

Comparative Genomics

Fredrik Ronquist
Steve Thompson
TA: Clemens Lakner

1. INTRODUCTION

2. THE SCIENTIFIC METHOD

3. COMPARATIVE GENOMICS

1. INTRODUCTION

Fredrik Ronquist

- PhD at Uppsala University, Sweden, in 1994 on "Comparative morphology, phylogeny and evolution of cynipoid wasps"
- Senior Curator at the Swedish Museum of Natural History in Stockholm, 1993-1996
- Assistant Professor of Biology; Professor of Systematic Zoology at Uppsala University 1996-2003
- Associate Professor at the School of Computational Sciences at FSU since August 2003

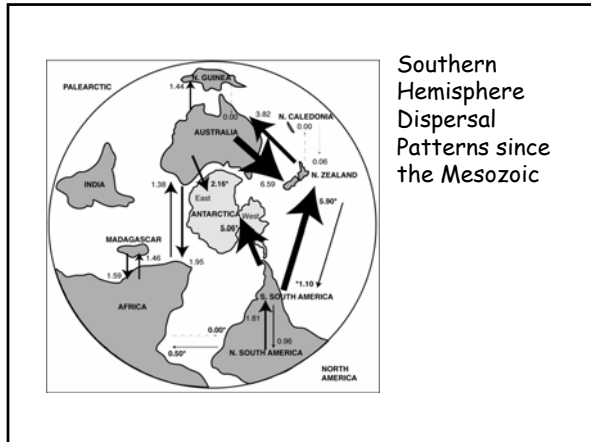



Research Interests

- Phylogeny and evolution of Hymenoptera (NSF: Assembling the Tree of Life)
- Biological image databases (www.morphbank.net)
- Parsimony methods for reconstructing host-parasite coevolution and past organism distributions and dispersal patterns (TreeFitter)
- Bayesian inference of phylogeny (www.mrbayes.net)



An undescrbed figitid wasp. Its larvae develop inside anthomyiid (fly) larvae

MrBayes: Bayesian Inference of Phylogeny

MrBayes is a program for the Bayesian estimation of phylogeny. Bayesian inference of phylogeny is based upon a quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observations. The conditioning is accomplished using Bayes's theorem. The posterior probability distribution of trees is impossible to calculate analytically; instead, MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees.

[Home](#)
[Download](#)
[Manual](#)
[Authors](#)
[Links](#)

The program takes as input a character matrix in a NEXUS file format. The output is several files with the parameters that were sampled by the MCMC algorithm. MrBayes can summarize the information in these files for the user. The program features include:

- A common command-line interface for Macintosh, Windows, and UNIX operating systems;
- Extensive help available via the command line;
- Ability to analyze nucleotide, amino acid, restriction site, and morphological data;
- Mixing of data types, such as molecular and morphological characters, in a single analysis;
- A general method for assigning parameters across data partitions;
- An abundance of evolutionary models, including 4 X 4, doublet, and codon models for nucleotide data and many of the standard rate matrices for amino acid data;
- Estimation of positively selected sites in a fully hierarchical Bayes framework;
- The ability to spread jobs over a cluster of computers using MPI (for Macintosh and UNIX environments only).

The most recent version of the program, version 3.0, has many improvements over the last version. Many bugs were fixed and improvements in the program's design have resulted in about a two times speed improvement.

- ### Overview of the Course
- Introduction, The Scientific Method
 - Crash Course in Comparative Genomics
 - Lectures and labs, group project and individual project
 - Individual Project:
 - Project Proposal - Counseling
 - Individual Research
 - Write a Scientific Report
 - Oral Presentation

- ### Course Books
- *Writing Papers in the Biological Sciences*, 4th edition (Victoria E. McMillan)
 - Instructions for writing proposals and research papers (and much more)
 - *Phylogenetic Trees Made Easy, A How-To Manual*, 2nd edition (Barry G. Hall)
 - Instructions for finding genetic data, aligning them and analyzing them

- ### Grading Expectations
- Lab Assignments, 1 page (8x2 p = 16 p)
 - Project Proposal, 2 pages (14 p)
 - Project Report, 10-20 pages (50 p)
 - Oral Presentation, 8 min. (20 p)
- Detailed description of what is required is found on the course website.
- 90-100 = A
 80-89 = B
 70-79 = C
 60-69 = D

- ### Attendance
- Attendance Required:
 - First lecture (FSU policy)
 - Counseling before individual project started
 - Oral Presentation
 - Attendance Highly Recommended:
 - Lectures (PowerPoint presentations will be available on Blackboard, campus.fsu.edu)
 - Labs (instructions will be available on Blackboard, campus.fsu.edu, you should be able to complete them from home)

Plagiarism

- As long as you cite the source, you can:
 - Use information from the internet
 - Use ideas of other students, given their permission to do so
- You must:
 - Contribute something substantial and unique to the material you present
- You must not:
 - Copy material from the internet or from other sources without citing the source (you will fail the course and face disciplinary action)

Practical Things

- Classroom and Computer Access
 - Swipe your FSU card to open classroom door
 - You can use the classroom computers on a first come first serve basis during weekdays 8 am to 5.30 pm when no other activities are scheduled (see the web for calendar).
 - Work from anywhere: Your user account will allow you to log into the classroom or Mendel accounts from anywhere.
 - Use a supercomputer (cluster): You can use classroom computers as a cluster by logging into Condor.
- Office Hours
 - Fredrik: Thursdays 1 pm - 2pm. Someone will be in or near the classroom to take care of you during the individual project period Thursdays 1.00 pm-6.00 pm.

Practical Things (cont'd)

- Time to:
 - Check attendance; collect FSU card info
 - Get your classroom user account (see general info at <http://campus.fsu.edu> if you do not have one)
 - Log into your computer
 - Launch Firefox (single-click on Firefox symbol in the bar at the bottom of the desktop) and explore the course web site (<http://campus.fsu.edu>)
 - Find the calendar for the classroom this week (link on course web pages)

2. THE SCIENTIFIC METHOD

The Scientific Method

1. Ask a question
2. Formulate scientific hypothesis
3. Derive testable predictions
4. Collect data
5. Try to disprove (falsify) the hypothesis
6. If the hypothesis is falsified, go to 2, else go to 3.

EXAMPLE

1. Q: What is the shape of Earth?
2. H: Earth is flat
3. Pred.: Horizon is a straight line
4. Data: Observe horizon
5. Conclusion: Horizon is curved, hypothesis falsified
6. Find new hypothesis (step 2).

The Scientific Method (2)

- Ask question
- Formulate alternative hypotheses
- Collect relevant data
- Use explicit or implicit probability reasoning to choose among alternative hypotheses
- Find new hypotheses if observations are unlikely under any of the existing hypotheses

EXAMPLE 2

1. Q: What is the shape of Earth?
2. H0: Earth is flat; H1: Earth is spherical
3. Pred.: Horizon is a straight line or horizon curves according to the curvature of Earth
4. Data: Observe horizon
5. Conclusion: Horizon is curved, Earth is likely to be spherical
6. Find new testable predictions (step 3).

What is a scientific hypothesis?

- You can derive testable predictions from it
- If not, it is not a scientific hypothesis (or scientific question)
- Definitions of words are often important in determining whether a hypothesis is testable
 - Example: *God answers prayers*

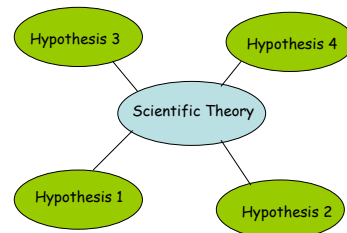
Are these scientific hypotheses?

- Clemens Lakner is immortal
- The 'Noles are a better football team than the Gators
- God exists
- There are 60 minutes in an hour
- HIV is not transmitted by sex
- Body odor is important in human mate choice
- Flowers are beautiful
- Johnny Depp is attractive to women
- Jesus is a historical person

Scientific Theory

- A coherent group of general propositions used as principles of explanation for a class of phenomena
- A scientific theory should be compatible with a large number of hypotheses that have withstood critical testing
- In comparative biology, alternative hypotheses are often all based on the theory of evolution (descent with modification)

Theory and Hypotheses



A good scientific theory should be supported by a large number of well-tested hypotheses and contradicted by few if any such hypotheses

HELP! The best hypothesis or theory is not obvious!

- Cookbook methods
- The parsimony principle
- Statistical inference

Cookbook methods

- Follow a predetermined recipe
- Work well in many cases
- Often simple and fast
- Characteristic feature: if you change the recipe, you are no longer using the same method
- **EXAMPLE:** Always choose the hypothesis with the smallest number of words

The Parsimony Principle

- Also known as Occam's razor (after William of Occam, died 1349?)
- The simplest explanation is the best
- Use it to choose among alternative hypotheses or scientific theories
- Example:
 - Leave this room for five minutes
 - Come back and find everything in the same place
 - H0: Nothing happened; H1: Dave Swofford came in through the back door and traded the places of two computers
 - H0 is more parsimonious than H1

Statistical Inference

- Uses probability theory to augment the parsimony principle
- Two major kinds:
 - **Maximum Likelihood** (classical statistical inference): choose the most likely hypothesis given the data and some probability model
 - **Bayesian Inference:** update your prior beliefs given the data and some probability model

EXAMPLE 3

1. Q: Are the 'Noles or the Gators a better football team?
2. H0: 'Noles better than Gators; H1: 'Noles worse than Gators.
3. Pred.: If 'Noles are better than they are more likely to win than the Gators when they play each other.
4. Data: The 'Noles play the Gators ten times. The 'Noles win three times, the Gators seven times.
5. Conclusion: H1 most likely to be true but it is still possible that H0 is correct and that the 'Noles just had some bad days. Use either Maximum Likelihood or Bayesian Inference.
6. Tentatively accept H1. Find new testable predictions (step 3).

3. WHAT IS COMPARATIVE GENOMICS?

What is a Genome?

- **Gene:** Piece of DNA that determines the composition of a polypeptide, often associated with particular traits
- **Genome:** An organism's or cell's entire complement of genetic material (DNA). For example, human sex cells have about 3×10^9 base pairs (nucleotides: A C G T) of DNA, containing about 25-30,000 genes.
- Can you fit the human genome into a book (say 500 pages with 80 characters per line and 30 lines per page)?
- Can you fit the human genome onto a CD-ROM? A DVD?

What does Comparative mean?

- Comparison of genes (or genomes) within or among species
- Usually requires an evolutionary tree (phylogeny) that describes the genealogy (family tree) of the gene (DNA or protein sequence) and its close relatives

Structure of a Comparative Genomics Research Project

- Find an interesting question and one or more alternative hypotheses
- Write a proposal and find funding for the project
- Collect relevant data from web databases of proteins, DNA sequences, or genome maps
- Analyze how the data relate to the postulated hypotheses using an appropriate method
- Write a scientific report

Skills You Will Learn (1)

Finding Sequences

- Find a sequence (in public databases) that is cited in a published paper
- Find sequences (in public databases) with specific organism, gene name or other tags
- Find genes (in public databases) implied in a particular disease (sometimes)
- Find sequences (in public databases) similar to (related to) a specific sequence

Skills You Will Learn (2)

Analyzing Sequences

- Find structural motifs in a protein sequence
- Find, view and use published 3D structures of a protein
- Align sequences, matching comparable (homologous) sites
- Infer the evolutionary (phylogenetic) tree for aligned sequences

Skills You Will Learn (3)

Using Evolutionary Trees

- Evaluate a biological classification
- Find the closest relatives of a group of sequences
- Find the rate of evolution in different organism groups, different genes or different gene segments
- Identify where and when specific traits originated, and if they correlate with specific sequence changes
- Identify the date of origin and place of origin of a group of sequences (representing, for instance, a disease or a specific trait)
- Find duplication of genes

Difficult Projects

- Analyses requiring patient data (typically not available but there may be some exceptions)
- Studying the evolution of traits that you cannot easily associate with sequences in public databases
- Hypothesis tests requiring experimental data
- Comparisons involving only human sequences (unless the sequences represent a gene family)

Some Advice

- Start thinking about the subject for your individual project now
- Take time to formulate your hypothesis: Asking the right question is half the answer
- Be skeptical: Don't trust publications
- Ask others (students and teachers) for ideas
- Use only one or a few of the methods we will cover and make sure you understand them well