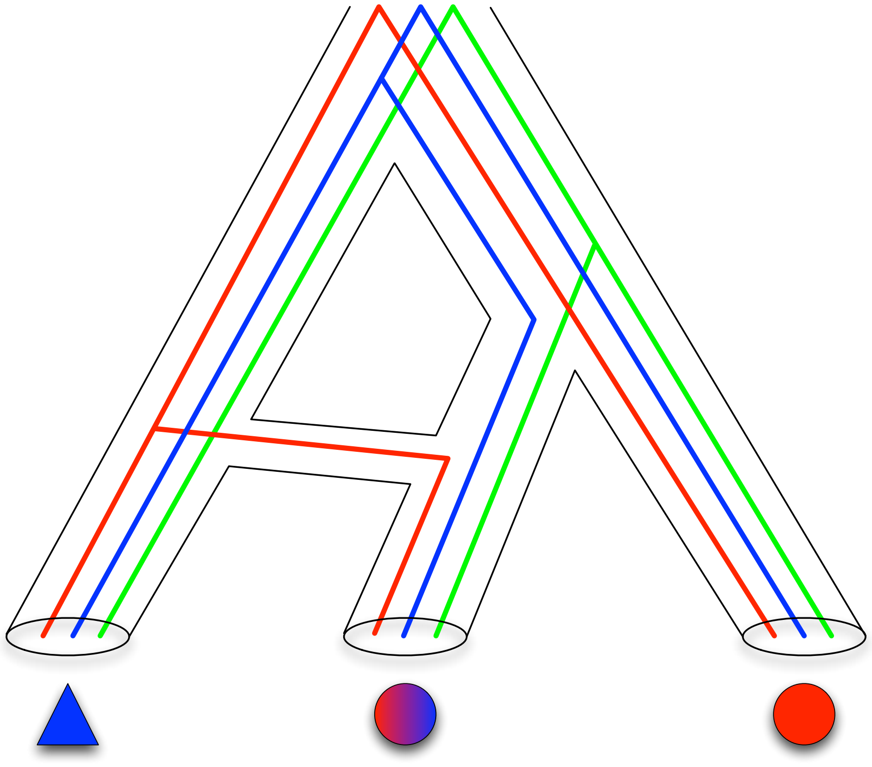
- Modeling: Combine species networks and hidden Markov models into one unified framework, PhyloNet-HMM.

- Inference: Using genomic sequence data, the task is to learn the model.

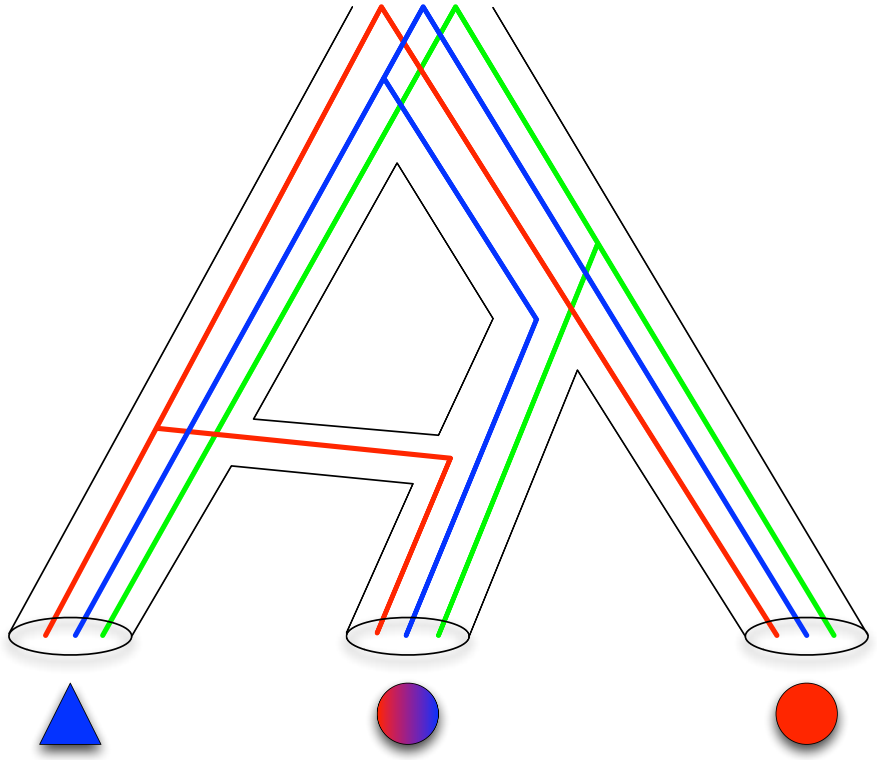
Gene Trees with Different Trajectories in a Species Network



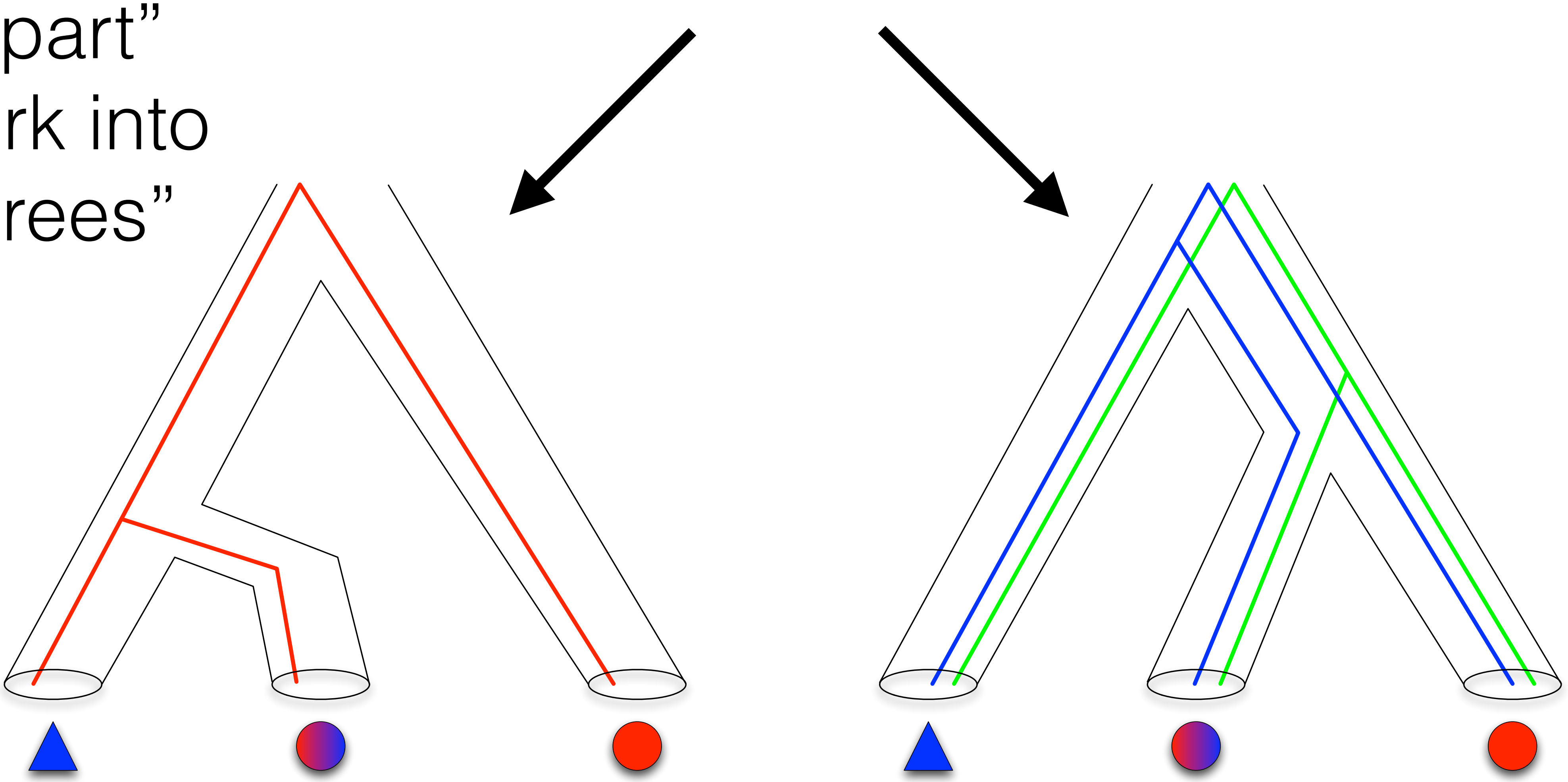
Disentangling Gene Tree Trajectories



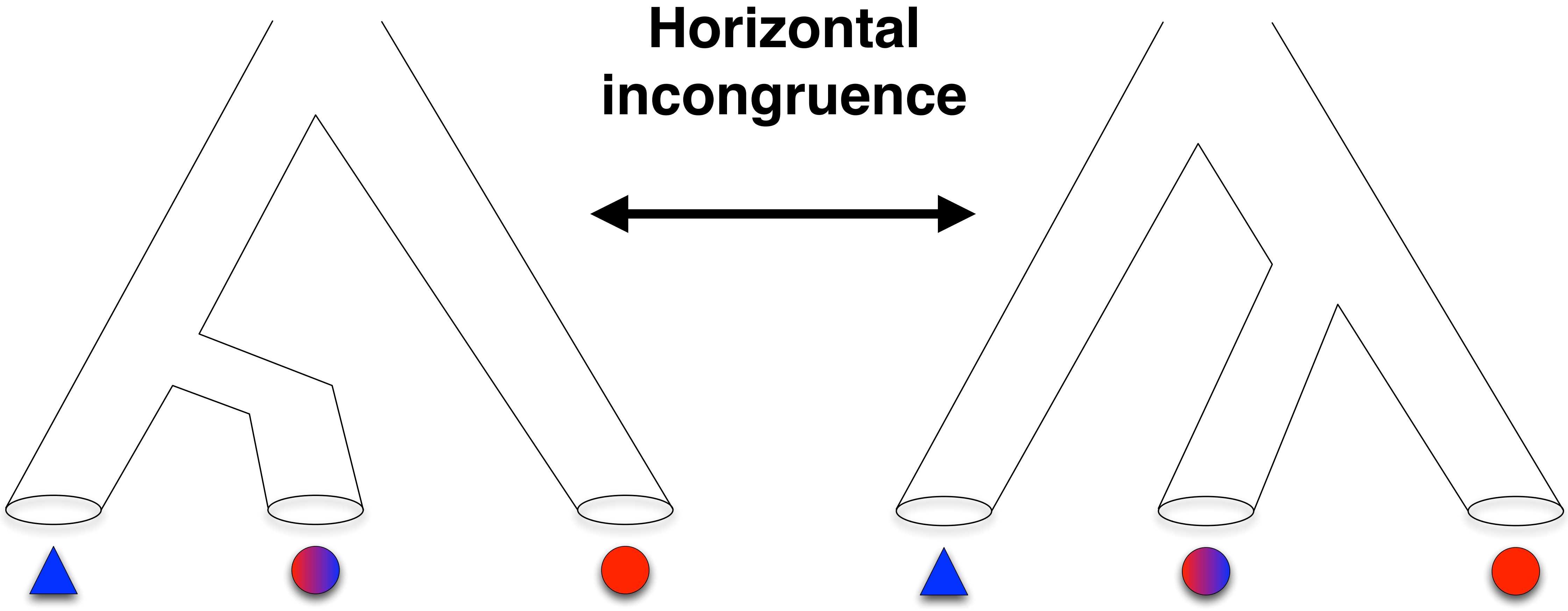
Disentangling Gene Tree Trajectories



Insight: “Pull apart” species network into two “parental trees”

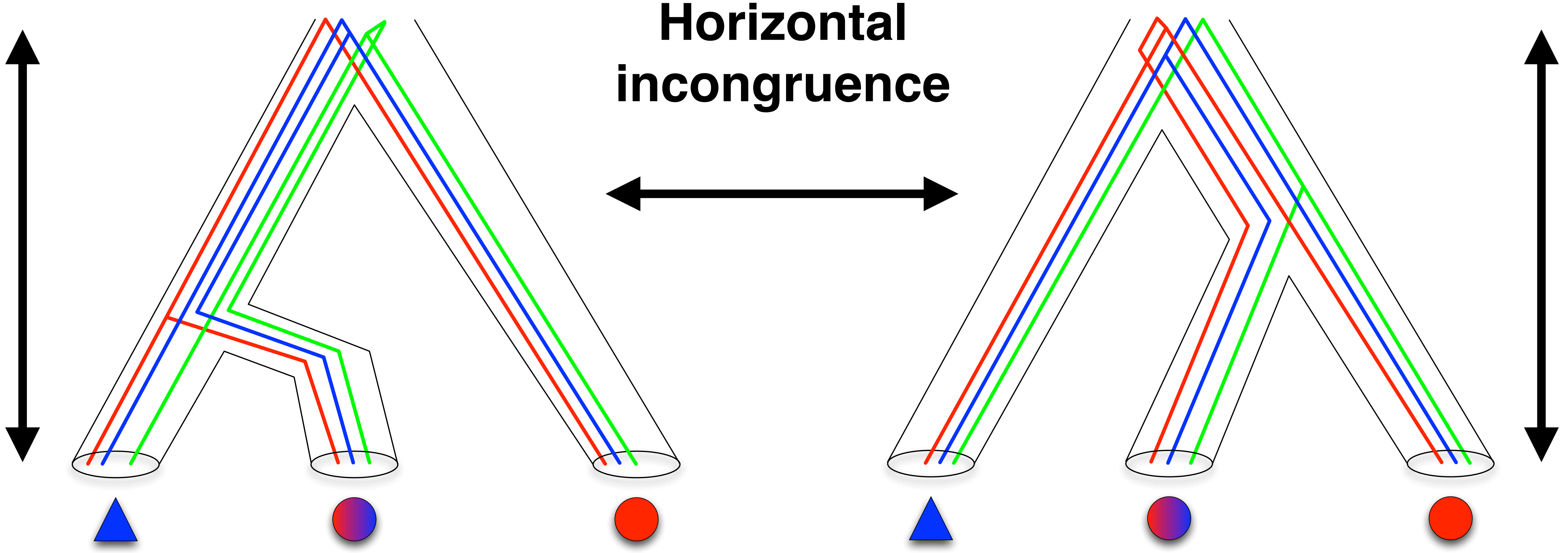


“Horizontal” and “Vertical” Incongruence



Horizontal incongruence

“Horizontal” and “Vertical” Incongruence

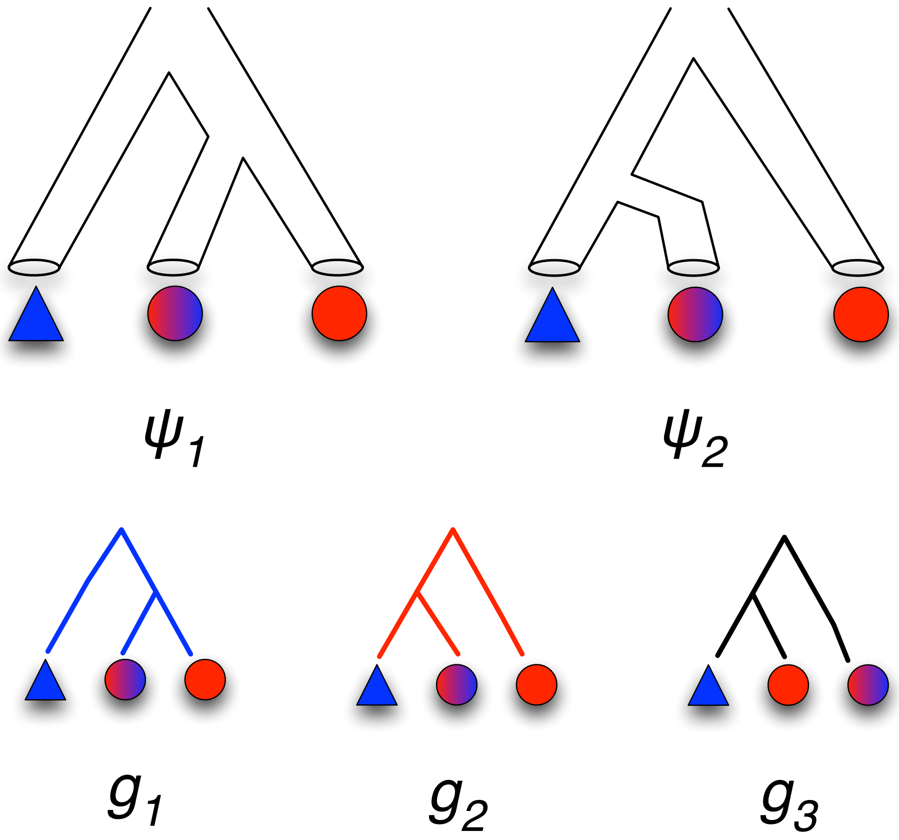


Vertical incongruence

Horizontal incongruence

Vertical incongruence

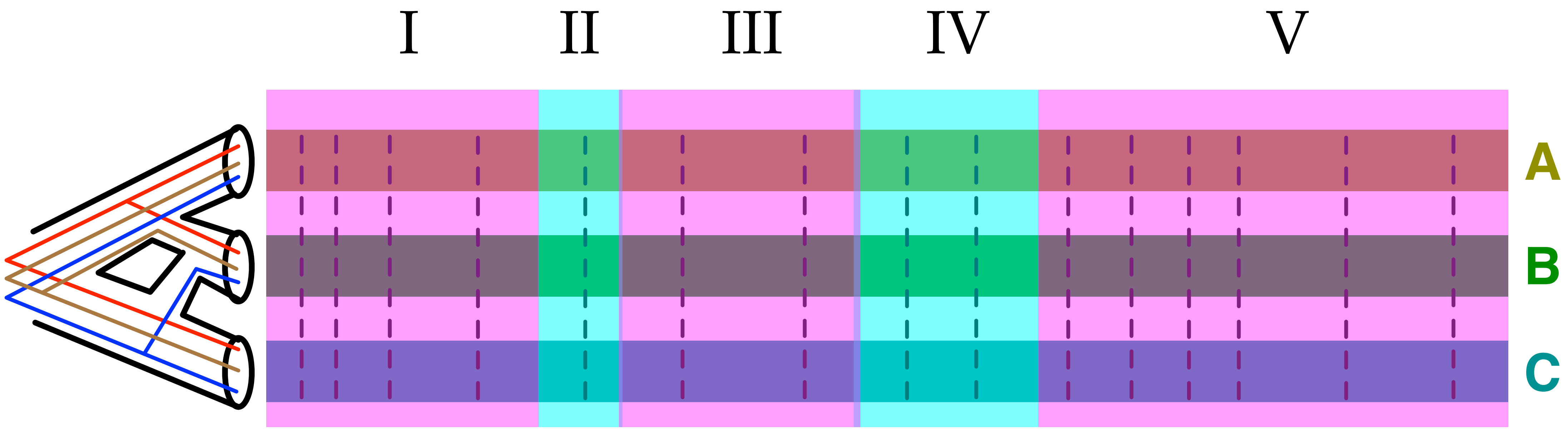
A Sequence-Level View of Local Incongruence



ψ_1 region 🟪

ψ_2 region 🟦

Gene-tree-switching breakpoint - - - - -



Insight #1

- “Horizontal” and “vertical” incongruence between neighboring gene trees represent two different types of dependence.
- Model the two dependence types using two classes of transitions in a graphical model.

Insight #2

- DNA sequences are observed, not gene trees.
- Under traditional models of DNA sequence evolution, the probability $P(s|g)$ of observing DNA sequences s given a gene tree g can be efficiently calculated using dynamic programming.

Insight #1 + Insight #2 = Use a Hidden Markov Model (HMM)

Hidden Markov Model (HMM) Example

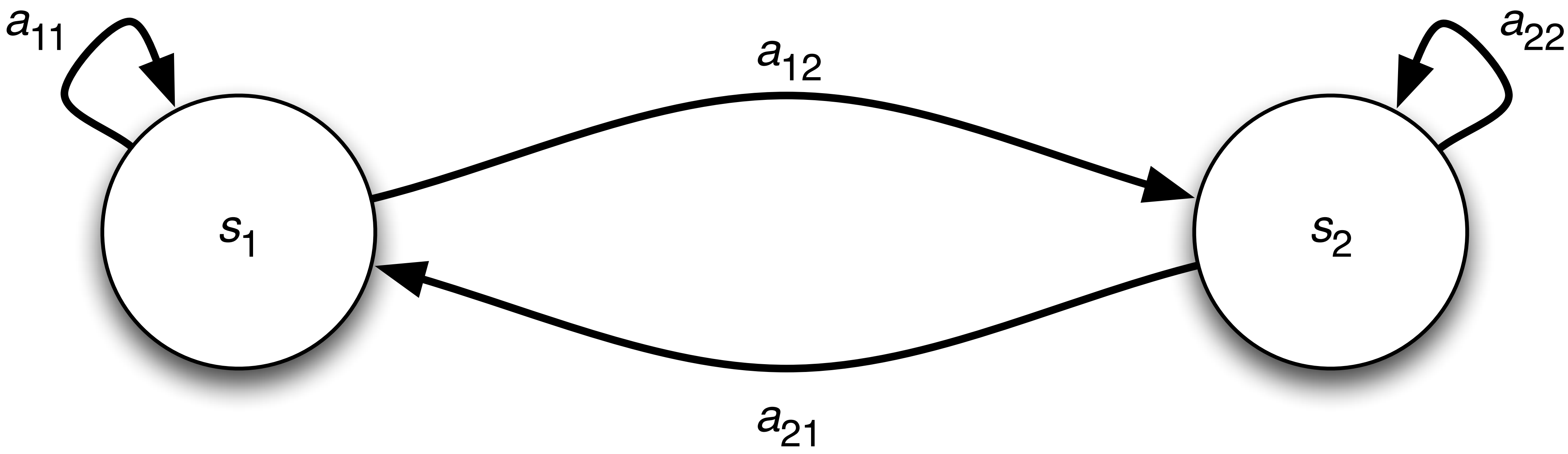
- Coin tossing experiment:

  1. An experimenter flips one of two hidden coins with unknown bias and tells you the result.

  2. Repeat for a total of *k* trials, resulting in observation sequence *O*.

Example adapted from Rabiner (1989).

Hidden Markov Model (HMM) Example



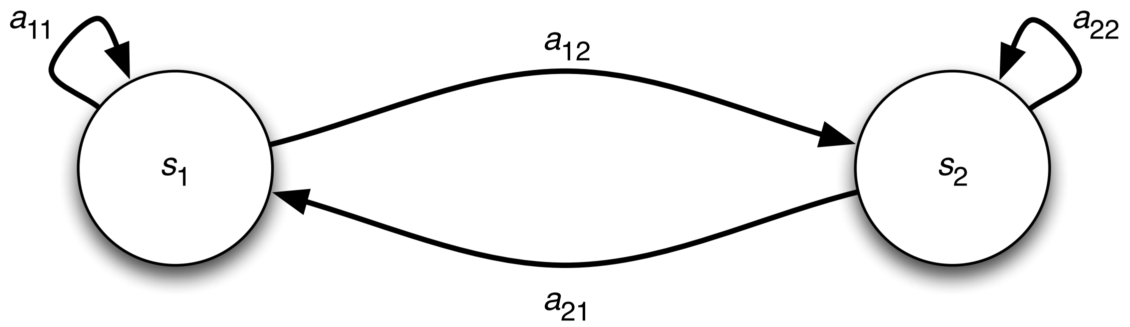
$\mathbf{P}(H|q_t = s_1) = b_1$

$\mathbf{P}(T|q_t = s_1) = 1 - b_1$

$\mathbf{P}(H|q_t = s_2) = b_2$

$\mathbf{P}(T|q_t = s_2) = 1 - b_2$

Hidden Markov Model (HMM) Example



- The HMM has N=2 states.
- The HMM is in state q_t at time t, where $1 \leq t \leq k$.
- The set of HMM parameters λ consists of:
    - The transition probability matrix $A = \{a_{ij}\}$
    - The emission probabilities $B = \{b_i\}$
    - The initial state distribution $\pi_i = \mathbf{P}(q_1 = s_i)$

Three Problems Addressed Using HMMs

1. What is the likelihood of the model given the observation sequence?

   - Forward algorithm calculates prefix probability $\alpha_t(i) = \mathbf{P}(O_1, O_2, \ldots, O_t, q_t = S_i|\lambda)$

   - Backward algorithm calculates suffix probability $\beta_t(i) = \mathbf{P}(O_{t+1}, O_{t+2}, \ldots, O_k|q_t = S_i, \lambda)$

   - Model likelihood is $\mathbf{P}(O|\lambda) = \sum_{i=1}^{N} \alpha_k(i)$

2. Which sequence of hidden states best explains the observation sequence?

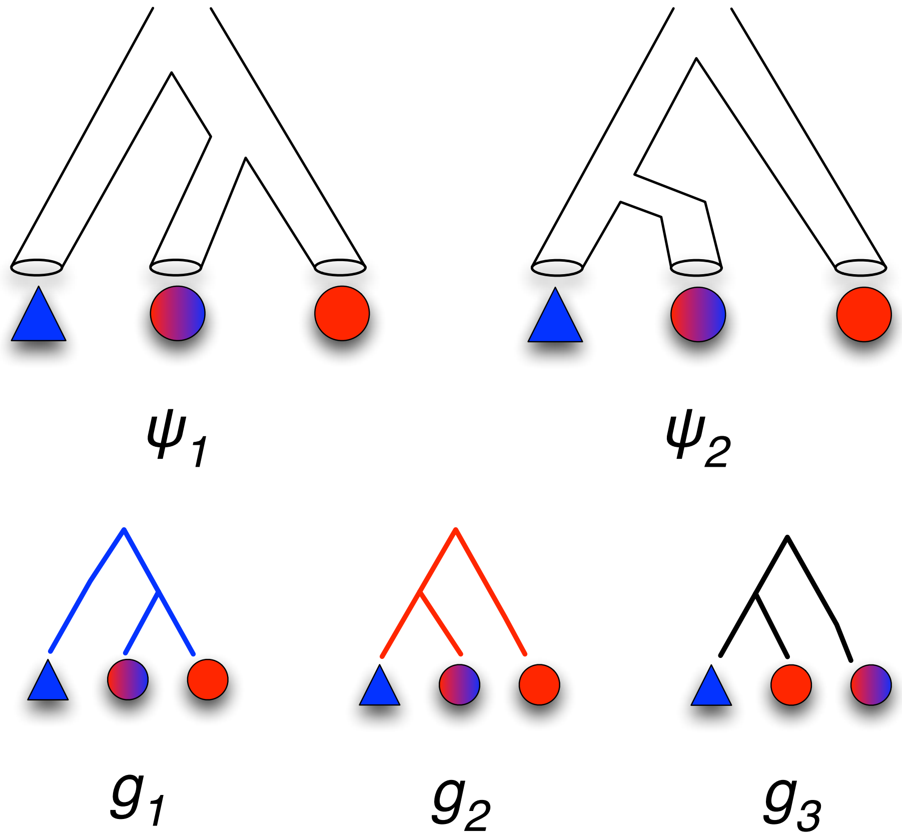
   - Posterior decoding probability $\gamma_t(i)$ is the probability that HMM is in state s_i at time t, calculated as:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\mathbf{P}(O|\lambda)}$$

3. How do we choose parameter values that maximize the model likelihood?

   - Apply Baum-Welch algorithm to search for $\arg\max_{\lambda} \mathbf{P}(O|\lambda)$

PhyloNet-HMM: Problem Definition



For each site $1 \leq i \leq k$, let π_i be a random variable that takes a value from the set $(g_x, \psi_y) : g_x \in G(n), \psi_y \in \Psi$.

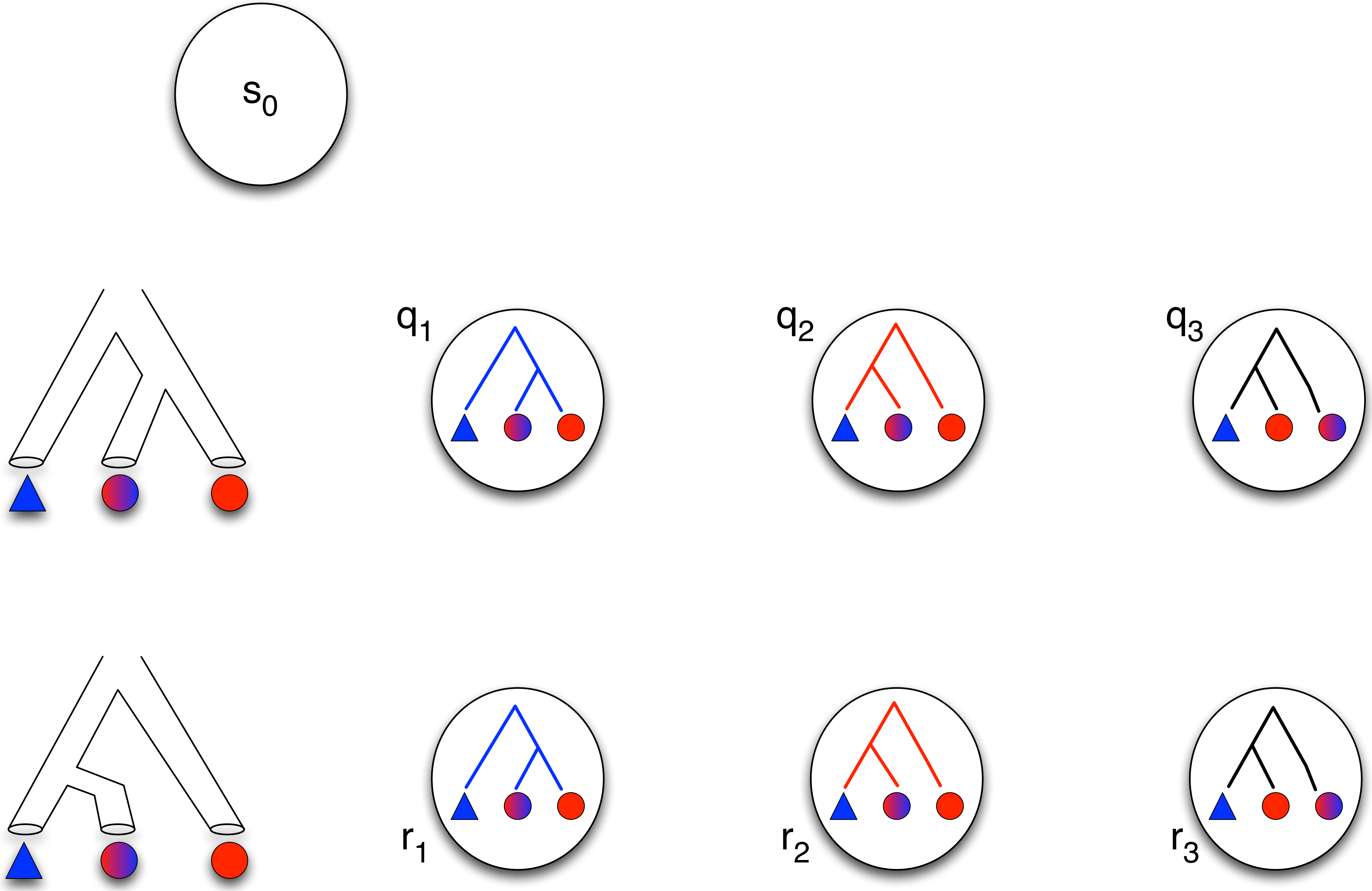
Input: A set S of n aligned genomes, each of length k, and a set Ψ of parental trees corresponding to a species network.

Output: For each site $1 \leq i \leq k$, the probability

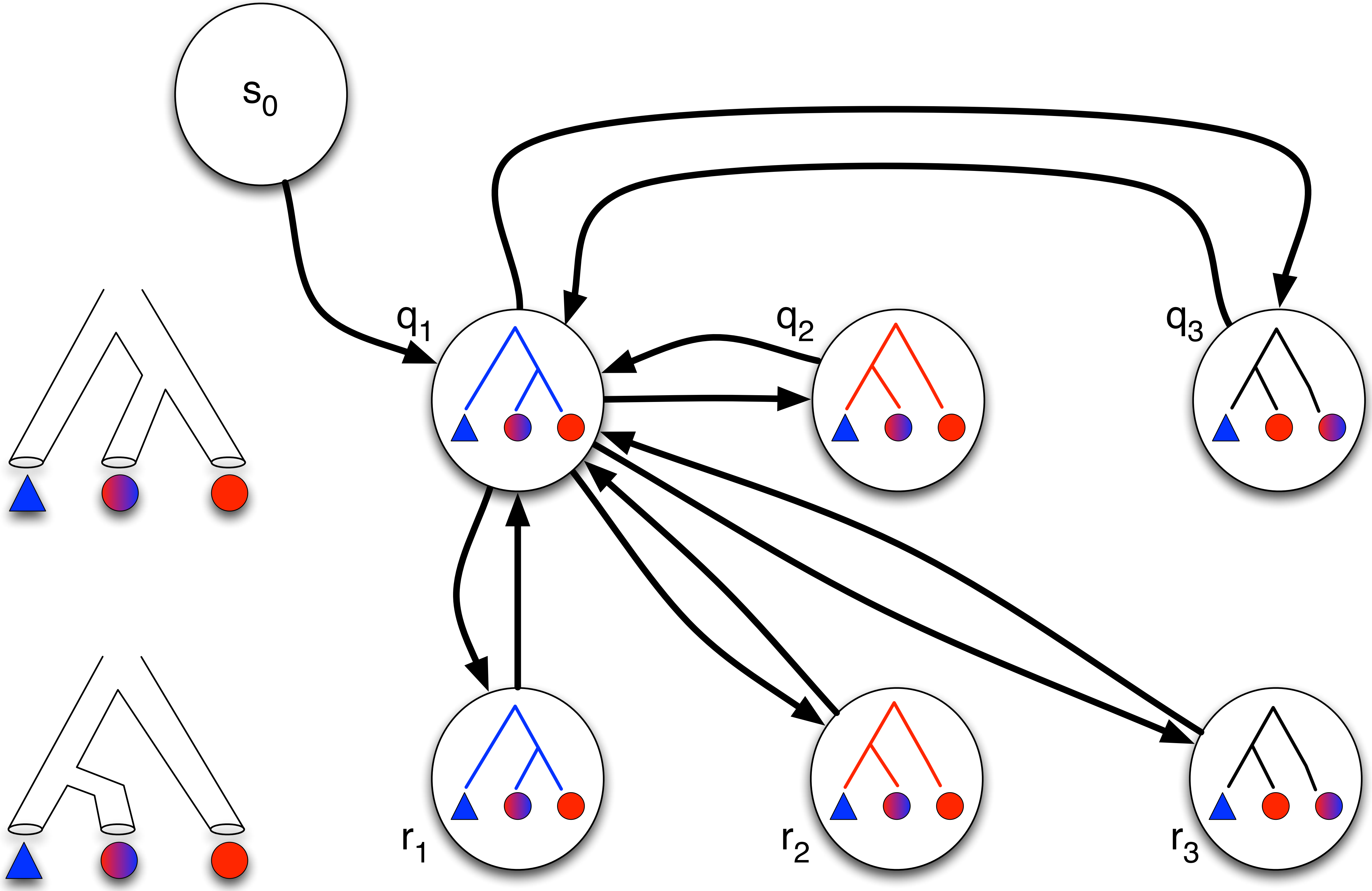
$$\mathbf{P}(\pi_i = (g_x, \psi_y)|S)$$

for every $g_x \in G(n)$ and $\psi_y \in \Psi$.

PhyloNet-HMM: Hidden States



PhyloNet-HMM: Hidden States and Transitions Involving *q_1*



PhyloNet-HMM

- Each hidden state s_i is associated with a gene tree $g(s_i)$ contained within a “parental” tree $f(s_i)$
- The set of HMM parameters λ consists of
  - The initial state distribution π
  - Transition probabilities

$$a_{ij} = \begin{cases} \mathbf{P}(g(s_i)|f(s_i)) \cdot \gamma & \text{if } s_i \text{ and } s_j \text{ in different rows} \\ \mathbf{P}(g(s_i)|f(s_i)) \cdot (1-\gamma) & \text{if } s_i \text{ and } s_j \text{ in same row} \end{cases}$$

where γ is the “horizontal” parental tree switching frequency.

  - The emission probabilities b_i = $\mathbf{P}(O_t|g(s_i))$

PhyloNet-HMM: Two Calculations

- The probability of a gene tree topology *g* given a containing species tree (*Ψ*,*λ*) (Degnan and Salter 2005):

$$P_{\psi,\boldsymbol{\lambda}}(G = g) = \sum_{\mathbf{h} \in H_{\psi}(g)} \frac{w(\mathbf{h})}{d(\mathbf{h})} \prod_{b=1}^{n-2} \frac{w_b(\mathbf{h})}{d_b(\mathbf{h})} p_{u_b(\mathbf{h}) v_b(\mathbf{h})}(\lambda_b).$$

- The probability of observing DNA sequences *S* given a gene tree (*g, ω*) can be efficiently computed using dynamic programming (Felsenstein 1981).

Three Problems Addressed Using PhyloNet-HMM

1. What is the likelihood of the model given the observed DNA sequences?

   - Forward algorithm calculates prefix probability $\alpha_t(i) = \mathbf{P}(O_1, O_2, \ldots, O_t, q_t = S_i|\lambda)$

   - Backward algorithm calculates suffix probability $\beta_t(i) = \mathbf{P}(O_{t+1}, O_{t+2}, \ldots, O_k|q_t = S_i, \lambda)$

   - Model likelihood is $\mathbf{P}(O|\lambda) = \sum_{i=1}^{N} \alpha_k(i)$

2. Which sequence of hidden states best explains the observed DNA sequences?

   - Posterior decoding probability $\gamma_t(i)$ is the probability that HMM is in state s_i at time t, calculated as:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\mathbf{P}(O|\lambda)}$$

3. How do we choose parameter values that maximize the model likelihood?

   - Apply E-M to optimize $\arg\max_{\lambda} \mathbf{P}(O|\lambda)$

Related Methods

- Current methods for inference under species networks fall into two classes:

  1. Methods that work for at most three genomes, e.g.

     - D-statistic (Durand *et al.* 2012)

     - CoalHMM (Mailund *et al.* 2012)

  2. Methods that consider vertical incongruence or horizontal incongruence but not both, e.g.

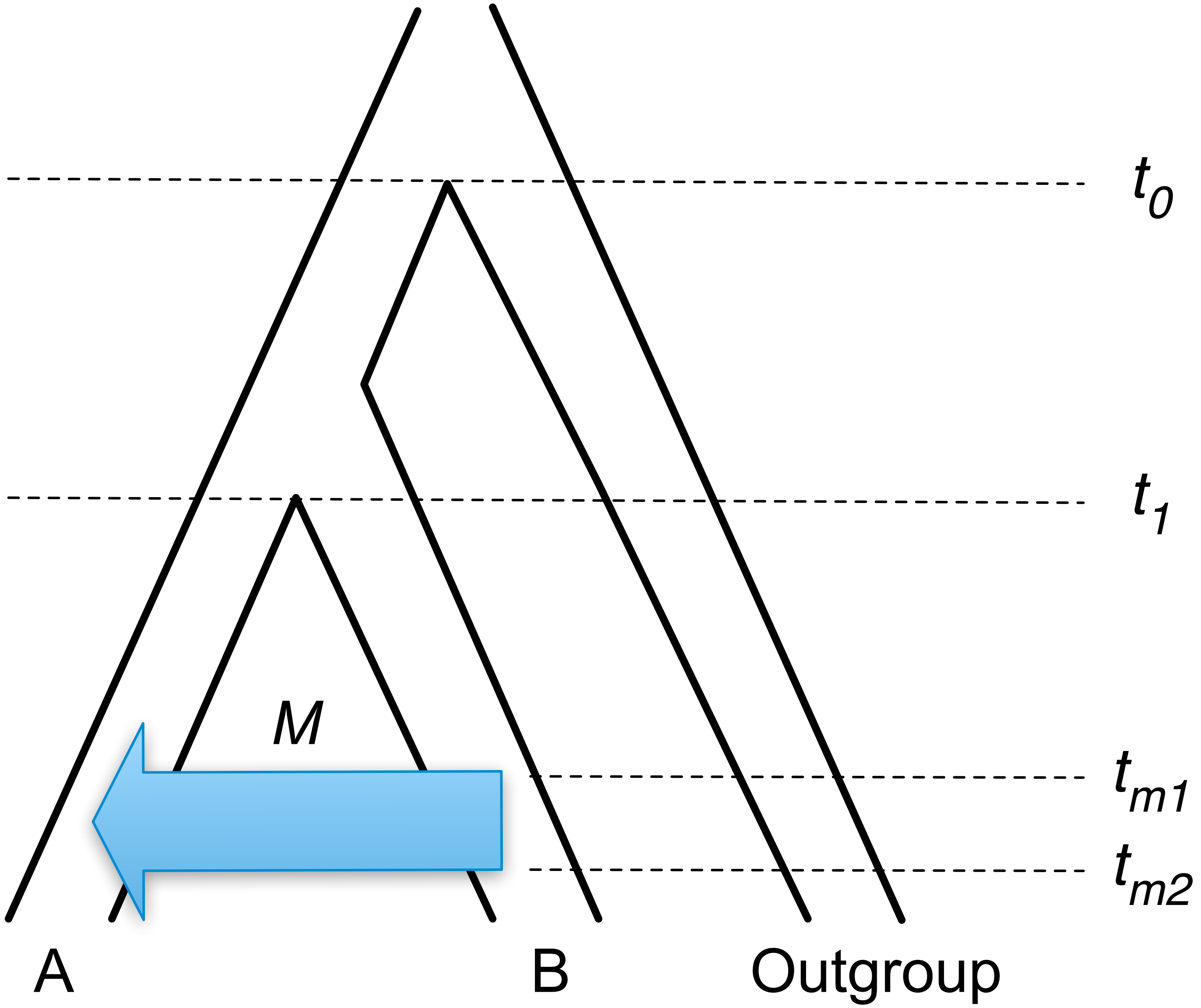
     - CoalHMM (Hobolth *et al.* 2007, Schierup *et al.* 2009)

     - RecHMM (Westesson and Holmes 2009)

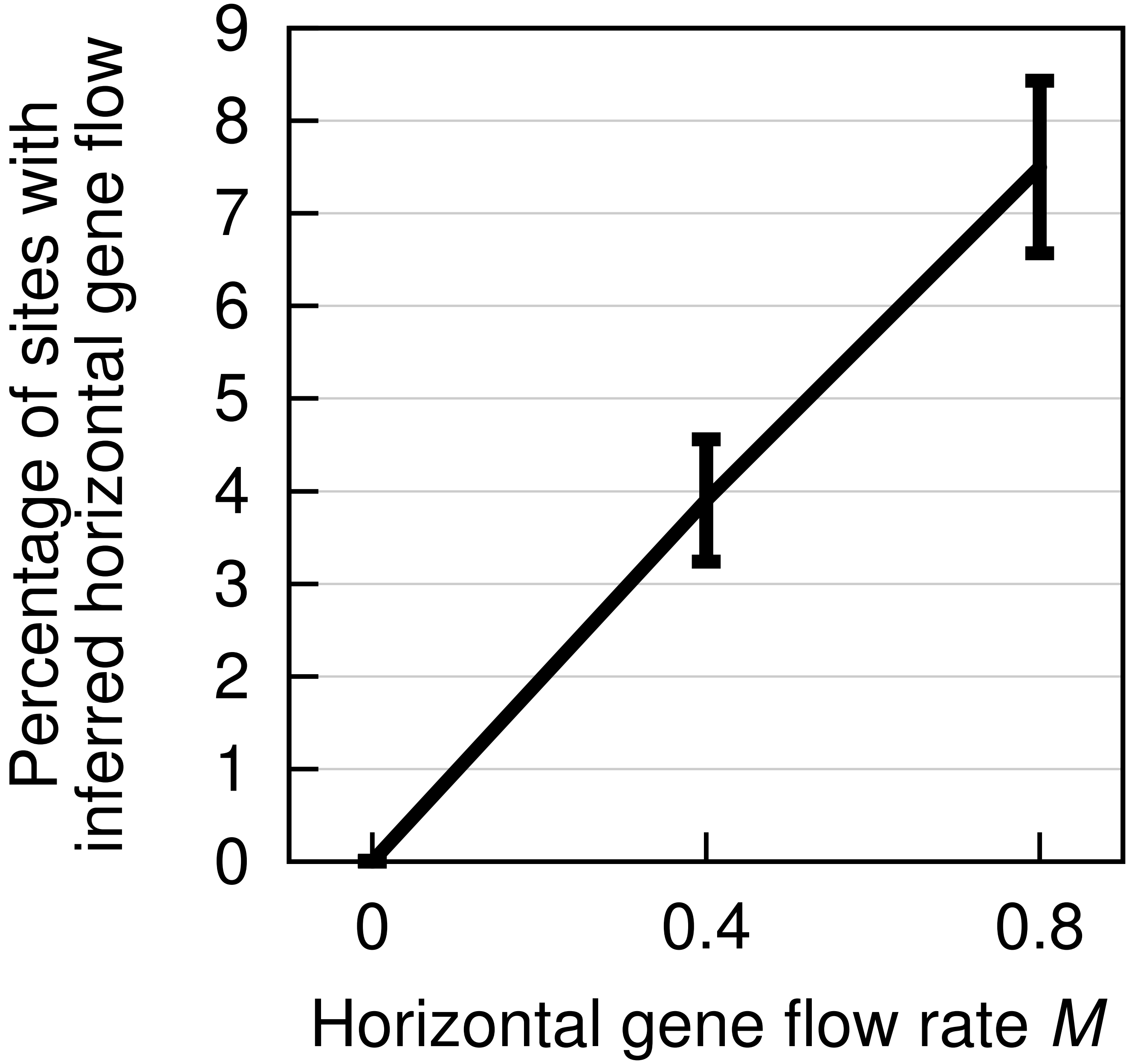
Evaluating PhyloNet-HMM

- Simulation study using:
    - Species tree model
    - Species network model
- Empirical study of different sets of mouse genomes:
    - Controls: lab mice, wild mice from populations that lacked gene flow
    - Additional wild mice from populations where gene flow was suspected

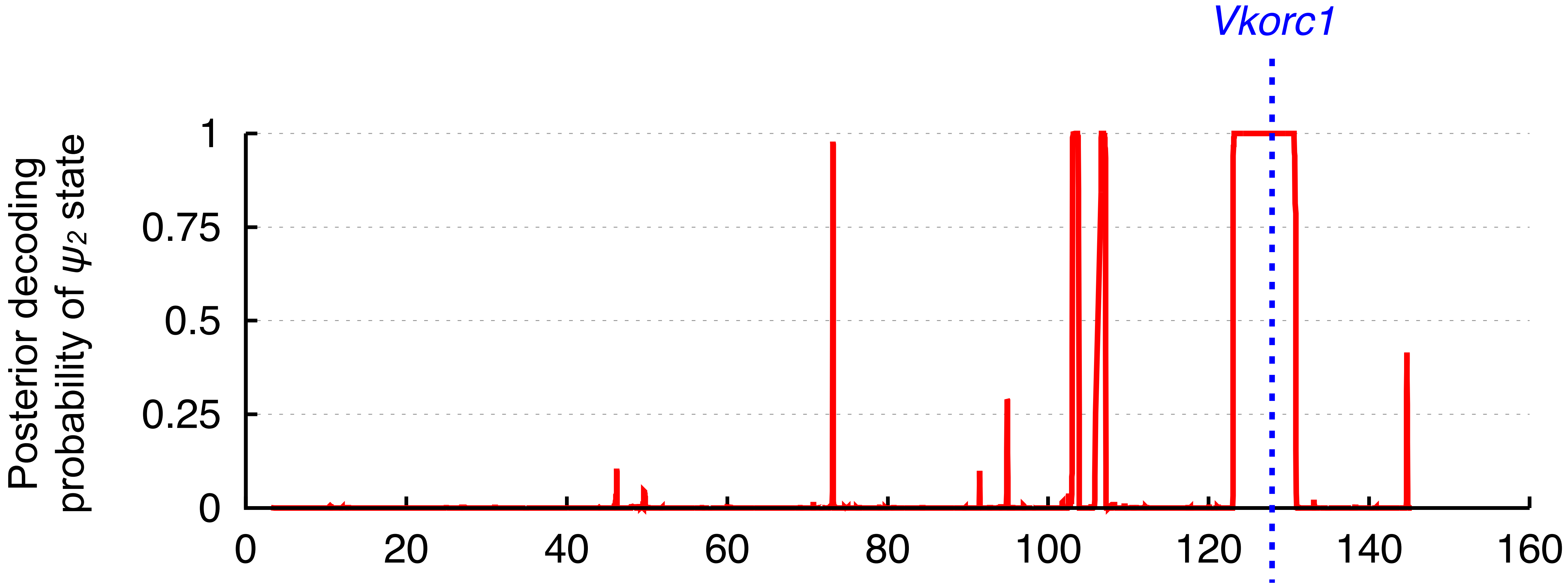
Simulation Model



Simulation Study Results



Empirical Study: Non-control Mice (Chromosome 7)



Liu *et al.*, revision under review, PLoS Computational Biology.

The *Vkorc1* Gene and Personalized Warfarin Therapy

nature
International weekly journal of science
www.nature.com/nature
Warfarin target protein
Linking vitamin K disorders to anticoagulation
Binary asteroids
Partner-swapping in the Kuiper belt
Ammonia synthesis
Nitrogen fixation dissected
Bird flu
How worried should we be?
naturejobs systems biology

- Mutant *Vkorc1* gene contributes to warfarin resistance
- Warfarin resistant individuals require larger-than-normal dose to prevent clotting complications (like stroke)

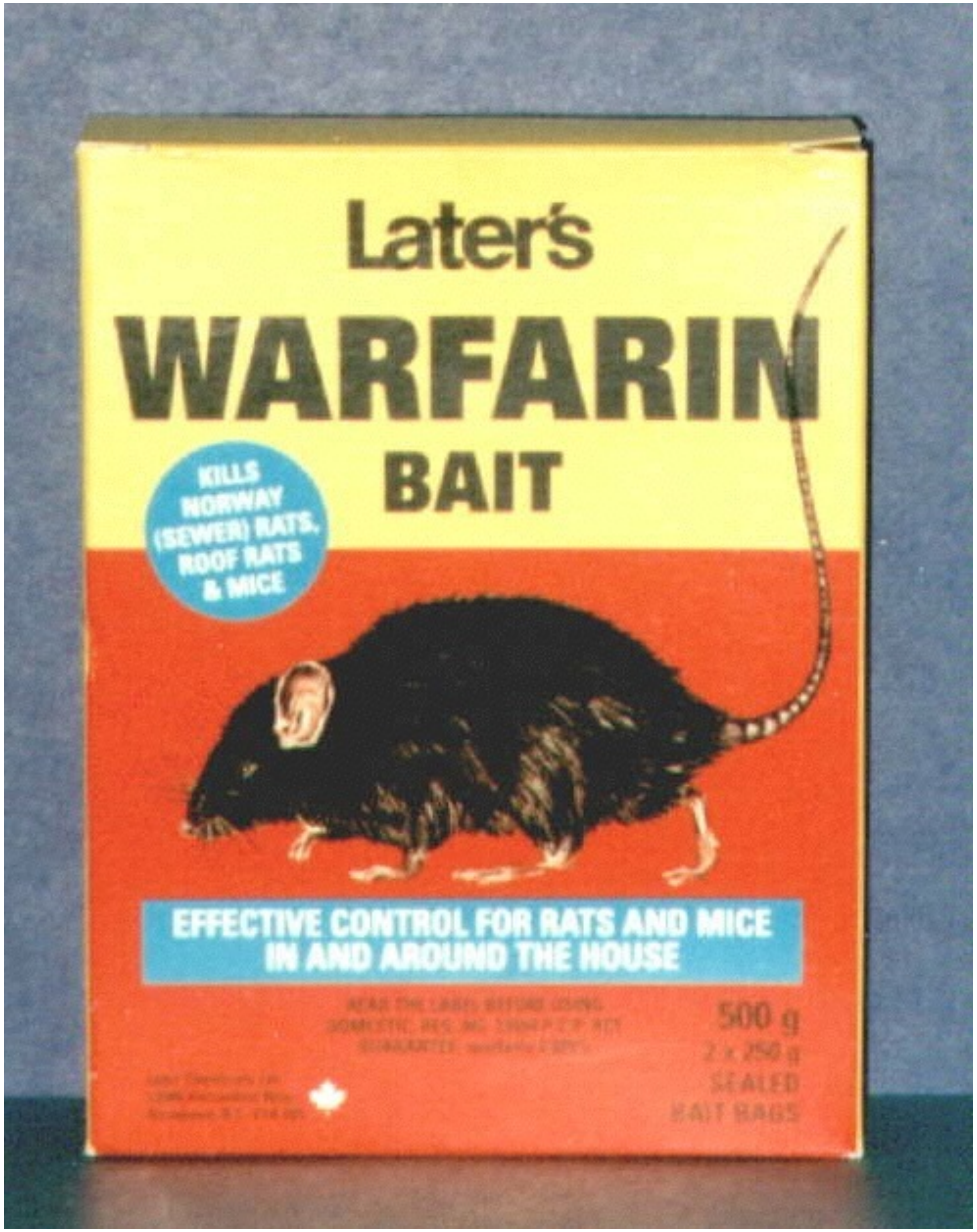
Rost *et al.* Nature 427, 537-541 2004.

Warfarin and Adverse Events

- Warfarin is the most widely prescribed blood thinner
- Treatment is complicated because every patient is different
  - Gene mutations confer resistance or susceptibility
- Annually,
  - 85,000 serious bleeding events
  - 17,000 strokes
  - Cost: $1.1 billion

McWilliam et al. AEI-Brookings Joint Center 2006.

Warfarin is Really Glorified Rodent Poison


Later's
WARFARIN
BAIT
KILLS NORWAY (SEWER) RATS, ROOF RATS & MICE
EFFECTIVE CONTROL FOR RATS AND MICE IN AND AROUND THE HOUSE
500 g
2 x 250 g
SEALED
BAIT BAGS

Reproduced from UTMB.

The Spread of Warfarin Resistance in Wild Mice

- Humans inadvertently started a gigantic drug trial by giving warfarin to mice in the wild
- Mice shared genes (including one that confers warfarin resistance) to survive (Song *et al.* 2011)
    - Gene sharing occurred between two different species (introgression)
- To find out results from the drug trial, we just need to analyze the genomes of introgressed mice and locate the introgressed genes

Summary of Part II

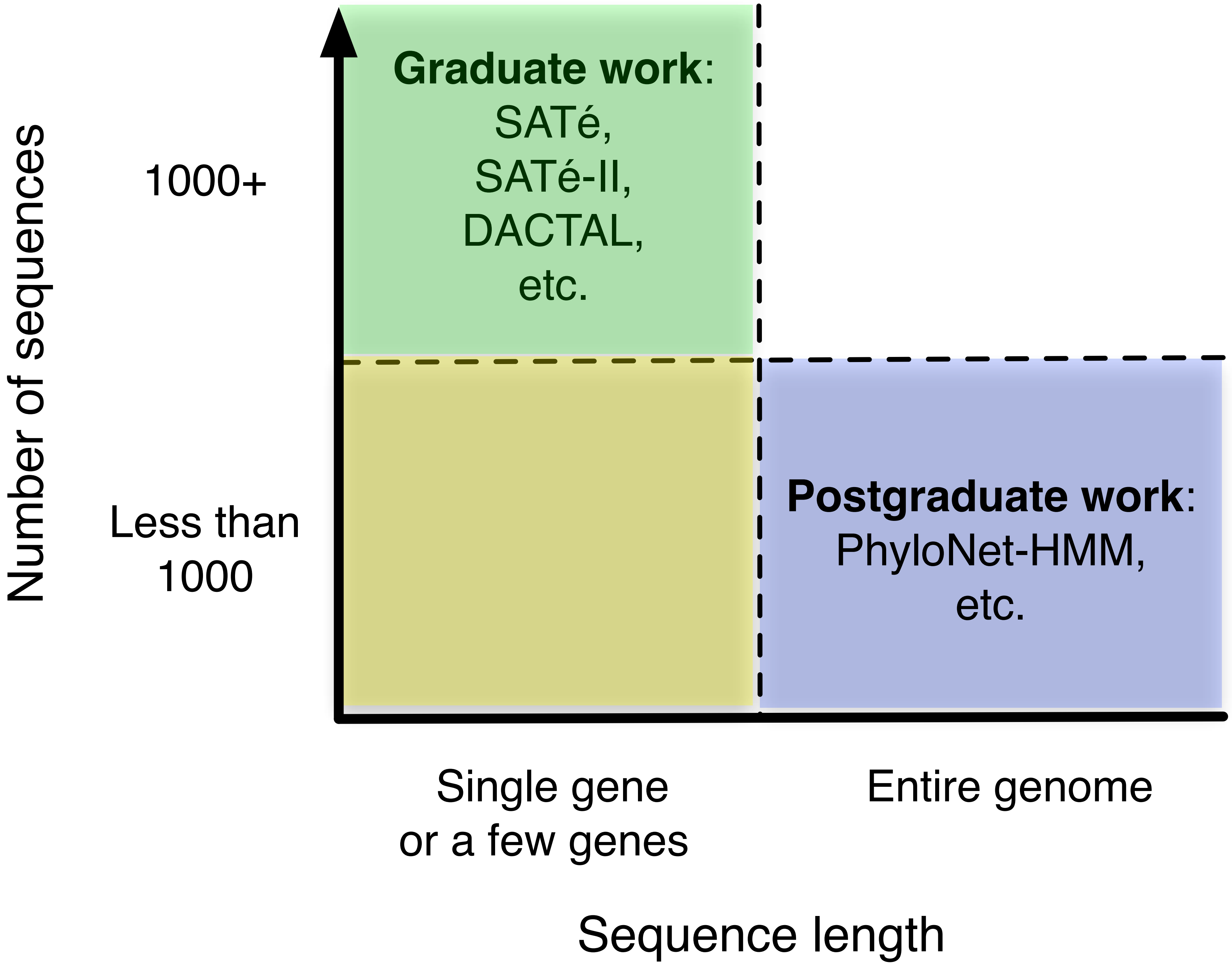
- PhyloNet-HMM generalizes the basic coalescent model, one of the most widely used models in population genetics, by using a DAG in place of a tree
- Simulated and empirical data sets with tree-like and non-tree-like evolution were used to validate PhyloNet-HMM
- PhyloNet-HMM found non-tree-like evolution in multiple mouse chromosomes
    - Introgressed mouse genes confer warfarin resistance, many with related human genes
    - New candidate genes to target for improved personalization of warfarin therapy
- Study of non-tree-like evolution is a fundamentally important research topic in biology

Future Research and Summary

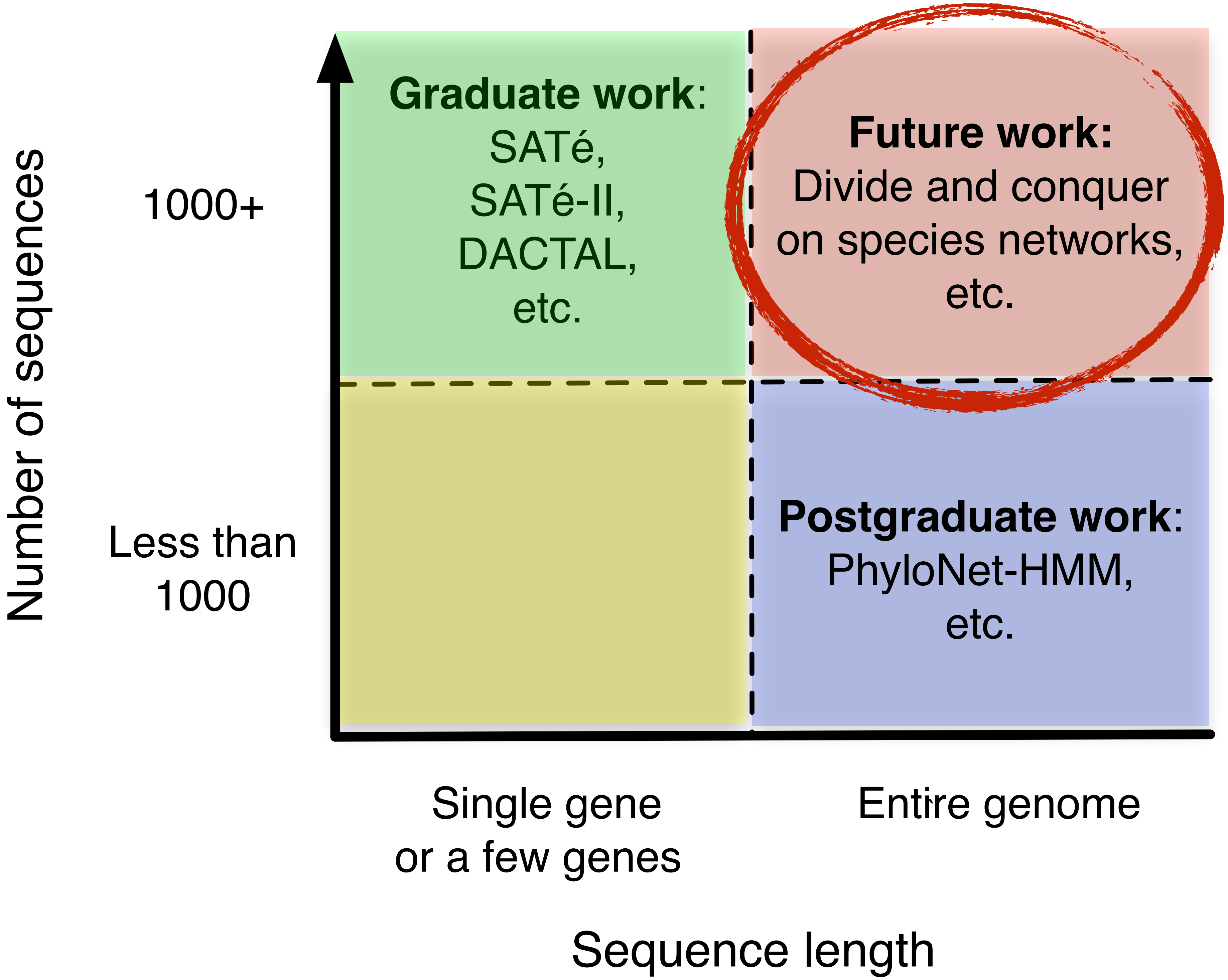
Future Direction #1

- Previous analyses (at most five genomes and a single network edge) required more than a CPU-month on a large cluster
- Problem is combinatorial in both the number of genomes and the number of network edges
- Challenge: efficient and accurate network-based inference from hundreds of genomes or more

My Contributions: A Big Data Perspective

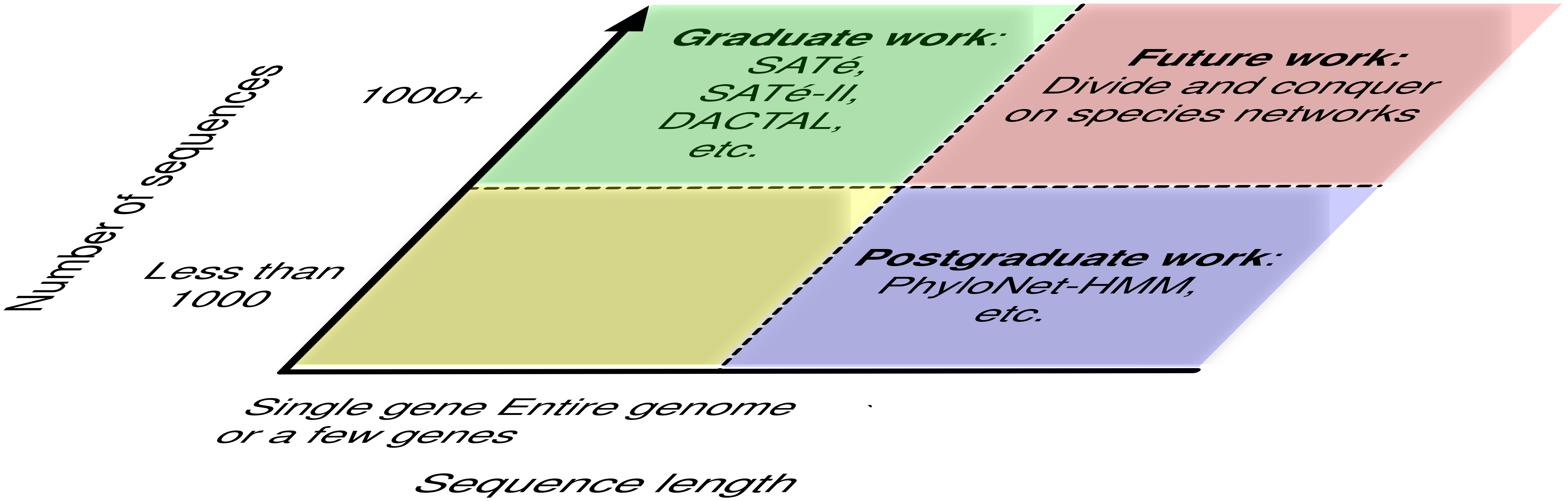


Future Direction #1

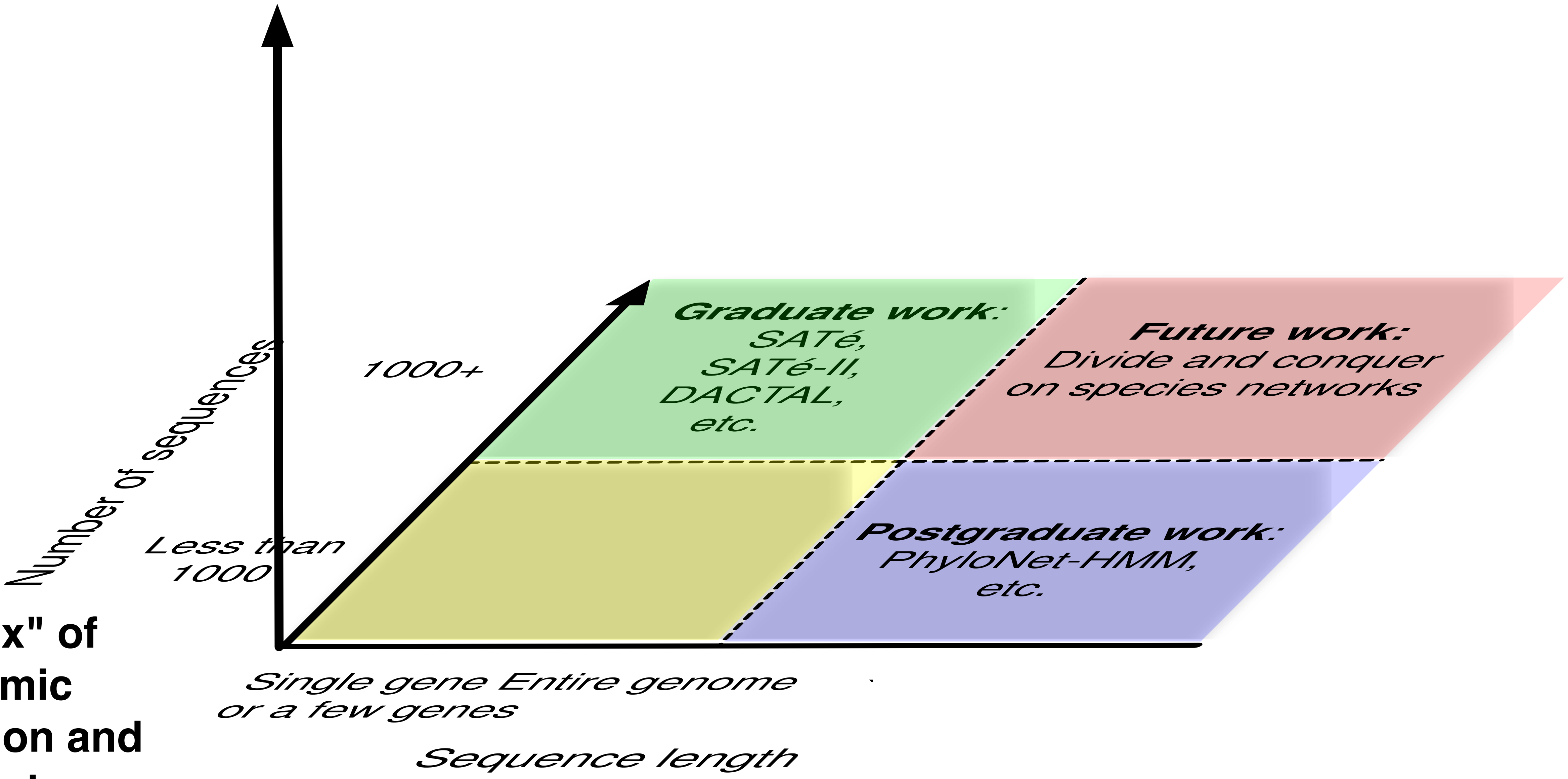


Number of sequences
1000+
Less than 1000
Graduate work: SATé, SATé-II, DACTAL, etc.
Future work: Divide and conquer on species networks, etc.
Postgraduate work: PhyloNet-HMM, etc.
Single gene or a few genes
Entire genome
Sequence length

Future Direction #2



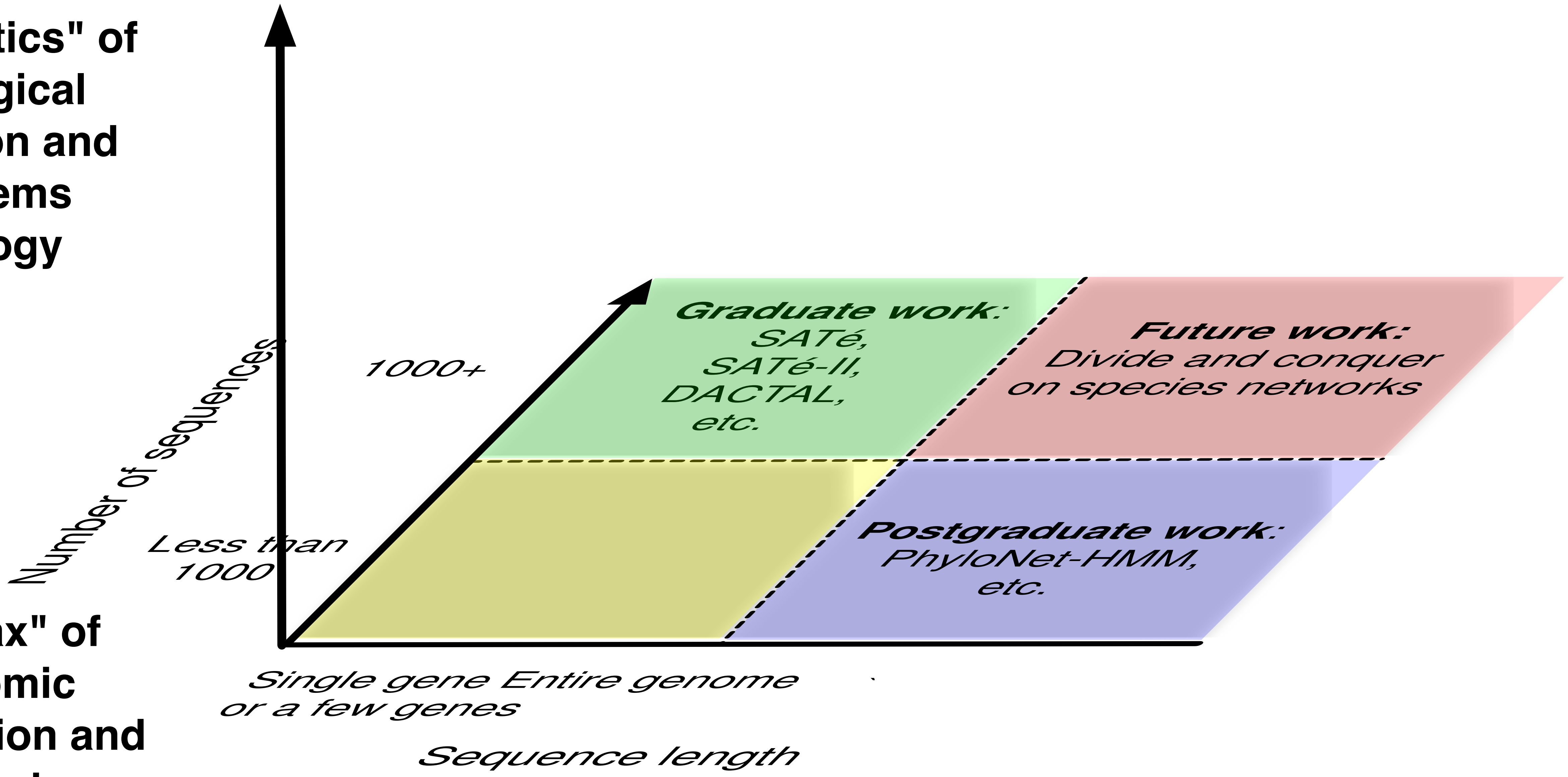
Future Direction #2



"Syntax" of Genomic Information and Architecture

Future Direction #2

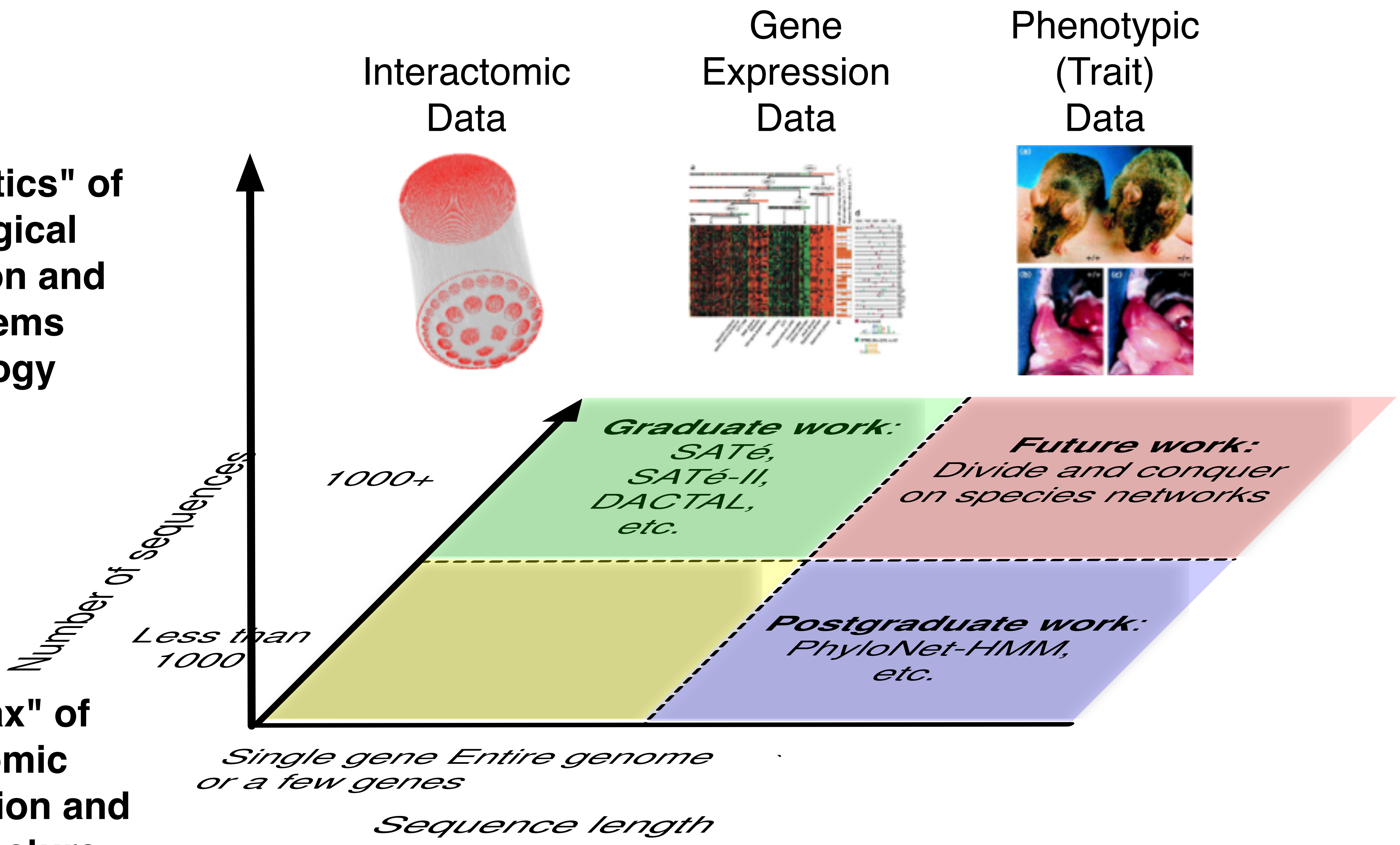
"Semantics" of Biological Function and Systems Biology



"Syntax" of Genomic Information and Architecture

Future Direction #2

"Semantics" of Biological Function and Systems Biology



"Syntax" of Genomic Information and Architecture

Funding Opportunities for My Work

- Computational approaches constitute basic research of interest to NSF (IIS, ABI)
- Wide range of applications of interest to different funding agencies, including:
    - The role of introgression in the spread of pesticide resistance in wild mice, with applications to personalized warfarin therapy (NIH)
    - The role of horizontal gene transfer in the spread of antibiotic resistance in bacteria (NIH)
    - Bacterial genomics (DOE)
    - Hybridization in plants (USDA)

Summary

- I have created:
  - new iterative divide-and-conquer techniques, which were used to develop methods for fast and accurate inference of alignments and trees from large-scale data sets, and
  - PhyloNet-HMM, a new inference method utilizing a DAG-based stochastic model, which is capable of disentangling “vertical” and “horizontal” evolution.
- My future research directions include:
  - developing divide-and-conquer methods for fast and accurate analysis of non-tree-like evolution using large-scale genomic data sets, and
  - synthesizing evolutionary analysis with interactomic and other functional analyses.

Acknowledgments

RICE

THE UNIVERSITY OF TEXAS AT AUSTIN

Luay Nakhleh
CS

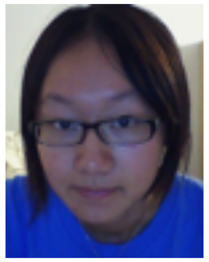
Michael Kohn
Biology

Tandy Warnow
CS (UIUC)

Yun Yu
CS

Ying Song
Biology

C. Randal Linder
Biology



Eileen Dai
CS

Kathy Truong
CS

Serita Nelesen
CS (Calvin College)

- Supported in part by:
  - A training fellowship from the Keck Center of the Gulf Coast Consortia, on Rice University's NLM Training Program in Biomedical Informatics (Grant No. T15LM007093).
  - NLM (Grant No. R01LM00949405 to Luay Nakhleh)
  - NHLBI (Grant No. R01HL09100704 to Michael Kohn)

Questions?

- My website:
  http://www.cs.rice.edu/~kl23

- Nakhleh lab website:
  http://bioinfo.cs.rice.edu

- Warnow lab website:
  http://www.cs.utexas.edu/~phylo