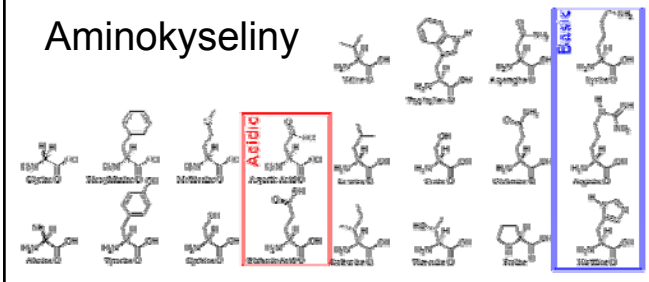
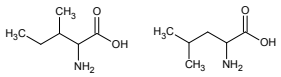
Aminokyseliny



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

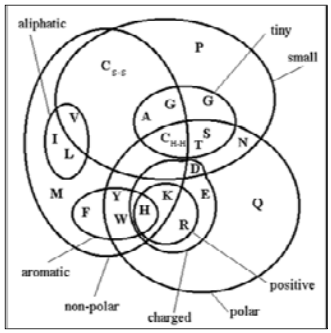
Třídění aminokyselin

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

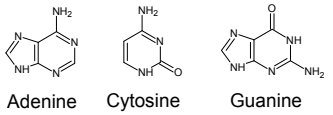


Isoleucine

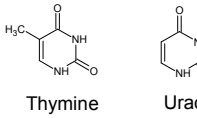
Leucine



Nukleové báze



Adenine Cytosine Guanine

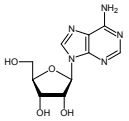


Thymine Uracil

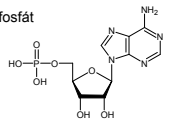
adenin	cytosin	guanin	thymin	uracil
A	C	G	T	U

Nukleová báze
Adenin

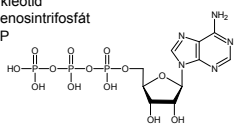
Nukleosid
Adenosin



Nukleotid
Adenosinmonofosfát
AMP



Nukleotid
Adenosintrifosfát
ATP



Polysacharidy

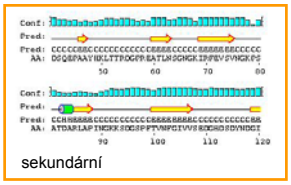
Komplikované sekvence – alignment se neprovádí

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

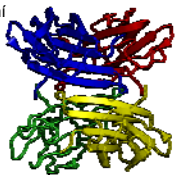
primární
(sekvence)



sekundární

terciární

kvartérní



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

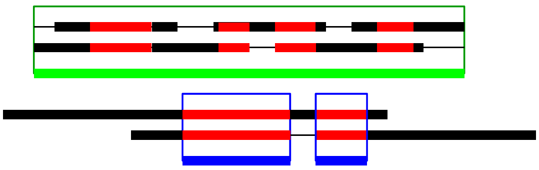
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM

Multiple sequence alignment – více sekvencí

WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
WLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAMWLAKALKYLMETAQASSISTELARHHPRAVDAKRKSEMKRKTAM
...

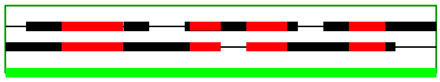
Pair-wise alignment

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



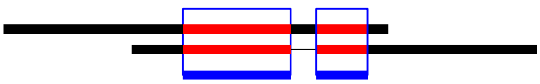
Global alignment

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

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 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 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).

PAM – Point Accepted Mutation

Constructed by Margaret Dayhoff in 1978.

Zahrnuje pravděpodobnost záměny jedné aminokyseliny v druhou během evoluce

Předpokládá, že každá další mutace nezávisí na předchozí.

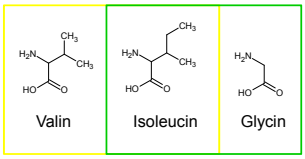
Odvozena z globálního alignmentu 71 rodin proteinů (Podobnost sekvencí v rodině > 85%)

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

Vypočtena pravděpodobnost s jakou jedna AA se změní na jakoukoliv jinou

Matice PAM 250

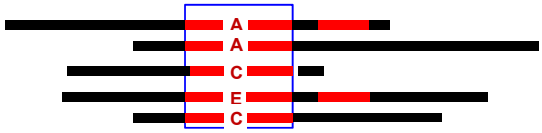
	C	G	P	S	A	T	D	E	N	Q	H	K	R	V	M	I	L	F	Y	W
C	12																			
G	-3	5																		
P	-3	-1	6																	
S	0	1	1	1																
A	-2	1	1	1	2															
T	-2	0	0	1	1	3														
D	-5	1	-1	0	0	0	4													
E	-5	0	-1	0	0	0	3	4												
N	-4	0	-1	1	0	0	2	1	2											
Q	-5	-1	0	-1	0	-1	2	2	1	4										
H	-3	-2	0	-1	-1	-1	1	1	2	3	6									
K	-5	-2	-1	0	-1	0	0	0	1	1	0	5								
R	-4	-3	0	0	-2	-1	-1	-1	0	1	2	3	6							
V	-2	-1	-1	-1	0	0	-2	-2	-2	-2	-2	-2	-2	4						
M	-5	-3	-2	-2	-1	-1	-3	-2	0	-1	-2	0	0	2	6					
I	-2	-3	-2	-1	-1	0	-2	-2	-2	-2	-2	-2	-2	4	2	5				
L	-6	-4	-3	-3	-2	-2	-4	-3	-3	-2	-2	-3	-3	2	4	2	6			
F	-4	-5	-5	-3	-4	-3	-6	-5	-4	-5	-2	-5	-4	-1	0	1	2	9		
Y	0	-5	-5	-3	-3	-3	-4	-4	-2	-4	0	-4	-5	-2	-2	-1	-1	7	10	
W	-8	-7	-6	-2	-6	-5	-7	-7	-4	-5	-3	-3	2	-6	-4	-5	-2	0	0	17



BLOSUM (Blocks Amino Acid Substitution)

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

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



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

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

Číslování BLOSUM jde v obráceném pořadí oproti PAM

- čím menší číslo, tím odlišnější sekvence byly použity

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

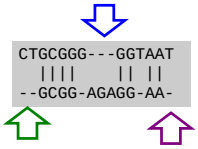
Similarity column gives range of similarities that the matrix is able to best detect.

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:

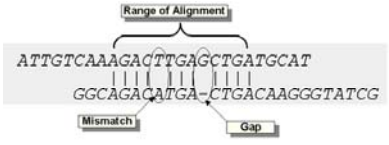
```
ATCTTCAGTGTTTCCCCTGTTTTGCCC-ATTTAGTTCGCTC
|||||||||||||||||||||||||||| ||||||||||||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGATTTAGTTCGCTC
```

Dlouhá mezera:

```
ATCTTCAGTGTTTCCCCTGTTTTGCCC----------ATTTAGTTCGCTC
|||||||||||||||||||||||||||          ||||||||||||||
ATCTTCAGTGTTTCCCCTGTTTTGCCCGCCCCCCCCCCCCCCCCCCATTTAGTTCGCTC
```

Skóre

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



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

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

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 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

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.

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

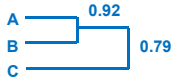
Progresivní multiple alignment

- Používá ho většina programů
- Vznik – 1987

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

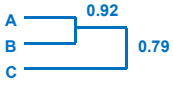
1) sestavení příbuzenského stromu (guide tree) na základě distanční matice (distance matrix) z nepřiložených sekvencí

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



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

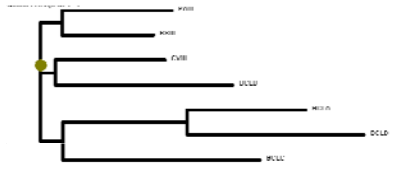
```
(
(
PAIIL:0.16435,
RSIIL:0.13654)
:0.03384,
(
CVIIL:0.16563,
BCLB:0.26800)
:0.02264,
(
(
BCLA:0.17899,
BCLD:0.26633)
:0.18717,
BCLC:0.29707)
:0.03484);
```

Phylogram a cladogram

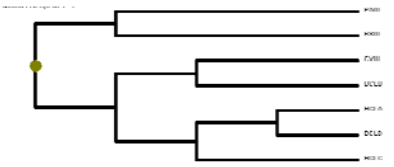
- **Phylogram** (phylogeny tree) – je rozvětvený diagram (strom), který naznačuje fylogenezi (postupný vývoj). Délka jednotlivých větví je úměrná **velikosti změny** v průběhu evoluce.
- **Cladogram** – rovněž strom, v němž však všechny větve mají **stejnou délku**. Ukazuje tak sice „společné předky“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

Phylogram a cladogram

Phylogram



Cladogram



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)

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

emboss 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

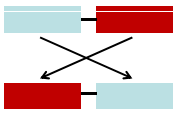
Needle – globální pairwise alignment, Needleman-Wunsch algoritmus

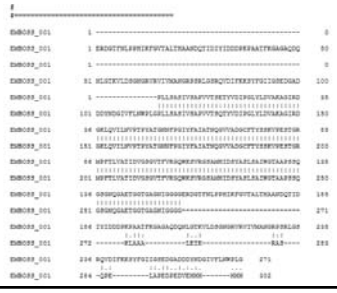
Water – lokální pairwise alignment, Smith-Waterman algoritmus

Nelze však spoléhat na zdánlivě dobrá řešení

PLLSASIVSAPVVTSETYVDIPGLYLDVAKAGIRDGKLQVILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKV
PESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAMHIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIG
GGGERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGR
PSRLGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLG

ERDGTFNLPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVIVMANGRPSR
LGSRQVDIFKKSYFGIIGSEDGADDDYNDGIVFLNWPLGPLLSASIVSAPVVTSQTYVDIPGLYLDVAKAGIRDGKLQ
VILNVPTPYATGNNFPGIYFAIATNQGVVADGCFTYSSKVPESTGRMPFTLVATIDVGSGVTFVKGQWKSVRGSAM
HIDSYASLSAIWGTAAPSSQGSGNQGAETGGTGAGNIGGGGKLAAALEIKRASQPELAPEDPEDVEHHHHHH

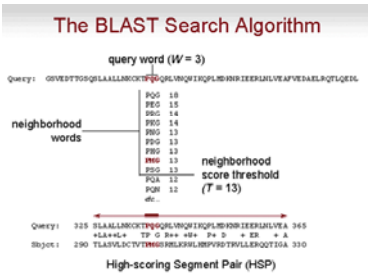




BLAST algoritmus

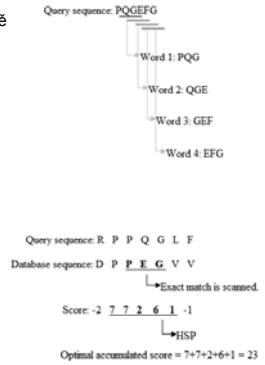
BLAST (Basic Local Alignment Search Tool)

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.

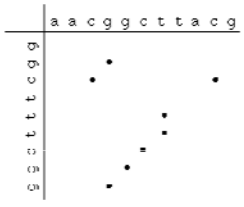


FASTA algoritmus

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 kriterií.



Příklad porovnání sekvencí GGCTTTCGG a AACGGCTTACG

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.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

```
IPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGFRE--ATLNSGNGKIRFE
LPPNTAFKAIFYANAADRQDLKLFIDDAPEPAATFVGNSEDGVRL--FTLNSKGGKIRIE
LPPNIAFGVTALVNSSAPQTIEVFVDDNPKPAATFQGAGTQDANLNTQIVNSGKGKVRVV
LPPHIKFGVTALTHAANDQTIDIYIDDDPKPAATFKGAGAQDQNLGTKVLDSGNGRVRVI
:**:  *.   .:::  * :.:::.*  :*** .    :. .     ::*  *::*.
```

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.

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** !!!