2.
Sequence alignments and searches

Sequence homology

Two sequences are homologous, if they have a common ancestor

Two homologous sequences are more likely to have similar function than two unrelated sequences.

orthologs vs paralogs

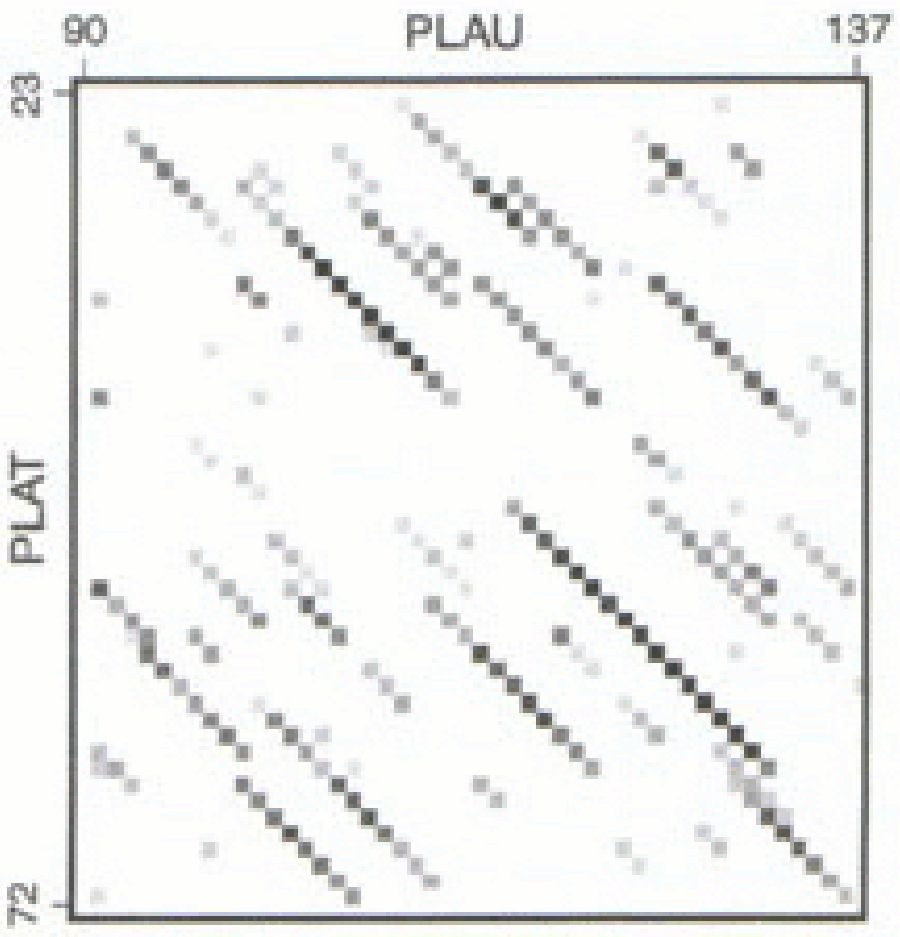
Comparing sequence pairs

- **Dot-matrix**

– One sequence in the row,

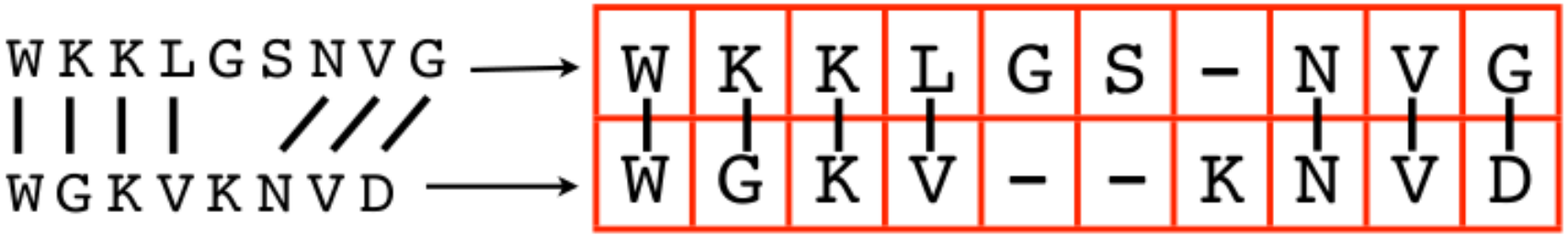
the other in the column

– a dot where they agree



Sequence alignment

Alignment: Finding equivalent regions of two or more sequences to maximize their similarity



Often represented using a grid/matrix:

One sequence per row

Residues in the same column are 'equivalent'

Gap characters (usually "-") indicate that the sequence contains no residues 'equivalent' to other residues in that column

Sequence alignment

Score: to evaluate similarity
- Substitution matrices

Gap penalty: to account for insertions/deletion

Algorithms can be used to maximize the overall score

– dynamic programming

– heuristic algorithms

Substitution matrices

Assign scores to aligned sequence positions

- Simplest: +1, if identical, 0 otherwise
- Certain amino acid replacements are more common, because they are better tolerated, these should get better scores
- The amino acid substitution matrices assign a score to each possible amino acid replacements (20x20)
- Built from curated multiple sequence alignment of known closely related protein families.
- Most common: PAM, BLOSUM
- Optimal matrix choice depends on task

BLOSUM 62

S_{ij} – The substitution score for aligning two residues i and j.

p_{ij} – The probability that we observe residues a and b aligned in homologous sequence alignments.

q_i and q_j – are background frequencies

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C		9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D			6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E				5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F					6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G						6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H							8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I								4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K									5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L										4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M											5	-2	-2	0	-1	-1	-1	1	-1	-1
N												6	-2	0	0	1	0	-3	-4	-2
P													7	-1	-2	-1	-1	-2	-4	-3
Q														5	1	0	-1	-2	-2	-1
R															5	-1	-1	-3	-3	-2
S																4	1	-2	-3	-2
T																	5	0	-2	-2
V																		4	-3	-1
W																			11	2
Y																				7

$$S_{ij} = \left(\frac{1}{\lambda}\right) \log\left(\frac{p_{ij}}{q_i * q_j}\right)$$

Negative-positive values

Sequence cutoff in original alignments was 62% identity

Gap

Indicates an insertion in one sequence or deletion in all other

- Gaps are usually characterized by two values:

– Gap opening

  Penalizing the occurrence of a gap in the alignment

– Gap extension

  if an already existing gap is extended

Gap extension is usually smaller than gap opening

Dynamic programming

Dynamic programming is a slow method that guarantees to find a mathematically optimal solution, not necessary the biological correct solution.

Objective: to find optimal alignment between sequences a1...an and b1...bn.

Be careful! DP will happily align completely unrelated sequences, the input depends on you!

Sequence identity

How similar are two sequences?:

- Calculated from the sequence alignment
- Simplest: number of identities compared to the total length of the alignment
- Number of identities in random cases ???

When is the similarity significant?:

- Generally above 30% we can say there is significant similarity
- between 20% and 30%: twilight zone
- If the number of identities is higher the number of similarities,
  - it is not likely to be significant
- *Depends on sequence length!!!!*

BLAST – Basic Local Alignment Search Tool

Goal: A fast search for homologous proteins or DNA/RNA sequences in a huge database

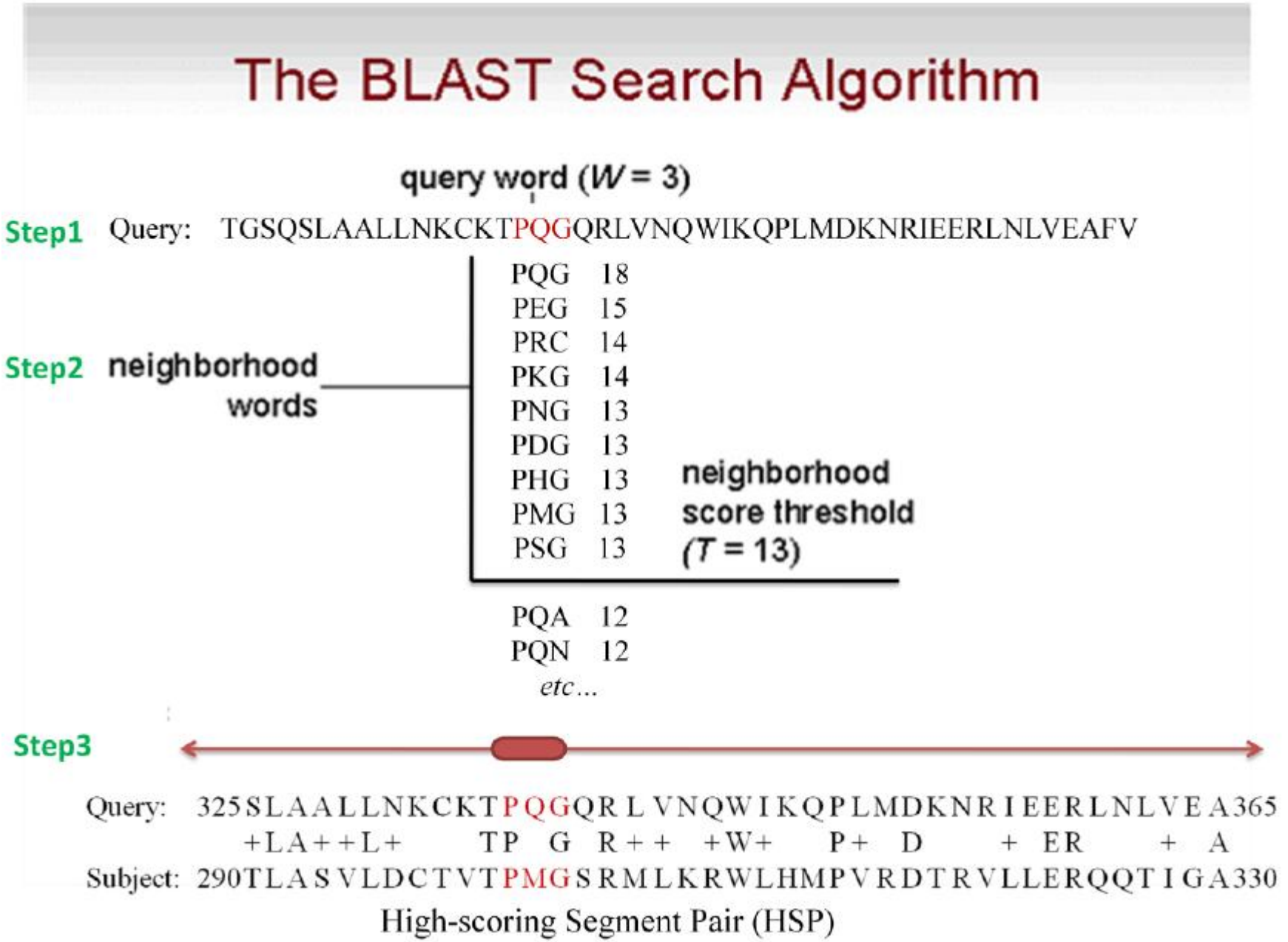
Key concept: Homologous sequences expected to contain several very similar short segments without gaps.

Heuristic: Technique designed to solve a problem that ignores whether the solution can be proven to be correct, but which usually produces a good solution...

Statistics: Provides statistical significance of alignments based on score distribution of random local alignments

Altschul, S.F.,Gish, W., Miller, W., Myers, E.W., and Lipman,D.J(1990) “basic local alignment search tool” J. Mol. Biol. 215: 403-410

BLAST


The BLAST Search Algorithm
query word (W = 3)
Step1 Query: TGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFV
Step2 neighborhood words
PQG 18
PEG 15
PRC 14
PKG 14
PNG 13
PDG 13
PHG 13
PMG 13
PSG 13
neighborhood score threshold (T = 13)
PQA 12
PQN 12
etc…
Step3
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L+ TP G R++ +W+ P+ D + ER + A
Subject: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330
High-scoring Segment Pair (HSP)

The statistics of local sequence comparison

E-value (Expectation Value) The Expect value (E) is a parameter that describes the number of hits one can "***expect" to see by chance*** when searching a database of a particular size. The lower the E-value, the more "significant" a match to a database sequence is.

For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

$$E = Kmn\, e^{-\lambda S}$$

m and n are the sequences lengths, S is the score found by BLAST
K and λ are scales for the search space size and the scoring system.

BLAST Statistics

```
 Score = 18.5 bits (36),  Expect = 47992
 Identities = 5/5 (100%), Positives = 5/5 (100%), Gaps = 0/5 (0%)

Query   1    ELVIS   5
             ELVIS
Sbjct   8    ELVIS   12
```

- Number of chance alignments = 48 thousand!
- Indistinguishable from chance

The most important statistic: Expect value (e-value)

Expected number of random alignments with a particular score or better

```
 Score = 89.7 bits (204),  Expect = 7e-18
 Identities = 50/103 (49%), Positives = 54/103 (52%), Gaps = 18/103 (17%)

Query   1    MKLLAATVL---LLTICSLEGALVR
             MK L   VL   LL +CSLEGA V
Sbjct   1    MKVL---VLAMVLLCVCSLEGAVVM

Query   54   SPELQAEAKSYFEKSKEQLTPLIKKAGTELVNFLSYFVELGTQ  96
               E   +AK Y E   EQ  P  K   TE       F +L TQ
Sbjct   5
```

- Number of chance alignments = 7×10^{-18}
- Not due to chance

- The e-value depends directly on the size of the search space (database)
- Search the smallest database likely to contain the sequence of interest

What is a significant hit?

E-value (Expectation Value) depends

- Database
- Amino acid composition of the sequence
- Length
- ….

There is no absolute cutoff!

Generally 0.001 or 0.0001 is a good starting point

PSI-BLAST
Position-Specific-Iterated BLAST

PSI-BLAST is a cycling/iterative method that provides increased sensitivity for detecting distantly related proteins, using the evolutionary information from the protein “family”.

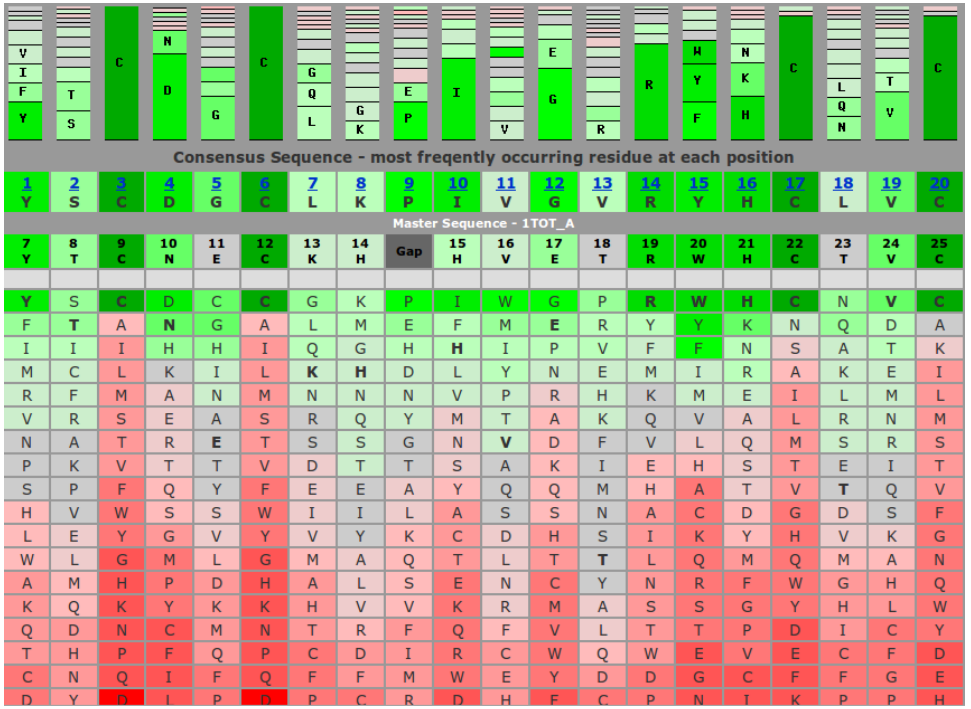
PSI-BLAST is still fast – still based on BLAST methods and simple to use

Position-Specific Scoring Matrix PSSM

PSSM amino acid substitution scores are position dependent in a protein multiple sequence alignment. Thus, a Tyr-Arg substitution may score different for different alignment positions.

In PSI-BLAST , a PSSM replaces BLOSSUM in the second iteration of BLAST

```
                      10        20        30        40        50
                 ....*....|....*....|....*....|....*....|....*....|....
Feature 1              #  #           #  #        #  #          #   #
1TOT_A           7 YTCNECKh--HVe-TRWHCTVCe--DYDLCINCYNTk------SH--THKMVKW 47
gi 7023094      95 ISCDGCDe--IApwHRYRCLQCs--DMDLCKTCFLGgvk--peGHgdDHEMVNM 142
gi 50750334    201 VRCRVCKtfpITg-LRYRCLKCl--NFDLCQVCFFTgrh--skPHksSHPVVEH 249
gi 50257626    582 SECTICLtalFS--NRFKCVSCp--KFDLCRSCYQKvd----eIHp-AHAFLSL 626
gi 47222763     98 IICDSCKkhgIMg-MRWKCKVCf--DYDLCTQCYMNn------KHdlSHAFERY 142
gi 40743717   1022 RVCNNCLk-eFDegKMVSCADCd--DFDLCITCILGhk---hgHHp-SHTFVLL 1068
gi 51261627     15 PPCKGCSs-yLMe-PYIKCAECgppEFLLCLQCFSGfe---ykKHqsDHSYEIM 63
gi 16944480    367 RTCNCCIq-dLPeaEFVHCQTCd--DFDLCKVCFAKnr----hGHhpKHAFSPI 413
gi 42546497    336 RTCNCCVq-eHPeaEFLHCRMCe--DFDLCQSCFARds----hGHhpKHSFAPA 382
gi 40745179    373 IICDGCNaegLA--VQYHCADCe--DYDLCQSCYKAgtrcgykGHt-YHLEFNA 421
```



PSI-BLAST Algoritmus

I. A standard BLAST search is performed against a database using a substitution matrix (e.g. BLOSUM62).

II. A PSSM (checkpoint) is constructed automatically from a multiple alignment of the highest scoring hits of the initial BLAST search.

III. The PSSM replaces the initial matrix to perform a second BLAST iteration search.

IV. Steps 2 and 3 can be repeated and the new found sequences included to build a new PSSM at each iteration.

Low complexity regions

Statistical estimates don't work well from low complexity regions during sequence searches

```
QQQQQQQQQQ
||||||||||
QQQQQQQQQQ
```

10/10 id

```
IDENTITIES
||||||||||
IDENTITIES
```

10/10 id

```
QQQQQQQQQQ
||||||||||
QQQQQQQQQQ
```

Shuffled: 10/10 id

```
IDENTITIES
   | |
SIINDIETTE
```

Shuffled: 2/10 id

Low complexity regions

It is possible to filter out low complexity regions using the SEG algorithm:

1) Find low complexity segments within a sequence window
2) Neighboring regions are merged

Eliminated common basic, acidic and proline rich regions

Improvements: Composition based statistics

(Wootton and Federhen, 1993)

Types of alignments

Global vs Local

Global alignment: Includes all characters from each sequence

Local alignment: Includes only the most similar local regions, typically not the whole sequence.

Pairwise vs Multiple

Pairwise alignment: A sequence alignment of two sequences.

Multiple sequence alignment: A sequence alignment of three or more biological sequences.

Exact vs Heuristic

Exact alignment: Generated by dynamic programming which guarantees optimal alignment given scores and gap penalties

Heuristics alignment: Good enough alignment (Blast)

Multiple Sequence Alignment - MSA

Given a set of proteins, one often wants to find the evolutionary relationship among them, by aligning them.

We expect from a good MSA to highlight the functional regions: conserved residues, motifs, secondary structure tendency etc.

The quality of the alignment is very important.

NP hard problem, heuristic approaches

CLUSTALW

CLUSTALW was introduced in 1994 and had become commonly used among biologists.

Three steps:

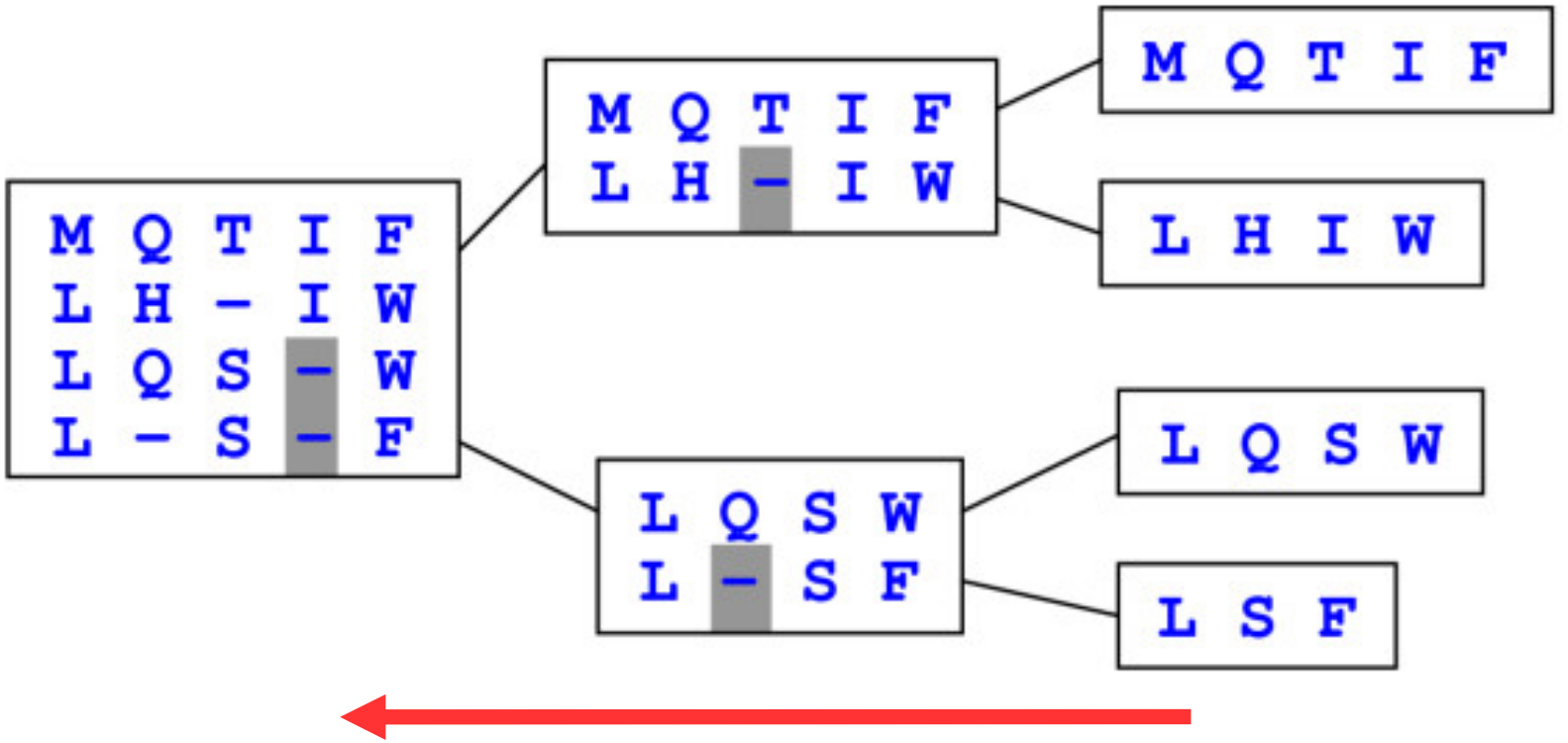
1. Pairwise sequence alignment

2. Based on the distances tree is generated

3. Based on the tree, sequence-sequence, sequence-profile, profile-profile alignments

Clustalw Omega, other methods

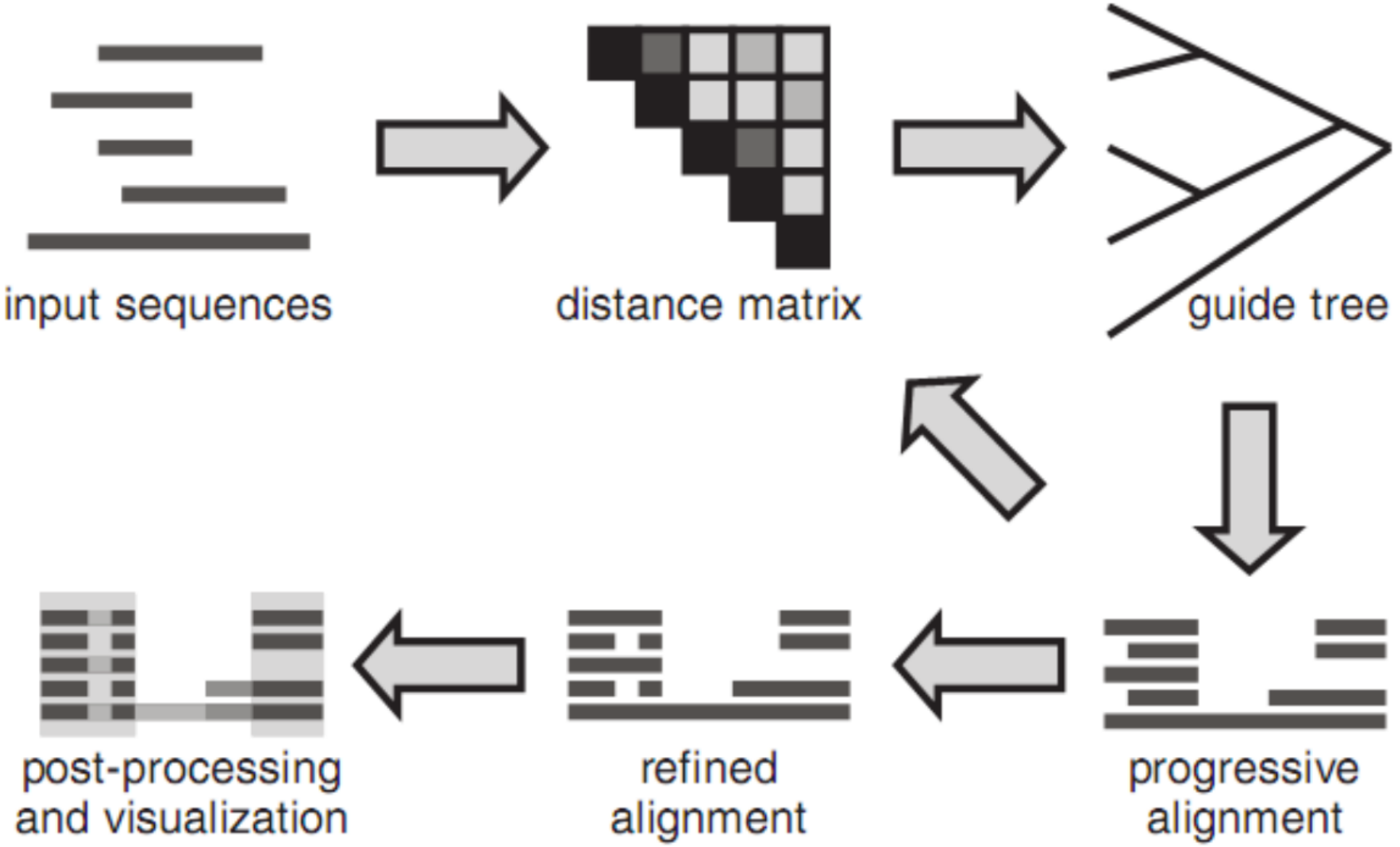
ClustalW Main idea: Progressive alignment



The sequence are assigned to the leaves of binary trees. At every inner points, we merge two child profiles based on the alignment.

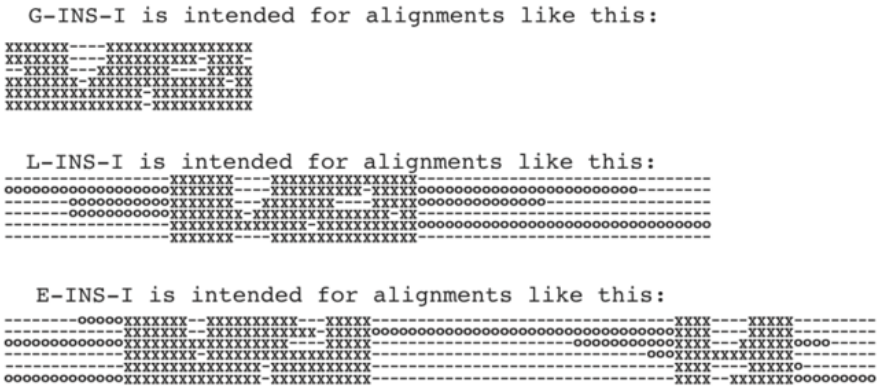
BMC Bioinformatics 2004, 5:113

Main Steps in Modern Multiple Sequence Alignment



MUSCLE and MAFFT

- Progressive alignment
- Based on word comparison (approximation)
- Iterative: new distance matrix
- MAFFT: Multiple options



ALIGNMENT editing

- Pruning
  - Removing badly alignable regions can improve further analyzes
  - Removing columns based on variability, gap ratio
- Manual
  - Some sequence cannot be aligned well
  - Manual editing can help, requires expertise

Jalview

A workbench for multiple sequence alignment and analysis