Rekonštrukcia evolučných histórií s duplikáciami pomocou simulovaného žíhania

Matej Krajčovič, doc. Mgr. Bronislava Brejová, PhD.

2018-06-14 Thu

Outline

Pojmy (1)

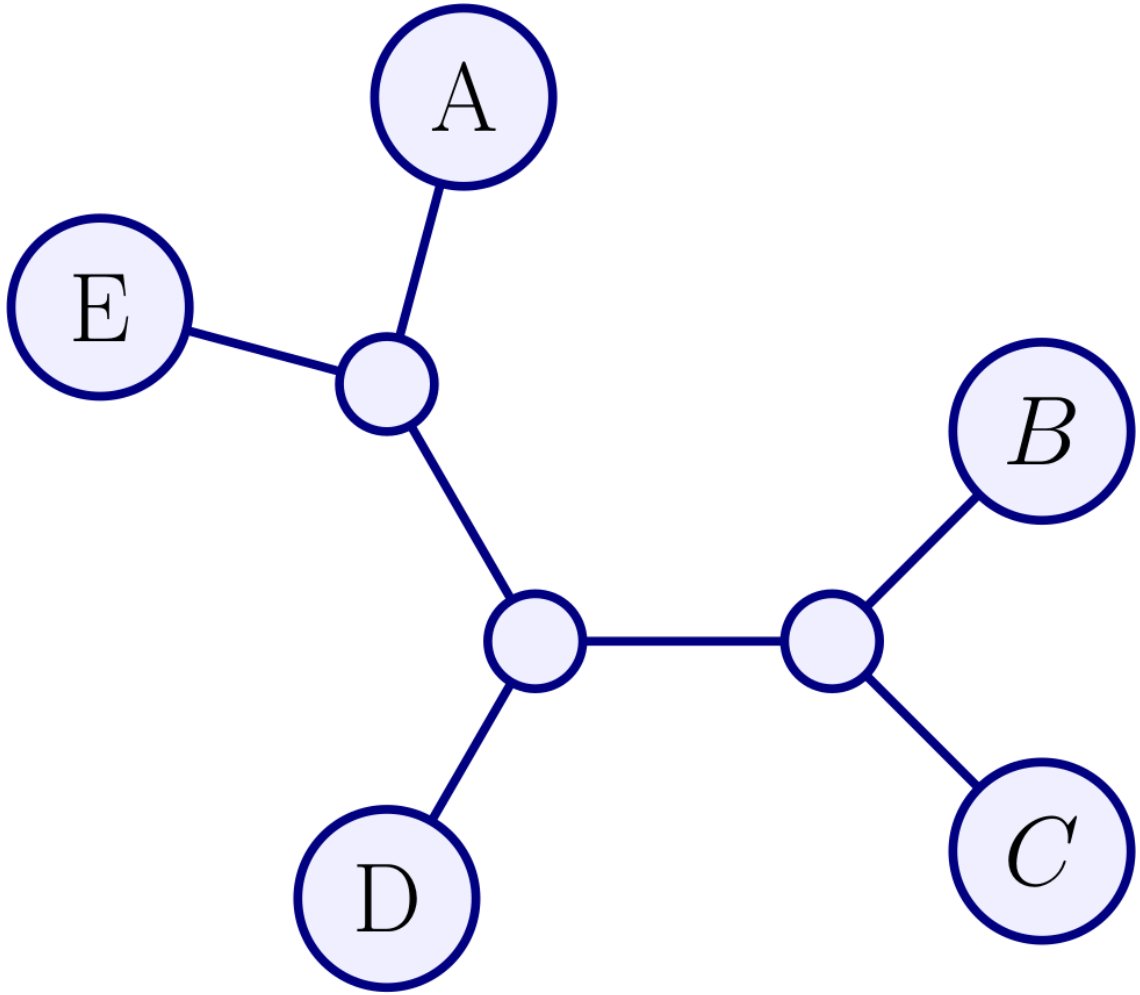
- DNA sekvencia - postupnosť znakov A, T, C, G
- atóm
  - dlhšia podsekvencia DNA
  - ak sa v DNA sekvencia nachádza viackrát s malými odlišnosťami, označia sa ako rovnaká trieda atómov
  - 20 až 200 atómov v porovnaní s 200 000 až 400 000 bázami DNA sekvencie

Atómy

AACGT	CGATCG	CGTTCG	TGCAA
1	2	2	-1

- stromy atómov
  - nezakorený strom
  - strom pre každú triedu atómov
  - podobné varianty atómov sú blízko seba $\rightarrow$ zrejme nedávno vznikli duplikáciou
- postupnosť atómov

"Čerešňovitosť"



Pojmy (2)

- udalosť - duplikácia, duplikácia s inveziou, delécia
- súčasná sekvencia - pracujeme s ňou
- ancestrálna sekvencia - pôvodná, hľadáme ju
- história - postupnosť udalostí, ktoré zmenia ancestrálnu sekvenciu na súčasnú

Udalosti

AACGT	CGATCG	CGTTCG	TGCAA
1	2	2	-1

Možné udalosti:

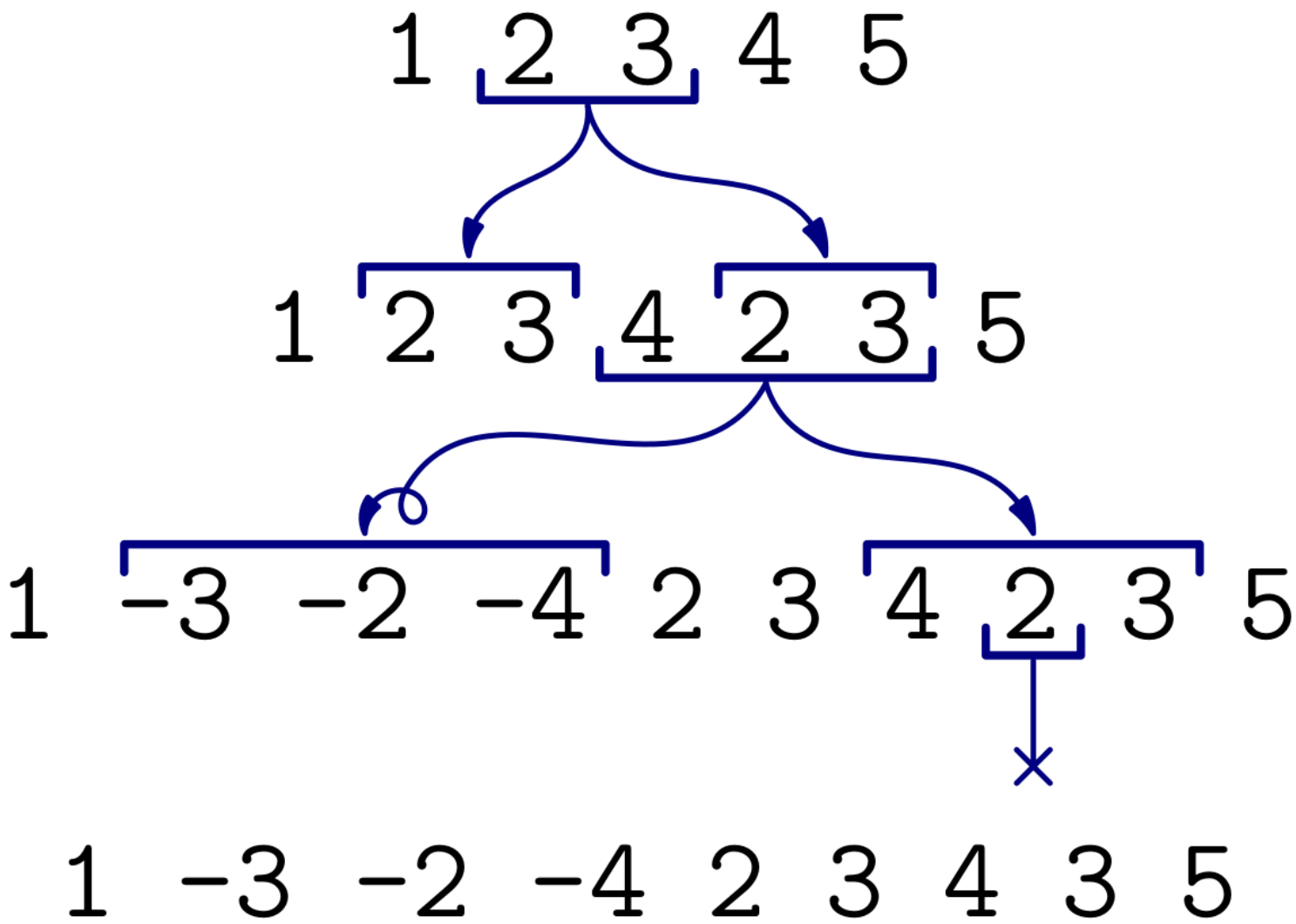
- skopíruj a otoč 1 na -1
- skopíruj a otoč -1 na 1
- skopíruj 2 na 2

Všetkých duplikačných udalostí s jednou deléciou je $\Theta(n^5)$.

Definícia problému - rekonštrukcia histórie

- Vstup:
  - súčasná postupnosť atómov
  - stromy atómov
- Výstup:
  - "najlepšia" história vzhľadom na zadanú postupnosť atómov a stromy atómov
  - čo je najlepšia história?
    - ideálne zhodná so správnou

Ukážka rekonštrukcie histórie



Existujúce práce

- problému sa už dlhodobo venujeme na fakulte a vzniklo viacero článkov a prác
- Ján Hozza, "Rekonštrukcia histórií génových zhlukov" (2016)
- Michal Anderle, "Časovanie udalostí pri inferencii duplikačných histórií" (2014)

Rekonštrukcia histórií génových zhlukov

- pokiaľ sa v sekvencii nevyskytuje každý atóm práve raz:
  - nájdi najlepšiu udalosť
  - odstráň ju a dostaň novú sekvenciu
- nájdenie najlepšej udalosti:
  - jednoduché navzorkovanie malého množstva udalostí
  - zložitejšie hodnotenie a výber tej najlepšej pomocou logistickej regresie (dĺžka atómov, vzdialenosť medzi nimi, . . .)
- generuje veľa rekonštrukcií a výstupom sú tie s najmenším počtom udalostí

Časovanie udalostí pri inferencii duplikačných histórií

- pravdepodobnostný model evolúcie
  - zmeny jednej bázy - Jukes-Cantorov substitučný model
  - modeluje aj udalosti
- aká je pravdepodobnosť, že model vygeneruje práve túto históriu?
- hľadá také časy medzi udalosťami, aby bola pravdepodobnosť najvyššia
- vyberie náhodnú hranu a upraví čas

Hodnotenie správnosti rekonštrukcie - Jaccardov index

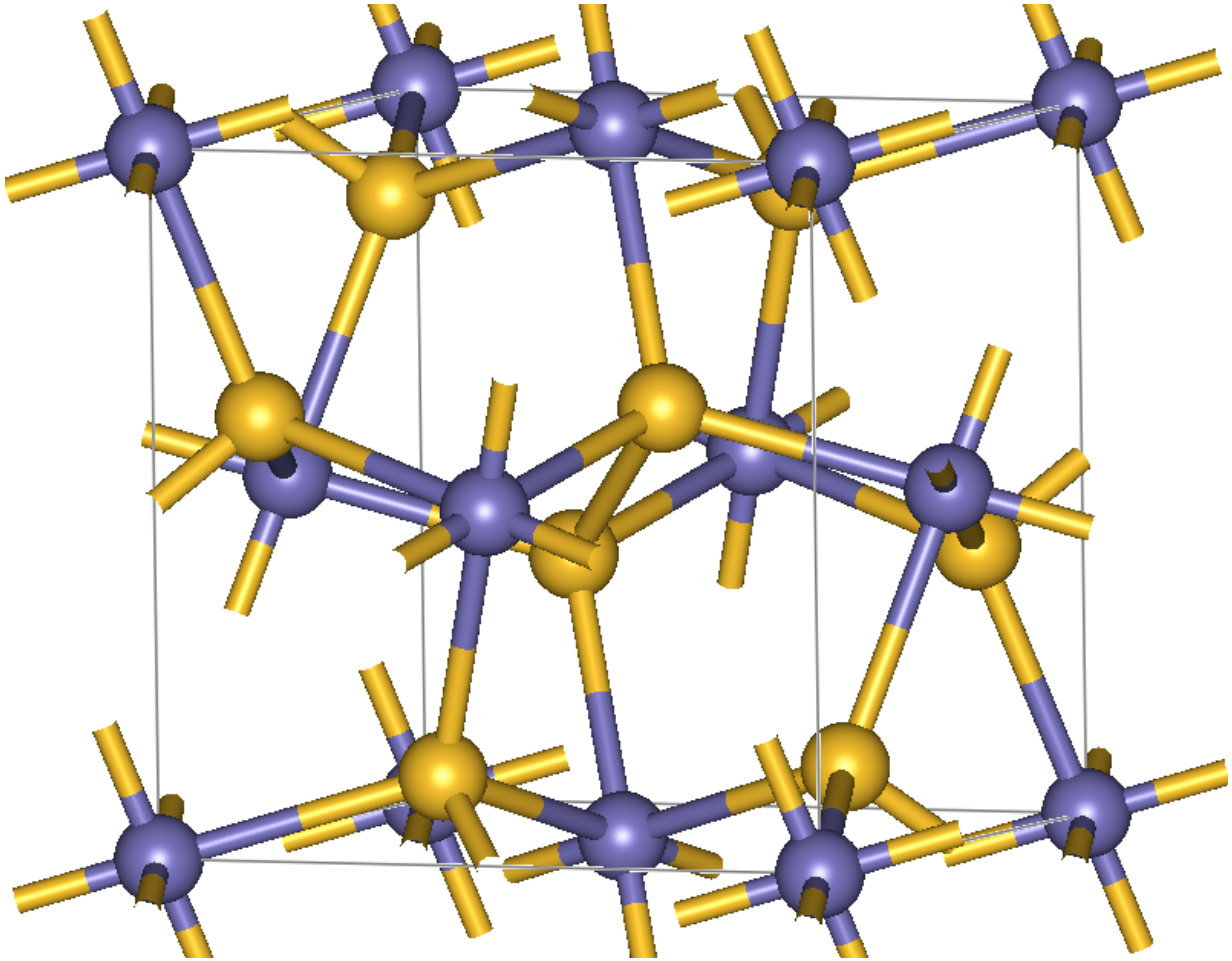
- Čo je správna rekonštrukcia? Niektoré udalosti sú navzájom nezávislé.
- A - množina udalostí (kopírovaných alebo mazaných postupností atómov) v správnej histórii
- B - množina udalostí v rekonštruovanej histórii

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$$0 \leq J(A, B) \leq 1$$

Žíhanie

- kryštalická štruktúra



Simulované žíhanie

- metaheuristika - nemusí nájsť najlepšie riešenie, ale často vie nájsť pomerne rýchlo dosť dobré riešenie
- parametre problému:
  - stavový priestor
  - skórovacia funkcia - energia stavu
  - generovanie susedov - mierna zmena stavu
- parametre metódy:
  - počiatočná teplota
  - chladnúci rozvrh - ako závisí teplota systému od času?
  - akceptačná pravdepodobnosť
    - lepší stav akceptujem vždy, horší niekedy
    - závisí od teploty a rozdielu energií dvoch stavov

Problém obchodného cestujúceho

- stavový priestor - permutácie miest
- skórovacia funkcia - cena potrebná na prechod mestami v danom poradí
- generovanie susedov - vymením niektoré dve po sebe idúce mestá a zmením smer cesty medzi nimi
- pre 20 miest obsahuje stavový priestor 20! stavov, ale je možné dostať sa do ľubovoľného stavu na 20 krokov

Aplikovanie na rekonštruovanie histórií

- stavový priestor - všetky histórie pre danú postupnosť atómov
- skórovacia funkcia:
  - počet udalostí - jednoduchý výpočet
  - vierohodnosť histórie
    - pomalý výpočet (nezávisí od počtu atómov, ale od počtu báz)
    - parametrom sa určuje počet vylepšení hrán
- generovanie susedov:
  - pri generovaní novej histórie preferujem udalosti použité v predošlej akceptovanej histórii
  - zmením jednu udalosť a zvyšné dopočítam
    - rozdelenie na tri fázy

Testovacie dáta

- 20 simulovaných histórií trvajúcich čas 0.04
- 20 simulovaných histórií trvajúcich čas 0.06

Porovnanie koeficientov korelácie skórovacích funkcií so Jaccardovým indexom

História	**Počet udalostí**	**Vierohodnosť**
F4100	-0.65	0.65
F4104	-0.56	0.47
F6100	-0.69	0.76
F6101	-0.67	0.60

Pomery znovu použitých udalostí pri metóde zvýhodňovania minulých udalostí

Pravdepodobnosť	0	0.3	0.6	0.9
Minimum	0.25	0.24	0.17	0.16
Maximum	0.72	0.62	0.68	0.65
Priemer	0.40	0.40	0.40	0.36

Pomery znovu použitých udalostí pri metóde dopočítania zvyšných udalostí

Fáza	1	2	3
Minimum	0.21	0.59	0.92
Maximum	0.88	1.00	1.00
Priemer	0.50	0.81	0.97

Porovnanie metód generovania susedov a parametrov pre histórie dlhé 0.04 (porovnanie počtom udalostí)

	lepšie	rovnaké	horšie
likelihood-100	5	3	12
likelihood-prob-0.3	1	13	6
likelihood-prob-0.6	2	8	10
likelihood-prob-0.9	2	4	14
num_events-0.04	9	9	2
num_events-prob-0.3	2	9	9
num_events-prob-0.6	4	6	10
num_events-prob-0.9	4	11	5

Výsledky

- spravili sme porovnanie rekonštrukcií aj podľa vierohodnosti a pre histórie dlhé 0.06
- výsledky boli porovnateľné
- záver
  - metóda vierohodnosť je v porovnaní s počtom udalostí pomalšia a nedáva výrazne lepšie výsledky
  - o metóde zmenenia udalosti a dopočítania zvyšku histórie sa v porovnaní so zvýhodňovaním použitých udalostí jednoduchšie uvažuje a má očakávateľné výsledky

Implementácia

- simulované žíhanie doimplementované do kódu Jána Hozzu
- počítanie pravdepodobnosti od Michala Anderleho ako knižnica
- bolo opravených veľa chýb v pôvodnom kóde

Záver

Ďakujem za pozornosť.

Otázky

- Aké iné skórovacie funkcie by bolo možné použiť tak, aby korelovali s Jaccardovým indexom?
- V práci na strane 24 píšete, že čas behu 100 rekonštrukcií sa zvýšil na hodiny. Skúšali ste nejakým spôsobom výpočet zrýchliť a zefektívniť? Ako?