Chemoinformatika a bioinformatika

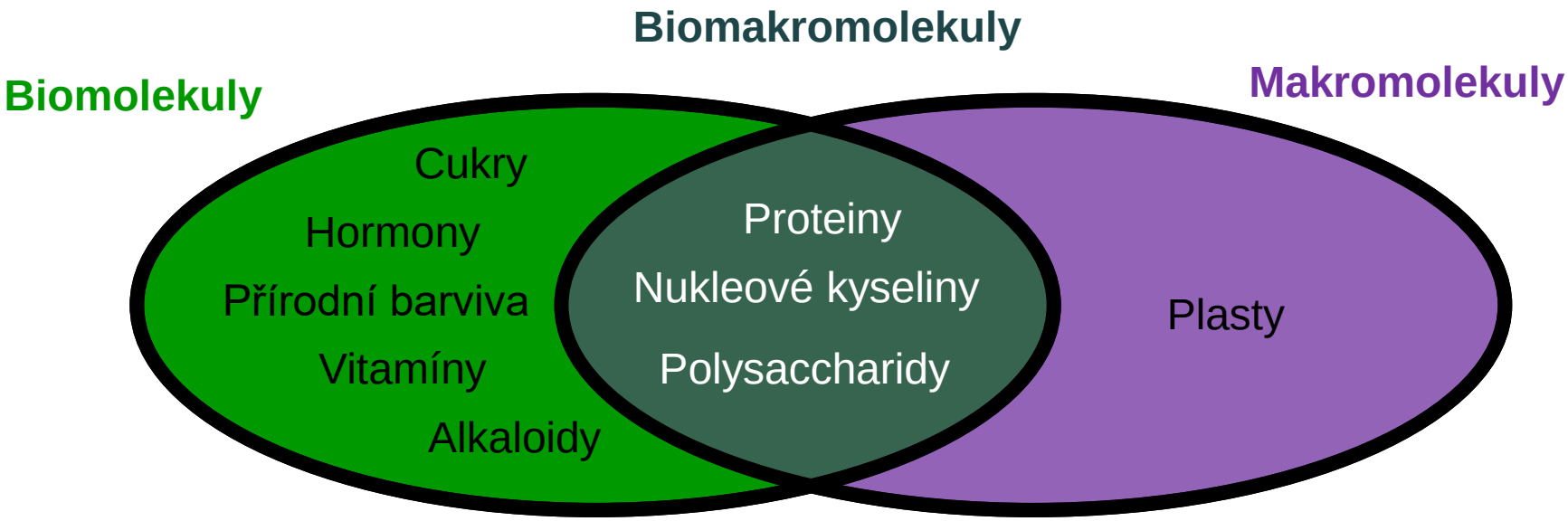
Sequence alignment

Biomakromolekuly

Biomolekuly jsou přirozenou součástí živých organismů.

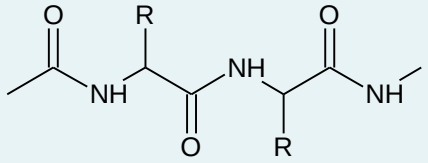
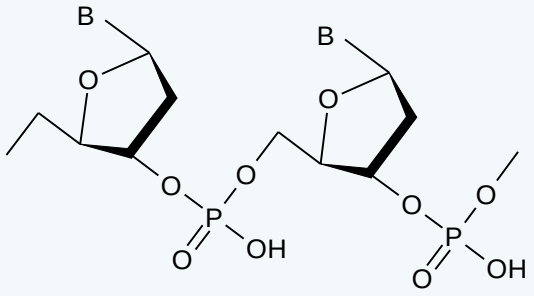
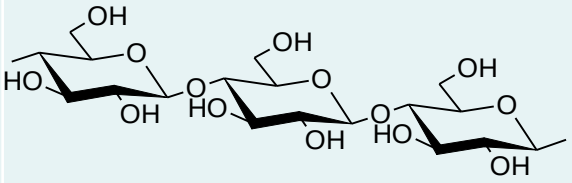
Velké molekuly. Typické malé molekuly jsou tvořeny několika atomy až několika sty atomů. Makromolekuly tvoří tisíce až miliony atomů.

Základní stavební jednotky hmoty. Jsou tvořeny atomy, které navzájem spojují kovalentní vazby.

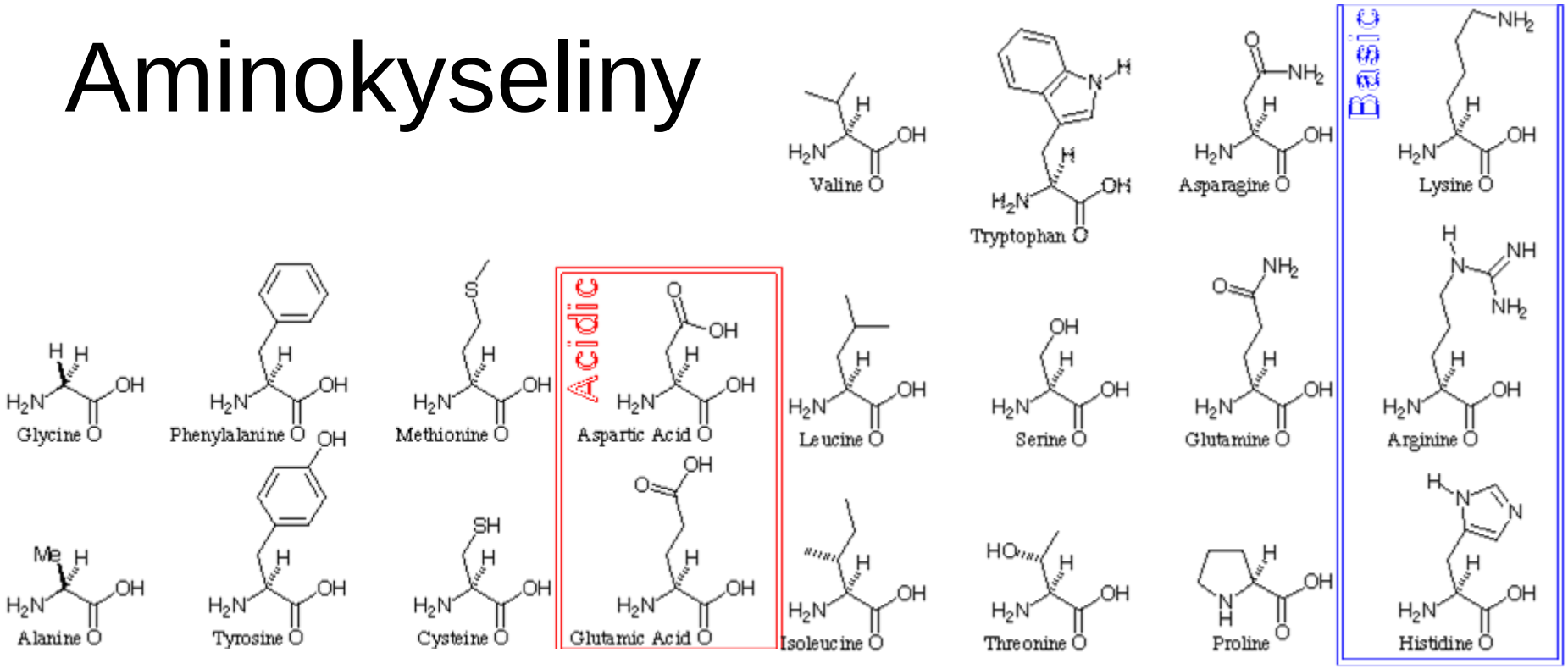


Složení biomakromolekul

- Vznikají spojováním velkého množství několika málo typů podjednotek

Makromolekula	Stavební jednotky	Typ vazby	Schéma
Protein	Aminokyseliny	Peptidová	
Nukleová kyselina	Nukleotidy	Esterová	
Polysacharid	Monosacharidy	Glykosidická	

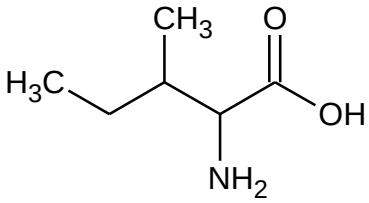
Aminokyseliny



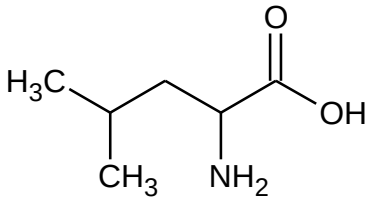
glycin	alanin	valin	leucin	izoleucin	asparagová kys.	asparagin	glutamová kys.	glutamin	arginin	lysin	histidin	fenylalanin	serin	threonin	tyrozin	tryptofan	methionin	cystein	prolin	selenocystein	pyrolysin
Gly	Ala	Val	Leu	Ile	Asp	Asn	Glu	Gln	Arg	Lys	His	Phe	Ser	Thr	Tyr	Trp	Met	Cys	Pro	Sec	Pyr
G	A	V	L	I	D	N	E	Q	R	K	H	F	S	T	Y	W	M	C	P	U	O

Třídění aminokyselin

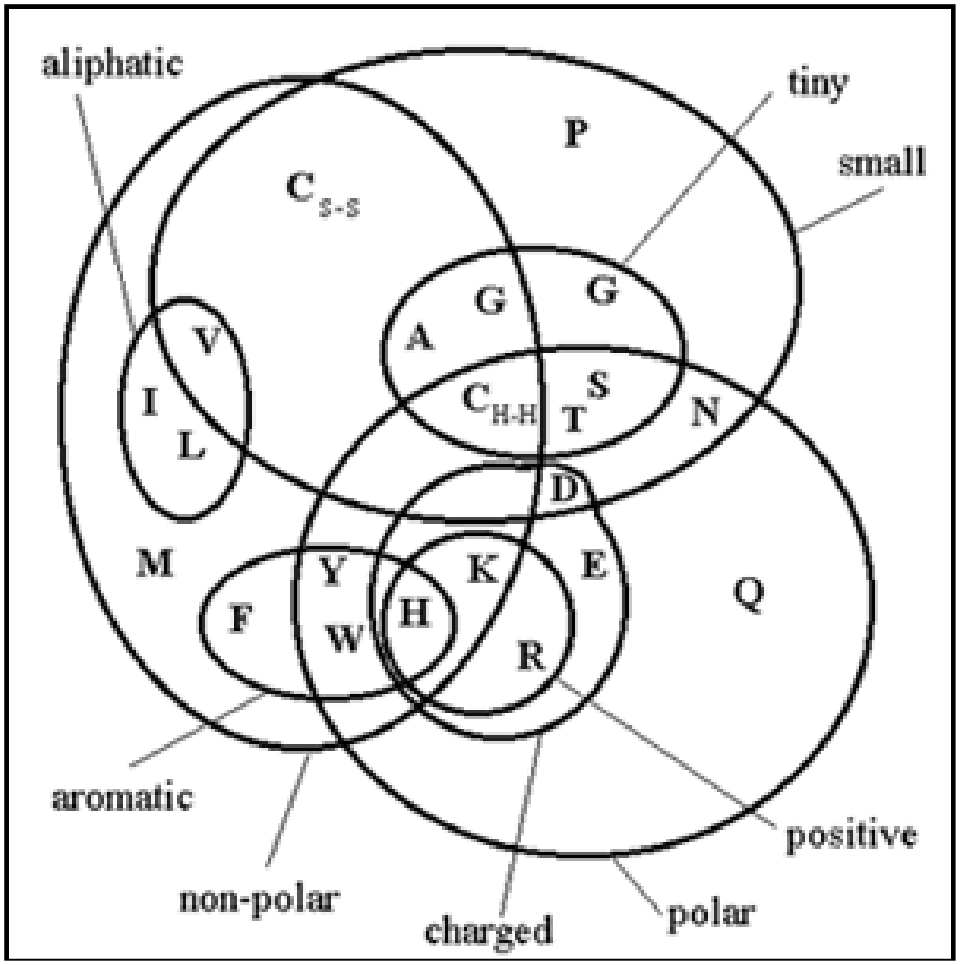
Aminokyseliny s podobnými vlastnostmi mohou plnit v proteinu stejné funkce – bývají vzájemně zastupitelné



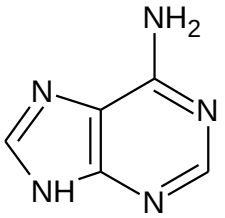
Isoleucine



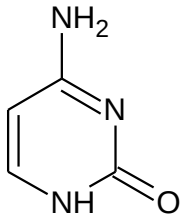
Leucine



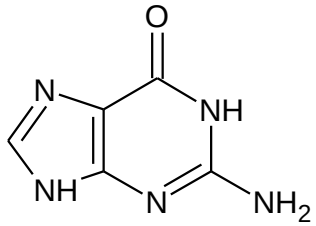
Nukleové báze



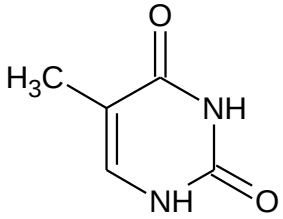
Adenine



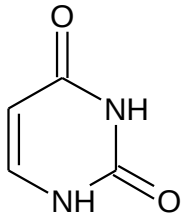
Cytosine



Guanine



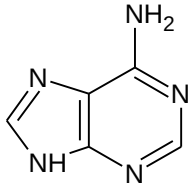
Thymine



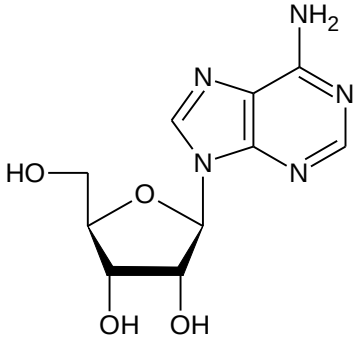
Uracil

adenin	cytosin	guanin	thymin	uracil
A	**C**	**G**	**T**	**U**

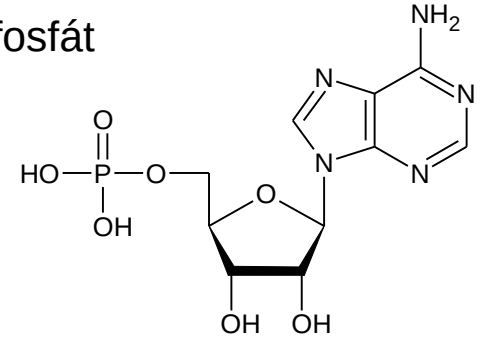
Nukleová báze
Adenin



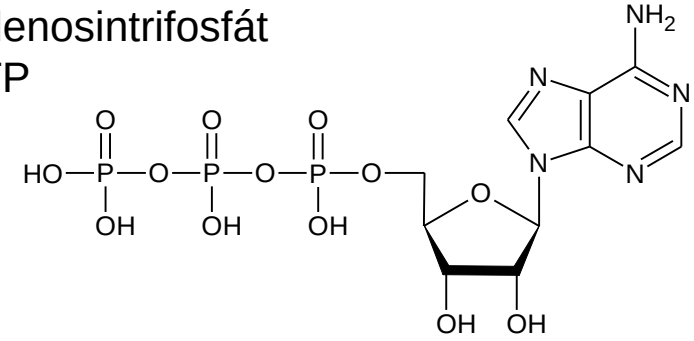
Nukleosid
Adenosin



Nukleotid
Adenosinmonofosfát
AMP



Nukleotid
Adenosintrifosfát
ATP



Polysacharidy

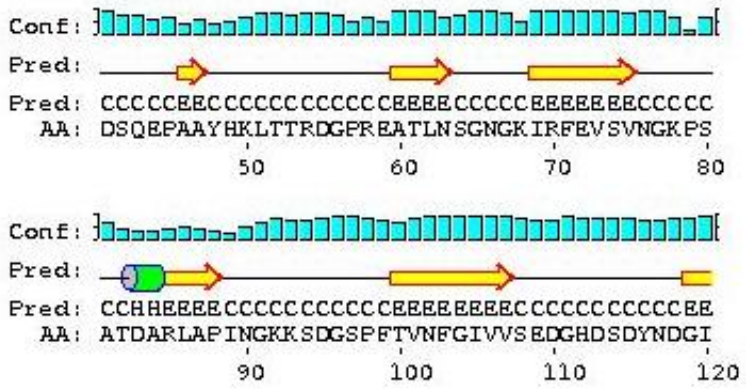
Komplikované sekvence – alignment se neprovádí

Polymer	**Protein**	**Nukleová kyselina**	**Polysacharid**
Počet druhů základních stavebních jednotek	20 (22)	4 (DNA) 4 (RNA)	desítky
Počet typů vzájemných vazeb	1	1	2 x 4 (pro hexosu)

Struktura proteinů (NK)

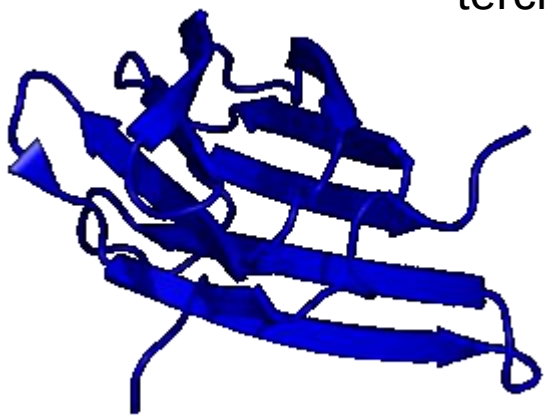
ADSQTSSNRAGEFSIPPNTDFRAIF
FANAAEQQHIKLFIGDSQEPAAYHK
LTTRDGPREATLNSGNGKIRFEVSV
NGKPSATDARLAPINGKKSDGSPF
TVNFGIVVSEDGHDSDYNDGIVVL
QWPIG

primární
(sekvence)

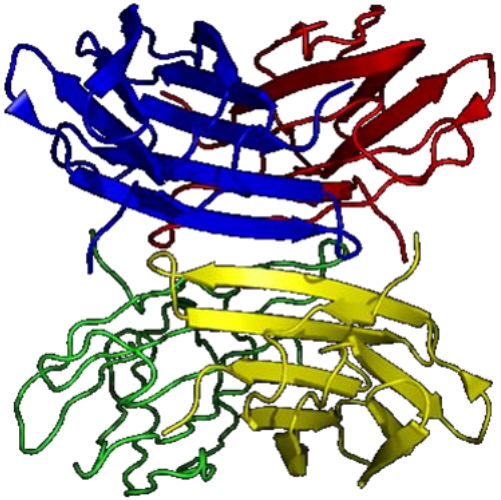


sekundární

terciární



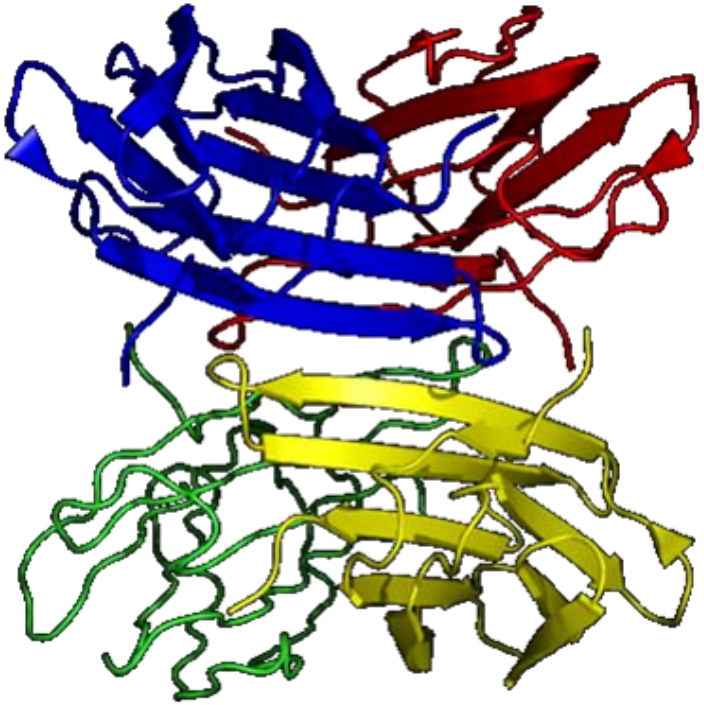
kvartérní



Kvartérní struktura proteinů

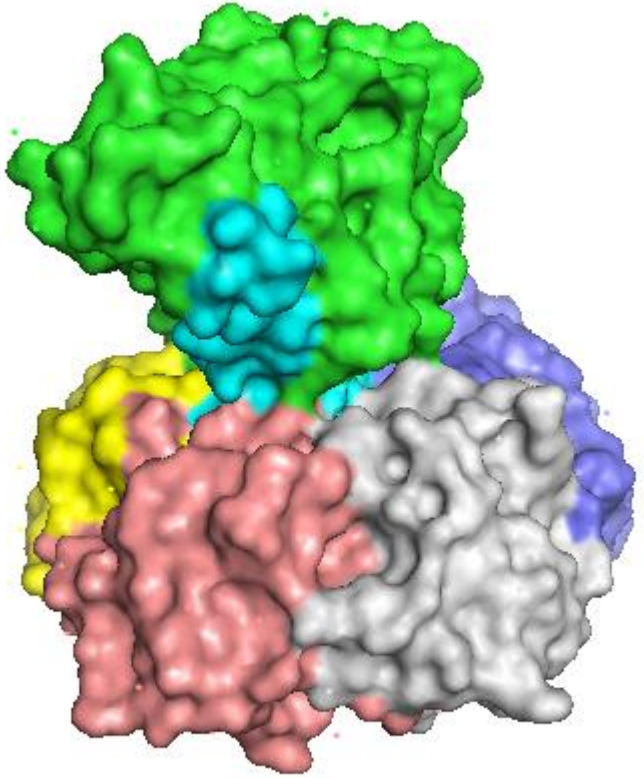
Homooligomer

Homotetramer



Heterooligomer

AB5 toxin



Jsou sekvence stejné, podobné či zcela odlišné?

ATGTCTACTCCTGGAGCACAGCAAGTCCTCTTCCGCACCGGAATTGCCGCGGTCAACTCAACCAACCATCTCCGTGTTTACTTCCAGGATGTCTATGGCAG
TATTCGCGAGAGTCTCTACGAGGGCAGCTGGGCTAACGGCACCGAAAAGAACGTTATCGGCAATGCTAAGCTTGGCAGCCCTGTGGCCGCGACTTCTAAG
GAGCTGAAGCATATCCGTGTCTACACCCTCACTGAAGGAAACACCCTACAGGAGTTCGCCTACGACTCCGGAACCGGATGGTACAACGGCGGGCTGGGC
GGTGCAAAGTTCCAAGTCGCACCCTACTCTCGCATTGCTGCCGTGTTCCTAGCCGGAACAGATGCATTGCAGTTGCGAATCTATGCACAGAAGCCAGATAA
CACAATCCAGGAGTATATGTGGAACGGCGATGGCTGGAAGGAGGGCACCAACCTGGGAGGTGCTCTCCCCGGCACTGGAATCGGAGCCACCTCCTTCCG
CTATACCGACTACAATGGCCCAAGCATCCGGATCTGGTTCCAAACTGACCTCAAACTCGTCCAAAGAGCCTACGACCCGCACAAAGGCTGGTACCCGGAC
CTCGTCACCATCTTTGACAGGGCACCGCCACGTACGGCCATTGCAGCCACCAGCTTTGGAGCCGGCAACAGTTCCATCTACATGCGTATCTACTTTGTCAA
TTCGGACAACACTATCTGGCAGGTCTGCTGGGACCACGGCAAGGGCTATCACGACAAGGGAACCATCACCCCAGTCATTCAGGGCTCGGAGGTCGCCATT
ATCAGCTGGGGCAGTTTCGCCAATAACGGGCCGGATCTGCGTCTGTACTTTCAGAATGGAACATACATTAGTGCTGTGAGCGAGTGGGTTTGGAATCGGG
CACATGGGTCGCAGTTGGGCAGAAGTGCTCTTCCTCCTGCTTGA

ATGGCTGATTCTCAAACGTCATCCAACCGCGCCGGCGAATTCTCGATTCCGCCGAATACCGATTTCCGCGCGATTTTCTTCGCGAATGCCGCCGAGCAACA
GCACATCAAATTGTTCATCGGCGACAGCCAGGAACCCGCCGCGTATCACAAGCTGACGACGCGCGACGGCCCGCGCGAAGCCACGCTGAATTCCGGCAA
CGGCAAGATCCGTTTCGAGGTGTCGGTGAACGGCAAGCCGTCGGCGACCGACGCGCGTCTCGCGCCGATCAACGGCAAGAAGTCGGACGGCTCGCCGT
TCACGGTCAACTTCGGGATCGTCGTGTCGGAAGACGGCCACGACAGCGACTACAACGACGGCATCGTCGTGCTCCAGTGGCCGATCGGCTGA

ATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCCCCGACCGTGATTGATGGTCGCCACCTGTATG
TTGTTAGCCCGGGCGATGCCGCGCAGCTGGGCCATAACGATAGCCGTCTGTTTACCGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGC
TGGCGCTGCGCGCGGAAGTGAGCGTGCTGTTTATTCGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCCGATCGAACTGGAAGTGCGTGATGCCGCCAC
CGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGATCATTATTGGCGCAGCGATGTGCTGGCGGCGGGCGCGACCACCTG
TACCGCCGATTTTGCGGTGTGCGATCGTGATGGCACCGTGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGGATACCAAA
CAGCCGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCGAATACCGCCTTTAAAGCGATCTTCTATGCGAACGCGGCGGATCGTC
AGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCGCCACCTTTGTGGGTAACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGT
GGTAAAATTCGTATTGAAGCGAGCGCGAACGGCCGTCAGAGCGCGACCGATGCCCGTCTGGCGCCGCTGAGCGCGGGCGATACCGTGTGGCTGGGCTG
GCTGGGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGATTACCTAA

ATGTCGAGCGTTCAAACCGCTGCCACTTCGTGGGGAACCGTACCGTCGATCCGTGTGTACACGGCCAATAATGGCAAGATCACCGAGCGATGCTGGGACG
GGAAGGGGTGGTACACCGGTGCCTTCAACGAGCCCGGCGATAACGTCTCCGTAACCAGCTGGCTGGTCGGCAGCGCGATCCATATCCGCGTCTATGCAA
GCACCGGCACCACGACCACGGAGTGGTGCTGGGACGGCAACGGCTGGACCAAGGGCGCCTACACCGCCACGAACTGA

ATGCCGCTGCTGAGCGCCAGTATCGTGAGCGCGCCGGTGGTGACCAGCGAAACCTATGTGGATATTCCGGGCCTGTATCTGGATGTTGCGAAAGCCGGTA
TCCGTGATGGCAAACTGCAGGTTATCCTGAATGTGCCGACCCCGTATGCGACGGGCAATAACTTTCCGGGTATTTATTTTGCGATCGCCACCAACCAGGGC
GTGGTGGCGGATGGTTGCTTTACGTATAGTAGCAAAGTGCCGGAAAGTACGGGCCGTATGCCGTTTACCCTGGTTGCGACCATTGATGTGGGTAGCGGTG
TTACCTTCGTGAAAGGTCAGTGGAAATCTGTTCGCGGCTCTGCGATGCATATTGATAGCTATGCAAGCCTGAGTGCGATTTGGGGCACCGCGGCACCGAGT
TCTCAGGGTTCTGGTAACCAGGGTGCGGAAACGGGTGGCACCGGTGCCGGTAATATTGGTGGCGGCGGTGAACGTGATGGCACCTTTAATCTGCCGCCG
CATATTAAATTCGGTGTTACCGCGCTGACCCACGCGGCGAACGATCAGACCATTGATATTTATATTGATGATGATCCGAAACCGGCAGCCACCTTTAAAGGC
GCGGGCGCGCAGGATCAGAACCTGGGTACCAAAGTGCTGGATTCTGGCAATGGCCGTGTTCGCGTTATCGTTATGGCGAACGGCCGTCCGAGCCGCCTG
GGTTCTCGTCAGGTGGATATTTTTAAAAAATCTTATTTCGGTATTATTGGCTCTGAAGATGGTGCGGATGATGATTATAACGATGGCATCGTGTTTCTGAACT
GGCCGCTGGGCTAA

ATGCCGCTCCTGAGCGCCAGTATCGTGAGCGCGCCGGTGGTGACCAGCCAAACCTATGTGGATATTCCGGGCCTGTATCTGGATGTTGCGAAAGCCGGTA
TCCGTGATGGCAAACTGCAGGTTATCCTGAATGTGCCGACCCCGTATGCGACGGGCAATAACTTTCCGGGTATTTATTTTGCGATCGCCACCAACCAGGGC
GTGGTGGCGGATGGTTGCTTTACGTATAGTAGCAAAGTGCCGGAAAGTACGGGCCGTATGCCGTTTACCCTGGTTGCGACCATTGATGTGGGTAGCGGTG
TTACCTTCGTGAAAGGTCAGTGGAAATCTGTTCGCGGCTCTGCGATGCATATTGATAGCTATGCAAGCCTGAGTGCGATTTGGGGCACCGCGGCACCGAGT
TCTCAGGGTTCTGGTAACCAGGGTGCGGAAACGGGTGGCACCGGTGCCGGTAATATTGGTGGCGGCGGTAAGCTTGCGGCCGCACTCGAGATCAAACGG
GCTAGCCAGCCAGAACTCGCCCCGGAAGACCCCGAGGATGTCGAGCACCACCACCACCACCACTGA

Jsou sekvence stejné, podobné či zcela odlišné?

MSTPGAQQVLFRTGIAAVNLTNHLRVYFQDVYGSIRESLYEGSWANGTEKNVIGNAKLGSPVAATSKELKHIRVYTLTEGNTLQEFAYDSGTGWYNGGLGGAKFQ
VAPYSRIAAVFLAGTDALQLRIYAQKPDNTIQEYMWNGDGWKEGTNLGGALPGTGIGATSFRYTDYNGPSIRIWFQTDDLKLVQRAYDPHKGWYPDLVTIFDRAPP
RTAIAATSFGAGNSSIYMRIYFVNSDNTIWQVCWDHGKGYHDKGTITPVIQGSEVAIISWGSFANNGPDLRLYFQNGTYISAVSEWVWNRAHGSQLGRSALPPA

MADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPREATLNSGNGKIRFEVSVNGKPSATDARLAPINGKKSDGSPFTVNFGIV
VSEDGHDSDYNDGIVVLQWPIG

MLVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVSVLFIRFALKDAGIVAPIELEVRDAATAVPDADDLL
HPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAA
TFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWLGWLGAEDGADADYNDGIVILQWPIT

MSSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTEWCWDGNGWTKGAYTATN

MPLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQW
KSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVL
DSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

MPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQW
KSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGKLAAALEIKRASQPELAPEDPEDVEHHHHHH

Alignment

Srovnání (přiložení) dvou či více sekvencí (aminokyselinových, nukleotidových) na základě jejich vzájemné podobnosti.

Význam alignmentu

- Identifikace sekvence v databázi
- Hledání podobných sekvencí v databázi
- Detekce mutací
- Hledání konzervovaných částí sekvence
- Odhalování příbuzenských vztahů
- Předpověď funkce makromolekuly
- Předpověď vyšších struktur

Typy alignmentu

Pairwise alignment – dvě sekvence

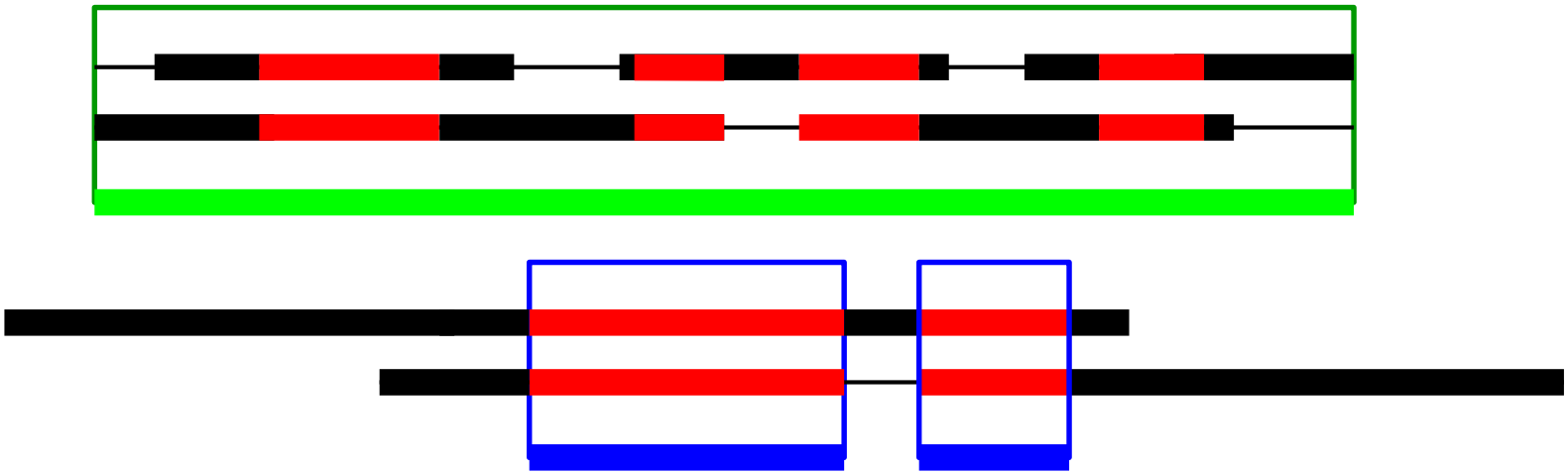
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM

Multiple sequence alignment – více sekvencí

WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
...

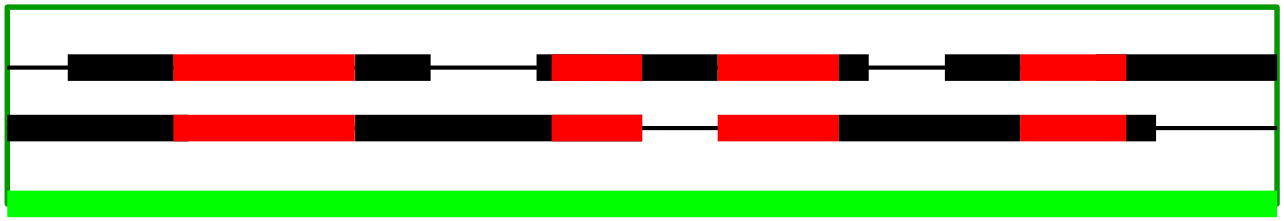
Pair-wise alignment

- Srovnání dvou sekvencí
- Sekvence mohou být přiloženy v celé své délce (global alignment) nebo jen v určitém regionu (local alignment).



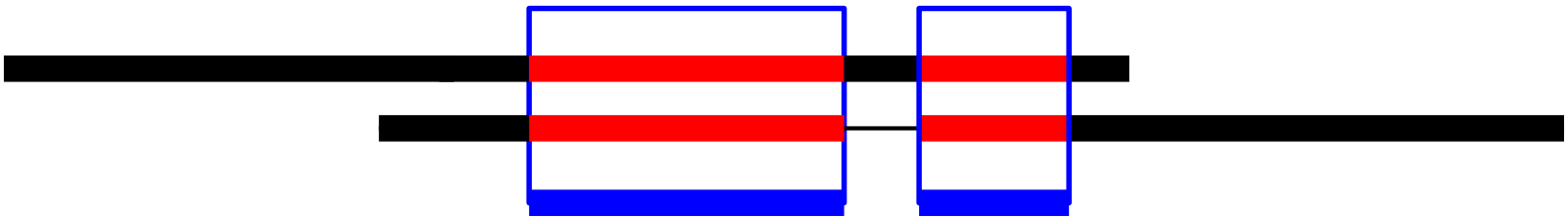
Global alignment

Vychází z předpokladu, že obě srovnávané sekvence jsou víceméně shodné v celé své délce. Alignment k sobě přikládá celé sekvence (od počátku do konce) a to včetně částí, které si příliš neodpovídají.



Local alignment

Hledá úseky dvou sekvencí, které si podle zvolených kritérií dobře odpovídají. Nesnaží se zahrnout celé sekvence, pokud si jejich některé části neodpovídají.



Algoritmy

- Téměř výhradně se užívají **heuristické algoritmy** – nalezení výsledku v dostatečně krátkém čase
- Vývoj algoritmů je prováděn v návaznosti na srovnávání výsledků s tzv. zlatým standardem – alignment na základě známých 3D struktur

Vstupní data

Sekvence AK (nt) v určitém formátu – dnes desítky formátů, mnohé obsahují kromě sekvence i doplňující data

Bližší např.
http://emboss.sourceforge.net/docs/themes/SequenceFormats.html

- **FASTA formát**

>název(␣popis dle vlastní volby)↵
SEKVENCESEKVENCESEKVENCESEKVENCES
EKVENCESEKVENCE↵

POVINNÉ VOLITELNÉ

>AFL
MSTPGAQQVLFRTGIAAVNLTNHLRVYFQDVYGSIRESLYEGSWANGTEKNVIGNAKLGSPVAATSKELKHIRVYT
LTEGNTLQEFAYDSGTGWYNGGLGGAKFQVAPYSRIAAVFLAGTDALQLRIYAQKPDNTIQEYMWNGDGWKEGT
NLGGALPGTGIGATSFRYTDYNGPSIRIWFQTDDLKLVQRAYDPHKGWYPDLVTIFDRAPPRTAIAATSFGAGNSS
IYMRIYFVNSDNTIWQVCWDHGKGYHDKGTITPVIQGSEVAIISWGSFANNGPDLRLYFQNGTYISAVSEWVWNR
AHGSQLGRSALPPA

>BC2LA
MADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPREATLNSGNGKIRFEVSV
NGKPSATDARLAPINGKKSDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG

> BC2LD
MLVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVSVL
FIRFALKDAGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGTVSGYFR
WETSIEIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRLF
TLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWLGWLGAEDGADADYNDGIVILQWPIT

>RSL
MSSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTE
WCWDGNGWTKGAYTATN

>gi|444369855|ref|ZP_21169562.1| fucose-binding lectin II [Burkholderia cenocepacia K56-2Valvano]
MPLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSS
KVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAG
NIGGGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMA
NGRPSRLGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

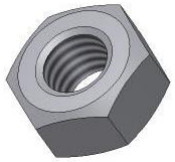
>gi|283806765|pdb|2WQ4|A Chain A
MPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSS
KVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAG
NIGGGGKLAAALEIKRASQPELAPEDPEDVEHHHHHH

Jak ale poznám dobré přiložení?

```
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
```

Scoring matrix (skórovací matice)

- Dvě sekvence považujeme za **příbuzné**, vycházejí-li ze společného předka; pak dobu potřebnou k jejich evoluci můžeme odvodit z množství rozdílů mezi nimi



- **Záměna** aa je častější než inserce/delece. Pravděpodobnost změny jedné aminokyseliny na jinou **je** přímo **úměrná podobnosti** obou aminokyselin.

- **Matice** vzniká přiřazením hodnoty (pravděpodobnosti) jednotlivým dvojicím aminokyselin v závislosti na jejich vzájemné „zastupitelnosti“ – pravděpodobnosti substituce

Skórování proteinového přiložení

Substituční matice (a z nich odvozeny **skórovací matice**)

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

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

Počet přiložených reziduí s podobnými vlastnostmi

Celkové délky obou sekvencí

Substituční matice

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

Typy matic

- **PAM** (Point Accepted Mutation) – založena na mutacích v rámci globálního alignmentu, tj. ve vysoce konzerovovaných i mutabilních oblastech. *PAM 250 znamená, že 250 mutací na 100 AA může nastat, PAM 10 akceptuje pouze 10 na 100, takže pouze velice podobné sekvence dosáhnou na pozitivní skóre.*
- **BLOSUM** (Blocks Substitution Matrix) – je odvozena z vysoce konzervovaných oblastí neobsahujících mezery - z těch počítá relativní zastoupení aa a pravděpodobnost jejich substitucí → lepší pro lokální alignment. *Je využívána v blastp, vhodná pro identifikaci neznámé nukleotidové sekvence. BLOSUM matrice s vysokými čísly je dobrá pro porovnání vysoce příbuzných sekvencí, zatímco nízké pro relativně vzdálené podobnosti*
- **GONNET** – vytvořena 1992, postupným opakováním cyklu: pairwise alignment – nová matice – nový pairwise alignment – nová matice…
- **DNA identity** matrix

V rámci jednoho typu matic existuje **více** jednotlivých **matic** založených na stejném principu, které se však liší konkrétními hodnotami a tedy i **oblastí použití** (vysoce příbuzné nebo naopak velmi vzdálené sekvence).

PAM – Point Accepted Mutation

- Vytvořila Margaret Dayhoff roku 1978.
- Zahrnuje pravděpodobnost záměny jedné aminokyseliny v druhou během evoluce
- Předpokládá, že každá další mutace nezávisí na předchozí.
- Odvozena z globálního alignmentu 71 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

PAM1

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

PAM1 reflektuje průměrnou záměnu 1% všech aminokyselinových pozic

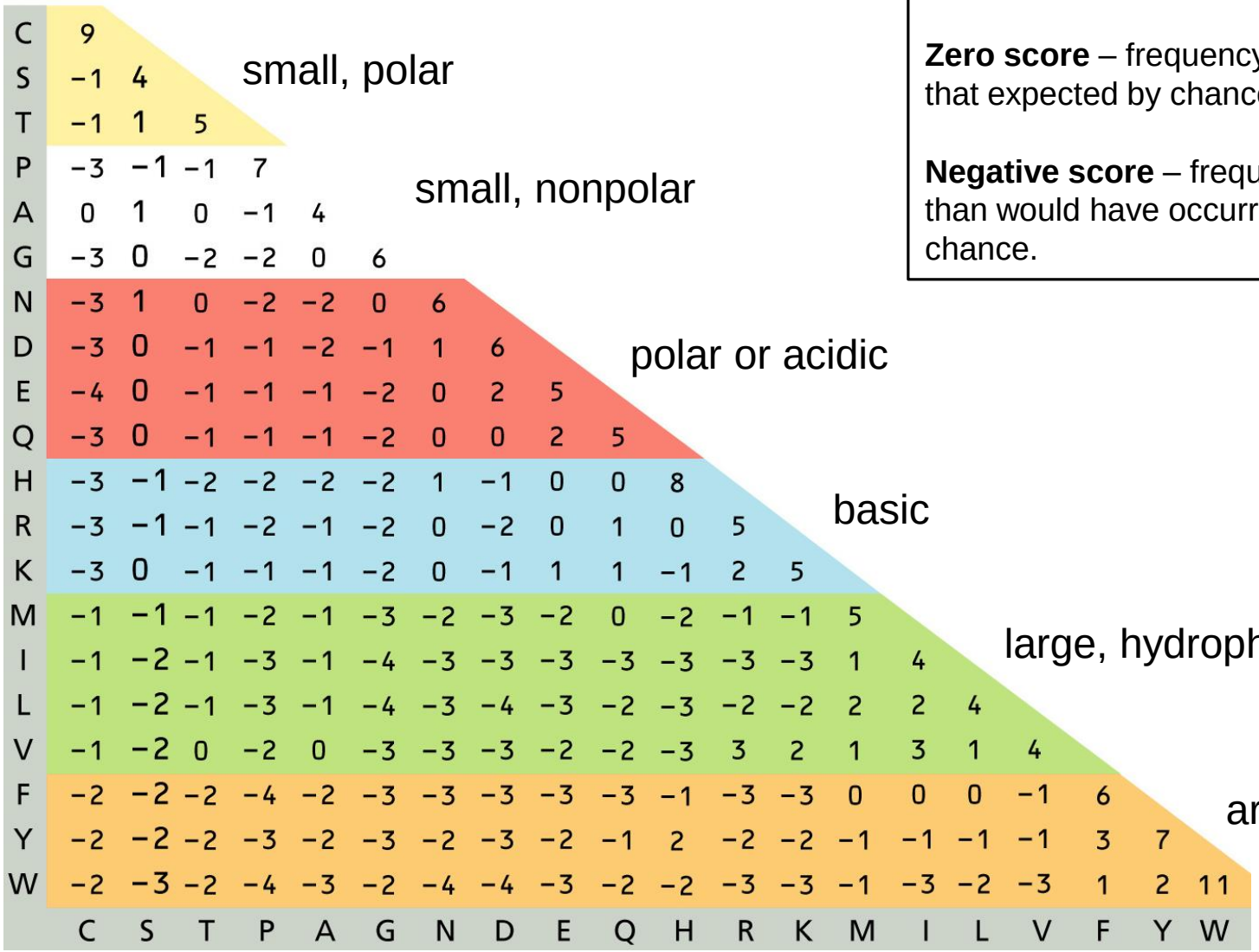
PAM250 (20% identita) je odvozena od PAM1 její 250-tinásobnou multiplikací (250 mutací na 100 aminokyselin)

Vyšší číslo PAM matrice znamená větší evoluční vzdálenost

PAM matice

	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

PAM250 matrice



	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

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.

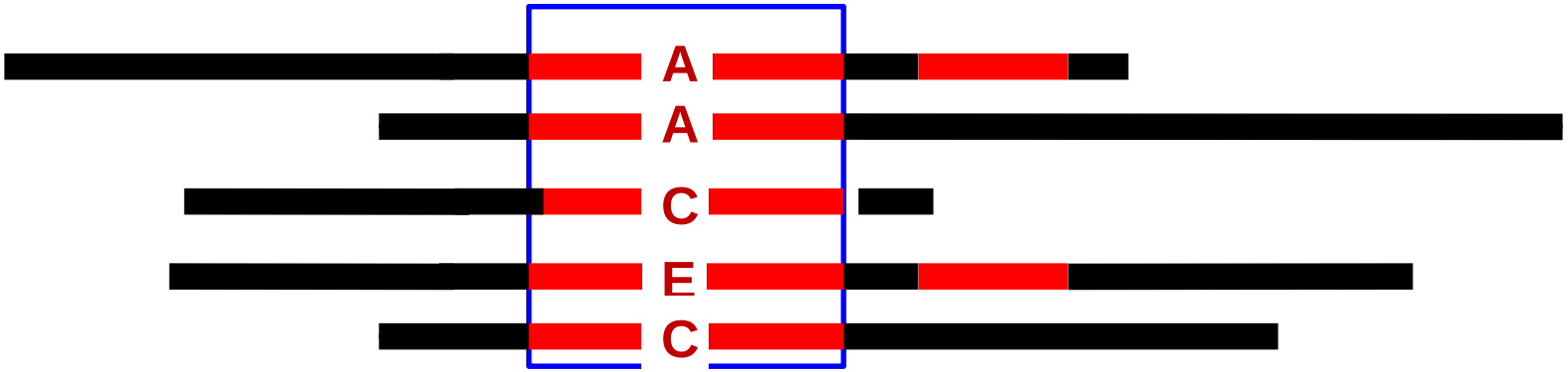
- Pozor na zjednodušení v matici PAM:
  - Mutace AA je nezávislá na předchozích mutacích v téže pozici (Markov process requirement).
  - Pouze matice PAM1 byla “změřena”, všechny ostatní jsou extrapolace (tj. jsou založeny na stejném modelu).
  - Všechna místa podléhají mutacím rovnoměrně.
  - Mutace nezávisí na okolních residuích.
  - Kratkodobé a dlouhodobé vlivy na evoluci sekvencí jsou stejně účinné.
  - PAM matice je založená na proteinových sekvencích dostupných v roce 1978 (bias vzhledem k malým globulárním proteinům)
    - Nová generace Dayhoff-type – např. PET91

BLOSUM (Blocks Amino Acid Substitution)

- 1992, Henikoff and Henikoff
- database BLOCKS – používá koncept „bloků“ k identifikaci proteinových rodin
- **sekvenční motiv**
  - konzervovaný aminokyselinový úsek 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

Matice BLOSUM 62

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

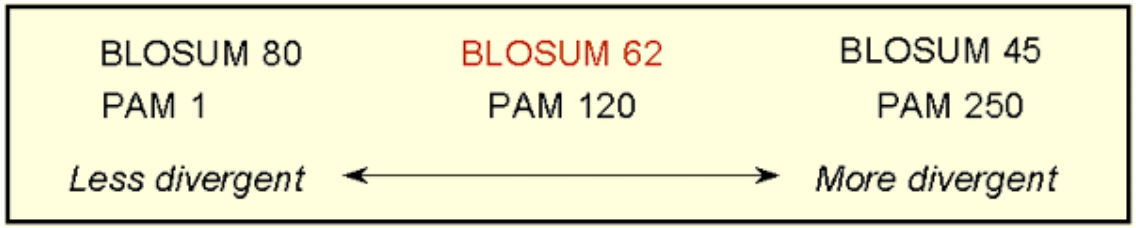
Číslování BLOSUM jde v obráceném pořadí oproti PAM

- 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 matice 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 čísla PAM a vyšší BLOSUM, pro vzdálenější vyšší čísla 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.

GONNETova matice

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	..
0.6	0.125	-0.075	0	-0.575	0.125	-0.2	-0.2	-0.1	-0.3	-0.175	-0.075	0.075	-0.05	-0.15	0.275	0.15	0.025	-0.9	-0.55	A
	2.875	-0.8	-0.75	-0.2	-0.5	-0.325	-0.275	-0.7	-0.375	-0.225	-0.45	-0.775	-0.6	-0.55	0.025	-0.125	0	-0.25	-0.125	C
		1.175	0.675	-1.125	0.025	0.1	-0.95	0.125	-1	-0.75	0.55	-0.175	0.225	-0.075	0.125	0	-0.725	-1.3	-0.7	D
			0.9	-0.975	-0.2	0.1	-0.675	0.3	-0.7	-0.5	0.225	-0.125	0.425	0.1	0.05	-0.025	-0.475	-1.075	-0.675	E
				1.75	-1.3	-0.025	0.25	-0.825	0.5	0.4	-0.775	-0.95	-0.65	-0.8	-0.7	-0.55	0.025	0.9	1.275	F
					1.65	-0.35	-1.125	-0.275	-1.1	-0.875	0.1	-0.4	-0.25	-0.25	0.1	-0.275	-0.825	-1	-1	G
						1.5	-0.55	0.15	-0.475	-0.325	0.3	-0.275	0.3	0.15	-0.05	-0.075	-0.5	-0.2	0.55	H
							1	-0.525	0.7	0.625	-0.7	-0.65	-0.475	-0.6	-0.45	-0.15	0.775	-0.45	-0.175	I
								0.8	-0.525	-0.35	0.2	-0.15	0.375	0.675	0.025	0.025	-0.425	-0.875	-0.525	K
									1	0.7	-0.75	-0.575	-0.4	-0.55	-0.525	-0.325	0.45	-0.175	0	L
										1.075	-0.55	-0.6	-0.25	-0.425	-0.35	-0.15	0.4	-0.25	-0.05	M
											0.95	-0.225	0.175	0.075	0.225	0.125	-0.55	-0.9	-0.35	N
												1.9	-0.05	-0.225	0.1	0.025	-0.45	-1.25	-0.775	P
													0.675	0.375	0.05	0	-0.375	-0.675	-0.425	Q
														1.175	-0.05	-0.05	-0.5	-0.4	-0.45	R
															0.55	0.375	-0.25	-0.825	-0.475	S
																0.625	0	-0.875	-0.475	T
																	0.85	-0.65	-0.275	V
																		3.55	1.025	W
																			1.95	Y

DNA matice

A	1			
T	-10000	1		
G	-10000	-10000	1	
C	-10000	-10000	-10000	1
	A	T	G	C

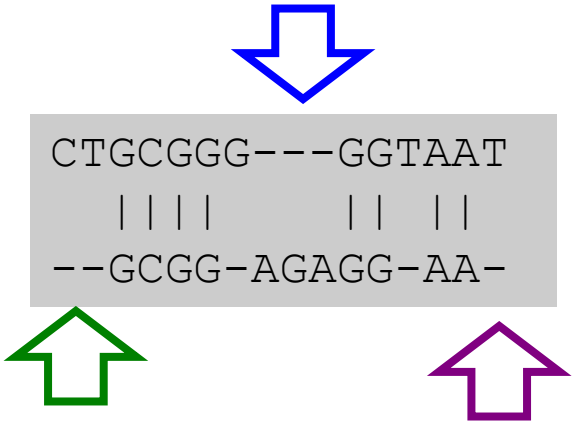
Jako pozitivní je uvažována pouze shoda, jakákoliv substituce je vysoce penalizována; jsou však povoleny mezery.

Mezery (Gaps)

Příčiny vzniku mezer:

- **Bodová mutace** (velmi častá příčina)
- Nepřesný crossover při meióze (inzerce nebo delece řetězce bází)
- DNA slippage během replikace (vzniká repetice – opakující se sekvence v řetězci)
- Inzerce retroviru
- Translokace DNA mezi chromozomy

Mezery nacházíme na **začátku** řetězce, **uprostřed** nebo na jeho **konci**.



Mezery umožňují alignment sekvencí, kdy v jedné z nich došlo k deleci. Zvyšují však také možnost alignmentu náhodných sekvencí. Jejich přítomnost je proto vždy „**penalizována**“, často více než substituce.

Čím nižší je penalizace mezer, tím lepší (dokonalejší) bude alignment, ovšem z biologického hlediska může jít o nesmysl.

Jednotlivé programy obvykle penalizují **přítomnost mezery** (gap open) a také zvyšují penalizaci s **délkou mezery** (gap ext).

Krátká mezera:

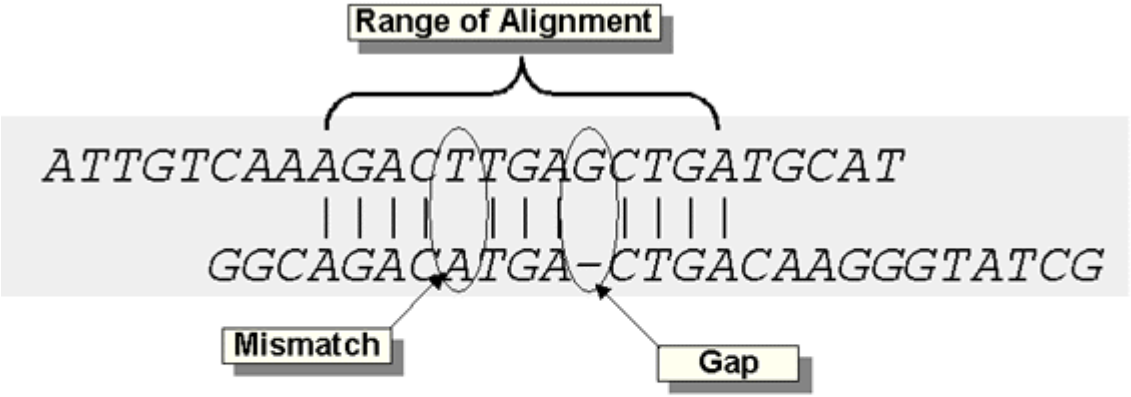
```
ATCTTCAGTGTTTCCCCTGTTTTGCCC-ATTTAGTTCGCTC
|||||||||||||||||||||||||||| ||||||||||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGATTTAGTTCGCTC
```

Dlouhá mezera:

```
ATCTTCAGTGTTTCCCCTGTTTTGCCC--------------------ATTTAGTTCGCTC
|||||||||||||||||||||||||||                    ||||||||||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGCCCCCCCCCCCCCCCCCCCATTTAGTTCGCTC
```

Skóre

Každé dvojici sekvencí je ve výsledku přiřazeno číslo – skóre, které **určuje míru** jejich **podobnosti**



$S = \sum(\text{identities, mismatches}) - \sum(\text{gap penalties})$

$\text{Score} = \text{Max}(S)$

Čím vyšší je skóre, tím vyšší je podobnost.

Podle použité matice může být skóre i záporné.

Příklad výpočtu

AAEECCDDEEF
AADDKKKEFGG

Ve chvíli, kdy zafixujeme pozici dvou sekvencí, pak můžeme snadno vypočítat skóre pro dané přiložení (příklad BLOSUM 62):

skóre na úrovni jednotlivých aa pro nesprávně přiložené sekvence:

```
A A E E C C D D E E F
A A D D K K K E F G G
4+4+2+2-3-3-1+2-3-2-3          = -1
```

Příklad výpočtu

AAEECCDDEEF
AADDKKKEFGG

Ve chvíli, kdy zafixujeme pozici dvou sekvencí, pak můžeme snadno vypočítat skóre pro dané přiložení (příklad BLOSUM 62):

skóre pro dané přiložení = skóre na bázi jednotlivých aa + celková penalizace

Například, celkové pozitivní skóre na úrovni jednotlivých aa

```
A A E E C C D D - - E E F
A A - - - - D D K K K E F G G
4+4         +6+6     +1+5+6          = 32
```

Naopak, pro každou mezeru (-) je dána penalizace: první výskyt zleva -10, každá následující -1.

```
A A E E C C D D - - E E F
A A - - - - D D K K K E F G G
   -10-1-1-1   -10-1              = -24
```

Celkové skóre 32 – 24 = 8

Příklad výpočtu

AAEEYYDDEEF
AADDFFKEFGG

Ve chvíli, kdy zafixujeme pozici dvou sekvencí, pak můžeme snadno vypočítat skóre pro dané přiložení (příklad BLOSUM 62):

skóre pro dané přiložení = skóre na bázi jednotlivých aa + celková penalizace

Například, celkové pozitivní skóre na úrovni jednotlivých aa

```
A A E E Y Y D D - - E E F
A A - - - - D D F F K E F G G
4+4         +6+6     +1+5+6          = 32
```

Naopak, pro každou mezeru (-) je dána penalizace: první výskyt zleva -10, každá následující -1.

```
A A E E Y Y D D - - E E F
A A - - - - D D F F K E F G G
   -10-1-1-1   -10-1              = -24
```

Celkové skóre 32 – 24 = 6

Příklad výpočtu

AAEEYYDDEEF
AADDFFKEFGG

Ve chvíli, kdy zafixujeme pozici dvou sekvencí, pak můžeme snadno vypočítat skóre pro dané přiložení (příklad BLOSUM 62):

skóre pro dané přiložení = skóre na bázi jednotlivých aa + celková penalizace

Například, celkové pozitivní skóre na úrovni jednotlivých aa

```
A A E E Y Y D D E E F
A A D D F F - - K E F G G
4+4+2+2+3+3     +1+5+6          = 30
```

Naopak, pro každou mezeru (-) je dána penalizace: první výskyt zleva -10, každá následující -1.

```
A A E E Y Y D D E E F
A A D D F F - - K E F G G
          -10-1                 = -11
```

Celkové skóre 30 – 11 = 19

Multiple sequence alignment - MSA

(mnohonásobné přiložení)

Multiple alignment slouží k:

- Nalezení „diagnostického vzoru“ (diagnostic patterns) na jehož základě jsou **charakterizovány proteinové rodiny**
- Odhalení či dokázání **homologie** mezi novou sekvencí a sekvencemi v databázích
- Určení vzájemné příbuznosti sekvencí v rámci skupiny – tvorba **fylogenetických stromů**
- **Predikci** sekundární a terciární **struktury** nových proteinů
- Navržení primerů (oligonukleotidů) pro PCR

Metody MSA

- Dynamické programování (dynamic programming) – rozšíření pairwise alignmentu - náročné na paměť a čas, nevhodné pro více než 3-4 sekvence (n=rozměrný prostor)
- **Progresivní alignment** (progressive sequence alignment) – nejčastěji používaný k vytvoření alignmentu; využívá fylogenetické informace – hierarchický, nejdříve identifikuje nejpodobnější sekvence a následně inkorporuje ostatní
- Iterativní alignment (iterative sequence alignment) – odstraňuje problémy progresivního alignmentu, který je závislý na prvotním přiložení nejpodobnějších sekvencí pomocí opakování alignmentu pro podskupiny sekvencí následující po globálním alignmentu
- Hledání motivů – nalezení částí konzervovaných sekvenčních motivů pomocí globálního přiložení a následně „hodnocení“ těcto úseků nezávisle na celé sekvenci

Dynamické programování

- **Simultánní alignment všech sekvencí** - analogické pairwise alignmentu
- Programové balíky: MSA (Lipman et al., 1989) a DCA (Stoye et al., 1997), založené na Carrilově a Lipmanově algoritmu (1988)
- Využívá skórovací matice, ale vytváří n-rozměrný prostor (n = počet sekvencí)
- Extrémně **náročný na výpočetní kapacity**
- I při zjednodušení nepoužitelné pro více než cca 20 sekvencí

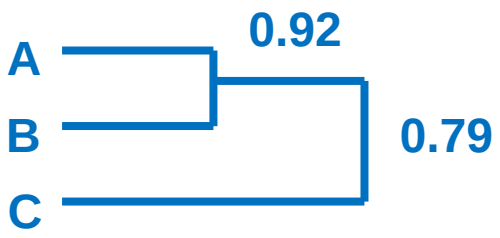
Progresivní multiple alignment

- Používá ho většina programů
- Vznik – 1987

  Feng, D.-F. and Doolittle, R.F. (1987) J. Mol. Evol. 25, 351-360.

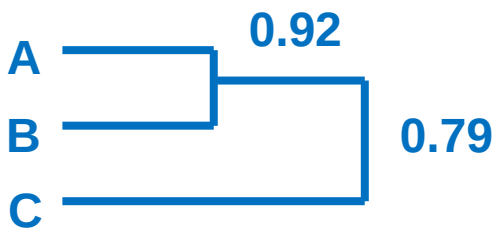
1) sestavení příbuzenského stromu (guide tree) na základě distanční matice (distance matrix) z jednotlivých sekvencí

A	-		
B	0.92	-	
C	0.65	0.79	-
	A	B	C



Počet exaktně stejných shod dělená celkovou délkou sekvence (ignoruje mezery)

Progresivní multiple alignment



Nejdříve provede pairwise alignment A a B
Pak přidá sekvenci C do předešlého alignmentu
(inzerce mezer, pokud je potřeba)

2) tvorba párových alignmentů postupně podle příbuznosti (topologie guide tree)

- Dnes obsahuje často iterativní smyčku

Guide tree vs. phylogenetic tree

- **Guide tree** je vypočítán na základě matice vzdáleností (distance matrix) vytvořené podle skóre pairwise alignmentů. Výstupem je .dnd soubor. NEMÁ fylogenetický význam

- **Phylogenetic tree** je vypočten na základě vytvořeného MSA. Vzdálenosti mezi sekvencemi jsou vypočteny a uloženy jako .ph soubor. Následně je možno je využít pro konstrukci fylogenetického stromu (soubory .nj, .ph, .dst) pomocí zvolené metody (nj, phylip, dist).

.dnd soubor

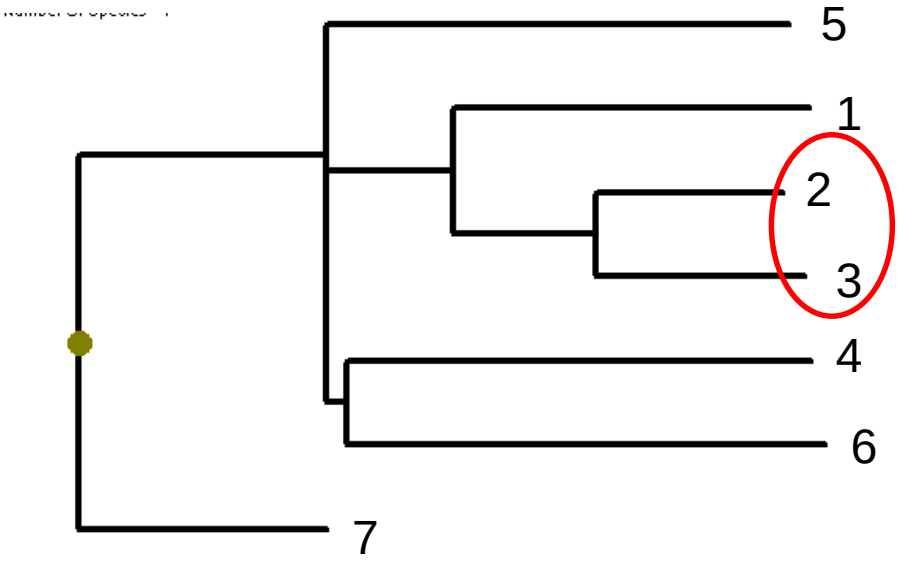
```
(
(
PAIIL:0.16435,
RSIIL:0.13654)
:0.03384,
(
CVIIL:0.16563,
BCLB:0.26800)
:0.02264,
(
(
BCLA:0.17899,
BCLD:0.26633)
:0.18717,
BCLC:0.29707)
:0.03484);
```

DIST = percentage divergence (/100)
Length = number of sites used in comparison
1 vs. 2 DIST = 0.6491; length = 114
1 vs. 3 DIST = 0.6842; length = 114
1 vs. 4 DIST = 0.9298; length = 114
1 vs. 5 DIST = 0.9035; length = 114
1 vs. 6 DIST = 0.9386; length = 114
1 vs. 7 DIST = 0.9825; length = 114
2 vs. 3 DIST = 0.3772; length = 114
2 vs. 4 DIST = 0.9123; length = 114
2 vs. 5 DIST = 0.8947; length = 114
2 vs. 6 DIST = 0.9123; length = 114
2 vs. 7 DIST = 0.9386; length = 114
3 vs. 4 DIST = 0.9123; length = 114
3 vs. 5 DIST = 0.9386; length = 114
3 vs. 6 DIST = 0.9298; length = 114
3 vs. 7 DIST = 0.9474; length = 114
4 vs. 5 DIST = 0.9211; length = 114
4 vs. 6 DIST = 0.9035; length = 114
4 vs. 7 DIST = 0.9649; length = 114
5 vs. 6 DIST = 0.9561; length = 114
5 vs. 7 DIST = 0.9211; length = 114
6 vs. 7 DIST = 0.9649; length = 114

Neighbor-joining Method

Saitou, N. and Nei, M. (1987) The Neighbor-joining Method:
A New Method for Reconstructing Phylogenetic Trees.
Mol. Biol. Evol., 4(4), 406-425
This is an UNROOTED tree
Numbers in parentheses are branch lengths
Cycle 1 = SEQ: 2 (0.17807) joins SEQ: 3 (0.19912)
Cycle 2 = SEQ: 1 (0.34101) joins Node: 2 (0.13706)
Cycle 3 = SEQ: 5 (0.44298) joins SEQ: 7 (0.47807)
Cycle 4 = SEQ: 4 (0.44518) joins SEQ: 6 (0.45833)
Cycle 5 (Last cycle, trichotomy):
Node: 1 (0.12171) joins
Node: 4 (0.01864) joins
Node: 5 (0.02083)

.nj soubor



.dst soubor

7

PAIIL	0.000	0.649	0.684	0.930	0.904	0.939	0.982
RSIIL	0.649	0.000	0.377	0.912	0.895	0.912	0.939
CVIIL	0.684	0.377	0.000	0.912	0.939	0.930	0.947
BCLA	0.930	0.912	0.912	0.000	0.921	0.904	0.965
BCLB	0.904	0.895	0.939	0.921	0.000	0.956	0.921
BCLC	0.939	0.912	0.930	0.904	0.956	0.000	0.965
BCLD	0.982	0.939	0.947	0.965	0.921	0.965	0.000

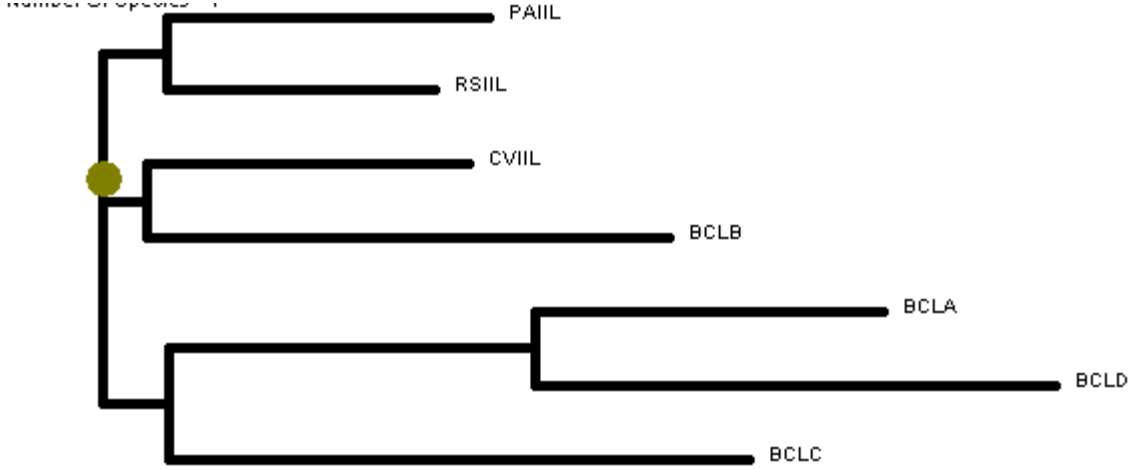
Fylogram a kladogram

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

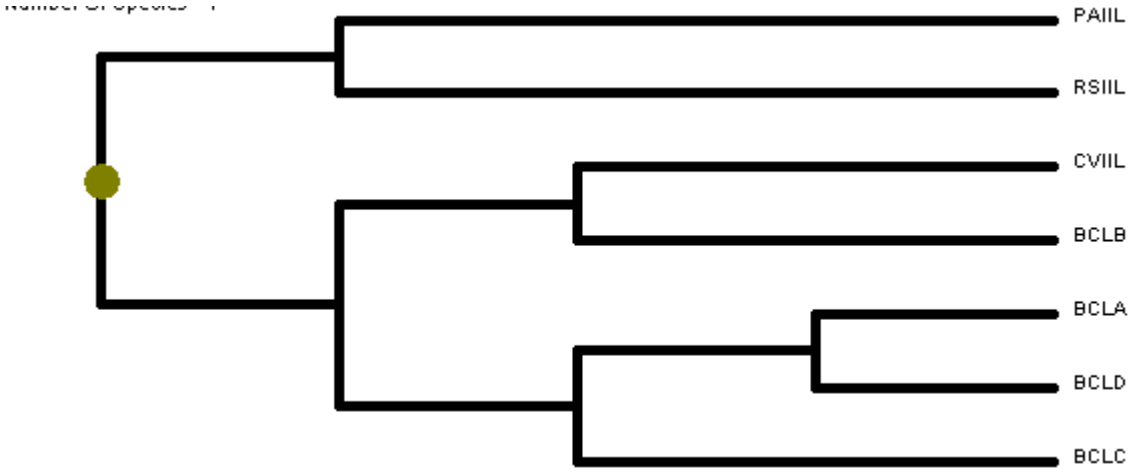
- **Kladogram** – rovněž strom, v němž však všechny větve mají **stejnou délku**. Ukazuje tak sice „společné předky“ pro jednotlivé sekvence, ale ne množství změn, jež od té doby prodělaly (evoluční dobu).

Fylogram a kladogram

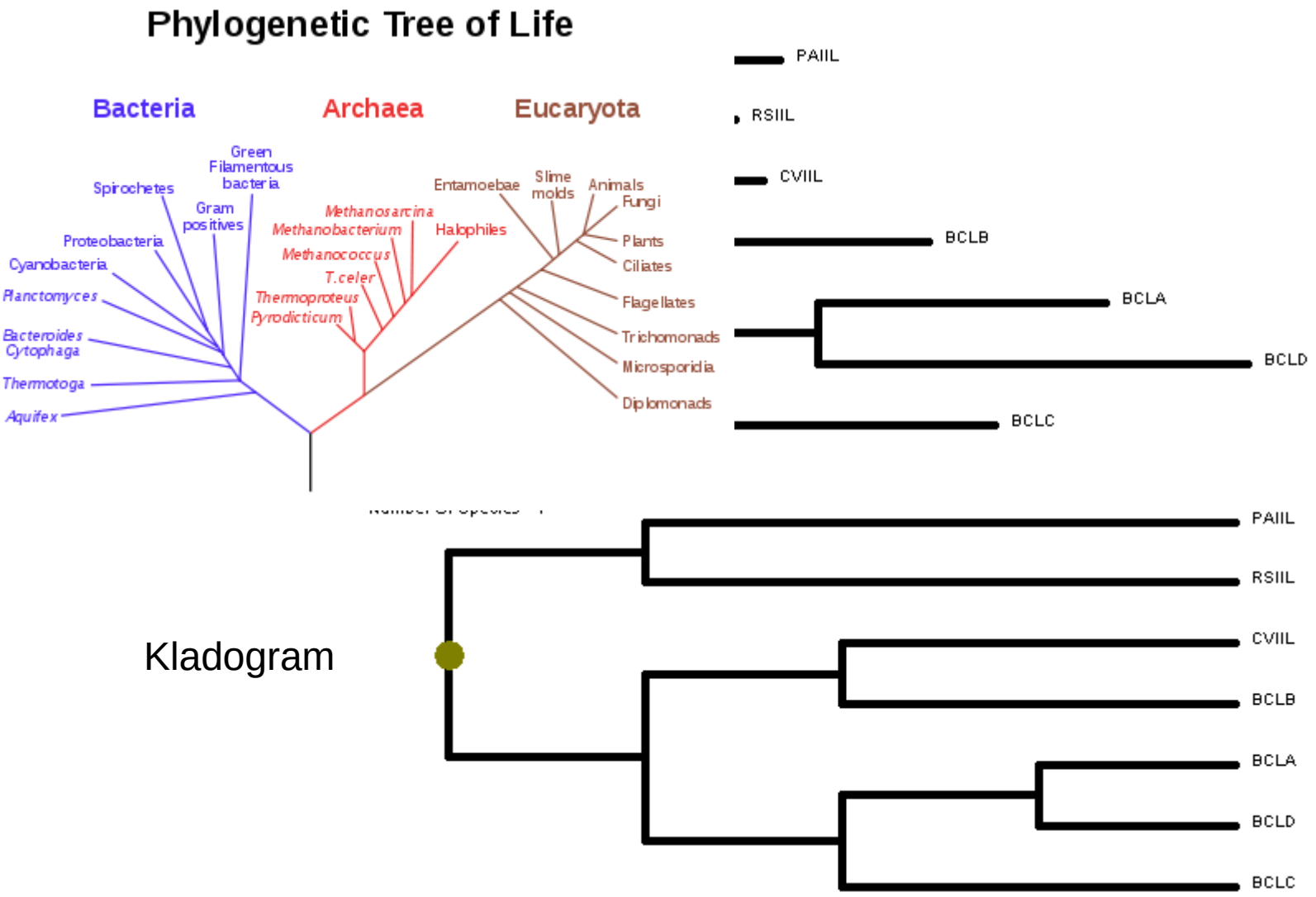
Fylogram



Kladogram



Fylogram a kladogram



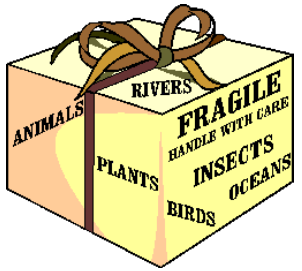
Kladogram

Výstup - .aln soubor

```
CLUSTAL 2.0.10 multiple sequence alignment


PAIIL           ------------------------------------------------------------
RSIIL           ------------------------------------------------------------
CVIIL           ------------------------------------------------------------
BCLB            ---LVEKLPQYDVFVDIATIPYSFDVGSWQNKVKTDAAGEVVACTVTWAGAPGVLPGAAA
BCLC            AIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDS
BCLA            ------------------------------------------------------------
BCLD            LRETALALRAEVSVLFIRFALKDAGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYW


PAIIL           ----------------------------------------------------ATQGVFT
RSIIL           ----------------------------------------------------AQQGVFT
CVIIL           ----------------------------------------------------AQQGVFT
BCLB            KFGVGAVVN----------------YFSKATPQPVQPAPVP--------TGGGERDGIFT
BCLC            YASLSAIWG----------------TAAPSSQGSGNQGAETGGTGAGNIGGGGERDGTFN
BCLA            ----------------------------------ADSQT---------SSNRAGEFS
BCLD            RSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTKQPGFKPSSDRNGNFS
                                                                  * *.

PAIIL           LPANTRFGVTAFANSSGTQTVNVLVNNETA--ATFSGQSTNNAVIGTQVLNSGSSGKVQV
RSIIL           LPANTSFGVTAFANAANTQTIQVLVDNVVK--ATFTGSGTSDKLLGSQVLNSGS-GAIKI
CVIIL           LPARINFGVTVLVNSAATQHVEIFVDNEPR--AAFSGVGTGDNNLGTKVINSGS-GNVRV
BCLB            LPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGK-GKVRV
BCLC            LPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGN-GRVRV
BCLA            IPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPRE--ATLNSGN-GKIRF
BCLD            LPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRL--FTLNSKG-GKIRI
                :*..  * .   .:::  * :.: :.:     * .      .       ::*   * ::.
```


BioEdit Sequence Alignment Editor
File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help
D:\Skola\Vyuka\MSA - data\BCLlectins seq.aln
8 total sequences
Courier New
11
Mode: Select / Slide
Selection: 0
Position:
Sequence Mask: None
Numbering Mask: None
Start ruler at: 1
Scroll speed slow fast
PAIIL
RSIIL
CVIIL
BCLB
BCLC
BCLA
BCLD
Clustal Cons
Snímek 34 z 64
Výchozí návrh
Jalview 2.3
File Tools Help Window
D:\Skola\Vyuka\MSA - data\BCLlectins seq.aln
File Edit Select View Format Colour Calculate Web Service
PAIIL/1-114
RSIIL/1-113
CVIIL/1-113
BCLB/1-243
BCLC/1-271
BCLA/1-128
BCLD/1-288
Conservation
Quality
Consensus
TLPPNTAFGVTA+ANAA+TQTI+VFVDDEPKPAATF+GAGT+DANLGTQVLNSGS-GKVRVQVSANG
Sequence position 247 5.460428

Programové balíky



- Existují programy pro pairwise alignment i pro MSA
- Využívají lokální nebo globální alignment nebo příp. kombinaci obou
- Neexistuje univerzální „nejlepší“ program – záleží na konkrétním použití

Pairwise alignment „programy“

Oblasti použití:

- Přímé porovnání dvou sekvencí
- Vyhledávání podobných sekvencí v databázích

Needle & Water

- vytvořeny 1970
  Needleman S.B. and Wunsch C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48: 443-453.*
- využívají dynamické programování
- umožňují vložení mezer

Needle – globální pairwise alignment, Needleman-Wunsch algoritmus

Water – lokální pairwise alignment, Smith-Waterman algoritmus

Globálně podobné sekvence

Needle

```
PA-IIL   1 ATQGVFTLPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVI 50
           |.||||||||||.||||||||::.|||:.|||:|...|||:|..|::.::
RS-IIL   1 AQQGVFTLPANTSFGVTAFANAANTQTIQVLVDNVVKATFTGSGTSDKLL 50
PA-IIL  51 GTQVLNSGSSGKVQVQVSVNGRPSDLVSAQVILTNELNFALVGSEDGTDN 100
           |:|||||| ||.:::|||||||||||||.|.||.|:||||:|||||||||
RS-IIL  51 GSQVLNSG-SGAIKIQVSVNGKPSDLVSNQTILANKLNFAMVGSEDGTDN 99
PA-IIL 101 DYNDAVVVINWPLG 114
           ||||.:.|:|||||
RS-IIL 100 DYNDGIAVLNWPLG 113
```

Water

```
PA-IIL   1 ATQGVFTLPANTRFGVTAFANSSGTQTVNVLVNNETAATFSGQSTNNAVI 50
           |.||||||||||.||||||||::.|||:.|||:|...|||:|..|::.::
RS-IIL   1 AQQGVFTLPANTSFGVTAFANAANTQTIQVLVDNVVKATFTGSGTSDKLL 50
PA-IIL  51 GTQVLNSGSSGKVQVQVSVNGRPSDLVSAQVILTNELNFALVGSEDGTDN 100
           |:|||||| ||.:::|||||||||||||.|.||.|:||||:|||||||||
RS-IIL  51 GSQVLNSG-SGAIKIQVSVNGKPSDLVSNQTILANKLNFAMVGSEDGTDN 99
PA-IIL 101 DYNDAVVVINWPLG 114
           ||||.:.|:|||||
RS-IIL 100 DYNDGIAVLNWPLG 113
```

Lokálně podobné sekvence

Needle

```
1   ------------------------------ADSQTSSN-----------      8
                                   ..||.|.|
101 TFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAG    150

9   -------RAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHK-     50
           |.|.|::||:..|......:||..|.|.::|.|..:|||..|
151 NIGGGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKG    200

51 -------LTTRDGPREATLNSGNGKIRFEVSVNGKPSATDARLAPINGKK     93
         |.|:      .|:||||::|..|..||:||...:|...| .||
201 AGAQDQNLGTK------VLDSGNGRVRVIVMANGRPSRLGSRQVDI-FKK    243

94  SDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG    128
    |       .|||:.||||.|.|||||||.|.||:|
244 S-------YFGIIGSEDGADDDYNDGIVFLNWPLG    271
```

Water

```
9   RAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHK--------     50
    |.|.|::||:..|......:||..|.|.::|.|..:|||..|
158 RDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQN    207

51  LTTRDGPREATLNSGNGKIRFEVSVNGKPSATDARLAPINGKKSDGSPFT    100
    |.|:      .|:||||::|..|..||:||...:|...| .|||
208 LGTK------VLDSGNGRVRVIVMANGRPSRLGSRQVDI-FKKS------    244

101 VNFGIVVSEDGHDSDYNDGIVVLQWPIG    128
     .|||:.||||.|.|||||||.|.||:|
245 -YFGIIGSEDGADDDYNDGIVFLNWPLG    271
```

Global vs. local alignment

Pairwise 314 vs. 90 aa protein
Obsahuje repetice

```
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 357
# Identity:      33/357 ( 9.2%)
# Similarity:    33/357 ( 9.2%)
# Gaps:         310/357 (86.8%)
# Score: 57.5
#
#
#=======================================

EMBOSS_001         1 STPGAQQVLFRTGIAAVNLTNHLRVYFQDVYGSIRESLYEGSWANGTEKN     50
EMBOSS_001         1 --------------------------------------------------      0
EMBOSS_001        51 VIGNAKLGSPV--AATSKELKH------IRVYT-----LTE----GNTLQ     83
                        |.|  ||||           |||||     .||    |.
EMBOSS_001         1 --------SSVQTAATS-----WGTVPSIRVYTANNGKITERCWDGK---     34
EMBOSS_001        84 EFAYDSGTGWYNGGLGGAKFQVAPYSRIAAVF------------LAGTDA    121
                        |||.    ||             |            |.|. |
EMBOSS_001        35 --------GWYT----GA-------------FNEPGDNVSVTSWLVGS-A     58
EMBOSS_001       122 LQLRIYAQKPDNTIQEYM----------WNGDGWKEG----TNLGGALPG    157
                     ...|.||                    |.|.||..|    ||
EMBOSS_001        59 IHIRVYA-----------STGTTTTEWCWDGNGWTKGAYTATN-------     90
EMBOSS_001       158 TGIGATSFRYTDYNGPSIRIWFQTDDLKLVQRAYDPHKGWYPDLVTIFDR    207
EMBOSS_001        91 --------------------------------------------------     90
EMBOSS_001       208 APPRTAIAATSFGAGNSSIYMRIYFVNSDNTIWQVCWDHGKGYHDKGTIT    257
EMBOSS_001        91 --------------------------------------------------     90
EMBOSS_001       258 PVIQGSEVAIISWGSFANNGPDLRLYFQNGTYISAVSEWVWNRAHGSQLG    307
EMBOSS_001        91 --------------------------------------------------     90

EMBOSS_001       308 RSALPPA    314
EMBOSS_001        91 -------     90
```

Skore:57.5

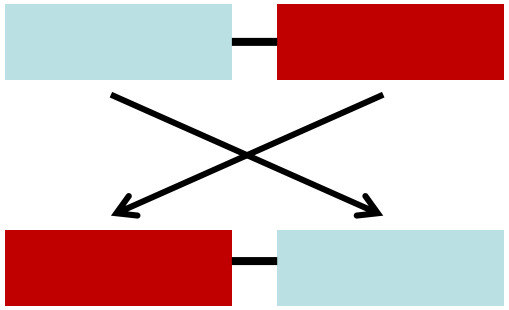
```
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 321
# Identity:      36/321 (11.2%)
# Similarity:    49/321 (15.3%)
# Gaps:         238/321 (74.1%)
# Score: 108
#
#
#=======================

EMBOSS_001         1 STPGAQQVLFRTGIAAVNLTNHLRVYFQDVYGSIRESLYEGSWANGTEKN     50
EMBOSS_001         1 --------------------------------------------------      0
EMBOSS_001        51 VIGNAKLGSPV--AATS-KELKHIRVYTLTEGNTLQEFAYDSGTGWYNGG     97
                        |.|  |||| ..:..|||||...|. :.|..:| |.|||.|.
EMBOSS_001         1 --------SSVQTAATSWGTVPSIRVYTANNGK-ITERCWD-GKGWYTGA     40
EMBOSS_001        98 LGGAKFQVAPYSRIAAVFLAGTDALQLRIYAQKPDNTIQEYMWNGDGWKE    147
                     .......|:..|     :|.|: |:.:|:|| ....|..|:.|:|:||.:
EMBOSS_001        41 FNEPGDNVSVTS-----WLVGS-AIHIRVYA-STGTTTTEWCWDGNGWTK     83
EMBOSS_001       148 G----TNLGGALPGTGIGATSFRYTDYNGPSIRIWFQTDDLKLVQRAYDP    193
                     |    ||
EMBOSS_001        84 GAYTATN----------------------------------------------     90
EMBOSS_001       194 HKGWYPDLVTIFDRAPPRTAIAATSFGAGNSSIYMRIYFVNSDNTIWQVC    243
EMBOSS_001        91 --------------------------------------------------     90
EMBOSS_001       244 WDHGKGYHDKGTITPVIQGSEVAIISWGSFANNGPDLRLYFQNGTYISAV    293
EMBOSS_001        91 --------------------------------------------------     90

EMBOSS_001       294 SEWVWNRAHGSQLGRSALPPA    314
EMBOSS_001        91 ---------------------     90
```

Skore:108

Nelze však spoléhat na zdánlivě dobrá řešení

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKV
PESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIG
GGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGR
PSRLGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSR
LGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLGPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRDGKLQ
VILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAM
HIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGKLAAALEIKRASQPELAPEDPEDVEHHHHHH

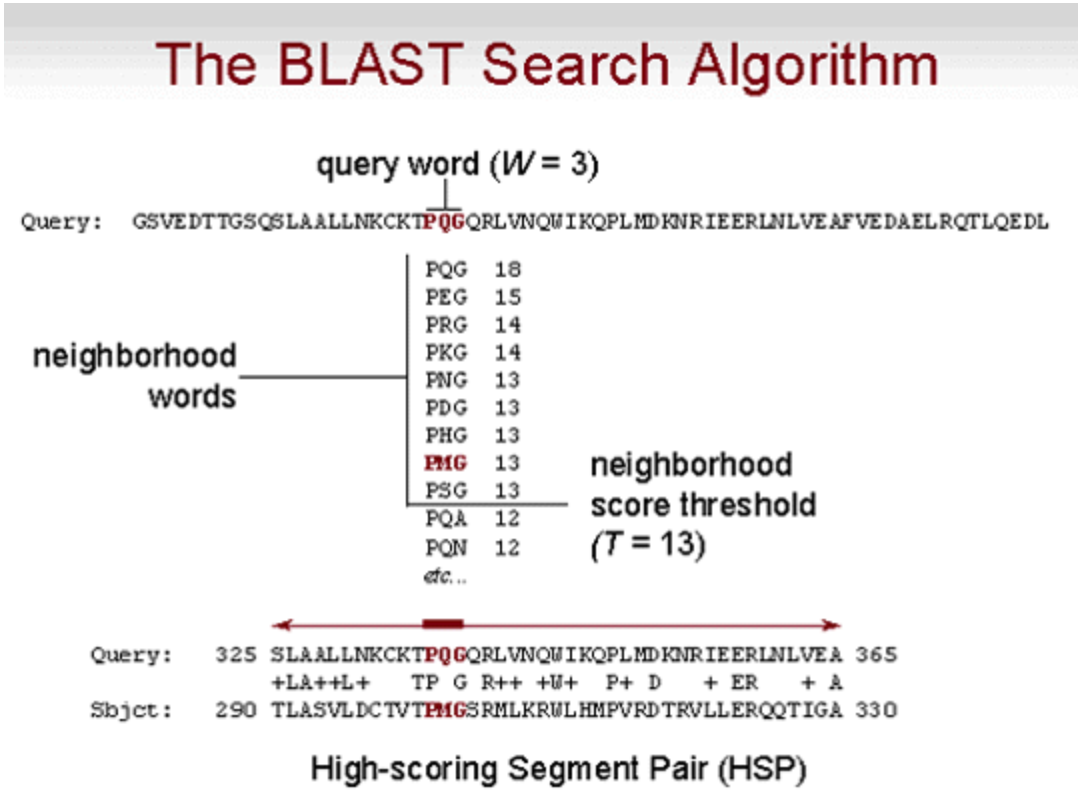


```
#
#=======================================

EMBOSS_001         1 --------------------------------------------------      0

EMBOSS_001         1 ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQ     50

EMBOSS_001         1 --------------------------------------------------      0

EMBOSS_001        51 NLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKSYFGIIGSEDGAD    100

EMBOSS_001         1 ---------------PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRD     35
                                    |||||||||||||||:||||||||||||||||||
EMBOSS_001       101 DDYNDGIVFLNWPLGPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRD    150

EMBOSS_001        36 GKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGR     85
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       151 GKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGR    200

EMBOSS_001        86 MPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQ    135
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||
EMBOSS_001       201 MPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQ    250

EMBOSS_001       136 GSGNQGAETGGTGAGNIGGGGERDGTFNLPPHIKFGVTALTHAANDQTID    185
                     ||||||||||||||||||||
EMBOSS_001       251 GSGNQGAETGGTGAGNIGGGG----------------------------    271

EMBOSS_001       186 IYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSRLGS    235
                            |.||:          |..|                 |:|
EMBOSS_001       272 -------KLAAA----------LEIK----------------RAS----    283

EMBOSS_001       236 RQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG    271
                      |.:         ::.||..|.:.:.       ...
EMBOSS_001       284 -QPE---------LAPEDPEDVEHHH-------HHH    302
```

BLAST algoritmus

BLAST (Basic Local Alignment Search Tool)

Heuristický algoritmus jehož základem je **hledání slov** (několikapísmenných sekvencí), s dostatečnou podobností (poskytují dostatečně vysoké skóre v substituční matici)

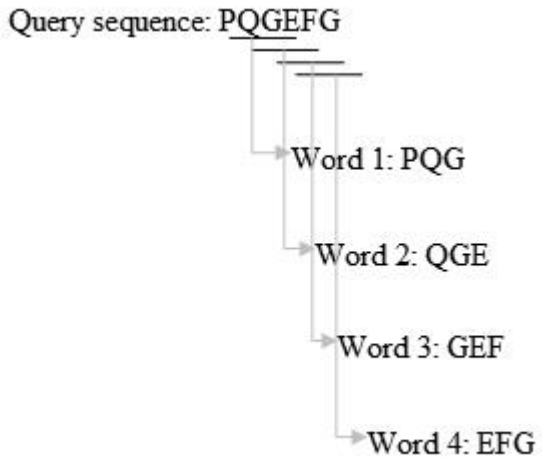


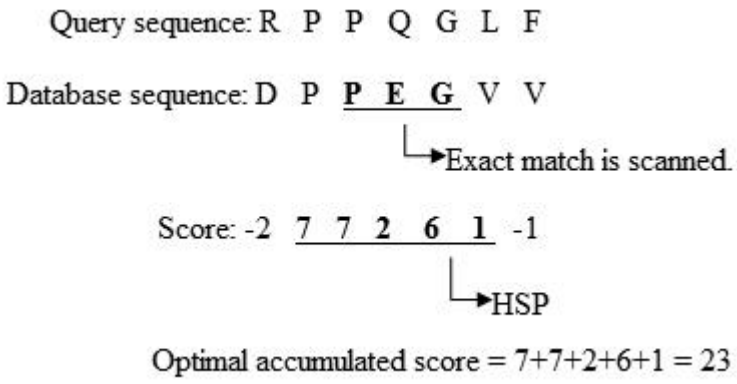
- **Tvorba k-písmenných slov ze vstupní sekvence**
  pro proteiny typicky 3-písmenných (v případě DNA 11-písmenných)

- **Porovnání slov na základě substituční matice**
  algoritmus BLAST hledá na základě vloženého skóre slova, která jsou podobná každému slovu v zadané sekvenci. Vyhovující slova jsou následně uspořádána.

- **Prohledání databázových sekvencí**
  Je hledána shoda s nalezenými vysoce podobnými slovy.

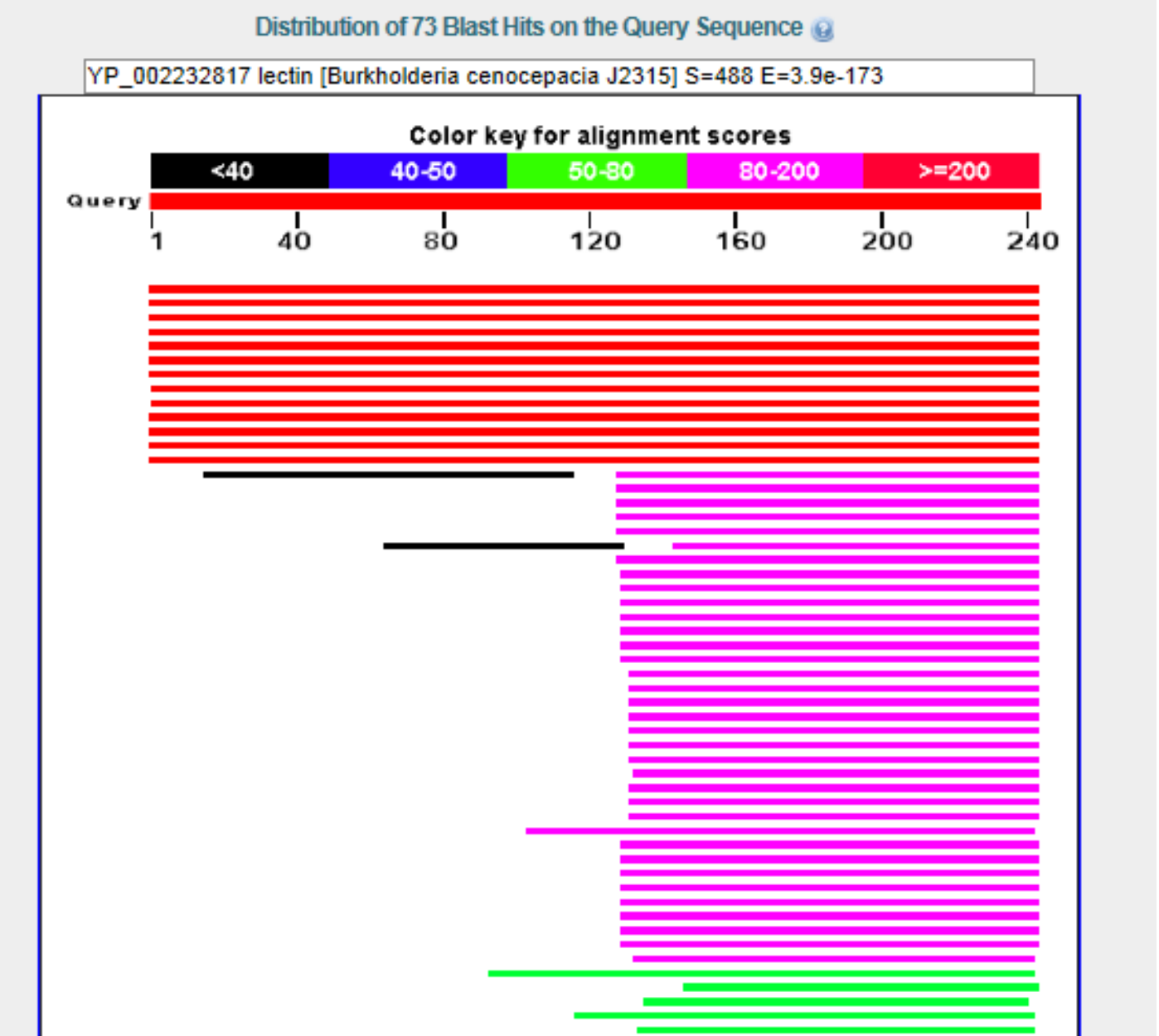
- **Rozšíření slov na segmenty**
  Přesné shody slov s databázovými sekvencemi jsou rozšiřovány oběma směry. To pokračuje dokud skóre pro tuto dvojici sekvencí je dostatečně vysoké.





Novější verze BLASTu (BLAST2) má mj. níže nastavenu hladinu pro hledání podobných slov, což rozšiřuje možnost nalezení vzdálenějších homologů.

Výstup z BLASTu



Výstup z BLASTu

Download GenPept Graphics

fucose-binding lectin II [Burkholderia multivorans ATCC BAA-247]

Sequence ID: ref|ZP_15916739.1| Length: 274 Number of Matches: 1

See 1 more title(s)

Range 1: 31 to 274 GenPept Graphics

Score	Expect	Method
443 bits(1140)	4e-155	Compositional matrix adjus

```
Query  2    QPFTHDDLYALLQLAGNDATAVQ
            QPFTHDDLYALLQLAGNDA AVQ
Sbjct  31   QPFTHDDLYALLQLAGNDAKAVQ

Query  62   SFDVGSWQNKVKTDAAGEVVACT
            SFDVGSWQNKVKTDAAG+VVACT
Sbjct  91   SFDVGSWQNKVKTDAAGQVVACT

Query  120  PAPVPTGGGERDGIFTLPPNIAF
             P     GGGERDG+F LPPNIAF
Sbjct  151  PDTATAGGGERDGVFNLPPNIAF

Query  180  LNTQIVNSGKGKVRVVVTANGKF
             LNTQIVNSG GKVRVVVT NGKF
Sbjct  211  LNTQIVNSGNGKVRVVVTVNGKF

Query  240  WPLG  243
             WPLG
Sbjct  271  WPLG  274
```

Download GenPept Graphics

sugar-binding lectin protein [Ralstonia solanacearum PSI07]

Sequence ID: ref|YP_003750856.1| Length: 114 Number of Matches: 1

See 3 more title(s)

Range 1: 3 to 114 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
124 bits(312)	2e-32	Compositional matrix adjust.	62/114(54%)	80/114(70%)	2/114(1%)

```
Query  130  RDGIFTLPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGK  189
            + G+FTLP N  FGVTA  N++  QTI+V VD+  K  ATF G+GT D  L +Q++NSG+
Sbjct  3    QQGVFTLPANTNFGVTAFANAANTQTIKVLVDNVVK--ATFSGSGTSDKLLGSQVLNSGR  60

Query  190  GKVRVVVTANGKPSKIGSRQVDIFKKTYFGLVGSEDGGDGDYNDGIAILNWPLG  243
            G V++ V+ NGKPS + S Q  +  K  F +VGSED  D DYNDGIA+LNWPLG
Sbjct  61   GAVQIQVSVNGKPSDLVSNQTILANKLNFAMVGSEDSSDNDYNDGIAVLNWPLG  114
```

Download GenPept Graphics

fucose-binding lectin PA-IIL [Pseudomonas aeruginosa ATCC 25324]

Sequence ID: ref|ZP_15618368.1| Length: 115 Number of Matches: 1

See 1 more title(s)

Range 1: 5 to 115 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
117 bits(294)	7e-30	Compositional matrix adjust.	61/113(54%)	77/113(68%)	3/113(2%)

```
Query  132  GIFTLPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGK-G  190
            G+FTLP N  FGVTA  NSS  QT+ V V  N + AATF G  T +A + TQ++NSG  G
Sbjct  5    GVFTLPANTQFGVTAFANSSGTQTVNVLV--NNETAATFSGQSTNNAVIGTQVLNSGSSG  62

Query  191  KVRVVVTANGKPSKIGSRQVDIFKKTYFGLVGSEDGGDGDYNDGIAILNWPLG  243
            KV+V V+ NG+PS + S QV +  +  F LVGSEDG D DYND + ++NWPLG
Sbjct  63   KVQVQVSVNGRPSDLVSAQVILTNELNFALVGSEDGTDNDYNDAVVVINWPLG  115
```

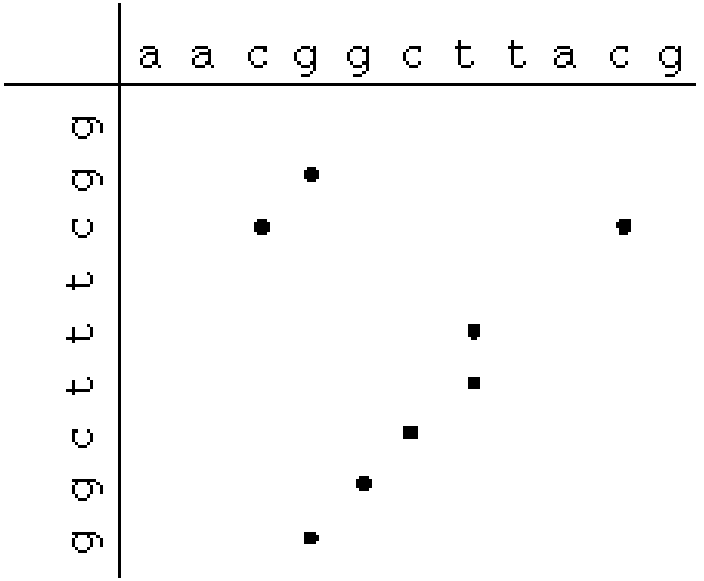
FASTA algoritmus

Na rozdíl od algoritmu BLAST jsou zde tolerovány mezery.

Proces:

Obě porovnávané sekvence tvoří horizontální a vertikální osu grafu.

Následně jsou jednotlivá slova z jedné sekvence porovnávána se slovy sekvence druhé. Odpovídající páry pak vytvoří sadu bodů. Body na úhlopříčce signalizují významnou shodu (či podobnost). Cílem je nalezení nejdelšího shodného úseku (úseku s nejvyšším skóre).



Příklad porovnání sekvencí GGCTTTCGG a AACGGCTTACG

V dalších krocích jsou zahrnuty konzervativní změny pro nejlepší úseky z prvního prohledání. Program pak vyhledává možnost spojení více takových úseků (může mezi nimi být mezera, či jsou na různých diagonálách) a tyto spojené úseky jsou posouzeny z hlediska zadaných kriterií.

MSA „programy“

- Za posledních 15 let vzniklo pres 50 MSA programových balíků (Wallace, I. M., O'Sullivan, O., Higgins, D. G. and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. **34**, 1692-1699.)

- Clustal W (Thompson et al., 1994)
- Clustal X (Thompson et al., 1997)
- Dialign2 (Morgenstern, 1999)
- T-Coffee (Notredame et al., 2000)
- MAFFT (Katoh et al., 2002)
- MUSCLE (Edgar, 2004)
- Kalign (Lassmann, 2005)
- …

Clustal http://www.clustal.org/

- V současné době **nejužívanější** program
- První verze 1988
  Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, 73, 237–244.
- Dnes používané verze:
  Clustal W (Thompson et al., 1994)
  Clustal X (Jeanmougin et al., 1998)
  Clustal Ω (Sievers et al., 2011)
- Využívá progresivní alignment

ClustalW: Jednotlivým sekvencím přiřazuje **váhy** (weight – W) podle četnosti zastoupení (čím více jsou si sekvence podobné, tím nižší mají váhu a naopak) a penalizuje přítomnost mezer v závislosti na jejich pozici (position-specific gap penalties)

Clustal – postup

1. Provedení **pairwise alignmentů** pro každou dvojici sekvencí a určení jejich podobnosti – v závislosti na množství neodpovídajících residuí a mezer

2. Sestavení **příbuzenského stromu** (similarity tree)

3. **Kombinace** alignmentů (viz. 1.) v pořadí dle příbuznosti – od nejvíce podobných k nejméně příbuzným (viz. 2.). Jednou vložené mezery jsou zachovány.

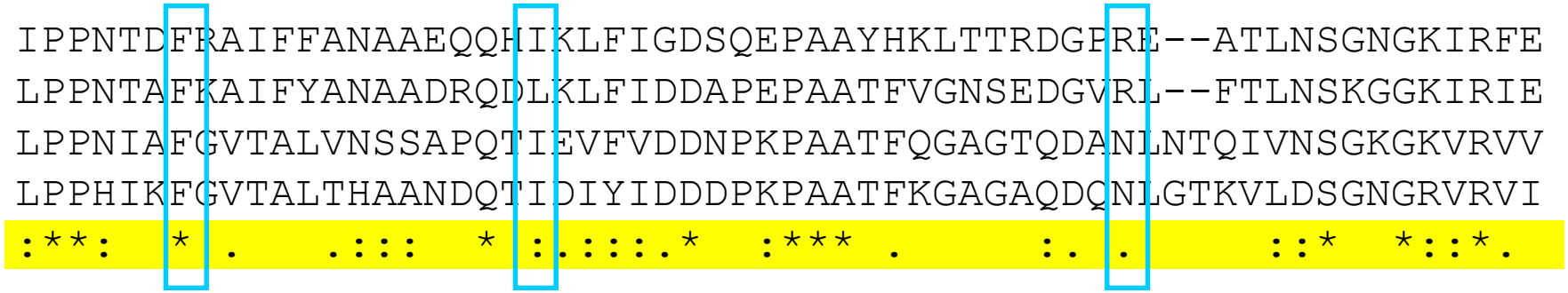
Clustal – výstup

Pod alignmentem je uváděn tzv. **consensus** – dohodnuté symboly vyjadřující „konzervovanost“ každého sloupce:

* - identické residuum ve všech sekvencích

: - silně konzervovaný sloupec

. - slabě konzervovaný sloupec



MUSCLE

(**MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation)

http://www.drive5.com/muscle

Rychlejší určení „vzdálenosti“ dvou sekvencí

Tzv. log-expectation skórovací funkce

Refinement metodou restricted partitioning

Vhodný i pro velký počet sekvencí (5000 seq po 350 bp za 7 min na PC – rok 2004)

Postup:

1. Sestavení matice pro každou dvojici sekvencí, určení jejich „vzdálenosti“ a sestavení matice vzdáleností (distance matrix)
2. Na základě distance matrix je sestaven první příbuzenský strom (tree1)
3. Skládání sekvencí v pořadí dle tree1 od větví ke kmenu – v každém rozvětvení je vytvořen profil, který při dalším porovnávání nahrazuje původní sekvence – výsledkem je první MSA

Algoritmus MUSCLE (podobne PRRP a MAFFT)

4. Přepočítání vzdáleností sekvencí na základě vzniklého MSA1 – tvorba druhé distance matrix (D2)
5. Na základě D2 sestaven vylepšený příbuzenský strom (tree2)
6. Progresivní alignment (viz bod 3) na základě tree2 – vytvoření druhého MSA

---

7. **Refinement** – rozdělení vzniklého stromu na dvě části a vytvoření MSA pro každou z nich. Pokud je výsledný alignment lepší, je zachován. Toto se opakuje do konvergence (žádná další změna nevede k lepšímu výsledku) nebo do určeného počtu kroků

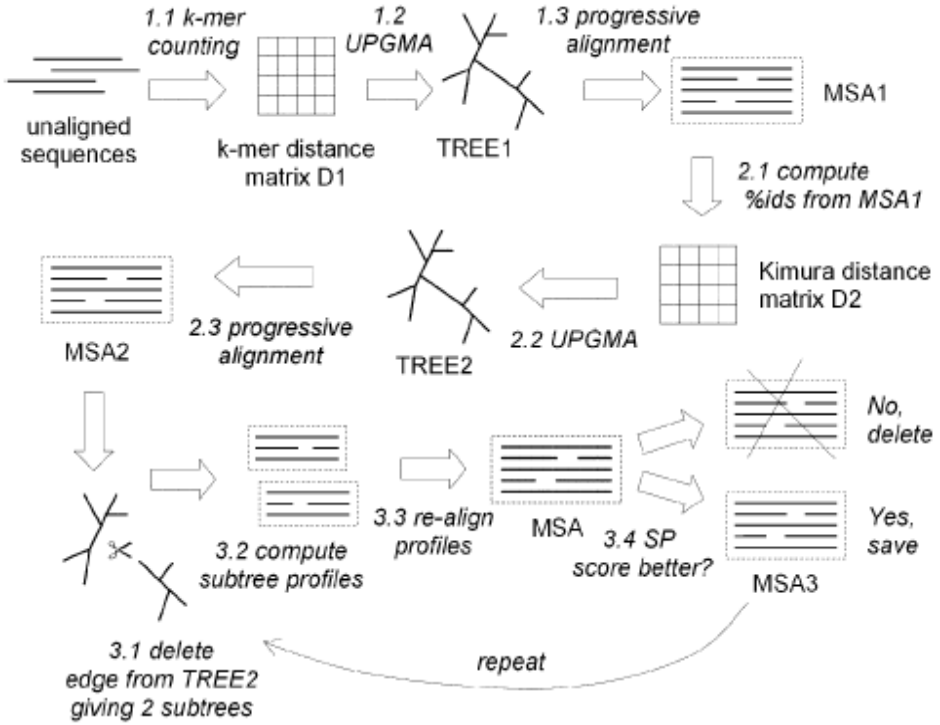


Figure 2. This diagram summarizes the flow of the MUSCLE algorithm. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). A multiple alignment is available at the completion of each stage, at which point the algorithm may terminate.

Další skórovací schémata (scoring schemes) pro pairwise alignment

Algoritmy založené na matici (matrix-based algorithms) – např. ClustalW, MUSCLE; pomocí substituční matice je příslušné dvojici (AK) přiřazena hodnota. Rozhoduje pouze **identita** těchto dvou **AK**, případně jejich **nejbližší okolí** (viz. např. BLAST)

Schémata založená na konzistenci (consistency-based schemes) – poprvé v T-Coffee, dále v PCMA, ProbCons, MUMMALS, MAFFT, aj. Vychází z nejlepších možných alignmentů každé dvojice sekvencí. Využívá často i **data z různých zdrojů** (např. strukturní informace). Cílem je dosáhnout maximální konzistence (vnitřní shody). Výsledek je přesnější, ale výpočet je časově náročnější.

T-Coffee

http://www.tcoffee.org

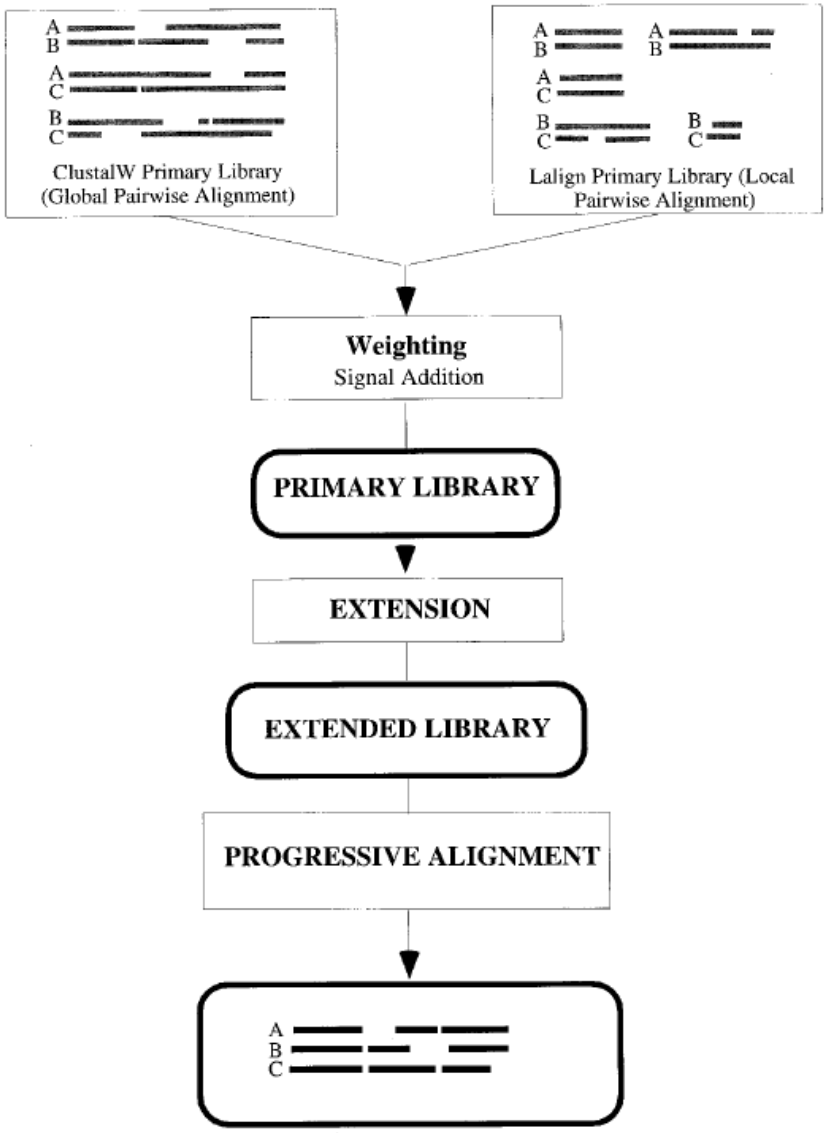
(Tree-based Consistency Objective Function for alignment Evaluation)

- Pomalejší ale výrazně přesnější než ClustalW
- Je schopen kombinovat data z více předchozích alignmentů, které mohly být vytvořeny různými postupy (lokální, globální, strukturní podobnost,…)

Hlavním rozdílem oproti tradičním metodám progresivního alignmentu je použití pozičně specifického skórovacího schématu (**extended library**) namísto substituční matice.

T-Coffee

1) Provedení pairwise alignmentů pro všechny dvojice sekvencí pomocí **globálního** a pomocí **lokálního alignmentu** (dvě primární knihovny).
2) Jednotlivým pairwise alignmentům je přiřazena **váha** podle poměru počtu identických residuí k celkovému počtu residuí.
3) Kombinace obou knihoven. Pokud je rozdíl v globálním a lokálním alignmentu, jsou zachovány oba s příslušnou váhou. Vzniká **pozičně specifická matice** (extended library), která je dále použita pro vlastní progresivní alignment.



Clustal Ω

1. Provedení **pairwise alignmentů** urychleno použitím modifikovaného algoritmu mBed – převedení sekvencí na n-rozměrný vektor a následný alignment vektorů

2. Sestavení **příbuzenského stromu** (similarity tree)

3. **Sestavení** alignmentů užitím přesného algoritmu HHalign (využití skrytých Markovových modelů).

Určen pro obsáhlé alignmenty.

V roce 2011 přiloženo 190 000 sekvencí během několika hodin.

Zlepšení přesnosti – strukturní informace

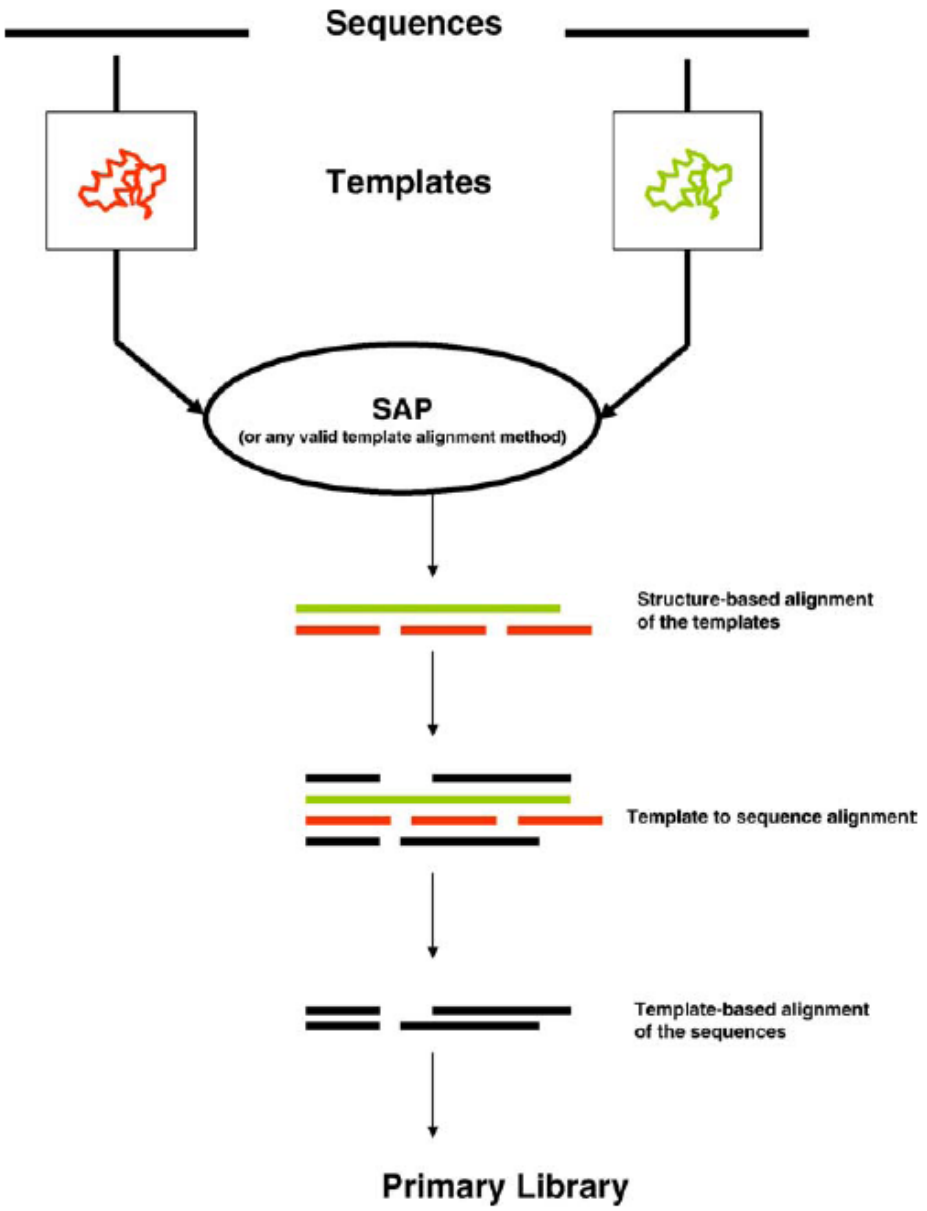
- Sekvence s vyšší homologií (>40%) – vysoká přesnost alignmentu
- Bez homologie – nepoužitelné
- Tzv. twilight zone – málo podobné sekvence (nižší než 20% homologie) = špatná (méně než 30%) přesnost alignmentu

Řešení: nejčastěji využití znalosti strukturní podobnosti (2D nebo 3D), která se behem evoluce zachovává více než sekvence AK.

Rozšíření konzistentního modelu

Template-based alignment metody – využití známých homologních proteinů (srovnání dle jejich struktury nebo tvorba profilu homologních sekvencí)

Výhoda: vyšší přesnost



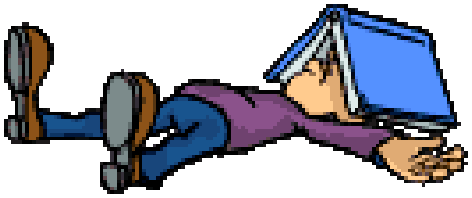
Expresso

- Je založeno na T-Coffee

Expresso: MSA server, který srovnává sekvence za užití strukturní informace. Po zadání sekvencí vyhledá v databázi struktur (PDB) pomocí BLASTu homology a použije je jako templáty pro následný alignment zadaných sekvencí pomocí metod MSA založených na struktuře (např. SAP, Fugue).

Zopakování / shrnutí

- **Alignment** – přiložení sekvencí (2 nebo více) na základě podobnosti
- **Využití** pro hledání příbuznosti sekvencí, tvorba profilů proteinových rodin, aj.
- Řada **programů** využívajících rozdílné přístupy – použití závisí na vstupních datech a účelu
- Nejčastěji používaný (ClustalW) neznamená nejpřesnější – každý program je **kompromisem mezi přesností a rychlostí**
- Každý alignment potřebuje **lidskou kontrolu** !!!



Benchmark (srovnávací testy)

BAliBASE - První vytvořená sada benchmarkových testů pro multiple alignment programy (Thompson et al., 1999) – byla vytvořena pomocí manuálně provedeného alignmentu

Na základě srovnání 3D struktur byly vytvořeny další sety:
HOMSTRAD [Mizuguchi *et al.*, 1998].
OxBench [Raghava *et al.*, 2003]
PREFAB [Edgar, 2004]

Benchmark (srovnávací testy)

Existují i **specificky zaměřené benchmarkové sety**, např.

IRMBASE [Subramanian *et al*, 2005] – náhodné (nepřiložitelné) sekvence s vloženými motivy. Slouží k testování metod pro lokální alignment

BAliBASE [Thompson *et al.*, 1999] contains eight reference sets, each dealing with a different type of alignment problem. Ref1 deals with test cases containing small numbers of equidistant sequences, and is further subdivided by percent identity. Ref2 alignments contain "orphan", or unrelated, sequences. Ref3 test cases contain a pair of divergent subfamilies, with less than 25% identity between the two groups. Ref4 is concerned with long terminal extensions, while Ref5 test cases contain large internal insertions and deletions. Test sets from References 6-8 deal with problems like transmembrane regions, inverted domains, and repeat sequences. In previous versions of BAliBASE, test cases were confined to homologous regions. In practice, the boundaries of such regions may be unknown. The current version [Thompson *et al.*, 2005] now also provides duplicate test cases containing full-length sequences. Only the first five reference sets are used here, as they have been corrected and verified in the latest release.

OxBench [Raghava *et al.*, 2003] comprises 3 related datasets. Test cases in the MASTER set deal with isolated domains derived exclusively from sequences of known structure. The FULL set was generated from suitable MASTER test cases, using full-length sequence data. High scoring homologous sequences were added to each MASTER test case to generate the EXTENDED set. The results from this third set, however, are not used here. It was found that some of the test cases in the EXTENDED set proved too large for some programs, and aborted due to excessive memory requirements. Of the 276 test cases selected from EXTENDED, T-COFFEE returned 235 alignments, and Align-m was only able to align 107, using a single processor with 4GB of RAM.

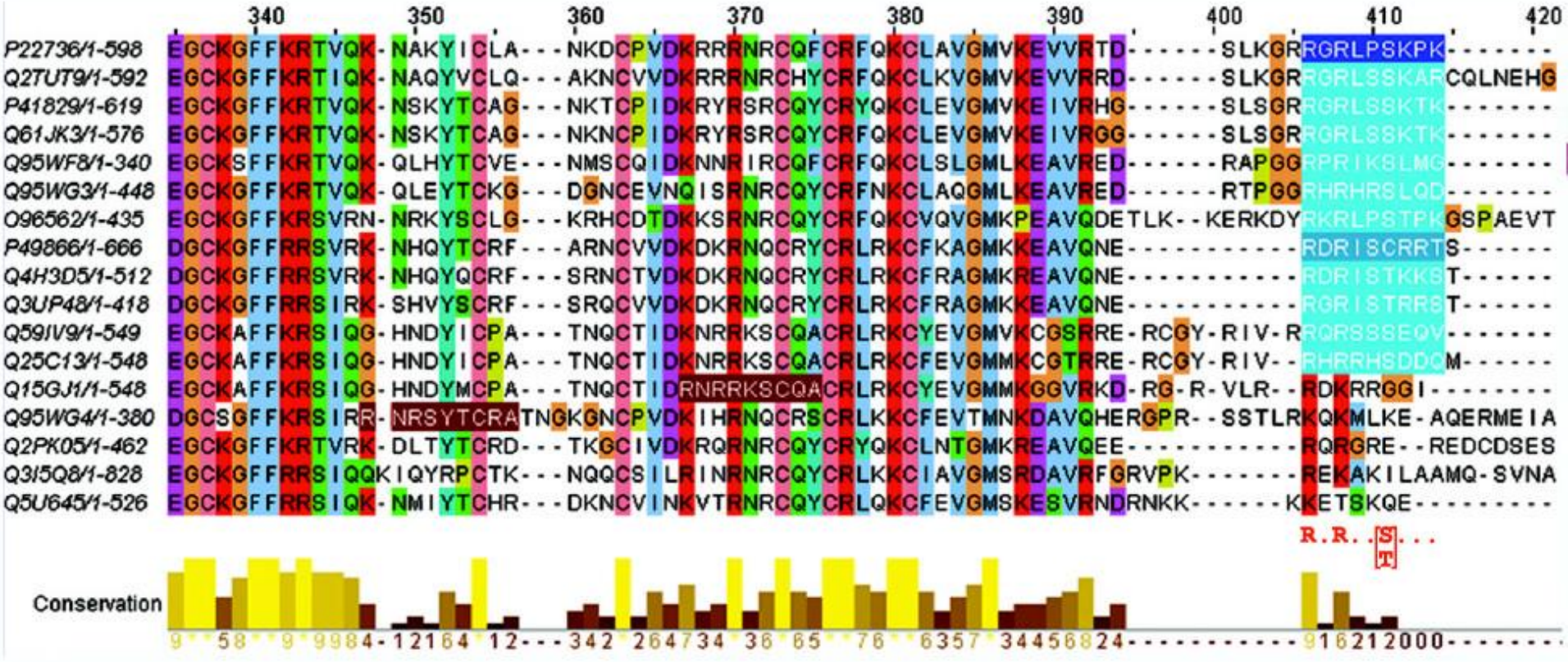
PREFAB [Edgar, 2004] test cases are generated by taking a pairwise alignment of sequences of known 3D structure, and adding up to 24 high scoring homologues for each sequence. Accuracy is assessed on the structural alignment of the original pair alone.

SABmark [Van Walle *et al.*, 2005] is divided into two subsets. Each test group in the SUPERFAMILY set represents a SCOP superfamily, whose sequences are 25-50% identical. Each test group in the TWILIGHT set represents a common SCOP fold and sequences are 0-25% identical. In addition, these two subsets are also provided with non-homologous (false positive) sequences included within each group. Instead of a single alignment acting as a reference, SABmark provides multiple pairwise references for each test, and it is the average score from each of these references that is taken here as a score for each test case.

IRMBASE [Subramanian *et al.*, 2005] test cases contain a number of simulated motifs [Stoye *et al.*, 1998] inserted into otherwise random (unalignable) sequences, and as such is entirely different to the other benchmarks used in this study. Test cases are designed to examine whether a method can detect isolated motifs within sequences, and so are tailored to a local alignment approach.

HOMSTRAD [Mizuguchi *et al.*, 1998] is a database exclusively based on protein structures derived from the PDB, arranged into homologous protein families. It was not specifically designed as a benchmark database, although it is regularly employed as such.

BaliBASE – ukázka alignmentu



Perrodou *et al. BMC Bioinformatics* 2008 **9**:213 doi:10.1186/1471-2105-9-213

Table 1: Programs used in this investigation.

Method	OVERVIEW
Align_m (2.3) [Van Walle et al., 2004]	http://bioinformatics.vub.ac.be/software/software.html Local, specialised for highly divergent sequences.
ClustalW (1.8) [Thompson et al., 1994]	http://www.ebi.ac.uk/clustalw/ Global, progressive alignment package.
Dialign2 (2.2) [Morgenstern, 1999]	http://bibiserv.techfak.uni-bielefeld.de/dialign/ Local, aligns segments of sequences rather than individual residues.
Dialign-t (0.1.3) [Subramanian et al., 2005]	http://dialign-t.gobics.de/ Local, progressive alignment. Recent re-implementation of Dialign2.
MAFFT (5.531) [Katoh et al., 2002]	http://www.biophys.kyoto-u.ac.jp/~katoh/programs/align/mafft/ Suite of alignment programs:
FFTNS	Global, uses Fast Fourier Transform to generate tree.
FFTNSi	As FFTNS, but with iteration step to refine alignment.
NWNS	Global, uses traditional Needleman-Wunsch algorithm.
NWNSi	As NWNS, but with iteration step to refi
FINSi	Local, iterative, uses local pairwise alig
GINSi	Global, iterative, uses global pairwise alignment information.
MUSCLE (3.6) [Edgar, 2004]	http://www.drive5.com/muscle/ Global, iterative, progressive alignment program that uses Log Expectation as scoring function.
ProbCons (1.09) [Do et al., 2005]	http://probcons.stanford.edu/ Global, uses posterior-probabilities from HMMs and pairwise alignment consistency.
PCMA (2.0) [Pei et al., 2003]	ftp://iole.swmed.edu/pub/PCMA/ Global, switches alignment strategies dependent on sequence data. ClustalW is used to align highly similar sequences and to form pre-aligned groups. T-COFFEE is used to align the more divergent groups.
POA (v2) [Lee et al., 2002]	http://www.bioinformatics.ucla.edu/poa/ Local; uses Partial Order graphs.
T-COFFEE (1.37) [Notredame et al., 2000]	http://igs-server.cnrs-mrs.fr/~cnotred/Projects_home_page/t_coffee_home_page.html Combines both global and local methods; uses consistency.

Blackshield 2006 oznacil ProbCons jako nejlepsi na zaklade 6 benchmarkovych testu

Local alignment

- For two-sequence comparisons, there is the well-known Smith and Waterman (1981) algorithm. Here we use Lalign
- For multiple sequences, the Gibbs sampler (Lawrence et al., 1993) and Dialign2 (Morgenstern, 1999) are the main automatic methods. These programs often perform well when there is a clear block of ungapped alignment shared by all of the sequences. They perform poorly, however, on general sets of test cases when compared with global methods

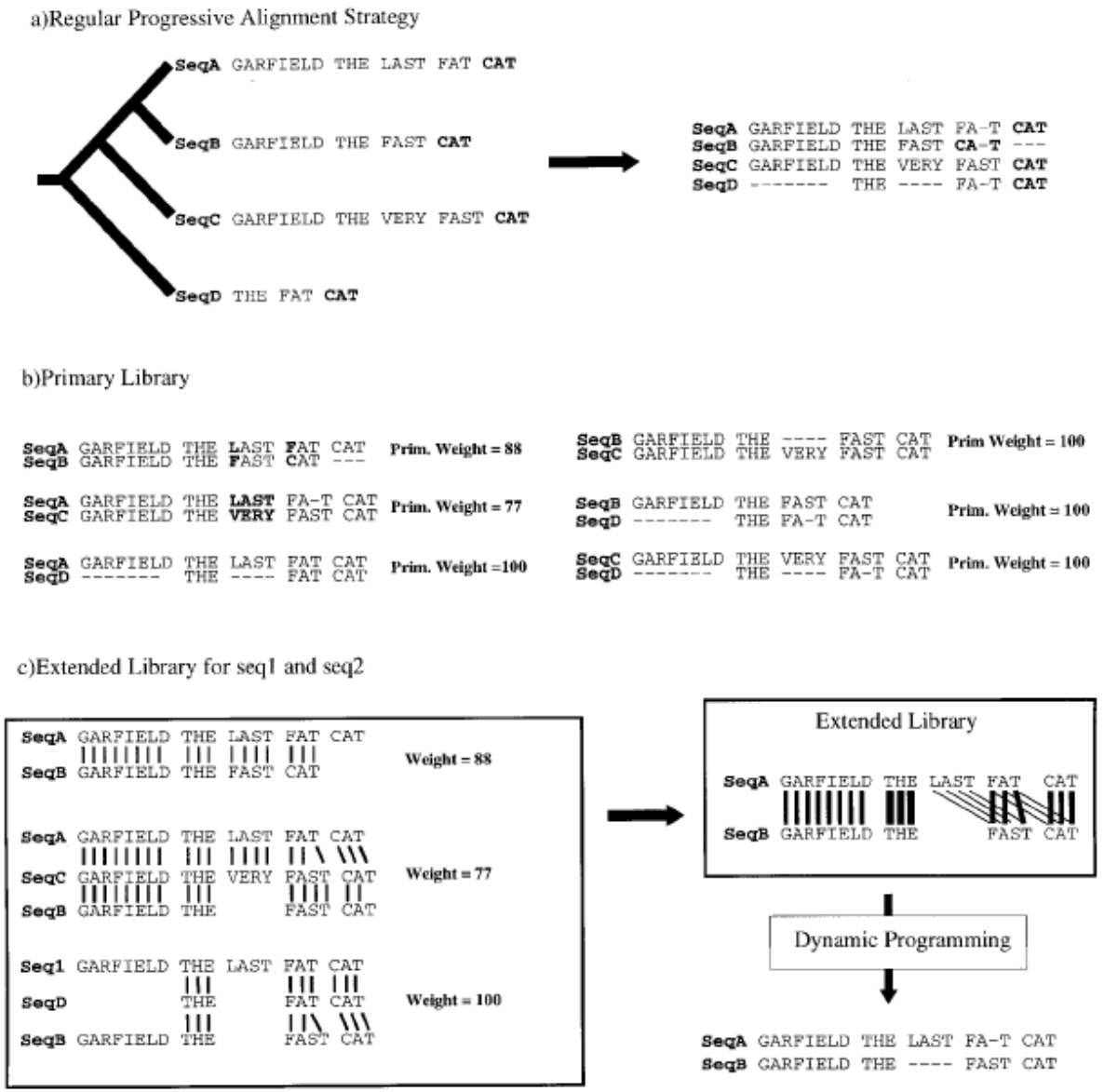


Figure 2. The library extension. (a) Progressive alignment. Four sequences have been designed. The tree indicates the order in which the sequences are aligned when using a progressive method such as ClustalW. The resulting alignment is shown, with the word CAT misaligned. (b) Primary library. Each pair of sequences is aligned using ClustalW. In these alignments, each pair of aligned residues is associated with a weight equal to the average identity among matched residues within the complete alignment (mismatches are indicated in bold type). (c) Library extension for a pair of sequences. The three possible alignments of sequence A and B are shown (A and B, A and B through C, A and B through D). These alignments are combined, as explained in the text, to produce the position-specific library. This library is resolved by dynamic programming to give the correct alignment. The thickness of the lines indicates the strength of the weight.

Method	Score	Templates	Validation Values		Server
			PreFab	HOMSTRAD	
ClustalW [14]	Matrix	—	61.80 [12]	—	http://www.ebi.ac.uk/clustalw/
Kalign	Matrix	—	63.00 [18]	—	http://msa.cgb.ki.se/
MUSCLE [6]	Matrix	—	68.00 [16]	45.0 [9]	http://www.drive5.com/muscle/
T-Coffee [10]	Consistency	—	69.97 [12]	44.0 [9]	http://www.tcoffee.org/
ProbCons [7]	Consistency	—	70.54 [12]	—	http://probcons.stanford.edu/
MAFFT [8]	Consistency	—	72.20 [12]	—	http://align.genome.jp/mafft/
M-Coffee [12]	Consistency	—	72.91 [12]	—	http://www.tcoffee.org/
MUMMALS [16]	Consistency	—	73.10 [16]	—	http://prodata.swmed.edu/mummals/
DbClustal [24]	Profiles	—	—	—	http://bips.u-strasbg.fr/PipeAlign/
PRALINE [9]	Matrix	Profiles	—	50.2 [9]	http://zeus.cs.vu.nl/programs/pralinewww/
PROMALS [16]	Consistency	Profiles	79.00 [16]	—	http://prodata.swmed.edu/promals/
SPEM [28]	Matrix	Profiles	77.00 [28]	—	http://sparks.informatics.iupui.edu/Softwares-Services_files/spem.htm
Expresso [13]	Consistency	Structures	—	71.9 [11][a]	http://www.tcoffee.org/
T-Lara [29]	Consistency	Structures	—	—	https://www.mi.fu-berlin.de/w/LiSA/

Validation values were compiled from several sources, and selected for comparability. PreFab validations were made using PreFab version 3. HOMSTRAD validations were made on datasets having less than 30% identity. The source of each value is indicated by the accompanying reference citation.

[a]The Expresso value comes from a slightly more demanding subset of HOMSTRAD (HOM39) made of sequences less than 25% identical.

doi:10.1371/journal.pcbi.0030123.t001