

Chemoinformatika a bioinformatika

Sequence alignment



Osnova

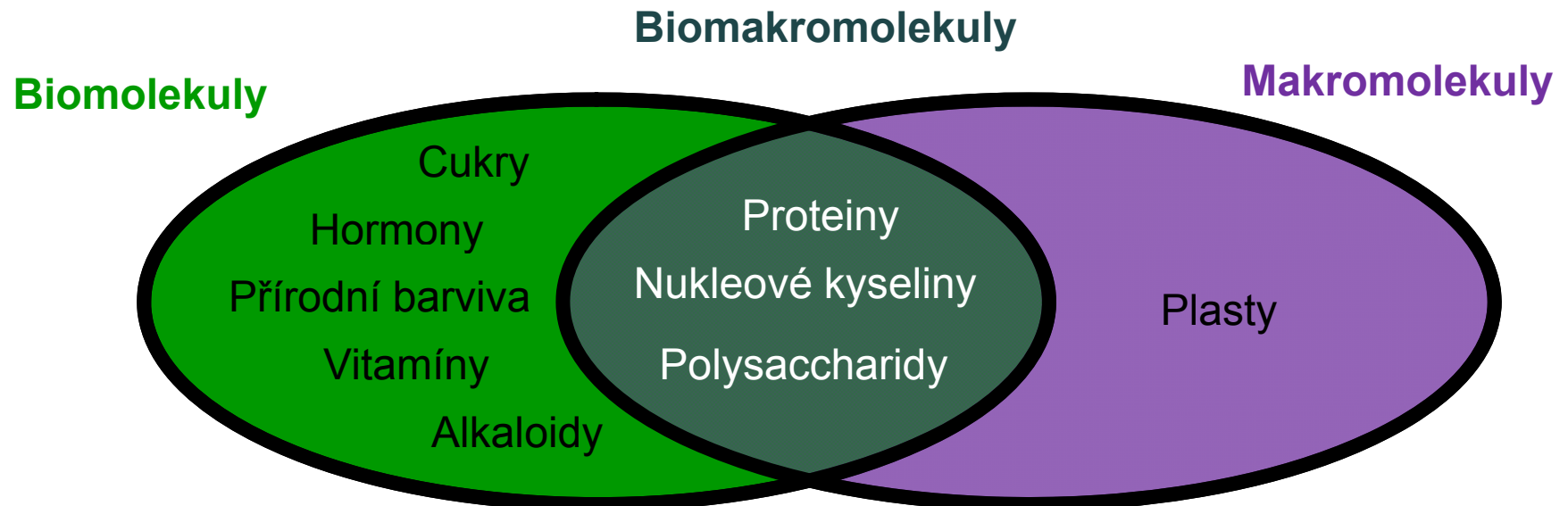
1. Struktura biomakromolekul – sekvence
2. Alignment a jeho typy
3. Užívané algoritmy
4. Multiple sequence alignment
5. Programové balíky

Biomakromolekuly

Biomolekuly jsou přirozenou součástí živých organismů.

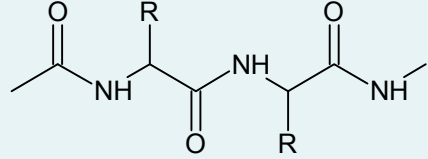
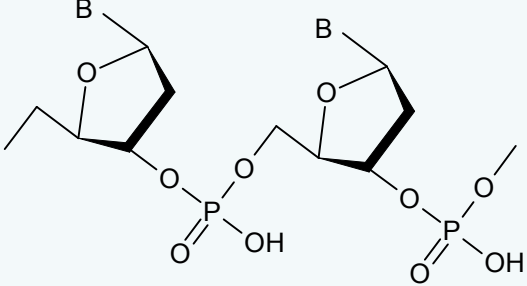
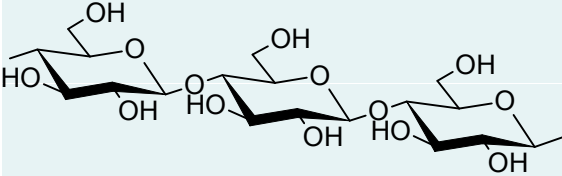
Velké molekuly. Typické malé molekuly jsou tvořeny několika atomy až několika sty atomy. Makromolekuly tvoří tisíce až miliony atomů.

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

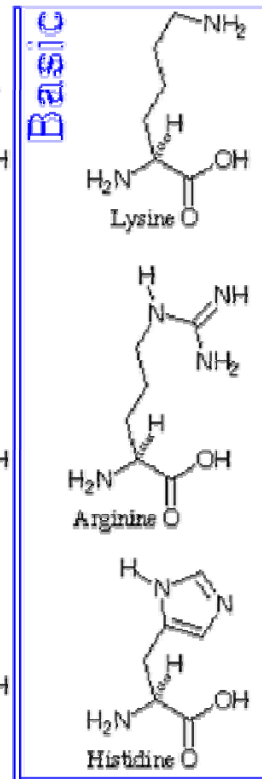
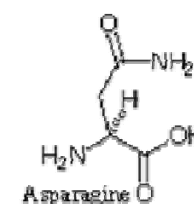
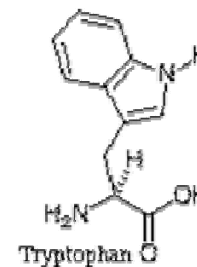
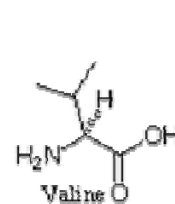
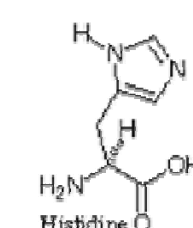
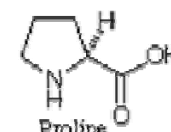
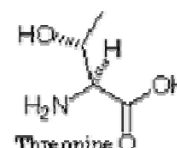
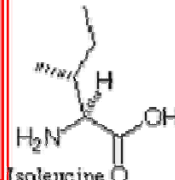
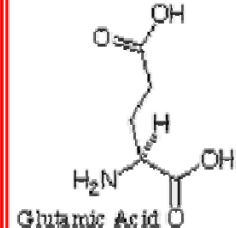
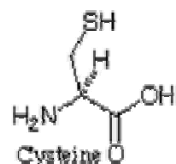
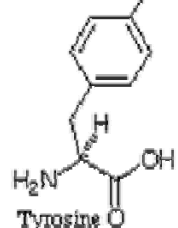
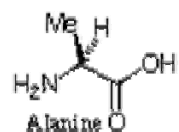
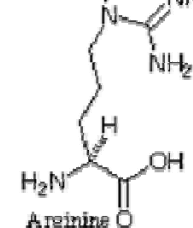
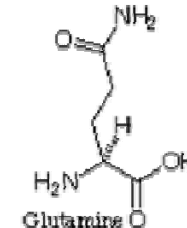
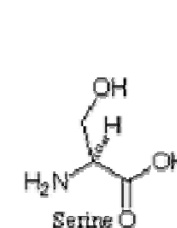
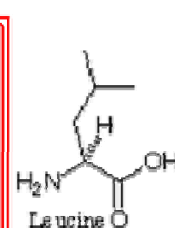
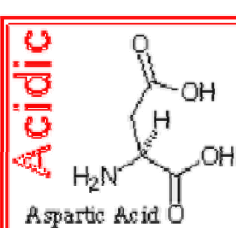
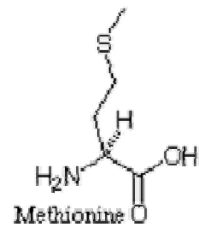
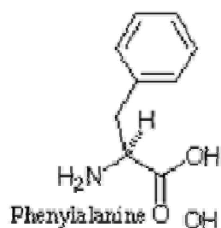
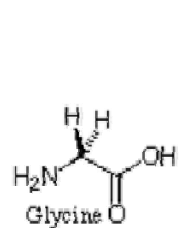


Složení biomakromolekul

- Vznikají spojováním velkého množství několika málo typů podjednotek

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

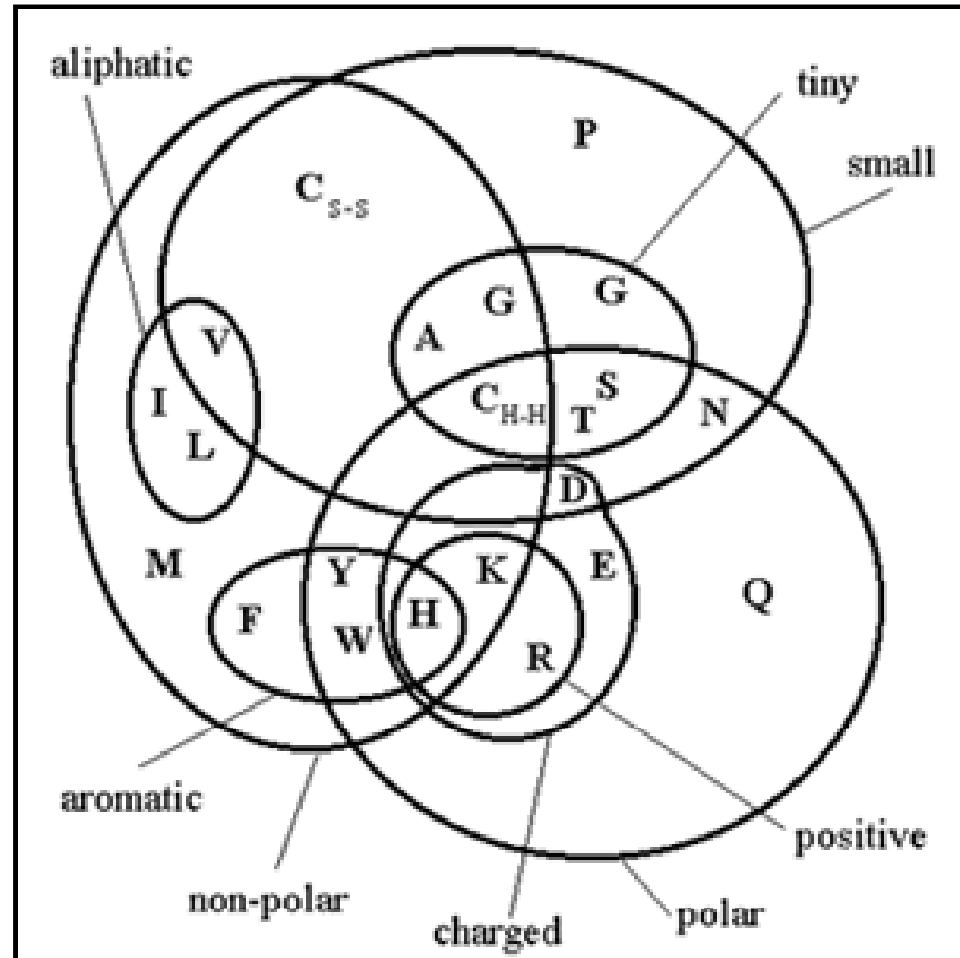
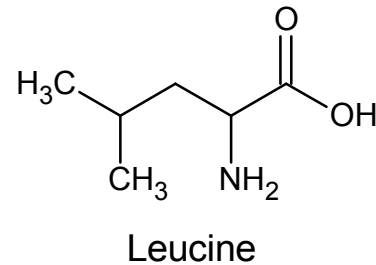
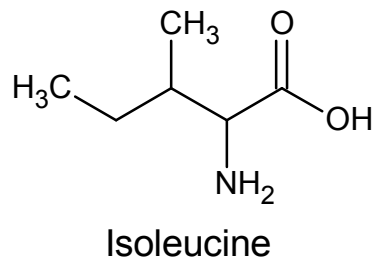
Aminokyseliny



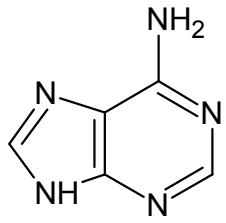
glycin	alanin	valin	leucin	izoleucin	asparagová kys.	asparagin	glutamová kys.	glutamin	arginin	lysin	histidin	fenylalanin	serin	threonin	tyrozin	tryptofan	methionin	cystein	prolin	selenocystein	pyrolysin
Gly	Ala	Val	Leu	Ile	Asp	Asn	Glu	Gln	Arg	Lys	His	Phe	Ser	Thr	Tyr	Trp	Met	Cys	Pro	Sec	Pyr
G	A	V	L	I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

Třídění aminokyselin

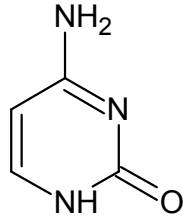
Aminokyseliny s podobnými vlastnostmi mohou plnit v proteinu stejné funkce – bývají vzájemně zastupitelné



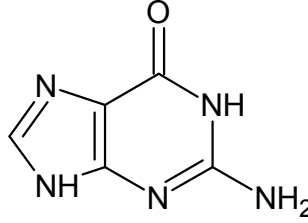
Nukleové báze



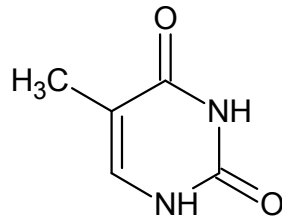
Adenine



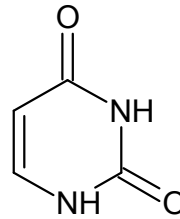
Cytosine



Guanine



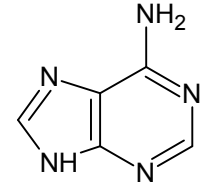
Thymine



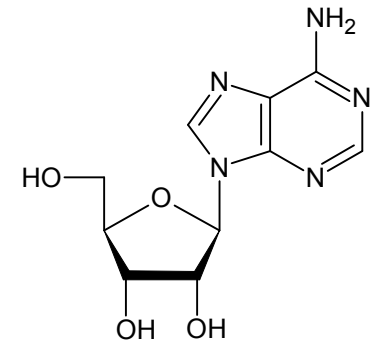
Uracil

adenin	cytosin	guanin	thymin	uracil
A	C	G	T	U

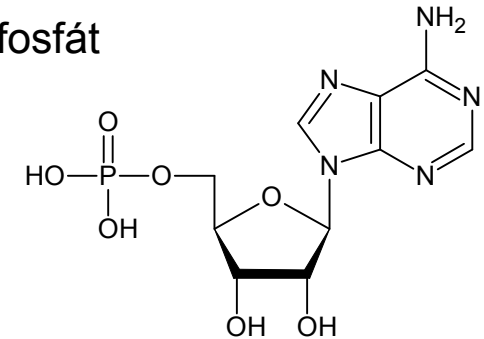
Nukleová báze
Adenin



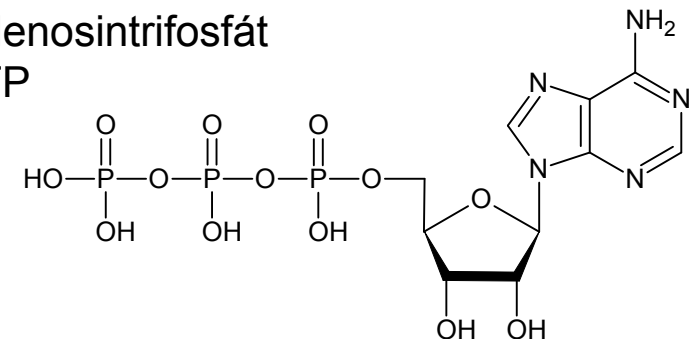
Nukleosid
Adenosin



Nukleotid
Adenosinmonofosfát
AMP



Nukleotid
Adenosintrifosfát
ATP



Polysacharidy

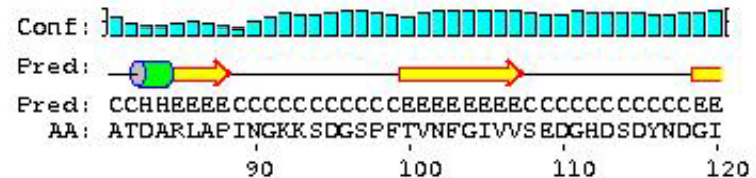
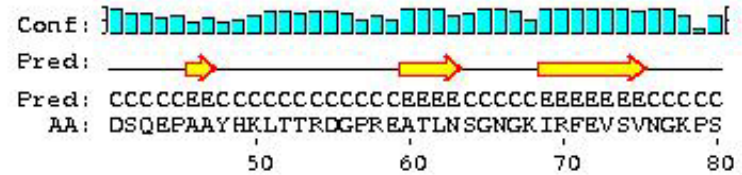
Komplikované sekvence – alignment se neprovádí

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)

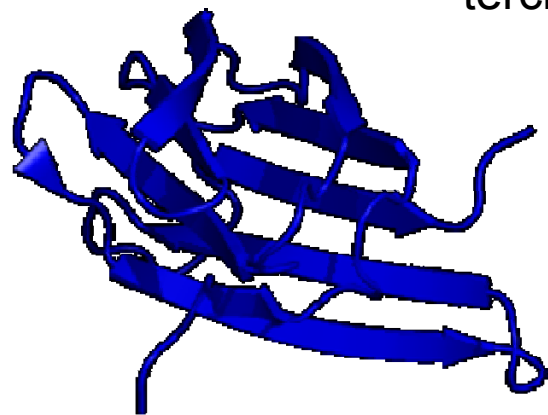
Struktura proteinů (NK)

ADSQTSSNRAGEFSIPPNTDFRAIF
FANAAEQQHKLFIGDSQEPAAYHK
LTTRDGPREATLN SGNGKIRFEVSV
NGKPSATDARLAPINGKKSDGSPF
TVNFGIVVSEDGHDSYNDGIVVL
QWPIG

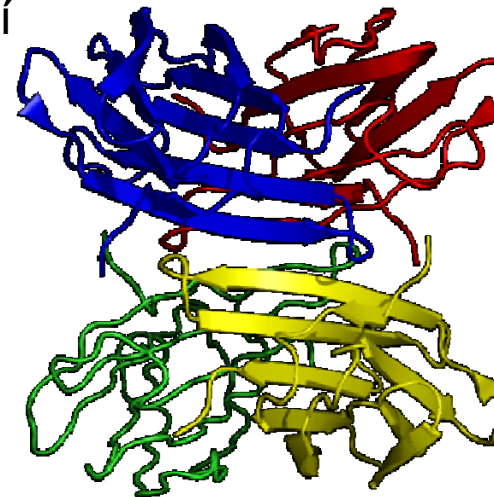
primární
(sekvence)



sekundární



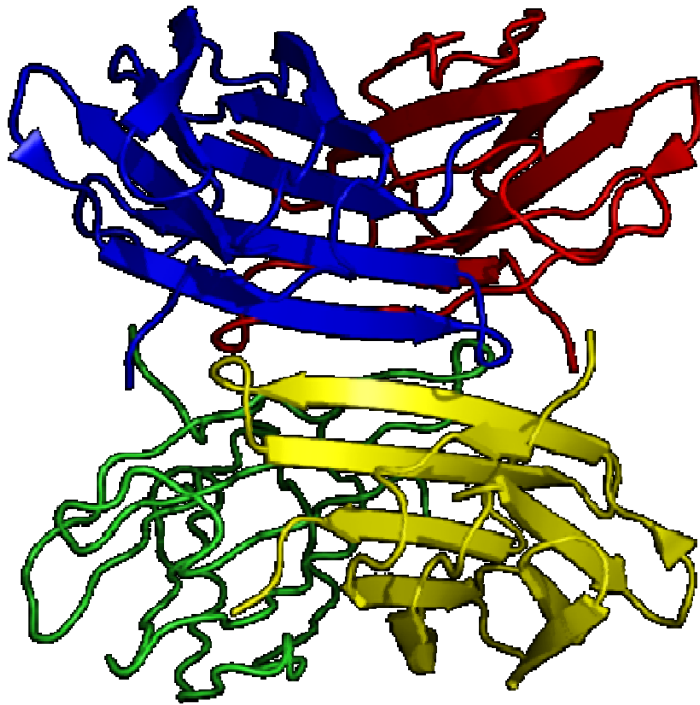
kvarterní



Kvartérní struktura proteinů

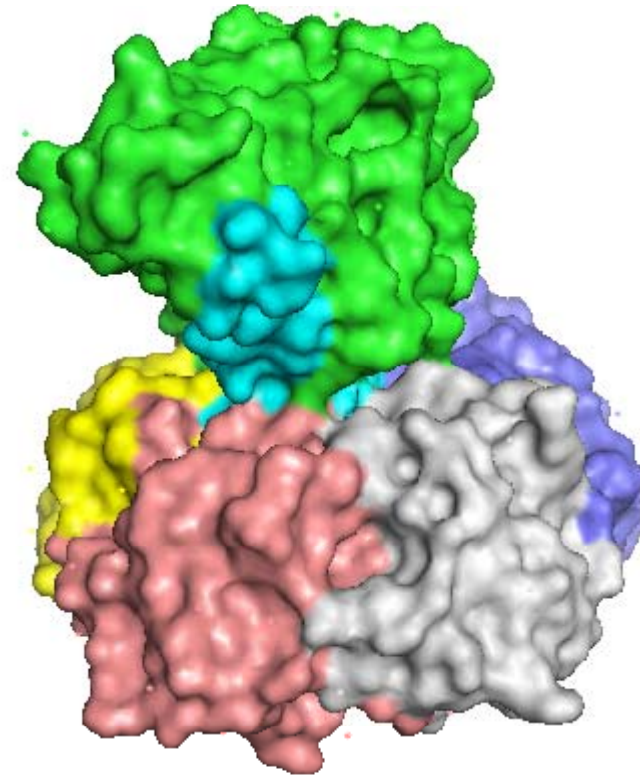
Homooligomer

Homotetramer



Heterooligomer

AB5 toxin



Jsou sekvence stejné, podobné či zcela odlišné?

ATGTCTACTCCTGGAGCACAGCAAGTCCTCTTCCGCACCGGAATTGCCGCGGTCAACTCAACCAACCATCTCCGTGTTTACTTCCAGGATGTCTATGGCAG
TATTCGCGAGAGTCTCTACGAGGGCAGCTGGGCTAACGGCACCCGAAAAGAAGCTTATCGGCAATGCTAAGCTTGGCAGCCCTGTGGCCGCGACTTCTAAG
GAGCTGAAGCATATCCGTGTCTACACCCTACTGAAGGAAACACCTACAGGAGTTCGCCTACGACTCCGGAACCGGATGGTACAACGGCGGGCTGGGC
GGTGCAAAGTTCCAAGTCGCACCCCTACTCTCGCATTGCTGCCGTGTTCCCTAGCCGGAACAGATGCATTGCAGTTGCGAATCTATGCACAGAAGCCAGATAA
CACAATCCAGGAGTATATGTGGAACGGCGATGGCTGGAAGGAGGGCACCAACCTGGGAGGTGCTCTCCCCGGCACTGGAATCGGAGCCACCTCCTTCCG
CTATACCGACTACAATGGCCCAAGCATCCGGATCTGGTTCAAACCTGACCTCAAACCTCGTCCAAAGAGCCTACGACCCGCACAAAGGCTGGTACCCGGAC
CTCGTCACCATCTTTGACAGGGCACCGCCACGTACGGCCATTGCAGCCACCAGCTTTGGAGCCGGCAACAGTTCCATCTACATGCGTATCTACTTTGTCAA
TTCGGACAACACTATCTGGCAGGTCTGCTGGGACCACGGCAAGGGCTATCACGACAAGGGAACCATCACCCAGTCATTACAGGGCTCGGAGGTGCCATT
ATCAGCTGGGGCAGTTTCGCCAATAACGGGCCGGATCTGCGTCTGTACTTTCAGAATGGAACATACATTAGTGCTGTGAGCGAGTGGGTTTGAATCGGG
CACATGGGTCGCAGTTGGGCAGAAGTGCTCTTCTCCTGCTTGA

ATGGCTGATTCTCAAACGTCATCCAACCGCGCCGGCGAATTCTCGATTCCGCCGAATACCGATTTCCGCGCGATTTTCTTCGCGAATGCCGCCGAGCAACA
GCACATCAAATTGTTTCATCGGCGACAGCCAGGAACCCGCCGCGTATCACAAGCTGACGACGCGCGACGGCCCGCGCGAAGCCACGCTGAATTCGGCAA
CGGCAAGATCCGTTTCGAGGTGTCGGTGAACGGCAAGCCGTCGGCGACCGACGCGCGTCTCGCGCCGATCAACGGCAAGAAGTCGGACGGCTCGCCGT
TCACGGTCAACTTCGGGATCGTGTGTCGGAAGACGGCCACGACAGCGACTACAACGACGGCATCGTGTGCTCCAGTGGCCGATCGGCTGA

ATGCTGGTGATTGTGGATGCCGTTACCCTGCTGAGCGCCTATCCGGAAGCCAGCCGTGATCCGGCCGCCCCGACCGTGATTGATGGTCGCCACCTGTATG
TTGTTAGCCCGGGCGATGCCGCGCAGCTGGGCCATAACGATAGCCGTCTGTTTACCAGGTCTGAGCCCGGGTGATCAGCTGCATCTGCGCGAAACCGCGC
TGCGCTGCGCGCGGAAGTGAGCGTGCTGTTTATTGCTTTGCCCTGAAAGATGCCGGCATTGTTGCCCGATCGAACTGGAAGTGCCTGATGCCGCCAC
CGCCGTTCCGGATGCGGATGATCTGCTGCATCCGAGCTGTCGTCCGCTGAAAGATCATTATTGGCGCAGCGATGTGCTGGCGGCGGGCGCGACCACCTG
TACCGCCGATTTTTCGGTGTGCGATCGTATGGCACCGTGAGCGGTTATTTTCGTTGGGAAACCAGCATTGAAATTGCGGGCAGCCAGCCGATACCAA
CAGCCGGGCTTTAAACCGAGCAGCGATCGCAATGGCAACTTTAGCCTGCCGCCAATACCGCCTTTAAAGCGATCTTCTATGCGAACGCGGCCGATCGTC
AGGATCTGAAACTGTTTATTGATGATGCGCCGGAACCGGCCACCTTTGTGGGTAACAGCGAAGATGGTGTGCGTCTGTTTACCCTGAATAGCAAAGGT
GGTAAAATTCGTATTGAAGCGAGCGCAACGGCCGTCAGAGCGCGACCGATGCCGCTGTCGCGCGCTGAGCGCGGGCGATACCGTGTGGCTGGGCTG
GCTGGGCGCGGAAGATGGTGCCGATGCGGATTATAATGATGGCATTGTTATTCTGCAGTGGCCGATTACCTAA

ATGTCGAGCGTTCAAACCGCTGCCACTTCGTGGGGAACCGTACCGTGCATCCGTGTGTACACGGCCAATAATGGCAAGATCACCGAGCGATGCTGGGACG
GGAAGGGGTGGTACACCGGTGCCTTCAACGAGCCCGGCGATAACGTCTCCGTAACCAGCTGGCTGGTCGGCAGCGCGATCCATATCCGCGTCTATGCAA
GCACCGGCACCACGACCACGGAGTGGTGTGCGGACGGCAACGGCTGGACCAAGGGCGCCTACACCGCCACGAACTGA

ATGCCGCTGCTGAGCGCCAGTATCGTGAGCGCGCCGGTGGTGACCAGCGAAACCTATGTGGATATTCGGGCCTGTATCTGGATGTTGCGAAAGCCGGTA
TCCGTGATGGCAAACCTGCAGGTTATCCTGAATGTGCCGACCCCGTATGCGACGGGCAATAACTTTCCGGGTATTTATTTTTCGATCGCCACCAACCAGGGC
GTGGTGGCGGATGGTTGCTTTACGTATAGTAGCAAAGTGCCGGAAAGTACGGGCCGTATGCCGTTTACCCTGGTTGCGACCATTGATGTGGGTAGCGGTG
TTACCTTCGTGAAAGGTCAGTGAAATCTGTTCCGCGGCTCTGCGATGCATATTGATAGCTATGCAAGCCTGAGTGCATTTGGGGCACCGCGGCACCGAGT
TCTCAGGGTTCTGGTAACCAGGGTGCGGAAACGGGTGGCACCGGTGCCGTAATATTGGTGGCGGCGGTGAACGTGATGGCACCTTTAATCTGCCGCCG
CATATTAATTCGGTGTACCGCGCTGACCCACGCGGCGAACGATCAGACCATTGATATTTATATTGATGATGATCCGAAACCGGCAGCCACCTTTAAAGGC
GCGGGCGCGCAGGATCAGAACCTGGGTACCAAAGTGCTGGATTCTGGCAATGGCCGTGTTCCGCTTATCGTTATGGCGAACGGCCGTCCGAGCCGCGCTG
GGTTCTCGTCAGGTGGATATTTTTAAAAATCTTATTTCCGGTATTATTGGCTCTGAAGATGGTGCAGGATGATGATTATAACGATGGCATCGTGTCTGAACT
GGCCGCTGGGCTAA

ATGCCGCTCCTGAGCGCCAGTATCGTGAGCGCGCCGGTGGTGACCAGCCAAACCTATGTGGATATTCGGGCCTGTATCTGGATGTTGCGAAAGCCGGTA
TCCGTGATGGCAAACCTGCAGGTTATCCTGAATGTGCCGACCCCGTATGCGACGGGCAATAACTTTCCGGGTATTTATTTTTCGATCGCCACCAACCAGGGC
GTGGTGGCGGATGGTTGCTTTACGTATAGTAGCAAAGTGCCGGAAAGTACGGGCCGTATGCCGTTTACCCTGGTTGCGACCATTGATGTGGGTAGCGGTG
TTACCTTCGTGAAAGGTCAGTGAAATCTGTTCCGCGGCTCTGCGATGCATATTGATAGCTATGCAAGCCTGAGTGCATTTGGGGCACCGCGGCACCGAGT
TCTCAGGGTTCTGGTAACCAGGGTGCGGAAACGGGTGGCACCGGTGCCGTAATATTGGTGGCGGCGGTAAAGCTTGGCGCCGCACTCGAGATCAAACGG
GCTAGCCAGCCAGAACTCGCCCCGGAAGACCCCGAGGATGTCGAGCACCAACCACCACTGA

Jsou sekvence stejné, podobné či zcela odlišné?

MSTPGAQQVLFRTGIAAVNLTNHLRVYFQDVYGSIRESLYEGSWANGTEKNVIGNAKLGSPVAATSKELKHIRVYTLTEGNTLQEFAYDSGTGWYNGGLGGAKFQ
VAPYSRIA AVFLAGTDALQLRIYAQKPDNTIQEYMWNGDGWKEGTNLGGALPGTGIGATSFRTYDYNGPSIRIWFQTDLKLVRAYDPHKGWYPDLVTIFDRAPP
RTAIAATSFAGAGNSSIYMRIYFVNSDNTIWQVCWDHKGKGYHDKGTITPVIQSEVAIISWGSFANNGPDLRLYFQNGTYISAVSEWVWNRAHGSQ LGRSALPPA
MADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHKLFIGDSQEPAAYHKL TTRDGPREATLNSGNGKIRFEVSVNGKPSATDARLAPINGKKS D GSPFTVNF GIV
VSE DGHDSYNDGIVVLQWPIG

MLVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVS PGDAAQLGHNSRLFTGLSPGDQLHLRETALALRAEVSVLFIRFALKDAGIVAPIELEVRDAATAVPDADDLL
HPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGT VSGYFRWETSIEIAGSQPDTKQPGFKPSSDRNGNFS LPPNTAFKAIFYANAADRQDLKLFIDDAPEPAA
TFVGNSEDGVRLFTLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWL GWLGAEDGADADYNDGIVILQWPIT

MSSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGA FNEPGDNVSVTSWLVGSAHIRVYASTGTTTTEWCWDGNGWTKGAYTATN

MPLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQW
KSVRGSAMHIDSYASLSAIWGTAAPSSQGSNGQAETGGTGAGNIGGGGERDGT FNLPPIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGA QDQNLG TKVL
DSGNGRVRVIVMANGRPSRLGSRQVDIFKKS YFGIIGSEDGADDDYNDGIVFLNWPLG

MPLLSASIVSAPVVT SQTYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQW
KSVRGSAMHIDSYASLSAIWGTAAPSSQGSNGQAETGGTGAGNIGGGGKLA AAEIKRASQPELAPEDPEDVEHHHHHH

Alignment

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

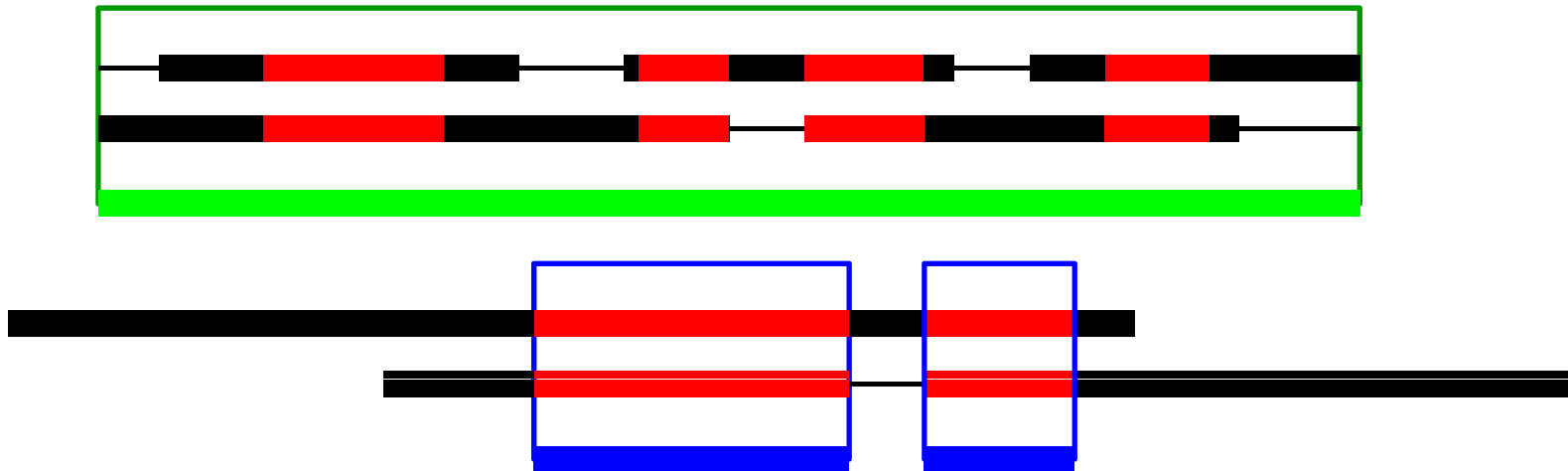


Význam alignmentu

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

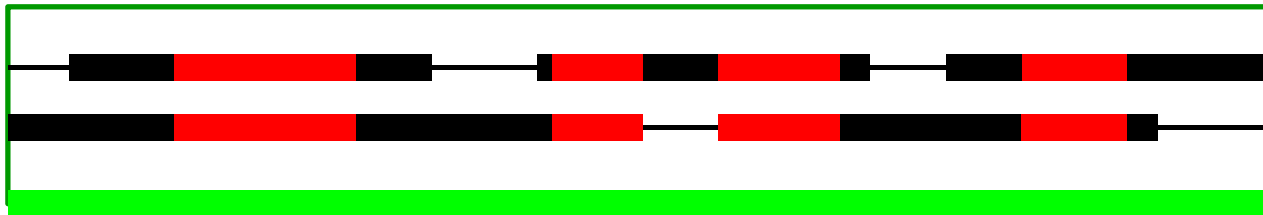
Pair-wise alignment

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



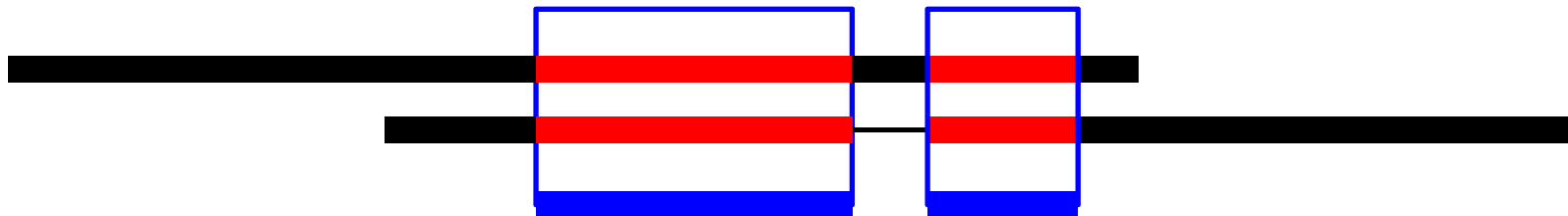
Global alignment

Vychází z předpokladu, že obě srovnávané sekvence jsou víceméně shodné v celé své délce. Alignment k sobě příkládá 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

- Těmeř 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
LTEGNTLQEFAYDSGTGWYNGGLGGAKFQVAPYSRIA AVFLAGTDALQLRIYAQKPDNTIQEYMWNGDGWKEGT
NLGGALPGTGIGATSFRTDYNGPSIRIWFQTDDLKLVQRAYDPHKGWYPDLVTIFDRAPPRTAIAATSFGAGNSS
IYMRIYFVNSDNTIWQVCWDHKGKGYHDKGTITPVIQGSEVAIISWGSFANNGPDLRLYFQNGTYISAVSEWWNR
AHGSQLGRSALPPA

>BC2LA

MADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHILFIGDSQEPAAAYHKLTTTRDGPREATLN SGNGKIRFEVSV
NGKPSATDARLAPINGKKS DGS PFTVNF GIVVSE DGHDS DYNDGIVVLQWPIG

> BC2LD

MLVIVDAVTLLSAYPEASRDPAAPTVIDGRHLYVVSPGDAAQLGHNDSRLFTGLSPGDQLHLRETALALRAEVS VL
FIRFALKDAGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYWRSDVLAAGATTCTADFAVCDRDGT VSGYFR
WETSIEIAGSQPDTKQPGFKPSSDRNGNFSLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNS EDGVRLF
TLNSKGGKIRIEASANGRQSATDARLAPLSAGDTVWLGWLGAE DGADADYNDGIVILQWPIT

>RSL

MSSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTE
WCWDGNGWTKGAYTATN

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

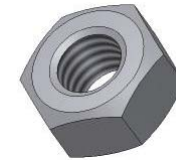
MPLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSS
KVPESTGRMPFTLVATIDVGSVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSNGQAETGGTGAG
NIGGGGERDGT FNLPPIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGA QDQNLG TKVLD SGN GRVRVIVMA
NGRPSRLGSRQVDIFKKS YFGIIGSE DGADDDYNDGIVFLNWPLG

>gi|283806765|pdb|2WQ4|A Chain A

MPLLSASIVSAPVVT SQTYVDIPGLYLDVAKAGIRDGKLVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSS
KVPESTGRMPFTLVATIDVGSVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSNGQAETGGTGAG
NIGGGGKLAAALEIKRASQPELAPEDPEDVEHHHHHH

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



Typy matic

- **PAM** (Point Accepted Mutation) – založena na mutacích v rámci globálního alignmentu, tj. ve vysoce konzervovaných 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 matrices 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 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).

Matrice BLOSUM 62

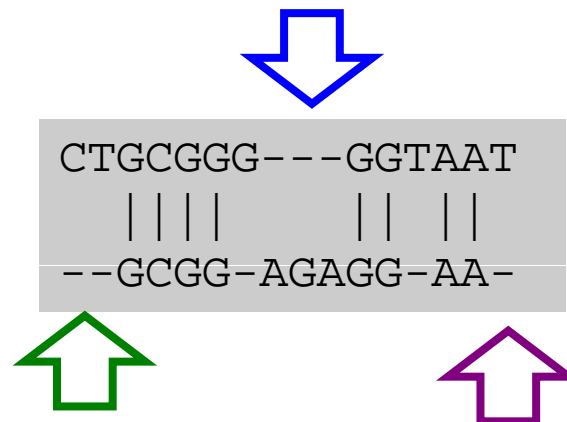
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3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

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 delecí. 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:

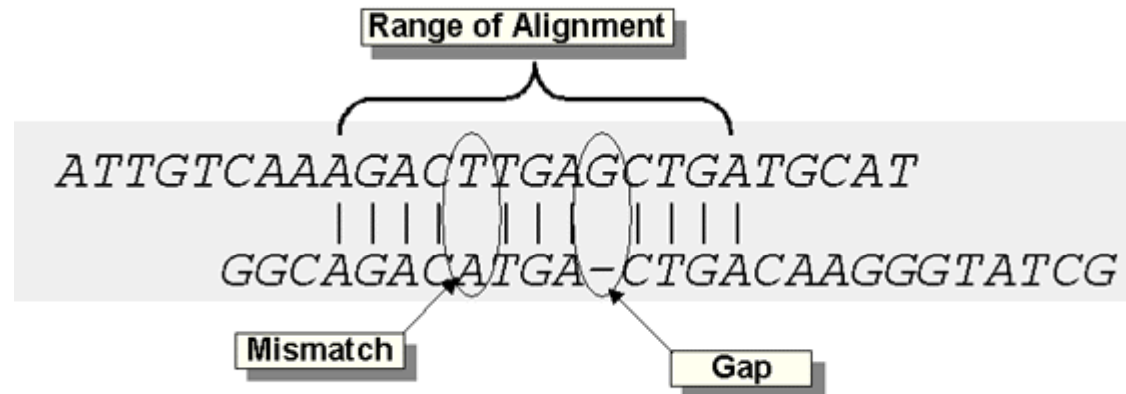
```
ATCTTCAGTGTTTCCCCTGTTTTGCCCG-ATTTAGTTCGCTC
|||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGATTTAGTTCGCTC
```

Dlouhá mezera:

```
ATCTTCAGTGTTTCCCCTGTTTTGCCCG-----ATTTAGTTCGCTC
|||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGCCCCCCCCCCCCCCCCCCCCCATTTAGTTCGCTC
```

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

AAEECCDDEEFF
AADDKKKEFFGG

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

AAYYCCDDEEF
AADDFFKKEFGG

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
-10-1-1-1-1 -10-1 = -24

Celkové skóre 32 – 24 = 8

Příklad výpočtu

AAYYCCDDEEF
AADDFFKKEFGG

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      = 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      = -11
```

Celkové skóre 32 – 11 = 21

DNA matice

A	<u>1</u>			
T	-10000	<u>1</u>		
G	-10000	-10000	<u>1</u>	
C	-10000	-10000	-10000	<u>1</u>
	A	T	G	C

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

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“ těchto 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



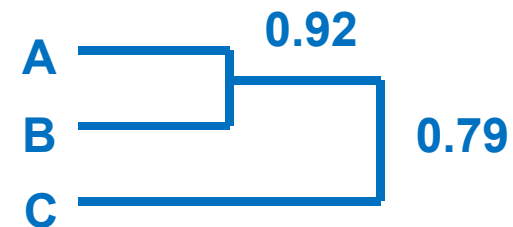
Progresivní multiple alignment

- 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 nepřiložených sekvencí

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

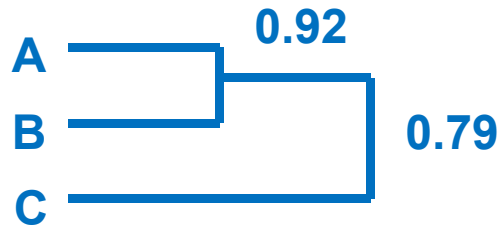


Počet exaktně stejných shod dělená celkovou délkou sekvence (ignoruje mezery)

Progresivní multiple alignment

- 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)
z nepřiložených sekvencí
- 2) tvorba párových alignmentů postupně podle příbuznosti
(topologie guide tree)
- Dnes obsahuje často iterativní smyčku

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,  
  RSIL: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)

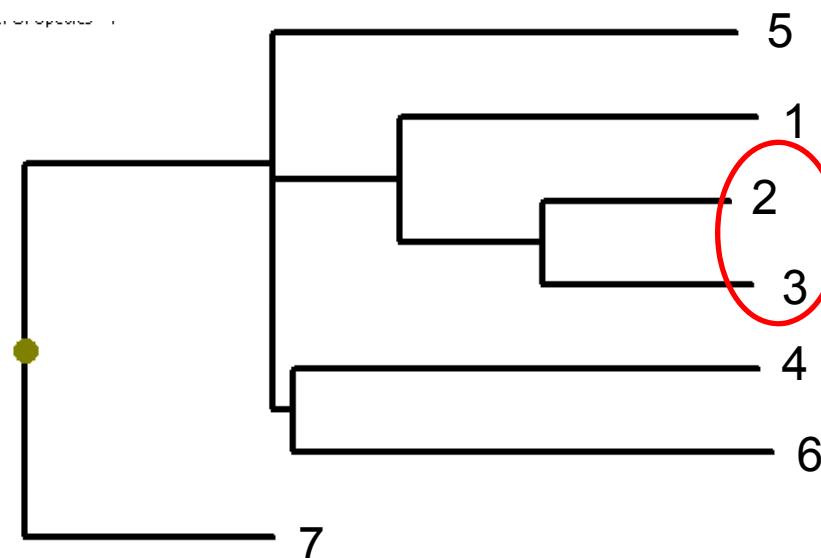
Cycle 5 (Last cycle, trichotomy):

Node: 1 (0.12171) joins

Node: 4 (0.01864) joins

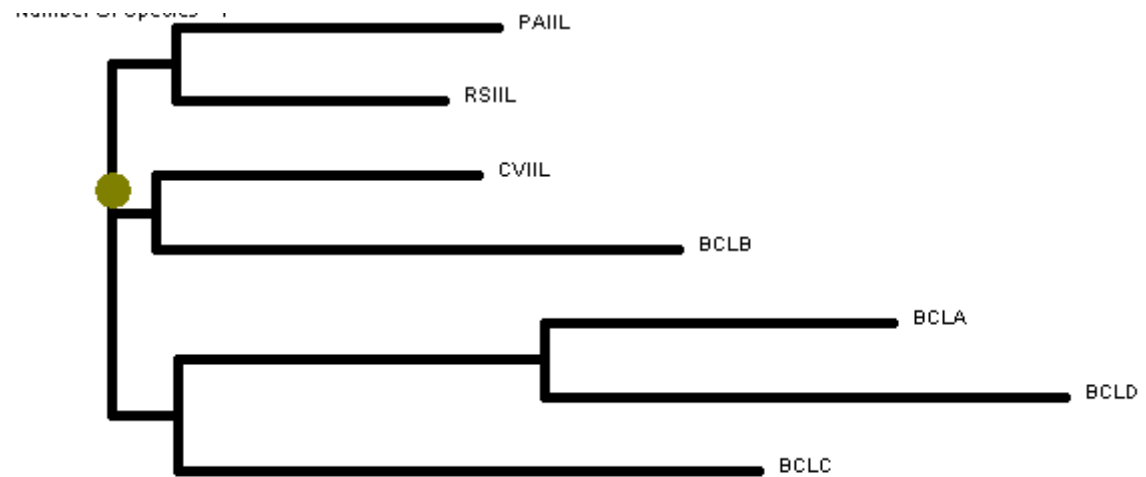
Node: 5 (0.02083)

.nj soubor



.ph soubor

```
(  
(  
(  
  PAIL:0.16435,  
  RSIL:0.13654)  
:0.03384,  
(  
(  
  BCLA:0.17899,  
  BCLD:0.26633)  
:0.18717,  
  BCLC:0.29707)  
:0.03484)  
:0.02264,  
  CVIL:0.16563,  
  
  BCLB:0.26800);
```



.dst soubor

7

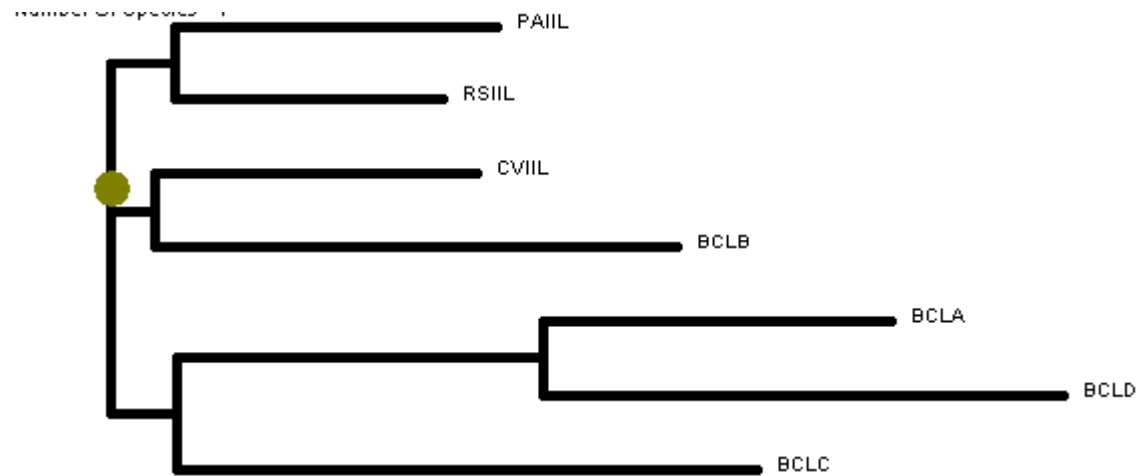
PAIL	0.000	0.649	0.684	0.930	0.904	0.939	0.982
RSIL	0.649	0.000	0.377	0.912	0.895	0.912	0.939
CVIL	0.684	0.377	0.000	0.912	0.939	0.930	0.947
BCLA	0.930	0.912	0.912	0.000	0.921	0.904	0.965
BCLB	0.904	0.895	0.939	0.921	0.000	0.956	0.921
BCLC	0.939	0.912	0.930	0.904	0.956	0.000	0.965
BCLD	0.982	0.939	0.947	0.965	0.921	0.965	0.000

Phylogram a cladogram

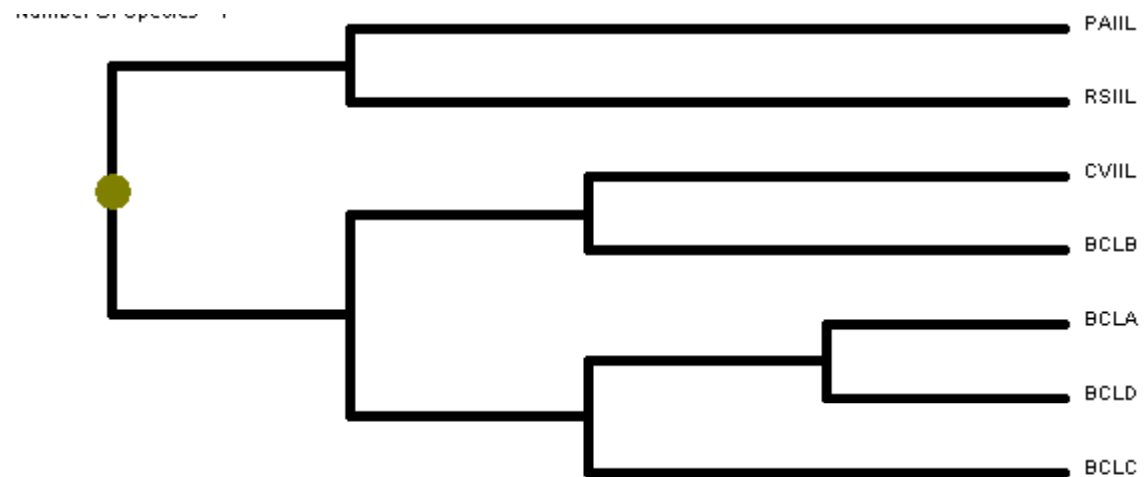
- **Phylogram** (phylogeny tree) – je rozvětvený diagram (strom), který naznačuje fylogenezi (postupný vývoj). Délka jednotlivých větví je úměrná **velikosti změny** v průběhu evoluce.
- **Cladogram** – rovněž strom, v němž však všechny větve mají **stejnou délku**. Ukazuje tak sice „společné předky“ pro jednotlivé sekvence, ale ne množství změn, jež od té doby prodělaly (evoluční dobu).

Phylogram a cladogram

Phylogram

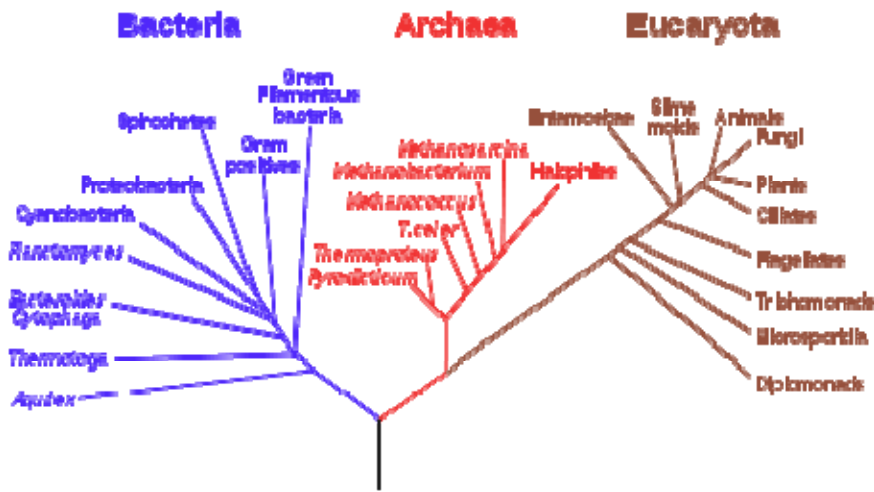


Cladogram

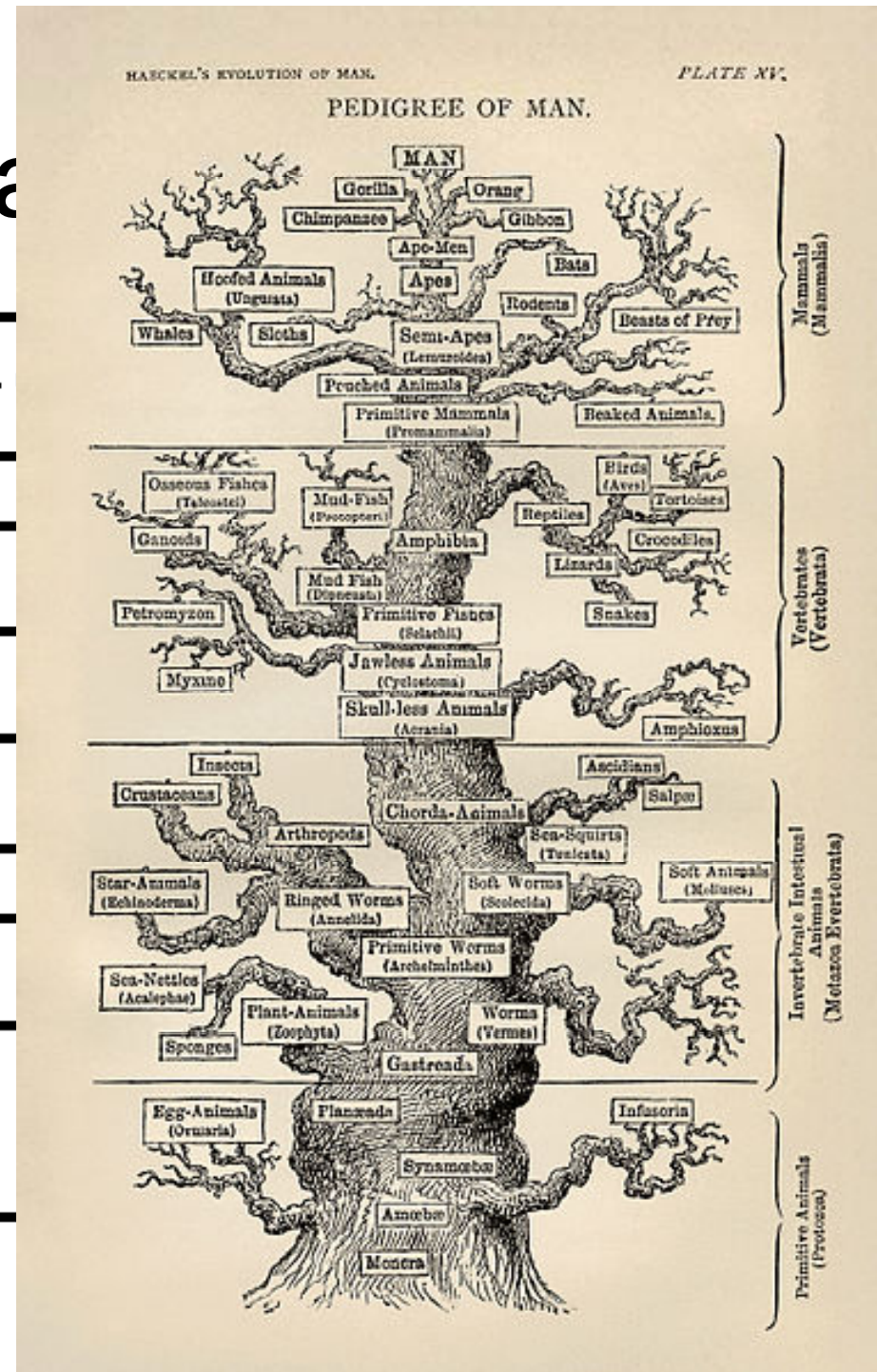
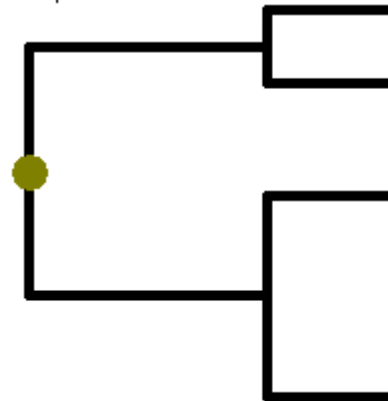


Phylogram and

Phylogenetic Tree of Life



Cladogram



Iterativní přístup

(Gotoh, 1996; Notredame & Higgins, 1996)

Vzniklý strom i alignment jsou následně **optimalizovány** do konvergence. Jinak jsou chyby vzniklé při prvním alignmentu (tvorba stromu) zachovány i ve výsledku.

Nezaručuje nalezení nejlepšího výsledku, ale – na rozdíl od deterministických alternativ – je dostatečně **robustní** a dobře použitelný i pro velký počet sekvencí.

Kombinace local a global alignment

- S výhodou lze kombinovat lokální a globální alignment.
- Lokální alignment může být reprezentován sadou kotvících bodů v místě dobré shody
- Následný globální alignment pak tyto odpovídající úseky sekvencí zahrnuje (využito např. v ClustalW2)

Výstup

Výstupem je sada sekvencí (případně s vloženými mezerami)

Různé formáty, nejčastěji používán **.aln soubor**, ale též **.fasta**, aj.

Mnoho programů sloužících pro zobrazení a/nebo editaci

- **Bioedit**
- **JalView**
- CINEMA 2.1...
- JavaShade
- ...

Výstup - .aln soubor

CLUSTAL 2.0.10 multiple sequence alignment

```
PAIIL      -----
RSIIL      -----
CVIIL      -----
BCLB      ---LVEKLPQYDVFVDIATIPYSFDVGSWQNKVKTDAAAGEVVACTVTWAGAPGVLPGAAA
BCLC      AIATNQGVVADGCFYSSKVPESTGRMPFTLVATIDVGSVTFVKQWKSVRGSAMHIDS
BCLA      -----
BCLD      LRETALALRAEVSVLFIRFALKDAGIVAPIELEVRDAATAVPDADDLLHPSCRPLKDHYW
```

```
PAIIL      -----ATQGVFT
RSIIL      -----AQQGVFT
CVIIL      -----AQQGVFT
BCLB      KFGVGAVVN-----YFSKATPQPVPAPVP-----TGGGERDGIFT
BCLC      YASLSAIWG-----TAAPSSQSGNQGAETGGTGAGNIGGGGERDGTFN
BCLA      -----ADSQT-----SSNRAGEFS
BCLD      RSDVLAAGATTCTADFAVCDRDGTVSGYFRWETSIEIAGSQPDTKQPGFKPSSDRNGNFS
```

* * .

```
PAIIL      LPANTRFGVTAFANSSGTQTVNVLVNNETA--ATFSGQSTNNAVIGTQVLNSGSSGKVQV
RSIIL      LPANTSFGVTAFANAANTQTIQVLVDNVVK--ATFTGSGTSDKLLGSQVLNSGS-GAIKI
CVIIL      LPARINFGVTVLVNSAATQHVEIFVDNEPR--AAFSGVGTGDNNLGTKVINSGS-GNVRV
BCLB      LPPNIAFGVTALVNSSAPQTEIEVFVDDNPKPAATFQAGTQDANLNTQIVNSGK-GKVRV
BCLC      LPPHIKFGVTALTHAANDQTIIDYIDDDPKPAATFKGAGAQQDQNLGTKVLDSEGN-GRVRV
BCLA      IPPNTDFRAIFFANAAEQQHILKFIGDSQEPAAAYHKL'TTRDGPRE--ATLNSGN-GKIRF
BCLD      LPPNTAFKAIIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRL--FTLNSKG-GKIRI
```

```
:*.. * . .::: * ::: ::: * . . .::* * ::.
```

BioEdit Sequence Alignment Editor

File Edit Sequence Alignment View Accessory Application RNA World Wide Web Options Window Help

D:\SkolaWyukaMSA - data\BCL lectins seq.aln

8 total sequences

Mode: Select / Slide Selection: 0 Position: Sequence Mask: None Numbering Mask: None Start ruler at: 1

Scroll speed slow fast

10 20 30 40 50 60 70 80 90 100 110 120

```

PAILL
RSIIL
CVIIL
BCLB
BCLC
BCLA
BCLD
Clustal Cons

```

-----SQQPPTHDDLYALLQ--LAGNDATAVQANGDQAVLDRMRQFMTAQ-----LVKLELPQYDVPVDIATIPYSFDVGSWQNKVKRIDAAGEVVACTVTWAGAPGVLPGAAAKFGVGA
-----PILBASIVSAPVVTSEYVDIPGLYLDVARAGIRDGKLVILNVPTPYATGNNPFGIYPAIATNGGVADGCTYSKVPBSTGRMPFTLVATIDVGSQVTFVRGQKRSVRGSAMHIDSYASLSAI
LVIVDAVTLLSAYPEASRDEPAFTVIDGRHLYVVSFGDAQLGHND SRLFTGLSPGDQLHRE TALALRAEVSVLFIRFALKDAGIVAPIELEVRDAATAVDPDADDLHPSCRPLKDHVWRSVLA

Snímek 34 z 64 Výchozí návrh

Jalview 2.3

File Tools Help Window

D:\SkolaWyukaMSA - data\BCL lectins seq.aln

File Edit Select View Format Colour Calculate Web Service

190 200 210 220 230 240

```

PAIL/1-114 TLPANTRFGVTAFAANS SGTQTIVNLVNNETA--ATFSGQSTNNAVIGTQVLNSGSSGKVRVQVSVNG
RSIL/1-113 TLPANTISFGVTAFAANAANTQTIQVLVDNVVK--ATFTGSGTSDKLLGSQVLNSGS-GAIKIQVSVNG
CVIL/1-113 TLPARINFGVTVLVNSAATQHVEIFVDNEPR--AAFSGVGTGDNNLGTVKINSGS-GNVRVQITANG
BCLB/1-243 TLPNINIAFGVTALVNS SAPQTIIEVFVDDNPKPAATFQGGAGTQDANLNTQIVNSGK-GKVRVVVTANG
BCLC/1-271 NLPPHIKFGVTALTHAANDQTIDIIYIDDDPKPAATFKGAGAQQNLGTVKVLDSGN-GRVRIVMANG
BCLA/1-128 SIPPNTDFRAIFFANAAEQHIKLFIGDSQEPAAAYHKLTTTRDGPRE--ATLNSGN-GKIRFEVSVNG
BCLD/1-288 SLPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVFGNSEDGVRLL--FTLNSKG-GKIRIEASANG

```

Conservation

8+ * 7 6 6 3 * 4 8 6 6 7 6 6 8 8 4 4 * 4 9 5 9 7 9 5 9 3 5 3 - - * 5 7 3 7 4 7 3 4 5 2 4 5 - - 3 6 9 9 * 5 4 - * 4 9 7 7 4 7 6 8 * *

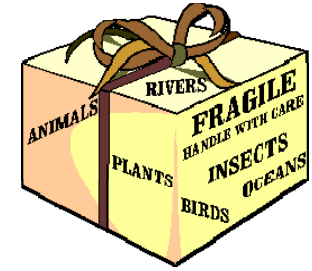
Quality

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

Globální alignment - Needle

A	1	MPTEFLYTSKIAAISWAATGGRQQRVYFQDLNGKIREAQRGGDNPWTGGS	50
B	1	-----	0
A	51	SQNVIGEAKLFSPLAAVTWKSAQGIQIRVYCVNKDNILSEFVYDGSKWIT	100
B	1	-----	0
A	101	GQLGSVGVKVGSNKLAALQWGGSESAPPNIRVYYQKSNGSGSSIHEYVW	150
B	1	-----	0
A	151	SGKWTAGASFGSTVPGTGIGATAIGPGRLRIYYQATDNKIREHCWDSNSW	200
	 : : . : : : .::	
B	1	--MQTAAISWGTT-PS-----IRV-YTANGNKITERCYDGSNW	34
A	201	YVGGFSASASAGVSIAAISW--GSTPNIRVYWQKGREELYEAAYGGSWNT	248
		. . : . .::::	
B	35	YTGA FN---QAGDNVSATCWLSGSAVHIRVY-----	62
A	249	PGQIKDASRPTPSLPDTFIAANSSGNIDISVFFQASGVSLQWQWISGKG	298
	:	
B	63	-----ATSGGSTTEWCW-DGDG	78
A	299	WSIGAVVPTGTPAGW	313
		: . . .	
B	79	WTRGAY--TGL-----	87

Lokální alignment - Water

A 155 TAGASFGSTVPGTGIGATAIGPGRLRIYYQATDNKIREHCWDSNSWYVGG 204
||..|:|:| |. :|: |.|..|||.|. |:|.::|||.|. |.
B 3 TAAISWGTT-----PS-IRV-YTANGNKITERCYDGSNWYTGA 38

A 205 FSASASAGVSIAAISW--GSTPNIRVYWQKREELYEAAYGGSWNTPGQI 252
|: .||.:::|..| ||..:||||
B 39 FN---QAGDNVSATCWLSGSAVHIRVY----- 62

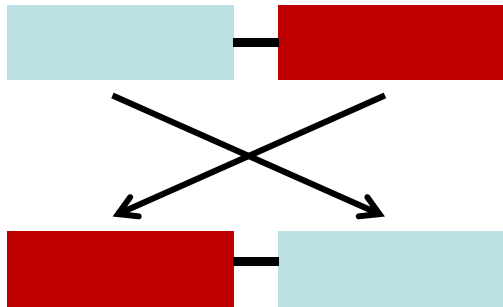
A 253 KDASRPTPSLPDTFIAANSSGNIDISVFFQASGVSLQQWQWISGKGWSIG 302
..|||.|.:.:|.|. |.|.||:|. |
B 63 -----ATSGGSTTEWCW-DGDGWTRG 82

A 303 A 303
|
B 83 A 83

Nelze však spoléhat na zdánlivě dobrá řešení

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQG VVADGCFTYSSKV
 PESTGRMPFTLVATIDVGSVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAA PSSQGSNGNQAETGGTGAGNIG
 GGERDGT FNLPPIHKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGA QDQNLGTKVLDSGNGRVRVIVMANGR
 PSRLGSRQVDIFKKS YFGIIGSEDGADDDYNDGIVFLNWPLG

ERDGT FNLPPIHKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGA QDQNLGTKVLDSGNGRVRVIVMANGRPSR
 LGSRQVDIFKKS YFGIIGSEDGADDDYNDGIVFLNWPLGPLLSASIVSAPVVT SQTYVDIPGLYLDVAKAGIRDGKLQ
 VILNVPTPYATGNNFPGIYFAIATNQG VVADGCFTYSSKVPESTGRMPFTLVATIDVGSVTFVKGQWKSVRGSAM
 HIDSYASLSAIWGTAA PSSQGSNGNQAETGGTGAGNIGGGKLA AALEIKRASQPELAPEDPEDVEHHHHH



```

#
#=====
EMBOSS_001      1 ----- 0
EMBOSS_001      1 ERDGT FNLPPIHKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGA QDQ 50
EMBOSS_001      1 ----- 0
EMBOSS_001     51 NLGTKVLDSGNGRVRVIVMANGRPSRLGSRQVDIFKKS YFGIIGSEDGAD 100
EMBOSS_001      1 ----- PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRD 35
EMBOSS_001     101 DDYNDGIVFLNWPLGPLLSASIVSAPVVT SQTYVDIPGLYLDVAKAGIRD 150
EMBOSS_001     36 GKLQVILNVPTPYATGNNFPGIYFAIATNQG VVADGCFTYSSKVPESTGR 85
EMBOSS_001     151 GKLQVILNVPTPYATGNNFPGIYFAIATNQG VVADGCFTYSSKVPESTGR 200
EMBOSS_001     86 MPFTLVATIDVGSVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAA PSSQ 135
EMBOSS_001     201 MPFTLVATIDVGSVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAA PSSQ 250
EMBOSS_001     136 GSGNQAETGGTGAGNIGGGGERDGT FNLPPIHKFGVTALTHAANDQTID 185
EMBOSS_001     251 GSGNQAETGGTGAGNIGGGG----- 271
EMBOSS_001     186 IYIDDDPKPAATFKGAGA QDQNLGTKVLDSGNGRVRVIVMANGRPSRLGS 235
EMBOSS_001     272 -----KLA A A-----LEIK-----RAS----- 283
EMBOSS_001     236 RQVDIFKKS YFGIIGSEDGADDDYNDGIVFLNWPLG 271
EMBOSS_001     284 -QPE-----LAPEDPEDVEHHH-----HHH 302
    
```

Lokálně podobné sekvence



Needle

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

9 -----RAGEFSIPPNTDFRAIFFANAAEQQHILFIGDSQEPAAYHK----- 50
      |.|.::||:..|.....:|..|.|.::|.|.:.|||.
151 NIGGGGERDGTFLNPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKG 200

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

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



Water

```
9 RAGEFSIPPNTDFRAIFFANAAEQQHILFIGDSQEPAAYHK----- 50
  |.|.::||:..|.....:|..|.|.::|.|.:.|||.
158 RDGTFLNPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQQN 207

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

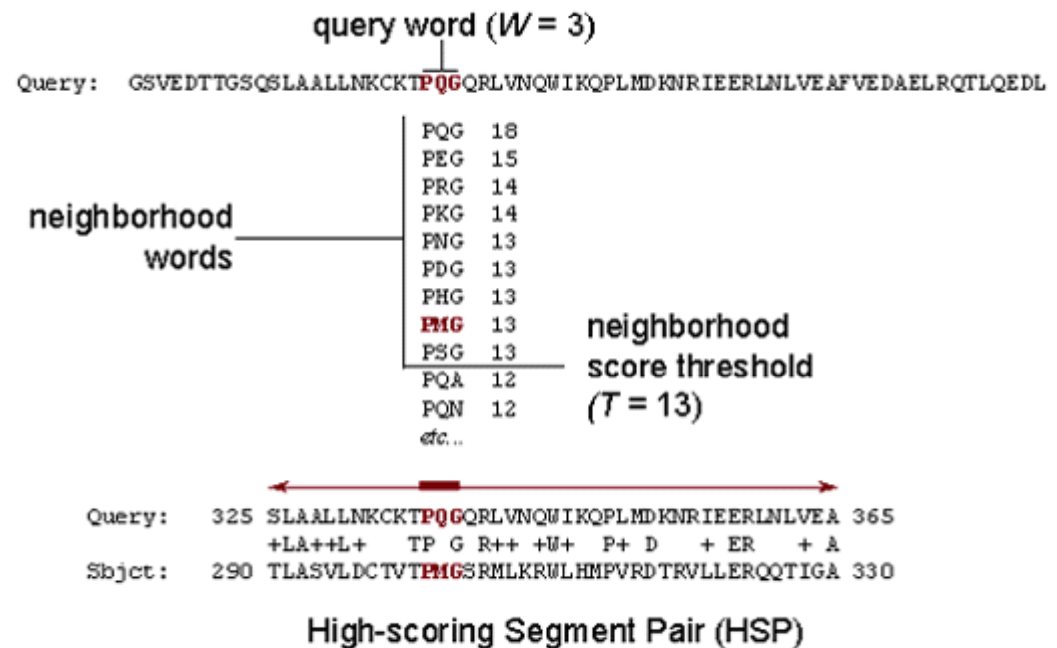
101 VNFVGVVSEDGHDSYNDGIVVLQWPIG 128
     .||||:|.|||||.|.|||||.|.||:|
245 -YFGIIGSEDGADDDYNDGIVFLNWPLG 271
```


BLAST algoritmus

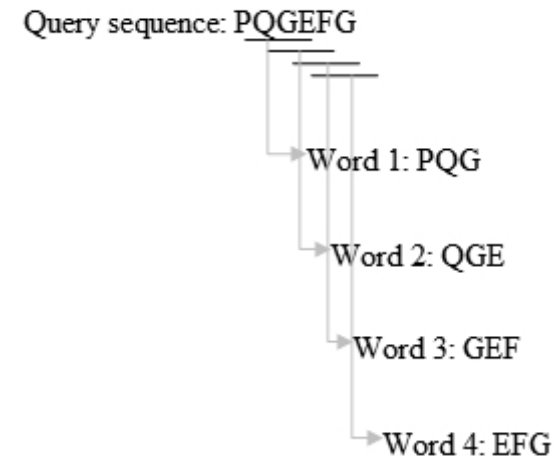
BLAST (Basic Local Alignment Search Tool)

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

The BLAST Search Algorithm



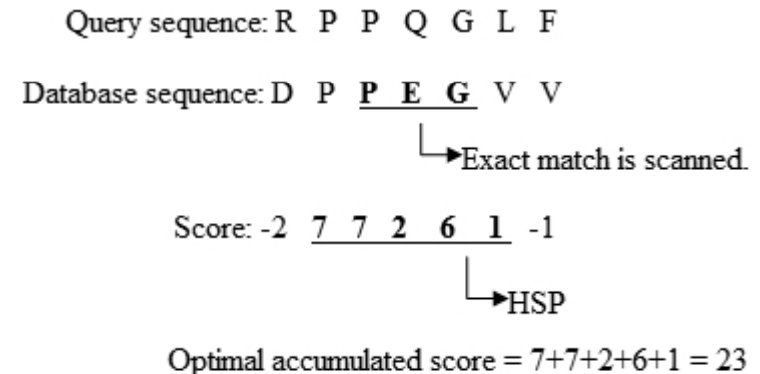
- **Tvorba k-písmenných slov ze vstupní sekvence**
pro proteiny typicky 3-písmenných (v případě DNA 11-písmenných)



- **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.

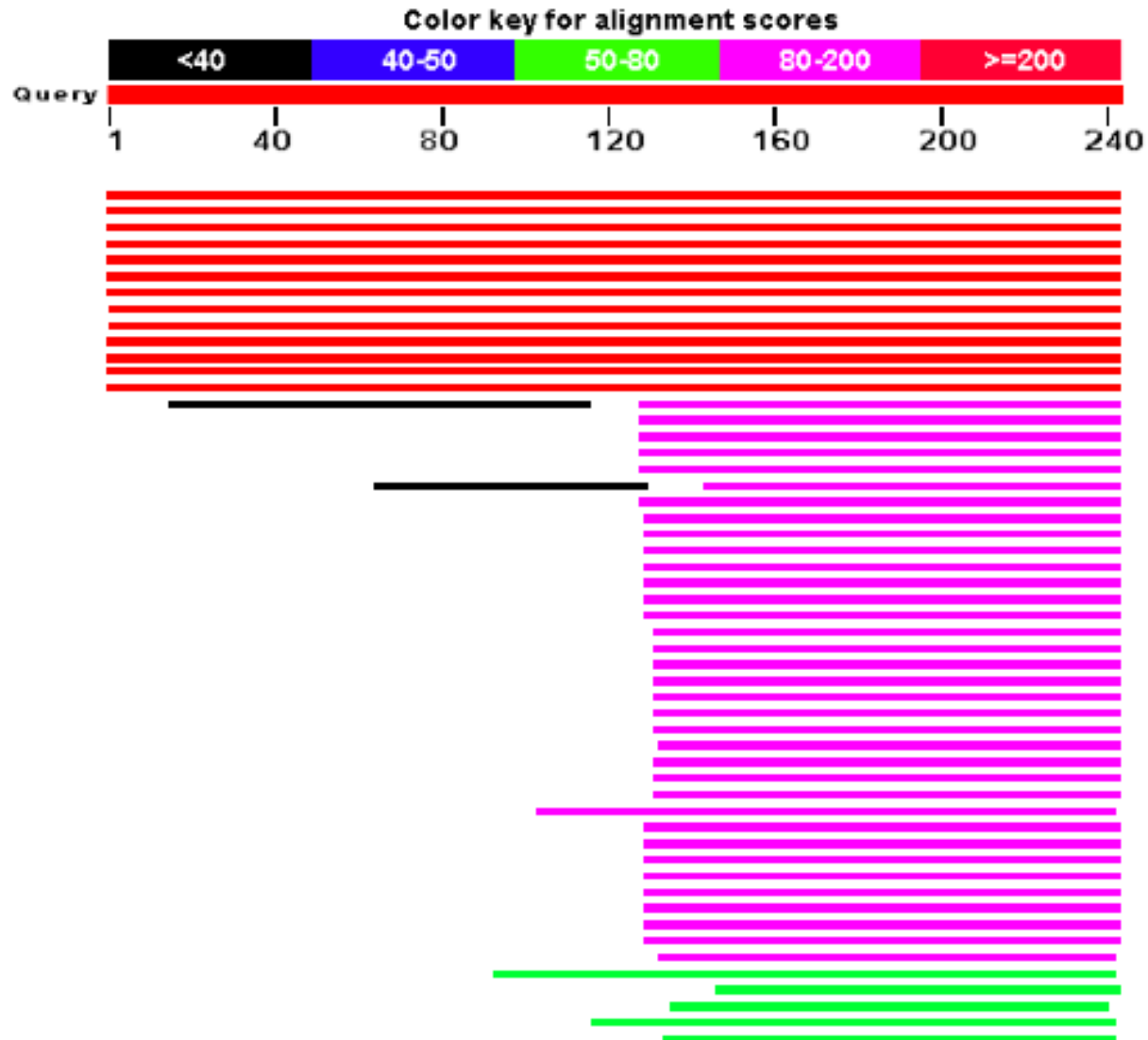


Novější verze BLASTu (BLAST2) má mj. níže nastavenou hladinu pro hledání podobných slov, což rozšiřuje možnost nalezení vzdálenějších homologů.

Výstup z BLASTu

Distribution of 73 Blast Hits on the Query Sequence

YP_002232817 lectin [Burkholderia cenocepacia J2315] S=488 E=3.9e-173



FASTA algoritmus

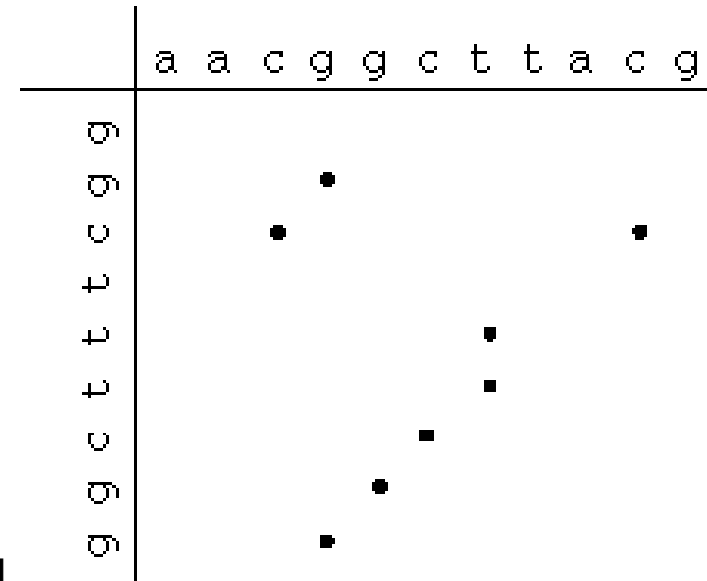
Na rozdíl od algoritmu BLAST jsou zde tolerovány mezery.

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).

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 kritérií.



Příklad porovnání sekvencí
GGCTTTCGG a
AACGGCTTACG

MSA „programy“

- Za posledních 15 let vzniklo přes 50 MSA programových balíčků (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 (Kato et al., 2002)
- MUSCLE (Edgar, 2004)
- Kalign (Lassmann, 2005)
- ...

Clustal <http://www.ebi.ac.uk/clustalw/>

- 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)
- 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)

ClustalW2 – 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 W/Clustal X



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

- * - identické residuum ve všech sekvencích
- : - silně konzervovaný sloupec
- . - slabě konzervovaný sloupec

```
IPPNTDFRAIFFANAAEQQHIFIGDSQEPAAAYHKLTTTRDGPREE--ATLNSGNGKIRFE
LPPNTAFKAIIFYANAADRQDLKLFIDDAPEPAATFVGNSDGVRL--FTLNSKGGKIRIE
LPPNIAFGVTALVNSSAPQTIIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGKGVV
LPPHIKFGVTALTHAANDQTIIDYIDDDPKPAATFKGAGAQQNLGKVLDSGNGRVRVI
***: * . .::: * :.:::. * :*** . :. . ::* *:::*
```




MUSCLE

(**M**ultiple **S**equences **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

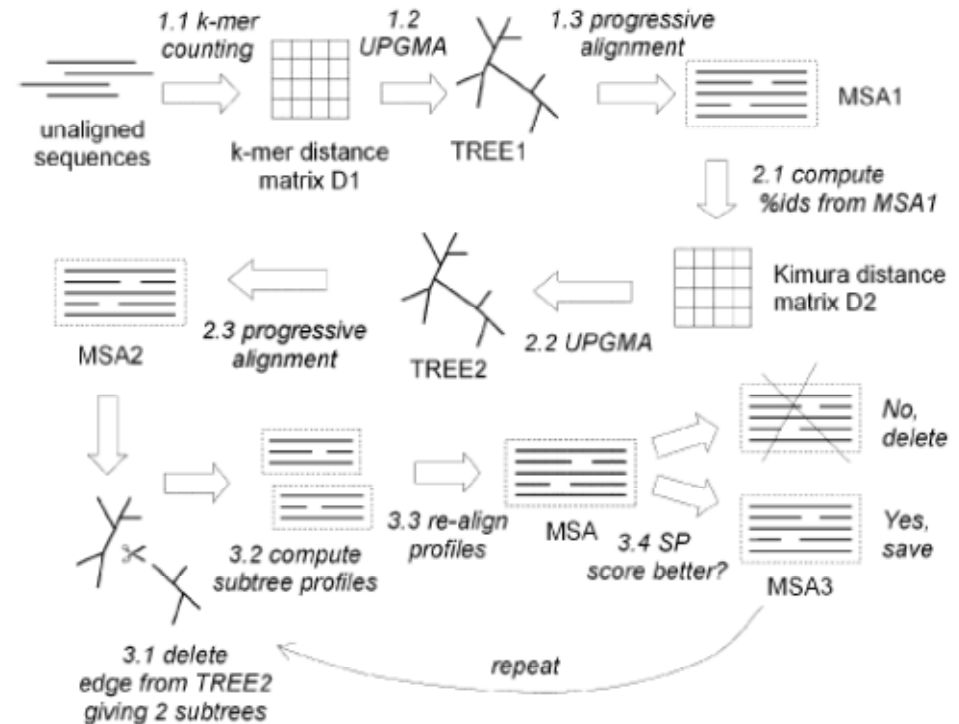


Figure 2. This diagram summarizes the flow of the MUSCLE algorithm. There are three main stages: Stage 1 (draft progressive), Stage 2 (improved progressive) and Stage 3 (refinement). A multiple alignment is available at the completion of each stage, at which point the algorithm may terminate.

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)

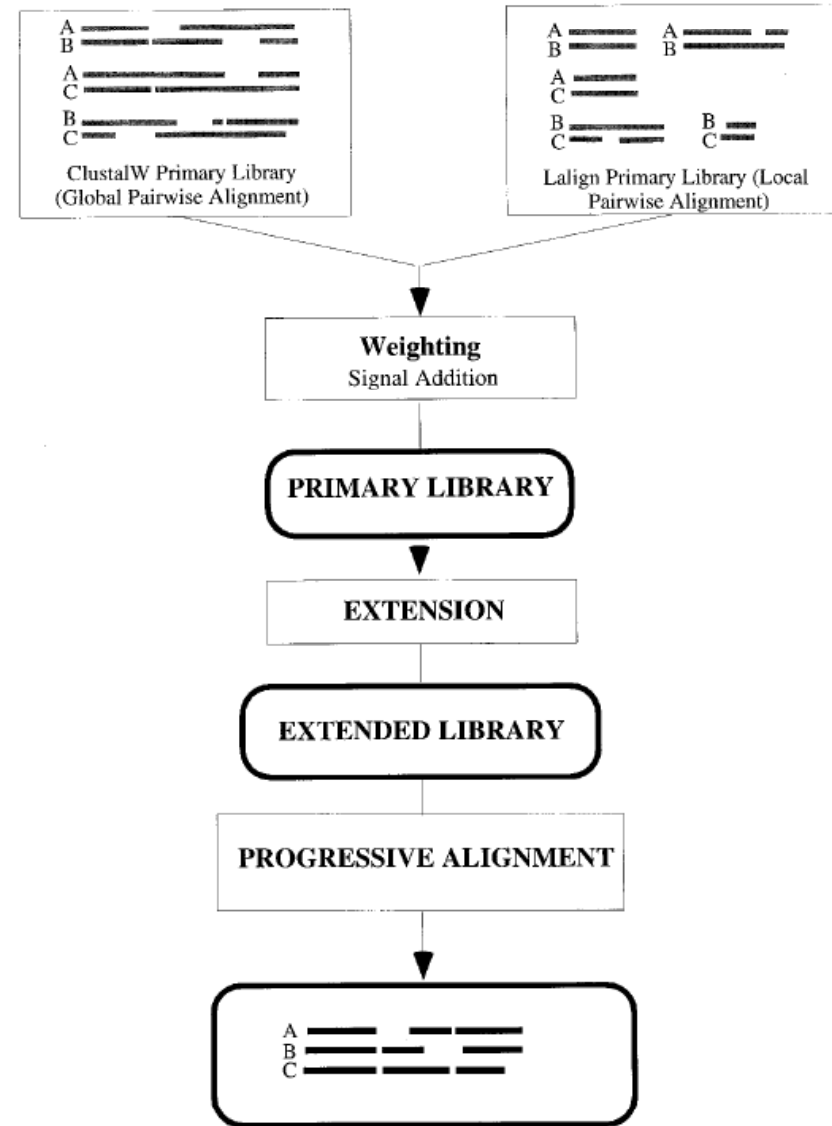


- Pomalejší ale výrazně přesnější než ClustalW
- Je schopen kombinovat data z více předchozích alignmentů, které mohly být vytvořeny různými postupy (lokální, globální, strukturní podobnost,...)

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 residuů k celkovému počtu residuů.
- 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.



Zlepšení přesnosti – strukturní informace

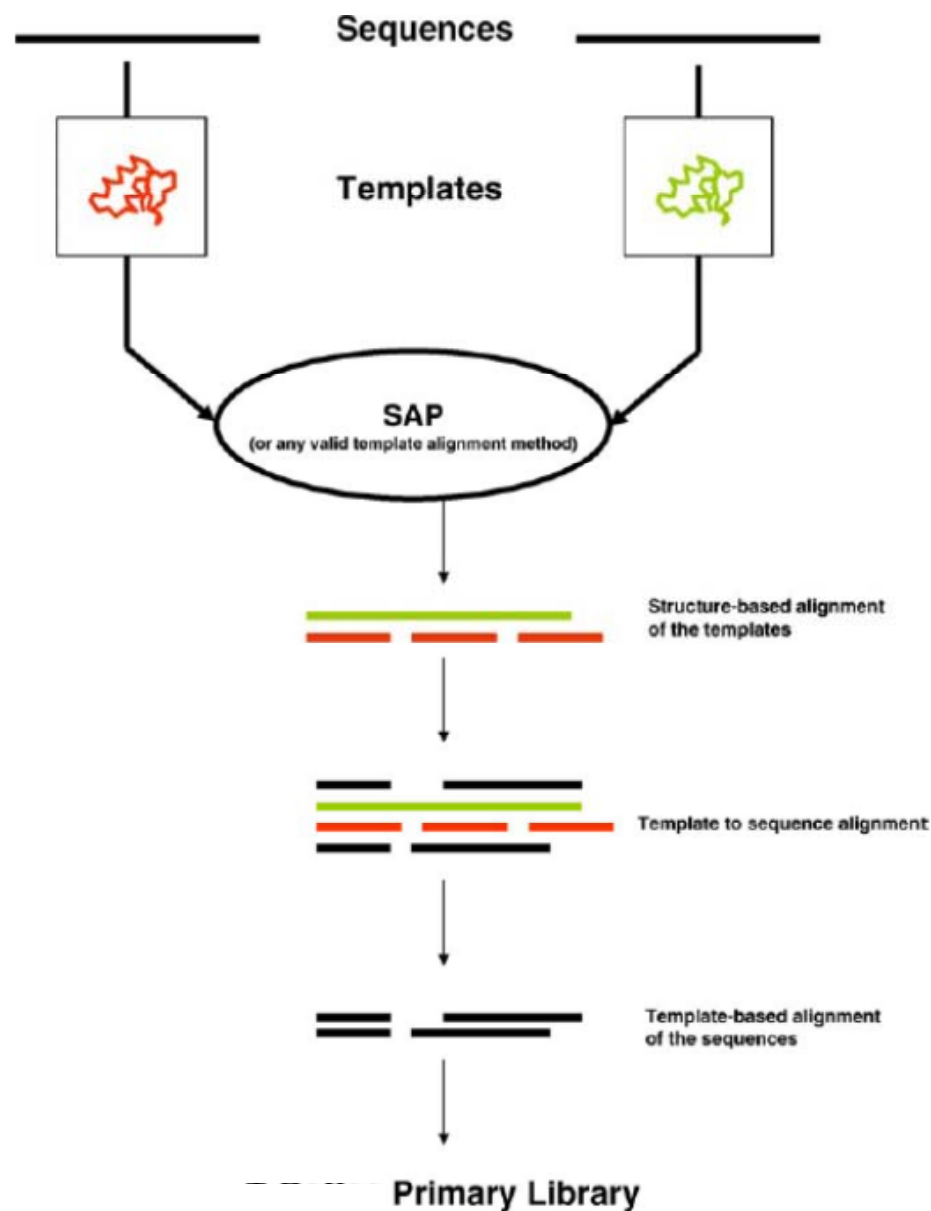
- Sekvence s vyšší homologií (>40%) – vysoká přesnost alignmentu
- Bez homologie – nepoužitelné
- Tzv. twilight zone – málo podobné sekvence (nižší než 20% homologie) = špatná (méně než 30%) přesnost alignmentu

Řešení: nejčastěji využití znalosti **strukturní podobnosti** (2D nebo 3D), která se během 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



Espresso

- Je založeno na T-Coffee

Espresso: MSA server, který srovnává sekvence za užití strukturní informace. Po zadání sekvencí vyhledá v databázi struktur (PDB) pomocí BLASTu homology a použije je jako šablony 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 !!!**

