BOOTSTRAP CONSENSUS TREE IN PHYLOGENY RECONSTRUCTION

Karel Sedlář
Bachelor Degree Programme (3), FEEC BUT
E-mail: xsedla74@stud.feec.vutbr.cz

Supervised by: Helena Škutková
E-mail: skutkova@feec.vutbr.cz

Abstract: In recent decades, phylogenetic reconstruction went through huge development coupled with better availability of molecular data. We can expect its wider use in the future due to increasing speed of sequencing of genomic data. At the same time the question how to evaluate the quality of phylogenetic trees raises. Bootstrap appears to be an effective tool for this operation. However there are different approaches in the bootstrap technique and there is no agreement on the interpretation of values it provides.

Keywords: phylogenetics, bootstrap, consensus tree, support

1. ÚVOD

Při konstrukci fylogenetických stromů máme na výběr z celé řady rekonstrukčních metod. Jejich použití je vázáno několika faktory jako výpočetní náročnost, předpokládaná evoluční vzdálenost organismů nebo samotný typ dat, na kterých rekonstrukci stavíme. Jednotlivé metody jsou ve své podstatě matematickými algoritmy, které sestrojí fylogenetický strom na základě jakýchkoliv dat. I takových, jež neobsahují hodnotnou informaci o evolučním vývoji. Velice často také nemáme dostatečné apriorní znalosti pro volbu nejoptimálnějšího algoritmu a různé algoritmy vedou k odlišným fylogramům. Tyto důvody vedou k nutnosti použití statistických testů na každý nově vytvořený fylogenetický strom. Jako nejlepší se prokázaly resamplingové testy a mezi nimi i nejpoužívanější bootstrapping. Ten můžeme použít k tvorbě konsenzuálního stromu, který dovede prezentovat informaci o vývoji s lepší statistickou podporou.

2. BOOTSTRAPPING VE FYLOGENETICE

Bootstrapping byl představen B. Efronem v roce 1979 [1]. Jedná se o test založený na předpokladu nezávisle stejně rozdělených náhodných veličin. Přesně takovými veličinami jsou genomické a proteomické sekvence. Proto se bootstraping hodí na použití ve fylogenetice.

2.1. BOOTSTRAPOVÝ VÝBĚR

Bootstrapping patří mezi vzorkovací statistické testy. Vzorkováním zde rozumíme opakovaný náhodný výběr z původních dat. V prvním kroku je potřeba analyzované sekvence rozdělit na jednotlivé znaky. Toho dosáhneme podélným rozdělením výchozího mnohočetného zarovnání. Na obrázku je vidět rozdělení vzorků čtyř sekvencí nukleotidů na jednotlivých 10 znaků.

OTU\znaky	1	2	3	4	5	6	7	8	9	10
Seq1	G	C	G	A	A	T	C	C	G	A
Seq2	G	C	G	A	C	T	G	C	G	A
Seq3	C	G	G	A	A	G	T	C	G	A
Seq4	C	G	T	A	A	T	T	C	G	A

Obrázek 1: Výchozí vzorek dat rozdělený na znaky.

Nukleotidy jednotlivých OTU (*Operational Taxonomic Unit*) [2] v pozicích nad sebou zůstávají po celou dobu v původním zarovnání. Opakovaným náhodným výběrem znaků se pak vytvoří dostatečný počet pseudovzorků. Ty jsou stejně dlouhé jako vzorek původní, ale některé znaky jsou v něm zastoupeny vícekrát a jiné zase vypuštěny. Samozřejmě je také změněno pořadí znaků. Dva takové příklady vycházející z původního souboru dat jsou k vidění níže.

OTU\znaky	10	5	10	2	3	9	1	9	10	7
Seq1	A	A	A	C	G	G	G	G	A	C
Seq2	A	C	A	C	G	G	G	G	A	G
Seq3	A	A	A	G	G	G	C	G	A	T
Seq4	A	A	A	G	T	G	C	G	A	T

OTU\znaky	5	9	1	9	4	6	5	2	3	7
Seq1	A	G	G	G	A	T	A	C	G	C
Seq2	C	G	G	G	A	T	C	C	G	G
Seq3	A	G	C	G	A	G	A	G	G	T
Seq4	A	G	C	G	A	T	A	G	T	T

Obrázek 2: Bootstrapové pseudovzorky.

Za dostatečný počet je obvykle považováno vytvoření alespoň 500 pseudovzorků. Vzhledem k tomu, že z každého se následně konstruuje předem zvoleným algoritmem nový fylogram, může být toto množství výpočetně neúnosné a je potřeba se spokojit s méně bootstrapovými výběry.

2.2. BOOTSTRAPOVÁ HODNOTA

Výsledkem analýzy je bootstrapová hodnota (*BP bootstrap percentage, bootstrap p-value*) [3]. Jedná se o hodnotu, která vyjadřuje stupeň podpory jednotlivých větvení vzhledem k našim vstupujícím datům. Výsledných hodnot tedy máme tolik, kolik je ve stromu větvení, tj. počet uzlů a umisťujeme je právě k příslušným uzlům. Existují dva přístupy k výpočtu bootstrapové podpory. V prvním případě vezmeme fylogram vytvořený na základě původních dat. A srovnáváme jej s nově vytvořenými stromy za použití stejného konstrukčního postupu. Bootstrapová hodnota je pak procentuální hodnotou počtu stromů obsahujících stejný uzel. Stejným uzlem rozumíme takový, který se dále větví na stejné OTU. Přitom vůbec nezáleží na pořadí a způsobu větvení v rámci daných větví za sledovaným uzlem. Bootstrapovou hodnotu tak můžeme vyjádřit [3]:

$$BP = \frac{s}{n} \cdot 100\% \qquad (1)$$

Kde *s* je počet stromů se stejným uzlem a *n* je počet všech bootstrapových stromů.

2.3. BOOTSTRAPOVÝ KONSENZUÁLNÍ STROM

Druhý přístup bootstrappingu pracuje na základě konsenzuálního stromu vytvořeného ze všech jednotlivých stromů sestrojených z pseudovzorků. Ten přitom plně nahrazuje fylogram sestrojený přímo z originálních dat, který není vůbec konstruován. Tedy výsledkem takového přístupu není pouze BP připsaná k uzlům původního stromu, ale kompletně nový strom včetně uvedených BP pro jednotlivé uzly.

Konsenzuální fylogram kombinuje a vyjadřuje v jednom schématu všechny stromy, z nichž je získán. Konkrétněji bootstrapping využívá většinově konsenzuálního stromu. To je fylogram obsahující uzly, které byly ve výchozích pseudostromech zastoupeny nejvíce. Z toho vychází i BP, ukazující procentuální zastoupení těchto uzlů vztažené k celkovému počtu dílčích bootstrapových stromů. Na rozdíl od prvního přístupu, kdy je bootstrapová hodnota počítána vzhledem k původnímu stromu, je zde vypočítána pouze na základě uměle vytvořených vzorků.

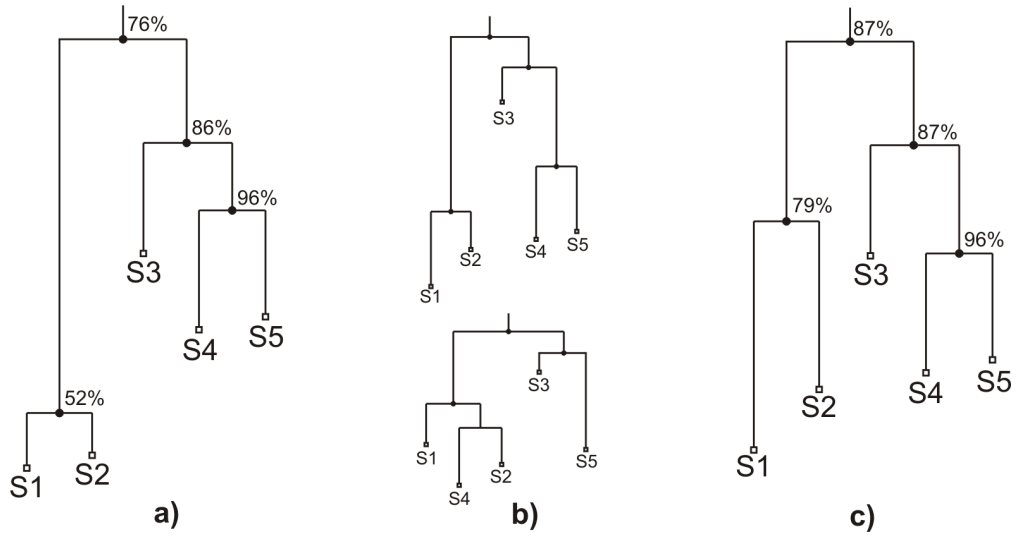
2.4. INTERPRETACE KONSENZUÁLNÍHO FYLOGRAMU

Klasické konsenzuální stromy vyjadřují míru shody mezi fylogramy sestrojenými různými metodami. Odlišné předpoklady jednotlivých metod však zapříčiní, že výsledný strom není vhodný k interpretaci evolučních vztahů. Uplatnění má při srovnávání jednotlivých konstrukčních postupů [4]. Naproti tomu bootstrapový konsenzuální fylogram kombinuje výhradně stromy, jež byly sestrojeny stejným konstrukčním algoritmem, navíc všechny vycházející z jednoho původního souboru dat. Ten je pouze pozměněn náhodným vypuštěním části informací a znásobením jiných. A to nezávisle

v mnoha jednotlivých vzorcích. Proto se dá předpokládat, že jako celek obsahuje informaci celou. Takový konsenzuální strom pak můžeme postavit na místo původního stromu, sestrojeného z výchozího souboru dat a při jeho interpretaci vycházet ze stejných předpokladů.

2.5. PRINCIP METODY

Malá BP původního stromu je způsobena nedostatečným fylogenetickým signálem. Tedy jednotlivé sekvence obsahují málo fylogeneticky informativních pozic. To jsou znaky nesoucí informaci o evolučním vývoji. Podpora uzlů v takových případech klesá i hluboko pod 50%. Z takových výsledků lze vyvodit, že topologie stromu není ideální. Bootstrappingem dosáhneme znásobení znaků informativních i planých. Ovšem znaky informativní budou vykazovat shodu, na základě které sestrojíme konsenzuální strom. Naopak znaky bez informace povedou k mnoha různým topologiím a nezapočítají se tak do výsledného konsenzu. Větvení takového stromu pak mohou být odlišná od větvení ve stromu původním za současného nárůstu podpory jednotlivých uzlů. Délku konsenzuálních větví zjistíme váhovaným součtem délek větví započítaných pseudostromů. Porovnání dvou přístupů bootstrapingu je vidět na následujícím obrázku.



Obrázek 3: Srovnání dílčích bootstrappingových kroků a) Standardně sestrojený strom s vyznačenými BP, podpora uzlů je nižší b) Příklady dvou bootstrapových pseudostromů c) Konsenzuální strom sestrojený bootstrappingem, podpora je vyšší než u prvního stromu.

3. ZÁVĚR

Spojením vlastností konsenzuálních stromů a základních principů bootstrappingu se nám povedlo získat nový pohled na využití bootstrappingu ve fylogenetice. Použitá metoda vede nejen k získání topologie stromu s vyšší statistickou podporou, ale zároveň jsme aplikací bootstrappingu při rekonstrukci stromu omezili vliv jednotlivých nevhodných vstupních vzorků na výsledek analýzy.

REFERENCE

[1] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7(1), 1-26

[2] FLEGR, J. Evoluční biologie. 2. vyd. Praha: ACADEMIA, 2009, IBSN 978-80-200-1767-3

[3] SOLTIS, P.S. Applying the Bootstrap in Phylogeny Reconstruction, Statist. Sci. Volume 18, Issue 2 (2003), 256-267

[4] Forey P. L. (2007) Cladistics: Consensus trees and tree support. Palaeontology Newsletter 64:28–34.