DNA v genomu určitého organismu obsahuje 30% thyminu. Jaký bude obsah cytosinu?

RNA v transkriptomu určitého organismu obsahuje 18% guaninu. Jaký bude obsah cytosinu v cDNA?

RNA v transkriptomu určitého organismu obsahuje 32% adeninu. Jaký bude obsah uracilu v cDNA?

myšleno jako % bází

Doplňte druhé vlákno DNA, tak aby vznikla dvouvláknová DNA:

GGATATCCGA

Jaké je komplementární vlákno DNA k sekvenci:

AAGTTCC

Jaká je komplementární sekvence k sekvenci GGATCC?

Co je na této sekvenci zajímavé? Jaké vlastnosti by mohl mít protein, který se váže na DNA s takovou sekvencí?

Definujte:

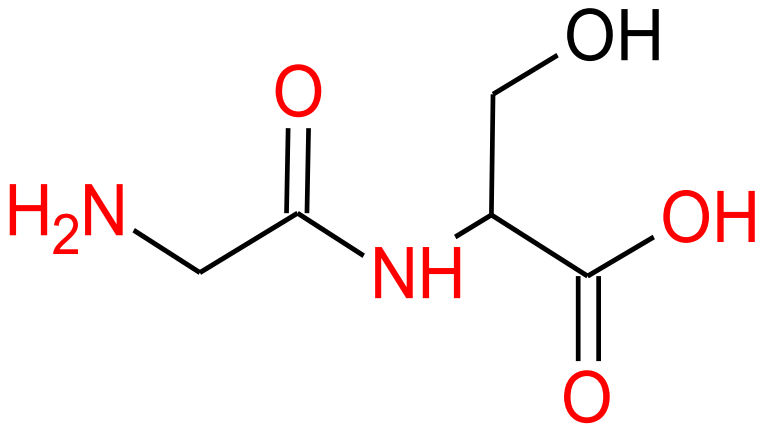
nukleotid
nukleosid

V jakém směru se zapisují sekvence:

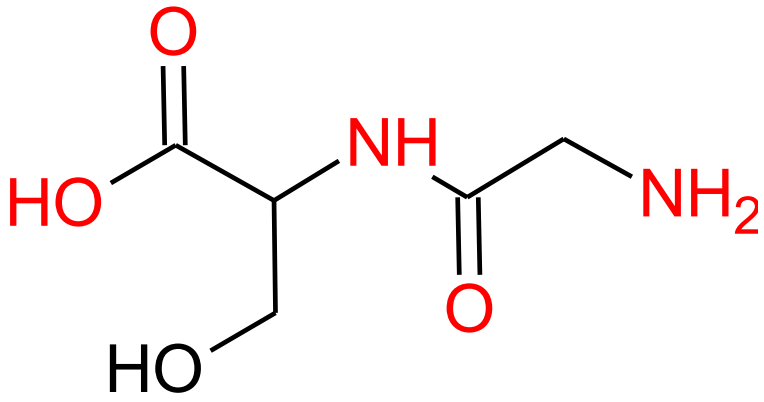
nukleových kyselin
aminokyselin
cukrů Glc-Gal-Man

Jak se jmenuje dipeptid:

A)



B)



Která sekvence koduje peptid Gly-Arg-Glu-Glu-Asn
(použijte kodovací tabulku)

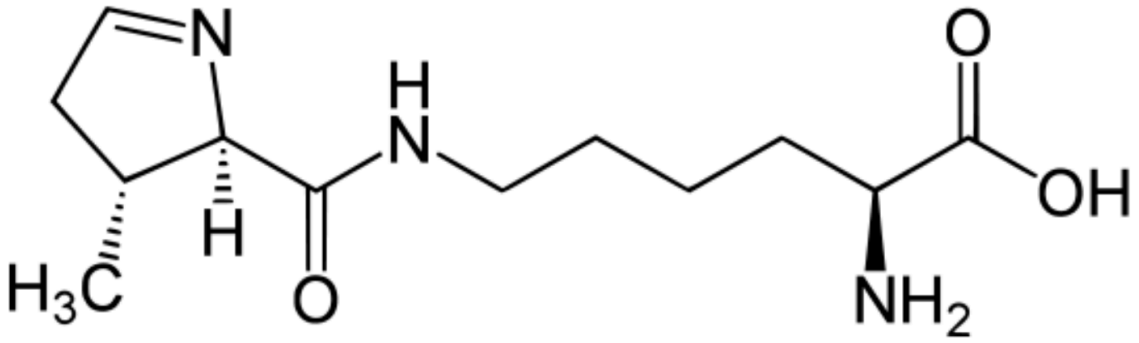
A) AGC AGA CCT TAC
B) GTT ATA AAT TAT CAT
C) GGC AGA GAG GCC TAA
D) GGC CGC GAA GAG AAC TAA

	Druhá báze							
První báze	U		C		A		G	
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

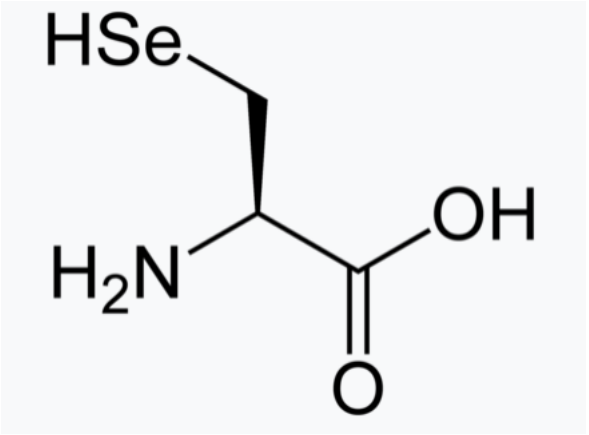
Znáte jednopísmenné zkratky aminokyselin?

A	**C**	**G**
K	**D**	**N**
L	**E**	**Q**
P	**I**	**V**
Y	**F**	**S**
W	**T**	**X**

pyrrolysin Pyl (O), UAG



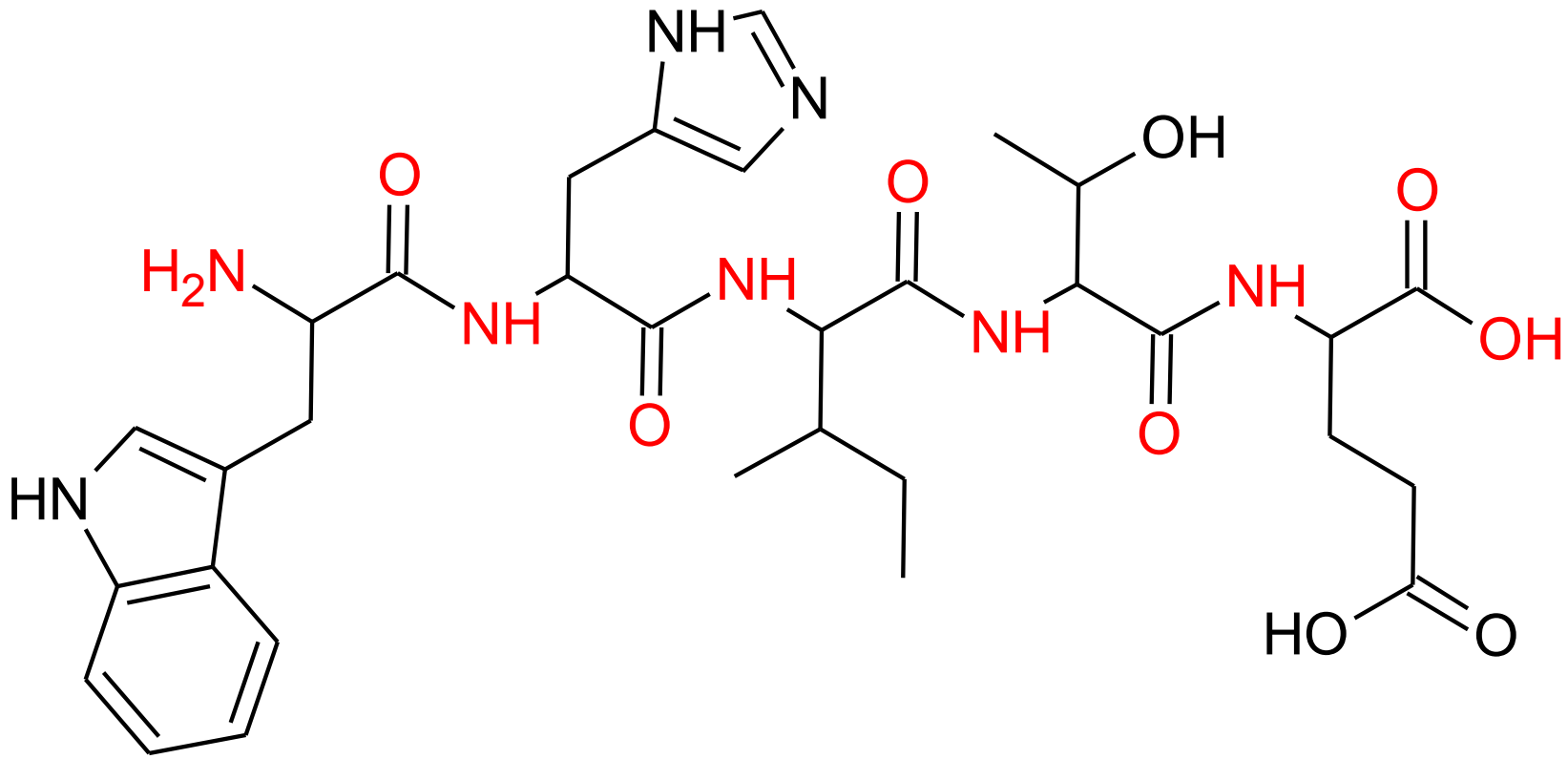
selenocystein Sec (U), UGA



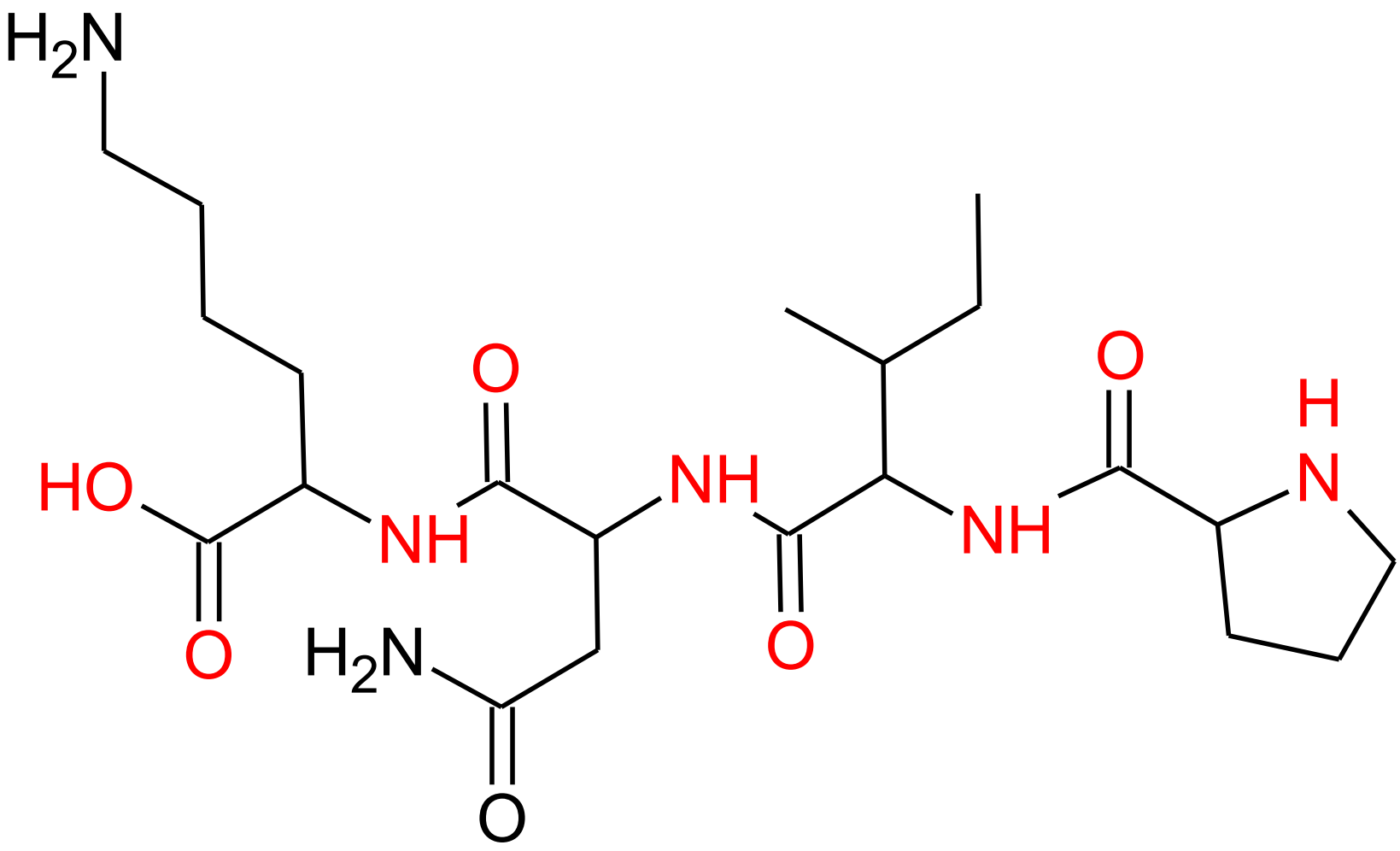
N-formylmethionin fMet, AUG

ASX-B (D/N)
GLX-Z (E/Q)
Xle-J (I/L)

Zapište aminokyselinovou sekvenci u následujících struktur



Zapište aminokyselinovou sekvenci u následujících struktur



Jak by asi vypadal alignment těchto dvou sekvencí:

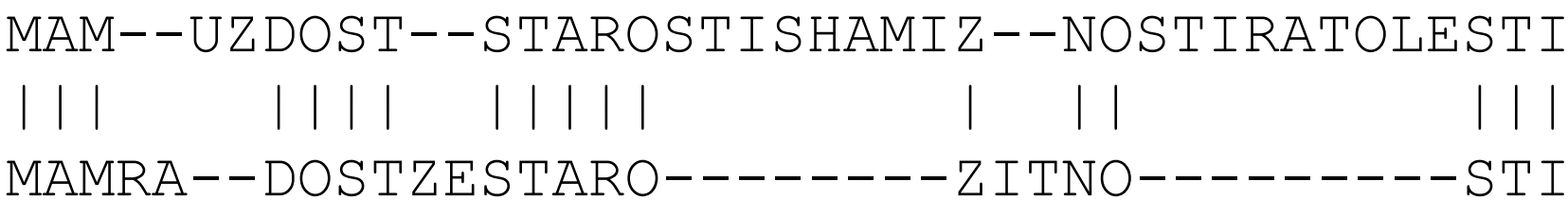
MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI
MAMRADOSTZESTAROZITNOSTI

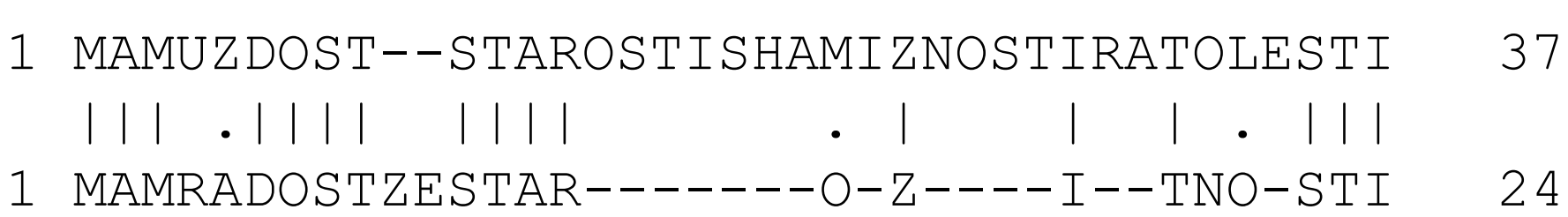
při absolutnim preferování
A) globálního alignmentu

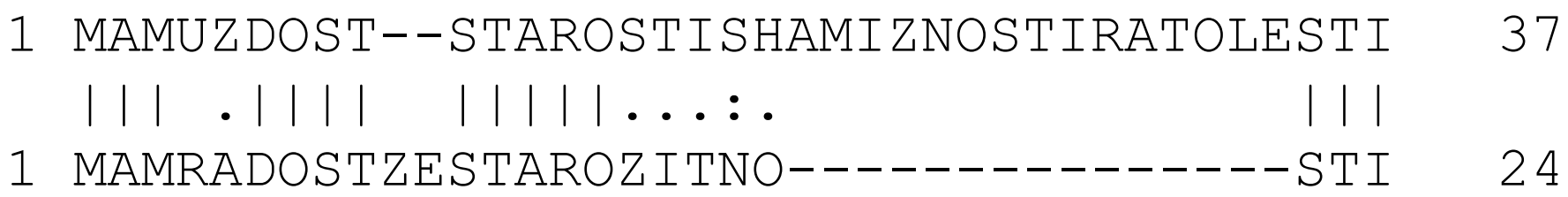
MAM--UZDOST--STAROSTISHAMIZ--NOSTIRATOLESTI
MAMRA--DOSTZESTARO---------ZITNO---------STI

B) lokálního alignmentu

MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI
MAMRADOSTZESTAROZ------ITNOSTI


MAM--UZDOST--STAROSTISHAMIZ--NOSTIRATOLESTI
MAMRA--DOSTZESTARO--------ZITNO---------STI


1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI 37
1 MAMRADOSTZESTAR--------O-Z----I--TNO-STI 24


1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI 37
1 MAMRADOSTZESTAROZITNO---------------STI 24

Co na to EMBOSS stretcher?

```
MAM--UZDOST--STAROSTISHAMIZ--NOSTIRATOLESTI
|||    ||||  |||||         |  ||          |||
MAMRA--DOSTZESTARO---------ZITNO----------STI


1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI   37
  ||| .|||||  ||||        . |    |  | . |||
1 MAMRADOSTZESTAR--------O-Z----I--TNO-STI   24
```

Gap_penalty: 1

Extend_penalty: 2

Score: 55

```
1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI   37
  ||| .|||||  |||||...:.                |||
1 MAMRADOSTZESTAROZITNO----------------STI   24
```

Gap_penalty: 12

Extend_penalty: 2

Score: 4

```
1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI   37
  ||| .|||| |||| . |  | | . |||
1 MAMRADOSTZESTAR-------O-Z----I--TNO-STI   24
```

Gap_penalty: 1

Extend_penalty: 2

Score: 55

```
1 MAMUZDOST--STAROSTISHAMIZNOSTIRATOLESTI   37
  ||| .|||| |||||...:.                |||
1 MAMRADOSTZESTAROZITNO---------------STI   24
```

Gap_penalty: 12

Extend_penalty: 2

Score: 4

```
1 MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI     37
  ||| .| ||           |. :..... ..|||
1 MAMRADOST----------ZESTAROZITNOSTI        24
```

Gap_penalty: 25

Extend_penalty: 2

Score: -11

Je tedy vhodnější:

Vysoká penalizace mezer:

Hledání sekvencí velmi striktně zaměřených na podobnost s hledanou sekvencí - najde oblasti velmi příbuzných sekvencí

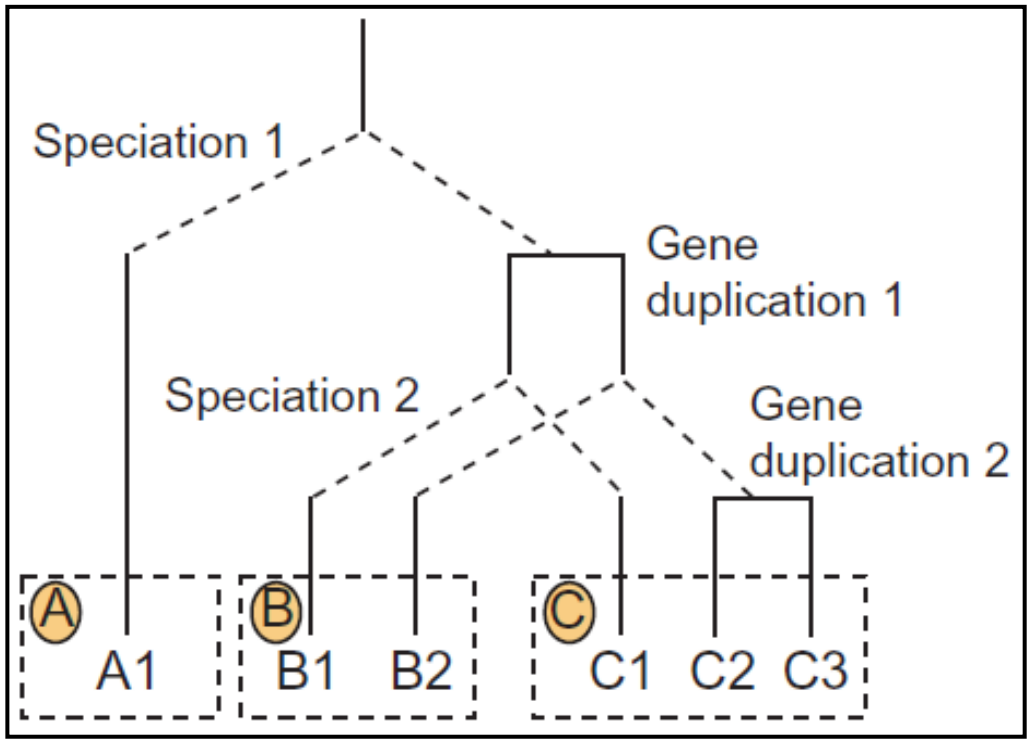
Nízká penalizace mezer:

Hledání podobností mezi sekvencemi vzdáleně příbuzných.

Jaký je rozdíl mezi:
„homology“ a „similarity“

MAMUZDOSTSTAROSTISHAMIZNOSTIRATOLESTI
MAMRADOSTZESTAROZITNOSTI

Jaký je rozdíl mezi:
„ortholog“ a „paralog“



Na čem je založeno vyhodnocení „kvality“ sekvenčního přiložení? Scoring alignments

snaha o co nejvyšší skóre:

1. identita (identity)
2. podobnost (similarity)
3. mezery (gaps)

Platí u nukleových kyselin i proteinů stejná pravidla ?

Nukleové kyseliny nemá smysl posuzovat podobnost:

sice **tranzice** (R→R or Y →Y) je mnohem častější než **transverze**(R→Y or Y →R), což ale není pro alignement užitečné

Frekvence mutací všech bází je obdobná, takže nejjednodušší hodnocení je: shoda (1), neshoda (0)

tím se nerozliší výborný alignment krátkých a mizerný dlouhých sekvencí: proto penalizace záměn:

match score +5

mismatch score -4

gap penalty (changeable parameter) opening -10, extending -2

Proteiny (similarity vs. identity): proč je bereme v úvahu?

$$S = [(L_s \times 2)/(L_a + L_b)] \times 100$$

number of aligned residues with similar characteristics

total lengths of each sequence

Skórování proteinového přiložení

substituční matrice (a z nich odvozeny **skórovací matrice**)

Reflektuje fyzikálně chemické vlastnosti jednotlivých aminokyselin ale zároveň i pravděpodobnost, že dojde k substituci konkrétní aminokyseliny za jinou konkrétní v průběhu evoluce.

Napadají Vás některé, které budou pravděpodobně vysoce „penalizovány“?

Substituční matrice

víceméně dva typy:

1. založené na záměnnosti genetického kódu nebo vlastností aminokyselin
2. odvozené z **empirických** studií aminokyselinových substitucí (přesnější)

Nejvíce používané jsou empirické matrice

PAM a BLOSUM

PAM – Point Accepted Mutation

Constructed by Margaret Dayhoff in 1978.

Zahrnuje pravděpodobnost záměny jedné aminokyseliny v druhou během evoluce

Předpokládá, že každá další mutace nezávisí na předchozí.

Odvozena z globálního alignmentu rodin proteinů (Podobnost sekvencí v rodině > 85%)

- vysoká spolehlivost alignmentu
- vysoká pravděpodobnost, že záměna aminokyseliny je dána jedinou mutací

Vypočtena pravděpodobnost s jakou jedna AA se změní na jakoukoliv jinou

PAM matrice

	A	R	N	D	C
A	9867	2	9	10	3
R	1	9913	1	0	1
N	4	1	9822	36	0
D	6	0	42	9859	0
C	1	1	0	0	9973

All entries × 10^4

PAM1

Byla vypočtena na základě 1572 změn v aminokyselinovém složení v 71 proteinových rodinách

PAM1 reflektuje průměrnou záměnu 1% všech aminokyselinových pozic

PAM250 (20% identita) je odvozena od PAM1 její 250-tinásobnou multiplikací (250 mutací na 100 aminokyselin)

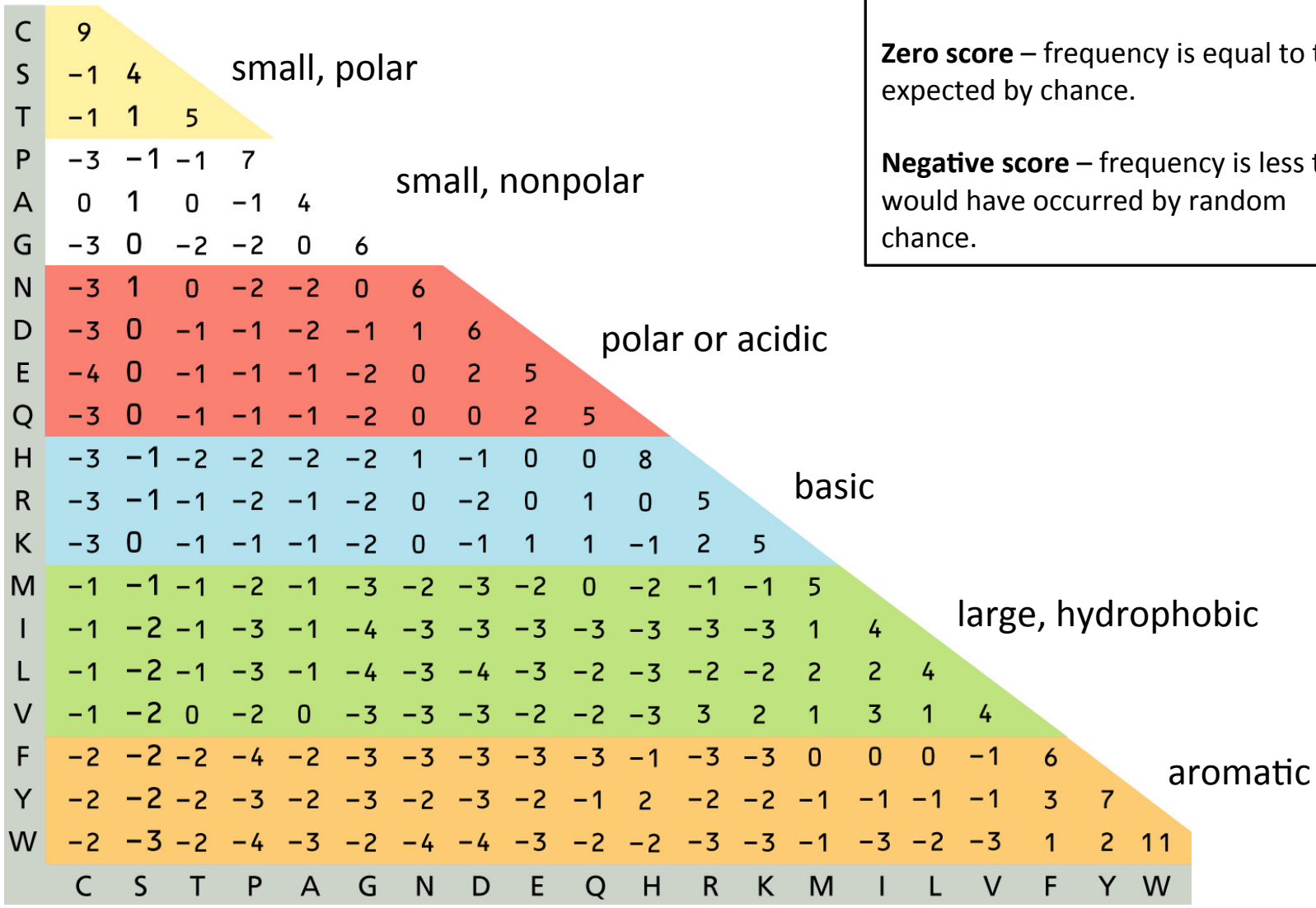
Vyšší číslo PAM matrice znamená větší evoluční vzdálenost

PAM250 matrice

Positive score – frequency of substitutions is greater than would have occurred by random chance.

Zero score – frequency is equal to that expected by chance.

Negative score – frequency is less than would have occurred by random chance.



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

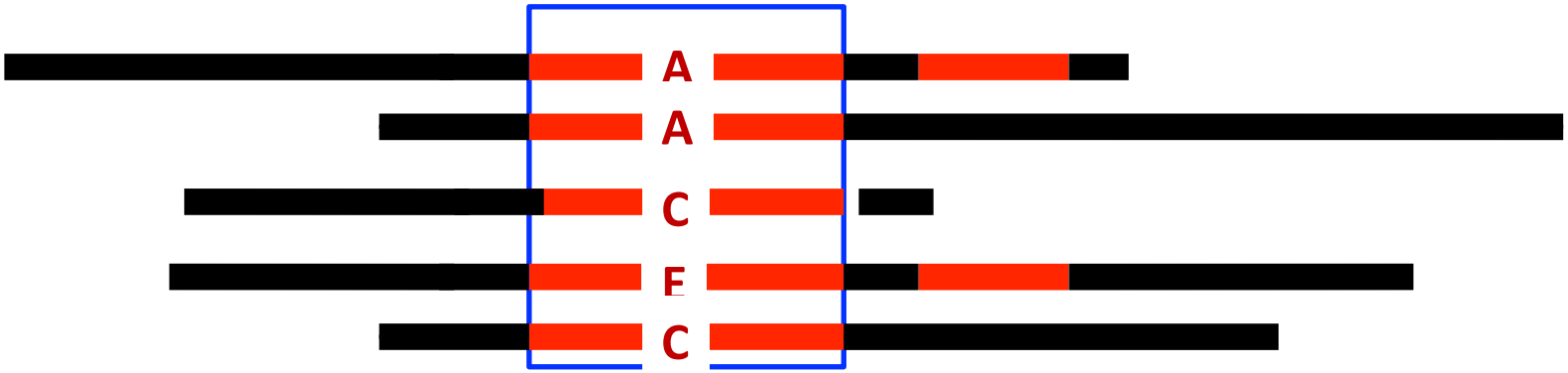
- Several assumptions you should be aware of:
  - Mutation of AA is independent of previous mutations on the same position (Markov process requirement).
  - Only PAM1 was “measured”, all other are extrapolations (i.e. predictions based on some model).
  - All sites are assumed to be mutable equally.
  - Mutations are assumed to be independent of surrounding residues.
  - Forces responsible for sequence evolution over short time are the same as these over longer times.
  - PAM matrices are based on protein sequences available in 1978 (bias towards small, globular proteins)
    - New generation of Dayhoff-type – e.g. PET91

BLOSUM (Blocks Amino Acid Substitution)

- 1992, Henikoff and Henikoff
- database BLOCKS– používá koncept „bloků“ k identifikaci proteinových rodin
- **sekvenční motiv**
  - konzervovaný aminokyselinový úsek conserved stretch of amino acids spojený se specifickou funkcí proteinu
- **sekvenční blok**
  - spárované motivy ze stejné proteinové rodiny bez mezer
- BLOSUM matrice byly vytvořeny na základě substitučních vzorů více než > 2 000 bloků (< 60 residuí) z 500 skupin proteinů

- nebere v potaz evoluci

- BLOSUM62 – znamená, že ke konstrukci matrice byly použity proteiny s průměrnou identitou 62%.




```
A - C = 4
A - E = 2
C - E = 2
A - A = 1
C - C = 1
```

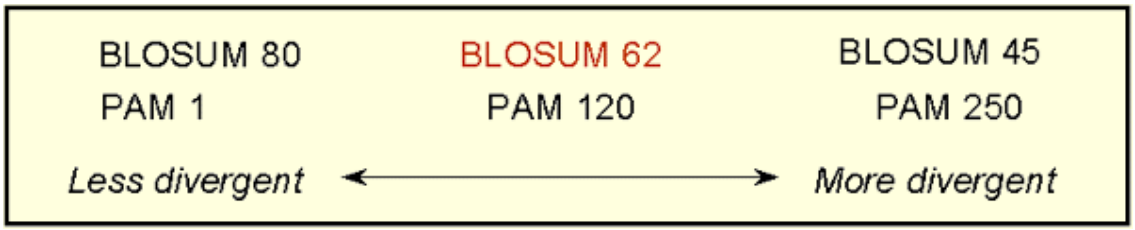
- výskyt každého páru AA v každém sloupci každého bloku je sečten
- čísla získána ze všech bloků slouží pro výpočet BLOSUM matricí

- Číslování BLOSUM jde v obráceném pořadí oproti PAM
  - čím menší číslo, tím odlišnější sekvence byly použity

Matrix	Best use	Similarity (%)
Pam40	Short highly similar alignments	70-90
PAM160	Detecting members of a protein family	50-60
PAM250	Longer alingments of more divergent sequences	~30
BLOSUM90	Short highly similar alignments	70-90
BLOSUM80	Detecting members of a protein family	50-60
BLOSUM62	Most effective in finding all potential similarities	30-40
BLOSUM30	Longer alingments of more divergent sequences	<30

Similarity column gives range of similarities that the matrix is able to best detect.

Odlišné substituční matrice jsou pro odlišné účely



more stringent

less stringent

- BLOSUM matrice pracují obvykle lépe než PAM pro lokální vyhledávání podobností (Henikoff & Henikoff, 1993)
- Pro porovnání blízce příbuzných proteinů by se měla používat nižší číslo PAM a vyšší BLOSUM, pro vzdálenější vyšší číslo PAM a nižší BLOSUM
- Pro prohledávání databází je nejběžnější BLOSUM62

Jak statisticky významné je skóre?

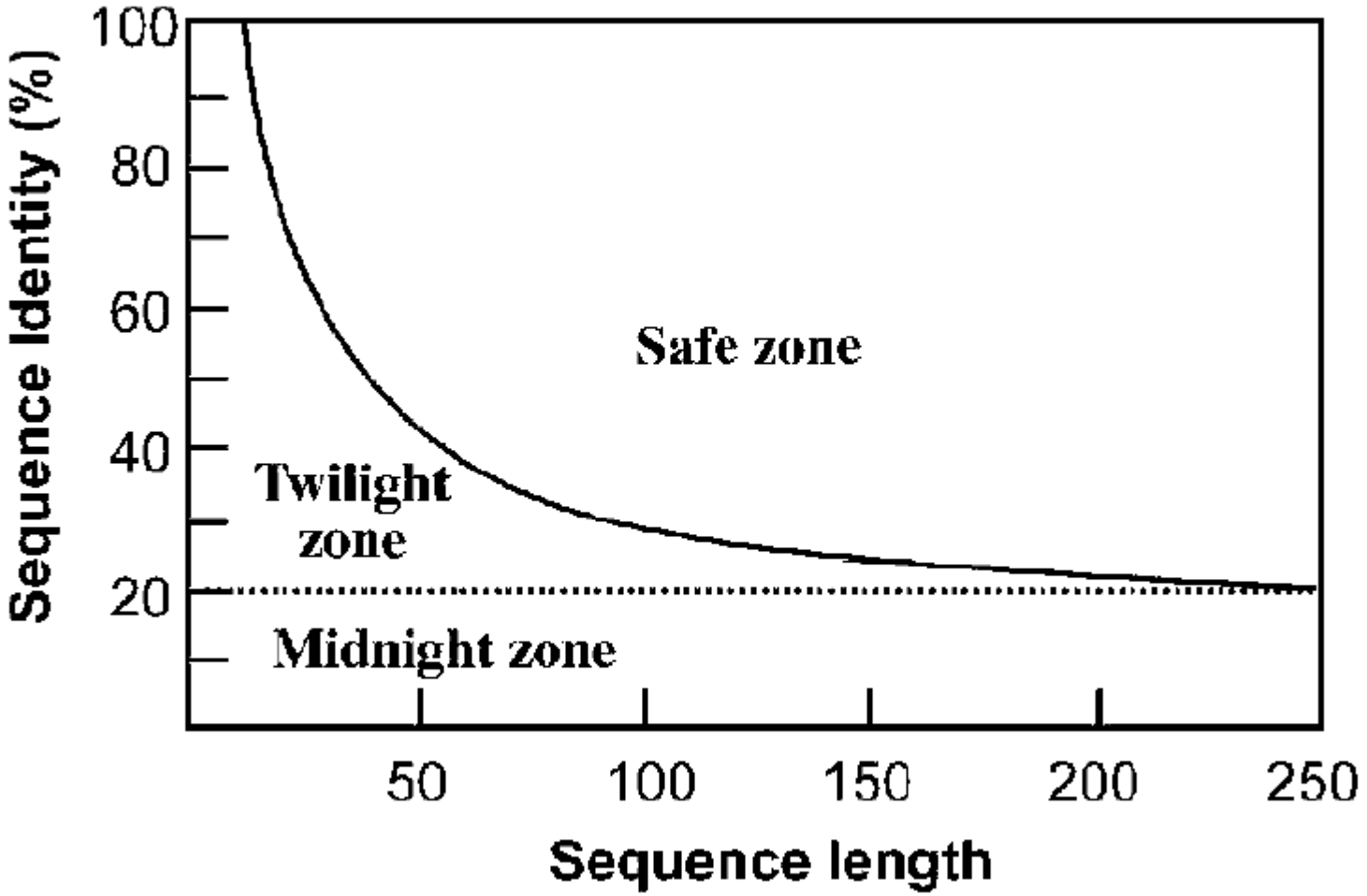
Pokud je podobnost dostatečně významná lze usuzovat na společné evoluční vztahy . Ale co je DOSTATEČNĚ?

závisí na typu sekvence a její délce

- Pravděpodobnost, že dvě rezidua v nepříbuzných sekvencích jsou identické?

  25% v NA, 5% v proteinech

- Vliv délky sekvence
  - čím kratší sekvence, tím větší je šance, že alignment je dán náhodnou shodou. Čím delší, tím je méně pravděpodobné, že je stejná úroveň podobnosti výsledkem náhody.
  - kratší sekvence vyžadují vyšší cut-off pro zjištění příbuznosti než u delších sekvencí



Essential bioinformatics, Xiong

Co to jsou oblasti sekvencí tzv.
„low complexity regions“

proč se definují a jak se používají?

- vysoce repetitivní krátké segmenty
  AAATAAAAAAAATAAAAAAT
- Hojně zastoupeny v databázích (cca 15% proteinů)
- Mohou vést k uměle vysokým hodnotám výsledných skóre nepříbuzných sekvencí
- Proto je nezbytné je vyjmout ze zadávacího dotazu stejně jako ze sekvenčních databází.

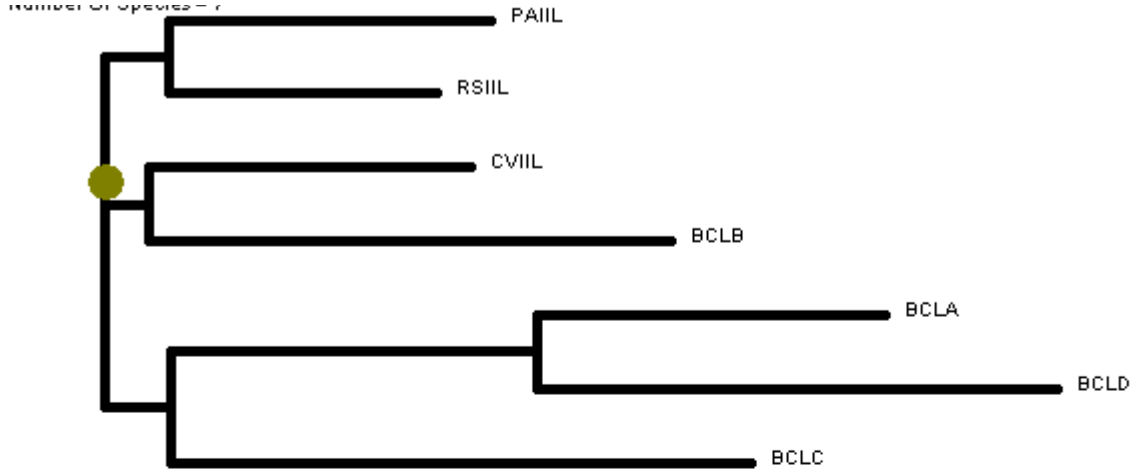
Phylogram a cladogram

- **Phylogram** (phylogeny tree) – je rozvětvený diagram (strom), který naznačuje fylogenezi (postupný vývoj). Délka jednotlivých větví je úměrná **velikosti změny** v průběhu evoluce.

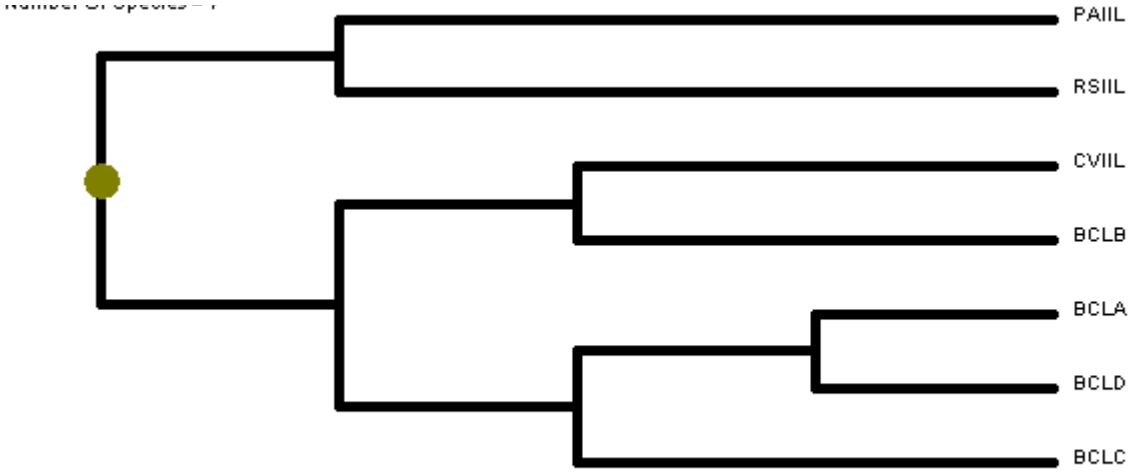
- **Cladogram** – rovněž strom, v němž však všechny větve mají **stejnou délku**. Ukazuje tak sice „společné předky“ pro jednotlivé sekvence, ale ne množství změn, jež od té doby prodělaly (evoluční dobu).

Phylogram a cladogram

Phylogram



Cladogram



Doplňte distanční matrici:

1 ATGTTTCTCCAACGCTGCTG
2 ATGTTTCTCCAGCGCTGCTG
3 ATGTTCCTTCAACGTTGTTG
4 ATGTTCCTTCAACGTTGCTG

	1	2	3	4
1				
2				
3				
4				

Který strom nejlépe popisuje fylogenezi?

1 ATGTTTCTCCAACGCTGCTG
2 ATGTTTCTCCAGCGCTGCTG
3 ATGTTCCTTCAACGTTGTTG
4 ATGTTCCTTCAACGTTGCTG

	1	2	3	4
1		0,05	0,2	0,15
2			0,25	0,2
3				0,05
4