Plant Biosystematics Group

Masaryk University
Brno, Czech Republic

PŘÍRODOVĚDECKÁ FAKULTA · MASARYKOVA UNIVERZITA · BRNO

ÚSTAV BOTANIKY A ZOOLOGIE

Bi6589

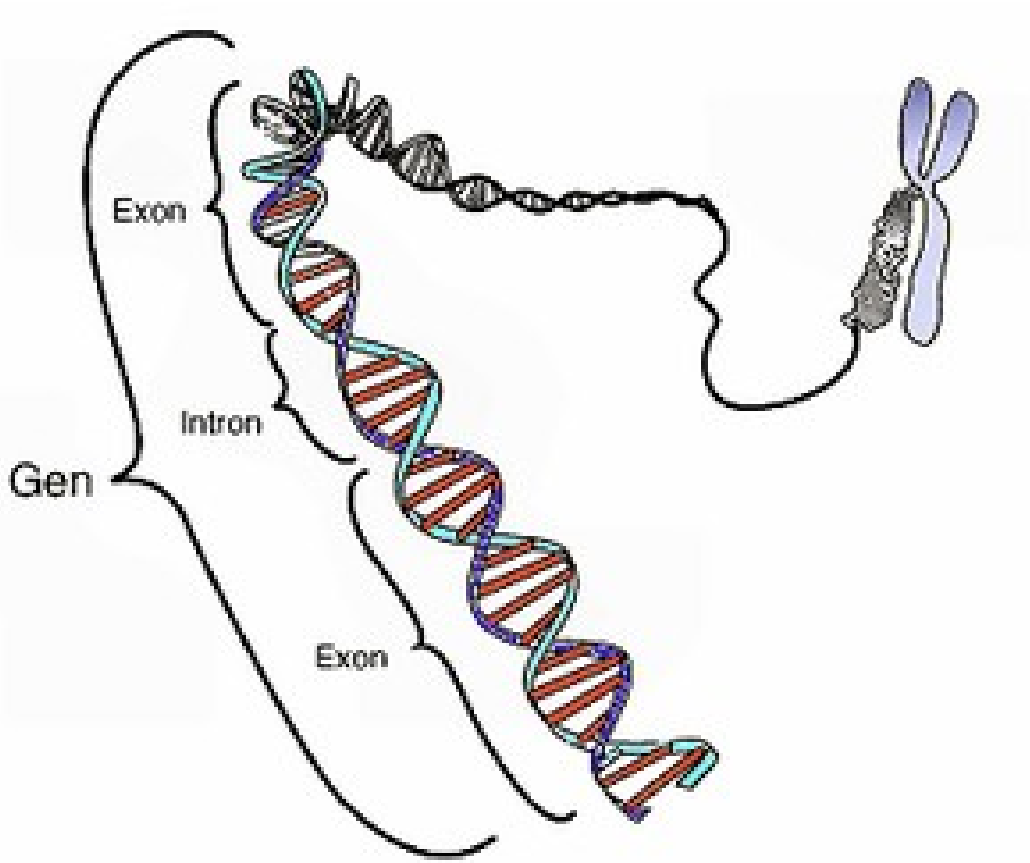
Laboratorní a bioinformatické metody rostlinné biosystematiky

Sekvenační data

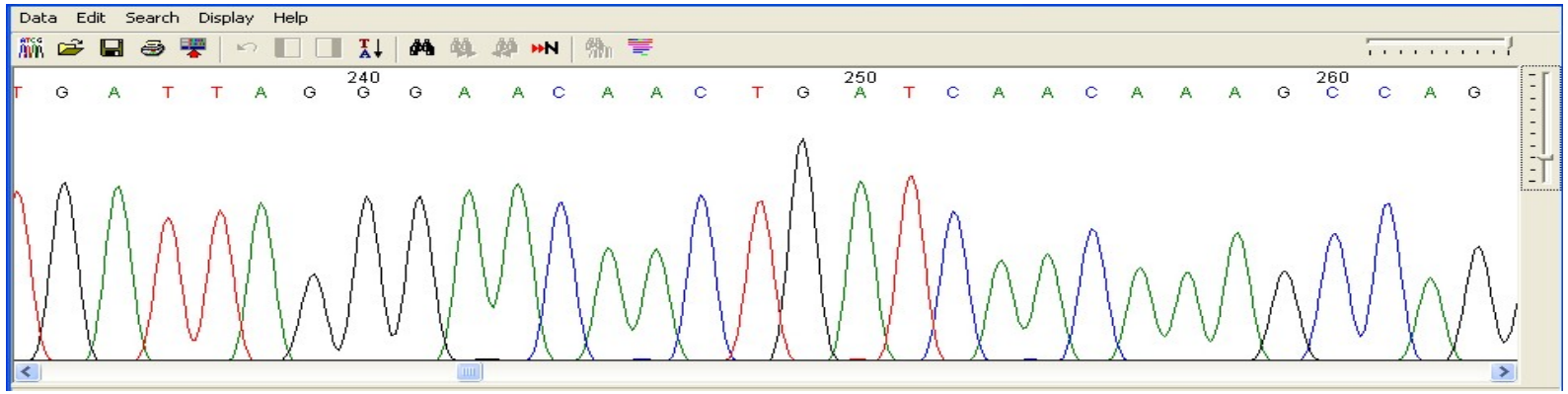
Jednotlivé sekvence – obvykle známe historii a původ

Mikrosatelity – délkový vs. sekvenční polymorfismus

NGS – mnoho sekvencí náhodně vybraných z celého genomu



Sekvence

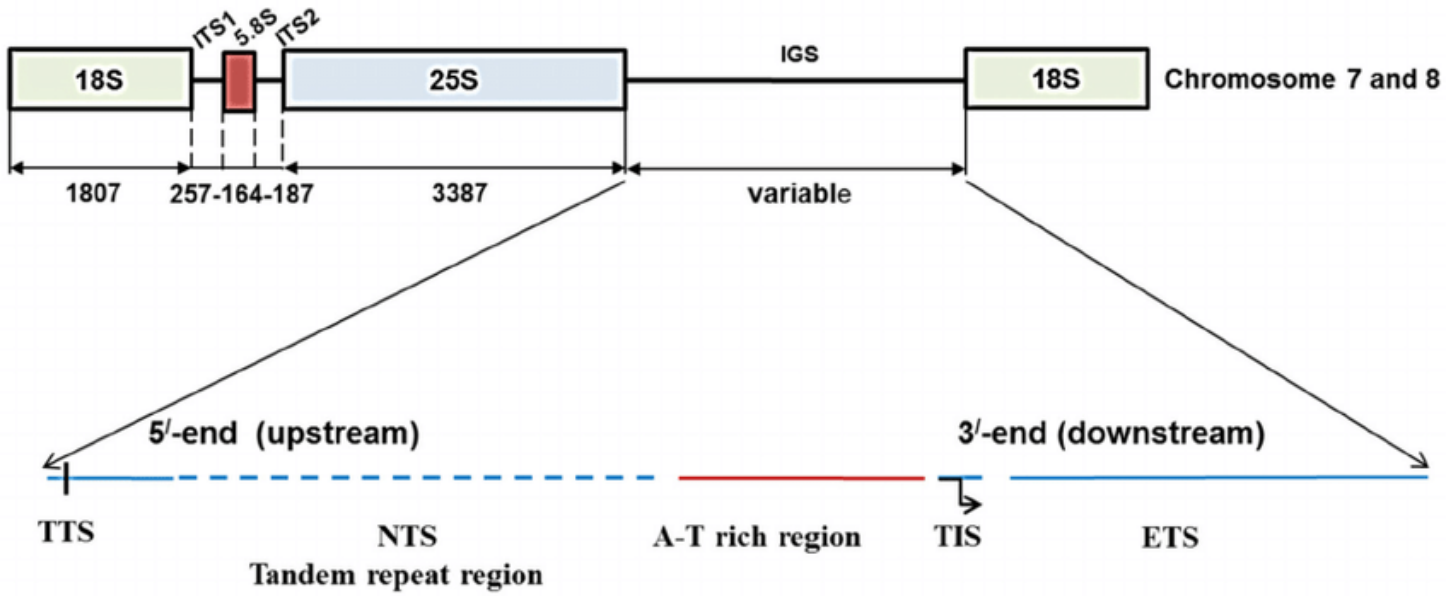

Data Edit Search Display Help
240
250
260
T G A T T A G G G A A C A A C T G A T C A A C A A A G C C A G

Sekvence

Dědičnost!

Jaderné: ITS, ETS, geny (introny)

Chloroplastové: TrnL – TrnF, rbcL, matK, ndhF



Primery obvykle v konzervativních sekvencích – přepisována a studována variabilní (nekódující) část

Sekvence

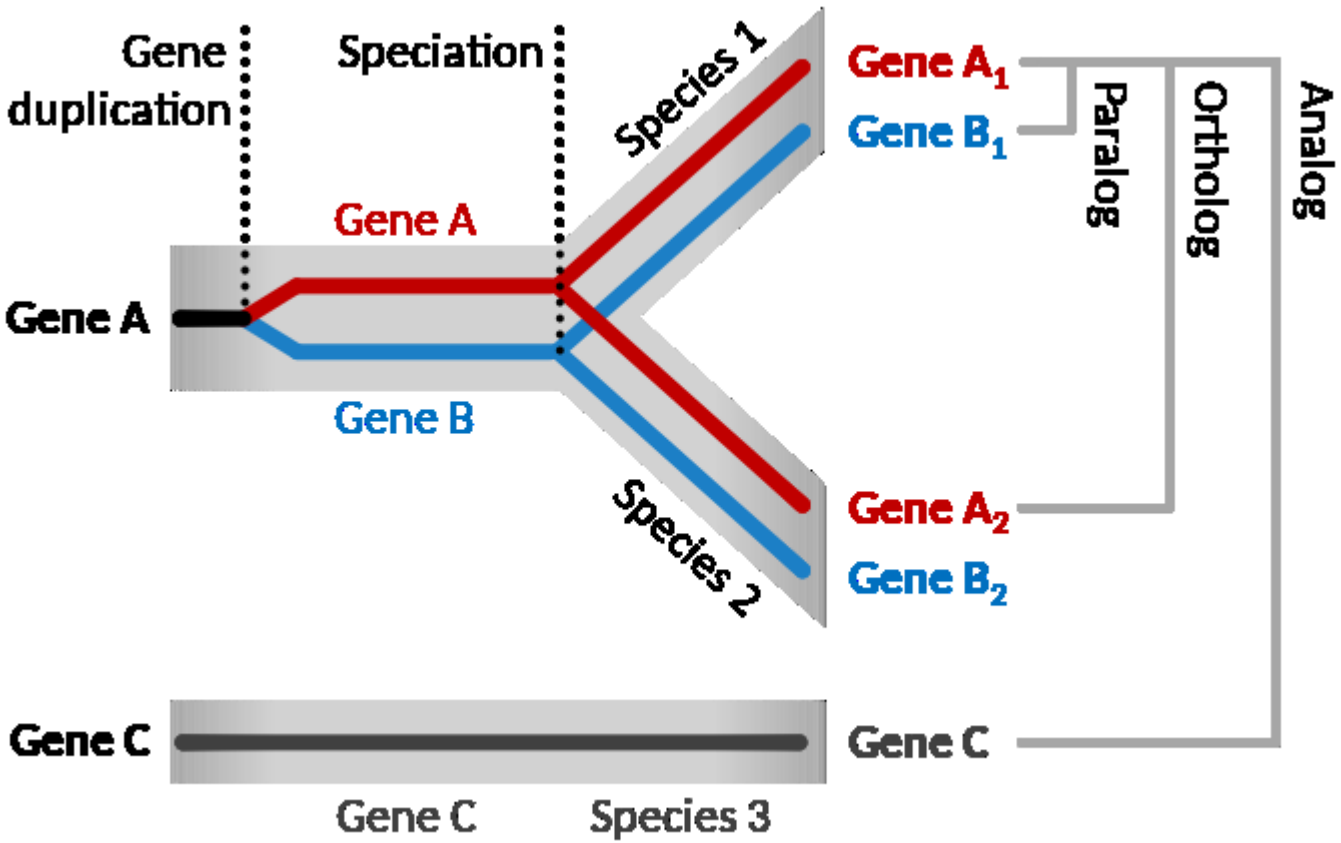
HOMOLOG = dva geny evolučně příbuzné

Dva homologní segmenty DNA mohou mít společný původ na základě:

1) speciační události (mezi druhy+ obvykle mají stejnou funkci) = ORTOLOG
2) duplikace genu (uvnitř jednoho druhu; obvykle jinou funkci) = PARALOG

3) Gen získaný horizontálním přenosem= XENOLOG

ANALOG = Gen s obdobnou funkcí, ale bez společného evolučního předka



Sekvence

HOMOLOG = dva geny evolučně příbuzné

Dva homologní segmenty DNA mohou mít společný původ na základě:

1) speciační události (mezi druhy+ obvykle mají stejnou funkci) = ORTOLOG
2) duplikace genu (uvnitř jednoho druhu; obvykle jinou funkci) = PARALOG

3) Gen získaný horizontálním přenosem= XENOLOG

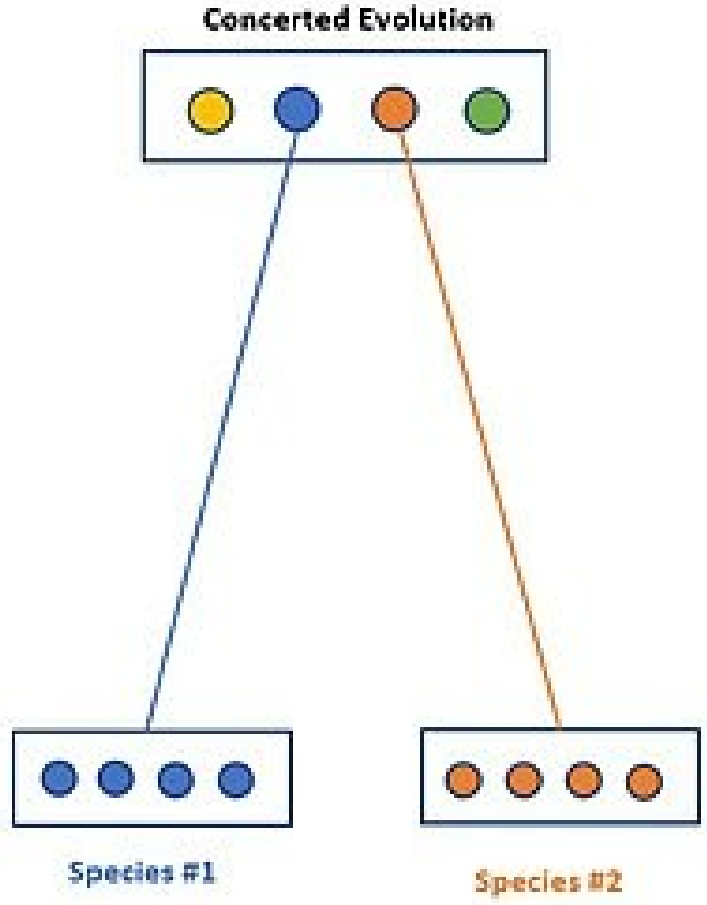
ANALOG = Gen s obdobnou funkcí, ale bez společného evolučního předka

Sekvence

Concerted Evolution

PARADOX: některé jednotky/geny se vyskytují v mnoha kopiích, přesto si udržují podobnou sekvenci (např.: 45S rDNA, 5S rDNA).

Jednotlivé jednotky se nevyvíjejí nezávisle, ale koordinovaně!



Sekvence

Databáze

Databáze NCBI

https://www.ncbi.nlm.nih.gov/

Prohledávání databáze (věrohodnost sekvencí/autorů)

Nahrávání/stažení sekvencí

Porovnávání vaší sekvence s databází atd.

BLAST (Basic Local Alignment Search Tool)

http://blast.ncbi.nlm.nih.gov/Blast.cgi

IUPAC nukleotidové kódy

Nucleic acid code	Common display	Meaning	Mnemonic
A	A	A	Adenine
C	C	C	Cytosine
G	G	G	Guanine
T (or U)	T	T (or U)	Thymine (or Uracil)
R	A/G	A or G	Purine
Y	C/T	C or T (or U)	Pyrimidines
K	G/T	G or T (or U)	Bases which are ketones
M	A/C	A or C	Bases with amino groups
S	C/G	C or G	Strong interaction
W	A/T	A or T (or U)	Weak interaction
B	C/G/T	Not A (i.e. C, G, T or U)	B comes after A
D	A/G/T	Not C (i.e. A, G, T or U)	D comes after C
H	A/C/T	Not G (i.e. A, C, T or U)	H comes after G
V	A/C/G	Neither T nor U (i.e. A, C or G)	V comes after U
N	A/C/G/T	A or C or G or T (or U)	Nucleic acid
. or -	-	Gap of indeterminate length	

Sekvence

Podobnost sekvencí
(distanční matice)

Distanční matice

Zarovnání (alignment) sekvencí

- Vstupní formát (doporučený FASTA)

```
>EU143268.1 Cirsium palustre sequence
GGTGAACCTGCGGAAGGATCATTGTCGAAGCCTGCACAGCAGAACGACCCGTGGACACGTAAT
CACAGCCGGGCGTCGAGGGGGTCGGGCGTCAGCTCGGTGCCCGCGATGCCTCGTCGACGTGC
GTCCATGATGCTTCGTTTTGAAGCGTCGTGGATGTTGCGTCGGCACCTAAACAAACCCCGGCAC
GGCATGTGCCAAGGAAAACAAAA
```

- Kódující vs. nekódující sekvence (= nutnost řešit synonymní x nesynonymní mutace)
- Substituce penalizovány
- Mezery povoleny, ale penalizovány (za otevření mezery, za zvětšování mezery)

=> hledá se nejlepší skóre mezi sekvencemi

Distanční matice

Evoluční vzdálenost (distance) mezi párem sekvencí se obvykle měří počtem nukleotidových (nebo aminokyselinových) substitucí vyskytujících se mezi nimi.

Distance jsou

1) zásadní pro studium molekulární evoluce

2) užitečné pro fylogenetické rekonstrukce a odhad časů divergence.

Distance

a) Nukleotidové substituční modely (nejjednodušší p-distance = počet rozdílů/počet všech nukleotidů; ostatní řeší rychlost různých typů substitucí).
b) Aminokyselinové substituční modely
c) Synonymní a nesynonymní substituční modely

Sekvence

Tvorba fylogenetických stromů

Tvorba fylogenetických stromů

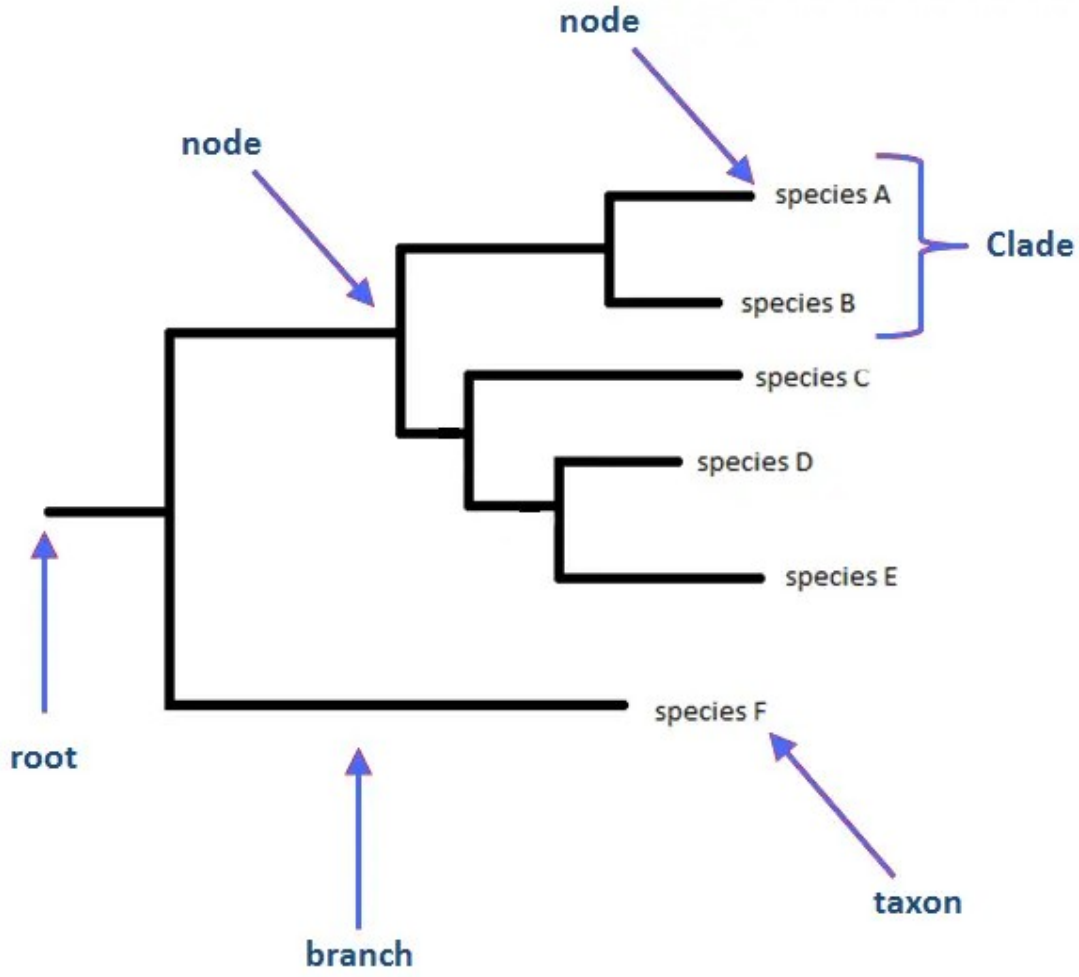
Fylogeneze je historický proces, který popisuje vývoj genů/druhů.

Fylogenetické vztahy genů nebo organismů jsou obvykle prezentovány ve formě stromu.

Kořen stromu = společný předek sdílený všemi taxony; pomáhá ilustrovat evoluční vztahy mezi taxony

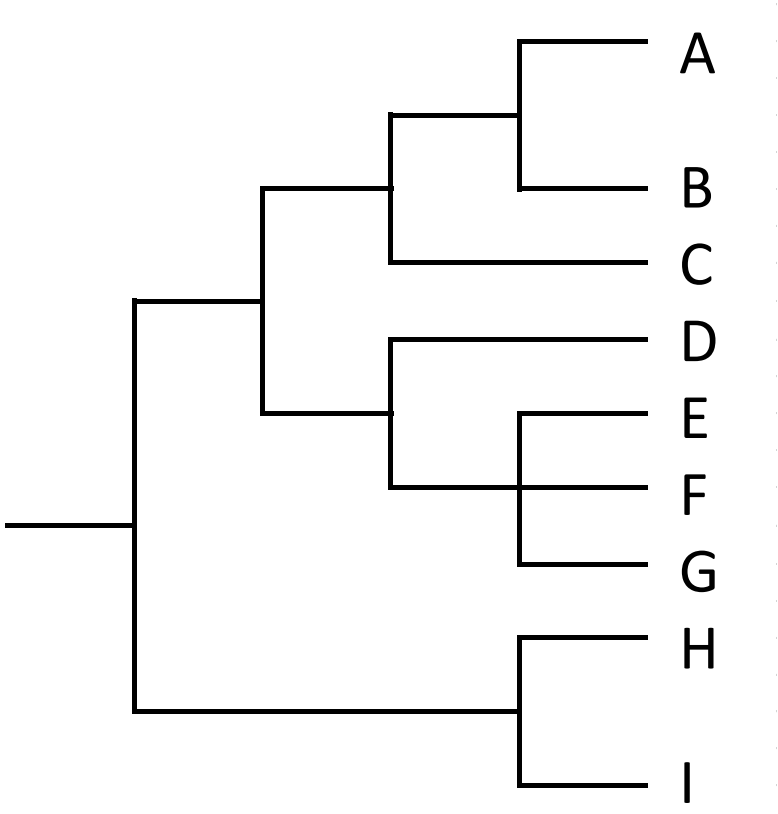
Délka větve = reprezentuje množství evolučních změn

Klady = monofyletické entity složené z předka a všech jeho potomků. Všechny taxony sdílejí jedinečné vlastnosti, které jsou odvozeny od společného předka



Tvorba fylogenetických stromů

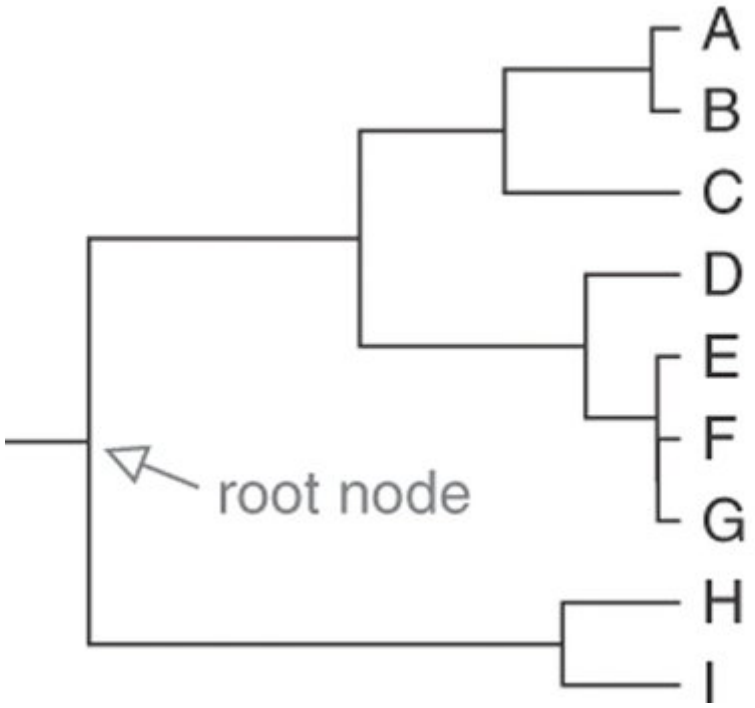
kladogram



Větvení ilustruje pouze evoluční vztahy jednotlivých taxonů, resp. pořadí divergence. Neobsahuje parametr času ani množství změn.

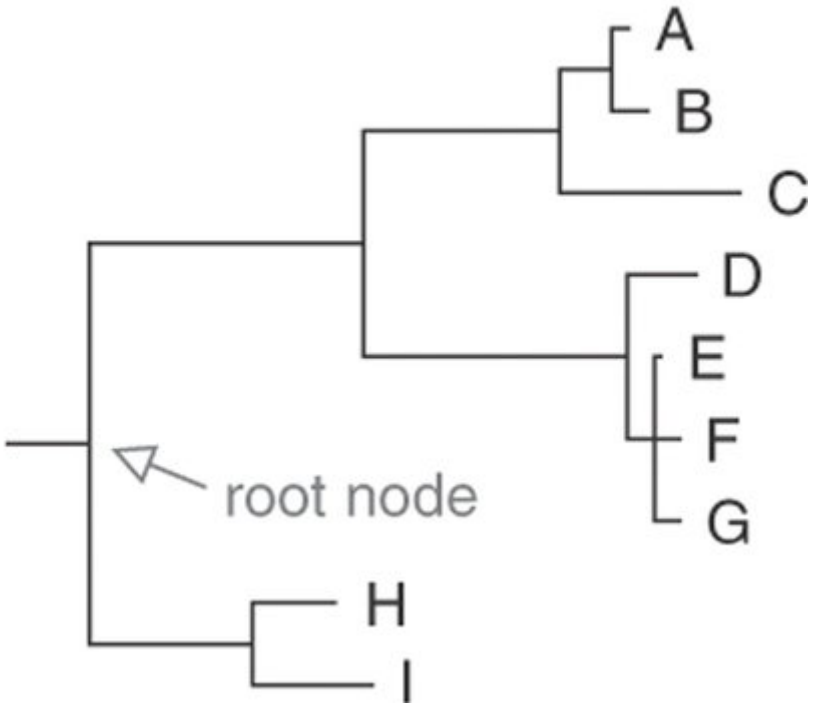
Tvorba fylogenetických stromů

chronogram



V chronogramu jsou délky větví úměrné času a délky cesty od kořene ke špičce jsou stejné.

fylogram



Ve fylogramu odvozeném jsou délky větví úměrné počtu substitucí podél větví a délky cesty od kořene ke špičce jsou obvykle nestejné.

Tvorba fylogenetických stromů

Hierarchické shlukování

- Výpočetně nenáročné
- Postupuje/shlukuje od nejpodobnějších sekvencí k nejméně podobným
- Vyžaduje distanční matici jako zdroj dat = distanční metody: **NJ, UPGMA, Minimum evolution (ME)**

Heuristické shlukování

- algoritmus typicky obsahuje možnost volby pokračování výpočtu, tj. vytvoří se náhodný strom, spočítají se evoluční změny, prohodí se dvě větve, spočítají změny atd. Celé se to opakuje s mnoha stromy a vybere se ten nejlepší.
- Znakové metody = snaží se o minimalizaci počtu změn znakového stavu (**Maximum Parsimony (MP), Maximum Likelihood (ML))** nebo o maximalizaci pravděpodobnosti pozorovaných dat (**Bayesian Inference (BI))**

Bootstrap (stanovení spolehlivosti stromu)

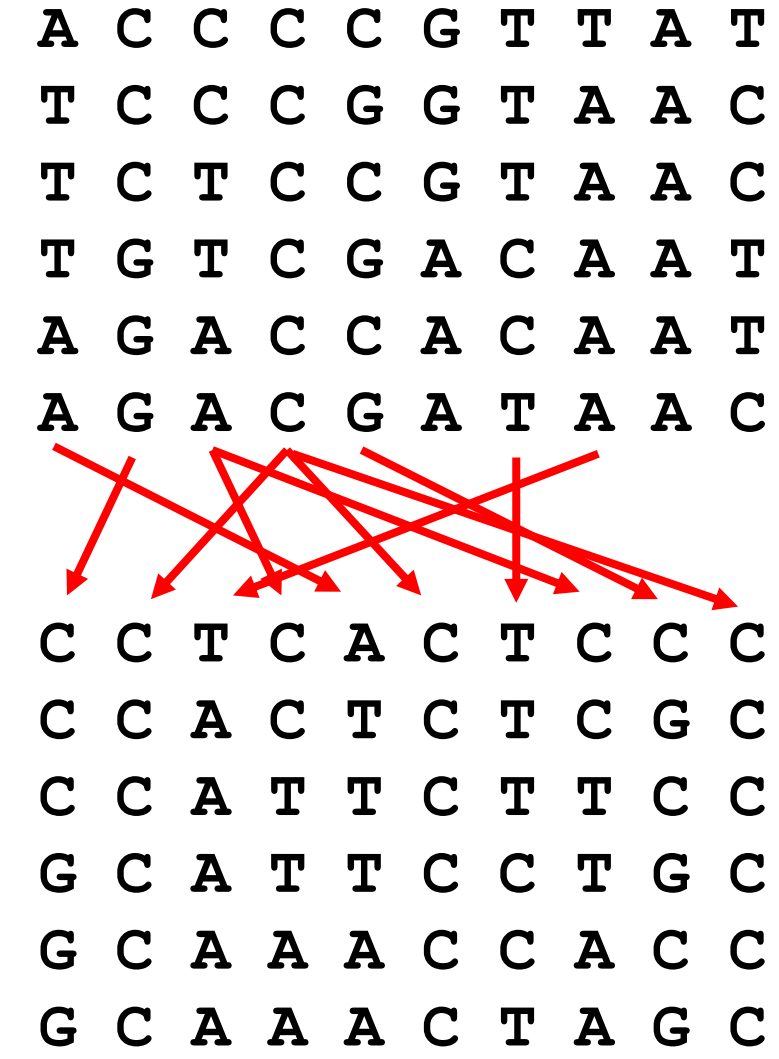
Založeno na převzorkovávání vstupního datasetu

Náhodný (i opakovaný) výběr sloupců

Zadává se počet replikací

Zjišťujeme odlišnosti nových stromů od původního

Bayesian Inference (**posteriorní pravděpodobnosti** místo bootstrapu; vyjadřují míru pravděpodobnosti/spolehlivosti uzlu)



Tvorba fylogenetických stromů

Datace stromu

Kombinace

1) **Molekulárních hodin** = informace o rychlosti mutací, vyjádřená v jednotkách substitucí za rok
2) **Fosilního záznamu**, který poskytuje kalibrační body (víme, kdy došlo ke štěpení taxonů = lokalizace události na časové ose)