BI694

Bioinformatics & Phylogenetics

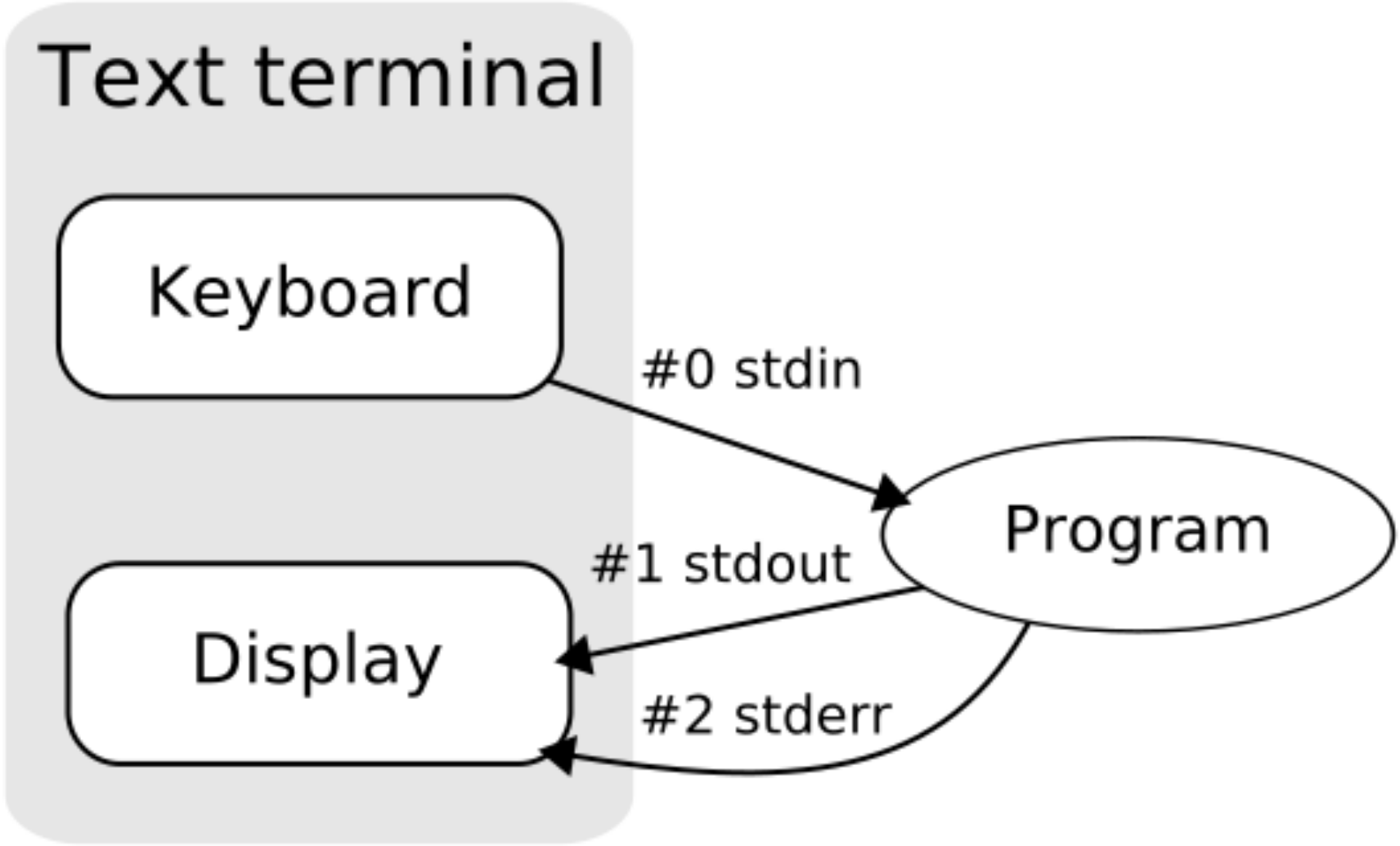
Winter Semester 2017

WEEK 3

(several slides courtesy of Maya Schushan and John R. Rose)

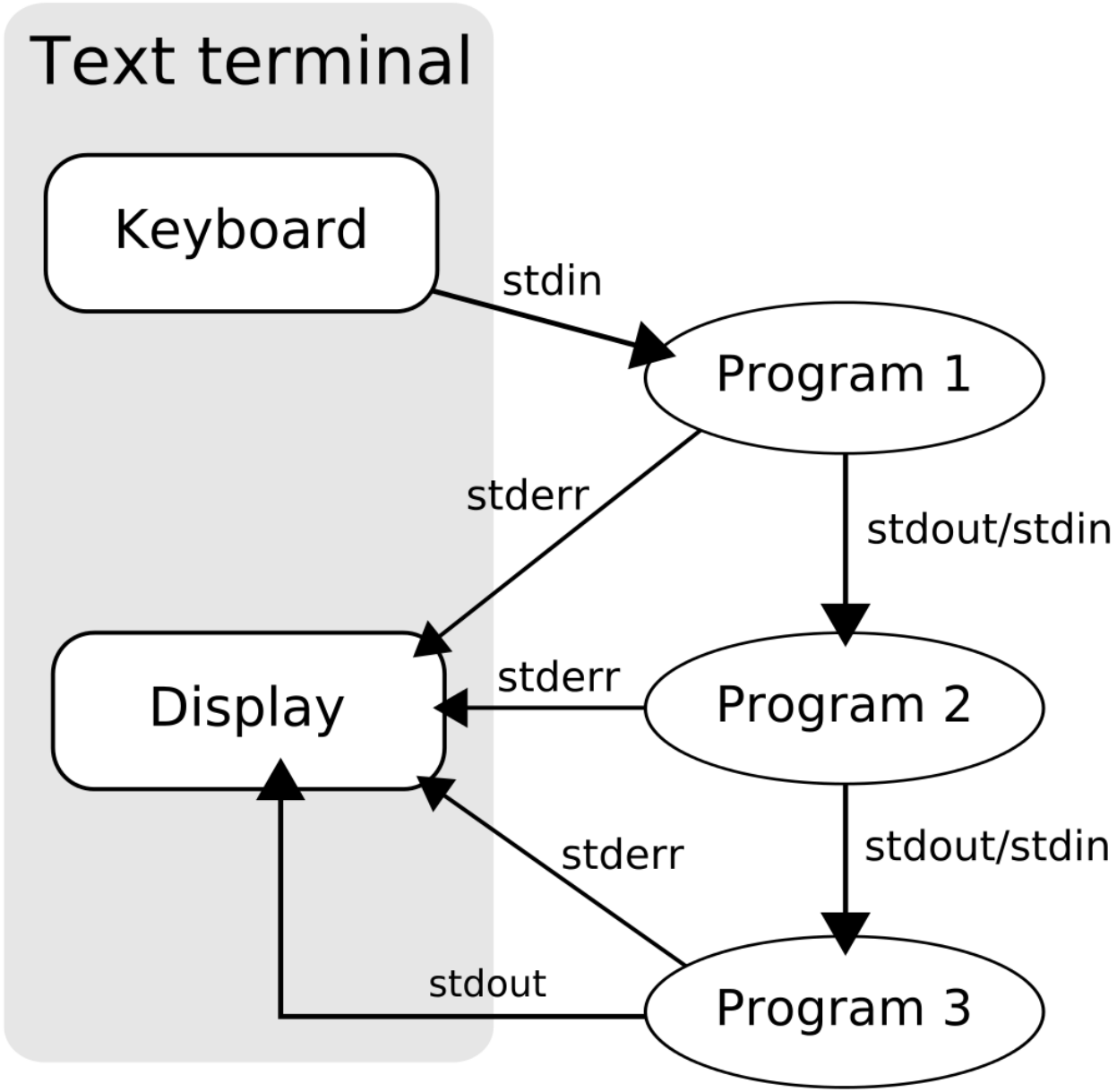
Streams and pipes

- pipes – |
- redirects - > < << >>
- Separating streams – 1> 2>



Streams and pipes

- pipes – |
- redirects - > < << >>
- Separating streams – 1> 2>



Text editors

- **less** – text viewer
- nano – intuitive and easy but limited
- **vi**(m) – available everywhere and powerful
- emacs – powerful but too bloated for my taste
- geany – light weight IDE

Let’s get some data…

- Navigate to week 3 of the github course repository
- Download the **coral catalog csv** (through a browser or the terminal with wget)
- Let’s look at it with **less**, **head**, **tail**
- How can we extract information quickly? (grep and awk)

grep, sed, awk

- **grep** – powerful pattern searches, including regex
- **sed** – stream editor, search and replace (perl…)
- **awk** – scanning and processing programming language

Let’s look at some sequence data

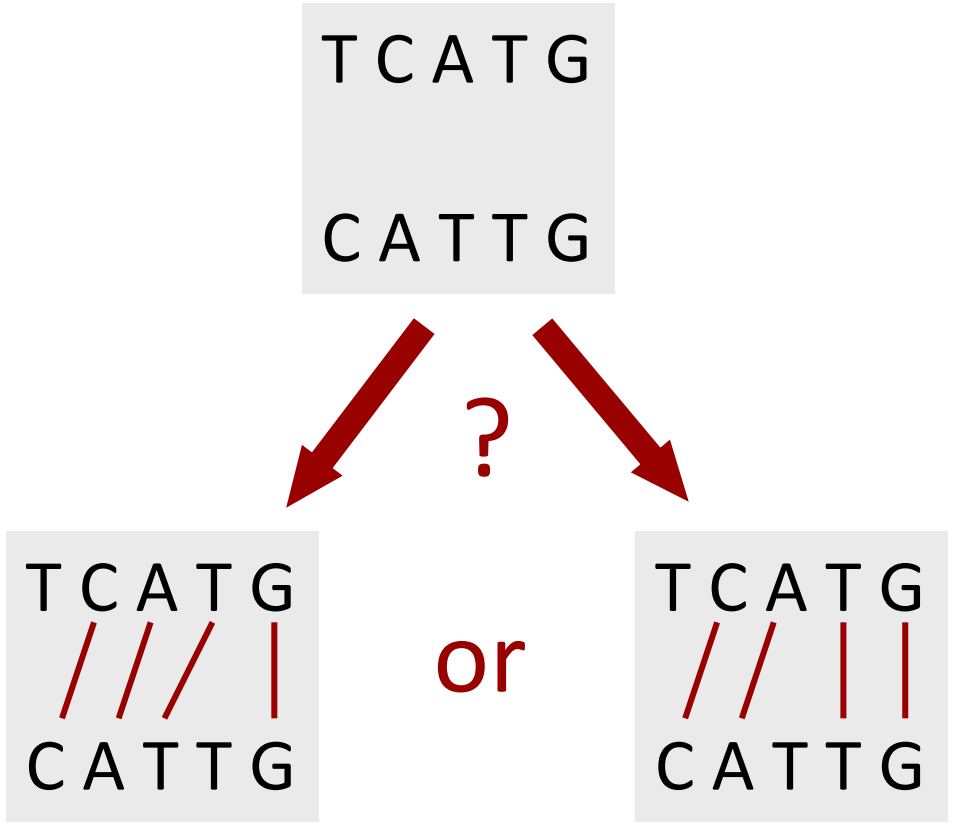
- Navigate to week 3 of the github course repository
- Download OTUs.txt
- Let’s examine it? What file format is this?
- grep –c ‘>’
- Seaview
- Phylip and Nexus formats

Aligning sequences

- Navigate to week 3 of the github course repository
- Download Jellys.fst
- Are these sequences aligned?

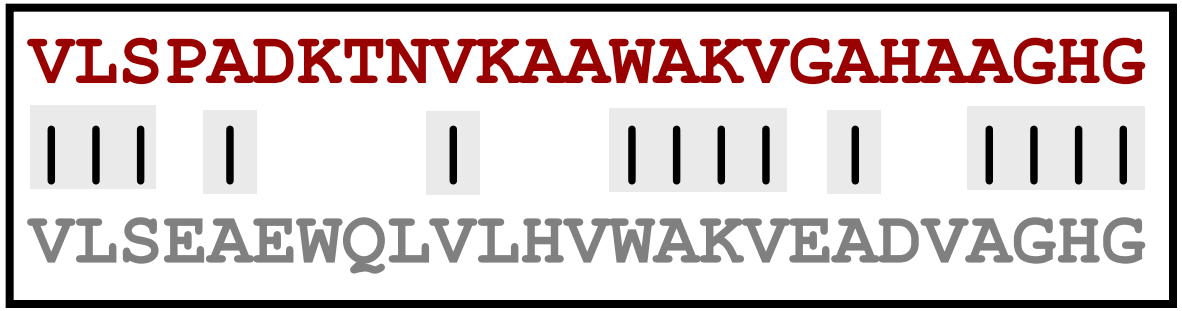
What is a sequence alignment?

Process of lining-up 2 or more sequences to achieve maximum level of identity, in order to find homologies.



What is a multiple sequence alignment?

- Comparing 2 (pairwise) or more (multiple) sequences.
- Searching for a series of identical or similar characters in the sequences.



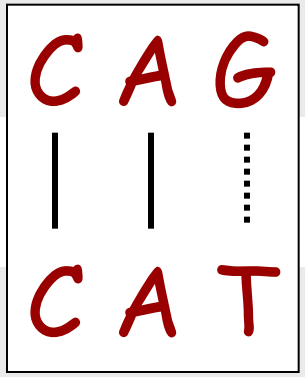
Defining Terms

- **Homology:**

Relation of sequences which is a result of shared from common ancestry

- **Identity:**

Sequences or Sub-sequences that are invariant.



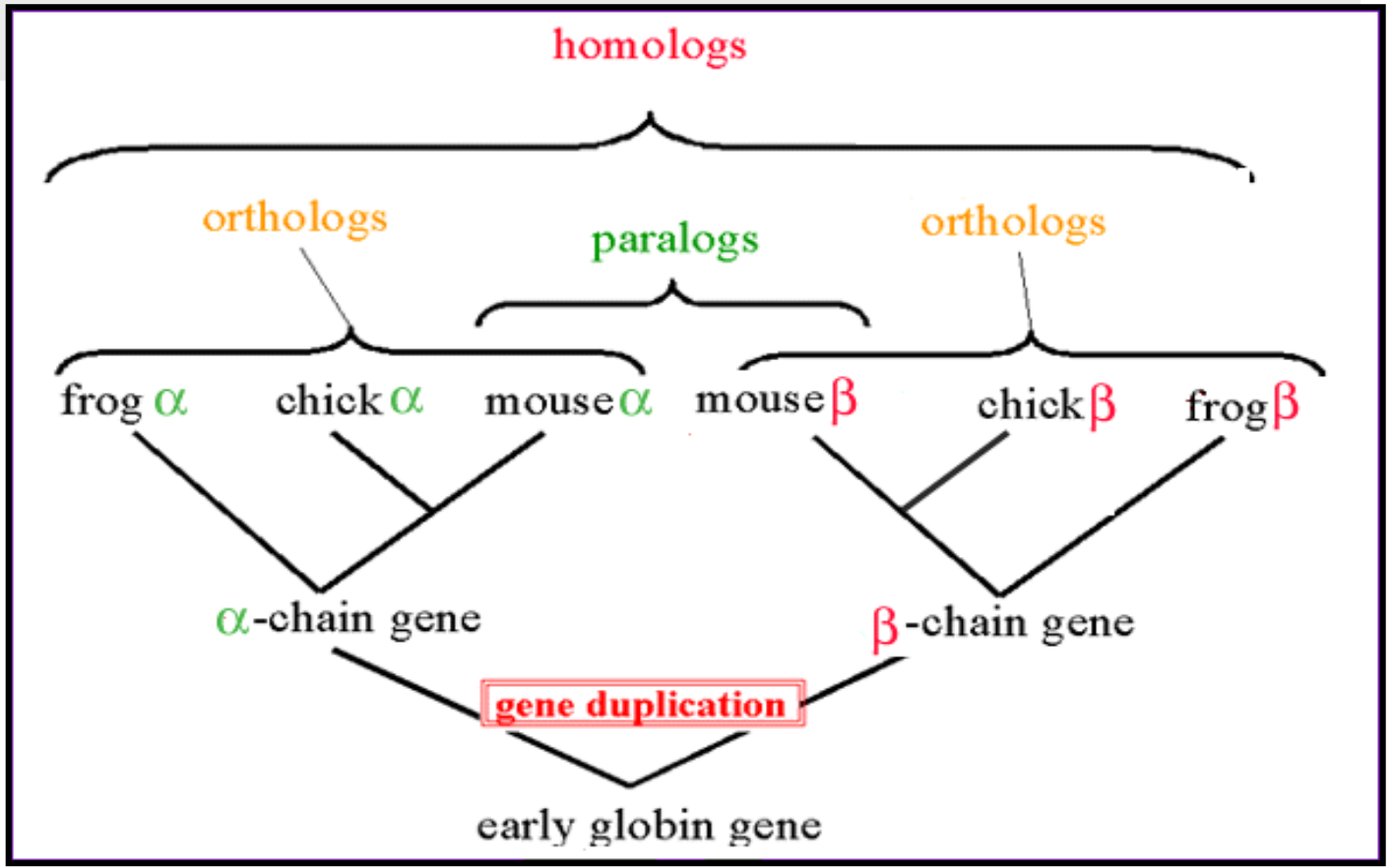
- **Similarity:**

Sequences or Sub-sequences that are related.

Defining Terms

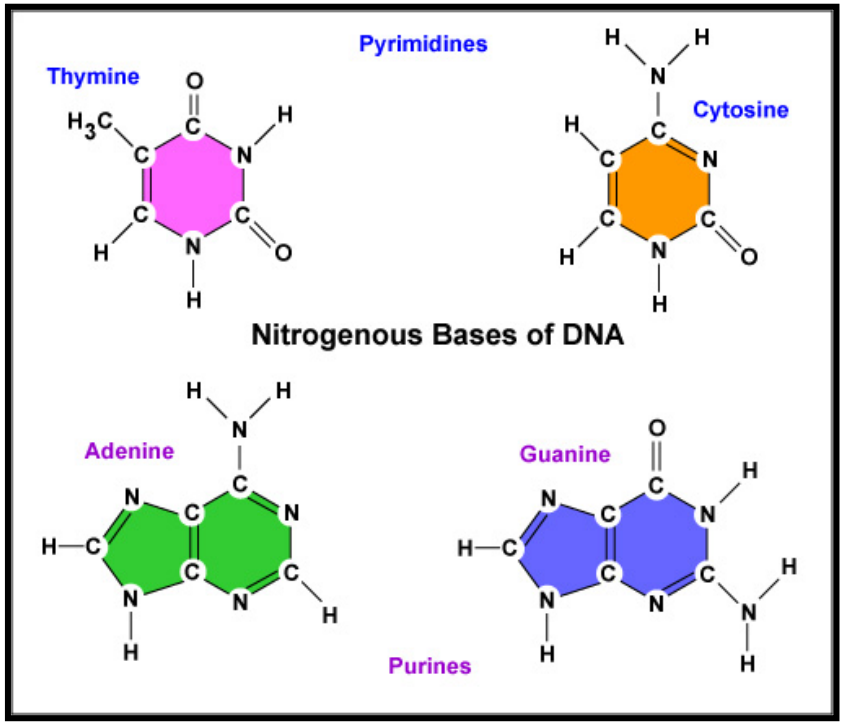
- **<u>Homology:</u>**

Relation of sequences which is a result of shared from common ancestry



Reproduced from NCBI education website

Defining Terms



- **Identity:**

Sequences or Sub-sequences that are invariant.

C A G
| | ⋮
C A T

- **Similarity:**

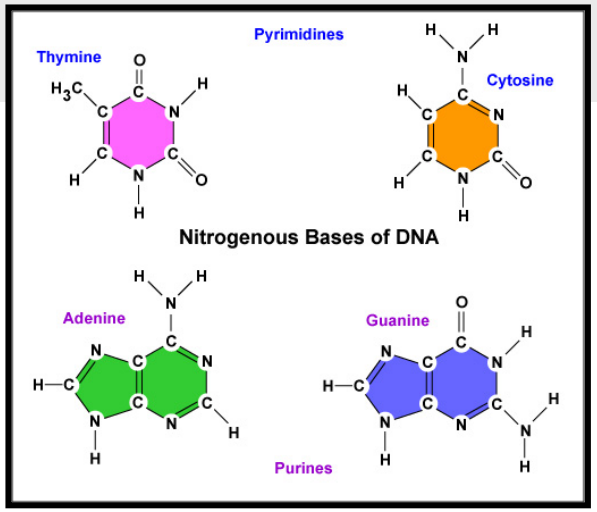
Sequences or Sub-sequences that are related.

DNA (RNA) similarity scoring

- Transitions – purine to purine or pyrmidine to pyrmidine

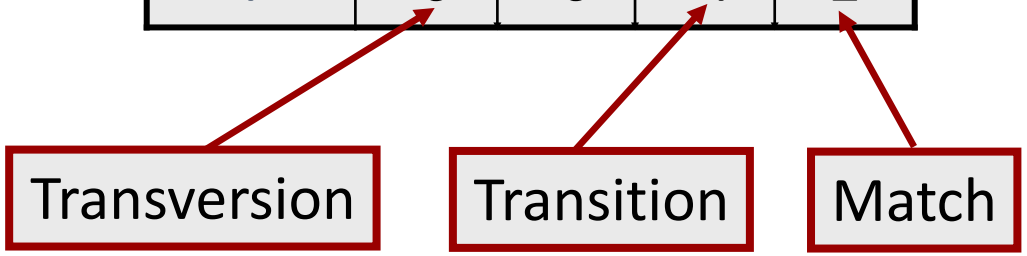
  (4 possibilities)
- Transversions – purine to pyrmidine or pyrmidine to purine

  (8 possibilities)
- By chance alone transversions should occur twice as often as transitions.
- De-facto transitions are more frequent than transversions.



DNA (RNA) similarity scoring

From To	A	G	C	T
A	2			
G	-4	2		
C	-6	-6	2	
T	-6	-6	-4	2



Protein similarity scoring

- Observation: some substitutions are more frequent than others, e.g., chemically similar amino acids
- As for DNA, protein matrices define the probabilities of change between the different amino acids
- Popular matrices are based on empirical data: PAM & BLOSUM

Protein similarity scoring

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Why align sequences?

Predict characteristics of a protein

VTIS**CTGS**SSNIGAG-NHVK**WYQQ**LPG

VTIS**CTGS**SSNIGS--ITVN**WYQQ**LPG

LRLS**CTGS**GFIFSS--YAMY**WYQQ**APG

LSLT**CTGS**GTSFDD-QYYST**WYQQ**PPG

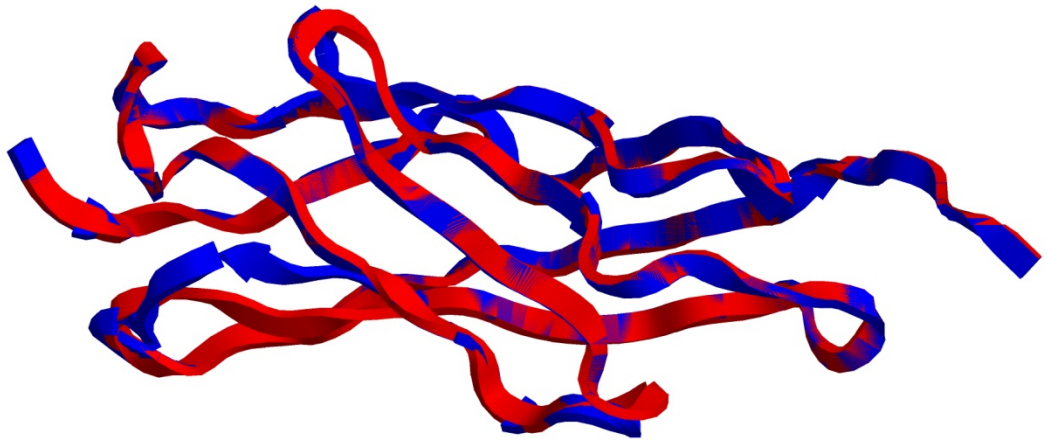
Why align sequences?

```
cad23_dom3  NLAIIDVQDMDPIFINLPYSTNIYEHSPPGTTVRIITAIDQDKGR---PRGIGYTIVSGN
1EDH        EIVITDQNDNRPEFTQEVFEGSVAEGAVPGTSVMKVSATDADDDVNTYNAAIAYTIVSQD

cad23_dom3  ----TNSIFALDYISGVL-TLNGLLDRENPLYSHGFILTVKGTELNDDRTPSDATVTTTF
1EDH        PELPHKNMFTVNRDTGVISVLTSGLDRES--YPT-YTLVVQAADLQGE----GLSTTAKA

cad23_dom3  NILVIDINDNAPEFNSSEYSVAITELAQVGFALPLF
1EDH        VITVKDINDNAPVFNPS------------------
```

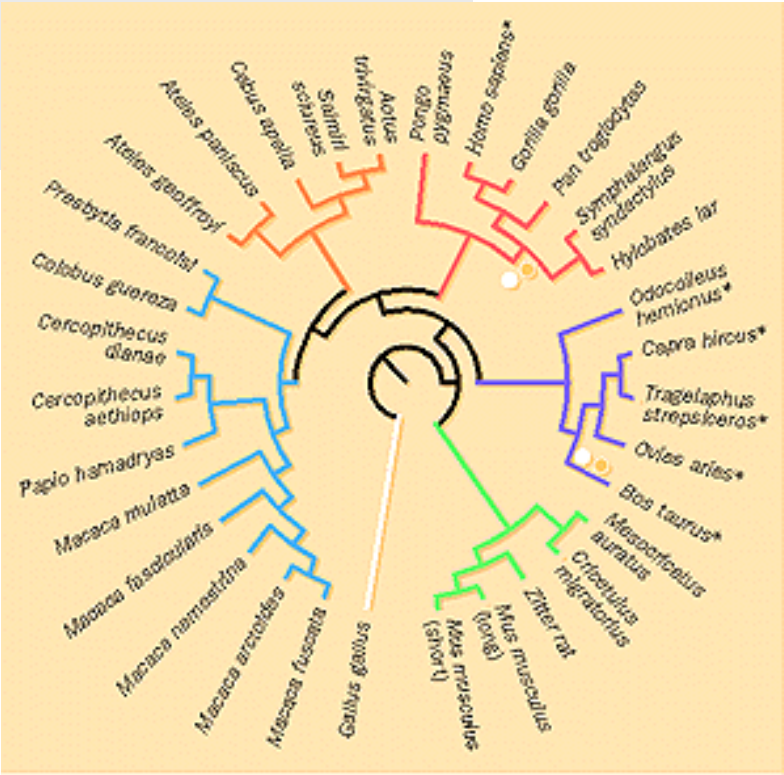
A model is generated according to a template structure of a homologous protein



Why align sequences?

Learn about evolutionary relationships –

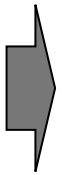
- Two sequences from different organisms are similar ➔ they may have a common ancestor.
- Needed for construction of phylogenetic trees



Pairwise versus Multiple Sequence Alignment

Pairwise:
For 2 sequences

```
F G K – G K G
F G K F G K G
```



MSA:
For more than 2 sequences

```
F G K – G K G
F G K F G K G
– G K Q G K G
– – K F G K G
```

Multiple sequence alignment

- By definition a multiple sequence alignment (MSA) is an alignment of 3 or more sequences (amino acid or nucleotide)
- Pairwise alignments align 2 sequences – useful for searchinga database to find sequence matching query best (eg, BLAST)
- Generally 3 approaches to MSA:
  - direct
  - progressive (hierarchial)
  - iterative

Direct sequence alignment

- Amino acids (AAs) are aligned using gap penalty and substitution model based on chemical properties of AAs
- Nucleotides are aligned using gap penalty and mismatch score
- Problem:
  - computationally expensive
  - for $n > 2$ sequences n-dimensional matrix of pairwise alignments
  - scales exponentially

Progressive alignment

Basic method:

1. Distance matrix is calculated
   - Distances are pairwise alignment scores
   - Gives divergence of each pair of sequences
2. Guide tree built from distance matrix
3. Progressive alignment according to guide tree
   - Branching order of tree specifies alignment order
   - Alignment progresses from leaves to root.

Progressive alignment

Distance matrix/pairwise alignments phase

- Fast approximation:
  - *k*-tuple (ordered list of length *k*) matches for identical residues (AAs or nucleotides), typically
    - 1 to 2 for proteins
    - 2 to 4 for nucleotide sequences
  - Scores are calculated as: (*k*-tuple matches) – fixed penalty per gap
  - Score is initially calculated as a percent identity score.
  - Distance = 1.0 – (score/100)

Clustal alignments

- Based on the idea that the sequences we want to align are phylogenetically related: a pairwise alignment algorithm is used iteratively, first to align the most closely related pair of sequences, then the next most closely related one to that pair.

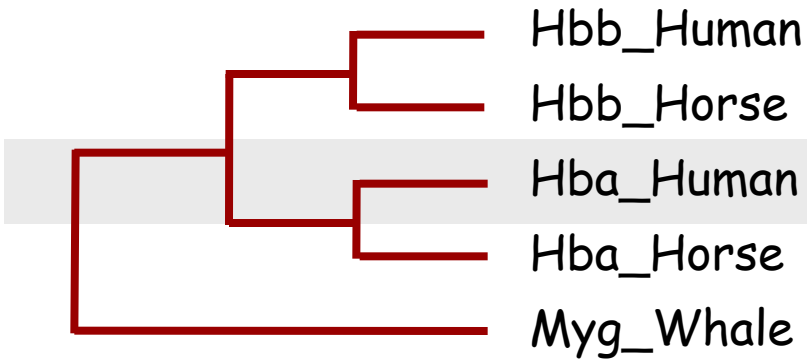
- Rule "once a gap, always a gap": The gaps between more similar pairs of sequences should not be affected by more distantly related ones.

Clustal alignments

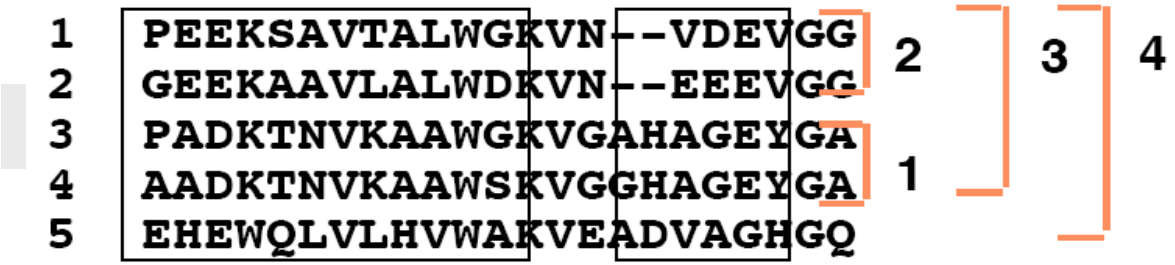
Hbb_Human 1
Hbb_Horse 2
Hba_Human 3
Hba_Horse 4
Myg_Whale 5

1	2	3	4	5
-				
17	-			
59	60	-		
59	59	13	-	
77	77	75	75	-

1. Quick pairwise alignment calculate distance matrix

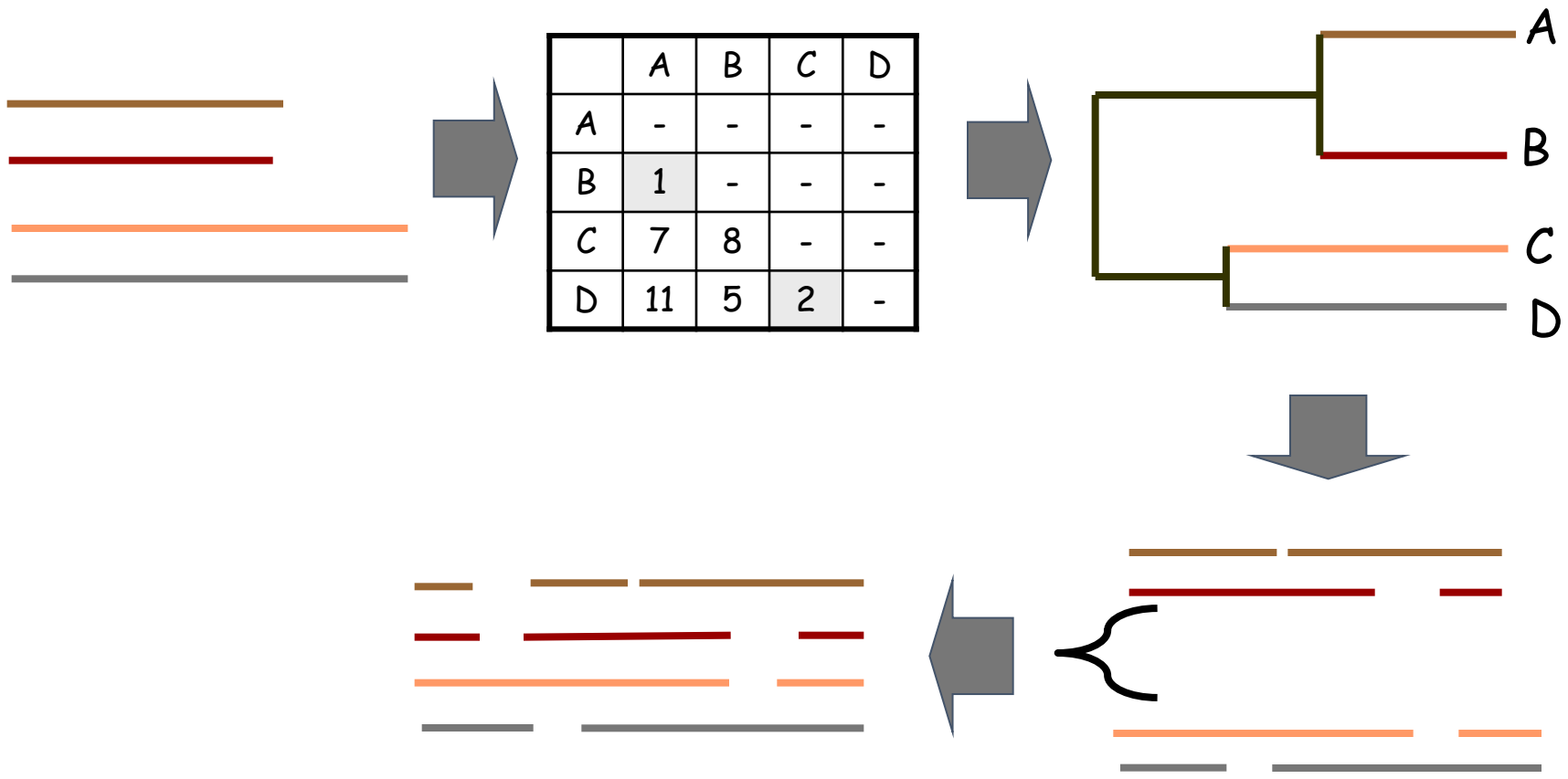


2. Build a guide tree using clustering or NJ



3. Progressive alignment following guide tree

Clustal alignments

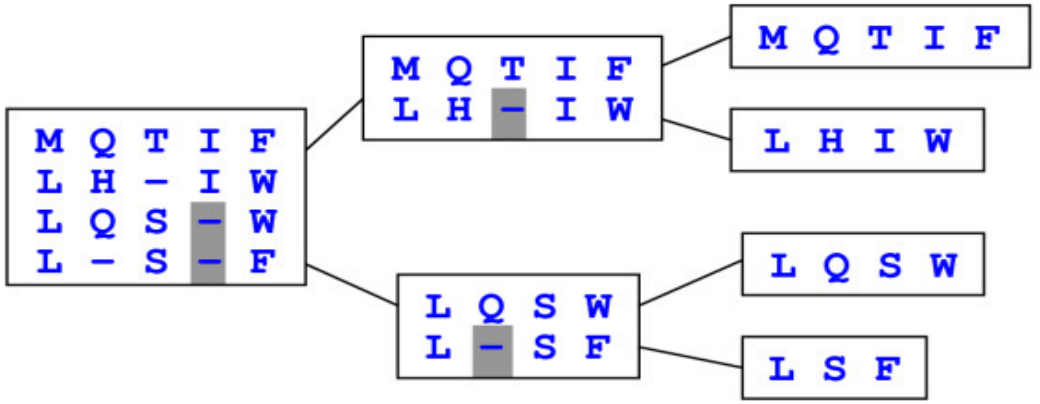


Clustal alignments

- Progressive alignment not guaranteed to find global optimum
- Errors made at any stage in growing the MSA are propagated through to the final result

Iterative alignment

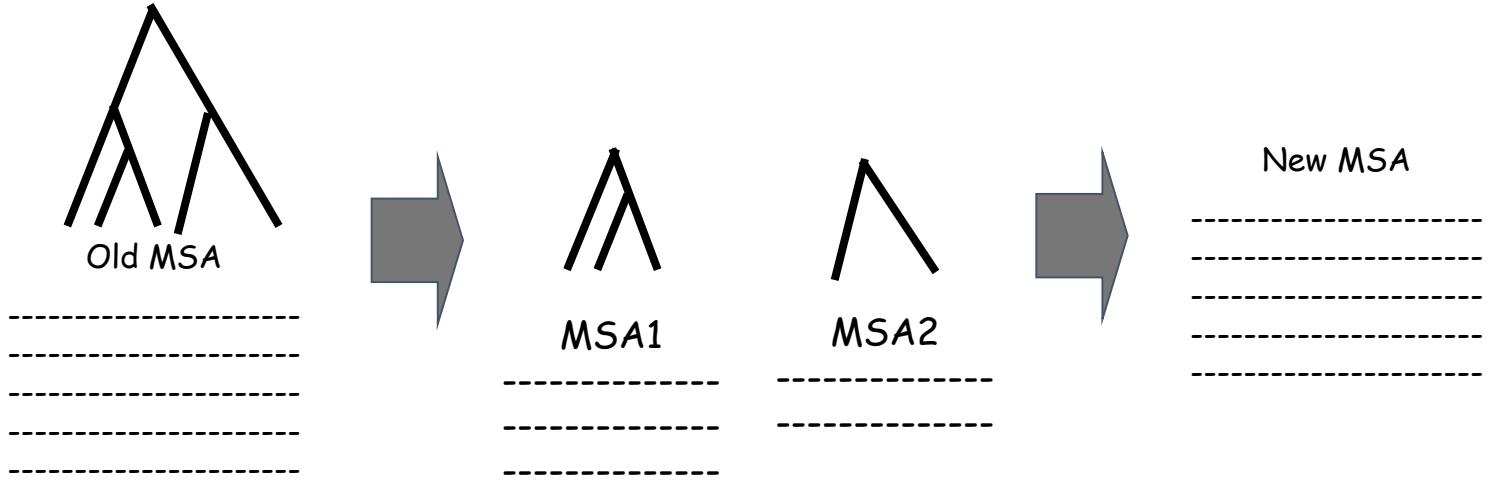
- MUSCLE: popular MSA software
- Considered highly accurate MSA software
- The basic idea: iterative progressive alignment → refinement of initial results



Edgar, R.C., 2004

MUSCLE alignment

- An edge is chosen from the progressive alignment tree.
- The tree is divided into two subtrees by deleting this edge.
- The MSA from each subtree is computed by progressive alignment.
- The two MSAs are aligned, generating an entire new MSA
- If the new MSA achieves higher score than the previous ➔ keep it



MUSCLE alignment

It's a bit more complicated…