Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Vizualizácia evolučných histórií

Bakalárska práca

2016
Dávid Simeunovič

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

Vizualizácia evolučných histórií

Bakalárska práca

Študijný program:	Informatika
Študijný odbor:	2508 Informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	doc. Mgr. Bronislava Brejová, PhD.

Bratislava, 2016
Dávid Simeunovič

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:
Študijný program: informatika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.2.1. informatika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský

Názov:

Cieľ:

Literatúra:

Kľúčové slová:

Vedúci:
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: doc. RNDr. Daniel Olejár, PhD.

Dátum zadania:

Dátum schválenia: doc. RNDr. Daniel Olejár, PhD.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie:

Abstrakt

Počas evolúcie dochádza v DNA k lokálnym mutáciám, ktoré menia jeden alebo niekoľko susedných nukleotidov, ale aj k väčším zmenám, ktoré menia poradie alebo počet výskytov dlhších oblastí. Cieľom práce je implementovať systém na vizualizáciu evolučnej histórie jednej alebo viacerých DNA sekvencií s dôrazom na tieto väčšie zmeny. Samotná história je daná na vstupe a cieľom je zobraziť ju tak, aby sa dali prehľadne sledovať jednotlivé mutácie a tiež vzťahy rôznych častí sekvencie.

Kľúčové slová: vizualizácia,evolučná história,poradie génov

Abstract

Abstract in the English language (translation of the abstract in the Slovak language).

Keywords:

Obsah

Zoznam obrázkov

Úvod

Úvod je prvou komplexnou informáciou o práci, jej cieli, obsahu a štruktúre. Úvod sa vzťahuje na spracovanú tému konkrétne, obsahuje stručný a výstižný opis problematiky, charakterizuje stav poznania alebo praxe v oblasti, ktorá je predmetom školského diela a oboznamuje s významom, cieľmi a zámermi školského diela. Autor v úvode zdôrazňuje, prečo je práca dôležitá a prečo sa rozhodol spracovať danú tému. Úvod ako názov kapitoly sa nečísluje a jeho rozsah je spravidla 1 až 2 strany.

Kapitola 1

Úvod do problematiky

V tejto kapitole si vysvetlíme základne pojmy potrebné pre túto bakalársku prácu.

1.1 Biologické pozadie

1.1.1 DNA,Gén,Genóm

DNA Deoxyribonukleová kyselina je nositeľom genetickej informácie bunky.
Má štruktúru dvojzávitnice, skladajúcej sa z dvoch komplementárnych vlákien. Vlákno je tvorené nukleotidmy, ktoré obsahujú jednu zo štyroch báz Adenín,Guanín,Tymín a Cytozín. DNA zvykneme zapisovať ako postupnosť týchto báz, kde každú bázu kódujeme jej počiatočným písmenom - A,G,T,C.

Gén Gén je súvislý úsek DNA ktorý kóduje tvorbu proteínu. Gén je základnou jednotkou dedičnosti.

Genóm Genóm je súbor DNA molekúl organizmu, ktoré sa väčšinou nachádzajú v chromozómoch. Napríklad v ľudskom tele sa jedná o 46 molekúl DNA, jedna v každom chromozóme[6].

1.1.2 Evolučná história

Evolučná história je postupnosť udalostí, ktoré sa odohrali na nejakej DNA sekvencii. Pre potreby tejto práce sú podstatné udalosti odohrávajúce sa na dlhých úsekoch DNA - génoch.
Možné udalosti sú:

- *Duplikácia* - skopírovanie génu na iné miesto v DNA.

- *Inzercia* - vloženie nového génu.
- *Delécia* - odstránenie génu.
- *Inverzia* - zmena poradia a orientácie génu alebo génov.
- *Translokácia* - zmena poradia génu alebo génov
- *Speciácia* - špeciálna udalosť, ktorá označuje vznik nového druhu. Vzniká nová vetva v evolučnej histórii.

Krok evolučnej histórie

Krok evolučnej histórie, ďalej len *krok e.h* je pre nás známa sekvencia génov. Medzi jednotlivími krokmi došlo k jednej alebo viacerím udalostiam.

1.1.3 Fylogenetický strom

Fylogenetický strom je grafické znázornenie, ktoré zobrazuje evolučné vzťahy medzi sadou objektov. Pokiaľ si za objekty zvolíme druhy, jedná sa o takzvaný *Druhový strom*. Jednotlivé druhy sú pospájané hranami, ktoré reprezentujú evolučný vzťah.
Druhy, ktoré sa nachádzajú na *listoch* stromu sú buď existujúce druhy, z ktorých sa nevyvinuli nové druhy, alebo vyhynuté druhy bez potomkov.
Vnútorné vrcholy predstavujú predchodcov, o ktorých sa predpokladá že sa vyskytli počas evolúcie.
Pokiaľ je v strome známy posledný spoločný predok, nazveme ho *koreň*, a takýto strom označíme ako *zakorenený*.
V *zakorenenom* strome je zrejmá orientácia vnútorných hrán, ktorá určuje ktorý druh sa vyvinul z ktorého.

1.2 Ostatné pojmy

Kapitola 2

Implementácia programu

V tejto kapitole sa budeme venovať niektorým významnejším črtám implementácie nášho programu

2.1 Formát vstupu

Evolučná história je tvorená postupnosťou riadkov, podobnej akú vidíme v tabulke 2.1 . Prvý riadok opisuje počiatočný stav sekvencie. Každý ďalší riadok, opisuje niektorú z udalostí, popísanú v sekcii 1.1.2.
Riadok obsahuje niekoľko reťazcov a čísel, oddelených medzerou alebo viacerými medzerami, pre lepšiu prehľadnosť.

Význam stĺpcov:

Prvý stĺpec je názov druhu, ktorého sa týka daný riadok.

Druhý stĺpec je id riadku.

Tretí stĺpec je id predchodcu, prvý riadok má špeciálneho predchodcu s hodnotou root.

Štvrtý stĺpec je čas, v ktorom sa daná udalosť odohrala. Koreň je v čase 0, a čas je rastúci.

Piaty stĺpec je skratka niektorej z udalostí, popísaných v sekcii 1.1.2 alebo špeciálna udalosť. Root je udalosť slúžiaca na identifikáciu koreňa a leaf slúži na určenie času v ktorom sa daná vetva končí.

Nasledujúce stĺpce obsahujú postupnosť génov. Každý gén je celé číslo, pričom znamienko určuje jeho orientáciu. To znamená že gén 2 je rovnaký ako gén -2, iba opačne orientovaný.

Znak # slúži ako ukončenie zoznamu génov.

Zvyšné stĺpce , ktoré pre každý gén určujú poradie predka génu v predchodcovi jeho riadku. Ak tento gén nemá predchodcu, obsahuje riadok hodnotu -1.

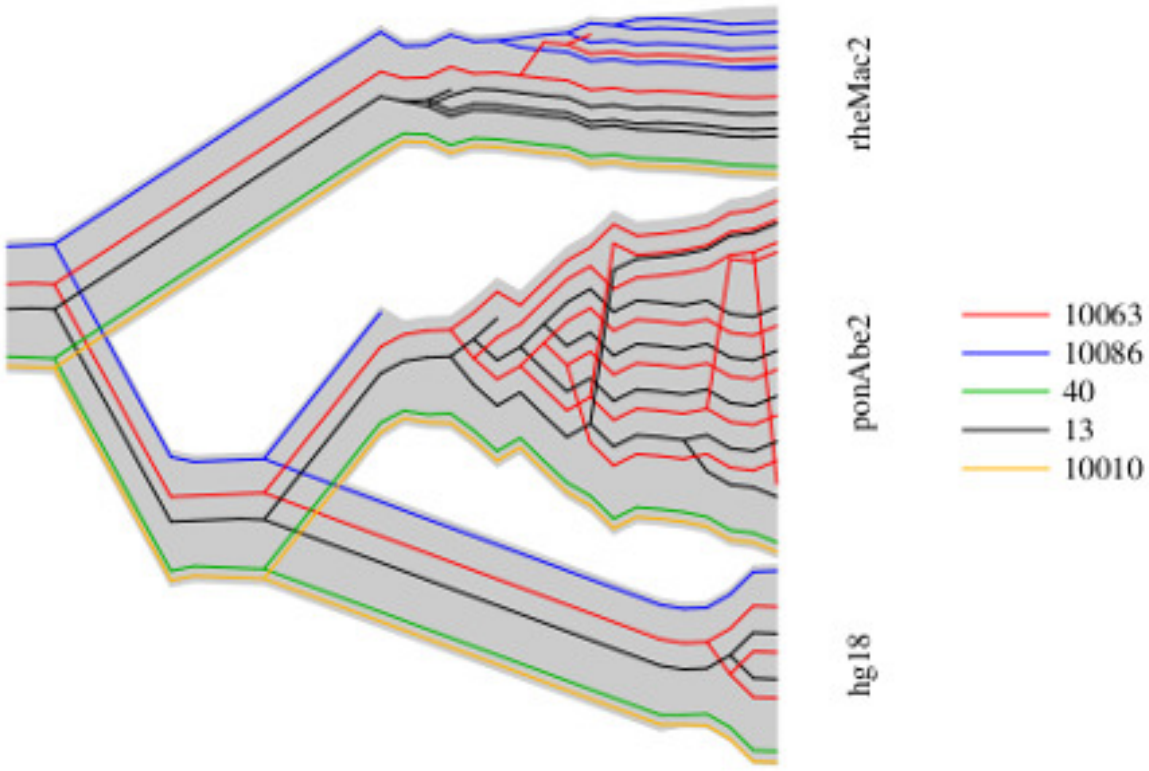
predok	e1	root	0	root	1 2 1 5 4 3 2	#	-1 -1 -1 -1 -1 -1 -1
predok	e2	e1	0.05	dup	1 2 1 2 5 4 3 2	#	0 1 2 1 3 4 5 6
clovek	e3	e2	0.12	sp	1 2 1 2 5 4 3 2	#	0 1 2 3 4 5 6 7
clovek	e4	e3	0.13	del	1 2 1 2 4 3 2	#	0 1 2 3 5 6 7
clovek	e5	e4	0.14	ins	1 2 1 6 7 2 4 3 2	#	0 1 2 -1 -1 3 4 5 6
clovek	e6	e5	0.2	inv	1 -1 -2 6 7 2 4 3 2	#	0 2 1 3 4 5 6 7 8
clovek	e7	e6	0.25	leaf	1 -1 -2 6 7 2 4 3 2	#	0 1 2 3 4 5 6 7 8
simpanz	e8	e2	0.12	sp	1 2 1 2 5 4 3 2	#	0 1 2 3 4 5 6 7
simpanz	e9	e8	0.2	leaf	1 2 1 2 5 4 3 2	#	0 1 2 3 4 5 6 7

Tabuľka 2.1: Ukážka vymysleného vstupu v súčasnom formáte [4]

2.1.1 Navrhované zmeny

2.2 Návrh výstupu

Výstupom programu je obrázok 2.1 zakoreneného fylogenetický stromu , ktorý zobrazuje evolučné vzťahy medzi rôznymi druhmi na základe vzťahov medzi ich génmi. X-ová os reprezentuje čas, v ktorom sa jednotlivé udalosti odohrali.
Strom druhov slúži ako pozadie pre gény.
Gény sú znázornené farebnými čiarami, ktoré idú vodorovne až kým nenastane nejaká udalosť.
Duplikácia je znázornená rozvetvením génu.
Speciácia rozvetvením všetkých génov, a na rozdiel od duplikácie sa vetví aj strom druhov.
Inzercia génu je znázornená ako pridanie novej čiary, na prislúchajúce miesto do stromu druhov.
Delécia je ukončenie čiary, ktorá znázorňuje gén.
Inverzia je znázornená ako kríženie čiar preusporiadaných génov.
Leaf je znázornení ako ukončenie stromu druhov a všetkých génov v tejto vetve.
Root je začiatok stromu druhov a aj všetkých génov, ktoré sa nachádzajú v počiatočnom predkovi.



Obr. 2.1: Možný vzhľad fylogenetického stromu [5]

2.2.1 Možné zmeny

Súčasný návrh vizualizácie nezobrazuje všetky informácie zo vstupu. Jedným z údajov, ktorý sa stráca je orientácia génu. Tak isto z obrázku nieje zrejmé, v ktorom chromozóme sa gén nachádza.

Kapitola 3

Problém množinového pokrytia a výber génov

V predchádzajúcej kapitole sme si predstavili základné prvky nášho programu ktorý dostane na vstupe súbor, popisujúci evolučnú históriu a na výstupe nám vykreslí fylogenetický strom reprezentujúci danú históriu. Problém nastane, pokiaľ v histórii nachádza príliš veľa génov. Výsledný vygenerovaný obrázok sa stáva neprehľadným, a získanie informácie z neho obtiažne. Potrebujeme teda vybrať iba niektoré gény na zobrazenie tak, aby na obrázku zostali zachované podstatné informácie. V tejto kapitole si predstavíme spôsob, akým budeme vyberať ktoré gény zobrazíme, využitie *Problému množinového pokrytia* pri hľadaní daných génov a dva algoritmy ktoré riešia daný problém.

3.1 Výber génov

Najpodstatnejšou informáciou pri analýze fylogenetického stromu je pre nás to, aké udalosti sa v ňom odohrali. Budeme sa teda snažiť nájsť podmnožinu všetkých génov tak, aby všetky udalosti ostali na obrázku zachované. Zvyšné gény následne z obrázku odstránime, čo môže viesť k strate informácií ktoré považujeme za menej podstatné, ako napríklad to, koľko a ktoré geny sa nachádzajú v danej histórii, ako aj koľko a ktoré gény sú ovplyvné danou udalosťou.

3.1.1 Blok

Blok predstavuje postupnosť génov ktoré sa pred aj po kroku e.h. nachádzali vedľa seba v rovnakom poradí a jednotlivé gény nemenili svoju orientáciu. Jedná sa teda o súvislý úsek DNA ktorý počas kroku e.h. nebol prerušený. Ak sa pri delécii alebo inzercii odobralo alebo pridalo viacero génov, a nenachádza sa medzi nimi žiaden iný gén, tvoria jeden blok. Pri duplikácii gény tvoria blok ak sa nachádzali pri sebe pred

duplikáciou a rovnako aj po nej vo všetkých zduplikovaných inštanciách. Pri Inverzii sa gény nachádzajú v bloku pokiaľ sa všetkým zmení orientácia, t.j. zrotuje celý blok. Napr blok génov (4,5,-6) bude po inverzii vyzerať ako (6,-5,-4).

Pokrytie blokov

Blok považujeme za pokrytý pokiaľ sa na obrázku vyskytuje aspoň jeden gén patriaci do daného bloku. Pokrytie všetkých blokov jedného kroku e.h nám zaručí zobrazenie všetkých udalostí, aj keď nie v úplnom rozsahu, ktoré sa v danom kroku vyskytli. Musíme preto nájsť také gény, ktoré pokryjú všetky bloky v kompletnej evolučnej histórii, a tým si zaistiť zobrazenie všetkých udalostí vo výslednom fylogenetickom strome. Ako cieľ si zvolíme aby bola daná množina génov čo najmenšia.

3.2 Problém Množinového Pokrytia

Definícia Máme dané univerzum$U$, ktoré obsahuje n prvkov, a systém jeho podmnožín $S = \{P_i : P_i \subseteq U\}$, ktorý pokrýva celé univerzum $\cup_{Pi \epsilon S} Pi = U$, vybrať čo najmenšiu množinu podmnožín $C \subseteq S$ takú ktorá tiež pokryje celé Univerzum $\cup_{Pi \epsilon C} Pi = U$. Problém Množinového Pokrytia (anglicky Set Cover Problem), ďalej len *PMP*. patrí medzi NP-úplne problémy.

Príklad Pre Univerzum$U = \{1, 2, 3, 4, 5, 6\}$
a systém jeho podmnožín $S = \{\{1, 2, 3\}, \{2, 3\}, \{3, 4\}, \{3, 4, 6\}, \{5\}\}$
je riešením množina podmnožín $C = \{\{1, 2, 3\}, \{3, 4, 6\}, \{5\}\}$

3.2.1 Výber génov pomocou Problému Množinového Pokrytia

Výber takých génov ktoré pokryjú všetky bloky v celej evolučnej histórii vieme formulovať ako Problém Množinového Pokrytia Univerzum predstavuje všetky bloky ktoré sa nachádzajú v našom fylogenetickom strome. Každý gén predstavuje jednu podmnožinu, v ktorej sa nachádzajú tie bloky, cez ktoré gén prechádza. Riešením je taká množina génov, ktorých zjednotenie pokrýva všetky prvky Univerza, v našom prípade všetky bloky nachádzajúce sa v evolučnej histórii.

3.3 Riešenie Problému Množinového Pokrytia

Keďže *PMP* patrí medzi NP-ťažké problémy, znamená to že zatiaľ neexistuje, a možno nikdy ani nebude existovať algoritmus ktorý by dokázal nájsť riešenie v polynomiálnom čase. Potrebujeme sa teda rozhodnúť, či je pre nás výhodnejšie hľadať najlepšie riešenie *PMP* čo môže byť časovo náročné, alebo sa uspokojíme s približním riešením, ktoré

sme schopný nájsť aproximačným algorimtom v polynomiálnom čase, a ktoré môže taktiež predstavovať dostatočné odstránenie prebytočných génov z obrázku. Predstavíme si jeden spôsob ktorým budeme hladať úplné riešenie, jeden spôsob na nájdenie približného riešenia a v nasledujúcej kapitole porovnáme výsledky ktoré produkujú.

3.3.1 Greedy algoritmus

Greedy algoritmus patrí medzi najjednoduchšie aproximačné algoritmy, umožnuje nám v polynomiálnom čase nájsť približné riešenie *PMP*. Greedy algoritmus v každom kroku pridá do riešenia takú podmnožinu, ktorá obsahuje najviac zatiaľ nepokrytých prvkov univerza. Riešenie teda hľadáme nasledovným spôsobom:
Všetky podmnožiny zoradíme na základe toho, koľko prvkov obsahujú. Do riešenia vyberieme najvačšiu podmnožinu a prvky, ktoré sa v nej nachádzajú odstránime z univerza aj zo zvyšných podmnožín. Zvyšné podmnožiny opäť zoradíme podľa veľkosti, a postup opakujeme až pokým nie je Univerzum prázdne.
Tesná analýza podľa Slavik

3.3.2 ILP

Náš Set Cover problém zapíšeme/pretransformujeme ako *Integer Linear Programming* problém

Záver

V závere je potrebné v stručnosti zhrnúť dosiahnuté výsledky vo vzťahu k stanoveným cieľom. Rozsah záveru je minimálne dve strany. Záver ako kapitola sa nečísluje.

Literatúra

[1] Albert Herencsár. An improved algorithm for ancestral gene order reconstruction. Master's thesis, Comenius University in Bratislava, 2014. Supervised by Broňa Brejová.

[2] Ján Hozza. Rekonštrukcia duplikačných histórií pomocou pravdepodobnostného modelu. Bachelor thesis, Comenius University in Bratislava, 2014. Supervised by Tomáš Vinař.

[3] Jakub Kovac, Brona Brejova, and Tomas Vinar. A Practical Algorithm for Ancestral Rearrangement Reconstruction. In Teresa M. Przytycka and Marie-France Sagot, editors, *Algorithms in Bioinformatics, 11th International Workshop (WABI)*, volume 6833 of *Lecture Notes in Computer Science*, pages 163–174, Saarbrücken, Germany, September 2011. Springer.

[4] Tomas Vinar and Brona Brejova. Biowiki, 2014.

[5] Tomas Vinar, Brona Brejova, Giltae Song, and Adam C. Siepel. Reconstructing Histories of Complex Gene Clusters on a Phylogeny. *Journal of Computational Biology*, 17(9):1267–1279, 2010. Early version appeared in RECOMB-CG 2009.

[6] David P. Williamson and David B. Shmoys. *The Design of Approximation Algorithms*. Cambridge University Press, New York, NY, USA, 1st edition, 2011.

[7] M.J. Zvelebil and J.O. Baum. *Understanding Bioinformatics*. Garland Science, 2008.