BIOINFORMATIKA V PRAXI – CVIČENÍ 2

SEQUENCE ALIGNMENT

STUDIJNÍ MATERIÁLY

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

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:

```
MNGTEGPNFYVPFSNKTGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKNVRTPLNYI
LLNLAVANHFMVFGGFTTTLYTSLHGYFVFGSTGCNLEGFFATLGGEIALWSLVVLAIERYVVVCKPMSNFRFGE
NHAIMGVAFTWVMALACAAPPLVGWSRYIPEGMQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPMTIIFFCYGQ
LVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTHQGSDFGPILMTLPAFFAKSSAI
YNPVIYIMMNKQFRNCMLTTICCGKNPFGEEEGSTTASKTETSQVAPA
```

Název genu/proteinu:
Číslo záznamu v databázi NCBI:
Míra shody zadané a nalezené sekvence:

Sekvence 2:

```
atgtcgagcgttcagaccgctgccacttcgtggggaaccgtcccgtcgatccgtgtgtacacggccaataatggc
aagatcaccgagcgttgctgggacgggaaggggtggtacacgggtgccttcaacgagcccggcgataacgtctcc
gtgaccagctggctggtcggcagcgcgatccatatccgcgtctatgcaagcaccggcaccacgaccacagagtgg
tgctgggacggcaacggctggaccaagggcgcctacaccgccaccaactga
```

Název genu/proteinu:
Číslo záznamu v databázi NCBI:
Míra shody zadané a nalezené sekvence:

Sekvence 3:

```
MGLSDGEWQMVLNIWGKVEGDLAGHGQEVLISLFKAHPETLEKFDKFKNLKSEEEMKSSEDLKKHGCTVLTALGT
ILKKKGQHAAEIQPLAQSHATKHKIPVKYLEFISEVIIQVLKKRYSGDFGADAQGAMSKALELFRNDIAAKYKEL
GFQG
```

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

Nalezněte 4 nejpodobnější sekvence k sekvenci zadané. Použijte aplikaci BLAST.

Sekvence:

```
MNTRSFHRIDVHKARELLQRPDTVLLDCRHPSDFRAGHIAGASPLGDYNADDHVLNIAKHRPVLIYCYHGNASQM
RAQLFADFGFAEVYSLDGGYEAWRKVHTPANSQLTEALQCWLMAQEFPAADIHARTRDGVTPLMRAAGEGDPARV
AELLAAGADPHQRNNDGNQALWFACVSENLDTLDLLVAVGAHLNHQNDNGATCLMYAASAGKTAVVERLLAFGAD
RSLLSLDDFTALDMAANLECLNLLRETPRRIKAVT
```

Číslo záznamu	Protein	Organismus	Score

VLIV POUŽITÉ MATICE NA VÝSLEDEK ALIGNMENTU

Dalším parametrem, 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 a výsledky porovnejte. Která matice je použita při základním nastavení?

Sekvence 1:

```
MPGIRLRYALLALVFAIYYYIVLSYRDQFSDIKKCFSSIRAKIEDYDSSKKGQPKLASNSYLEADMLYRD
RTQVGIENATMVMLVRNRELEGALSSMRSLEDRFNRQYKYPWVFLNDEPFTEEFIEQTMLMASSQTFYEL
IPSSDWNMPDFIDNERVEQNIANSTDVIYGFSKSYRNMCHFNSGYFYKQKRLLNYDWYFRVEPDVEYMCD
FQYDPFTLLRTNNKIYGFVIAIHEYENTIPTLWPTVEKFMQTYPDLIHANNSLKFITTNESSLNHWVTPI
QSSSDYNLCHFWSNFEIGNLNFFRGEAYNKYFDFLDRAGGFYYERWGDAPVHSIGLNLLADKNSIHHFED
IGYYHPPYLACPSSKDVIAAKRCVCRKRGNDGEVIDSAIDVNVFSCLSRWWRYGAGKRFLNEIDYTFNN
```

Sekvence 2:

```
MPESGVPAGRRRLLAPLLLVTSAAALLAVTMANPPDLVDFHVYMLGGVALDRPDTLYSFAYSDQSPDQPL
PFVYPPFAAILFRPLTALPFVVAGVLWQLGILAAVYGIVRISQRLLGGGSHRTAMLWTAGLIWLEPVRVA
LDLGQVGIFLTLAVLYAVCSRRWWLSGLLVGLAAGVKLTPAITGLYFLGVRRWTAAAFSAVVFVATIGLS
YLVVGDQVRNYFTRVMGDTSINPIGIALNQSWRGAISRFLGHDAGGSALVIAAIAGTAVLALFAWAALGA
GSRPRDRLGSVLVVQLFGLLMSPISWVHHWAWVVPLVLWLLSGPWRNEPGARVLGWGWLALTFVGVPSQL
SLLEPSLWEISRPWYLAWAGLAYVVAAVATLGWIVLTGRRNAAPTPPVRRRFARVVGTRGKQIACEEHRA
GR
```

Sekvence 3:

```
MELQSLIDTVSLQKLLLLGALLRLILIAYAFFHDQWFRVKYTDIDYMIVVDGARHMWNGGSPFDRTTFRY
TPLLAALVMPSIWIANPMGKLIFASSDLGAAWYCYGVLKSFAKERSAKWMVSLFILFNPIVLSVSTRGNS
DMLVTFMSLMVLSKFARRKCYQAAAVLGFAVHFKIYPIIYALPLTLGVWEQSVAASTNTWRRVVKTAVVV
SICALMAAISFAVPTVLCYMKYGQQYLNEAFIYHVYREDHRHNFSPYWLLMYLNMARRHLGQGVDFSPRL
VAFAPQAVVLSFVSYKLRRNTAHACCVQTVLFVAFNKVCTVQYFVWFIPFLAFLFCEPKEVEDDESGGSG
AFKFFSWVKALGVVLMWAATIPLWVTTAVPLEFHGYSDFAQLWIVSCLFFLAMVVLASMLARIAYRVQCT
KCSAKSIKVA
```

Sekvence 4:

```
MDASTPNPPTLGTDAVASTIVFFHPDLGIGGAERLVVDAAVGLQTRGHKVVIFTNHCDPTHCFDECRDGT
LDVRVRGNSIVPPSIFSRLTILCAILRHIHLLLTIHLTGELAALSPRAFIVDQLSAGLPLMRFLAPDVPV
LFYCHFPDLLLAQGRQSLVKRLYRVPFDRLEEWSMGFAHAVAVNSKFTRGIVGNTWPALQNKVPINVVYP
CVDTHTTHETAPDEAKLAAGKKLILSINRFERKKDIGLAIRAFAQIPEEQRRGARLVLAGGYDARVSENV
LYHAELQALATSLSLAHHTLTPAELGSAAAPPDAQHFGIVPLEAMLARVPVLAANTGGPVETVADRETGW
LRDPADAPAWTDVMARCLALPDDQLAAMGDAGRRRVRELFGRDKMAQTLDESLVQIAGLAEERRVSGAAG
FGVLAAFIAACAVLAAWFAF
```

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

```
RIQVRLGEHNIEVLEGNEQFINAAKIIRHPKYDRKTLNNDIMLIKLSSRAVINARVSTISLPTAPPATGTKCLIS
GWGNTASSG
```

Sekvence 2:
FlgA [Agrobacterium tumefaciens] – AAB71791.1

```
MRFGRNNSSCRTALVRMCLASAFSLGALAPALAQAPMALVPVRTIYPGEAISPEQVKSVEVTNPNISAGYASDIS
EVEGMISKQTLLPGRTIPIAALREPSLVVRGTSVKLVFHIGNMTLMASGTPMSDGSLGEVVRVRNIDSGVMVSGT
VMKDGTIQVMAK
```

Sekvence 3:
collagen [Caenorhabditis elegans] – CAA35955.1

```
MSEDLKQIAQETESLRKVAFFGIAVSTIATLTAIIAVPMLYNYMQHVQSSLQSEVEFCQHRSNGLWDEYKRFQGV
SGVEGRIKRDAYHRSLGVSGASRKARRQSYGNDAAVGGFGGSSGGSCCSCGSGAAGPAGSPGQDGAPGNDGAPGA
PGNPGQDASEDQTAGPDSFCFDCPAGPPGPSGAPGQKGPSGAPGAPGQSGGAALPGPPGPAGPPGPAGQPGSNGN
AGAPGAPGQVVDVPGTPGPAGPPGSPGPAGAPGQPGQAGSSQPGGPGPQGDAGAPGAPGAPGQAGAPGQDGESGS
EGACDHCPPPRTAPGY
```

Gap open = 1
Vzájemně příbuznější sekvence:

Gap open = 100
Vzájemně příbuznější sekvence:

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 proveďte 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 proveďte 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 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 b2

```
atg tgc gca gtg cgc agg gcc ggc tcg 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:

```
atggctgattctcaaacgtcatccaaccgcgccggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgccgagcaacagcacatcaaattgttcatcggcgacagccaggaacccgccgcgtatcac
aagctgacgacgcgcgacggcccgcgcgaagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgttc
acggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga
```

Sekvence 1:

```
atggctgattctcaaacgtcatccaaccgcgccggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgccgagcaacagcacatcaaaaatttgttcatcggcgacagccaggaacccgccgcgtat
cacaagctgacgacgcgcgacggcccgcgcgaagccacgctgaattccggcaacggcaagatccgtttcgaggtg
tcggtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccg
ttcacggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgactacaacgacggcatcgtcgtgctc
cagtggccgatcggctga
```

Sekvence 2:

```
atggctgattctcaaacgtcatccaaccgcgccggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgccgagcaacagcacatcaaattgttcatcggcgacagccaggaacccgccgcgtatcac
aagctgacgacgcgcgacggcccgcgcgaagccacgctgaattccggcaaaacggcaagatccgtttcgaggtgt
cggtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgt
tcacggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgactacaacgacggcatcgtcgtgctcc
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:

```
atggctgattctcaaacgtcatccaaccgcgccggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgccgagcaacagcacatcaaattgttcatcggcgacagccaggaacccgccgcgtatcac
aagctgacgacgcgcgacggcccgcgcgaagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgttc
acggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga
```

Sekvence 1:

atggctgattctcaaacgtcatccaaccgcgccggcgagttctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgccgagcaacagcacatcaaattgttcatcggcgacagccaggaaccagccgcgtatcac
aagctgacgacgcgcgacggcccgcgcgaagccacgttgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgttc
acggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence 2:
atggctgattctcaaacgtcatccaaccgcgccggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgccgagcaacagcacatcaaattgttcatcggcgacagccaggaacccgccgcgtatcac
aagctgacgacgcgcgacggcccgcgcgtagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgttc
acggtcaactttgggatcgtcgtgtcggaagacggccacgacagcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence 3:
atggctgattctcaaacgtcatccaaccgcgccggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgccgagcaacagcacatcaaattgttcatcggcgacagccaggaacccgccgcgtatcac
aagctgacgacgcgcgacggcccgcgctaagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgttc
acggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence 4:
aaggctgattctcaaacgtcatccaaccgcgccggcgaattctcgattccgccgaataccgatttccgcgcgatt
ttcttcgcgaatgccgccgagcaacagcacatcaaattgttcatcggcgacagccaggaacccgccgcgtatcac
aagctgacgacgcgcgacggcccgcgcgaagccacgctgaattccggcaacggcaagatccgtttcgaggtgtcg
gtgaacggcaagccgtcggcgaccgacgcgcgtctcgcgccgatcaacggcaagaagtcggacggctcgccgttc
acggtcaacttcgggatcgtcgtgtcggaagacggccacgacagcgactacaacgacggcatcgtcgtgctccag
tggccgatcggctga

Sekvence	Charakter mutace z hlediska genu	Charakter mutace z hlediska proteinu	Použitelná pro další práci (ANO/NE)
1			
2			
3			
4			

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.

ÚKOL 8

Proveďte alignment následujících dvou sekvencí programem Align (http://www.ebi.ac.uk/Tools/emboss/align) s použitím algoritmu needle (globalní alignment) a water (lokální alignment). V obou případech nastavte parametr Gap open na 15.0 a výsledky porovnejte.

Sekvence 1
PTEFLYTSKIAAISWAATGGRQQRVYFQDLNGKIREAQRGGDNPWTGGSSQNVIGEAKLFSPLAAVTWKSAQGIQ
IRVYCVNKDNILSEFVYDGSKWITGQLGSVGVKVGSNSKLAALQWGGSESAPPNIRVYYQKSNGSGSSIHEYVWS

GKWTAGASFGSTVPGTGIGATAIGPGRLRIYYQATDNKIREHCWDSNSWYVGGFSASASAGVSIAAISWGSTPNI
RVYWQKGREELYEAAYGGSWNTPGQIKDASRPTPSLPDTFIAANSSGNIDISVFFQASGVSLQQWQWISGKGWSI
GAVVPTGTPAGW

Sekvence 2

SSVQTAATSWGTVPSIRVYTANNGKITERCWDGKGWYTGAFNEPGDNVSVTSWLVGSAIHIRVYASTGTTTTEWC
WDGNGWTKGAYTATN

	Identické ak	Podobné ak	Mezery
Needle			
Water			

Výše uvedené sekvence jsou příkladem repetic, tj. opakujících se podobných (homologních) úseků v rámci jedné sekvence. Přítomnost repetic 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 repetic zjištěných u každé sekvence:

Sekvence 1:
Sekvence2:

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.