

Two Sequence Alignment & Scoring Matrices

BME 110: CompBio Tools

Todd Lowe

April 08, 2008

Admin

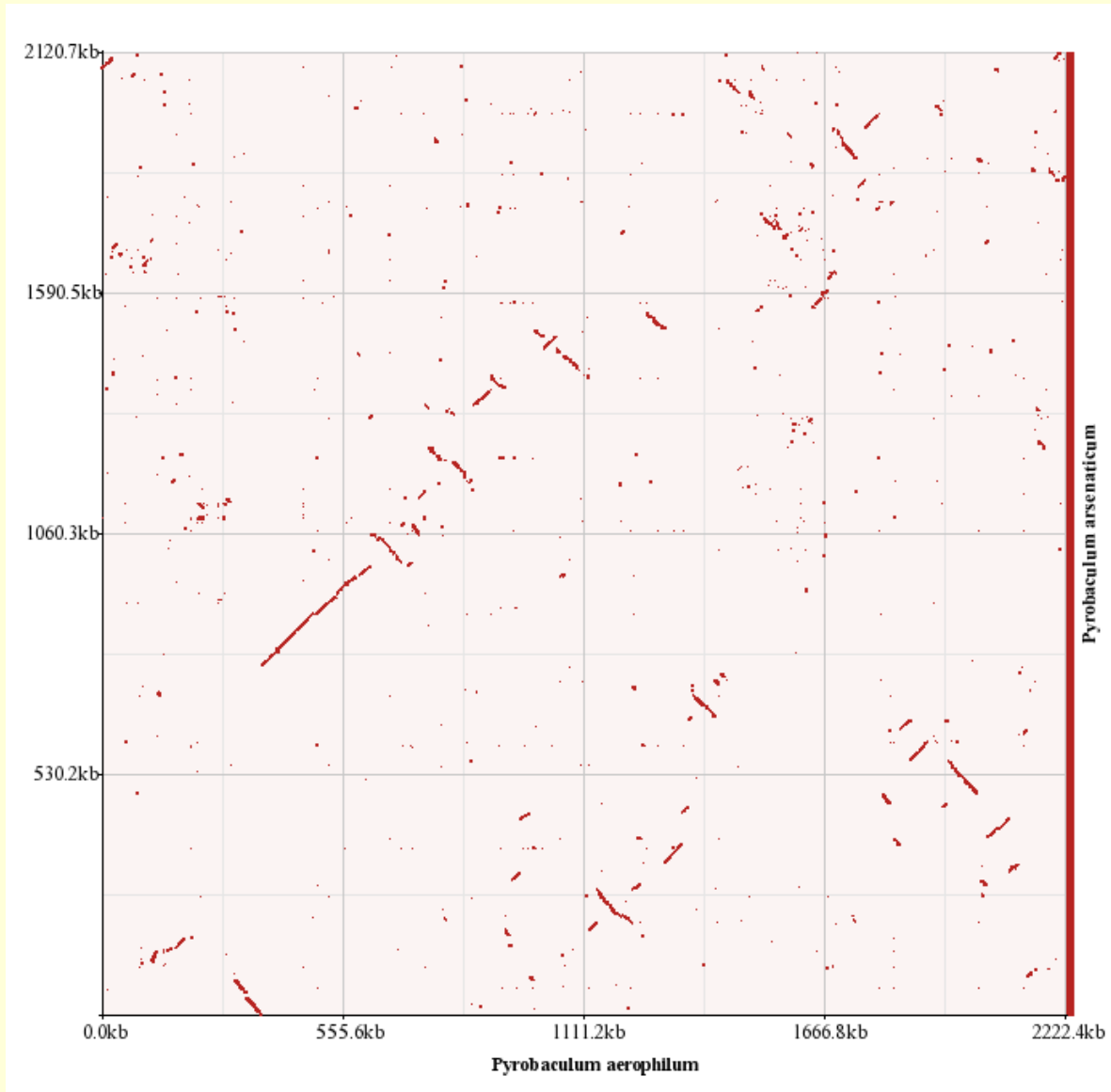
- Reading:
 - Chapter 3 should be completed
 - Chapter 5 for tuesday
- Homework #1 due tomorrow (Fri) 5pm
- Homework #2 assigned Tuesday

Dot-plot: Most Basic Sequence Comparison

- Put sequence on X & Y axes
- Mark “X” for nucleotide match
- Self-sequence comparison:
 - Find repeats, insertions, deletions, palindromes
- Two-sequence analysis
 - Find similarity, inversions, transversions, on large scale

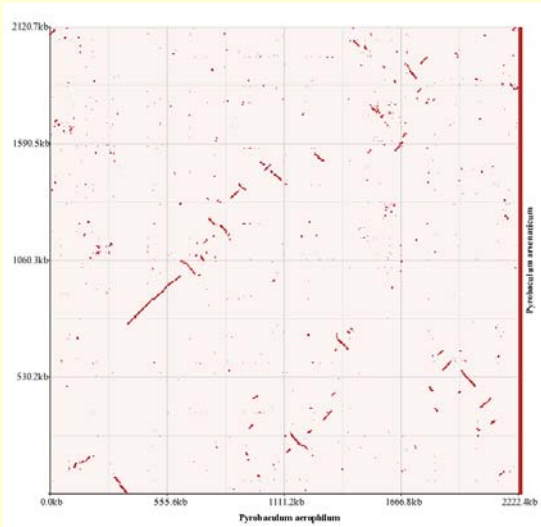
Many tools out there: Dotlet is good java app

Full Genome Dot-Plot



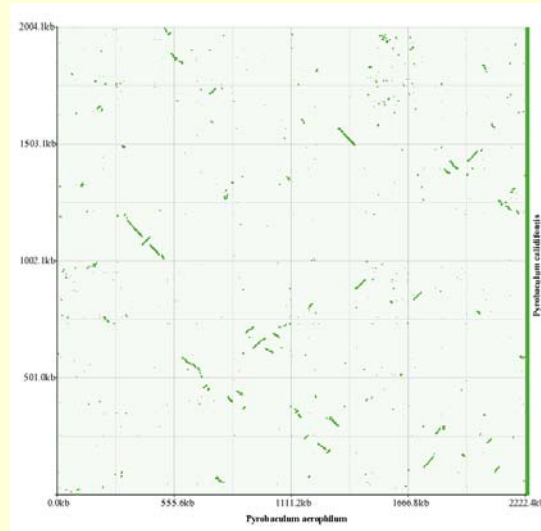
Multiple Genome Alignment Dot Plots

P.arsenaticum



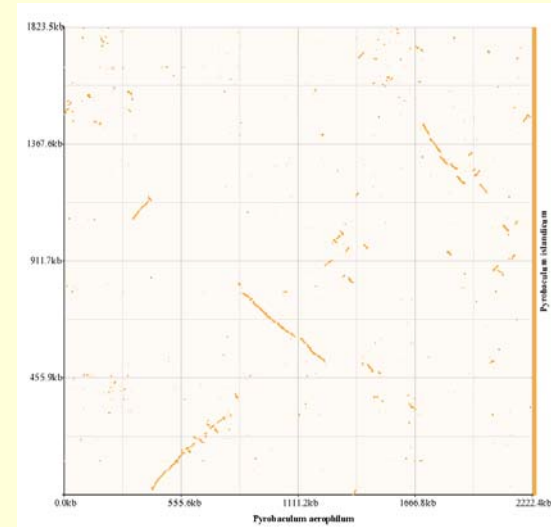
P.aerophilum

P.calidifontis



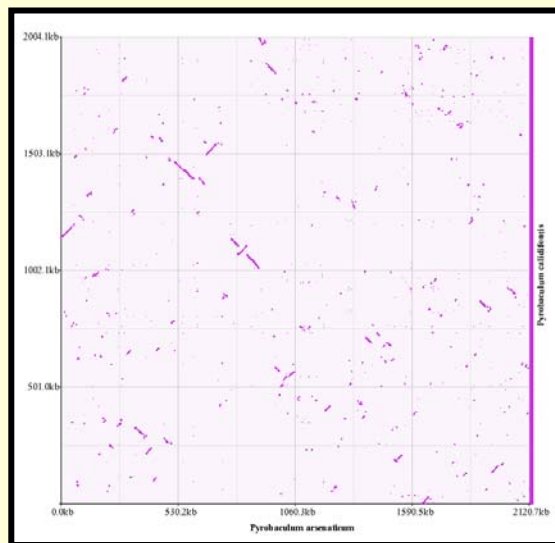
P.aerophilum

P.islandicum



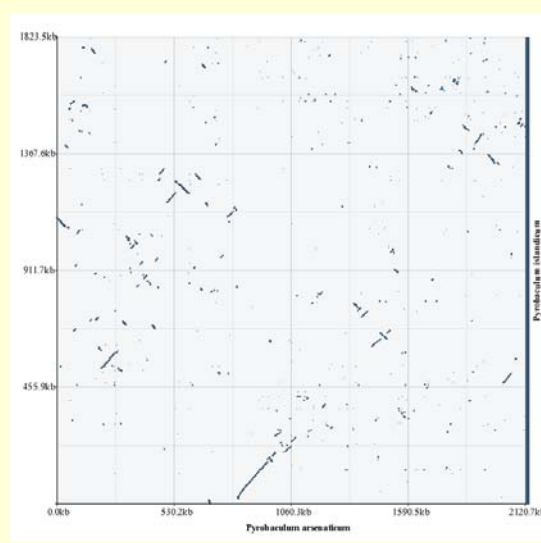
P.aerophilum

P.calidifontis



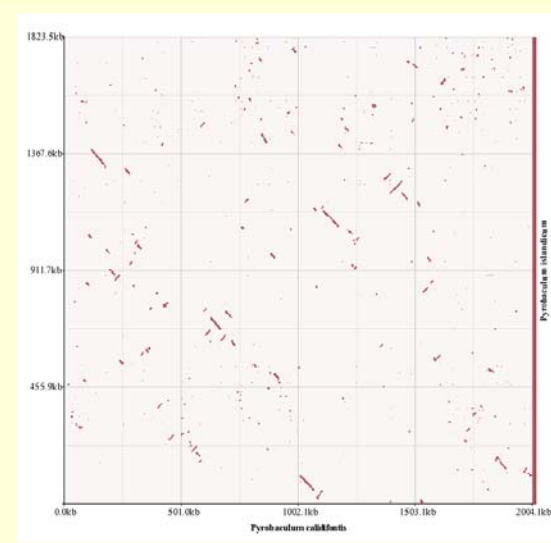
P.arsenaticum

P.islandicum



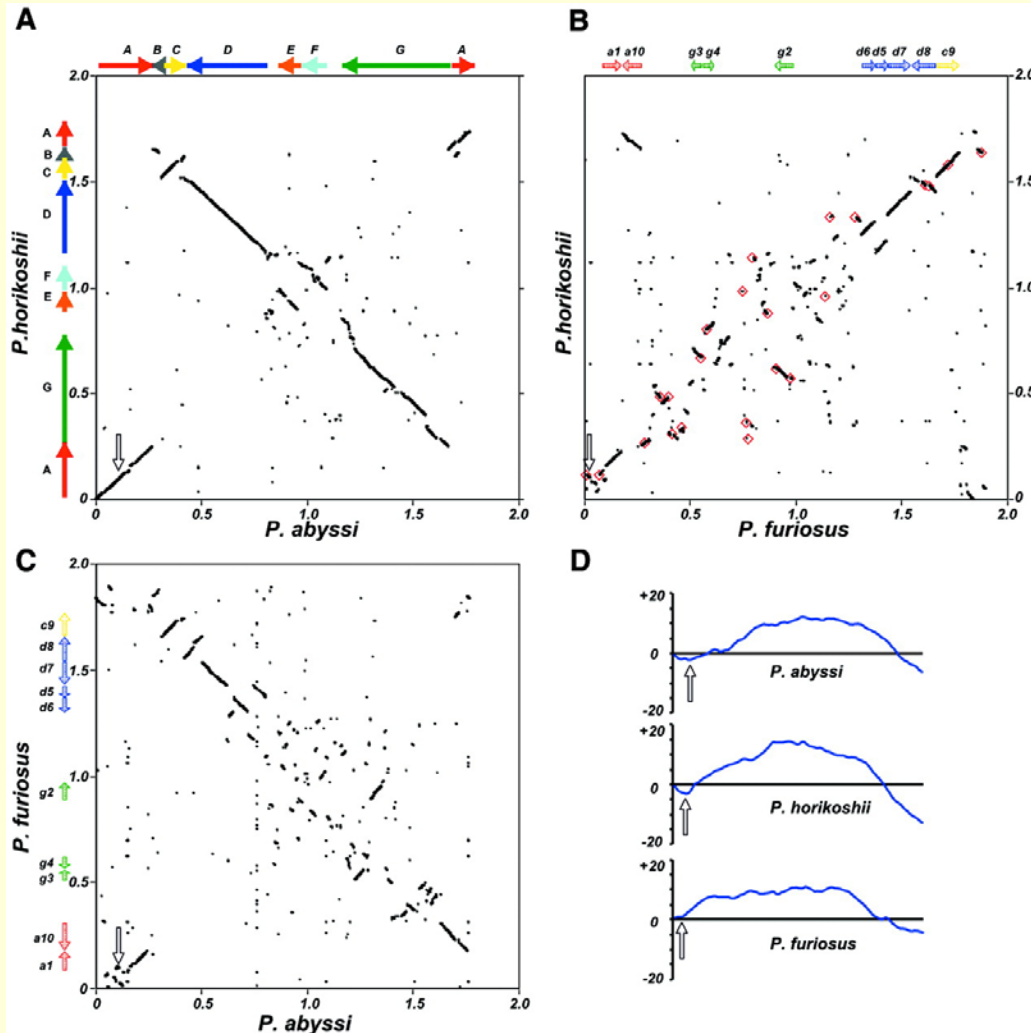
P.arsenaticum

P.islandicum



P.calidifontis

Full-genome Comparisons



Each dot is a
gene match

From **Zivanovic**
et al., NAR
30: 1902-10

Pair-wise Sequence Comparison

- Basis for relating biological information from a well-studied gene to a new sequence
- Many programs exist for pairwise comparison
- Some specialize in fast database searching and get “good” alignments
 - One sequence v. many thousands:
 - BLAST or FASTA
- Some are much slower, but guarantee the “optimal alignment”
 - Smith-Waterman is the de facto standard

Dot-plots: Dotlet

<http://myhits.isb-sib.ch/cgi-bin/dotlet>

Example: In Archaeal Genome browser, bring up *Pyrobaculum aerophilum*

Select CRISPR2 region (chr:45,423-46,754) to compare to CRISPR6-7 region (chr:1,898,656-1,899,678)

Get DNA, paste into Dotlet one at a time, giving descriptive labels,
Zoom 1:5,

Are there direct or inverted repeats in each CRISPR (against itself?)
Relative to each other, are these direct or inverted repeats?

Dot plots: Need to Adjust Stringency to see Patterns

- For last example, set use:
- window size = 21, adjust slider contrast bar to eliminate most gray, leaving black and white

Assessing Alignment Significance

Most Basic Rules of thumb:

Two nucleotide sequences – at least 70% identical, they are likely homologous

Two protein sequences – at least 25% identical over 100 amino acid alignment

Does not take into account precise length of alignment, or number of gaps!

Not sufficient to quantitatively rank hits from a database search

The “Twilight Zone”

- Less than 25% sequence identity for two protein sequences
- May still be homologous, but only similarity of 3-D protein structures can verify similar function (structural comparison tools to detect these discussed later in quarter)
- Must have a good / near optimal alignment for most distantly related proteins

What is an Optimal Alignment??

- How do we get an “optimal” alignment
- Optimal to who?
- Optimal based on scoring model:
 - Substitution scoring matrix
 - Insertion / deletion scoring (penalties)
- Caution: Just because it is optimal for a given scoring scheme, doesn't mean it is *biologically* correct!!

Dynamic Programming

- Fancy term for type of algorithm used to get the “optimal” or best possible alignment between two sequences
- Needleman and Wunsch (1970) most basic method
 - Gives the “global” (end to end) best alignment
- Smith-Waterman based closely on this algorithm, but allows for “local” alignments (best subsequence match only)
- See simple example of Global v. Local alignments in book, Figure 3.1 p. 71

Basic Example

- Find best global alignment of two sequences:

G A T C

G T G C

Which is better?

Match +1, Mismatch -1, Gap -2

G A T C +1 -1 -1 +1

| | | OR (Score = 0)

G T G C

G A T - C +1 -2 +1 -2 +1

| | | | (Score = -1)

G - T G C

Which is better?

Match +1, Mismatch -1, Gap -1

G A T C +1 -1 -1 +1

| | OR (Score = 0)

G T G C

G A T - C +1 -1 +1 -1 +1

| | | (Score = 1)

G - T G C

Moral: Scoring Model Matters!!

- For DNA, model can be very simple:
- +1 match, -1 mismatch

- However, not all mutations have equal likelihood:
- Transition: $A \leftrightarrow G$ or $C \leftrightarrow T$
 - more likely
- Transversion: $A \leftrightarrow C$ or $G \leftrightarrow T$
 - less likely

Kimura Two-parameter Scoring Matrix

	A	C	G	T
A	0.6	0.1	0.2	0.1
C	0.1	0.6	0.1	0.2
G	0.2	0.1	0.6	0.1
T	0.1	0.2	0.1	0.6

Actual values not important, only values relative to each other

Same Matrix (*10)

	A	C	G	T
A	6	1	2	1
C	1	6	1	2
G	2	1	6	1
T	1	2	1	6

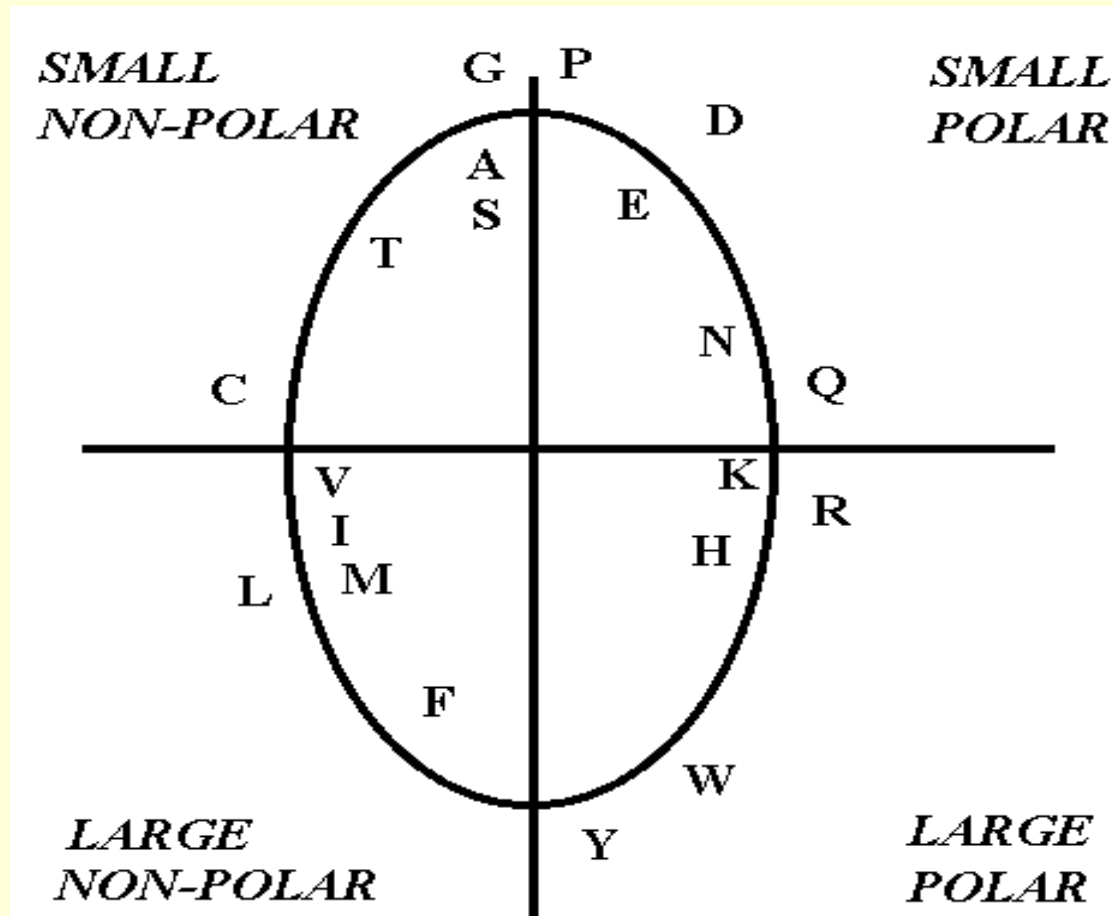
Actual values not important, only values relative to each other

Protein Matrices, Same Idea

- Original: Dayhoff matrix aka PAM
- PAM = Percent accepted mutations
- Based on small number of *correctly* aligned proteins
- Simply count how often each amino acid is substituted for another
- Frequency of substitutions based on properties of amino acids relative to each other

Newer “Version” of Protein Matrices: BLOSUM

- By Henikoff & Henikoff (1992), based on a much larger group of aligned protein sequences in the Blocks database
- BLOSUM = Blocks substitution matrix
- Used most commonly today



- Closer two amino acids are, more similar in properties

Versions of Matrices

- Should use different substitution matrices based on expected evolutionary distance between two sequences
- PAM1 original matrix
- Can derive matrices of varying evolutionary distances from original PAM 1 matrix

PAM1 < PAM120 < PAM250

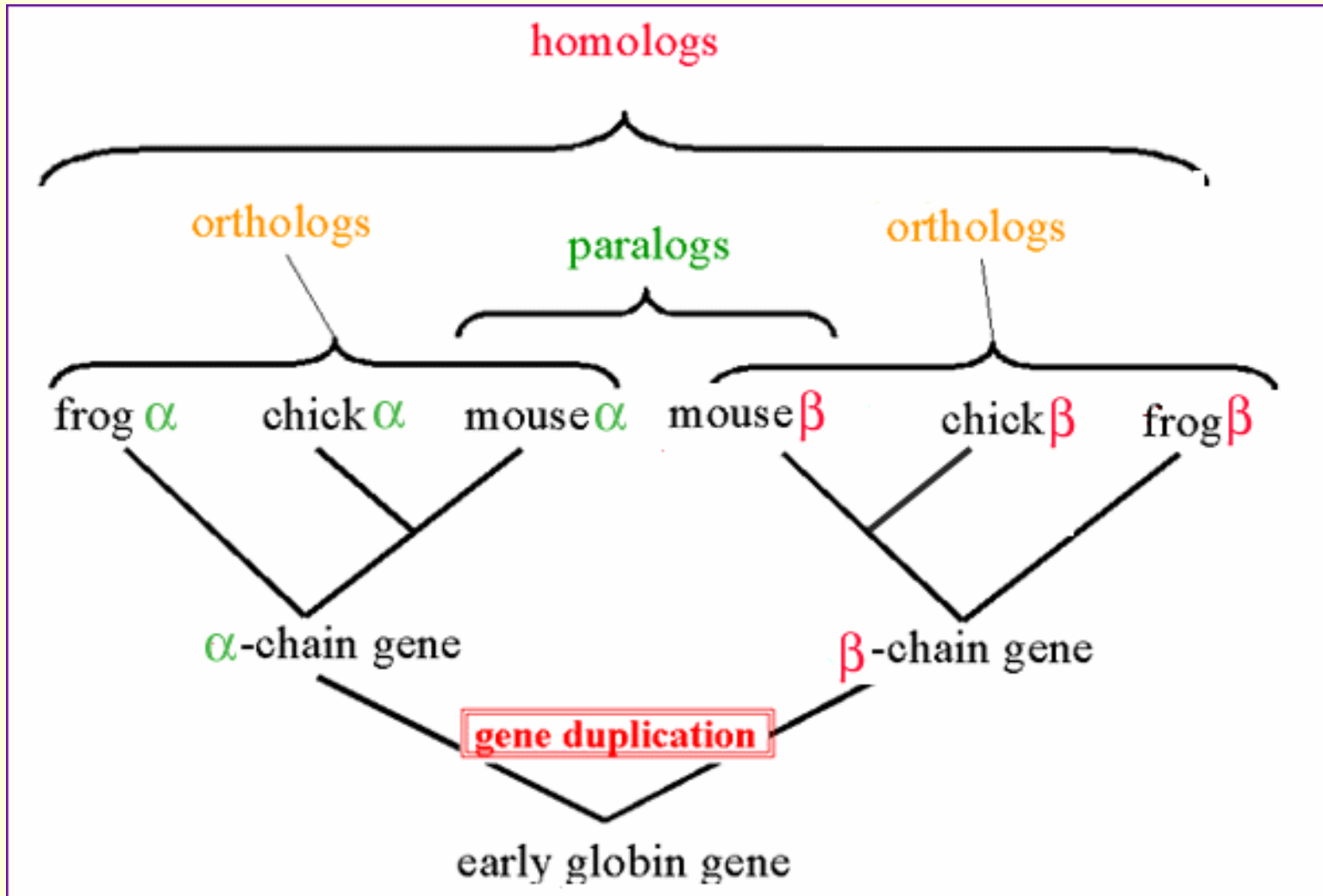
A Related Note: Homology

- Based on inference that two sequences are ancestrally derived from same molecule
- If two sequences have high *similarity*, they may be *inferred* to be homologous
- It is **WRONG** to say two sequences or genes are 80% homologous (they either are related, or they are not)

Homology: Same Function?

- Even if two sequences are ancestrally derived from same molecule, they may or may not still have the same function
 - Orthologs: homologous genes created by speciation
 - Generally implies function remains the same
 - Paralogs: homologous genes created by a gene duplication event (in same species)
 - Implies function may have changed

Homology Diagram



Convergent Evolution

- Two genes have independently evolved to have same function
- Figure 3.3 C
- Because genes are not homologous, convergent evolution is usually not detectable by sequence analysis (no link by evolution)

Horizontal or Lateral Transfer

- DNA is transferred between species, not inherited by descent
- Figure 3.3 D
- A jump between leaves of a species tree
- Sequences are *more similar* than one would expect from distance between species
- Example: transposons, or other “mobile” elements carried by viruses