

Biologický korespondenční seminář



Biozvěst

Ročník 0

Serie 4

Milí řešitelé,

Máme tu poslední sérii nultého ročníku Biozvěstu. Doufám, že Vás úlohy opět překvapí. A doufám, že spíše obsahem, než obtížností. V seriálu na pokračování završíte Vaše samostudium bioinformatiky a směle se budete moci nazývat „pokročilými“. Omlouvám se Vám za dvou-týdenní zpoždění při vydání série, ale na odškodněnou je uzávěrka prodloužena až do 8.5. Slavnostní vyhlášení výsledků proběhne během terénního soustředění a vítězové se mohou těšit na sladkou odměnu.



TERÉNNÍ SOUSTŘEDĚNÍ

Koná se od **16. do 21.5.2013 (středa - úterý)**

Účastníci se budou vybírat dle dosažených bodů za 1. - 3. sérii a měly by získat více než 0 bodů a řešit alespoň 2 série (počítá se i čtvrtá).

Hlaste se do konce dubna na adrese

biozvest@gmail.com

Do přihlášky napište, zda-li se budete účastnit celou dobu nebo specifikujte, od kdy do kdy přijedete. Po 1.5. vybereme účastníky a pošleme jim podrobné informace.

Místo: Čtvrtníkův srub <http://araneus.skauti.net/srub/>

Leží u Vacova v Pošumaví. Není zde elektřina, tudíž bude akce mimořádně autentická.

Náplň: Výpravy do okolí, rekognoskace orchidejových luk mimořádné kvality, determinace nalezených organismů (zejména budeme zkoušet práci s determinacním klíčem), diskuse nad tématy úloh Biozvěstu aj.

Strava: Z doneseného či uloveného připravena svépomocí.

Ubytování: Ve spacácích. PřF UK na akci přispívá 5 000 Kč, což pokryje náklad na ubytování a zbude něco málo na stravu. Dopravu si každý hradí sám.



Jak řešit

Veškeré pokyny k řešení semináře získáte na internetové stránce Biozvěstu

<http://web.natur.cuni.cz/~vosolsob/krouzek/bios.html>

(nebo zadejte „Biozvěst“ do Google). Nejdříve je třeba přihlásit se ke Google skupině „Řešitelé Biozvěstu“

Biozvest-resitele@googlegroups.com,

<https://groups.google.com/d/forum/biozvest-resitele>, vyplnit přihlášku a následné odesílání úloh se provádí prostřednictvím služby „Disk“ (dřívější „Dokumenty“) na Google.

Kdybyste měli jakýkoliv problém s uvedeným postupem, pošlete řešení na adresu biozvest@gmail.com.

V případě, že byste se ocitli bez internetu, můžete využít i klasickou poštu

Stanislav Vosolsobě

Katedra experimentální biologie rostlin

Přírodovědecká fakulta Univerzity Karlovy v Praze

Viničná 5

128 44 Praha 2

opravené řešení však dostanete naskenované e-mailem.

Nelekejte se, když Vám přijdou úlohy na první pohled příliš těžké, ponořte se do informačních zdrojů a uvidíte, že na vše lze někde nalézt odpověď. Dobré tipy k řešení naleznete také na stránce Biozvěstu v sekci „Návody“. Není nutné abyste vyřešili všechny úlohy, stačí odeslat libovolně velký fragment. Oceníme, pokud přiložíte jakékoliv připomínky (např. úloha byla příliš lehká/těžká, nesrozumitelná, nudná), úlohy se pokusíme tvořit k Vaší maximální spokojenosti.

Veškeré dotazy či připomínky směřujte na adresy

biozvest@gmail.com či vosolsob@natur.cuni.cz

Uzávěrka 4. série proběhne ve středu 8.5.2013 o půlnoci.

Úlohy můžete vypracovat přímo do zadání jednotlivých úloh, které se objeví ve Vašich sdílených dokumentech. Hoto-ovou úlohu pojmenujte

Ročník-Série-Úloha-Jméno_Příjmení,

např. **0-1-2-Bioslav_Biomilný** v případě druhé úlohy první série aktuálního ročníku. Poslední Vámi provedená změna by měla být ze dne uzávěrky. V případě opožděného odevzdání úloh se strhává za každý celý den jeden bod s výjimkou zvláště závažných situací. Z technických důvodů ukládejte úlohu ve formátu .doc, abychom Vám mohli přidávat do řešení komentáře.

Neboť nelze systémově blokovat kooperaci mezi řešiteli a navíc kooperace je základem úspěchu při vědecké práci, akceptujeme i skupinově řešené úlohy, v názvu a hlavičce úlohy však vyjmenujte všechny řešitele a tito řešitelé mohou odevzdat dohromady pouze jednu úlohu, všem bude přičten identický počet bodů. Případné spoluautory mimo řešitelů semináře taktéž uvádějte do hlavičky úlohy. Zdroje informací taktéž do řešení připište.

V každé sérii se můžete těšit na jednu „zelenou“, jednu „bílou“, studijní, praktickou a naučnou úlohu, ta naučná je v tomto ročníku věnována bioinformatice.

Mnoho zdaru při řešení Vám za kolektiv autorů přeje

Stanislav Vosolsobě

Úloha 1: Zmrzlinový pohár

Autor: Stanislav Vosolobě

Počet bodů: 20

Bioslav byl koncem zimy na lyžařském kurzu v Tatrách a náhodou slavil svátek (30.2.). Spolužáci mu připravili zmrzlinový pohár. Neboť byly na horách odříznuty od civilisace, museli si vystačit s tím, co okolní příroda dala a tak použili k přípravě poháru různobarevný sníh z přírody. Pohár se skládal z těchto složek, které jsme pracovníě pojmenovali podle vzhledu (Bioslav snědl vždy jen část a zbytek podrobil analýze):

Stracciatella

Obsahovala tisíce drobných tmavých organismů, které od pohledu patřili do skupiny Pancrustacea. Bioslav si již dříve všiml, že tyto živočichové poskakují všude po sněhu v obrovských počtech.

Melounová

Pod mikroskopem byly patrné jednotlivé oranžové buňky. Bioslav izoloval ze zmrzliny pigmenty, rozdělil je chromatograficky a určil absorpční spektrum. Identifikoval dvě látky s absorpčními maximy 430 a 663 nm, respektive 463 a 648 nm a velké množství látky s absorpčními mezi 450 a 500 nm. Jiný zástupce tohoto rodu je již delší dobu kompletně osekvenován.

Citronová

Bioslav měl silné tušení hned od prvního pohledu. Pod mikroskopem nic nepozoroval. Nechal vzorek odpařit a získal bezbarvou látku tvořenou jehličkovitými krystaly. Svě tušení potvrdil poté, co trochu látky natavil nad vařičem a poté přidal k Fehlingovu činidlu - vzniklo temně modré zabarvení. Při zahřívání roztoku nenatavené látky ve Fehlingově činidle se nic nestalo.

Čokoládová

Pod mikroskopem Bioslav spatřil neidentifikovatelnou změť. Potom preparát obarvil barvivem DAPI a pod fluorescenčním mikroskopem s ultrafialovou excitací spatřil veliké množství drobných modrých teček (objektiv 60x) a občas i větší kolečka. Bioslav zajásal, provedl izolaci DNA a amplifikoval různé mitochondriální geny pomocí PCR. Toto zjistil sekvencí:

>sekvence 1

```
AACTATTCCCTGATCTCCTCTCCCCACATTTTAATTCATATAT
TTAATAACATCTACTGTACCTCCCCAGTATGTACTCTTTCCAC
CCCTATGTATATCGTGCATTATGGTTTGGCCCATGCATATAA
GCATGTACATACTATGCTTGATTTTGCATTCGTGCACCTCACCT
AGACCACGAGCTTGATCACCAGCCTCGAGAAACCATCAACCTT
TGCTAGACGTGTACCTCTTCTCGCTCCGGGCCATAGCATGTGG
GGGTTTCTAGCCTGAACTATACCTGGCATCTGGTTCTTACTTC
AGGGCCATGAAAGTCTCAATCCAATCCTACTAACCTCTCAAAT
GGGACATCTCGATGGACTAATGACTAATCAGCCCATGATCACAC
ATAACTGTGGTGCATGCATTTGGTATTTTTTAATTTTATAGGGG
GGGGGGACTGGTATCACTCAGCTATGGCCGTAAGGCCCTCGTAG
CAGTCAAATAACTTGTAGCTGGGCTTATCCTTCATCATT
```

>sekvence 2

```
ATGACAAACATTCGAAAAACACCCCCCTACTAAAAATTATTA
TCACTCTTTTCATCGACTTACCAGCCCCATCCAATATCTCATCAT
GATGAAACTTTGGCTCCCTCTTAGGAATCTGCCAATAATCCAA
ATCCTTACAGGTCTATTTCTAGCAATACACTATACATCAGACAC
AATAACAGCATCTCTTTCAGTAACCTCATATCTGCCGAGACGTAA
ATTATGGATGACTAATTCGATATTTACACGCAACGGAGCCTCA
ATATTCTTCAATTTGCTTATTTTACATGTAGGACGAGGAATATA
TTACGGATCATATACCTTCATAGAAACATGAAACATCGGTGTAA
TTCTCCTATTTCGAGTAATAGCCACAGCATTCATAGGTTATGTC
```

```
CTTCCATGGGGACAAATATCCTTCTGAGGAGCAACAGTAATTAC
AAATCTATTATCAGCAATTCCATACATCGGAACACCCTAGTAG
AATGAATCTGAGGAGGATTCTCAGTAGATAAAGCTACACTAACA
CGTTTCTTCGCCTTCCACTTTATCTTTACCATT
CATCATTTGCCGCCCTAGTAATCGTCCATCTCCTATTTCTCCATG
AAACTGGATCAAATAACCCTACAGGCCTTAACCTCAGACGCCGAC
AAAATTCATTTTACCCTGACTATACAAATAAAGACATTCTAGG
AGTACTCATAATAGTGTCTATTCCTAATAACCTTAGTTCTCTTCT
TCCCAGACCTTCTAGGTGACCCGGACAACATATACCTGCCAAC
CCACTAAATACCCACCACATATTAACCAGAATGGTACTTCTCT
ATTTGCATACGCAATTCTACGATCCATCCCTAACAAATTAGGTG
GAGTCTAGCCCTAATTTTATCTATCCTTATTCTAGCCCTATTA
CCATTCTACACACCTCTAAACAACGCAGCCTAATATTCCGTCC
AATCACTCAAACCTTACTGAATCCTAGTTGCTAACCTTCTCG
TCCTAATTGAATTGGAGGTCAACCAGTAGAACACCCATTTCATC
ATCATTGGCCAACTAGCCTCAATCAGCTACTTCTCTATTATTCT
AATCCTCATACCCATCTCAGGCATTATTGAAG
ATAAAATA
```

Limetková

Tentokrát Bioslav našel organismu příbuzný organismu z melounové zmrzliny, který se i v přírodě vyskytuje v podobném prostředí. Akorát nebyly útvary kulaté a oranžové, ale zelené rohlíčkovité s dlouze zašpičatělými konci.

1. Rozeberte podstatu všech výše zmíněných důkazů a identifikujte složení všech pěti zmrzlin.
2. Popište, jak vzniká limetkový, melounový a stracciatellový sníh, kde se organismy ve sněhu berou. Co by to znamenalo, kdyby po zahřívání i nenatavené látky z citronové zmrzliny došlo k reakci (červená sraženina)?
3. Jak první umělá syntéza hlavní látky z citronové zmrzliny změnila náhled na živý svět?
4. Co jsou nejpravděpodobněji malé tečky a větší kolečka obarvená DAPI v čokoládové zmrzlině?
5. Zdůvodněte přítomnost dvou zmíněných sekvencí v čokoládové zmrzlině (při identifikaci sekvencí využijte zkušenosti z 5. úlohy 2. série).

Úloha 2: Sláva Mendelovi!

Autor: Jiří Hadrava

Počet bodů: 12

Předpokládejme dostatečně velkou panmiktickou populaci diploidních jednoletých rostlin. Tyto rostliny mají gen *A* ve formě tří různých alel. Alela *A1* má frekvenci 60%, *A2* 30% a *A3* 10%. Enzym kódovaný genem *A* je životně důležitý. Recesivní alela *A3* je poškozenou variantou alel předchůzích. Rostliny, které jsou recesivními homozygoty, nevyklíčí. Rostliny, které mají alespoň jednu alelu *A1* nebo *A2* nejsou vůči sobě nijak selektivně znevýhodněny. Alela *A2* se od alely *A1* liší mutacemi v promotoru a neexprimuje se v květech. Gen *B* má dvě alely, *B1* a *B2*. Každá z alel má v populaci frekvenci 50% a není proti nim žádný selekční tlak. Tento gen neleží na stejném chromosomu s genem *A*. Alela *B2* je recesivní, její produkt je nefunkční. Enzym kódovaný alelou *B1* přeměňuje produkt enzymu *A* v pigment způsobující červené zbarvení.

1. Kolik procent rostlin v populaci bude mít červené květy?
2. Jaká je pravděpodobnost, že náhodně vybraná rostlina nebude nést alelu *A2*?
3. Jaká bude v této populaci heterozygotnost za dva roky?

4. Jaké procentuální zastoupení alel *A1*, *A2* a *A3* najdeme v populaci za 10 let? (Doporučuji řešit pomocí Excelu nebo podobného tabulkového editoru.)
5. Jsou alely *A* vzájemně orthology, nebo paralogy?

Úloha 3: Nový, dosud neznámý druh

Autor: Albert Damaška

Počet bodů: 10

Bioslav rád loví v terénu myšice. Ty mají ve svém kožíšku plno blech, pročež se Bioslav často baví i chytáním těchto skákajících potvůrek. V tomto koníčku je už pěkně zkušený, dělá ho už dlouho. Když byl v roce 2007 na Muránské planině na Slovensku, chytil v trávě vedle uskakujících blech také jednoho dřepčíka. Zkušeně v něm už tehdy poznal zástupce rodu *Dibolia*. Jaké bylo ale jeho překvapení, když se doma ukázalo, že jde o nový, dosud neznámý druh. Popsat ho bohužel nestihl. V tomtéž roce totiž stejného dřepčíka popsal ze sběrů, provedených na Slovensku o několik let dříve, entomolog Čížek. Bioslav se ale rozhodl, že si popis druhu zkusí alespoň nanečisto. Tehdy netušil, kolik s tím bude mít práce.

Entomolog Čížek v roce 2007 nový druh dřepčíka rodu *Dibolia* ze Slovenska popsal v časopise *Klapalekiana*. Kdo tento časopis vydává, jaký je impakt faktor tohoto časopisu a jak se jmenuje onen nový druh dřepčíka, jehož popis v časopise vyšel?

Možná by někoho mohlo udivit, že mohl Bioslav najít nový druh pro vědu i v tak prozkoumané oblasti, jako je střední Evropa, ba dokonce Slovensko. Nové druhy se však ze střední Evropy stále popisují, ba co více, popisují se i z České republiky. Najděte alespoň dva druhy živočichů, popsáné ve 21. století z území České republiky.

Aby v popisech druhů a v jejich názvech nebyl úplný guláš, koordinují tyto aktivity *Mezinárodní pravidla zoologické nomenklatury*, která vydává Mezinárodní komise zoologické nomenklatury ICZN. Tato pravidla jsou online k dispozici v anglickém jazyce na stránce <http://www.nhm.ac.uk/hosted-sites/iczn/code/>, v českém překladu je vydala knižně Česká společnost entomologická. Kdy vyšlo zatím poslední vydání Pravidel v anglickém jazyce a kdy v jazyce českém?

Aby mohl být popis druhu platný, je od roku 2000 nutné, aby byl vedle publikovaného popisu k dispozici také typový exemplář, který se nazývá holotypus. To je jeden jedinec, podle kterého byl popis druhu učiněn. S holotypy se často můžete setkat, budete-li nahlížet do muzejních entomologických sbírek. Jedinec tam bývá označen červeným lístkem s nápisem holotypus. Vedle holotypu se však ve sbírkách často setkáte i s červenými lístky s nápisem *paratypus*. Jaký je rozdíl mezi holotypem a paratypem?

Představme si, že by se Bioslav nedozvěděl o tom, že by stejný druh, který našel, už někdo popsal, a popis nového druhu by v roce 2008 publikoval. V tu chvíli by na světě byly dva popisy stejného druhu! To se vzhledem k velkému množství odborných časopisů stává celkem často. Pokud nezávisle na sobě vyjde více popisů stejného druhu, který je platný?

Když v literatuře vidíme název nějakého druhu, obvykle sestává jen z rodového a druhového jména (např. *Argopus ahrensi*), ale za druhovým jménem je uvedeno ještě jakési

jméno a jakýsi letopočet (např. *Argopus ahrensi* Germar, 1817). Někdy navíc za názvem není uvedeno žádné jméno, ale pouze zkratka L. (např. *Apodemus sylvaticus* L. 1758). Či jméno a jaký letopočet se za názvy druhů uvádí a co znamená zkratka L.?

Když byl Bioslav v Mediteránu, doslechl se, že v macchii může snadno narazit na takové věci, jako je *Bacillus* a *Iris*. Bylo mu řečeno, že *Bacillus* se před ním bude snažit schovat, zatímco *Iris* ho bude chtít kousnout kusadly. „To je divné,“ pomyslel si, „že by se přede mnou chtěla schovávat infekční prokaryota a že by mi hrozilo kousnutí od mokřadní kytky v suchém křoví!“ Později však zjistil, že obojí je v macchii možné. Jak to, že se živočich může jmenovat stejně jako prokaryot nebo stejně jako rostlina? A do jakých skupin zmínění živočichové patří?

Úloha 4: Nedotýkati se!

Autor: Stanislav Vosolsobě, inspirace Miloš Duchoslav

Počet bodů: 15

Rostliny vnímají mnoho podnětů ze svého okolí. Mimo jiné reagují i na dotek, ať již je způsoben větrem, půdou či jinými organismy. Pokud byly rostliny huseníčku dvakrát denně vystaveny dotyku, mělo to velmi dramatický vliv na jejich růst, vizte obrázek.



Obrázek je převzat z přehledného článku

<http://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.2004.01263.x/full>

1. Pokuste se reprodukovat tento pokus s jinými dostupnými rychle rostoucími druhy rostlin. Popište podrobně metodiku a nezapomeňte na statistické zhodnocení, v tomto případě byste měli vystačit i s t-testem, který spočítá i Excel... Pokud by vám rostliny rostly pomalu, je tuto úlohu možné odevzdat až do 14.5.

Úloha 5: Strom

Autor: Stanislav Vosolsobě

Počet bodů: 12

V minulé sérii jsme si připravili mnohočetné přiřazení sekvencí. Nyní celý keratinový příběh uzavřeme výpočtem fylogenetického stromu. Jak již jsem zmínil při výrobě alignmentu, nejodpovědnější fylogenetickou práci jsme provedli při vzájemném přiřazování aminokyselinových posic, kdy jsme se rozhodli, které posice budeme považovat za homologické. Vlastní tvorba stromu spočívá pouze v seskupení sekvencí na základě kritéria, podle něhož hodnotíme podobnost sekvencí. Existuje však nepřehledné množství výpočetních strategií. Pokusím se je nastínit jen v nejhrubších rysech:

Distanční algoritmické metody

Vezmeme sekvenční celek a spočítáme vzájemné podobnosti mezi všemi sekvencemi. Vytvoříme matici distancí (tabulku vzdáleností) mezi sekvencemi. Nejjednodušším kritériem pro výpočet podobnosti v případě nukleotidové sekvenční je počet shodných posic. Takto vypočítaná distance mezi sekvencemi by odpovídala evoluční vzdálenosti mezi nimi za předpokladu, že by všechny posice mutovaly náhodně, ke všem typům mutací by docházelo stejně často a konkrétní posice mutovala jen jednou. Ve skutečnosti je evoluce sekvencí složitější a pro výpočet distancí se používá složitější vzorec, **substituční model**. Existuje mnoho různých modelů, které se liší počtem parametrů, které berou v úvahu, základní je korekce na zpětné mutace, to jsou případy, kdy např. A mutuje na C, G a pak zase zpět na A a po dlouhé době dojde k tzv. substituční saturaci, kdy je původní signál zcela překryt mladšími mutacemi. Pravděpodobnost zpětných mutací lze přesně vyčíslit. Dále můžeme zavést korekci na mutační pravděpodobnost (empiricky zjistíme, že např. častěji mutuje A na T, než A na C a tak budeme brát sekvenční, kde se vyskytuje záměna A/C za vzdálenější, než sekvenci se záměnou A/T). Každá část sekvenční také může mutovat odlišnou rychlostí a naopak některé posice mohou být neměnné (invariantní), proto se v modelech můžeme setkat s parametrem G, který jednoduše řečeno říká, kolik rozdílných mutačních rychlostí v rámci zkoumané sekvenční máme brát v úvahu, a parametr I, který znamená proporcii invariantních posic. Zvláště složitý substituční model je potřeba pro proteinové sekvenční, kdy je nutné brát v úvahu pravděpodobnosti záměn mezi 20 aminokyselinami. Substituční model se nejčastěji počítá z dat samotných na základě frekvencí konkrétních substitucí. Ne vždy znamená, že komplexnější model je lepší, neboť čím složitější model použijeme, tím musíme odhadnout více parametrů. Složitý model proto lze použít pouze v případě, že máme dostatečný počet dlouhých sekvencí.

Pokud máme určeny vzdálenosti, provedeme na jejich základě klastrování (shlukování) sekvencí. Primitivnější metoda **UPGMA** začíná seskupovat od větví, které mají nejmenší vzdálenost. Po seskupení každého páru sekvencí jsou jejich distance nahrazeny průměrnou hodnotou. Pokročilejší metoda **neighbour-joining (NJ)** seskupuje větve podle distancí tak, aby se po každém kroku zkrátila délka celého stromu. Metoda UPGMA se nedokáže vypořádat s divnými rychle mutujícími taxony, které mají velkou distanci od sesterských druhů. Místo, aby je spojila se sesterskými, přikládá je ke kořeni stromu (artefakt LBA - long branch attraction). Výsledný strom pak nesplňuje parsimonické kritérium nejmenšího počtu evolučních změn (tj. strom s nejkratší délkou větví). Metoda NJ se s tímto artefaktem dokáže vypořádat, je však nutné, aby byly distance dobře spočítány, což může být problematické, pokud se sekvenční vyvíjí jinak, než předpokládá substituční model.

Heuristické znakové metody

Zatímco předchozí metody počítaly strom pomocí jasného algoritmu (např. postupně seskupovat větve od nejbližších), heuristika spočívá v prohledávání všech možných stromů dle daného kritéria a výběru toho nejoptimálnějšího. Zpravidla se při heuristickém hledání nepočítají distance, ale pracuje se přímo se znaky (tj. výskyt konkrétních bází či aminokyselin na konkrétní posici v alignmentu. Tradiční

metoda **maximum parsimony (MP)** pro každý hypotetický strom spočítá počet evolučních změn, ke kterým muselo dojít a vybírá se nejparsimonnější strom. Nebere v úvahu žádné substituční modely. Naopak **maximum likelihood (ML)** je založena na substitučním modelu. Metodu lze nejlépe pochopit na konkrétním příkladu:

- máme alignment tří sekvencí X, Y, Z
- metodou ML se hodnotí každá posice (sloupec) alignmentu separátně, mějme na první posici nukleotidy A, C, C
- generujeme všechny hypotetické topologie stromu:

$$X,(Y,Z) \quad (X,Y),Z \quad (X,Y),Z$$
 (sekvence mimo závorku je sesterská k sekvencím v závorce)
- pro každou topologii stromu generujeme všechny hypotetické kombinace nukleotidů, které mohly být v minulosti na každém z uzlů stromu. Pro každý strom to bude 16(!) možností (4 možné nukleotidy v kořeni stromu x 4 nukleotidy v prvním uzlu). Tudíž i když Y a Z mají shodně bázi C, operujeme s možností, kdy jejich nejbližší společný předek měl jinou bázi. Tím se značně lišíme od metody MP, která by operovala s hypotézou, že během evoluce došlo pouze k jedné záměně, a to konkrétně $A \leftrightarrow C$. Díky tomuto přístupu se metoda ML oprostila od artefaktu LBA.
- je zřejmé, že v případě konkrétní topologie musela evoluce proběhnout jednou z 16 možných cest. Pro každou z topologií můžeme určit kritérium likelihood (věrohodnost), které říká, jak pravděpodobně každá konkrétní topologie vysvětluje pozorované rozložení znaků a získá se jako součet pravděpodobností všech šestnácti cest.
Příklad uvažování: Pozorované rozložení znaků je, že Bioslav je nachlazen. Testujeme dvě různé topologie - nachladil se při pobytu doma / při terénním výzkumu. Konkrétní cesty k nachlazení doma by byly (v závorce jsou pravděpodobnosti, že se tento jev během dne stane): rozbilo se topení (1:200), zapomněl se obléknout, když šel z vany (1:7, Bioslav se myje jednou týdně a vždy zapomene), do okna vletěla ledová kometa (1:10²⁰). Nachlazení venku se mohlo přivodit takto: zapadl do závěje (1:1000), zabral se do pokusu (1:2), nedostatečně se oblékl z roztržitosti (1:3). Pravděpodobnost, že se Bioslav nachladil poruchou topení *nebo* neoblečením *nebo* kometou je rovna součtu dílčích pravděpodobností (tedy 0,14786), v případě druhé hypotézy je to 0,834. Vidíme, že naše data věrohodněji vysvětluje hypotéza venkovní topologie, tudíž bychom ji přijali. Také vidíme, že i když bereme v úvahu krkolomné evoluční cesty, přispívají k celkové věrohodnosti nepatrně.
- pravděpodobnost každé z evolučních cest počítáme na základě substitučního modelu
- na závěr kombinujeme výstupy ze všech znaků a vybíráme ten strom, který vyšel nejvěrohodněji pro všechny znaky.

Metoda ML je výpočetně nejnáročnější, ale dává zpravidla nejlepší výsledky bez LBA. Podmínkou je však odhadnout substituční strom. Tudíž pokud máme nedostatečná data, může ML selhat.

Která metoda je nejlepší?

Zpravidla nelze odpovědět obecně. Pro kvalitní datové soubory zřejmě ML, ale pro slabší data může být lepší i jedno-

dušší metoda. Osvědčeným postupem je použít více metod a porovnat výsledek. Pokud nějaká část stromu má oporu ve všech metodách, je možno ji věřit. Pokud poskytuje každá metoda zcela jiný výsledek, je nutné dělat závěry opatrně.

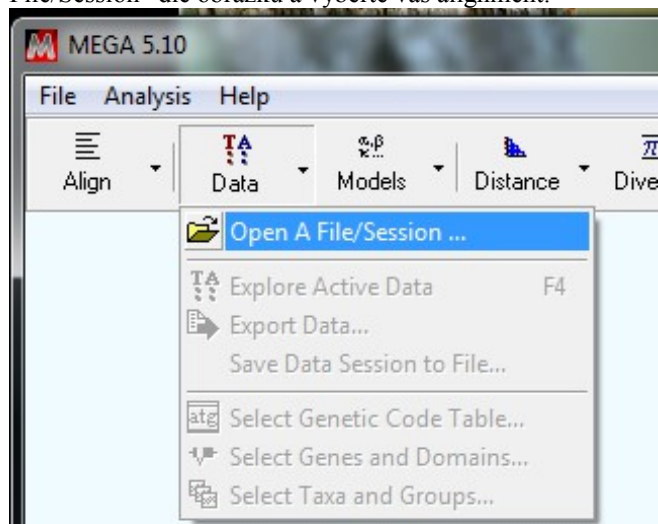
Bootstrap

Pro zhodnocení výsledku je nezbytné provést vám již známé bootstrapování, tj. např. v případě alignmentu sekvencí délky 100 bází vybereme náhodné sloupce alignmentu (s tím, že některé se opakují, jiné budou vynechány) a provedeme výpočet znovu. To opakujeme 100x - 1000x. Pokud ve většině případů vychází stejná topologie stromu, znamená to, že všechny posice poskytují konsistentní signál a náš alignment není zatížen konvergenčními. Pak můžeme topologii důvěřovat. Už jen výpočet stromu může trvat minuty. Pokud bootstrapování přepočítá strom stokrát, strávíte výpočtem notnou dobu. V historických dobách to byly i týdny.

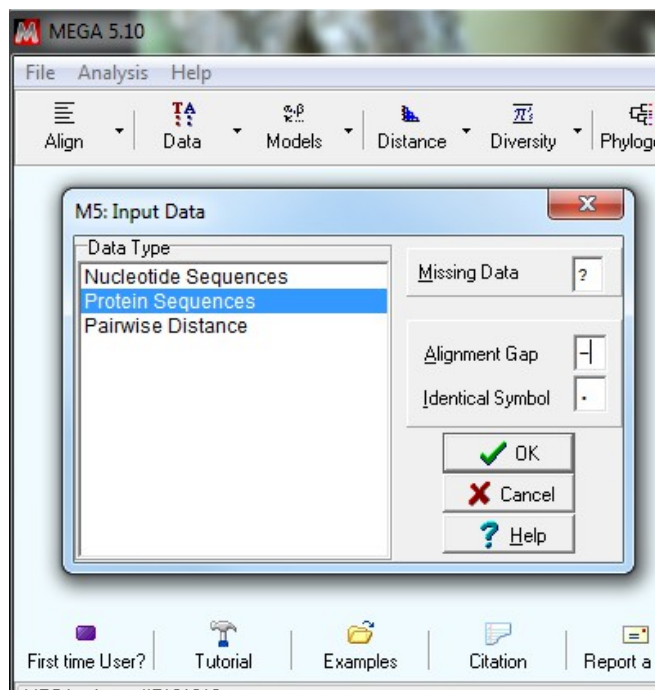
Pokud byste chtěli nastudovat více, doporučuji <http://web.natur.cuni.cz/~vlada/moltax/>

A nyní prakticky

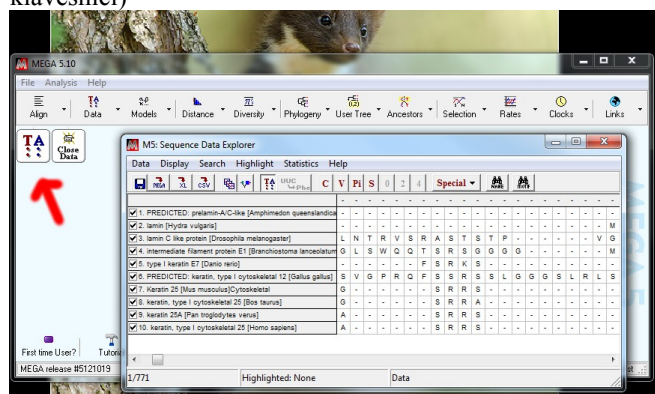
Nejdříve si stáhněte program MEGA, který je pro tvorbu stromečků nejpřívětivější. Do Google zadávejte heslo „mega sequence“, či <http://www.megasoftware.net/>. Když kliknete na stahovací logo pro Windows (omlouvám se Linuxovým uživatelům, podpora není, ale zkuste emulaci), požádá vás server o vaše údaje a stahovací odkaz přijde e-mailem. Vše je zdarma. Po instalaci (pouze odklikáte „Ano“ a „Vpřed“) spustíte program a nahrajte data alignmentu ve formátu .fasta. Klikněte na tlačítko „Data“ a „Open A File/Session“ dle obrázku a vyberte váš alignment.



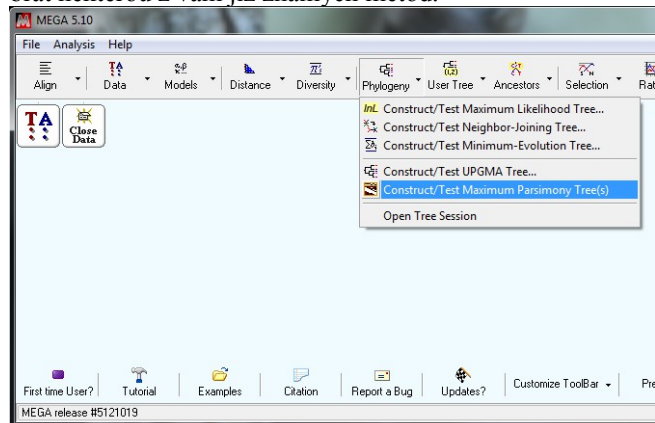
Dále musíte specifikovat, zda hodláte dělat alignment (to už máte hotovo, ale Mega to umí skvěle), či analýzu (to je náš záměr).



Následně specifikujte typ dat (DNA či proteiny) a hlavně typ symbolu pro mezeru v alignmentu. Buď to bude pomlčka či ~ (píše se ve Windows „pravý alt + 1“ na hlavní klávesnici)



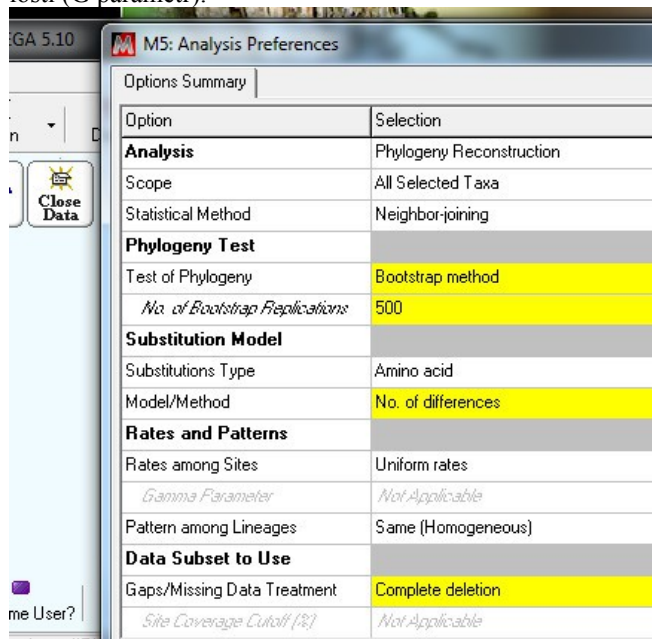
Klepnutím na ikonku dat si můžete prohlédnout alignment. Analýzu spustíte tlačítkem „Phylogeny“, kde si můžete vybrat některou z vám již známých metod.



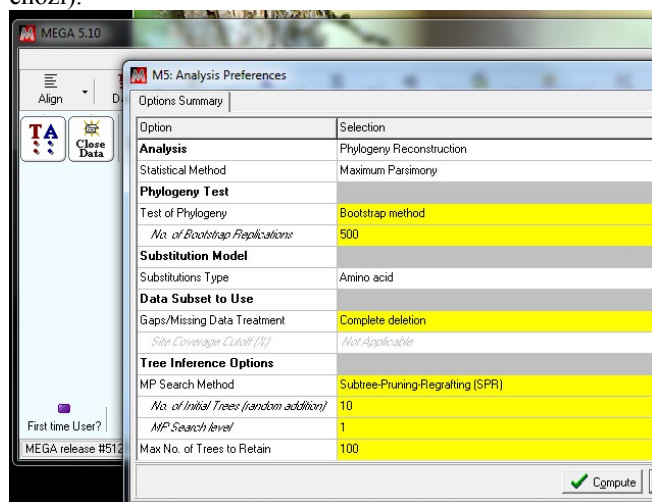
Program se vás vždy bude ptát na podrobnosti analýzy, které nastavujete ve žlutých políčkách. Vždy nastavte Bootstrap a hodnotu 500 (výpočet bude trvat poněkud déle, pokud byste měli 50 sekvencí a počítali ML, může to trvat i

den až dva. Raději nejdřív vypočítejte strom bez bootstrapu a pak zvolte bootstrap dle vašich výpočetních možností. Dále se vždy program ptá, co se sloupečky alignmentu, kde jsou mezery. Na mezeru nelze totiž snadno aplikovat substituční model a vyžaduje jiné vyhodnocení, což zneprůšňuje výpočet. Doporučuji kompletní delecí těchto posic.

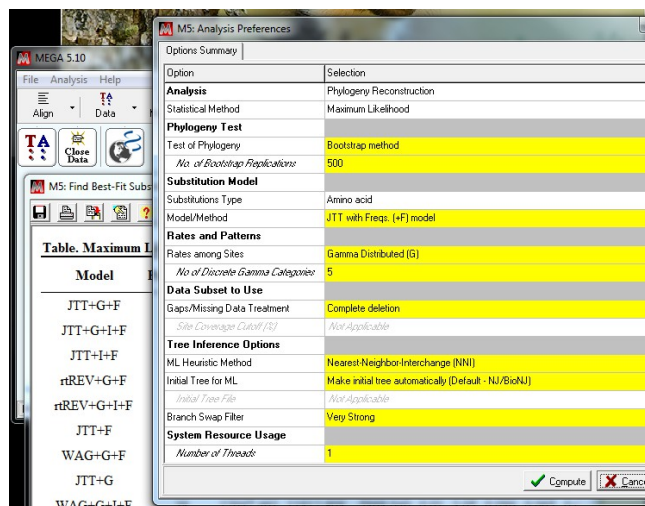
Pokud budete počítat NJ, můžete vybrat substituční model a v případě složitějšího modelu i počet různých mutačních rychlostí (G parametr).



Metoda MP nevyžaduje nastavit nic dalšího (lze nastavit parametry heuristického prohledávání lesa všech možných stromů - Tree inference option, ale doporučuji nechat výchozí).

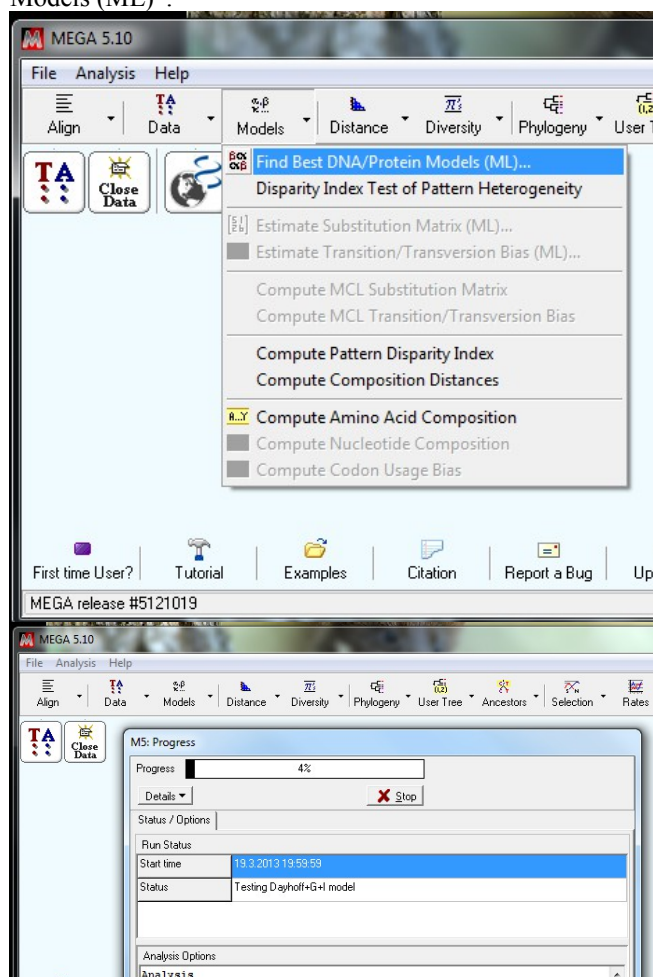


U metody ML opět ignorujte nastavování hledacích parametrů a specifikujte pouze substituční model.



Jak poznat který model vybrat?

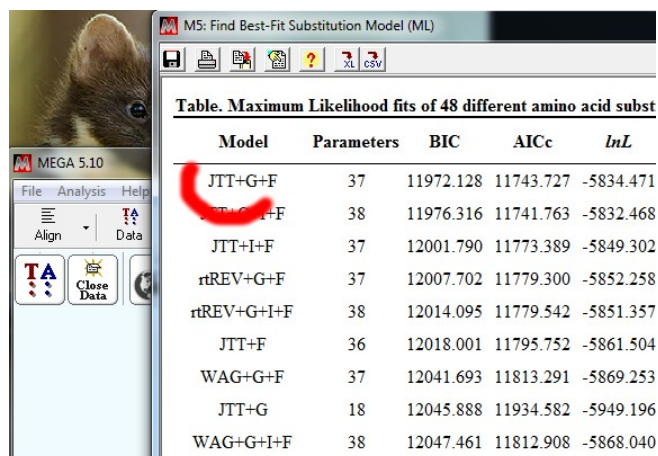
Při výběru modelu se ptáme takto: „vyplatí se použít složitější model za cenu nutnosti odhadnout více parametrů pro naše data, nebo nikoliv?“. Slouží k tomu v programu MEGA tlačítko „Models“ a „Find Best DNA/Protein Models (ML)“.



Vlastní výpočet zabere delší dobu a na konci počítání vypadne tabulka, kde jsou modely seřazeny od nejvhodnějšího dle informačního kritéria, které rozhoduje, zda-li složitější model poskytne signifikantně lepší výsledek.

Názvy modelů odpovídají těm, které jsou uvedeny v nabídce pro výpočet ML. Pokud je v modelu uveden parametr +G, znamená to, že máte nastavit parametr G. Doporučuje

se hodnota 5 (tj. operujeme, že napříč sekvencí může být pět odlišných mutačních rychlostí).



MEGA 5.10

File Analysis Help

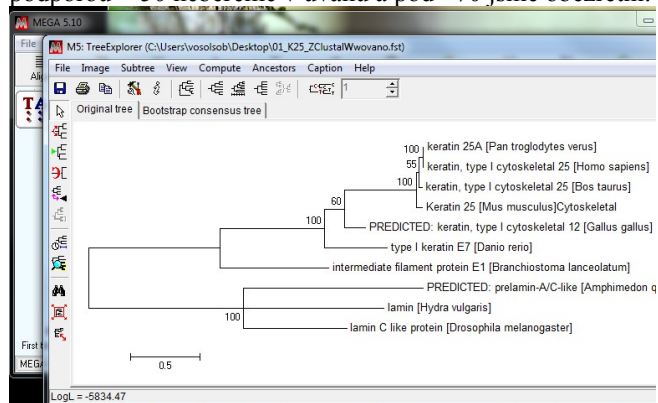
Align Data

Close Data

Table. Maximum Likelihood fits of 48 different amino acid subst

Model	Parameters	BIC	AICc	lnL
JTT+G+F	37	11972.128	11743.727	-5834.471
JTT+G+I+F	38	11976.316	11741.763	-5832.468
JTT+I+F	37	12001.790	11773.389	-5849.302
rtREV+G+F	37	12007.702	11779.300	-5852.258
rtREV+G+I+F	38	12014.095	11779.542	-5851.357
JTT+F	36	12018.001	11795.752	-5861.504
WAG+G+F	37	12041.693	11813.291	-5869.253
JTT+G	18	12045.888	11934.582	-5949.196
WAG+G+I+F	38	12047.461	11812.908	-5868.040

Pokud vytrváte, vypadne strom v novém okně. Čísla v uzlech jsou bootstrapy v %, 100 znamená, že ve všech případech (tj. náhodně přeuspořádaných alignmentech) vycházela tato větev, pokud obdržíme hodnotu 50, znamená to, že u poloviny stromů chyběla v rámci této větve alespoň jedna sekvence ve srovnání s hlavním stromem. Obecně větve s podporou < 50 nebereme v úvahu a pod ~70 jsme obezřetní.



Okno se stromem nabízí mnoho nástrojů k úpravě stromů, zejména můžeme otáčet větve v uzlech (tak aby byl strom přehlednější) a měnit polohu kořene, neboť výpočet samotný nemá zadáno, která sekvence je outgroup. Kořen bychom měli umístit mezi outgroupem a zbytkem. Pro získání ilustračního obrázku jsem si vypůjčil sekvence od Doubravky.

Úkoly

1. Vezměte alignment, kde jsou všechny různé keratiny ze všech zkoumaných organismů a spočítejte strom NJ, MP a ML. U ML vyberte vhodný substituční model dle výše uvedeného postupu. U NJ použijte stejný model jako u ML a pak pro porovnání některý jednodušší model.
2. Pokuste se zjistit, čím se modely odlišují.
3. Hotové stromy můžete pomocí tlačítka pro kopírování vložit do prezentace v Power-Pointu či uložit jako obrázek. Pokud vložíte strom do prezentace, chová se jako obrázek sloučených grafických prvků. Pokud vyberete v nabídce pravého tlačítka myši „rozpojit“, můžete ve stromu editovat text i barvy. V řešení by měly být celkem čtyři stromy. Přejmenujte sekvence, aby byl strom co nejprehlednější a případně je i nějak barevně rozlište.

4. Diskutujte všechny rozdíly, které pozorujete a posuďte, jak je strom směrodatný.

5. Zkuste nalézt, co je publikováno o fylogenezi keratinu pomocí databáze NCBI. Pokud by vám článek nešel stáhnout, pošlete mi jeho název a já se o to pokusím (vosolsob@natur.cuni.cz).

Pokud jste se dostali až sem, vězte, že disponujete vysoce nadprůměrnými schopnostmi v oboru bioinformatiky ☺.

