

BIOINFORMATIKA V PRAXI – CVIČENÍ 2

SEQUENCE ALIGNMENT

STUDIJNÍ MATERIÁLY

Studijní materiály předmětu C2130 Úvod do chemoinformatiky a bioinformatiky, přednáška Sequence alignment.

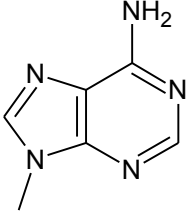
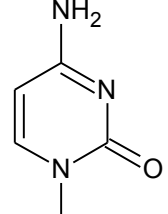
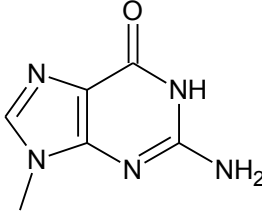
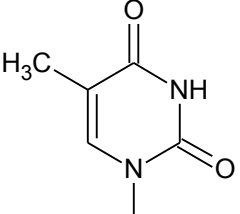
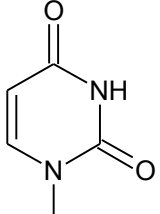
ÚVOD

Sekvence v biochemii

Při práci s biologickými makromolekulami (nukleové kyseliny, proteiny, sacharidy) hovoříme často o jejich sekvenci. Pod tímto pojmem rozumíme pořadí a identitu jednotlivých stavebních prvků, z nichž se makromolekula skládá (v případě sacharidů i typ vazby mezi sousedními jednotkami). Pro jednotlivé stavební prvky existují obvykle třípísmenné zkratky, v bioinformatice však častěji používáme jednopísmenné zkratky. V tomto předmětu se budeme zabývat jen problematikou nukleových kyselin a proteinů. V případě sacharidů je celá situace komplikovaná množstvím různých stavebních jednotek a možností větvení řetězce. S tím souvisí větší náročnost určení sekvence oligo-/poly-sacharidu a též menší počet dostupných sacharidových sekvencí v databázích.

Nukleové kyseliny

Nukleové kyseliny (DNA, RNA) jsou tvořeny kombinací čtyř prvků (nukleových bází) spojených prostřednictvím tzv. cukr-fosfátové kostry. V případě DNA se jedná o adenin, cytosin, guanin a thymin, v případě RNA pak adenin, cytosin, guanin a uracil. Tyto báze jsou uvedeny v přiložené tabulce.

Adenin	Cytosin	Guanin	Thymin	Uracil
Ade	Cyt	Gua	Thy	Ura
A	C	G	T	U
DNA, RNA	DNA, RNA	DNA, RNA	DNA	RNA
				

Pozn.: Používání zkratk a názvů pro různé části nukleových kyselin (nukleotidy, nukleosidy, báze) je poměrně komplikované. V případě zájmu lze detailní doporučení názvoslovné komise nalézt např. na stránkách: <http://www.chem.qmul.ac.uk/iupac/misc/naabb.html>

Sekvenci nukleové kyseliny zapisujeme v pořadí **od 5'-konce** (nukleotid s volnou OH skupinou na 5. uhlíku) směrem k 3'-konci (nukleotidu s volnou OH skupinou na uhlíku číslo 3). V tomto směru je také v živých organismech DNA resp. RNA syntetizována.

Proteiny

Proteiny jsou sestaveny z 20 standardních aminokyselin (22, počítáme-li selenocystein a pyrrolysin) – viz tabulka. Dle různých charakteristik můžeme proteinogenní aminokyseliny rozdělit do několika skupin – nejčastěji rozlišujeme aminokyseliny nabitě (kyselé a bazické),

nenabité polární, hydrofobní a malé (s krátkým postranním řetězcem). Při analýze proteinových sekvencí obvykle neuvažujeme možné posttranslační modifikace, jako je např. hydroxylace či glykosylace.

	NONPOLAR, HYDROPHOBIC	R GROUPS	POLAR, UNCHARGED	
Alanine Ala A MW = 89	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_3 \end{matrix}$		$\begin{matrix} \text{H} - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Glycine Gly G MW = 75
Valine Val V MW = 117	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}(\text{CH}_3)_2 \end{matrix}$		$\begin{matrix} \text{HO} - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Serine Ser S MW = 105
Leucine Leu L MW = 131	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}(\text{CH}_3)_2 \end{matrix}$		$\begin{matrix} \text{OH} \\ \\ \text{CH}_3 - \text{CH} - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Threonine Thr T MW = 119
Isoleucine Ile I MW = 131	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}(\text{CH}_3) - \text{CH}_2 - \text{CH}_3 \end{matrix}$		$\begin{matrix} \text{HS} - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Cysteine Cys C MW = 121
Phenylalanine Phe F MW = 131	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}_6\text{H}_5 \end{matrix}$		$\begin{matrix} \text{HO} - \text{C}_6\text{H}_4 - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Tyrosine Tyr Y MW = 181
Tryptophan Trp W MW = 204	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}_8\text{H}_6\text{N}_2 \end{matrix}$		$\begin{matrix} \text{NH}_2 \\ \\ \text{O} = \text{C} - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Asparagine Asn N MW = 132
Methionine Met M MW = 149	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{S} - \text{CH}_3 \end{matrix}$		$\begin{matrix} \text{NH}_2 \\ \\ \text{O} = \text{C} - \text{CH}_2 - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Glutamine Gln Q MW = 146
Proline Pro P MW = 115	$\begin{matrix} ^- \text{OOC} \\ \\ \text{CH} - \text{CH}_2 - \text{CH}_2 \\ \quad \\ \text{HN} - \text{CH}_2 \end{matrix}$		POLAR BASIC $\begin{matrix} \text{NH}_3^+ - \text{CH}_2 - (\text{CH}_2)_3 - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Lysine Lys K MW = 146
Aspartic acid Asp D MW = 133	POLAR ACIDIC $\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{C}(=\text{O})\text{O}^- \end{matrix}$		$\begin{matrix} \text{NH}_2 \\ \\ \text{N H}_2^+ = \text{C} - \text{NH} - (\text{CH}_2)_3 - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+ \end{matrix}$	Arginine Arg R MW = 174
Glutamine acid Glu E MW = 147	$\begin{matrix} ^- \text{OOC} \\ \\ \text{H}_3\text{N}^+ - \text{CH} - \text{CH}_2 - \text{CH}_2 - \text{C}(=\text{O})\text{O}^- \end{matrix}$		$\begin{matrix} \text{C} = \text{N}^+ \\ \\ \text{HN} \end{matrix} - \text{CH}_2 - \text{CH} - \text{COO}^- \\ \\ \text{N H}_3^+$	Histidine His H MW = 155

Není-li uvedeno jinak, zapisujeme sekvenci proteinů od N-konce (volná NH₂- skupina) k C-konci (volná COOH- skupina). Tento směr je opět totožný se směrem, v němž je protein v živém organismu syntetizován.

Sekvenční příložen (Sequence alignment)

Porovnání dvou (nebo více) sekvencí mezi sebou označujeme jako sekvenční příložen (častěji anglicky sequence alignment). Cílem je určení vzájemné podobnosti těchto sekvencí. Vizualním výstupem je zarovnání sekvencí tak, aby sobě odpovídající residua ležela nad sebou. Detaily viz. přednáška Sequence alignment předmětu C2130.

VYUŽITÍ SEKVENČNÍHO PŘILOŽENÍ PRO IDENTIFIKACI GENU V ONLINE DATABÁZÍCH

Pro vyhledávání v internetových databázích lze použít několik přístupů (viz. Bioinformatika v praxi – cvičení 1). Pokud máme jako vstupní údaj sekvenci genu/proteinu, využíváme hledání na základě podobnosti sekvencí. Typickou ukázkou je aplikace **BLAST** na serveru NCBI (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>).

ÚKOL 1

Pomocí aplikace BLAST identifikujte následující sekvence:

Sekvence 1:

```
MNGTEGPNFYVPFSNKTGVVRS PF EY P Q Y Y L A E P W Q F S M L A A Y M F L L I V L G F P I N F L T L Y V T V Q H K K L R T P L N Y I
L L N L A V A N H F M V F G G F T T T L Y T S L H G Y F V F G S T G C N L E G F F A T L G G E I A L W S L V L A I E R Y V V V C K P M S N F R F G E
N H A I M G V A F T W V M A L A C A A P P L V G W S R Y I P E G M Q C S C G I D Y Y T L K P E V N N E S F V I Y M F V V H F T I P M T I I F F C Y G Q
L V F T V K E A A A Q Q Q E S A T T Q K A E K E V T R M V I I M V I A F L I C W V P Y A S V A F Y I F T H Q G S D F G P I L M T L P A F F A K S S A I
Y N P V I Y I M M N K Q F R N C M L T T I C C G K N P F G E E E G S T T A S K T E T S Q V A P A
```

Název genu/proteinu:

Číslo záznamu v databázi NCBI:

Míra shody zadané a nalezené sekvence:

Sekvence 2:

```
a t g t c g a g c g t t c a g a c c g c t g c c a c t t c g t g g g g a a c c g t c c c g t c g a t c c g t g t g t a c a c g g c c a a t a a t g g c
a a g a t c a c c g a g c g t t g c t g g g a c g g g a a g g g g t g g t a c a c g g g t g c c t t c a a c g a g c c c g g c g a t a a c g t c t c c
g t g a c c a g c t g g t g g t c g g c a g c g c g a t c c a t a t c c g c g t c t a t g c a a g c a c c g g c a c c a c g a c c a c a g a g t g g
t g c t g g g a c g g c a a c g g c t g g a c c a a g g g c g c c t a c a c c g c c a c c a a c t g a
```

Název genu/proteinu:

Číslo záznamu v databázi NCBI:

Míra shody zadané a nalezené sekvence:

Sekvence 3:

```
M G L S D G E W Q M V L N I W G K V E G D L A G H G Q E V L I S L F K A H P E T L E K F D K F K N L K S E E M K S S E D L K K H G C T V L T A L G T
I L K K K G Q H A A E I Q P L A Q S H A T K H K I P V K Y L E F I S E V I I Q V L K K R Y S G D F G A D A Q G A M S K A L E L F R N D I A A K Y K E L
G F Q G
```

Název genu/proteinu:

Číslo záznamu v databázi NCBI:

Míra shody zadané a nalezené sekvence:

VYHLEDÁNÍ PODOBNÝCH SEKVENCÍ A URČENÍ PŘÍBUZNOSTI

Výhodou použití sequence alignmentu je schopnost nalezení nejen shodného záznamu, ale i záznamů podobných. Tak lze na základě podobných sekvencí identifikovat i dosud neznámou sekvenci a odhadnout její „příbuzenské“ vztahy.

ÚKOL 2

Identifikujte zadanou sekvenci a nalezněte 4 další nejpodobnější sekvence. Použijte aplikaci BLAST.

Sekvence:

MNTRSFHRI DVHKARELLQRPDTVLLDCRHPSDFRAGHIAGASPLGDYNADDHVLNIAKHRPVLIYCYHGNASQM
 RAQLFADFGFAEVYSLDGGYEAWRKVHTPANSQLEALQCWLMAQEFPAADIHARTRDGVTPLMRAAGEGDPARV
 AELLAAGADPHQRNNDGNQALWFACVSENLDTLDLLVAVGAHLNHQNDNGATCLMYAASAGKTAVVERLLAFGAD
 RSLLSLDDFTALDMAANLECLNLLRETPRRIKAVT

Číslo záznamu	Protein	Organismus	Score

Volitelný ÚKOL

Z nalezených sekvencí v úkolu 2 sestavte Multiple sequence alignment a vytvořte fylogenetický strom (phylogenetic tree).

VLIV POUŽITÉ MATICE NA VÝSLEDEK ALIGNMENTU

Jedním z parametrů, který může ovlivnit výsledek alignmentu je použitá matice. Většina programů detekuje automaticky nukleotidovou sekvenci a použije příslušnou matici, v případě proteinových sekvencí je však situace komplikovanější.

ÚKOL 3

Následující sekvence **identifikujte** a přiložte v programu **ClustalW** (<http://www.ebi.ac.uk/Tools/clustalw2/index.html>). Poté proveďte alignment s použitím matice PAM, BLOSUM, Gonnet a ID pro multiple alignment a výsledky porovnejte. Která matice je použita při základním nastavení?

Sekvence 1:

MPGIRLRYALLALVFAIYYYIVLSYRDQFSDIKKCFSSIRAKIEDYDSSKKGQPKLASNSYLEADMLYRDRTOVG
 IENATMVMLVRNRELEGALSSMRSLEDRFNRQYKYPWVFLNDEPFTEEFIEQTMLMASSQTFYELIPSSDWNMPD
 FIDNERVEQNIANSTVDVIYGFSKSYRNMCHFNSEYFYKQKRLNLDWYFRVEPDVEYMCDFQYDPFTLLRRTNKKI
 YGFVIAIHEYENTIPTLWPTVEKFMQTYPDLIHANNSLKFITNESSLNHWVTPIQSSSDYNLCHFWSNFEIGNL
 NFFRGEAYNKYFDFLDRAGGFYERWGDAPVHSIGLNLADKNSIHHFEDIGYYHPPYLACPSSKDVIAAKRCVC
 RKRGNDEVIDSAIDVNVFSCLSRWWRYGAGKRFLNEIDYTFNN

Sekvence 2:

MPESGVPAGRRRLAPLLLVTSAALLAVTMANPPDLVDFHVYMLGGVALDRPDTLYSFAYSQSPDQPLPFVYP
 PFAAILFRPLTALPFVAVGLWQLGILAAVYGVIRISQRLGGGSHRTAMLWTAGLIWLEPVRVALDLGQVGI
 TLAVLYAVCSRRWWSGLLVGLAAGVKLTPAITGLYFLGVRRTAAAFSAVVFVATIGLSYLVVGDQVRNYFTRV
 MGDTSINPIGIALNQSWRGAI SRFLGHDAGGSALVIAAIAGTAVLALFAWAALGAGSRPRDRLGSVLVQVLFGLL
 MSPISWVHHWAVVPLVLWLLSGPWRNEPGARVLGWGWALTFVGVPSQLSLEPSLWEISR PWYLAWAGLAYV
 AAVATLGIWVLTGRRNAAPTTPVRRRFARVVGTRGKQIACEEHRAGR

Sekvence 3:

MELQSLIDTVSLQKLLLLGALLRLILIAAFFHDQWFRVKYTDIDYMI VVDGARHMWNGGSPFDRITFRYTPLLA
 ALVMPISIWIANPMGKLI FASSDLGAAWYCYGVLKSF AKERSAKWMVSLFILFNPIVLSVSTRGNSDMLVTFMSLM
 VLSKFARRKCYQAAAVLGFVAVFKIYPIIYALPLTLGVWEQSVAASTNTWRRVVKTA VVVSICALMAAISFAVPT
 VLCYMKYGOQYLNEAFIYHVYREDHRHNFSPYWLLMYLNMARRHLGQGVDFSPRLVAFAPQAVVLSFVSYKLRN

TAHACCVQTVL FVAFNKVCTVQYFVWFIPFLAFLFCEPKEVEDESGGSGAFKFFSWVKALGVVLMWAATIPLWV
TTAVPLEFHGYSDFACLWIVSCLFFLAMVVLASMLARIA YRVQCTKCSAKSIKVA

Sekvence 4:

MDASTPNPPTLGTDAVASTIVFFHPDLGIGGAERLVVDAAVGLQTRGHKVVIFTNHCDPTHCFDECRDGTLDVRV
RGNSIVPPSIFSRLTILCAILRHIHLLLTIIHLTGELAALS PRAFIVDQLSAGLPLMRFLAPDVPVLFYCHFDDL
LAQGRQSLVKRLYRVFPDRLEEWSMGFAHAVAVNSKFTRGIVGNTWPALQNKVPINVVYPCVDTHTTTETAPDEA
KLAAGKKLILSINRFERKKDIGLAIRAFQAIP EEQRRGARLVLAGGYDARVSENVLYHAELQALATSLSLAHHTL
TPAELGSAAAPPDAQHFGIVPLEAMLARVPVLAANTGGPVETVADRETGWLRDPADAPAWTDVMARCLALPDDQL
AAMGDAGRVRVRELFGRDKMAQTLDES LVQIAGLAEERRVSGAAGFGVLA AFIAACAVLAAWFAF

VLIV MEZER NA VÝSLEDEK ALIGNMENTU

Možnost vložení mezer významně zvyšuje šance na úspěšný alignment sekvencí. Při změně nastavení parametrů se mění skóre určující podobnost sekvencí a ty tak mají vliv i na určení vzájemné příbuznosti sekvencí. Při špatném nastavení pak umožňují provést alignment i u naprosto nepodobných sekvencí.

ÚKOL 4

Proveďte multiple alignment následujících sekvencí pomocí programu ClustalW. V prvním případě nastavte parametr GAP OPEN na 1, ve druhém případě na 100 a výsledky porovnejte.

Sekvence 1:

trypsin [*Homo sapiens*] – AAZ40216.1

RIQVRLGEHNIEVLEGEQFINAAKIRHPKYDRKTLNNDIMLIKLSRAVINARVSTISLPTAPPATGTKCLIS
GWGNTASSG

Sekvence 2:

FlgA [*Agrobacterium tumefaciens*] – AAB71791.1

MRFGRNSSCRTALVRMCLASAFSLGALAPALAQAPMALVPVRTIYPGEAISPEQVKSVEVTNPNISAGYASDIS
EVEGMISKQTLPLGRTIPIAALREPSLVVRGTSVKLVFHIIGNMTLMASGTPMSDGS LGEVVRNRNIDSGVMVSGT
VMKDGTIQVMAK

Sekvence 3:

collagen [*Caenorhabditis elegans*] – CAA35955.1

MSEDLKQIAQETESLRKVAFFFGIAVSTIATLTALIAVPMLYNYMQHVQSSLSQSEVEFCQHRSNGLWDEYKRFQGV
SGVEGRIKRDAYHRSLGVSGASRKARRQSYGNDAAVGGFGSSGGSCCSCGSGAAGPAGSPGQDGPNDGAPGA
PGNPGQDASEDQTAGPDSFCFDCPAGPPGSPGAPGQKGPAGPAGPQSGGAALPGPPGAPPPGAPQPGSNGN
AGAPGAPGQVVDVPGTPGPAGPPGSPGAPGAPGQPGQAGSSQPGGPGPQGDAGAPGAPGAPGQAGAPGQDGESGS
EGACDHCPPPRTAPGY

Gap open = 1, 1

Počet zcela identických residuí:

Gap open = 100, 100

Počet zcela identických residuí:

VOLITELNÝ ÚKOL

Pomocí programu ClustalW proveďte Multiple sequence alignment následujících sekvencí při základním (default) nastavení parametrů. V druhém okně prohlížeče proveďte totéž příložením, ale nastavte parametr Gap open na hodnotu „5“. Výsledky porovnejte.

Sekvence 1:

PTEFLYTSKIAAISWAATGGRQQRVYFQDLNGKIREAQRGGDNPWTGGSSQNVI GEAKLFSPLAAVTWKS AQGIQ
IRVYCVNKDNILSEFVYDGSKWITGQLG SVGVKVG SNSKLAALQWGGSESAPPNIRVYYQKSN GSGSSIHEYVWS
GKWTAGASFGSTVPGTGIGATAIGPGRRLRIYYQATDNKIREHCWDSNSWYVGGFSASASAGVSI AAI SWGSTPNI
RVYWQKGREELYEAAYGGSWNTPGQIKDASRPTPSLPDTFIAANSSGNIDISVFFQASGVSLQQWQWISGKGWSI
GAVVPTGT PAGW

Sekvence 2:

STPGAQQVLFRTGIAAVNSTNHLRVYFQDVYGSIRESLYEGSWANGTEKNVIGNAKLGS PVAATSKELKHIRVYT
LTEGNTLQEFAYDSGTGWYNGGLGGAKFQVAPYSCIAAVFLAGTDALQLRIYAQKPDNTIQEYMWNGDGWKEGTN
LGGALPGTGIGATSFRYTDYNGPSIRIWFQTDLKLQRAYDPHKGWY PDLVTIFDRAPPRTAIAATSFGAGNSS
IYMRIYFVNSDNTIWQVCWDHGKGYHDKGTITPVIQSEVAIISWGSFANNGPDLRLYFQNGTYISAVSEWVWNR
AHGSQLGRSALPPA

Sekvence 3:

SSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTEWC
WDGNGWTKGAYTSTN

Sekvence 4:

MQTA AISWGTTPSIRVYTANGNKITERCYDGSNWYTGAFNQAGDNVSATCWLSGSAVHIRVYATSGGSTTEWCWD
GDGWTRGAYTGL

ALIGNMENT NA GENOVÉ vs. PROTEINOVÉ ÚROVNI

Často se setkáváme se situací, kdy alignment na genové úrovni není pro naše potřeby vhodný. Je tedy zapotřebí výsledné sekvence porovnat i na úrovni proteinu.

ÚKOL 5

U následujících dvojic sekvencí proved'te sequence alignment na genové úrovni (program **lalign** – http://www.ch.embnet.org/software/LALIGN_form.html). Tyto sekvence přeložte do sekvence aminokyselin programem **Translate** – server ExPassy (<http://www.expasy.ch/tools/dna.html>) a proved'te alignment těchto – přeložených sekvencí. Porovnejte množství nespárovaných nukleotidů/aminokyselin (resp. procento identity) v obou případech.

Sekvence A1

atg tgc gca gtg cgc agg gcc ggc tcg aag cgc aag cag gaa gcg ttt gcg gtg atc
ccg gcg act gcg ctg gct aat gcg gta ccg gct agc gtg gct tct gca ccc cgc act
gcc cag cac ctt ccg ctc agt cct cgg cgc ccg cct gcc gcc tcc gga gcg cct gtg
tgg ttt cca aaa aaa gac tta cag caa aat gaa tac tcc agc cat caa gag aat agg
aaa tca cat tac caa gtc tcc tga

Sekvence A2

atg tgt gca gtg cgc cga gcc ggc tcc aag agg aag caa gaa gcg ttt gcg gtt atc
ccg gcg act gct ctg gct aat gca gta ccg gct agc gtg gct tct gca ccg cgc act
gcc cag cat tta ccg ctg agt cct cgc cgg ccg cct gca gct tcc gga gcg cca gtg
tgg ttc cca aaa aaa gat ttg cag caa aat gaa tat tcc agc cac cag gag aat agg
aag tcc cat tac caa gtc tca tga

Identita nt sekvencí a1-a2:

Identita ak sekvencí a1-a2:

Sekvence B1

atg tgc gca gtg cgc agg gcc ggc tgc aag cgc aag cag gaa gcg ttt gcg gtg atc
ccg gcg act gcg ctg gct aat gcg gta ccg gct agc gtg gct tct gca ccc cgc act
gcc cag cac ctt ccg ctc agt cct cgg cgc ccg cct gcc gcc tcc gga gcg cct gtg
tgg ttt cca aaa aaa gac tta cag caa aat gaa tac tcc agc cat caa gag aat agg
aaa tca cat tac caa gtc tcc tga

Sekvence B2

atg tgc gca gtg cgc agg gcc ggc tgc aag cgc aag cag gaa gcg tgt gcg gtg atc
ccg gcg act gcg ctg gct aat gcg gaa ccg gct agc gtg gct tct gca ccc cgc act
gcc cag cac ctt ccg ctc agt cct cgg cgc ccg cct gcc gcc tcc cga gcg cct gtg
tgg ttt cca aaa aaa gac tta cag caa aat gaa tac tcc agc cat taa gag aat agg
aaa tca cat tac caa gtc tcc tga

Identita nt sekvencí b1-b2:

Identita ak sekvencí b1-b2:

VYUŽITÍ ALIGNMENTU PRO INTERPRETACI VÝSLEDKŮ SEKVENACE

Běžným užitím sequence alignmentu je analýza výstupu po sekvenaci. Detekujeme tak mutace (inzerce, delece, substituce), které mohou mít vliv na sekvenci kódovaného proteinu – záměna aminokyseliny, posunutí čtecího rámce, vytvoření nebo odstranění STOP kodonu, atd. Můžeme aplikovat pairwise alignment nebo u více sekvencí multiple alignment.

ÚKOL 6

Následující sekvence obsahují inzerce. Určete, která z obou sekvencí je vhodnější pro budoucí práci s proteinem a proč. Pro alignment použijte vámi zvolený program (lalign, ClustalW, případně jiný).

Původní gen:

atggctgattctcaaacgtcatccaaccgcgcggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgcccagcaacagcacatcaaaattgttcacatcggcgacagccaggaacccgccgcgtatcac
aagctgacgacgcgcgacggcccgcgcgaagccacgctgaattccggcaacggcaagatccggttccgaggtgtcg
gtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgttc
acggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence 1:

atggctgattctcaaacgtcatccaaccgcgcggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgcccagcaacagcacatcaaaattgttcacatcggcgacagccaggaacccgccgcgtatcac
cacaagctgacgacgcgcgacggcccgcgcgaagccacgctgaattccggcaacggcaagatccggttccgaggtg
tcggtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccg
ttcacggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgcgactacaacgacggcatcgtcgtgctc
cagtggccgatcggctga

Sekvence 2:

atggctgattctcaaacgtcatccaaccgcgcggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgcccagcaacagcacatcaaaattgttcacatcggcgacagccaggaacccgccgcgtatcac
aagctgacgacgcgcgacggcccgcgcgaagccacgctgaattccggcaaaacggcaagatccggttccgaggtgt
cgggtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgt

tcacgggtcaacttcgggatcgtcgtgtcgggaagacgggccacgacagcgcgactacaacgacggcatcgtcgtgctcc
agtggccgatcggctga

Vhodnější sekvence:

Důvod:

ÚKOL 7

Následující sekvence obsahují různé mutace. Určete, které z těchto sekvencí jsou použitelné pro budoucí práci s proteinem a proč. Označte nejvhodnější sekvenci.

Originální sekvence:

atggctgattctcaaacgtcacccaaccgcgccggcgaattctcgattccgcccgaataaccgatttccgcgcgatt
ttcttcggaatgcccgcgagcaacagcacatcaaattgttcacgggcgacagccaggaaccgcccgcgtatcac
aagctgacgacgcgcgacggcccgcggaagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgctcggcgaccgacgcgcgctctcgcgccgatcaacggcaagaagtccgacggctcgccggtc
acgggtcaacttcgggatcgtcgtgtcgggaagacgggccacgacagcgcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence 1:

atggctgattctcaaacgtcacccaaccgcgccggcgaattctcgattccgcccgaataaccgatttccgcgcgatt
ttcttcggaatgcccgcgagcaacagcacatcaaattgttcacgggcgacagccaggaaccgcccgcgtatcac
aagctgacgacgcgcgacggcccgcggaagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgctcggcgaccgacgcgcgctctcgcgccgatcaacggcaagaagtccgacggctcgccggtc
acgggtcaacttcgggatcgtcgtgtcgggaagacgggccacgacagcgcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence 2:

atggctgattctcaaacgtcacccaaccgcgccggcgaattctcgattccgcccgaataaccgatttccgcgcgatt
ttcttcggaatgcccgcgagcaacagcacatcaaattgttcacgggcgacagccaggaaccgcccgcgtatcac
aagctgacgacgcgcgacggcccgcgtagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgctcggcgaccgacgcgcgctctcgcgccgatcaacggcaagaagtccgacggctcgccggtc
acgggtcaacttcgggatcgtcgtgtcgggaagacgggccacgacagcgcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence 3:

atggctgattctcaaacgtcacccaaccgcgccggcgaattctcgattccgcccgaataaccgatttccgcgcgatt
ttcttcggaatgcccgcgagcaacagcacatcaaattgttcacgggcgacagccaggaaccgcccgcgtatcac
aagctgacgacgcgcgacggcccgcgtaagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgctcggcgaccgacgcgcgctctcgcgccgatcaacggcaagaagtccgacggctcgccggtc
acgggtcaacttcgggatcgtcgtgtcgggaagacgggccacgacagcgcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence 4:

aaggctgattctcaaacgtcacccaaccgcgccggcgaattctcgattccgcccgaataaccgatttccgcgcgatt
ttcttcggaatgcccgcgagcaacagcacatcaaattgttcacgggcgacagccaggaaccgcccgcgtatcac
aagctgacgacgcgcgacggcccgcggaagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgctcggcgaccgacgcgcgctctcgcgccgatcaacggcaagaagtccgacggctcgccggtc
acgggtcaacttcgggatcgtcgtgtcgggaagacgggccacgacagcgcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence	Charakter mutace z hlediska genu	Charakter mutace z hlediska proteinu	Použitelná pro další práci (ANO/NE) a proč
1			
2			
3			

PROBLÉM REPETIC

Při porovnávání dvou celkově podobných sekvencí užíváme zpravidla metody globálního alignmentu. V případě sekvencí, které jsou podobné jen v určité své části (např. jedné z domén), je vhodnější použít lokální alignment. Ten má svůj význam i v případě proteinů s tzv. repeticemi, tj. opakujícími se úseky, které jsou si navzájem podobné.

ÚKOL 8

Proveďte alignment následujících dvou sekvencí programem **Align** (<http://www.ebi.ac.uk/Tools/psa/>) s použitím algoritmu **Needle** (globální alignment) a **Water** (lokální alignment). V obou případech nastavte parametr Gap open na 15.0 a výsledky porovnejte.

Sekvence 1

PTEFLYTSKIAAISAATGGRQQRVYFQDLNGKIREAQRGGDNPWTGGSSQNVIGEAKLFSPLAAVTWKSQAQGIQ
IRVYCVNKDNLSEFVYDGSKWITGQLGSGVGVKVGSNKLAALQWGGSESAPPNIRVYYQKSNVSGSSIHEYVWS
GKWTAGASFGSTVPGTGIGATAIGPGLRLRIYYQATDNKIREHCWDSNSWYVGGFSASASAGVSIAAISWGSTPNI
RVYWQKGREELYEAAYGGSWNTPGQIKDASRPTPSLPDTFIAANSNGNIDISVFFQASGVSLQQWQWISGKGWSI
GAVVPTGTAPAGW

Sekvence 2

SSVQTAATSWGTVPsirvyTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTTEWC
WDGNGWTKGAYTATN

	Identické ak	Podobné ak	Mezery
Needle			
Water			

Výše uvedené sekvence jsou příkladem repetice, tj. opakujících se podobných (homologních) úseků v rámci jedné sekvence. Přítomnost repetice lze zjistit/ověřit programem **RADAR** (<http://www.ebi.ac.uk/Tools/Radar/>).

ÚKOL 9

V sekvencích z úkolu 8 detekujte repetice pomocí programu Radar. Uveďte počet repetice zjištěných u každé sekvence:

Sekvence 1:

Sekvence 2:

Sekvenci s více repeticemi rozdělte na jednotlivé repetice a proveďte multiple alignment pomocí programu ClustalW. Která z residuí jsou v repeticích konzervována (zcela, částečně)? Využijte tzv. consensus.