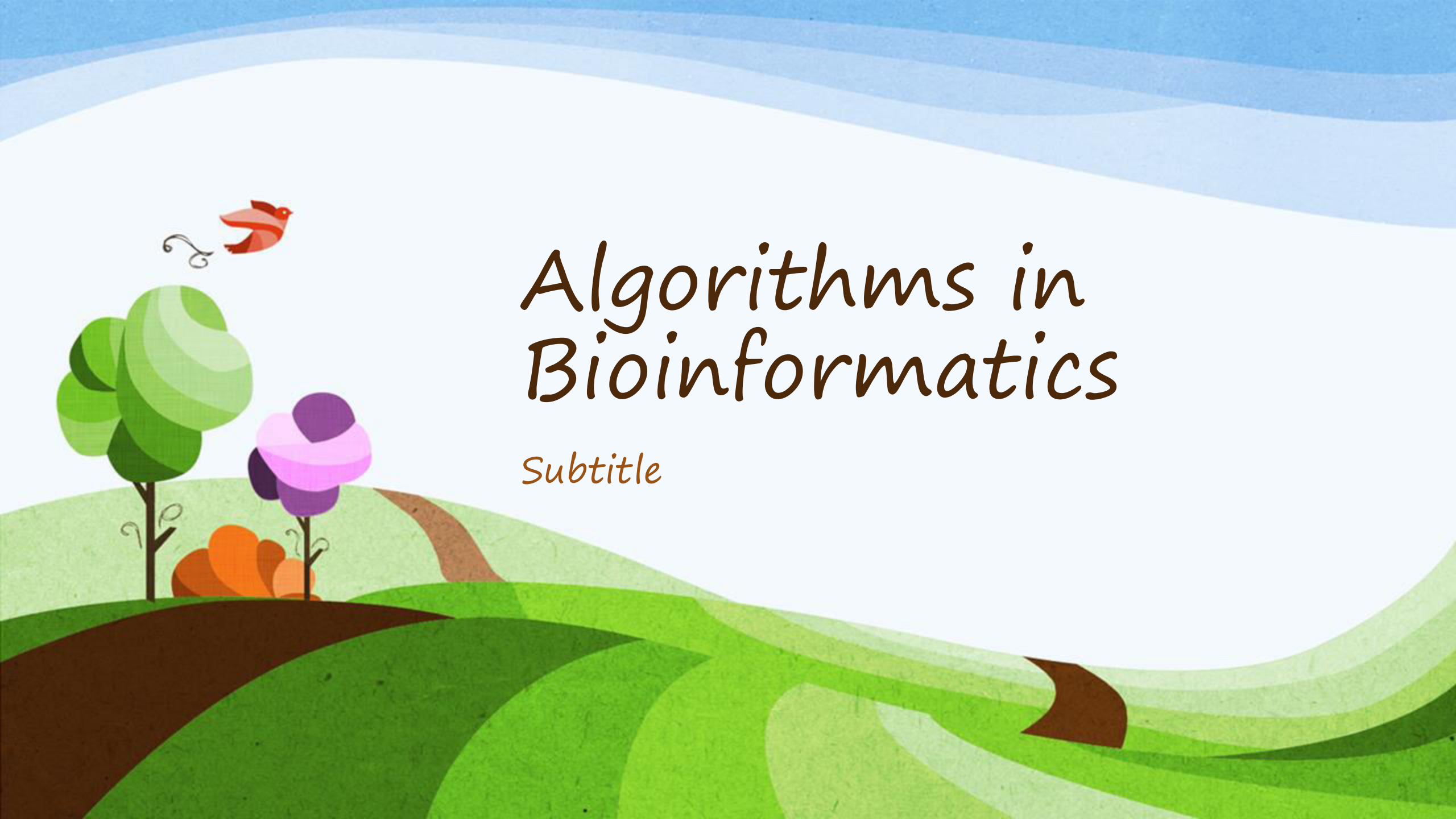


Algorithms in Bioinformatics

Subtitle

- Baum–Welch algorithm
- Binarization of consensus partition matrices
- BLAST
- Blast2GO
- Bowtie (sequence analysis)
- De novo sequence assemblers
- High-performance Integrated Virtual Environment
- Hirschberg's algorithm

Island algorithm

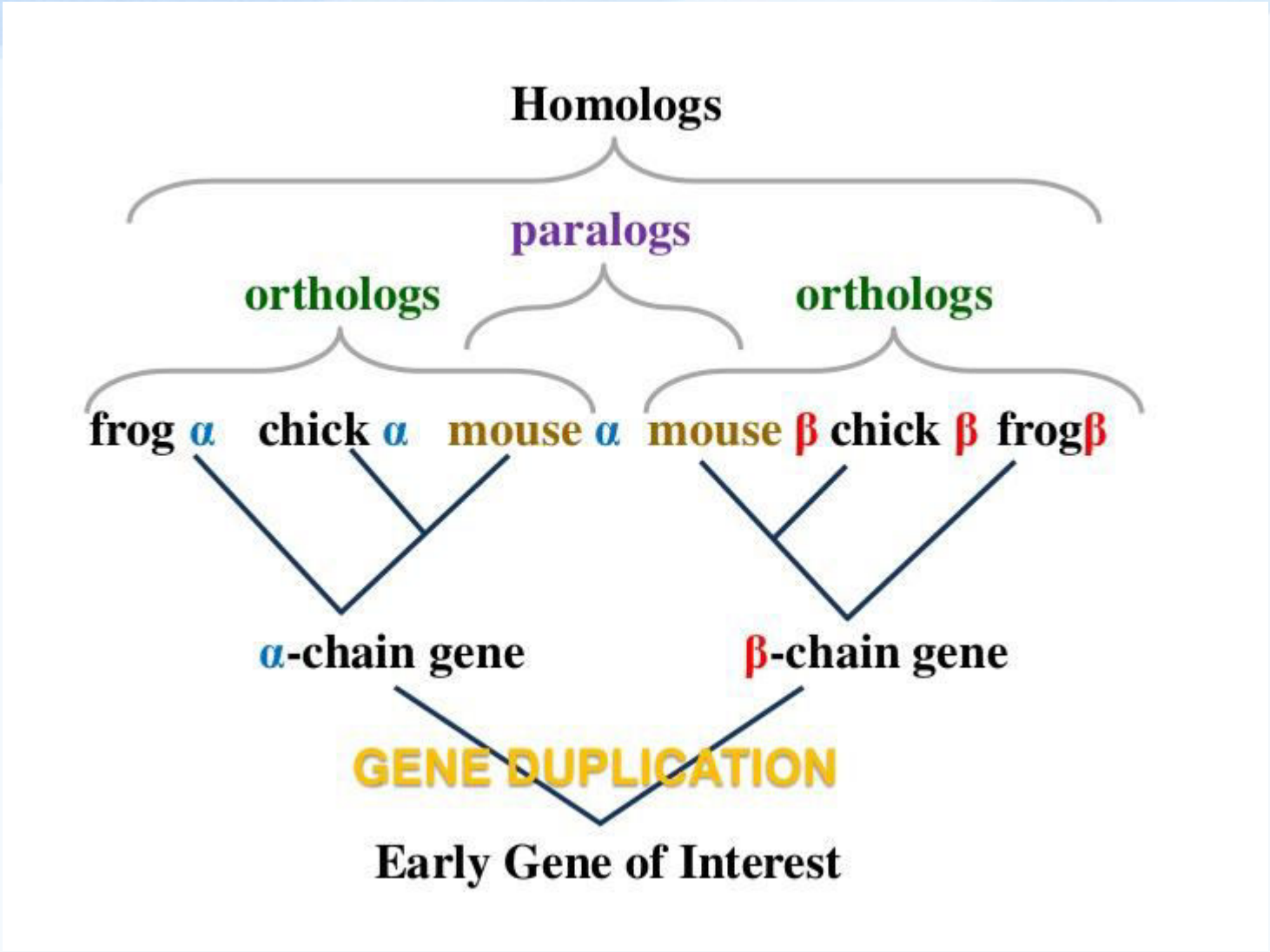
- Kabsch algorithm
- Microarray analysis techniques
- Needleman–Wunsch algorithm
- Neighbor joining
- Pairwise Algorithm
- Pseudo amino acid composition
- PSI Protein Classifier

Quartet distance

- Quasi-median networks
- Robinson–Foulds metric
- SAMtools
- SCHEMA (bioinformatics)
- Sequential pattern mining
- Smith–Waterman algorithm
- SPAdes (software)

TopHat (bioinformatics)

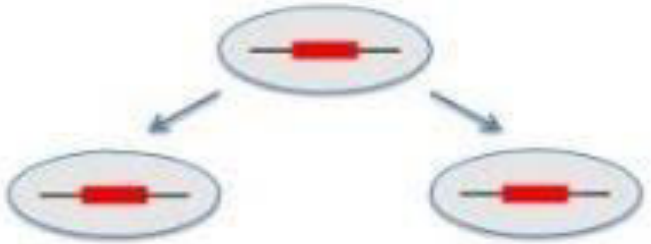
- UCLUST
- Ukkonen's algorithm
- UPGMA
- Velvet assembler
- ViennaRNA Package
- WPGMA
- Z curve


Homologs
paralogs
orthologs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frogβ
α-chain gene
β-chain gene
GENE DUPLICATION
Early Gene of Interest

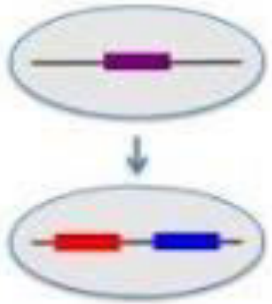
Homology, orthology, paralogy

Homology: descent from a common ancestor

Orthology: descent from a common ancestor by genome division



Paralogy: descent from a common ancestor by duplication within a genome



Global and Local Alginment

- Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. A general global alignment technique is the Needleman–Wunsch algorithm, which is based on dynamic programming. (BLAST & FASTA)
- Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith–Waterman algorithm is a general local alignment method also based on dynamic programming.

Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

Global versus Local Alignment

- A **global** alignment covers the **entire lengths of the sequences** involved

  The **Needleman-Wunsch** algorithm finds the **best global alignment** between 2 sequences

- A **local** alignment only covers **parts of the sequences**

  The **Smith-Waterman** algorithm finds the **best local alignment** between 2 sequences

Global alignment

```
  Q K E S G P S S S Y C
  |   | | |           |
V Q Q E S G L V R T T C
```

Local alignment

```
    E S G
    | | |
    E S G
```

Pair wise and Multiple sequence alignment

- Pairwise sequence alignment methods are used to find the best-matching piecewise (local or global) alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time.
- Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set. Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related.

Pairwise versus Multiple Alignment

- So far we have considered the alignment of two sequences ('**pairwise alignment**')

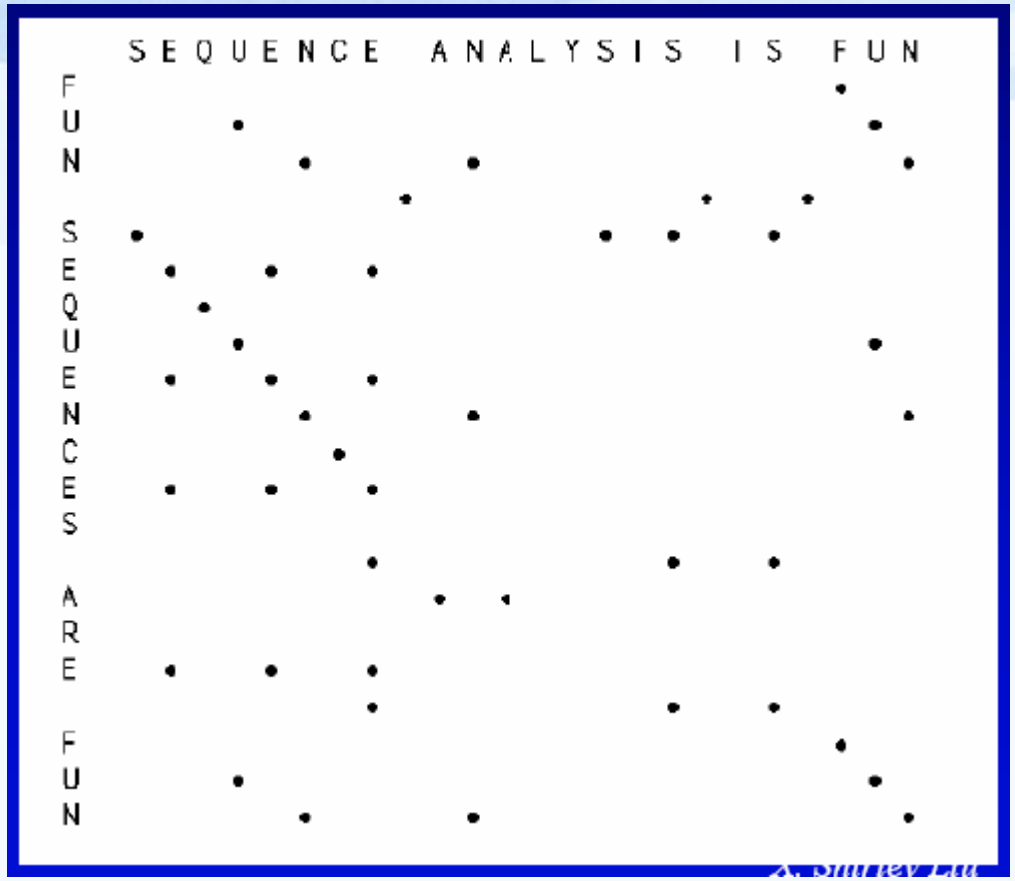
```
  Q K E S G P S S S Y C
  |   | | |             |
V Q Q E S G L V R T T C
```

- Alignment can be performed between three or more sequences ('**multiple alignment**')

```
  Q K E S G P S S S Y C
  |   | | |             |
V Q Q E S G L V R T T C
  |   | |   | | |   | |
V Q K E S L L V R S T C
```

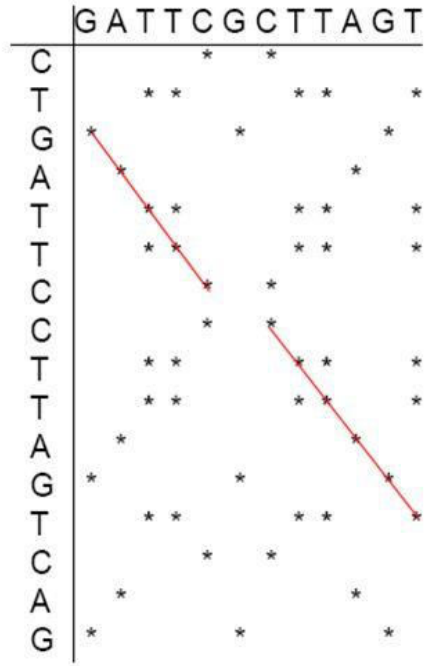
Dot-matrix

- To construct a dot-matrix plot, the two sequences are written along the top row and leftmost column of a two-dimensional matrix and a dot is placed at any point where the characters in the appropriate columns match—this is a typical recurrence plot. Some implementations vary the size or intensity of the dot depending on the degree of similarity of the two characters, to accommodate conservative substitutions.



Dot Matrices (cont'd)

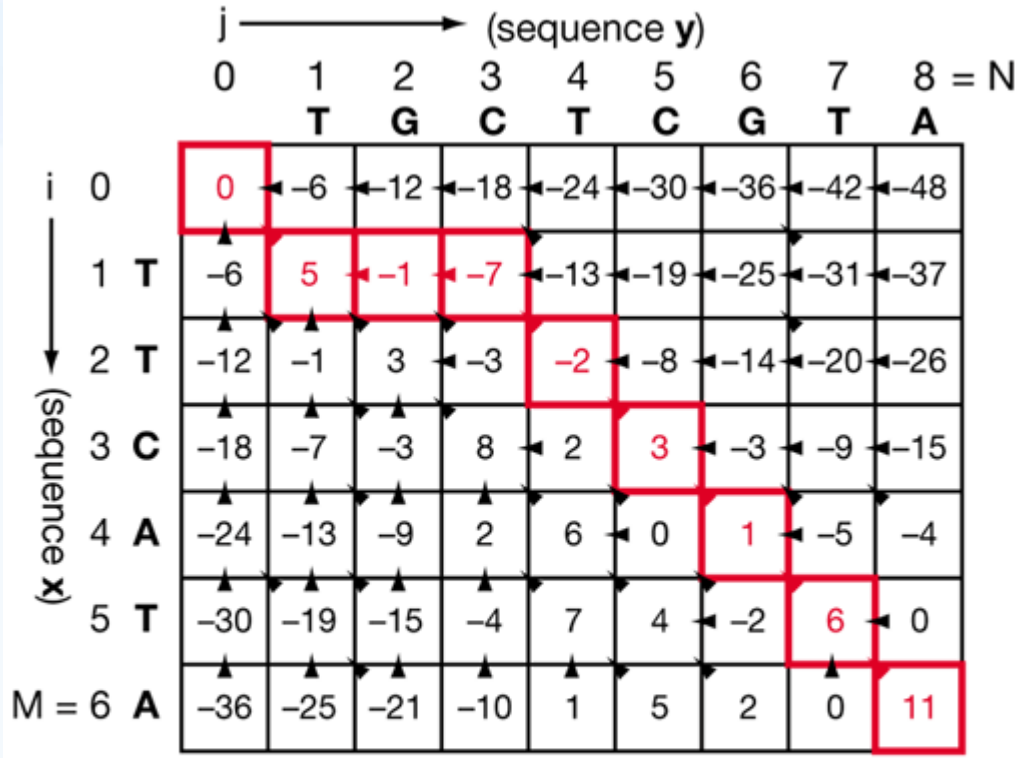
- Identify diagonals above a threshold length
- Diagonals in the dot matrix indicate exact substring matching



Dynamic programming

The technique of dynamic programming can be applied to produce global alignments via the Needleman–Wunsch algorithm, and local alignments via the Smith–Waterman algorithm. In typical usage, protein alignments use a substitution matrix to assign scores to amino-acid matches or mismatches, and a gap penalty for matching an amino acid in one sequence to a gap in the other.

Dynamic programming matrix:



Optimum alignment scores 11:

T	**–**	**–**	**T**	**C**	**A**	**T**	**A**
T	**G**	**C**	**T**	**C**	**G**	**T**	**A**
+5	−6	−6	+5	+5	−2	+5	+5

Substitution Matrix – PAM & BLOSUM

PAM

- One of the first amino acid substitution matrices, the PAM (Point Accepted Mutation) matrix was developed by Margaret Dayhoff in the 1970s.
- This matrix is calculated by observing the differences in closely related proteins. The PAM1 matrix estimates what rate of substitution would be expected if 1% of the amino acids had changed. There are many PAM matrices like PAM 1 to PAM30)

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

BLOSUM (BLOck SUbstitution Matrix)

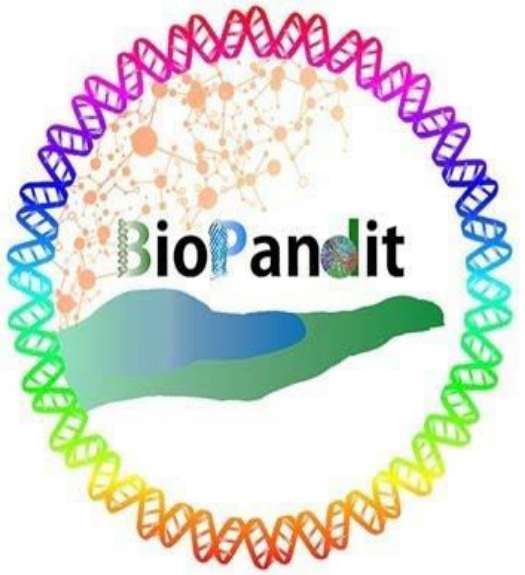
Henikoff constructed these matrices using multiple alignments of evolutionarily divergent proteins.

The probabilities used in the matrix calculation are computed by looking at "blocks" of conserved sequences found in multiple protein alignments.

These conserved sequences are assumed to be of functional importance within related proteins and will therefore have lower substitution rates than less conserved regions. There are many BLOSUM (1-70)

BLOSUM Substitution Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7	-3	-3	-3	-1	-2	-2	0	-3	-3	-3	-1	-2	-4	-1	2	0	-5	-4	-1
R	-3	9	-1	-3	-6	1	-1	-4	0	-5	-4	3	-3	-5	-3	-2	-2	-5	-4	-4
N	-3	-1	9	2	-5	0	-1	-1	1	-6	-6	0	-4	-6	-4	1	0	-7	-4	-5
D	-3	-3	2	10	-7	-1	2	-3	-2	-7	-7	-2	-6	-6	-3	-1	-2	-8	-6	-6
C	-1	-6	-5	-7	13	-5	-7	-6	-7	-2	-3	-6	-3	-4	-6	-2	-2	-5	-5	-2
Q	-2	1	0	-1	-5	9	3	-4	1	-5	-4	2	-1	-5	-3	-1	-1	-4	-3	-4
E	-2	-1	-1	2	-7	3	8	-4	0	-6	-6	1	-4	-6	-2	-1	-2	-6	-5	-4
G	0	-4	-1	-3	-6	-4	-4	9	-4	-7	-7	-3	-5	-6	-5	-1	-3	-6	-6	-6
H	-3	0	1	-2	-7	1	0	-4	12	-6	-5	-1	-4	-2	-4	-2	-3	-4	3	-5
I	-3	-5	-6	-7	-2	-5	-6	-7	-6	7	2	-5	2	-1	-5	-4	-2	-5	-3	4
L	-3	-4	-6	-7	-3	-4	-6	-7	-5	2	6	-4	3	0	-5	-4	-3	-4	-2	1
K	-1	3	0	-2	-6	2	1	-3	-1	-5	-4	8	-3	-5	-2	-1	-1	-6	-4	-4
M	-2	-3	-4	-6	-3	-1	-4	-5	-4	2	3	-3	9	0	-4	-3	-1	-3	-3	1
F	-4	-5	-6	-6	-4	-5	-6	-6	-2	-1	0	-5	0	10	-6	-4	-4	0	4	-2
P	-1	-3	-4	-3	-6	-3	-2	-5	-4	-5	-5	-2	-4	-6	12	-2	-3	-7	-6	-4
S	2	-2	1	-1	-2	-1	-1	-1	-2	-4	-4	-1	-3	-4	-2	7	2	-6	-3	-3
T	0	-2	0	-2	-2	-1	-2	-3	-3	-2	-3	-1	-1	-4	-3	2	8	-5	-3	0
W	-5	-5	-7	-8	-5	-4	-6	-6	-4	-5	-4	-6	-3	0	-7	-6	-5	16	3	-5
Y	-4	-4	-4	-6	-5	-3	-5	-6	3	-3	-2	-4	-3	4	-6	-3	-3	3	11	-3
V	-1	-4	-5	-6	-2	-4	-4	-6	-5	4	1	-4	1	-2	-4	-3	0	-5	-3	7



	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	

(a)

Target sequence	A	L	E	D	N
Functionally Identical or Related Enzymes (property A proteins)	H	L	E	P	L
Sequence-related Proteins (~A proteins)	L	L	P	D	L
Score	-1	0	6	-7	0

(b)