Základy bioinformatiky
Tutorial 1: Hledání sekvencí v databázích

Cílem tohoto tutoriálu bude nalézt aminokyselinovou a nukleotidovou sekvenci mamutí cytochrom-c-oxidasy. Dále nalezneme jí podobné sekvence. Sekvenci z mamuta spolu s další vybranou sekvencí poté binárně zarovnáme. Pro tyto účely budeme nejčastěji používat prostředky Evropského bioinformatického institutu (EBI, http://www.ebi.ac.uk). Nejprve navštivte tyto stránky. Tento tutoriál je napsán tak, že vystihuje stav stránek v únoru 2014 s tím, že se stránky mohou samozřejmě v průběhu času měnit. Hlavní stránce dominuje textové pole s nadpisem „*Explore the EBI:*“. Vpravo je několik odkazů s nadpisem „*Popular*“. Nejprve se pokusíme nalézt všechny molekulárně biologické informace o sibiřském mamutovi *Mammuthus primigenius*. Zadejte tento latinský název do pole „*Explore the EBI:*“ a zmáčkněte knoflík „*Search*“. Portál Vám nalezne, nebo alespoň v únoru 2014 nalezl, 36 odkazů na odbornou literaturu, 2 380 nukleotidových a 356 proteinových sekvencí, tři prostorové struktury a další odkazy. Je nutné zmínit, že ne všechny nalezené sekvence musí nutně patřit sibiřskému mamutovi. Například jeden ze článků se zabývá sekvenováním mitochondriálního genomu sibiřského mamuta a jeho srovnáváním s genomem slonů. Proto je možné mezi výsledky hledání nalézt i sloní geny, a to proto, že jejich záznamy obsahují název článku a ten obsahuje jméno *Mammuthus primigenius*.

Nejprve se podíváme na aminokyselinové sekvence. Ty získáte tak, že kliknete na odkaz „*View all 356 results for Protein sequences*“. Mezi nalezenými proteinovými sekvencemi naleznete sekvenci označenou:

COX1_MAMPR (Q38PS0)
Cytochrome c oxidase subunit 1
***Mammuthus primigenius* (Reviewed)**

Enzym cytochrom-c-oxidasa je vícepodjednotkový mitochondriální enzym, jehož některé podjednotky jsou kódovány mitochondriálním a jiné jaderným genomem. Pro účely tohoto tutoriálu se budeme zabývat podjednotkou jedna, která nese výše uvedené označení. Označení **`COX1_MAMPR`** je označení, které nese protein v databázi *Swiss-Prot*, která se přetransformovala v databázi *UniProt*. Z tohoto označení je možné odhadnout, že se jedná o první podjednotku cytochrom-c-oxidasy (**`COX1`**) a že se jedná o mamuta (**`MAMPR`**). Alternativním označením je **`Q38PS0`**, které pochází z databáze *TrEMBL* (přeložené nukleotidové sekvence z databáze Evropské molekulárně-biologické laboratoře). Pokud na tento odkaz kliknete, dozvíte se, že se jedná o první podjednotku cytochrom-c-oxidasy (anglický název je bez druhé pomlčky), že pochází ze sibiřského mamuta, dále získáte taxonomické odkazy na mamuta, délku sekvence a informace o katalyzované reakci a biologické funkci. Dále jsou pro strukturu předpovězeny transmembránové úseky a vazebná místa kofaktorů. Následuje aminokyselinová sekvence. Pokud Vám z nějakého důvodu vadí přítomnost číslic vyjadřujících pořadí v aminokyselinové sekvenci, pak můžete kliknout na odkaz „*FASTA*“. Označení „*FASTA*“ nese jednak program pro prohledávání sekvenčních databází (dnes méně populární než jeho konkurent BLAST) a rovněž i jeden z nejjednodušších formátů pro zápis sekvencí. Po kliknutí na odkaz „*FASTA*“ získáte sekvenci ve formáty FASTA, tedy tento nebo podobný výstup:

```
>sp|Q38PS0|COX1_MAMPR Cytochrome c oxidase subunit 1 OS=Mammuthus primigenius GN=MT-CO1 PE=3 SV=1
MFANRWLYSTNHKDIGTLYLLFGAWAGMVGTAFSILIRAELGQPGSLLGDDQIYNVIVTA
HAFVMIFFMVMPIMIGGFGNWLIPLMIGAPDMAFPRMNNMSFWLLPPSFLLLLASSMVEA
GTGTGWTVYPPLAGNLAHAGASVDLTIFSLHLAGVSSILSAINFITTIINMKPPAMSQYH
MPLFVWSILVTAVLLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGH
PEVYILILPGFGMVSHIVTYYSGKKEPFGYMGMVWAMMSIGFLGFIVWAHHMFTVGMDVD
TRAYFTSATMIIAIPTGVKVFSWLATLHGGNIKWSPAMMWALGFIFLFTIGGLTGIVLAN
SSLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFIHWFPLFSGYTLNHTWAKIQFLVMFIG
VNLTFFPQHFLGLSGMPRRYSDYPDAYTAWNTASSMGSFISLVAVILMVFMIWEAFASKR
EVSVMELTTTNVEWLNGCPPPHHTFEEPAYVKSNS
```

První řádek obsahuje identifikátor sekvence a vždy začíná zobáčkem (znakem >). Na dalších řádcích následuje vlastní sekvence. Tím je řečeno vše nutné o tomto formátu. Formát FASTA, alespoň v obecném použití, si neklade specifické nároky na počet znaků v jednom řádku nebo počtu řádků. Je možné jej použít pro aminokyselinové i nukleotidové sekvence a jak pro jednu tak i pro více sekvencí (stačí pospojovat více sekvencí za sebou tak, aby každou předcházel řádek se zobáčkem a názvem). Dále je možné tento formát použít i na zarovnání sekvencí, jak bude ještě ukázáno.

Pokud se vrátíte na předchozí stránku, ještě zde můžete nalézt odkazy na literaturu. Oba články pojednávají o isolaci mitochondriální DNA z mamutí kosti a jejím sekvenování. Dále zde naleznete odkazy na nukleotidové sekvence, modely trojrozměrných struktur, metabolické informace, předpokládanou doménovou strukturu a další údaje. Odkaz s nukleotidovou sekvencí **DQ188829** můžete navštívit a prohlídnout si stránku. Tato stránka obsahuje kompletní mamutí mitochondriální genom. Dole se Vám objeví sekvence s rozsahem „*Base range: 1 - 1000 of 16770*“, kterou změňte na „*Base range: 1 - 20000 of 16770*“ a zmáčkněte „*Apply*“, aby se zobrazila celá sekvence.

Nukleotidovou sekvenci si můžeme přeložit do aminokyselinové sekvence. Pro tento účel si otevřete nové okno prohlížeče a v něm jděte na stránku www.ebi.ac.uk a zde pokračujte na odkaz „*Services*“, poté „*Proteins*“ a nakonec „*EMBOSS tools*“. Zde vyberte nástroj „*Transseq*“. Do okna vložte sekvenci mamutí mitochondriální DNA, položku „*FRAME*“ změňte na „*6 (All six frames)*“ a „*CODON TABLE*“ změňte na „*Vertebrate Mitochondrial*“. Rozdíl mezi univerzální kodonovou tabulkou a kodonovou tabulkou mitochondrií obratlovců můžete nalézt v literatuře. Šest čtecích rámců představuje tři čtecí rámce v jednom a další tři v druhém směru. Pokud spustíte program, po chvíli se vám ukáže šest různých aminokyselinových sekvencí. Geny je možné alespoň do určité míry nalézt jako úseky, které nejsou přerušovány hvězdičkami, které představují stop-kodony. Důvodem je fakt, že zatímco proteiny mívají obvykle délku několika stovek aminokyselinových zbytků, nekódující sekvence a sekvence ve špatném čtecím rámci obsahují jeden stop-kodon v průměru na přibližně dvacet aminokyselin. Sekvenci cytochrom-c-oxidasy najdete pomocí vyhledávání v textu zadáním její počáteční sekvence, tedy `MFANRWLYST`. Je možné se přesvědčit, že sekvence, která tento řetězec následuje, není přerušována stop-kodony. Celá kódující sekvence končí nejbližším stop-kodonem. Nalezení počátku genu je rovněž možné a ukážeme si jej v jednom z příštích tutoriálů. Dále si můžete všimnout, že různé geny jsou kódovány v různých čtecích rámcích. Mitochondriální geny, na rozdíl od jaderných, zpravidla neobsahují introny.

Pokud bychom tuto sekvenci získali sekvenací mamutí DNA, ale nevěděli bychom, že se jedná o cytochrom-c-oxidasu, nebo bychom nevěděli že se jedná o mamuta, pak bychom nejspíše použili

program *BLAST*. Stejný program bychom použili na nalezení podobných proteinů s cílem předpovědět prostorovou strukturu, zjistit které části proteinu jsou evolučně konzervované a které nikoliv, zjistit jestli existují nějaké patenty týkající se tohoto enzymu a podobně. Použití programu *BLAST* si nyní ukážeme, konkrétně si ukážeme nalezení podobných aminokyselinových sekvencí aminokyselinové sekvence mamutí cytochrom-c-oxidasy. Vraťte se nyní na stránku s aminokyselinovou sekvencí a v novém okně si otevřete stránku www.ebi.ac.uk. Pro názornost použijeme lehce složitější postup (jednodušší spočívá v tom, že vedle sekvence vyberete položku „*BLAST*“ a zmáčknete „*go*“). V rámci složitějšího postupu klikněte na odkaz *FASTA*, myší vyberte sekvenci a zkopírujte jí do schránky. Pak v novém okně na stránce EBI nalezněte odkaz na program *NCBI BLAST*. Program *BLAST* existuje v několika variantách podle toho jakou sekvenci zadáváte a jakou databázi chcete prohledávat. Program *blastp* prohledává proteinové databáze na základě proteinového dotazu. Program *blastn* prohledává nukleotidové databáze na základě dotazu ve formě nukleotidové sekvence. Program *blastx* vezme nukleotidovou sekvenci, přeloží ji do proteinové sekvence (ve všech šesti čtecích rámcích) a prohledá s ní proteinové sekvence. Naproti tomu program *tblastn* vezme proteinovou sekvenci a prohledá s ní nukleotidové sekvence přeložené do všech šesti čtecích rámců. Nakonec, program *tblastx* vezme nukleotidovou sekvenci, přeloží jí do všech šesti čtecích rámců a s výsledkem prohledá nukleotidové sekvence přeložené do všech šesti čtecích rámců. My použijeme program *blastp*.

Program *BLAST* umožňuje vybrat databázi kterou chceme prohledávat. V defaultní nastavení je vybraná databáze *UniProt Knowledgebase*. Jedná se o obecnou databázi, v případě programu *blastp* proteinových. Ostatní databáze obsahují různé zúžené výběry, například dobře charakterizované proteiny, proteiny pouze určitých organizmů, proteiny se známou prostorovou strukturou, patentované sekvence a podobně. Do dalšího pole zkopírujte sekvenci a zmáčkněte tlačítko *Submit*. Program nějakou dobu bude chroustat a pak vám vytiskne seznam nalezených výsledků. Na prvních třech místech program nalezl sekvence, které jsou identické se zadanou sekvencí. Konkrétně se jedná o jeden protein z *Mammuthus columbi* a dva z *Mammuthus primigenius*. Mezi nimi je i námi zadaná sekvence (aniž bychom to programu předem říkali). Zmáčknutím záložky *Tool Output* se zobrazí zarovnávání sekvencí. Další mamutí sekvence obsahuje jednu mutaci. Dále následují Slon africký, Slon indický a různá další více nebo méně očekávaná zvířátka.

Nalezené sekvence jsou si tak podobné, že neobsahují ani jednu mezeru (gap). Pokud bychom si chtěli ukázat jak bude vypadat zarovnání sekvencí s mezerami, pak se musíme podívat na nějakou vzdálenější skupinu organismů. Proto se vraťte na úvodní stránku programu *BLAST* a vypněte databázi *UniProt Knowledgebase* a místo ní vyberte databázi *UniProtKB Taxonomic Subsets* a dále *UniProtKB Viridiplantae*. Díky tomu můžete prohledávat pouze rostlinné sekvence. Jako první nám program nalezne nějakou vodní rostlinu, kapradí a další. Kliknutím na odkaz s označením sekvence získáte záznam podobný záznamu mamutí sekvence.

Nakonec si ukážeme binární zarovnání sekvencí. Pro tento účel najděte v nástrojích *EBI* položku *EMBOS tools* a dále 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*. Jedná se o vhodný příklad pro ukázku nastavení parametrů *Matrix*, *Gap open penaly*, *Gap extension penalty* a dalších. K ním se můžete dostat před spuštěním programu zmáčknutím tlačítka *More options* … V dalším tutoriálu si ukážeme tvorbu vícečetného zarovnání sekvencí.