From Alignment to Phylogeny

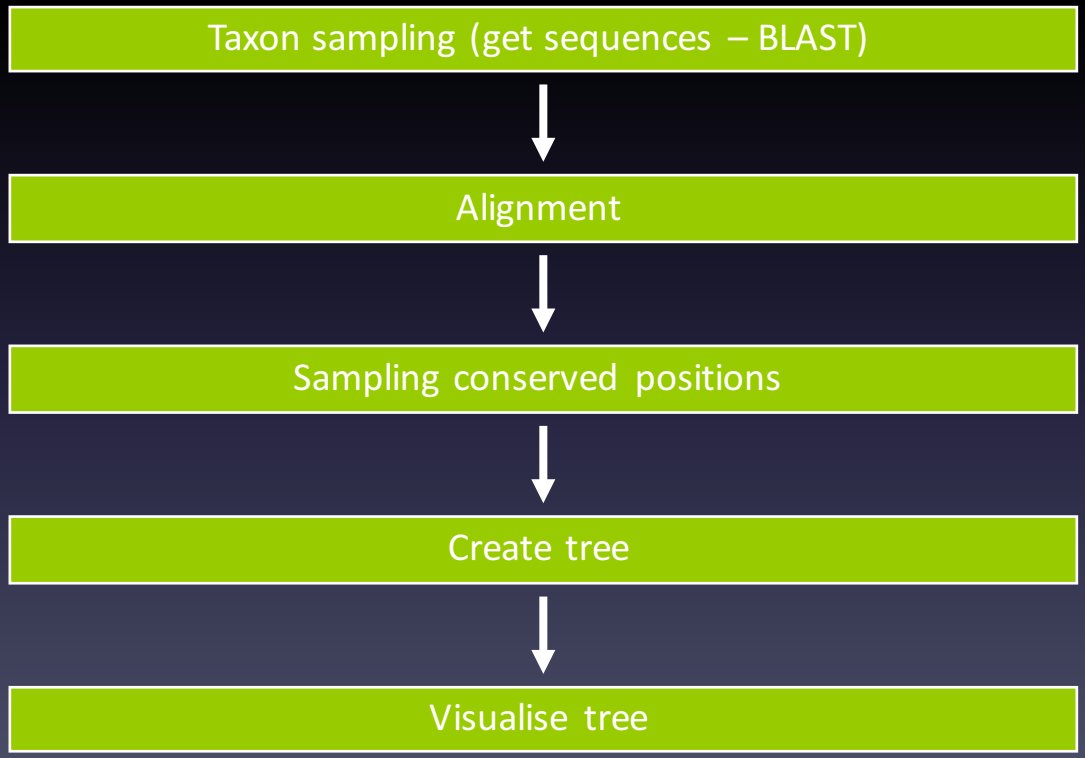
Introduction to NGS Data and Analysis

Lecture 4

Jia-Ming Chang

政大資訊科學系
Department of Computer Science, National Chengchi University

Outline


Taxon sampling (get sequences – BLAST)
Alignment
Sampling conserved positions
Create tree
Visualise tree

SEQUENCE ALIGNMENT – TWO SEQUENCES

Why Does It Make Sense To Align Sequences ?

-Evolution is our Real Tool.

-Nature is LAZY and Keeps re-using Stuff.

-Evolution is mostly DIVERGEANT

Same Sequence ⇔ Same Ancestor

Comparing Is Reconstructing Evolution

Courtesy American Rivers

Adapted from Cedric Notredame

An Alignment is a **STORY**

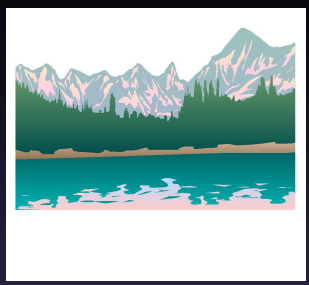
ADKPKRPLSAYMLWLN

ADKPKRPLSAYMLWLN

ADKPRRPLS-YMLWLN

Mutations
+
Selection

ADKPKRPLSAYMLWLN



ADKPKRPKPRLSAYMLWLN

Insertion

Deletion

ADKPRRP---LS-YMLWLN
ADKPKRPKPRLSAYMLWLN

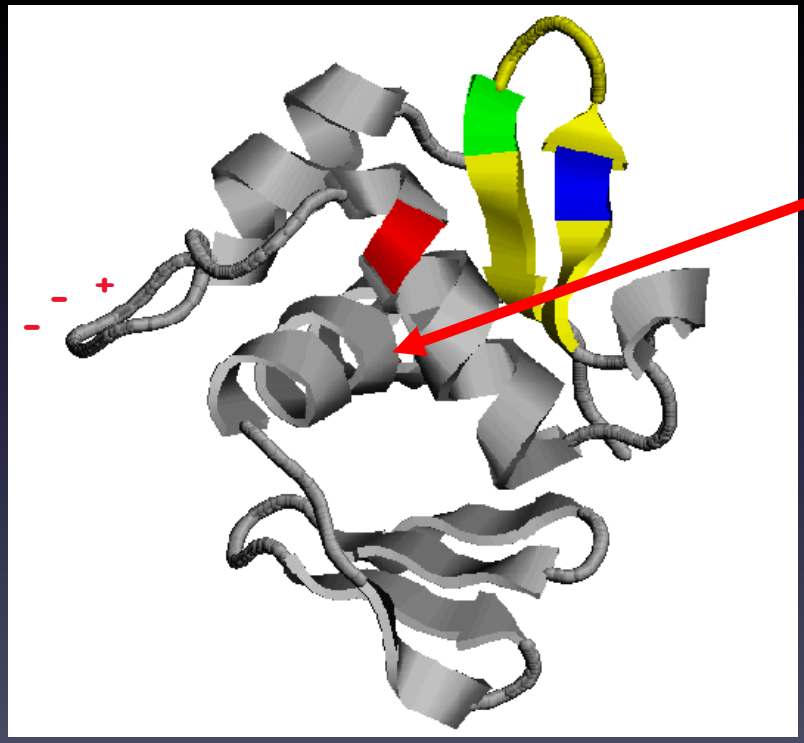
Mutation

Adapted from Cedric Notredame

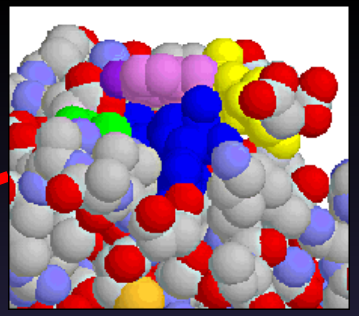
How Do Sequences Evolve ?

In a structure, each Amino Acid plays a Special Role

On the surface,
CHARGE MATTERS



OmpR, Cter Domain

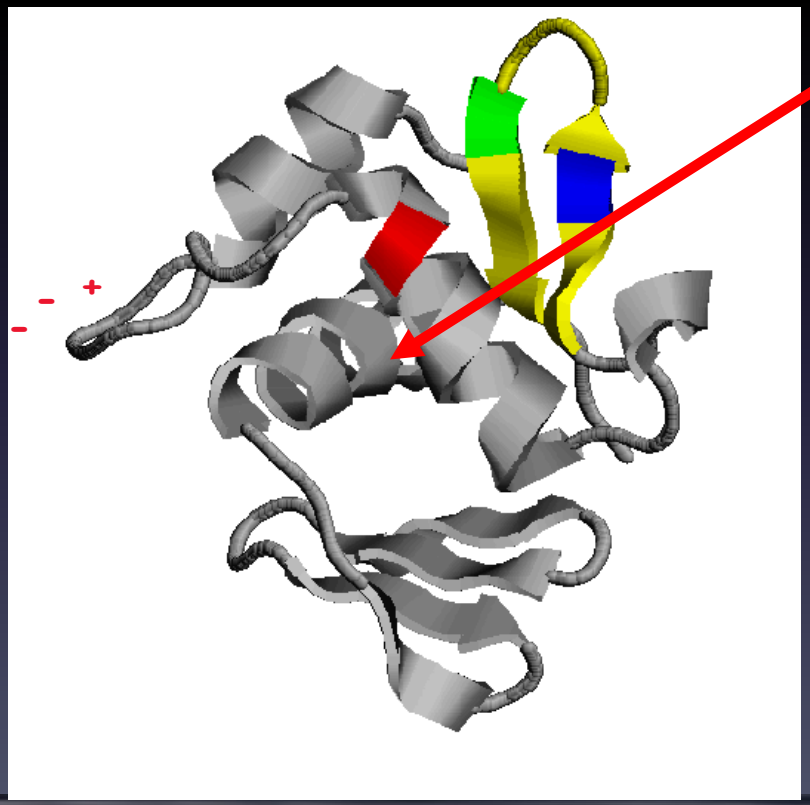


In the core,
SIZE MATTERS

How Do Sequences Evolve ?

Accepted Mutations Depend on the Structure

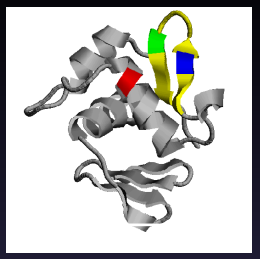
Charged -> Charged
Small <-> Big or Small
DELETIONS



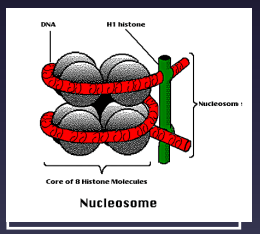
Big -> Big
Small ->Small
NO DELETION

How Can We Compare Sequences ?

To Compare Two Sequences, We need:



Their Structure



Their Function

We Do Not
Have Them !!!

How Can We Compare Sequences ?

To Compare Sequences, We need to Compare Residues

We Need to Know How Much it COSTS to SUBSTITUTE

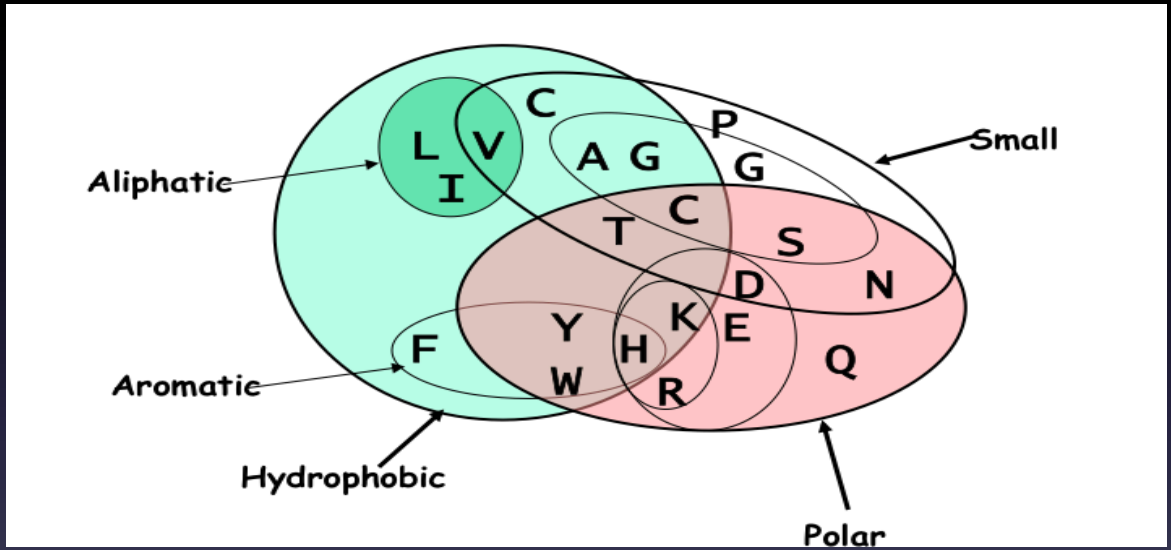
an Alanine into an Isoleucine
a Tryptophan into a Glycine
…

The table that contains the costs for all the possible substitutions is called the SUBSTITUTION MATRIX

How to derive that matrix?

How Can We Compare Sequences ?

Using Knowledge Could Work



But we do not know enough about Evolution and Structure.

Using Data works better.

Adapted from Cedric Notredame

How Can We Compare Sequences ?
Making a Substitution Matrix

- Take 100 nice pairs of Protein Sequences, easy to align (80% identical).
- Align them…
- Count each mutations in the alignments
  - 25 Tryptophans into phenylalanine
  - 30 Isoleucine into Leucine
  - …

- For each mutation, set the substitution score to the log odd ratio:

$$\text{Log}\left(\frac{\text{Observed}}{\text{Expected by chance}}\right)$$

$$\log\left(\frac{p_{ij}}{q_i * q_j}\right)$$

Adapted from Cedric Notredame

How Can We Compare Sequences ?
Making a Substitution Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

The Diagonal Indicates How Conserved a residue tends to be. W is VERY Conserved

Some Residues are Easier To mutate into other similar

PAM units

- PAM (point accepted mutation) is a unit of evolutionary distance between 2 amino acid sequences*
- 1 PAM = 1 accepted point-mutation (no insertions or deletions) event per 100 aa
- 200 PAM = 200 point mutations/100 aa (assumes mutations occur multiple times at any given position)
- 2 sequences diverged by 200 PAM ≅ 25% identity
- *PAM is also sometimes defined as "percent accepted mutation"

Adapted from Cedric Notredame

Scoring an Alignment

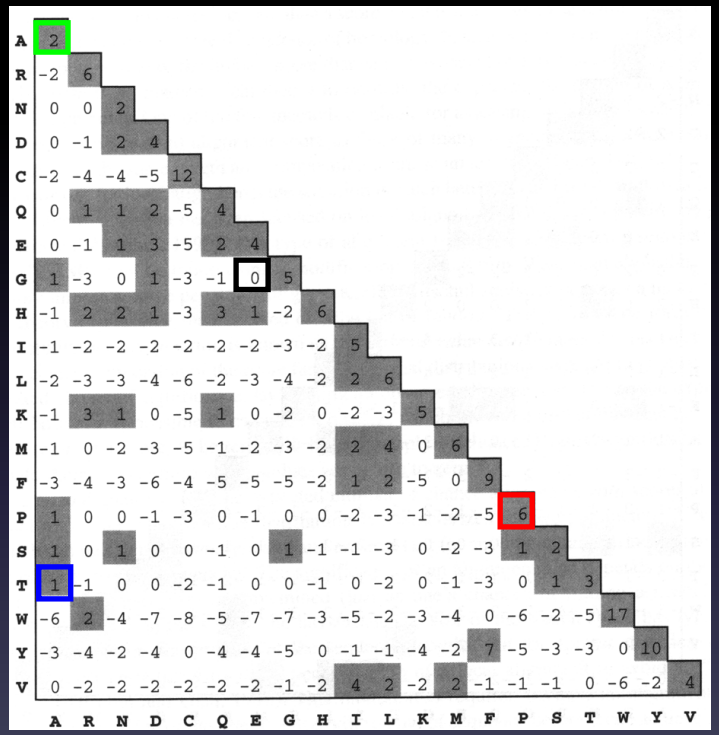
Most popular Subsitution Matrices

- PAM250
- Blosum62 (Most widely used)

Raw Score

```
T P E A
| | | |
A P G A
```

Score = 1 + 6 + 0 + 2 = 9

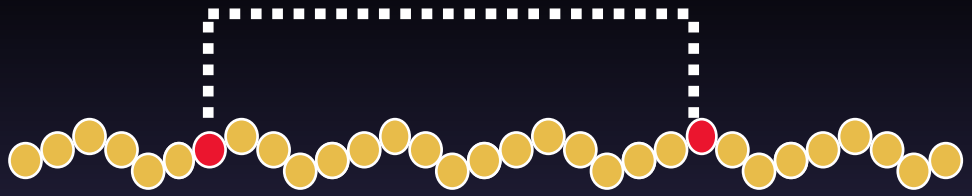


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

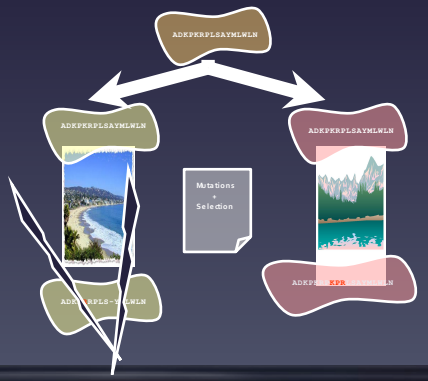
- Question: Is it possible to get such a good alignment by chance only?

How Can We Compare Sequences ?
Limits of the substitution Matrices

They ignore non-local interactions and Assume that identical residues are equal



They assume evolution rate to be constant



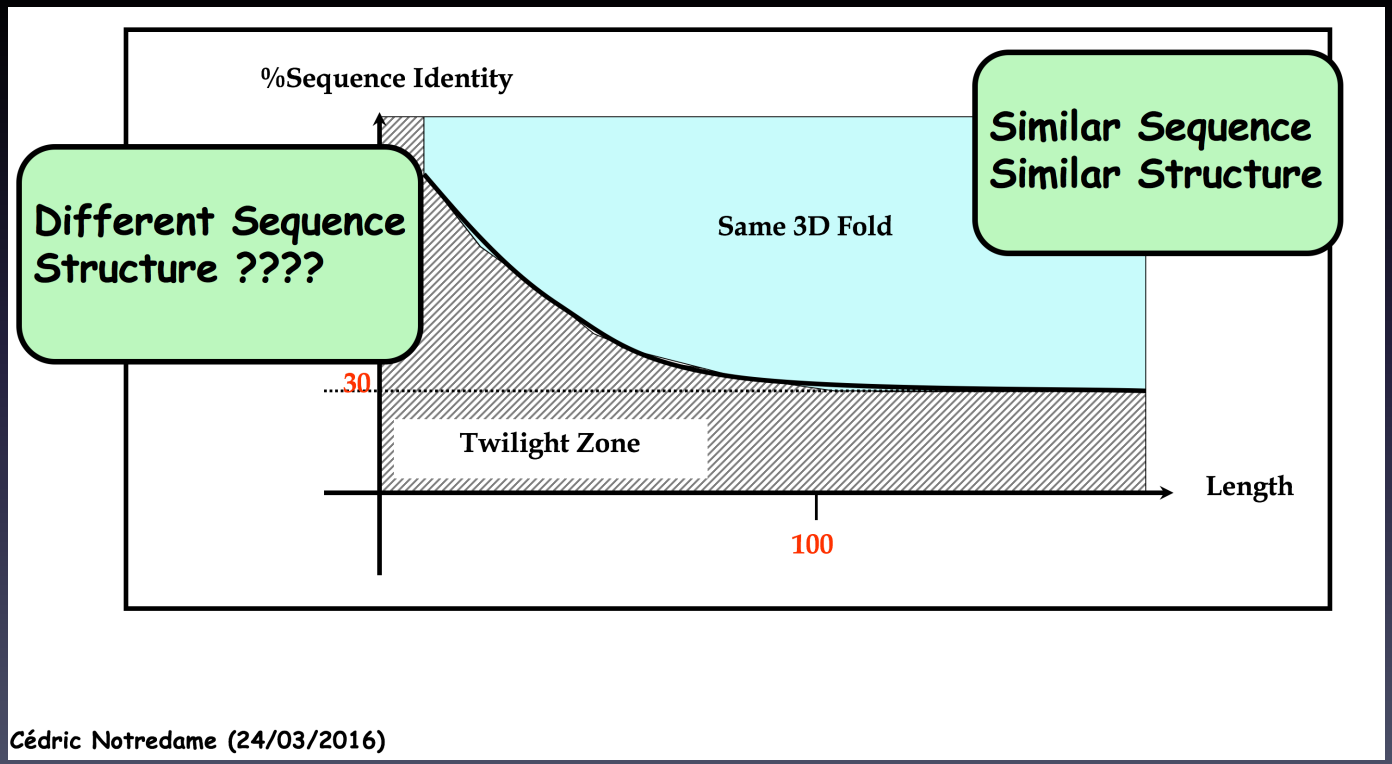
Adapted from Cedric Notredame

How Can We Compare Sequences ?
Limits of the substitution Matrices

I know… But at least, could I get some idea of when they are likely to do all right

How Can We Compare Sequences ? *The Twilight Zone*

Substitution Matrices Work Reasonably Well on Sequences that have more than 30 % identity over more than 100 residues



How Can We Compare Sequences ?
Which Matrix Shall I use

PAM: Distant Proteins ⇔High Index (PAM 350)
BLOSUM: Distant Proteins ⇔ Low Index (Blosum30)

Choosing The Right Matrix may be Tricky…

- GONNET 250> BLOSUM62>PAM 250.
- But This will depend on:
  - The Family.
  - The Program Used and Its Tuning.
- Insertions, Deletions?

Adapted from Cedric Notredame

HOW Can we Align Two Sequences ?

Dot Matrices
Global Alignments
Local Alignment

Different types of pairwise comparisons

Method name	*Situation*
Dot-plot	**General exploration of your sequence** Discovering repeats Finding long insertion/deletions Extracting portions of sequences to make a multiple alignment
Local alignments	**Comparing sequences with partial homology** Making high quality alignments Making residue-per-residue analysis
Global alignments	**Comparing two sequences over their entire length** Identifying long insertion/deletions Checking the quality of your data Identifying every mutation in your sequences

Adapted from Cedric Notredame

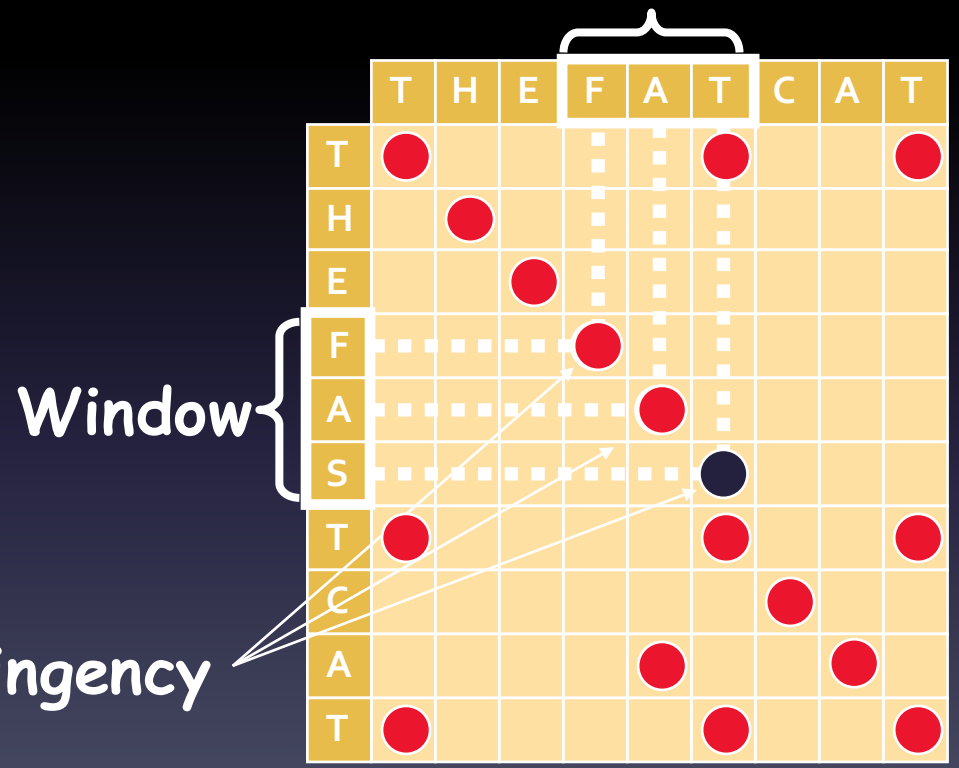
Dot Matrices

QUESTION

What are the elements shared by two sequences ?

Dot Matrices

>Seq1
THEFATCAT
>Seq2
THELASTCAT



Dot Matrices

Limits

-Visual aid

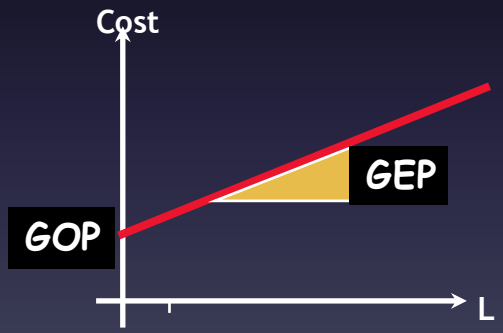
-Best Way to EXPLORE the Sequence Organisation

-Does NOT provide us with an ALIGNMENT

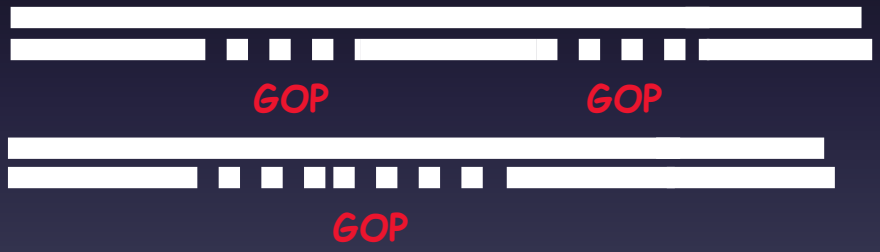
```
wheat   --DPNKPKRAMTSFVFFMSEFRSEFKQKHSKLKSIVEMVKAAGER
          | | |||||||| || | |||    |||  |  |||| ||||
?????   KKDSNAPKRAMTSFMFFSSDFRS----KHSDL-SIVEMSKAAGAA
```

Global Alignments

- Take 2 Nice Protein Sequences
- A good Substitution Matrix (blosum)
- A Gap opening Penalty (GOP)
- A Gap extension Penalty (GEP)



Afine Gap Penalty

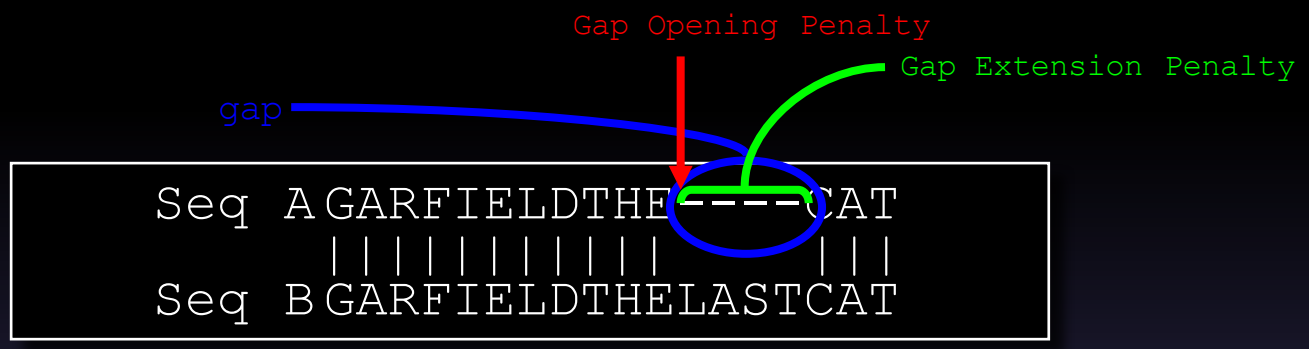


Parsimony:
Evolution takes the simplest path
(So We Think…)

Adapted from Cedric Notredame

Insertions and Deletions

Gap Penalties

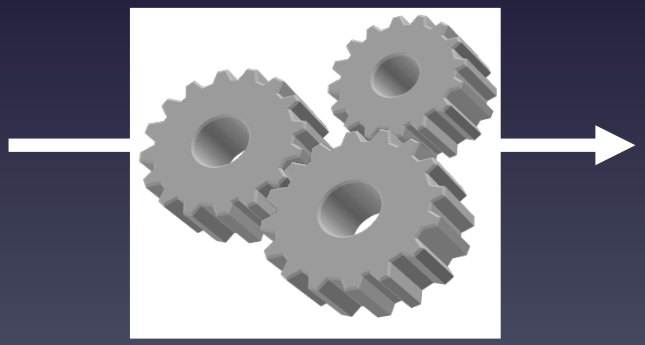


- Opening a gap is more expensive than extending it

Global Alignments

- Take 2 Nice Protein Sequences
- A good Substitution Matrix (blosum)
- A Gap opening Penalty (GOP)
- A Gap extension Penalty (GEP)
- DYNAMIC PROGRAMMING

```
>Seq1
THEFATCAT
>Seq2
THEFASTCAT
```



DYNAMIC PROGRAMMING

```
THEFA-TCAT
THEFASTCAT
```

Using Dynamic Programming To Align Sequences

- Understanding the DP concept
- Coding a Global and a Local Algorithm
- Aligning with Affine gap penalties

A bit of History…

-DP invented in the 50s by Bellman

-Programming ⇔ Tabulation

-Re-invented in 1970 by Needlman and Wunsch

-It took 10 year to find out…

The Foolish Assumption

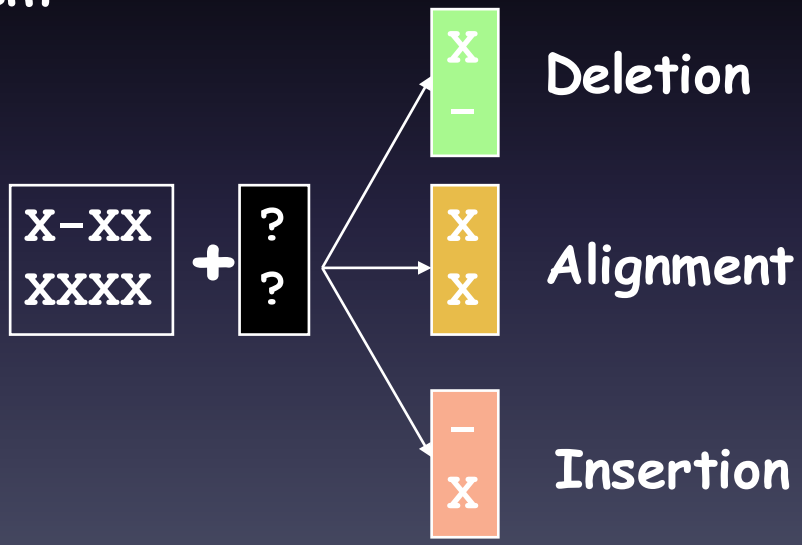
The score of each column of the alignment is independent from the rest of the alignment

It is possible to model the relationship between two sequences with:

- A substitution matrix
- A simple gap penalty

The Principal of DP

If you extend optimally an optimal alignment of two sub-sequences, the result remains an optimal alignment

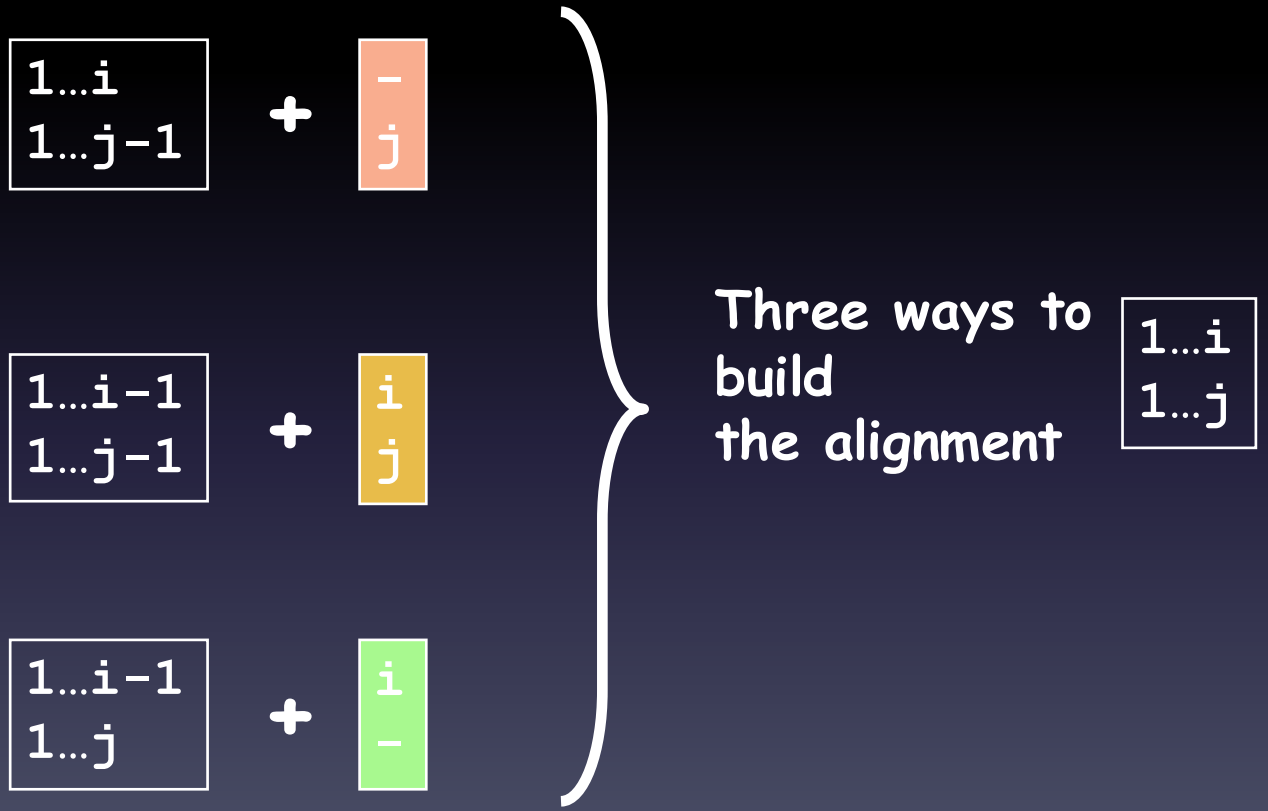


Finding the score of i,j

-Sequence 1: [1-i]
-Sequence 2: [1-j]

-The optimal alignment of [1-i] vs [1-j] can finish in three different manners:

Finding the score of i,j



Finding the score of i,j

In order to Compute the score of

1…i
1…j

All we need are the scores of:

1…i-1
1…j

1…i-1
1…j-1

1…i
1…j-1

Formalizing the algorithm

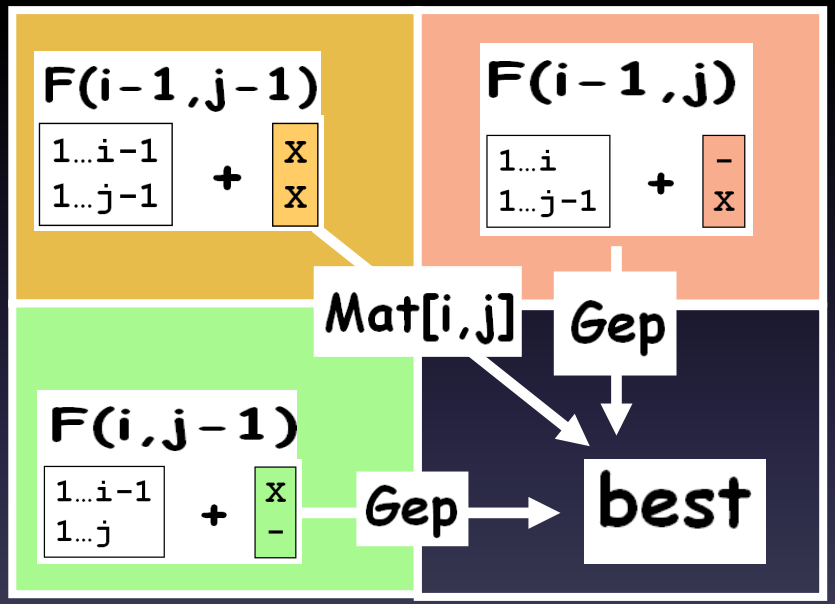
$$F(i,j) = \text{best} \begin{cases} F(i,j-1) + Gep & \begin{matrix}1\dots i \\ 1\dots j-1\end{matrix} + \begin{matrix}- \\ x\end{matrix} \\ F(i-1,j-1) + Mat[i,j] & \begin{matrix}1\dots i-1 \\ 1\dots j-1\end{matrix} + \begin{matrix}x \\ x\end{matrix} \\ F(i-1,j) + Gep & \begin{matrix}1\dots i-1 \\ 1\dots j\end{matrix} + \begin{matrix}x \\ -\end{matrix} \end{cases}$$

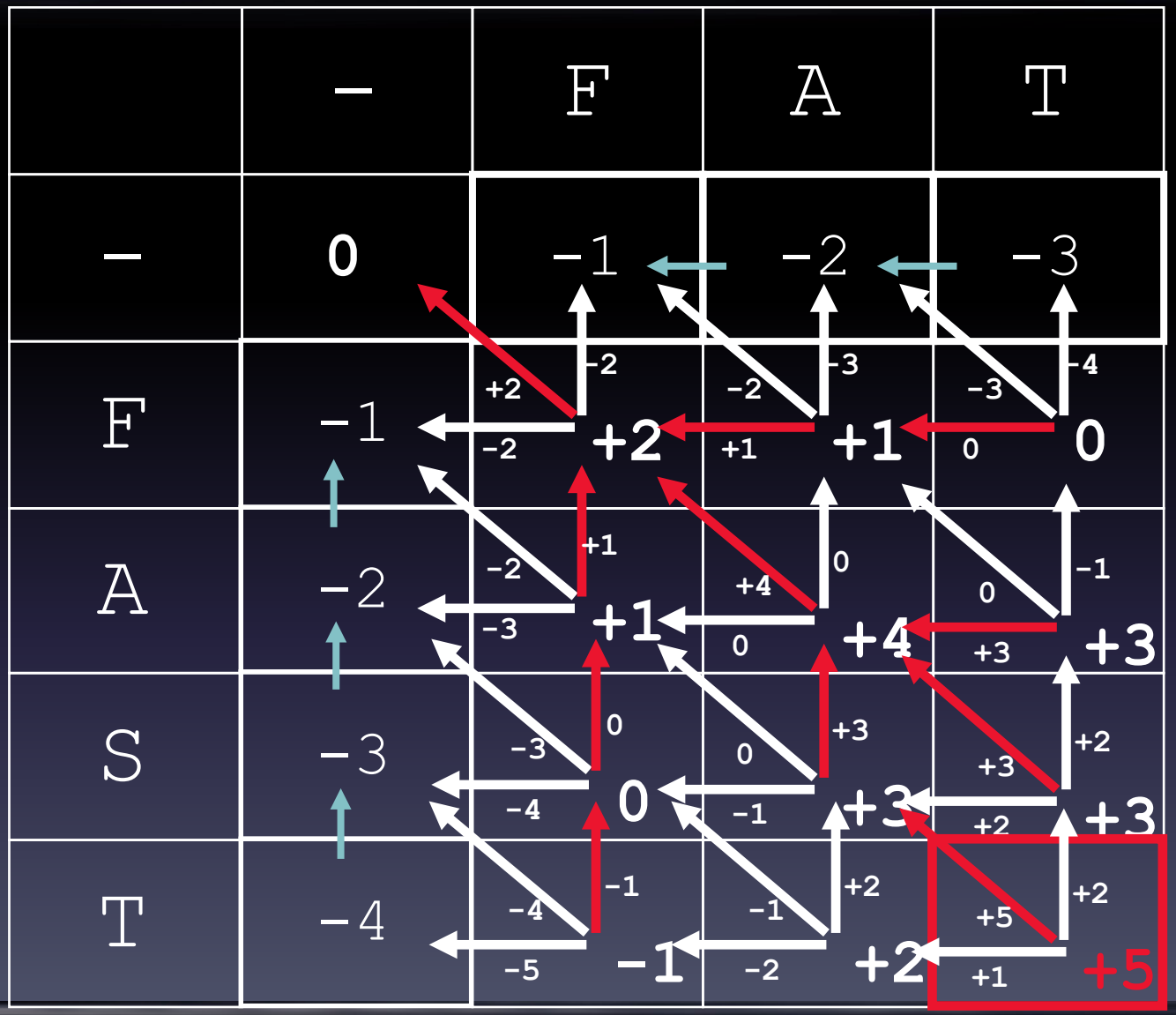
Arranging Everything in a Table

	–	F	A	T
–				
F		1…I-1 1…J-1	1…I 1…J-1	
A		1…I-1 1…J	1…I 1…J	
S				
T				

Filing Up The Matrix

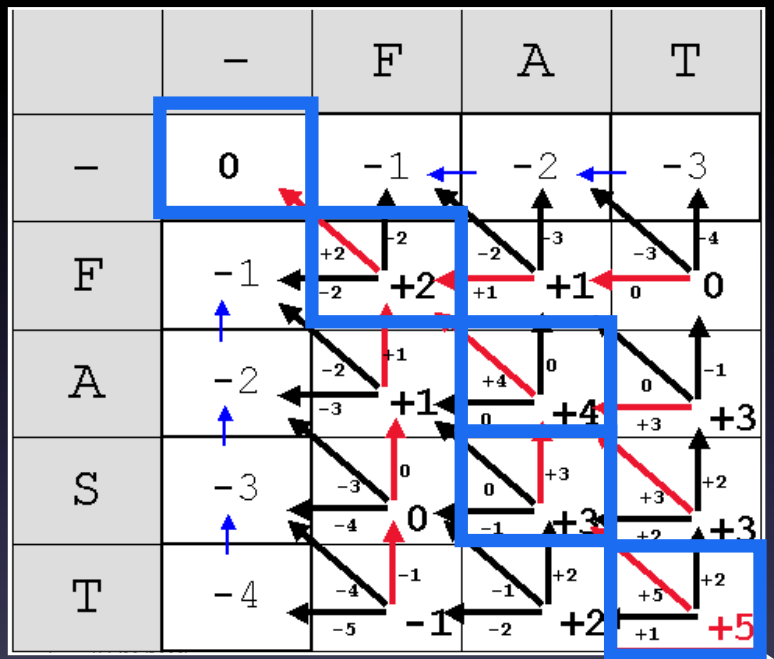
$$F(i,j) = \text{best} \begin{cases} F(i,j-1) + Gep & \begin{matrix} 1\ldots i \\ 1\ldots j-1 \end{matrix} + \begin{matrix} - \\ x \end{matrix} \\ F(i-1,j-1) + Mat[i,j] & \begin{matrix} 1\ldots i-1 \\ 1\ldots j-1 \end{matrix} + \begin{matrix} x \\ x \end{matrix} \\ F(i-1,j) + Gep & \begin{matrix} 1\ldots i-1 \\ 1\ldots j \end{matrix} + \begin{matrix} x \\ - \end{matrix} \end{cases}$$





Adapted from Cedric Notredame

Delivering the alignment: Trace-back



T	S	A	F
T	–	A	F

Score of 1…3 Vs 1…4
⇔
Optimal Aln Score

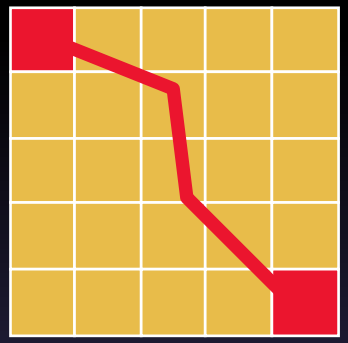
Adapted from Cedric Notredame

Trace-back: possible implementation

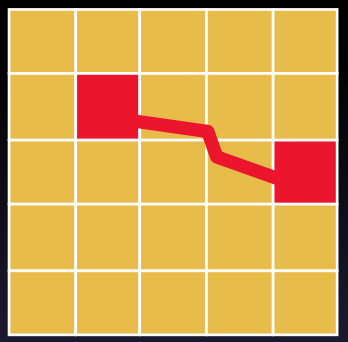
```
while (!($i==0 && $j==0))
      { if ($tb[$i][$j]==$sub)     #SUBSTITUTION
          { $alnI[$aln_len]=$seqI[--$i];
                $alnJ[$aln_len]=$seqJ[--$j];
          }
        elsif ($tb[$i][$j]==$del)    #DELETION
          { $alnI[$aln_len]='-';
            $alnJ[$aln_len]=$seqJ[--$j];
          }
        elsif ($tb[$i][$j]==$ins)    #INSERTION
          { $alnI[$aln_len]=$seqI[0][--$i];
            $alnJ[$aln_len]='-';
          }
      $aln_len++;
      }
```

Adapted from Cedric Notredame

Local Alignments



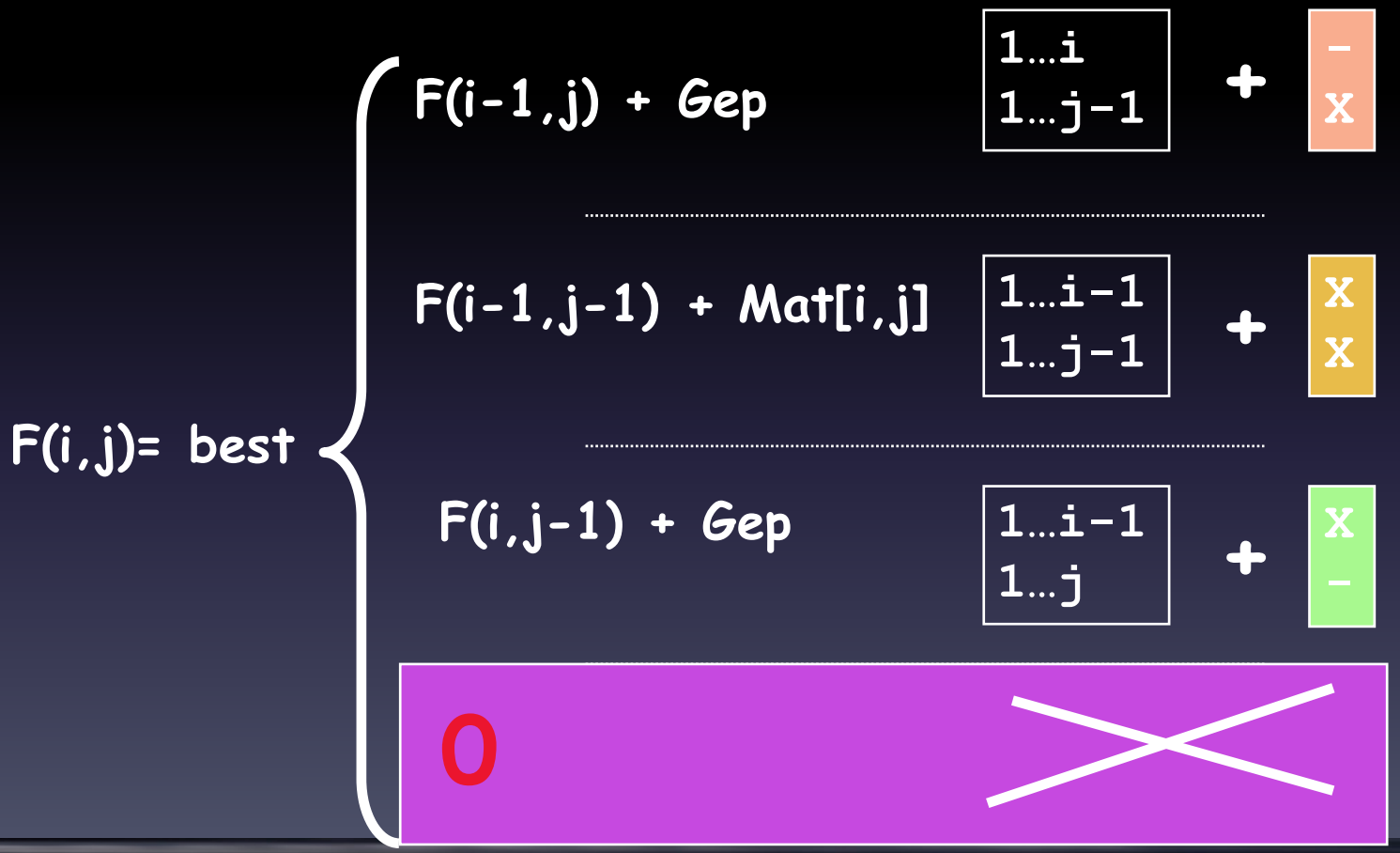
GLOBAL Alignment



LOCAL Alignment

Smith And Waterman (SW)=LOCAL Alignment

The Smith and Waterman Algorithm



The Smith and Waterman Algorithm

$$F(i,j) = \text{best} \begin{cases} F(i-1,j) + Gep \\ F(i-1,j-1) + Mat[i,j] \\ F(i,j-1) + Gep \\ 0 \end{cases}$$

Adapted from Cedric Notredame

The Smith and Waterman Algorithm

0

⇔

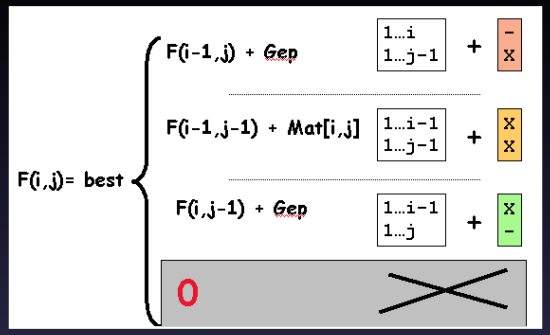
Ignore The rest of the Matrix

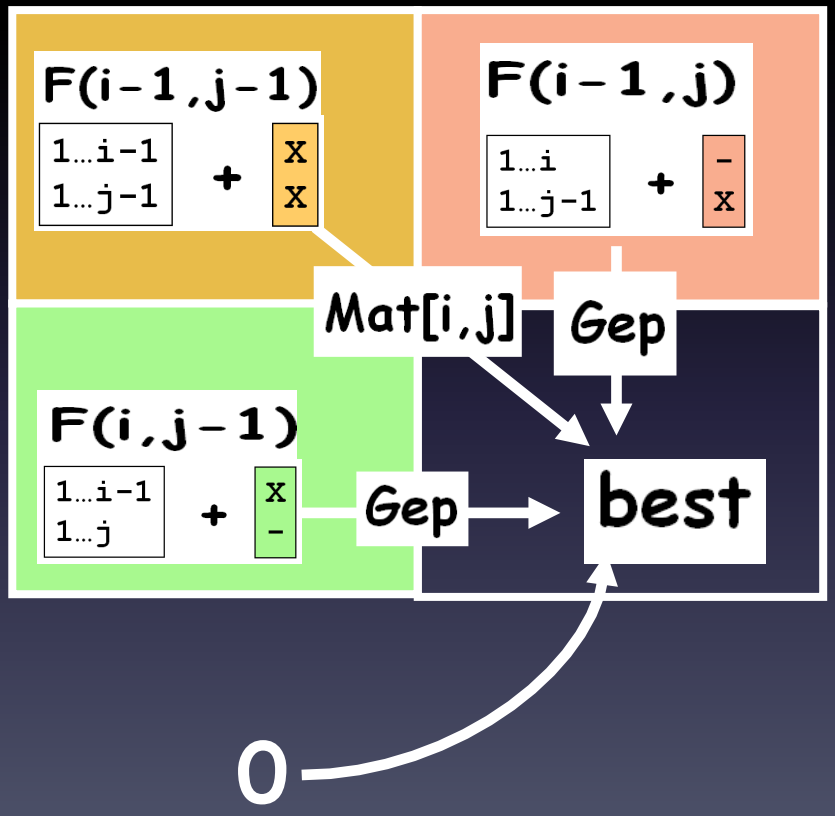
⇔

Terminate a local Aln

Adapted from Cedric Notredame

Filing Up a SW Matrix





Adapted from Cedric Notredame

Filling up a SW matrix: borders

		A	N	I	C	E	C	A	T
	0	0	0	0	0	0	0	0	0
C	0								
A	0								
T	0								
A	0								
N	0								
D	0								
O	0								
G	0								

Easy:
Local alignments NEVER start/end with a gap…

Filling up a SW matrix

		A	N	I	C	E	C	A	T
-	0	0	0	0	0	0	0	0	0
C	0	0	0	0	2	0	2	0	0
A	0	2	0	0	0	0	0	4	0
T	0	0	0	0	0	0	0	2	6
A	0	2	0	0	0	0	0	0	4
N	0	0	4	2	0	0	0	0	2
D	0	0	2	2	0	0	0	0	0
O	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0

Best Local score

⇔

Beginning of the trace-back

Adapted from Cedric Notredame

Turning NW into SW

Prepare Trace back

```
for ($i=1; $i<=$len0; $i++) {
    for ($j=1; $j<=$len1; $j++) {
        if ($res0[0][$i-1] eq $res1[0][$j-1]){$s=2;}
        else {$s=-1;}
        $sub=$mat[$i-1][$j-1]+$s;
        $del=$mat[$i ][$j-1]+$gep;
        $ins=$mat[$i-1][$j ]+$gep;
        if ($sub>$del && $sub>$ins && $sub>0)
            {$smat[$i][$j]=$sub;$tb[$i][$j]=$subcode;}
        elsif($del>$ins && $del>0 )
            {$smat[$i][$j]=$del;$tb[$i][$j]=$delcode;}
        elsif( $ins>0 )
            {$smat[$i][$j]=$ins;$tb[$i][$j]=$inscode;}
        else {$smat[$i][$j]=$zero;$tb[$i][$j]=$stopcode;}

        if ($smat[$i][$j]> $best_score) {
            $best_score=$smat[$i][$j];
            $best_i=$i; $best_j=$j;
        }
    }
}
```

More than One match…

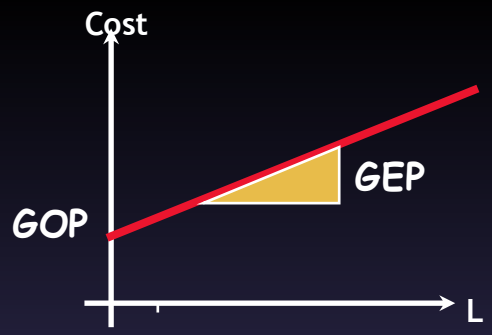
-SW delivers only the best scoring Match

-If you need more than one match:
- -SIM (Huang and Millers)
- Or
- -Waterman and Eggert (Durbin, p91)

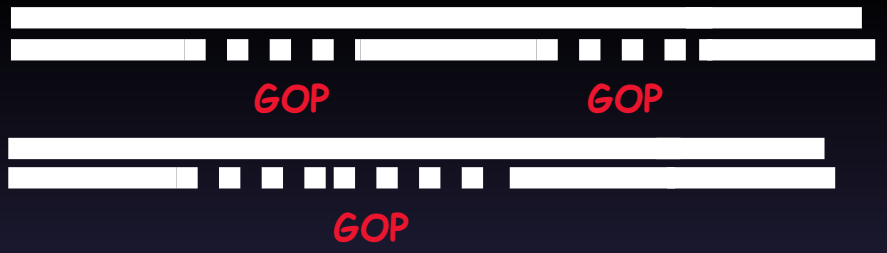
The Gotoh Algorithm

- Adding Affine Gap Penalties
- Forcing a bit of Biology into your alignment

Why Affine gap Penalties are Biologically better



Afine Gap Penalty



Cost=gop+L*gep

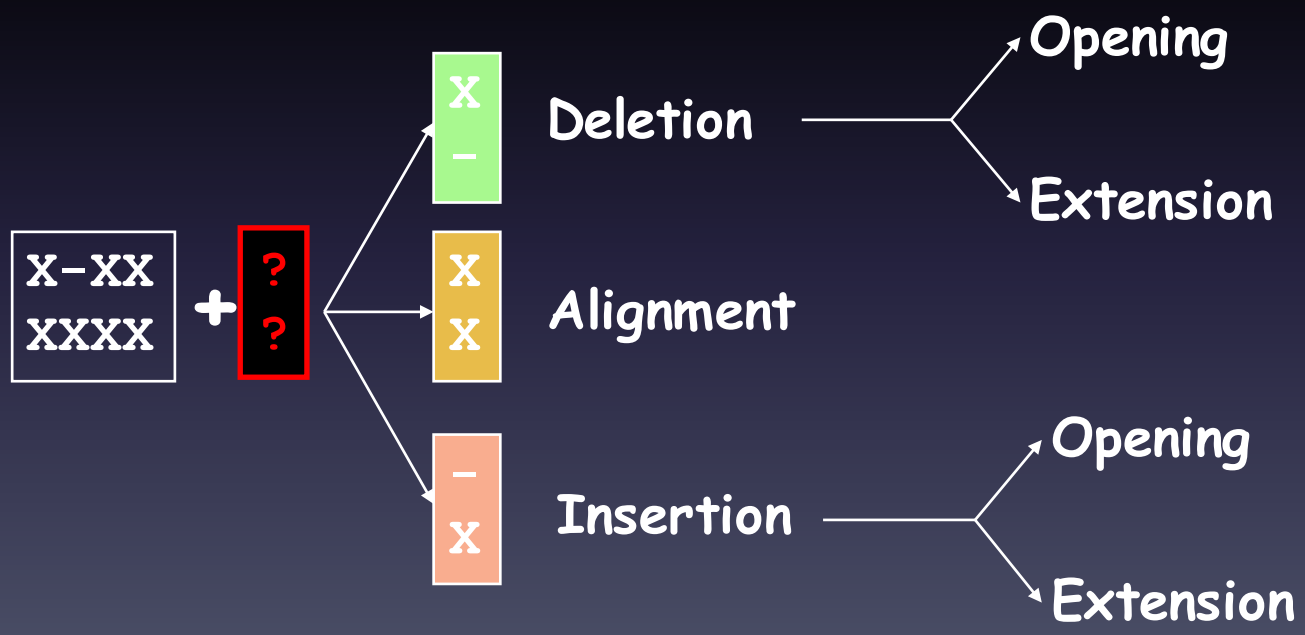
Or Cost=gop+(L-1)*gep

Parsimony:
Evolution takes the simplest path
(So We Think…)

Adapted from Cedric Notredame

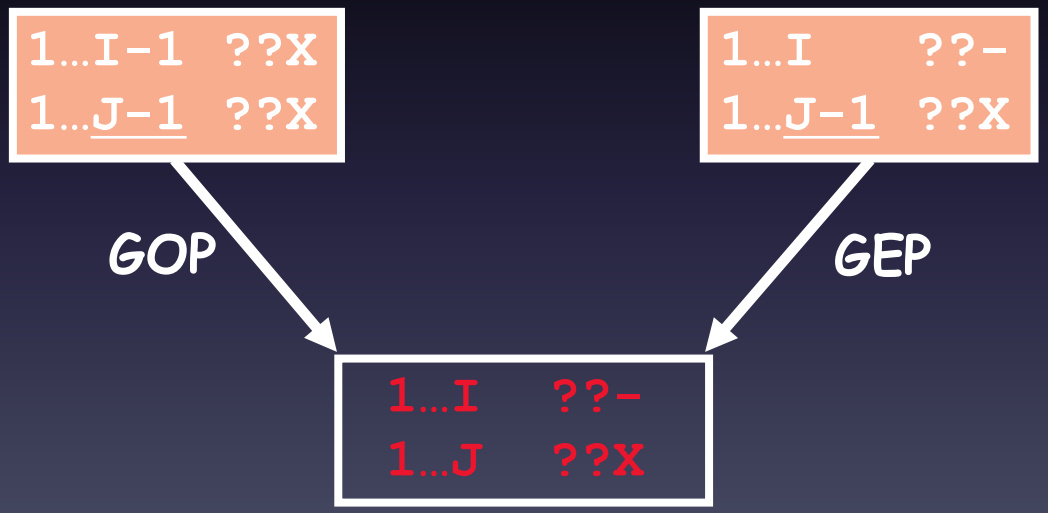
But Harder To compute…

More Than 3 Ways to extend an Alignment



More Questions Need to be asked

For instance, what is the cost of an insertion ?



Solution:Maintain 3 Tables

Ix: Table that contains the score of every optimal alignment 1…i vs 1…j that finishes with an Insertion in sequence X.

Iy: Table that contains the score of every optimal alignment 1…I vs 1…J that finishes with an Insertion in sequence Y.

M: Table that contains the score of every optimal alignment 1…I vs 1…J that finishes with an alignment between sequence X and Y

The Algorithm

$$M(i,j) = \text{best} \begin{cases} M(i-1,j-1) + Mat(i,j) \\ Ix(i-1,j-1) + Mat(i,j) \\ Iy(i-1,j-1) + Mat(i,j) \end{cases}$$

$$\begin{matrix} 1\ldots i-1 \\ 1\ldots j-1 \end{matrix} + \begin{matrix} x \\ x \end{matrix}$$

$$Ix(i,j) = \text{best} \begin{cases} M(i-1,j) + gop \\ Ix(i-1,j) + gep \end{cases}$$

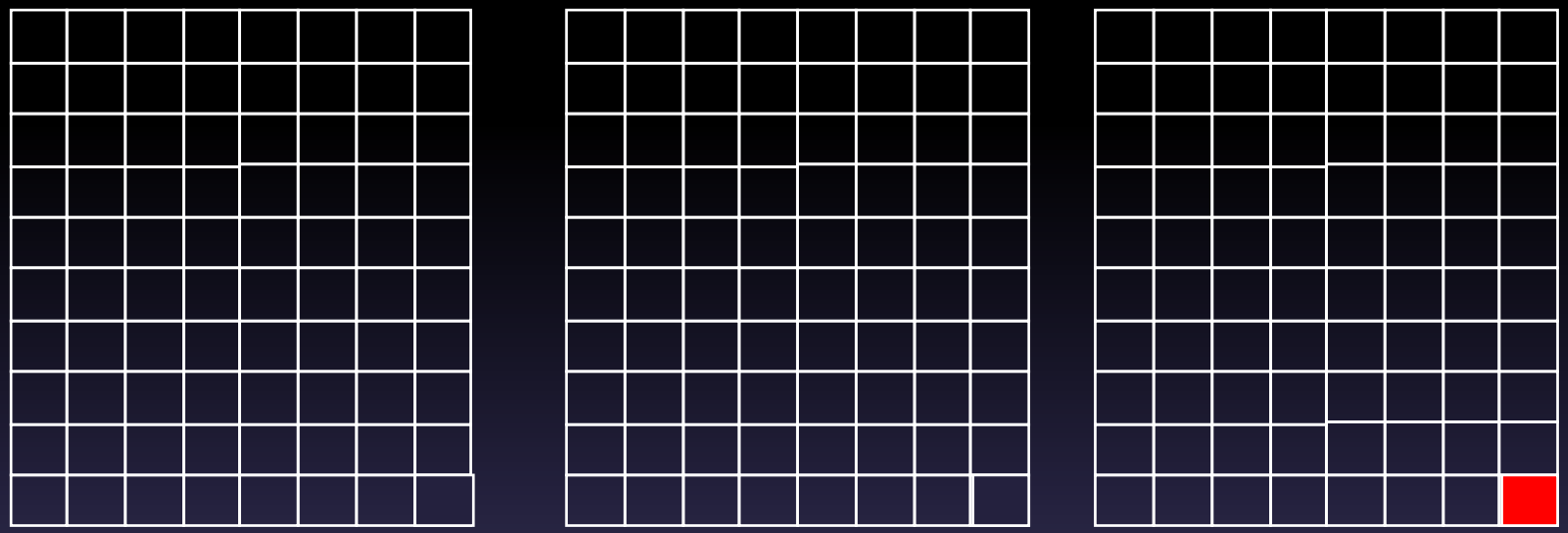
$$\begin{matrix} 1\ldots i-1 & x \\ 1\ldots j & x \end{matrix} + \begin{matrix} x \\ - \end{matrix} \qquad \begin{matrix} 1\ldots i-1 & x \\ 1\ldots j & - \end{matrix} + \begin{matrix} x \\ - \end{matrix}$$

$$Iy(i,j) = \text{best} \begin{cases} M(i,j-1) + gop \\ Iy(i,j-1) + gep \end{cases}$$

$$\begin{matrix} 1\ldots i & x \\ 1\ldots j-1 & x \end{matrix} + \begin{matrix} - \\ x \end{matrix} \qquad \begin{matrix} 1\ldots i & - \\ 1\ldots j-1 & x \end{matrix} + \begin{matrix} - \\ x \end{matrix}$$

Adapted from Cedric Notredame

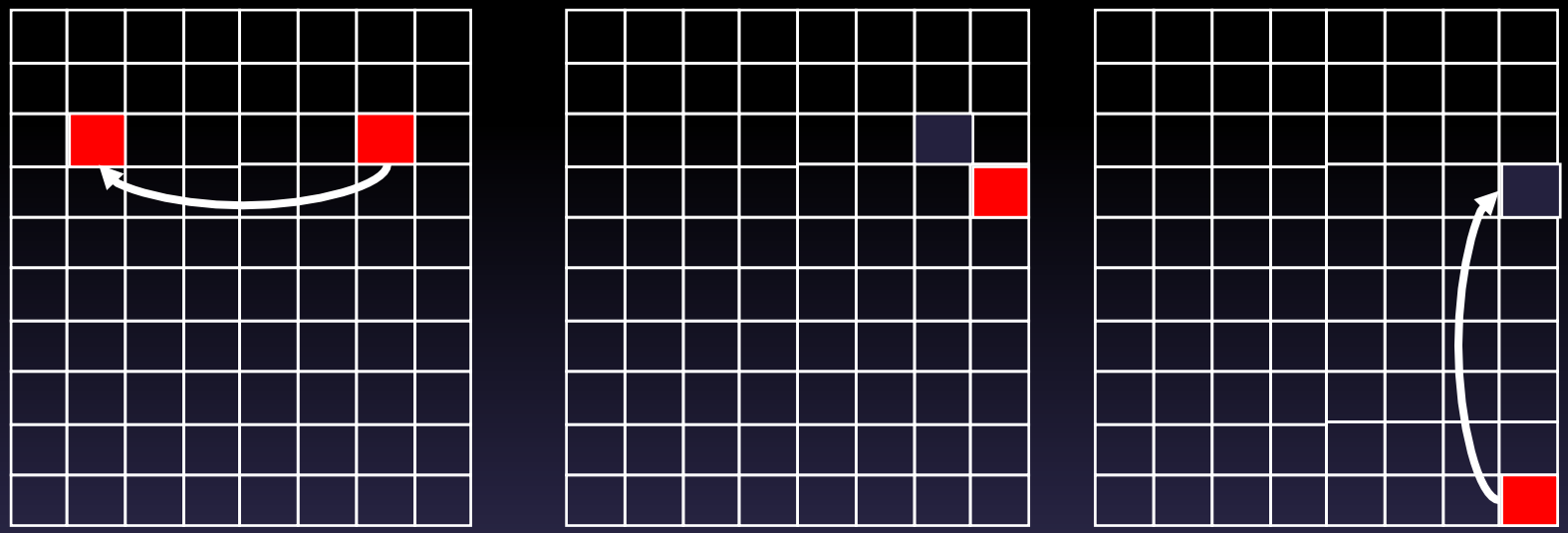
Trace-back?



Ix M Iy

Start From BEST { M(i,j) Ix(i,j) Iy(i,j)

Trace-back?

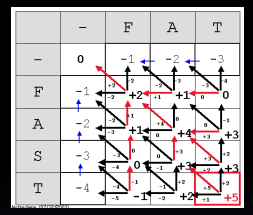


Ix M Iy

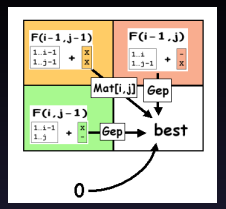
Navigate from one table to the next, knowing that a gap always finishes with an aligned column…

SUMMARY: Dynamic Programming

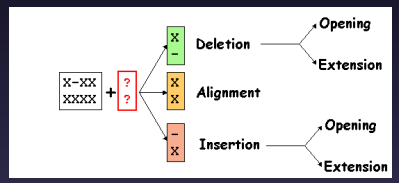
Needleman and Wunsch: Delivers the best scoring global alignment



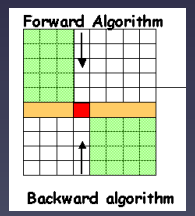
Smith and Waterman: NW with an extra state 0



Affine Gap Penalties: Making DP more realistic



Linear space: Using Divide and Conquer Strategies Not to run out of memory



Adapted from Cedric Notredame

Local Alignments

We now have a PairWise Comparison Algorithm,

We are ready to search Databases

Using BLAST to Search Sequence Databases

- Evolution and Sequence Similarity
- The inside of BLAST
- Using BLAST
- Adapting BLAST to your needs
- Searching Protein Domains with BLAST
- Digging Genomes

Adapted from Cedric Notredame

A few basic definitions

- Query : Your sequence
- Subject: The database against which you search
- Heuristic: Algorithm that does not guaranty the optimal solution

Other Important Definitions

- Identity
  - Proportion of IDENTICAL residues between two sequences.
  - Depends on the Alignment. Unit: the % id
- Similarity
  - Proportion of SIMILAR residues
  - Two residues are similar if their substitution cost is higher than 0.
  - Depends on the matrix. Unit: the %similarity
- Homology
  - Sequences SIMILAR enough are sometimes HOMOLOGOUS
  - HOMOLOGY ~ COMMON ANCESTOR
  - Unit: Yes or No!
  - DIFFERENT sequences can also be Homologous

Adapted from Cedric Notredame

More Important Definitions

- Hit : A sequence that matches your sequence and reported by BLAST.
- E-Value
  - Expectation value
  - How many times would you expect to find a hit by chance only?
    - Depends on the alignment.
    - Depends on the matrix
    - Depends on the database
    - Sensitive to Low complexity regions
  - Unit: must be lower than 0.0001 to mean something

A Good Hit Is Something You Would Not Expect by Chance

Adapted from Cedric Notredame

What is BLAST?

- Basic Local Alignment Search Tool
- BLAST is a Program Designed for RAPIDLY Comparing Your Sequence With every Sequence in a database and REPORT the most SIMILAR sequences

Adapted from Cedric Notredame

BLAST

A Bit of History

Smith and Waterman

- Exact Local Dynamic Programming, 1981

FASTA

- Lipman and Pearson, 1985
- Looks for similar words (k-tup) on the same diagonal.
- Comparison on the sequences one by one…

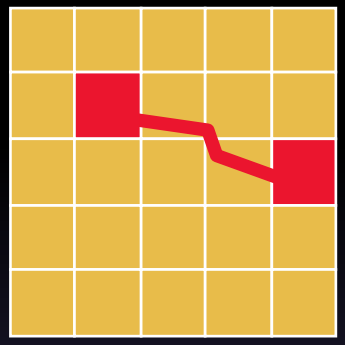
BLAST

- Altschul *et al.*, 1990
- The most widely cited tool in Biology
- www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html

Database Search

1-Query

2-Comparison Engine



LOCAL Alignment

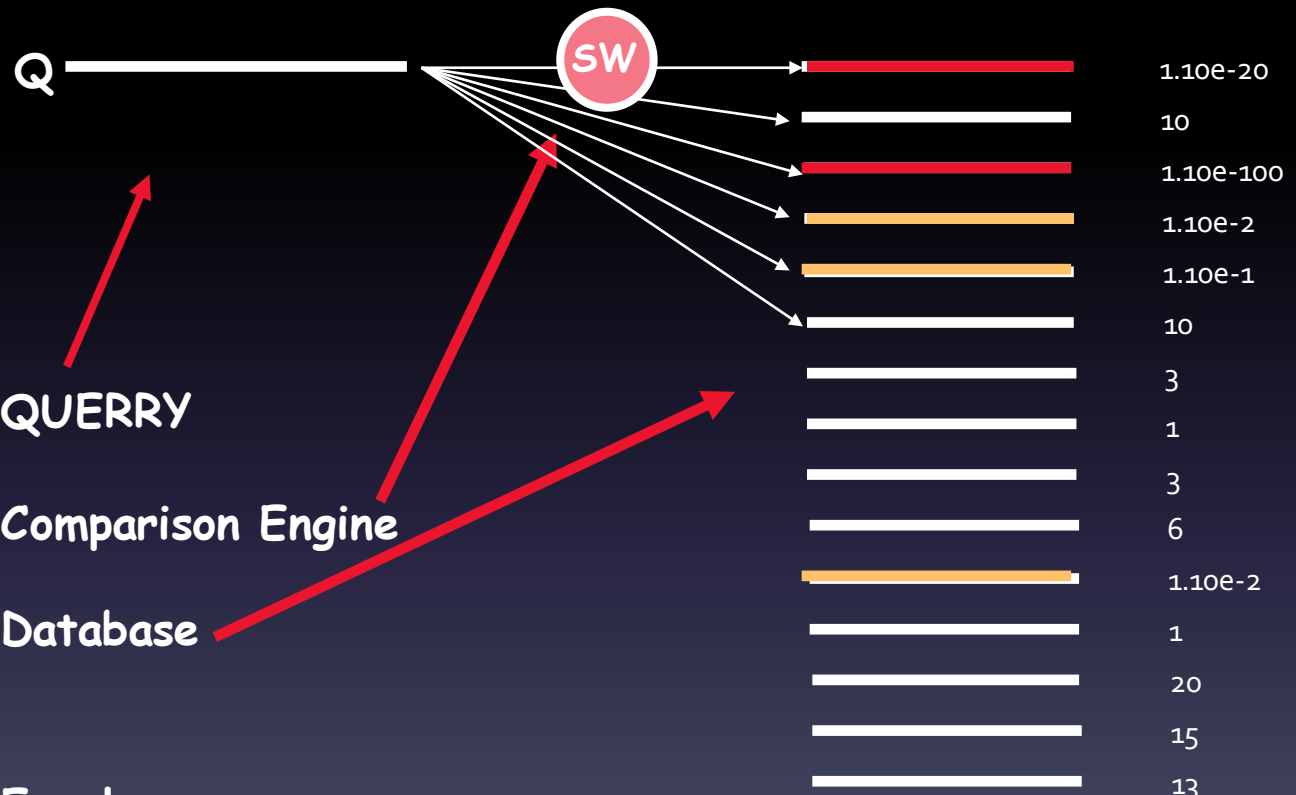
3-Database

4-Statistical Evaluation (E-Value)

PROBLEM: LOCAL ALIGNMENT (SW)TOO SLOW

Adapted from Cedric Notredame

Database Search



QUERRY

Comparison Engine

Database

E-values
How many time do we expect such an Alignment by chance?

Adapted from Cedric Notredame

BLAST

Basic Local Alignment Search Tool

BLAST is a Heuristic Smith and Waterman

BLAST = 3 STEPS

1-Decide who will be compared

This is where Blast SAVES TIME

This is where it LOSES HITS

Most BLAST parameters refer to this step

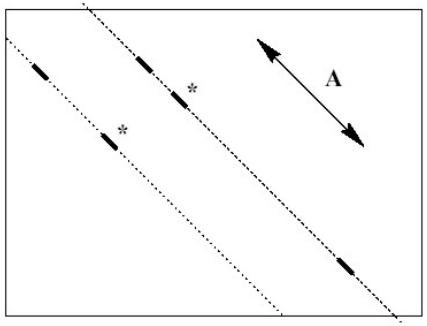


Figure 3: The "two-hits" requirement: those hits that are going to trigger an ungapped extension are labelled with a star

Adapted from Cedric Notredame

BLAST

Basic Local Alignment Search Tool

BLAST is a Heuristic Smith and Waterman

BLAST = 3 STEPS

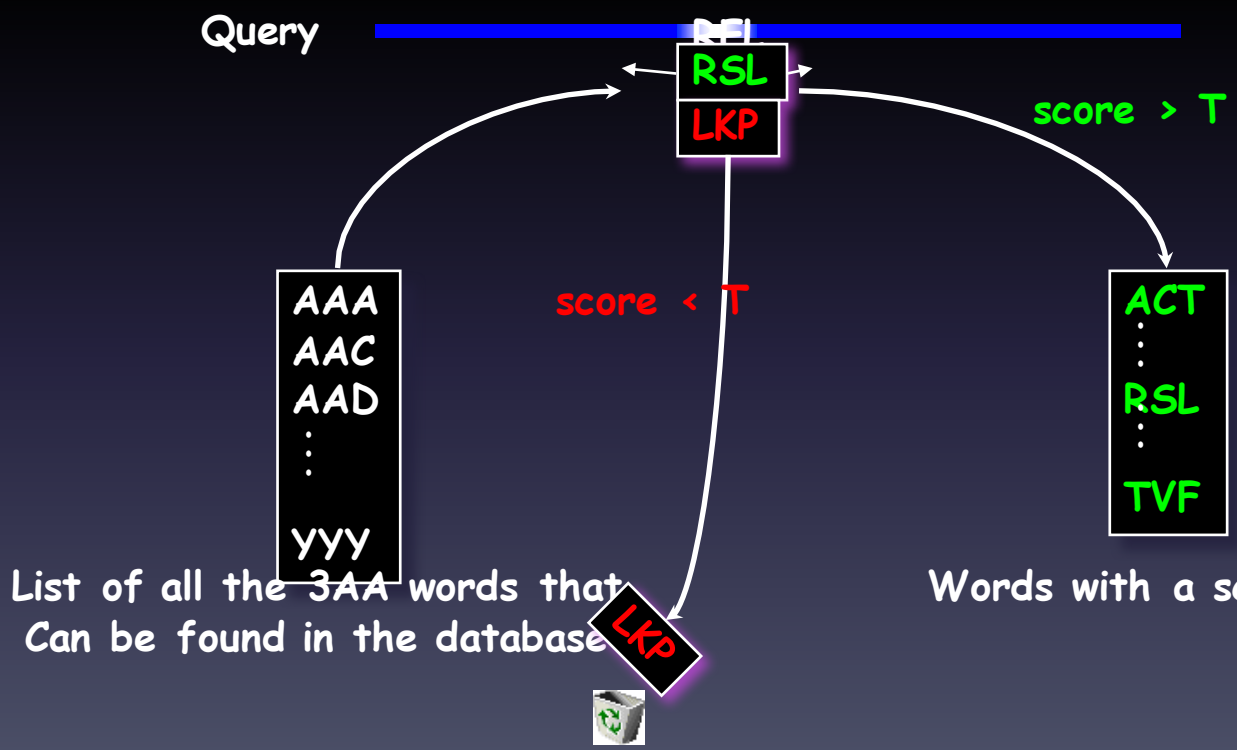
1-Decide who will be compared

2-Check the most promising Hits

3-Compute the E-value of the most interesting Hits

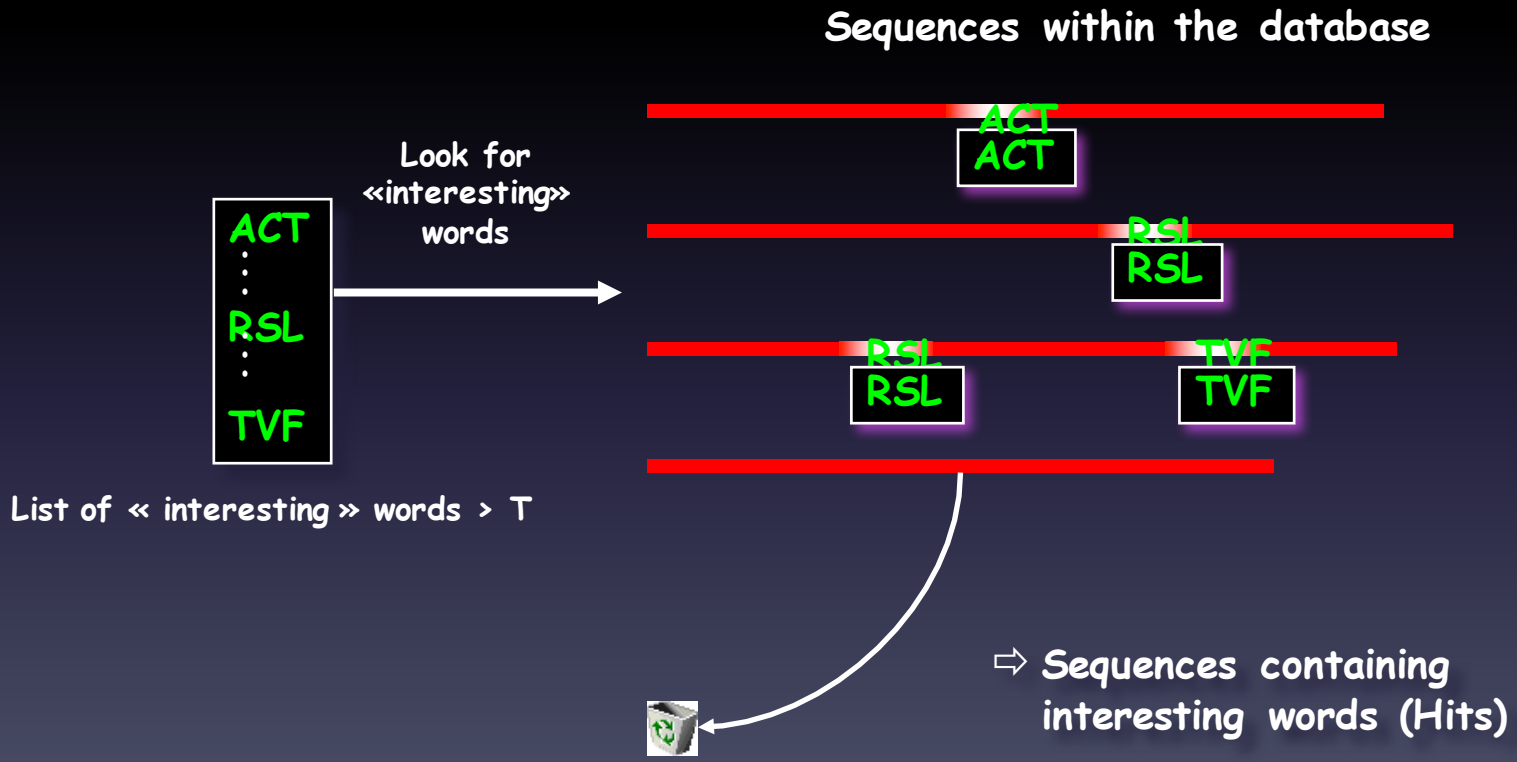
Inside BLAST

Step 1: finding the worthy words



Inside BLAST

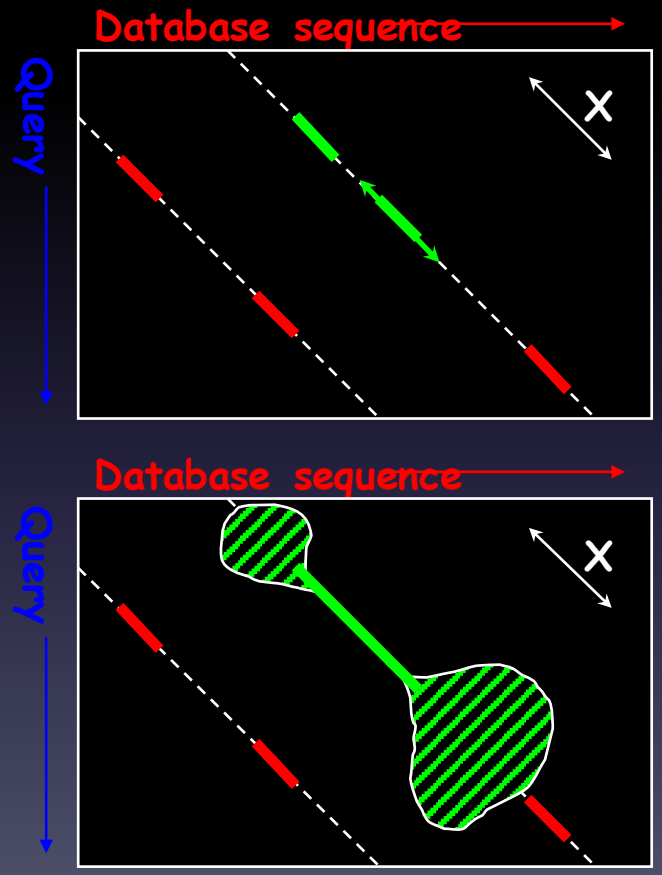
Step 2: Eliminate the database sequences that do not contain any interesting word



Adapted from Cedric Notredame

Inside BLAST: the end

Step 3: Extension of the Hits



- 2 "Hits" on the same diagonal distant by less than X

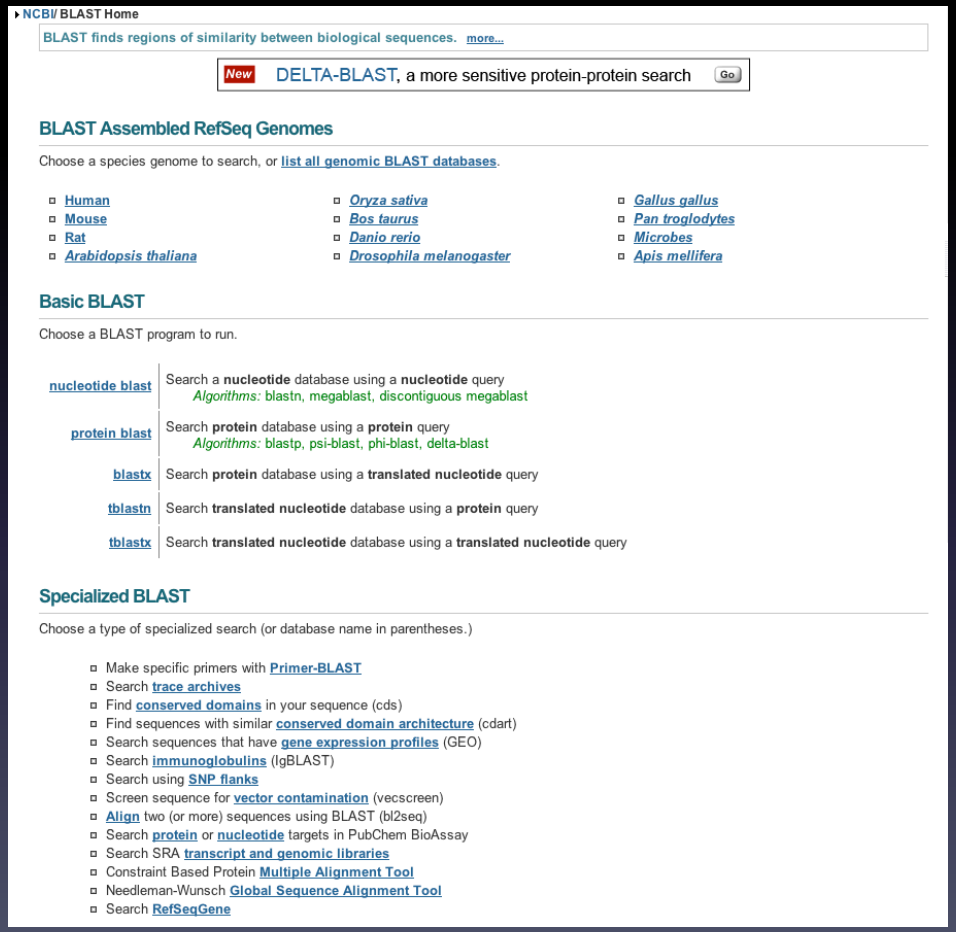
Extension by limited Dynamic Programming

BLAST Statistics

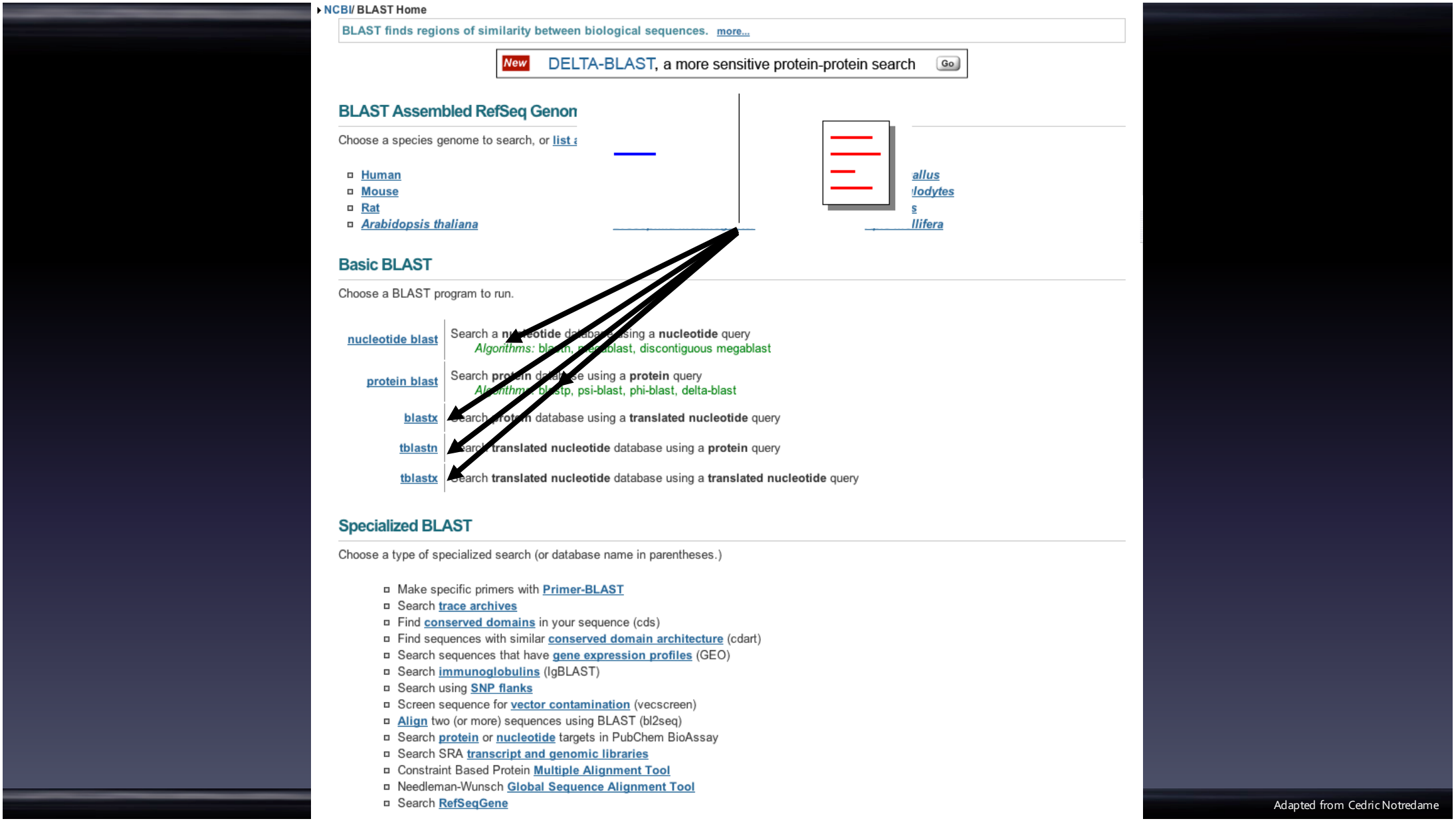
- Raw Score
  - Sum of the substitutions and gap penalties.
  - Not very informative
- p-value (Derived Statistics)
  - Probability of finding an alignment with such a score, by chance.
  - The lower, the better

Adapted from Cedric Notredame

The Many Flavors of BLAST



Adapted from Cedric Notredame



Adapted from Cedric Notredame

The Many Flavors of BLAST

Program	Query	Database
blastp	protein	protéine
blastn	nucleotide	nucleotide
blastx	nucleotide → protein	vs protein
tblastn	protein	vs nucleotide → protein
tblastx	nucleotide → protein	vs nucleotide → protein

Adapted from Cedric Notredame

NCBI/ BLAST Home

BLAST finds regions of similarity between biological sequences. more...

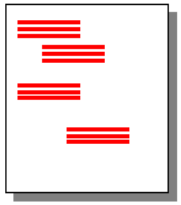
New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled RefSeq Gen

Choose a species genome to search, or li

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*

MSA



lus
dytes
fera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

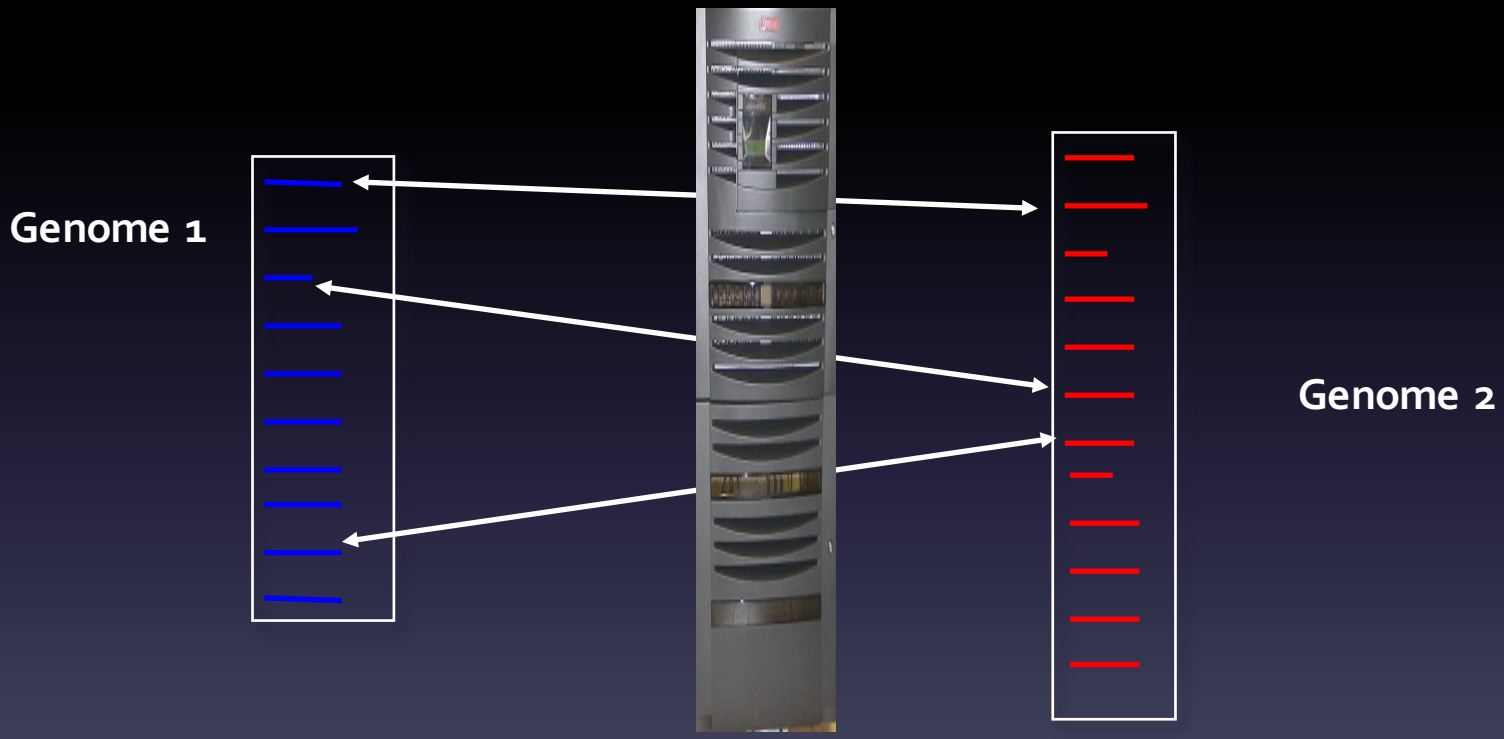
Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search using SNP flanks
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript and genomic libraries
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene

Adapted from Cedric Notredame

Database Against Database: Farm-Blast »



Ideal for finding Orthologues

Adapted from Cedric Notredame

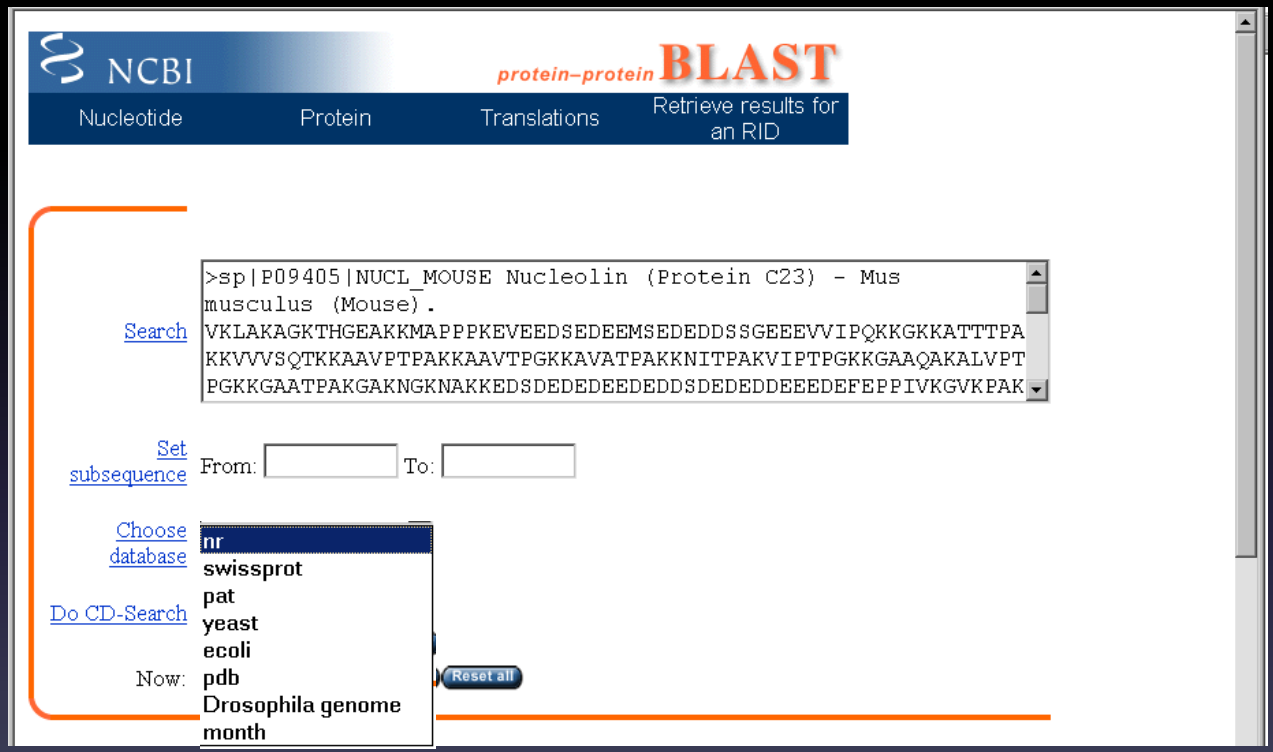
Database Search

Database Search Result=Prediction

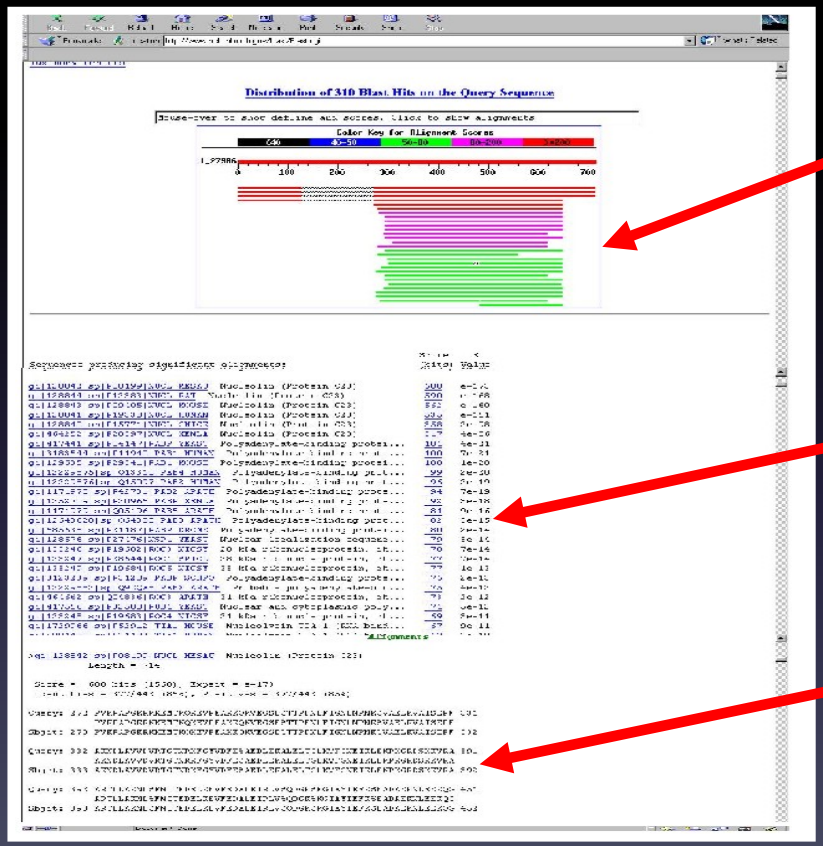
Protein X IS or IS NOT homologous to the QUERRY.

Adapted from Cedric Notredame

Submitting your Query



Understanding the BLAST Output



Graphic Display

Hit List

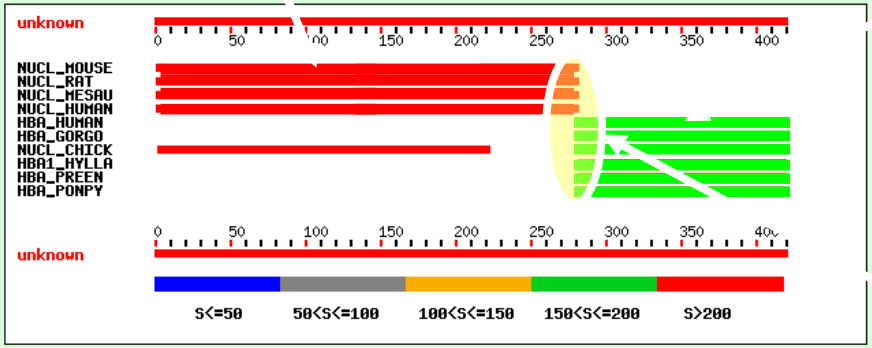
Alignments

Using BLAST: Trouble Shooting

Adapted from Cedric Notredame

Domain 1

Domain 2



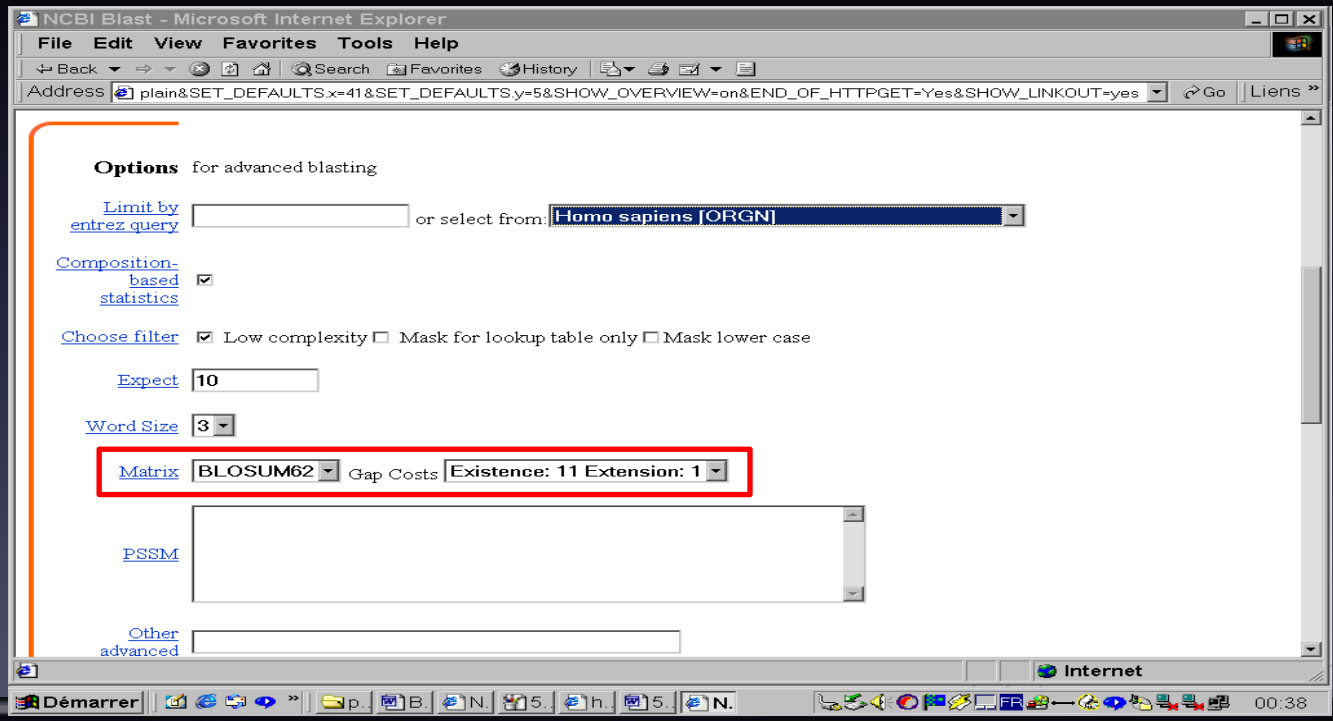
No Overlap

Sequences producing significant alignments:	Score (bits)	E Value
sp|P09405|NUCL_MOUSE|89505EE39C89F832 (NCL..)Nucleolin (Protein ...	425	e-119
sp|P13383|NUCL_RAT|68774A214E550F90 (NCL..)Nucleolin (Protein C2...	407	e-113
sp|P08199|NUCL_MESAU|79DDCF724CED7DB4 (NCL)Nucleolin (Protein C2...	397	e-110
sp|P19338|NUCL_HUMAN|85A2F2CA22EA03DB (NCL)Nucleolin (Protein C2...	371	e-102
sp|P01922|HBA_HUMAN|34D13618E62A33C1 (HBA1..)Hemoglobin alpha ch...	285	2e-76
sp|P01923|HBA_GORGO|25D13618E72A3306 (HBA)Hemoglobin alpha chain...	283	4e-76
sp|P15771|NUCL_CHICK|7996C504BE9459A1 Nucleolin (Protein C23).[G...	283	5e-76
sp|Q9TS35|HBA1_HYLLA|25D13618E36A7706 (HBA1)Hemoglobin alpha-1 c...	281	2e-75
sp|P01924|HBA_PREEN|3771361D402A35C7 (HBA)Hemoglobin alpha chain...	280	5e-75
sp|P06635|HBA_PONPY|37DE74049545CE88 (HBA)Hemoglobin alpha chain...	279	7e-75

Adapted from Cedric Notredame

Some reasons to change BLAST default parameters

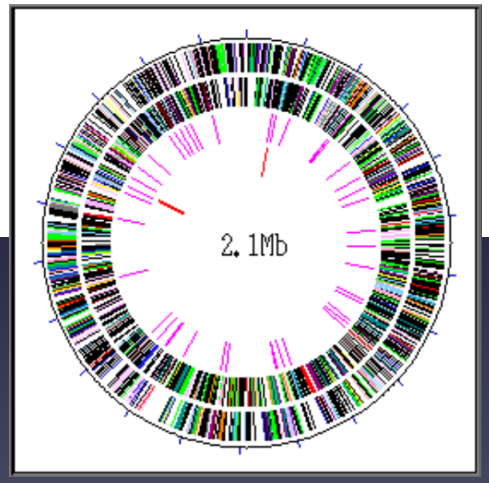
Reason	*Parameters to change*
BLAST does not report any result	Change the *substitution matrix* or the *gap penalties*.
Your match has a borderline E-value	Change the *substitution matrix* or the *gap penalties* to check the match robustness.



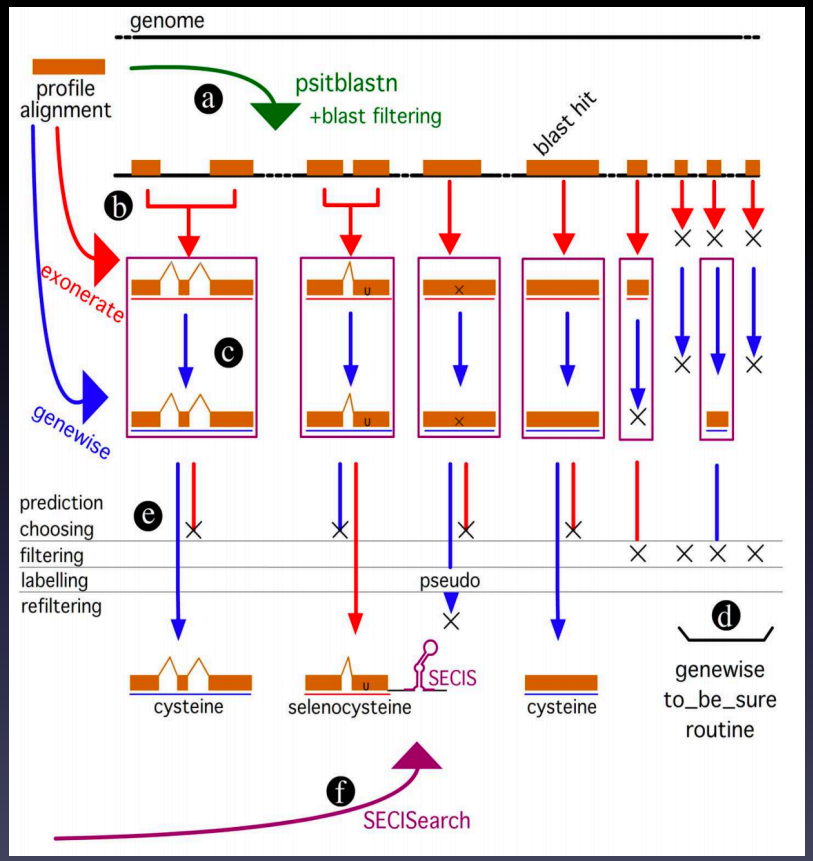
Adapted from Cedric Notredame

Adapting BLAST To your Problem

Table 8.4 Asking biological questions with BLAST

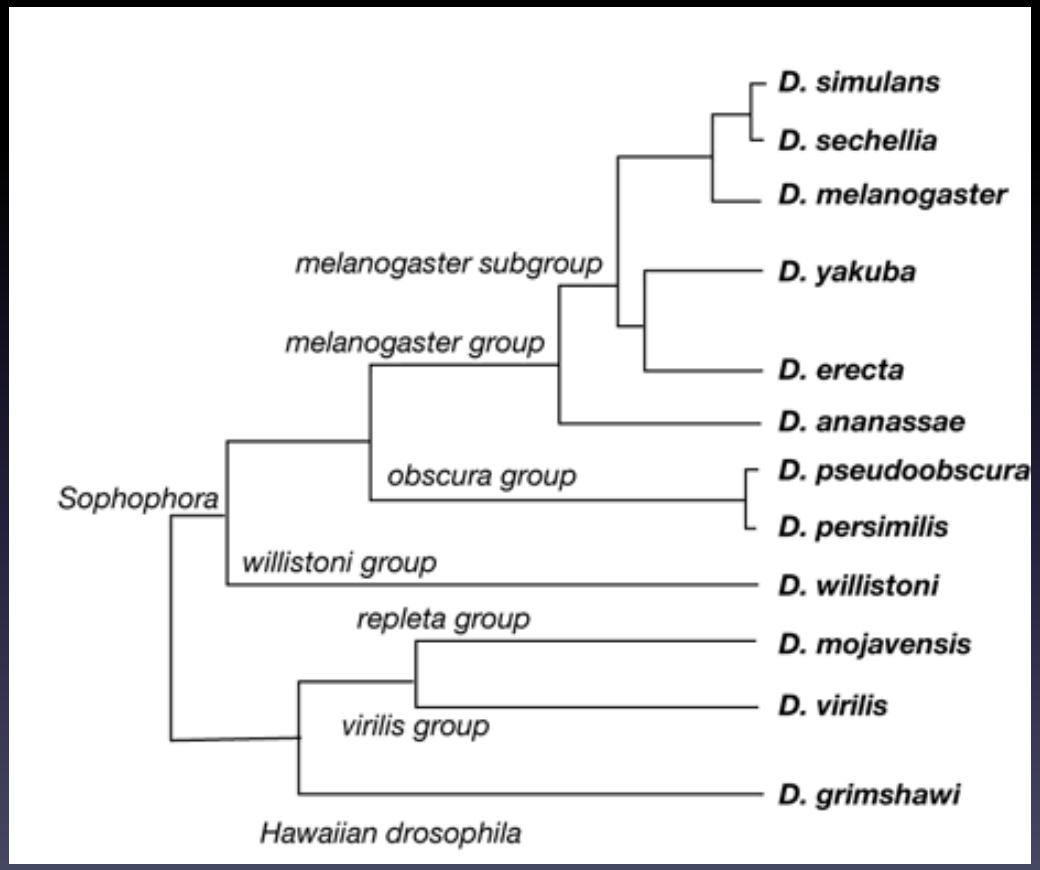
What you need	*The BLAST way*
Finding genes in a genome 	Cut your genome sequence in little (2-5kb) overlapping sequences. Use blastx to BLAST each piece of genome against NR (the Non Redundant Protein database). This works better if you have no introns (bacteria). The complicated alternative is to run a gene prediction software.

Pipeline for profile-based protein finding in genomes



Mariotti, M. & Guigó, R. Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *Bioinformatics* **26,** 2656–63 (2010).

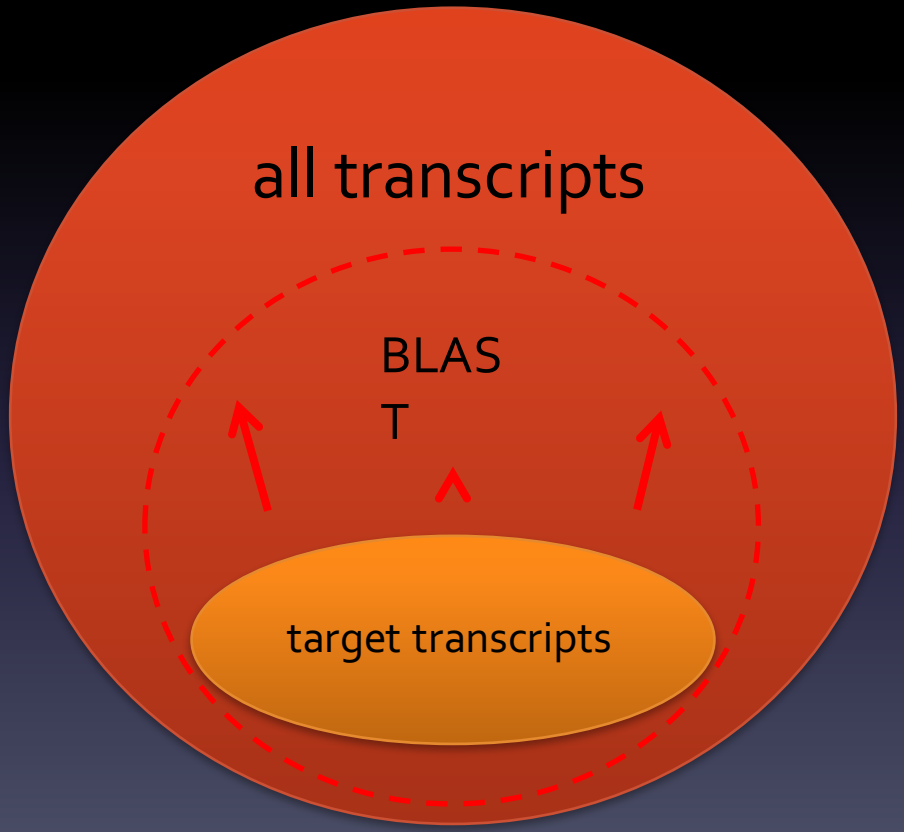
Target vs All transcriptomes



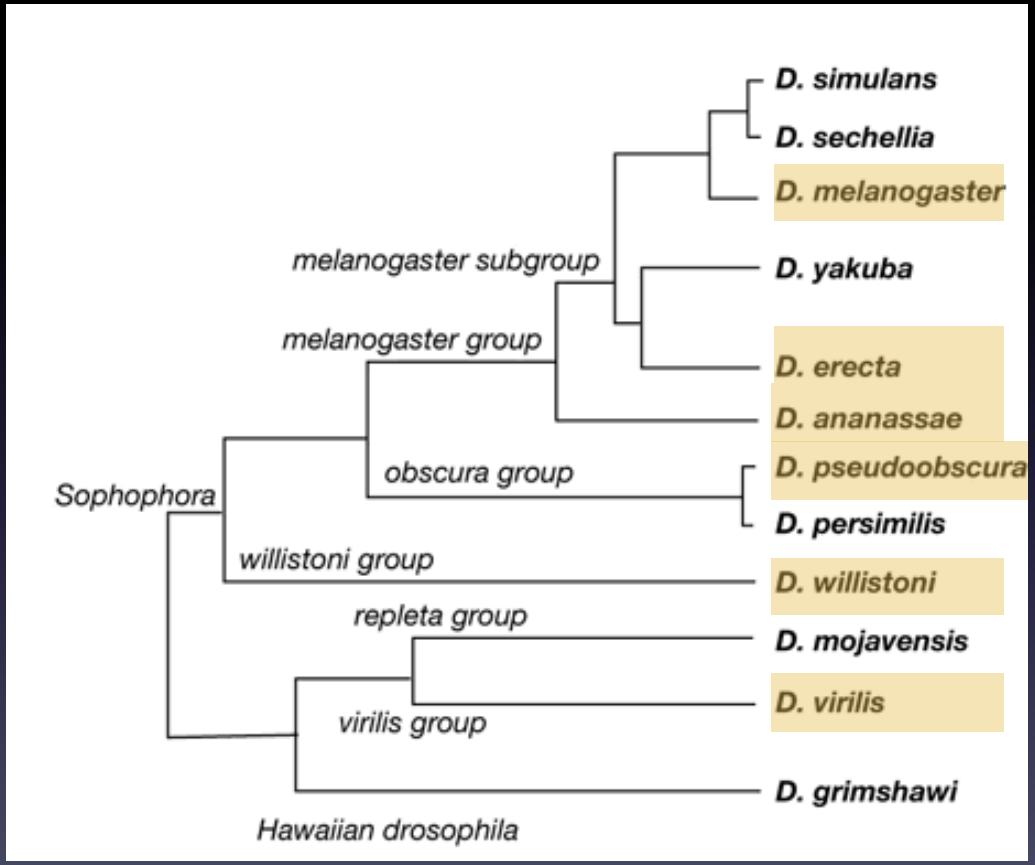
	#	#
D. melanogaster	261	21,909
D. erecta	233	15,048
D. ananassae	212	15,070
D. pseudoobscura	232	16,181
D. willistoni	203	15,513
D. virilis	175	14,491

http://rana.lbl.gov/drosophila/graphics/tree.gif

Where are missing transcripts?



Homologous extension



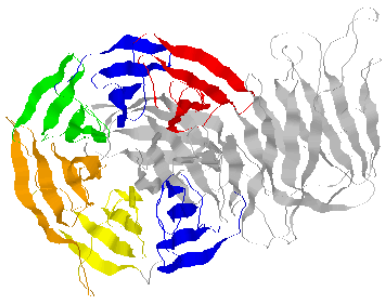
http://rana.lbl.gov/drosophila/graphics/tree.gif

	Before #	After #
D. melanogaster	261	282
D. erecta	233	273
D. ananassae	212	258
D. pseudoobscura	232	259
D. willistoni	203	242
D. virilis	175	222

Table 8.4 Asking biological questions with BLAST

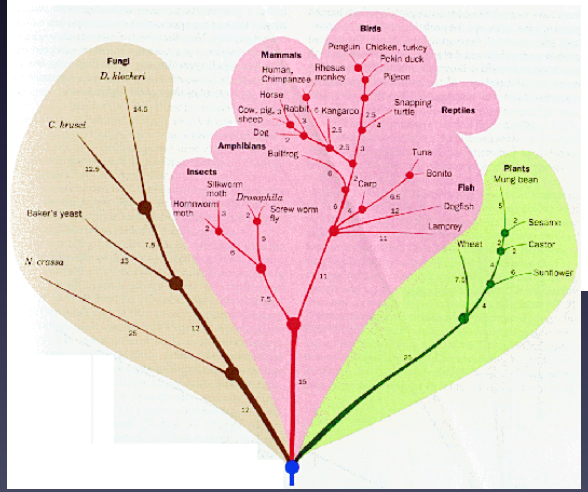
What you need	*The BLAST way*
Predicting a Protein Function	Use blastp to BLAST your protein sequence against SwissProt. If you get a good hit (more than 25% identity) over the complete length of the protein, then you have solved your problem and you know that your protein has the same function as the SwissProt protein. The complicated alternative is to do domain analysis or wet-lab experiments.

Table 8.4 Asking biological questions with BLAST

What you need	*The BLAST way*
Predicting a Protein 3D structure 	Use blastp to BLAST your protein against PDB (the database of protein structure). If you get a good hit, (more than 25% identity), then you know that your protein and this good hit have a similar 3D structure. The complicated alternative is to do homology modeling, Xray or NMR analysis of your protein.

Asking biological questions with BLAST

What you need	*The BLAST way*
Finding a protein family members	Use blastp (or its more powerful cousin Psi-BLAST) and run it on NR the non-redundant protein family. Once you have all the members of the family, you can make a multiple sequence alignment (see Chapter 11) and draw a phylogenic tree.



Use blastp (or its more powerful cousin Psi-BLAST) and run it on NR the non-redundant protein family. Once you have all the members of the family, you can make a multiple sequence alignment (see Chapter 11) and draw a phylogenic tree.

The Complicated alternative is to use PCR for Clonning your sequences

BLAST latest Flavor

PSI-BLAST

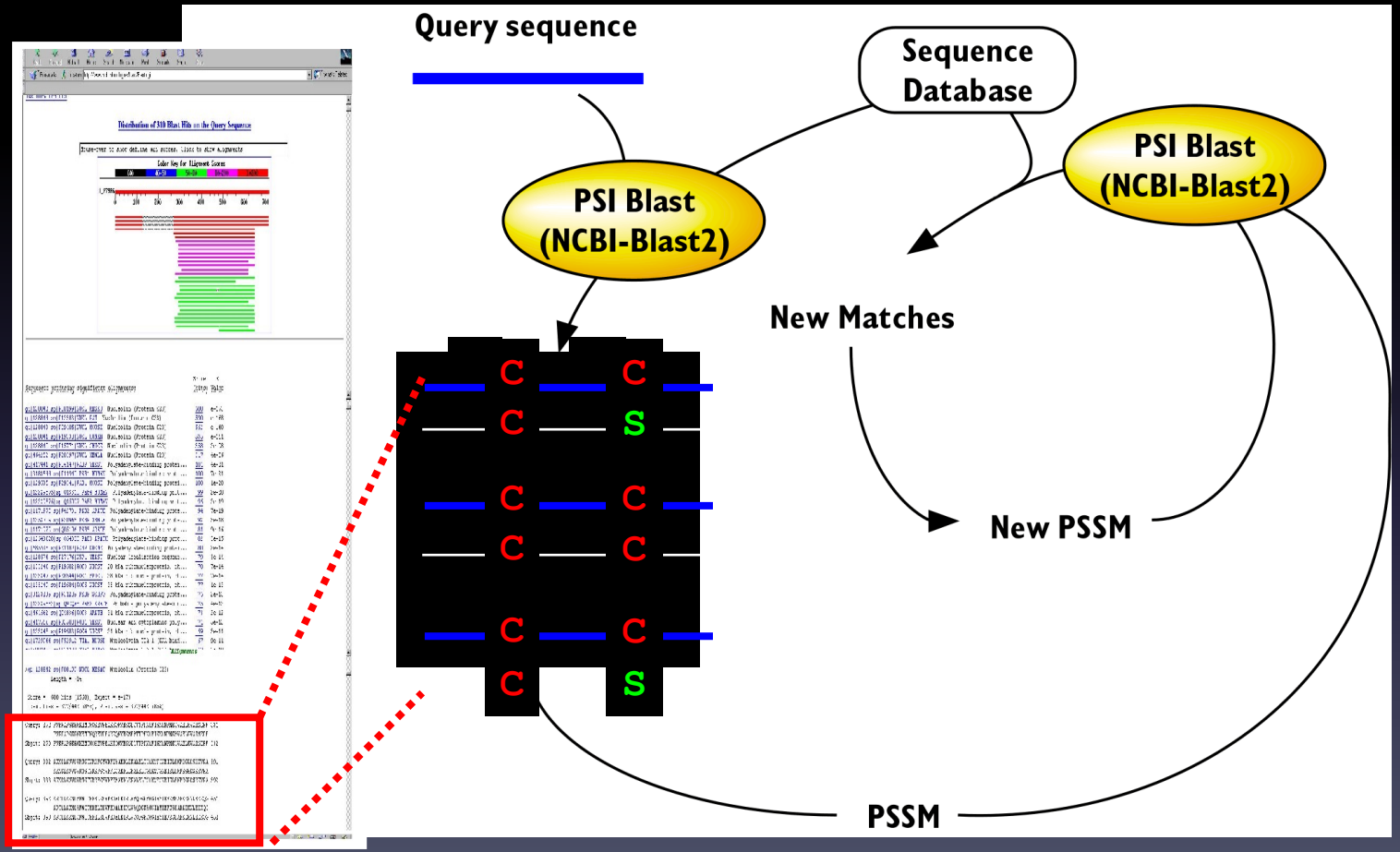
-Position Specific Iterated Version of BLAST.

-Uses Profiles.

-More Sensitive.

Adapted from Cedric Notredame

Psi-BLAST Iteration



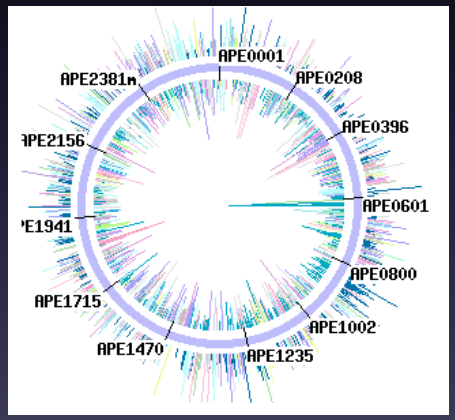
Adapted from Cedric Notredame

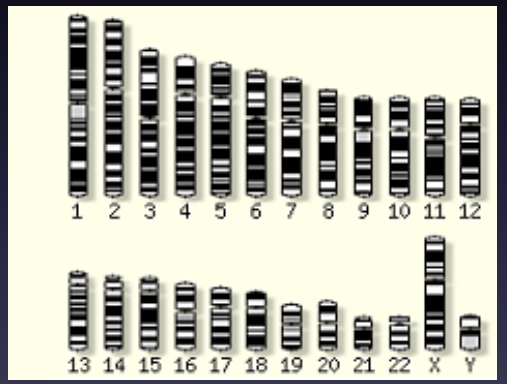
BLAST PSSM or weight matrix

	M	Y	C	E	Q	U	E	N	C	E	S	.	.
A	0	2	-1	0	0	0	0	-1	0	-1	3		
S	-1	-1	-1	0	-1	0	0	0	**5**	-1	-1		
C	-1	-1	**10**	1	-1	0	0	5	**5**	4	-1		
.													
.													
Y	-1	6	-1	-1	-1	0	-1	-1	-1	-1	-1		
V	-1	1	-1	-1	-1	0	-1	-1	-1	1	-1		

Adapted from Cedric Notredame

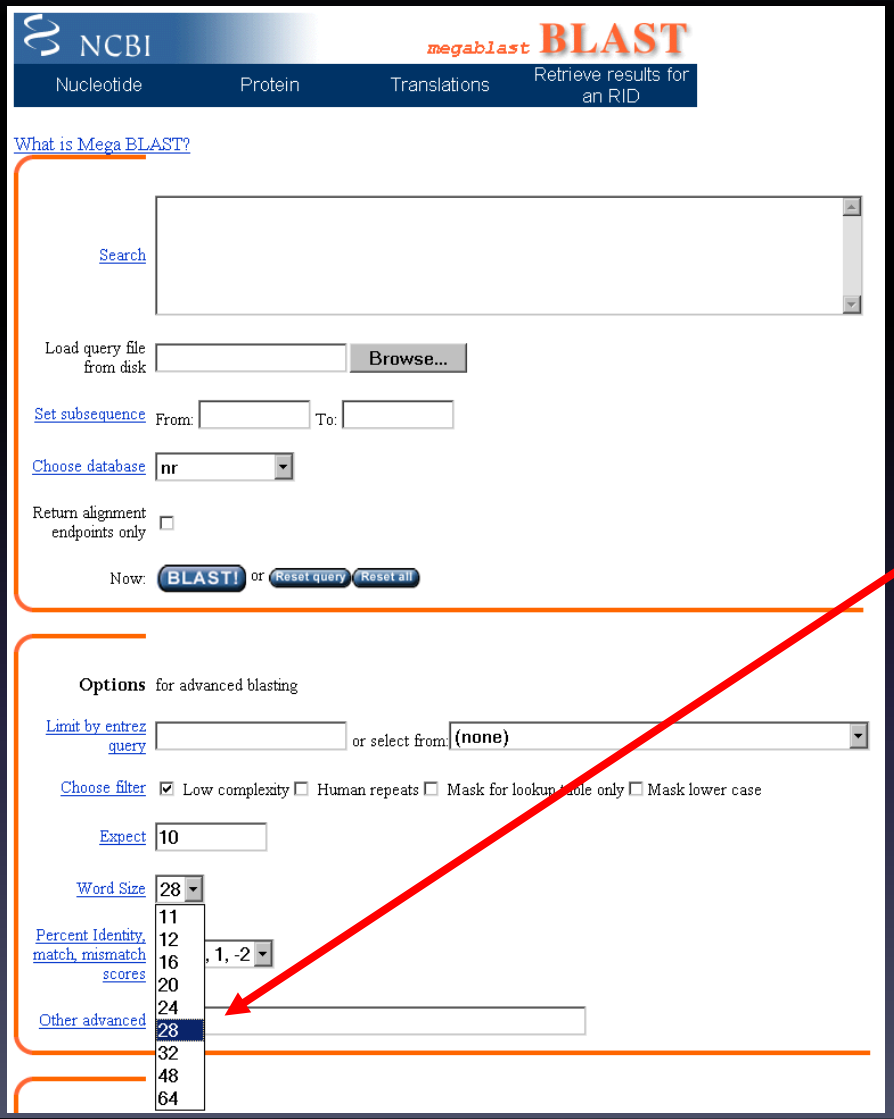
Genome Flavored BLAST





Adapted from Cedric Notredame

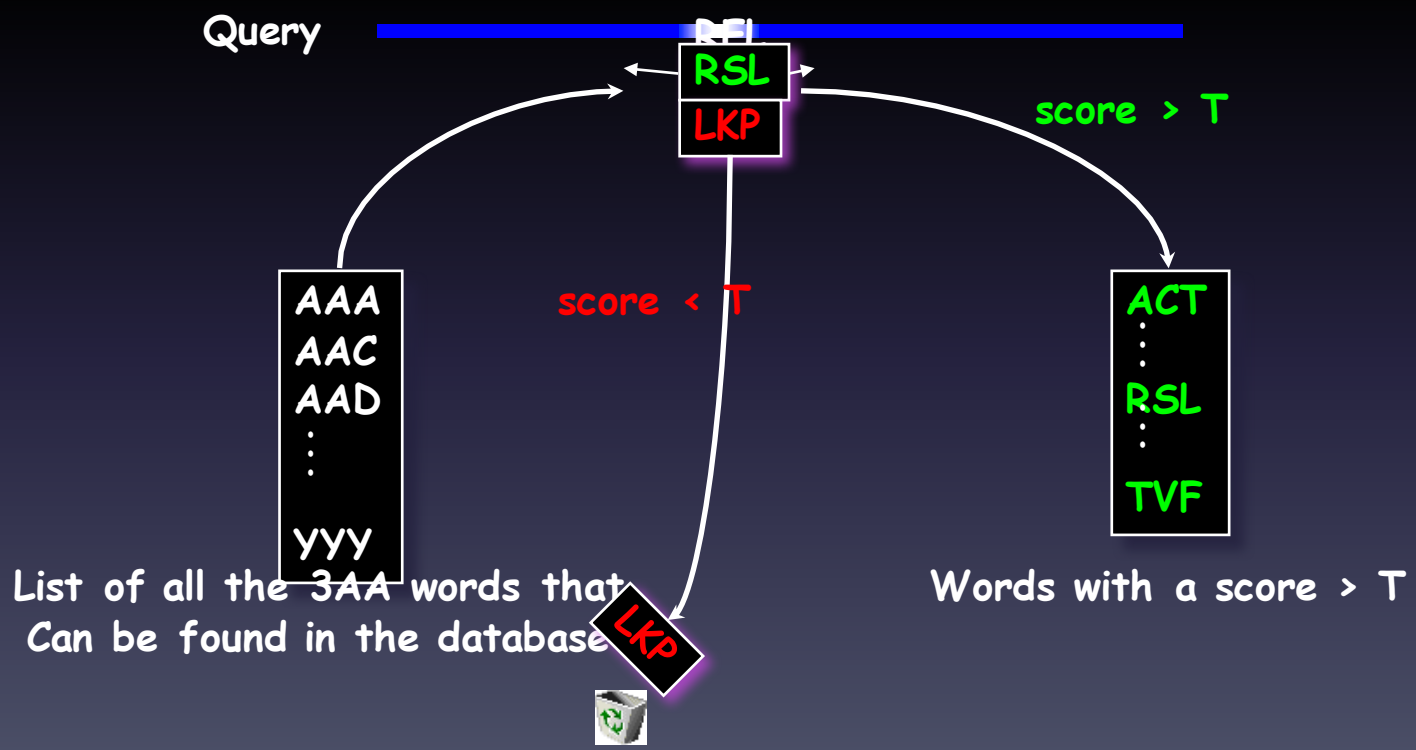
Adapted from Cedric Notredame



Standard Blastn
with long word size

MegaBLAST=Longer Words

Faster BUT Less sensitive



Adapted from Cedric Notredame

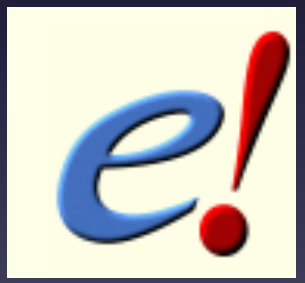
Adapted from Cedric Notredame

The NcBi BlAsT GEnoMe SecTion is MesSy

Adapted from Cedric Notredame

When it comes to BLASTing Eukaryotic Genomes:

WWW.ENSEMBL.ORG

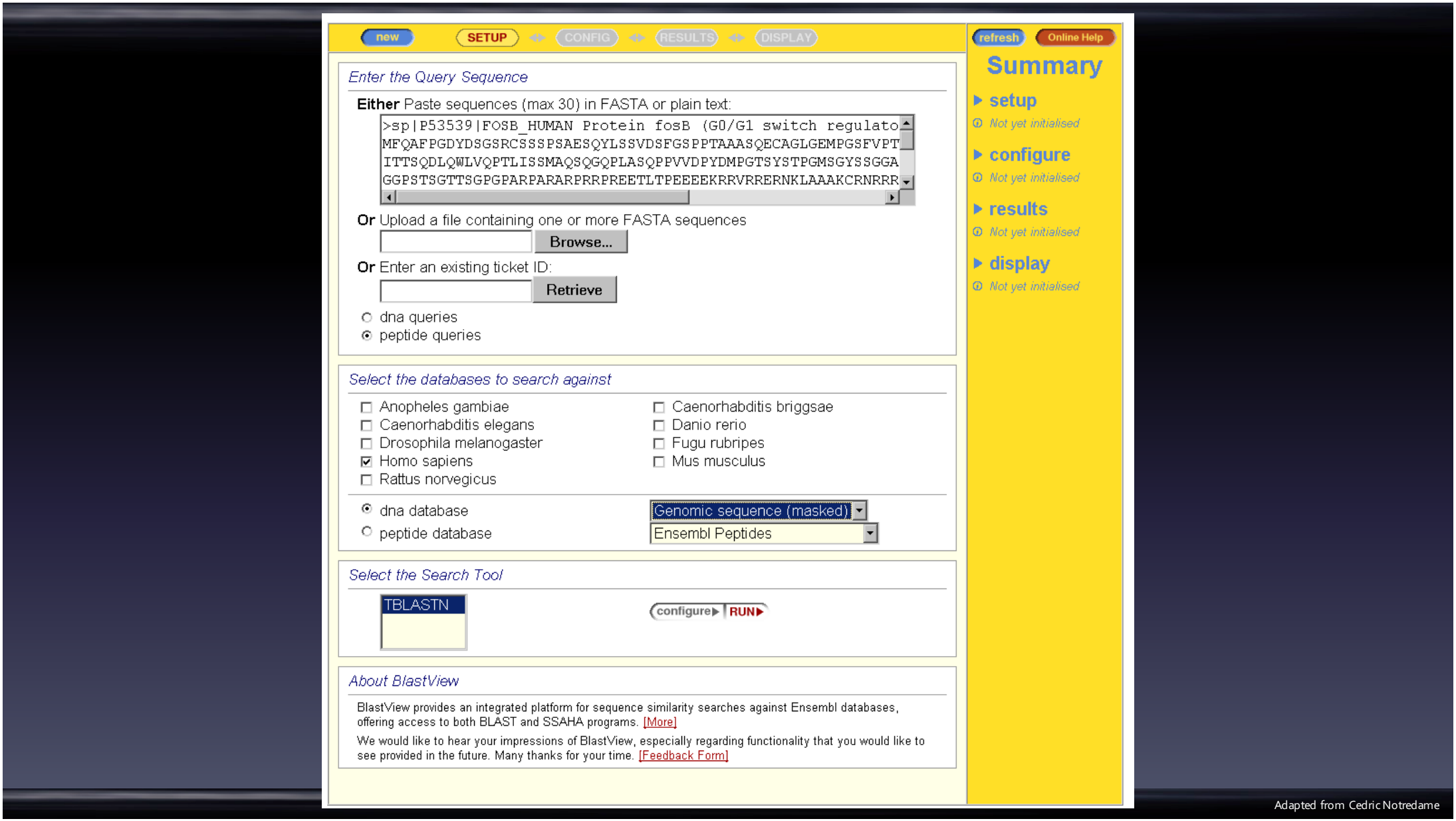


Adapted from Cedric Notredame

Asking a Question With ENSEMBL-BLAST

ENSEMBL:

***WHERE** are located the genes coding for Homologues of my protein*



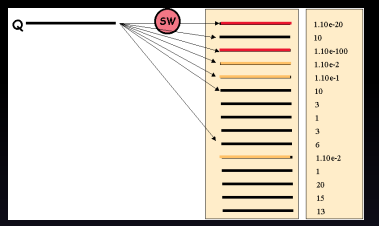
Adapted from Cedric Notredame

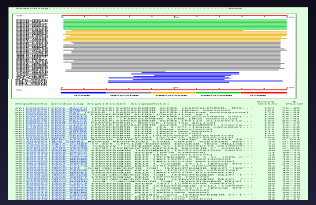
Conclusion: Searching Databases

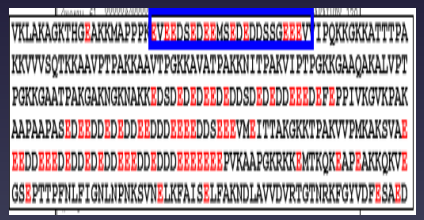
-BLAST is a fast approximation for the Full Local Dynamic Programming. It is convenient to scan Databases.

-BLAST computes the Statistical Significance of the Alignments (E-Value, P-Value).

-The main pitfall to avoid are low complexity regions

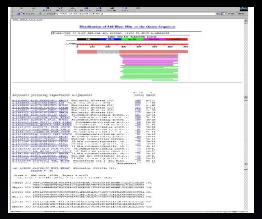




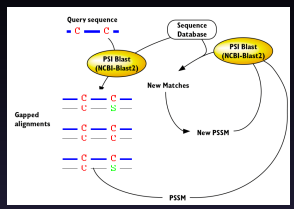


Searching Databases

-USE **blastp** the best educated blast to discover the function of your protein

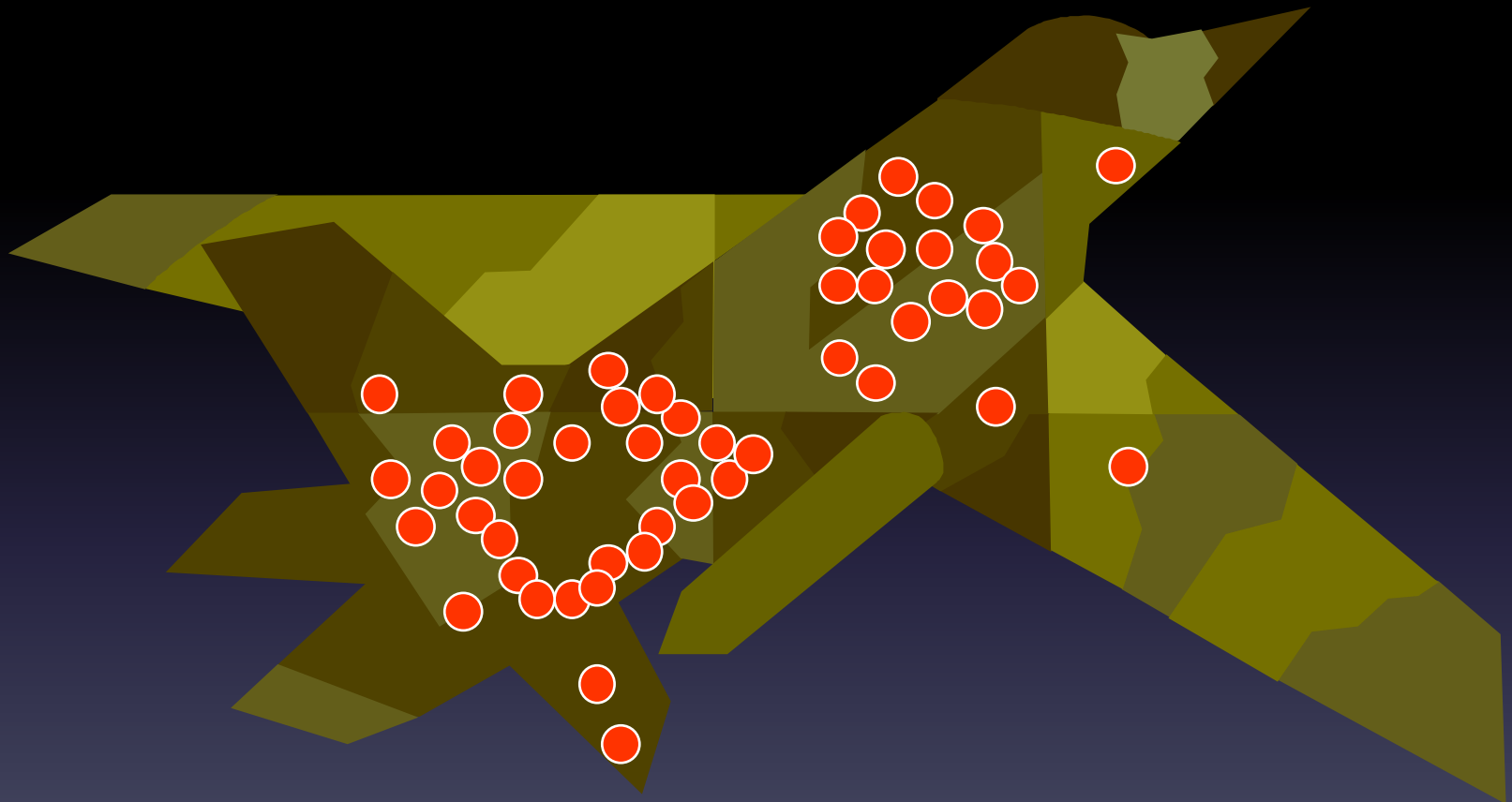


-USE **Psi-Blast** to find remote homologues



-USE **ENSEMBL-Blast** for the human Genome

SEQUENCE ALIGNMENT – MULTIPLE SEQUENCES



Manguel M, Samaniego F.J., ***Abraham Wald's Work on Aircraft Survivability,*** J. American Statistical Association. 79, 259-270, (1984)

Adapted from Cedric Notredame

Why Do We Need Multiple Sequence Alignment ?

Adapted from Cedric Notredame

Sometimes Two Sequences Are Not Enough…

The man with TWO watches
NEVER knows the time

Adapted from Cedric Notredame

What is A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

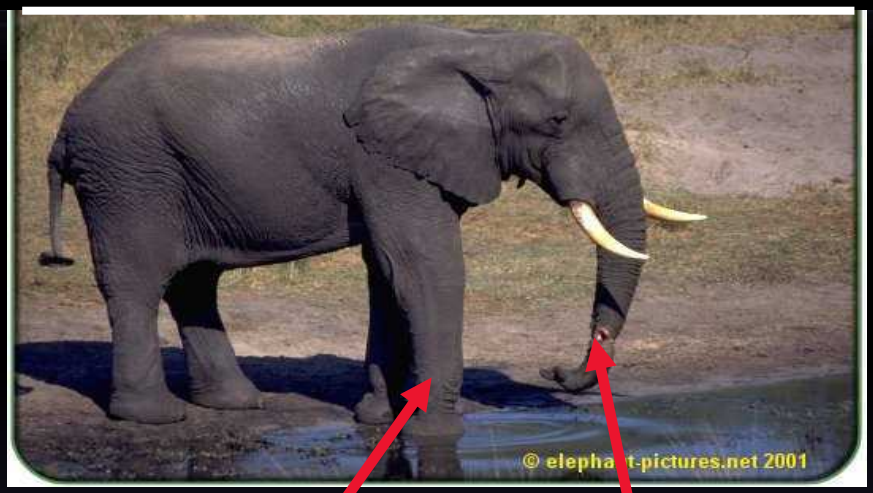
Structural Criteria:

Residues are arranged so that those playing a similar role end up in the same column.

Evolution Criteria:

Residues are arranged so that those having the same ancestor end up in the same column.

Adapted from Cedric Notredame



Phylogenic Relation

Functional Relation

Adapted from Cedric Notredame

Main Criteria for building a multiple sequence alignment

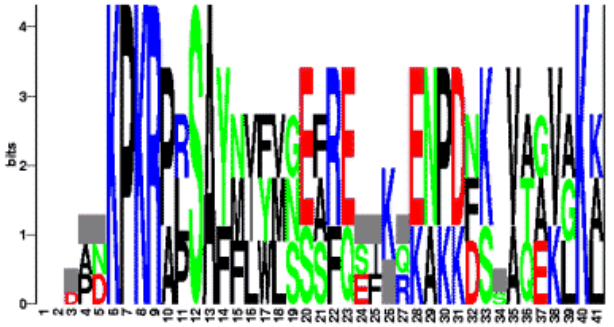
Criterion	*Meaning*
Structure similarity	**Amino acids that play the same role in each structure are in the same column.** Structure superposition programs are the only ones that use this criterion.
Evolutionary similarity	**Amino acids or nucleotides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column.** No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	**Amino acids or nucleotides with the same function are in the same column.** No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it or you can edit your alignment manually.
Sequence similarity	**Amino acids in the same column are those that yield an alignment with maximum similarity.** Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, structure, evolutionary and functional similarities are equivalent to sequence similarity.

How Can I Use A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..   .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Extrapolation

Prosite Patterns



Adapted from Cedric Notredame

How Can I Use A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .     *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Extrapolation

Prosite Patterns

P-K-R-[PA]-x(1)-[ST]...

How Can I Use A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-IQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .     *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

L?

K>R

Extrapolation

Prosite Patterns

Profiles And HMMs

-More Sensitive

-More Specific

Adapted from Cedric Notredame

How Can I Use A Multiple Sequence Alignment?

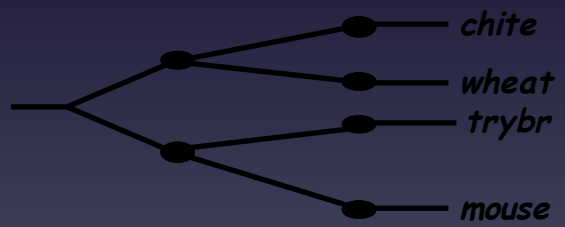
```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .     *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Extrapolation

Motifs/Patterns

Profiles

Phylogeny



-Evolution
-Paralogy/Orthology

Adapted from Cedric Notredame

How Can I Use A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..      :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Extrapolation

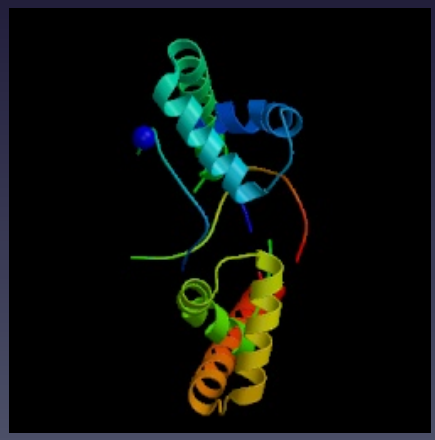
Motifs/Patterns

Profiles

Phylogeny

Struc. Prediction

Column Constraint
⇔
Evolution Constraint
⇔
Structure Constraint



Adapted from Cedric Notredame

How Can I Use A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..  .    :  . .     *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Automatic Multiple Sequence Alignment methods are not always perfect…

You know better…With your big BRAIN

Adapted from Cedric Notredame

Main applications of multiple sequence alignments

Application	*Procedure*
Extrapolation	A good multiple alignment can help convincing you that an uncharacterized sequence is really a member of a protein family.
Phylogenetic analysis	If you carefully chose the sequences to include in your multiple alignment, you can reconstruct the history of these proteins.
Pattern Identification	By discovering very conserved positions you can identify a region that is characteristic of a function (in proteins or in nucleic acid sequences).
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain. You can use this profile to scan databases for new members of the family.
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potential similar binding sites.
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for proteins or RNA. Sometimes it can also help building a 3-D model.
PCR analysis	A good multiple alignment can help you identifying the less degenerated portions of a protein family
nsSNP	Identify the nsSNP that are the most likely to alter the function

Adapted from Cedric Notredame

The Biological Problem.

Same as PairWise Alignment Problem

We do NOT know how Sequences Evolve.

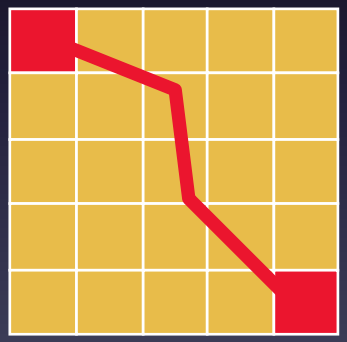
We do NOT understand the Relation Between Structures and Sequences.

We would NOT recognize the Correct Alignment if we had it IN FRONT of our eyes…

Adapted from Cedric Notredame

The COMPUTATIONAL Problem.
Producing the Alignment

- A nice set of Sequences
- Substitution Matrix (Blosum)
- Gap Penalties.
- An Evaluation Function
- An Alignment Algorithm



GLOBAL Alignment

Making An Alignment

Any Exact Method would be TOO SLOW

We will use a Heuristic Algorithm.

Progressive Alignment Algorithm is the most Popular

-ClustalW

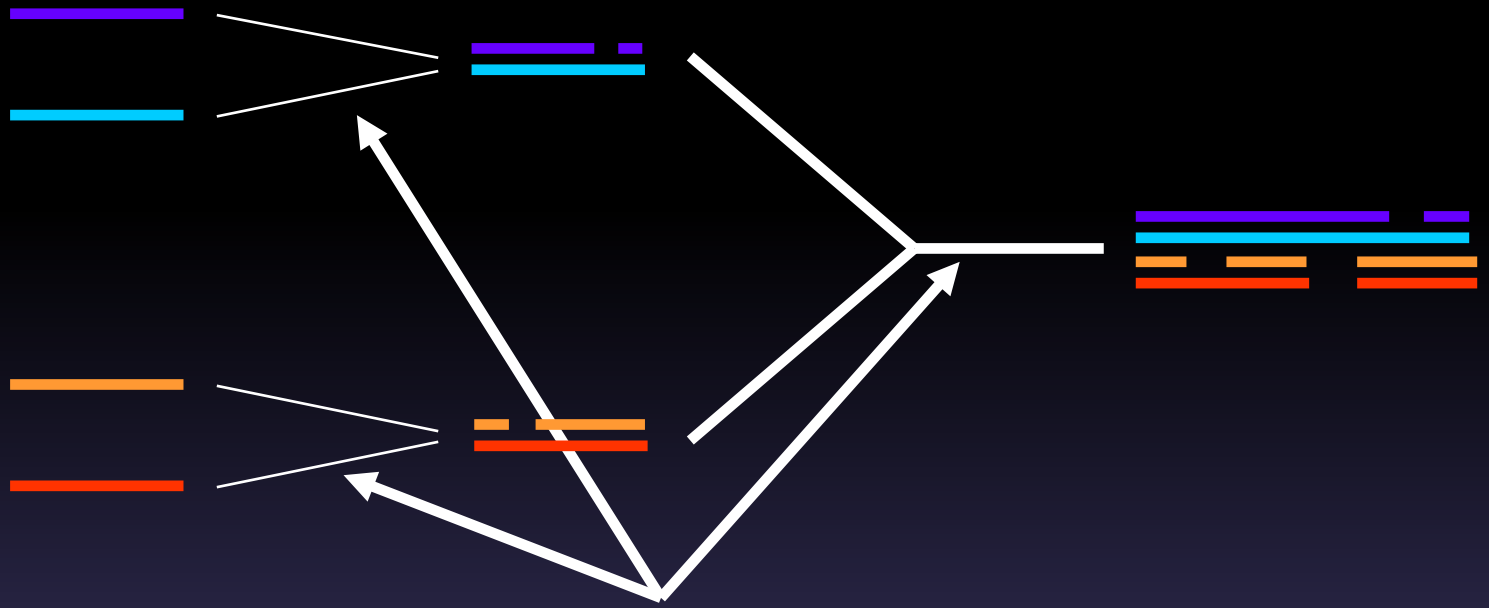
-Greedy Heuristic (No Guarranty).

-Fast

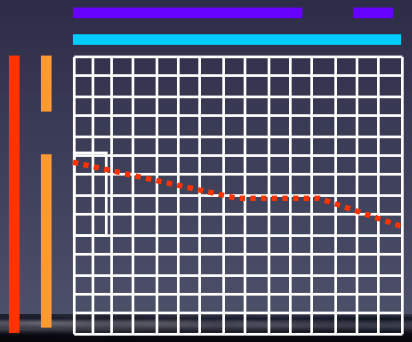
Progressive Alignment

Feng and Dolittle, 1988; Taylor 1989

Progressive Alignment

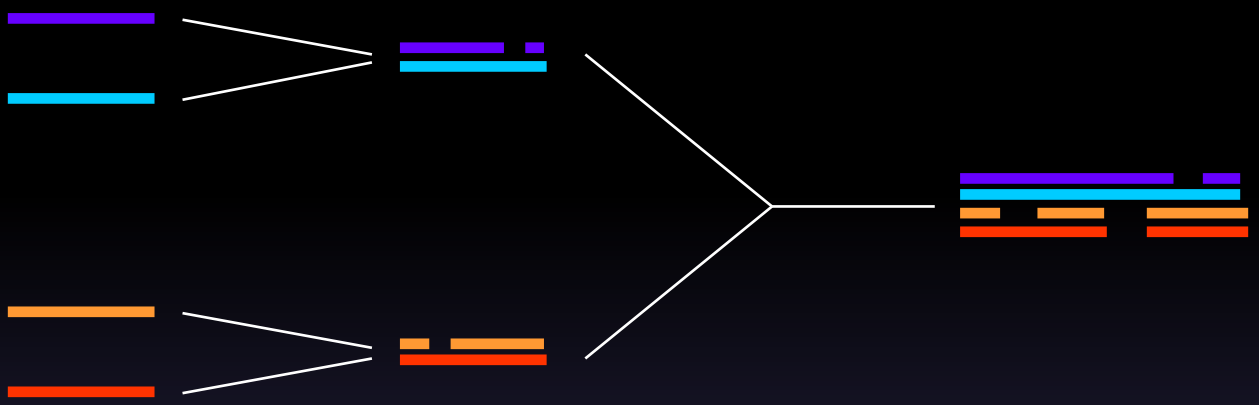


Dynamic Programming Using A Substitution Matrix



Adapted from Cedric Notredame

Progressive Alignment



- Depends on the CHOICE of the sequences.
- Depends on the ORDER of the sequences (Tree).
- Depends on the PARAMETERS:
  - Substitution Matrix.
  - Penalties (Gop, Gep).
  - Sequence Weight.
  - Tree making Algorithm.

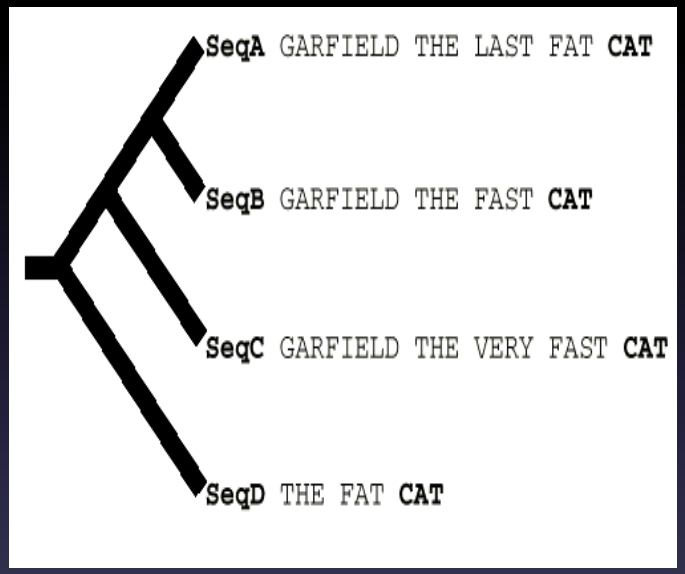
Progressive Alignment
When Does It Work

Works Well When Phylogeny is Dense

No outlayer Sequence.

Adapted from Cedric Notredame

Progressive Alignment
When Doesn't It Work




```
CLUSTALW (Score=20, Gop=-1, Gep=0, M=1)

SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD -------- THE ---- FA-T CAT
```

```
CORRECT (Score=24)

SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST ---- CAT
SeqC GARFIELD THE VERY FAST CAT
SeqD -------- THE ---- FA-T CAT
```

Common Mistake:
Sequences Too Closely Related

```
PRVA_MACFU      SMTDLLNAEDIKKAVGAFSAIDSFDHKKFFQMVGLKKKSADDVKKVFHILDKDKSGFIEE
PRVA_HUMAN      SMTDLLNAEDIKKAVGAFSATDSFDHKKFFQMVGLKKKSADDVKKVFHMLDKDKSGFIEE
PRVA_GERSP      SMTDLLSAEDIKKAIGAFAAADSFDHKKFFQMVGLKKKTPDDVKKVFHILDKDKSGFIEE
PRVA_MOUSE      SMTDVLSAEDIKKAIGAFAAADSFDHKKFFQMVGLKKKNPDEVKKVFHILDKDKSGFIEE
PRVA_RAT        SMTDLLSAEDIKKAIGAFTAADSFDHKKFFQMVGLKKKSADDVKKVFHILDKDKSGFIEE
PRVA_RABIT      AMTELLNAEDIKKAIGAFAAAESFDHKKFFQMVGLKKKSTEDVKKVFHILDKDKSGFIEE
                :**::*.*******:***:*  :***************..::******:**********

PRVA_MACFU      DELGFILKGFSPDARDLSAKETKTLMAAGDKDGDGKIGVDEFSTLVAES
PRVA_HUMAN      DELGFILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES
PRVA_GERSP      DELGFILKGFSSDARDLSAKETKTLLAAGDKDGDGKIGVEEFSTLVSES
PRVA_MOUSE      DELGSILKGFSSDARDLSAKETKTLLAAGDKDGDGKIGVEEFSTLVAES
PRVA_RAT        DELGSILKGFSSDARDLSAKETKTLMAAGDKDGDGKIGVEEFSTLVAES
PRVA_RABIT      EELGFILKGFSPDARDLSVKETKTLMAAGDKDGDGKIGADEFSTLVSES
                :***  ******.******.****  *:************.:******:**
```

-IDENTICAL SEQUENCES BRING NO INFORMATION FOR THE MULTIPLE SEQUENCE ALIGNMENT

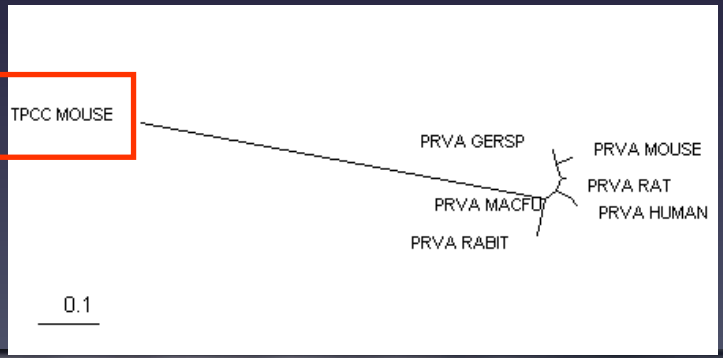
-MULTIPLE SEQUENCE ALIGNMENTS THRIVE ON DIVERSITY…

Adapted from Cedric Notredame

Respect Information!

```
PRVA_MACFU      ----------------------------------------SMTDLLN----AEDIKKA
PRVA_HUMAN      ----------------------------------------SMTDLLN----AEDIKKA
PRVA_GERSP      ----------------------------------------SMTDLLS----AEDIKKA
PRVA_MOUSE      ----------------------------------------SMTDVLS----AEDIKKA
PRVA_RAT        ----------------------------------------SMTDLLS----AEDIKKA
PRVA_RABIT      ----------------------------------------AMTELLN----AEDIKKA
TPCC_MOUSE      MDDIYKAAVEQLTEEQKNEFKAAFDIFVLGAEDGCISTKELGKVMRMLGQNPTPEELQEM
                                                       :  :*.     .*::::

PRVA_MACFU      VGAFSAIDS--FDHKKFFQMVG------LKKKSADDVKKVFHILDKDKSGFIEEDELGFI
PRVA_HUMAN      VGAFSATDS--FDHKKFFQMVG------LKKKSADDVKKVFHMLDKDKSGFIEEDELGFI
PRVA_GERSP      IGAFAAADS--FDHKKFFQMVG------LKKKTPDDVKKVFHILDKDKSGFIEEDELGFI
PRVA_MOUSE      IGAFAAADS--FDHKKFFQMVG------LKKKNPDEVKKVFHILDKDKSGFIEEDELGSI
PRVA_RAT        IGAFTAADS--FDHKKFFQMVG------LKKKSADDVKKVFHILDKDKSGFIEEDELGSI
PRVA_RABIT      IGAFAAAES--FDHKKFFQMVG------LKKKSTEDVKKVFHILDKDKSGFIEEEELGFI
TPCC_MOUSE      IDEVDEDGSGTVDFDEFLVMMVRCMKDDSKGKSEEELSDLFRMFDKNADGYIDLDELKMM
```

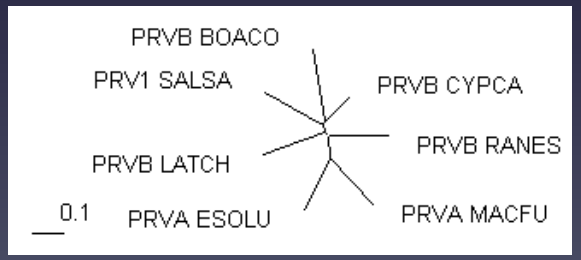
This Alignment Is not Informative about the relation Between TPCC MOUSE and the rest of the sequences.



-A better Spread of the Sequences is needed

Selecting Diverse Sequences (Opus II)

```
PRVB_CYPCA      -AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAIIDQDKSGFIE
PRVB_BOACO      -AFAGILSDADIAAGLQSCQAADSFSCKTFFAKSGLHSKSKDQLTKVFGVIDRDKSGYIE
PRV1_SALSA      MACAHLCKEADIKTALEACKAADTFSFKTFFHTIGFASKSADDVKKAFKVIDQDASGFIE
PRVB_LATCH      -AVAKLLAAADVTAALEGCKADDSFNHKVFFQKTGLAKKSNEELEAIFKILDQDKSGFIE
PRVB_RANES      -SITDIVSEKDIDAALESVKAAGSFNYKIFFQKVGLAGKSAADAKKVFEILDRDKSGFIE
PRVA_MACFU      -SMTDLLNAEDIKKAVGAFSAIDSFDHKKFFQMVGLKKKSADDVKKVFHILDKDKSGFIE
PRVA_ESOLU      --AKDLLKADDIKKALDAVKAEGSFNHKKFFALVGLKAMSANDVKKVFKAIDADASGFIE
                      :    *:   .: . .* .:*. * **    *:    *  :    *  :* * **:**

PRVB_CYPCA      EDELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVKA-
PRVB_BOACO      EDELKKFLQNFDGKARDLTDKETAEFLKEGDTDGDGKIGVEEFVVLVTKG
PRV1_SALSA      VEELKLFLQNFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ-
PRVB_LATCH      DEELELFLQNFSAGARTLTKTETETFLKAGDSDGDGKIGVDEFQKLVKA-
PRVB_RANES      QDELGLFLQNFRASARVLSDAETSAFLKAGDSDGDGKIGVEEFQALVKA-
PRVA_MACFU      EDELGFILKGFSPDARDLSAKETKTLMAAGDKDGDGKIGVDEFSTLVAES
PRVA_ESOLU      EEELKFVLKSFAADGRDLTDAETKAFLKAADKDGDGKIGIDEFETLVHEA
                 :**   .*:.*    .* *:   **  ::   .* **** **::**   **
```



- A REASONABLE Model Now Exists.
- Going Further: Remote Homologues.

Adapted from Cedric Notredame

Aligning Remote Homologues

```
PRVA_MACFU    ----------------------------------------SMTDLLNA----EDIKKA
PRVA_ESOLU    ----------------------------------------AKDLLKA----DDIKKA
PRVB_CYPCA    ----------------------------------------AFAGVLND----ADIAAA
PRVB_BOACO    ----------------------------------------AFAGILSD----ADIAAG
PRV1_SALSA    ----------------------------------------MACAHLCKE----ADIKTA
PRVB_LATCH    ----------------------------------------AVAKLLAA----ADVTAA
PRVB_RANES    ----------------------------------------SITDIVSE----KDIDAA
TPCS_RABIT    -TDQQAEARSYLSEEMIAEFKAAFDMFDADGG-GDISVKELGTVMRMLGQTPTKEELDAI
TPCS_PIG      -TDQQAEARSYLSEEMIAEFKAAFDMFDADGG-GDISVKELGTVMRMLGQTPTKEELDAI
TPCC_MOUSE    MDDIYKAAVEQLTEEQKNEFKAAFDIFVLGAEDGCISTKELGKVMRMLGQNPTPEELQEM
                                                                :       ::

PRVA_MACFU    VGAFSAIDS--FDHKKFFQMVG------LKKKSADDVKKVFHILDKDKSGFIEEDELGFI
PRVA_ESOLU    LDAVKAEGS--FNHKKFFALVG------LKAMSANDVKKVFKAIDADASGFIEEEELKFV
PRVB_CYPCA    LEACKAADS--FNHKAFFAKVG------LTSKSADDVKKAFAIIDQDKSGFIEEDELKLF
PRVB_BOACO    LQSCQAADS--FSCKTFFAKSG------LHSKSKDQLTKVFGVIDRDKSGYIEEDELKKF
PRV1_SALSA    LEACKAADT--FSFKTFFHTIG------FASKSADDVKKAFKVIDQDASGFIEVEELKLF
PRVB_LATCH    LEGCKADDS--FNHKVFFQKTG------LAKKSNEELEAIFKILDQDKSGFIEDEELELF
PRVB_RANES    LESVKAAGS--FNYKIFFQKVG------LAGKSAADAKKVFEILDRDKSGFIEQDELGLF
TPCS_RABIT    IEEVDEDGSGTIDFEEFLVMMVRQMKEDAKGKSEEELAECFRIFDRNADGYIDAEELAEI
TPCS_PIG      IEEVDEDGSGTIDFEEFLVMMVRQMKEDAKGKSEEELAECFRIFDRNMDGYIDAEELAEI
TPCC_MOUSE    IDEVDEDGSGTVDFDEFLVMMVRCMKDDSKGKSEEELSDLFRMFDKNADGYIDLDELKMM
              :   .  .:  .. . *:             *  :    *  :* : .*:*: :**  .

PRVA_MACFU        LKGFSPDARDLSAKETKTLMAAGDKDGDGKIGVDEFSTLVAES-
PRVA_ESOLU        LKSFAADGRDLTDAETKAFLKAADKDGDGKIGIDEFETLVHEA-
PRVB_CYPCA        LQNFKADARALTDGETKTFLKAGDSDGDGKIGVDEFTALVKA--
PRVB_BOACO        LQNFDGKARDLTDKETAEFLKEGDTDGDGKIGVEEFVVLVTKG-
PRV1_SALSA        LQNFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ--
PRVB_LATCH        LQNFSAGARTLTKTETETFLKAGDSDGDGKIGVDEFQKLVKA--
PRVB_RANES        LQNFRASARVLSDAETSAFLKAGDSDGDGKIGVEEFQALVKA--
TPCS_RABIT        FR---ASGEHVTDEEIESLMKDGDKNNDGRIDFDEFLKMMEGVQ
TPCS_PIG          FR---ASGEHVTDEEIESIMKDGDKNNDGRIDFDEFLKMMEGVQ
TPCC_MOUSE        LQ---ATGETITEDDIEELMKDGDKNNDGRIDYDEFLEFMKGVE
                  ::      .. ::  :   ::  .* :.** *. :**  ::
```

Adapted from Cedric Notredame

WHAT MAKES A GOOD ALIGNMENT…

-THE MORE DIVERGEANT THE SEQUENCES, THE BETTER

-THE FEWER INDELS, THE BETTER

-NICE UNGAPPED BLOCKS SEPARATED WITH INDELS

-DIFFERENT CLASSES OF RESIDUES WITHIN A BLOCK:

- Completely Conserved
- Conserved For Size *and* Hydropathy
- Conserved For Size *or* Hydropathy

-THE ULTIMATE EVALUATION IS A MATTER OF PERSONNAL JUDGEMENT AND KNOWLEDGE.

DO NOT OVERTUNE!!!

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..   .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

DO NOT PLAY WITH PARAMETERS IF YOU KNOW THE ALIGNMENT YOU WANT: MAKE IT YOURSELF!

```
chite   ---ADKPKRPL-SAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAP-SAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPR-SAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. :*: .: ..   .    :  . .      *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

Adapted from Cedric Notredame

KEEP A BIOLOGICAL PERSPECTIVE

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
              ***. ::: .: ..   .    :  . .      *  .  *: *
```

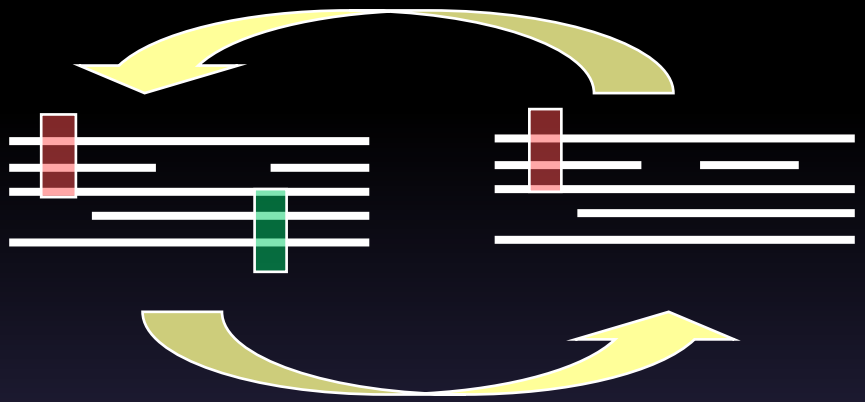
DIFFERENT PARAMETERS

```
chite AD--K----PKR-PLYMLWLNS-ARESIKRENPDFK-VT-EVAKKGGELWRGL-
wheat -DPNK----PKRAP-FFVFMGE-FREEFKQKNPKNKSVA-AVGKAAGERWKSLS
trybr -K--KDSNAPKR-AMT-MFFSSDFR-S-KH-S-DLS-IV-EMSKAAGAAWKELG
mouse ----K----PKR-PRYNIYVSESFQEA-K--D-D-S-AQGKL-KLVNEAWKNLS
          *    *** .:: ::... :   *  . . .     : *  .  *: *
```

WRONG ALIGNMENT !!!

Adapted from Cedric Notredame

Iterative Methods



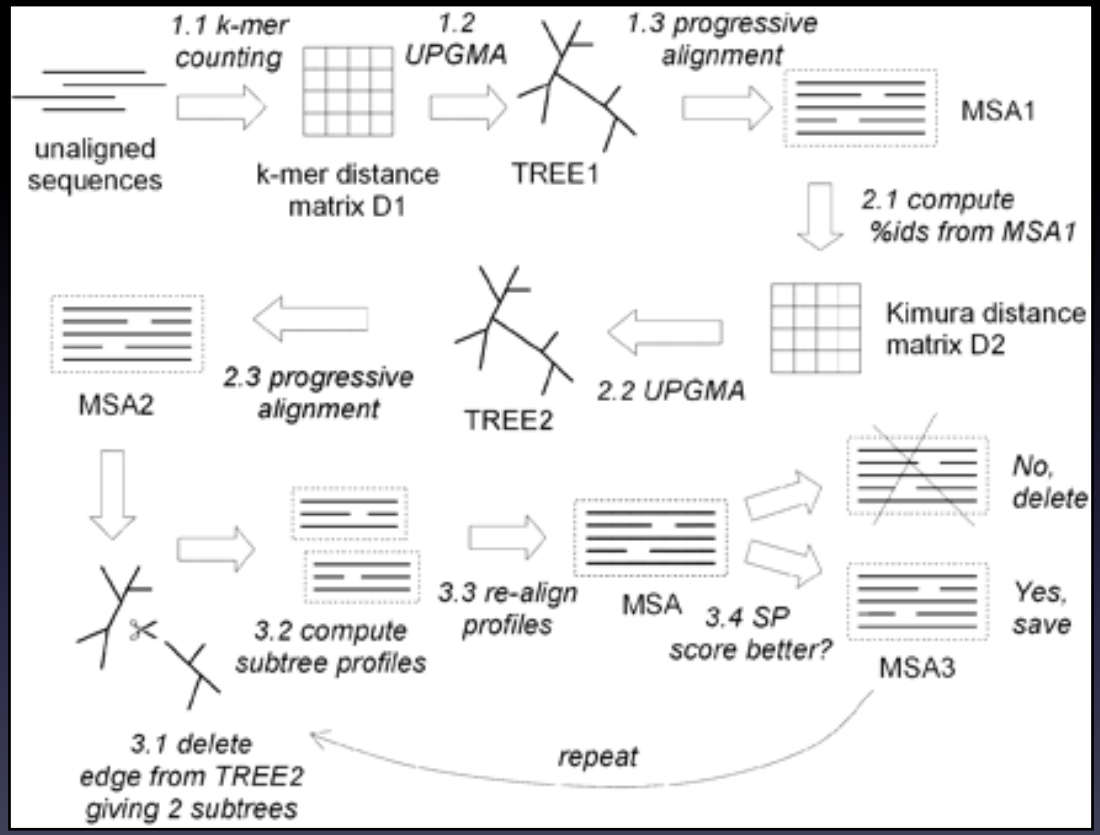
-HMMs, HMMER, SAM, MUSCLE

−-Slow, Sometimes Inaccurate

+-Good Profile Generators

Adapted from Cedric Notredame

MUSCLE



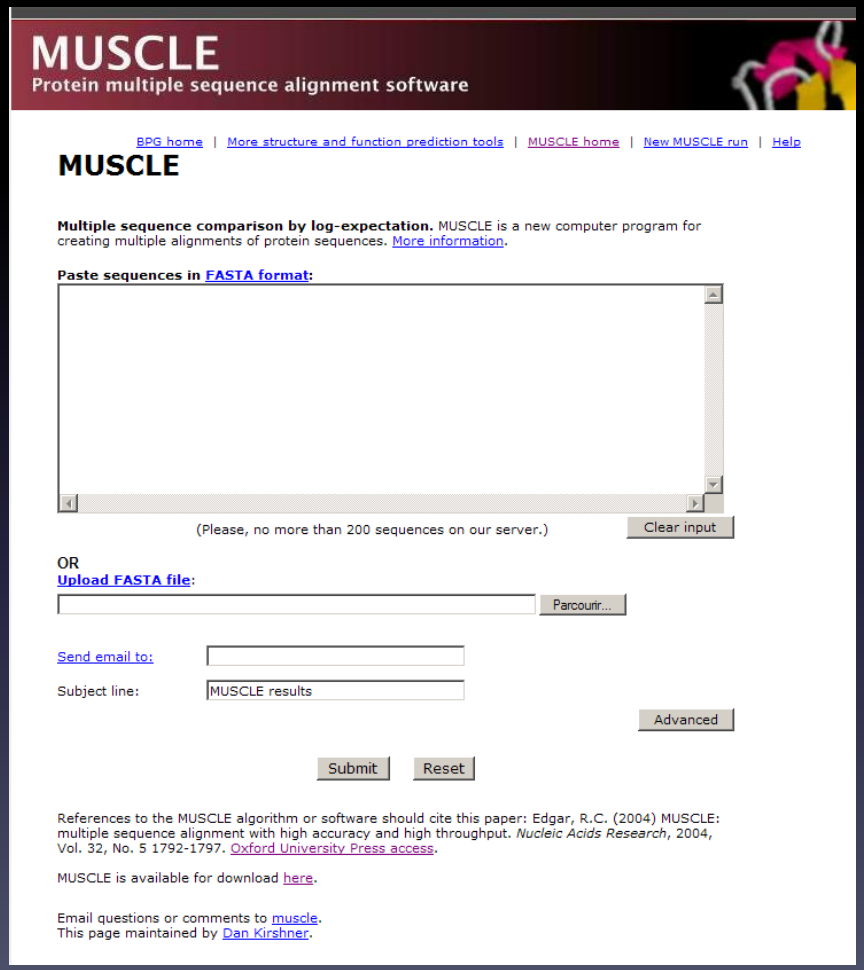
1: Edgar RC.

MUSCLE: a multiple sequence alignment method with reduced time and space complexity.
BMC Bioinformatics. 2004 Aug 19;5(1):113.
PMID: 15318951 [PubMed - indexed for MEDLINE]

Adapted from Cedric Notredame

MUSCLE

phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py



Adapted from Cedric Notredame

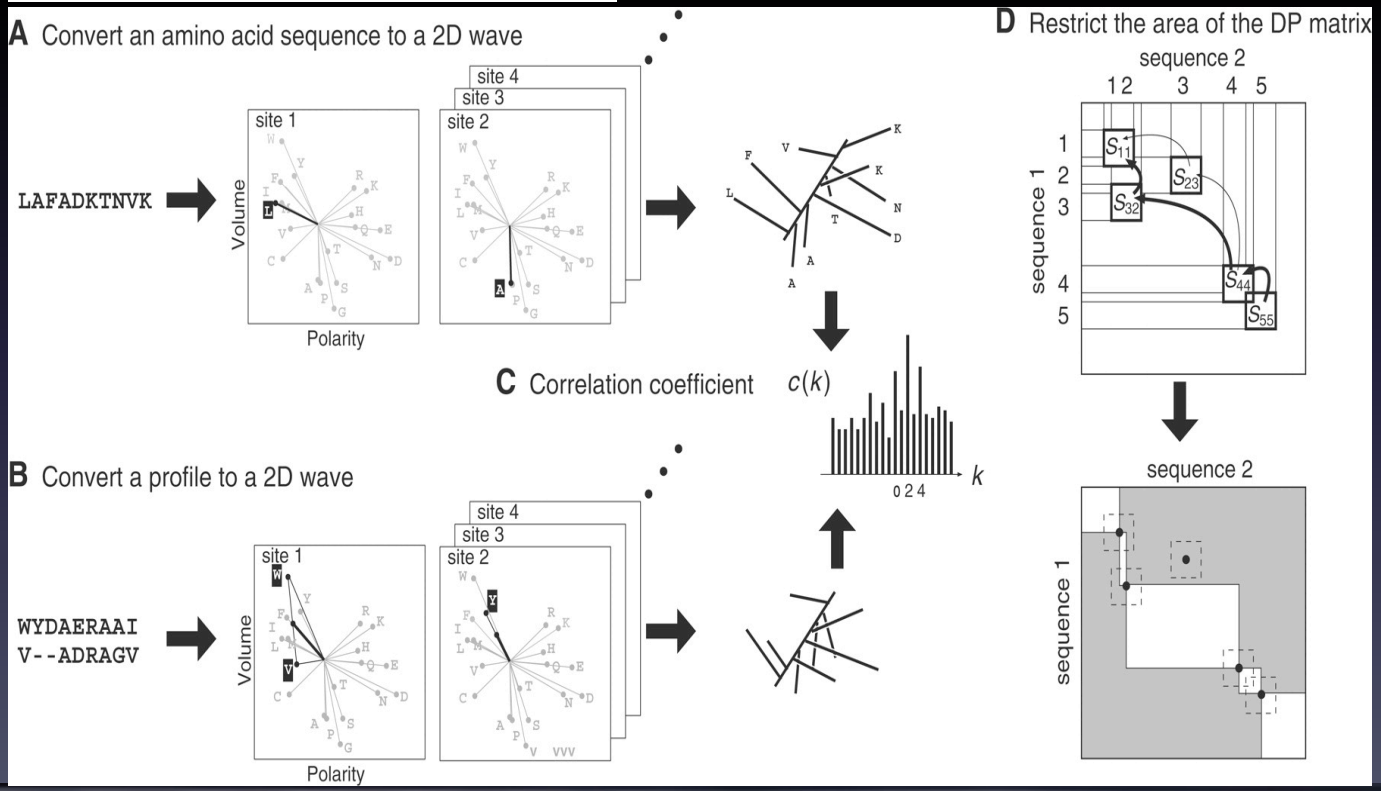
MAFFT
Fast Fourrier Transformé

BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 4. 286–298
Advance Access publication March 27, 2008
doi:10.1093/bib/bbn003

Recent developments in the MAFFT multiple sequence alignment program

Kazutaka Katoh and Hiroyuki Toh

Submitted: 14th January 2008; Received (in revised form): 26th February 2008



Adapted from Cedric Notredame

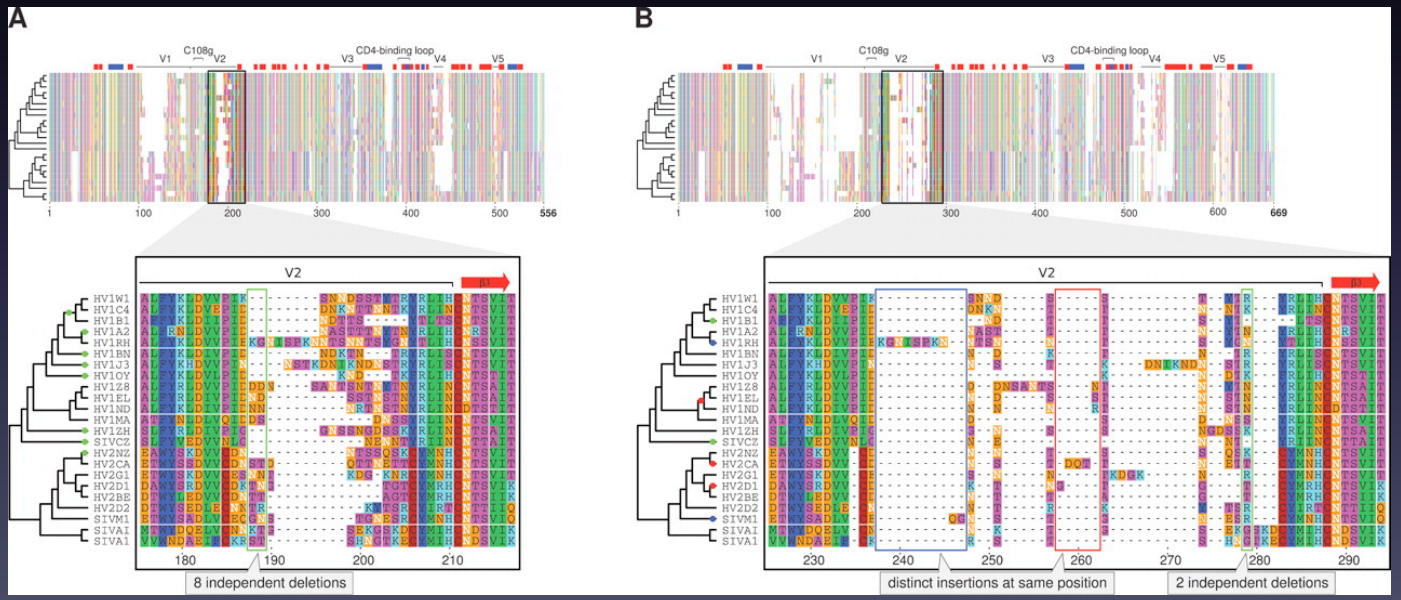
Prank

Science. 2008 Jun 20;320(5883):1632-5.

Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis.

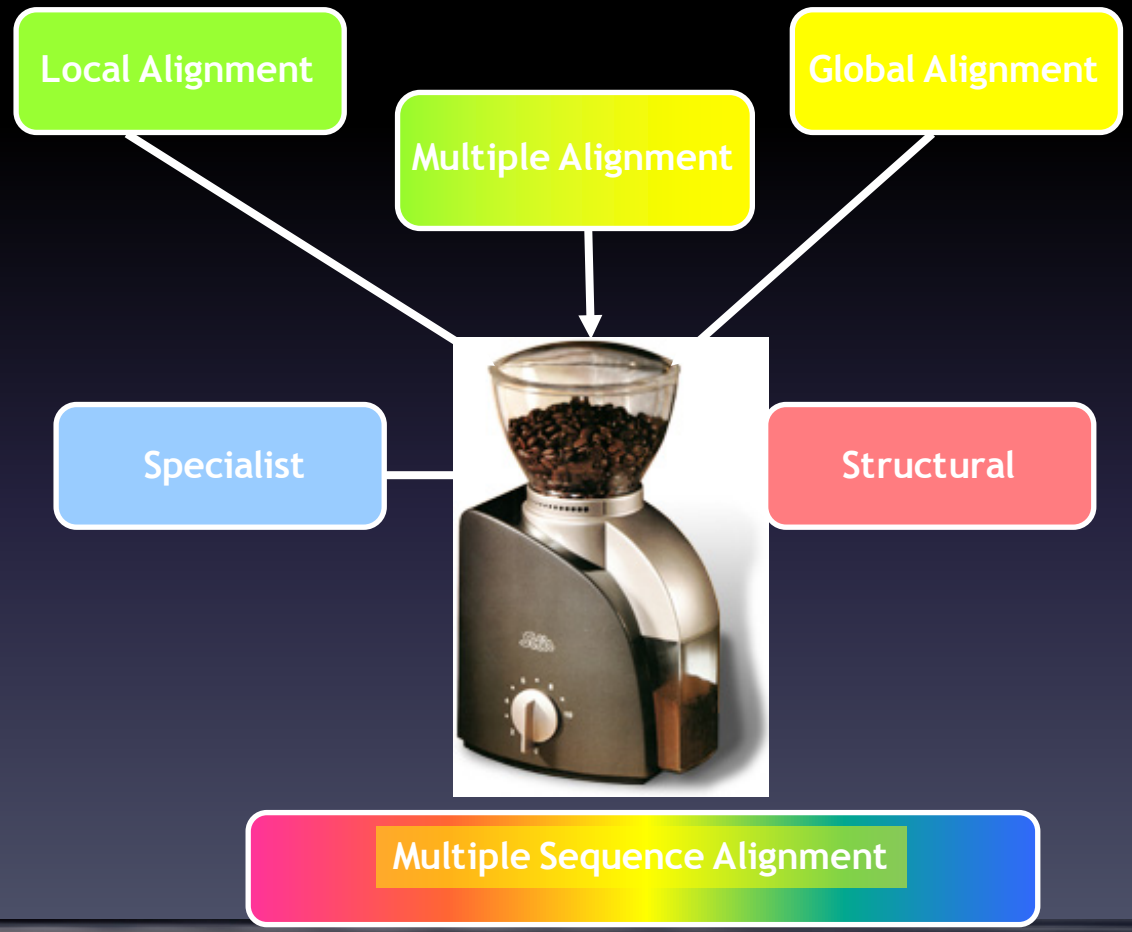
Löytynoja A, Goldman N.

European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK. ari@ebi.ac.uk



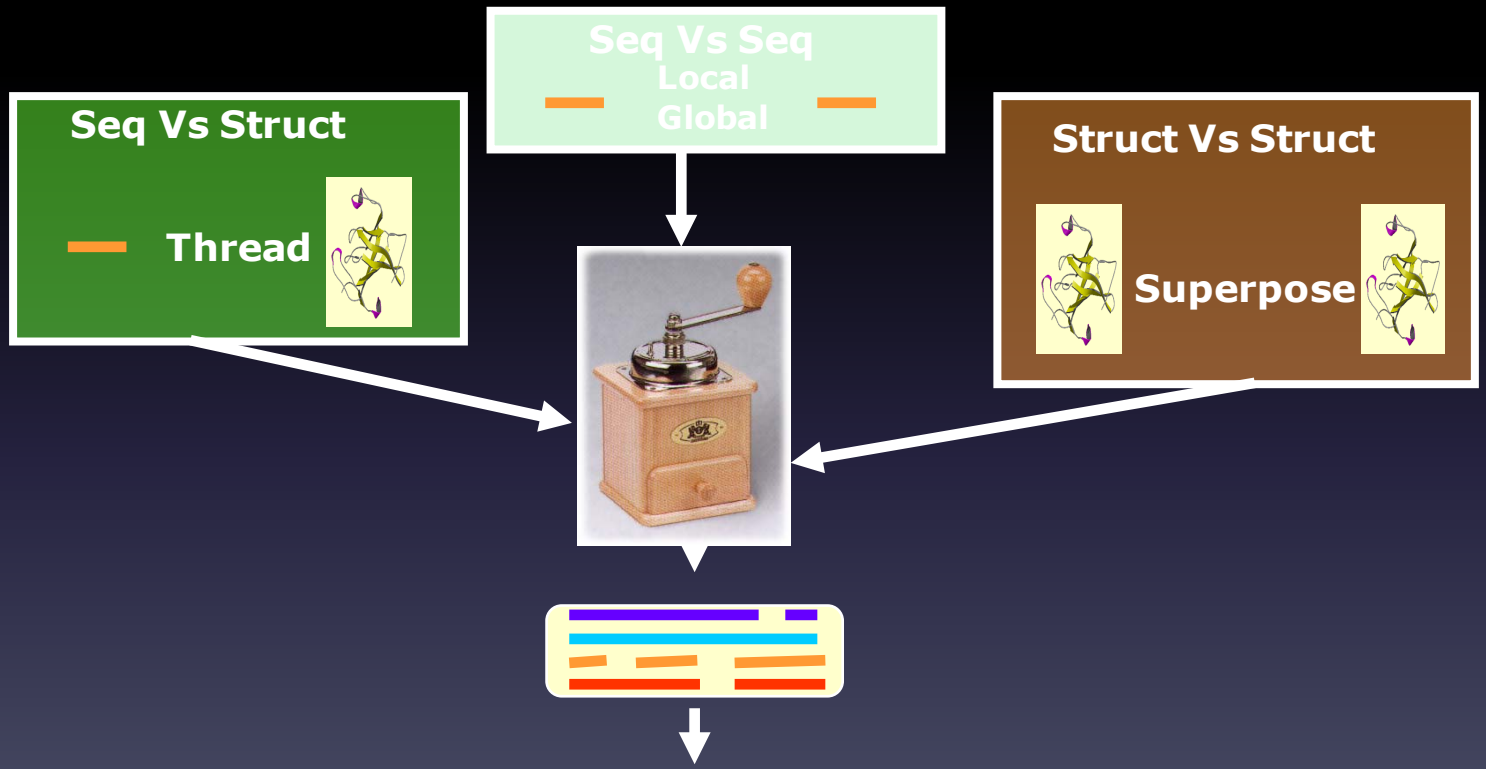
Adapted from Cedric Notredame

Mixing Heterogenous Data With T-Coffee



Adapted from Cedric Notredame

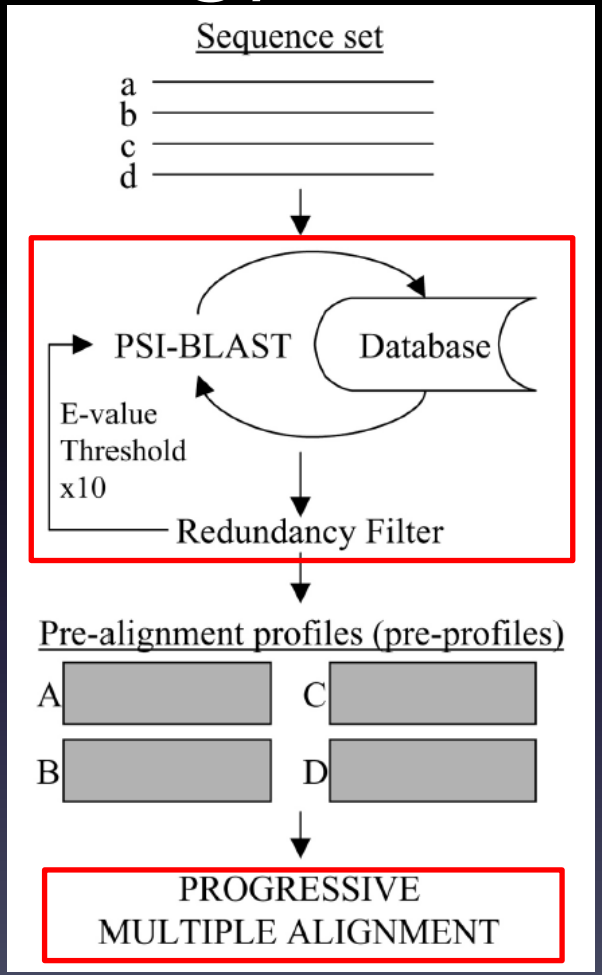
Mixing Sequences and Structures with T-Coffee



Evaluation on Homestrad

Adapted from Cedric Notredame

Homology-extended



Que1: how to build a profile?

Que2: how to score profiles?

Simossis VA, Kleinjung J, Heringa J: **Homology-extended sequence alignment**. *Nucleic Acids Res* 2005, **33**(3):816-824.

Adapted from Cedric Notredame

Que1: how to build a profile?

- Database Size
- Searching parameters
  - E-value : most used, anything else???
    1. Matrix file : -M
    2. Filter the query sequence for low-complexity subsequence : -F
    3. Neighborhood word threshold : -f
    4. Truncates the report to number of alignments: -b

Adapted from Cedric Notredame

Word hit & Neighborhood

Table 5-2. The neighborhood near RGD

BLOSUM62		PAM200	
Word	Score	Word	Score
RGD	17	RGD	18
KGD	14	RGE	17
QGD	13	RGN	16
RGE	13	KGD	15
EGD	12	RGQ	15
HGD	12	KGE	14
NGD	12	HGD	13
RGN	12	KGN	13
AGD	11	RAD	13
MGD	11	RGA	13
RAD	11	RGG	13
RGQ	11	RGH	13
RGS	11	RGK	13
RND	11	RGS	13
RSD	11	RGT	13
SGD	11	RSD	13
TGD	11	WGD	13

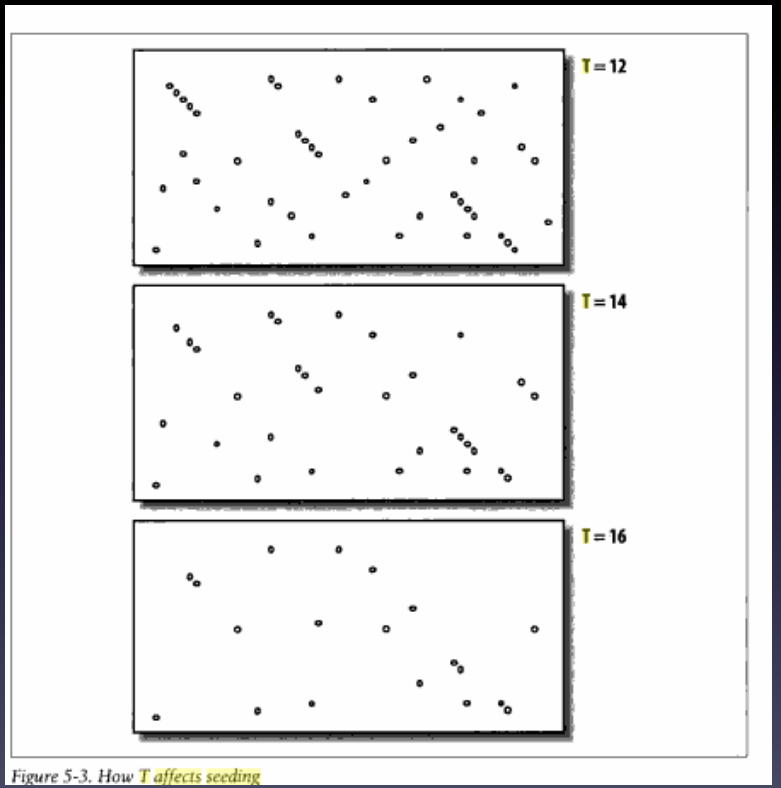


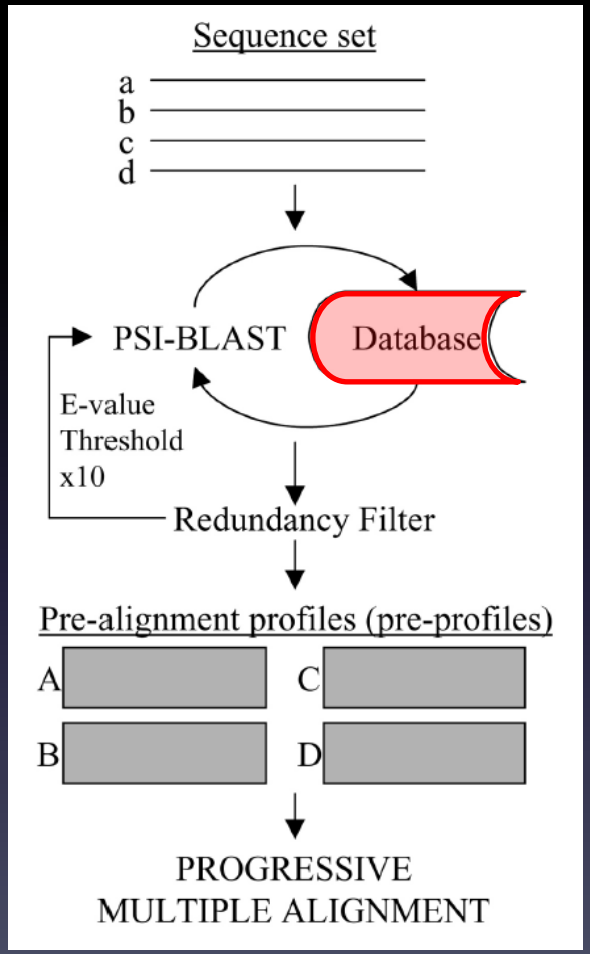
Figure 5-3. How T affects seeding

Adapted from Cedric Notredame

Searching parameters

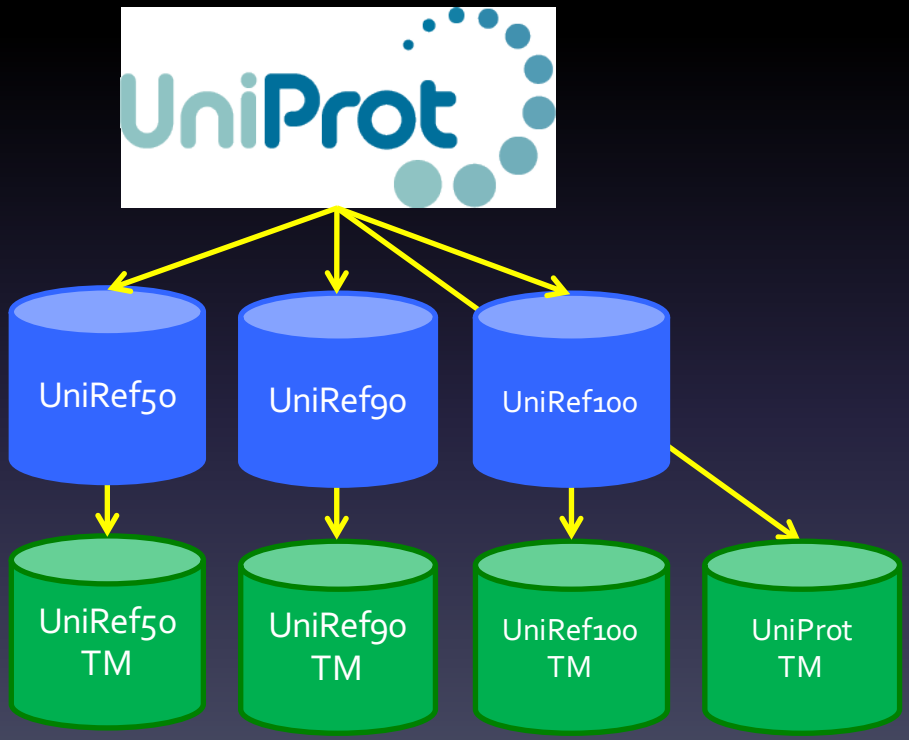
- Fast, Insensitive search
  - High percent identity
  - blastp –F "m S" –f 999 –M BLOSUM80 –G 9 –E 2 –e 1e-5
- Slow, Sensitive search
  - Increase sensitivity, decrease specificity
  - blastp –F "m S" –f 9 –M BLOSUM45 –e 100 –b 10000 –v 10000
- Book "BLAST", page 146, 147

Different database



NCBI non-redundant (NR)

UniProt (release 15.15 – 2010)



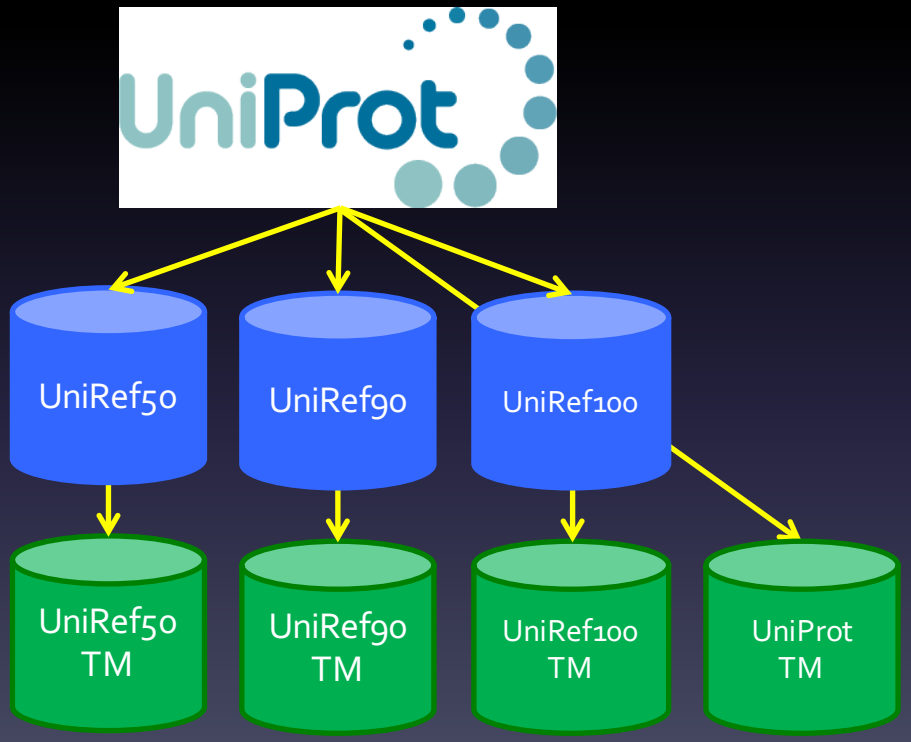
keyword:"Transmembrane [KW-0812]"

Adapted from Cedric Notredame

Database Size

NCBI non-redundant (NR)

UniProt (release 15.15 – 2010)



keyword:"Transmembrane [KW-0812]"

Data Set	No.
UniRef50-TM	87,989
UniRef90-TM	263,306
UniRef100-TM	613,015
UniProt-TM	818,635
UniRef50	3,077,464
UniRef90	6,544,144
UniRef100	9,865,668
UniProt	11,009,767
NCBI NR	10,565,004

Adapted from Cedric Notredame

Performance comparison of different database sizes for the BAliBASE2-ref7.

database	# of seqs	SP	TC	extension(s)	total(s)
default T-Coffee	0	0.911	0.498	0	2,735
UniRef50-TM	87,989	0.916	0.561	1,483	8,177
UniRef90-TM	263,306	0.918	0.548	3,343	9,610
UniRef100-TM	613,015	0.925	0.545	6,499	12,111
UniProt-TM	818,635	0.923	0.536	7,871	13,285
UniRef50	3,077,464	0.920	0.553	19,087	26,442
UniRef90	6,544,144	0.924	0.561	40,448	46,478
UniRef100	9,865,668	0.922	0.554	66,696	71,895
UniProt	11,009,767	0.923	0.563	66,964	72,199
NCBI NR	10,565,004	0.921	0.554	65,201	70,375

UniRef50-TM contains about 100 times fewer sequences than the full UniProt.

The level accuracy is comparable and even superior to that achieved with the default PSI-Coffee while the CPU time requirements are dramatically decreased by a factor 10.

Adapted from Cedric Notredame

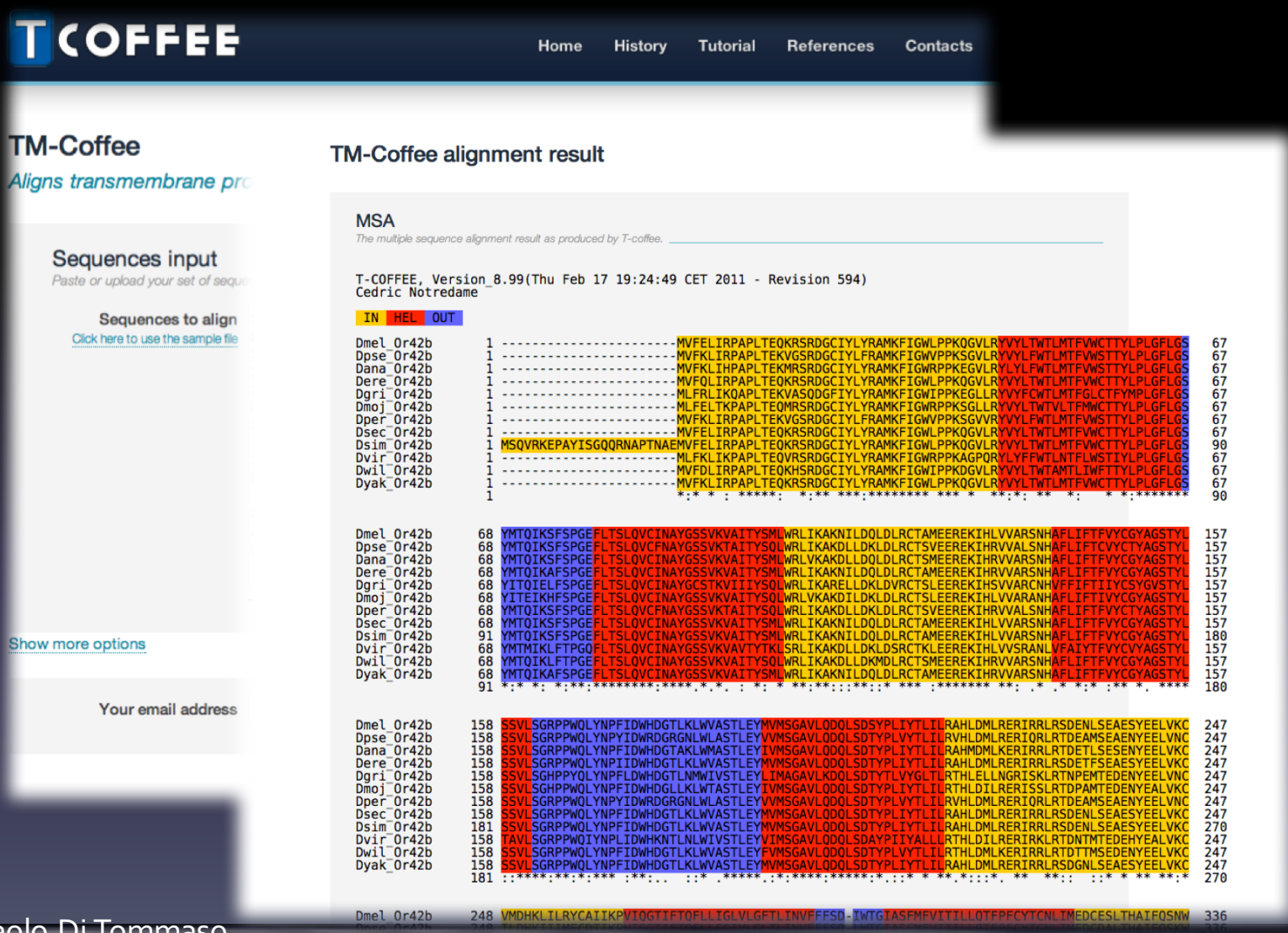
10% more columns are correctly aligned when compared with PRALINE™ .

family	Kalign	PROMALS	MAFFT	ProbCons	PRALINE™	PSI-Coffee
SP						
7TM	0.938	0.985	0.962	0.978	0.983	**0.986**
Nat	0.765	**0.815**	0.797	0.777	0.732	0.779
ACR	0.969	0.964	**0.994**	0.989	0.987	0.992
DTD	0.961	0.965	0.975	0.967	0.960	**0.977**
ION	0.810	0.761	0.788	**0.862**	0.837	0.783
MSL	0.936	**1.000**	0.980	0.958	0.986	0.971
PHOTO	0.928	0.954	0.949	0.957	**0.965**	0.955
PTGA	0.826	0.863	0.886	0.903	0.808	**0.926**
AVG	0.892	0.913	0.916	**0.924**	0.907	0.921
Pairs	3,014,033	3,109,227	3,093,269	3,108,377	3,080,356	**3,124,007**
TC						
7TM	0.360	**0.690**	0.440	0.550	0.560	0.620
Nat	0.190	0.100	0.110	0.180	0.180	**0.250**
ACR	0.620	0.530	**0.890**	0.830	0.810	0.880
DTD	0.580	0.400	0.540	0.520	0.580	**0.620**
ION	0.130	0.260	0.260	**0.320**	0.000	0.210
MSL	0.850	**1.000**	0.950	0.900	0.960	0.930
PHOTO	0.440	0.510	0.510	0.540	**0.690**	0.520
PTGA	0.320	0.280	0.370	**0.400**	0.270	**0.400**
AVG	0.436	0.471	0.509	0.530	0.506	**0.554**
Cols	863	814	1,057	1,054	1,058	**1,146**

The rows, *Pairs* and *Cols*, denote the sum of corrected aligned pairs and columns, respectively. The number of pairs and columns in the reference alignments are 3,294,102 and 1,781, respectively.

Adapted from Cedric Notredame

http://tcoffee.crg.cat/tmcoffee



Paolo Di Tommaso

Adapted from Cedric Notredame

What is The Best Method

Upcoming challenges for multiple sequence alignment methods in the high-throughput era.

Kemena C, **Notredame** C.

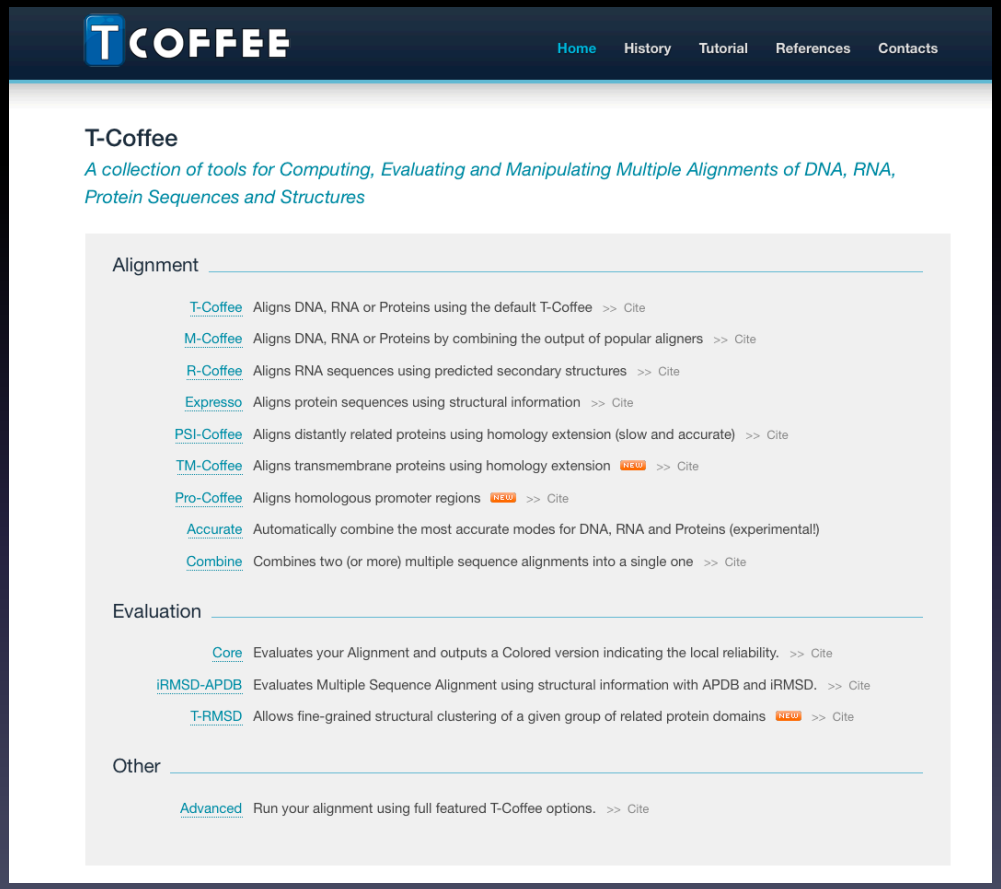
Bioinformatics. 2009 Oct 1:25(19):2455-65. Epub 2009 Jul 30. Review.

Method	Method	Template	Score	Comment
ClustalW-2	Progressive	NO	**22.74**	
PRANK	Gap	NO	**26.18**	Science2008
MAFFT	Iterative	NO	**26.18**	
Muscle	Iterative	NO	**31.37**	
ProbCons	Consistency	NO	**40.80**	
ProbCons	MonoPhasic	NO	**37.53**	
T-Coffee	Consistency	NO	**42.30**	
M-Coffe4	Consistency	NO	**43.60**	
PSI-Coffee	**Consistency**	**Profile**	**53.71**	
PROMAL	**Consistency**	**Profile**	**55.08**	
PROMAL-3D	**Consistency**	**PDB**	**57.60**	
3D-Coffee	**Consistency**	**PDB**	**61.00**	**Expresso**

Score: fraction of correct columns when compared with a structure based reference (BB11 of BaliBase).

Adapted from Cedric Notredame

www.tcoffee.org



Adapted from Cedric Notredame

A better Question…

- What is the Best Alignment ?
- What is the best bit of my alignment ?

Situation ⇔ Solution

	MUSCLE	MAFFT	PROBCONS	T-COFFEE	CLUSTALW
Accuracy	++	+++	+++	+++	+
<100 Seq.	++	++	+++	+++	+
>100 Seq.	+++	+++	-	+	+
Remote Homologues	++	+++	+++	+++	+
MSA vs Seq.	-	-		+++	+++
MSA vs MSA	-	-	-	+++	+++
>2 MSAs	-	-	-	+++	-
Seq. vs Struc.	-	-	-	+++	+
Splicing Var.	-	+++	-	+++	-
Reformat	-	-	-	+++	++
Phylogeny	-	-	-	+	++
Evaluation	-	-	+	+++	-
Speed	+++	+++	+	+	++

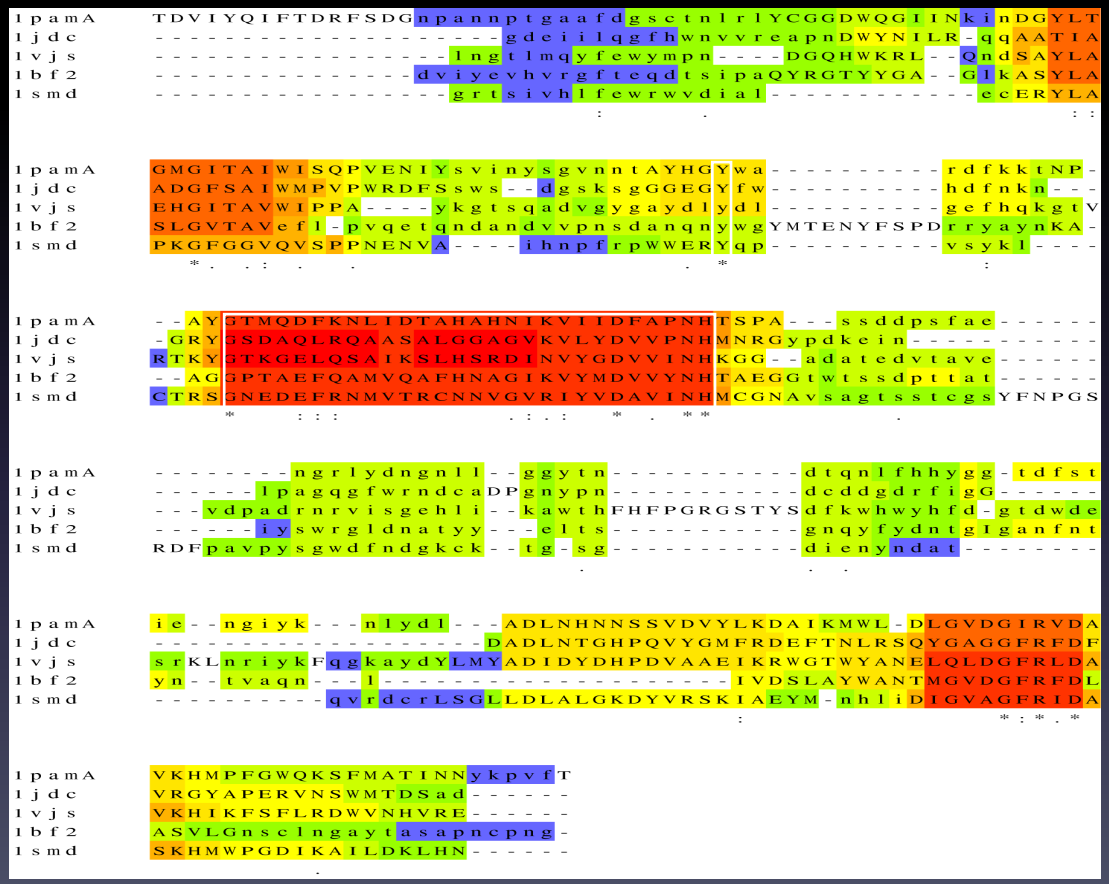
Adapted from Cedric Notredame

Purpose ⇔ Solution

	MUSCLE	MAFFT	PROBCONS	T-COFFEE	CLUSTALW
Dist Based Phylogeny	+++	+++	++	++	++
ML or MP Phylogeny	++	+++	+++	+++	++
Profile Construction	++	+++	+++	+++	++
3D Modeling	++	++	++	+++	+
Secondary Structure P	+++	+++	++	++	++

Adapted from Cedric Notredame

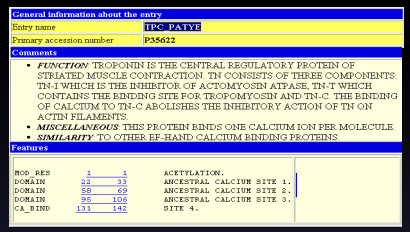
What is the Local Quality of my Alignment ?



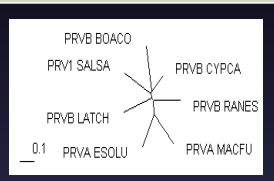
Multiple Alignment

-The BEST alignment Method:
Your Brain
The Right Data

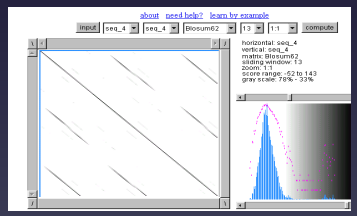
-The Best Evaluation Procedure:
Experimental Data (SwissProt)



-Choosing The Sequences Well is Important



-Beware of repeated elements



Know Your Problem: What do you want to do with your MSA

Problem \ Method	ClustalW	T-Coffee	dbClusta
One Seq Only			
Many Sequences			
3D Information			
Distant Sequences			

Highly Cited Articles

Some molecular-evolution-related articles are **very** highly cited

Method/Software	Year	Original Citation	# of Citations§
MEGA3	2004	Kumar et al. 2004 Brief Bioinform.;5(2):150-63. PMID: 15260895	6630
MRBAYES	2001	Huelsenbeck and Ronquist 2001 Bioinformatics;17(8):754-5 PMID: 11524383	5707
CLUSTALW **	1994	Thompson et al. 1994 Nucleic Acids Res.;22(22):4673-80 PMID: 7984417	29658
BLAST *	1990	Altschul et al. 1990 J Mol Biol.;215(3):403-10 PMID: 2231712	27660
Neighbor-Joining Algorithm	1987	Saitou and Nei 1987 Mol Biol Evol.;4(4):406-25 PMID: 3447015	20523
Non-Parametric Bootstrap in Phylogenetics	1985	Felsenstein 1985 Evolution;39(4):783-91 PMID: N/A	14566

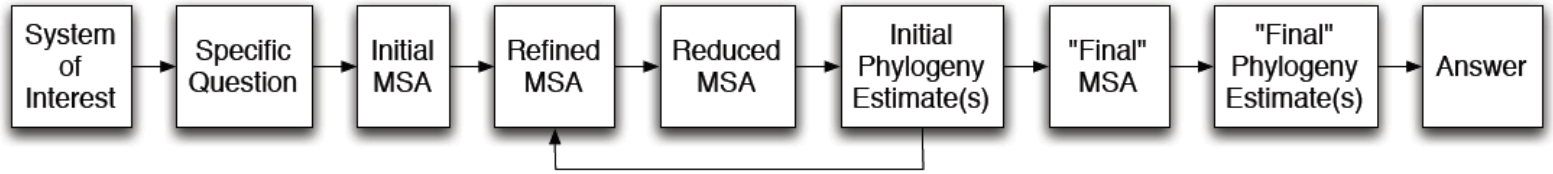
as of 2006 in the ISI web of knowledge:

§ Source: ISI Web of Knowledge, as of 29.03.2010

* most cited paper that year, 26th most cited in the entirety of science

** second most cited paper that year, 31st most cited paper in the entirety of science

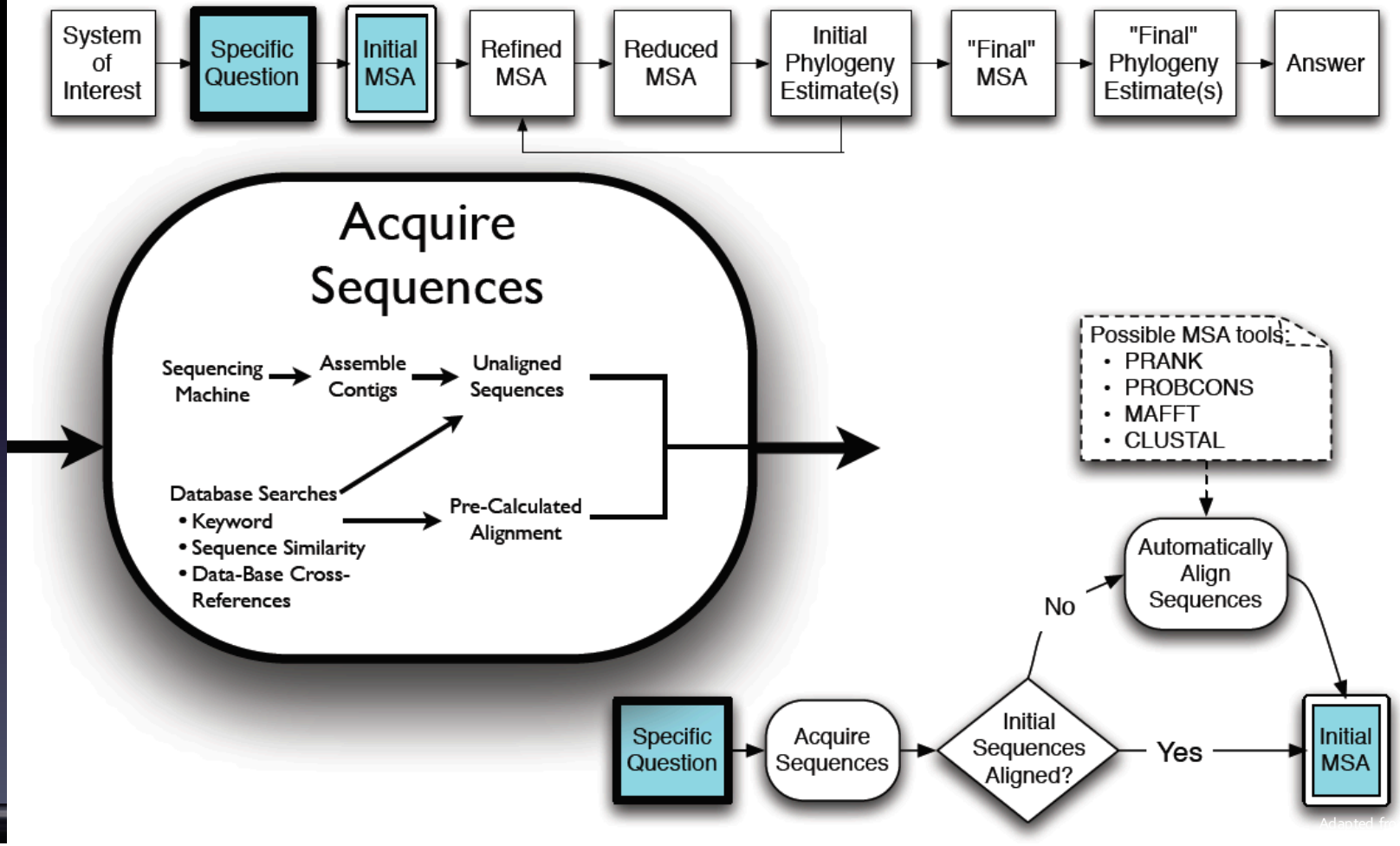
Example Phylogeny Estimation Workflow



Aims:

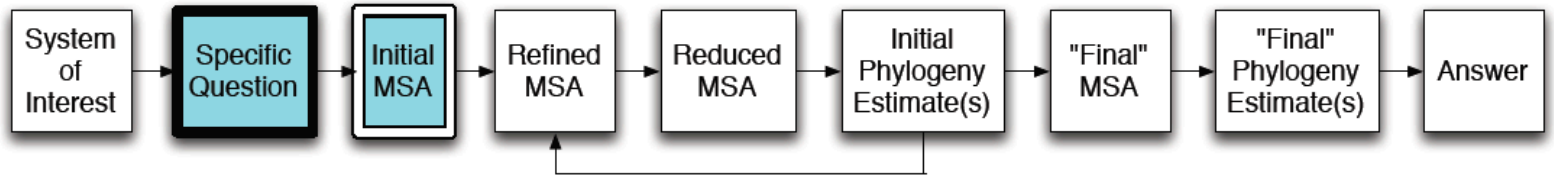
- Provide guidance/reference for planning your own analyses
- Show how different tools/stages in an analysis can link together
  - i.e. providing a context in which to place what you learn later on
- Provide a "target" to criticise
  - once you know more, what would you do differently?

Building an Initial MSA: Acquiring Sequences



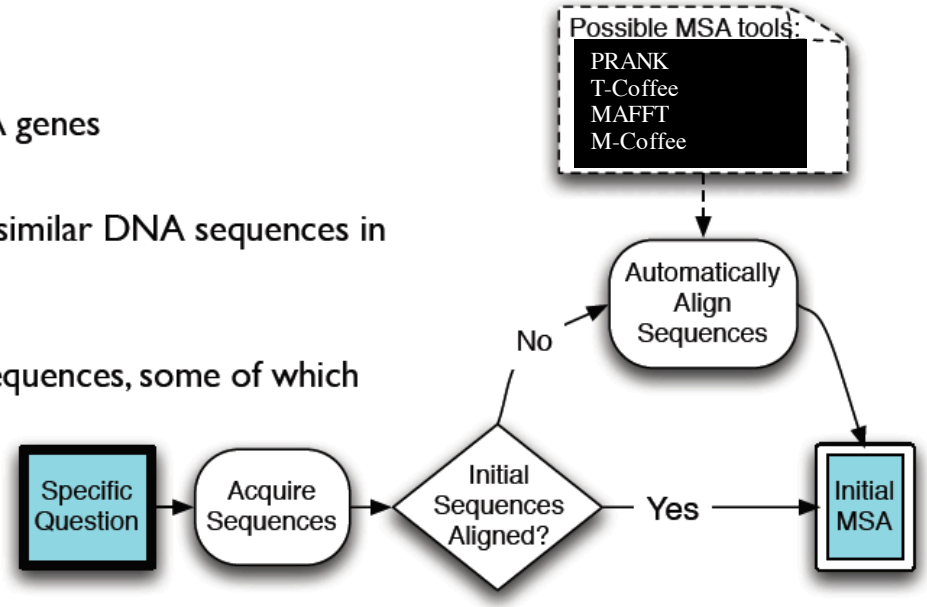
Adapted from Paiden Budd's Lecture on Phylogeny

Building an Initial MSA: Automatically Aligning Sequences

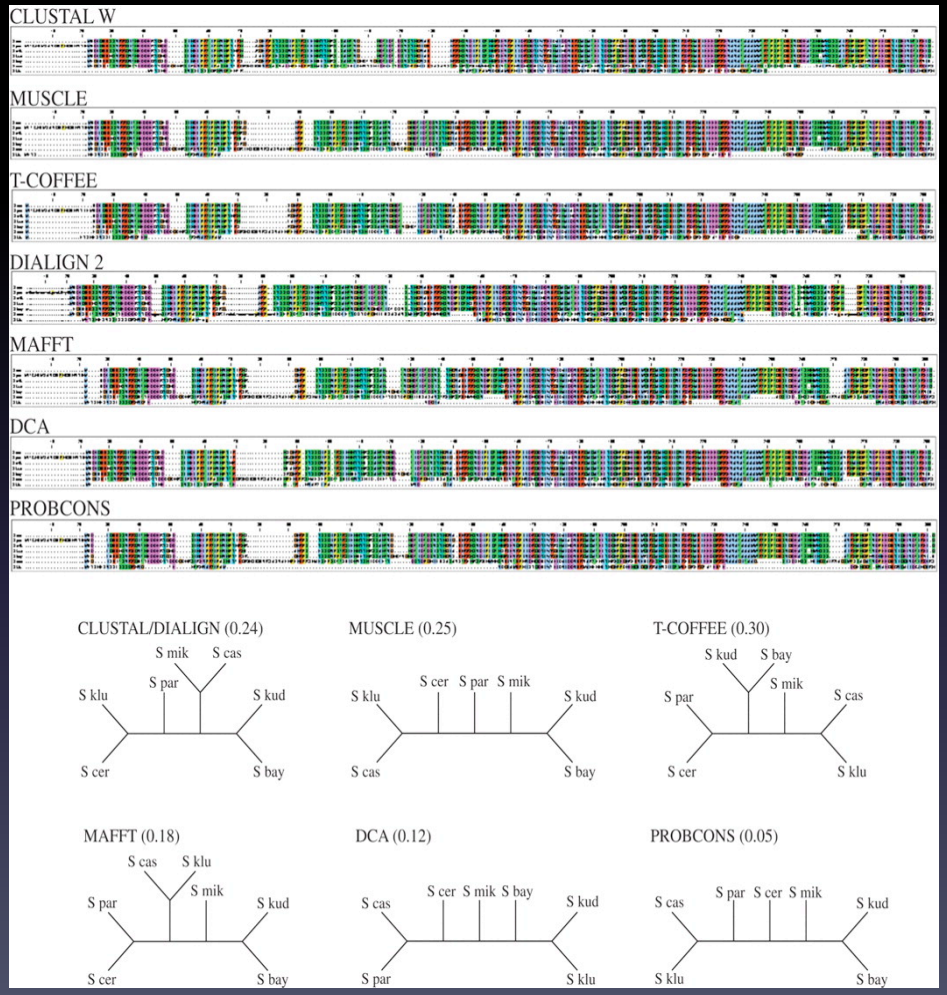


Different automatic MSA tools are designed for different tasks

- CLUSTALX, MUSCLE, PROBCONS
  divergent protein sequences
- NAST
  multiple alignment of 16S rRNA genes
- PRANK
  multiple alignment of relatively similar DNA sequences in an evolutionary context
- EXPRESSO(3DCoffee)
  multiple alignment of protein sequences, some of which have 3D structural information
- MAUVE, Enredo
  multiple alignment of genomes
- and many others...



Which MSA Method ?



Science. 2008 Jan 25;319(5862):473-6.

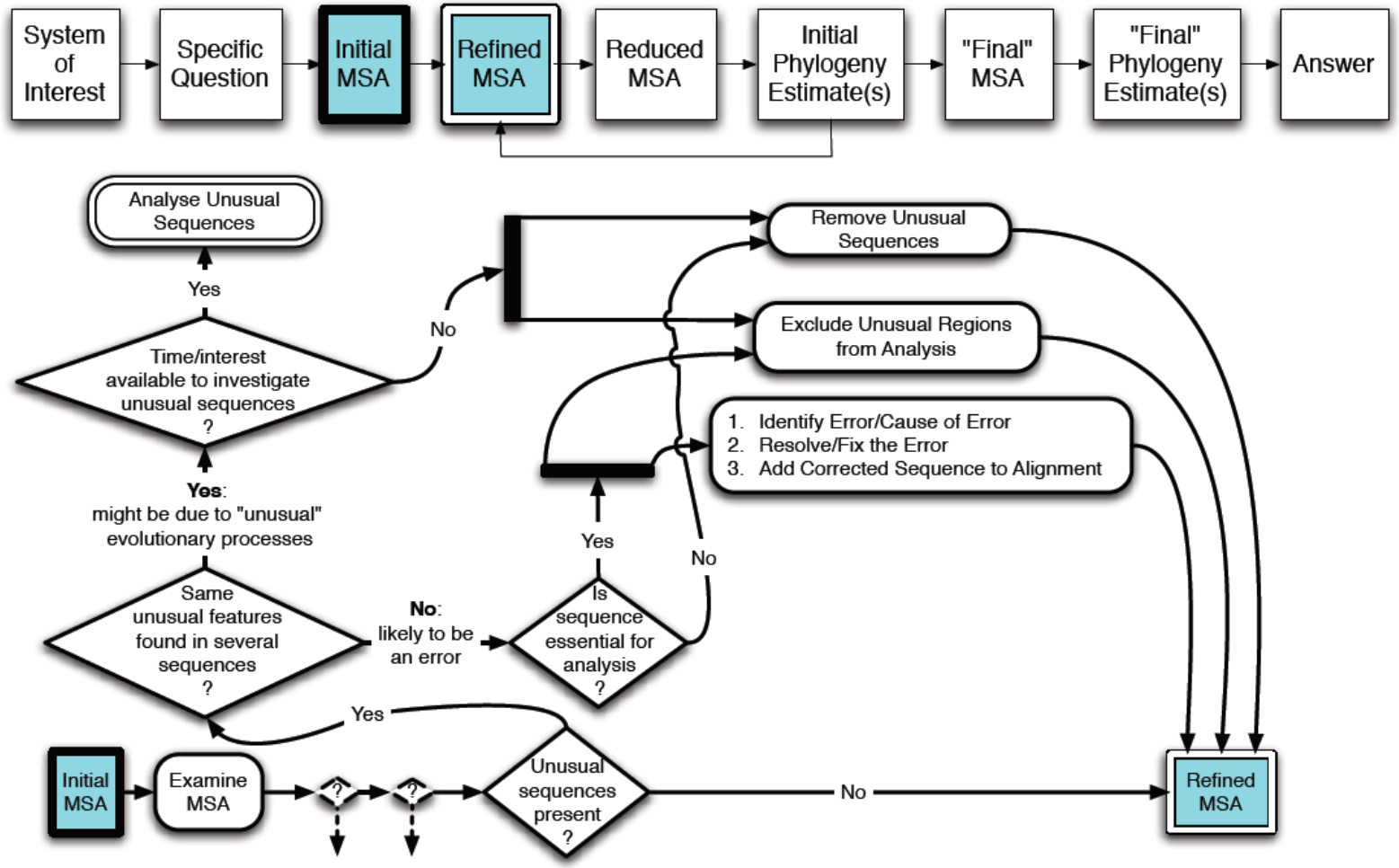
Alignment uncertainty and genomic analysis.

Wong KM, Suchard MA, Huelsenbeck JP.

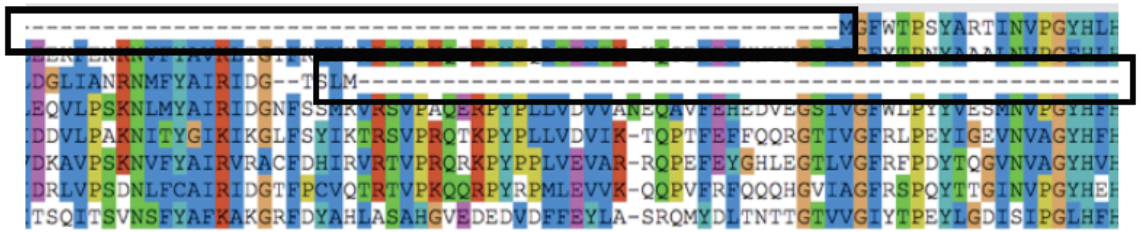
Section of Ecology, Behavior and Evolution, University of California, San Diego, La Jolla, CA 92093, USA.

Adapted from Cedric Notredame

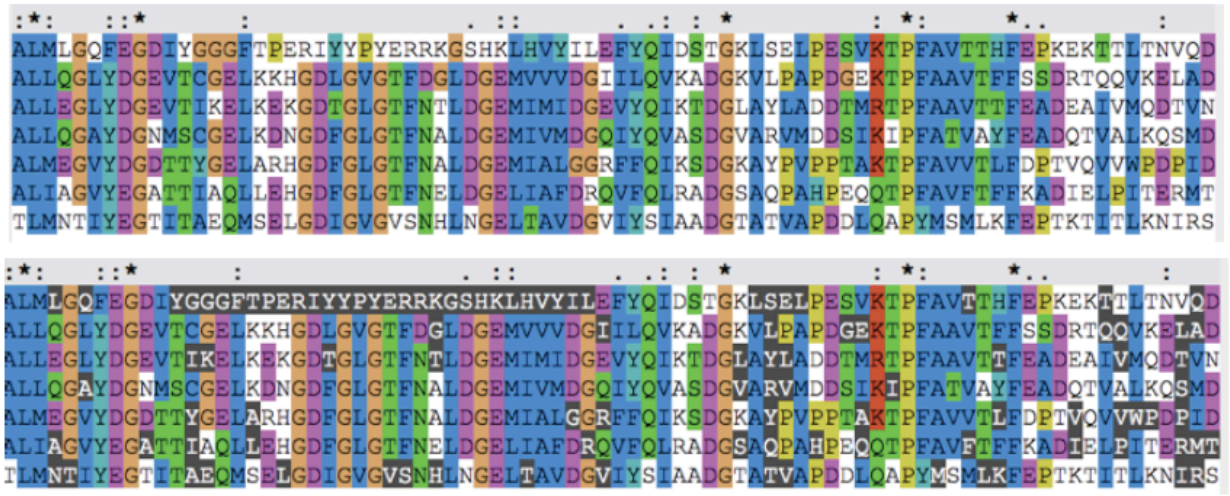
Refining Initial MSA: Unusual Sequences



Unusual Sequences: Examples



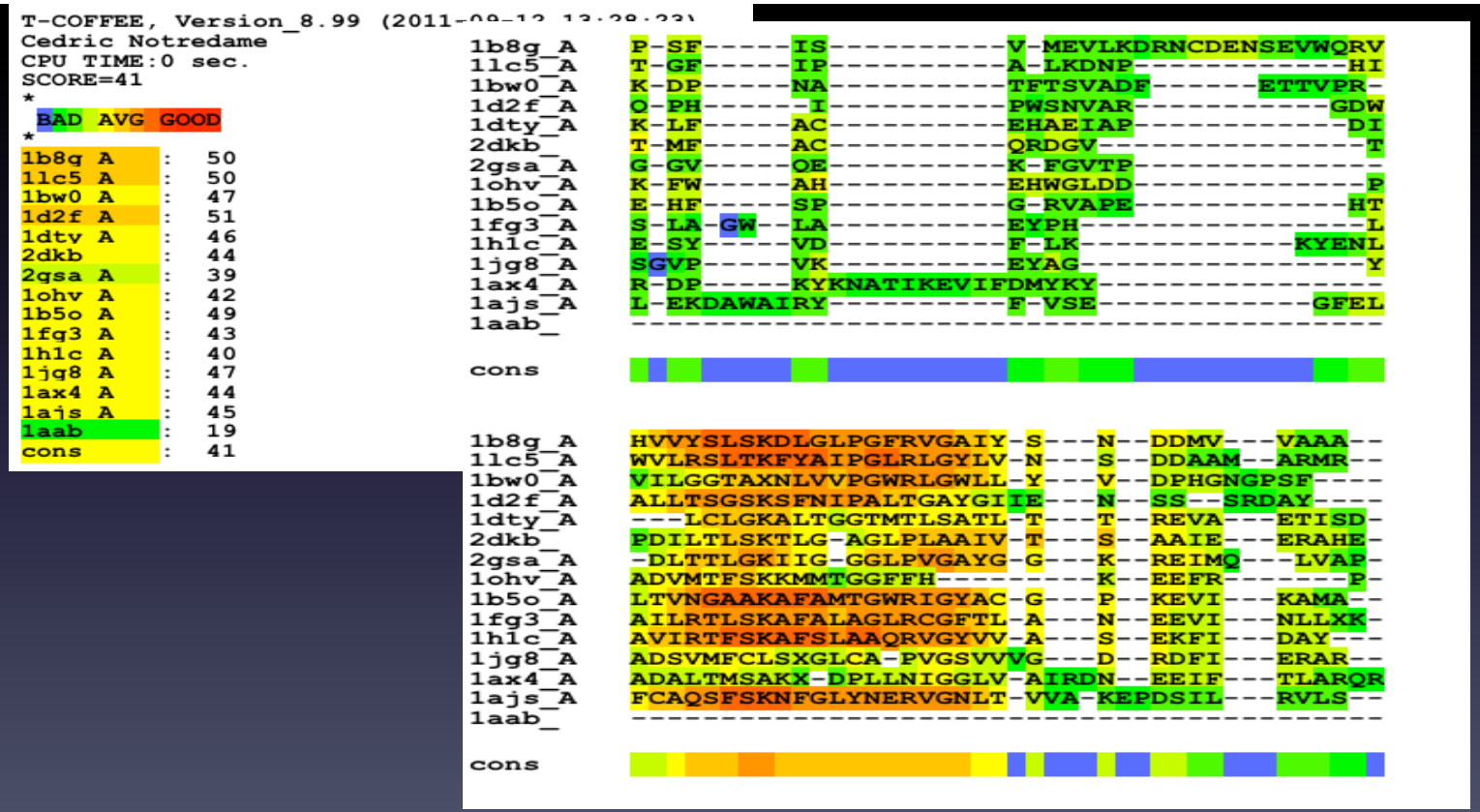
Short/fragmented sequences



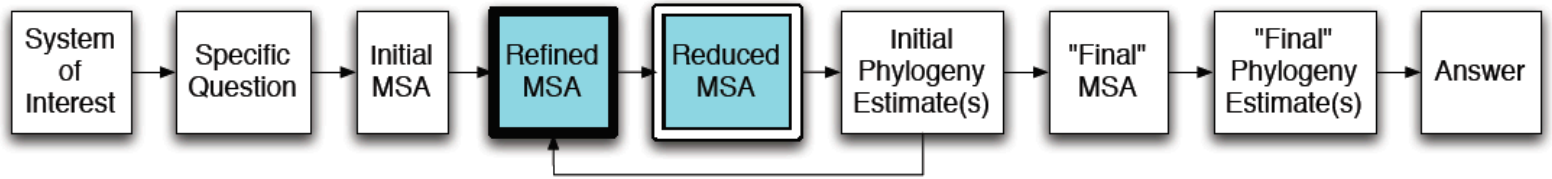
With CLUSTALX ""Quality"->"Show Low-Scorring Segments" switched on

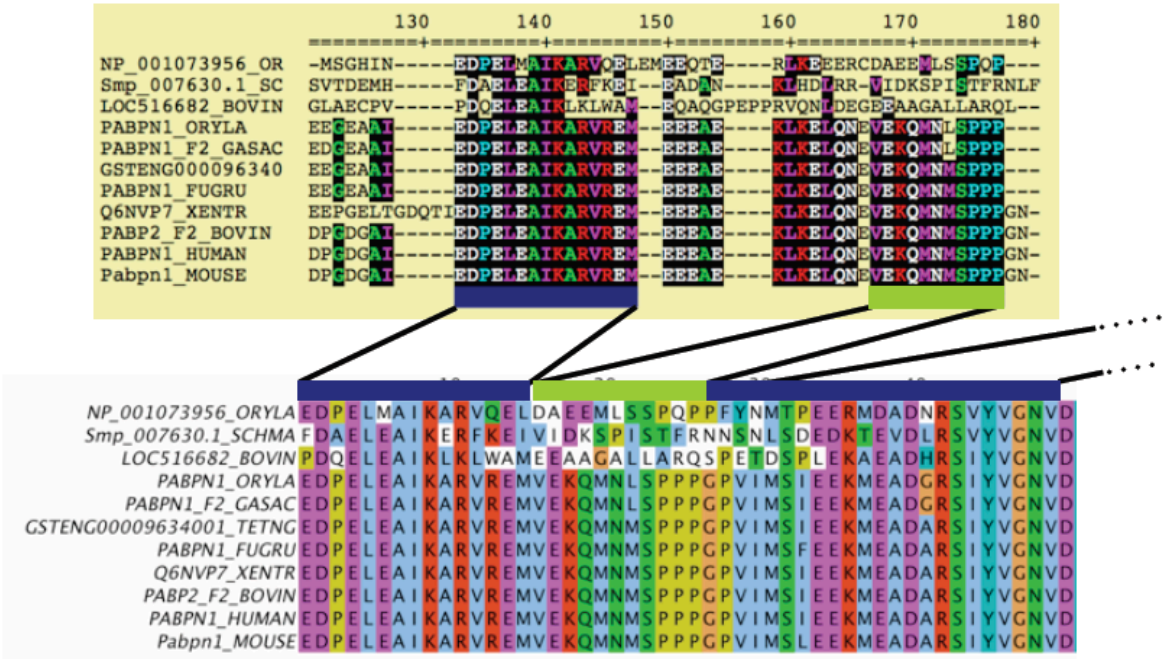
Unusual pattern of "conservation"

Unusual Sequences: Examples

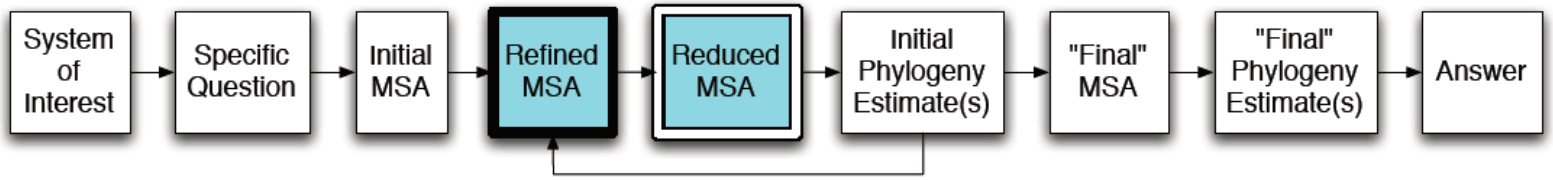


What is a "Reduced" MSA?

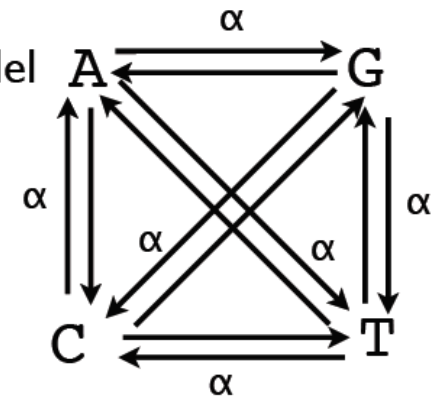




Why Use "Reduced" Alignments?



(Almost) all phylogeny estimation software ONLY model **point substitutions**



Analysing data (alignment columns) **related by any other process** introduces **systematic error** in the phylogeny estimate

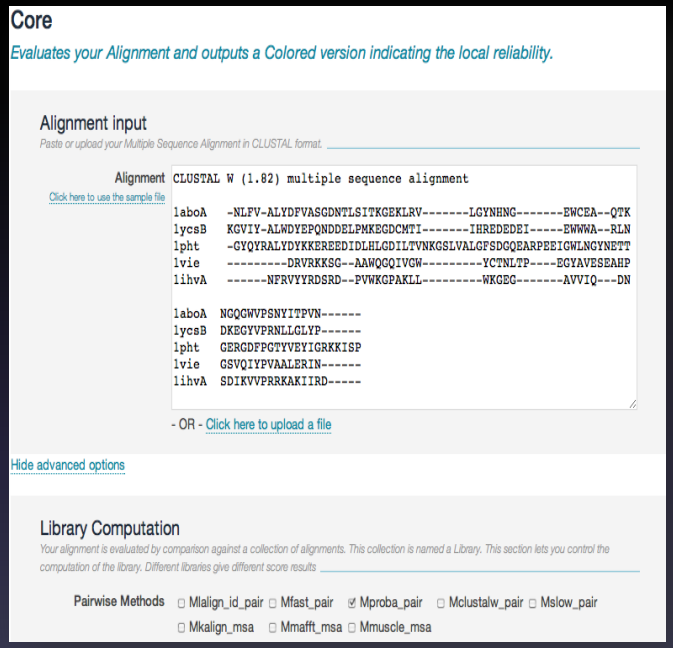
Preparing "Reduced" Alignments

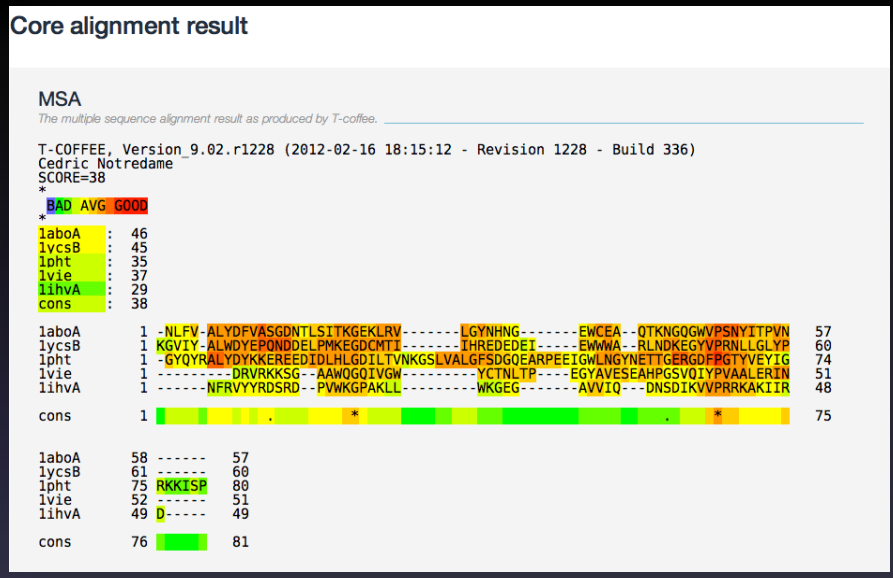
Demonstration and Exercise

Choose which columns to remove:

- "by eye" (using an alignment editor e.g. JalView)
- automatically (e.g. using GBLOCKS)
- Using trimAl
- Using seq_reformat and the T-Coffee TCS Index

TCS: A new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction





- http://www.tcoffee.org/Packages/Stable/Latest
- http://tcoffee.crg.cat/tcs

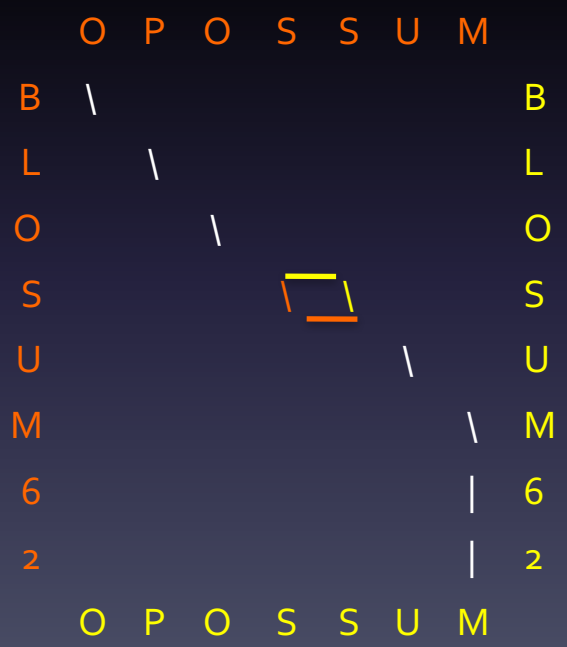
Jia-Ming Chang, Paolo Di Tommaso, and Cedric Notredame TCS: A new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction, Mol Biol Evol first published online April 1, 2014, doi:10.1093/molbev/msu117

alignment uncertainty - data

```
Aln1
OPOSSUM--
BLOS-UM62
```

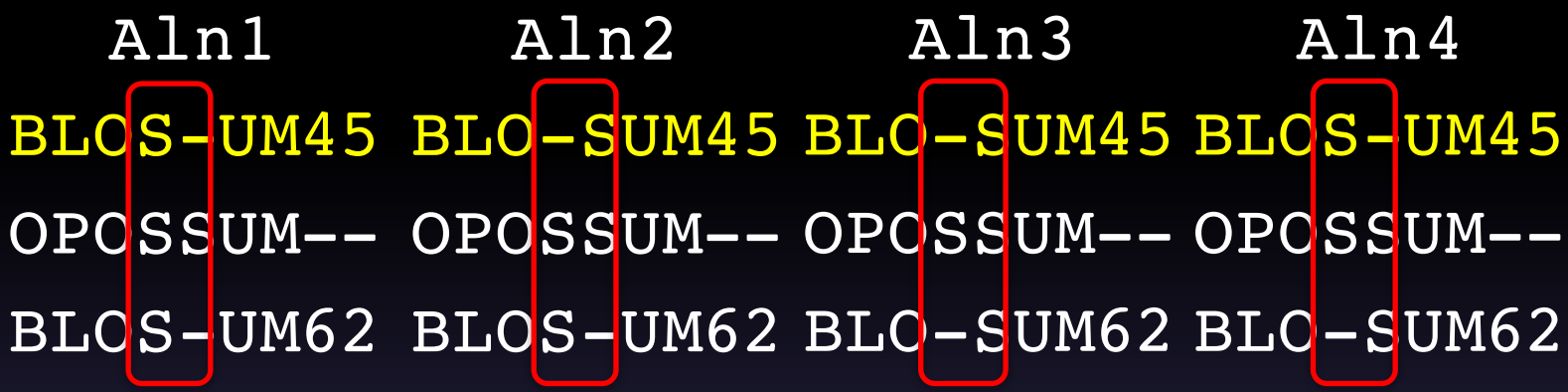
```
Aln2
OPOSSUM--
BLO-SUM62
```

If there are *two* paths
{
  chooses low-road;
}



Landan G, Graur D (2007) Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. Molecular Biology and Evolution 24: 1380–1383.

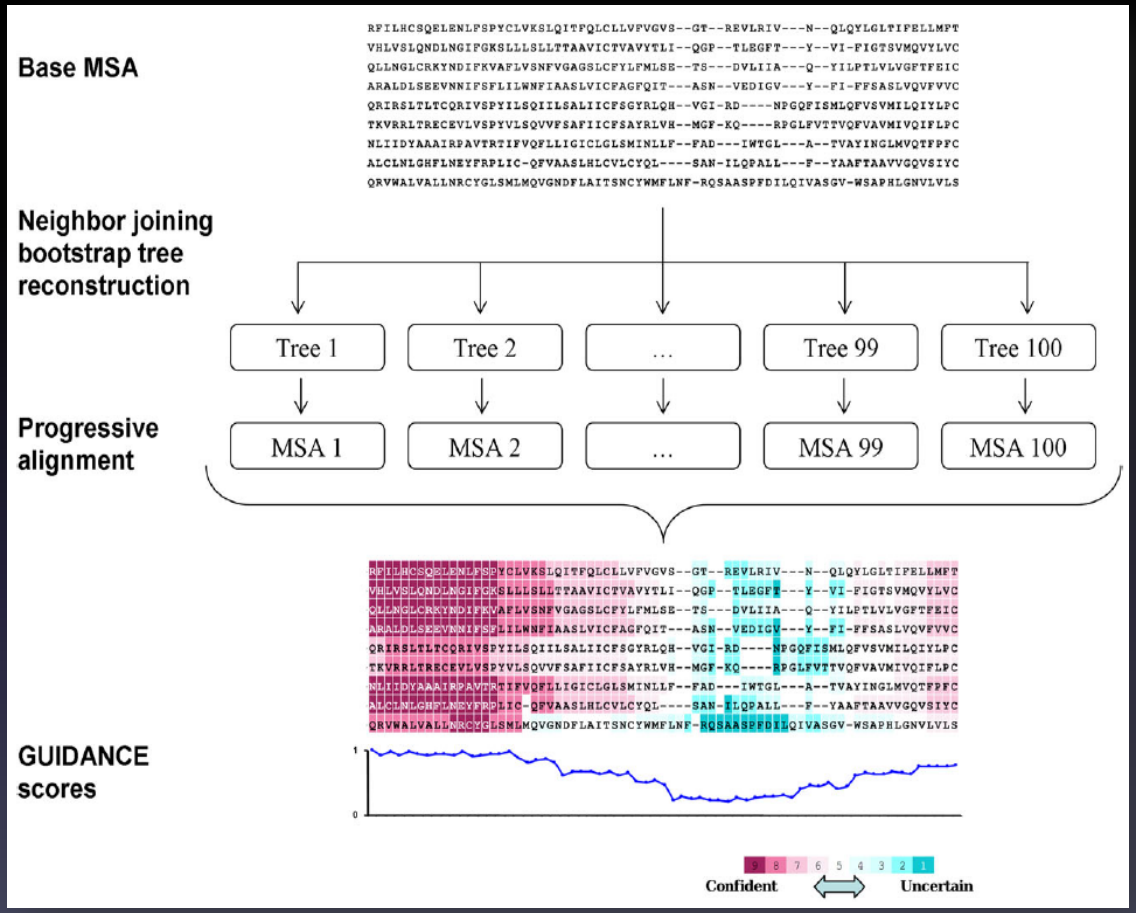
alignment uncertainty - data



It gets worse with a multiple sequence alignment.

Telling apart Uncertainty parts of the alignment is more important than the overall accuracy.

Guidance

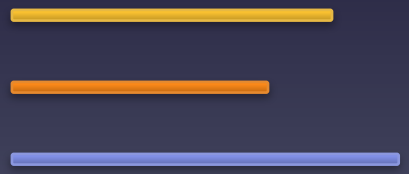


Penn O, Privman E, Landan G, Graur D, Pupko T (2010) An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol 27: 1759–1767.

Which alignment task is difficult?

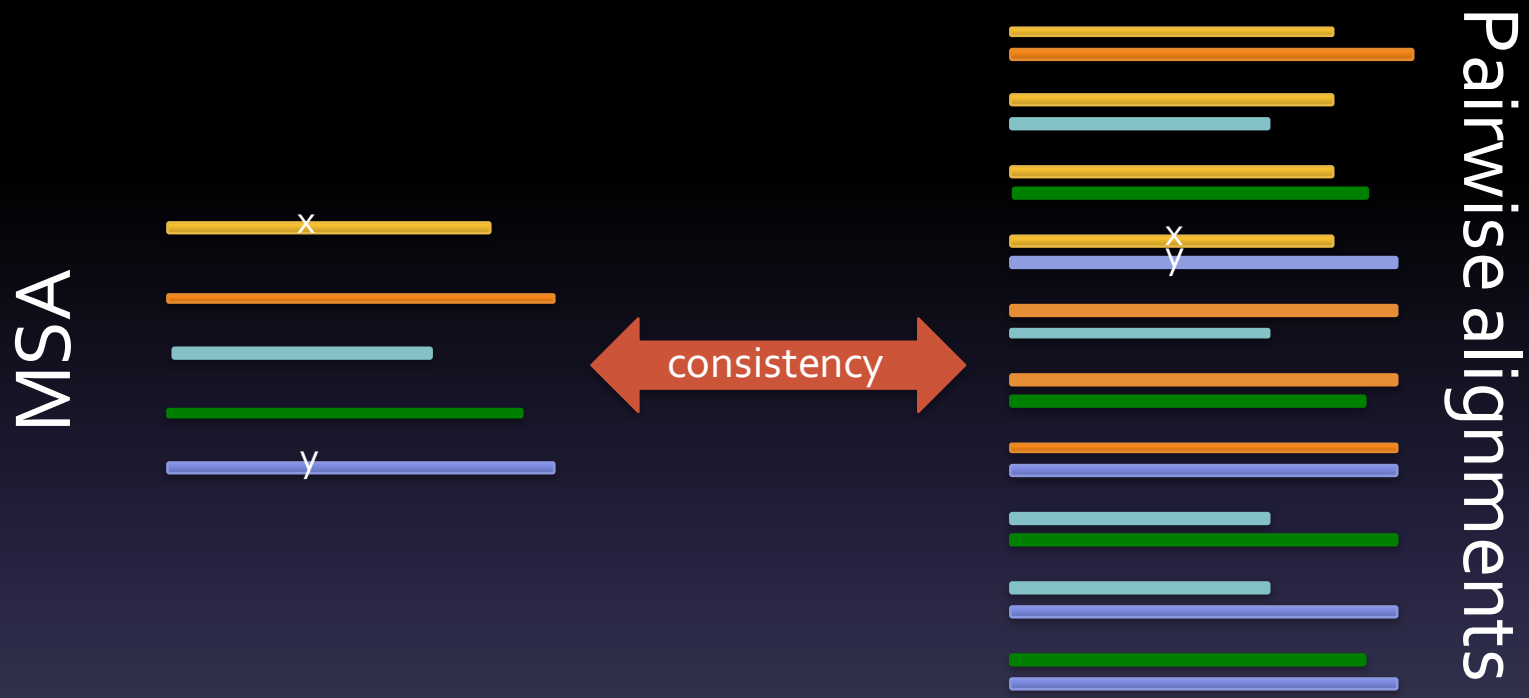
pairwise alignment $3*l^2$

multiple sequence alignment l^3



If l = 200, the second is 66 times slower than the first

Where are samples?



Consistency between MSA & pairwise alignment : **0/1**

How can we increase the resolution of confidence?

Transitive relation

In mathematics, a binary relation *R* over a set *X* is transitive if whenever an element *a* is related to an element *b*, and *b* is in turn related to an element *c*, then *a* is also related to *c*.

-WikiPedia

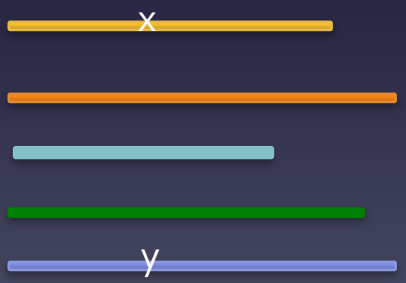
$$\forall a,b,c \in X : \left(aRb \wedge bRc\right) \Rightarrow aRc$$

Transitive relation in alignment scene

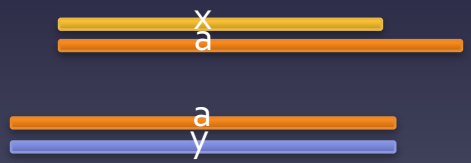
$$\forall a,b,c \in X : (aRb \wedge bRc) \Rightarrow aRc$$

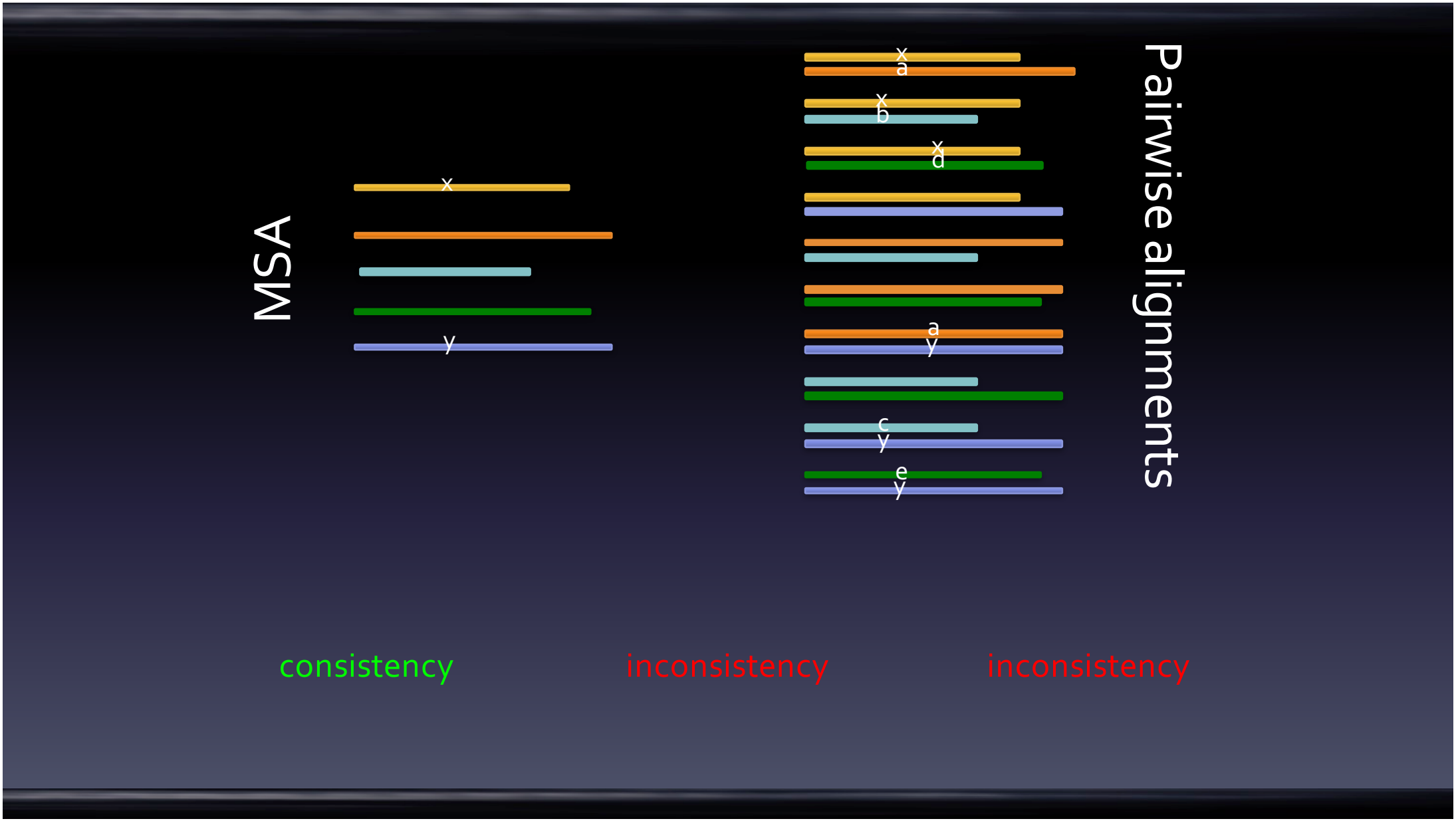
$$\forall x,y,z \in \text{alned} : (x\text{Aln}\,z \wedge z\text{Aln}\,y) \Rightarrow x\text{Aln}\,y$$

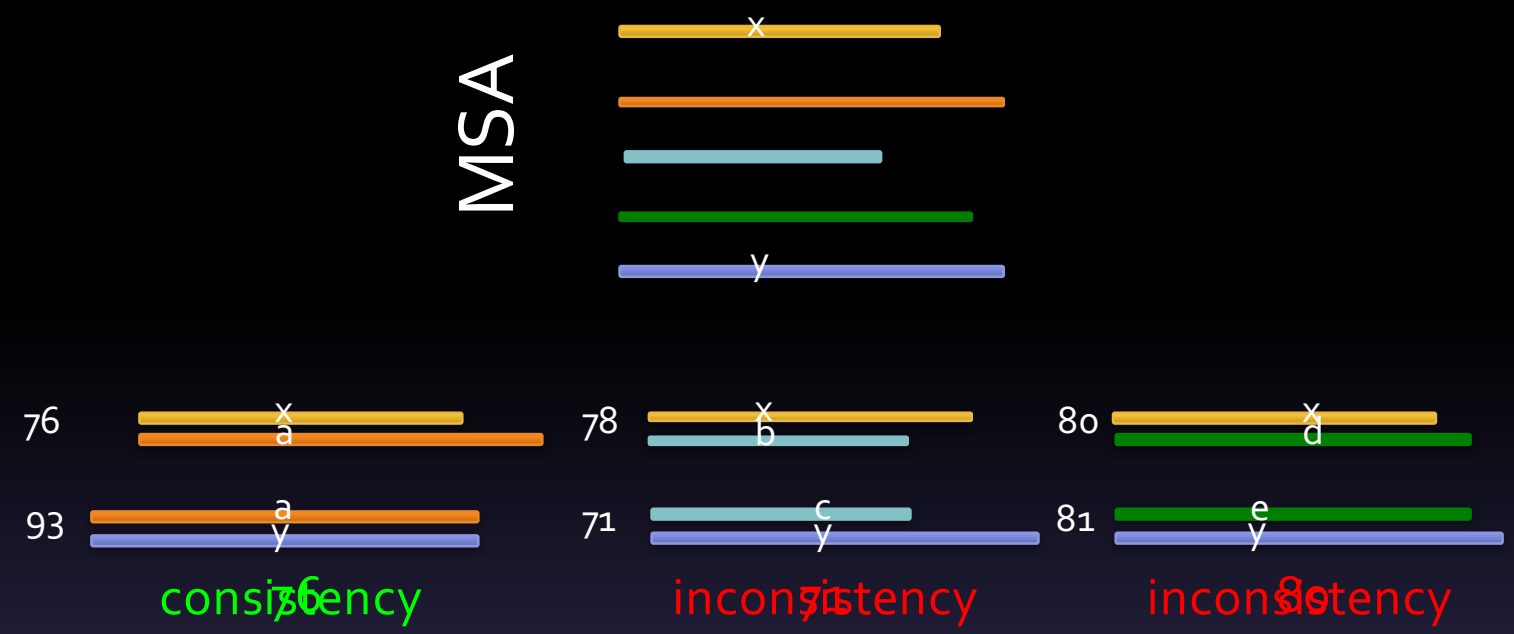
multiple sequence alignment



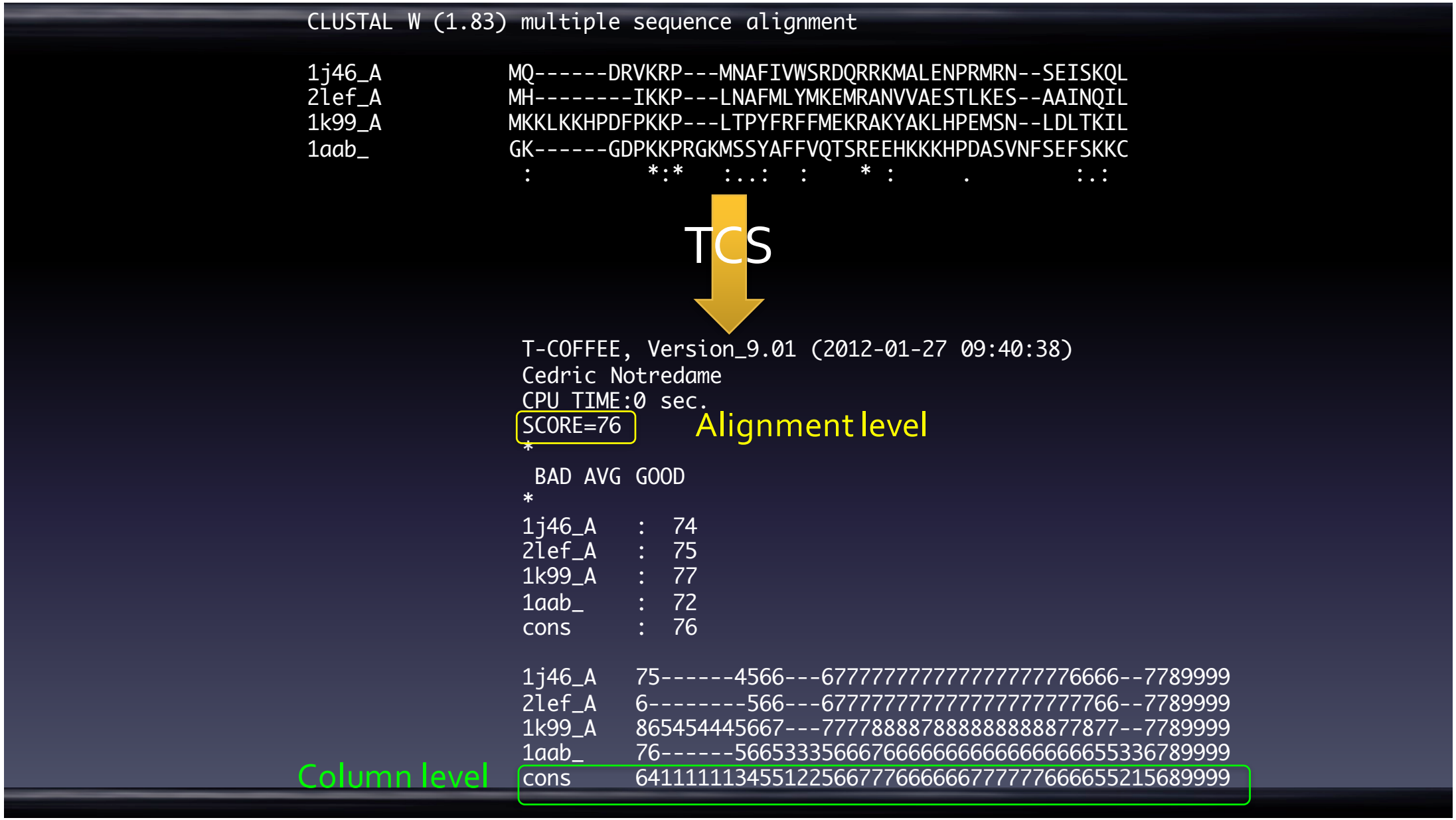
pairwise alignment




MSA
x
y
Pairwise alignments
x
a
x
b
x
d
a
y
c
y
e
y
consistency
inconsistency
inconsistency



$$TCS(x,y) = \frac{76}{76 + 71 + 80}$$

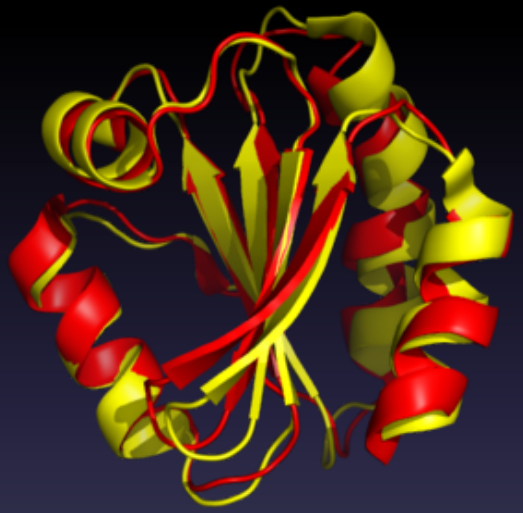

CLUSTAL W (1.83) multiple sequence alignment
1j46_A MQ------DRVKRP---MNAFIVWSRDQRRKMALENPRMRN--SEISKQL
2lef_A MH--------IKKP---LNAFMLYMKEMRANVVAEESTLKES--AAINQIL
1k99_A MKKLKKHPDFPKKP---LTPYFRFFMEKRAKYAKLHPEMSN--LDLTKIL
1aab_ GK------GDPKKPRGKMSSYAFFVQTSREEHKKKHPDASVNFSEFSKKC
TCS
T-COFFEE, Version_9.01 (2012-01-27 09:40:38)
Cedric Notredame
CPU TIME:0 sec.
SCORE=76
Alignment level
BAD AVG GOOD
1j46_A : 74
2lef_A : 75
1k99_A : 77
1aab_ : 72
cons : 76
1j46_A 75------4566---67777777777777777776666--7789999
2lef_A 6--------566---67777777777777777777766--7789999
1k99_A 865454445667---777788887888888888877877--7789999
1aab_ 76------566533356667666666666666666655336789999
cons 641111113455122566777666666777777666655215689999
Column level

```
T-COFFEE, Version_9.01 (2012-01-27 09:40:38)
Cedric Notredame
CPU TIME:0 sec.
SCORE=76
*
 BAD  AVG  GOOD
*
1j46_A    :  74
2lef_A    :  75
1k99_A    :  77
1aab_     :  72
cons      :  76
```

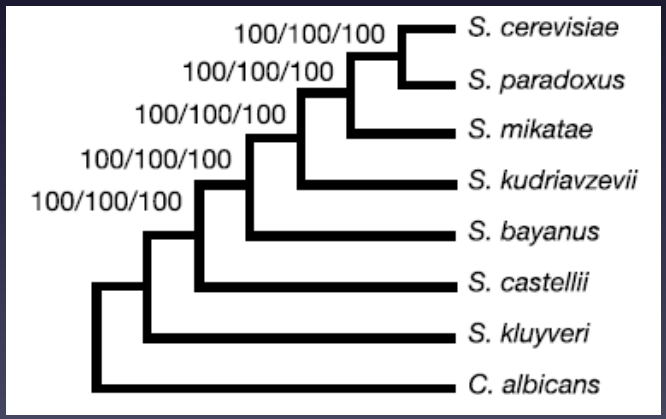
Alignment level

Column level

```
1j46_A    75------4566---67777777777777777776666--7789999
2lef_A    6--------566---67777777777777777777766--7789999
1k99_A    865454445667---77778888788888888877877--7789999
1aab_     76------566533356667666666666666666655336789999
cons      64111111345512256677766666677777766665521568999
```



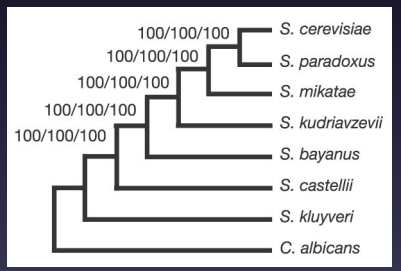
Structural modeling



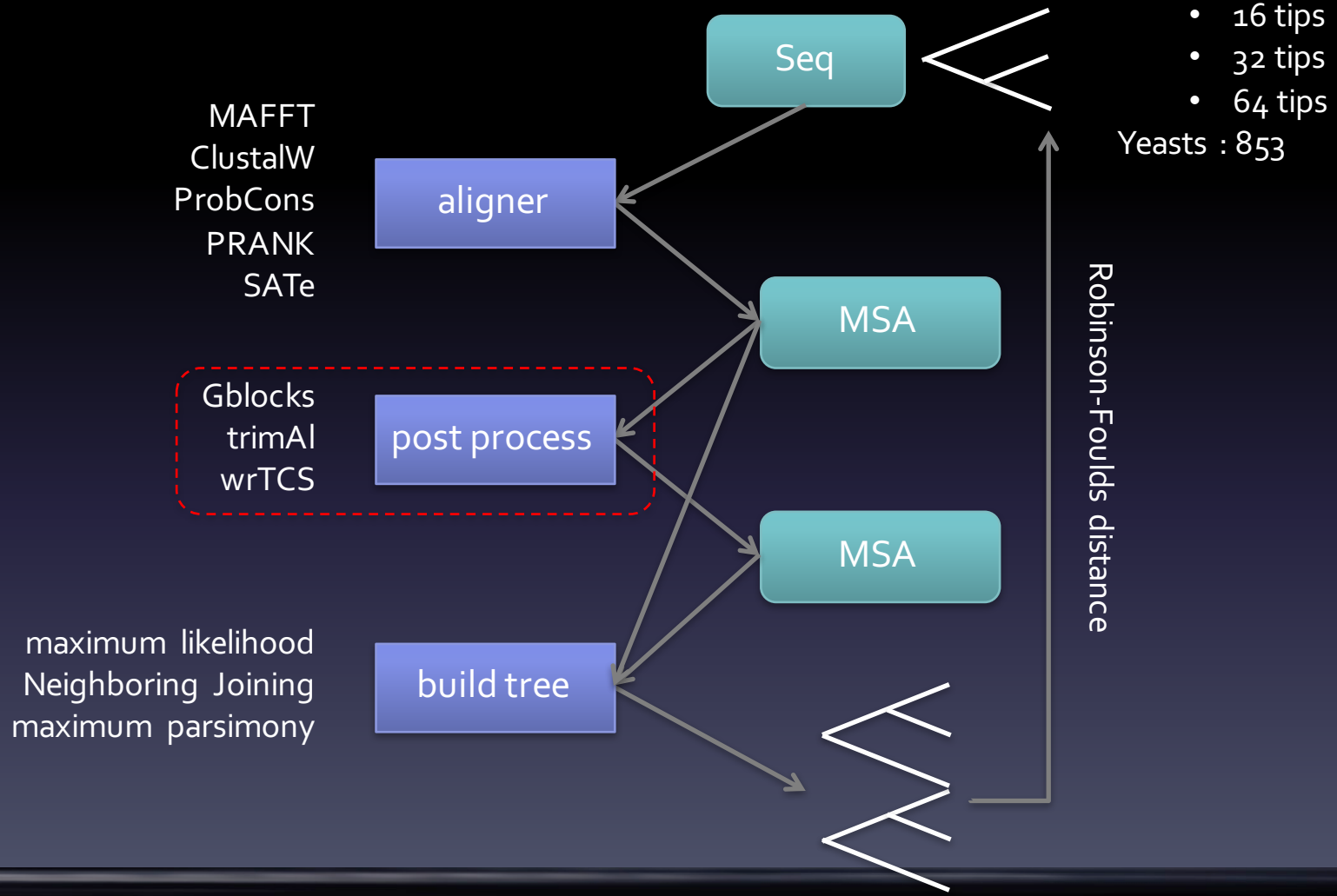
Evolutionary modeling

Q3:Does Transitive Consistency Score help phylogenetic reconstruction?

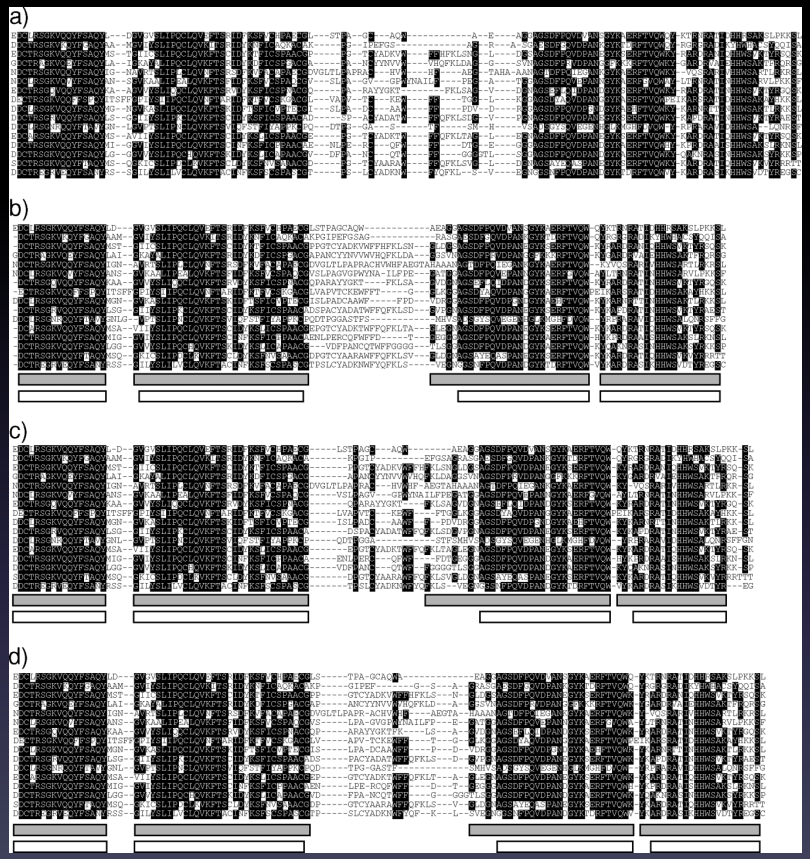
```
T-COFFEE, Version_9.03.r1318 (2012-07-12 19:05:45 - Revision 1318 - Build 366)
Cedric Notredame
SCORE=38
*
 BAD AVG GOOD
*
1aboA    :  44
1ycsB    :  42
1pht     :  40
1vie     :  31
1ihvA    :  27
cons     :  38

1aboA       1 -NLFV-ALYDFVASGDNTLSITKGEKLRV-------LGYNHNG-------EWCEA--QTKNGQGWVPSNYITPVN   57
1ycsB       1 KGVIY-ALWDYEPQNDDELPMKEGDCMTI-------IHREDEDEI-----EWWWA--RLNDKEGYVPRNLLGLYP   60
1pht        1 -GYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPEEIGWLNGYNETTGERGDFPGTYVEYIG   74
1vie        1 ----------DRVRKKSG--AAWQGQIVGW---------YCTNLTP----EGYAVESEAHPGSVQIYPVAALERIN   51
1ihvA       1 ------NFRVYYRDSRD--PVWKGPAKLL---------WKGEG-------AVVIQ---DNSDIKVVPRRKAKIIR   48

cons        1 [                .       *          *       .       *                ]   75


1aboA      58 ------   57
1ycsB      61 ------   60
1pht       75 RKKISP   80
1vie       52 ------   51
1ihvA      49 D-----   49

cons       76 [    ]   81
```



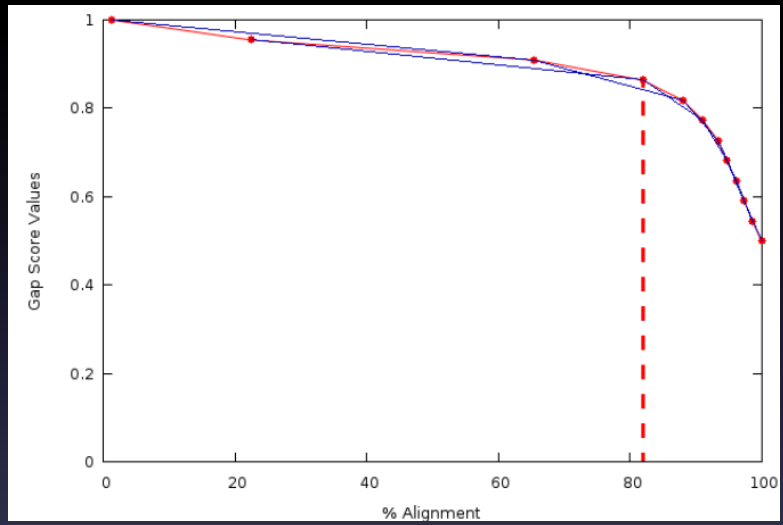
Test3 - Evolutionary Benchmark



Gblocks



trimAl



Talavera G, Castresana J (2007) Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. Syst Biol 56: 564–577.

419 citation by Google

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

104 citation by Google

Replication instead of filtering

gaps carry substantial phylogenetic signal, but are poorly exploited by most alignment and tree building programs;

Dessimoz C, Gil M: Phylogenetic assessment of alignments reveals neglected tree signal in gaps. Genome Biol 2010, 11(4):R37.

Original align.

```
1aboA   -NLFV-ALYDFVASGDNTLSITKGEKLRV-------LGYNHNG-----
1ycsB   KGVIY-ALWDYEPQNDDELPMKEGDCMTI-------IHREDEDEI---
1pht    -GYQYRALYDYKKEREEDIDLHLGDILTVNKGSLVALGFSDGQEARPE
1vie    ---------DRVRKKSG--AAWQGQIVGW---------YCTNLTP---
1ihvA   ------NFRVYYRDSRD--PVWKGPAKLL---------WKGEG-----
```

TCS scores

```
1aboA   -4445-6666667666545556665566 6-------6565544-----
1ycsB   33444-666666777755566666666 6-------655554434---
1pht    -54447766656566556666665555434446666666655445555
1vie    ---------33344444--5555555555---------5555555---
1ihvA   ------33344444444--4555554433---------33344-----
cons    133332444343443333444455433331111223332221111111
```

TCS enrich align

```
1aboA   -NNNLLL                                  ...
                                                          -
1ycsB   KGGGVVV                                  ...
                                                          -
1pht    -GGGYYY                                  ...
                                                          E
1vie    -------                                  ...
```

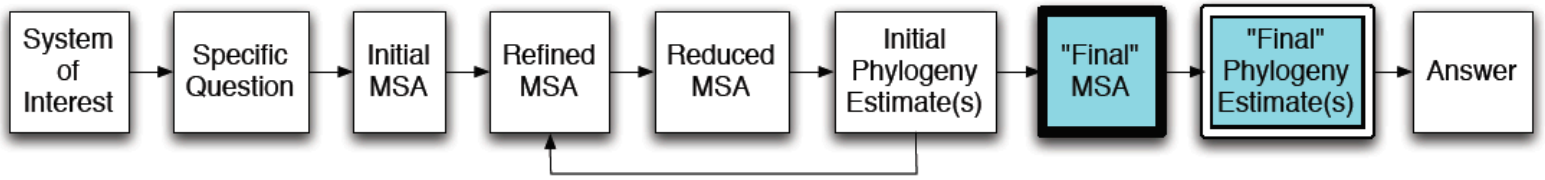
853 Yeast ToL

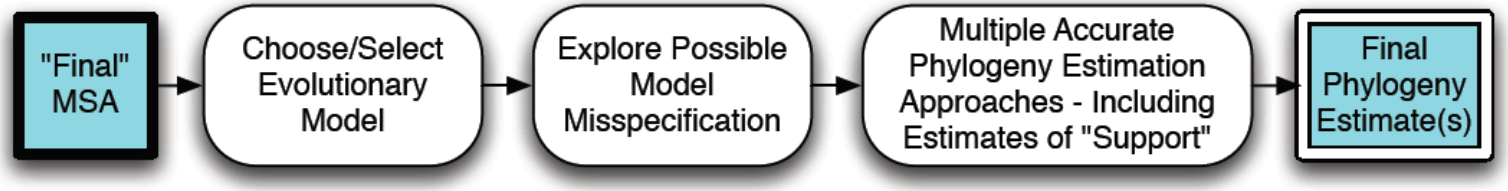
RF: average Robinson-Foulds distance respect to Yeast ToL.
TPs: the number of genes whose tree topology is identical with yeast ToL.

	Original		Gblocks relaxed		Gblocks stringent		trimAl gappyout		trimAl strictplus		TCS replicate	
	RF	TPs	RF	TPs	RF	TPs	RF	TPs	RF	TPs	RF	TPs
ClustalW	**0.90**	643	0.99	629	1.24	584	0.95	628	1.31	561	0.91	**649**
MAFFT	0.80	665	0.83	653	1.26	573	0.83	657	1.28	562	**0.76**	**669**
Muscle	0.95	639	0.91	646	1.26	578	0.96	633	1.29	559	**0.84**	**662**
PRANK	**0.79**	**665**	0.88	642	1.28	565	0.84	648	1.19	575	0.81	662
SATe	0.86	660	0.87	650	1.28	578	0.85	655	1.25	567	**0.79**	**666**
AVE	0.86	654	0.896	644	1.26	575	0.88	644	1.26	565	**0.82**	**661**

Adapted from Cedric Notredame

Final(ish) Phylogeny Estimate

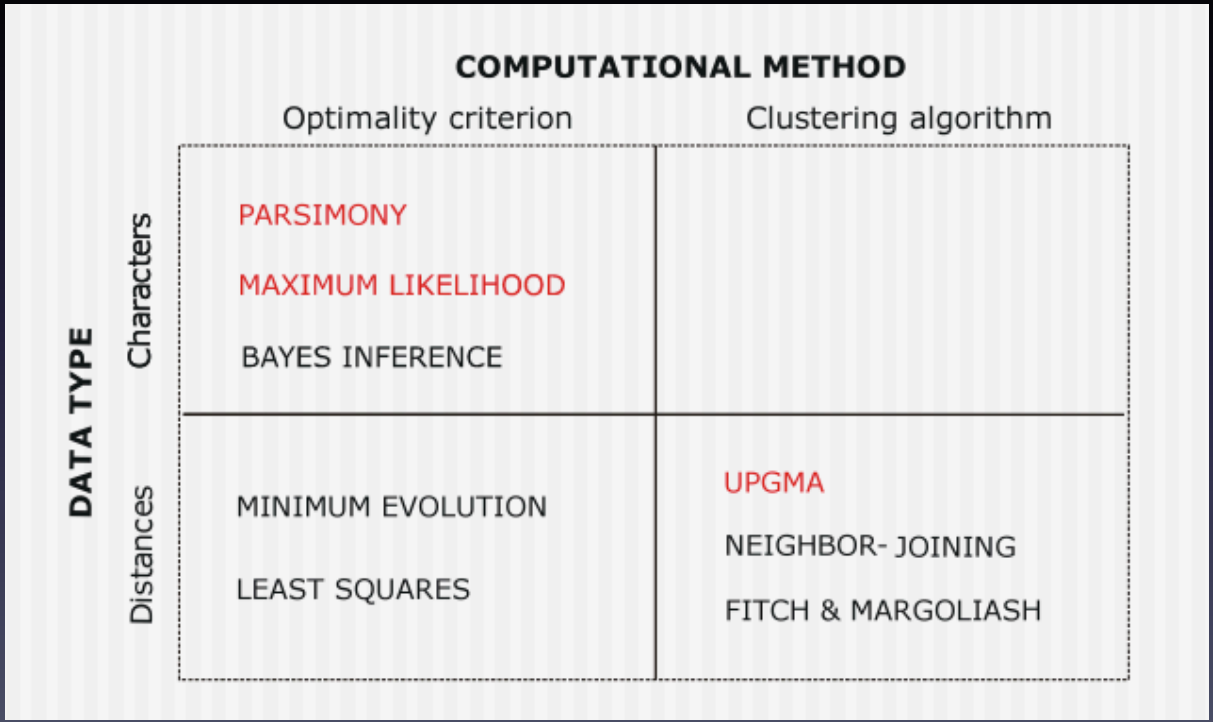




- Use several (more) accurate phylogeny estimation methods and implementations
  - Bayesian - MrBayes
  - ML - RAxML, PhyML
- Estimate using different parameter values within each implementation
  - Models
  - Specifics of tree search algorithm
  - Support values

Methods for Phylogenetic reconstruction

Main families of methods :



Adapted from Cedric Notredame

Number of possible tree topologies for *n* taxa

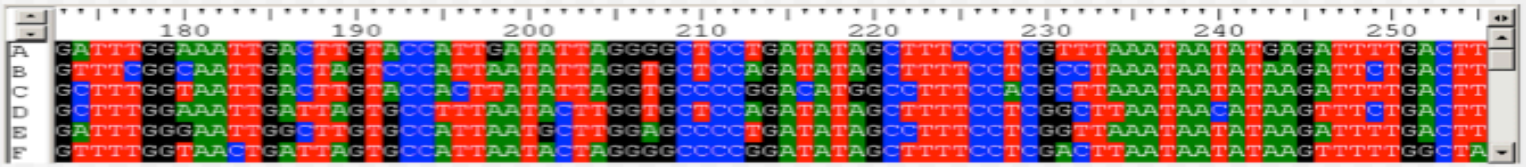
$$N_{trees} = 3.5.7...(2n-5) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

n	N_{trees}
4	3
5	15
6	105
7	945
...	...
10	2,027,025
...	...
20	~ 2×10^{20}

Adapted from Cedric Notredame

UPGMA

- Very Unreliable Method
- Very Simple Principle



	A	B	C	D	E	F	G
A	-						
B	63	-					
C	94	79	-				
D	111	96	47	-			
E	67	16	83	100	-		
F	23	58	89	106	62	-	
G	107	92	43	20	96	102	-

Adapted from Cedric Notredame

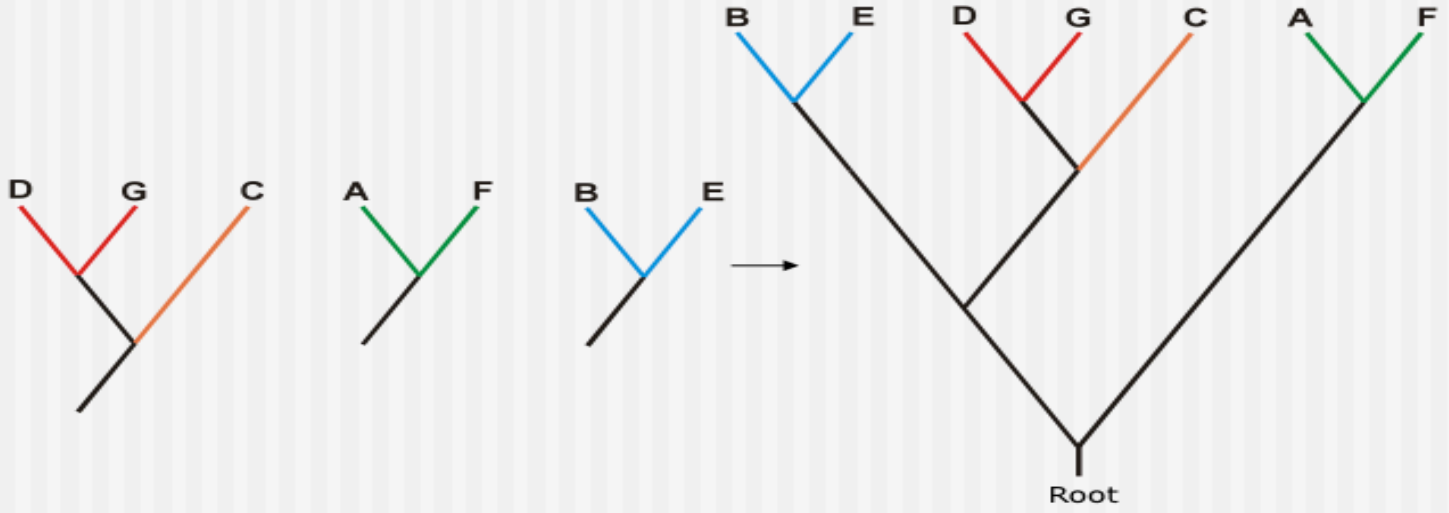
UPGMA

	A	B	C	E	F	DG
A	-					
B	63	-				
C	94	79	-			
E	67	16	83	-		
F	23	58	89	62	-	
DG	94	84	35	88	94	-

Adapted from Cedric Notredame

UPGMA

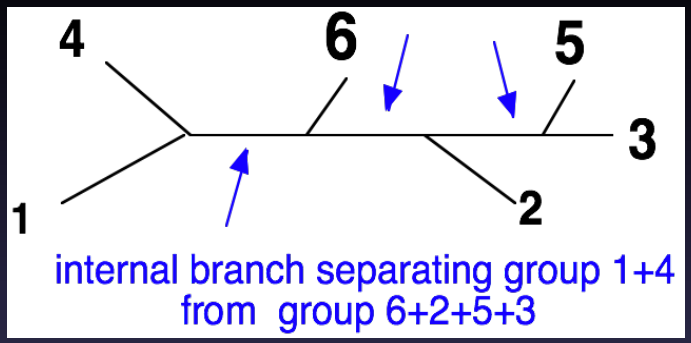
	AF	BE	CDG
AF	-		
BE	188	-	
CDG	112	108	-



Adapted from Cedric Notredame

Reliability of phylogenetic trees: the bootstrap

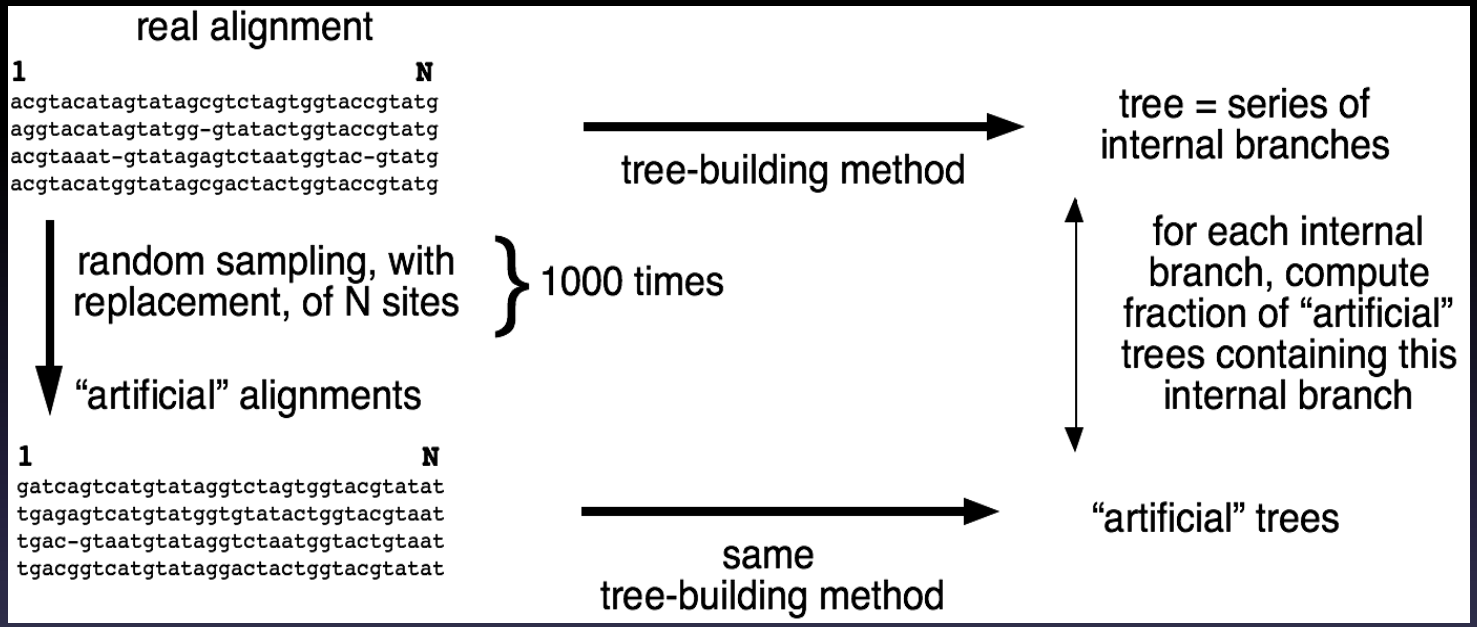
- The phylogenetic information expressed by an unrooted tree resides entirely in its internal branches.



- The tree shape can be deduced from the list of its internal branches.
- Testing the reliability of a tree = testing the reliability of each internal branch.

Adapted from Cedric Notredame

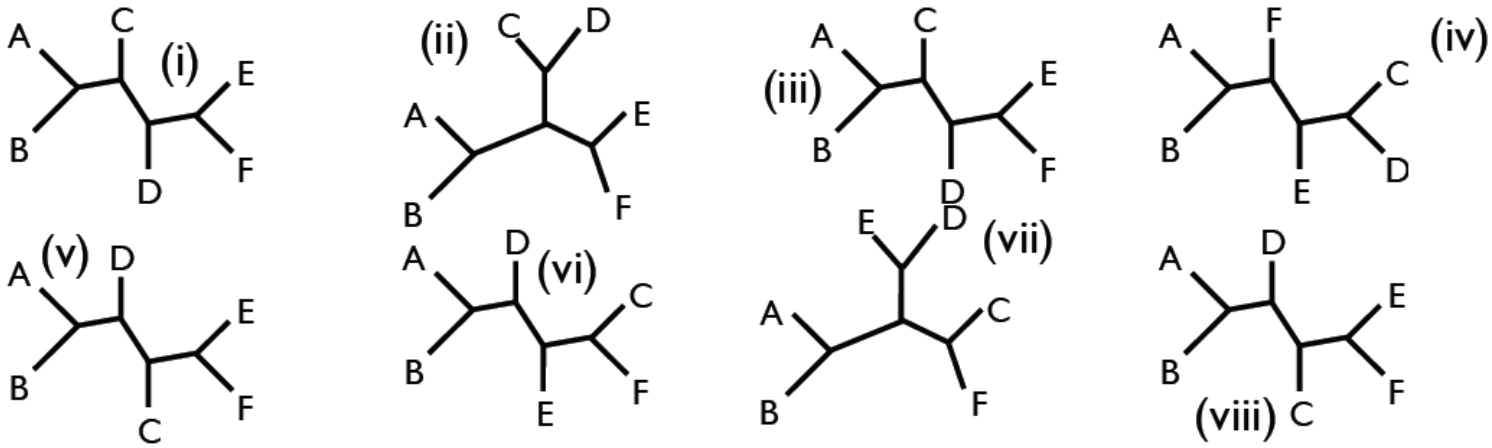
Bootstrap procedure



The support of each internal branch is expressed as percent of replicates.

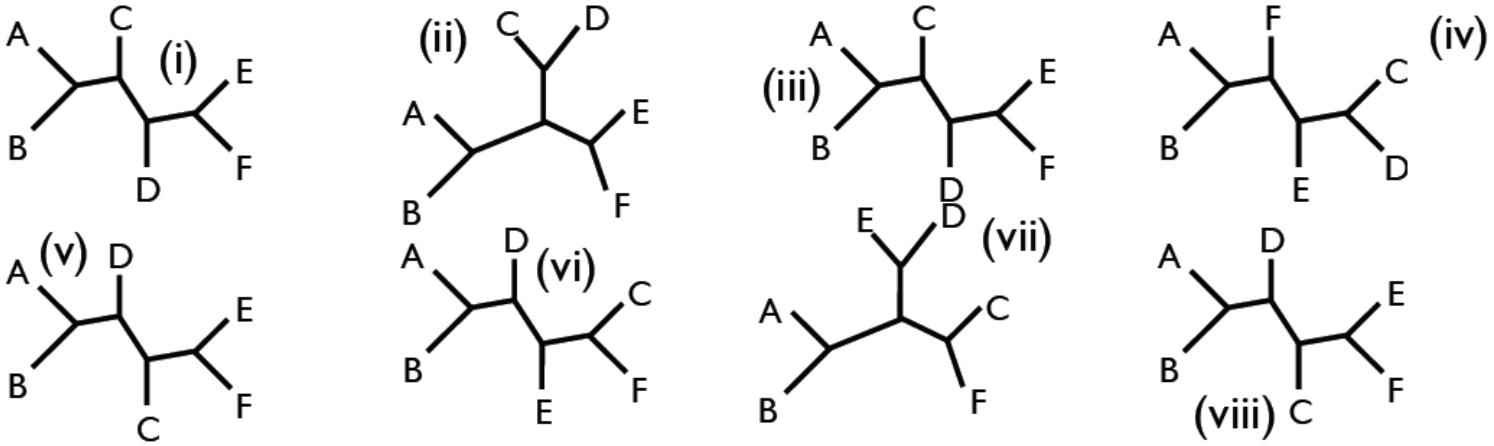
Adapted from Cedric Notredame

50%/Majority Rule Consensus Trees

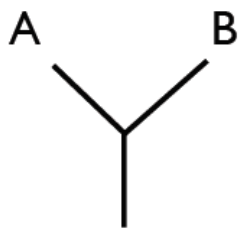


	i	ii	iii	iv	v	vi	vii	viii	
AB | CDEF	*	*	*	*	*	*	*	*	8
CD | ABEF		*		*					2
EF | ABCD	*	*	*		*			*	5
ABC | DEF	*		*						2
DE | ABCF							*		1
CF | ABED						*	*		2
ABD | ECF					*	*		*	3
ABF | CDE				*					1

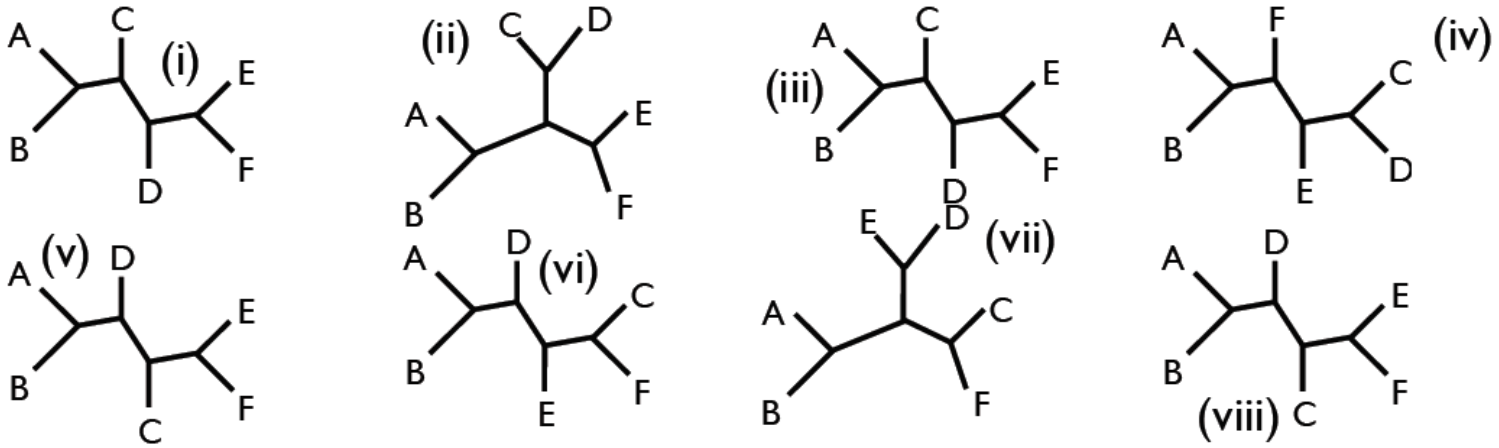
50%/Majority Rule Consensus Trees



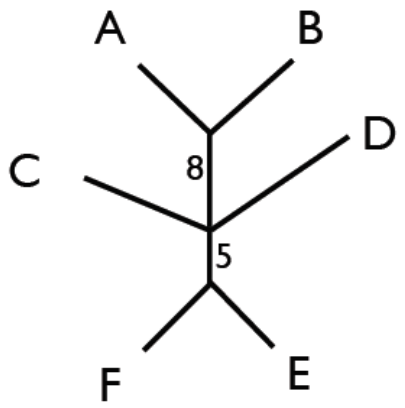
	i	ii	iii	iv	v	vi	vii	viii	
AB | CDEF	*	*	*	*	*	*	*	*	8
CD | ABEF		*		*					2
EF | ABCD	*	*	*		*			*	5
ABC | DEF	*		*						2
DE | ABCF							*		1
CF | ABED						*	*		2
ABD | ECF					*	*		*	3
ABF | CDE				*					1



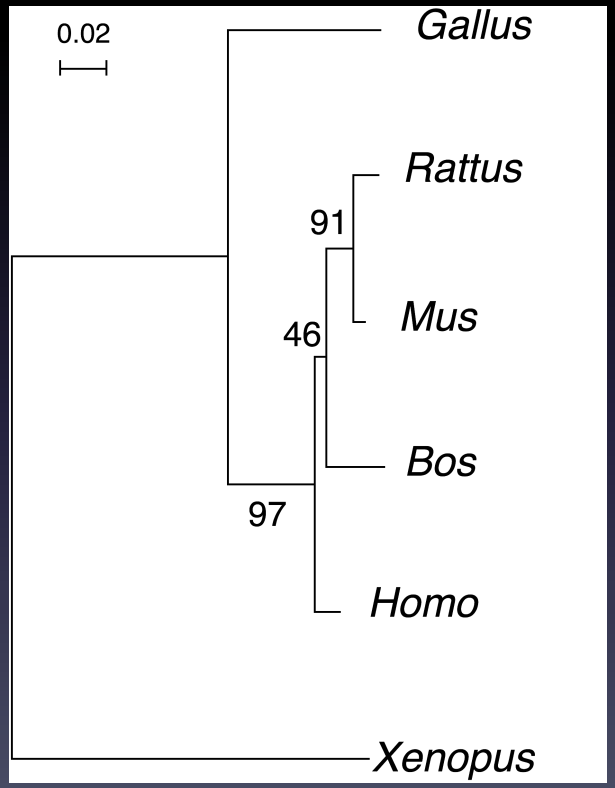
50%/Majority Rule Consensus Trees



	i	ii	iii	iv	v	vi	vii	viii	
AB | CDEF	*	*	*	*	*	*	*	*	8
CD | ABEF		*		*					2
EF | ABCD	*	*	*		*			*	5
ABC | DEF	*		*						2
DE | ABCF							*		1
CF | ABED						*	*		2
ABD | ECF					*	*		*	3
ABF | CDE				*					1



"bootstrapped" tree



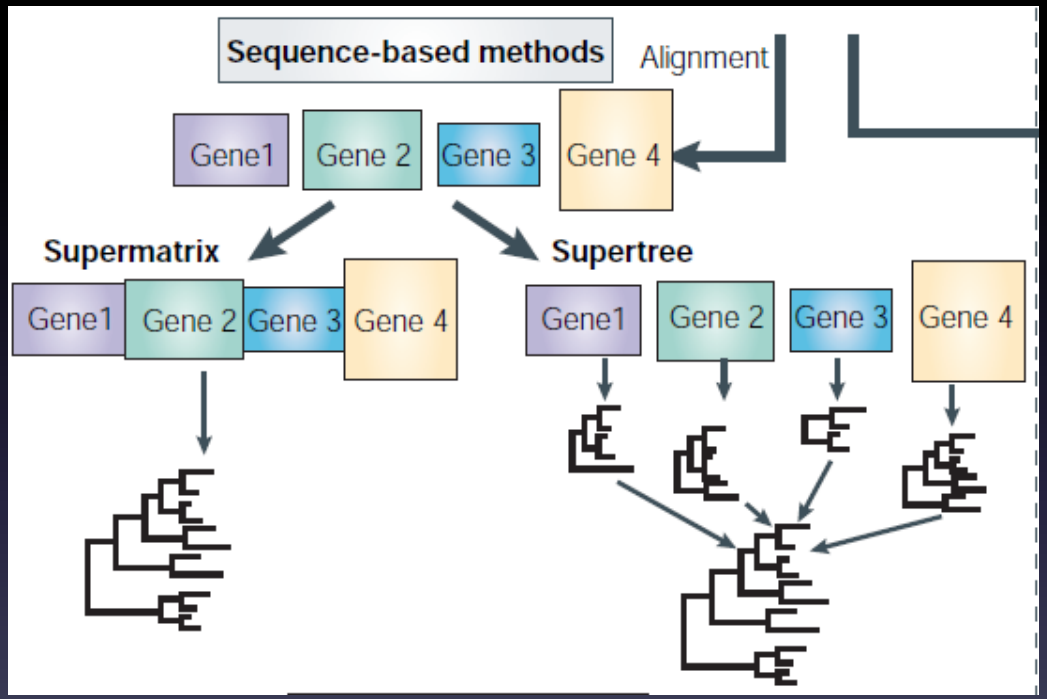
Adapted from Cedric Notredame

Bootstrap procedure : properties

- Internal branches supported by ≥ 90% of replicates are considered as statistically significant.
- The bootstrap procedure only detects if sequence length is enough to support a particular node.
- The bootstrap procedure does not help determining if the tree-building method is good. A wrong tree can have 100 % bootstrap support for all its branches!

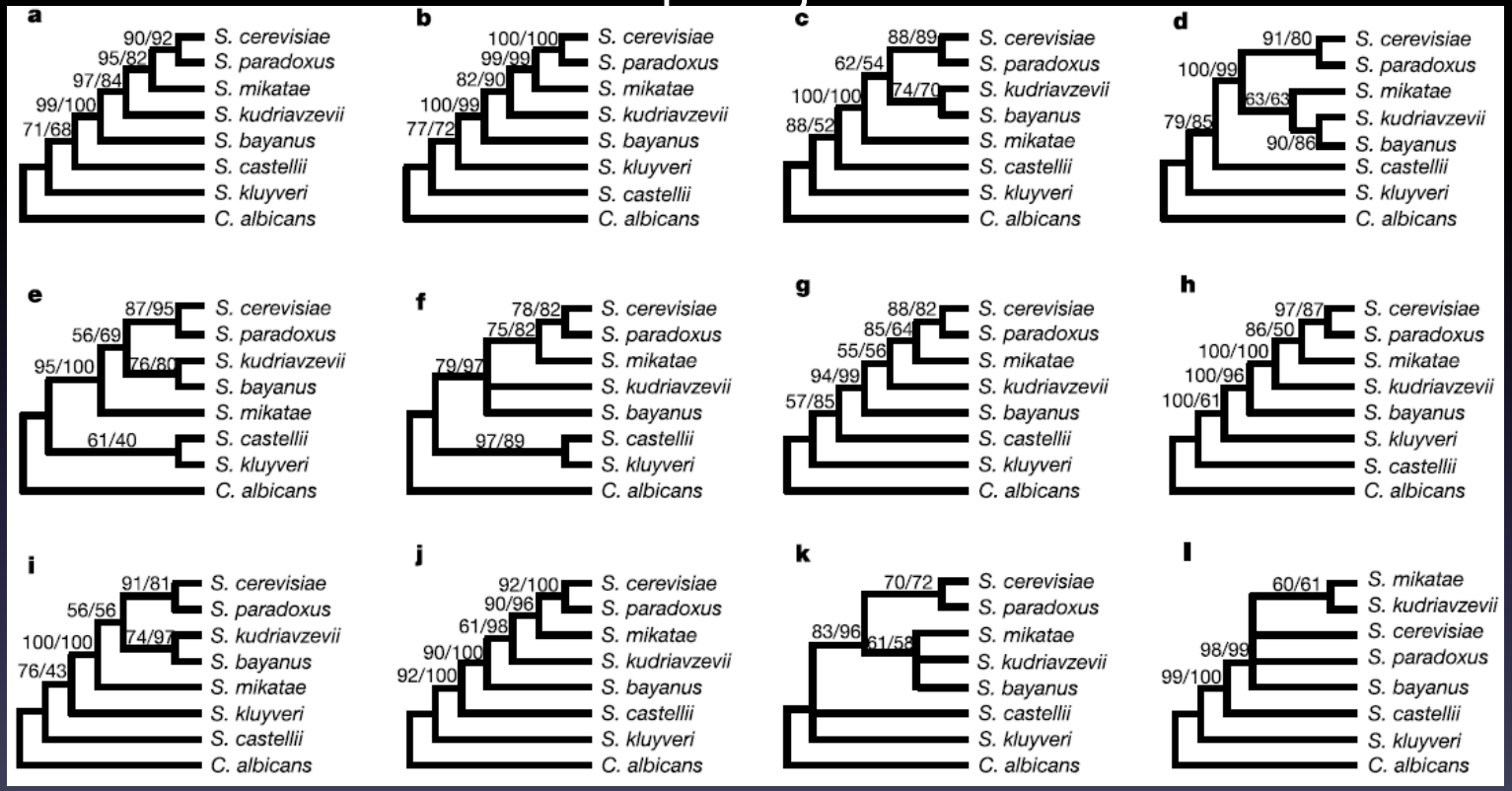
Adapted from Cedric Notredame

Supermatrix



Phylogenomics and the reconstruction of the tree of life. Frédéric Delsuc, Henner Brinkmann, Hervé Philippe (2005) *Nature reviews. Genetics* 6 (5) p. 361-75

Single-gene data sets generate multiple, robustly supported alternative topologies.



Genome-scale approaches to resolving incongruence in molecular phylogenies.
Antonis Rokas, Barry L Williams, Nicole King, Sean B Carroll (2003)
Nature 425(6960) p. 798-804

Concatenation

Right Topology ✚ High Bootstrap Support

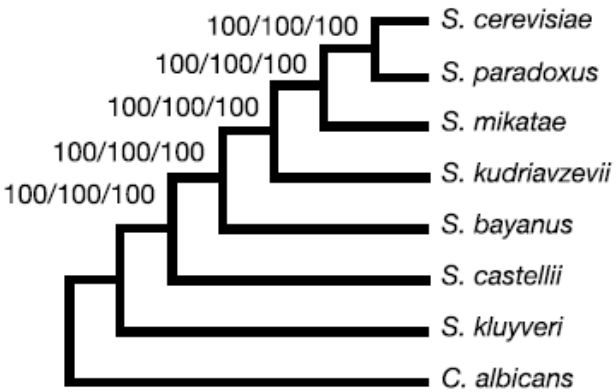


Figure 4 Phylogenetic analyses of the concatenated data set composed of 106 genes yield maximum support for a single tree, irrespective of method and type of character evaluated. Numbers above branches indicate bootstrap values (ML on nucleotides/MP on nucleotides/MP on amino acids).

Genome-scale approaches to resolving incongruence in molecular phylogenies. Antonis Rokas, Barry L Williams, Nicole King, Sean B Carroll (2003) *Nature* 425 (6960) p. 798-804

If you want to build a tree,

Clustal

Dalign

DCA

Mafft

Muscle

ProbCons

T-Coffee

Which guy should I trust?

panel D). Most studies ignore that these scores are based on a fixed sequence alignment that supports the tree in the first place; they may thus make us overly confident of its accuracy.

Ari Löytynoja and Nick Goldman, “Uniting Alignments and Trees,” *Science* 324, no. 5934 (June 19, 2009): 1528 -1529.

Alignment uncertainty - aligner

YPL077C with 6 topologies

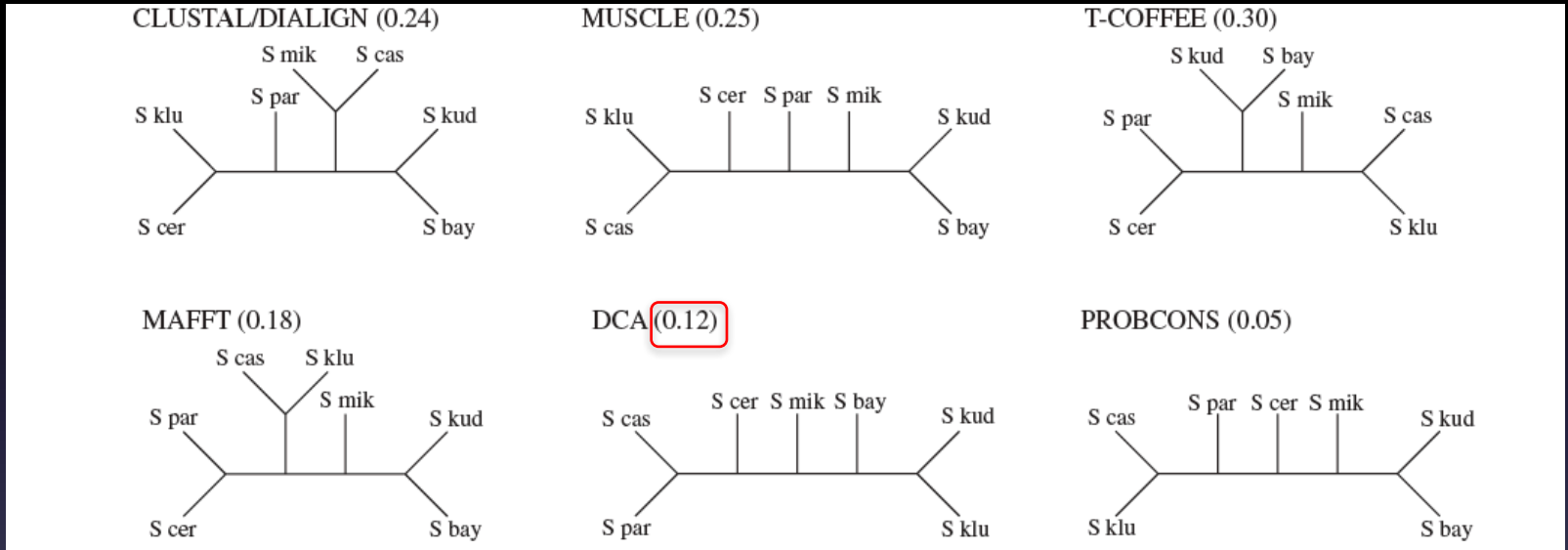


Fig. 1. An example, involving ORF YPL077C, in which alignments produced by seven different alignment methods produce six different estimated trees, albeit with low bootstrap support (bootstrap proportions shown parenthetically for each tree).

Karen M Wong, Marc A Suchard, and John P Huelsenbeck, "Alignment uncertainty and genomic analysis", Science 319, no. 5862 (January 25, 2008): 473-476.

Thank you!
Any question?

WWW resources for molecular phylogeny (1)

- Compilations
  - A list of sites and resources:

    http://www.ucmp.berkeley.edu/subway/phylogen.html
  - An extensive list of phylogeny programs

    http://evolution.genetics.washington.edu/phylip/software.html
- Databases of rRNA sequences and associated software
  - The rRNA WWW Server - Antwerp, Belgium.
    http://rrna.uia.ac.be
  - The Ribosomal Database Project - Michigan State University

    http://rdp.cme.msu.edu/html/

WWW resources for molecular phylogeny (2)

- Database similarity searches (Blast) :
  http://www.ncbi.nlm.nih.gov/BLAST/
- http://www.infobiogen.fr/services/menuserv.html
- http://bioweb.pasteur.fr/seqanal/blast/intro-fr.html
- http://pbil.univ-lyon1.fr/BLAST/blast.html
- Multiple sequence alignment
- ClustalX : multiple sequence alignment with a graphical interface (for all types of computers).
  http://www.ebi.ac.uk/FTP/index.html and go to ‘software’
- Web interface to ClustalW algorithm for proteins:
- http://pbil.univ-lyon1.fr/ and press “clustal”

WWW resources for molecular phylogeny (3)

- Sequence alignment editor
  - SEAVIEW : for windows and unix

    http://pbil.univ-lyon1.fr/software/seaview.html
- Programs for molecular phylogeny
  - PHYLIP : an extensive package of programs for all platforms
    http://evolution.genetics.washington.edu/phylip.html
  - PAUP : a very performing commercial package
    http://paup.csit.fsu.edu/index.html
  - PHYLO_WIN : a graphical interface, for unix only
    http://pbil.univ-lyon1.fr/software/phylowin.html
  - MrBayes : Bayesian phylogenetic analysis http://morphbank.ebc.uu.se/mrbayes/
  - PHYML : fast maximum likelihood tree building
    http://www.lirmm.fr/~guindon/phyml.html
  - WWW-interface at Institut Pasteur, Paris
    http://bioweb.pasteur.fr/seqanal/phylogeny

WWW resources for molecular phylogeny (4)

- Tree drawing

  NJPLOT (for all platforms)

  `http://pbil.univ-lyon1.fr/software/njplot.html`

- Lecture notes of molecular systematics

  `http://www.bioinf.org/molsys/lectures.html`

WWW resources for molecular phylogeny (5)

- Books
  - Laboratory techniques

    Molecular Systematics (2nd edition), Hillis, Moritz & Mable eds.; Sinauer, 1996.
  - Molecular evolution

    Fundamentals of molecular evolution (2nd edition); Graur & Li; Sinauer, 2000.
  - Evolution in general

    Evolution (2nd edition); M. Ridley; Blackwell, 1996.