Základy bioinformatiky
Tutorial 2: Zarovnávání sekvencí

V tomto tutoriálu si ukážeme binární a vícečetné zarovnávání nukleotidových a aminokyselinových sekvencí. Zarovnávání sekvencí představuje optimalizační problém. Tento problém řeší sofistikované algoritmy v různých programech a webových serverech, ale my si jej na jednoduché ukázce můžeme vyřešit i „ručně“.

Při binárním zarovnáváním sekvencí máme k dispozici dvě nukleotidové či aminokyselinová sekvence, které chceme zarovnat tak, aby bylo dosaženo maximální skóre. Toto skóre je dáno tím jaké báze nebo aminokyseliny jsou v zarovnání pod sebou a kolik a jak dlouhých mezer se v zarovnání nachází. Před tím, než provedeme porovnání musíme vybrat vhodnou podobnostní matici a zvolit hodnoty „cen mezer“ (*gap open penalty* a *gap extension penalty*). Podobnostní matice pro zarovnávání nukleotidových sekvencí může vypadat například takto:

	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

Tato matice vyjadřuje, že například adenin je nejméně variabilní bází (skóre A-A = 10), zatímco nejméně pravděpodobná je záměna cytosinu za guanin (skóre G-C = C-G = –5). Skóre můžeme vypočítat jako součet příspěvků jednotlivých dvojic přes celé zarovnání. Pokud se v zarovnání nachází mezery, pak se od skóre odečítají ceny mezer. Některé programy mají jedinou cenu mezer. Pak se od skóre odečítá tato hodnota pro každou mezeru. Jiné programy mají dvě ceny mezer (*gap open penalty* a *gap extension penalty*). Pak se od skóre odečítá *gap open penalty* pro každý začátek a *gap extension penalty* pro každé pokračování mezery. Touto dvojí cenou mezer je možné řídit, jestli v zarovnání bude málo krátkých, málo dlouhých, hodně krátkých nebo hodně dlouhých mezer. Pokuste se vypočítat skóre pro tato dvě alternativní zarovnání sekvencí s výše uvedenou podobnostní maticí a s *gap penalty* rovnou 5 (pouze jediná cena mezer):

```
sekvence1   AGACTAGTTAC
sekvence2   AAGTT----AC

sekvence1   AGACTAGTTAC
sekvence2   AAG----TTAC
```

Pokuste se nalézt optimální zarovnání sekvencí. Dále vypočtěte skóre zarovnání sekvence1 samu se sebou a sekvence2 samu se sebou.

Binární zarovnání sekvencí v reálu si ukážeme na sekvenci mamutí cytochrom-*c*-oxidasy a homologní sekvenci rostlinného enzymu. Najděte si tyto sekvence postupem představeným v minulé hodině. Dále najděte v nástrojích *EBI* položku *EMBOS tools* a program *Needle*. Tento program využívá algoritmus podle Needlemana a Wunsche. Zkopírujete si sekvenci mamutího a vybraného rostlinného enzymu do každého ze dvou políček a zmáčkněte *Submit*. Postup opakujte s různým nastavením parametrů *Matrix*, *Gap open penaly*, *Gap extension penalty*, případně dalších parametrů. K ním se můžete dostat před spuštěním programu zmáčknutím tlačítka *More options*.

Vícečetné zarovnávání představuje složitější problémem než zarovnání binární. Zatímco výše použitý *Needlemanův-Wunschův* algoritmus z principu vždy nalezne optimální zarovnání (i když mu to trvá velmi dlouho), v případě vícečetných zarovnání představuje nalezení optima velký, většinou neřešitelný problém. My si ukážeme zarovnání na programu *ClustalW2*, který zdaleka ne vždy nalezne optimální zarovnání. Program funguje tak, že nejprve zarovná každou sekvenci s každou, vypočte jejich podobnost, vytvoří improvizovaný fylogenetický strom a nakonec na

základě stromu vytvoří zarovnání. Nalezněte sekvenci mamutí cytochrom-c-oxidasy a homologní enzymů z rostlin. Z nich vyberte vzorek například pěti sekvencí. To je možné provést tak, že ve výstupu z programu *BLAST* zmáčknete tlačítko *Clear* a pak zaškrtnete políčka u jednotlivých sekvencí. Nemá cenu zaklikávat více stejných sekvencí. Pak vyberte formát sekvencí *Fasta* a zmáčkněte *Download*. Sekvence zkopírujete do textového editoru. Je praktické zkrátit označení sekvencí za „zobáčkem“, například je zredukovat na kód databáze *UniProt*. K rostlinným sekvencím můžeme přidat i sekvenci z mamuta. Nyní z nástrojů vyberte program *ClustalW2*, do hlavního okna vložte záznam sekvencí a zmáčkněte *Submit*. Program vám vytvoří vícečetné zarovnání sekvencí ve formátu vlastním programům rodiny *Clustal*. My budeme pro další použití chtít zarovnání sekvencí ve formátu *FASTA*. Ten získáme tak, že opět spustíme program *ClustalW2* s tím, že před jeho spuštěním zmáčkneme v položce *STEP 3 - Set your Multiple Sequence Alignment Options* tlačítko *More options* a vybereme formát výstupu *Pearson/FASTA*. Získáme výstup ve formátu *FASTA*.

Pro získání krásného zarovnání sekvencí do odborné publikace nebo absolventské práce je možné použít program *ESPript3*, který jako první z námi používaných serverů není součástí portálu *EBI*. Tento server najdete na stránce http://espript.ibcp.fr/ESPript/ESPript/. Tento program nepočítá zarovnání sekvencí, pouze ze zadaného zarovnání vytvoří jeho grafickou reprezentaci. Pro použití programu musíte uložit zarovnání z programu *ClustalW3* ve formátu *FASTA*. Ten zkopírujte do textového souboru a uložte v čistě textovém formátu (např. *txt*, nikoliv *doc*, *docx* a podobně). Dále pokračujte zmáčknutím *Run ESPript*, uploadujte sekvenci a zmáčkněte *Submit*. Program vyžaduje povolení Pop-up oken. V jednom z nich se vám objeví odkaz na výsledek. Je možné vybrat prezentaci ve formátu *Postscript*, *PDF* a *PNG* s různým rozlišením. Dále je možné vybrat velikost písma, výšku řádků, barevné provedení, velikost a orientaci stránky a mnoho dalších parametrů.