

Web-base and standalone blast

Vitor Pavinato
correapavinato.1@osu.edu

Troubleshooting

- Is there any question about the installation?
 - Homebrew (macOS only)
 - Git (windows installation)
 - Blast+ (windows installation)
- Is there any problem on how to make a copy of the repository?
 - Windows users: I noticed there are some problems on how to navigate between volumes (C:\ to D:\)?
- Do you want me to review some useful command-line commands?

Before we start

- Experience with BLAST and computational biology (in general);
- There are five exercises we can go through today (or less, depending on the time spent in each one);
- The goal is to follow the instructions and try to answer the questions;
- We are going to split the class in four groups (each group should have at least one student with a standalone version of blast installed);
- Each exercise 20-30 min;
- Choose one person to report back at discussion at the end of the activity;
- 5-10min break after ~1:20h
- After class (more likely next week, I will post an exercise in the repository page);

Background

- Did you watch the videos linked in the repository page?
- Why align sequences?
- Which alignments are significant?
- What is local alignment?
- What does BLAST tell you?
- How are scores and e-values related?
- Can e-values be directly compared when searching databases of different sizes? Why or why not? What about the bit scores with the same question? What if you change the scoring matrix? (we are going to have an example below);
- Is there a magic cutoff for poor e-values? How might you determine an e-value cutoff?

Motivation

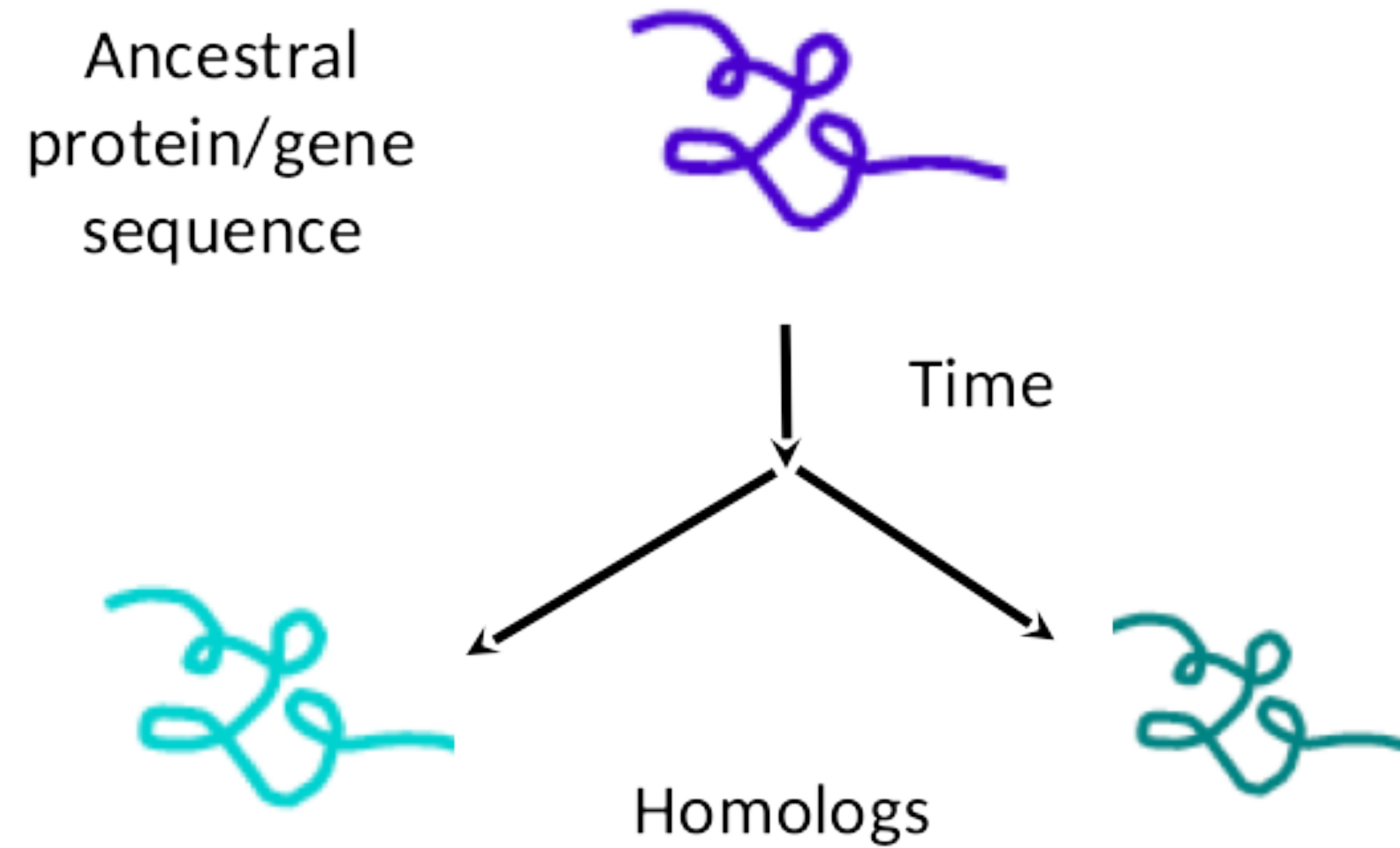
Let's say you want to know which genes are differently expressed in your biological system when it is exposed to different conditions (blood-fed mosquitos vs. non-blood-fed ones). You extracted the whole transcriptome, sequenced (RNA-cDNA) and you use that data to assemble a whole transcriptome (RNAseq lecture). You accessed which genes were differentially expressed between two conditions, and you find one transcript highly expressed in condition A (blood-fed mosquitos). You take the sequence that represents the sequence of the mRNA and you search it against the Mosquito genome using BLASTN.

Is the alignment you found significant?

Is this likely to represent an homologous RNA?

- **Local Alignment:**
 - **Find stretches of high similarity between two sequences;**
 - **Don't require the alignment of the whole sequence**

Importance of Similarity



Homology: based on sequence identity

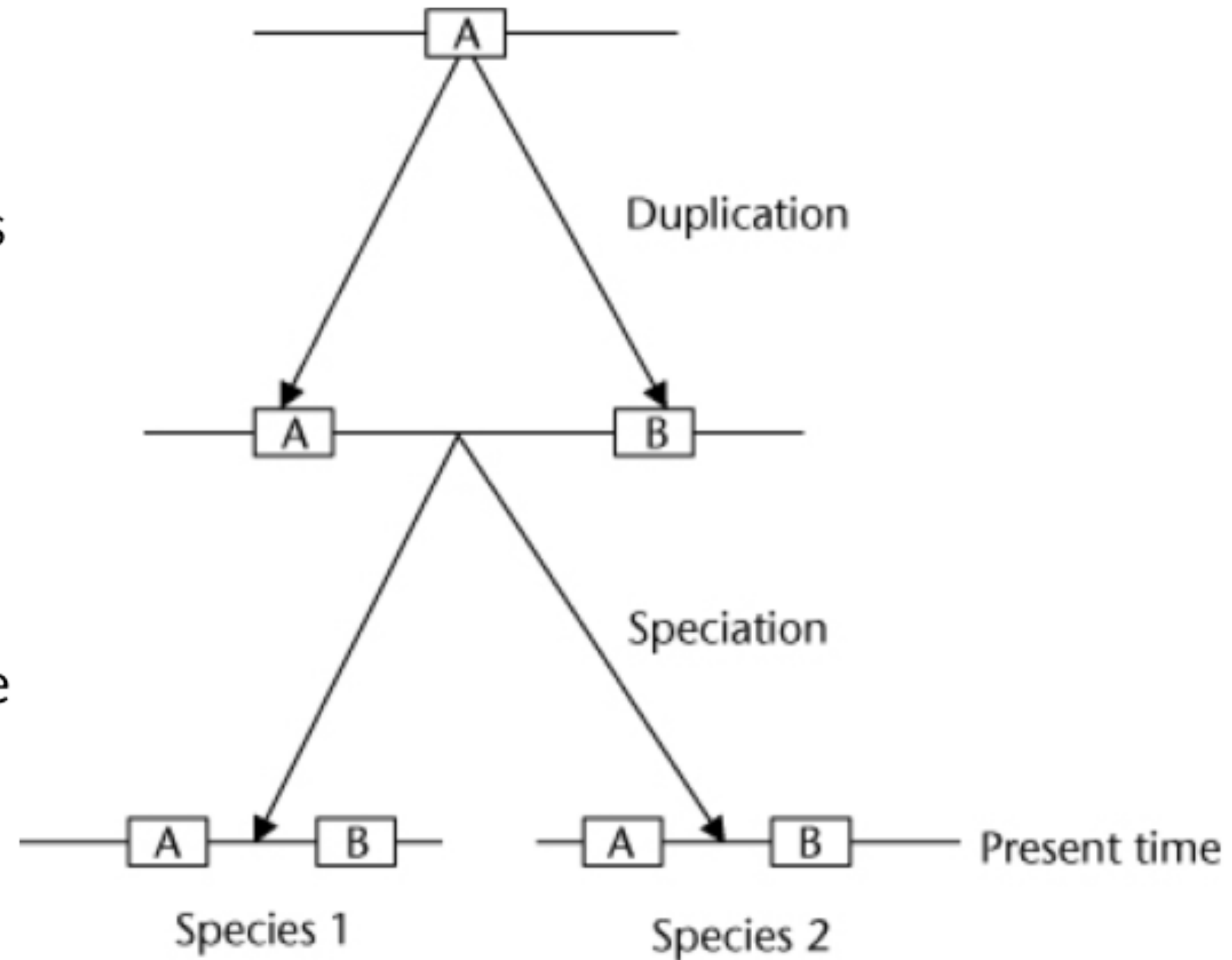
Bordoli L. "Similarity Searches on Sequence Databases".
Powerpoint presentation. EMBnet Course, Basel, October 2003.



Types of Homologs

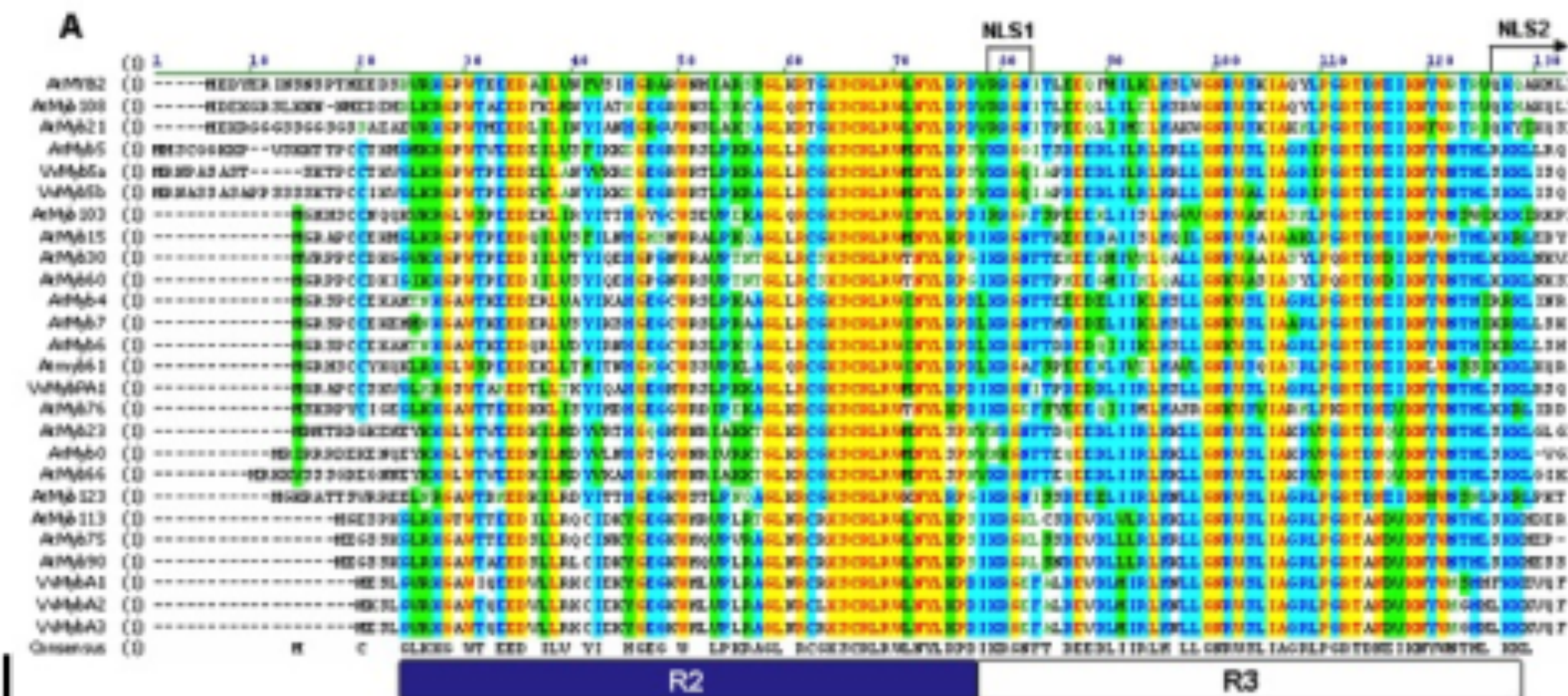
Paralogs: duplicate genes within one species

Orthologs: the same gene within different species



Sequence Conservation

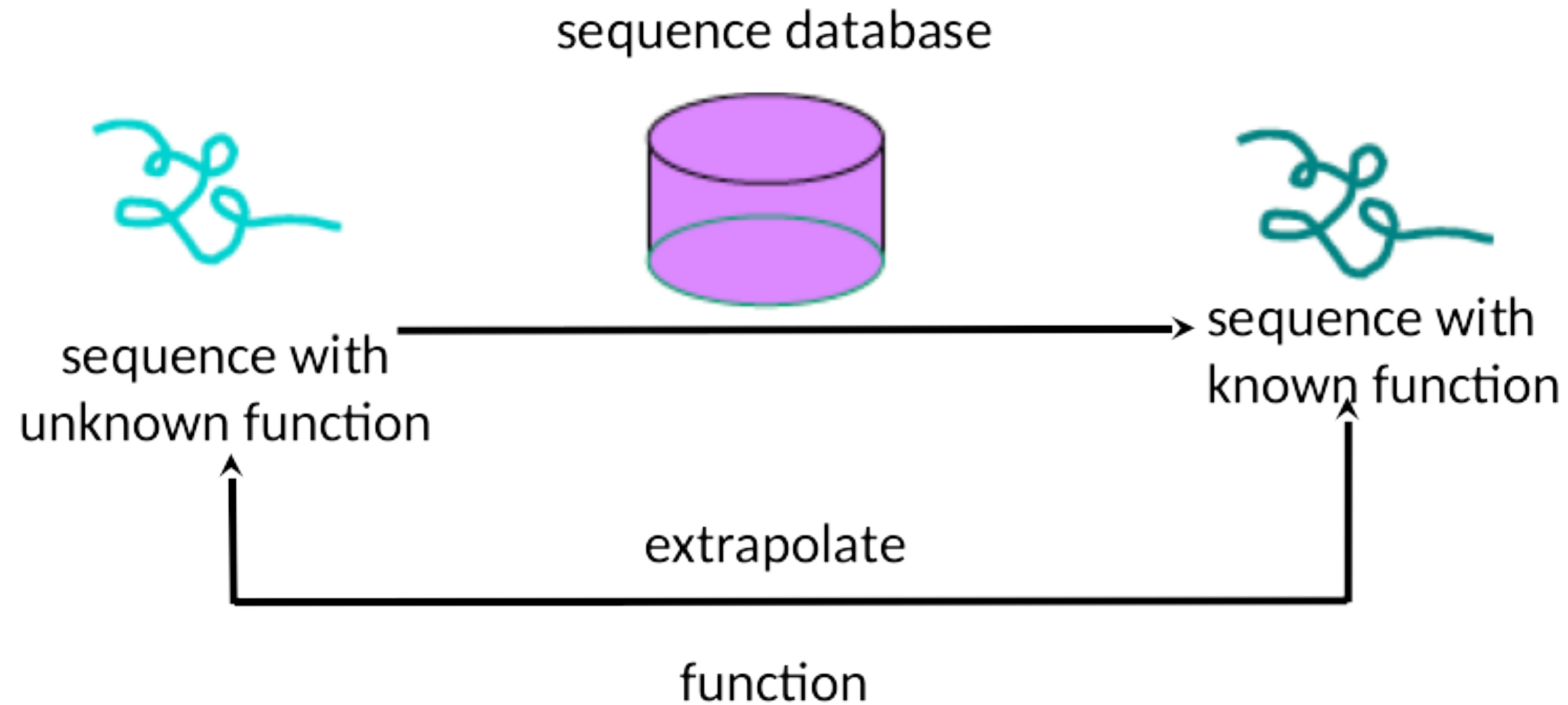
- A sequence of amino acids or nucleotides that is similar across species
- Pattern of most frequent residues (polypeptide) or bases (DNA or RNA) found at a particular position
- Consensus sequence may define a motif of biological significance



Matus JL, et al. (2008). *BMC Plant Biol.* **8**:83

GENOME SOLVER

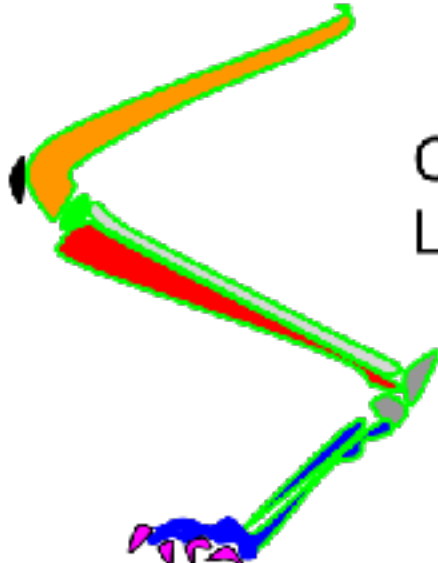
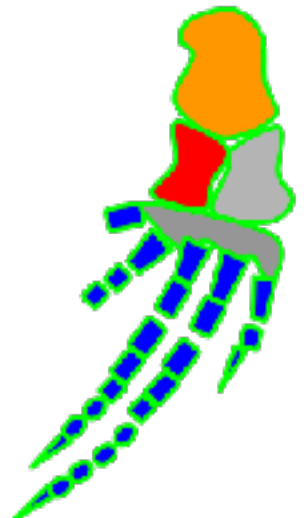

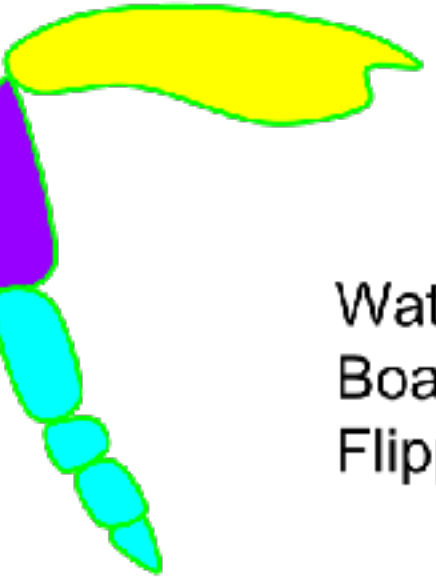
Sequence Similarity



Bordoli L. "Similarity Searches on Sequence Databases". Powerpoint presentation. EMBnet Course, Basel, October 2003.



Convergent Evolution

	Analogous Leg	Analogous Flipper
Homologous: Mammals	 <p>Cat Leg</p>	 <p>Whale Flipper</p>
Homologous: Insects	 <p>Preying Mantis Leg</p>	 <p>Water Boatman Flipper Leg</p>

Same idea for proteins: similar structure but no significant similarity in sequence

