National Institute of Allergy and Infectious Diseases

Phylogenetics and Sequence Analysis

Lecture 1
Introduction and Sequence Assembly

Kurt Wollenberg, PhD
Phylogenetics Specialist
Bioinformatics and Computational Biosciences Branch
Office of Cyber Infrastructure and Computational Biology

Fall 2015

NIH National Institute of Allergy and Infectious Diseases

NIAID

We Are BCBB!

NIH National Institute of Allergy and Infectious Diseases

- Group of 37
  - Bioinformatics Software Developers
  - Computational Biologists
  - Project Management & Analysis Professionals

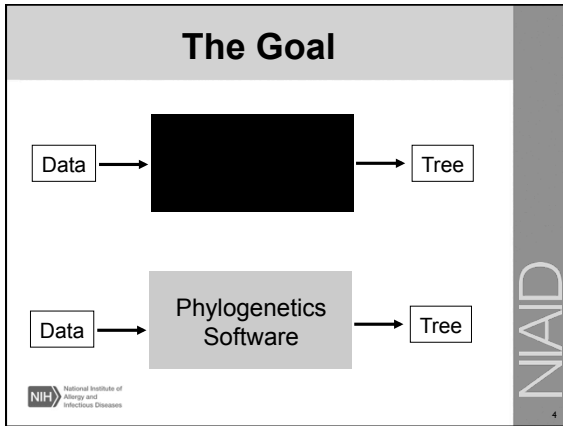NIH National Institute of Allergy and Infectious Diseases

NIAID

2

Contact BCBB…

- "NIH Users: Access a menu of BCBB services on the NIAID Intranet:
  - http://bioinformatics.niaid.nih.gov/
- Outside of NIH –
  - search "BCBB" on the NIAID Public Internet Page: www.niaid.nih.gov

    – or – use this direct link

- Email us at:
  - ScienceApps@niaid.nih.gov

NIH National Institute of Allergy and Infectious Diseases

NIAID

3

The Goal

Data → ▮ → Tree

Data → Phylogenetics Software → Tree

NIAID

4

Course Organization

- Building a clean sequence
- Collecting homologs
- Aligning your sequences
- Building trees
- Further analyses

NIAID

5

Biological sequences

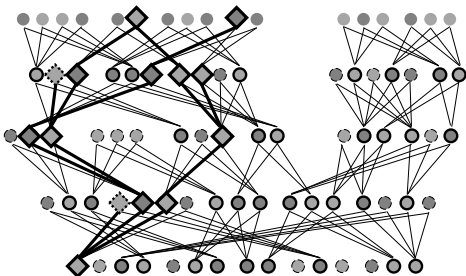# Why analyze biological sequences?

NIAID

6

Biological sequences

- DNA contains the information basic to every process in a cell
- Proteins (and RNA) are the machines performing cellular processes
- Passed from one generation to the next

NIAID

National Institute of Allergy and Infectious Diseases

7

Sequence data are genealogical

NIAID

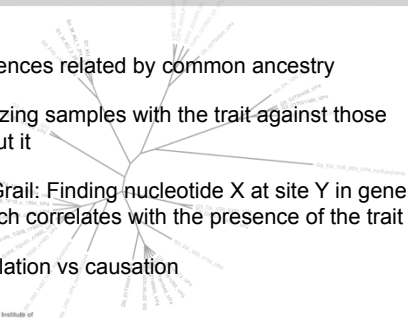National Institute of Allergy and Infectious Diseases

8

Comparative Methods

Why analyze sequences using comparative methods?

NIAID

National Institute of Allergy and Infectious Diseases

9

Comparative Methods

- Sequences related by common ancestry

- Analyzing samples with the trait against those without it

- The Grail: Finding nucleotide X at site Y in gene Z which correlates with the presence of the trait

- Correlation vs causation

NIAID

10

Hierarchy of Life

- Carl Linnaeus (1707 - 1778)
  - Swedish physician/naturalist
  - Hierarchical organization of life
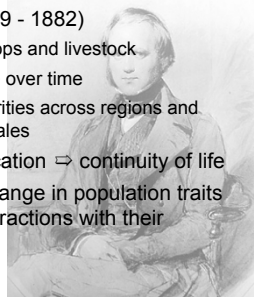 - Binomial system of scientific names

NIAID

11

Common Ancestry

- Charles Darwin (1809 - 1882)
  - Artificial selection: crops and livestock
  - Fossil record: change over time
  - Biogeography: similarities across regions and differences within locales
- Descent with modification ⇨ continuity of life
- Natural selection: change in population traits due to individual interactions with their environment
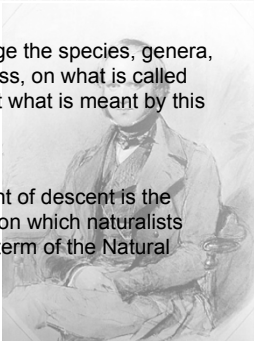
NIAID

12

Common Ancestry

"Naturalists try to arrange the species, genera, and families in each class, on what is called the Natural System. But what is meant by this system?" p.413

"… I believe this element of descent is the hidden bond of connexion which naturalists have sought under the term of the Natural System" p. 433
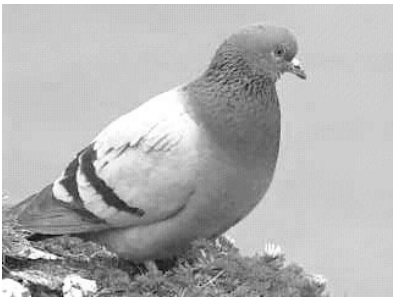
NIAD

NIH National Institute of Allergy and Infectious Diseases

13

Artificial Selection

NIAD

NIH National Institute of Allergy and Infectious Diseases

14

Artificial Selection

De La Nuez Photo *De La Nuez Photo* *De La Nuez Photo*

NIAD

NIH National Institute of Allergy and Infectious Diseases

15

Course Organization

- **Building a clean sequence**
- Collecting homologs
- Aligning your sequences
- Building trees
- Further analyses

NIAID

NIH National Institute of Allergy and Infectious Diseases

16

I've cloned a sequence ...

- Getting from raw sequence data to a clean sequence for further analysis.
  - Removing low-quality peaks
  - Removing plasmid sequence
  - Combining reads into a contig

NIAID

NIH National Institute of Allergy and Infectious Diseases

17

Welcome to Sequencer

The complete software solution for sequencing DNA

SEQUENCHER™
5.3 - Build 43228
Registration Number : 2902016

(C) 1991 - 2014
Gene Codes Corporation and its licensors
All rights reserved.

T C A GENE
A G T CODES

NIAID

NIH National Institute of Allergy and Infectious Diseases

18

Alternatives to Sequencher

- Lasergene
  - NIAID site license
  - Suite of molecular sequence analysis programs
- Mesquite
  - Open source, freely distributed
  - Java program, using modules for analysis
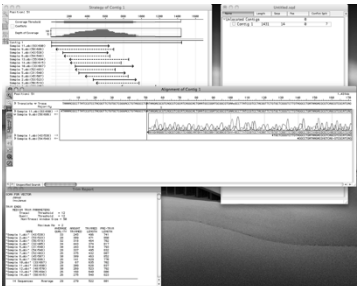 - Requires a local installation of Phred and Phrap

NIH National Institute of Allergy and Infectious Diseases
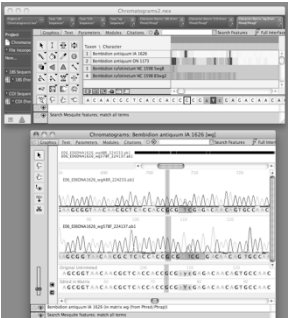
19

Contig Assembly Software

Lasergene:
SeqMan

NIH National Institute of Allergy and Infectious Diseases

20

Contig Assembly Software

Mesquite:
Chromaseq

NIH National Institute of Allergy and Infectious Diseases

21

Recapitulation

- Hierarchical and genealogical data
- Comparative sequence analysis
- Generating clean sequence
  - Trim vector contamination
  - Trim low-quality ends
  - Align fragment overlap to build contig
  - Export contig (consensus)

NIH National Institute of Allergy and Infectious Diseases

22

Seminar Follow-Up Site

- For access to past recordings, handouts, slides visit this site from the NIH network: http://collab.niaid.nih.gov/sites/research/SIG/Bioinformatics/

Recommended Browsers:
- IE for Windows,
- Safari for Mac (Firefox on a Mac is incompatible with NIH Authentication technology)

Login
- If prompted to log in use "NIH\" in front of your username

NIH National Institute of Allergy and Infectious Diseases

23

Retrieving Slides/Handouts

NIH National Institute of Allergy and Infectious Diseases

24

Retrieving Slides/Handouts

Questions?

ScienceApps@niaid.nih.gov

Next Lecture

**BLAST Searching
and
Multiple Sequence Alignment**