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:

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

Sekvence 2:

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

Sekvence 3:

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:

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. Poté proveďte alignment s použitím matice PAM, BLOSUM, Gonnet a ID a výsledky porovnejte. Která matice je použita při základním nastavení?

Sekvence 1:

Sekvence 2:

Sekvence 3:

Sekvence 4:

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** (http://www.ebi.ac.uk/Tools/clustalw2/index.html). V prvním případě nastavte parametr GAP OPEN na 1, ve druhém případě na 100 a výsledky porovnejte.

Sekvence 1:

Sekvence 2:

Sekvence 3:

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

Sekvence a2

Identita nt sekvencí a1-a2:
Identita ak sekvencí a1-a2:

Sekvence b1

Sekvence b2

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).

Sekvence 1:

Sekvence 2:

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.

Sekvence 1:

Sekvence 2:

Sekvence 3:

Sekvence 4:

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

Sekvence 2

	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.