Základy bioinformatiky
Tutorial 5: Studium molekulární fylogeneze

Koncept fylogenetických stromů („stromu života“) je velmi intuitivní pro lidi s biologickým nebo biochemickým základem. Jako určité zobecnění fylogenetických stromů do ostatních vědních a technických oborů můžeme považovat hierarchickou shlukovou (clusterovou) analysu. Fylogenetické vztahy je možné studovat na mnoha úrovních s různými vstupními daty. Tradičně biologové porovnávají různé morfologické a další nemolekulární parametry organismů, například délky ploutví u ryb nebo tvary křídel motýlů. S rozvojem sekvenování DNA se začaly vytvářet fylogenetické stromy na základě jejich sekvencí. V dnešní době je možné fylogeneticky analysovat i celé genomy nebo informace z jiných -omických projektů. Kromě porovnávání různých organismů na základě sekvencí určitého proteinu nebo DNA je možné porovnávat i sekvence podobných proteinů nebo podobných DNA v jednom organismu, například proteinkinas, receptorů atd.

Pokud chceme vytvořit fylogenetický strom na základě vybraných sekvencí, pak si musíme položit otázku k čemu má tento strom sloužit. V tomto tutoriálu si ukážeme tvorbu základních stromů, které ilustrují vzájemné podobnosti mezi sekvencemi. Pokud bychom chtěli vytvářet skutečné fylogenetické stromy, tedy detailně studovat evoluční vztahy mezi druhy, například vypočítat před kolika miliony let došlo k oddělení druhů ze společného předka, pak na to jeden tutoriál nestačí. Případné zájemce je možné odkázat na učebnice evoluční biologie, např. J. Flegr: Evoluční biologie. Academia Praha, 2009.

Na fylogenetickém stromě je každý analysovaný organismus representován jako konec větve. Stromy počítají s existencí společného předka, který je representován rozvětvením dvou větví. Vznik oddělených linií nazýváme **kladogenesí**. Naopak hromadění genotypových změn v rámci větve nazýváme **anagenesí**. **Fylogenese** je tedy kombinací kladogenese a anagenese. Kromě těchto procesů může docházet k **horizontálnímu přenosu** a **syngenesi**. Společný předek všem analysovaným organismům nazýváme **kořen**. Fylogenetické stromy mohou být buď **zakořeněné** nebo **nezakořeněné**. Fylogenetické stromy také mohou být **kvantitativní** nebo **kvalitativní** podle toho, jestli linie oddělující jednotlivé rozvětvení a vrcholy mají délku, která vyjadřuje evoluční vzdálenost, nebo jestli mají uniformní délky.

Pro otestování pochopení koncepce fylogenetických stromů si můžeme vyzkoušet „ručně“ vytvořit jednoduchý kvalitativní nezakořeněný strom pro tyto sekvence:

```
>A
ATGCCGTTGCTA
>B
AACCCGTTGCTA
>C
ATGCCGTGCGTC
>D
ATGCCGTGCGGC
```

Prakticky všechny metody fylogenetické analysy vycházejí z mnohočetného zarovnání sekvencí. Mezi nejjednodušší metody pro konstrukci fylogenetických stromů patří například metoda *Unweighted Pair Group Method with Arithmetic Mean* (**UPGMA**). Ta spočívá v tom, že se porovnají sekvence každá s každou. Mezi sekvencemi vyjádříme vzdálenost, například jako (1 – (*identita v procentech*/100)). Získáme tak vzdálenostní matici. V matici nalezneme dvojici s nejmenší vzdáleností. Ty spojíme formu dvou větví a jednoho uzlu. Řádky matice, které odpovídají těmto organismům, nahradíme aritmetickým průměrem vzdáleností a takto pokračujeme dále dokud nevytvoříme celý strom.

Další metodou je metoda *Neighbour Joining* (**NJ**), kterou budeme využívat. Tato metoda vychází z hvězdicového fylogenetického stromu kde mají všechny organismy stejnou vzdálenost od společného předka. Tento strom je iterativně vylepšován dokud co nejlépe neodpovídá vzdálenostní

matici.

Mezi další metody patří *Maximum Parsimony* (**MP**). Zatímco předchozí dvě metody nejprve porovnají sekvence, vypočtou jejich vzdálenosti a pak už pracují jen se vzdáleností maticí, metoda MP pracuje se samotnými sekvencemi tak, že hledá minimální počet mutací pro nalezeni společných předků. Poslední metodou je *Maximum Likehood* (**ML**), která generuje všechny možné topologie stromů, počítá jejich pravděpodobnosti a pak generuje strom jako jejich konsensus.

V případě „ruční“ konstrukce fylogenetického stromu jste se mohli přesvědčit, že ani v tak jednoduchém případě nebylo možné vytvořit strom takový, aby délky větví přesně odpovídaly vzdálenostem sekvencí. Z tohoto důvodu fylogenetický strom vyjadřuje vzdálenosti pouze přibližně. Pokud chceme otestovat přesnost fylogenetického stromu, pak je možné použít metody *bootstrapping*. Pokud analysujeme například deset sekvenci o délce sto aminokyselin, pak nejprve provedeme jejich zarovnání. Pokud zarovnání sekvencí obsahuje mezery, pak je většinou pro fylogenetickou analysu ignorujeme. Zarovnání sekvencí analysujeme pomocí zvolené metody a získáme fylogenetický strom. Pak vezmeme zarovnání sekvencí a nastříháme ho na sto proužků s deseti aminokyselinami pod sebou. Pak je sesypeme do klobouku a náhodně taháme a zapisujeme do nového zarovnání. Po každém zápisu vrátíme proužek do klobouku a zamícháme, takže v novém zarovnání se může jeden proužek objevit vícekrát nebo některý proužek může chybět. Pro toto zarovnání vypočteme nový fylogenetický strom. Tento postup opakujeme mnohokrát, například stokrát. Získáme tak sto stromů, které se budou lišit ve vzdálenostech a mohou se lišit i v topologii. K větvením originálního stromu pak můžeme dopsat *bootstrap* hodnoty, které budou 100 pokud se dané větvení vyskytuje ve všech stromech a například 50 pokud se vyskytuje v 50 ze 100 stromů.

Je nutné zmínit, že spousta programů vytváří fylogenetické stromy s kořenem, i když tento kořen jako společný předek je pochybný. V tom případě je vhodnější zobrazit strom jako nezakořeněný. Pokud chceme opravdu získat kořen, pak je možné k sekvencím přidat nějakou pořád podobnou, ale od ostatních vzdálenou sekvenci, neboli **outgroup**. Například při tvorbě fylogenetického stromu sekvencí 16S RNA bakteriích rodu *Klebsiella*, pak je jako outgroup možné použít sekvenci bakterie *E. coli*, která bude sekvencím podobná, ale vzdálenost mezi jakoukoliv Klebsiellou a *E. coli* bude vždy vyšší než mezi Klebsiellami. Potom větev pro *E. coli* můžeme považovat za kořen stromu.

Pro praktické provedení si otevřete prohlížeč na stránce *EBI* a nalezněte sekvenci první podjednotky mamutí cytochrom-c-oxidasy. Pak pomocí *NCBI BLASTu* nalezněte několik (například 10-20) sekvencí podobných proteinů. Je vhodné vybrat databasi SwissProt a rozšířit výstup na více sekvencí ve výstupu, například 100. Vyberte vhodné sekvence živočichů a zakřížkujte je. Pak zadejte *Download entries* ve formátu *FASTA*. Pak si otevřete program *ClustalW2*, který naleznete mezi nástroji EBI. Program *ClustalW2* vám zobrazí porovnání. Tento program funguje tak, že vezme sekvence, porovná každou s každou, vytvoří jednoduchý strom (*guide tree*) a ten pak použije pro vytvoření zarovnání. Tento strom je jen pomůckou pro tvorbu zarovnání a nikoliv opravdovým fylogenetickým stromem. Opravdový fylogenetický strom získáte kliknutím na záložku Phylogenetic tree. Objeví se vám jednak grafické vyobrazení stromu a dále záznam stromu v textovém formátu. K jeho ilustraci je možné použít program Phylodendron (http://iubio.bio.indiana.edu/treeapp/treeprint-form.html).

Program pro metodu *Maximum Likehood* si můžeme ukázat na webovém portálu (http://www.phylogeny.fr/version2_cgi/simple_phylogeny.cgi) ve Francouzském Montpellier.