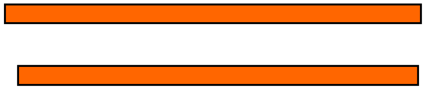
Lecture 6:
SEQUENCE COMPARISON

✓PAIRWISE ALIGNMENT

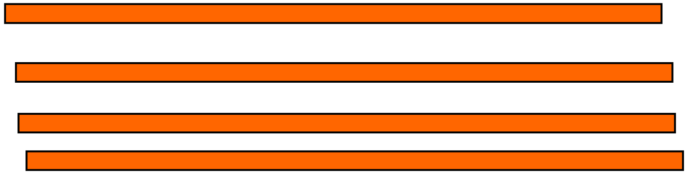
Part I: Basic Terminology

Types of Sequence Comparison

- Pairwise Alignment
  - Comparison of two sequences



- Multiple Alignment
  - Comparison of more than two sequences



What is sequence alignment or sequence comparison?

⌘ Given two sequences of letters, find best pairing from one sequence to letters of the other sequence.

1. THISISANEWVIRUSFORCOVID9
   THISISAVIRUSFORSARS

2. THISISANEWVIRUSFORSARSCOV
   THISISAVIRUSFORHUMANSARS

Aligning biological sequences

- DNA (4 letter alphabet)

1. TTGACAC

   TTTACAC

2. TTGACAC

   TTCACAC

3. TTGACAC

   TTACAC

Proteins (20 letter alphabet)

1. RKVA
   RKIA

2. RKVA
   RKKA

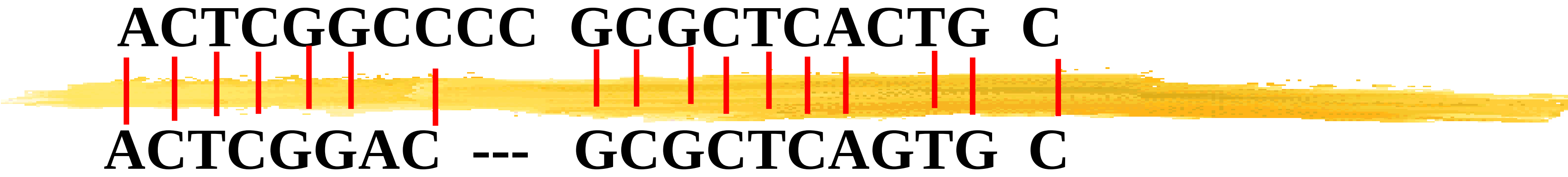
3. RKVA
   RKSA

4. RKVA
   RKQA

CONCEPTS IN SEQUENCE COMPARISON

1.IDENTITY

- Percentage identity between sequences means that they have a certain number of residues (nucleotide /amino- acids) that are identical at that particular position after aligning both sequences.



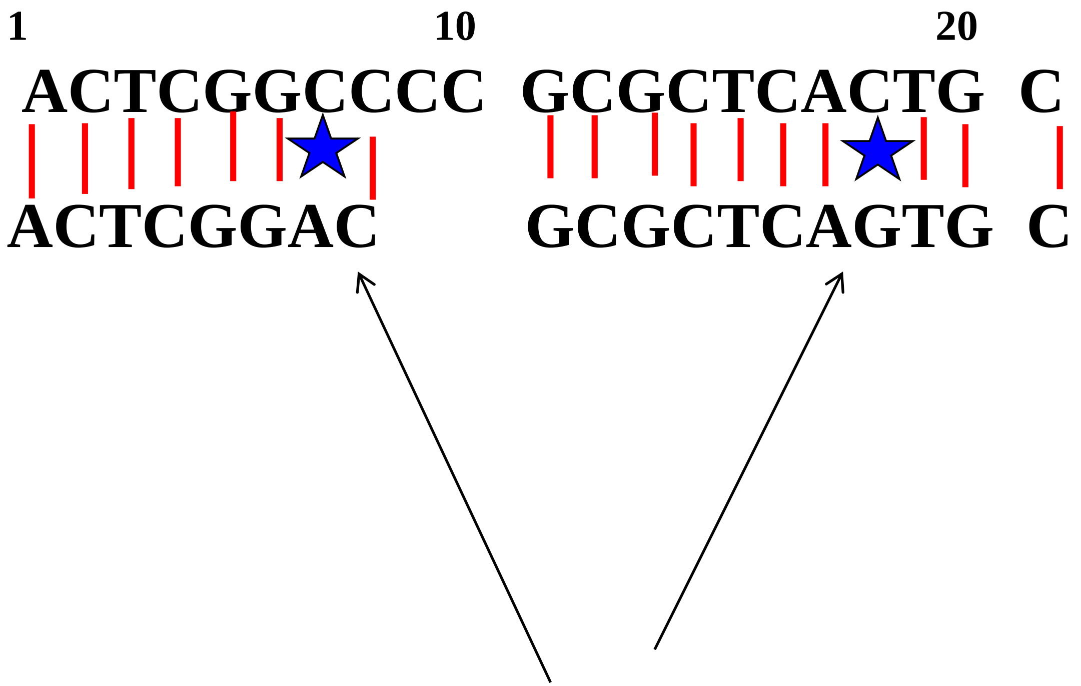
ACTCGGCCCC GCGCTCACTG C

ACTCGGAC --- GCGCTCAGTG C

- **Exact match (shown by |) : identical residues**
-
- **Percentage of identity:**

no. of identical matches

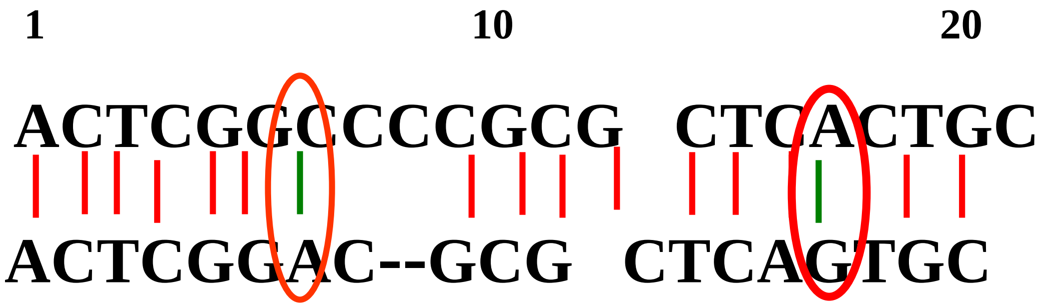
residues in the aligned sequence



```
1                  10                     20
 ACTCGGCCCC  GCGCTCACTG  C
ACTCGGAC      GCGCTCAGTG  C
```

MISMATCH HERE

2. Mismatch:different characters



```
1           10          20
ACTCGGCCCCGCG  CTCACTGC
ACTCGGAC--GCG  CTCAGTGC
```

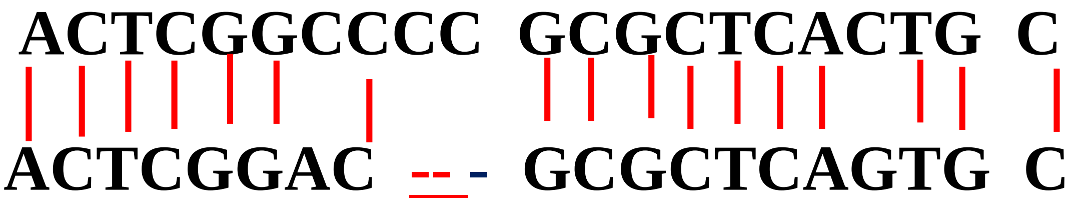
➢Substitution:

•Transversion : purine-pyrimidine or vice versa

•Transition : purine-purine or pyrimidine- pyimidine

•Substitution : less score than identical match

For eg: +1 per substitution



```
ACTCGGCCCC  GCGCTCACTG  C
ACTCGGAC --- GCGCTCAGTG  C
```

- 3. Gaps: no characters
- Gaps have penalties:
- Insertion of first gap(GAP OPENING) :
    - high penalty (For eg. -2)
- Insertion of consecutive gaps (GAP EXTENSION):
    - less penalty (For eg. -1 for each gap)
- More no. of gaps lower the score of the alignment

4. Similarity

```
1        10        20
ACTCGGCCCCGCGCTCACTGC
||||||||  |||||||||||
ACTCGGAC - -  GCGCTCAGTGC
```

- Similarity :
- Identical matches + Substitutions.

  No. of aligned residues

Similarity Vs. Homology

- 5. Homology: When two similar proteins come from a common ancestor.

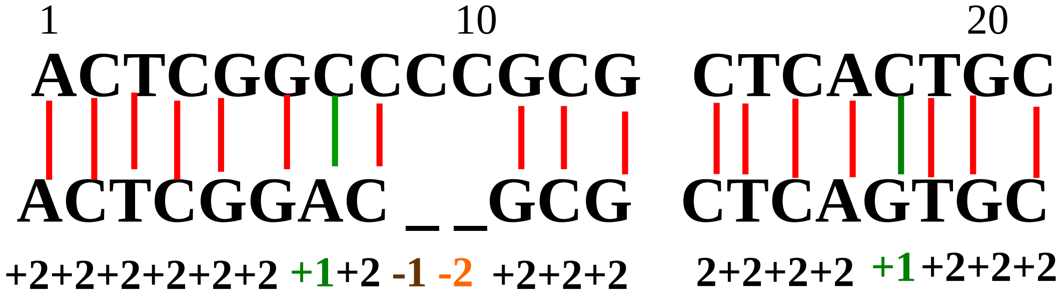
- Homology is inferred from Similarity.
- If two sequences are similar, then they are known as homologous sequences.
- Usually, at least 30% identity over 400 bp for DNA sequences and over 125 amino acids for proteins.

HOW TO SCORE ALIGNMENTS?

Alignments are scored to get idea which is the best possible alignment between two sequences.

Match: +2; Gap opening: -1: Gap extension: -2

Substitution: +1



1 10 20

ACTCGGCCCCGCG CTCACTGC

ACTCGGAC _ _GCG CTCAGTGC

+2+2+2+2+2+2 +1+2 -1 -2 +2+2+2 2+2+2+2 +1 +2+2+2

Score: ?

Orthologs   & paralogs

Evolution by sequence mutation

Gene sequence

ATGAAGGCTGCCTACGACTGCCGTG
ATGCAGGCTGCCTACGACTGCCGTG
ATGCAGGCTGCCAACGACTGCCGTG

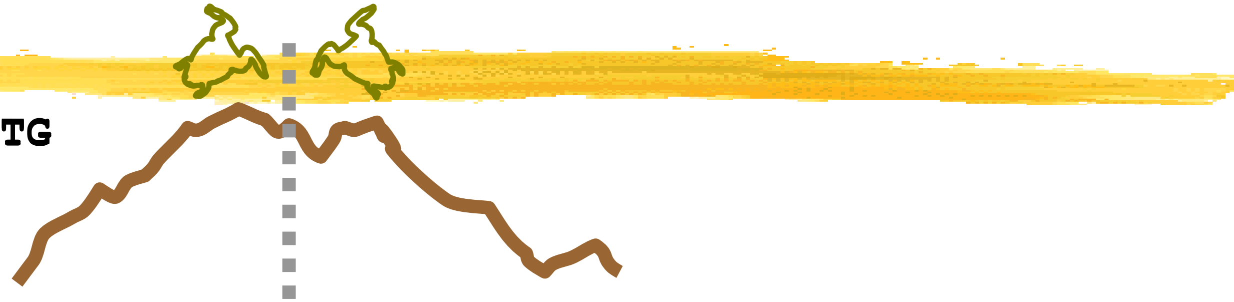
Imagine one mutation gets fixed every 100,000 years in this gene sequence…

ATG CGTG
ATGCATGCTGCCAACGACTGCCCTG
ATGCATGCTGCCAACGGCTGCCCTG
ATGCATGCTGCCAACGGATGCCCTG
ATGCATGCCGCCAACGGATGCCCTG
ATGCATGCCGCCAACGGATGTCCTG

Speciation

Gene A

ATGAAGGCTGCCTACGACTGCCGTG



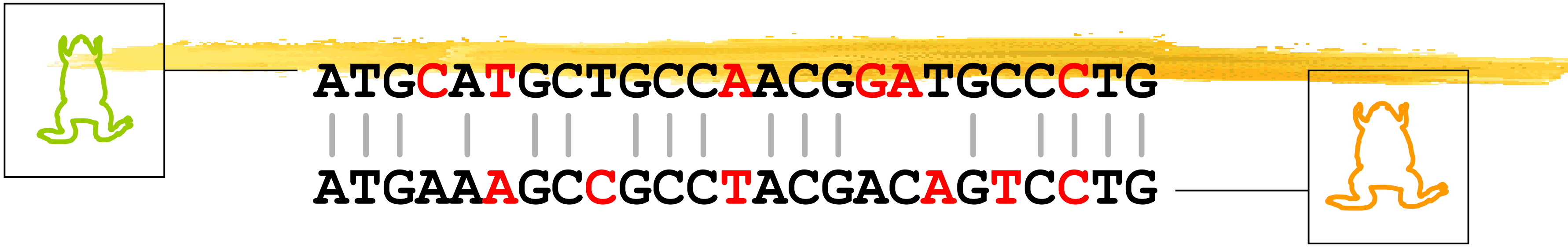
```
ATGCAGGCTGCCTACGACTGCCGTG
ATGCAGGCTGCCAACGACTGCCGTG
ATGCATGCTGCCAACGACTGCCGTG
ATGCATGCTGCCAACGACTGCCCTG
ATGCATGCTGCCAACGGCTGCCCTG
ATGCATGCTGCCAACGGATGCCCTG
```

```
ATGAAGGCCGCCTACGACTGCCGTG
ATGAAGGCCGCCAACGACTGTCGTG
ATGAAAGCCGCCAACGACTGTCGTG
ATGAAAGCCGCCAACGACAGTCGTG
ATGAAAGCCGCCTACGACAGTCGTG
ATGAAAGCCGCCTACGACAGTCCTG
```

```
ATGCATGCTGCCAACGGATGCCCTG
||| | || ||| |||   | ||||
ATGAAAGCCGCCTACGACAGTCCTG
```

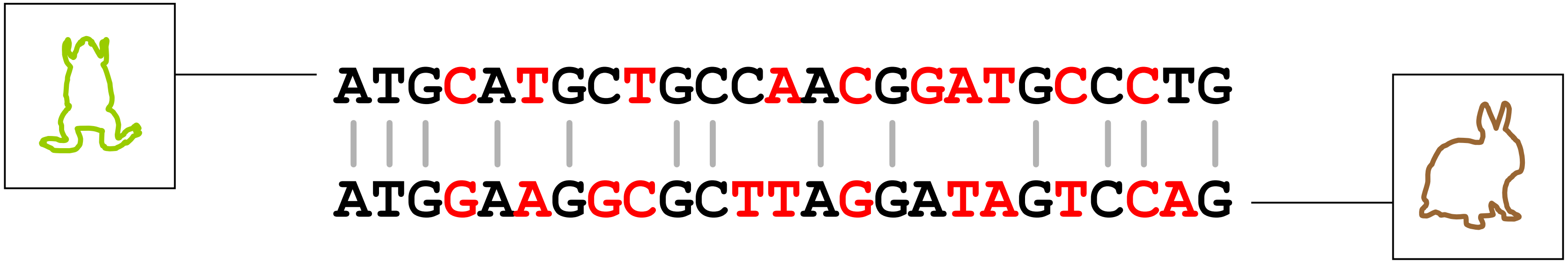
If the genetic difference means they can no longer interbreed, with fertile offspring – then we have a new species…

Residual Similarity



```
ATGCATGCTGCCAACGGATGCCCTG
||| | || ||| |||   | ||||
ATGAAAGCCGCCTACGACAGTCCTG
```

We can still easily detect residual similarity between these sequences, this is what we call ***homology*** – detectable similarity because of common evolutionary origin.

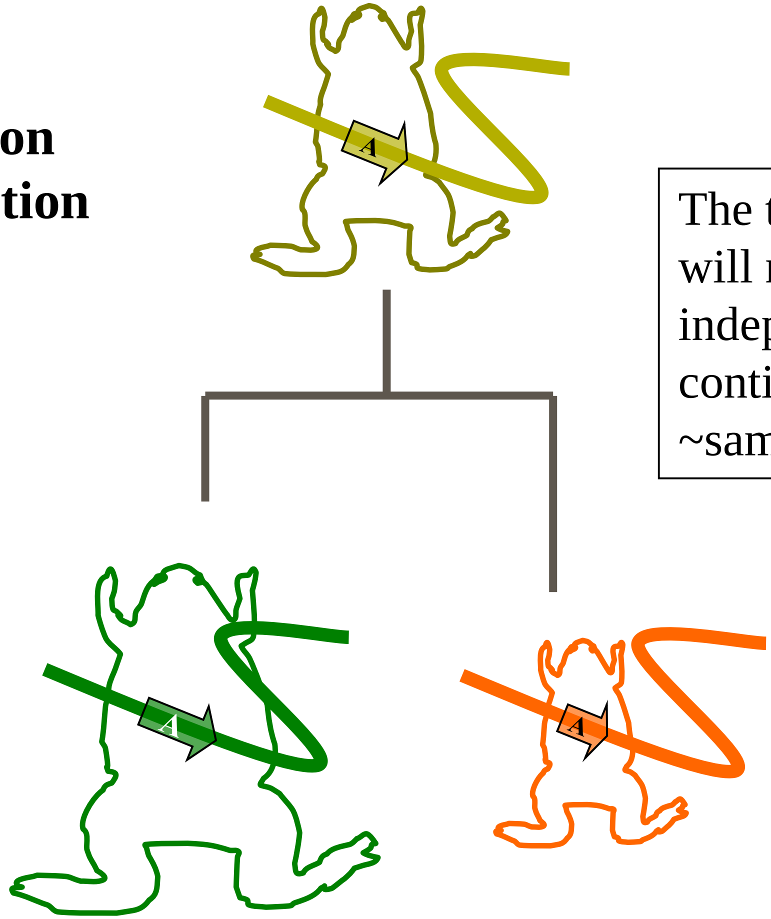


```
ATGCATGCTGCCAACGGATGCCCTG
||| | |  ||  | |   | || |
ATGGAAGGCGCTTAGGATAGTCCAG
```

After longer periods of evolution, homology may no longer be detectable in the DNA sequence…

Orthologs

Gene duplication through speciation



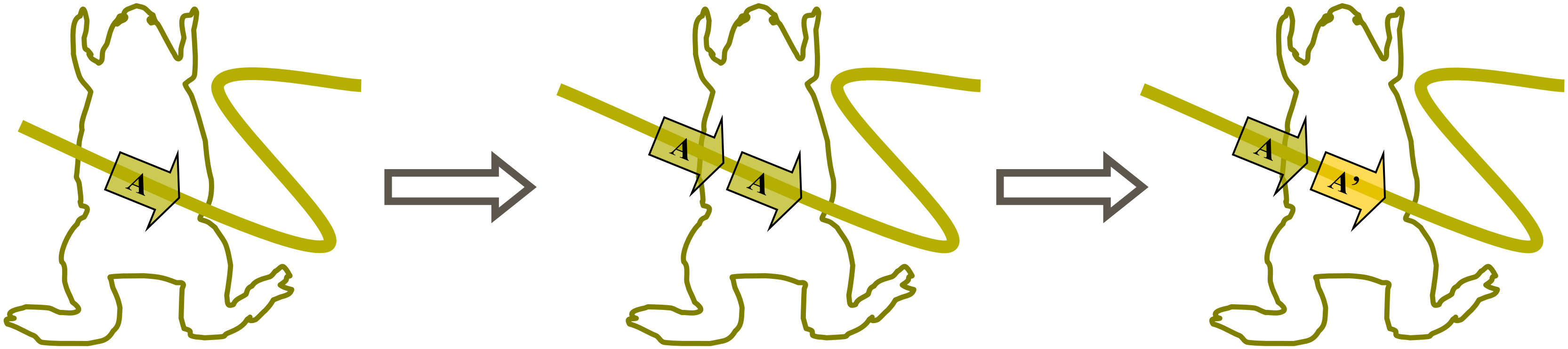
The two copies of Gene A will now evolve independently, but will continue to have the ~same function

They are ORTHOLOGS

Paralogs

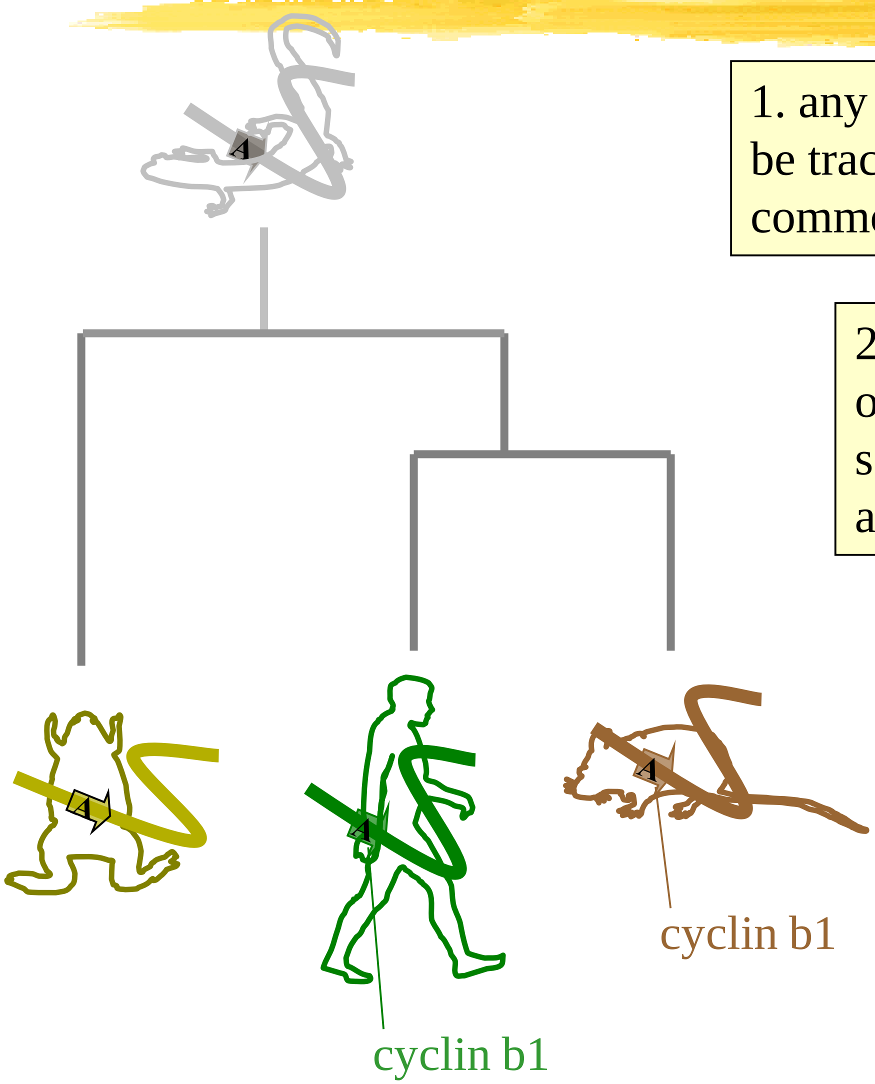
Gene duplication through internal genome duplication

The two copies of Gene A will now evolve independently, but will *probably not* continue to have exactly the same function



They are PARALOGS

The Essential Paradigm

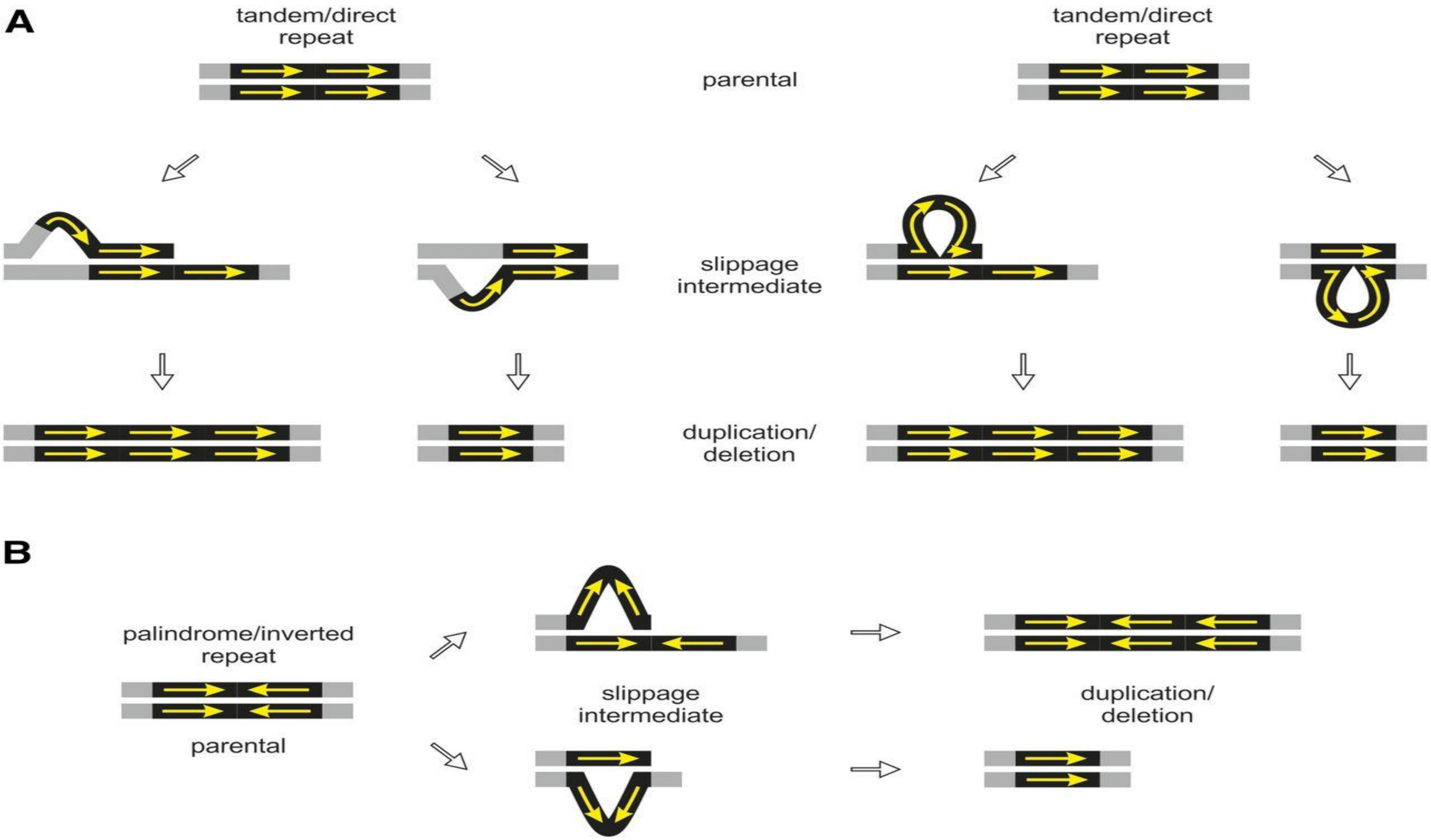


1. any group of modern species can be traced back to some extinct common ancestor

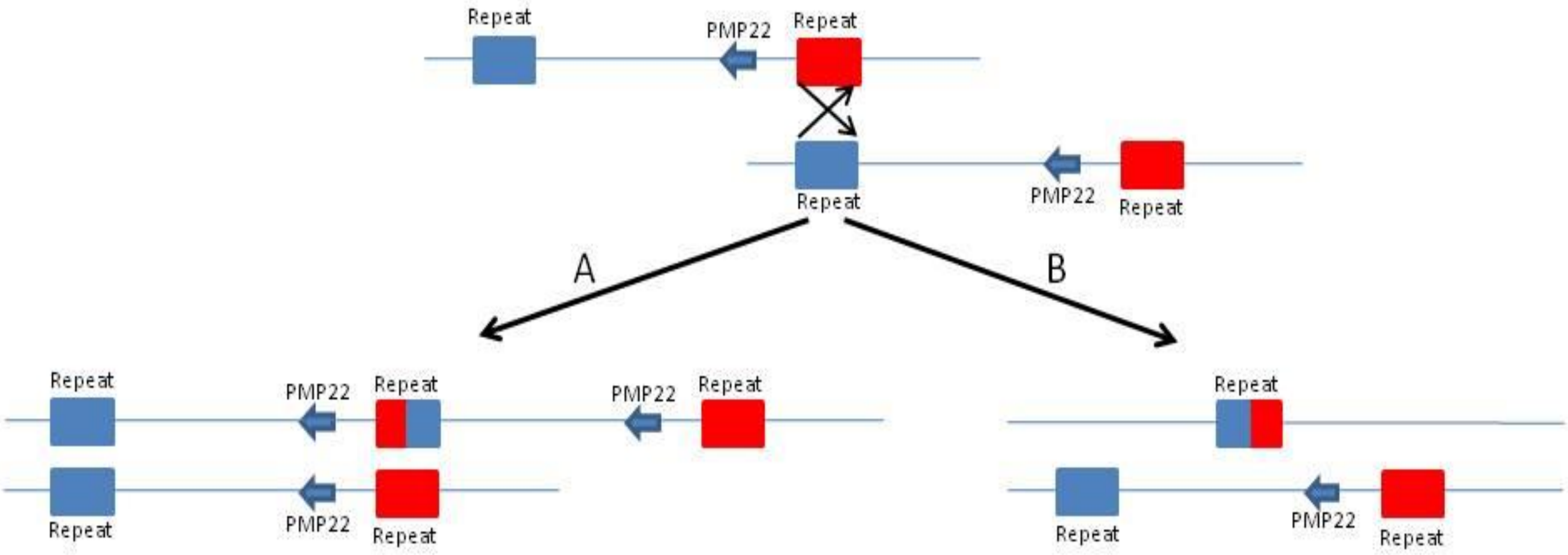
2. in all likelihood they share orthologous genes which have the same function in the modern animal as in the extinct ancestor

3. If we can experimentally determine the function of a gene in one of these organisms, then there is a good chance the ORTHOLOGOUS gene in another organism will have the same function

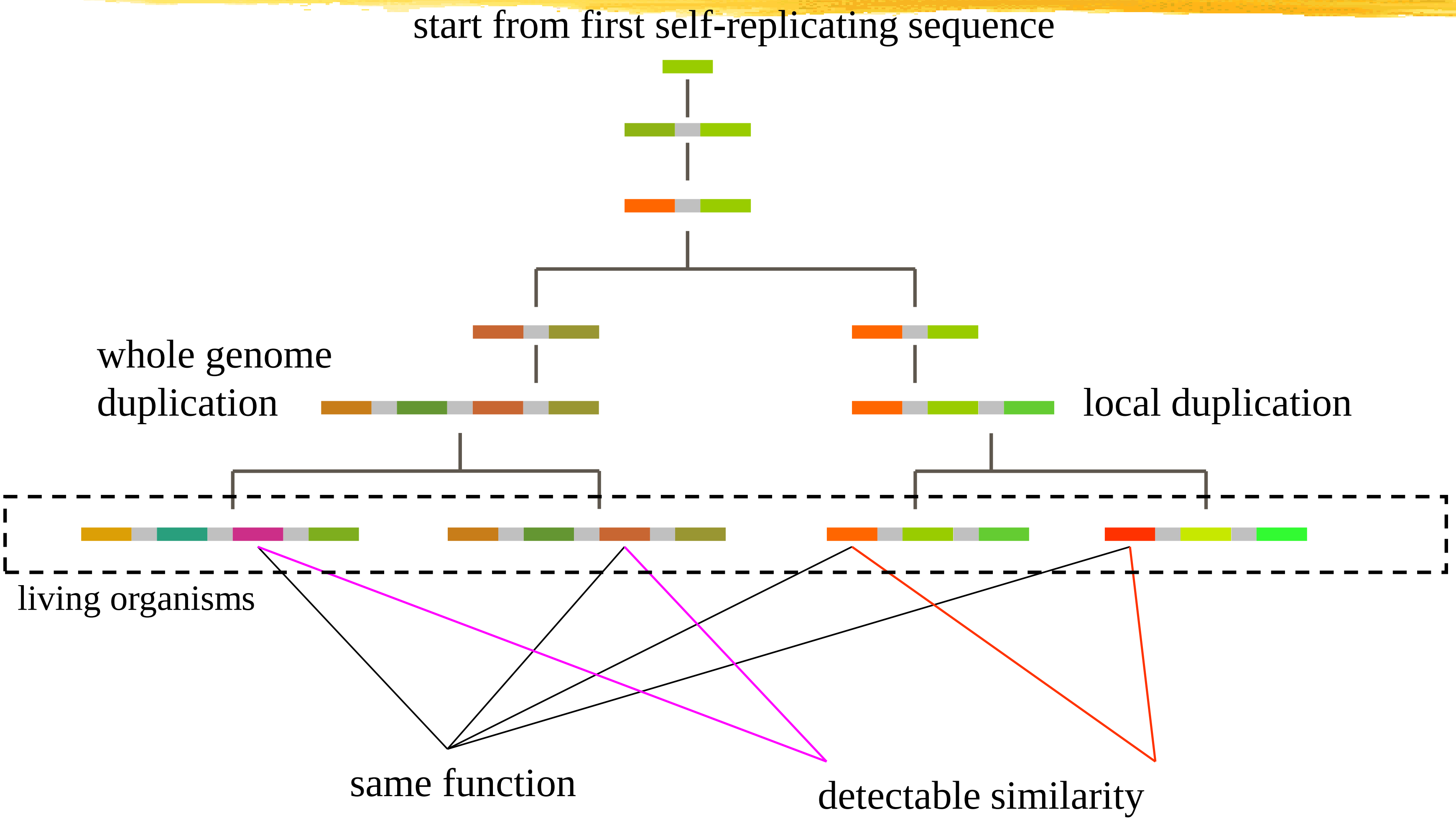
DNA duplication



DNA duplication



Function Conserved Longer than Detectable Similarity



Redundancy in the Genetic Code

GCA	A	alanine
GCC	A	
GCG	A	
GCT	A	
TGC	C	cystine
TGT	C	
GAC	D	aspartate
GAT	D	
GGA	G	glycine
GGC	G	
GGG	G	
GGT	G	

‘Synonymous’ or ‘silent’ mutations in the third position of the codon triplets have no effect on the amino acid coded for – so there is no evolutionary pressure against this…

Protein Similarity Persists Longer

```
CTATCACGAGAACCTGTG
CTATCCCGAGAACCTGTG
CTATCCCGAGAACCAGTG
CTATCCCGTGAACCAGTG
CTATCCCGTGAGCCAGTG
CTATCCCGTGAGCCAGTT
CTGTCCCGTGAGCCAGTT
```

```
CTATCACGAGAACCTGTG          LSREPV
|| || || || || ||           ||||||
|| || || || || ||   67%     ||||||   100%
CTGTCCCGTGAGCCAGTT          LSREPV
```

```
CTATCACGAGAACCTGTG          LSREPV
 | ||  |    || ||           ||| ||
 | ||  |    || ||   44%     ||| ||   80%
TTGTCCCGGTCGCCAGTT          LSRFPV
```

Always Compare Protein Sequences

DNA comparison   amino acid comparison

```
ATGAATGCAGCCTATGATTGCCGAGCCAGAATGCTAAGG      MNAAYDCRARMLR
||||| || || || || || || || ||||| || ||       | ||||||||+||
ATGAAGGCCGCATACGACTGTCGTGCTAGAATCCTGAGA      MKAAYDCRARILR
```

The DNA sequence can change while the amino acid sequence stays the same, so always look for similarities by comparing amino acid sequences.

Scoring matrix

Position Independent Matrices

PAM Matrices (Percent Accepted Mutation)

- Derived from observation; small dataset of alignments
- Implicit model of evolution
- All calculated from PAM1
- PAM250 widely used

BLOSUM Matrices (BLOck SUbstitution Matrices)

- Derived from observation; large dataset of highly conserved blocks
- Each matrix derived separately from blocks with a defined percent identity cutoff
- *BLOSUM62* - default matrix for BLAST

Position Specific Score Matrices (PSSMs)

PAM matrix

- PAM矩阵（Point Accepted Mutation） 基于进化的**点突变模型**，如果两种氨基酸替换频繁，说明自然界接受这种替换，那么这对氨基酸替换得分就高。
- 一个PAM就是一个进化的变异单位, 即**1%的氨基酸改变，**但这并不意味100次PAM后，每个氨基酸都发生变化，因为其中一些位置可能会经过多次突变，甚至可能会变回到原来的氨基酸。
- **一个PAM-N矩阵元素（i，j）的值：**
  反应两个相距N个PAM单位的序列中第i种氨基酸替换第j种氨基酸的频率。

序列相似度 = 40% 50% 60% | | |
打分矩阵 = PAM120 PAM80 PAM60

PAM250 → 14% – 27%

摘自hugeheadhuge博客

BLOSUM matrix

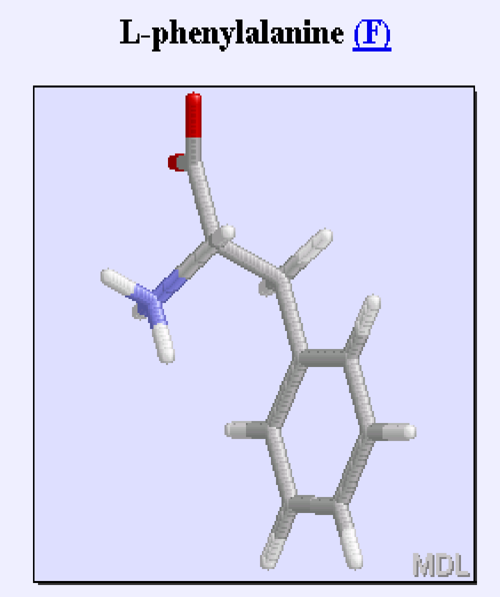
- 源于蛋白质模块（BLOCKS） 数据库
- BLOSUM：首先寻找氨基酸模式，即有意义的一段氨基酸片断（如一个结构域及其相邻的两小段氨基酸序列），分别比较相同的氨基酸模式之间氨基酸的保守性（某种氨基酸对另一种氨基酸的取代数据）

- 以所有 60％保守性的氨基酸模式之间的比较数据为根据，产生BLOSUM60；

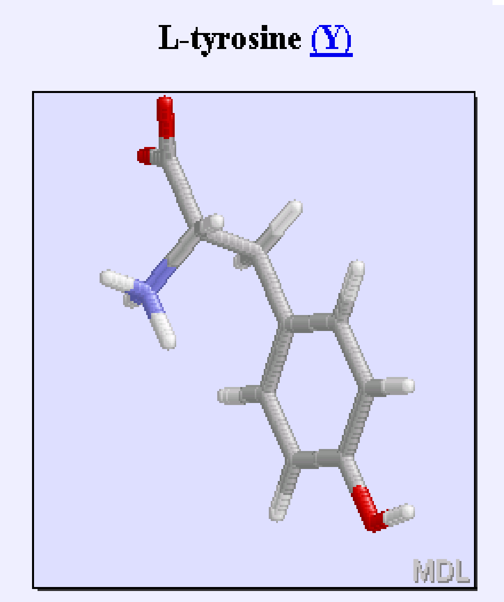
- 以所有80％保守性的氨基酸模式之间的比 较数据为根据，产生BLOSUM80。

BLOSUM62 Substitution Matrix

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



L-phenylalanine (F)



L-tyrosine (Y)

mo

low weights

ar

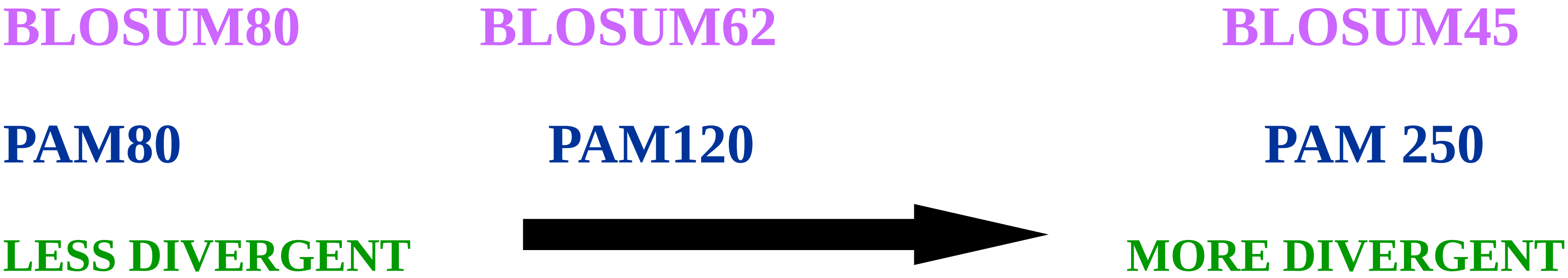
h weights

Negative for less likely substitutions

Positive for more likely substitutions

SCORING SCHEMES FOR PROTEIN SEQUENCE LIGNMENTS

- Scoring matrices used are: **PAM** and **BLOSUM**



BLOSUM80 BLOSUM62 BLOSUM45

PAM80 PAM120 PAM 250

LESS DIVERGENT → MORE DIVERGENT

USE OF DIFFERENT MATRICES WILL GIVE DIFFERENT RESULTS

TRY AND SEE!!!

SUMMARY

- **TODAY WE LOOKED :**
- Concepts
  - Identity & Similarity;
  - mismatch & gap;
  - homolog, ortholog, paralog
- how to score
- Scoring matrix
  - PAM
  - BLOSUM

课堂讨论思考题

- How to find the best alignment among tons of sequences?