Hands-on Session 2: Obtaining Data from On-line Sources

Katherine St. John
Lehman College and the Graduate Center
City University of New York
stjohn@lehman.cuny.edu

Session Organization

- Goal: To be comfortable building trees from real data
- Lecture:
  - Standard Software Packages
  - Details on Web-based Software
  - Motivating Problem
- Lab:
  - Organized so you can use the DIMACS lab, or your own laptop
  - Welcome to work singly or in groups

Lecture Outline

- Motivating Problem

Lecture Outline

- Motivating Problem
- Building Trees Overview

Lecture Outline

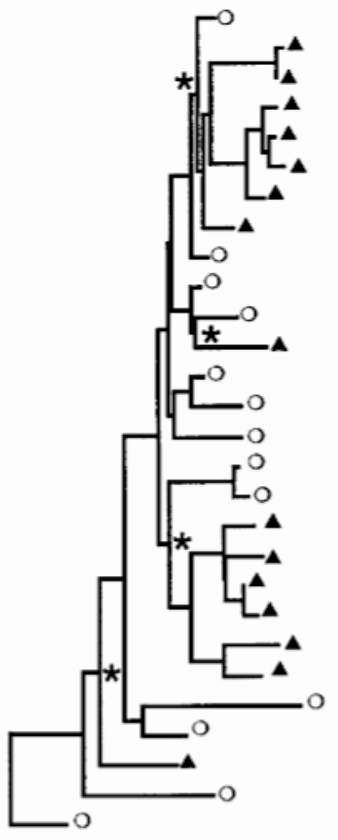
- Motivating Problem
- Building Trees Overview
- Using Sequence Databases

Lecture Outline

- Motivating Problem
- Building Trees Overview
- Using Sequence Databases
- Aligning Sequences

Motivating Problem: Building Trees with Serial Data?

Rodrigo *et al.*,
"Coalescent estimates of
HIV-1 generation time in vivo."
PNAS '99



Motivating Problem: Using Serial Data

- Rodrigo *et al.* includes 55 HIV-env partial sequences, all from the same patient

Table 1. Summary statistics for each sequence sample set

Sample	Days from first sample	No. of sequences	Average pairwise diversity, %	θ	N
1	0	13	3.6	0.088	1100
2	214	15	3.9	0.106	1325
3	671	15	5.0	0.074	925
4	699	9	4.2	0.144	1800
5	1005	8	4.1	0.092	1150

- Starting question: what is the genealogy samples (from the same patient) taken at different times?

Building Trees

1. Get data (from wet lab, authors, genBank, etc).

Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.

Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
3. If needed, choose the appropriate model of evolution.

Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
3. If needed, choose the appropriate model of evolution.
4. Use software program(s) to build trees.

Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
3. If needed, choose the appropriate model of evolution.
4. Use software program(s) to build trees.
5. Analyze Results.

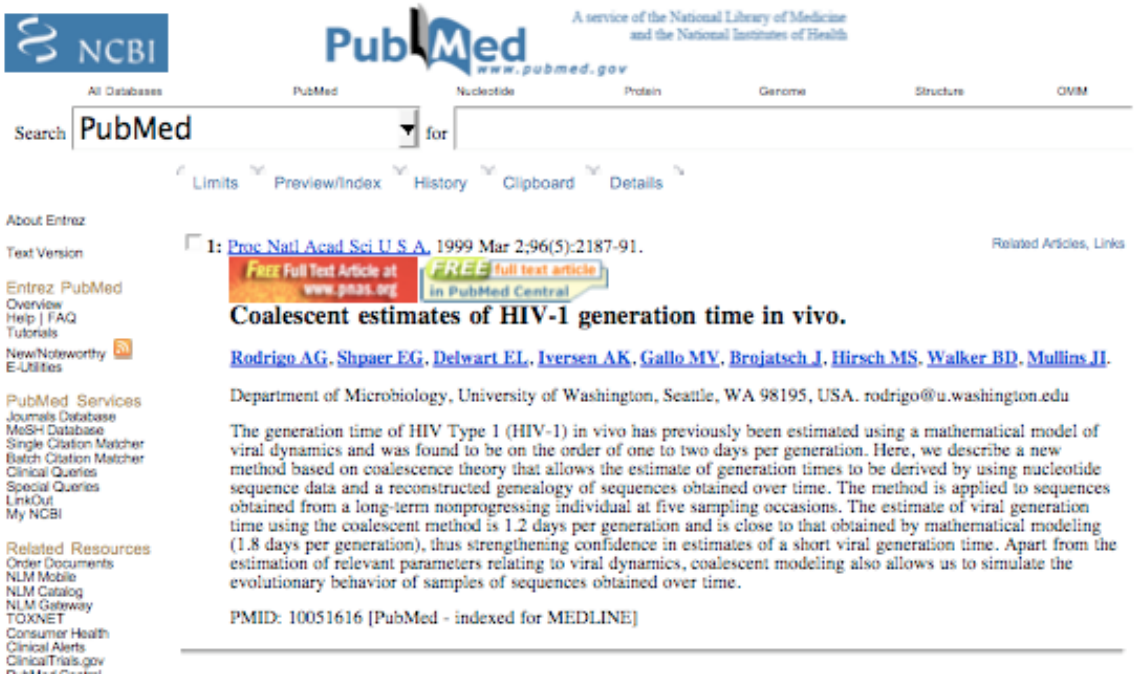
Building Trees

1. Get data (from wet lab, authors, genBank, etc).
2. Align and/or filter data.
3. If needed, choose the appropriate model of evolution.
4. Use software program(s) to build trees.
5. Analyze Results.

We'll focus on the first two today.

Using PubMed

An on-line index of scientific papers:



Can search by all standard methods...

Sequence Databases

- GenBank: repository of sequences from NCBI (NIH).
- As of August 2005, GenBank had 100 gigabases of sequences.
- Almost all sequences from published articles are there, and can be located by their unique accession number or PubMed ID.

LANL HIV Databases

- Los Alamos National Laboratory maintains databases of sequences, resistance, immunology, and vaccine trials.
- Can be searched in numerous ways including accession number or PubMed ID.

Aligning Sequences

- Before building a tree, the similar regions of the sequences need to be aligned.

Aligning Sequences

- Before building a tree, the similar regions of the sequences need to be aligned.
- One of the most common alignment programs is ClustalW:
  - Available via multiple servers including EBI & the Pasteur Institute
  - Does a global multiple sequence alignment

Getting Started

- Find the Rodrigo *et al.* paper on PubMed.

Getting Started

- Find the Rodrigo *et al.* paper on PubMed. Download the paper, and note it's PubMed ID (PMID).
- Use the PMID to find the sequences in the HIV Sequence Database.

Getting Started

- Find the Rodrigo *et al.* paper on PubMed. Download the paper, and note it's PubMed ID (PMID).
- Use the PMID to find the sequences in the HIV Sequence Database.
- Use ClustalW to align the sequences.

Getting Started

- Find the Rodrigo *et al.* paper on PubMed. Download the paper, and note it's PubMed ID (PMID).
- Use the PMID to find the sequences in the HIV Sequence Database.
- Use ClustalW to align the sequences.
- Using your favorite phylogenetic reconstruction method, build a tree from the sequences.

Getting Started

- Find the Rodrigo *et al.* paper on PubMed. Download the paper, and note it's PubMed ID (PMID).
- Use the PMID to find the sequences in the HIV Sequence Database.
- Use ClustalW to align the sequences.
- Using your favorite phylogenetic reconstruction method, build a tree from the sequences.
- Analyze resulting trees

Hints:

- Choose the "fast" tree building option for `ClustalW`.

Hints:

- Choose the "fast" tree building option for `ClustalW`.
- To use a distance based method, you need to create a distance matrix (dnadist) to give to the method (ie `BioNJ` or `QuickTree`).

Hints:

- Choose the "fast" tree building option for `ClustalW`.
- To use a distance based method, you need to create a distance matrix (dnadist) to give to the method (ie `BioNJ` or `QuickTree`).
- At the Pasteur Institute site, at each step, you can choose the next step, without reloading the file.

Hints:

- Choose the "fast" tree building option for `ClustalW`.
- To use a distance based method, you need to create a distance matrix (dnadist) to give to the method (ie `BioNJ` or `QuickTree`).
- At the Pasteur Institute site, at each step, you can choose the next step, without reloading the file.

  For example, after returning the distance matrix, you have the option of applying a method to the matrix.

Helpful Websites

- Dataset for this tutorial:

  `http://comet.lehman.cuny.edu/stjohn/dimacsTutorial`

- PubMed & Genbank:

  `http://www.ncbi.nlm.nih.gov/entrez`

- HIV Sequence Database:

  `http://hiv-web.lanl.gov/content/index`

- The Pasteur Institute:

  `http://bioweb.pasteur.fr/intro-uk.html`