

UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA
ŠTUDIJ BIOTEHNOLOGIJE

Aljaž MAJER

**VERJETNOST APOMORFIJE KOT FILOGENETSKI
OPTIMALNOSTNI KRITERIJ**

DIPLOMSKO DELO

Univerzitetni študij

Ljubljana, 2009

UNIVERZA V LJUBLJANI
BIOTEHNIŠKA FAKULTETA
ŠTUDIJ BIOTEHNOLOGIJE

Aljaž MAJER

**VERJETNOST APOMORFIJE KOT FILOGENETSKI
OPTIMALNOSTNI KRITERIJ**

DIPLOMSKO DELO
Univerzitetni študij

**PROBABILITY OF APOMORPHIC SUPPORT AS A
PHYLOGENETIC OPTIMALITY CRITERION**

GRADUATION THESIS
University studies

Ljubljana, 2009

Diplomsko delo je zaključek Univerzitetnega študija biotehnologije. Opravljeno je bilo na Univerzi v Ljubljani, Biotehniški fakulteti, Oddelku za biologijo, Katedri za zoologijo.

Študijska komisija medoddelčnega dodiplomskega študija biotehnologije je za mentorja diplomskega dela imenovalaizr. prof. dr. TRONTELJ Petra.

Komisija za oceno in zagovor:

Predsednik: prof. dr. Branka JAVORNIK
Univerza v Ljubljani, Biotehniška fakulteta,
Oddelek za agronomijo

Član: izr. prof. dr. Peter TRONTELJ
Univerza v Ljubljani, Biotehniška fakulteta,
Oddelek za biologijo

Član: doc. dr. Jernej JAKŠE
Univerza v Ljubljani, Biotehniška fakulteta,
Oddelek za agronomijo

Datum zagovora: 7. september 2009

Naloga je rezultat lastnega raziskovalnega dela. Podpisani se strinjam z objavo svoje naloge v polnem tekstu na spletni strani Digitalne knjižnice Biotehniške fakultete. Izjavljam, da je naloga, ki sem jo oddal v elektronski obliki, identična tiskani verziji.

Aljaž MAJER

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

ŠD	Dn
DK	UDK 575:8:57.08(043.2)=163.6
KG	Taksična metoda/sinapomorfija/verjetnost apomorfne podprtosti/ simulirana nukleotidna zaporedja
AV	MAJER, Aljaž
SA	TRONTELJ, Peter (mentor)
KZ	SI-1000 Ljubljana, Jamnikarjeva 101
ZA	Univerza v Ljubljani, Biotehniška fakulteta, Študij biotehnologije
LI	2009
IN	VERJETNOST APOMORFIJE KOT FILOGENETSKI OPTIMALNOSTNI KRITERIJ
TD	Diplomsko delo (univerzitetni študij)
OP	XIII, 38, [1] str., 6 pregl., 16.sl, 40 vir.
IJ	sl
JI	sl/en
AI	Utemeljitelj filogenetske sistematike Willi Hennig je predvidel, da je treba za določitev filogenij razločiti med apomorfnimi in pleziomorfnimi stanji znakov, saj lahko le na podlagi prvih sklepamo na skupni izvor organizmov. V tem delu obravnavana filogenetska taksična metoda (PTM) temelji samo na osnovnih Hennigovih načelih. Za mero podpore določenemu kladu uporablja verjetnost apomorfne podpore (probability of apomorphic support - PAS) znakov, ki se izračuna izključno na podlagi frekvenc stanj znaka v kladu ter izven klada. S tem postopkom se izognemo potrebi po upoštevanju dodatnih a priori predpostavk, ki bremenijo preostale filogenetske metode. Namen raziskave je bil pridobiti osnovno poznavanje uporabnosti nove probabilistične taksične metode za filogenetske analize. Preverili smo, če lahko metodo uporabimo za določitev pravega kladograma, in raziskali, kako se filogenetski signal odraža v podpori določeni topologiji drevesa. Želeli smo tudi ugotoviti, kako se metoda obnaša v primerih, ko druge filogenetske metode odpovedo. Ugotovili smo, da metoda PTM v večini primerov izbere napačen kladogram, zato je za ta namen neuporabna. Pri določanju vrednosti PAS posameznim razvejiščem poznanih topologij pa smo ugotovili, da gre za ustrezen način izračuna podpore tem razvejiščem.

KEY WORDS DOCUMENTATION

DN	Dn
DC	UDC 575.8:57.08(043.2)=163.6
CX	Taxic method/sinapomorphy/probability of apomorphic support/ simulated nucleotide sequences
AU	MAJER, Aljaž
AA	TRONTELJ, Peter (supervisor)
PP	SI-1000 Ljubljana, Jamnikarjeva 101
PB	University of Ljubljana, Biotechnical Faculty, Academic Study Program in Biotechnology
PY	2009
TI	PROBABILITY OF APOMORPHIC SUPPORT AS A PHYLOGENETIC OPTIMALITY CRITERION
DT	Graduation Thesis (University Studies)
NO	XIII, 38, [1] p., 6 tab., 16 fig., 40 ref.
LA	sl
AL	sl/en
AB	The founder of phylogenetic systematics Willi Hennig identified the need to distinguish between apomorphic and plesiomorphic character states in order to infer phylogenies, as only the former convey information of a common evolutionary origin of taxons. The probabilistic taxic method (PTM) tested in this experiment is founded only on basic Hennigian principles. Calculation of its measure of clade support by characters, termed probability of apomorphic support (PAS), is based exclusively on the relative frequencies of character states inside and outside the clade of interest. This allows us to avoid the need to consider additional background assumptions typical of other phylogenetic methods. The goal of this experiment was to gather basic knowledge of the applicative value of PTM for phylogenetic analyses. We investigated the ability of PTM to infer the true cladogram. Afterwards, we investigated if and in what manner the method uses phylogenetic signal to gather clade support in predetermined tree topologies. We also tried to determine the performance of PTM in cases when other phylogenetic methods fail. We observed that PTM chooses the wrong cladogram in most cases, therefore it cannot be used in tree inference. Analyses of PAS for individual nodes of predetermined topologies revealed that it is a potentially valid measure of clade support.

KAZALO VSEBINE

	str.
Ključna dokumentacijska informacija (KDI)	III
Key words documentation (KWD)	IV
Kazalo vsebine	V
Kazalo preglednic	VII
Kazalo slik	VIII
Okrajšave in simboli	X
Slovarček	XI
Dodatni material	XIII
1 UVOD	1
1.1 OPREDELITEV PROBLEMA	2
1.2 HIPOTEZA	3
1.3 NAMEN RAZISKAVE	3
1.4 PREDPOSTAVKE V POSTOPKU KLADISTIČNE ANALIZE	4
1.5 RAZLIČNI POGLEDI NA IZVOR RAZLIK V STANJIH ZNAKOV	5
1.6 POGOSTO UPORABLJANE METODE DOLOČANJA IN IZBORA DREVES	5
1.6.1 Metoda največjega verjetja	5
1.6.2 Analiza posteriorne verjetnosti porazdelitve dreves z Bayesovim teoremom (Bayesov pristop)	6
1.6.3 Metoda največje varčnosti	7
1.7 TAKSIČNE FILOGENETSKE METODE	7
1.7.1 Analiza hipotez treh taksonov	7
1.7.2 Probabilistična taksična metoda	7
2 MATERIAL IN METODE	9
2.1 PROBABILISTIČNA TAKSIČNA METODA (PTM)	9
2.1.1 Apomorfni in pleziomorfni signal	9
2.1.2 Verjetnost apomorfne podprtosti (PAS – Probability of apomorphic support)	10
2.1.3 Kriterij izbire najboljšega kladograma	11
2.2 PREVERJANJE USTREZNOSTI PTM KOT METODE ZA ODKRIVANJE ODNOSOV MED TAKSONI	12
2.2.1 Priprava naborov simuliranih nukleotidnih zaporedij	12
2.2.2 Pridobitev filogenetskih dreves, ki najbolje pojasnijo pripravljene nabore nukleotidnih zaporedij	12

2.2.3	Statistično ovrednotenje topologije	12
2.3	PODPORA PTM POSAMEZNIM RAZVEJIŠČEM	
	DOLOČENIH TOPOLOGIJ	13
2.3.1	Priprava naborov simuliranih nukleotidnih zaporedij	13
2.3.2	Izračun PAS na podlagi naborov simuliranih nukleotidnih zaporedij za znane topologije dreves	14
2.3.3	Statistična analiza podatkov	14
3	REZULTATI	15
3.1	USTREZNOST PTM ZA DOLOČANJE TOPOLOGIJ DREVES	15
3.2	PAS KOT MERA PODPORE SKUPIN NA KLADOGRAMU	16
3.2.1	Podpora filogenetske taksične metode posameznim razvejiščem	19
3.2.1.1	Topologija simetričnega drevesa	19
3.2.1.2	Topologija drevesa z dolgimi zunanjimi vejami	21
3.2.1.3	Topologija asimetričnega drevesa	24
3.2.2	Navzkrižna podpora razvejiščem dveh popolnoma različnih topologij dreves	27
3.2.3	Navzkrižna podpora razvejiščem dreves s podobnostmi v topologiji	29
4	DISKUSIJA	30
5	SKLEPI	34
6	POVZETEK	35
7	VIRI	36
	ZAHVALA	

KAZALO PREGLEDNIC

	str.
Preglednica 1: Parametri, po katerih smo pripravili posamezne nabore simuliranih nukleotidnih zaporedij	13
Preglednica 2: Pari naborov simuliranih sekvenc ter topologij dreves, uporabljeni v nadaljnjih analizah	14
Preglednica 3: Povprečne vrednosti podpor testiranih filogenetskih metod za posamezna razvejišča vseh izhodiščnih dreves (N=20; glej Metode 2.3.2). Nizke vrednosti podpor s strani Bayesovega pristopa in ML so podčrtane	17
Preglednica 4: Pregled razlik povprečij PAS in rPAS med posameznimi nabori podatkov, izračunanimi na podlagi topologije A. Izrazito negativne vrednosti so obarvane rdeče, izrazito pozitivne pa modro. Oznake a A, d A in r A pomenijo, da so vrednosti izračunane za topologijo A z uporabo naborov a, d oz. r.	20
Preglednica 5: Pregled razlik PAS in rPAS med posameznimi nabori podatkov, pridobljenih na podlagi topologije B. Izrazito negativne vrednosti so obarvane rdeče, izrazito pozitivne pa modro. Oznake b B, e B in r B pomenijo, da so vrednosti mer izračunane za topologijo B z uporabo naborov b, e oz. r.	23
Preglednica 6: Pregled razlik PAS in rPAS med posameznimi nabori podatkov, pridobljenih na podlagi topologije C. Oznake c C, e C in r C pomenijo, da so vrednosti izračunane za topologijo C z uporabo naborov c, f oziroma r.	26

KAZALO SLIK

	str.
Slika 1: Koreninjeno filogenetsko drevo. Črke označujejo končne taksone A-E, odebeljene pike pa razvejišča. Številke v razvejiščih predstavljajo hierarhične ravni kladograma; razvejišče 2 je hierarhično višje od razvejišča 3 in hierarhično nižje od razvejišča 1. RK – razvejišče s korenino.	1
Slika 2: Postopek izračuna AS (a) in PS (b) v posameznih razvejiščih kladograma.	9
Slika 3: Izhodiščna drevesa, uporabljena za simuliranje naborov nukleotidnih zaporedij: simetrično drevo (A), drevo z dolgimi zunanjimi vejami (B) in asimetrično drevo (C). Oznake S_1 - S_{17} predstavljajo končne taksone, rdeče obarvane številke 1-13 pa posamezna razvejišča. V nadaljnjem besedilu so ta drevesa poimenovana s pripadajočimi oznakami (A, B, C).	13
Slika 4: Primerjava pravilne topologije drevesa (A) in enega od dreves, izbranih s PTM (B). Prikazan je izračun mere podobnosti obeh topologij. Obe drevesi se razlikujeta le v enem razvejišču; skladnost topologij obeh dreves znaša 3. Števila v razvejiščih spodnjega drevesa so vrednosti PAS teh razvejišč.	15
Slika 5: Škatlični diagrami porazdelitve skladnosti topologij topologij dreves, izbranim z vsako od uporabljenih filogenetskih metod, izrisani na podlagi naborov sekvenc s 5000 (A), 10000 (B), 25000 (C) in 50000 (D) znaki (v vseh primerih za vsak par PTM-ML in PTM-MP velja: $p \gg 0,01$).	16
Slika 6: Porazdelitev podpore PAS (A) in rPAS (B) posameznim razvejiščem simetričnega drevesa (glej Slika 3A) s strani na podlagi istega drevesa pripravljenih naborov simuliranih sekvenc a (modro, $N=100$) in d (rdeče, $N=100$) ter naborov naključno generiranih sekvenc r (črno, $N=100$; glej Preglednico 2). Vrednost podpore s strani enega nabora sekvenc danemu razvejišču predstavlja točko na grafu. Vse razlike med skupinami na posameznih razvejiščih so statistično značilne ($p < 0,01$).	20
Slika 7: Simetrično drevo A z označenimi razvejišči s pozitivno (modro) in z negativno (rdeče) razliko $rPAS(a A) - rPAS(r A)$ (glej Preglednico 4). Drevo je simetrično in ima zelo enakomerne dolžine posameznih vej.	21

- Slika 8: Porazdelitev podpore PAS (A) in rPAS (B) posameznim razvejiščem dolgovejnatega drevesa (glej Slika 3B) s strani na podlagi istega drevesa pripravljenih naborov simuliranih sekvenc b (modro, N=100) in e (rdeče, N=100) ter naključno generiranih naborov sekvenc r (črno, N=100; glej Preglednico 2). Vrednost podpore enega nabora sekvenc danemu razvejišču predstavlja točko na grafu. Vse razlike med skupinami na posameznih razvejiščih so statistično značilne ($p < 0,01$). 22
- Slika 9: Dolgovejnato drevo B z označenimi razvejišči s pozitivno (modro) in z negativno (rdeče) razliko $rPAS(b|B) - rPAS(r|B)$. Drevo ima izrazito dolge zunanje veje in kratke notranje veje. 24
- Slika 10: Porazdelitev podpore PAS (A) in rPAS (B) posameznim razvejiščem asimetričnega drevesa (glej Slika 3C) s strani na podlagi istega drevesa pripravljenih naborov simuliranih sekvenc c (modro, N=100) in f (rdeče, N=100) ter naključno generiranih sekvenc r (črno, N=100; glej Preglednico 2). Vrednost podpore enega nabora sekvenc danemu razvejišču predstavlja točko na grafu. 25
- Slika 11: Asimetrično drevo C z označenimi razvejišči s pozitivno (modro) in z negativno (rdeče) razliko $rPAS(c|C) - rPAS(r|C)$. 26
- Slika 12: Kladogram R. 27
- Slika 13: Porazdelitev podpore rPAS, izračunanih na podlagi topologij drevesa A (modro) in popolnoma drugačnega, naključno pridobljenega drevesa R (rdeče) s strani nabora sekvenc a. Opozorilo: enako oštevilčena razvejišča ne predstavljajo istih razvejišč pri obeh drevesih. 28
- Slika 14: Škatlična diagrama mer podpore topologije PAS (A) in rPAS (B) po celotnih topologijah drevesa A ter naključno pridobljenega kladograma R (za oba grafa velja: $p < 0,01$). 28
- Slika 15: Porazdelitev podpore rPAS, izračunanih na podlagi topologij drevesa A (modro) in drevesa C (rdeče), z uporabo nabora sekvenc c. Opozorilo: enako oštevilčena razvejišča ne predstavljajo istih razvejišč pri obeh drevesih. 29
- Slika 16: Škatelni diagram $rPAS(c|A)$ in $rPAS(c|C)$ ($p < 0,01$). 29

OKRAJŠAVE IN SIMBOLI

AS	apomorfni signal
BS	bootstrap podpora (bootstrap support)
ML	standardna metoda največjega verjetja (maximum likelihood)
MP	standardna metoda največje varčnosti (maximum parsimony)
PAS	celokupna verjetnost apomorfne podprtosti
pas_j	verjetnost apomorfne podprtosti za znak j
$PAS(a A)$	verjetnost apomorfne podprtosti nabora sekvenc a za topologijo A (ustrezen simbol se uporablja tudi za druge pare nabor-topologija)
PP	posteriorna verjetnost (posterior probability)
PS	pleziomorfni signal
PTM	filogenetska taksična metoda
rPAS	relativna verjetnost apomorfne podprtosti

SLOVARČEK

Apomorfija	Stanje znaka, podedovano od skupnega prednika in značilno za končne taksonne danega klada.
Bayesov pristop	Filogenetska metoda, ki vrednoti hipoteze sorodnosti na podlagi posteriorne verjetnosti
Bootstrap podpora (BS)	Mera podpore razvejiščem filogenetskega drevesa, ki jo dobimo s kvantificiranjem prisotnosti razvejišč na drevesih, dobljenih na podlagi naključno tvorjenih zaporedij izhodiščnih znakov.
Filogenetski signal	Razporeditev stanj znaka med taksoni, ki je posledica dejanske evolucije taksonov in ne naključnih pretvorb teh stanj.
Klad	Skupina, ki združuje vse potomce nekega predvidenega skupnega prednika; sinonim za monofiletsko skupino.
Kladistika	Veja biološke sistematike, ki organizme razvršča v skupine na podlagi skupnega evlucijskega izvora
Kladogram	Filogenetsko drevo, ki prikazuje le odnose med taksoni, filogenetskih razdalj pa ne
Kriterij varčnosti	Filozofski pristop, ki daje najbolj enostavni razlagi nekega problema prednost pred bolj kompleksnimi razlagami.
Metoda največje varčnosti	(Maximum Parsimony; MP) filogenetska metoda, ki na podlagi razporeditve stanj znakov izbere hipotezo sorodstvenih odnosov, ki predvideva najmanj sprememb stanj znakov.
Metoda največjega verjetja	(Maximum Likelihood; ML) filogenetska metoda, ki izbere hipotezo sorodstvenih odnosov z največjim verjetjem glede na izbrani model evolucije
Model evolucije	Sistem predpostavk o poteku evolucije nukleotidnih zaporedij. Podaja npr. hitrosti pretvorb med nukleotidi in končno pogostnost posameznih nukleotidov v sekvencah.
Nabor nukleotidnih zaporedij	Datoteka, ki vsebuje nukleotidna zaporedja vseh analiziranih končnih taksonov.
Pleziomorfija	Predniško stanje znaka.
Posteriorna verjetnost (PP)	Verjetnost opažene razporeditve stanj znakov v končnih taksonih, če privzamemo resničnost nekega filogenetskega drevesa.
Sinapomorfija	Stanje znaka, prisotno pri dveh opazovanih sorodnih taksonih in njunem najbližjem skupnem predniku.
Šum	Razporeditev stanj znaka, ki je posledica naključnih sprememb stanj znaka (homoplazij)
Taksična metoda	Filogenetska metoda, ki temelji na vrednotenju hipotez skupin končnih taksonov.

Topologija	Vzorec razvejenosti filogenetskega drevesa, torej oblika, ki jo določa lega razvejišč na drevesu.
Transformacijska metoda	Filogenetska metoda, ki temelji na vrednotenju hipotez pretvorb med posameznimi stanji znaka.

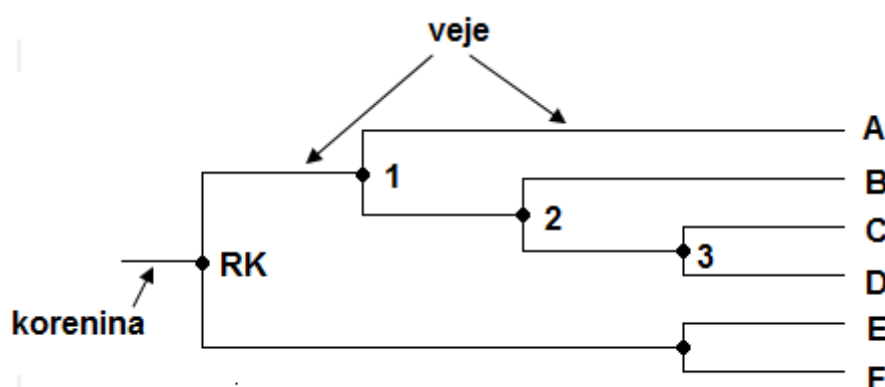
DODATNI MATERIAL

Vhodne datoteke z nabori simuliranih nukleotidnih zaporedij za vse v tem delu uporabljene računalniške programe so dostopne na spletnem mestu:

http://web.bf.uni-lj.si/jjakse/dn_majer_aljaz_vhodne.rar

1 UVOD

Filogenetika je znanstvena disciplina, ki na podlagi morfoloških, molekularnih ali drugih podatkov proučuje sorodstvene odnose med organizmi. Cilj filogenetske analize je poznavanje filogenij, t.j. verjetnega poteka evolucije opazovanih skupin organizmov (taksonov) od najbližjega skupnega prednika dalje (Wägele, 2005). Sorodstvene odnose med taksoni navadno ponazorimo s filogenetskimi drevesi (dendrogrami, Slika 1), med katerimi sta največkrat uporabljena **kladogram** (prikazuje le sorodstvene odnose) in **filogram** (prikazuje sorodstvene odnose in evolucijsko razdaljo med taksoni; Page in Holmes, 1998).



Slika 1: Koreninjeno filogenetsko drevo. Črke označujejo končne taksoni A-E, odebeljene pike pa razvejišča. Številke v razvejiščih predstavljajo hierarhične ravni kladograma; razvejišče 2 je hierarhično višje od razvejišča 3 in hierarhično nižje od razvejišča 1. RK – razvejišče s korenino.

Filogenetske metode tradicionalno temeljijo na prepoznavanju homologij med organizmi. Določen **znak** (t.j. kakršna koli lastnost organizma, ki se lahko obravnava kot spremenljivka z vsaj dvema nezdružljivima stanjema; DeSalle, 2006) neke skupine organizmov je **homologen**, če je bil prisoten pri skupnem predniku vseh organizmov; nasprotno temu pa imajo **analogni** znaki enaka stanja, čeprav so podedovani od različnih predniških organizmov (Phillips, 2006). Dva filogenetska znaka sta si lahko podobna, vendar ne homologna, ali pa zelo različna, a vseeno homologna; važen je le izvor znaka, ne njegovo stanje. Stanje znaka, ki je podedovano od skupnega prednika in značilno za končne taksoni danega klada, imenujemo **apomorfija**, predniško stanje znaka pa **pleziomorfija**. Posebno obliko homologije, pri kateri je pri dveh taksonih neko stanje znaka prisotno v enaki obliki kot pri najbližjem skupnem predniku, pri katerem se je ta oblika tudi prvič pojavila, imenujemo **sinapomorfija**.

Že na začetku filogenetske analize se je treba odločiti, kaj predstavljajo znaki neke skupine taksonov. Homologiji, opaženi pri vnaprej prepoznanih znakih, pravimo primarna homologija (Phillips, 2006), po analizi celotnih naborov znakov in določitvi odnosov med taksoni na podlagi podanega optimalnostnega kriterija pa razpoznamo sekundarno homologijo (de Pinna, 1991). Če se stanja nekega znaka ne skladajo s topologijo filograma, lahko analizo ponovimo tudi brez upoštevanja teh znakov.

Kladistika je veja filogenetike, katere metode temeljijo na principu varčnosti (t.j. za določitev filogenije uporabljajo najmanjše možno število predpostavk; Phillips, 2006). Kladistična metoda kot najbolj ustreznega izbere tisti kladogram, ki predvideva najmanjše število sprememb med različnimi stanji posameznih znakov (najbolj »varčno« združi taksoni v monofilije).

Z **računalniškimi simulacijami** procesa evolucije iz enega predniškega zaporedja po danem evolucijskem modelu lahko za vse končne taksoni pridobimo nabor nukleotidnih sekvenc, ki ustrezajo topologiji in dolžinam vej filograma, na podlagi katerega je ustvarjen (Page in Holmes, 1998). Pri izboru modela evolucije je potrebna previdnost. Če je model evolucije, uporabljen pri simulacijah, preveč podoben modelu evolucije, ki ga uporablja na simuliranih podatkih testirana filogenetska metoda, pridemo do pristranskih zaključkov (Wägele, 2005). Analiza simuliranih naborov nukleotidnih sekvenc omogoča ovrednotenje filogenetskih metod na podlagi natančno poznanih odnosov med analiziranimi taksoni. Primerjava izhodiščnega filogenetskega drevesa z drevesom, ki ga izberejo na simuliranih sekvencah izvedene filogenetske metode, nam omogoča statistično ovrednotenje zanesljivosti teh metod (Simmons in Webb, 2006).

1.1 OPREDELITEV PROBLEMA

Kladistična analiza razvršča organizme v skupine na podlagi sinapomorfij, t.j. izpeljanih stanj znakov, ki so skupni pripadnikom neke monofilije in najbližjemu skupnemu predniku, pri drugih taksonih pa jih ni. Bang in sod. (2000) razlikujejo dva glavna pristopa h kladistični analizi: taksičnega in transformacijskega. Rezultat uporabe taksičnega pristopa so skupine taksonov. Če analiziramo štiri taksoni z ozirom na neko stanje znaka in opazimo pri A in B stanje znaka X, pri taksonih C in D pa stanje znaka X', je po taksičnem pristopu najbolj verjetno drevo ((A,B),(C,D)). Transformacijski pristop pa se sploh ne ukvarja s hipotezami skupin; iz prejšnjih podatkov dobimo tezo o pretvorbi $X \leftrightarrow X'$. (Carine in Scotland, 1999). Transformacijski pristop se lahko uporabi kot soliden vir hipotez o sorodstvenih odnosih, vendar pa je treba za potrditev teh hipotez uporabiti taksični pristop (Bang in sod., 2000). Vseeno pa je večina filogenetskih metod v uporabi danes transformacijskih. Med diskretnimi filogenetskimi metodami so standardna metoda največjega verjetja, standardna metoda največje varčnosti in metode, ki uporabljajo Bayesov pristop, eksplicitno transformacijske (Scotland, 2000).

Več avtorjev se strinja, da je kladistična metoda boljša od filogenetskih metod, ki temeljijo na predpostavkah poteka evolucije (Farris in sod., 1970; Brady, 1985; Brower, 2000). Zato je po njihovem mnenju pomembno, da se iz kladistike v čim večji meri odstrani predpostavke o poznavanju poteka evolucije v opazovanih skupinah.

Že Hennig (1966) je prepoznal tri vzroke za podobnost znakov – sinapomorfije (znaki, značilni za najbližnjega skupnega prednika taksonov), simpleziomorfije (stanja znakov, značilna že za bolj oddaljene skupne prednike) in homoplazije (naključno enaka stanja znakov pri dveh nesorodnih kladih) – in opazil, da samo sinapomorfije nosijo signal o filogenetskih odnosih med taksoni. Razvil je metodo združevanja taksonov samo po sinapomorfijah, ki je uporabljala določene predpostavke evolucije. Platnick (1979) in

Nelson (1989) sta pokazala, da grupiranje taksonov izključno po sinapomorfijah ni nujna predpostavka kladističnih metod, čeprav nekateri avtorji to še zmeraj trdijo (e.g. Siddall in Kluge 1997).

Princip varčnosti (parsimony) razlaga, da je treba pri interpretaciji opažanj uporabiti najmanjše možno število predpostavk, ki še zadostuje za razlago podatkov. Vsako dodatno predpostavko je treba zavreči, saj otežuje ločitev pravilne razlage od drugih razlag, ki slabše pojasnijo opažanja. Popper (1959, cit. po Brower 2000) si je razlagal uporabo principa varčnosti v znanstvenih metodah kot izogibanje vsem odvečnim predpostavkam, razen če povečajo zanesljivost preverjanja osrednje hipoteze.

Več avtorjev je že poskusilo postaviti najbolj varčen nabor osnovnih hipotez. Wiley (1975) je iz Hennigovih načel kladistike kot prvi izpeljal naslednje aksiome kladistike: (1) evolucija se odvija, (2) obstaja samo ena filogenija vseh organizmov, (3) znaki se lahko spremenjeni ali nespremenjeni dedujejo iz predhodne v naslednjo generacijo. Wileyevi aksiomi so premalo natančni za uporabo v kladistiki (Brower, 2000). Osnovne temeljne predpostavke kladistike so predlagali še Farris in sod. (1970), Platnick (1979), Kluge (1997) in Brower (2000). Slednji je predlagal naslednje aksiome kladistike: (1) osnova kladističnih analiz so opažene razlike med stanji znakov, (2) kladogram je ustrezen način prikazovanja odnosov med taksoni in (3) pri postavljanju predpostavk je treba zadostiti principu varčnosti (Brower, 2000).

1.2 HIPOTEZA

Testirana metoda je poskus izvedbe filogenetske metode z najmanjšim možnim številom predpostavk. Osvobojena je vsakršnih predpostavk o evoluciji taksonov, filogenetski signal pa meri na popolnoma nov način, s čimer opušča tudi kriterij varčnosti. Pričakujemo, da bomo z metodo tudi pri problematičnih topologijah filogenetskih dreves uspeli izluščiti filogenetski signal z razpoznavanjem sinapomorfne narave znakov in vsaj do neke mere pravilno določiti odnose med taksoni.

1.3 NAMEN RAZISKAVE

Glavni namen raziskave je bil pridobiti osnovno poznavanje uporabnosti nove probabilistične taksične metode za filogenetske analize. V prvem delu poskusa smo preverili, če lahko metodo uporabimo za odkrivanje pravih odnosov med taksoni, in jo primerjali z uspešnostjo popularnih filogenetskih metod na istih naborih podatkov. V nadaljevanju poskusa smo želeli spoznati, kako se filogenetski signal odraza v podpori določeni topologiji drevesa, še posebno nas je zanimala uspešnost razločevanja signala v nesimetričnih filogenetskih drevesih. Obenem pa smo želeli preverili, kako se metoda obnaša v primerih, v katerih pogosto uporabljane filogenetske metode odpovedo.

1.4 PREDPOSTAVKE V POSTOPKU Kladistične analize

Običajen potek kladistične analize je sledeč: opazanje razlik in podobnosti med taksoni, vnos opazanj v podatkovno matriko (data matrix) in ovrednotenje takšnih podatkov s postopkom, ki lahko določi odnose med organizmi in izbere kladogram (Brower, 2000; glej tudi Jenner, 2003). Za vsakega od teh korakov potrebujemo določeno število predpostavk.

Opazovanje razlik temelji na čutnem zaznavanju in kogniciji (prepoznavanju), torej je subjektivno. Opažanje postane dejstvo (postane objektivno), ko dobi dovolj veliko podporo s strani različnih opazovalcev, oz. ko se vsi strinjajo z njim; takrat ga ni več treba testirati (Popper, 1959). Celotna primerjalna biologija je prežeta z idejo evolucije; gre torej za splošno sprejeto dejstvo. Podobno je s principom varčnosti, za katerega je Popper (1979) trdil, da je lasten opazovalčevemu umu. Vsak kladist torej oba principa nevede že uporablja pri svojih opazovanjih, zato jih ni treba dodajati v metodo. Brower (2000) zato trdi, da so opažene razlike in podobnosti edina potrebna postavka za določanje odnosov med organizmi.

Tudi razvrščanje opazanj v podatkovne matrike naj ne bi temeljilo na nobenih nepotrebnih predpostavkah (npr. a priori določanju polarnosti znakov). Skladnost hipotez o podobnosti stanj znakov je podlaga za razlago razvoja taksonov iz skupnega prednika, nikakor pa ne velja obratno (Brower, 2000). Nekateri podatkovne matrike zavračajo kot neučinkovit način prikaza podatkov (Cao in sod., 2007).

Uporaba filogenetskih dreves je splošno uveljavljen način prikazovanja odnosov med taksoni. V uporabi je že od zgodnjega 19. stoletja (Brady, 1985). Hierarhija taksonov v takšnem kladogramu je, podobno kot zgoraj obravnavan princip varčnosti, uporaben kriterij za nedvoumno razvrščanje taksonov na podlagi znanja o njih. V veliko pomoč pri določanju odnosov neke ravni je obsežno vzorčenje na višjih hierarhičnih ravneh glede na predlagano drevo (med predvidoma manj sorodnimi taksoni), kar nam pomaga pri testiranju monofilij v notranji skupini (ingroup) kot tudi na nižjih nivojih (bolj sorodni taksoni; Brower, 2000).

Še pred določanjem skupin taksonov je nujno treba ločiti apomorfna in pleziomorfna stanja znakov, kar v enem od korakov analize zahteva koreninjenje kladograma. Koreninjenje z zunanjo skupino (outgroup rooting) je ustrezna metoda, saj temelji samo na preverljivi hipotezi, da so vsi taksoni notranje skupine klad glede na dano zunanjo skupino (Brower, 2000).

Sistematika teži k določitvi ene same dobro podprte hierarhije vseh živih bitij, za odkrivanje te hierarhije pa obstaja mnogo metod. Do sedaj pa so bile za razložitev odnosov v posameznih vejah »drevesa življenja« uporabljene različne metode, ki temeljijo na zelo različnih predpostavkah, razlike med metodami (in s tem vpliv predpostavk na končni rezultat) pa niso dovolj dobro poznane (Brower, 2000). Nekateri avtorji vztrajajo, da je treba za potrditev neke sorodstvene hipoteze uporabiti več metod na istih podatkih (e.g. Kim 1993), kar pa nikakor ne more privedi do večkratne

neodvisne potrditve rezultata in tudi ne povečuje zaupanja v rezultat. Za zadovoljivo razlago te hierarhije je – že zaradi principa varčnosti – treba uporabiti zgolj eno zadovoljivo uspešno filogenetsko metodo.

1.5 RAZLIČNI POGLEDI NA IZVOR RAZLIK V STANJIH ZNAKOV

Glavni cilj filogenetske sistematike je iskanje zgodovinskega poteka evolucije na podlagi razlik med opazovanimi organizmi, vir katerih so spremembe v zaporedjih njihove DNA (Bang in sod., 2000). Za ustrezno razlago evlucijske zgodovine neke skupine organizmov je treba vzpostaviti ustrezno paradigmo, ki samo na podlagi razlik in podobnosti med njimi najbolj nepristransko določi hierarhijo odnosov med njimi. Grant in Kluge (2004) si razlagata posamezne znake kot zgodovinske entitete, stanja znakov pa kot njihove lastnosti, ki odražajo zgodovinski signal znaka, ki je vzrok za homologije. Na znake gledata kot na serije transformacij stanj. Avtorja enačita koncepta homologije in sinapomorfije, zato trdita, da prisotnost sinapomorfnihih stanj znakov ne sme biti kriterij določanja homologije, saj to vodi v krožno dokazovanje.

Farris (1983) vidi kriterij varčnosti kot osnovo transformacijskega pogleda na evolucijo organizmov; po njigovem mnenju so vsa stanja, ki kažejo sinapomorfno naravo, nastala le enkrat. Izključuje pa možnost, da so med nadaljnjo evolucijo taksonov, ki jih določajo, izginila in se spet pojavila. Bang in sod (2000) zavračajo transformacijski pristop kot osnovo za določanje sorodnosti med taksoni, ker zahteva hipoteze o zgodovinskem spreminjanju znakov. Ne strinjajo se tudi s predpostavko, da je najenostavnejša razlaga zgodovinskega spreminjanja tudi najbolj enostavna (varčna), saj obstaja tudi možnost, da je sinapomorfno stanje znaka v zgodovini izginilo in se nato spet pojavilo.

Transformacijski pristop vedno podaja pozitivne razlage podobnosti (razlik v stanjih dveh znakov ne pojasnjuje). Zato lahko z njim le predlagamo topologije filogenetskih dreves, ne smemo pa jih tudi potrditi. Taksični pristop vedno upošteva tudi porazdelitev razlik stanj znakov posameznih taksonov, s čimer omogoča direktno testiranje hipotez sorodnosti med dvema taksonoma, zato poda precej boljše razlago sorodnosti med njimi (sinapomorfij). Assis in Brigandt (2009) podpirata ta pogled in vidita transformacijski koncept homologije kot dinamični (razlaga zgodovine), taksično homologijo pa kot statični koncept (trenutno stanje). Zgodovinski razvoj taksonov je hipoteza sorodnosti, ki je ne smemo privzeti kot pravilno brez ovrednotenja glede na dejansko stanje, kar nam omogoča le taksični pogled. Vsi znaki predstavljajo tako sinapomorfije na nekem nivoju taksonomske hierarhije kot transformacije stanj znakov bolj predniških sekvenc (Nelson in Platnick, 1981).

1.6 POGOSTO UPORABLJANE METODE DOLOČANJA IN IZBORA DREVES

1.6.1 Metoda največjega verjetja

Metoda največjega verjetja je eksperimentalno dobro podprta in široko uporabljana filogenetska metoda. Temelji na principu največjega verjetja (maximum likelihood), ki

razlaga, da je najustreznejša hipoteza sorodnosti tista, ki se najbolj sklada z danimi vhodnimi podatki.

$$L_D = P(D|H) \quad \dots(1)$$

Verjetje danih filogenetskih podatkov (npr. nukleotidnih sekvenc) je verjetnost, da so dani podatki resnični, če neke odnose med taksoni (npr. filogram) privzamemo kot resnične. Filogenetsko drevo, za katerega so naši podatki najbolj verjeten rezultat evolucije, lahko prepoznamo kot najbolj ustrezno (Felsenstein, 1981; Page in Holmes, 1998).

Metoda največjega verjetja poleg poznavanja razlik v znakih med opazovanimi taksoni in privzetega hipotetičnega filogenetskega drevesa z znano topologijo in dolžinami vej zahteva uporabo modela evolucije (model substitucij baz v nukleotidnih zaporedjih). Na podlagi danega modela lahko izračunamo verjetje, s katerim dana topologija filogenetskega drevesa podpira porazdelitev stanj znakov v končnih taksonih (Wägele, 2005). Problem metode je, da je treba za vsak par taksonov izračunati verjetje vseh možnih sprememb stanj znakov za vsako možno topologijo dreves. Število možnih dreves eksponentno narašča z večanjem števila taksonov, zato je ta metoda omejena na majhne nabore taksonov (Page in Holmes, 1998). Problematična je tudi nujnost uporabe modelov evolucije (Brower, 2000).

Kot mera podpore posameznim razvejiščem se pri tem postopku uporablja bootstrap podpora (BS, bootstrap support); biti mora izrazito visoka (npr. >90%). Pred izračuni se za vse nabore zaporedij naključno izbere zaporedje posamernih znakov, na podlagi katerih se generirajo nove sekvence z ustreznimi stanji znakov. Na podlagi te sekvence se izbere najboljše drevo; BS nekega klada je delež takšnih dreves, ki vsebujejo ta klad.

1.6.2 Analiza posteriorne verjetnosti porazdelitve dreves z Bayesovim teoremom (Bayesov pristop)

Bayesov pristop omogoča iskanje najbolj verjetnih filogenetskih dreves ob izbranem modelu evolucije in danih taksonomskih znakih (na primer zaporedjih DNA). Ta pristop iskanja filogenij temelji na Bayesovem teoremu:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)} \quad \dots(2)$$

Ta omogoča izračunavanje posteriorne verjetnosti $P(H|D)$ neke porazdelitve znakov v nukleotidnih zaporedjih v množici taksonov ob predpostavki resničnosti dane topologije drevesa. Posteriorna verjetnost je tudi optimalnostni kriterij te metode, biti mora čim večja. Za iskanje optimalnega drevesa se uporablja postopek MCMCMC (Metropolis-Coupled Markov Chain Monte Carlo), ki naključno generira drevesa, le nekoliko različna od predhodnega drevesa. Tudi Bayesov pristop zahteva uporabo modela evolucije za primerjanje vhodnih podatkov (sekvence) s topologijami dreves. Na začetku analize se izračuna posteriorna verjetnost dane porazdelitve znakov glede na neko naključno generirano drevo. Nato se na podlagi prejšnjega drevesa generira množico nekoliko

spremenjenih dreves; ta drevesa se ovrednoti s t.i. Metropolis-Hastingsovim algoritmom, ki temelji na izračunavanju posteriornih verjetnostih vhodnih podatkov, vendar dovoljuje možnost, da se najbolj verjetna hipoteza ne sprejme, kar omogoča bolj ohlapno iskanje pravega drevesa v večjem številu podobnih dreves, z vsakim korakom pa se povečuje $P(D|H)$. Vsak tak korak imenujemo generacija. V filogenetskih analizah navadno dovoljujemo, da markovska veriga poteče čez stotisoče generacij (Wägele, 2005).

1.6.3 Metoda največje varčnosti

Optimalnostni kriterij metode največje varčnosti (maximum parsimony - MP) je število pretvorb znaka v celotni filogeniji, za kar ne potrebuje predpostavk modela evolucije (Hagstrom in sod., 2004). MP za vsako možno drevo izračuna število mutacij, ki bi rezultirale v nukleotidnih sekvencah taksonov. Najbolj varčno drevo predvidi najmanj mutacij.

Podobno kot ML mora MP preveriti vse možne topologije dreves, zato je zelo potratna pri analizi velikega števila taksonov. To težavo lahko rešimo z uporabo hevrističnih oblik metode. Poznano je tudi, da metoda včasih predlaga napačne topologije, denimo zaradi privlaka dolgih vej (Felsenstein, 1978).

1.7 TAKSIČNE FILOGENETSKE METODE

1.7.1 Analiza hipotez treh taksonov

Edina bolj raziskana taksična metoda je analiza hipotez treh taksonov (three taxon statement analysis, 3TA; Nelson in Platnick, 1991). Temelji na postavljanju hipotez tripletov dveh taksonov z istim, sinapomorfim stanjem znaka, in dodatnega taksona s pleziomorfim stanjem znaka. Apomorfim stanjem znaka dodeli metoda vrednost 1, pleziomorfim pa 0; za zunanjo skupino pri analizi uporablja hipotetični takson, kjer imajo vsi znaki pleziomorfno stanje. Na podlagi analize vseh dobljenih tritaksonskih hipotez se določi možne kladograme in v skladu s principom varčnosti izbere najboljšega. Originalna metoda 3TA predvideva zgolj dve stanji znaka – prisotnost (1) in odsotnost (0); med več različnimi stanji znaka (npr. nukleotidi v zaporedju nukleinskih kislin) pa ne more razlikovati (Carine in Scotland, 1999). Scotland (2000) je predlagal prilagojeno obliko 3TA, pri kateri se še pred analizo binarno zakodira stanja znakov tako, da med analizo ostane informativna; ta metoda in standardna metoda največje varčnosti sta dali neujemajoče rezultate. Obe metodi 3TA sta slabo eksperimentalno podprti in se v sistematiki ne uporabljata (Farris in sod., 2001).

1.7.2 Probabilistična taksična metoda (PTM)

Ta metoda (Trontelj, osebni vir) temelji na alternativnem pojmovanju verjetnosti ter zaupanja. Hennig (1966) je svojo filogenetsko metodo želel utemeljiti na predpostavki, da so odnosi med organizmi resnični. Taksoni na višjih nivojih torej posedujejo enako stanje znakov, ker pripadajo isti skupini (de Queiroz in Gauthier, 1990), in ne obratno. Ta predpostavka omogoča probabilistični pristop k izračunavanju filogenetskega signala, ki

ga imajo znaki v dani kladistični hierarhiji. Tako se znebimo potrebe po definiranju hipotez o monofilijah skupin po podobnosti znakov. Zato se lahko vprašamo, kakšna je verjetnost, da dana monofilija podpira domnevno sinapomorfno naravo nekega znaka. To verjetnost lahko izračunamo kot relativno frekvenco stanj znaka v predlagani monofiliji. Uporabna vrednost metode za filogenetsko analizo še ni bila preverjena.

2 MATERIAL IN METODE

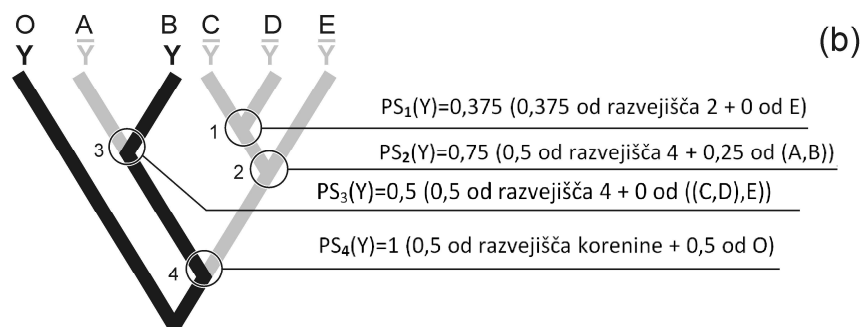
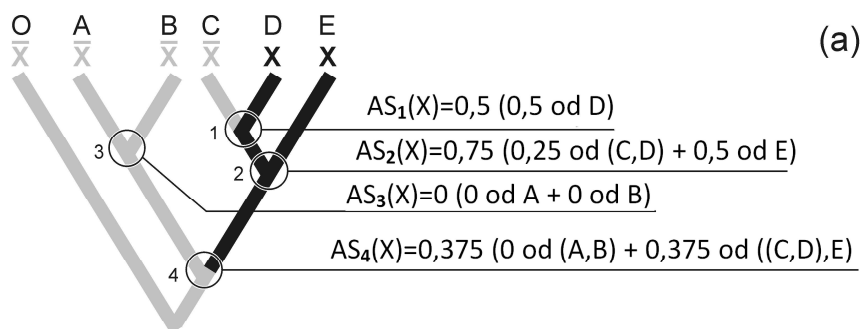
2.1 PROBABILISTIČNA TAKSIČNA METODA (PTM)

Osrednji spremenljivki predlagane metode (Trontelj, osebni vir) sta apomorfni signal (AS) in pleziomorfni signal (PS). Določiti ju je treba za vsak znak na vsakem razvejišču kladograma, razen korene. AS danega stanja znaka je relativna frekvenca tega znaka v skupini, ki jo določa razvejišče, PS istega stanja znaka pa je frekvenca tega znaka izven skupine. Razvejišče korenine se pri izračunih ne upošteva.

2.1.1 Apomorfni in pleziomorfni signal

V AS sta izraženi pogostnost in razporeditev stanja znaka v skupini končnih taksonov, ki jih določa neko razvejišče. AS nekega stanja danega znaka predstavlja verjetnost, da končni taksoni, ki imajo to stanje znaka, nosijo originalni signal razvejišča, ki določa monofilijo vseh teh končnih taksonov. Ta verjetnost se ne nanaša na verjetnost ali količino sprememb stanj znakov, obenem pa ni odvisna od modelov ali predpostavk. Odvisna je le od vzorca razvejanosti kladograma.

Apomorfni signal stanja znaka X na koreninjenem kladogramu mora biti izračunan za vsako razvejišče posebej za poljubno število stanj znakov, ki se pojavljajo na končnih taksonih. Najprej se izračuna AS končnih taksonov. Za vsako hierarhično višje ležeče razvejišče se AS izračuna kot aritmetična sredina AS razvejišč naslednjega nižjega nivoja (Slika 2).



Slika 2: Postopek izračuna AS (a) in PS (b) v posameznih razvejiščih kladograma.

Na vpliv AS nekega končnega taksona na AS nekega višje ležečega razvejišča močno vpliva oblika kladograma. Asimetrični kladogrami imajo več ravni razvejišč (so globlji) kot simetrični. AS bolj bazalnih taksonov v asimetričnih drevesih imajo zato relativno večji vpliv na AS še nižje ležečih razvejišč kot AS taksonov z razvejiščem v nižjih nivojih razvejišč kladograma. Pri simetričnih drevesih je razporeditev AS končnih taksonov enakomerna ($n^{-1}, n^{-1}, \dots, n^{-1}$), v popolnoma asimetričnih drevesih pa je zelo neenakomerna ($2^{-1}, 2^{-2}, \dots, 2^{2-n}, 2^{1-n}, 2^{1-n}$). Močnejši AS terminalnega taksona z razvejiščem globlje v asimetričnem kladogramu lahko prispeva več šuma v razvejišče pod njim.

Apomorfija lahko poda filogenetsko informacijo samo v razvejišču, pod katerim je nastala. Na naslednjem, nižje ležečem razvejišču apomorfija izgubi ves signal in postane pleziomorfija. Navadno samo z AS ne moremo zanesljivo odkriti tega razvejišča, zato potrebujemo neko količino, komplementarno AS, ki jo imenujemo **pleziomorfni signal** (PS). PS stanja znaka na danem razvejišču se ne nanaša na s slednjim določeno monofilijo, ampak na preostalo drevo izven razvejišča. Izračuna se ga podobno kot AS, le da začnemo pri korenini kladograma namesto pri končnih taksonih. PS korenine je enak AS zunanje skupine. Za vsako razvejišče se PS izračuna kot povprečje PS za nivo višjega razvejišča in AS sestrške skupine (Slika 2).

2.1.2 Verjetnost apomorfne podprtosti (PAS – Probability of apomorphic support)

Taksični pristop narekuje, da mora optimalnostni kriterij izbire pravega drevesa odražati verjetnost, da taksoni pripadajo monofiliji. To se lahko izpelje iz verjetnost, da je neko dano stanje znaka apomorfno, ter iz števila znakov, za katere je ta verjetnost znana. Za stanje znaka D in kladogram κ s kladi $C_1, C_2, \dots, C_n$ lahko verjetnost, da opazovani znak daje apomorfno podporo določenemu kladu na kladogramu κ , izrazimo kot verjetje tega klada za ta kladogram

$$L(C_i, \kappa) = P(D|C_i, \kappa) \quad \dots(3)$$

ter jo lahko izračunamo kot vsoto pogojnih verjetnosti apomorfne podpore vseh možnih stanj znaka (s_i) znaka D_j . Imenujemo jo verjetnost apomorfne podpore j -tega znaka ($pas_j(C)$).

$$pas_j(C) = P(D_j | C) = \sum_{i=1}^m P(s_i | C) \quad \dots(4)$$

Če za neko stanje znaka s_j predpostavimo, da je apomorfno na danem razvejišču, potem je eno od preostalih stanj pleziomorfno na istem razvejišču, vsa preostala stanja pa predstavljajo avtapomorfije na hierarhično nižjih nivojih razvejišč. Verjetnosti, da je neko stanje apomorfno v danem razvejišču lahko izračunamo iz AS in PS danega znaka za to razvejišče. Verjetnost, da je stanje znaka s_1 apomorfno, stanje s_2 pa pleziomorfno, lahko izračunamo kot:

$$AS(s_1) \cdot PS(s_2) \cdot (1 - AS(s_3)) \cdot (1 - AS(s_4)) \cdot \dots \cdot (1 - AS(s_m)) \quad \dots(5)$$

Verjetnost apomorfne podprtosti klada C s strani stanja znaka s_1 je seštevek vseh možnih scenarijev, pri katerih je stanje s_1 apomorfno, eno od preostalih stanj ($s_2, s_3, \dots, s_m$) pa pleziomorfno:

$$\begin{aligned} P(s_1 | C) = & AS(s_1) \cdot PS(s_2) \cdot (1 - AS(s_3)) \cdot (1 - AS(s_4)) \cdot \dots \cdot (1 - AS(s_m)) \\ & + AS(s_1) \cdot PS(s_3) \cdot (1 - AS(s_4)) \cdot (1 - AS(s_5)) \cdot \dots \cdot (1 - AS(s_m)) \cdot (1 - AS(s_2)) \\ & \vdots \\ & + AS(s_1) \cdot PS(s_m) \cdot (1 - AS(s_{m-1})) \cdot (1 - AS(s_{m-2})) \cdot \dots \cdot (1 - AS(s_2)) \end{aligned} \quad \dots(6)$$

To enačbo lahko preuredimo v:

$$P(s_1 | C) = AS(s_1) \prod_{i=2}^m (1 - AS(s_i)) \sum_{i=2}^m \frac{PS(s_i)}{1 - AS(s_i)} \quad \dots(7)$$

Z enačbama 2 in 4 lahko izračunamo pas_j za dano razvejišče zgolj iz AS in PS posameznih stanj znakov v tem razvejišču; pas_j za neinformativne znake (imajo le eno stanje) je 0, zato jih lahko izključimo iz analize.

Celokupno verjetnost apomorfne podpore opazovanega razvejišča - $PAS(C)$ – lahko izračunamo kot seštevek $pas_j(C)$ vseh v analizi uporabljenih znakov znakov:

$$PAS(C) = \sum_{i=1}^n pas_i(C) \quad \dots(8)$$

Ta izračun je implementiran v programih EntroPhy (avtorja Marko Pistotnik in Martin Turjak) in Newt (avtor Martin Turjak), uporabljenih v spodaj opisanih poskusih.

Relativno verjetnost apomorfne podpore v nekem razvejišču $rPAS(C)$ izračunamo kot količnik $PAS(C)$ in števila znakov, za katere velja $pas_j(C) > 0,5$.

2.1.3 Kriterij izbire najboljšega kladograma

Najboljši kladogram, dobljen s to metodo, mora izkazovati čim manjšo mero negotovosti. Ustrezna mera negotovosti je entropija, ki jo za diskretno naključno spremenljivko v splošnem izračunamo kot:

$$H(X) = - \sum_i p_i(X) \log p_i(X) \quad \dots(9)$$

Verjetnost apomorfne podprtosti nekega znaka ($pas_j(C)$) lahko obravnavamo kot binarno naključno spremenljivko, predstavlja pa mero gotovosti. Ustrezna mera negotovosti je vrednost $1-pas_j(C)$, ki združuje zmanjšanje podpore stanja znaka zaradi pleziomorfij in prikritje apomorfne stanja znakov zaradi avtapomorfij na nižjih nivojih taksonov. Entropijo sinapomorfne podprtosti j-tega znaka opazovanemu kladu C izračunamo kot:

$$H_j(C) = -(pas_j(C) \log(pas_j(C)) + (1-pas_j(C)) \log(1-pas_j(C))) \quad \dots(10)$$

Stanje znaka s podporo $pas_j < 0,5$ po definiciji ne vsebuje nikakršne informacije o verjetnosti apomorfije, zato mu lahko določimo entropijo 1, za znake s $pas_j > 0,5$ pa izračunamo ustrezno $H_j(C)$. Ta način izračuna entropije je implementiran v programu EntroPhy. Najbolj ustrezna topologija ima najmanjšo skupno entropijo, ki jo izračunamo kot vsoto entropij po celotni topologiji.

2.2 PREVERJANJE USTREZNOSTI PTM KOT METODE ZA ODKRIVANJE ODNOSOV MED TAKSONI

2.2.1 Priprava naborov simuliranih nukleotidnih zaporedij

Pripravili smo 50 naključnih filogenetskih dreves s 7 končnimi taksoni in enim taksonom v zunanji skupini. S programom Evolver v programskem paketu PAML 4.2 (Yang, 2007) smo pripravili simulirane nabore nukleotidnih zaporedij za vsako od dreves (parametri: random rooted trees, birth rate: 2; death rate: 2; sampling fraction: 0,1; mutation rate: 1). Za vsako od dreves smo ustvarili simuliran nabor zaporedij štirih različnih dolžin (5000, 10000, 25000 in 50000 znakov).

2.2.2 Pridobitev filogenetskih dreves, ki najboljše pojasnijo pripravljene nabore nukleotidnih zaporedij

Vsak nabor simuliranih nukleotidnih zaporedij smo analizirali s programi PAUP* 4.0b10 (Swofford, 2003), PAML 4.2 in EntroPhy. V programu PAUP* smo izvedli analize po standardni metodi največje varčnosti (MP; prvi takson v newick formatu drevesa določen za zunanjo skupino, 500 ponovitev bootstrap). S PAML 4.2 smo pridobili najbolj verjetno drevo po standardni metodi največjega verjetja (ML; originalni parametri v datoteki baseml.ctl programa PAML 4.2). V programu EntroPhy (avtorja Marko Pistotnik in Martin Turjak) je implementirana v tem delu testirana probabilistična taksična metoda (Trontelj, osebni vir). Z vsakim od programov smo za vsak nabor zaporedij pridobili za dano metodo najboljše podprto filogenetsko drevo v Newick formatu.

2.2.3 Statistično ovrednotenje topologije

Vsa dobljena filogenetska drevesa smo primerjali z ustrežajočimi izhodiščnimi topologijami dreves v programu PEST: Precision Estimated by Sampling Trait 3.0 (Zujko-Miller in Miller, 2003). Grafe smo izrisali v programu SPSS 13.0.

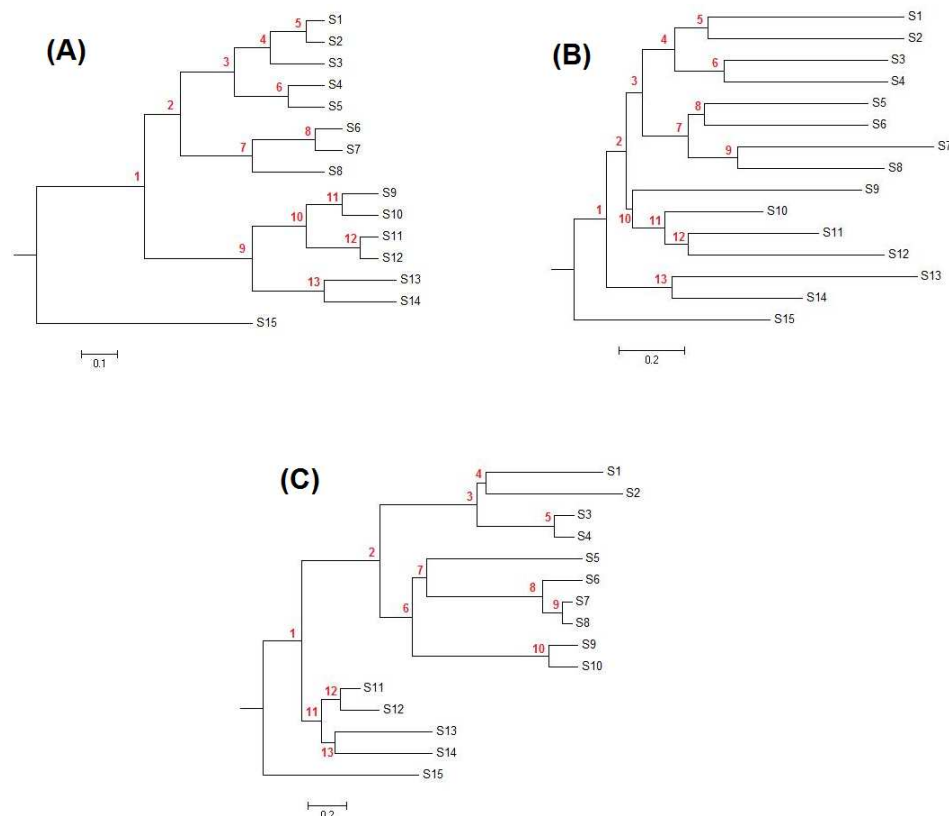
2.3 PODPORA PTM POSAMEZNIM RAZVEJIŠČEM DOLOČENIH TOPOLOGIJ

2.3.1 Priprava naborov simuliranih nukleotidnih zaporedij

Pripravili smo tri filogenetska drevesa z značilnimi razlikami v topologiji oz. v dolžinah vej (Slika 3). Na podlagi vsakega od teh dreves smo s programom Evolver pripravili nabore nukleotidnih zaporedij dolžine 10000 znakov za 15 hipotetičnih končnih taksonov. Za vsako od dreves smo samo s spreminjanjem začetnega naključnega števila (random number seed) pripravili po 100 naborov zaporedij za dva različna modela evolucije, in sicer JC69 (atributi: $\kappa=1$, $\alpha=0$, št. kategorij diskretne $\gamma=0$, frekvence baz: vse 0,25) in GTR (atributi, originalno nastavljeni v datoteki McBase.dat za PAML 4.2). Pripravili smo še nabor popolnoma naključno generiranih nukleotidnih zaporedij (nabor r; Preglednica 1) ter naključno pridobljen kladogram R (Slika 12).

Preglednica 1: Parametri, po katerih smo pripravili posamezne nabore simuliranih nukleotidnih zaporedij.

	Nabor zaporedij						
	a	b	c	d	e	f	r
Topologija	A	B	C	A	B	C	/
Model evolucije	JC69	JC69	JC69	GTR	GTR	GTR	/



Slika 3: Izhodiščna drevesa, uporabljena za simuliranje naborov nukleotidnih zaporedij: simetrično drevo (A), drevo z dolgimi zunanji vejami (B) in asimetrično drevo (C). Oznake S₁-S₁₇ predstavljajo končne taksoni, rdeče obarvane številke 1-13 pa posamezna razvejišča. V nadaljnjem besedilu so ta drevesa poimenovana s pripadajočimi oznakami (A, B, C).

2.3.2 Izračun PAS na podlagi naborov simuliranih nukleotidnih zaporedij za znane topologije dreves.

Vseh 7 skupin naborov simuliranih nukleotidnih zaporedij smo analizirali v programih Garli (Zwickl, 2006), MrBayes (Huelsenbeck in Ronquist, 2001) in Newt (avtor Martin Turjak). S programom Garli smo izvedli analize za 20 naborov zaporedij iz vsake skupine naborov (a-e in r) po standardni metodi največjega verjetja (ML). V programu MrBayes smo izvedli analize vseh naborov po metodi, ki uporablja Bayesov pristop (parametri: nst=6, rates=invgamma, mcmc ngen=10000, samplefreq=10, sumt burnin=250).

Program Newt je implementacija probablistične taksične metode. Z njim smo za vsak nabor sekvenc izračunali vrednosti PAS za vsa razvejišča ustreznih izhodiščnih topologij dreves. Izračunali smo tudi vrednosti PAS naključno generiranih naborov sekvenc (nabor r) za vsako od analiziranih topologij dreves. Program je zabeležil tudi število vseh znakov, ki vsebujejo apomorfni signal ($\text{pas} > 0,5$).

Dodatno smo za nabor c izračunali tudi podporo določeni slučajno izrisani topologiji R (Slika 12), za nabor c pa smo izračunali še podporo simetrični topologiji A, ki je nesimetrični topologiji C, po kateri smo pripravili nabor c, do neke mere podobna. Preglednica 2 prikazuje vse v analizah uporabljene nabore izračunanih vrednosti podpor PAS.

Preglednica 2: Pari naborov simuliranih sekvenc ter topologij dreves, uporabljeni v nadaljnjih analizah.

	Nabor simuliranih sekvenc						
topologija	a	b	c	d	e	f	r
A	PAS(a A)			PAS(d A)			PAS(r A)
B		PAS(b B)			PAS(e B)		PAS(r B)
C	PAS(a C)		PAS(c C)			PAS(f C)	PAS(r C)
R (Slika 12)	PAS(a R)						

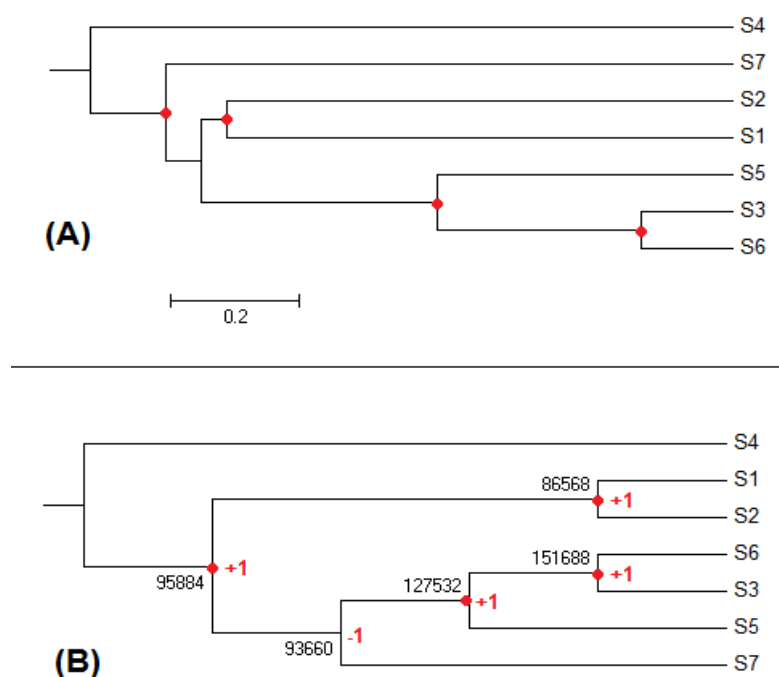
2.3.3 Statistična analiza podatkov

V tabele v programu Excel 2003 smo vnesli podatke, ki smo jih pridobili z vsemi tremi uporabljenimi računalniškimi programi. Na podlagi celokupne PAS in števila znakov s $\text{pas}_j > 0,5$ smo za vse nabore izračunali relativno PAS (rPAS), ki jo izračunamo kot količnik obeh količin in predstavlja povprečno PAS vseh informativnih znakov. Medsebojno smo primerjali rezultate analize PTM vseh simuliranih sekvenc, dobljenih na podlagi enake topologije dreves. Nabore PAS(a|A) smo primerjali še z nabori PAS(a|C) ter PAS(a|R). Statistične analize smo izvedli v programu R 2.9.1.

3 REZULTATI

3.1. USTREZNOST PTM ZA DOLOČANJE TOPOLOGIJ DREVES

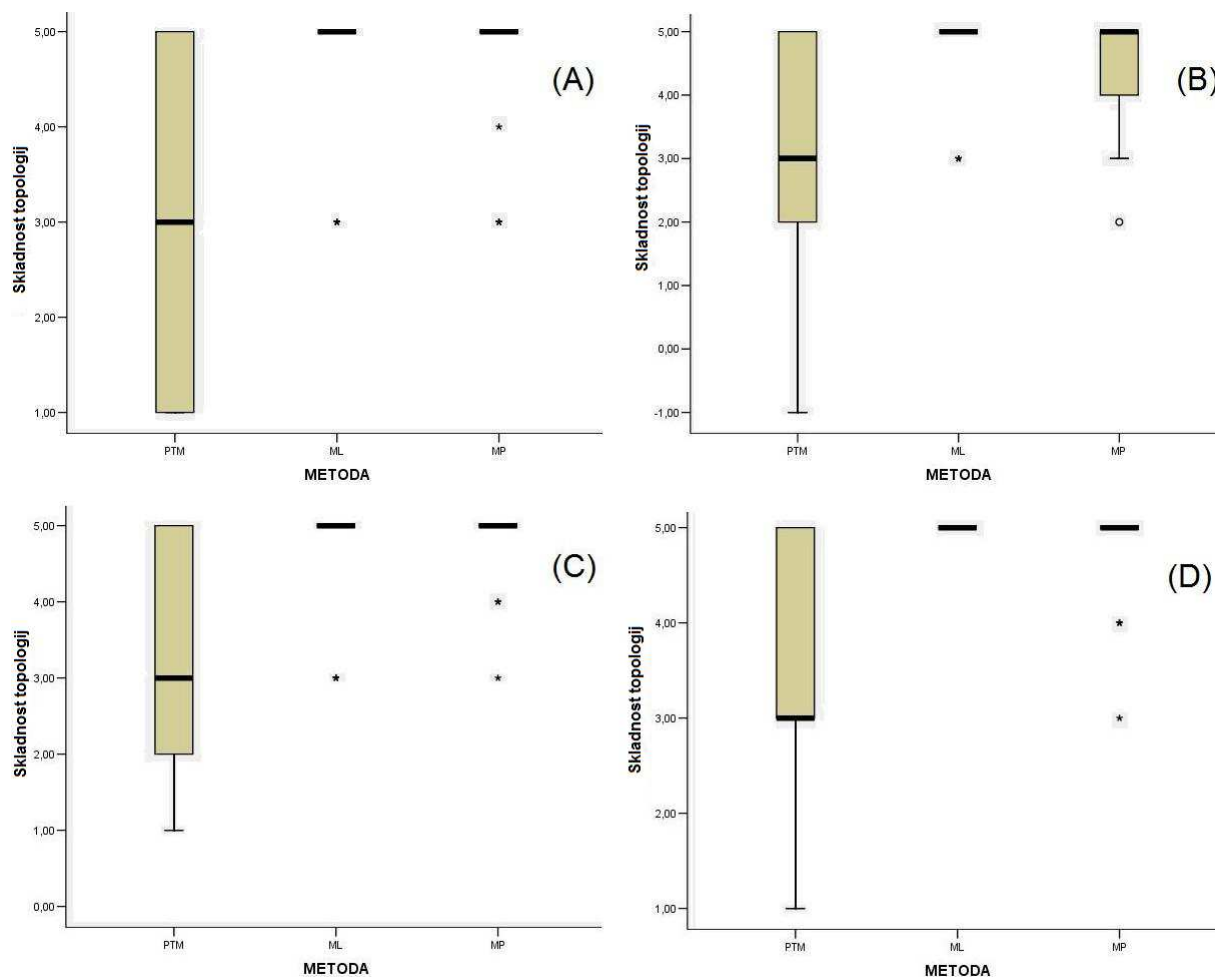
V programu PEST (Zujko-Miller in Miller, 2003) smo preverili skladnost topologij v analizi pridobljenih dreves z izhodiščnim drevesom in to prikazali z enostavno mero podpore posameznih dreves na podlagi enakosti kladov, ki jih kažejo posamezna razvejišča drevesa. Za vsako razvejišče z visoko podporo k skupni vrednosti podpore se po tej metodi skupni vrednosti prišteje 1 točko, za razvejišče, ki ima pri danem drevesu izrazito slabo podprto glede na druga drevesa v skupini dreves, ki jo analiziramo, se ne prišteje ničesar, za vsako manjkajoče razvejišče pa se odšteje 1 točko (Slika 4). Skladnost topologij 5 torej pomeni, da vseh je vseh 5 možnih kladov vsakega od dreves s sedmimi taksoni enakih in dobro podprtih; podpora 1 pa npr. pomeni, da je v dveh kladih prišlo do neujemanja. Skladnost topologij 4 pomeni, da en klad ni dovolj dobro podprt. Skladnost topologij -5 bi pomenila popolno neujemanje kladov.



Slika 4: Primerjava pravilne topologije drevesa (A) in enega od dreves, izbranih s PTM (B). Prikazan je izračun mere podobnosti obeh topologij. Obe drevesi se razlikujeta le v enem razvejišču; skladnost topologij obeh dreves znaša 3. Števila v razvejiščih spodnjega drevesa so vrednosti PAS teh razvejišč.

Probabilistična taksična metoda (PTM) se je v primerjavi z metodo največje varčnosti in metodo največjega verjetja za vse primerjane dolžine sekvenc obnesla slabo (Slika 5), saj v večini primerov ni določila pravilne topologije dreves. Metodi MP in ML sta v skoraj vseh primerih izbrali ustrezno topologijo. Srednja vrednost skladnosti topologij, določenih s PTM, je 3, kar ustreza enemu napačno določenemu kladu. Praktično vse

vrednosti skladnosti topologij za PTM so v intervalu med vrednostima 1 in 5. Pri ponovnem pregledu dreves se je izkazalo, da so bila pri skoraj vseh drevesih pravilno določena hierarhično najnižja razvejišča (razvejišča dveh končnih taksonov) in hierarhično najvišje razvejišče, razvejišča vmesnih ravni pa so prejela zelo slabo podporo PAS, zato je metoda pogosto izbrala napačen klad.



Slika 5: Škatlični diagrami porazdelitve skladnosti topologij topologij dreves, izbranim z vsako od uporabljenih filogenetskih metod, izrisani na podlagi naborov sekvenc s 5000 (A), 10000 (B), 25000 (C) in 50000 (D) znaki (v vseh primerih za vsak par PTM-ML in PTM-MP velja: $p \gg 0,01$).

Z večanjem števila analiziranih znakov se je nekoliko zmanjšala pogostnost napačne izbire obeh višjih razvejišč, vendar je srednja vrednost podobnosti topologij v vseh primerih 3. To pomeni bodisi eno napačno določeno višje razvejišče, bodisi dve slabo podprti, a pravilno izbrani višji razvejišči. V vseh primerih ima le majhen delež dreves pravilno določeno topologijo (vrednost 4 ali več). Celokupna PAS posameznih kladov morebiti ni ustrezna mera za izbor optimalnega drevesa.

3.2. PAS KOT MERA PODPORE SKUPIN NA KLADOGRAMU

ML in Bayesov pristop sta v vseh primerih za vseh šest naborov sekvenc določili izhodiščni enako topologijo dreves s povečini popolno podporo posameznim razvejiščem razen na nekaterih problematičnih razvejiščih (Preglednica 3).

Obe metodi sta slabo podprli ista razvejišča. Razvejišče 10 dolgovejnatega drevesa B združuje štiri taksone z visokim številom razlik v nukleotidnem zaporedju, ki so od preostalega drevesa ločeni z izredno kratko vejo, zato prihaja do dodeljevanja posameznih taksonov alternativnim skupinam. V primeru razvejišč 4, 7 in 11 asimetričnega drevesa C prihaja do enakega pojava (primerjaj Sliko 3 B,C).

Preglednica 3: Povprečne vrednosti podpor testiranih filogenetskih metod za posamezna razvejišča vseh izhodiščnih dreves (N=20; glej Metode 2.3.2). Nizke vrednosti podpor s strani Bayesovega pristopa in ML so podčrtane.

		Podpora setom, simuliranim po JC69				Podpora setom, simuliranim po GTR			
	Razvejišče	Bayes (PP)	ML (BS)	PTM (PAS)	PTM (rPAS)	Bayes (PP)	ML (BS)	PTM (PAS)	PTM (rPAS)
Drevo 1	1	1,00	1,00	2426	0,68	1,00	1,00	1398	0,72
	2	0,99	0,99	2022	0,68	1,00	1,00	950	0,66
	3	1,00	1,00	2029	0,72	1,00	1,00	1010	0,69
	4	1,00	1,00	1599	0,77	1,00	1,00	956	0,75
	5	1,00	1,00	1512	0,80	1,00	1,00	1043	0,82
	6	1,00	1,00	1879	0,79	1,00	1,00	1138	0,79
	7	1,00	1,00	2162	0,76	1,00	1,00	1126	0,74
	8	1,00	1,00	2112	0,84	1,00	1,00	1277	0,84
	9	1,00	1,00	2432	0,71	1,00	1,00	1186	0,69
	10	1,00	1,00	2213	0,74	1,00	1,00	1134	0,72
	11	1,00	1,00	1574	0,80	1,00	1,00	981	0,80
	12	1,00	1,00	1939	0,81	1,00	1,00	1282	0,80
	13	1,00	1,00	2192	0,79	1,00	1,00	1138	0,79
Drevo 2	1	1,00	1,00	1878	0,62	1,00	1,00	1233	0,67
	2	0,99	0,97	949	0,59	1,00	0,99	505	0,61
	3	1,00	1,00	683	0,57	1,00	1,00	352	0,59
	4	1,00	1,00	1142	0,62	1,00	1,00	585	0,62
	5	1,00	1,00	1598	0,75	1,00	1,00	855	0,75
	6	1,00	1,00	1821	0,75	1,00	1,00	935	0,75
	7	1,00	1,00	1293	0,62	1,00	1,00	636	0,63
	8	0,99	0,99	1668	0,73	1,00	1,00	841	0,73
	9	1,00	1,00	1691	0,75	1,00	1,00	874	0,76
	10	<u>0,82</u>	<u>0,81</u>	1827	0,61	0,97	<u>0,94</u>	945	0,60
	11	1,00	1,00	2304	0,68	1,00	1,00	1237	0,66
	12	1,00	1,00	1277	0,82	1,00	1,00	771	0,80
	13	1,00	1,00	1865	0,81	1,00	1,00	1074	0,79
Drevo 3	1	1,00	1,00	1706	0,60	1,00	1,00	1209	0,68
	2	1,00	1,00	1187	0,60	1,00	1,00	790	0,64

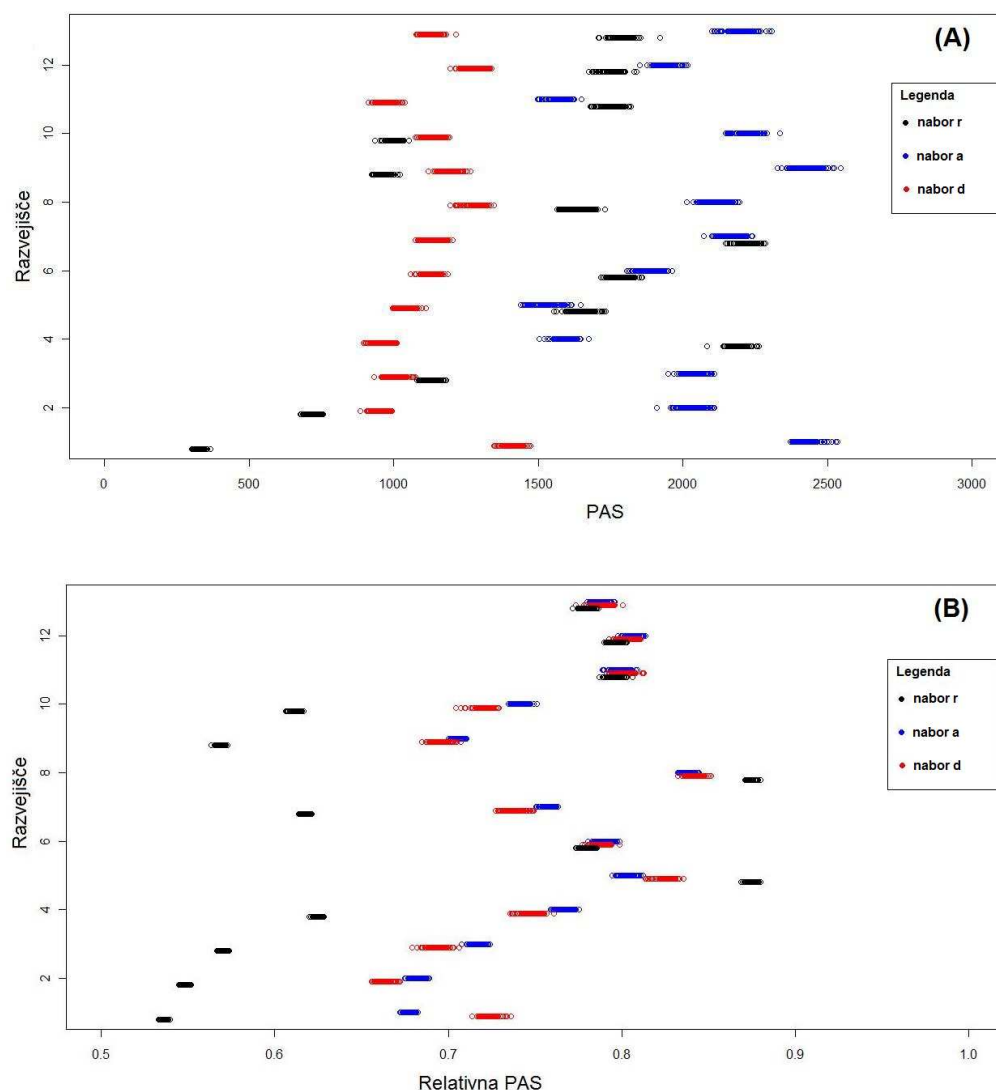
	3	1,00	1,00	2447	0,66	1,00	1,00	1242	0,68
	4	<u>0,83</u>	<u>0,92</u>	1559	0,82	<u>0,94</u>	<u>0,94</u>	868	0,79
	5	1,00	1,00	4277	0,77	1,00	1,00	2009	0,77
	6	1,00	1,00	1904	0,63	1,00	1,00	980	0,63
	7	<u>0,88</u>	<u>0,90</u>	1901	0,72	0,99	0,98	1092	0,69
	8	1,00	1,00	3965	0,83	1,00	1,00	2032	0,80
	9	1,00	1,00	1917	0,85	1,00	1,00	1235	0,84
	10	1,00	1,00	4377	0,78	1,00	1,00	2060	0,79
	11	<u>0,93</u>	<u>0,94</u>	2628	0,70	1,00	1,00	1225	0,68
	12	1,00	1,00	2719	0,75	1,00	1,00	1313	0,76
	13	0,99	1,00	1215	0,79	1,00	0,99	708	0,79

3.2.1 Podpora filogenetske taksične metode posameznim razvejiščem

3.2.1.1 Topologija simetričnega drevesa

Drevo A ima zelo enakomerno topologijo z majhnimi razlikami dolžin posameznih zunanjih in notranjih vej. Na podlagi tega drevesa smo simulirali nabora sekvenc a in d (Preglednica 2). Vidimo, da z drevesom skladne skupine naborov simuliranih zaporedij (nabor a in nabor d) rezultirajo v zelo podobnih vrednostih PAS za vsa razvejišča za oba uporabljena modela evolucije teh sekvenc; vzorec, dobljen za nabor PAS(d|A) je nekoliko bolj variabilen, vendar se za vsa razvejišča giblje v intervalu PAS med 1500 in 2500, večina vrednosti PAS vseh razvejišč za celoten nabor a pa se giblje med 1000 in 1500. Nabor PAS(r|A), ki pravzaprav predstavlja podporo topologiji s strani sekvenc z maksimalnim šumom, izkazuje zelo širok interval vrednosti PAS. Vrednosti rPAS hierarhično višjih razvejišč (glej Slika 3A) za nabore rPAS(a|A) in rPAS(d|A) so izrazito večje od vrednosti rPAS(r|A), vrednosti rPAS istih naborov na hierarhično najnižjih razvejiščih pa so enake ali celo nižje od rPAS(r|A).

V preglednici 4 so prikazane razlike v PAS in rPAS med posameznimi analiziranimi nabori podatkov. Razlika rPAS povprečij celotnih naborov a|A in d|A je zelo nizka, medtem ko sta razliki rPAS a|A kot tudi d|A z ozirom r|A zelo visoki. Oba na podlagi topologije pripravljena nabora sekvenc torej nosita približno enak AS po celotni dolžini drevesa, nabor naključnih sekvenc pa ga vsebuje manj.



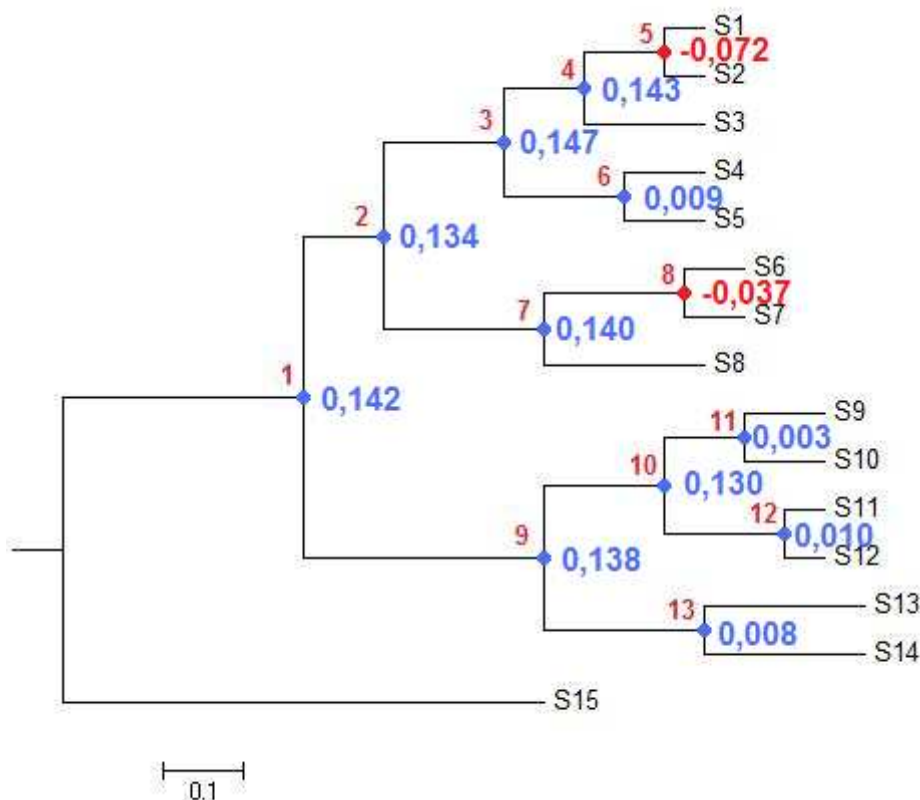
Slika 6: Porazdelitev podpore PAS (A) in rPAS (B) posameznim razvejiščem simetričnega drevesa (glej Slika 3A) s strani na podlagi istega drevesa pripravljenih naborov simuliranih sekvenc a (modro, N=100) in d (rdeče, N=100) ter naborov naključno generiranih sekvenc r (črno, N=100; glej Preglednico 2). Vrednost podpore s strani enega nabora sekvenc danemu razvejišču predstavlja točko na grafu. Vse razlike med skupinami na posameznih razvejiščih so statistično značilne ($p < 0,01$)

Preglednica 4: Pregled razlik povprečij PAS in rPAS med posameznimi nabori podatkov, izračunanimi na podlagi topologije A. Izrazito negativne vrednosti so obarvane rdeče, izrazito pozitivne pa modro. Oznake a|A, d|A in r|A pomenijo, da so vrednosti izračunane za topologijo A z uporabo naborov a, d oz. r.

razlika	(a A)-(d A)		(a A)-(r A)		(d A)-(r A)	
razvejišče	PAS	rPAS	PAS	rPAS	PAS	rPAS
1	1028	-0,046	2103	0,142	1075	0,188
2	1074	0,018	1308	0,134	234	0,116
3	1024	0,023	909	0,147	-115	0,124
4	639	0,020	-594	0,143	-1232	0,123
5	485	-0,021	-119	-0,072	-604	-0,051
6	754	0,003	96	0,009	-658	0,007

7	1034	0,018	-56	0,140	-1090	0,121
8	836	-0,004	479	-0,037	-358	-0,033
9	1237	0,011	1462	0,138	226	0,127
10	1090	0,021	1227	0,130	137	0,109
11	597	-0,003	-180	0,003	-777	0,006
12	658	0,002	190	0,010	-468	0,007
13	1061	0,000	407	0,008	-654	0,008
Celotna topologija	886	0,003	556	0,069	-330	0,066

Vsa hierarhično višja razvejišča so prejela višjo podporo rPAS s strani naborov PAS(a|A) in PAS(d|A), ki sta bila izhodiščno pripravljena na podlagi topologije A, zaradi česar nosita informacijo o sinapomorfijah na vseh nivojih te topologije. Slika 7 prikazuje topologijo drevesa A z označenimi razvejišči z zelo visoko ali zelo nizko vrednostjo razlike rPAS(a|A)-rPAS(r|A).



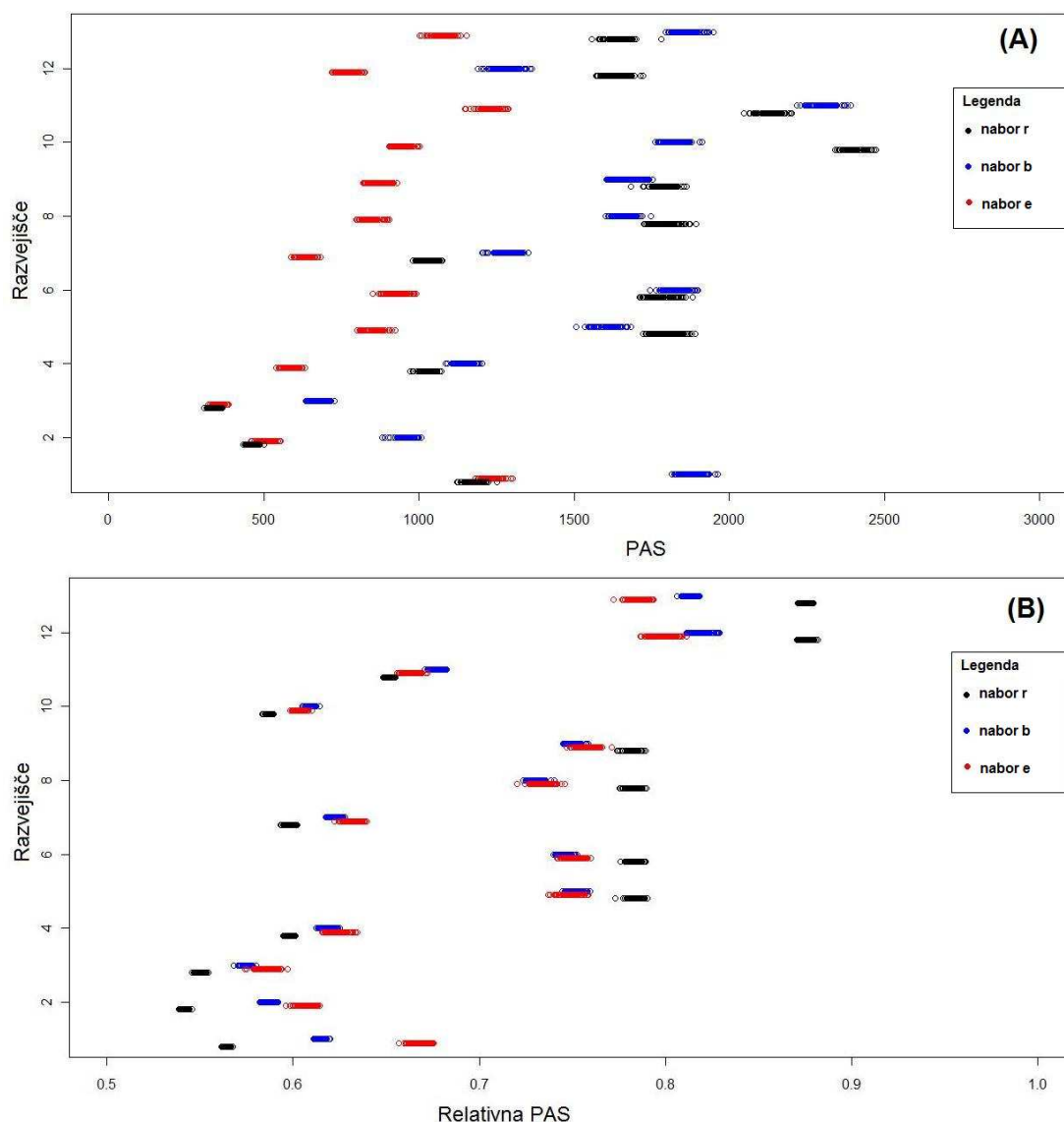
Slika 7: Simetrično drevo A z označenimi razvejišči s pozitivno (modro) in z negativno (rdeče) razliko rPAS(a|A) - rPAS(r|A) (glej Preglednico 4). Drevo je simetrično in ima zelo enakomerne dolžine posameznih vej.

3.2.1.2 Topologija drevesa z dolgimi zunanji vejami

Drevo B ima izrazito dolge zunanje veje, notranje pa kratke. Nabora sekvenc b in e sta simulirana na podlagi tega drevesa. Vrednosti PAS(e|B) so podobno razporejene kot pri zgoraj obravnavanem naboru PAS(d|A), le da so precej nižje (Slika 8). Vrednosti

PAS(b|B) so v povprečju nekoliko nižje in nekoliko bolj variabilne med razvejišči kot pri naboru sekvenc a za topologijo A; njihova razporeditev razen za razvejišče 1 (razvejišče na najvišjem hierarhičnem nivoju) je podobna kot za nabor naključno generiranih sekvenc r.

Razporeditev vrednosti rPAS med posameznimi razvejišči za vse tri nabore pa nakazuje precej drugačno situacijo. Na zunanjih razvejiščih so vrednosti rPAS(b|B) in rPAS(e|B) izrazito nižje kot rPAS(r|B), na višjih razvejiščih pa slednjega presegajo, in sicer najbolj za hierarhično najvišje razvejišče. Razlika med naboroma rPAS(b|B) in rPAS(e|B) je na večini razvejišč majhna, precejšnja je le na hierarhično najvišjem razvejišču.



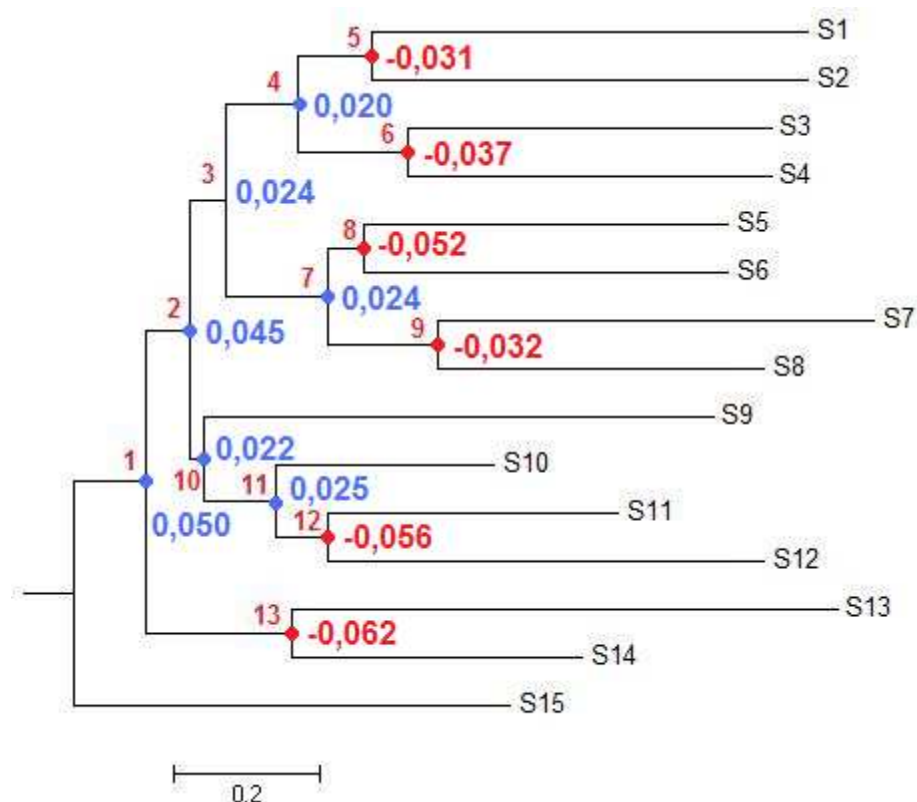
Slika 8: Porazdelitev podpore PAS (A) in rPAS (B) posameznim razvejiščem dolgovejnatega drevesa (glej Slika 3B) s strani na podlagi istega drevesa pripravljenih naborov simuliranih sekvenc b (modro, N=100) in e (rdeče, N=100) ter naključno generiranih naborov sekvenc r (črno, N=100; glej Preglednico 2). Vrednost podpore enega nabora sekvenc danemu razvejišču predstavlja točko na grafu. Vse razlike med skupinami na posameznih razvejiščih so statistično značilne ($p < 0,01$).

Razlike v rPAS naborov PAS(b|B) in PAS(e|B) v odnosu do PAS(r|B) so precej nižje kot v primeru analiz ustreznih naborov za topologijo A, medsebojne razlike povprečnih rPAS celotnih naborov pa so izrazito majhne (Preglednica 5).

Preglednica 5: Pregled razlik PAS in rPAS med posameznimi nabori podatkov, pridobljenih na podlagi topologije B. Izrazito negativne vrednosti so obarvane rdeče, izrazito pozitivne pa modro. Oznake b|B, e|B in r|B pomenijo, da so vrednosti mer izračunane za topologijo B z uporabo naborov b, e oz. r.

razlika	(b B)-(e B)		(b B)-(r B)		(e B)-(r B)	
razvejišče	PAS	rPAS	PAS	rPAS	PAS	rPAS
1	654	-0,053	706	0,050	52	0,103
2	448	-0,019	487	0,045	40	0,064
3	327	-0,011	343	0,024	16	0,036
4	556	-0,006	118	0,020	-438	0,027
5	754	0,003	-188	-0,031	-942	-0,034
6	905	-0,005	47	-0,037	-857	-0,031
7	645	-0,009	253	0,024	-392	0,034
8	823	-0,003	-124	-0,052	-947	-0,048
9	803	-0,007	-105	-0,032	-908	-0,024
10	879	0,005	-584	0,022	-1463	0,017
11	1066	0,013	162	0,025	-904	0,012
12	509	0,020	-359	-0,056	-869	-0,076
13	797	0,028	222	-0,062	-574	-0,090
skupaj	705	-0,004	75	-0,004	-630	-0,001

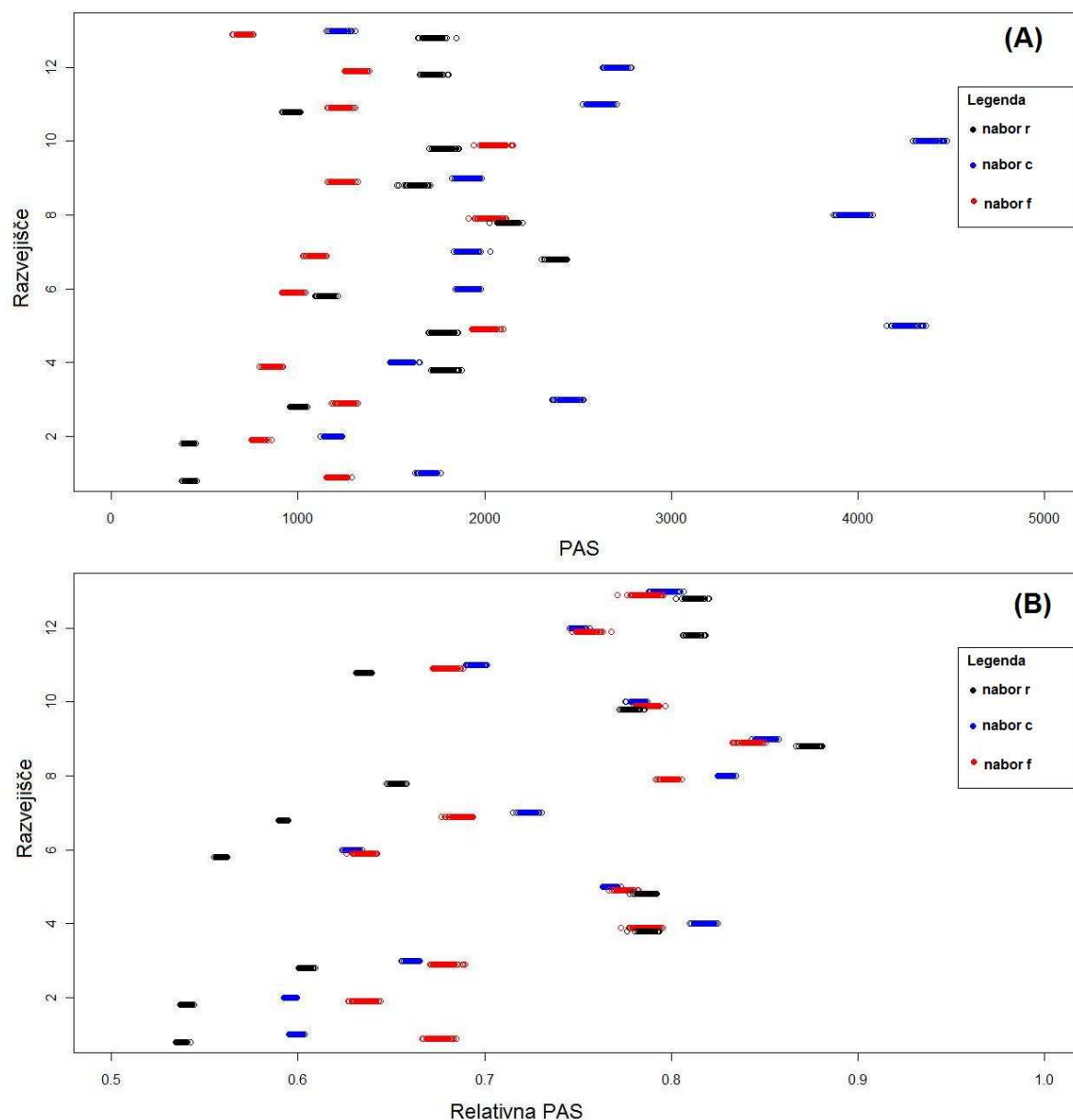
Imamo podobno situacijo kot pri topologiji A, le da je resolucija obravnavane metode na nivoju višjih razvejišč manjša zaradi izrazito dolgih vej končnih taksonov, kjer je potekla večina evolucijskih sprememb v simuliranih sekvencah. To se odraža v nizki podpori tem razvejiščem s strani nabora simuliranih sekvenc b glede na naključne sekvence r.



Slika 9: Dolgovejnato drevo B z označenimi razvejišči s pozitivno (modro) in z negativno (rdeče) razliko $rPAS(b|B) - rPAS(r|B)$. Drevo ima izrazito dolge zunanje veje in kratke notranje veje.

3.2.1.3 Topologija asimetričnega drevesa

Drevo C je izrazito asimetrično. Zunanje veje drevesa so izrazito daljše kot notranje. Nabor $PAS(c|C)$ v treh razvejiščih (5, 8 in 10) doseže izredno visoke vrednosti PAS na račun kopičenja skupnih stanj znakov zaradi dolge skupne evolucije (Slika 10). Za višja razvejišča dobimo pri naboru $PAS(r|C)$ malo, a značilno nižje vrednosti $rPAS$ kot za nabora $PAS(c|C)$ in $PAS(f|C)$ (Preglednica 6).

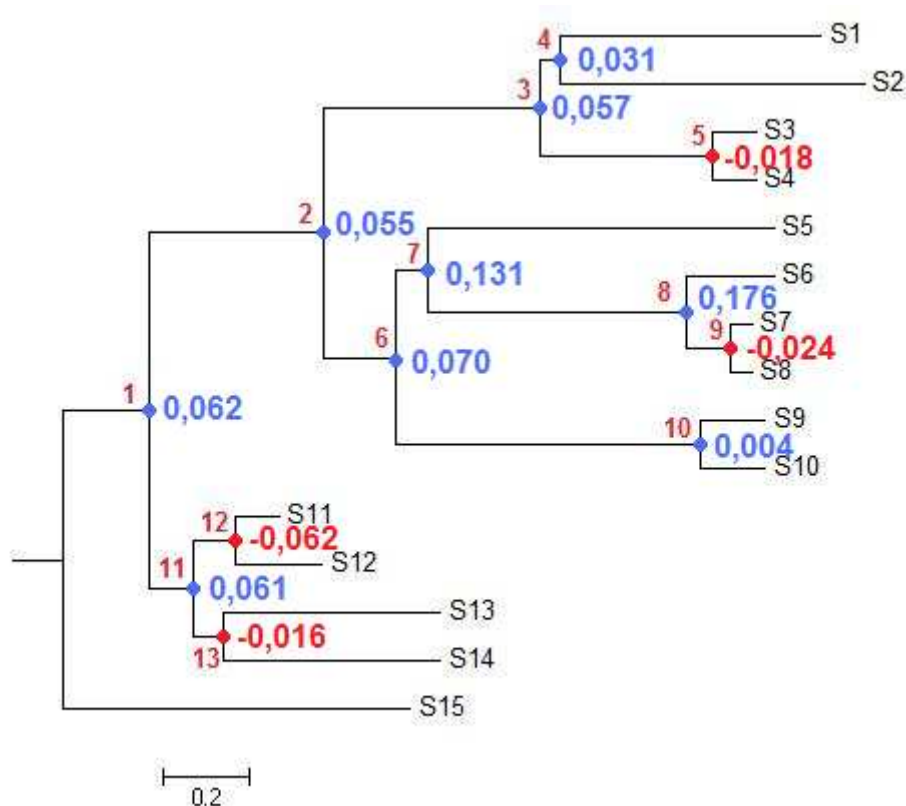


Slika 10: Porazdelitev podpore PAS (A) in rPAS (B) posameznim razvejiščem asimetričnega drevesa (glej Slika 3C) s strani na podlagi istega drevesa pripravljenih naborov simuliranih sekvenc c (modro, N=100) in f (rdeče, N=100) ter naključno generiranih sekvenc r (črno, N=100; glej Preglednico 2). Vrednost podpore enega nabora sekvenc danemu razvejišču predstavlja točko na grafu.

V skladu s pričakovanji je v tem primeru šum v veliki meri prevladal nad filogenetskim signalom, vendar je slednjega ostalo dovolj za nekaj podpore višjim razvejiščem (Slika 11).

Preglednica 6: Pregled razlik PAS in rPAS med posameznimi nabori podatkov, pridobljenih na podlagi topologije C. Oznake $c|C$, $e|C$ in $r|C$ pomenijo, da so vrednosti izračunane za topologijo C z uporabo naborov c , e oziroma r .

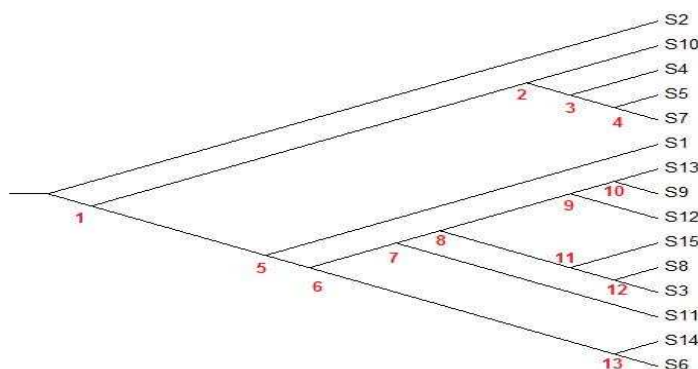
razlika	(c C)-(f C)		(c C)-(r C)		(e c)-(r C)	
razvejišče	PAS	rPAS	PAS	rPAS	PAS	rPAS
1	489	-0,076	1286	0,062	797	0,138
2	396	-0,040	774	0,055	379	0,096
3	1205	-0,018	1447	0,057	242	0,075
4	686	0,032	-232	0,031	-918	0,000
5	2253	-0,007	2488	-0,018	235	-0,012
6	922	-0,006	754	0,070	-169	0,076
7	817	0,037	-469	0,131	-1286	0,095
8	1942	0,030	1838	0,176	-103	0,146
9	679	0,009	279	-0,024	-400	-0,033
10	2325	-0,005	2592	0,004	267	0,009
11	1379	0,016	1643	0,061	264	0,045
12	1402	-0,006	1001	-0,062	-401	-0,056
13	508	0,010	-502	-0,016	-1011	-0,026
skupaj	1154	0,002	992	0,041	-162	0,042



Slika 11: Asimetrično drevo C z označenimi razvejišči s pozitivno (modro) in z negativno (rdeče) razliko $rPAS(c|C)-rPAS(r|C)$.

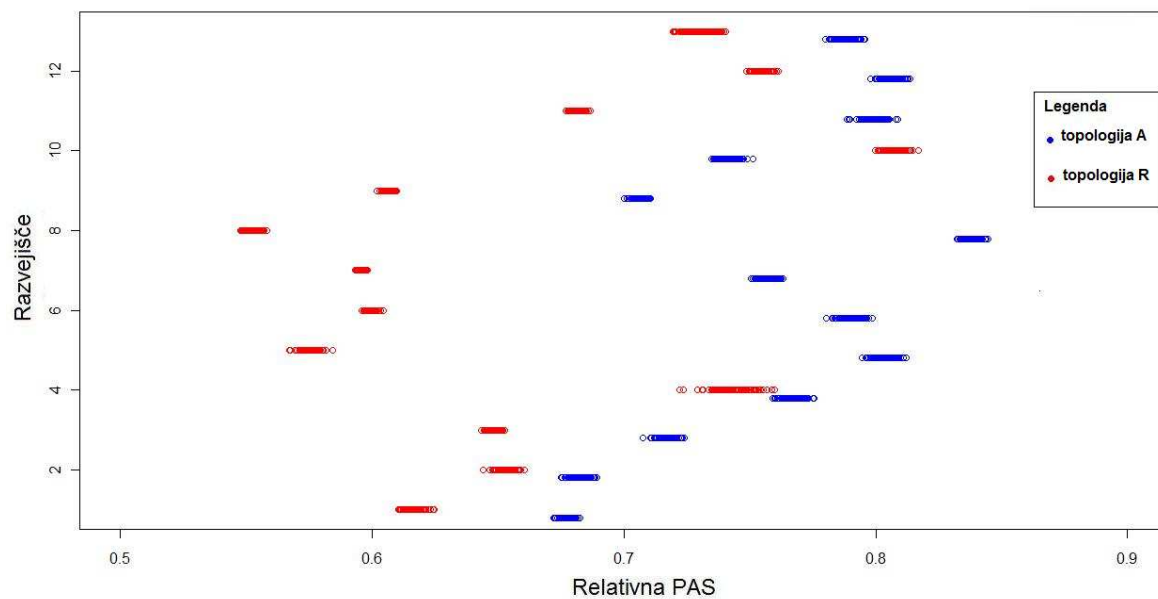
3.2.2 Navzkrižna podpora razvejiščem dveh popolnoma različnih topologij dreves

Preverili smo še, kako skupina naborov sekvenc, simulirana na podlagi neke topologije drevesa, podpre neko naključno, izrazito drugačno topologijo. Za primerjavo smo uporabili topologijo kladograma R (Slika 12) in zanjo izračunali podpore razvejiščem na podlagi nabora sekvenc a.

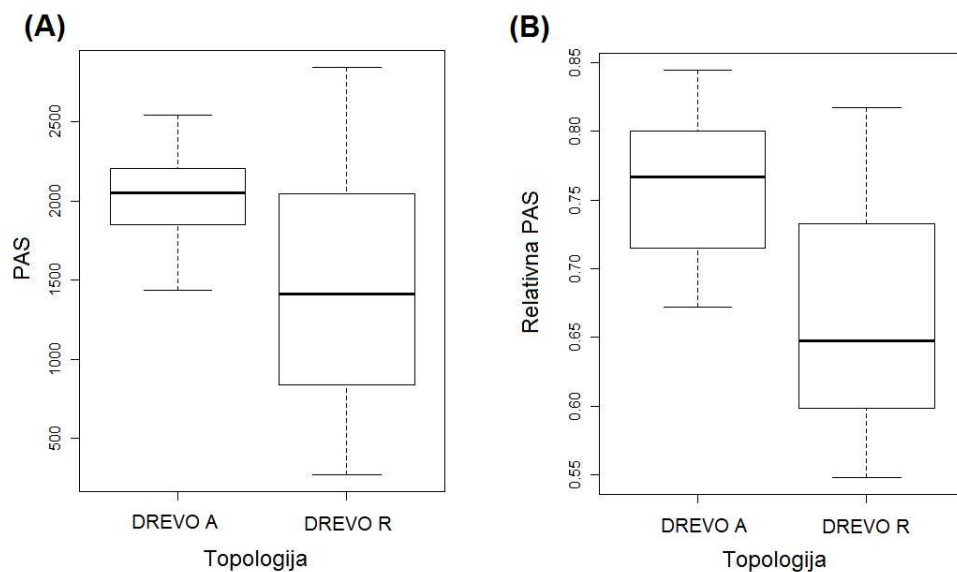


Slika 12: Kladogram R.

Izračunane vrednosti rPAS za topologijo R so v večini primerov nižje kot pri topologiji A (Slika 13, Slika 14). Le eno razvejišče (10) je bolje podprto pri topologiji R, vendar je razlika majhna. Povprečna vrednost celokupnih PAS in rPAS nabora PAS(a|A) je precej višja in mnogo manj variabilna kot pri naboru PAS(a|R). Nabori sekvenc, simulirani na podlagi topologije A, torej po pričakovanjih ne nosijo apomorfnege signala za podporo višjih razvejišč nesorodnih topologij. Povprečna razlika rPAS nabora sekvenc a hierarhično najnižjim razvejiščem topologije A (razvejišča 5, 6, 8, 11, 12, 13) in hierarhično najnižjim razvejiščem topologije R (razvejišča 4, 10, 12 in 13) znaša 0,045, medtem ko je povprečna razlika za vsa ostala, hierarhično višja razvejišča pri vsaki od topologij precej višja (0,121).



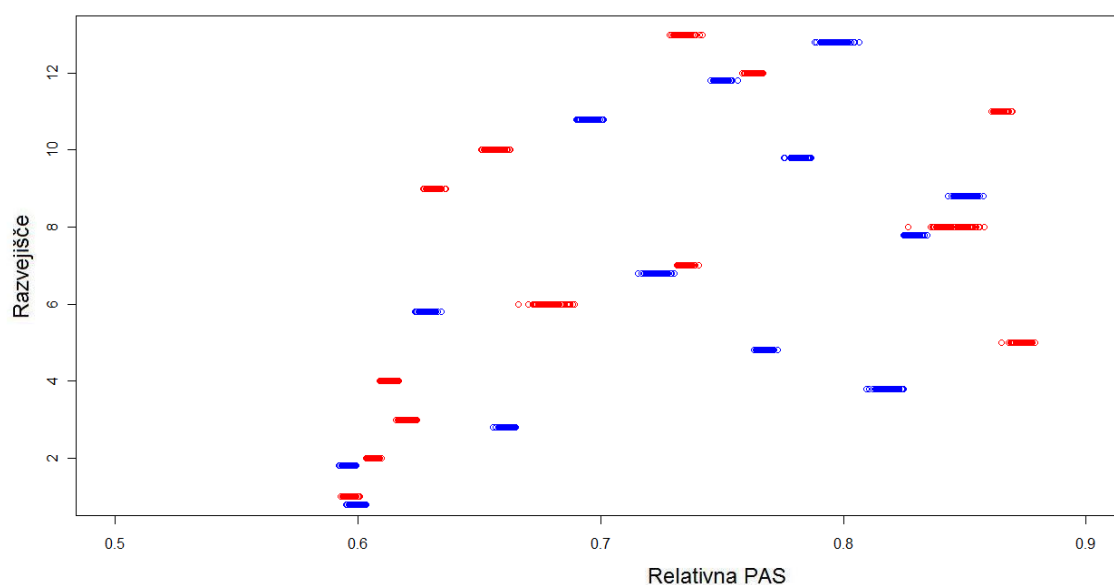
Slika 13: Porazdelitev podpore rPAS, izračunanih na podlagi topologij drevesa A (modro) in popolnoma drugačnega, naključno pridobljenega drevesa R (rdeče) s strani nabora sekvenc a. Opozorilo: enako oštevilčena razvejišča ne predstavljajo istih razvejišč pri obeh drevesih.



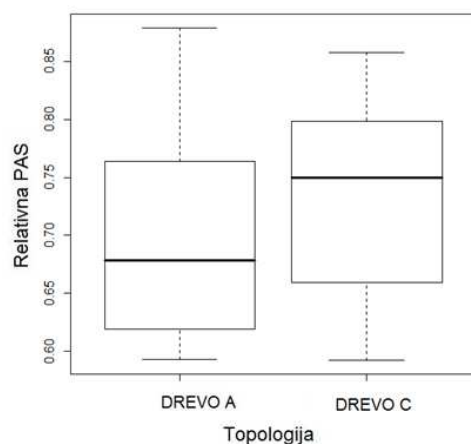
Slika 14: Škatlična diagrama mer podpore topologije PAS (A) in rPAS (B) po celotnih topologijah drevesa A ter naključno pridobljenega kladograma R (za oba grafa velja: $p < 0,01$).

3.2.3 Navzkrižna podpora razvejiščem dreves s podobnostmi v topologiji

Razporeditev rPAS(c|A) kaže na to, da nabor sekvenc c , ki je bil pripravljen na podlagi drevesa C , podaja nekaj apomorfnege signala za podporo topologije A (Slika 15). Obe topologiji izkazujeta zelo podobne odnose med končnimi taksoni, čeprav imata drastično različno hierarhijo (glej Sliko 3A,C). Razlika rPAS(c|C)-rPAS(c|A) znaša za hierarhično najnižja razvejišča znaša le 0,007, za vsa ostala, hierarhično višja razvejišča, pa znaša 0,039. Obe razliki sta statistično značilni ($p < 0,01$), a nizki, kar kaže na to, da PTM omogoča zaznavo podobnosti med dvema podobnima topologijama. PTM namreč podobnima topologijama za posamezna razvejišča dodeli podobnejše vrednosti rPAS kot v primeru popolnoma različnih topologij (Slika 16; primerjaj s Sliko 14B).



Slika 15: Porazdelitev podpore rPAS, izračunanih na podlagi topologij drevesa A (modro) in drevesa C (rdeče), z uporabo nabora sekvenc c . Opozorilo: enako oštevilčena razvejišča ne predstavljajo istih razvejišč pri obeh drevesih.



Slika 16: Škatelni diagram rPAS(c|A) in rPAS(c|C) ($p < 0,01$).

4 DISKUSIJA

Utemeljitelj filogenetske sistematike Willi Hennig je predvidel, da je treba za določitev filogenij razločiti med apomorfni in pleziomorfni stanji znakov, saj lahko le na podlagi prvih sklepamo na skupni izvor organizmov (Hennig, 1966). Transformacijske filogenetske metode, ki se obsežno uporabljajo v sistematiki, temeljijo na uporabi kompleksnih modelov evolucije nukleotidnih zaporedij (metoda največjega verjetja, Bayesov pristop) ali pa utemeljujejo izbor najboljših filogenij na podlagi kriterija varčnosti (metoda največje varčnosti; Farris, 1983). Noben od teh pristopov ne upošteva zgoraj zapisane Hennigove zahteve. Brower (2000) zato vidi transformacijske filogenetske metode le kot vir solidnih hipotez sorodnosti in predvideva uporabo taksičnih filogenetskih metod za dokončno ovrednotenje hipotez sorodnosti med taksoni.

Glavna prednost taksičnih filogenetskih metod je vrednotenje podpore posameznim kladom določene filogenije zgolj na podlagi stanj znakov pri posameznih obravnavanih taksonih. Edina bolj raziskana taksična filogenetska metoda – analiza hipotez treh taksonov (Nelson in Platnick, 1991) – razrešuje odnose med taksoni na podlagi predvidenih apomorfij v posameznih kladih. Vendar pa ta metoda in njena spremenjena izvedba (Scotland, 2000) še nista dovolj eksperimentalno podprti, nekateri avtorji pa ju imajo za neučinkoviti (e.g. Kluge in Farris, 1999).

Probabilistična taksična metoda deluje na drugačnem principu kot vse preostale filogenetske metode (Trontelj, osebni vir). Glavni kriterij za določanje podpore kladom je mera, ki temelji na zgoraj zapisani Hennigovi zamisli, saj omogoča izračun verjetnosti, da sta dva znaka sinapomorfna, na podlagi opaženih stanj znakov v celotni filogeniji. Ta izračun temelji na ločevanju med pleziomorfno in apomorfno naravo stanj znakov zaradi a priori koreninjenja analiziranih kladogramov (Grant in Kluge, 2004). Gre torej za poskus implementacije zgoraj opisanega Hennigovega kriterija v filogenetski metodi. Ostale metode namreč temeljijo bodisi na modelih evolucije (ML, Bayesov pristop), ali pa predpostavljajo uporabo kriterija varčnosti (MP), ki tudi ni skladen s Hennigovim pojmovanjem filogenetskih metod.

V tem diplomskem delu sem poskušal ugotoviti, če je probabilistična taksična metoda na podlagi izračunanih verjetnosti apomorfne podprtosti posameznim kladom zmožna uspešno določiti pravilne topologije dreves. Kriterij za izbor topologije je entropija vrednosti PAS, ki mora biti čim nižja (Shannon, 1984). Entropijo so v filogenetskem kontekstu uporabili tudi drugi avtorji (Brooks in Wiley, 1986), pri PTM pa je prvič uporabljena kot kriterij za izbor kladograma. Že pri majhnem številu analiziranih taksonov se je izkazalo, da metoda v predlagani obliki v večini primerov ne uspe pravilno določiti notranje topologije kladogramov. Tudi večja dolžina nukleotidnih zaporedij ni bistveno vplivala na učinkovitost metode PTM. Izračunana skupna entropija na podlagi vrednosti PAS po celotni topologiji torej ni ustrezna mera za izbiro prave topologije.

V nadaljevanju poskusa sem poskusil ugotoviti, v kakšni meri izračun PAS za posamezna razvejišča a priori podane pravilne topologije omogoča ločevanje filogenetskega signala od naključnega šuma. Izkazalo se je, da samo na podlagi vrednosti PAS, ki filogenetski

signal ovrednoti izključno na podlagi verjetnosti sinapomorfij, ne moremo dovolj zanesljivo predvideti, da je podpora posameznim razvejiščem dejansko odraz razporeditve stanj znakov na končnih taksonih, ki odraža potek evolucije. Tudi naključne sekvence so namreč prejele izrazito visoke vrednosti PAS. Vrednost PAS je namreč odvisna tudi od števila znakov, ki prispevajo k njenemu izračunu (t. j. vseh znakov s $pas > 0,5$). Zato smo kot dodatno mero uvedli relativno PAS (rPAS), t.j. povprečno PAS za vsak predvidoma sinapomorfen znak. Ta mera omogoča kvantificiranje razlik v podpori posameznim razvejiščem s strani različnih naborov podatkov. Izkazalo se je, da ima rPAS hierarhično najnižjih razvejišč za topologiji ustrezajoče nabore sekvenc (vsebujejo filogenetski signal) in naključne nabore sekvenc (nosijo le šum) v večini primerov zelo podobne vrednosti ali pa so te celo v prid naključnih naborov. Naključno enaka stanja znaka – gre torej za homoplazije – na končnih razvejiščih namreč izkazujejo visok apomorfni signal za dano razvejišče. Zaradi nizke verjetnosti, da se bo enako stanje znaka pojavilo še drugje v topologiji, kar za našo metodo pomeni odsotnost pleziomorfne signala, prejmejo ta razvejišča visoke vrednosti rPAS. Pri hierarhično višjih razvejiščih pa se zaradi večjega števila končnih taksonov v kladu močno zmanjša verjetnost, da bodo imeli vsi taksoni zaradi naključja enako stanje določenega znaka, zato naključni nabori sekvenc prejmejo nizke rPAS predvsem zaradi nizkih izračunanih vrednosti apomorfne signala. Nabori sekvenc, ki so pridobljeni v simuliranem evolucijskem procesu in zato nosijo razporeditev stanj znakov, ki odraža ustrezne odnose med taksoni, pa prejmejo na notranjih razvejiščih izrazito visoke vrednosti rPAS. rPAS torej odraža filogenetski signal, zato vsaj do neke mere ustreza kot mera podpore hierarhično višjim razvejiščem. Prav hierarhično višja razvejišča pa so tista, ki so s stališča filogenetske analize pomembna.

Analiza treh različnih topologij filogramov s PTM je pokazala, da lahko ta metoda tudi za drevesa, ki se ob uporabi drugih filogenetskih metod izkažejo za problematična (t. j. prejmejo nizko podporo notranjim razvejiščem oz. vejam) prepozna filogenetski signal za vsa notranja razvejišča. Največje razlike rPAS za hierarhično višja razvejišča med nabori sekvenc, dobljenimi na podlagi opazovanega drevesa, ter nabori naključnih sekvenc, so bile pridobljene za simetrično drevo z enakomernimi dolžinami vej, torej za najmanj problematično s stališča filogenetske analize. Razlika je za notranja razvejišča tega drevesa v povprečju znašala 0,13. Za drevo z dolgimi zunanjimi vejami so bile za hierarhično višja razvejišča pridobljene precej nižje vrednosti razlike rPAS med ustrezajočimi in naključnimi nabori sekvenc (povprečje 0,03). Pri tem drevesu namreč pride zaradi kopičenja razlik na končnih vejah drevesa in s tem večjih razlik med sorodnimi končnimi taksoni do zmanjšanja števila apomorfni stanj znakov v posameznih kladih. Za izrazito nesimetrično drevo z zmernimi razlikami v dolžinah vej smo po metodi PTM izračunali precej visoke vrednosti rPAS za topologiji ustrezajoče nabore sekvenc v primerjavi z naključnimi nabori sekvenc (povprečna razlika znaša 0,09). Metoda torej kljub težavni topologiji, zaradi katere imajo posamezni taksoni, ki se na drevo pripenjajo v hierarhično višjih razvejiščih topologije, izrazito višji vpliv na podporo topologije, uspe izluščiti filogenetski signal za vsa notranja razvejišča, kar je bil tudi eden od problemov, zaradi katerega je bila metoda sploh zasnovana (Trontelj, osebni vir).

Standardna metoda največjega verjetja (ML) in metoda, ki uporablja Bayesov pristop, sta za simetrično drevo za oba pri simulacijah uporabljena evolucijska modela podali ustrezno drevo z izredno dobrimi podporami vsem notranjim vejam. Za drevo z izrazito dolgimi zunanji vejami sta obe metodi podali zelo nizko podporo eni od notranjih vej; ta združuje štiri dolge zunanje veje z dvema izrazito kratkima notranjima vejama. Metoda PTM je ustrezajočemu razvejišču (razvejišče 10, glej Sliko 9) določila podobno vrednost podpor PAS in rPAS kot ostalim hierarhično višjim razvejiščem iste topologije, kar morda pomeni, da je metoda manj občutljiva na izrazite razlike v dolžinah posameznih vej v kladu, ki ga želimo ovrednotiti. Pri analizi naborov sekvenc, simuliranih na podlagi modela substitucij JC69, sta ML in Bayesov pristop slabo podprli dve notranji veji; ustrezajoči razvejišči (razvejišči 7 in 11; glej Sliko 11) sta s strani metod PTM prejeli ustrezno visoki vrednosti PAS in rPAS v primerjavi z ostalimi razvejišči iste topologije. Metoda PTM torej v za ostale filogenetske metode problematičnih razvejiščih dosega podobne vrednosti rPAS kot v neproblematičnih razvejiščih.

PTM je tudi v veliki meri neodvisna od substitucijskega modela, po katerem so bili pripravljeni nabori simuliranih sekvenc. Za enostaven model JC69, ki predvideva enakomerno razporeditev substitucij nukleotidov po celotni dolžini sekvenc, smo s PTM za vse tri topologije za hierarhično višja prejeli zelo podobne vrednosti rPAS kot za bolj kompleksen model GTR. Slednji predpostavlja različne hitrost substitucij za vsak par nukleotidov in različne ravnotežne frekvence posameznih nukleotidov v sekvenci. Pri tem modelu pride do nižjega števila skupnih sprememb znakov po celotni dolžini simuliranih nukleotidnih zaporedij, zato prejmejo nabori sekvenc, simulirani po tem modelu, močno nižjo vrednost celokupne PAS za vsako od razvejišč topologije, izračunane rPAS pa so za vse obravnavane topologije podobne kot pri substitucijskem modelu JC69.

Izračunali smo še podpore rPAS posameznim razvejiščem dveh različnih topologij z uporabo naborov sekvenc, simuliranih na podlagi ene od teh topologij. Ugotovili smo, da s PTM za drevo, ki topologiji, na podlagi katere so bili simulirani v analizi uporabljeni nabori sekvenc, izračunamo precej višje rPAS za celotno topologijo kot za drevo, ki ima glede na izhodiščno drevo popolnoma drugačno topologijo (primerjaj Sliki 14 in 16). Nabori simuliranih podatkov namreč nosijo filogenetski signal (informacijo o sinapomorfijah), ki lahko podpre klade s podobno razporeditvijo končnih taksonov kot pri pravem filogenetskem drevesu za ta nabor sekvenc, kladov z izrazito drugačno razporeditvijo končnih taksonov pa ne more podpreti.

Ta diplomska naloga je prvi poskus ovrednotenja delovanja probabilistične taksistične metode. V njej smo pokazali, da ta metoda omogoča izračun podpore hipotetičnim topologijam dreves, saj omogoča ločevanje med naključnim šumom in filogenetskim signalom na višjih hierarhičnih ravneh kladograma. Metoda bi se lahko izkazala za uporabno v primerih, ko ostale filogenetske metode odpovedo, npr. pri privlaku dolgih vej ali pri setih podatkov z zelo nizkim apomorfnim signalom in visokim homoplastičnim šumom. Na kakšen način lahko to značilnost metode uporabimo za ovrednotenje ustreznosti posameznih topologij pa še ni jasno. Potrebno bo ovrednotiti rPAS kot možno ustrezno mero podpore posameznim razvejiščem. Da bi pa določili popolnoma ustrezno

mero podpore, bi bilo treba raziskati, kako se razporeditev stanj vsakega znaka v nukleotidnem zaporedju odraža v izračunu celokupne PAS. To pa zahteva poznavanje prispevka k podpori ter razporeditve stanj znaka za vsak znak nukleotidnega zaporedja.

5 SKLEPI

PTM v večini primerov ne poda pravilne topologije kladogramov. Celokupna verjetnost apomorfne podprtosti za vse klade torej ni ustrezna mera za odbiro pravega kladograma.

Na podlagi celokupne verjetnosti apomorfne podprtosti (PAS) ne moremo nedvoumno predvideti, ali je ta vrednost posledica šuma ali pa sinapomorfnega signala. To pa lahko do neke mere storimo s primerjanjem vrednosti relativne PAS (rPAS), ki upošteva število predvidoma sinapomorfnihih znakov (znakov s $pas > 0,5$). Ta je pri naborih sekvenc, ki nosijo sinapomorfni signal, za vsa razvejišča - razen hierarhično najnižjih - značilno večja kot pri naključno generiranih sekvencah, kjer se enakost stanj znakov med taksoni pojavlja naključno.

S probabilistično taksično metodo lahko na podlagi verjetnosti sinapomorfij izmerimo filogenetski signal, vsebovan v naborih nukleotidnih zaporedij. Značilno podporo vsem hierarhično višjim razvejiščem prejmemo tudi za nabore podatkov, ki nosijo le malo filogenetskega signala (imajo dolge zunanje veje in kratke notranje), čeprav je ta precej nižja kot pri naborih podatkov z manj šuma. Metodo bi morda lahko uspešno uporabili v primerih, ko nabori podatkov nosijo veliko homoplastičnega šuma in malo filogenetskega signala, ter takrat, ko ne moremo ustrezno določiti modela evolucije, kar pa moramo potrditi z nadaljnjimi raziskavami.

6 POVZETEK

Utemeljitelj filogenetske sistematike Willi Hennig je predvidel, da je treba za določitev filogenij razločiti med apomorfni in pleziomorfni stanji znakov, saj lahko le na podlagi prvih sklepamo na skupni izvor organizmov. V tem delu obravnavana filogenetska taksična metoda (PTM) temelji samo na osnovnih Hennigovih načelih. Za mero podpore določenemu kladu uporablja verjetnost apomorfne podpore (PAS – probability of apomorphic support) znakov, ki se izračuna izključno na podlagi frekvenc stanj znaka v kladu ter izven klada, polarnost znakov pa določi na podlagi a priori koreninjenja drevesa. S tem postopkom se izognemo potrebi po upoštevanju dodatnih predpostavk, ki bremenijo ostale filogenetske metode.

Namen raziskave je bil pridobiti osnovno poznavanje uporabnosti nove probabilistične taksične metode za filogenetske analize. V prvem delu poskusa smo preverili, če lahko metodo uporabimo za določitev pravega kladograma, in jo primerjali z uspešnostjo drugih filogenetskih metod na istih naborih podatkov. V nadaljevanju poskusa smo želeli spoznati, kako se filogenetski signal odraža v podpori določeni topologiji drevesa, obenem pa smo želeli preveriti, kako se metoda obnaša v primerih, v katerih pogosto uporabljane filogenetske metode odpovedo.

Ugotovili smo, da metoda PTM v večini primerov izbere napačen kladogram, zato je za ta namen neuporabna. Pri določanju vrednosti PAS posameznim razvejiščem poznanih topologij pa smo ugotovili, da gre za ustrezen način izračuna podpore tem razvejiščem. Tudi v primeru drevesa z izrazito dolgimi zunanjimi vejami v primerjavi z notranjimi smo po metodi PTM uspeli izračunati sicer nizko, a značilno podporo za vsa razvejišča razen hierarhično najnižjih. Metodo bi morda lahko uspešno uporabili v primerih, ko podatki nosijo veliko homoplastičnega šuma, ali pa ko ne moremo ustrezno določiti modela evolucije.

VIRI

- Assis L., Brigandt I. 2009. Homology: Homeostatic Property Cluster Kinds in Systematics and Evolution. *Evolutionary Biology*, 36, 2: 248-255.
- Bang R., DeSalle R., Wheeler W. 2000. Transformationalism, Taxism and Developmental Biology in Systematics. *Systematic Biology*, 49, 1: 19-27.
- Brady R.H. 1985. On the independence of systematics. *Cladistics*, 1, 2: 113-126.
- Brooks D.R., Wiley E.O. 1986. *Evolution as entropy: Toward a unified theory of biology*. Chicago, Chicago University Press: 335 str.
- Brower A.V. 2000. Evolution Is Not a Necessary Assumption of Cladistics. *Cladistics*, 16, 1: 143-154.
- Carine M.A., Scotland R.W. 1999. Taxic and transformational homology: Different ways of seeing. *Cladistics*, 15, 2: 121-129.
- Cao N., Zaragüeta Bagils R, Vignes-Lebbe R. 2007. Hierarchical representation of hypotheses of homology. *Geodiversitas*, 29, 1: 5-15.
- DeSalle R. 2006. What's in a character. *Journal of Biomedical Informatics*, 39, 1: 6-17.
- de Pinna M.G. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics*, 7, 4: 367-394.
- De Queiroz K., Gauthier J. 1990. Phylogeny as a central principle in taxonomy: Phylogenetic definitions of taxon names. *Systematic Zoology*, 39, 2: 307-322.
- Farris J.S., Kluge A.G., Eckhardt M.J. 1970. A numerical approach to phylogenetic systematics. *Systematic Zoology*, 19, 2: 172-191
- Farris J.S., Kluge A.G., Laet J.E. 2001. Taxic Revisions. *Cladistics*, 17, 1: 79-103.
- Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27, 4: 401-410.
- Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17, 6: 368-376.
- Grant T., Kluge A.G. 2004. Transformation Series as an Ideographic Character Concept. *Cladistics*, 20, 1: 23-31.

- Hagstrom G.I., Hang D.H., Ofria C., Torng E. 2004. Using Avida to Test the Effects of Natural Selection on Phylogenetic Reconstruction Methods. *Journal of Artificial Life*, 10, 2: 157-166.
- Hennig W. 1966. *Phylogenetic Systematics*. Urbana, Univ. Illinois Press: 284 str.
- Huelsenbeck, J.P., Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17, 8: 754-755.
- Jenner R.A. 2004. The scientific status of metazoan cladistics: why current research practices must change. *Zoologica scripta*, 33, 4: 293-310.
- Kim J. 1993. Improving the accuracy of phylogenetic estimation by combining different methods. *Systematic Biology*, 42, 3: 331-340.
- Kluge A.G. 1997. Testability and the refutation and corroboration of cladistic hypotheses. *Cladistics*, 13, 1-2: 81-96.
- Kluge A.G., Farris J.S. 1999. Taxic homology = overall similarity. *Cladistics*, 15, 2: 205-212.
- Nelson G. 1989. Cladistics and evolutionary models. *Cladistics*, 5, 3: 275-289.
- Nelson G., Platnick N.I. 1991. Three-taxon statements: a more precise use of parsimony? *Cladistics*, 7, 4: 351-366.
- Page R.D., Holmes E.C. 1998. *Molecular Evolution: A Phylogenetic Approach*. Oxford, Blackwell Science Ltd: 346 str.
- Phillips A.J. 2006. Homology assessment and molecular sequence alignment. *Journal of Biomedical Informatics*, 39, 1: 18-33.
- Platnick N.I. 1979. Philosophy and the transformation of cladistics. *Systematic Zoology*, 28, 4: 537-546.
- Popper K.R. 1959. *The Logic of Scientific Discovery*. New York, Basic Books: 480 str.
- Popper, K.R. 1979. *Objective Knowledge – An Evolutionary Approach*. Oxford, Clarendon Press: 390 str.
- Scotland R.W. 2000. Taxic Homology and Three-Taxon Statement Analysis. *Systematic Biology*, 49, 3: 480-500.
- Shannon C. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27, 370-424 in 623-656.

Siddall M.E., Kluge A.G. 1997. Probabilism and phylogenetic inference. *Cladistics*, 13, 4: 313-336.

Simmons M.P., Webb C.T. 2006. Quantifications of the success of phylogenetic inference in simulations. *Cladistics*, 22, 3: 249-255.

Swofford D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Trontelj P. 2009. "Probabilistična taksična metoda". Ljubljana, Univerza v Ljubljani, Biotehnična fakulteta, Oddelek za biologijo (osebni vir, avgust 2009).

Wägele J.W. 2005. Foundations of Phylogenetic Systematics. München, Verlag dr. Friedrich Pfeil: 365 str.

Wiley E.O. 1975. Karl R. Popper, systematics and classification: A reply to Walter Bock and other evolutionary systematists. *Systematic Zoology*, 24, 2: 233-243.

Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 8: 1586-1591.

Zujko-Miller C., Miller, J.A. 2003. PEST: Precision estimated by sampling traits.

Zwickl D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Doktorska disertacija. Austin, University of Texas.

ZAHVALA

Mentorju izr. prof. dr. Petru Trontlju se zahvaljujem za vse koristne nasvete, optimizem in stalno spodbujanje k izvedbi poskusov in pisanju diplomske naloge.

Zahvaljujem se Martinu Turjaku za programersko podporo in debate.

Hvala staršema in bratu za potrpljenje in podporo tekom študija.