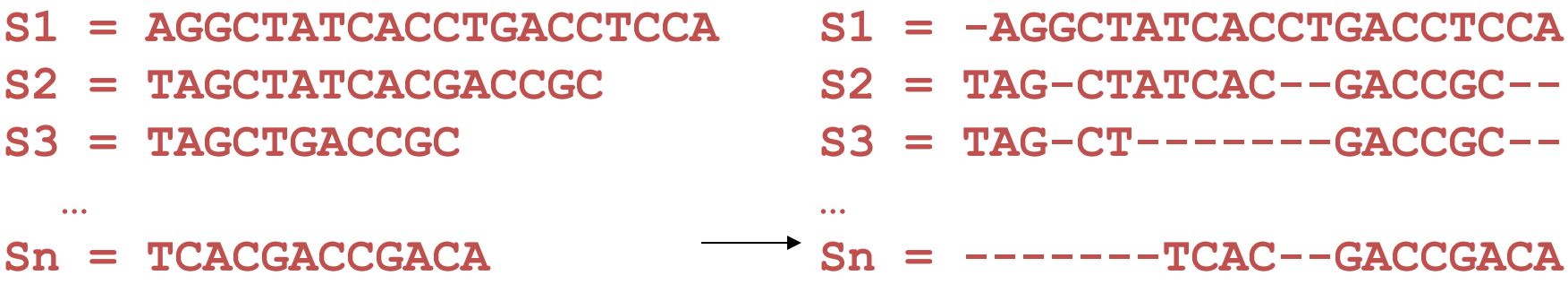
IPAM Workshop on Multiple Sequence Alignment

Tandy Warnow
The University of Illinois
at Urbana-Champaign

Multiple Sequence Alignment (MSA): *a major grand challenge*[1]



Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets

Current methods do not provide good accuracy

Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

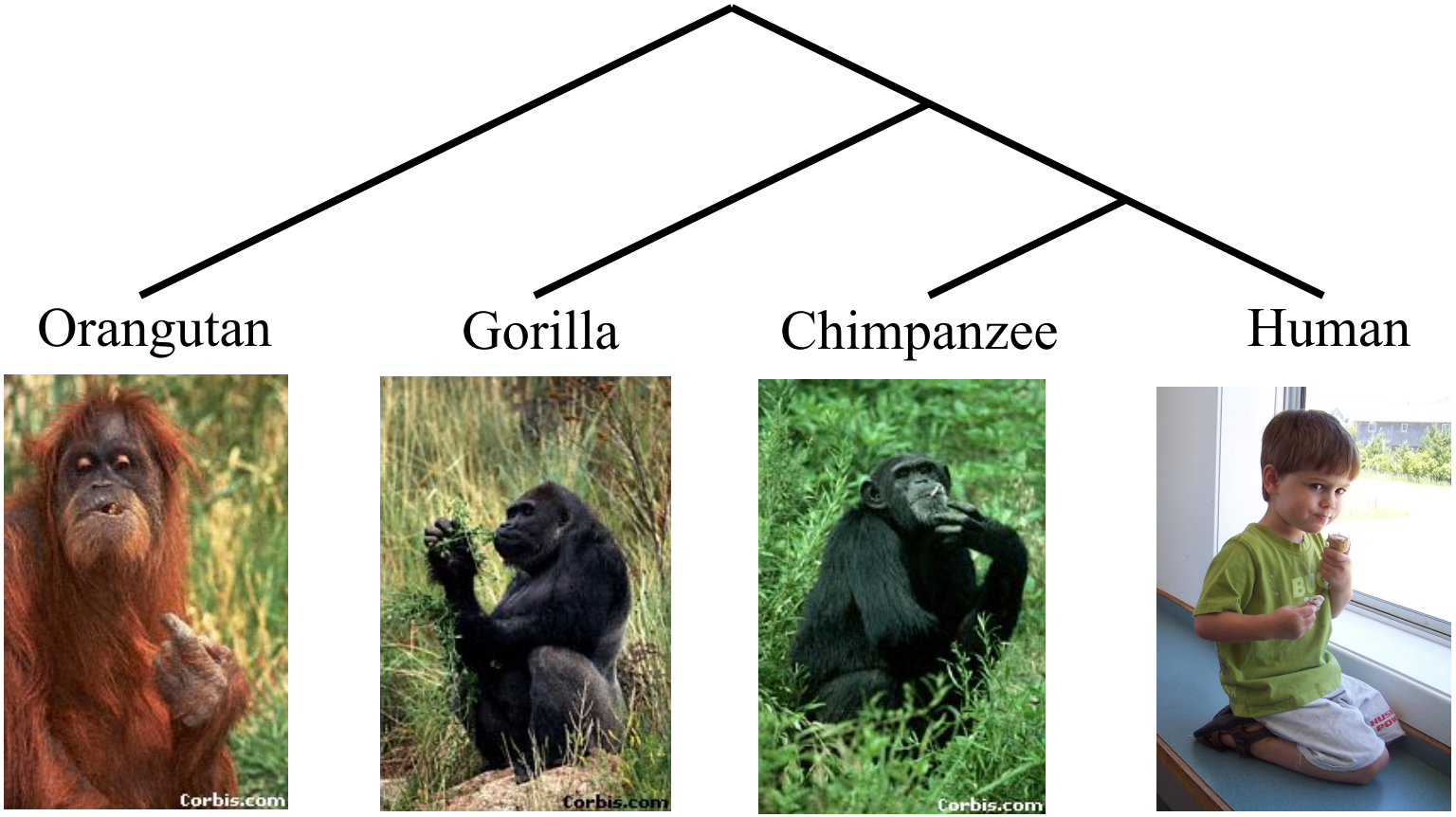
Multiple Sequence Alignment

- Multiple purposes (e.g., phylogeny estimation and molecular function/structure prediction)
- Multiple techniques, drawing from disparate communities (biophysics, statistical inference, computer science, discrete mathematics, etc.)
- Multi-disciplinary effort and communication needed

Multiple Sequence Alignment

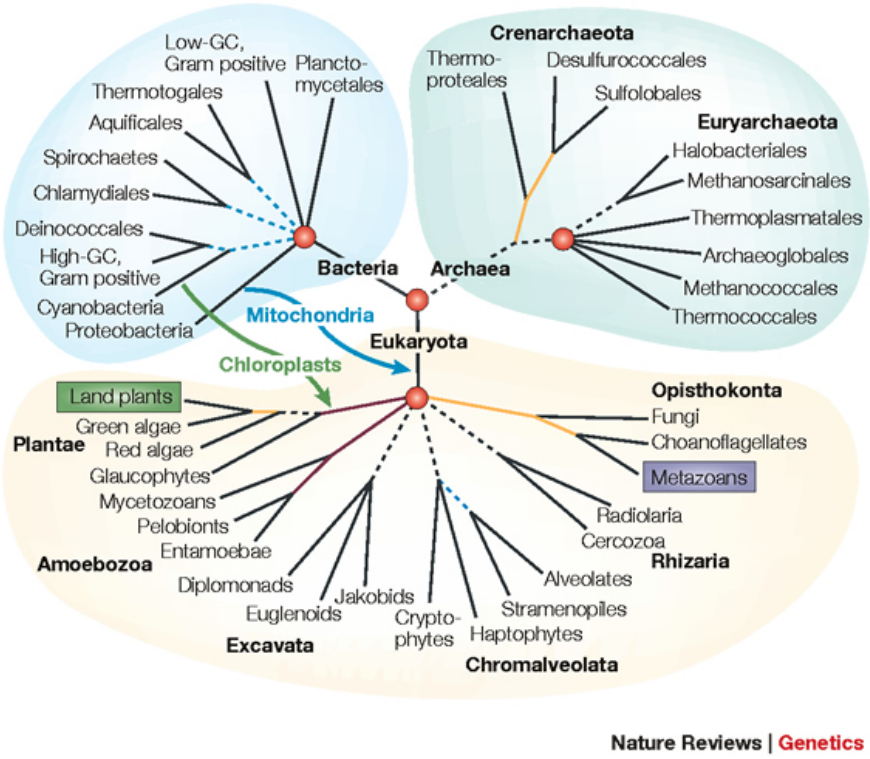
- Multiple purposes (e.g., phylogeny estimation and molecular function/structure prediction)
- Multiple techniques, drawing from disparate communities (biophysics, statistical inference, computer science, discrete mathematics, etc.)
- Multi-disciplinary effort and communication needed

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

Constructing the Tree of Life: Hard Computational Problems



NP-hard problems

Large datasets

- 100,000+ sequences
- thousands of genes

"Big data" complexity:

- model misspecification
- fragmentary sequences
- errors in input data
- streaming data

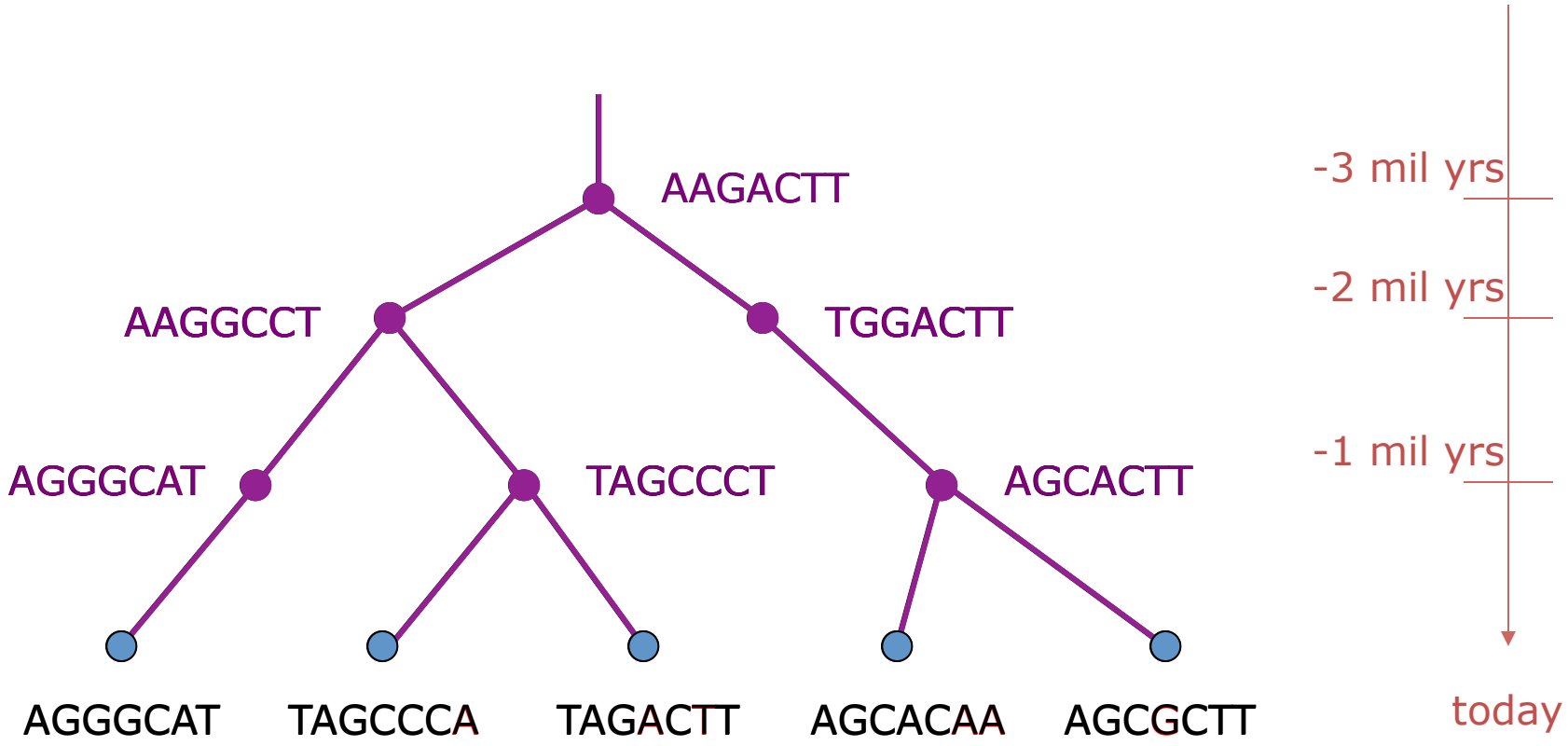
Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology

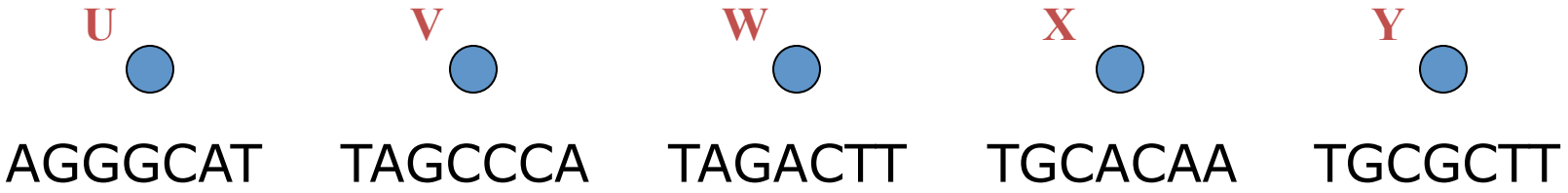
Phylogenomic pipeline

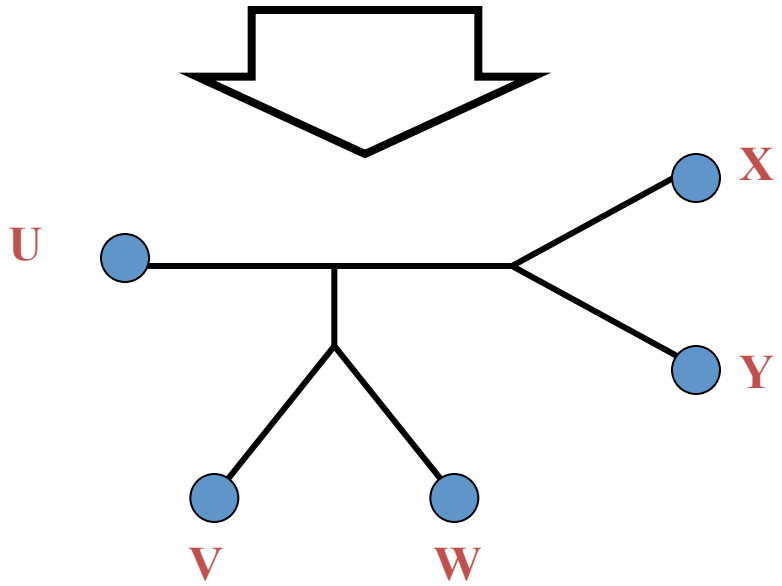
- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology

DNA Sequence Evolution



Phylogeny Problem





Performance criteria

- Running time
- Space
- Statistical performance issues (e.g., statistical consistency) with respect to a Markov model of evolution
- “Topological accuracy” with respect to the underlying ***true tree or true alignment,*** typically studied in simulation
- Accuracy with respect to a particular criterion (e.g. maximum likelihood score), on real data

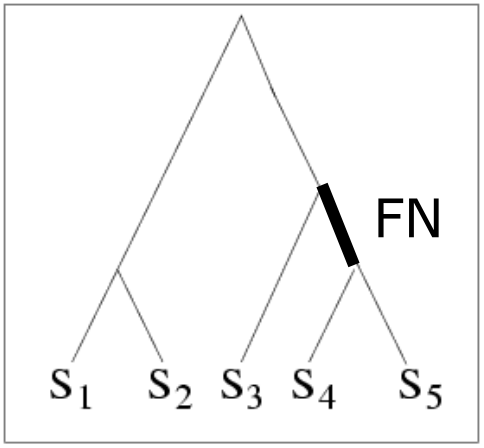
Markov models of site evolution

Simplest (Jukes-Cantor):

- The model tree is a pair (T,{e,p(e)}), where T is a rooted binary tree, and p(e) is the probability of a substitution on the edge e
- The state at the root is random
- If a site changes on an edge, it changes with equal probability to each of the remaining states
- The evolutionary process is Markovian

More complex models (such as the General Markov model) are also considered, with little change to the theory.

Quantifying Error



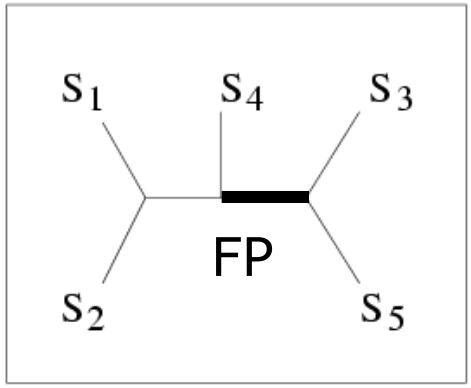
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

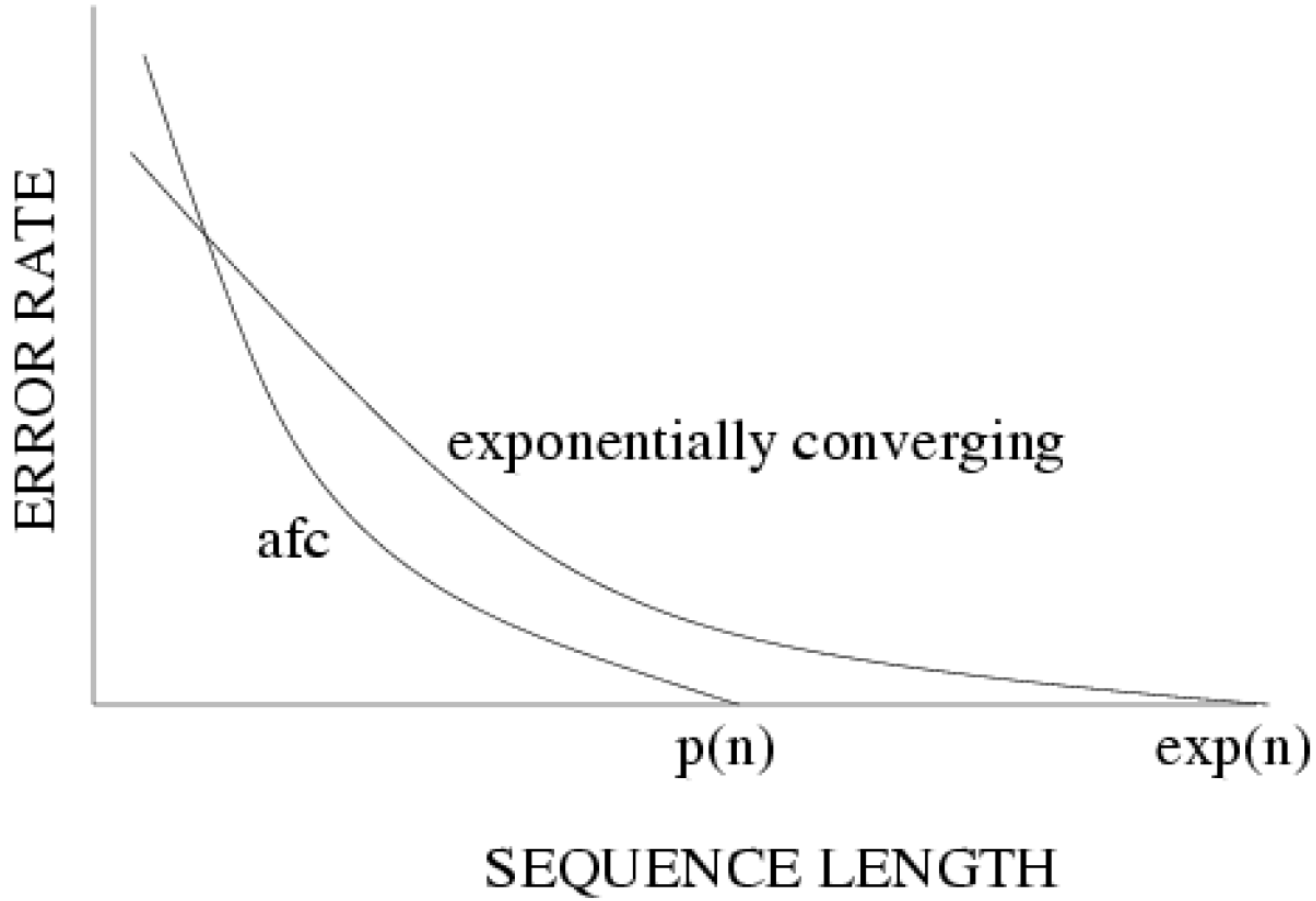
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



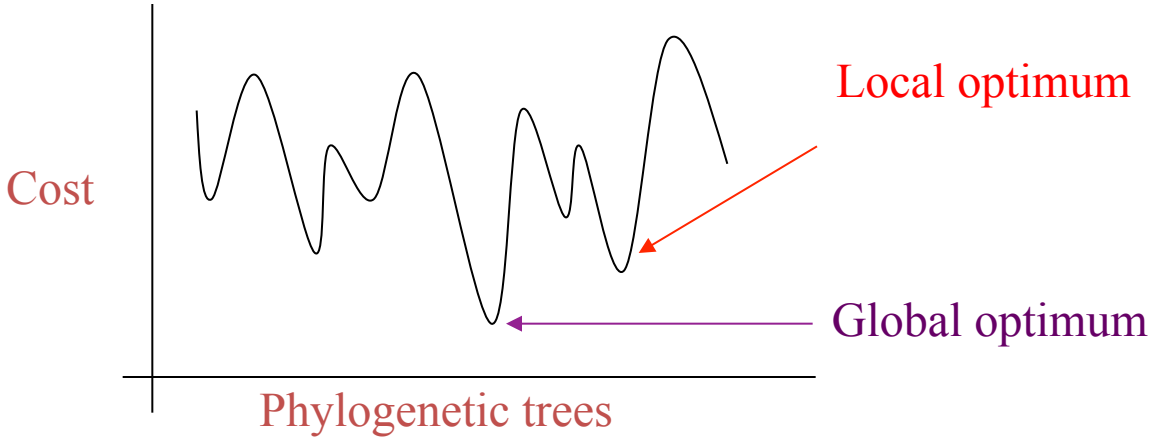
INFERRED TREE

Statistical consistency, exponential convergence, and absolute fast convergence (afc)



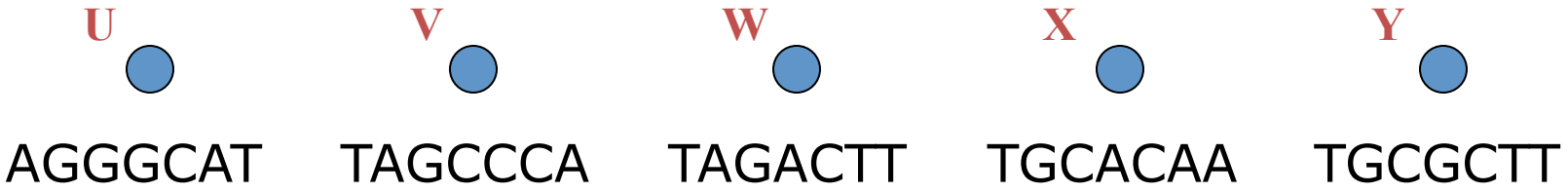
Phylogenetic reconstruction methods

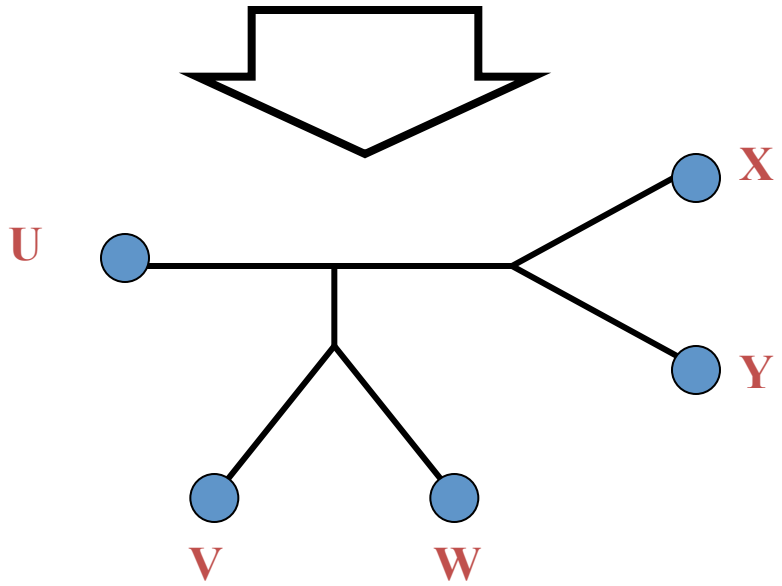
1. Hill-climbing heuristics for hard optimization criteria (Maximum Parsimony and Maximum Likelihood)



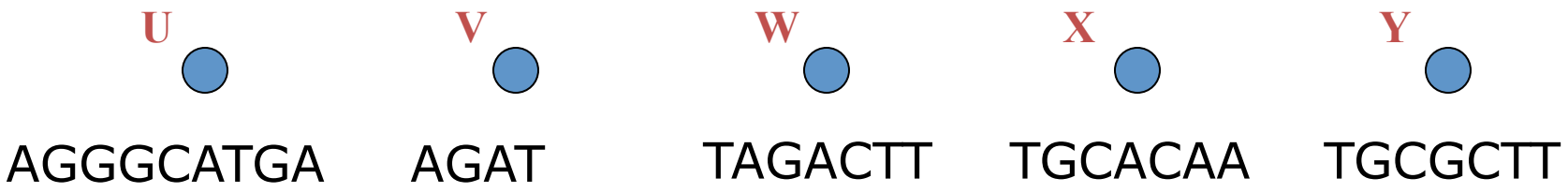
2. Polynomial time distance-based methods: Neighbor Joining, FastME, Weighbor, etc.
3. Bayesian methods

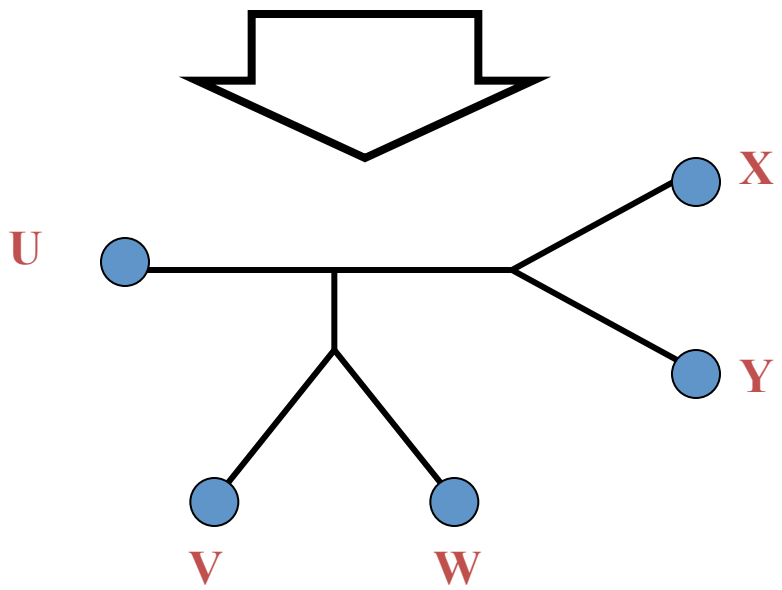
Phylogeny Problem



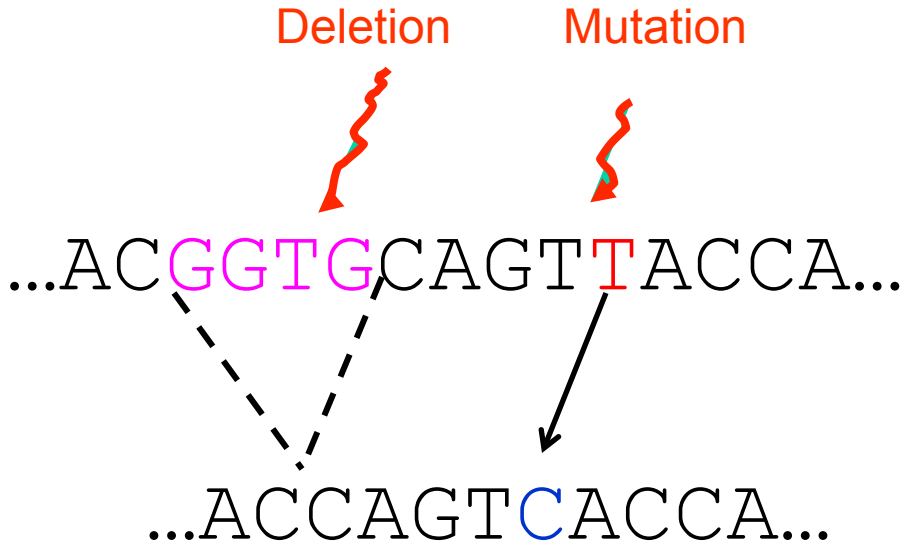


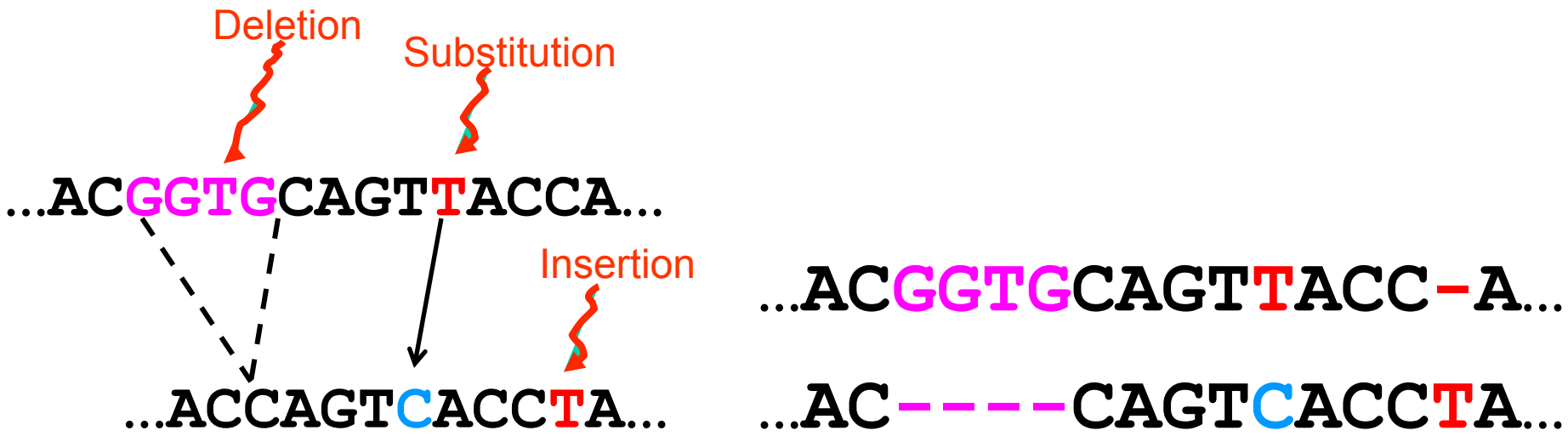
The “real” problem





Indels (insertions and deletions)





The true multiple alignment

- **Reflects historical substitution, insertion, and deletion events**
- **Defined using transitive closure of pairwise alignments computed on edges of the true tree**

Input: unaligned sequences

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

Phase 1: Alignment

```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



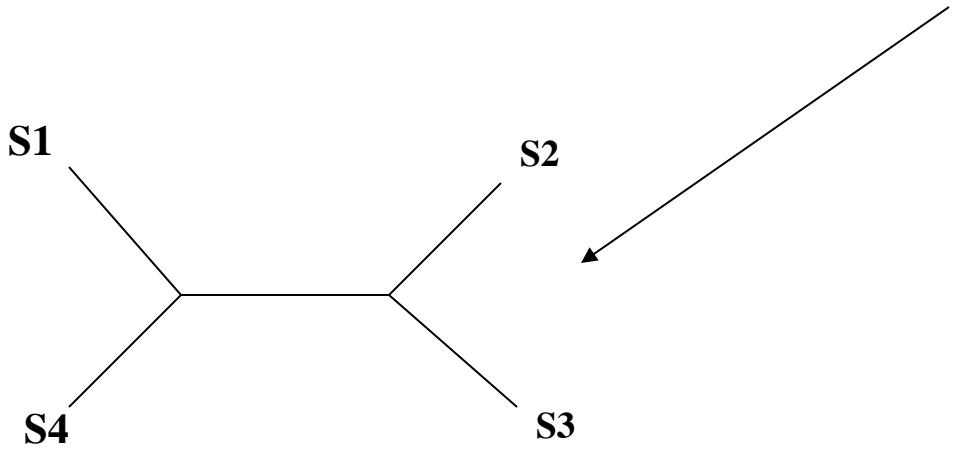
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct tree

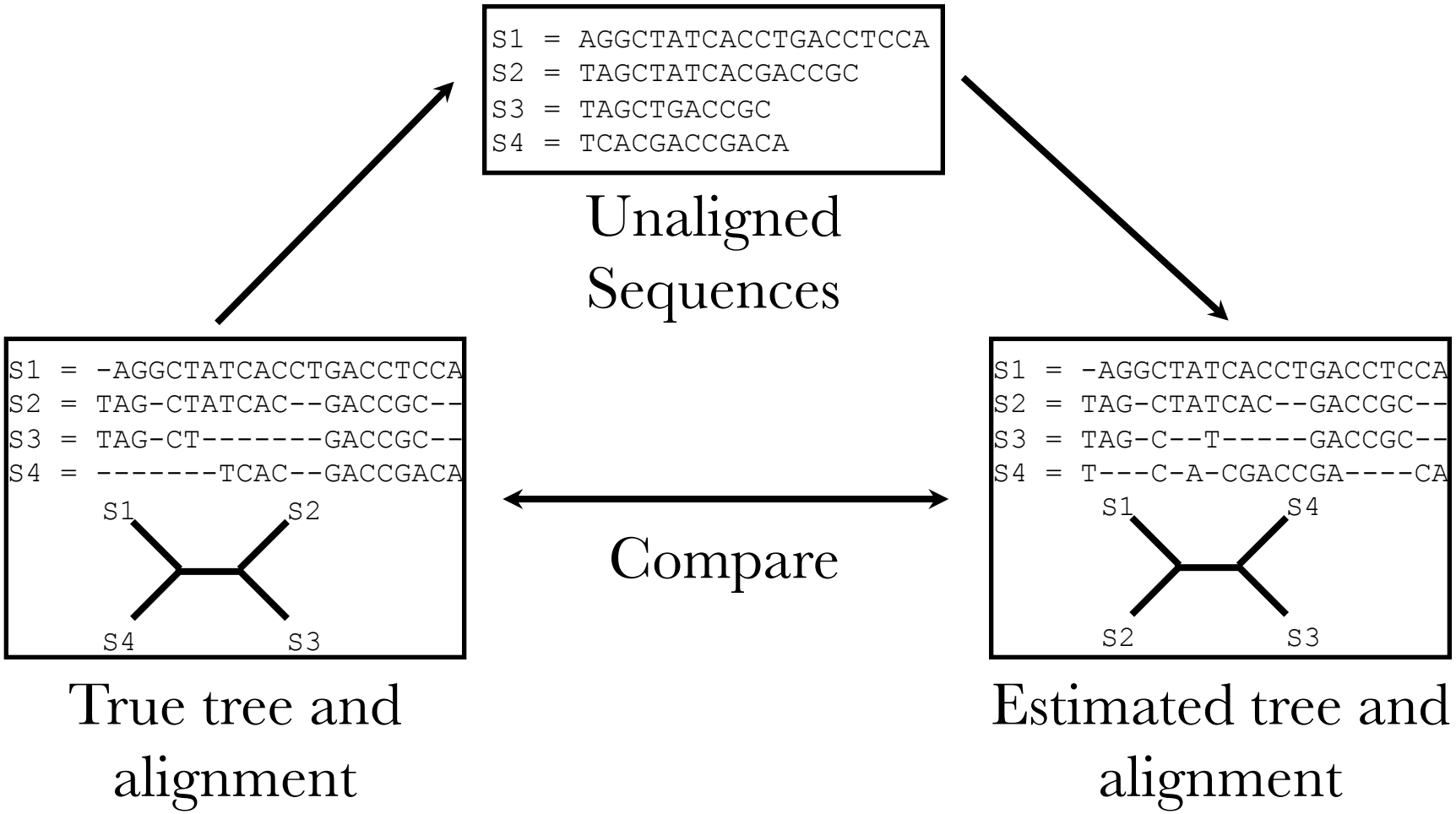
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

→

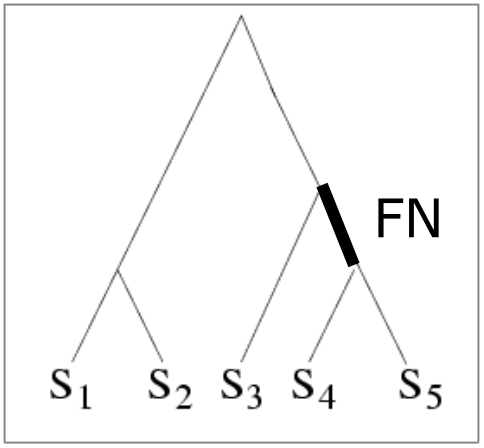
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Simulation Studies



Quantifying Error



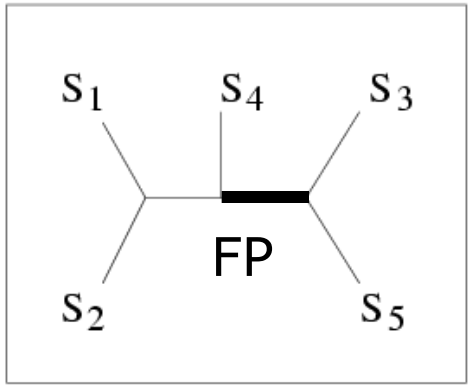
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

Two-phase estimation

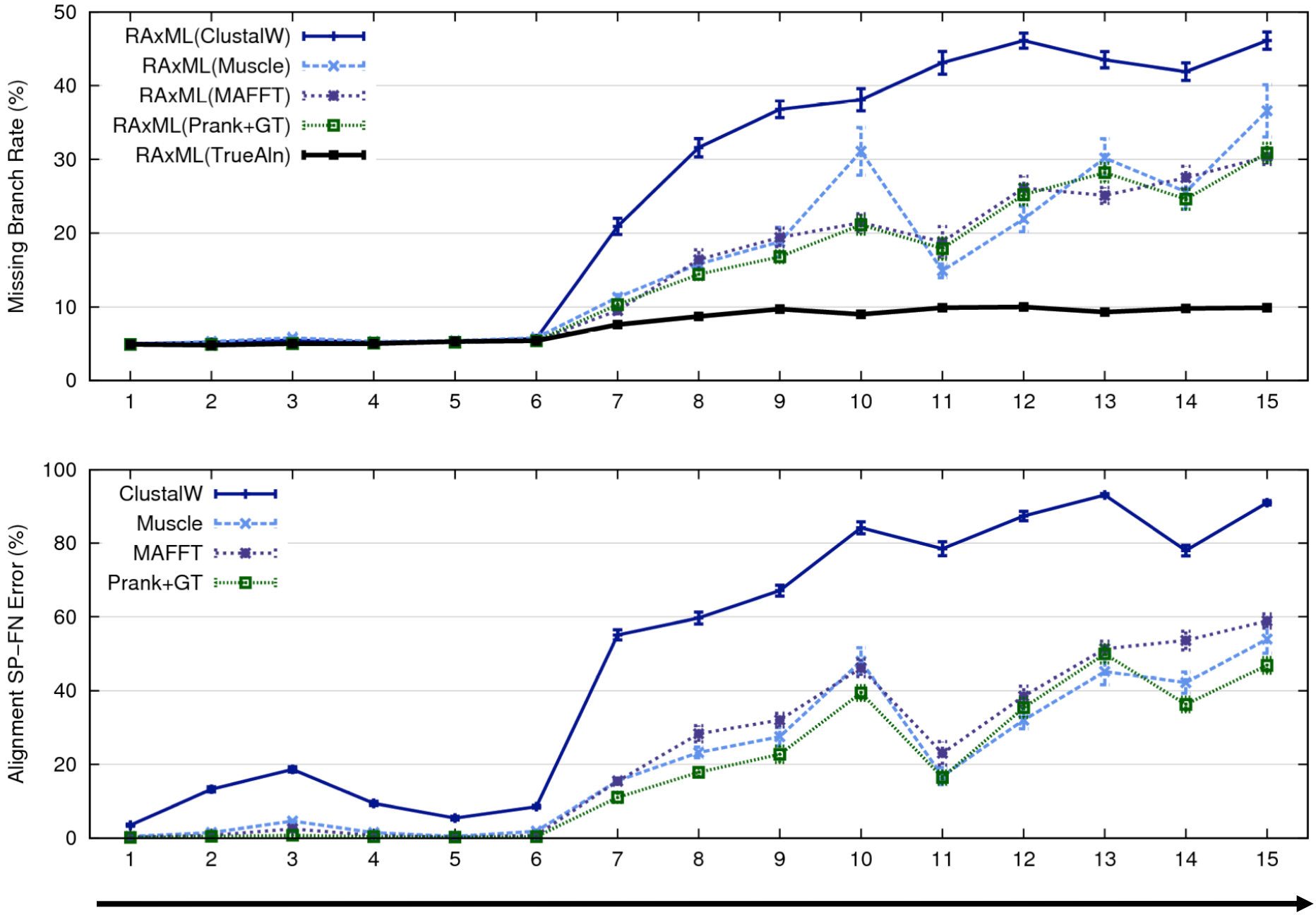
Alignment methods

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

Phylogeny methods

- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

***RAxML**: heuristic for large-scale ML optimization*


Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000-taxon models, ordered by difficulty (From Liu et al., Science 2009)

Large-scale Alignment Estimation

- Alignments of large datasets with high rates of evolution typically have high error, and trees estimated on these alignments also have high error
- Only a few methods can analyze large datasets

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UIUC

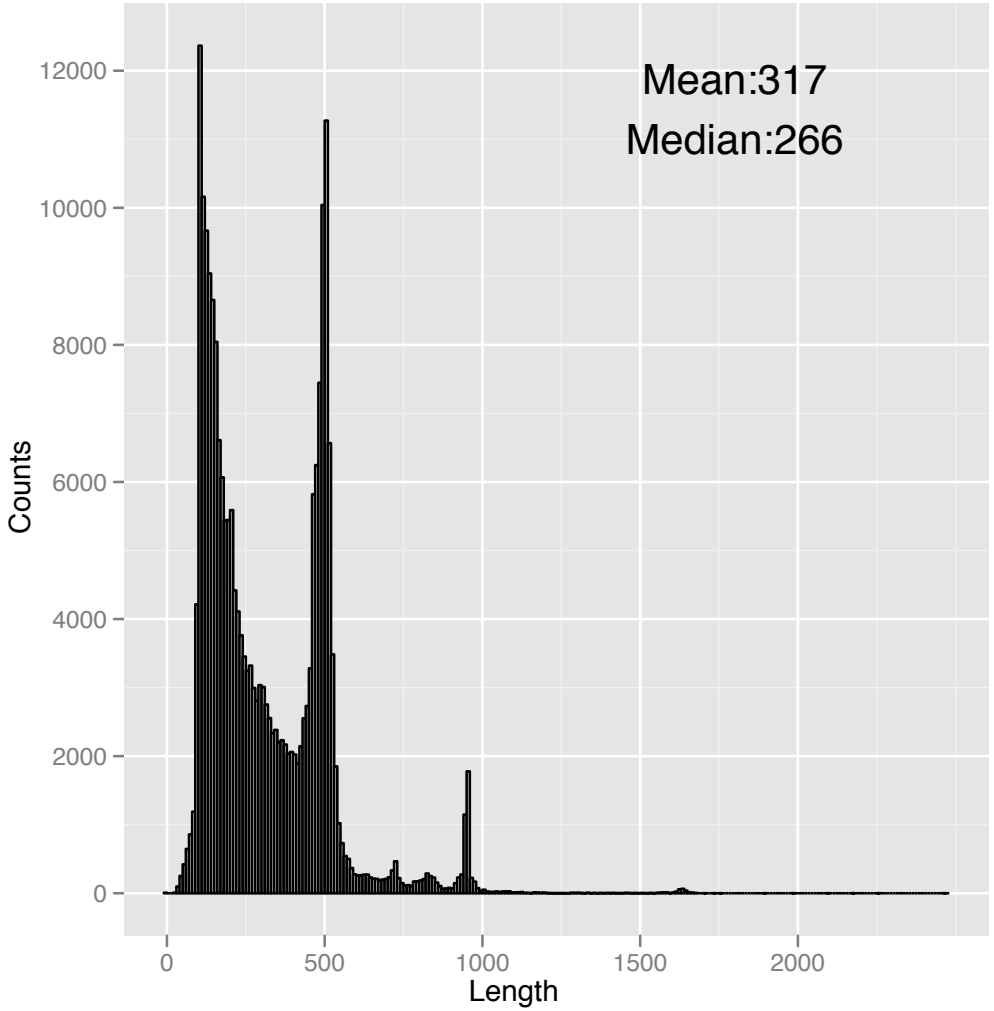
S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Plus many many other people…

- First study (Wickett, Mirarab, et al., PNAS 2014) had ~100 species and ~800 genes, gene trees and alignments estimated using SATe, and a coalescent-based species tree estimated using ASTRAL
- Second study: Plant Tree of Life based on transcriptomes of ~1200 species, and more than 13,000 gene families (most not single copy)

Upcoming Challenges:
Species tree estimation from conflicting gene trees
Alignment of datasets with > 100,000 sequences



1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary

All standard multiple sequence alignment methods we tested performed poorly on datasets with fragments.

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UIUC

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Plus many many other people…

- First study (Wickett, Mirarab, et al., PNAS 2014) had ~100 species and ~800 genes, gene trees and alignments estimated using SATe, and a coalescent-based species tree estimated using ASTRAL
- Second study: Plant Tree of Life based on transcriptomes of ~1200 species, and more than 13,000 gene families (most not single copy)

Upcoming Challenges:
Species tree estimation from conflicting gene trees
Alignment of datasets with > 100,000 sequences, and many fragmentary sequences!

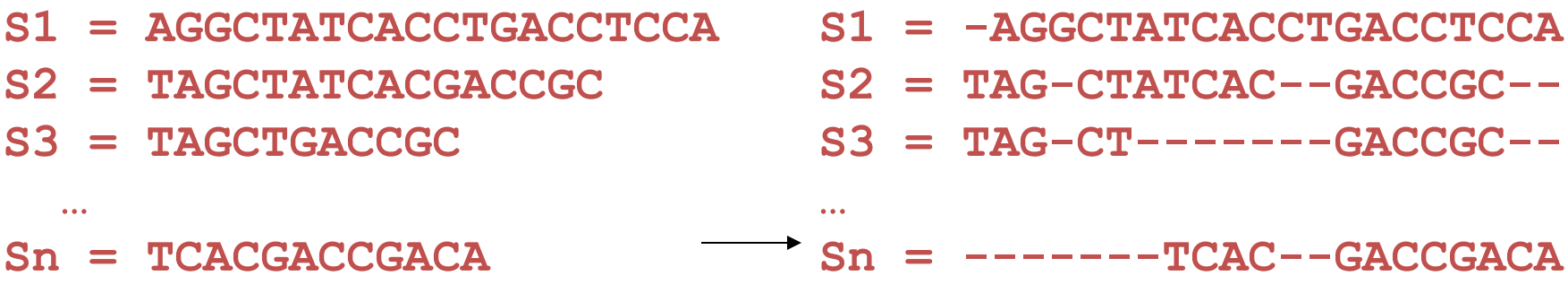
Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology

Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology
- Use multiple sequence alignment to understand biology

Multiple Sequence Alignment (MSA): *another grand challenge*[1]



Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets

Current methods do not provide good accuracy

Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

Research Questions

- What are good statistical models of sequence evolution that include insertions, deletions, and other events (rearrangements, duplications, etc.)? Are the model trees identifiable under these models?
- Can we co-estimate sequence alignments and trees with high accuracy?
- Can we improve alignments?
- Can we do large-scale alignment estimation with high accuracy?
- Can we do alignment-free phylogeny estimation?
- How should we measure alignment accuracy? Are common ways of measuring alignment accuracy predictive of tree accuracy?
- Are alignments for the purpose of structure/function prediction the same as alignments for phylogeny estimation?

Research Questions

- What are good statistical models of sequence evolution that include insertions, deletions, and other events (rearrangements, duplications, etc.)? Are the model trees identifiable under these models?
- Can we co-estimate sequence alignments and trees with high accuracy?
- Can we improve alignments?
- Can we do large-scale alignment estimation with high accuracy?
- Can we do alignment-free phylogeny estimation?
- How should we measure alignment accuracy? Are common ways of measuring alignment accuracy predictive of tree accuracy?
- Are alignments for the purpose of structure/function prediction the same as alignments for phylogeny estimation?
- What is the impact of alignment error on downstream biological analyses?

Patsy Babbit

Back to the Beginning: Which sequences to align?

- While much attention has been directed at mathematical and statistical issues for creating accurate multiple alignments, consideration of which sequences (or parts of sequences) and structures to align is less well explored. This issue is especially important for investigation of structure-function relationships in large sets of highly diverse homologs for which the proteins of unknown function are far greater than those that have been biochemically or structurally characterized.
- We discuss what we have learned about choosing representative sequences for creating MSAs from studies of several large and functionally diverse enzyme superfamilies and provide examples for how biologically informed questions can be framed using this context.
- Sequence similarity networks built to summarize on a large scale relationships among members of several of these superfamilies are used to illustrate new challenges for creating MSAs as the volume of sequence data continues to increase.

Alexandre Bouchard-Cote

MSA using Divide-and-Conquer Sequential Monte Carlo

- Divide-and-Conquer Sequential Monte Carlo (D&C SMC), a method for performing inference on a collection of auxiliary distributions organized into a tree.
- D&C SMC provides a simple method for approximating the posterior distribution of Bayesian MSA models

Noah Daniels

Structure-based multiple sequence alignments

- There exist several approaches to improving protein sequence alignment algorithms using structural information. However, how to best balance sequence alignment quality with structural alignment quality has not been clear.
- We will first demonstrate how sequence information can improve structural alignments, and then explore how advances in homology detection such as Markov random field approaches can improve sequence alignments.
- Finally, we will discuss how we might re-evaluate the tradeoffs between sequence and structural alignment quality when they are in disagreement.

Steve Evans

Recovering a tree from the lengths of random subtrees

- Suppose that we sample the leaves of edge-weighted tree with n leaves in a uniform random order and record the lengths of the subtrees spanned by the first k leaves (that is, in biological terms, the phylogenetic diversity of the first k taxa) for k between 2 and n.
- “Can we reconstruct the tree (up to isomorphism) from the joint probability distribution of this random increasing sequence of lengths?"
- The answer is affirmative if we know *a priori* that the tree belongs to one of a number of families, but the general question is still open.

Adam Godzik

Analysis and multiple alignments of periodic proteins

- Periodic proteins, characterized by the presence of multiple repeats of short sequence motifs, present unique challenges in their analysis and alignments. Especially interesting class of periodic proteins are irregular periodic proteins, where individual repeats can vary in length, sometimes considerably.
- We have developed a series of tools to classify regularity (or irregularity) of periodic proteins and to use such irregularity patterns to guide the multiple alignments.
- We present application of these tools to a group of Leucine Rich Repeat (LRR) proteins.

Nick Grishin

Pushing the limits of sequence profile similarity search and alignment

- Traditionally, sequence similarity search has been riding on negatives and deriving power from random models to be rejected for positive hits.
- Exploring the other side, we show that the search can be significantly improved by considering the positives, i.e., known homology relationships in a database of sequence profiles.
- Similar strategies have been widely used by most successful search engines, such as Google. This algorithm results in re-ranking of hits, but does not correct faulty alignments.
- The main focus in the sequence alignment field has been on alignment construction. However, many alignments are reasonably accurate with the exception of several mildly misaligned regions.
- We propose new approaches to refinement of existing alignments and show that successful a posteriori detection and correction of misaligned regions results in alignment improvement.

Jim Leebens-Mack

Challenges in plant phylogenomics

- Gene and genome duplications are rampant in plant genomes.
- I will discuss the challenge this presents for gene family circumscription, multiple sequence alignment, gene tree estimation, ortholog identification, and species tree estimation.

Olivier Lichtarge

Evolution versus disease: the calculus of life

- The relationship between genotype mutations and phenotype variations determines health in the short term and evolution over the long term, and it hinges on the action of mutations on fitness. A fundamental difficulty in determining this action, however, is that it depends on the unique context of each mutation, which is complex and often cryptic. As a result, the effect of most genome variations on molecular function and overall fitness remains unknown, and stands apart from population genetics theories linking fitness effect to polymorphism frequency.
- Here, we hypothesize that evolution is a continuous and differentiable physical process coupling genotype to phenotype.
- Thus elementary calculus and phylogenetics can be integrated into a perturbation analysis of the evolutionary relationship between genotype and phenotype that quantitatively links point mutations to function and fitness and that opens a new analytic framework for equations of biology. In practice, it explicitly bridges molecular evolution with population genetics with applications from protein redesign to the clinical assessment of human genetic variations.

Ari Loytyjoja

Phylogeny-aware alignment with sequence graphs

- PAGAN is a new program for phylogeny-aware multiple sequence alignment using partial-order sequence graphs, and Wasabi is a graphical front-end for phylogeny-aware alignment.
- In PAGAN, we use sequence graphs to model uncertainties in character presence/absence and thus make the phylogeny-aware algorithm less sensitive to errors in guide phylogeny or noisy input data.
- PAGAN can also extend existing alignments with new data: we have built applications of this for phylogenetic placement of marker gene data and for reference-based scaffolding of NGS data.

Cedric Notredame

Grabbing High Hanging Fruits From the Tree Of Life

- Results on simulated data suggest that some disagreements may exist between evolutionarily and structurally correct multiple sequence alignments
- I will introduce a recently developed confidence index for multiple sequence alignments, the TCS (Transitive consistency score) and show how this index can be used to both identify structurally correct positions in an alignment and evolutionary informative sites, thus suggesting more unity than initially thought between these two parameters.
- I will then introduce the structure based clustering method we recently developed to further test these hypothesis.

Jian Peng

Distances between protein sequence alignments

- We consider the problem on how to measure the similarity between sequence alignments.
- With good similarity metrics, we are able to use machine learning methods to learn more accurate alignment models than traditional ones, for structure prediction and/or homology search.

Mark Ragan

Phylogenetics without multiple sequence alignment

- Multiple alignment is computationally hard, and does not extend naturally to instances in which the sequences under consideration have been rearranged relative to each other, misassembled (or not assembled in the first place), or contain regions of lateral origin.
- I explore alternative approaches that begin with the extraction of short perfectly or near-perfectly matching character strings variously known as words, k-mers or n-grams.
- Using synthetic and empirical data I will survey the major alignment-free approaches in phylogenetics, consider their performance and robustness under various scenarios of sequence evolution, and comment on their computational scalability.

Benjamin Redelings

Erasing Errors Due to Alignment Ambiguity When Estimating Positive Selection

- Current estimates of diversifying positive selection rely on first having an accurate multiple sequence alignment. Simulation studies have shown that under biologically plausible conditions, relying on a single estimate of the alignment from commonly used alignment software can lead to unacceptably high false positive rates in detecting diversifying positive selection.
- We present a novel statistical method that eliminates excess false positives resulting from alignment error by jointly estimating the degree of positive selection and the alignment under an evolutionary model.
- We also show that samples taken from the posterior alignment distribution using the software BAli-Phy have substantially lower alignment error compared to MUSCLE, MAFFT, PRANK, and FSA alignments.

Sebastien Roch

A survey of theoretical results for the TKF91 model

- Thorne, Kishino, and Felsenstein (TFK91) developed a model of sequence evolution that included insertions and deletions in addition to substitutions.
- I will give a survey of theoretical results on the reconstruction of phylogenies under the TKF91 model. I will mostly discuss known consistency/ inconsistency results.

Scott Schmidler

The Cutoff Phenomenon in Evolutionary Models for Sequence Alignment

- We examine limits on inferring evolutionary divergence times using sequence evolution models arising as a consequence of the probabilistic "cutoff phenomenon", in which a Markov chain remains far from equilibrium for an extended period, followed by a rapid transition into equilibrium.
- We show that evolutionary sequence models exhibit a cutoff, which relates directly to increased uncertainty in evolutionary distance inferences.
- We derive the cutoff explicitly for symmetric models, and demonstrate empirically the behavior in models routinely used in the literature. We also show how to locate cutoffs for specific models and sequences.
- Finally, we show that the cutoff explains several previously reported problems with common default priors for Bayesian phylogenetic analysis, and we suggest a new class of priors to address these problems.

Martin Weigt

Coevolutionary modeling of protein sequences: Inference of 3D structure and mutational landscapes

- Direct-Coupling Analysis (DCA): a statistical-inference approach for detecting direct residue coevolution in large multiple-sequence alignments of homologous proteins.
- We will show how to predict tertiary and quaternary protein structures, reconstruct protein-protein interaction networks, and infer quantitative mutational landscapes.

Jinbo Xu

Graphical models of multiple protein sequence alignment

- This talk will present the modeling of multiple protein sequence alignment (MSA) by Markov Random Fields (MRF) and its applications to homology detection, evolutionary coupling analysis and protein folding.
- This talk will cover
  - modeling a set of related protein families (each represented as an MSA) by group graphical lasso and its application to joint evolutionary coupling analysis and protein contact prediction; and
  - aligning two MSAs by aligning their respective MRFs and its application to homology detection and fold recognition.

Research Questions

- What are good statistical models of sequence evolution that include insertions, deletions, and other events (rearrangements, duplications, etc.)? Are the model trees identifiable under these models?
- Can we co-estimate sequence alignments and trees with high accuracy?
- Can we improve alignments?
- Can we do large-scale alignment estimation with high accuracy?
- Can we do alignment-free phylogeny estimation?
- How should we measure alignment accuracy? Are common ways of measuring alignment accuracy predictive of tree accuracy?
- Are alignments for the purpose of structure/function prediction the same as alignments for phylogeny estimation?
- What is the impact of alignment error on downstream biological analyses?

Opportunities for presentations

- Today: 4-5 PM: Brief (3-5 minute) presentations by participants. (No need to request a slot…)

- Today: 5-6 PM: Poster session

- Tuesday and Thursday: 4-5 PM: short talks (approx. 15 minutes each) – *please see one of the organizers today to request to speak*.