http://is.muni.cz/de/21109/Cvika_Zadani.doc

Cv. 1 – Popisná statistika

V přiloženém souboru s morfometrickými daty z okruhu *Festuca pallens* (http://is.muni.cz/de/21109/Cv1Festuca.xls) v programu Excel spočtěte pro každý znak každého taxonu minimum, 5% kvantil, dolní kvartil, průměr, medián, průměr, horní kvartil, 95% kvantil, maximum, rozptyl, S.D. a CV.

Zodpovězte, který ze sledovaných znaků je nejvariabilnější.

Zodpovězte, který ze sledovaných taxonů je nejvariabilnější ve velikostech květních částí (délka klásku, pluchy a osiny)

Data exportujte do programu R.

Udělejte zde stejnou tabulku, jako jste dělali v MS Excel.

Vytvořte box and whisker ploty pro 3 vybrané kontinuální proměnné zobrazujících všechny taxony. Můžete zkusit jak jednotlivě, tak 3 grafy v jednom grafickém okně.

Vytvořte celkový histogram délky stébla napříč všemi taxony.

Vytvořte obrázek se dvěma společnými histogramy (oba v jednom grafickém okně) pro délku osiny taxonů „pal“ a „sc“.

Cv. 2 – testování rozdílů ve znacích

Otestujte normalitu rozložení v rámci taxonů u znaku délka pluchy a u znaku délka osiny pro taxony (mm+vd) a (pa+hu+sc+ss+st).

Otestujte, zda se sdružené taxony (mm+vd = *F. psammophila* s.l.) a (pa+hu+sc+ss+st = *F. pallens* s.l.) liší v délce osin. Použijte různé testy a sledujte, jak se liší obdržené pravděpodobnosti.

Otestujte, zda se taxony liší v délce pluchy a zda je tedy tento znak může být významný pro jejich rozlišování (ANOVA).

Pomocí post-hoc testu otestujte, které taxony se v délce pluchy liší a které jsou stejné. Nakreslete boxplot s označením (písmeny) skupin se statisticky stejně dlouhými pluchami.

Pomocí permutačního testu a nasimulování alespoň 1000 náhodných výběrů otestujte, zda se sdružené taxony (mm+vd = *F. psammophila* s.l.) a (pa+hu+sc+ss+st = *F. pallens* s.l.) liší v maximálním počtu cévních svazků na průřezu listem.

Nasimulujte si v R (MS Excelu) vektor (sloupec) 100 náhodných dat od 0 do 1 a korelujte ho s další stovkou takto náhodně vygenerovaných dat. Sledujte, kolik dostanete náhodně signifikantních výsledků (které by bylo potřeba korigovat Bonferoniho nebo jinou korekcí).

Cv. 3 - Korelace a regrese

Definujte hlavní hypotézy vašich experimentů/vašeho bádání a diskutujme nad nejlepším a nejefektivnějším designem sběrů nebo pokusů.

Vygenerujte si náhodná data s normálním rozložením s alespoň 1000 hodnotami a pozorujte (nakreslete graf), jak moc mimo byste byli od průměru těchto dat, kdybyste jejich průměr nepočítali ze všech 1000 dat, ale jen z 1, 2, 3….. hodnot. Pokud máte vlastní data, použijte vlastní (třeba průduchy).

Načtete soubor průduchy.txt . Vykreslete 2D graf závislosti mezi 2C velikostí genomu (X2C) a délkou pruduchů (l.stom). Otestujte korelaci obou proměnných pomocí různých korelačních koeficientů (Pearson, Spearman, Kendall). V případě parametrického Pearsonova korelačního koeficientu zkuste analýzu s netransformovanými i logaritmovanými daty.

Spočítejte LS regresi pro logaritmovaná data o velikosti genomu a délce průduchů. Obě proměnné zkuste dát jednou jako závislou (predikovanou) a jednou jako nezávislou proměnnou (prediktor). Regresní přímky z obou regresí zobrazte v 2D grafu s hodnotami proměnných.

Do grafu z předchozího odstavce přidejte křivku z vážené LS regrese obou proměnných vážených podle variability naměřených výsledků velikosti genomu (CV.2C).

Na stejných proměnných udělejte všechny 4 typy model II regresí pomocí balíčku „lmodel2“

Na stejných proměnných spočítejte kvantilovou regresi pomocí balíčku „quantreg“. Zobrazte 2D scatterplot hodnot s vyznačenými regresními křivkami pro kvantily (5%, 10%, 25%, 50%, 75%, 90%, 95%).

Zobrazte vývoj změn interceptu a odhadovaného sklonu regresních křivek v závislost na hodnotě zvoleného kvantilu regrese. Diskutujte, možnou interpretaci této závislosti na základě toho, co víte o velikosti průduchů a genomu.

Cv. 4 – Distanční matice

Načtěte do R soubor Matice_Festuca.txt. Soubor obsahuje výběr morfologických dat ze souboru FestucaCV1 bez NA hodnot.

Nakreslete histogramy a boxploty pro všechny proměnné. Zhodnoťte vizuálně jejich normalitu a případně je vhodně transformujte.

Z balíčku funkcí MorphoTools (soubor MorphoTools1-1.R; Koutecký et al. 2015) si do R načtěte následující funkce (text zkopírujte a vložte do R okna): read.morphodata, export.res, descr.tax, descr.all, charost, cormat (nebo klidně všechny). Nápověda k funkcí je v souboru „MorphoTools_manual_2016.pdf“ a „MorphoTools working protocol 1-1.R.txt“

Z MorphoTools zkuste na datech funkce descr.tax a descr.all.

Spočtěte matici párových korelací (Pearson i Kendall) pro všechny znaky v souboru (obsahujícím už transformované proměnné). Potom to samé zkuste s MorphoTools funkcí cormat. Všechny matice uložte do Excelu a obarvěte v Excelu pomocí podmíněného formátování hodnoty korelací. Zamyslete se, proč spolu některé proměnné korelují a které se vlastně vztahují víceméně k jednomu znaku.

Data standardizujte na rozpětí a směrodatnou odchylku. S takto standardizovanými daty spočtěte maticí Euklidovských vzdáleností mezi vzorky (ploidii vynechte; pokud by se měl v této fázi proměnné vážit, stačilo by je standardizované před počítáním vzdáleností vynásobit jejich váhou). Matice si uložte pro další výpočty ve cvičení 5.

Načtěte prezenčně absenční data v csv formátu z analýzy AFLP v rodě Cirsium (Cirsium_AFLP.csv) a s pomocí funkce vegdist v balíčku vegan spočtěte vzdálenosti mezi druhy pomocí Jaccardova indexu. Matice si uložte pro další výpočty ve cvičení 5.

Pomocí funkce read.dna balíčku ape načtěte alignovaná sekvenční data genu rbcL (Ribulose bisphosphate carboxylase large chain) v různých kvetoucích rostlinách v souboru Alignment_Angiosperms_rbcL.FAS (je ve formátu fasta). Pomocí funkce dist.dna spočtěte vzdálenosti druhů pomocí Jukes & Cantor 1969 a Kimura 1980 modelů. Matice si uložte pro další výpočty v dalších cvičeních.

Cv. 5 - Ordinace

Udělejte PCoA s uloženou maticí euklidovských vzdáleností vzorků kostřav (standardizovanými na rozsah). Zobrazte ordinační diagram.

Udělejte PCoA s uloženou maticí „Jaccardových“ vzdáleností vzorků pcháčů. Zobrazte výsledná skóre prvních tří os a barevně odlište jednotlivé taxony.

Udělejte PCA s transformovanými daty (logaritmovanými znaky) – viz cv. 4. Tyto data uložte pro další použití. Zobrazte PCA biplot (vzorky jako body a proměnné jako šipky; různé taxony různými barvami).

U provedené PCA rozhodněte o výpovědní hodnotě os a otestujte jejich „významnost“ pomocí broken stick modelu.

Tu samou PCA zkuste zobrazit pomocí dalších programů a funkcí v MorphoTools

Udělejte 2 PCA: jednu s daty standardizovanými na směrodatnou odchylku a druhou s daty standardizovanými na rozsah – porovnejte, jak se mění významnost jednotlivých proměnných.

Udělejte NMDS s uloženou maticí euklidovských vzdáleností vzorků kostřav (standardizovanými na rozsah). Zkontrolujte kolik dimenzí by se asi hodilo pomocí NMDS zobrazit (koukněte na stressplot pro NMDS s různým k). Výslednou ordinaci vhodně zobrazte.

Vypočtěte matici vzdáleností mezi vzorky na základě jejich skóre na významných osách provedené PCA. Porovnejte zobrazení prvních dvou os PCA se zobrazením této matice vzdáleností pro dvě osy (k=2) pomocí NMDS.

Cv. 6 – Diskriminační analýza (DA) s MorphoTools funkcemi

1. Spočtěte kanonickou DA s transformovanými daty kostřav. Otestujte významnost jednotlivých proměnných a zobrazte biplot
2. Spočtěte klasifikační DA a ukažte + diskutujte, jak moc se které taxony dají/nedají odlišit a které se nejvíce vzájemně zaměňují
3. Spočtete úspěšnost klasifikaci pomocí K-nearest neighbour metody
4. Pomocí DA i K-nearest neighbour klasifikujte tři neznámé vzorky v souboru Festuca_unclear.txt

Opakujte body 1-4 pro taxony sdružené na psamofilní (mm+vd) a saxikolní (pa+hu+sc+st)

Pro tuto DA se dvěma taxony najděte stepwice metodou nejlepší minimální kombinaci znaků k jejich rozlišování a udělejte na to vlastní diskriminační formuli

Opakujte body 1-4 pro taxony sdružené na psamofilní diploidy (mm+vd), saxikolní diploidy (pa) a saxikolní tetraploidy (hu+sc+st). Které z rozdělení taxonů (6 taxonů, 3 taxony, 2 taxony) vám dává nejlepší smysl?

Cv. 7 – (fenetické) klastrování a maticové testy

- Morfologická data o kostřavách klastrujte pomocí metod Single linkage, Complete linkage, UPGMA, WARD metody (incremental sum of squares). K výpočtu použijte skóre signifikantních os z PCA (vybrané pomocí Broken stick modelu) a nezávisle pak také matici euklidovských vzdáleností vzorků s daty standardizovanými na směrodatnou odchylku vytvořenou ve cv. 4. Pro stejná data také pro zajímavost spočtěte minimum evolution stromy, neighbor joining tree (NJ) a least square tree (LS). Vzorky v dendrogramech odlište různou barvou podle příslušností do jednotlivých taxonů a výsledné dendrogramy graficky porovnejte. Porovnejte jednak vhodnost minim evolution stromů k analýze morfologických dat, tak rozdíl mezi klastrováním vlastní matice vzdáleností a matice vzdáleností z PCA skóre, která jsou očesaná o korelované proměnné a šum v datech.
- Pro dendrogramy z PCA dat pomocí „tanglegramů“ porovnejte graficky rozdíly v pozicích vzorků v dendrogramech zhotovených pomocí UPGMA a (WARD, NJ a LS) a mezi NJ a LS pro oba zdroje dat.

- Porovnejte korelaci vzdáleností vzorků vzájemně mezi dendrogramy spočtenými různými metodami. Použijte k tomu Pearsonův korelační koeficient, Spearmanův korelační koeficient nebo prosté porovnání % shodných klastrů. Výsledky vhodně graficky zobrazte.

- Spočtěte kofenetické korelace (originální vzdálenosti vs vzdálenosti v dendrogramu) pro dendrogramy zhotovené jednotlivými metodami. Graficky zobrazte.

- Pomocí funkcí v balíčku ape bootstrapujte UPGMA strom kostřav z prvního úkolu. Zobrazte dendrogram s bootstrap hodnotami.

- V PCA diagramu s využitím Minimum spanning stromu zobrazte vzájemně nejbližší (nejpodobnější) vzorky.

- Pro surová data i PCA skóre spočtěte kmeans klastry pro vhodný počet skupin (odhadněte z podstaty problému nebo spočítejte pomocí nějakého balíčku). Klastry zobrazte v PCA vzorků.

- Opakujte první úkol pro matici Jaccardových vzdáleností s AFLP daty pro Cirsium (ne pro PCA i když i to by mohlo být zajímavý). Jako outgroup v případě NJ a LS stromů vyberte „Sylibum48“.

- Opakujte úkol se sekvenčními (rbcl) daty pro angiosperms ze cv. 4. Kdybyste na to nepřišli sami, tak jako outgroup v případě NJ a LS stromů vyberte „Ginkgo_biloba“ :O).

- NJ a LS strom z předchozího úkolu bootstrapujte – k tomu vám nedám skripty ale pár tipů: stromy se nesnažte předělat na dendrogramy nebo formát h.clust, ale nechte je ve stejném výstupu, jak je vyhodí nj a fasme.bal; stromy se při bootstrapování pro jednoduchost nesnažte kořenit ani ultrametrizovat.

---

cv. 8a – Alignment

V R: Načtěte soubor Angiosperms_rbcL.fasta obsahující cDNA sekvence chloroplastového genu *rbc*L (Rubisco). Alignujte nukleotidovou sekvenci pomocí algoritmu ClustalW a Muscle v balíku msa. Bonus pro toho, kdo u sebe rozchodí mafft a prank algoritmy z balíku ips. V případě zájmu o alignování koukněte na nepovinné cvičení v programu MEGA.

V MEGA: nepovinné

Soubor Angiosperms_rbcL.fasta obsahující cDNA sekvence chloroplastového genu *rbc*L (Rubisco) otevřete v programu MEGA (zadáním cesta v programu, kliknutím na ikonu souboru, nebo přetažením myší). Alignujte nukleotidovou sekvenci pomocí algoritmu Muscle; jako klastrovaní algoritmus použijte NJ.

Alignment zobrazte v aminokyseinovem zápisu a v alignmentu najděte počátek genu (start kodon s methioninem na začátku) a jeho konec (stop kodon = *). Nukleotidový alignment ořízněte a uložte jako fasta soubor.

Soubor Angiosperms_rbcL.fasta vlozte k alignovani pomoci algoritmu PRANK na adrese: http://www.ebi.ac.uk/goldman-srv/webprank/ (k alignovani použijte default nastaveni). Výsledný alignment zobrazte, porovnejte s výsledkem z MEGA a uložte.

Soubor Angiosperms_rbcL.fasta vložte k alignování pomoci algoritmu Muscle a 100 bootstrap replikaci na serveru GUIDANCE na adrese: http://guidance.tau.ac.il/. V souboru s Muscle alignmentem případně podle PRANK hodnocení vyřaďte sporné pozice sekvence.

Cv 8. b – fylogenetické stromy dokončení

Do R načtěte už hotový a ořezaný alignment „Alignment_Angiosperms_rbcL.FAS“. V balíku „phangorn“ vytvořte fylogenetický strom (zakořeněněný na Ginkgo biloba) na základě maximum parsimonie. Výsledný strom bootstrapujte a zobrazte.

Opakujte to samé s metodou maximum likelihood (strom nemusí být zakořeněný). Pro výpočet najděte nejvhodnější substituční model párů bazí. Výsledný strom bootstrapujte a zobrazte.

cv. 8c – Maticové testy

Načtěte soubor „"Festuca_matice_kultivace.txt", který obsahuje data měřená na těch samých rostlinách kostřav (trsech) v přírodě a poté za rok v kultivaci. Ordinujte pomocí PCA odděleně data měřená na volně rostoucích a kultivovaných trsech. Pomocí Mantel testu otestujte shodu obou ordinací (vzdáleností bodů v ordinačním prostoru).

To samé udělejte pomocí procrustes testu. Zde navíc koukněte na to, které vzorky se v přírodě a kultivaci nejvíce lišily.

Pomocí t testu otestujte, které znaky se v kultivaci nejvíce změnily. Graficky zobrazte. Může t být závislé na taxonu?