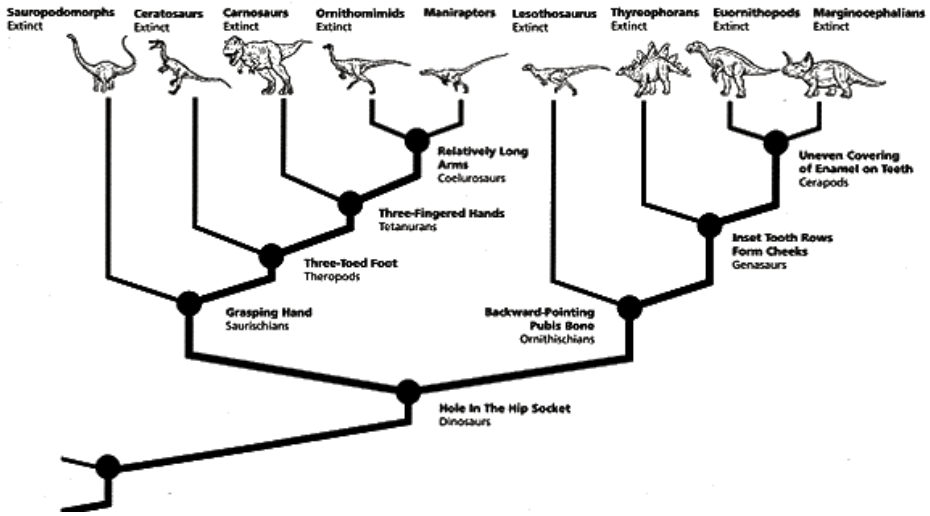
Introduction to Phylogeny

Lecture Notes for 9 October 2003
CMP 464/788: Introduction to Computational Biology

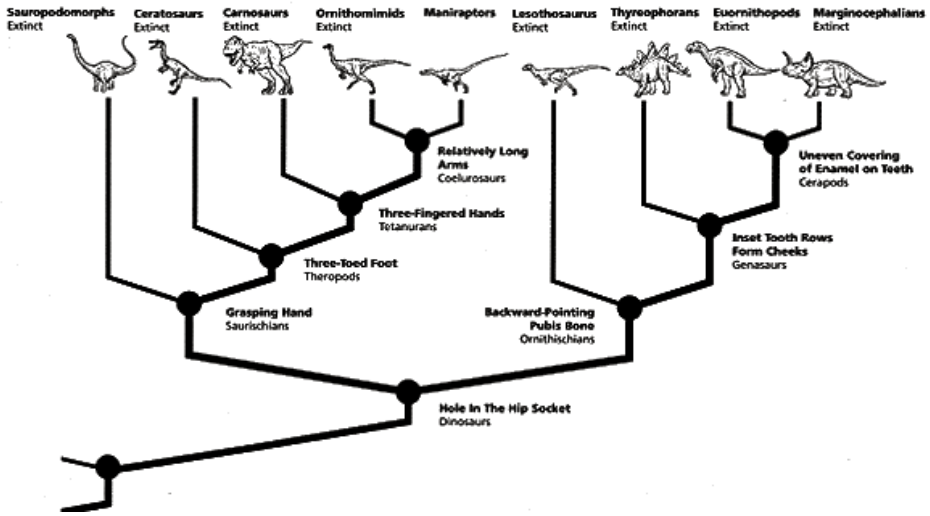
Prof. St. John
Lehman College
City University of New York

Goal: Reconstruct the Evolutionary History



(www.amnh.org/education/teacherguides/dinosaurs)

Goal: Reconstruct the Evolutionary History



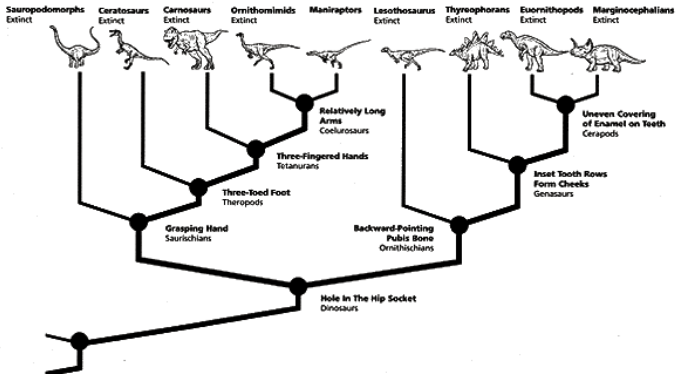
(www.amnh.org/education/teacherguides/dinosaurs)

The evolutionary process not only determines relationships among taxa, but allows prediction of structural, physiological, and biochemical properties.

Process for Reconstruction: Input Data

Start with information about the taxa. For example:

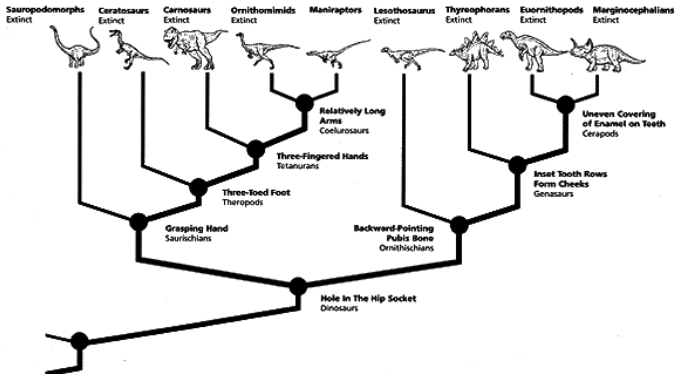
Morphological
Characters



Process for Reconstruction: Input Data

Start with information about the taxa. For example:

Morphological Characters



Biomolecular Sequences

A	GTTAGAAGGCGGCCAGCGAC. . .
B	CATTTGTCCTAACTTGACGG. . .
C	CAAGAGGCCACTGCAGAATC. . .
D	CCGACTTCCAACCTCATGCG. . .
E	ATGGGGCACGATGGATATCG. . .
F	TACAAATACGCGCAAGTTCG. . .

Process for Reconstruction

Process for Reconstruction

Input

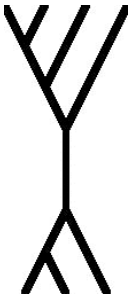
Data

A	GTTAGAAGGC...
B	CATTTGTCCT...
C	CAAGAGGCCA...
D	CCGACTTCCA...
E	ATGGGGCACG...
F	TACAAATACG...

Process for Reconstruction

Input Data			Reconstruction Algorithms
A	GTTAGAAGGC...	$\longrightarrow$	Maximum Parsimony
B	CATTTGTCCT...		Maximum Likelihood
C	CAAGAGGCCA...		Distance Methods: NJ,
D	CCGACTTCCA...		Quartet-Based,
E	ATGGGGCACG...		Fast Convering,
F	TACAAATACG...		⋮

Process for Reconstruction

Input Data		Reconstruction Algorithms	Output Tree
A	GTTAGAAGGC...	Maximum Parsimony	
B	CATTTGTCCT...	Maximum Likelihood	
C	CAAGAGGCCA...	Distance Methods: NJ,	
D	CCGACTTCCA...	Quartet-Based,	
E	ATGGGGCACG...	Fast Convering,	
F	TACAAATACG...	⋮	

Applications

In addition to finding the evolutionary history of species, phylogeny is also used for:

Applications

In addition to finding the evolutionary history of species, phylogeny is also used for:

- drug discovery: used to determine structural and biochemical properties of potential drugs

Applications

In addition to finding the evolutionary history of species, phylogeny is also used for:

- drug discovery: used to determine structural and biochemical properties of potential drugs
- determining origins of HIV infection

Applications

In addition to finding the evolutionary history of species, phylogeny is also used for:

- drug discovery: used to determine structural and biochemical properties of potential drugs
- determining origins of HIV infection
- origin of other virus and bacteria strains

Algorithms for Reconstruction

- Most optimization criteria are hard:

Algorithms for Reconstruction

- Most optimization criteria are hard:
  - Maximum Parsimony: (NP-hard: Foulds & Graham '82)
    find the tree that can explain the observed sequences with a minimal number of substitutions.

Algorithms for Reconstruction

- Most optimization criteria are hard:
  - Maximum Parsimony: (NP-hard: Foulds & Graham '82) find the tree that can explain the observed sequences with a minimal number of substitutions.
  - Maximum Likelihood Estimation: find the tree with the maximum likelihood: P(data|tree).

Algorithms for Reconstruction

- Most optimization criteria are hard:
  - Maximum Parsimony: (NP-hard: Foulds & Graham '82) find the tree that can explain the observed sequences with a minimal number of substitutions.
  - Maximum Likelihood Estimation: find the tree with the maximum likelihood: P(data|tree).
- For both, the best known algorithms require an exponential number of trees to be checked.

Algorithms for Reconstruction

- Most optimization criteria are hard:
  - Maximum Parsimony: (NP-hard: Foulds & Graham '82)
    find the tree that can explain the observed sequences with a minimal number of substitutions.
  - Maximum Likelihood Estimation: find the tree with the maximum likelihood: P(data|tree).
- For both, the best known algorithms require an exponential number of trees to be checked.
  This is not feasible for more than 20 taxa.

Approximation Algorithms

- Since calculating the exact answer is hard, algorithms that estimate the answer have been developed.

Approximation Algorithms

- Since calculating the exact answer is hard, algorithms that estimate the answer have been developed.
  - Heuristics for maximum parsimony and maximum likelihood estimation

    (use clever ways to limit the number of trees checked, while still sampling much of "tree-space")

Approximation Algorithms

- Since calculating the exact answer is hard, algorithms that estimate the answer have been developed.
  - Heuristics for maximum parsimony and maximum likelihood estimation

    (use clever ways to limit the number of trees checked, while still sampling much of "tree-space")
  - Polynomial-time methods, based on the distance between taxa

Distance-Based Methods

- These methods calculate the distance between taxa:

	B	D	A	C	F	E
B	0	0.496505	0.496505	0.444519	0.375798	0.268166
D	0.496505	0	0.496505	0.375798	0.275673	0.279728
A	0.496505	0.496505	0	0.362124	0.323812	0.496505
C	0.444519	0.375798	0.362124	0	0.496505	0.496505
F	0.375798	0.275673	0.323812	0.496505	0	0.496505
E	0.268166	0.279728	0.496505	0.496505	0.496505	0

and then determine the tree using the distance matrix.

Distance-Based Methods

- These methods calculate the distance between taxa:

	B	D	A	C	F	E
B	0	0.496505	0.496505	0.444519	0.375798	0.268166
D	0.496505	0	0.496505	0.375798	0.275673	0.279728
A	0.496505	0.496505	0	0.362124	0.323812	0.496505
C	0.444519	0.375798	0.362124	0	0.496505	0.496505
F	0.375798	0.275673	0.323812	0.496505	0	0.496505
E	0.268166	0.279728	0.496505	0.496505	0.496505	0

  and then determine the tree using the distance matrix.

- One way to calculate distance is to take differences divided by the length (the normalized Hamming distance).

Distance-Based Methods

- Popular distance based methods include

Distance-Based Methods

- Popular distance based methods include
  - Neighbor Joining (Saitou & Nei '87) which repeatedly joins the "nearest neighbors" to build a tree, and

Distance-Based Methods

- Popular distance based methods include
  - Neighbor Joining (Saitou & Nei '87) which repeatedly joins the "nearest neighbors" to build a tree, and
  - Quartet-based methods that decide the topology for every 4 taxa and then assemble them to form a tree (Berry *et al.* 1999, 2000, 2001).

Distance-Based Methods

- Popular distance based methods include
  - Neighbor Joining (Saitou & Nei '87) which repeatedly joins the "nearest neighbors" to build a tree, and
  - Quartet-based methods that decide the topology for every 4 taxa and then assemble them to form a tree (Berry *et al.* 1999, 2000, 2001).
- Many of these methods have good performance empirically, and some can be proven to have nice accuracy properties.

Testing Methods Empirically

- How accurate are the methods at reconstructing trees?

Testing Methods Empirically

- How accurate are the methods at reconstructing trees?
- In biological applications, the true, historical tree is almost never known, which makes assessing the quality of phylogenetic reconstruction methods problematic.

Testing Methods Empirically

- How accurate are the methods at reconstructing trees?
- In biological applications, the true, historical tree is almost never known, which makes assessing the quality of phylogenetic reconstruction methods problematic.
  (an exception: Hillis '92 created an evolutionary tree in the laboratory)

Testing Methods Empirically

- How accurate are the methods at reconstructing trees?
- In biological applications, the true, historical tree is almost never known, which makes assessing the quality of phylogenetic reconstruction methods problematic.
  (an exception: Hillis '92 created an evolutionary tree in the laboratory)
- Simulation is used instead to evaluate methods, given a model of evolution.

Simulation Studies

1. Construct a

"model" tree.

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

3. Reconstruct the tree using method.

A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

3. Reconstruct the tree using method.

A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

4. Evaluate the accuracy of the constructed tree.

Simulation Studies

1. Construct a "model" tree.

2. "Evolve" sequences down the tree.

3. Reconstruct the tree using method.

A	GTTAGAAGGCGGCCA...
B	CATTTGTCCTAACTT...
C	CAAGAGGCCACTGCA...
D	CCGACTTCCAACCTC...
E	ATGGGGCACGATGGA...
F	TACAAATACGCGCAA...

4. Evaluate the accuracy of the constructed tree.