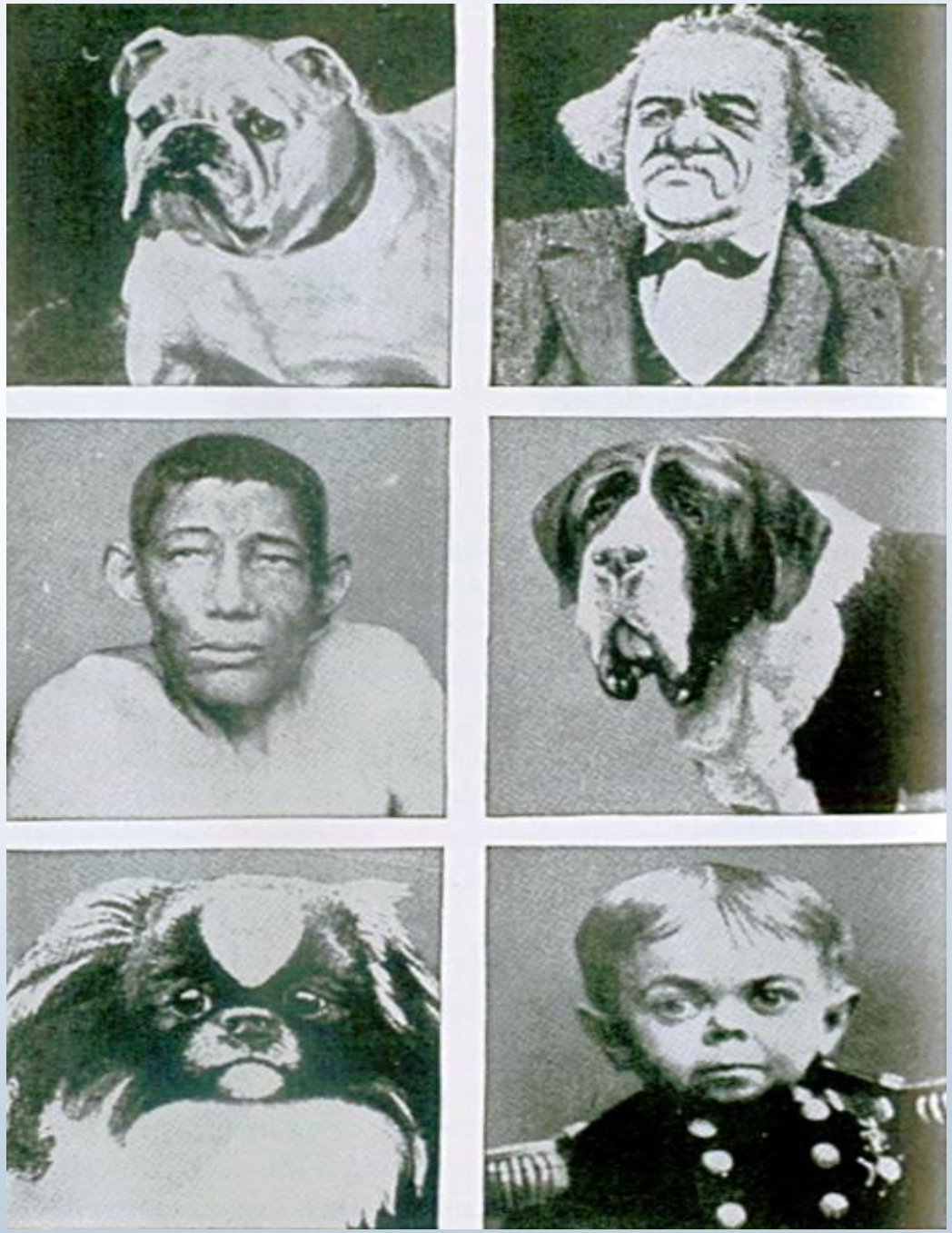
Alignment

Yazdan Asgari

2019

Sequence Alignment



Pairwise sequence alignment, Jonathan Pevsner

Complete DNA Sequences

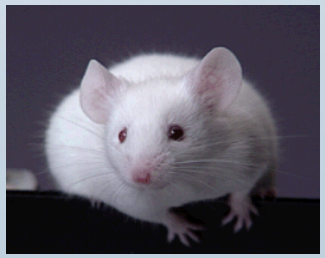





More than 400
complete genomes
have been sequenced

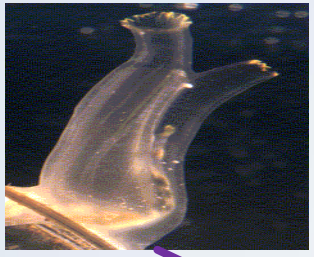




Pairwise sequence alignment, Jonathan Pevsner

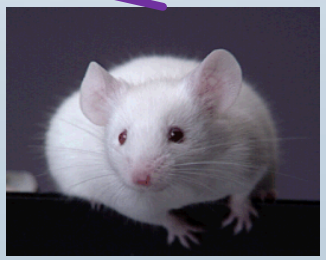
Evolution based on Sequence Alignment









Pairwise sequence alignment, Jonathan Pevsner

Sources of variation

- Nucleotide substitution
  - Replication error
  - Chemical reaction
- Insertions or deletions (indels)
  - Unequal crossing over
  - Replication slippage
- Duplication
  - a single gene (complete gene duplication)
  - part of a gene (internal or partial gene duplication)
    - Domain duplication
    - Exon shuffling
  - part of a chromosome (partial polysomy)
  - an entire chromosome (aneuploidy or polysomy)
  - the whole genome (polyploidy)

Homolog

A gene related to a second gene by descent from a common ancestral DNA sequence

Ortholog

Genes in different species that evolved from a common ancestral gene by speciation

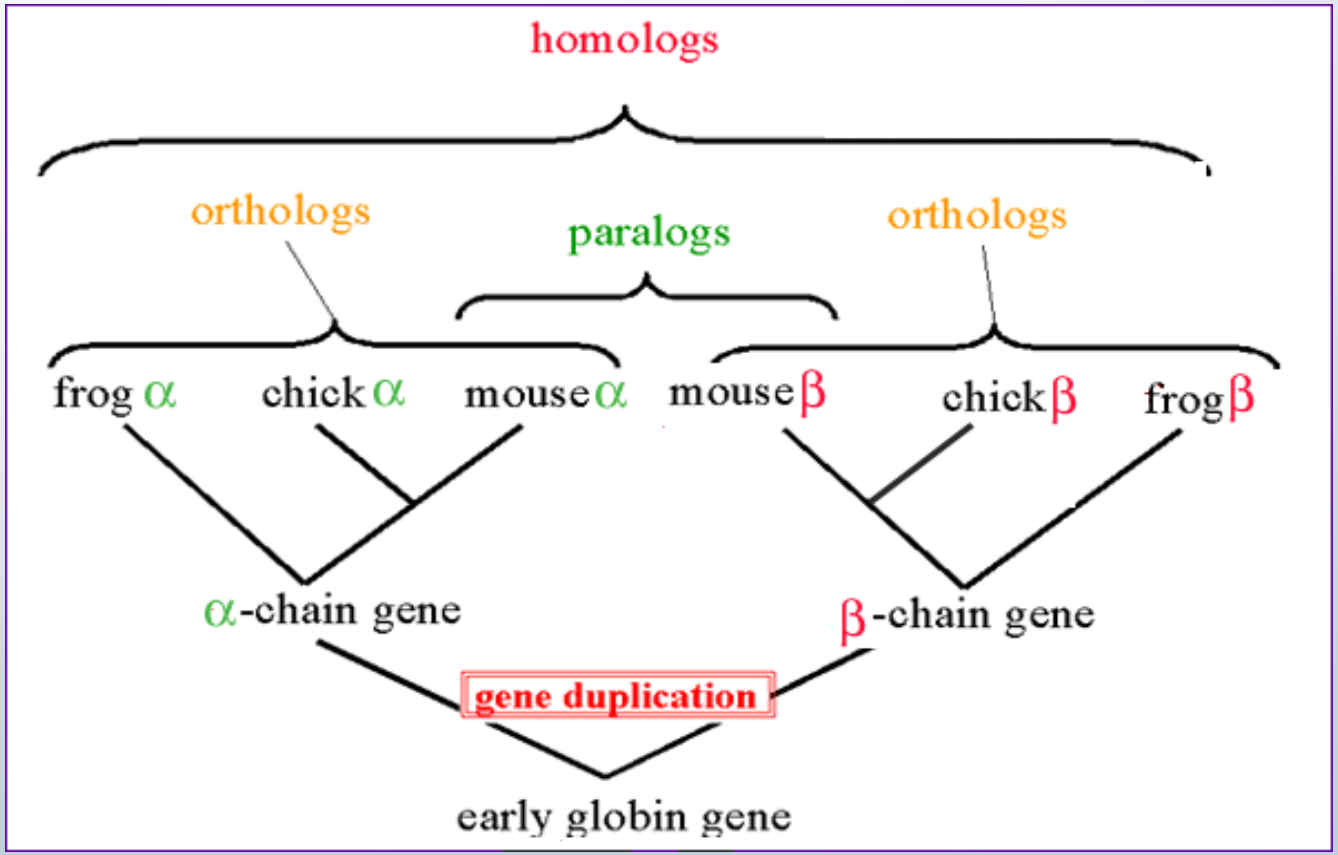
Paralog

Genes related by duplication within a genome

Orthologs retain the same function in the course of evolution, whereas Paralogs evolve new functions, even if these are related to the original one.

Orthologs are homologous genes that are the result of a **speciation event**.

Paralogs are homologous genes that are the result of a **duplication event**.



Pairwise sequence alignment, Jonathan Pevsner

Common mutations in DNA

Substitution:

A C G T T G A C

A C G A T G A C

Deletion:

A C G T T G A C

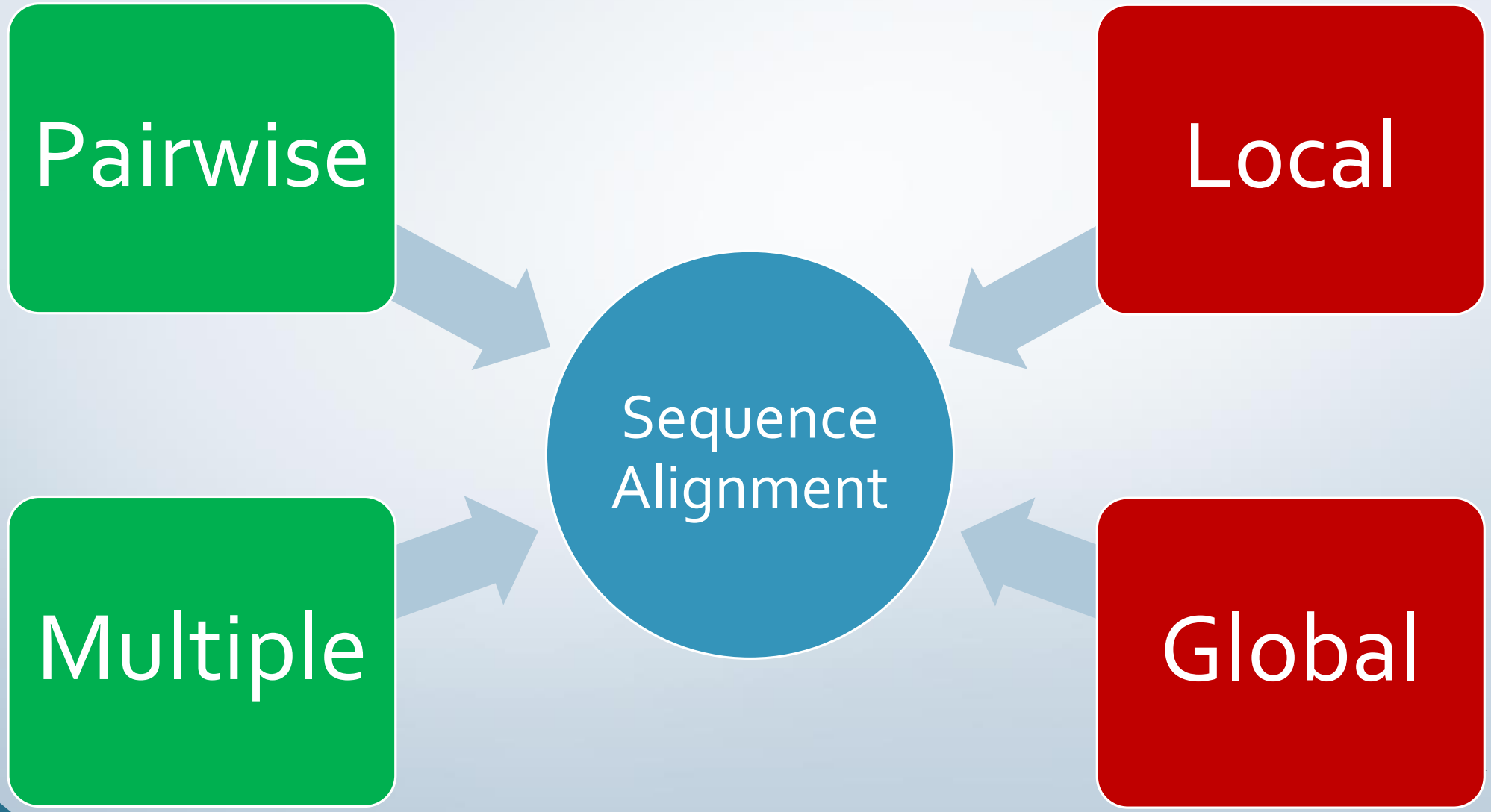
A C G A C

Insertion:

A C G T T G A C

A C G C A A T T G A C

Sequence Alignment

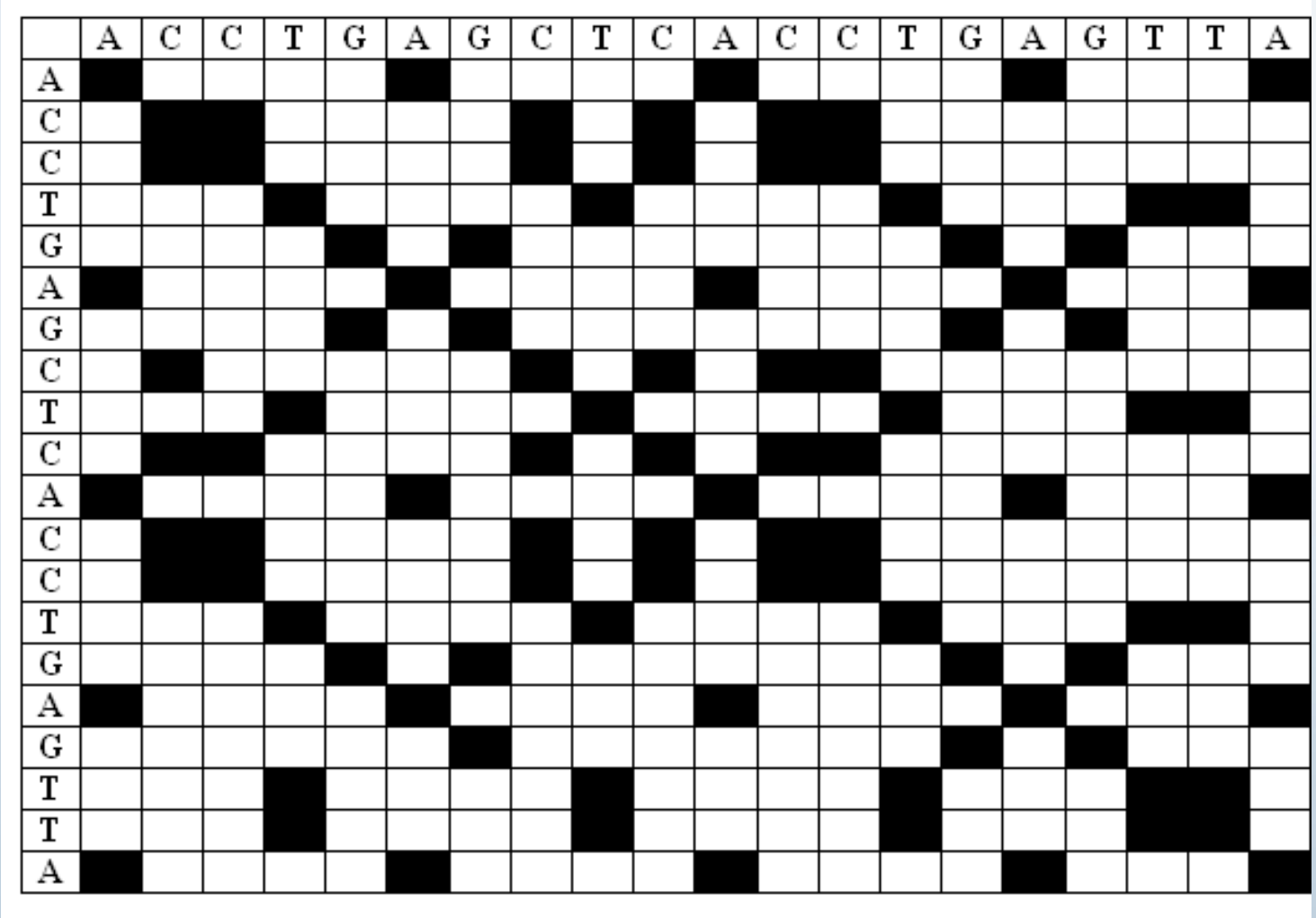


Visual Alignments

Percent Identity as a Measure for Quantifying Sequence Similarity

- Identity is the number of identical bases or amino acids matched between two aligned sequences
- Percent identity is obtained by dividing this number by the total length of the aligned sequences and multiplying by 100
- Sequence similarity based on identity is usually visualized using the dot-plot representation

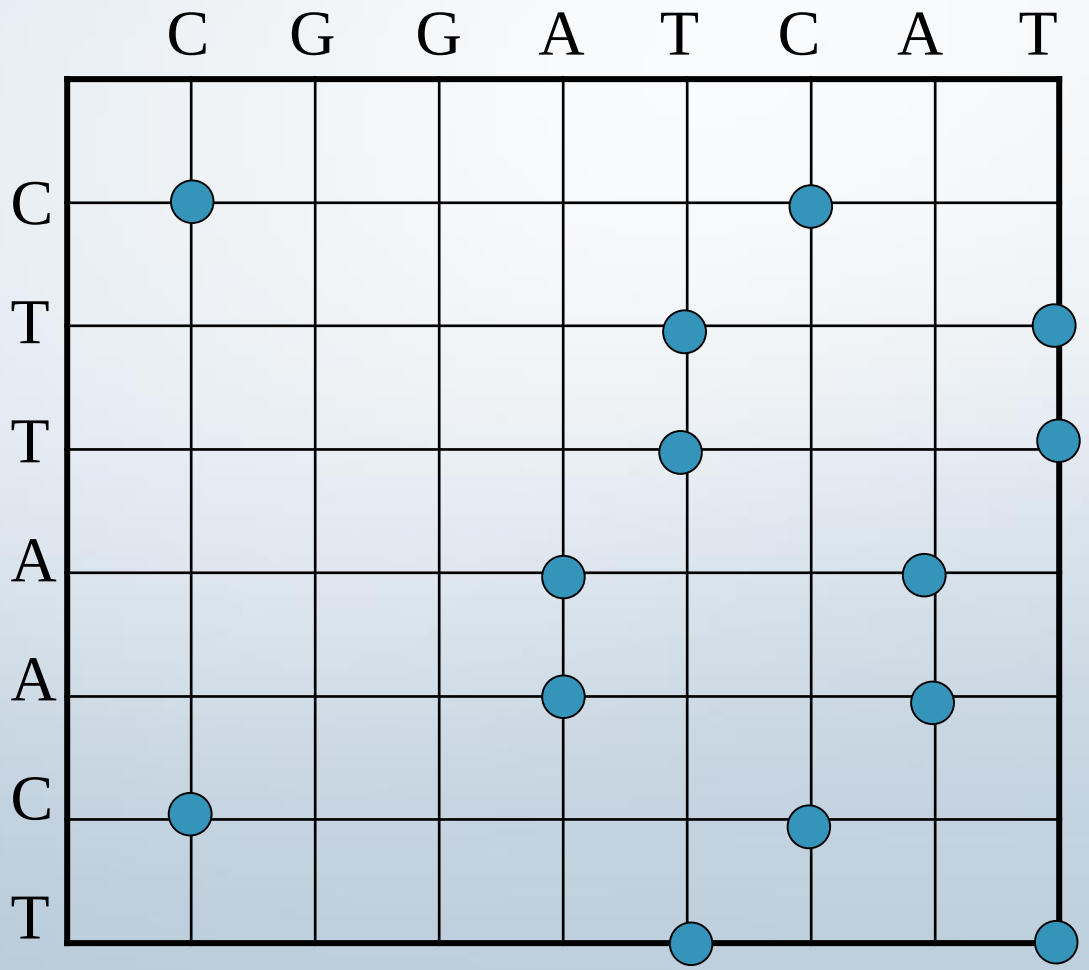
Example Dot Plot



Dot Matrix

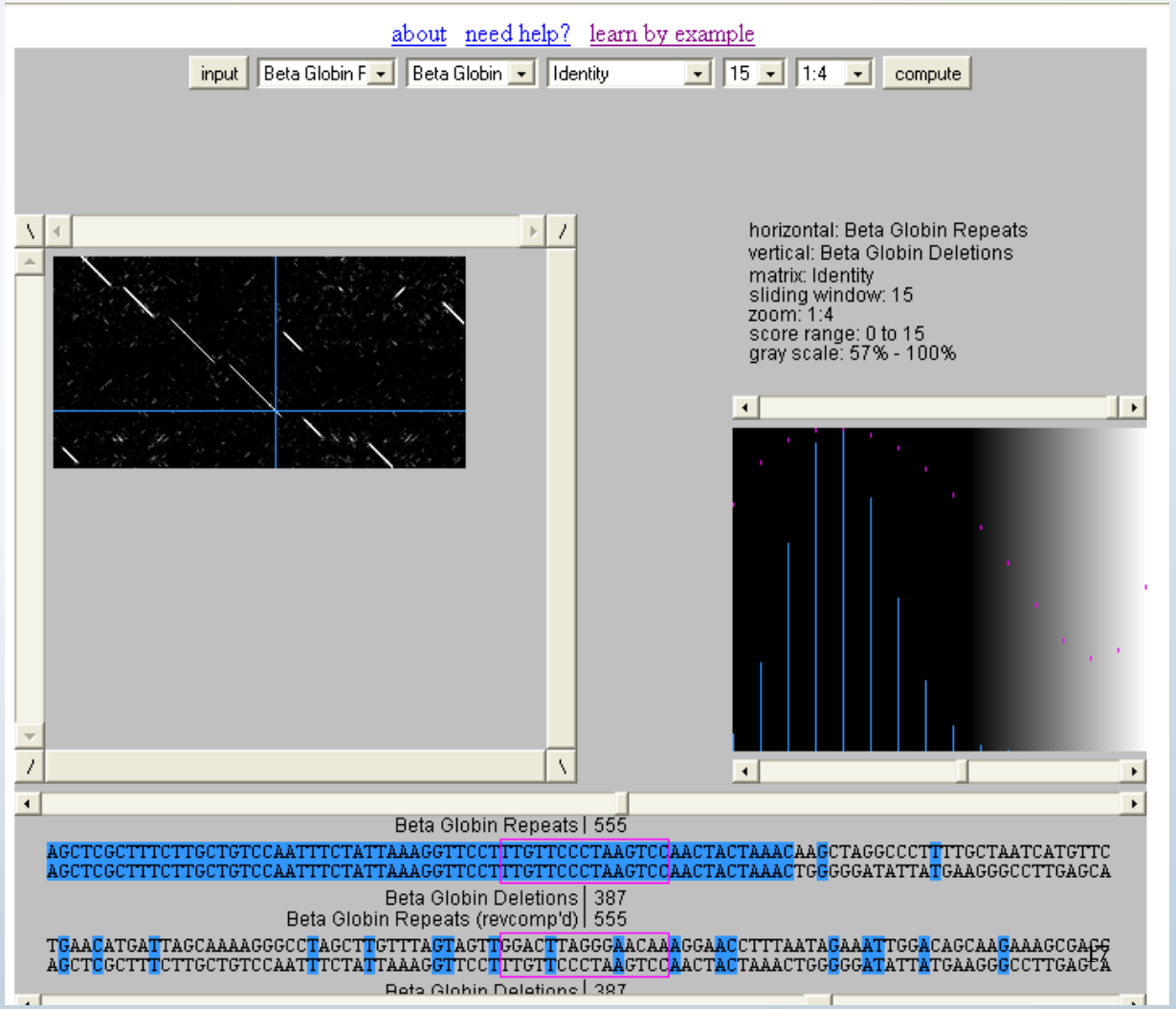
Sequence A：CTTAACT

Sequence B：CGGATCAT



Available Dot Plot Programs

- Dotlet (Java Applet)
- Dotter
- Compare
- DotPlot+
- DNA strider
- PipMaker



http://myhits.isb-sib.ch/cgi-bin/dotlet

Scoring Alignment

Sequence A: CTTAACT

Sequence B: CGGATCAT

An alignment of A and B:

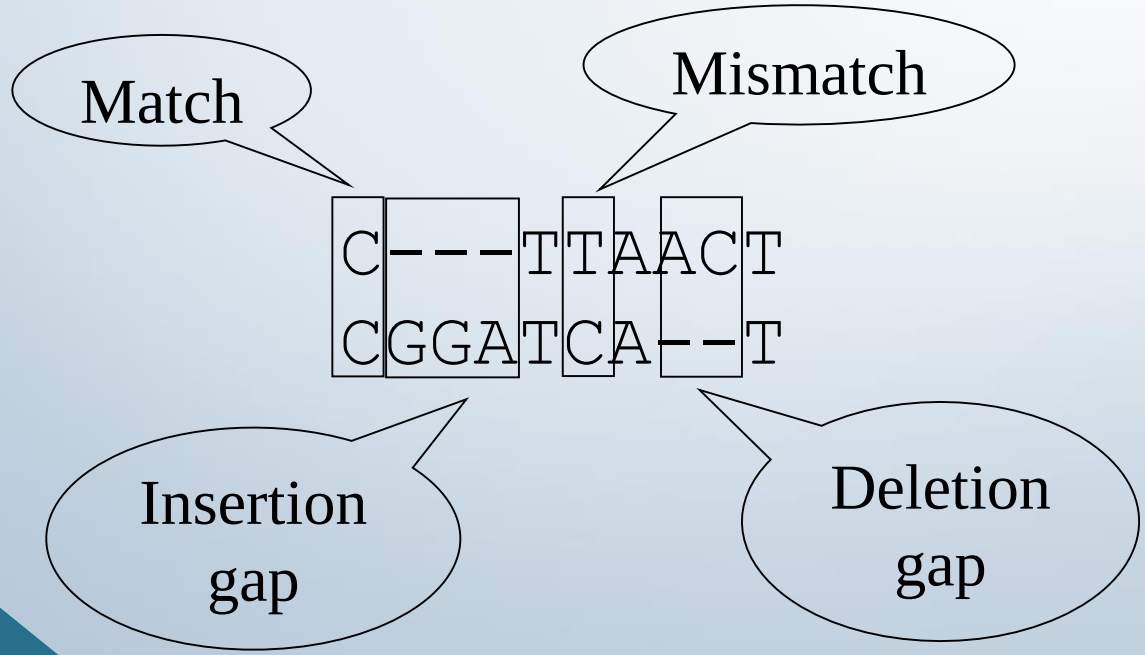
```
C---TTAACT  <---- Sequence A
CGGATCA--T  <---- Sequence B
```

Pairwise Alignment

Sequence A: CTTAACT

Sequence B: CGGATCAT

An alignment of A and B:



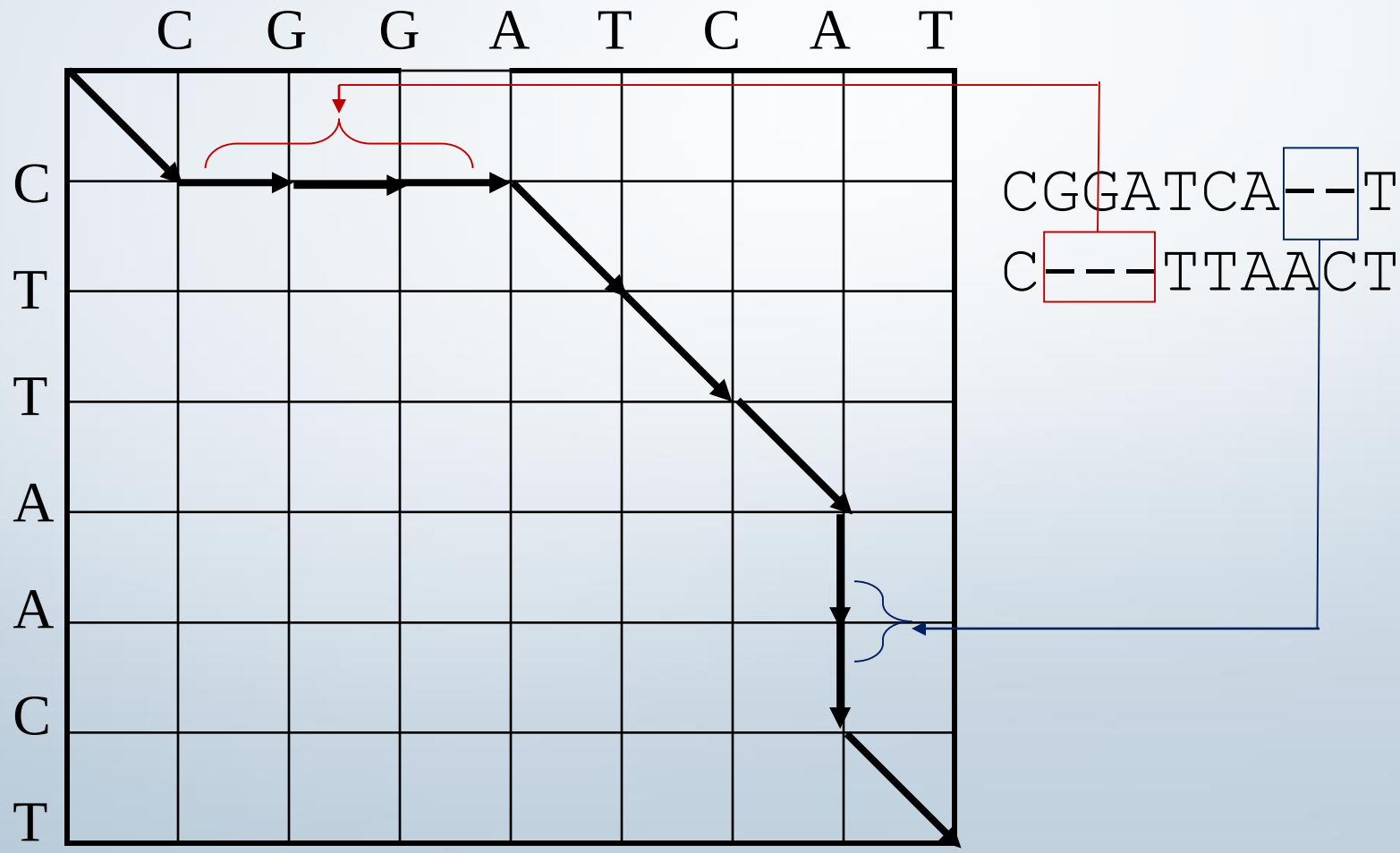
Alignment Graph (another example)

Sequence A: CGGATCAT

Sequence B: CTTAACT

Note:

- For **writing the alignment**, begin from **upper left** and follow the arrows.



A simple scoring scheme

- Match: +8 ($w(x, y) = 8$, if $x = y$)
- Mismatch: -5 ($w(x, y) = -5$, if $x \neq y$)
- Each gap symbol: -3 ($w(-,x) = w(x,-) = -3$)

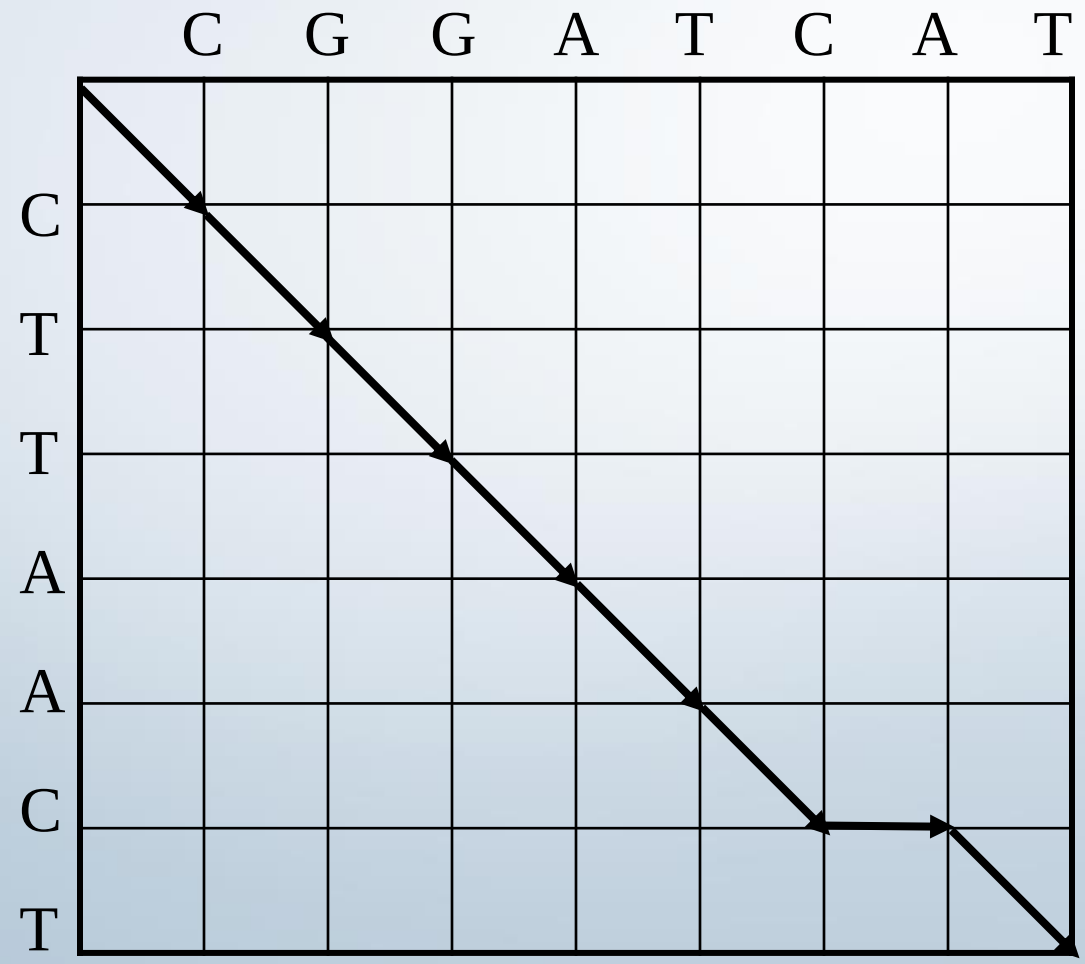
C	–	–	–	T	T	A	A	C	T	
C	G	G	A	T	C	A	–	–	T	
+8	-3	-3	-3	+8	-5	+8	-3	-3	+8	= +12

Alignment score

Alignment Graph (another example)

Note:

- For **writing the alignment**, begin from **upper left** and follow the arrows.

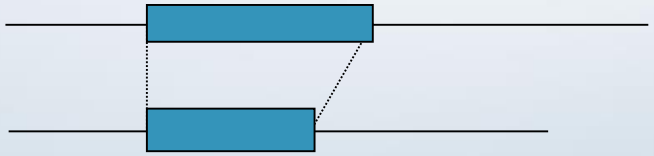



```
C G G A T C A T
C T T A A C - T
```

Global Alignment vs. Local Alignment

- global alignment:

- local alignment:



Global Alignment

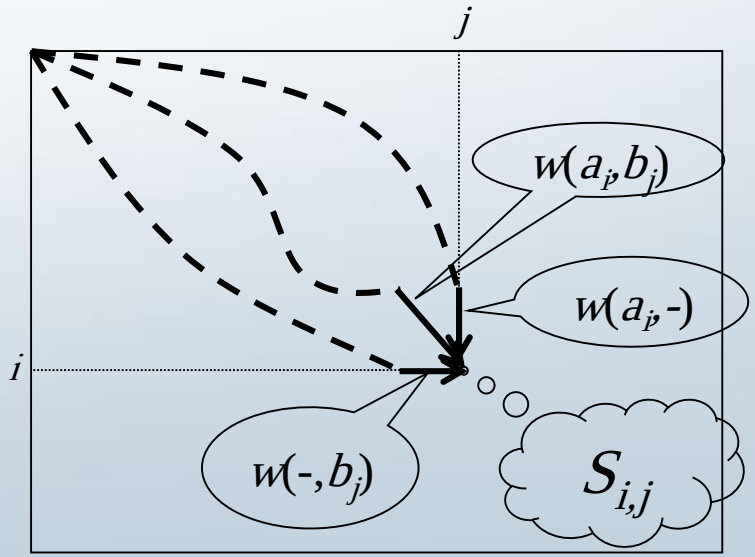
- Compares total length of two sequences
- Needleman, S.B. and Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48(3):443-53(1970).

An optimal alignment

the alignment of maximum score

- Let A $=a_1a_2\ldots a_m$ and $B=b_1b_2\ldots b_n$.
- $S_{i,j}$: the score of an optimal alignment between $a_1a_2\ldots a_i$ and $b_1b_2\ldots b_j$
- With proper initializations, $S_{i,j}$ can be computed as follows.

$$s_{i,j} = \max\begin{cases} s_{i-1,j} + w(a_i,-) \\ s_{i,j-1} + w(-,b_j) \\ s_{i-1,j-1} + w(a_i,b_j) \end{cases}$$



Initializations

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3								
T	-6								
T	-9								
A	-12								
A	-15								
C	-18								
T	-21								

$S_{3,5} = ?$

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5	2	-1	-4	-7	-10	-13
T	-6	5	3	0	-3	7	4	1	-2
T	-9	2	0	-2	-5	?			
A	-12								
A	-15								
C	-18								
T	-21								

$$S_{3,5} = 5$$

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5	2	-1	-4	-7	-10	-13
T	-6	5	3	0	-3	7	4	1	-2
T	-9	2	0	-2	-5	5	-1	-1	9
A	-12	-1	-3	-5	6	3	0	7	6
A	-15	-4	-6	-8	3	1	-2	8	5
C	-18	-7	-9	-11	0	-2	9	6	3
T	-21	-10	-12	-14	-3	8	6	4	14*

optimal score

```
C G G A T C A T
C T T A A C – T
8 –5 –5 +8 -5 +8 -3 +8 = 14
```

Note:

- For **finding path**, begin from **lower right**.
- For **writing the alignment**, begin from **upper left**.

		C	G	G	A	T	C	A	T
	0	-3	-6	-9	-12	-15	-18	-21	-24
C	-3	8	5	2	-1	-4	-7	-10	-13
T	-6	5	3	0	-3	7	4	1	-2
T	-9	2	0	-2	-5	5	-1	-1	9
A	-12	-1	-3	-5	6	3	0	7	6
A	-15	-4	-6	-8	3	1	-2	8	5
C	-18	-7	-9	-11	0	-2	9	6	3
T	-21	-10	-12	-14	-3	8	6	4	14

Local Alignment

- Compares segments of sequences
- Finds cases when one sequence is a part of another sequence, or they only match in parts.
- Smith, T.F. and Waterman, M.S. Identification of common molecular subsequences. J Mol Biol. 147(1):195-7 (1981)

An optimal local alignment

- $S_{i,j}$: the score of an optimal local alignment ending at a_i and b_j
- With proper initializations, $S_{i,j}$ can be computed as follows.

$$s_{i,j} = \max \begin{cases} 0 \\ s_{i-1,j} + w(a_i, -) \\ s_{i,j-1} + w(-, b_j) \\ s_{i-1,j-1} + w(a_i, b_j) \end{cases}$$

local alignment

Match: 8

Mismatch: -5

Gap symbol: -3

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	8	5	2	0	0	8	5	2
T	0	5	3	0	0	8	5	3	13
T	0	2	0	0	0	8	5	2	11
A	0	0	0	0	8	5	3	?	
A	0								
C	0								
T	0								

local alignment

Match: 8

Mismatch: -5

Gap symbol: -3

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	8	5	2	0	0	8	5	2
T	0	5	3	0	0	8	5	3	13
T	0	2	0	0	0	8	5	2	11
A	0	0	0	0	8	5	3	13	10
A	0	0	0	0	8	5	2	11	8
C	0	8	5	2	5	3	13	10	7
T	0	5	3	0	2	13	10	8	18

Note:

- For **finding path**, begin from **lower right**.
- For **writing the alignment**, begin from **upper left**.

		C	G	G	A	T	C	A	T
	0	0	0	0	0	0	0	0	0
C	0	8	5	2	0	0	8	5	2
T	0	5	3	0	0	8	5	3	13
T	0	2	0	0	0	8	5	2	11
A	0	0	0	0	8	5	3	13	10
A	0	0	0	0	8	5	2	11	8
C	0	8	5	2	5	3	13	10	7
T	0	5	3	0	2	13	10	8	18

```
A T C A T
A A C - T
8-5+8-3+8 = 16

- A T C A T
A - A C - T
-3-3-5+8-3+8 = 2
```

Moving Backward from here

Affine gap penalties

- Match: +8 ($w(x, y) = 8$, if $x = y$)
- Mismatch: -5 ($w(x, y) = -5$, if $x \neq y$)
- Each gap symbol: -3 ($w(-,x) = w(x,-) = -3$)
- Each gap is charged an extra gap-open penalty: -4 (-4 for gaps greater than 1)

		-4						-4		
C	–	–	–	T	T	A	A	C	T	
C	G	G	A	T	C	A	–	–	T	
+8	-3	-3	-3	+8	-5	+8	-3	-3	+8	= +12

Alignment score: 12 – 4 – 4 = 4

Gap Penalties Types

- Affine gap penalties
- Constant gap penalties
- Restricted affine gap penalties

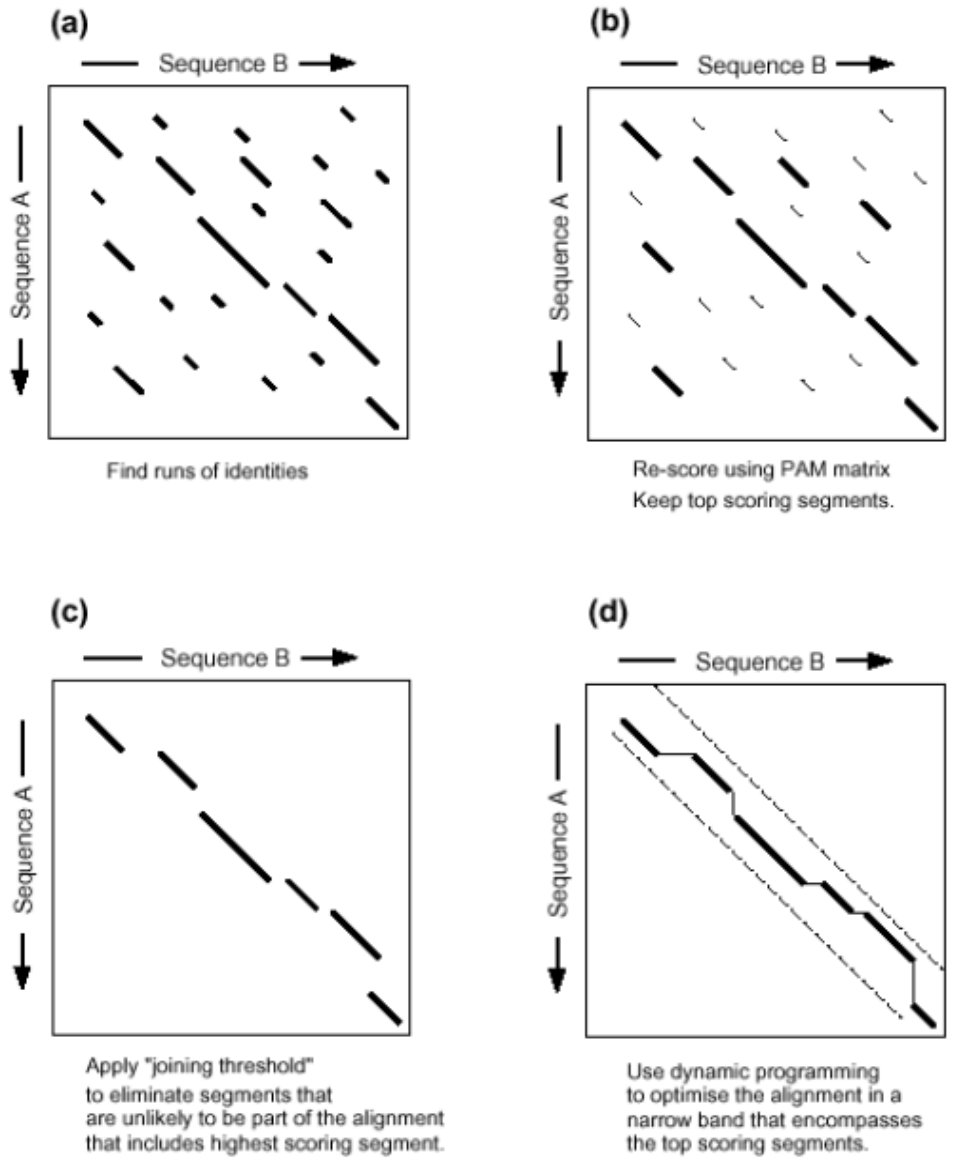
BLAST & FASTA

Heuristic search algorithms 1

FASTA (Pearson 1995)

Uses heuristics to avoid calculating the full dynamic programming matrix

Speed up searches by an order of magnitude compared to full Smith-Waterman



https://www.ctu.edu.vn/~dvxe/Bioinformatic/computer/sim-fasta.html

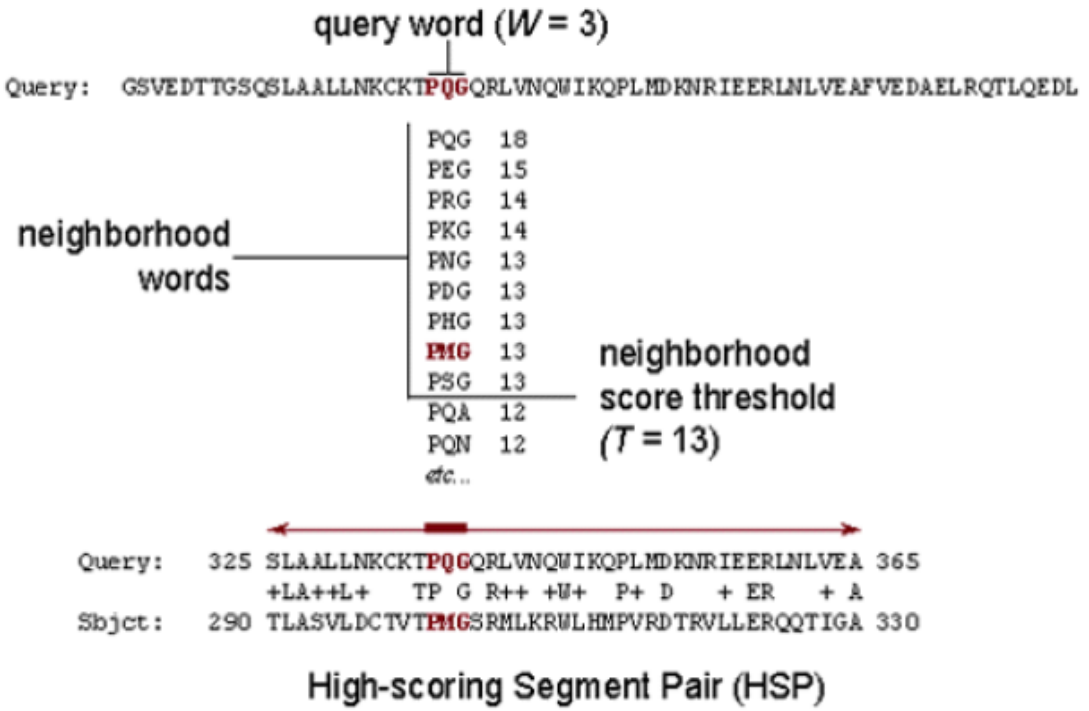
Heuristic search algorithms 2

BLAST (Altschul 1990, 1997)

Uses rapid word lookup methods to completely skip most of the database entries

Extremely fast:
- *One order of magnitude faster than FASTA*
- *Two orders of magnitude faster than Smith-Waterman*

Almost as sensitive as FASTA, but the statistical side of FASTA is still stronger than BLAST



https://www.ctu.edu.vn/~dvxe/Bioinformatic%20course/mod4/mod4_2.html

BLAST

- ✓ Basic Local Alignment Search Tool
  (by Altschul, Gish, Miller, Myers and Lipman)
- ✓ The central idea of the BLAST algorithm is that a statistically significant alignment is likely to contain a *high-scoring pair* of aligned words.

1) Build the hash table for Sequence A.
2) Scan Sequence B for hits.
3) Extend hits.

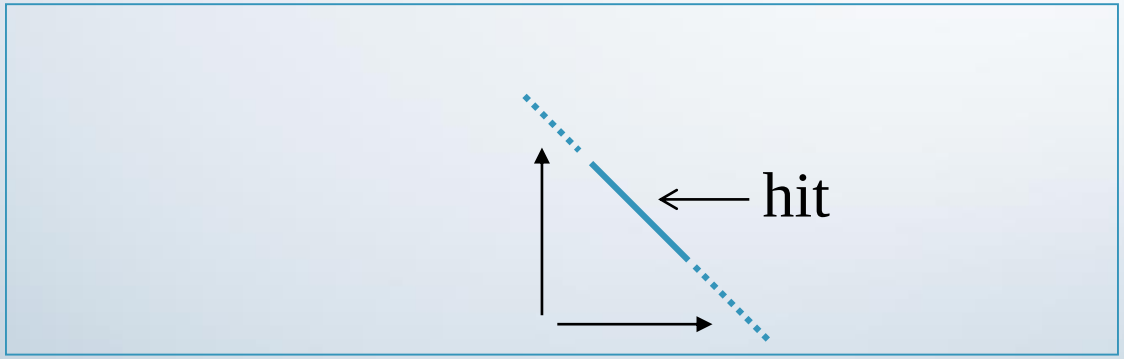
BLAST

Step2: Scan sequence B for hits.

BLAST

Step2: Scan sequence B for hits.

Step 3: Extend hits.



Terminate if the score of the extension fades away. (That is, when we reach a segment pair whose score falls a certain distance below the best score found for shorter extensions.)

Getting the most out of Alignment

1. What kind of Alignment?
2. Pick an appropriate database
3. Pick the right algorithm
4. Choose parameters

Substitution Matrices

Substitution Matrices

- PAM
  - Substitution matrix between amino-acids, no gaps.
  - Developed by Margaret Dayhoff and published in 1978
- BLOSUM
  - Substitution matrix between amino-acids.
  - Developed by Henikoff and Henikoff and published in 1992

PAM1 Matrix

- Constructed by Margaret Dayhoff
- Accepted point mutation
  - Replacement of one amino acid by another accepted by natural selection
  - Analyzed 1572 changes in 71 protein families
- One PAM
  - Unit of evolutionary divergence
  - Equivalent to distance where 1% of amino acids changed
  - Used protein families with > 85% identity

PAM1 and PAM250 for Phe -> X

X	PAM1	PAM250	X	PAM1	PAM250
Ala	0.0002	0.04	Leu	0.0013	0.13
Arg	0.0001	0.01	Lys	0.0000	0.02
Asn	0.0001	0.02	Met	0.0001	0.02
Asp	0.0000	0.01	Phe	0.9946	0.32
Cys	0.0000	0.01	Pro	0.0001	0.02
Gln	0.0000	0.01	Ser	0.0003	0.03
Glu	0.0000	0.01	Thr	0.0001	0.03
Gly	0.0001	0.03	Trp	0.0001	0.01
His	0.0002	0.02	Tyr	0.0021	0.15
Ile	0.0007	0.05	Val	0.0001	0.05

These are mutation probabilities!

PAM1 Matrix

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg	R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn	N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp	D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys	C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln	Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu	E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly	G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His	H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile	I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu	L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys	K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met	M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe	F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro	P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser	S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr	T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp	W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr	Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val	V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

PAM1 mutation probabilty matrix from the Dayhoff group. All values multipled by 10,000.

Extrapolating

- Matrix multiplication
  - Extrapolates long evolutionary distances
  - Based on short-term evolutionary changes
  - Assumes long-term changes extension of short-term changes
- Multiply PAM1 matrix by itself
  - Generate PAM2 matrix
  - Mathematical estimate of two units of evolutionary distance

PAM250 Matrix

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	**A**	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg	**R**	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn	**N**	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp	**D**	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys	**C**	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln	**Q**	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu	**E**	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly	**G**	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His	**H**	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile	**I**	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu	**L**	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys	**K**	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met	**M**	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe	**F**	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro	**P**	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser	**S**	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr	**T**	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	33	6
Trp	**W**	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr	**Y**	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val	**V**	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

A PAM250 matrix. Each column has been adjusted so that the columns sum to 100.

PAM250 Matrix

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

PAM250 Matrix

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	5	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

Bioinformatics: Life Sciences, Robert Lessick

PAM250 Matrix

	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	5	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

Scoring Matrix (Logodd Matrix)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	6	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Normalized Frequencies of Amino Acids

Normalized Frequencies of Amino Acids			
Ala	0.096	Asn	0.042
Gly	0.090	Pro	0.041
Lys	0.085	Ile	0.035
Leu	0.085	His	0.034
Val	0.078	Arg	0.034
Thr	0.062	Gin	0.032
Ser	0.057	Tyr	0.030
Asp	0.053	Cys	0.025
Glu	0.053	Met	0.012
Phe	0.045	Trp	0.012

***How often a given amino acid appears in a protein (determined by empirical analyses)*

$$S_{i,j} = 10 \times \log\left(\frac{q_{i,j}}{p_j}\right) \qquad S_{K,G} = 10 \times \log\left(\frac{0.06}{0.090}\right) = -1.76 \cong -2$$

Bioinformatics: Life Sciences, Robert Lessick

PAM250 Scoring 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

Bioinformatics: Life Sciences, Robert Lessick

For More Widely Divergent Sequences

- Matrices representing larger evolutionary distances may be derived from the PAM1 matrix by matrix multiplication.
- PAM250:
  - Corresponding to ~20% identity
  - The lowest sequence similarity for which we can hope to produce a correct alignment

PAM	0	30	80	110	200	250
% identity	100	75	50	60	25	20

BLOSUM

- Uses more data => more reliable stats
- Uses local alignment of multiple proteins of the same family as opposed to global alignment that is used in PAM => collects the stats from the regions that matter
- BLOSUM-X – constructed from proteins that are at most X% similar

BLOSUM Matrices

- PAM matrices were based on only a small number of observed substitutions (~1500)
- Perform best in identifying distant relationships
- BLOCKS database (BLOcks Subsitution Matrix)
- Regions of closely-related proteins alignable without gaps
- BLOSUM62 ≈ PAM150
- BLOSUM50 ≈ PAM250

Position-Specific Iterated BLAST

- Instead of 20 × 20 matrix, use m × 20 matrix, where m is the length of the query
- First, run normal BLAST
- Using the results, construct the position specific score matrix
- Next iterations use global alignment and the position specific score matrix

PAM1 DNA log-odds matrices

A. Model of uniform mutation rates among nucleotides.

	A	**G**	**T**	**C**
A	2			
G	-6	2		
T	-6	-6	2	
C	-6	-6	-6	2

B. Model of 3-fold higher transitions than transversions.

	A	**G**	**T**	**C**
A	2			
G	-5	2		
T	-7	-7	2	
C	-7	-7	-5	2

E-Values

Points

What does it mean if I said I scored 10 points in today's game?

- Depends on the game!
- In Ping-Pong, that's not a big accomplishment
- In Basketball, that's pretty good
- In soccer, that's almost unheard of

Number of Events

- If a lottery has a probability of one in 10 million
  - You will probably not win with one ticket
- If 20 million tickets are sold
  - There will likely be a winner
  - Expected number of winners = 2

Points and Events

- BLAST search yields alignment scores
  - How rare is that score?
  - Depends on substitution matrix, gap penalties
- Suppose score is very rare
  - How many sequences were in database?
  - With many sequences, a lottery-style "winner" may have unusually high score

E value

- Each sequence has ‘match score’
- E (expect) represents likelihood
  - Number of database sequences in database
    - Helps determine search space
  - Probability that score exceeds result
    - If score = 210 bits
    - P(score ≥ 210)
- Multiply probability, database size
  - Wind up with E value

The Monkey Problem

- Famous problem in probability courses
- Give all monkeys on earth keyboards in which they type eight hours a day for a year.
  - Keystrokes, number of monkeys all given
- What is probability of a monkey typing the word “Hamlet?”
- What is probability of a monkey typing Act I, Scene I of “Hamlet?”

Monkey Conclusions

- For word “Hamlet”
  - E value > 1
  - The monkey ***may have been*** lucky
- For Act I, Scene I of “Hamlet”
  - E value < 1
  - Even with misspellings!
  - The monkey was clearly not lucky!

E value interpretation

- E values much lower than 1:
  - Similarity between sequences significant
- Test for homology
  - Look for $E < 1 \times 10^{-4}$
  - If E value is lower, then sequence is a homolog
    - Very low error rate
- E values greater than 1×10^{-4}
  - **No conclusions!**
  - Never infer non-homology!

Bits

- Crucial E value calculations are done in bits
  - Base 2 numbers or powers of two
  - For instance, 32 = 5 bits
  - $2^5 = 32$
- E value = N/S'
  - N is the database size (search space) in bits
  - S' is the adjusted score in bits

Examples

- If search space is 35 bits and score is 32 bits, then
  - $E = N/S' = 35 \text{ bits}/32 \text{ bits} = 2^{35}/2^{32} = 2^3 = 8$
- If search space is 35 bits and score is 42 bits, then
  - $E = N/S' = 35 \text{ bits}/42 \text{ bits} = 2^{35}/2^{42} = 2^{-7} = 8 = 0.078$
- **Low E values indicate rarity of score overcomes search space**

Conculsions

- Think of two factors in E value
  - Rarity of score
  - Size of database
- If E value is less than 1e-04
  - Score is very rare, even with a large search space