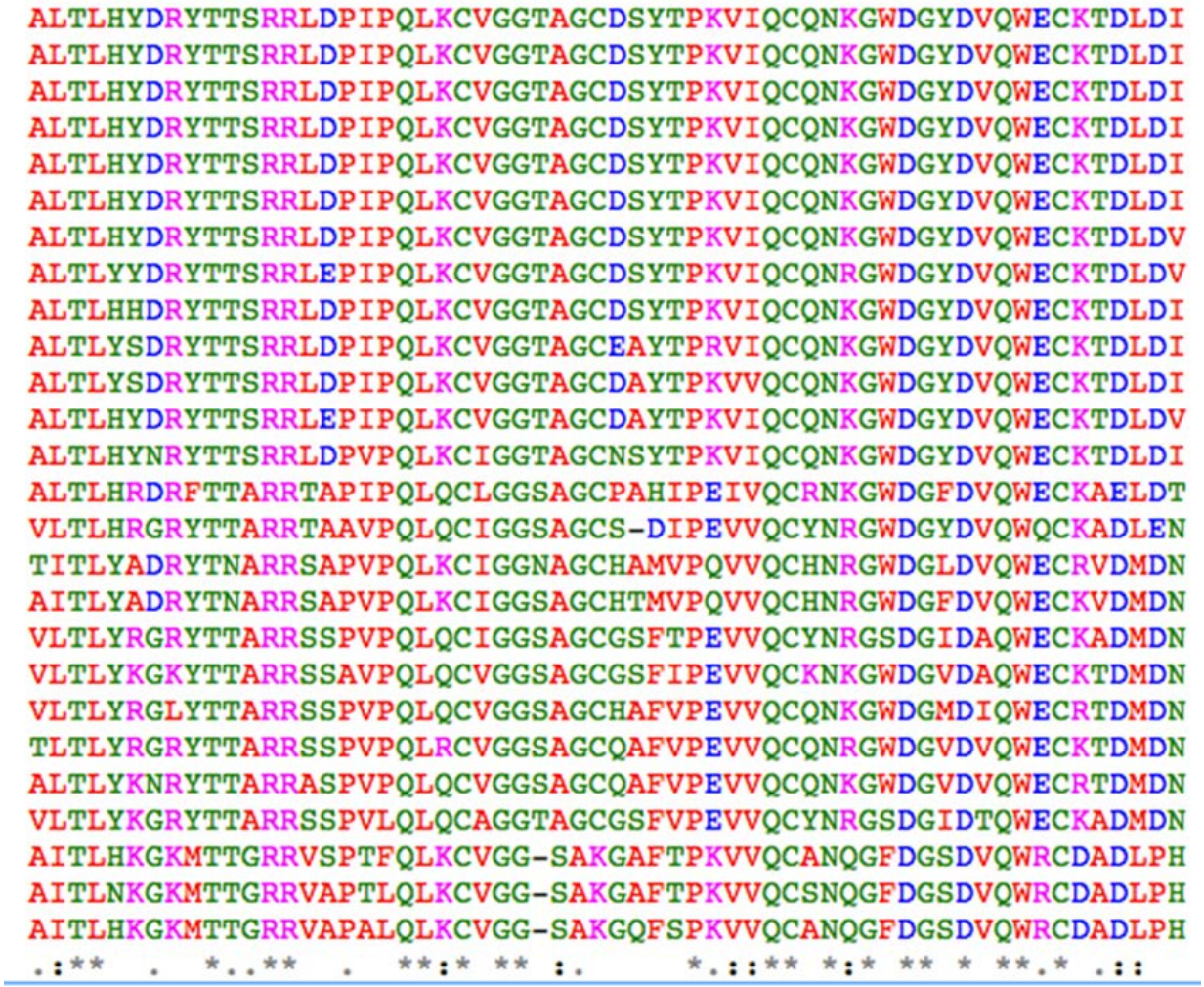
ASAB – Week 3
Intro to Sequence Alignment
Motif search

Algorithms for Sequence Analysis in Bioinformatics

Arnau Cordomí
arnau.cordomi@esci.upf.edu

Why do we align sequences?

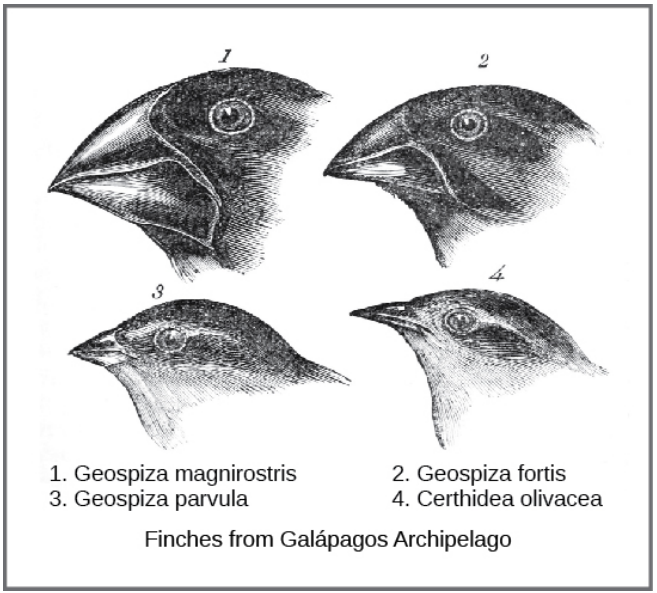

ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDV
ALTLYYDRYTTSRRLEPIPQLKCVGGTAGCDSYTPKVIQCQNRGWDGYDVQWECKTDLDV
ALTLHHDRYTTSRRLDPIPQLKCVGGTAGCDSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCEAYTPRVIQCQNKGWDGYDVQWECKTDLDI
ALTLYSDRYTTSRRLDPIPQLKCVGGTAGCDAYTPKVVQCQNKGWDGYDVQWECKTDLDI
ALTLHYDRYTTSRRLEPIPQLKCVGGTAGCDAYTPKVIQCQNKGWDGYDVQWECKTDLDV
ALTLHYNRYTTSRRLDPVPQLKCIGGTAGCNSYTPKVIQCQNKGWDGYDVQWECKTDLDI
ALTLHRDRFTTARRTAPIPQLQCLGGSAGCPAHIPEIVQCRNKGWDGFDVQWECKAELDT
VLTLHRGRYTTARRTAAVPQLQCIGGSAGCS-DIPEVVQCYNRGWDGYDVQWQCKADLEN
TITLYADRYTNARRSAPVPQLKCIGGNAGCHAMVPQVVQCHNRGWDGLDVQWECRVDMDN
AITLYADRYTNARRSAPVPQLKCIGGSAGCHTMVPQVVQCHNRGWDGFDVQWECKVDMDN
VLTLYRGRYTTARRSSPVPQLQCIGGSAGCGSFTPEVVQCYNRGSDGIDAQWECKADMDN
VLTLYKGKYTTARRSSAVPQLQCVGGSAGCGSFIPEVVQCKNKGWDGVDAQWECKTDMDN
VLTLYRGLYTTARRSSPVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGMDIQWECRTDMDN
TLTLYRGRYTTARRSSPVPQLRCVGGSAGCQAFVPEVVQCQNRGWDGVDVQWECKTDMDN
ALTLYKNRYTTARRASPVPQLQCVGGSAGCQAFVPEVVQCQNKGWDGVDVQWECRTDMDN
VLTLYKGRYTTARRSSPVLQLQCAGGTAGCGSFVPEVVQCYNRGSDGIDTQWECKADMDN
AITLHKGKMTTGRRVSPTFQLKCVGG-SAKGAFTPKVVQCANQGFDGSDVQWRCDADLPH
AITLNKGKMTTGRRVAPTLQLKCVGG-SAKGAFTPKVVQCSNQGFDGSDVQWRCDADLPH
AITLHKGKMTTGRRVAPALQLKCVGG-SAKGQFSPKVVQCANQGFDGSDVQWRCDADLPH
.:** . *..** . **:* ** :. *.::** *:* ** * **.* .::

Why comparting sequences?

Comparison of biological sequences is the most fundamental tool of Bioinformatics.

- Similar genes/proteins have **similar functions**
- Similar proteins have **similar structures**
- Identify **evolutionary relationships** between genes or proteins
- Classification

Charles Darwin: the Galapagos finches



Anatomy: beak sizes and shapes

- Gene Sequences: DNA, RNA, Proteins
- Structures of proteins, RNA and Genomes
- Genome Sequences: CNV and epigenetics
- Omics phenotypes: Transcriptomes, Proteomes, Metabolomes
- Phenotypes and Behaviors: Big Data

An alignment is a story

ADKPKRPLSAYMLWLN

ADKPKRPLSAYMLWLN

ADKPRRPLS-YMLWLN

Mutation

+

Selection

ADKPKRPLSAYMLWLN

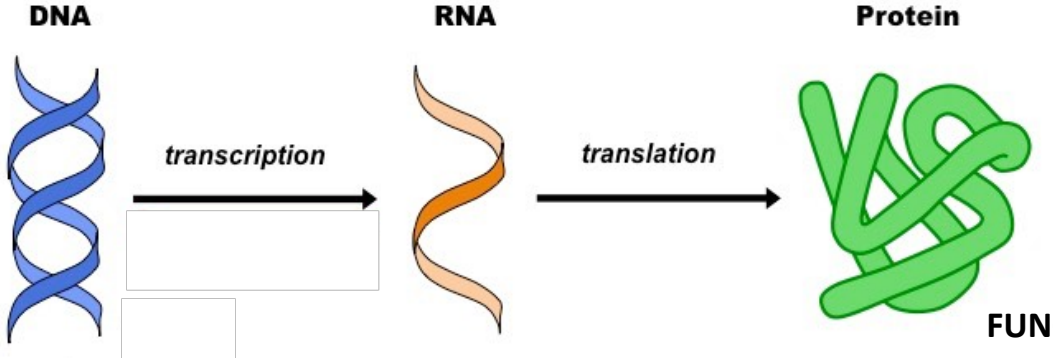
ADKPKRPKPRLSAYMLWLN

Insertion

Deletion

```
ADKPRRP---LS-YMLWLN
ADKPKRPKPRLSAYMLWLN
```

Mutation



Store information

FUNCTION

AGCTACGTACGTACC**GGT**
AGACGAGGATCGACTAG
TCGATAGTAGCTACG

GGUAGA

GR

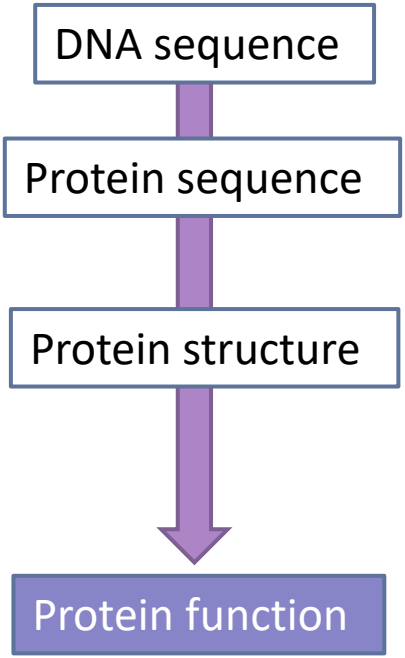
A, C, G, T

A, C, G, U

A, C, D, E, F, G, H, I, K, L
M, N, P, Q, R, S, T, W, Y, V

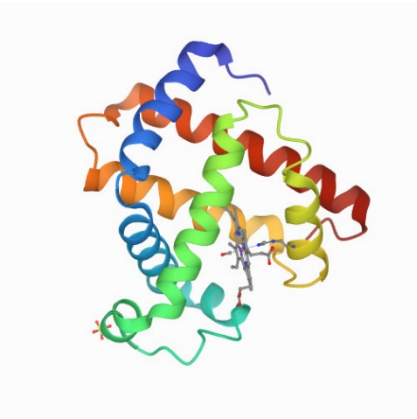
*The **central dogma of biology** describes the flow of genetic information within a biological system (Francis Crick in 1957)*

DNA sequence determines function



Anfinsen's dogma: the protein's sequence determines the native structure

Proteins are molecular machines



Myoglobin (MYG, Mg)

Human vs. Chimpanzee

```
MYG_HUMAN MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
MYG_PANTR MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE 60
          ************************************************************

MYG_HUMAN DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH 120
MYG_PANTR DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLHSKH 120
          ************************************************************:***

MYG_HUMAN PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
MYG_PANTR PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
          **********************************
```

>99%
Sequence identity

Sequence comparison (Myoglobins)

Human, Horse, Whale (ORCOR), Zebrafish (DANRE), Tuna (KATPE)

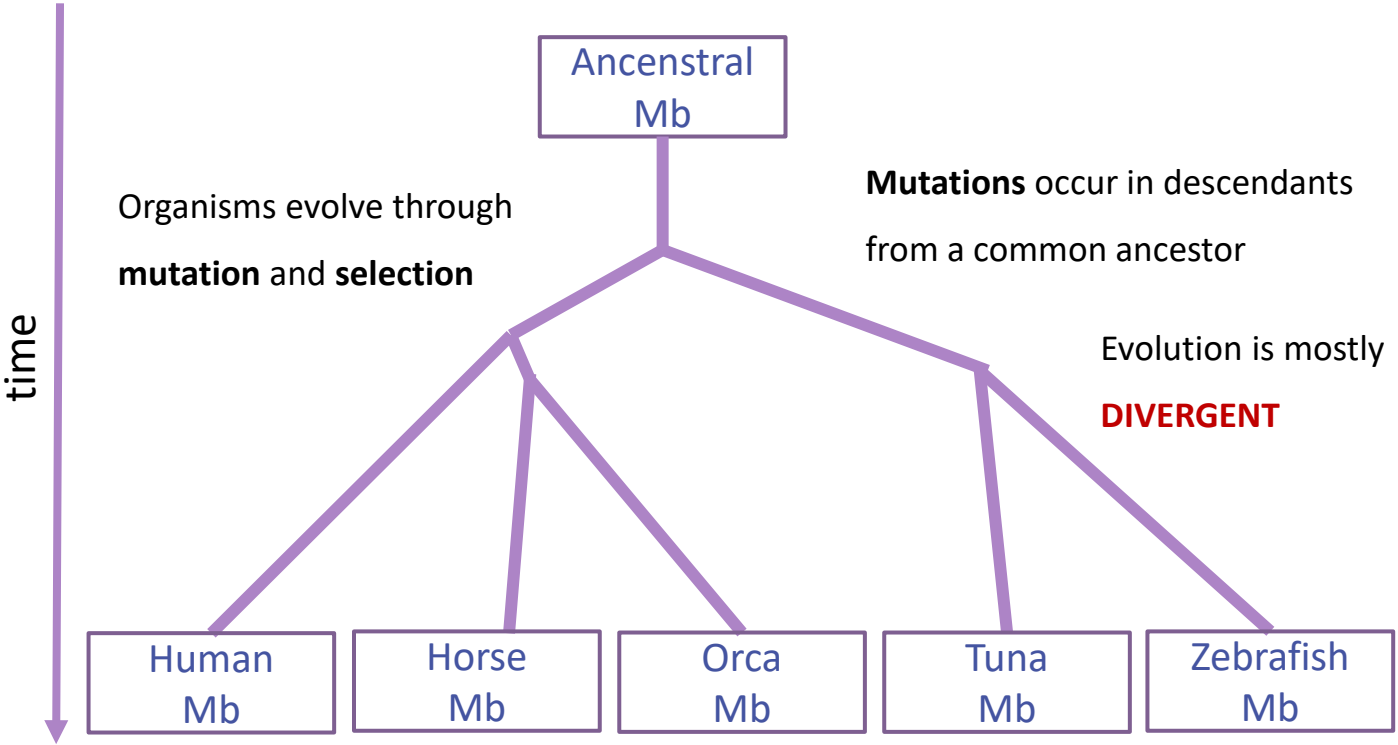
```
sp|P02144|MYG_HUMAN     MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASE   60
sp|P68082|MYG_HORSE     MGLSDGEWQQVLNVWGKVEADIAGHGQEVLIRLFTGHPETLEKFDKFKHLKTEAEMKASE   60
sp|P02173|MYG_ORCOR     MGLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPETLEKFDKFKHLKTEADMKASE   60
sp|Q6VN46|MYG_DANRE     ----MADHDLVLKCWGAVEADYAANGGEVLNRLFKEYPDTLKLFPKFSGIS-QGDLAGSP   55
sp|Q9DGI8|MYG_KATPE     ----MADLDAVLKCWGAVEADFNTVGGLVLARLFKDHPETQKLFPKFAGIT--GDIAGNA   54
                             .: : **: ** ****    *  :* ***. :*:* : * **  :.   :: ..

sp|P02144|MYG_HUMAN     DLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKH   120
sp|P68082|MYG_HORSE     DLKKHGTVVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSKH   120
sp|P02173|MYG_ORCOR     DLKKHGNTVLTALGAILKKKGHHDAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRH   120
sp|Q6VN46|MYG_DANRE     AVAAHGATVLKKLGELLKAKGDHAALLKPLANTHANIHKVALNNFRLITEVLVKVMAEKA   115
sp|Q9DGI8|MYG_KATPE     AVAAHGATVLKKLGELLKAKGNHAAIIKPLANSHAKQHKIPINNFKLITEALAHVLHEKA   114
                          :  ** .**. ** :** **.* * :****::**. **: :: :.:*:: : :*: .:

sp|P02144|MYG_HUMAN     PGDFGADAQGAMNKALELFRKDMASNYKELGFQG     154
sp|P68082|MYG_HORSE     PGDFGADAQGAMTKALELFRNDIAAKYKELGFQG     154
sp|P02173|MYG_ORCOR     PAEFGADAQGAMNKALELFRKDIAAKYKELGFHG     154
sp|Q6VN46|MYG_DANRE     --GLDAAGQGALRRVMDAVIGDIDGYYKEIGFAG     147
sp|Q9DGI8|MYG_KATPE     --GLDAAGQTALRNVMGIVIADLEANYKELGFTG     146
                           :.* .* *: ..:  .  *: . ***:** *
```

1:	sp|P02144|MYG_HUMAN	100.00	88.31	85.71	41.50	44.52
2:	sp|P68082|MYG_HORSE	88.31	100.00	88.96	40.14	43.84
3:	sp|P02173|MYG_ORCOR	85.71	88.96	100.00	41.50	45.21
4:	sp|Q6VN46|MYG_DANRE	41.50	40.14	41.50	100.00	70.55
5:	sp|Q9DGI8|MYG_KATPE	44.52	43.84	45.21	70.55	100.00

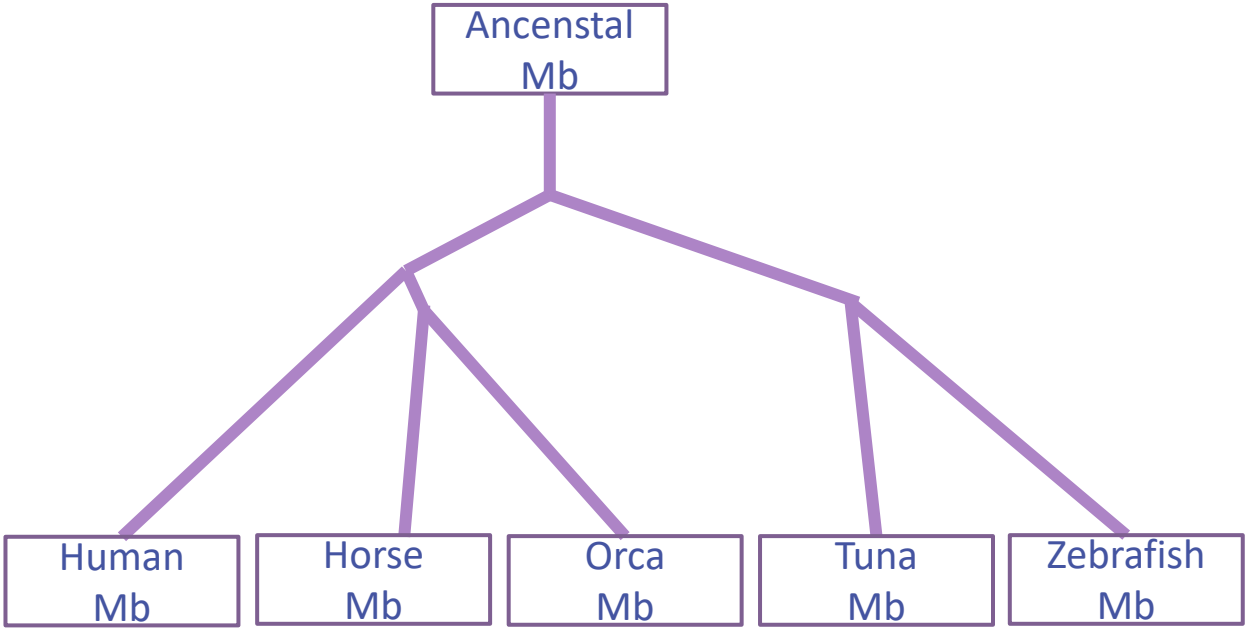
A sequence alignment is a story!



The biochemical properties and cellular functions tend to be preserved

Normally we only see the end of hidden the story!

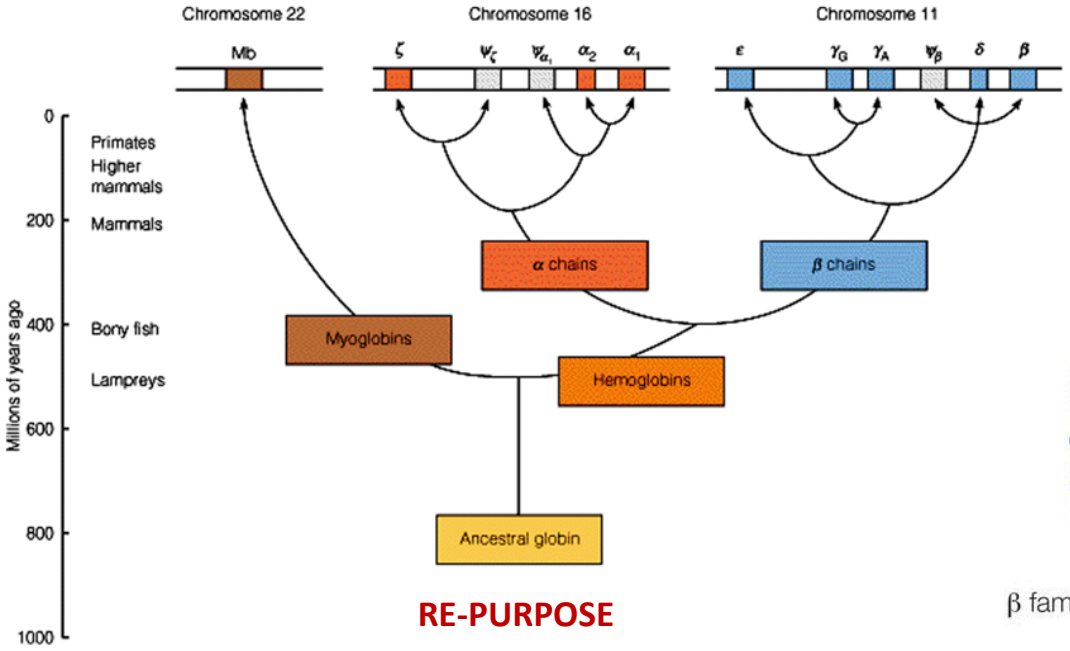
- Large sequence identity between 2 sequences is telling us that proteins diverged from a common **ancestor**.
- **Homology** refers to the similarity between characteristics of organisms due to a common origin from a common ancestor.
- 2 types of homologs: **orthologs** and **paralogs**



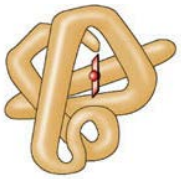
Orthologous sequences: inferred to be descended from the same ancestral sequence separated by ***speciation events***.

Orthologous proteins have the same function.

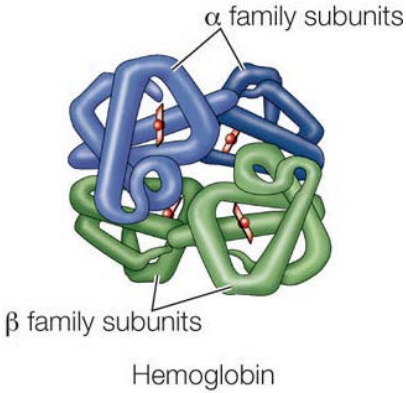
Globins evolution



RE-PURPOSE



Myoglobin

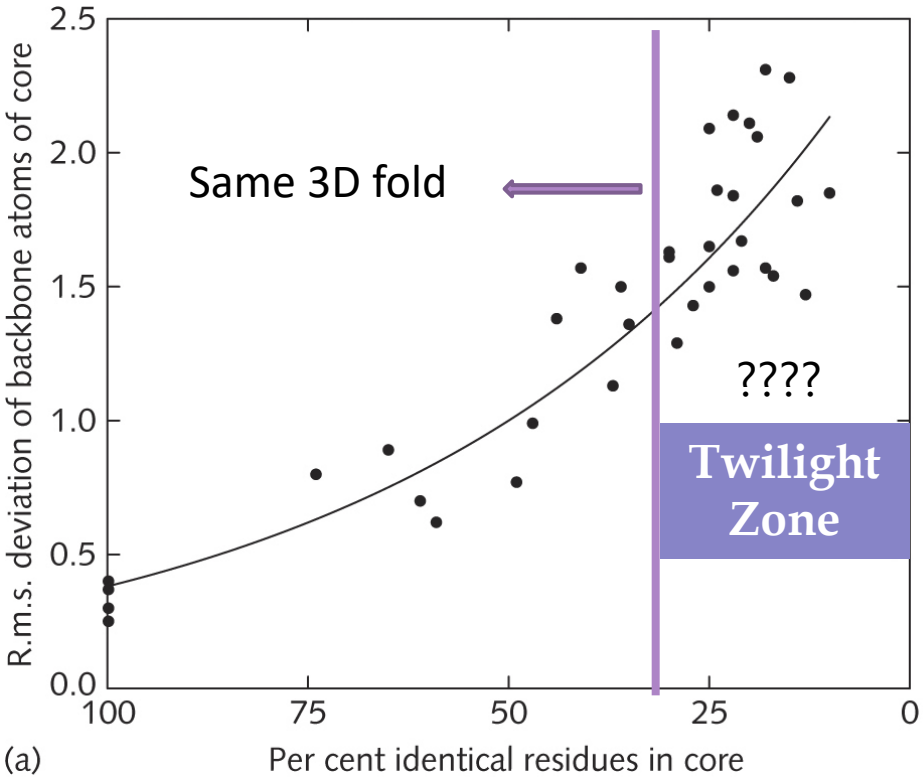


Hemoglobin

Paralogs: Descendants from a common ancestor separated by *a duplication event*. Often acquire new molecular functions.

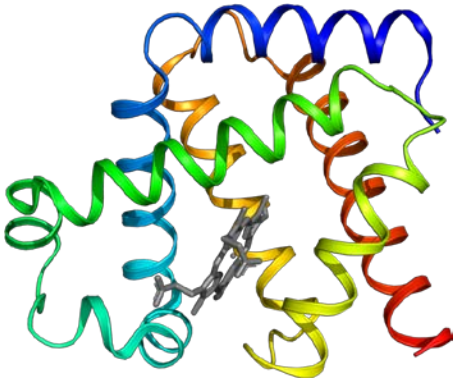
Ex: myoglobin and hemoglobin
alpha hemoglobin and beta hemoglobin

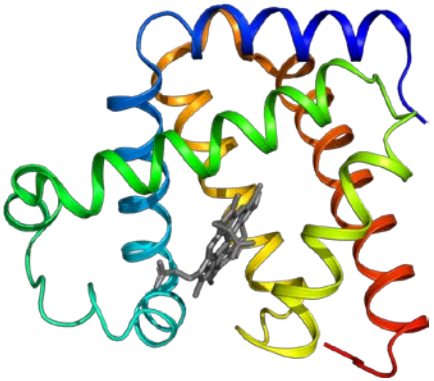
Sequence vs. structure conservation

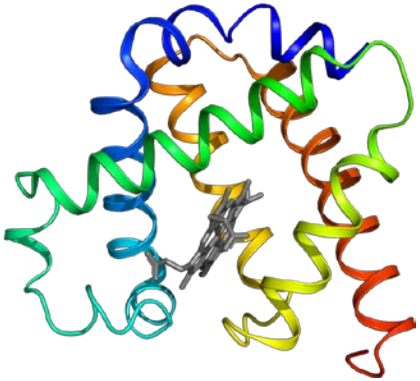


Sequence similarities can be low between proteins that share the same structure

Structure comparison (Myoglobins)







Human
Mb
PDB id: 3RGK

Whale
Mb
PDB id: 7CEN

Tuna
Mb
PDB id 2NRL

85%
identity

45%
identity

45%
identity

MYG: Myoglobin (human)
HBA: Hemoglobin alpha (human)
LGB2: Leghemoglobin (plant)

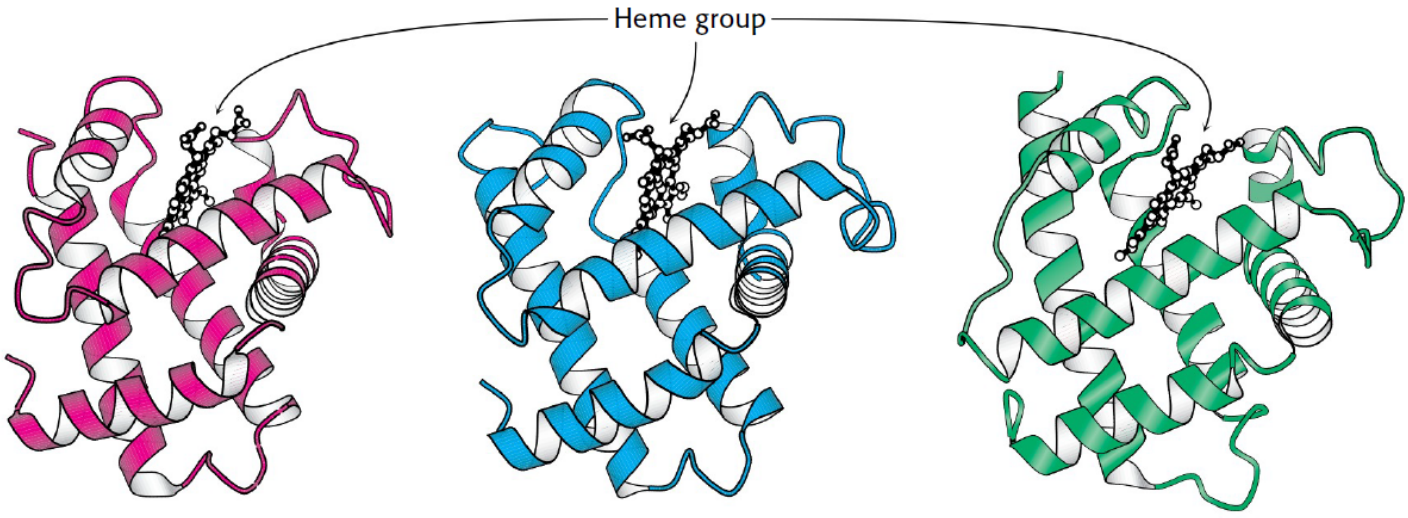
```
sp|P02240|LGB2_LUPLU      MGALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLK---GTSEVPQN  57
sp|P69905|HBA_HUMAN       -MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHG-------  52
sp|P02144|MYG_HUMAN       -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMK-A  58
                             *: .:   *   * :. *.   :  . :  ::   * :   *  :.   .

sp|P02240|LGB2_LUPLU      NPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKGV-ADAHFPVVKEAIL  116
sp|P69905|HBA_HUMAN       SAQVKGHGKKVADALTNA-----VAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLL  107
sp|P02144|MYG_HUMAN       SEDLKKHGATVLTALGGI-----LKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECII  113
                          . ::: *. .*   :        :         :. *.. *. *       : .:...::

sp|P02240|LGB2_LUPLU      KTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA------- 154
sp|P69905|HBA_HUMAN       VTLAAHLPAEFTPAVHAS----LDKFLASVSTVLTSKYR------ 142
sp|P02144|MYG_HUMAN       QVLQSKHPGDFGADAQGA----MNKALELFRKDMASNYKELGFQG 154
                           .:     ..:    :.:     ::    . . : .
```

```
1: sp|P02240|LGB2_LUPLU  100.00   17.52   18.18
2: sp|P69905|HBA_HUMAN    17.52  100.00   26.76
3: sp|P02144|MYG_HUMAN    18.18   26.76  100.00
```

Structural comparison (Globins)



Hemoglobin (α chain)

PDB id: 1HBB

Myoglobin

PDB id: 7CEN

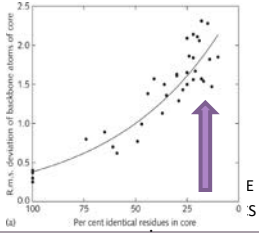
Leghemoglobin

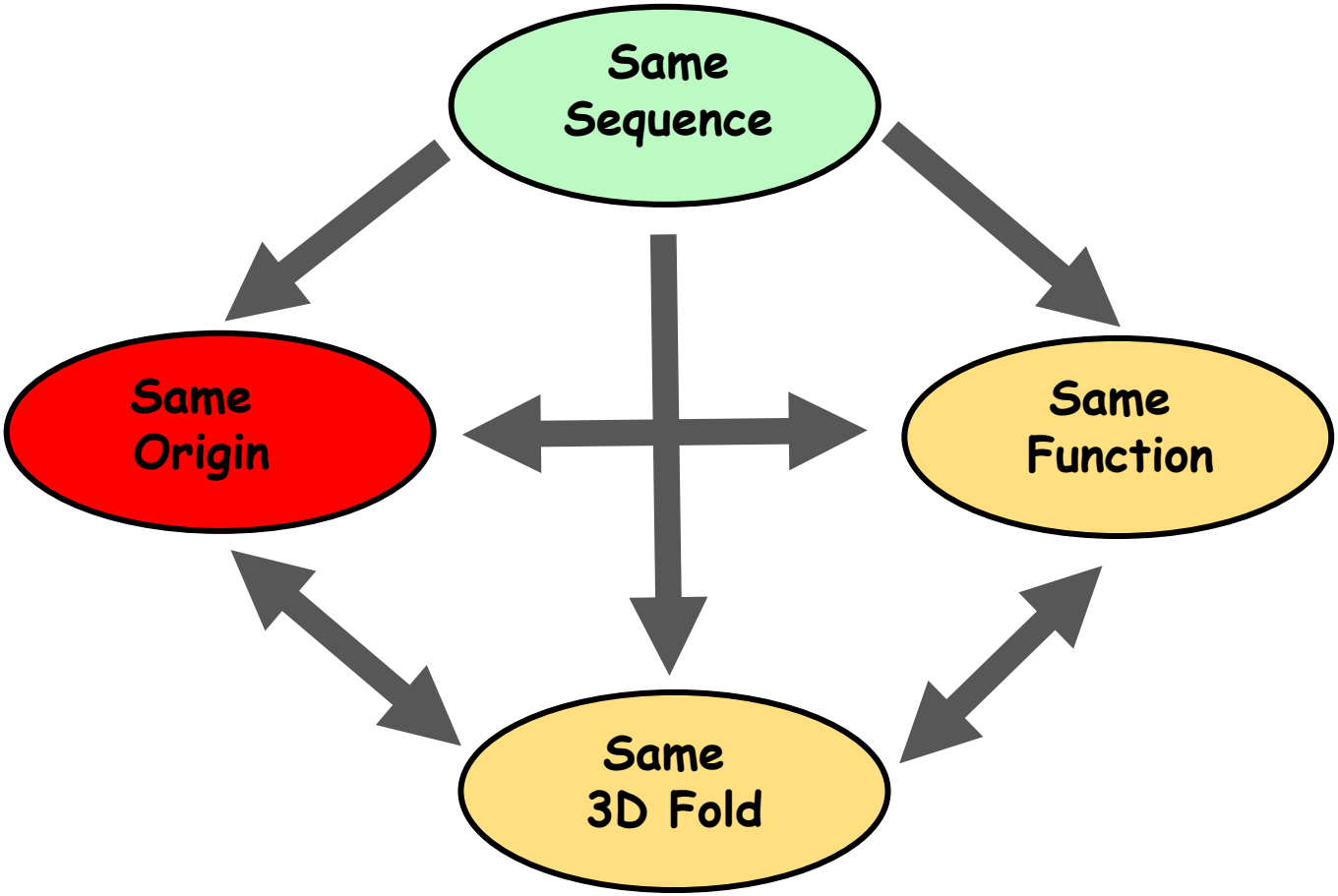
PDB id 1GDJ

26% identity

18% identity

18% identity





The money graph (by Cedric Notredame)

An alignment is a model

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

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

By putting amino acids in the same column, you make a ***hypothesis about their relationships.***

A relationship *may mean different things ….*

Different criteria to align residues

Sequence similarity: amino acids in the same column are those that yield and alignment with maximum similarity.
- *Most programs use this criteria because is the easiest.*

Evolution: ammino acids related to the same amino acid in the common ancestor of all the sequences are put in the same column.
- *Programs do not use it explicitly but respect it.*

Structural similarity: amino acids that play the same role in each structure are in the same column.
- *Used in structure superposition programs.*

Functional similarity: amino acids with the same function are in the same column.
- *No program use this criterion, but you may impose it*

The alignment problem

1o letters

THEFASTCAT

SCORING SYSTEM

match score = 1

mismatch score = -1

m: number of **m**atches
n: number of **n**ot matches

score = m · match score + n · mismatch score

	matches (m)	mismatches (n)	score
THE**L**AST**C**AT	9	1	8
THE**L**AST**R**AT	8	2	6
THEFASTCAT	10	0	10

Comparing sequences of different length (I)

Comparison requires an alignment

1o letters

```
THEFASTCAT
THEFATCAT        1
THEFATCA T       3
THEFATC AT       5
THEFAT CAT       7
THEFA TCAT       9
THEF ATCAT       7
THE FATCAT       5
TH EFATCAT       3
T HEFATCAT       1
 THEFATCAT       1
THEFATCAT
```

9 letters

SCORING SYSTEM

match score = 1

mismatch score = -1

gap scrore = 0

g: number of gaps

score = m · match score +
n · mismatch score +
g · gap score

$L_1 = 10$ (longest sequence)

g = 1

number of alignments (*shortest alignments*): $L_1 - g + 1 = 10$

Comparing sequences of different length (II)

1o letters

```
THEFASTCAT
AFASTCAT
AFASTCA  T
AFASTC  AT
AFAST  CAT
AFAS  TCAT
AFA  STCAT
AF  ASTCAT
A  FASTCAT
  AFASTCAT
```

```
THEFASTCAT
AFASTCA T
AFASTC AT
AFAST CAT
AFAS TCAT
AFA STCAT
AF ASTCAT
A FASTCAT
 AFASTCAT
```

8

```
THEFASTCAT
AFASTC A T
AFAST CA T
AFAS TCA T
AFA STCA T
AF ASTCA T
A FASTCA T
 AFASTCA T
```

7

…

6, 5, 4, 3, 2, 1

AFASTCAT

8 letters

$L_1 = 10$

g = 2 (two gaps together)

number of alignments:

$L_1 - g + 1 = 10 - 2 + 1 = 9$

total number of alignments (*shortest alignments*)

$L_1 \cdot (L_1 - 1) / 2 = 45$

Alignments up to $|s_1| + |s_2|$ characters long

```
THEFASTCAT---------
----------THEFATCAT
```

$$\binom{L_1 + L_2}{L_1} = \frac{(L_1 + L_2)!}{L_1!\, L_2!}$$

$L_1 = 10$

$L_2 = 9$

alignments = 92378

1o letters

```
THEFASTCAT
 THEFATCAT
  THEFATCAT
   THEFATCAT
    THEFATCAT
     THEFATCAT
      THEFATCAT
       THEFATCAT
        THEFATCAT
         THEFATCAT
```

```
THEFATCAT
```

9 letters

Write the possible alignments

1) Which alignments are possible between ACTG and TG with a maximum of two gaps? (shortest alignments; L = 4)

2) Which other alignments are possible between ACTG and TG (L = 6)?

Motif search

THEFASTCAT

CAT
CAT
CAT
CAT
CAT
CAT
CAT
CAT

THEFASTCAT

FAST
FAST
FAST
FAST
FAST
FAST
FAST

- *String S*
- *Substring* of S: a string occurring inside S

CAT

3 letters

$L_1 - L_2 + 1 = 8$

FAST

4 letters

$L_1 - L_2 + 1 + = 7$

1o letters

THEFASTCAT
THEFATCAT
THEFATCA T
THEFATC AT
THEFAT CAT
THEFA TCAT
THEF ATCAT
THE FATCAT
TH EFATCAT
T HEFATCAT
THEFATCAT

THEFASTCAT
AFASTCAT
AFASTCA T
AFASTC AT
AFAST CAT
AFAS TCAT
AFA STCAT
AF ASTCAT
A FASTCAT
AFASTCAT

THEFATCAT

9 letters

AFASTCAT

8 letters

Why motif search?

Table 7-1 Some Gene Regulatory Proteins and the DNA Sequences That They Recognize

	NAME		DNA SEQUENCE RECOGNIZED*
Bacteria	lac repressor	5′	AATTGTGAGCGGATAACAATT
		3′	TTAACACTCGCCTATTGTTAA
	CAP		TGTGAGTTAGCTCACT
			ACACTCAATCGAGTGA
	lambda repressor		TATCACCGCCAGAGGTA
			ATAGTGGCGGTCTCCAT
Yeast	Gal4		CGGAGGACTGTCCTCCG
			GCCTCCTGACAGGAGGC
	Matα2		CATGTAATT
			GTACATTAA
	Gcn4		ATGACTCAT
			TACTGAGTA
Drosophila	Kruppel		AACGGGTTAA
			TTGCCCAATT
	Bicoid		GGGATTAGA
			CCCTAATCT
Mammals	Sp1		GGGCGG
			CCCGCC
	Oct-1 Pou domain		ATGCAAAT
			TACGTTTA
	GATA-1		TGATAG
			ACTATC
	MyoD		CAAATG
			GTTTAC
	p53		GGGCAAGTCT
			CCCGTTCAGA

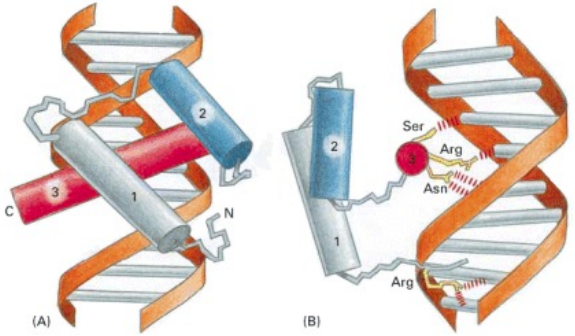
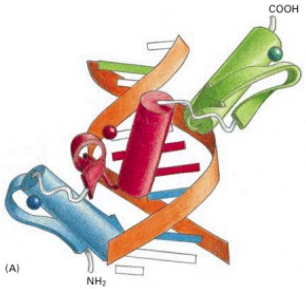


Figure 7-16 - A homeodomain bound to its specific DNA sequence



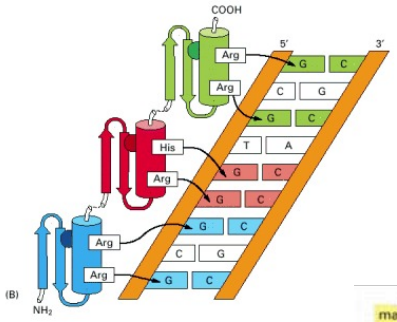


Figure 7-18 - DNA binding by a zinc finger protein

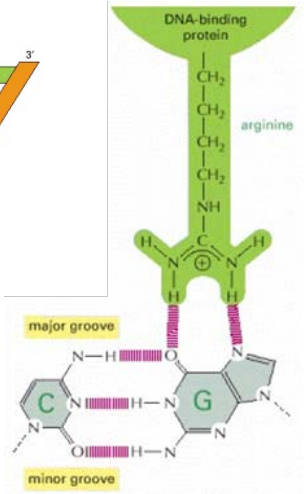


Figure 7-27 - One of the most common protein-DNA interactions

Molecular Biology of the Cell. 4th edition.
Alberts B, Johnson A, Lewis J, et al.
New York: Garland Science; 2002.

- **Basic Local Alignment Search Tool**
- **Compare** a query protein or nucleotide sequence **with a library of sequences**,

Google of biological research

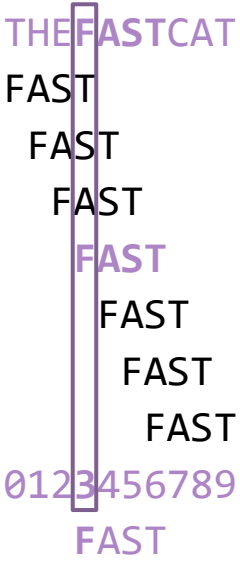
NCBI

http://www.ncbi.nlm.nih.gov/

***First step**: locating short words in common between the two sequences*

3 letter words

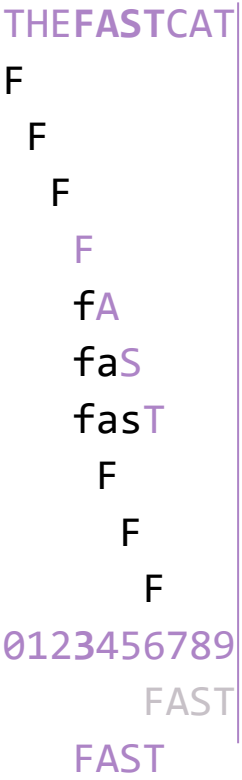
Naïve search algorithm



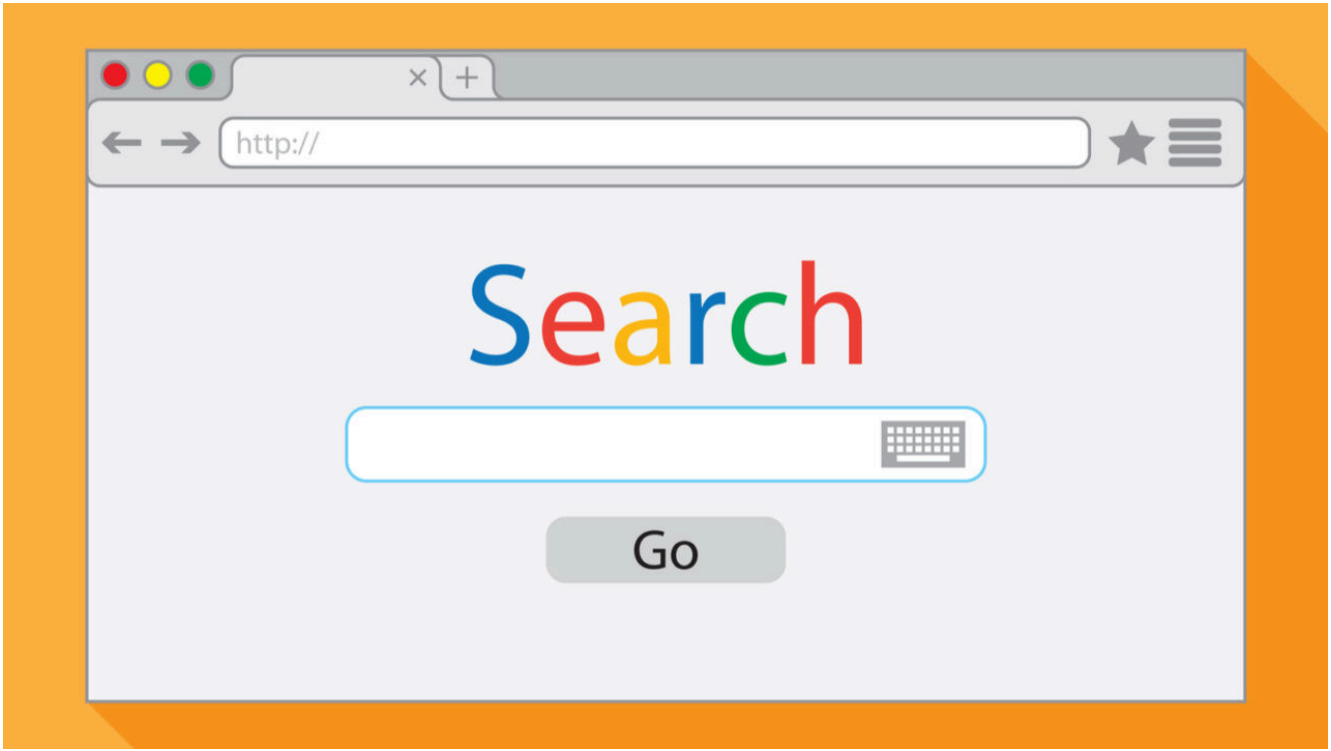
Positions within a string S are referred to with *offsets*

(in Python) leftmost offset = 0

The substring occurs at offset 3



Indexing



FATFASTRATFASTCAT
FAST

0	1	2	3	4	5	6	7	8	9	offset
THEFASTCAT	HEFASTCAT	EFASTCAT	FASTCAT	ASTCAT	STCAT	TCAT	CAT	AT	T	
THEFASTCA	HEFASTCA	EFASTCA	FASTCA	ASTCA	STCA	TCA	CA	A		
THEFASTC	HEFASTC	EFASTC	FASTC	ASTC	STC	TC	C			
THEFAST	HEFAST	EFAST	FAST	AST	ST	T				
THEFAS	HEFAS	EFAS	FAS	AS	S					
THEFA	HEFA	EFA	FA	A						
THEF	HEF	EF	F							
THE	HE	E								
TH	H									
T										

all substrings within

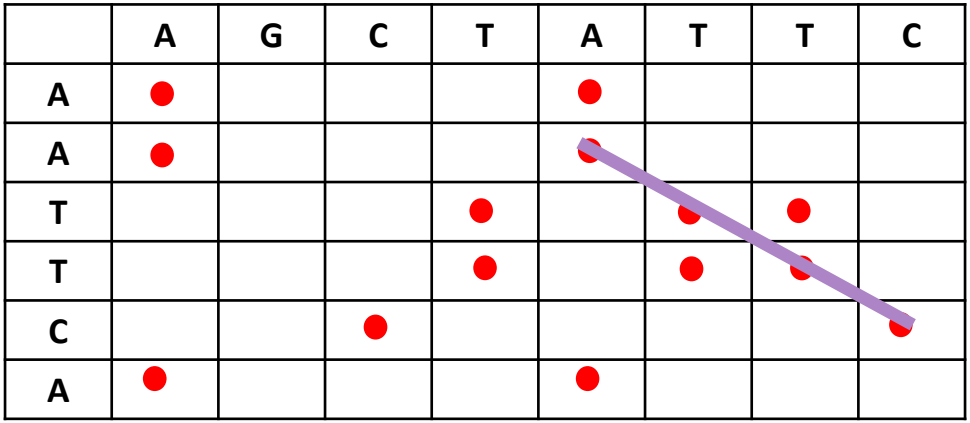
THEFASTCAT

all substrings in alphabetic order (and offset)

A	4	8	E	2	F	3	H	1	S	5		TH	0
AS	4		EF	2	FA	3	HE	1	ST	5		THE	0
AST	4		EFA	2	FAS	3	HEF	1	STC	5		THEF	0
ASTC	4		EFAS	2	FAST	3	HEFA	1	STCA	5		THEFA	0
ASTCA	4		EFAST	2	FASTC	3	HEFAS	1	STCAT	5		THEFAS	0
ASTCAT	4		EFASTC	2	FASTCA	3	HEFAST	1	T	0	6	9	THEFAST	0
AT	8		EFASTCA	2	FASTCAT	3	HEFASTC	1	TC	6		THEFASTC	0
C	7		EFASTCAT	2			HEFASTCA	1	TCA	6		THEFASTCA	0
CA	7						HEFASTCAT	1	TCAT	6		THEFASTCAT	0
CAT	7												

Dot plot

AGCTATTC vs. AATTCA



AGCTATTC

AATTCA

Dot-matrix (dot-plot) is one of the simplest (qualitative) ways to compare two sequences

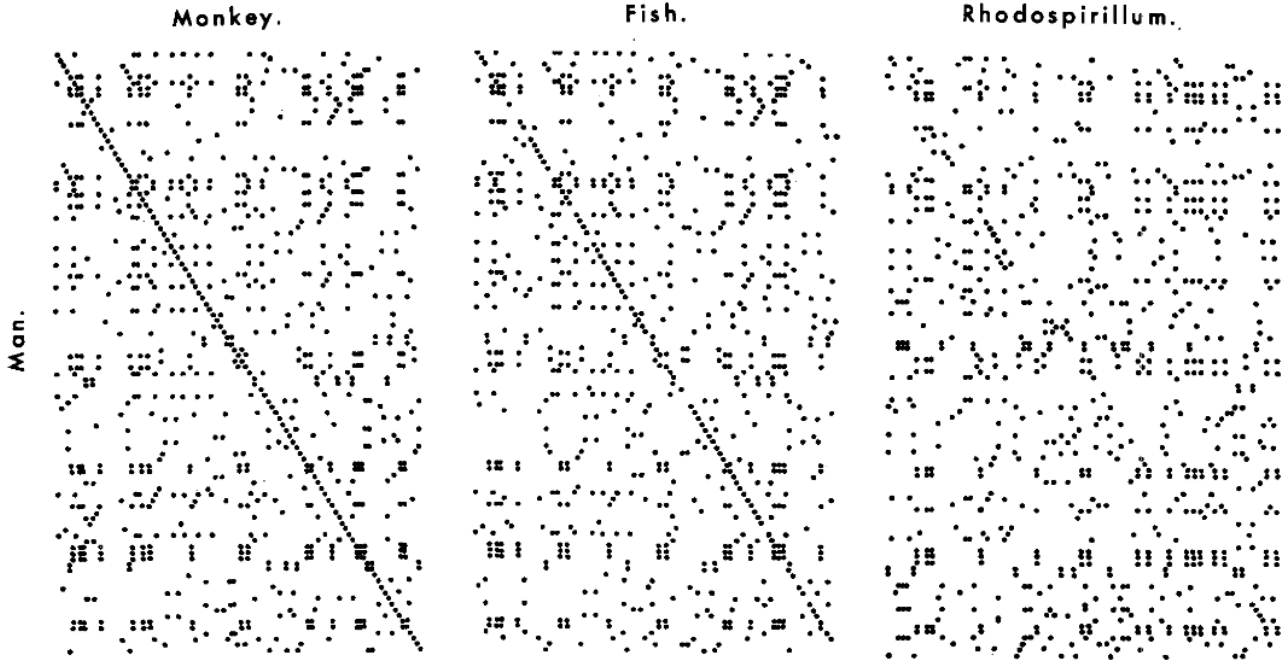


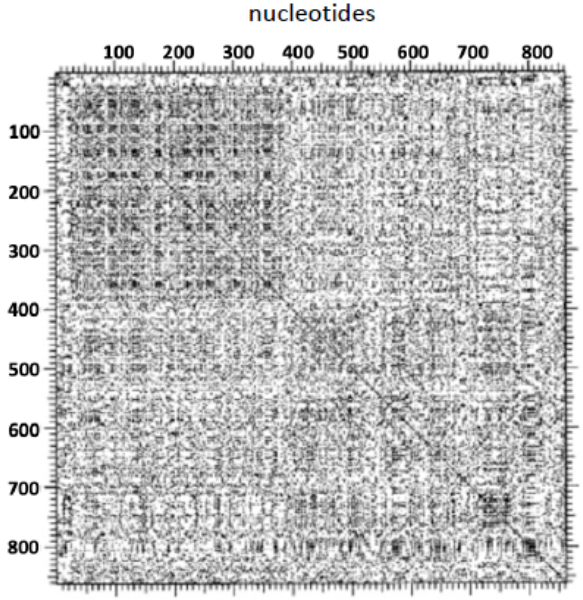
Fig. 1. *The diagrams obtained from comparisons of human cytochrome* c *(left margin of each diagram, N terminus at top) and the cytochromes* c *of monkey, fish and* Rhodospirillum *(upper margin, N termini at left end)*

Gibbs and McIntyre 1970
https://febs.onlinelibrary.wiley.com/doi/epdf/10.1111/j.1432-1033.1970.tb01046.x

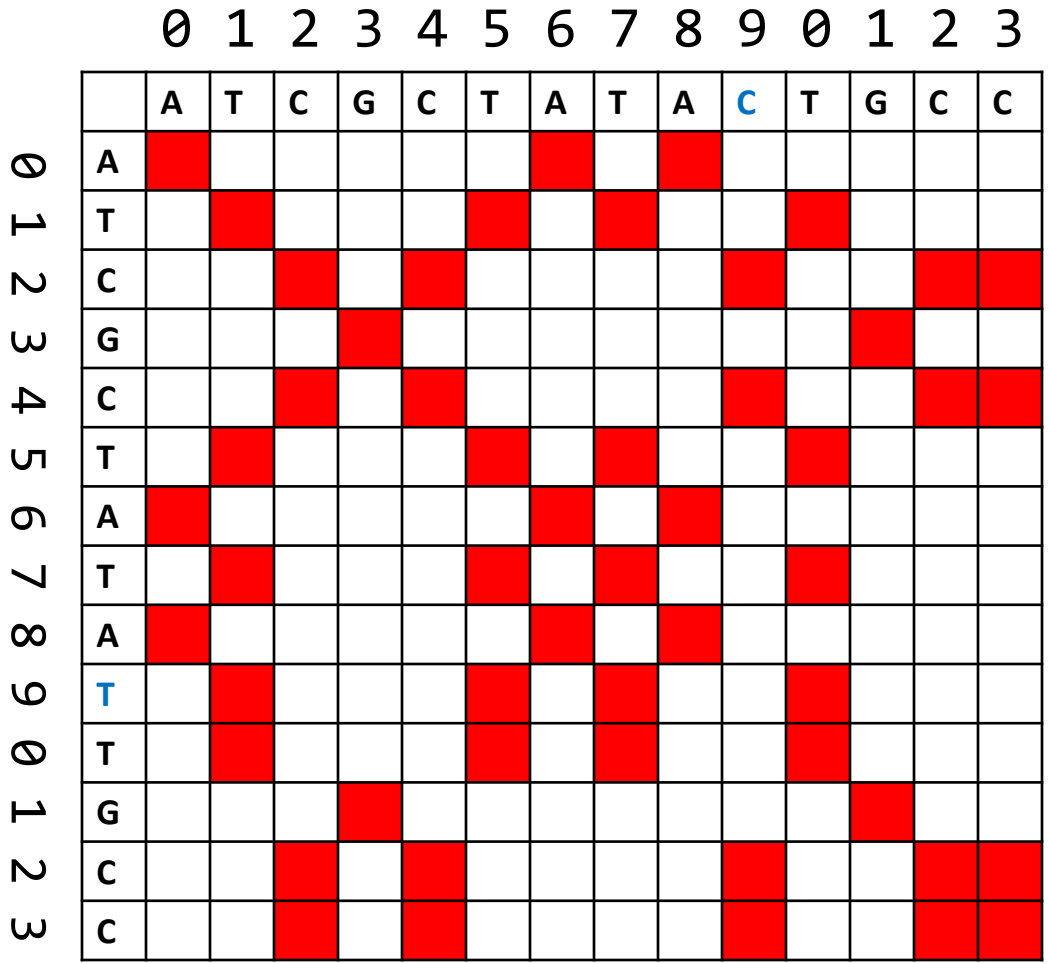
Dot Plots. Analysis

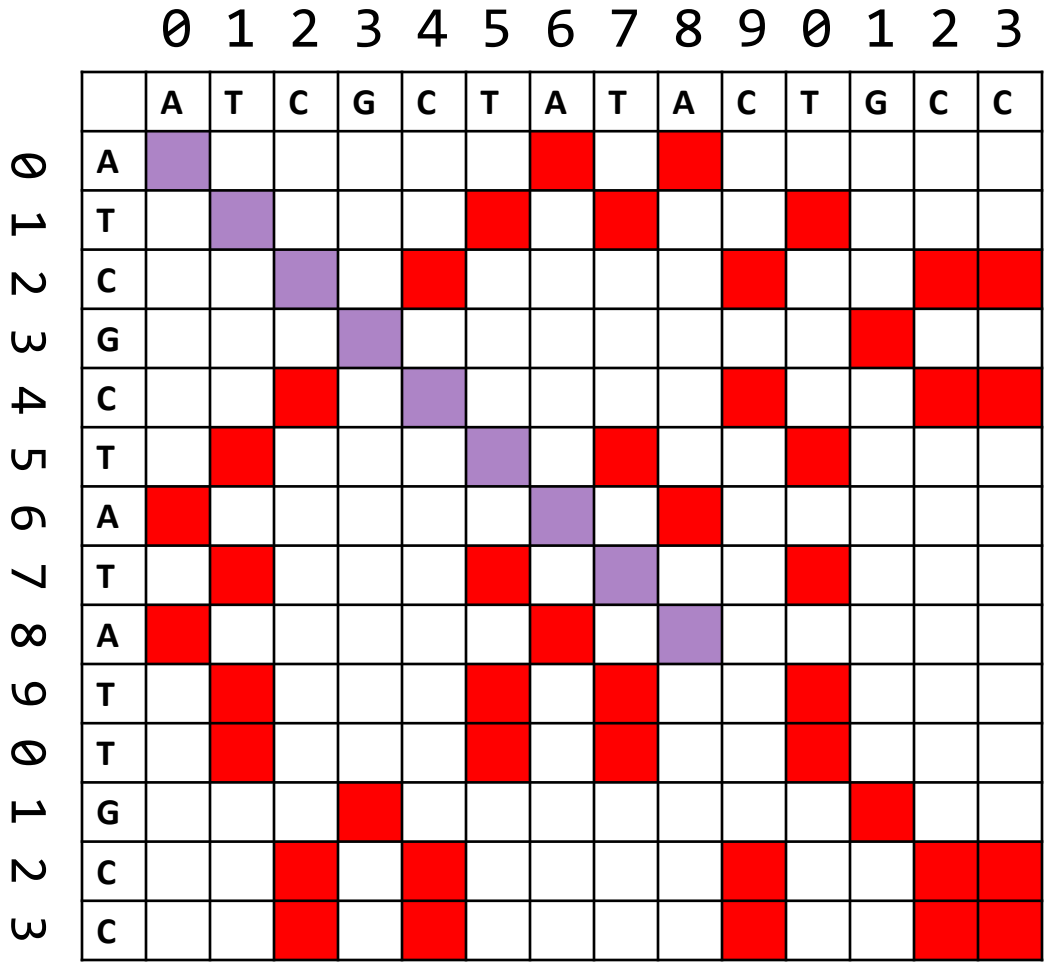
9 nucleotides

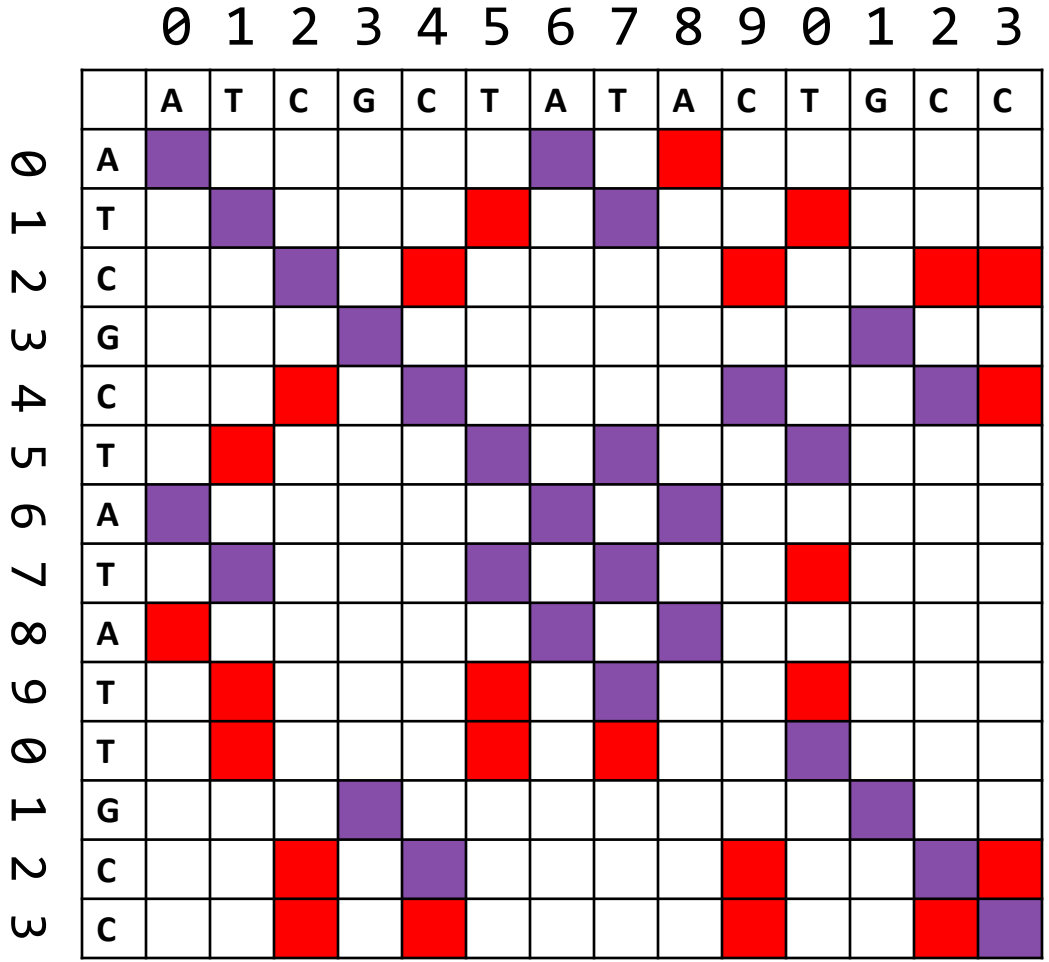
	C	T	A	G	A	A	C	T	T
C	●						●		
T		●						●	●
A			●		●	●			
G				●					
A			●		●	●			
A			●		●	●			
C	●						●		
T		●						●	●
T		●						●	●



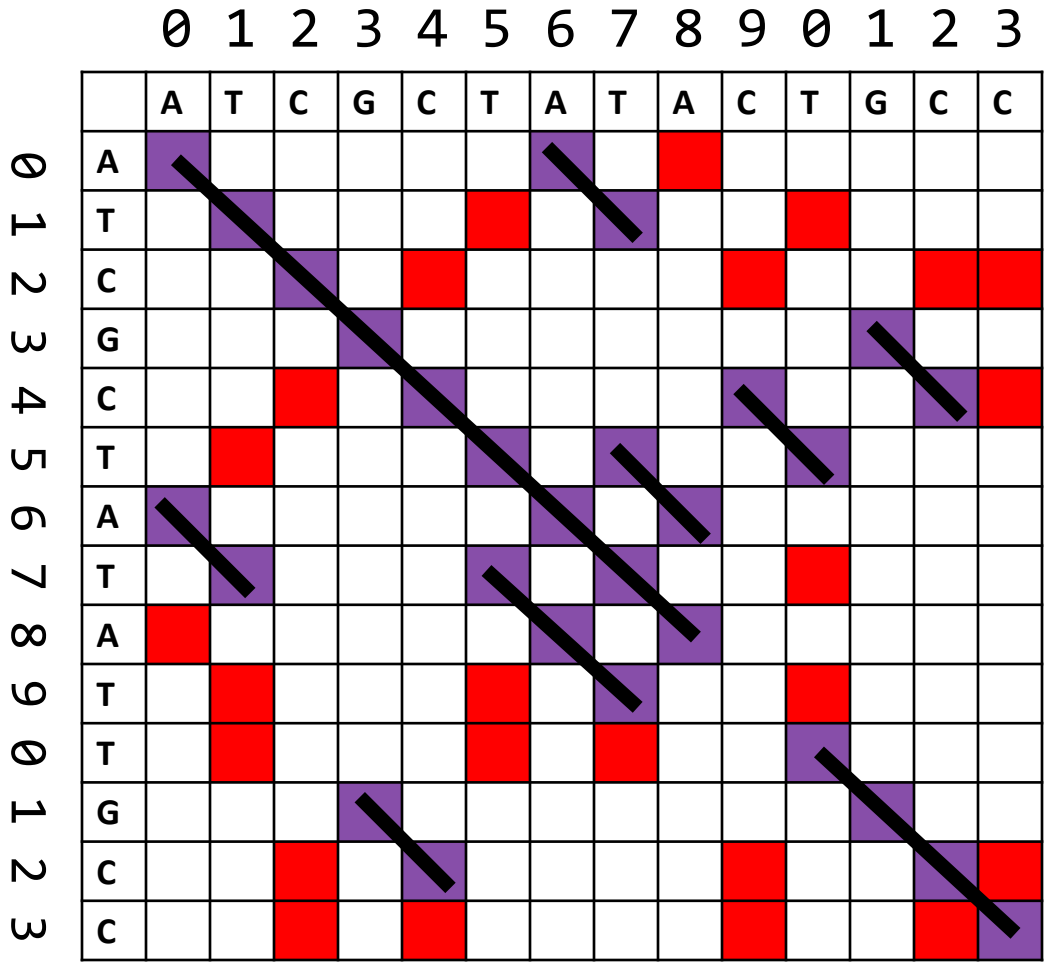
- For few nucleotides the noise is low
- For many nucleotides, there is a lot of noise

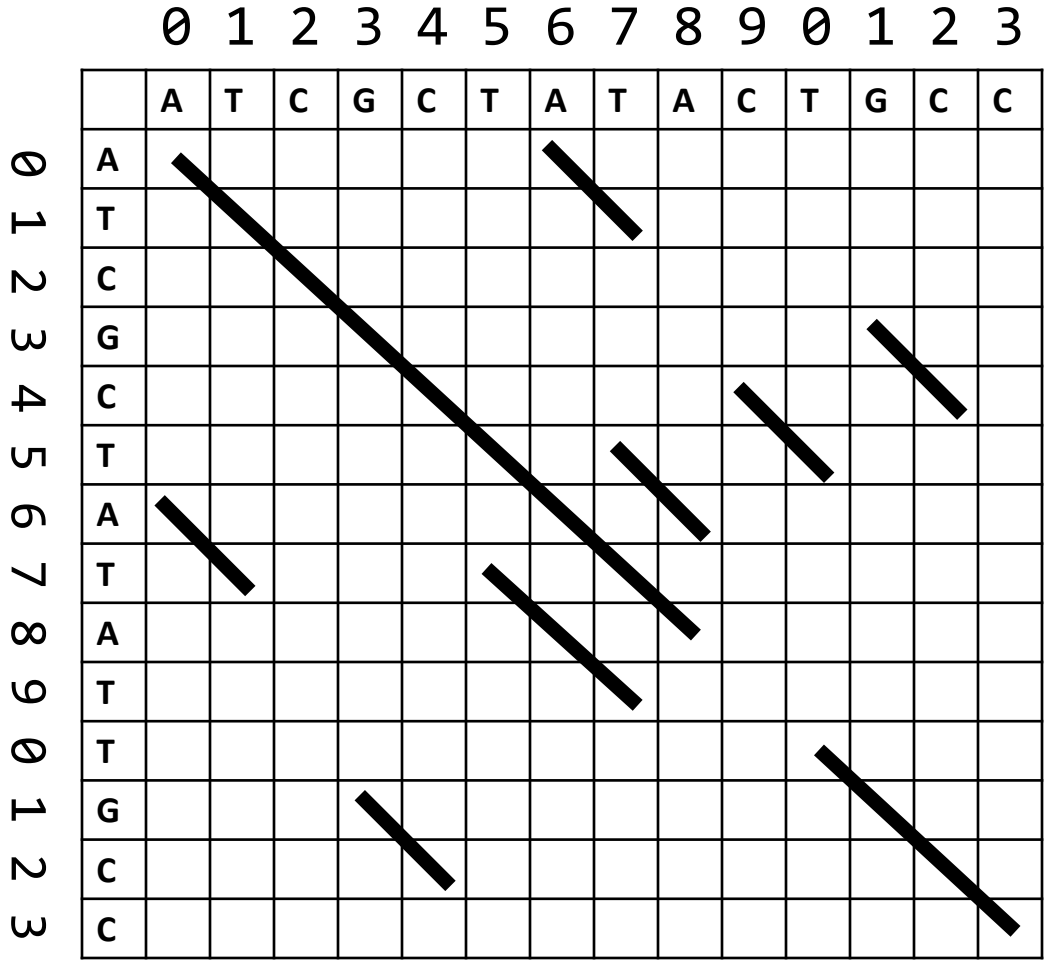

0 1 2 3 4 5 6 7 8 9 0 1 2 3
A T C G C T A T A C T G C C
0 A
1 T
2 C
3 G
4 C
5 T
6 A
7 T
8 A
9 T
0 T
1 G
2 C
3 C


0 1 2 3 4 5 6 7 8 9 0 1 2 3
A T C G C T A T A C T G C C
0 A
1 T
2 C
3 G
4 C
5 T
6 A
7 T
8 A
9 T
0 T
1 G
2 C
3 C

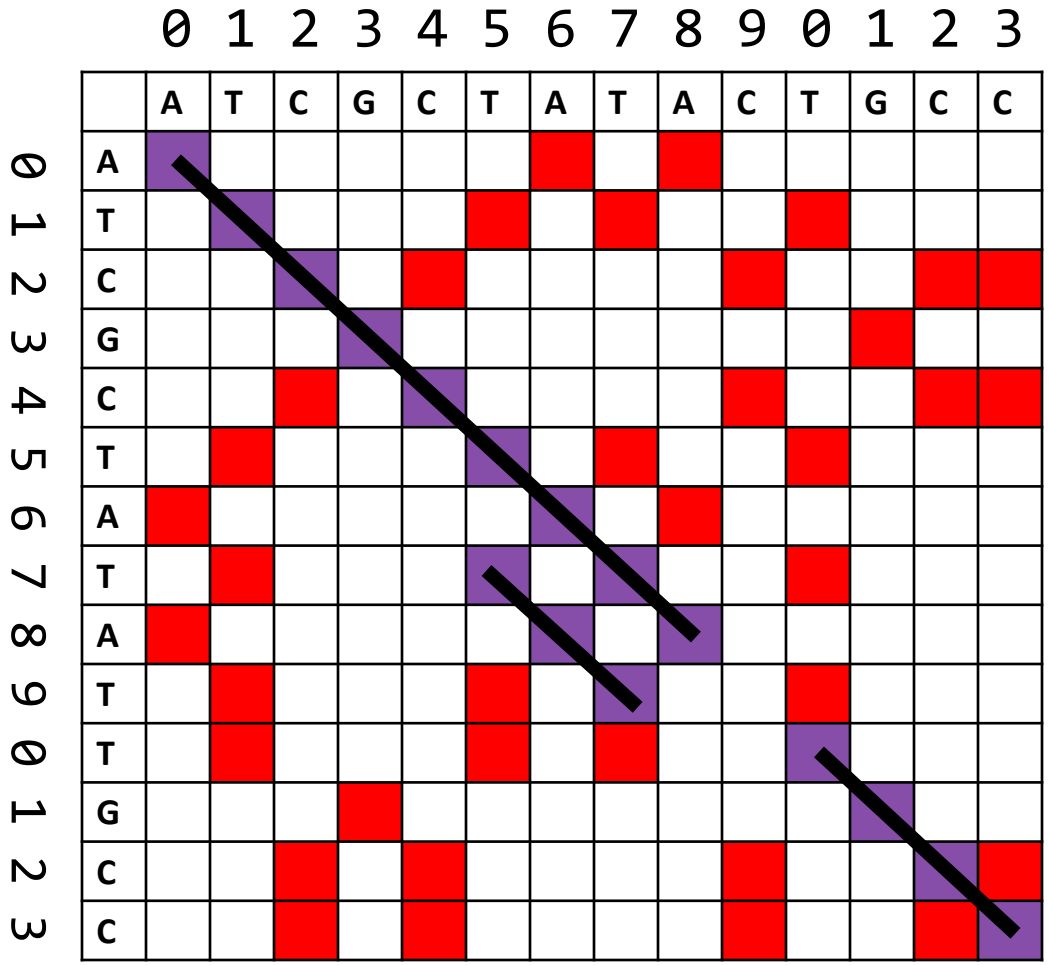


Words of length ≥ 2

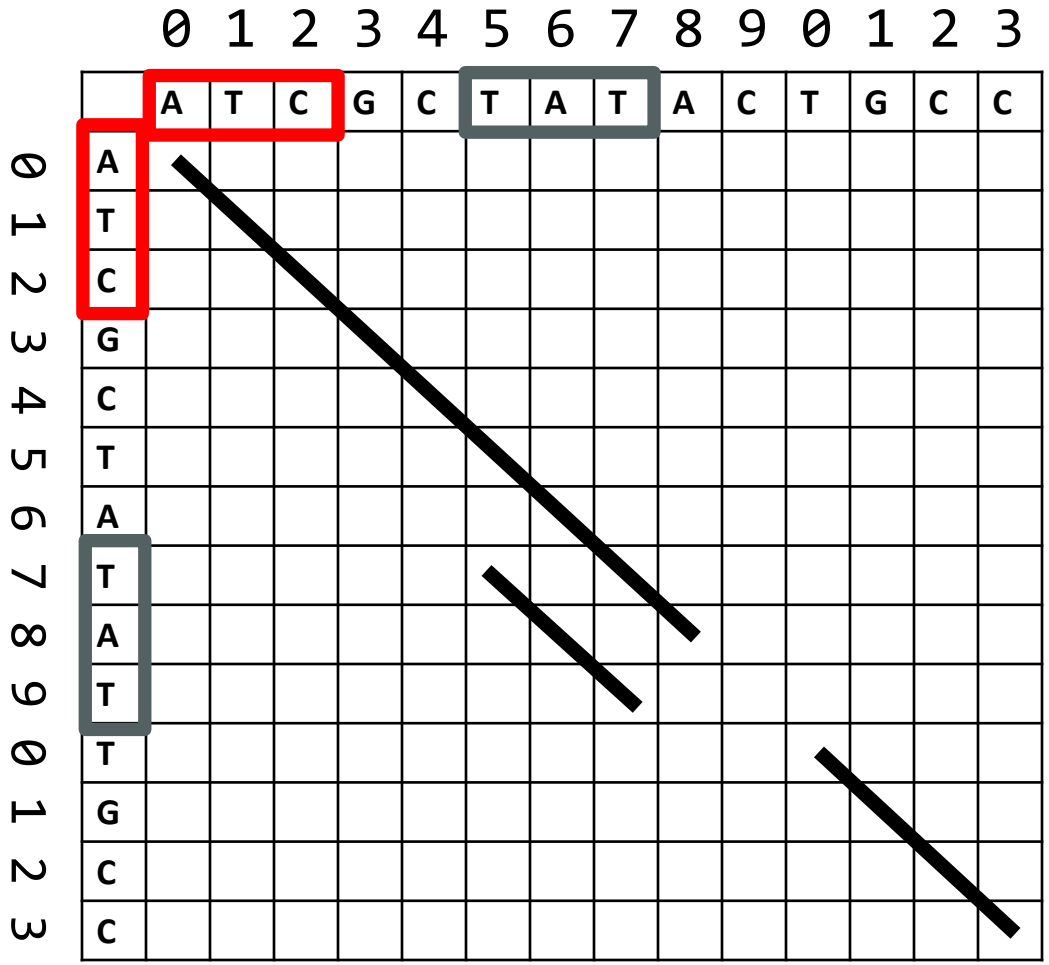

0 1 2 3 4 5 6 7 8 9 0 1 2 3
A T C G C T A T A C T G C C
0 A
1 T
2 C
3 G
4 C
5 T
6 A
7 T
8 A
9 T
0 T
1 G
2 C
3 C



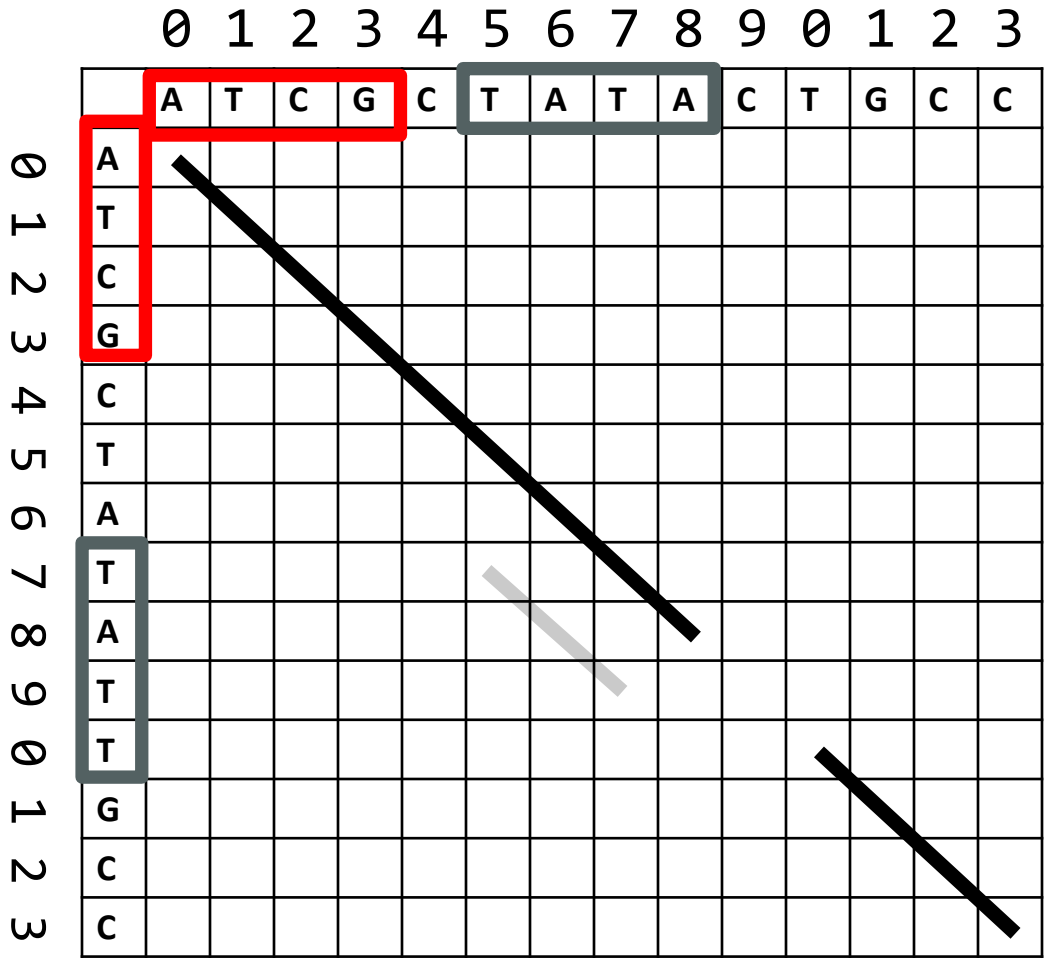
Window size = 2



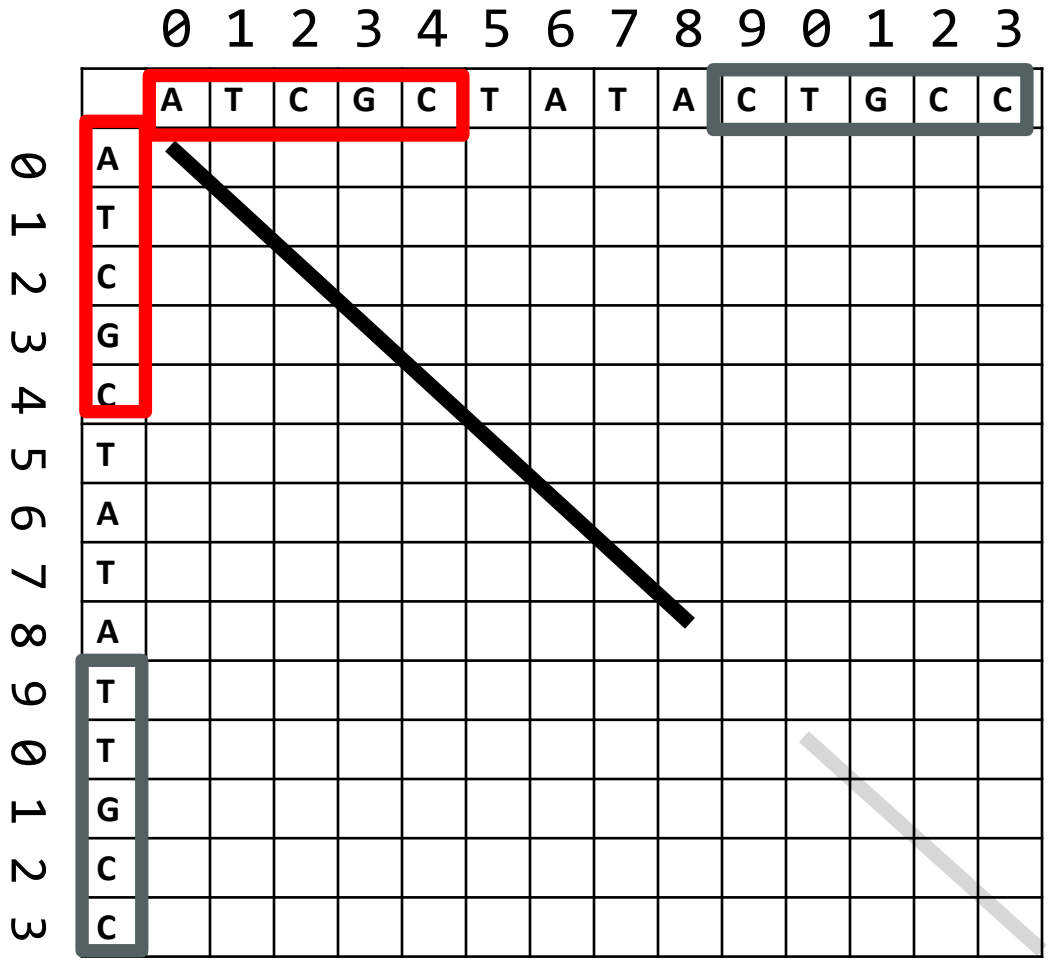
Window size = 3



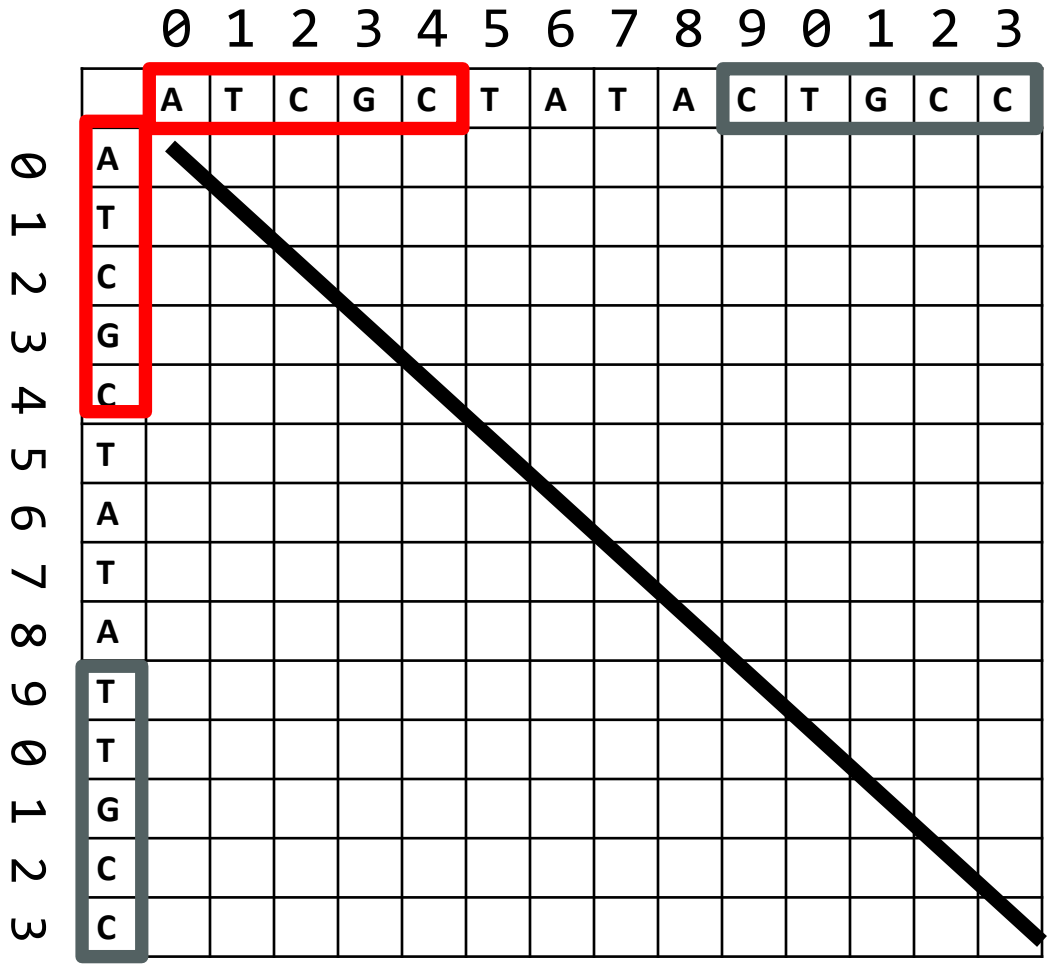
Window size = 3



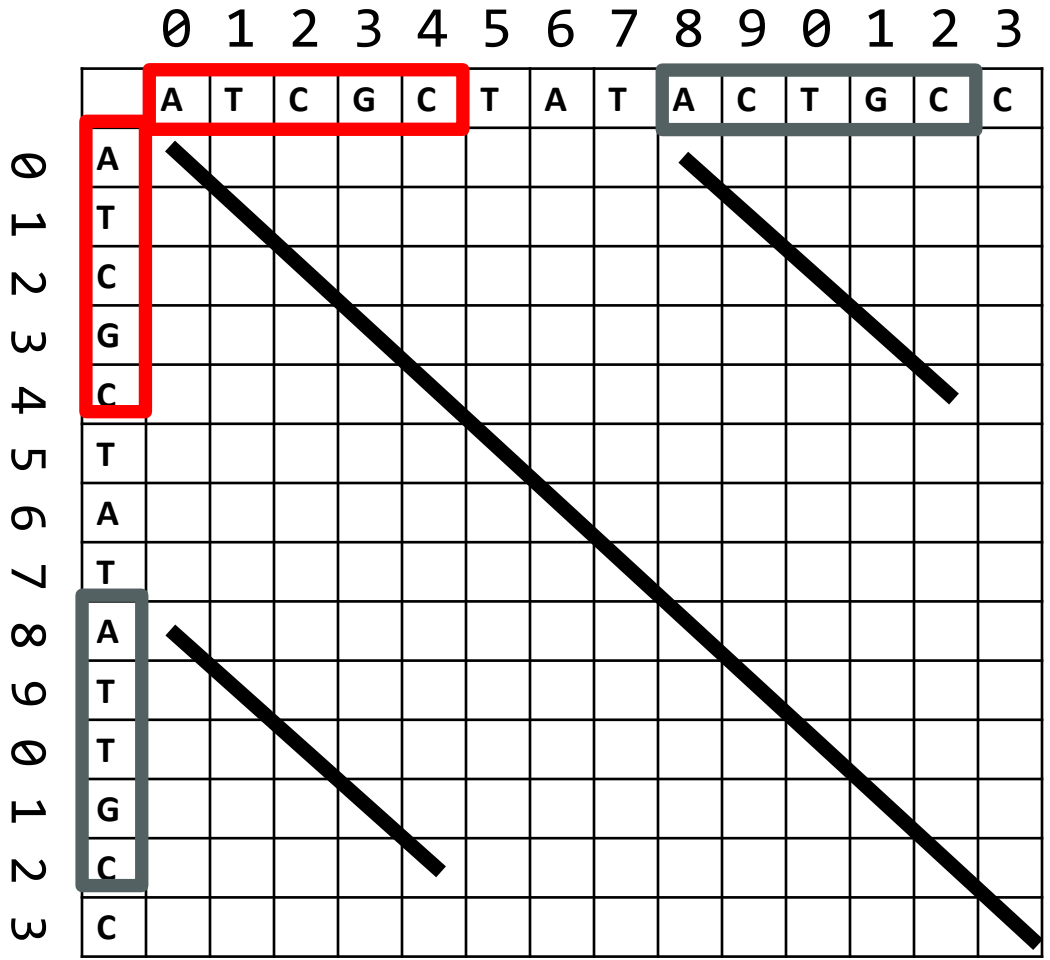
Window size = 4



Window size = 5



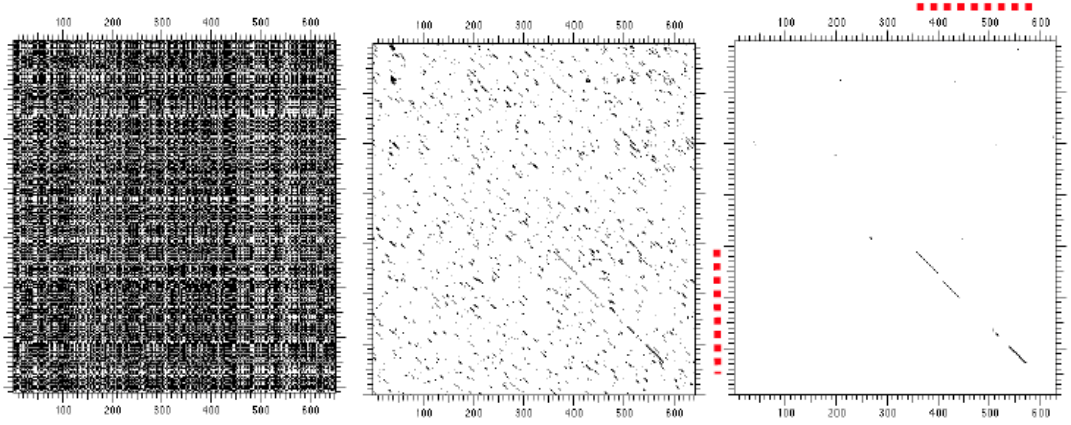
Threshold = 4 (allow 4 matches in a window) **Window size = 5**



Window size = 5 **Treshold = 4**

Window size and stringency

Dot Plots. Visualization



Window size=1
Stringency=1

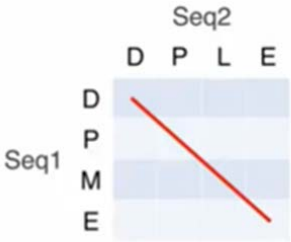
Window=11
Stringency=7

Window=25
Stringency=15

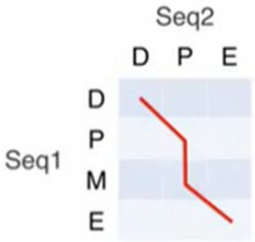
- With the **window size**, we can reduce the noise
- With the **treshold** (= ***stringency***), we can tolerate some changes for a certain window size

Constructing the alignment

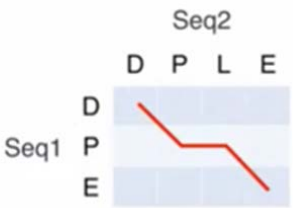
- Matches are represented by diagonal paths
- Indels are represented by horizontal or vertical paths




```
Seq1: D P L E
      | | : |
Seq2: D P M E
```

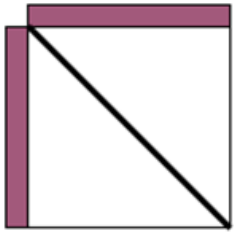



```
Seq1: D P M E
      | |   |
Seq2: D P - E
```

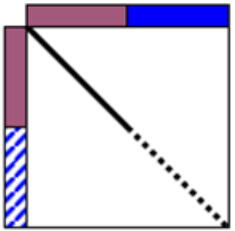



```
Seq1: D P - E
      | |   |
Seq2: D P L E
```

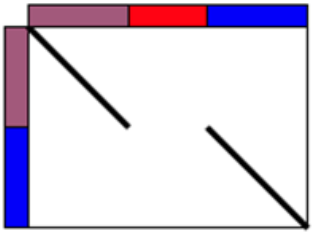
Patterns



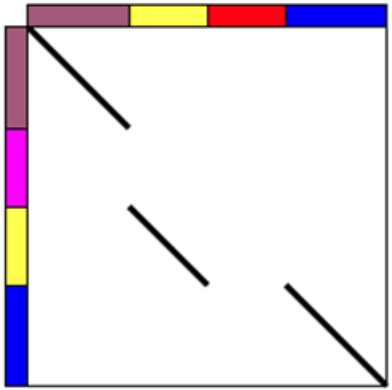
identical

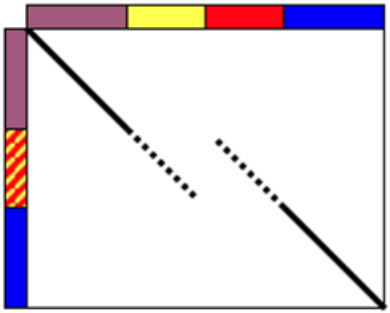


substitutions



gap





How can we compare two sequences?

- **Word methods**
- **Dot-matrix**
- **Evolutionary alignments (dynamic programming)**

PRACTICUM

- Implement algorithms
- Learn the foundations of some bioinformatics tools
- Practice programming

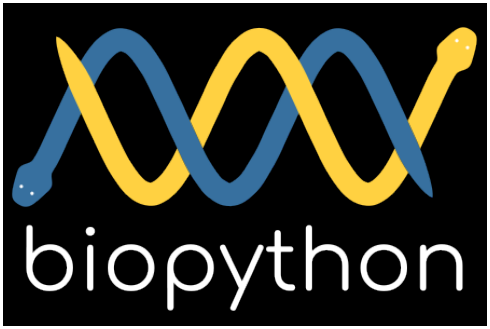
Getting the solution of proposed exercises is not important

What matters is the process!!!

How to read/write sequence files?

Biopython: a collection of free Python tools for computational biology

https://biopython.org/



- Access UniProt, NCBI resources
- Parse or run Clustal, Blast …
- Tools for sequence manipulation
- …

SeqIO module to read/write sequence files

Installation
Conda: $ conda install biopython
Ubuntu: $ sudo apt-get install python-biopython

Week 3 help notebook

Open **Jupyter Lab** or **Jupyter Notebook**

Run/complete: **help_week3.ipynb**

Topics:

- Read and write sequences (with and without Biopython)
- Exact matching (many finds)

PROGRAMMING EXERCICES ASAB

Requirements are indicated in parenthesis

Week 3

Scoring alignments

score_seqs

score_seqs_gap

move_gaps

move_gaps_scores

move_seq2

move_seq2_scores

Finding motifs

naive_match

many_finds

indexing

Dot matrix

dot_matrix_basic

https://acordomi.github.io/BDBI_ASAB/

Each exercise comes with tests (doctests)

```
$ python -m doctest score_seqs.py
$ python -m doctest -v score_seqs.py
```