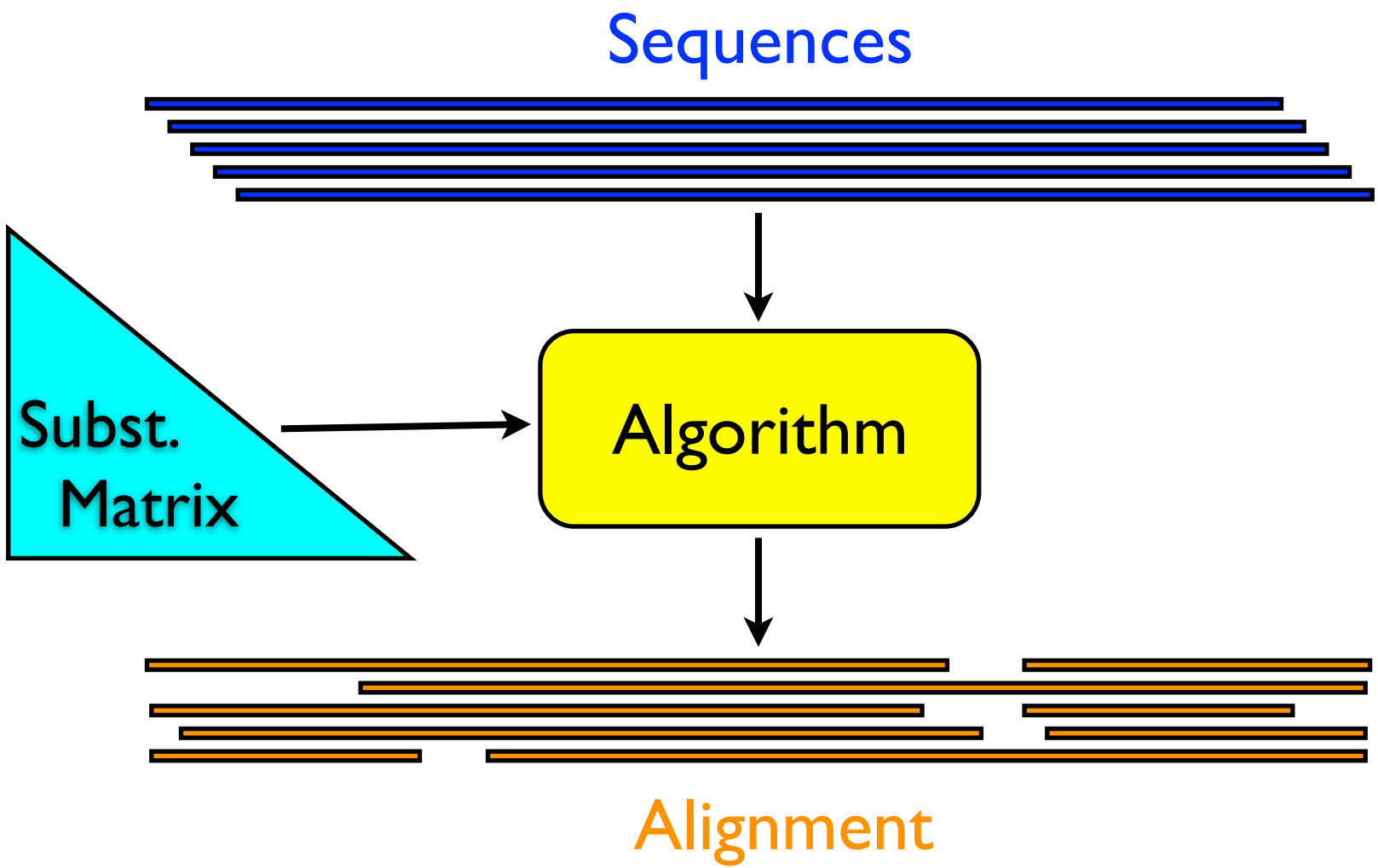
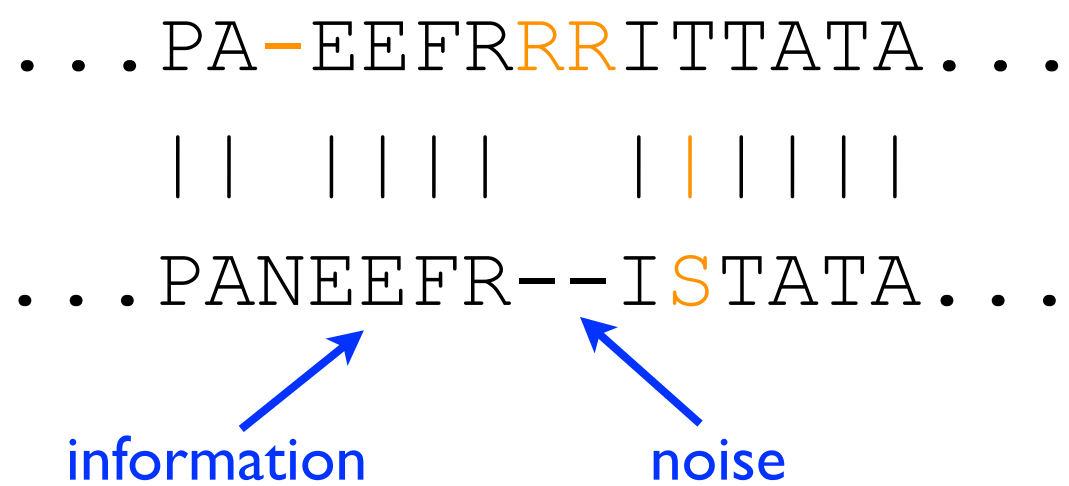
Lecture 3
Sequence and Structure Alignment

- Word Matching
- Dynamic Programming
- Substitution Matrix
- Gap Penalties
- DP Forward Algorithm
- DP Traceback
- Local/Global
- Alignment Quality
- Multiple Alignment
- HMMs
- References

Alignment Model



Conservation and Mutation: Information and Noise



- Pairs of identical or similar amino acid residues carry information to construct an alignment.

- In terms of alignability, mutations, in particular INDEL events, represent noise, because the original signal (amino acid residue) is lost.

Word Matching

Exact word matching

- GNU grep (grep 'GRGDS' *.seq)
- Many exact string matching algorithms exist
- Suffix tree

Approximate word matching (pattern matching)

- GNU grep with wildcards (grep '\<GR..S\>' *.seq)
- Pattern matching syntax in programming languages
- PROSITE pattern

Fast word matching with precise statistics

- Blast

Pattern Matching

PERL

```
while (<IN>) {
    if ($_ =~ /^\s{1,}\d{1,}.{8}\w{3,5}[C,O,N,S,H]/) {
```

C

```
#include <regex.h>

char matchPatternStr[] = ".*[[:digit:]]{1,5} [[:print:]]{1,5} NA";
regex_t matchPatternReg;
char searchPatternStr[] = " 1 LYS NA";

regcomp(&matchPatternReg, matchPatternStr, REG_EXTENDED) == 0);

regexec(&matchPatternReg, searchPatternStr, 0, NULL, 0);
```

Database of protein domains, families

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associa
PROSITE is complemented by **ProRule**, a collection of rules based on profiles and patterns, which increases the discri
acids [More details].

Release 20.63, of 20-Apr-2010 (1577 documentation entries, 1308 patterns, 886 profiles and 883 ProRule)

PROS

e.g: PDOC00022, PS50089, SH3, zinc finger

Search

add wildcard '*'

PRC

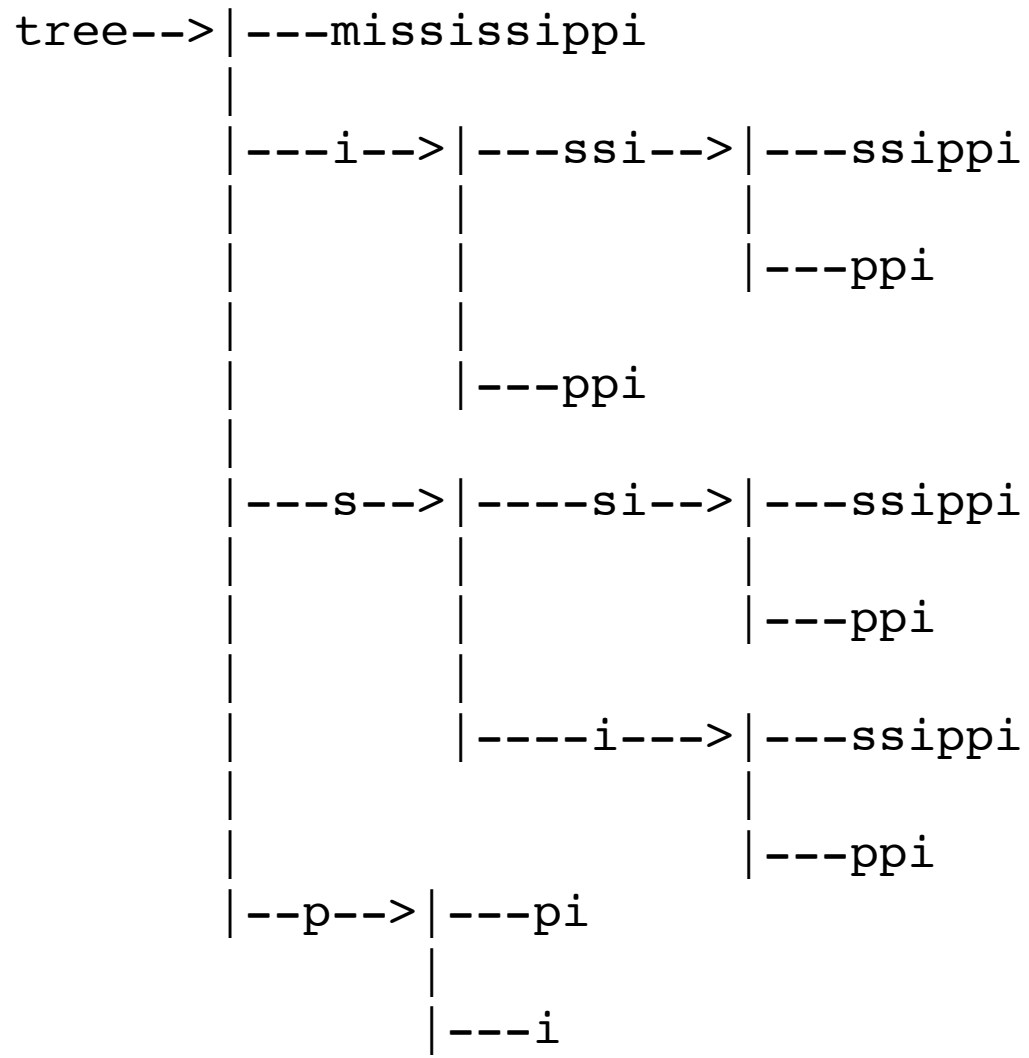
Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)

Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC [help]:

Scan Clear

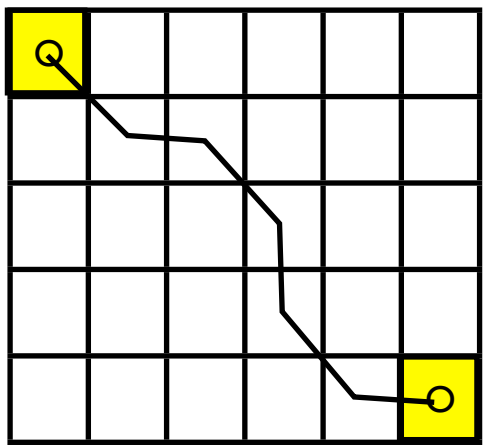
Suffix Tree of mississippi

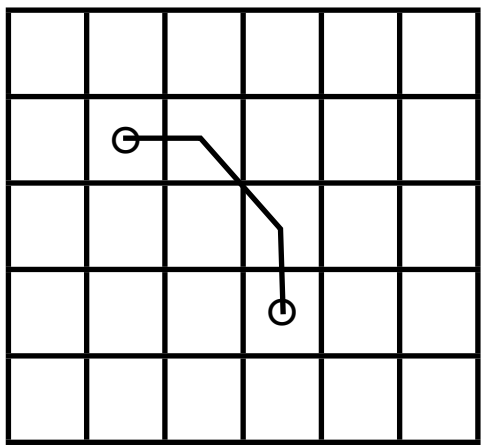


Dynamic Programming

Dynamic Programming is an iterative algorithm with square complexity and square memory usage that generates the optimal pairwise alignment given a scoring scheme, i.e. a substitution matrix and gap penalties.

Global and Local Alignment





Global Alignment:
highest scoring path
from top left cell
to bottom right cell

Local Alignment:
highest scoring segment
with all cell scores >0

Needleman & Wunsch 1970

Smith & Waterman 1981

3.7 Amino Acid Substitution Matrix

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

Similar amino acids (coloured blocks) have often positive substitution scores: the substitution occurs more often than randomly in proteins.

Substitution Matrices

PAM (Dayhoff 1978)
Original substitution matrix based on Markov process of amino acid substitution.

JTT (Jones, Taylor, Thornton 1992)
Modern version of PAM matrix.

GONNET (Gonnet, Cohen, Benner 1992)
Iterative refinement of alignment and PAM-like matrix.

BLOSUM (Henikoff & Henikoff 1992)
Uses the substitutions observed in conserved blocks of multiple sequence alignments.

Amino Acid Substitution Matrix

Substituting amino acid A with B:

$$Score_{AB} = \log [p(AB) / (p(A) * p(B))]$$

p(AB) : probability of aligned amino acid pair AB in trusted alignment.

p(A) * p(B): probability of observing AB in random alignment (= background probability).

This type of score (log [p(observed) / p(random)]) is called relative entropy and it is related to the mutual information.

3.8 Gap Penalties

One can envisage sequence alignment as placing gaps at the correct place.

A gap is the model of an INDEL event. There are many types of INDEL events, but usually only one gap penalty parametrisation.

Common is the 'affine' gap penalty scheme with a high gap-open (g_o) penalty and a low gap-extension (g_e) value.

$$\text{Score(gap)} = g_o + l * g_e$$

with l : gap length

3.9 The DP Forward Algorithm: Initialise

Task: align GAGGCGA with GAGTGA!

$$G[i,j] = \max \begin{cases} G[i-1, j-1] \pm 1 \\ G[i, j-1] - 2 \\ G[i-1, j] - 2 \end{cases}$$

diagonal up-left
+1 for match, -1 for mismatch

up = gap in j

left = gap in i

	j	
i	i-1, j-1	i-1, j
	i, j-1	i, j

DP algorithm

DP alignment matrix →

	-	G	A	G	T	G	A
-	0						
G							
A							
G							
G							
C							
G							
A							

The DP Algorithm: Fist Column and Row

Procedure: Fill the DP matrix using the DP algorithm!

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2						
A	-4						
G	-6						
G	-8						
C	-10						
G	-12						
A	-14						

The DP Algorithm: Neighbour Cell Scores

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2	?					
A	-4						
G	-6						
G	-8						
C	-10						
G	-12						
A	-14						

The DP Algorithm: The First Step

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2	1	-1				
A	-4						
G	-6						
G	-8						
C	-10						
G	-12						
A	-14						

-1 -2 -2

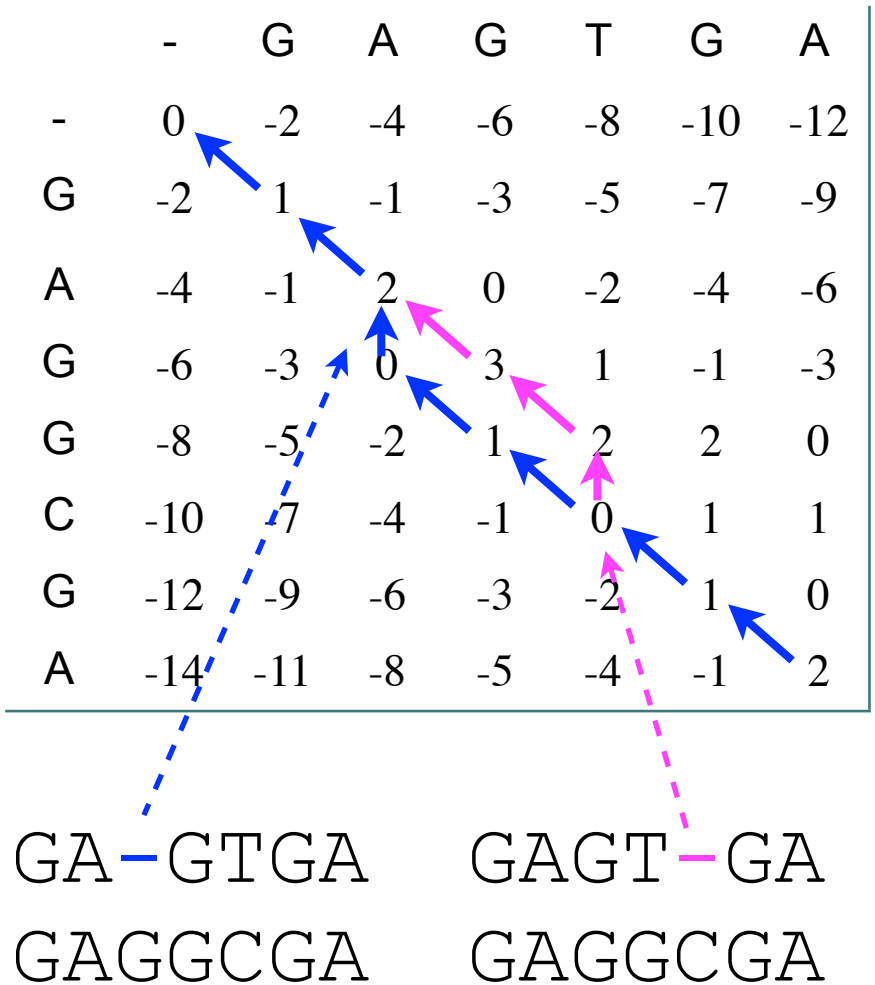
The DP Algorithm: and so on ...

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
			+1	-2			
G	-2	1	-1	-2	-3		
A	-4						
G	-6						
G	-8						
C	-10						
G	-12						
A	-14						

The DP Algorithm: The Complete DP Matrix

	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2	1	-1	-3	-5	-7	-9
A	-4	-1	2	0	-2	-4	-6
G	-6	-3	0	3	1	-1	-3
G	-8	-5	-2	1	2	2	0
C	-10	-7	-4	-1	0	1	1
G	-12	-9	-6	-3	-2	1	0
A	-14	-11	-8	-5	-4	-1	2

Dynamic Programming: Traceback



Join the optimal steps to an optimal path.

Local/Global Alignment

'Global alignment' aligns the entire length of two sequences. Traceback is started from the right-lower DP matrix cell.

'Local alignment' aligns the highest scoring sequence segments with positive cell scores. All negative cell scores are set to zero in the matrix construction phase; traceback is started at the highest cell score in the DP matrix and terminated at the first cell with zero score.

The Meaning of Alignments

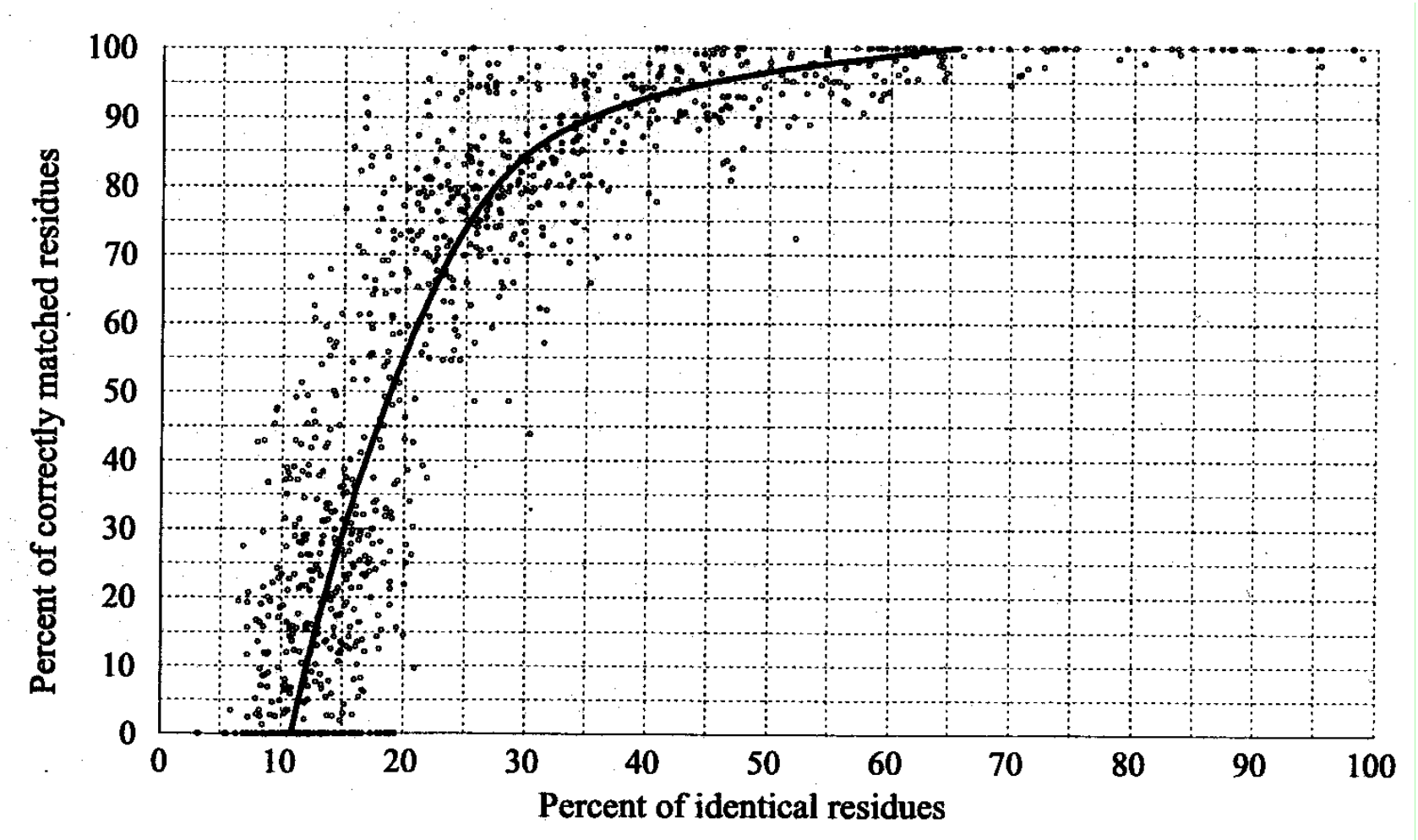
Local alignments select segments with exclusively positive scores, i.e. above-random information about biological similarity.

We assume that aligned residues have the same position in the phylogenetic history (homologues!).

Aligned residues have very similar roles in the molecular structure and function.

The standard of truth against which alignment programs are calibrated are structure alignments.

Pairwise Alignment Quality



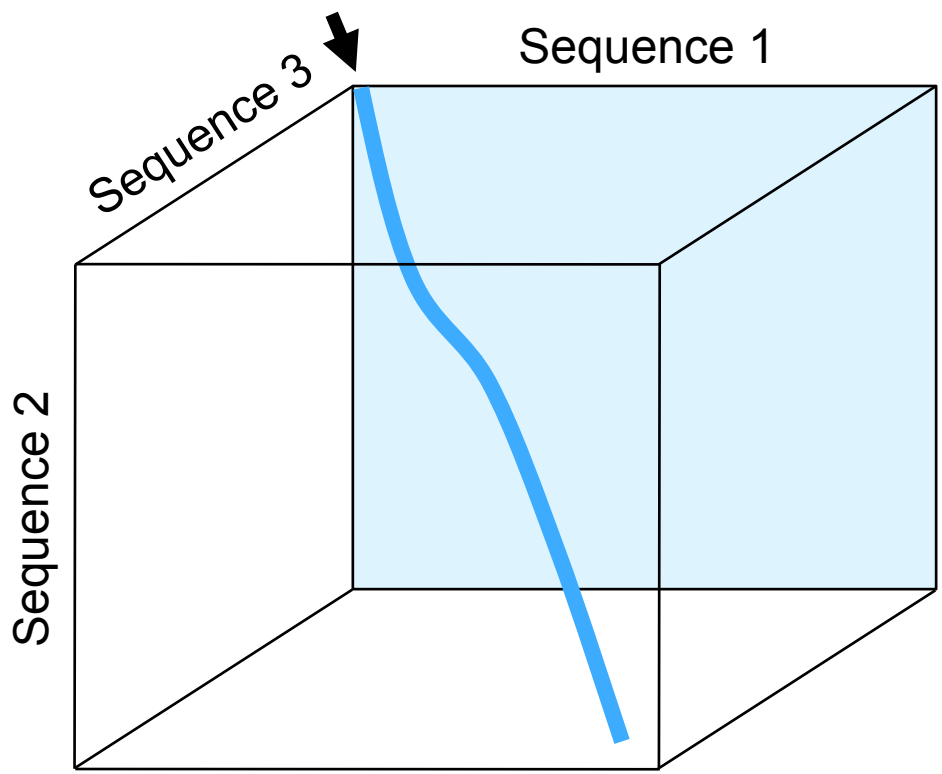
Vogt et al., JMB 249, 816-831,1995

3.13 Multiple Alignment

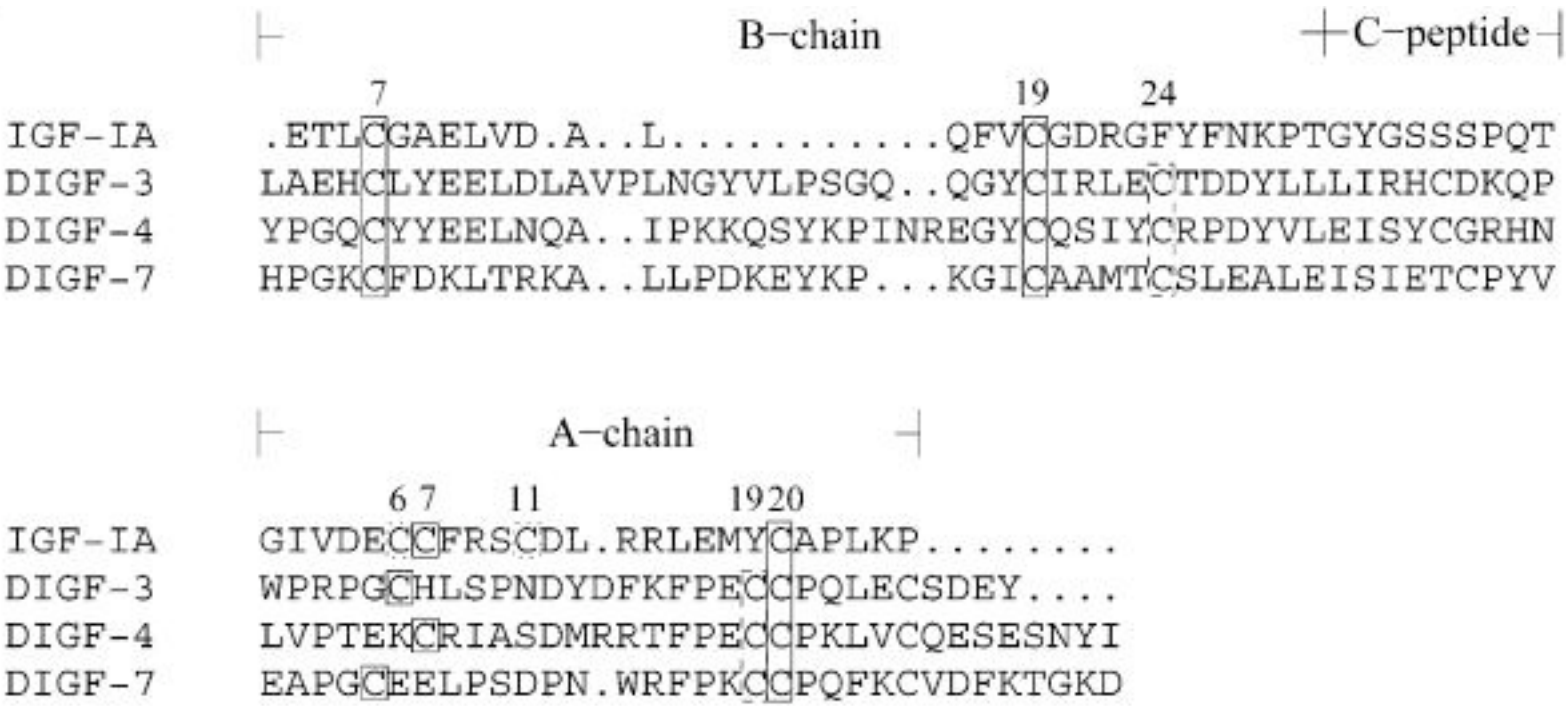
Multiple Alignment extends the concept of pairwise sequence alignment to >2 sequences. The aim is to align all evolutionary related amino acid or nucleotide residues in the same column.

Multiple DP

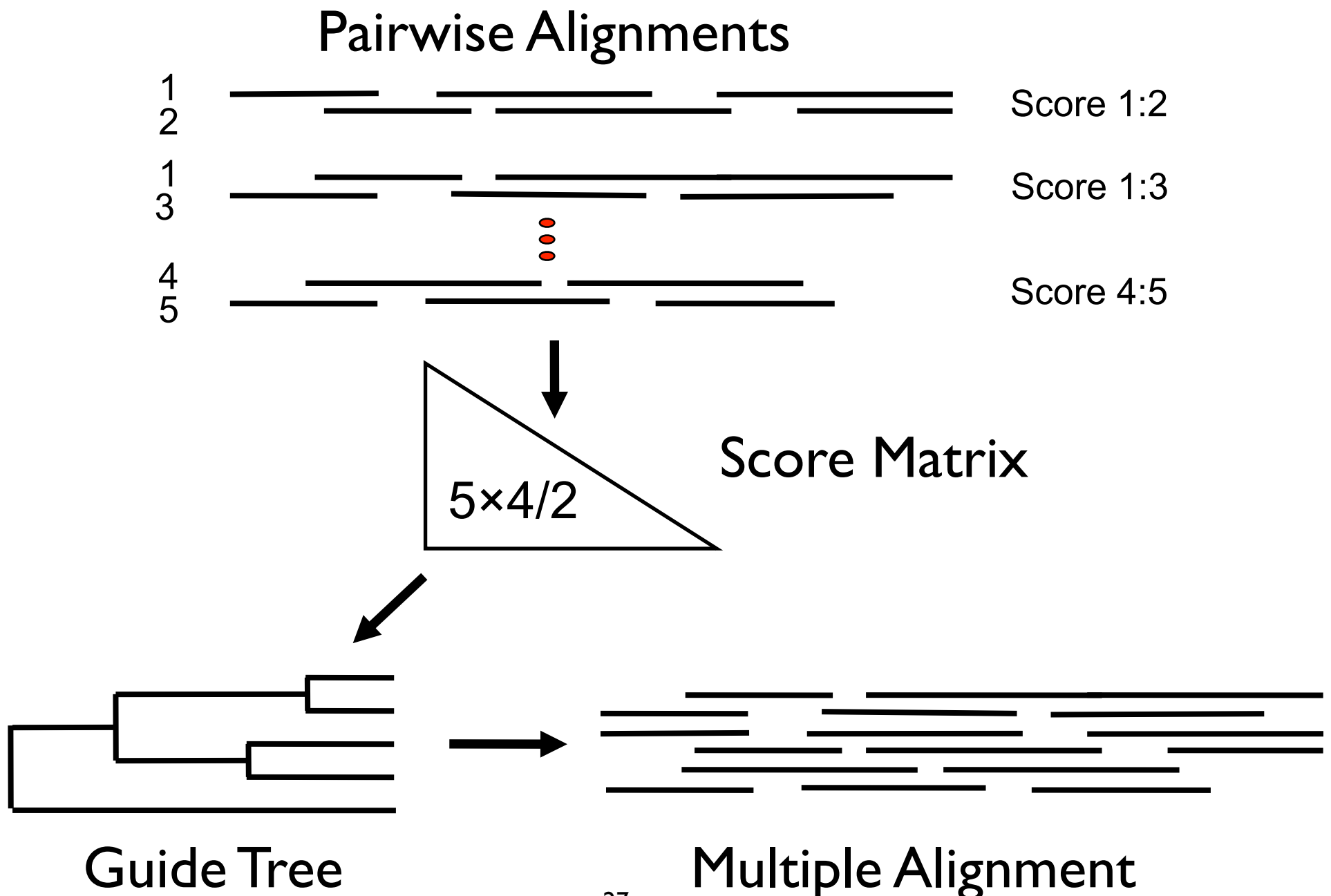
In principle one could perform multi-dimensional Dynamic Programming, but that becomes very slow for many sequences. Complexity l^n with l= sequence length and n = sequence number.



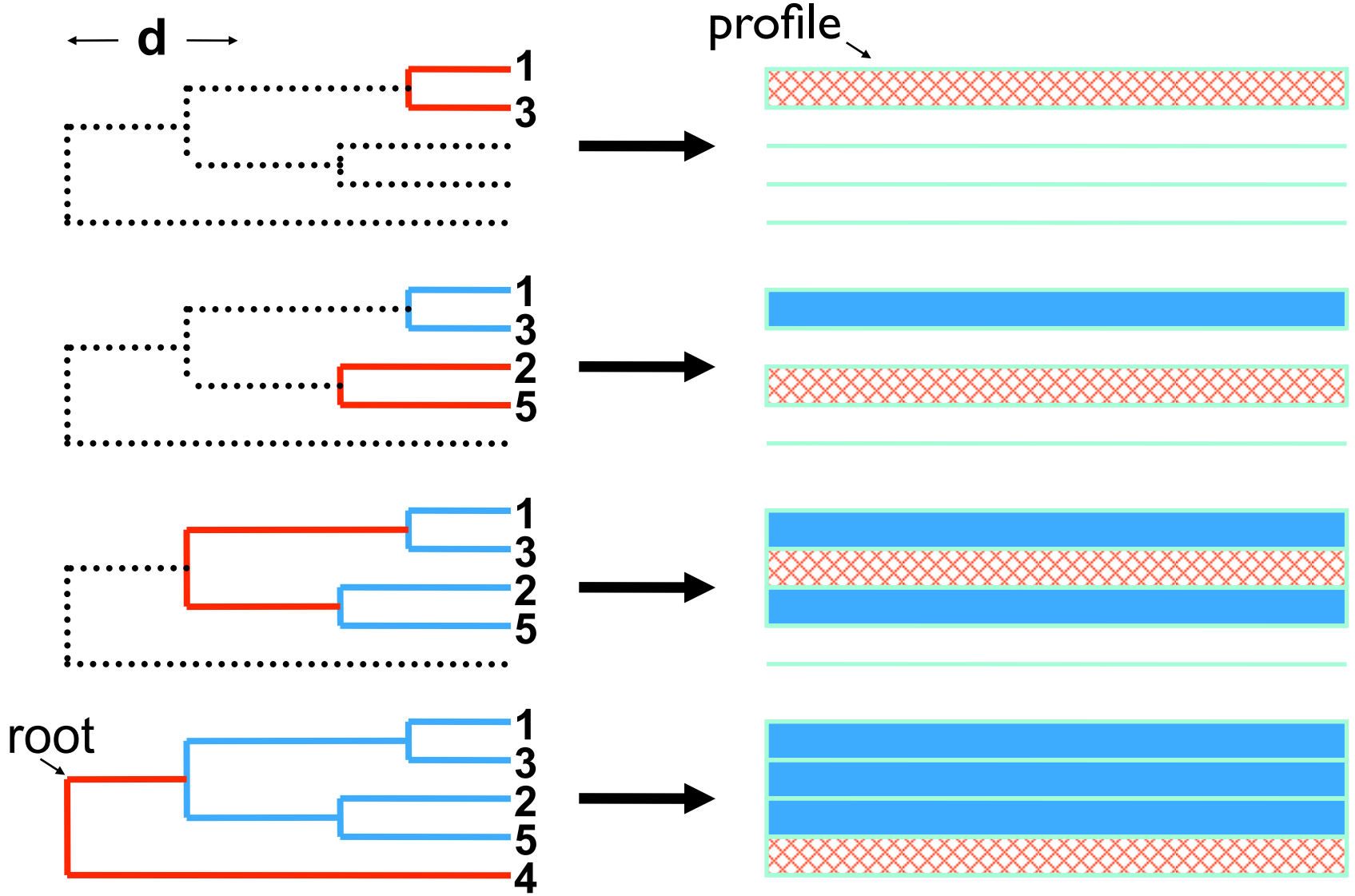
A Muliple Sequence Alignment



Progressive Alignment



Profiles from Guide Tree



Comments to Progressive Alignment

The idea is that the pairwise alignments between the closest (= highest scoring) sequences
have the least number of errors.

Errors in the pairwise alignments will not be corrected in the progressive phase!

At later stages in the scheme one needs to score and align sequences against profiles and profiles against profiles.

Improvement of Multiple Alignment Quality

Consistency check

Use consistency of matching (transitivity):
if A→B and B→C also A→C?
All the top-performing multiple alignment programs use consistency scores.

Homology information

Use profiles to enhance positional information.
Before the actual sequence alignment, collect all homologues from the database and use the profile for the alignment instead of the single sequence.

Multiple Alignment Programs

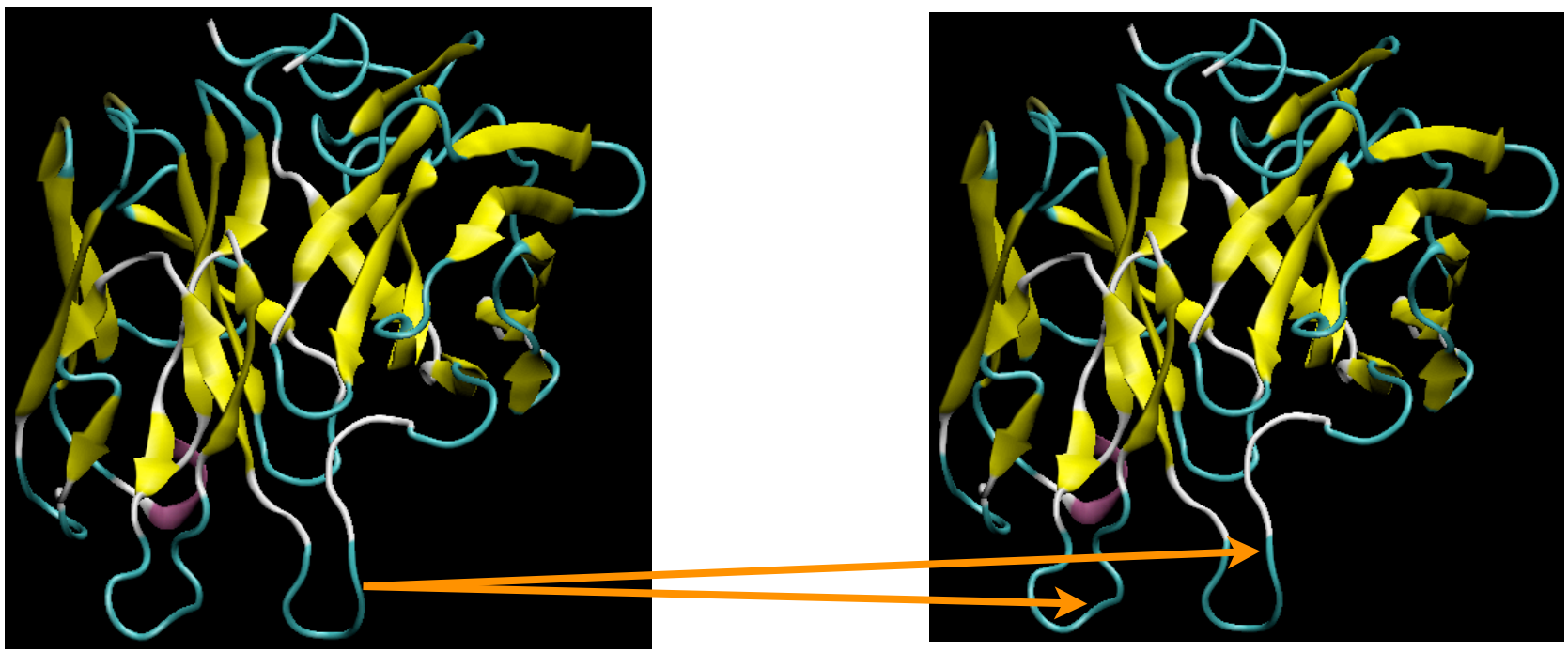
T-Coffee : versatile, combines local/global alignment

ProbCons : probabilistic (gap model) alignment

Muscle : fast, good for very large multiple alignments

SeqAln (C++ library) : alignment program templates

Structure Alignment



Structure alignment is a computationally hard problem. Residue matches are not independent (as in sequence alignment) because of the rigid 3D-structure.

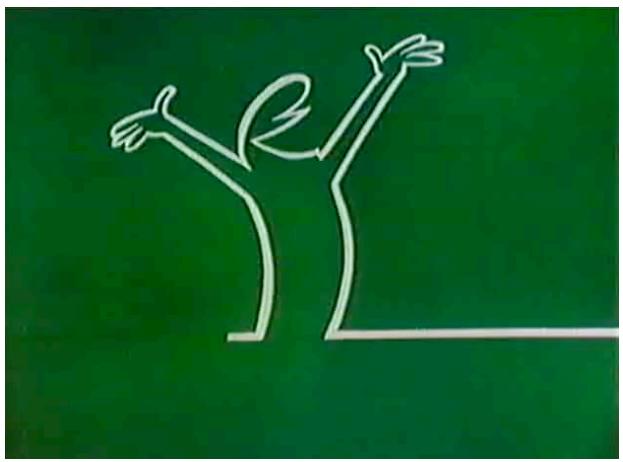
Possible Alignment Schemes

Align in all possible orientations (6D search space). Unfeasible for most structure pairs.

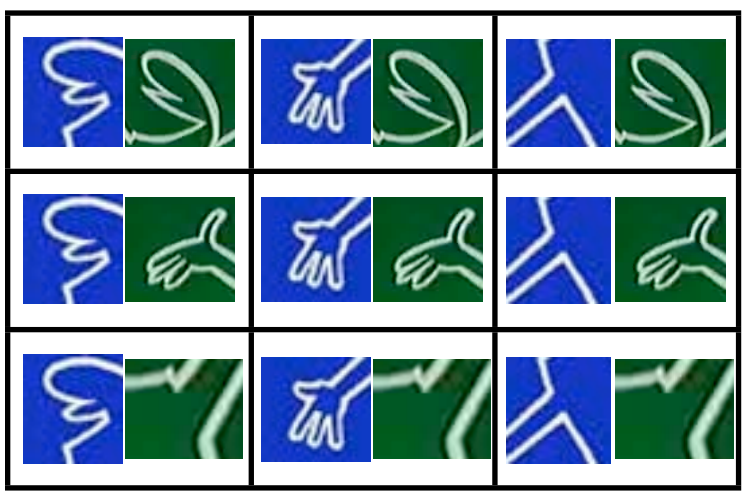
Use a coarse-grained representation (grid) and align in all possible orientations (reduced 6D search space). Limited to few pairwise comparisons.

Create optimal sub-solutions (fragment matches) and assemble these to near-optimal total solution. The search space is approximately n^2 with n = number of fragments.

A Structure Alignment Scheme



DP matrix with fragment matches



Structure Alignment Programs

TMalign (pairwise)

MAMMOTH (multiple)

Structure Searching

VAST (NCBI)

CE

Hidden Markov Model

"A first-order discrete HMM is a stochastic generative model for time series defined by a finite set S of states, a discrete alphabet A of symbols, a probability transition matrix $T = (t_{ji})$, and a probability emission matrix $E = (e_{iX})$. The system randomly evolves from state to state while emitting symbols from the alphabet."

P Baldi, S Brunak, 2001

Hidden Markov Model

Hidden Markov Models are similar to finite state automata.

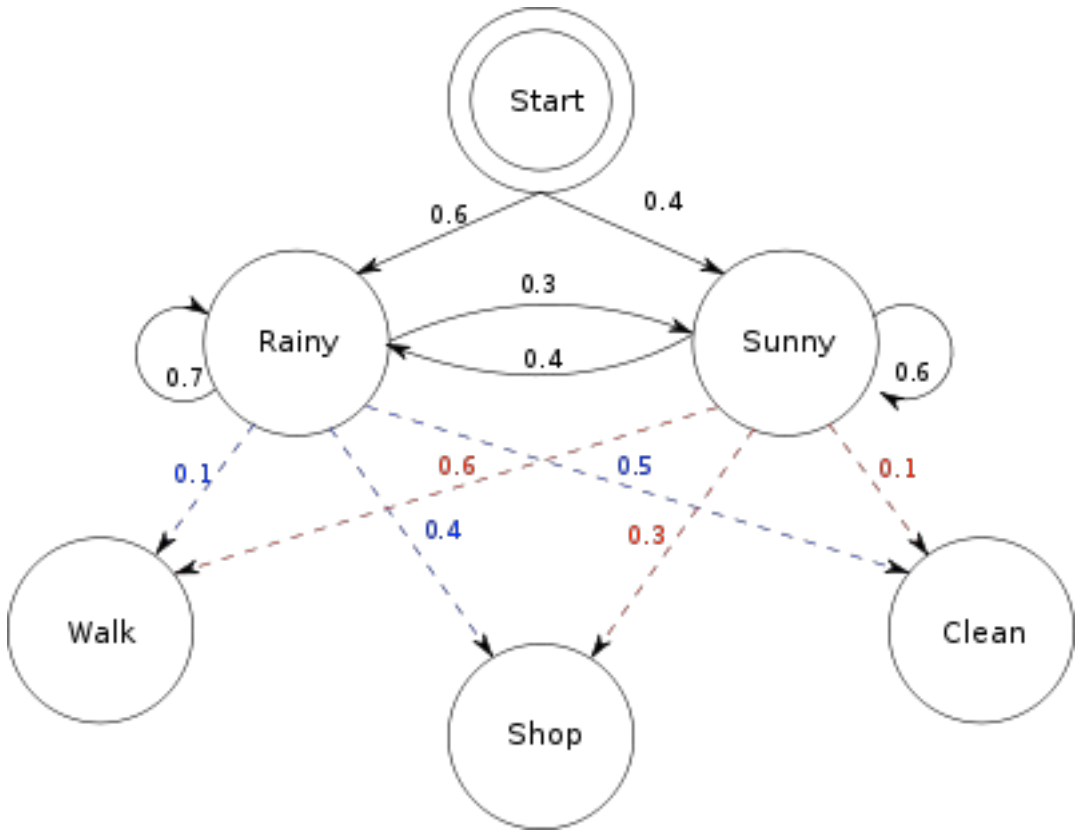
The system HMM represents a Markov chain with states and transitions between states. Walking along the Markov chain, each state emits a character (observable). The associated transition probabilities and emission probabilities of the Markov chain are hidden.

To determine the probabilities, a HMM has to be trained on representative data.

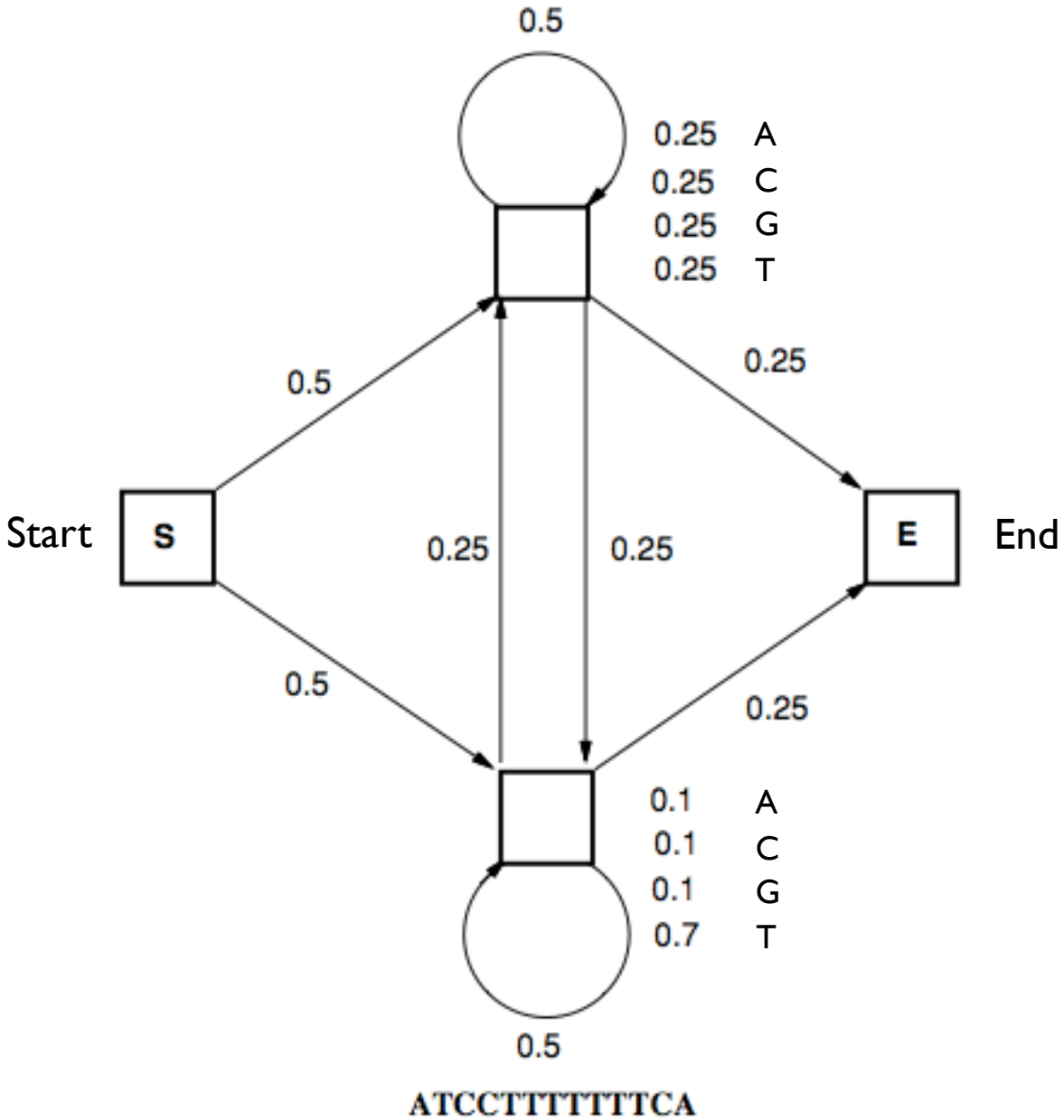
HMM Example

Transition probabilities (between 'states')

Emission probabilities (of 'actions' or 'characters')



Example taken from Wikipedia


0.5
0.25 A
0.25 C
0.25 G
0.25 T
0.5
0.25
Start
S
0.25
0.25
E
End
0.5
0.25
0.1 A
0.1 C
0.1 G
0.7 T
0.5
ATCCTTTTTTCA

What Can We Do With A HMM?

What are the best parameters given observed sequences (learning).

Compute the probability (likelihood) of a sequence.

Compute the most probable sequence of transitions and emissions (decoding). This yields the most probable sequence.

Given an alignment probability (Viterbi algorithm), decide whether a sequence belongs to a protein family.

PFAM : HMMs of protein families; very good tool to study the sequence properties of protein families

Learning Outcomes

- Word matching, Suffix Tree
- Alignment scheme
- Affine gap penalties
- Dynamic Programming algorithm
- Global / Local alignment
- Multiple sequence alignment
- Progressive strategy
- Hidden Markov Model