Úloha 2 - určovanie príbuznosti pomocou kompresie

V úlohe som používal dve rôzne kompresie na porovnanie príbuznosti organizmov (DNA sekvencií). Špeciálne program GenCompress a 7zip. Úlohou bolo vytvoriť vzdialenostnú maticu, z ktorej sa dá vygenerovať fylogenetický strom a z neho určiť príbuznosti organizmov.

Na vytvorenie vzdialenostnej matice $\mathcal{D}$ je možné (ako bolo prebrané na prednáške) použiť Kolmogorov complexity. Na výpočet vzdialenosti sekvencie w od z použijeme (w a z sú sekvencie, ktoré rozparsujeme z all_seq.fasta).

$$D(w,z) = 1 - R(w,z)$$

Obrázok 1 : Vzdialenosť dvoch sekvencií

$\mathcal{R}(w,z)$ sa zo vzorca dá vypočítať pomocou.

$$R(w,z) = \frac{K(w) - K(w \mid z)}{K(wz)} = \frac{K(z) - K(z \mid w)}{K(wz)}$$

Obrázok 2 : Výpočet podľa Kolmogorov Complexity

Keďže $\mathcal{K}$ nie je možné vypočítať, tak používame aproximáciu.

$$K(w) := Compress(w) := |\, GenCompress(w) \,|$$
$$K(w \mid z) := Compress(w \mid z) := Compress(zw) - Compress(z) =$$
$$|\, GenCompress(zw) \,| - |\, GenCompress(z) \,|$$

Obrázok 3 : Aproximácia Kolmogorov Complexity

V aproximácii je zložku Compress možno vypočítať použitím kompresného algoritmu na sekvencie a navrátiť dĺžky týchto skomprimovaných sekvencií.

Prepríprava dát

Parsovanie

Sekvencie sa nachádzali v súbore all_seq.fasta, ktorý bolo na výpočet nutné rozparsovať podľa mena sekvencie (skriptom vznikli súbory seqA, seqB,...). Taktiež bolo nutné získať sekvencie súbory, v ktorých by sa nachádzali 2 rôzne sekvencie za sebou spojené (skriptom vznikli subory seqAseqB,seqAseqC,...).

Kompresia pomocou GenCompress

Prvou úlohou bolo komprimovať súbory pomocou programu GenCompress.

- Na výpočet K(w) je nutné zavolať program GenCompress na každú individuálnu sekvenciu (seqA, seqB,...), čím dostaneme skomprimované sekvencie, ktoré si je dobré odložiť do nejakého adresára (gencompress/individual).
- Pomocou programu GenCompress je možné výpočet K(w|z) vykonať priamo tým že pri komprimovaní pridáme parameter **–c**, za ktorý pridáme referenčnú sekvenciu podľa ktorej komprimujeme. Command : *GenCompress(w|z) = GenCompress.exe w –c z.*Taktiež si ich uložime do nejakého adresára (gencompress/combinated)
- Na výpočet K(wz) zavoláme program GenCompress na každú dvojicu spojených súborov sekvencií, ktoré sme si predpripravili. Taktiež si ich uložíme do adresára (gencompress/concanated).

Kompresia pomocou 7zip

Na skomprimovanie je možné využiť možnosti TotalCommander-u a komprimovať súbory dávkovo. My sme pridali možnosť vykonať komprimáciu priamo zo skriptu.

- Na výpočet K(w) je nutné zavolať 7zip na každú individuálnu sekvenciu (seqA, seqB,...), čím dostaneme skomprimované sekvencie, ktoré si je dobré odložiť do nejakého adresára (zip/individual).
- Pomocou programu 7zip nevytvárame referenčnú kompresiu.
- Na výpočet K(wz) zavoláme program 7zip na každú dvojicu spojených súborov sekvencií, ktoré sme si predpripravili. Taktiež si ich uložíme do adresára (zip/concanated).

Vytvorenie vzdialenostných matíc

Na vytvorenie vzdialenostných matíc slúži hlavný skript, ktorý postupne pridáva dĺžky skomprimovaných súborov do slovníku/dictionary, kde kľúč sú mená sekvencií. Takýmto spôsobom vytvoríme 3 slovníky

- Individual
- Combinated
- Concanated

V prípade 7zip sme nevytvárali súbory s referenčnou komprimáciou. V takomto prípade je nutné do slovníku Combinated dopočítať vzdialenosti takejto komprimácie, tak ako je uvedené na obrázku 3. Všetky údaje máme prichystané, takže môžeme dopočítať vzdialenostnú maticu $\mathcal{D}$ podľa vzorca uvedeného na obrázkoch vyššie.

Zobrazenie matíc

Po vygenerovaní vypíšeme matice v takom formáte, aby sa ľahko dala určiť príbuznosť (obrázok 4).

```
                                GenCompress
7
seqA    0.00000 0.96539 1.00029 1.00015 0.99926 1.00015 1.00015
seqB    0.96519 0.00000 1.00072 1.00000 0.99971 1.00000 1.00000
seqC    1.00000 1.00000 0.00000 0.95350 0.98455 0.70933 0.97913
seqD    1.00000 0.99971 0.95409 0.00000 0.98351 0.94368 0.97677
seqE    0.99911 0.99927 0.98538 0.98378 0.00000 0.98006 0.96682
seqF    1.00000 1.00000 0.71042 0.94395 0.97992 0.00000 0.96788
seqG    1.00000 1.00000 0.97998 0.97677 0.96641 0.96760 0.00000
------------------------------------------------------------
                                    7z
7
seqA    0.00000 0.85212 0.96352 0.96369 0.95771 0.96580 0.96331
seqB    0.87069 0.00000 0.96637 0.96539 0.95815 0.96690 0.96423
seqC    0.96893 0.96933 0.00000 0.74406 0.87023 0.46047 0.84230
seqD    0.96681 0.96598 0.74778 0.00000 0.87774 0.75964 0.84900
seqE    0.96131 0.96064 0.87101 0.88468 0.00000 0.85926 0.81679
seqF    0.96747 0.96796 0.45894 0.75675 0.84792 0.00000 0.83753
seqG    0.96463 0.96589 0.86170 0.85589 0.80943 0.84584 0.00000
------------------------------------------------------------
```

Obrázok 4 : Vzdialenostné matice jednotlivých kompresií

Takýto formát zaručuje funkcia *printDistanceMatrix* nachádzajúca sa v skripte. Formát je zvolený nielen kvôli čitateľnosti, ale aj kvôli tomu, že na vygenerovanie fylogenetických stromov používam program nachádzajúci sa na stránkach http://iubio.bio.indiana.edu/treeapp/treeprint-form.html. Tento program dokáže pracovať iba s reprezentáciou stromov nazývajúcich sa Newickove stromy. Na vygenerovanie Newickovho stromu zase využijem program na stránke http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::bionj. Fylogenetické stromy vygenerované týmto programom je vidieť na obrázku 5 a obrázku 6. Obidva stromy berú do úvahy dĺžky hrán získané zo vzdialenostnej matice.

Výsledky

Z vytvorených fylogenetických stromov (phenogramov a cladogramov, pri ktorých berieme do úvahy vzdialenosti), ako aj zo vzdialenostných matíc je možné vidieť príbuzenské vzťahy.

1. seqC je najpríbuznejšia k sekvencii seqF.
2. Obidve sú spolu príbuzné k sekvencii seqD.
3. seqB je najpríbuznejšia k sekvencii seqA
4. skupina seqC, seqF, seqD je príbuzná k skupine sekvencií seqB, seqA
5. seqG je príbuzná so sekvenciou seqE

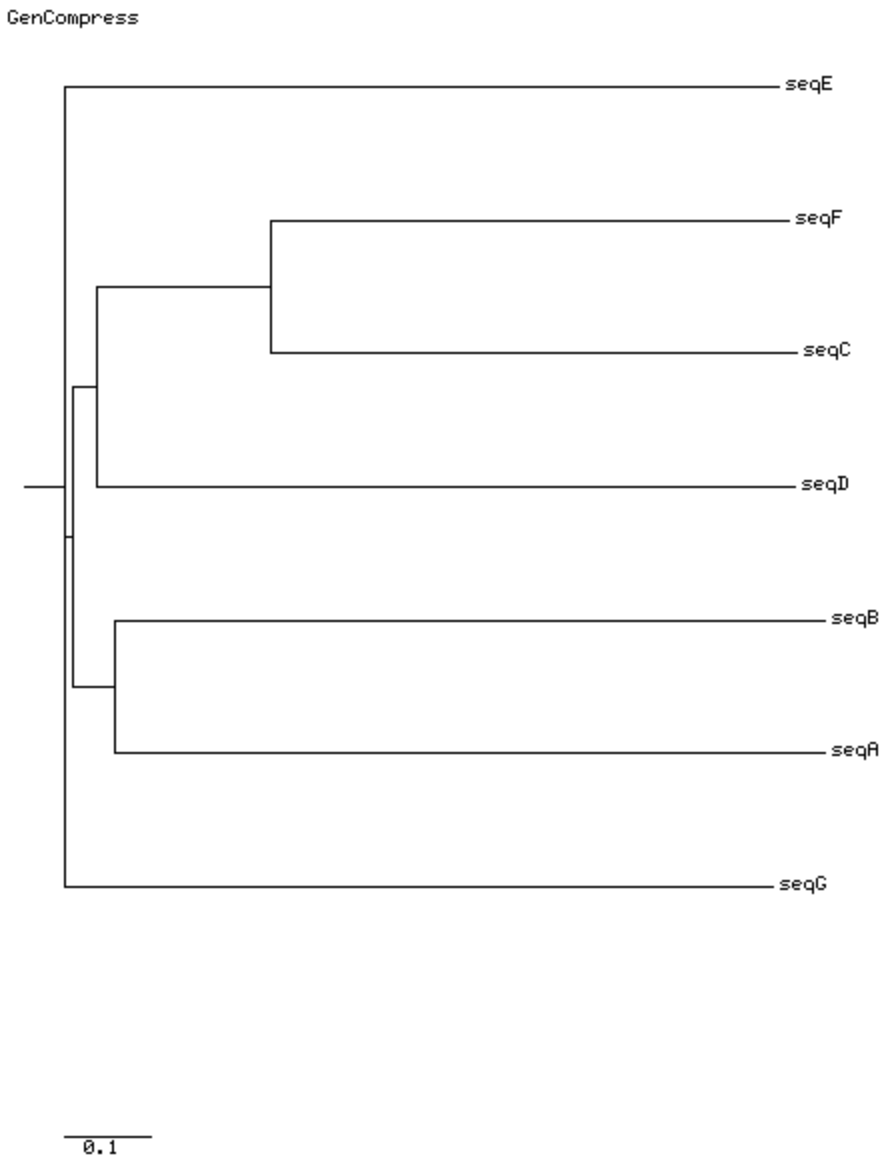
Taktiež sa dá na príbuzenské vzťahy pozerať z cladogramov na obrázkoch 8 a 9.

1. Na začiatku bola nejaká DNA sekvencia.

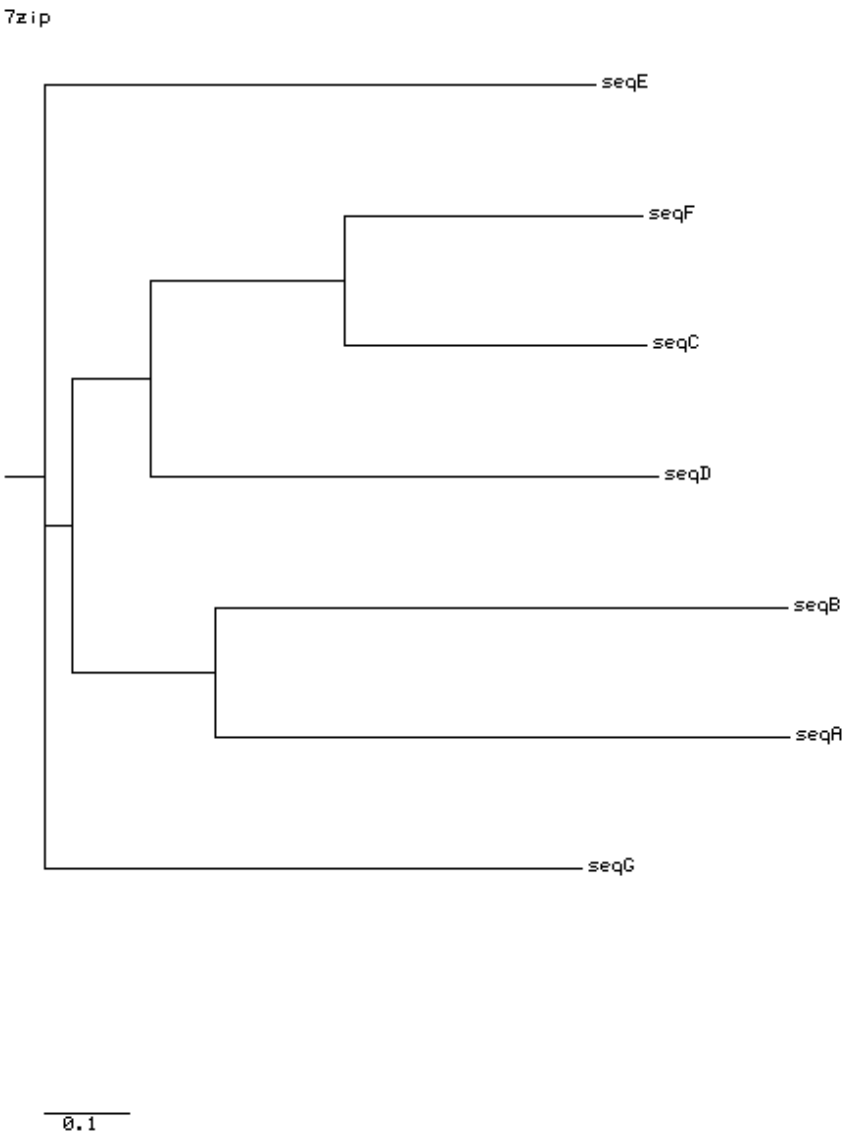
2. Vymedzila sa rodová línia a vznikol predok sekvencií seqE a seqG.
3. Ďalšia udalosť nastala pri rozvetvení línie na dve skupiny (nadchádzajúce sekvencie zdieľajú spoločného predka, a tie ďalej zdieľajú spoločného predka so sekvenciami seqE a seqG).
4. Ďalej vznikol predok sekvencií seqF,seqC,seqD, potom predok sekvencií seqA a seqB.
5. Z predka seqA a seqB vznikli sekvencie seqA a seqB
6. Z predka sekvencií seqF,seqC,seqD sa separovala rodová línia na sekvenciu seqD a líniu vedúcu k seqF a seqC.
7. Línia sa znova rozdelila a vytvorila seqF a seqC => zdieľajú spoločného predka.

Nami vytvorené fylogenetické stromy dávajú náhľad nato, ako sú postavené vzťahy medzi jednotlivými sekvenciami. Doposiaľ sme vytvorili iba také informácie o sekvenciách a ich vzťahoch, pri ktorých neberieme do úvahy vzdialenosti, dĺžky hrán stromu. Táto vlastnosť je pri GenCompress nie príliš viditeľná, keďže hrany stromu sa javia skoro rovnaké. Pri kompresii 7zip sú rozdiely očividnejšie a tým sa dajú utvrdiť ďalšie spozorované informácie, ako sú poradia kedy sa stali rozdelenia rodinných línii. Napríklad rozdelenie na seqA a seqB sa stalo skorej, ako rozdelenie na seqF a seqC

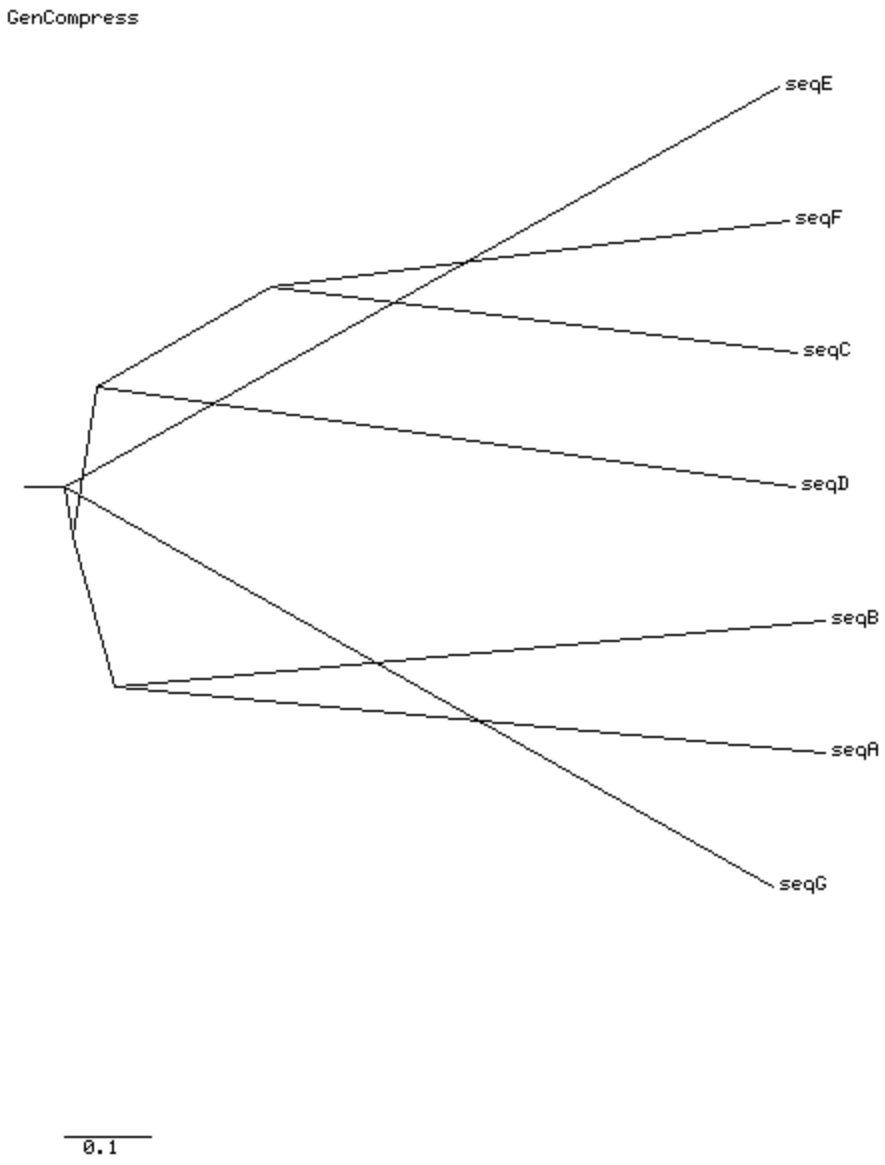
Stromy ani vzdialenostné matice nie sú rovnaké kvôli použitiu rôznych kompresných algoritmov. Štruktúra fylogenetických stromov síce zostáva rovnaká, no vzdialenosti medzi sekvenciami sú rôzne. Pre zistenie príbuznosti mi príde lepšie použiť Na konštrukciu stromov som použil programy nachádzajúce sa na internete, pretože som dostal požadované stromy bez inštalácii nadbytočných python-ových knižníc.



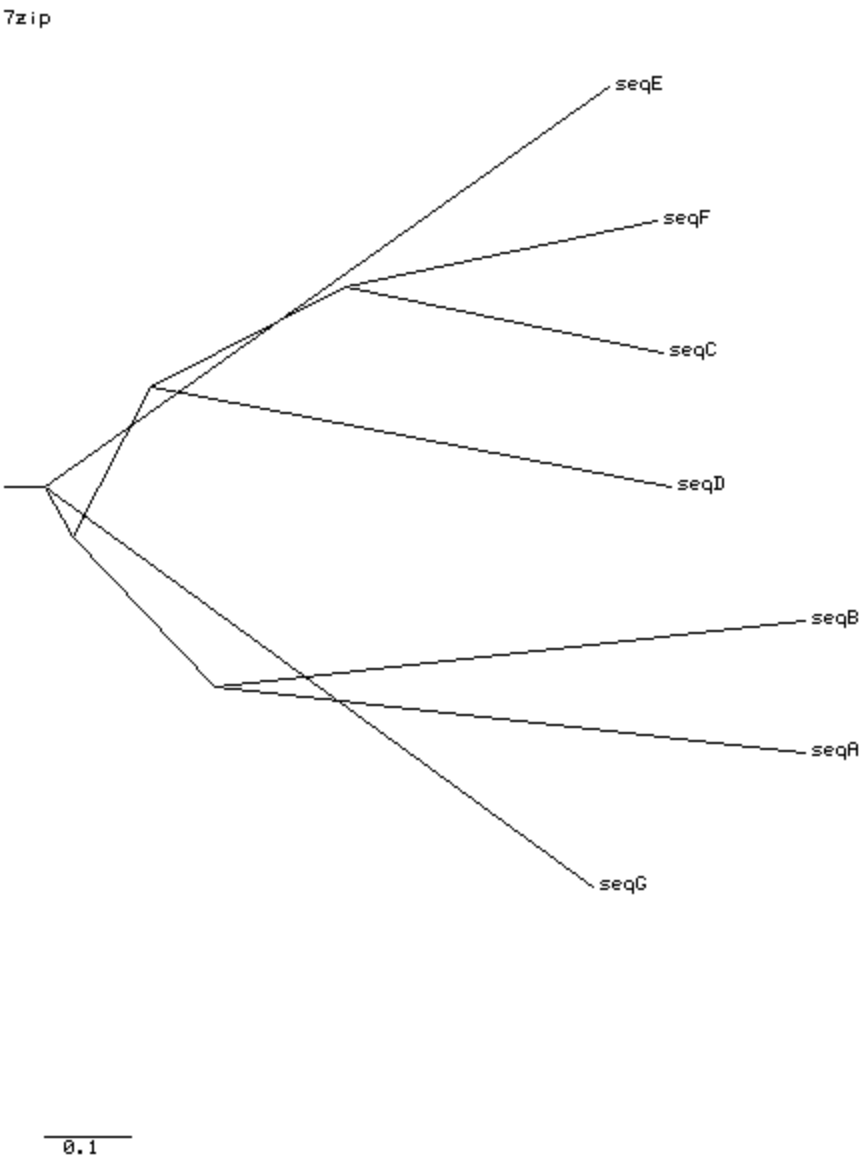
Obrázok 5 : Fylogenetický strom pri kompresii GenCompress



Obrázok 6 : Fylogenetický strom pri kompresii 7zip



Obrázok 7 : Cladogram GenCompress



Obrázok 8 : Cladogram 7zip-u

Bibliography

Baum, D. (2008). *Scitable*. Dostupné na Internete: Nature: http://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956

D.G.Gilbert. (January 1999). *iubio*. Dostupné na Internete: Phylodendron: http://iubio.bio.indiana.edu/treeapp/treeprint-form.html

Mobyle. Dostupné na Internete: Mobyl @ Pasteur: http://mobyle.pasteur.fr/cgi-bin/portal.py?#jobs::bionj.S16579746548891