Tvorba fylogenetických stromů

Fylogenetický strom

Fylogenetický strom je grafické znázornění příbuzenských vztahů mezi různými taxonomickými jednotkami, o nichž lze předpokládat, že mají společného předka. Příbuzenské vztahy se zde posuzují na základě morfologické či genetické podobnosti. Místo taxonomických jednotek mohou v některých stromech vystupovat přímo jednotlivé biologické druhy nebo i jednotlivé geny.

Pojem strom je převzat z teorie grafů, kde označuje neorientovaný souvislý acyklický graf. Vrcholy, které jsou spojeny hranami se dvěma a více dalšími vrcholy, označujeme jako *vnitřní vrcholy*. Zbývající vrcholy, které jsou spojeny pouze s jedním dalším vrcholem, se nazývají *listy*.

V případě fylogenetických stromů každý vrchol představuje určitou taxonomickou jednotku a hrana mezi dvěma vrcholy značí vztah mezi taxonomickými jednotkami, které tyto vrcholy reprezentují. V závislosti na typu fylogenetického stromu může délka hrany udávat dobu vývoje nebo míru podobnosti mezi příslušnými taxonomickými jednotkami.

Typy fylogenetických stromů

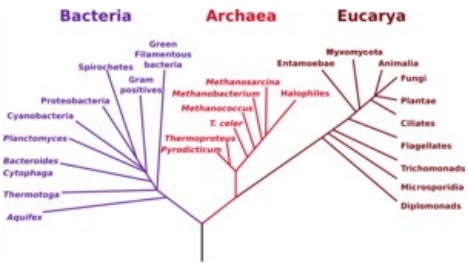
Nezakořeněný fylogenetický strom

Tento typ stromu znázorňuje vztahy mezi taxonomickými jednotkami, aniž by specifikoval jejich společného předka.

Zakořeněný fylogenetický strom

Zakořeněný fylogenetický strom je strom, u kterého byl jeden z vnitřních vrcholů označen jako *kořen*. Hrany stromu tím získaly přirozenou orientaci ve směru od kořene k listům. Kořen reprezentuje společného předchůdce všech taxonomických jednotek znázorněných stromem. Každý vnitřní vrchol představuje nejbližšího předchůdce svých potomků. Vnitřní vrcholy přitom obvykle představují hypotetické taxonomické jednotky, které v současnosti nelze pozorovat. Naproti tomu listy stromu zastupují reálné taxonomické jednotky.

Ze zakořeněného stromu je možné kdykoliv zkonstruovat nezakořeněný strom pouhým zrušením označení kořene, opačný postup je možný pouze s dodatečnými informacemi o průběhu evoluce.



Fylogenetický strom života

Porovnání taxonomických jednotek

Při tvorbě fylogenetických stromů se vychází z údajů o podobnosti mezi jednotlivými taxonomickými jednotkami. Existuje mnoho možností, jak tuto podobnost definovat. V poslední době se hojně využívají znalosti z oblasti molekulární biologie. Vycházíme ze sekvencí bází v genomech jednotlivých biologických druhů, případně lze použít i informace o příslušných aminokyselinových a proteinových produktech. Na základě těchto dat je možné určit genetické vzdálenosti mezi jednotlivými dvojicemi taxonomických jednotek. Přesný výpočet této vzdálenosti vyžaduje nejprve vhodné zarovnání porovnávaných DNA sekvencí – tzv. aligning. Jedná se o výpočetně velmi obtížnou úlohu (spadá do třídy NP-úplných úloh), v praxi se proto používá celá řada heuristických metod, které jsou schopny nalézt alespoň suboptimální řešení v přijatelném čase. U zarovnaných sekvencí je možné určit vzdálenost například na základě procenta odlišných bází mezi sekvencemi. Sofistikovanější metody se pokoušejí odhadnout počet mutací, které jsou zapotřebí pro přechod od jedné sekvence k druhé.

Mimo molekulárně biologických dat lze vycházet i z morfologických vlastností zkoumaných taxonomických jednotek. Výpočet vzdáleností v tomto případě závisí na sledovaných znacích a důležitosti, která se jednotlivým znakům přisoudí.

Metody konstrukce fylogenetických stromů

Distanční metody

Tyto metody vycházejí z matice distancí, která udává vzájemné vzdálenosti mezi všemi dvojicemi taxonomických jednotek, pro které konstruujeme fylogenetický strom. Jako vzdálenost se v tomto případě používá genetickou vzdálenost.

UPGMA (Unweighted Pair Group Method with Arithmetic mean)

UPGMA, zjednodušeně Shlukovací analýza, je nejjednodušší algoritmickou metodou konstrukce fylogenetického stromu. Postup je následující:

1. Nalézt v distanční matici nejmenší hodnotu (odpovídá dvojici taxonomických jednotek, které mají k sobě nejblíže).
2. Příslušné taxonomické jednotky sloučit do jedné skupiny a spočítat vzdálenost této nové skupiny ke všem ostatním taxonomickým jednotkám. Vzdálenost taxonomické jednotky T k této nové skupině S se spočítá jako aritmetický průměr vzdáleností mezi jednotkou T a všemi prvky skupiny S. Skupinu S lze dále považovat za hypotetickou taxonomickou jednotku.
3. Pokud máme k dispozici více než jednu taxonomickou jednotku, opakovat postup od 1. kroku.

Znázorníme-li graficky postup shlukování v průběhu popsaného algoritmu, získáme požadovaný fylogenetický strom. Hypotetická taxonomická jednotka, která vznikla jako poslední, je jeho kořenem.

Metoda nejmenších čtverců

V tomto případě konstruujeme všechny možné fylogenetické stromy a hodnotíme, který z nich je nejlepší. Ohodnocení můžeme provést podle následujícího předpisu:

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} (D_{i,j} - d_{i,j})^2,$$

kde $d_{i,j}$ je vzdálenost mezi vrcholy i a j v hodnoceném fylogenetickém stromě a $D_{i,j}$ je vzdálenost mezi odpovídajícími taxonomickými jednotky v distanční matici.

Tento postup vyžaduje konstrukci a ohodnocení všech možných fylogenetických stromů, což je podobně jako aligning NP-úplný problém.

Metoda minimální evoluce

Postup je stejný jako u metody nejmenších čtverců, jednotlivé stromy však porovnáváme podle součtu délek všech větví.

Neighbor-joining

Na začátku se vytvoří jeden hvězdicový strom, kde je jeden vnitřní vrchol, a všechny řešené taxonomické jednotky jsou reprezentovány pomocí listů. Tento strom se postupně rozkládá shlukováním nejbližších taxonomických jednotek tak, aby se v každém kroku co možná nejvíce zmenšila celková délka stromu.

Maximální parsimonie

Metoda maximální parsimonie se snaží nalézt takový fylogenetický strom, který vyžaduje co nejmenší množství evolučních událostí, ke kterým by muselo dojít, pokud by tento strom odpovídal průběhu evoluce. V některých případech se při hodnocení stromů přiřazuje jednotlivým evolučním událostem různá váha, například je-li známo, že některé nukleotidy či aminokyseliny mutují snáze či hůře než ostatní.

V základní variantě tato metoda opět vyžaduje konstrukci všech možných fylogenetických stromů a jejich následné ohodnocení. Pro zefektivnění prohledávání prostoru stromů lze použít například metodu Branch and bound, která je schopná prohledávání omezit pouze na „nadějné“ stromy.

Metoda maximální věrohodnosti

Zde se vychází ze statistických metod a aposteriorní pravděpodobnosti. Snažíme se odhadnout, jaká je pravděpodobnost, že platí statistická hypotéza představovaná konkrétním fylogenetickým stromem pro data, která máme k dispozici. Pro hypotézu H a data D lze tuto pravděpodobnost spočítat takto:

$$P(H|D) = P(H).\frac{P(D|H)}{P(D)},$$

kde P(D|H) je pravděpodobnost, že pozorujeme skutečná data D, za předpokladu, že je hypotéza H pravdivá.

Metoda vyžaduje substituční model, na základě kterého určujeme pravděpodobnost jednotlivých evolučních změn (mutací). Strom, který pro vysvětlení dostupných fylogenetických dat potřebuje těchto změn více, bude mít menší věrohodnost než strom, který si vystačí s menším počtem změn. Mimo toho si všímáme i délek jednotlivých větví.

Odkazy

Související články

- Fylogenetická systematika
- Evoluce

Externí odkazy

- Fylogenetický strom (anglická Wikipedie)

Zdroje

- Vladimír Hampl: Přednáška z molekulární taxonomie (http://web.natur.cuni.cz/~vlada/moltax/)
- Computational phylogenetics (anglická Wikipedie)