MOLEKULÁRNÍ TAXONOMIE - 8

Pro porovnávání topologií by bylo dobré vědět, jaká je pravděpodobnost hypotézy (tedy stromu) při datech (alignmentu), která pozorujeme. Zapisujeme to následovně

P (Hypotézy|Data) = P(H|D)

Bayéský teorém nám říká, že tuto pravděpodobnost spočítáme následovně

P(H|D) = P(H) x P (D|H) / P(D)

P(H).....apriorní pravděpodobnost hypotézy
P (D|H) …… likelihood (věrohodnost) hypotézy = pravděpodobnost, že bychom pozorovali pozorovaná data, pokud by hypotéza byla pravdivá
P(D).....pravděpodobnost dat.

Příklad ze života
O patro výš slyšíte zvuky.
Co to ____ může být?
Kamarád povídá:
„*Máš na půdě skřítky a hrají tam kuželky*".
Vy na to:„*Skřítci jsou jen v pohádkách*".
On na to: „*No jo, ale kdyby tam byli a hráli by kuželky, znělo by to přesně takhle*".
Vy: „*Moment, skočím si raději pro kalkulačku*"

Předchozí znalosti nám říkají, že pravděpodobnost existence skřítků (natož těch, co hrají kuželky) je velmi malá. Proto apriorní pravděpodobnost této hypotézy je malá

P(H) = P(Skřítci co umí hrát kuželky) = velmi malá

Přesto, kdyby skřítci byli a hráli, téměř jistě bychom slyšeli takovéto zvuky, takže likelihood této hypotézy je velký

P(D|H) = P(Slyšet zvuky kdyby skřítci hráli) = velká

Takže pravděpodobnost, že tyto zvuky vydávají skřítci je podle Bayéského teorému (člen P(D) zatím ignorujme)....

P(H|D) = P(H) x P (D|H) = malá x velká = malá
Např. 0,000001 x 1,0 = 0,000001

V tomto případě nás likelihood mohl zmást, pokud bychom si nebyli vědomi Bayéského teorému. Pokud ovšem nemáme žádné informace o apriorních pravděpodobnostech hypotéz, které porovnáváme, pak likelihood P(D|H) není špatným měřítkem pro porovnávání kvality hypotéz. Pokud P(D|H1) > P(D|H2) potom je smysluplné dát přednost hypotéze H1.

Příklad:

Pokud víte, že na půdě je hodně pavouků a kun jsou apriorní pravděpodobnosti hypotéz, že po půdě běhají pavouci a kuny srovnatelné
[$\mathbf{P(H_{pavouci}) \sim P(H_{kuny})}$]

Slyšíte-li zvuky, pravděpodobnost, že byste slyšeli zvuky běhajících pavouků je MENŠÍ než pravděpodobnost, že byste slyšeli zvuky běhajících kun. Jinými slovy, likelihood kun dělajících na půdě hluk je vyšší než likelihood pavouků dělajících hluk.

$\mathbf{P(Zvuky|H_{pavouci}) << P(Zvuky|H_{kuny})}$

Proto dáme přednost hypotéze kun.

Dříve než se pustíme do výpočtů likelihoodů si připomeneme dvě základní pravidla pravděpodobnostních počtů:

Pravděpodobnost, že se stane A a B se vypočítá **P(A a B) = PA x PB**
Pravděpodobnost, že se stane A nebo B se vypočítá **P(A nebo B) = PA + PB**

Maximum likelihood - maximální věrohodnost

Nadále zůstaneme mimo fylogenetiku a ukážeme si, jak likelihood počítat na jednoduchém případu hodu mincí.

Hypotéza: Pravděpodobnost, že při hodu mincí padne panna je 0,4 (p=0,4).

Předpokládejme nyní, že apriorní pravděpodobnost této hypotézy neznáme (nevíme, jak se tato mince chová). Proto potřebujeme data, takže si hodíme několikrát mincí:

PPOOPOPPOOO

Spočítejme likelihood naší hypotézy (p=0,4) pro tato data. Pravděpodobnost, že padne panna, značíme p, pravděpodobnost, že padne orel je tedy 1-p. Pravděpodobnost výše uvedených dat lze poté vyjádřit jako

$p*p*(1-p)*(1-p)*p*(1-p)*p*p*(1-p)*(1-p)*(1-p)=p^5*(1-p)^6$

Likelihood naší hypotézy, že p=0,4, je potom $0,4^5*(1-0,4)^6= 0,00047775744$
Je to tedy hypotéza s nejvyšším likelihoodem? Pokud ne, jaká hypotéza má nejvyšší likelihood? Abychom toto zjistili, musíme její likelihood porovnat s likelihoody jiných hypotéz a najít tu, která ho bude mít nejvyšší. V tomto případě, to znamená hledat vrchol křivky pro funkci $L=p^5*(1-p)^6$. Funkce dosahuje vrcholu pro p=5/11 (0,454545...), likelihood této hypotézy je 0,00051102 a je to nejvyšší hodnota. Hypotéza p=5/11 tedy vysvětluje naše data nejlépe.

V tomto případě bychom to dokázali vyřešit mnohem rychleji selským rozumem. Krása přístupu skrze likelihood je v jeho obecné použitelnosti.

Likelihood pro distance

Pomocí likelihoodu lze počítat genetické distance, správně řečeno porovnávat likelihoody hypotéz o konkrétních hodnotách genetické vzdálenosti. Pravděpodobnosti dat (alignmentu sekvencí) vypočítáme podle nám důvěrně známých substitučních modelů. Pro nejjednodušší model Jukes-Cantora by to vypadalo následovně

Data:

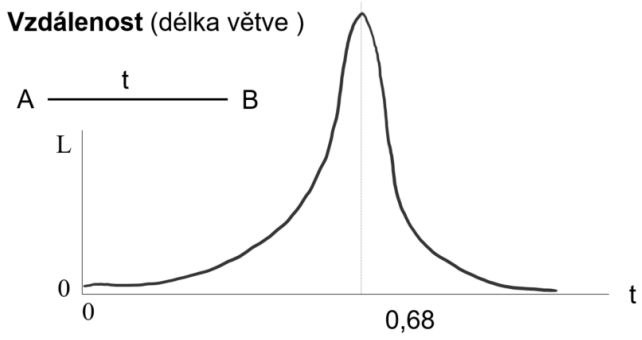
Taxon A	CCCTGG
Taxon B	ACTTGA

Hypotéza:

Sekvence substituují podle Jukes-Cantorova modelu a délka větve t (kdysi ut) je 0,6
Výpočet likelihoodu pro tuto hypotézu by vypadal následovně

$L = P(A|C,t) \times P(C|C,t) \times P(T|C,t) \times P(T|T,t) \times P(G|G,t) \times P(A|G,t) = (¼ - ¼\, e^{-4/3t}) \times (¼ + ¾\, e^{-4/3t}) \times (¼ - ¼\, e^{-4/3t}) \times (¼ + ¾\, e^{-4/3t}) \times (¼ + ¾\, e^{-4/3t}) \times (¼ - ¼\, e^{-4/3t})$

Členy v závorkách pocházejí z modelu Jukes-Cantora, u kterého je pravděpodobnost jakékoli změny rovna $¼ - ¼\, e^{-4/3t}$ a pravděpodobnost zachování stavu $¼ + ¾\, e^{-4/3t}$. Nyní stačí za t dosadit 0,6 a získáme hodnotu likelihoodu. Hledaná genetická vzdálenost bude taková hodnota t, která poskytne nejvyšší likelihood. Takže opět hledáme maximální hodnotu nějaké, tentokrát složitější funkce. Měli bychom dospět ke stejné hodnotě genetické vzdálenosti, jakou získáme, kdybychom použili vzorec, který jsme si představili v 5. přednášce.



Stejně tak můžeme předpokládat, že evoluce sekvencí běžela podle substitučního modelu GTR + Γ, tedy modelu umožňujícího 6 různých rychlostí pro jednotlivé typy substitucí a zároveň předpokládajícího, že pozice alignmentů substituují různě rychle a jejich frekvence rychlostí v alignmentu má rozložení funkce gama rozděleného do 4 diskrétních kategorií. V takovém případě postupujeme obdobně s tím, že pravděpodobnosti pro jednotlivé substituce, které potřebujeme do funkce dosadit jako P(A|C,t) apod. najdeme v substituční matici P(t), kterou spočítáme tak, jak jsme si ukázali v minulé přednášce

$$P(t) = 1/4\ e^{r1Qt} + 1/4\ e^{r2Qt} + 1/4\ e^{r3Qt} + 1/4\ e^{r4Qt}$$

kde Q je rychlostní matice, t je délka větve a r1, r2, r3 a r4 jsou průměrné rychlosti 4 rychlostních kategorií pro pozice v alignmetu. Jejich hodnoty jsou určeny tvarem funkce Γ a ten je dán parametrem α. Na rozdíl od likelihoodové funkce, která nám vznikla při modelu Jukes-Cantor, budou nyní ve funkci přítomny kromě neznámé t, kterou chceme určit hlavně, ještě další neznámé (π, rychlostní α,β,γ,δ,ε,ζ a α parametr funkce Γ). V dřívějších přednáškách jsme si ukázali, že všechny parametry π umíme odhadnout z dat. Jak to uděláme s těmi dalšími? Jelikož se pohybujeme v rámci maximum likelihoodu, tak je odpověď jednoduchá, použijeme takové hodnoty další parametrů, které nám budou maximalizovat hodnotu likelihoodu. Budeme tedy hledat maximum mnoharozměrné funkce s proměnnými t, α,β,γ,δ,ε,ζ a α parametr funkce Γ. Vlastně bychom totéž mohli nebo dokonce měli provést i s hodnotami parametrů π, ale buďme rádi, že jejich hodnotu máme již odhadnutou, protože měnit tolik parametrů naráz s cílem maximalizovat likelihood by bylo obtížné. Teoreticky je to však nejsprávnější řešení v likelihoodovém "světě". Může to znít jako podvod, dosazovat si parametry, které se nám "hodí do krámu". Ale pokud jejich hodnoty neznáme tak, proč by jejich hodnota nemohla být taková, pro kterou je pravděpodobnost pozorovaných dat nejvyšší? Ani v likelihoodovém světě si nemůžeme dělat, co se nám zlíbí. Pořád zůstáváme omezeni pozorovanými daty a "tvarem" funkce, jen její "náplň" hodnoty parametrů, můžeme měnit. Zatímco v případě modelu Jukes-Cantor jsme mohli k výsledku dojít i pomocí vzorce, který jsme si ukázali v 5. přednášce, tak v případě GTR modelu nemáme jinou možnost, než genetickou vzdálenost počítat takto přes likelihood. Žádné vzorce pro model GTR totiž zatím odvozeny nebyly.

Kontrolní otázka: Které t bude lepší? To co bychom vypočítali pomocí Jukes-Cantorova modelu nebo to, které bychom vypočítali podle modelu GTR + Γ? Pokud jste dávali v předchozí pasáži pozor je odpověď jasná, lepší bude to t a ten model, který nás dovede k vyššímu likelihoodu. Když se nad tím zamyslíme hlouběji, přijdeme na to, že Jukes-Cantor představuje speciální případ modelu GTR + Γ, který vznikne tak, že všechny parametry budou mít hodnotu 1. Proto likelihood, který získáme pomocí modelu GTR + Γ nemůže být horší, leda tak stejný, jako likelihood, který nám poskytne model Jukes-Cantor.

Likelihood pro topologie

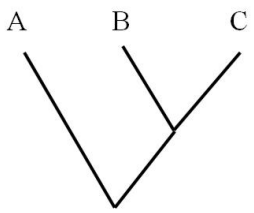
Jak počítat likelihood pro topologie? Ukážeme si to na velmi zjednodušeném příkladu, kdy znaky nabývají jen dvou forem **0** a **1** a na větvích, ať jsou jakkoli dlouhé, platí následující pravděpodobnosti

P0->1 = 0.1 a P0->0 = 0.9
P1->0 = 0.1 a P1->1 = 0.9

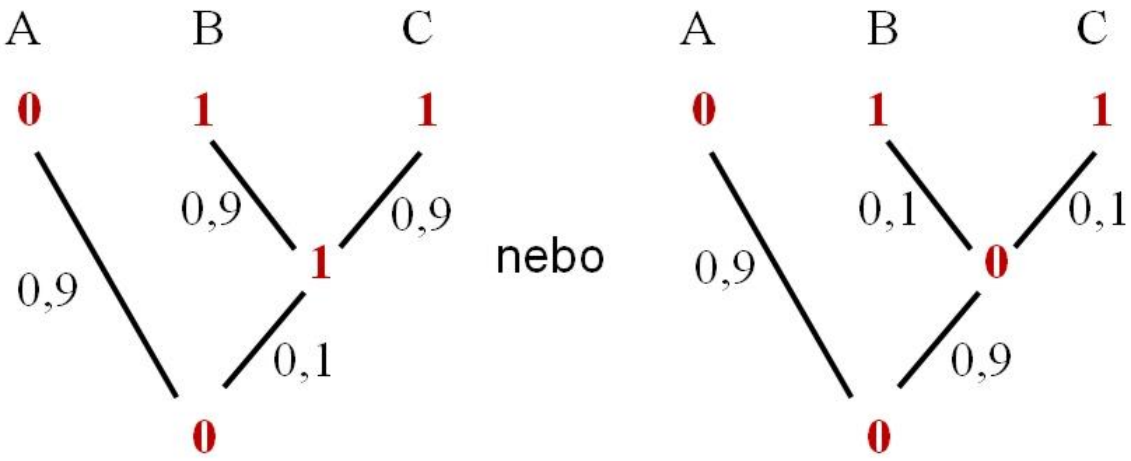
Navíc víme, že předek měl formu znaků 0. Jaká je pravděpodobnost níže uvedeného alignmentu

Druh A 0 0
Druh B 1 0
Druh C 1 0

při zakořeněné topologii



Při výpočtu pravděpodobnosti alignmentu jsme povinni brát v úvahu všechny cesty, kterými se evoluce mohla ubírat, ne jen ty nejpravděpodobnější. Protože jsme si řekli, že společný předek všech měl hodnotu znaku 0, jsou dvě možné cesty, které se liší formou znaku u společného předka B a C.



Na větví jsou uvedeny pravděpodobnosti událostí (změny či zachování stavu) tak, jak jsme si je na začátku stanovili. Spočítejme tedy pravděpodobnost 1. cesty. Ta sestává z následujících událostí (postupujeme zezdola nahoru)

P_{cesta1}= **$P_{0->0A}$ a $P_{0->1BC}$ a $P_{1->1B}$ a $P_{1->1C}$**= 0,9 x 0,1 x 0,9 x 0,9 = 0,0729

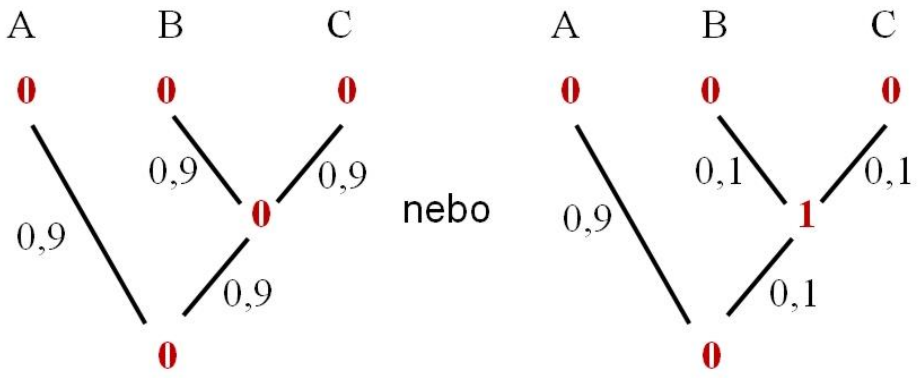
Pravděpodobnost 2. cesty bude obdobně

P_{cesta2}= **$P_{0->0A}$ a $P_{0->0BC}$ a $P_{0->1B}$ a $P_{0->1C}$**= 0,9 x 0,9 x 0,1 x 0,1 = 0,0081

Likelihood této topologie pro pozici 1 alignmentu je

P_{cesta1}+ P_{cesta2}= 0,081

Stejným způsobem spočítáme likelihood topologie pro 2. pozici alignmentu. Opět musíme uvažovat obě cesty



P_{cesta1}= 0,9 x 0,9 x 0,9 x 0,9 = 0,6561
P_{cesta2}= 0,9 x 0,1 x 0,1 x 0,1 = 0,0009

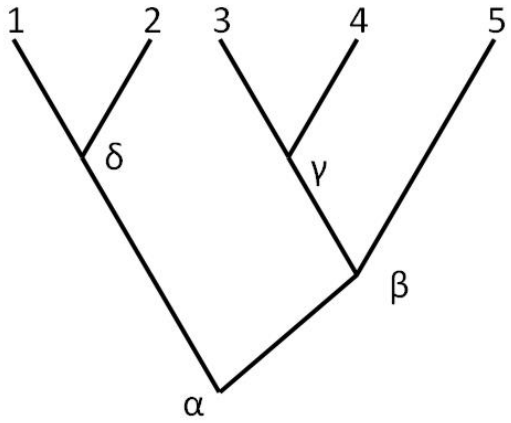
Likelihood tohoto stromu pro pozici 2 je

P_{cesta1}+ P_{cesta2}= 0,657

Likelihood tohoto stromu pro celý alignment je

L1 x L2 = 0,053217

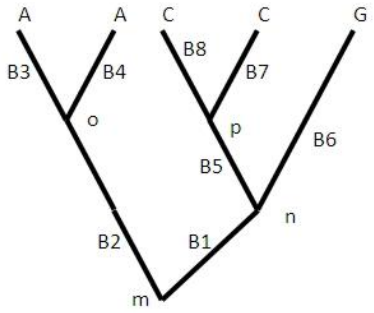
V případě sekvencí se změní to, že máme více forem znaků a že výpočet pravděpodobností je složitější a závisí na délce větve. Obecně vyjádřeno, likelihood níže uvedené topologie pro taxony 1-5 a jednu pozici alignmentu se vypočítá podle formule uvedené pod stromem.



1 A
2 A
3 C
4 C
5 G

$$P(Data \mid T) = \sum_{\alpha} \sum_{\beta} \sum_{\gamma} \sum_{\delta} P(A, A, C, C, G, \alpha, \beta, \gamma, \delta \mid T)$$

Tuto formuli je třeba číst tak, že budeme sčítat pravděpodobnosti pro všechny možné cesty, které vzniknou dosazováním nukleotidů A, C, T a G na vnitřní uzly α,β,γ a δ. V tomto případě existuje celkem 16 takových cest, takže to bude vypadat následovně



$$= P(m = A) \times P(n = A \mid m = A, B1) \times \ldots$$
$$+ P(m = C) \times P(n = A \mid m = C, B1) \times \ldots$$
$$\ldots \; 4^4 \text{ členů!}$$

kde B1-B8 jsou délky větví. No a jak spočítáme pravděpodobnostní členy výše uvedené rovnice? P(m=A) je vlastně πA (frekvence A v sekvencích) no a pravděpodobnosti změn přes každou z větví najdeme v pravděpodobnostní matici P(t), kterou získáme

$$P(t) = e^{-Qt}$$

Za t si musíme dosadit délku patřičné větve (B1...B8). (Všimněte si, že jsem tentokrát z lenosti použil model, který neuvažuje různou substituční rychlost pozic.) Princip zůstává stejný, budeme tak dlouho měnit parametry modelu až dosáhneme nejvyššího celkového likelihoodu pro všechny možné cesty a pozice alignmentu. Jsme omezeni jen tvarem topologie (který určuje "tvar" rovnice) a daty. Všechny proměnné jsou dovoleny měnit. V praxi by to bylo ovšem výpočetně náročné, takže se spokojíme s tím, že si parametry určíme předem a při výpočtu likelihoodu topologie měníme jen délky větví (B1-B8) i to dá docela zabrat.

Jak určíme? parametry předem. Dělá se to tak, že si nějakou rychlou metodou určíte topologii i délky větví, o kterých se domníváte, že nejsou daleko od pravdy. Třeba algoritmickou metodou Neighbour-joining. Tuto topologii a délky větví zafixujete a budete na ní optimalizovat hodnoty parametrů substitučního modelu, t.j. hledat jejich kombinaci, která poskytne nejvyšší likelihood pro danou zkušební topologii. Hodnoty parametrů pak použijete při skórování všech ostatních topologií.

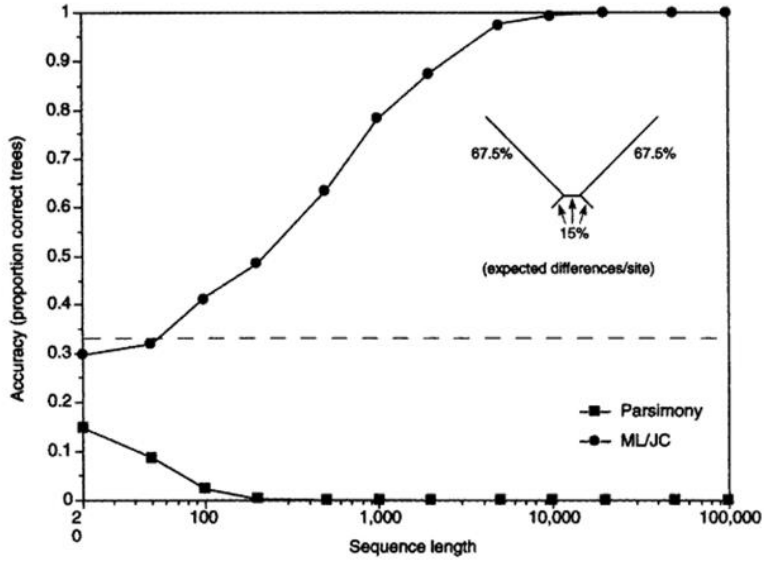
Protože celkový likelihood stromu vzniká násobením čísel menších než jedna, jedná se o velmi malé kladné číslo. Aby se s těmito čísly dobře pracovalo, tak se logaritmují, a protože logaritmus čísla menšího než jedna je záporné číslo a na ty se nehezky kouká, používá se často záporný logaritmus likelihoodu (-lnL). Jestliže u samotného likelihoodu hledáme hypotézu s nejvyšší hodnotou L, pak vyjádřeno v -lnL musíme hledat hypotézu s nejnižší hodnotou -lnL.

Zásadní rozdíly mezi parsimonií a likelihoodem

- V parsimonii jsme brali v potaz pouze nevhodnější stavy na vnitřních uzlech. V likelihoodu musíme uvažovat všechny možnosti a také jej počítáme pro všechny pozice, i ty, co nejsou informativní pro parsimonii.
- Používáme pravděpodobnostní substituční modely, které korigují na substituční saturaci.
- Všímáme si délek větví (ovlivňuje pravděpodobnosti), pokaždé je musíme optimalizovat – to je velmi náročné

Odměnou za tyto složitosti je konzistence likelihoodu. Pokud máme dobře nastavený substituční model (což nikdy nebude 100% pravda) a používáme hodnoty parametrů blízké skutečnosti, pak likelihood, na rozdíl od parsimonie netrpí inkonzistencí a je schopen potlačit artefakt

přitahování dlouhých větví. I když model nebude ideální a parametry nepřesné Felsensteinova zóna pro likelihood bude menší než pro parsimonii.



Bayéská metoda

Bayéská metoda je příbuzná metodě maximum likelihood. Jejím cílem je však určit přímo pravděpodobnost hypotéz a porovnávat hypotézy podle jejich pravděpodobnosti a ne podle likelihoodu. Připomínám, že likelihood není pravděpodobnost hypotézy, ale pravděpodobnost dat pro danou hypotézu. Vzpomeňte si na příklad se skřítky hrajícími bowling a uvidíte, že ne všechny hypotézy, které dokážou dobře vysvětlit naše pozorování jsou pravděpodobné. Některé mohou být doslova "hloupé", protože životní zkušenost nás učí, že něco takového, jako skřítci navíc hrající bowling přece neexistuje. Hypotéza skřítků má, slovy statistika, malou apriorní pravděpodobnost. Konečnou pravděpodobnost hypotézy (tzv. posteriorní pravděpodobnost) nám poskytne Bayéský teorém, se kterým jsme se setkali už na začátku

$$\text{Prob}(H \mid D) = \frac{\text{Prob}(H)\ \text{Prob}(D \mid H)}{\sum_H \text{Prob}(H)\ \text{Prob}(D \mid H)}$$

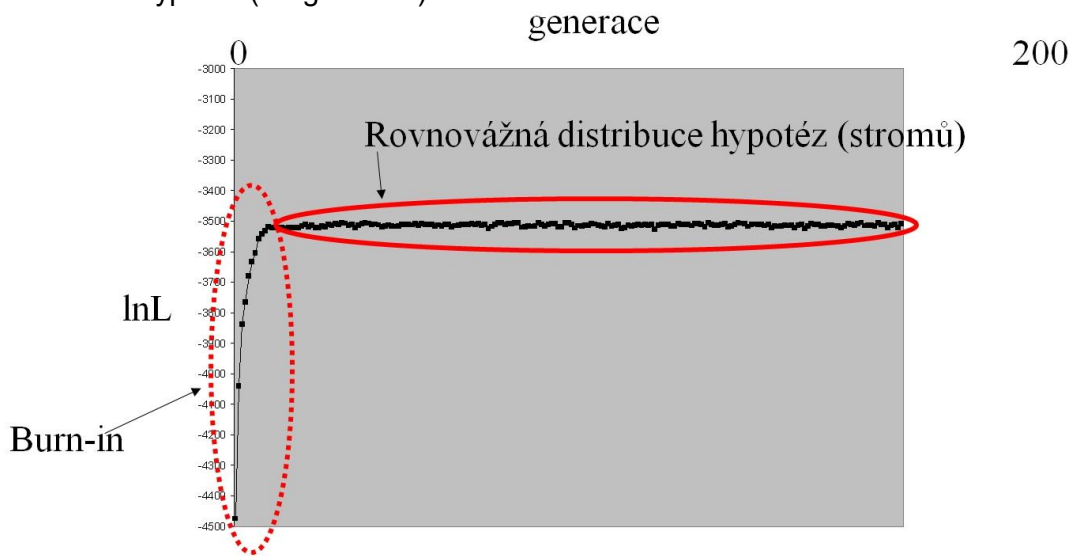
Prob (H) je apriorní pravděpodobnost hypotézy, Prob (D|H) je nám známý likelihood. Tento teorém je ovšem v podobně, jak stojí, prakticky nespočítatelný. Kromě toho, že obvykle neznáme apriorní pravděpodobnost, nedokážeme spočítat jmenovatel - sumu likelihoodu násobeného apriorní pravděpodobností pro všechny možné hypotézy. Bayésští statistici ovšem vymysleli způsob, jak posteriorní pravděpodobnost hypotézy odhadnout. Používají k tomu řetězec zvaný Marcov Chain Monte Carlo (MCMC).

Tento řetězec představuje chůzi prostorem všech možných hypotéz - v našem případě se hypotéza skládá z topologie, délek větví, typu substitučního modelu a hodnot jeho parametrů. Tato chůze prostorem probíhá podle následujícího pravidla. Stojím na hypotéze T1 (pokud je to

první hypotéza tak si ji zvolím náhodně). Podívám se na sousední hypotézu T2 a zvážím, zda na ni přejdu nebo ne. (Sousední hypotézu si řetězec vytvoří změnou topologie nebo délek větví nebo substitučního modelu.) Při této úvaze beru v potaz poměr pravděpodobnosti sousední hypotézy ku pravděpodobnosti té hypotézy, na které se zrovna nacházím. Spočítám si tedy zlomek

$$\frac{\text{Prob}(T2 \mid D)}{\text{Prob}(T1 \mid D)}$$

V tomto zlomku se elegantně zbavím nespočitatelného jmenovatele Bayéského teorému, který je stejný pro oba členy a pokud si řeknu, že apriorní pravděpodobnost všech hypotéz je stejná, tak se zbavím i tohoto členu a výsledek je dán pouze poměrem likelihoodů obou hypotéz. Pokud je výsledný poměr větší než 1, tj. sousední hypotéza má vyšší likelihood, přejdu na tuto hypotézu. Pokud je poměr nižší než jedna, přejdu na sousední hypotézu s pravděpodobností rovnou tomuto poměru, tj. bude-li poměr roven 0,9, tak si vylosuji jedno číslo z 10, pokud si vylosuji 1-9, přejdu na sousední hypotézu, pokud si vylosuji 10, zůstanu, kde jsem. Je zřejmé, že na rozdíl od heuristického algoritmu chůze stromovým prostorem MCMC ochotně a často přechází na horší hypotézy. Není tedy ani zdaleka tak chamtivý. Vlastností MCMC je, že se po nějaké době dostane do místa v prostoru hypotéz, kde se začne pohybovat mezi několika málo hypotézami - říkáme, že konvergoval. Je matematicky dokázáno, že frekvence s jakou během této fáze konvergence navštěvuje hypotézy je odhadem jejich posteriorní pravděpodobnosti - hypotéza, na které stráví nejvíce času má nejvyšší posteriorní pravděpodobnost. To, že se MCMC dostal do rovnovážného stavu, se pozná podle toho, že se přestanou zvyšovat hodnoty likelihoodů hypotéz (viz graf níže)



Dá se to hodnotit také statisticky (o tom více na praktických cvičeních). Pokud se řetězec dostane do rovnovážného stavu získáme posteriorní pravděpodobnost hypotézy tak, že spočítáme její frekvenci ve vzorku hypotéz z rovnovážné distribuce (plný ovál). Hypotéz, které navštívil na cestě k rovnovážnému stavu, si nevšímáme a “upálíme je” - burn in. Bohužel si však

nikdy nemůžeme být jisti, že stav, který považujeme za rovnovážný je skutečně rovnovážný a že, kdybychom řetězec nechali jít ještě dál, by se nestalo tohle...

Jistotu bychom měli jen pokud bychom MCMC nechali jít nekonečně dlouho a na to nikdo z nás nemá čas.
Závěrem bych shrnul že:

- Bayéská metoda je příbuzná metodě maximum likelihood.
- Používá stejné substituční modely na výpočet pravděpodobností.
- Na rozdíl od maximum likelihood se snaží získat posteriorní pravděpodobnost hypotézy a ne jen likelihood. K odhadu posteriorní pravděpodobnosti používá k MCMC.
- Výhodou je, že optimalizuje zároveň topologii, délky větví a hodnoty parametrů substitučního modelu. Čím více parametrů optimalizuje, tím více potřebuje času, než se dostane do rovnovážného stavu. Poznat s jistotou, že MCMC je v rovnovážném stavu nelze.
- Počítá statistickou podporu větvení (o tom příště).