Evaluating Methods

Tandy Warnow

You've designed a new method! Now what?

To evaluate a new method:

- Establish theoretical properties.
- Evaluate on data.
- Compare the new method to other methods.

How do you do this?

General Issues

- So far we have computed trees and we have computed alignments.
- How can we quantify accuracy or error? What datasets should we use?
- What are the issues?

Basic criteria

- Sensitivity = true positive rate = recall rate = TP/(TP+FN)
- Precision = positive predictive value = TP/(TP+FP)
- Specificity = true negative rate = TN/(TN+FP)
- False Discovery Rate = 1-PPV

True positives, false positives, etc.

- For these criteria, we need to understand the concepts of
  - true positive,
  - false positive,
  - true negative, and
  - false negative
- In other words, we need to have a “yes/no” classifier.

Simple example: HIV testing

- Sample space: HIV tests (Eliza)
  - True positive: the test comes out positive and the person does have HIV
  - True negative: the test comes out negative and the person does not have HIV
  - False positive: the test comes out positive but the person does not have HIV
  - False negative: the test comes out negative and the person does have HIV

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease
  - TP = 9, FP = 11, TN = 979, FN = 1

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease
  - TP = 9, FP = 11, TN = 979, FN = 1
  - Sensitivity = TP/(TP+FN) = 9/10 = 90%

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease
  - TP = 9, FP = 11, TN = 979, FN = 1
  - Sensitivity = TP/(TP+FN) = 9/10 = 90%
  - Specificity = TN/(TN+FP) = 979/990 = 98.9%

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease
  - TP = 9, FP = 11, TN = 979, FN = 1
  - Sensitivity = TP/(TP+FN) = 9/10 = 90%
  - Specificity = TN/(TN+FP) = 979/990 = 98.9%
  - Precision = TP/(TP+FP) = 9/20 = 45%

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease
  - What is the false positive rate?
  - What is the false negative rate?

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease
  - What is the false positive rate?
  - FP rate = # false positives divided by the number of total positives, so FP/(FP+TP) = 11/20 = 55%

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease
  - What is the false negative rate?

Hypothetical Example

- The population is 1,000 samples
- 10 of them have the disease, 990 do not
- The test is positive on 20: 9 of the 10 with the disease, and 11 of the 990 who do not have the disease
  - What is the false negative rate?
  - FN rate = # false negatives divided by the number of total negatives, so FN/(FN+TN) = 1/990 = 0.1%

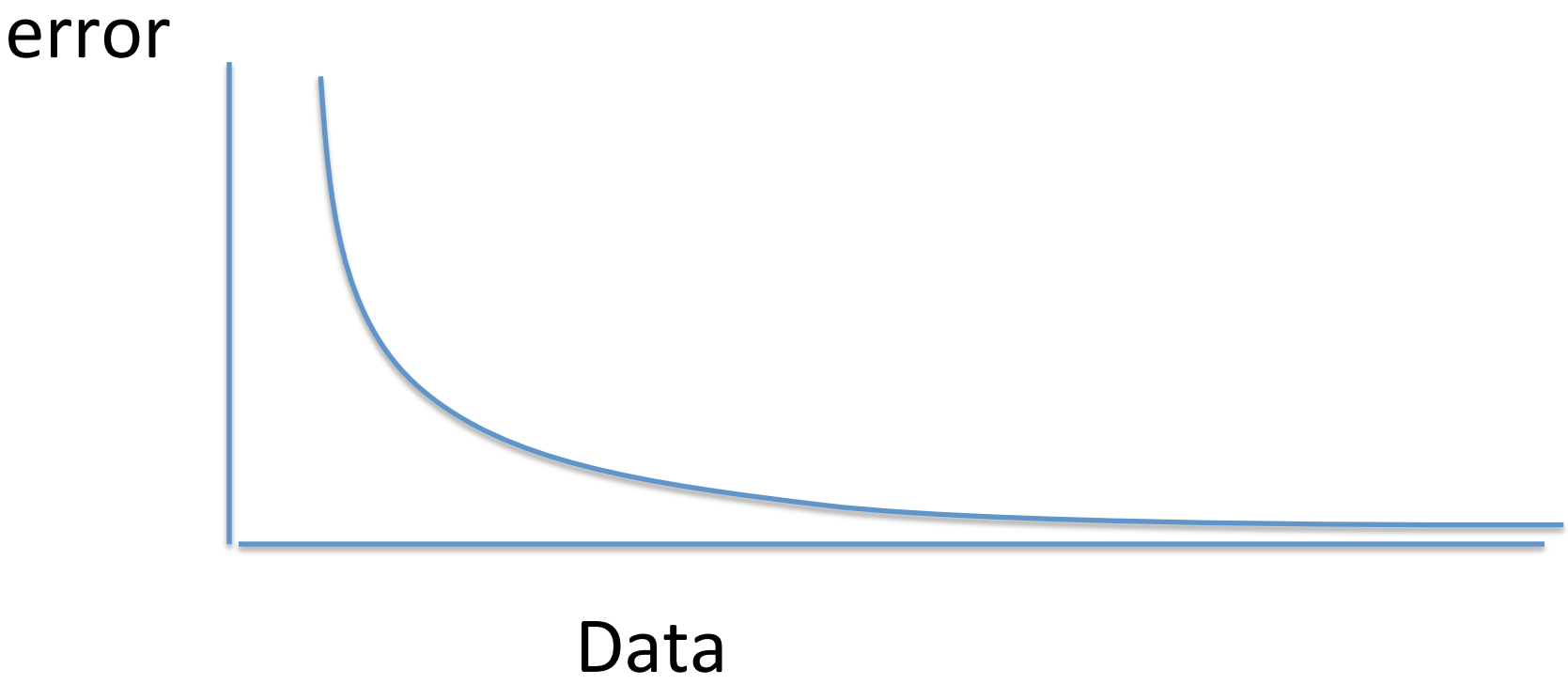
General Issues

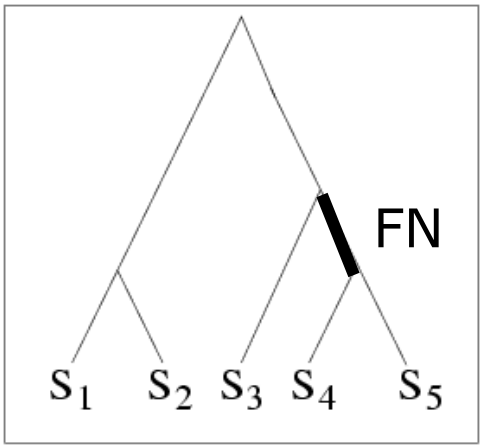
- So far we have computed trees and we have computed alignments.
- How can we quantify accuracy or error? What datasets should we use?
- What are the issues?

Performance criteria

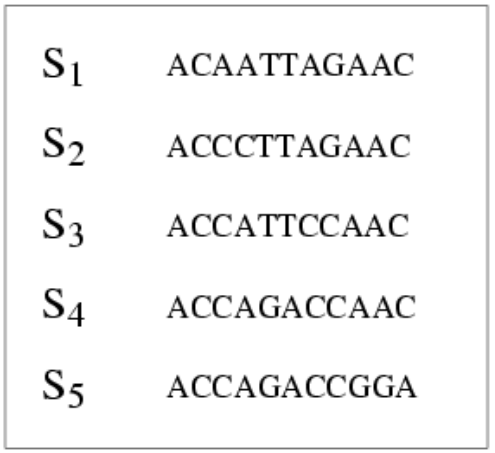
- Running time
- Space
- Statistical performance issues (e.g., statistical consistency and sequence length requirements)
- “Topological accuracy” with respect to the underlying true tree, typically studied in simulation.
- Accuracy with respect to a mathematical score (e.g. tree length or likelihood score) on real data

Statistical Consistency



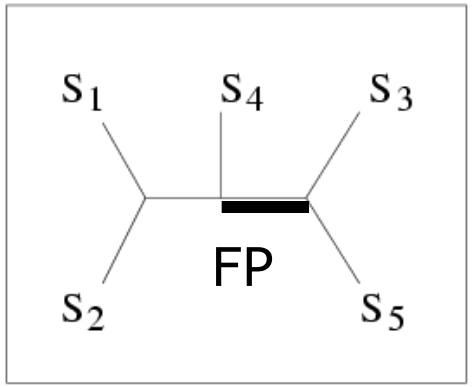


TRUE TREE



DNA SEQUENCES

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



INFERRED TREE

50% error rate

Alignment Error/Accuracy

- SPFN: percentage of homologies in the true alignment that are *not* recovered (false negative homologies)
- SPFP: percentage of homologies in the estimated alignment that are false (false positive homologies)

- TC: total number of columns correctly recovered
- SP-score: percentage of homologies in the true alignment that are recovered
- Pairs score: 1-(avg of SP-FN and SP-FP)

Other Alignment Estimation Criteria

- Tree topology error
- Tree branch length error

- Gap length distribution
- Insertion/deletion ratio
- Alignment length
- Number of indels

Studying Methods

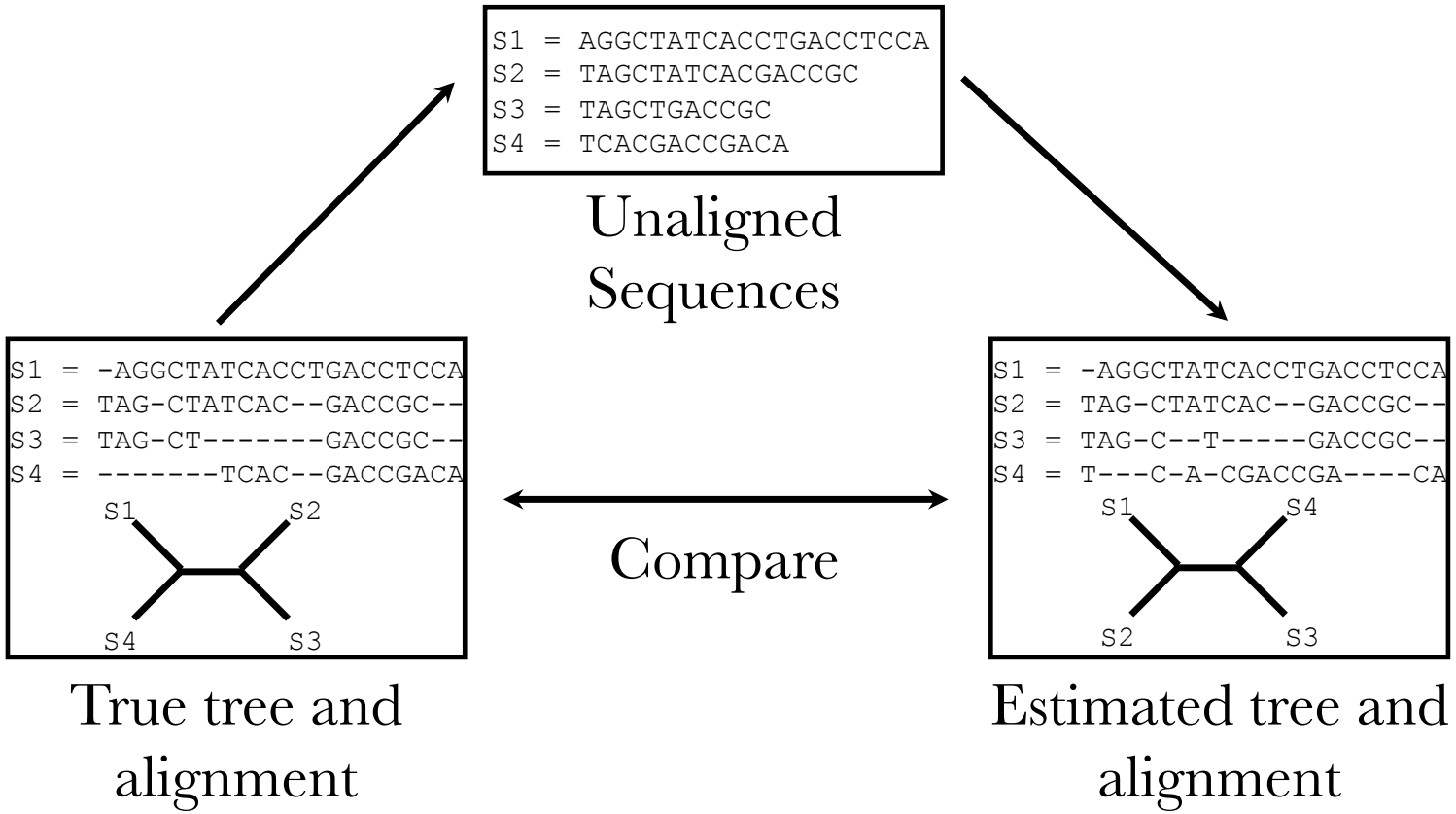
- The point is to evaluate a new method in comparison to prior methods.
- You need to do this on data, not just using theorems.
- How do you do this?

Benchmarks

- Simulations: can control everything, and true alignment is not disputed
  - Different simulators

- Biological: can't control anything, and reference alignment and reference tree might not be correct. Alignment benchmarks are also somewhat problematic, for various reasons:
  - BAliBASE, HomFam, Prefab
  - CRW (Comparative Ribosomal Website)

Simulation Studies



Designing a simulation study

- Consider the realism of the simulator.
- Consider whether the conditions are too easy or too difficult to be helpful.
- Consider the competing methods to explore.
- Consider statistical significance.
- Be concerned with repeatability.

Data

- Biological data:
  - How reliable are the reference alignments and trees?
- Simulated data:
  - How realistic are the simulation conditions?

Simulators

- Sequence evolution down a tree:
  - Indels? If so, what lengths?
  - Substitutions under what model?
  - How many substitutions? How many indels?
  - How is the tree topology and set of branch lengths defined?
  - Is the tree ultrametric?
  - How many leaves in the tree (i.e., # sequences)?
  - How long are the sequences?

Methods

- Are you picking the best competing methods?
- Are you running them in the best way?

Criteria

- Are you using criteria that are considered appropriate by the research community?
- If you are using new criteria, justify these criteria (and probably use the standard criteria anyway).

Repeatability

- Provide full details about how you ran the analyses so that the same experiment could be done by the person reading the paper.
- Save your data and make them available to the readers.

Writing Papers

Read

- Appendix C in Computational Phylogenetics for guidelines about writing papers about computational methods.
- “How to write your first paper” – on my homepage
- “Commonly encountered challenges in research ethics” – on my homepage