Univerzita Komenského, Bratislava
Fakulta Matematiky, Fyziky a Informatiky

MCMC algoritmus na rekonštrukciu duplikačných histórií

Bc. Martin Kravec

2011

Katedra Informatiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského, Bratislava

MCMC algoritmus na rekonštrukciu duplikačných histórií

(Diplomová práca)

Bc. Martin Kravec

(Kód: 03f033bc-7faf-4c5b-bcdb-c66f77a8ab36)

Študijný odbor: 9.2.1 informatika
Študijný program: informatika

Vedúci práce: Mgr. Tomáš Vinař, PhD

Bratislava, 2011

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Martin Kravec
Študijný program: informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.2.1. informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský

Názov: MCMC algoritmus na rekonštrukciu duplikačných histórií

Cieľ: Duplikované regióny genómov patria medzi najzložitejšie štruktúry analyzované metódami komparatívnej genomiky. K takejto analýze je potrebná rekonštrukcia histórie duplikácií v rámci regiónu. Cieľom práce je vytvoriť algoritmus na rekonštrukciu duplikačnej histórie s využitím metód MCMC vzorkovania, ako aj jeho implementácia.

Vedúci: Mgr. Tomáš Vinař, PhD.

Dátum zadania: 13.11.2009

Dátum schválenia: 18.02.2011

prof. RNDr. Branislav Rovan, PhD.
garant študijného programu

.....
študent

.....
vedúci

Čestne prehlasujem, že som túto diplomovú prácu vypracoval samostatne pod dohľadom vedúceho diplomovej práce a s použitím citovaných zdrojov.

. . . . . .

V Bratislave, dňa 6. 5. 2011 Martin Kravec

Poďakovanie

V prvom rade patrí moja vďaka Tomášovi Vinařovi a Broni Brejovej, bez ktorých by táto práca nikdy nevznikla, za čas, ktorý mi venovali a za cenné rady a pripomienky, ktoré mi poskytli. Patrí im môj obdiv za ich entuziazmus a inovatívny prístup k študentom, ktoré sú skvelou motiváciou a inšpiráciou pre celé ich okolie. Okrem toho im moja veľká vďaka patrí aj za možnosť byť členom ich bioinformatickej skupiny a za úžasnú atmosféru, ktorú v nej šíria.

Ďakujem tiež svojim rodičom za ich dlhoročnú podporu v štúdiu a priateľom, ktorí mi boli vždy ochotní pomôcť.

Abstrakt

Autor:	Bc. Martin Kravec
Názov práce:	MCMC algoritmus na rekonštrukciu duplikačných histórií
Škola:	Univerzita Komenského v Bratislave
Fakulta:	Fakulta matematiky, fyziky a informatiky
Katedra:	Katedra informatiky
Vedúci práce:	Mgr. Tomáš Vinař, PhD

V bunkách živých organizmov sa nachádza DNA, ktorá je nosičom genetickej informácie. Ak máme DNA sekvencie súčasných druhov, zaujímavou úlohou je predikcia ancestrálnej DNA sekvencie. Za predpokladu, že v rámci evolúcie uvažujeme iba proces jednoduchých mutácií, t.j. substitúcií jedného znaku za iný, je rekonštrukcia ancestrálnych sekvencií pomerne jednoduchá. Situácia sa komplikuje, ak začneme uvažovať zložitejšie operácie väčšieho rozsahu, duplikácie. V tejto práci zavedieme pravdepodobnostný model evolúcie duplikovaných úsekov DNA (génových zhlukov) a inferenciu vzorkovaním za pomoci MCMC algoritmu. Ukážeme algoritmus na výpočet vierohodnosti histórií, ale najmä predstavíme nový algoritmus pre navrhovanie jednotlivých duplikácií pomocou dynamického programovania a stochastického spätného prechodu. Predbežné výsledky ukazujú, že implementácia tohoto algoritmu umožňuje rýchle MCMC vzorkovanie a získanie kvalitných výsledkov pri aplikácii tejto metódy na reálne dáta.

Kľúčové slová: Markov chain Monte Carlo (MCMC), rekonštrukcia, duplikácia, história, evolúcia, bioinformatika.

Abstract

Author:	Bc. Martin Kravec
Caption:	MCMC algorithm to reconstruct duplication history
University:	Comenius University in Bratislava
Faculty:	Faculty of Mathematics, Physics and Informatics
Department:	Department of Computer Science
Supervisor:	Mgr. Tomáš Vinař, PhD

DNA found in the cells of living organisms is the carrier of genetic information. From DNA sequences of extant species, we can attempt prediction of ancestral DNA sequences. When considering the evolutionary process with simple mutations only (substitution of one character by another), the prediction of ancestral sequences is quite simple. However, including more complex, larger scale operations, such as duplications, leads to more complex problems. In this work, we introduce a probabilistic model of evolution of duplicated segments of DNA (gene clusters) and we propose inference using MCMC sampling algorithm. We show how to compute likelihood of a particular history in this model, but more importantly we introduce a novel algorithm for sampling duplications using dynamic programming and stochastic traceback. Preliminary results indicate that the implementation of this algorithm will allow for fast MCMC sampling and obtaining quality results by applying this method to real data.

Key words: Markov chain Monte Carlo (MCMC), reconstruction, duplication, history, evolution, bioinformatics.

Obsah

Úvod **1**

1 Biologické pozadie problému **2**

2 Pravdepodobnostný model evolúcie zhlukov génov **5**

2.1 Jukes-Cantorov model mutácií 5

2.2 Pravdepodobnostný model evolúcie sekvencie s duplikáciami 8

2.3 Počítanie vierohodnosti histórie 10

2.4 Obmedzenia modelu 13

3 MCMC vzorkovanie **15**

3.1 MCMC algoritmus 15

3.2 Atomizácia 17

3.3 Navrhovacie rozdelenie 19

3.4 Navrhovací algoritmus 19

4 Vzorkovanie duplikačnej udalosti **22**

4.1 Základný model 24

4.2 Základný model s reverziou 29

4.3 Rozšírený model s deléciami 32

4.4 Hlavný model 36

4.5 Speciácie 39

5 Aplikácia na dáta **40**

Záver **45**

Literatúra **47**

Zoznam obrázkov

1.1 Príklad duplikácie sekvencie $TGACTGG$. 3
1.2 Jednoduchá duplikačná história dvoch duplikačných udalostí a jednej mutácie. 3

2.1 Vľavo duplikačná história a vpravo jeden zo stromov evolúcie mutácií pre vyznačený znak. 10

3.1 Celková schéma MCMC algoritmu aplikovaná na duplikačné histórie. Blok označený ako A reprezentuje navrhovací algoritmus. Časť označená ako B reprezentuje rýchly algoritmus na vzorkovanie jednej duplikácie. 18
3.2 Príklad duplikácie sekvencie atómov b c. Na obrázku vidíme len príklad, ako by mohla byť DNA sekvencia rozdelená na atómy. Rovnaké písmená označujú rovnaký typ atómu. V skutočnosti majú atómy obvykle dlhšiu sekvenciu, väčšinou viac ako 500 báz. 19

4.1 Úlohou bolo nájsť duplikáciu v sekvencii $\langle\ a\ b\ c\ d\ e\ b'\ c'\ d'\ a'\ b''\ e'\ \rangle$. Na obrázku sú zobrazené len aktívne vrcholy matice S. Ostatné vrcholy mali hodnotu 0 a stupeň vrchola tiež 0. 26
4.2 Tri rozličné cesty nájdené spätným stochastickým prechodom. 28
4.3 Problematická neplatná duplikácia. 29
4.4 Príklad reverznej duplikácie $\langle\ b\ c\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ \overline{c'}\ \overline{b'}\ \rangle$. 30
4.5 Príklad základného modelu duplikácií s reverziou. 31
4.6 Príklad rozšíreného modelu duplikácií s deléciami. 34
4.7 Príklad cesty zobrazujúcej duplikáciu s deléciou. 35
4.8 Príklad hlavného modelu duplikácií s deléciami obsahujúci aj bonusové hrany. 38

5.1 Fylogenetický strom troch skúmaných druhov aj s dĺžkami hrán. . . . 40
5.2 Simulovaná história. Každá farba predstavuje jeden typ atómu. 43
5.3 Rekonštruovaná duplikačná história s viac penalizovanými deléciami. 43
5.4 Rekonštruovaná duplikačná história s menej penalizovanými deléciami. 44
5.5 Rekonštruovaná duplikačná história s menej penalizovanými deléciami a zvýhodnenými speciáciami. 44

Úvod

Jednou z nových moderných vied súčasnosti je bioinformatika, čo je vlastne veda na pomedzí biológie, informatiky a štatistiky. Príchodom sekvenovacích technológií sa pre biológov stalo nemožné spracovávať množstvo informácií o organizmoch iba ľudskými zdrojmi. Preto sa k slovu dostala informatika. V našej práci sa budeme venovať analýze biologických dát, konkrétne evolúcii; procesom, akým sa mení genóm organizmov v historii.

V rámci evolúcie môžeme uvažovať jednoduché (mutácie), ale aj zložitejšie procesy. Vinař et al. (2009)[VBSS09] uvažujú vo svojej práci operácie väčšieho rozsahu, duplikácie. Zavedú pravdepodobnostný model duplikácií a inferenciu vzorkovaním za pomoci MCMC algoritmu. Neoddeliteľnou súčasťou takéhoto algoritmu je počítanie vierohodnosti jednotlivých histórií a pravdepodobnostný algoritmus pre navrhovanie jednotlivých duplikácií.

Vinař et al. navrhujú iba veľmi jednoduchý algoritmus pre navrhovanie duplikacií, ktorého implementácia je pomerne neefektívna, čoho dôsledkom je značne pomalá iterácia v rámci MCMC algoritmu. V našej práci zavedieme veľmi podobný pravdepodobnostný model, ukážeme algoritmus na výpočet vierohodnosti histórií, ale najmä predstavíme nový prístup k navrhovaniu jednotlivých duplikácií. Kým v predchádzajúcej práci trval výber jednej duplikácie čas $O(n^5)$, náš novonavrhnutý algoritmus funguje v čase $O(n^2)$. Prototyp novonavrhnutého algoritmu sme implementovali a odskúšali na simulovaných DNA sekvenciách.

V 1. kapitole načrtneme biologické pozadie problému rekonštrukcie duplikačných histórií. V 2. kapitole zadefinujeme pravdepodobnostný model evolúcie. Ďalšie dve kapitoly popisujú MCMC algoritmus a jeho konkrétnu aplikáciu na náš problém. V rámci nich predstavíme navrhovací algoritmus a náš algoritmus na vzorkovanie duplikácií. Na záver ukážeme zopár výsledkov skúšobnej implementácie nášho algoritmu spustených na simulovaných dátach.

Kapitola 1

Biologické pozadie problému

V bunkách živých organizmov sa nachádza DNA, ktorá je nosičom genetickej informácie. Rôznymi biotechnologickými procesmi sa dá extrahovať z buniek a premeniť (osekvenovať) na reťazec nad abecedou $\{A, C, G, T\}$. Evolučná teória predpokladá, že každá dvojica organizmov rôznych druhov vznikla zo spoločného predka. Ak máme DNA sekvencie súčasných druhov, zaujímavou úlohou je predikcia ancestrálnej DNA sekvencie.

Za predpokladu, že v rámci evolúcie uvažujeme iba proces jednoduchých mutácií, t.j. substitúcií jedného znaku za iný, je rekonštrukcia ancestrálnych sekvencií pomerne jednoduchá. Existuje množstvo pravdepodobnostných modelov (napr. Jukes-Cantorov model), na základe ktorých je možné pre danú pozíciu predpovedať najpravdepodobnejší ancestrálny znak na danej pozícii.

Situácia sa komplikuje, ak začneme uvažovať zložitejšie operácie väčšieho rozsahu. Uvažujme preto problém rekonštrukcie ancestrálnych sekvencií za prítomnosti duplikačných udalostí:

- Duplikácia - skopírovanie určitej časti DNA sekvencie na iné miesto. Príklad duplikácie môžeme vidieť na obrázku 1.1.

- Reverzná duplikácia (reverzia) - duplikácia, v ktorej sa duplikovaná časť sekvencie kopíruje odzadu a komplementárne čo znamená, že

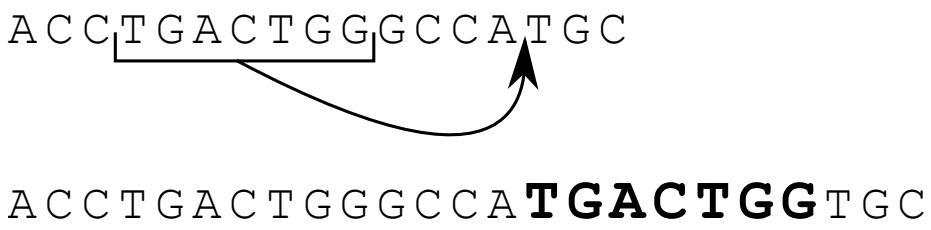
A sa kopíruje na T,

C sa kopíruje na G,

G sa kopíruje na C,

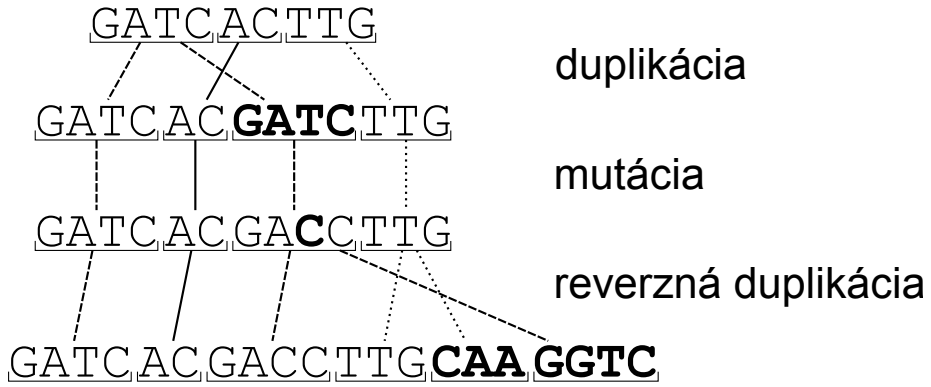
T sa kopíruje na A,

- Delécia - vymazanie časti DNA sekvencie.



Obr. 1.1: Príklad duplikácie sekvencie $TGACTGG$.

Zoznam duplikačných udalostí vytvorí duplikačnú históriu, ktorá je návodom, ako sa z jedného pôvodného organizmu vyvinuli súčasné druhy. Ukážku duplikačných udalostí a mutácií vidíme na obrázku 1.2. Pôvodnou sekvenciou v histórii bola DNA sekvencia $GATCACTTG$. Prvou udalosťou v histórii bola duplikácia sekvencie $GATC$ medzi sekvenciu $GATCAC$ a sekvenciu TTG. Následne sa udiala mutácia, zmena znaku T na C a poslednou udalosťou bola reverzná duplikácia sekvencie $GACCTTG$ na koniec sekvencie. Výsledná sekvencia bola potom $GATCACGACCTTGCAAGGTC$.



Obr. 1.2: Jednoduchá duplikačná história dvoch duplikačných udalostí a jednej mutácie.

Ak začneme skúmať históriu viacerých druhov organizmov, musíme zahrnúť ešte jednu dôležitú udalosť, speciáciu. Z biologického hľadiska je to proces, kedy sa z jedného druhu oddelila vetva iných organizmov, ktoré sa už s pôvodným druhom nemôžu

krížiť. Pokiaľ pôvodný alebo oddelený druh neprežil v nejakej podobe doteraz, nemáme možnosť zaznamenať túto speciáciu v histórii a preto nás budú zaujímať len také speciácie medzi organizmami, pri ktorých vieme získať ich DNA sekvenciu. Z informatického hľadiska je speciácia rozdelenie ancestrálnej sekvencie na dve totožné sekvencie, ktoré sa ďalej vyvíjajú nezávisle.

Našou úlohou je teda zo sekvencií súčasných organizmov získať históriu duplikácií a ancestrálne sekvencie o ktorých predpokladáme, že vznikli vyššieuvedeným procesom. Pripomína to návrat v čase, postupné nachádzanie udalostí, odstraňovanie častí sekvencií, ktoré tieto udalosti vyrobili (napríklad pri duplikácií vymazať duplikovanú časť sekvencie) a postupné získavanie ancestrálnych sekvencií organizmov na vstupe.

Ak by sme chceli analyzovať celé genómy, museli by sme náš model rozšíriť o ďalšie operácie, ktoré presúvajú veľké časti genómu, separujú alebo spájajú chromozómy. Takýmto modelom sa zaoberali napríklad Ma et al.(2008) [MRR+08]. Ich prístup má viacero predpokladov, ktoré nemožno aplikovať na zhluky génov, ktorými sa zaoberáme my, preto sme sa rozhodli použiť v tejto práci iný prístup. Hlavné rozdiely sú v tom, že náš prístup sa snaží vytvárať podstatne „podrobnejšie“ duplikačné histórie a namiesto minimalizovania počtu operácií sme zaviedli pravdepodobnostný model, ktorý popíšeme v nasledujúcej kapitole.

Kapitola 2

Pravdepodobnostný model evolúcie zhlukov génov

Základnou ideou generatívnych pravdepodobnostných modelov je vytvorenie procesu, ktorý by nám za normálnych okolností umožnil simulovať údaje (v našom prípade umožnil simulovať evolúciu). Taktiež sa generatívnym pravedpodobnostným modelom môžeme pozrieť na dáta (napríklad sekvencie súčasných organizmov) a pýtať sa, akým najpravdepodobnejším spôsobom nám model tieto dáta dokáže vygenerovať, prípadne porovnať rôzne histórie, akú pravdepodobnosť má vygenerovanie danej sekvencie touto históriou (tzv. vierohodnosť histórie). V našom prípade je odpoveďou na prvú otázku konkrétna postupnosť duplikácií, delécií a mutácií a príslušná ancestrálna sekvencia, z ktorej daná sekvencia mohla vzniknúť. V tejto kapitole zavedieme pravdepodobnostný model a ukážeme efektívny spôsob na výpočet vierohodnosti.

2.1 Jukes-Cantorov model mutácií

Základným biologickým procesom popisujúci zmeny v genóme jedného druhu sú mutácie. Tradičným pravdepodobnostným modelom mutácií je Jukes-Cantorov model [JC69], ktorý predpokladá, že všetky substitúcie znakov sa dejú s rovnakou pravdepodobnosťou.

Nech $P(b|a,t)$ je pravdepodobnosť, že ak začneme so znakom a, tak po čase t budeme mať znak b. Pre dané t môžeme takéto pravdepodobnosti usporiadať do matice S, kde $S_{a,b}(t) = P(b|a,t)$

$$S(t) = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & P(A|A,t) & P(A|C,t) & P(A|G,t) & P(A|T,t) \\ C & P(C|A,t) & P(C|C,t) & P(C|G,t) & P(C|T,t) \\ G & P(G|A,t) & P(G|C,t) & P(G|G,t) & P(G|T,t) \\ T & P(T|A,t) & P(T|C,t) & P(T|G,t) & P(T|T,t) \end{array}$$

Intuitívne platí, že s pribúdajúcim časom sa zväčšuje pravdepodobnosť zmeny znaku. Na začiatku je $S(0) = I$ (identická matica). Pre $t \to \infty$ má $S(t)$ všetky riadky rovnaké (napríklad $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$). Budeme uvažovať model, v ktorom pravdepodobnosť mutácie závisí len od aktuálneho znaku a nie od predchádzajúcich stavov. Takýto model nazývame multiplikatívny a platí v ňom, že ak máme časy t_1 a t_2, tak vieme spočítať pravdepodobnosť pre čas $t_1 + t_2$:

$$P(b|a, t_1 + t_2) = \sum_{x \in \sum} P(x|a, t_1) \cdot P(b|x, t_2)$$

V maticovej notácii:

$$S(t_1 + t_2) = S(t_1) \cdot S(t_2)$$

Ak by sme uvažovali iba diskrétne časy, stačilo by nám určiť hodnoty matice $S(1)$ a pravdepodobnosti pre všetky ostatné časy by sme dostali umocnením tejto matice. My však potrebujeme definovať maticu $S(t)$ pre reálne t. Pre veľmi malý čas ϵ máme maticu

$$S(\epsilon) = \begin{pmatrix} 1-3p & p & p & p \\ p & 1-3p & p & p \\ p & p & 1-3p & p \\ p & p & p & 1-3p \end{pmatrix}$$

kde p je pravdepodobnosť mutácie znaku a má tiež veľmi malú hodnotu. Pre čas 2ϵ dostaneme maticu

$$S(2\epsilon) = S(\epsilon^2) = \begin{pmatrix} 1-6p+12p^2 & 2p-4p^2 & 2p-4p^2 & 2p-4p^2 \\ \cdots & & & \end{pmatrix}$$

Členy s p^2 sú ale oveľa menšie ako členy s p, takže matica S je približne

$$S(2\epsilon) = S(\epsilon^2) \approx \begin{pmatrix} 1-6p & 2p & 2p & 2p \\ \cdots & & & \end{pmatrix}$$

Vytvoríme si teraz maticu rýchlosti R.

$$R = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

Dostávame, že pre veľmi malé časy platí $S(\epsilon) \approx I + R\epsilon$ (p používané vyššie by malo hodnotu $\frac{\alpha}{\epsilon}$)

$$S(t+\epsilon) = S(t)S(\epsilon) \approx S(t)(I + R\epsilon)$$
$$\frac{S(t+\epsilon) - S(t)}{\epsilon} \approx S(t)R$$

V limite pre $\epsilon \to 0$ dostávame diferenciálne rovnice $S'(t) = S(t)R$ s počiatočným stavom $S(0) = I$. Ak teraz diagonálne prvky $S(t)$ označíme $r(t)$ a nediagonálne $s(t)$, dostaneme, že diagonálny prvok $S(t)R$ je $(-3\alpha r(t) + 3\alpha s(t))$ a nediagonálny je $(-\alpha s(t) + \alpha r(t))$. Takže dostávame diferenciálne rovnice

$$r'(t) = 3\alpha s(t) - 3\alpha r(t) \qquad \text{a} \qquad s'(t) = \alpha r(t) - \alpha s(t)$$

Riešením týchto rovníc dostaneme

$$r(t) = \frac{1 + 3e^{-4\alpha t}}{4} \qquad \text{a} \qquad s(t) = \frac{1 - e^{-4\alpha t}}{4}$$

$$S(t) = \begin{pmatrix} (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 \\ (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1-e^{-4\alpha t})/4 & (1+3e^{-4\alpha t})/4 \end{pmatrix}$$

Aby sme nemali v modeli naraz α aj t, zvykneme maticu R normalizovať tak, aby priemerný počet substitúcií za jednotku času bol 1. V prípade Jukes-Cantorovho modelu je to pri $\alpha = \frac{1}{3}$. Dostali sme teda maticu pravdepodobností rýchlostí mutácií $S(t)$.

$$S(t) = \begin{pmatrix} (1+3e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 \\ (1-e^{-\frac{4}{3}t})/4 & (1+3e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 \\ (1-e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 & (1+3e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 \\ (1-e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 & (1-e^{-\frac{4}{3}t})/4 & (1+3e^{-\frac{4}{3}t})/4 \end{pmatrix} \tag{2.1}$$

Ak teraz máme strom T_x mutácií pôvodného znaku x, vieme vypočítať pravdepodobnosť tohto stromu v histórii ako súčin pravdepodobností zmien na jednotlivých vetvách stromu T_x.

V praxi sa používajú komplikovanejšie substitučné modely, ktoré majú všeobecnejšiu maticu rýchlostí R. Napríklad HKY model (Hasegawa et al.(1985) [HKY85]), ktorý použili vo svojej práci Vinař et al. [VBSS09], umožňuje definovať rôzne pravdepodobnosti znakov A, C, G a T. V našej práci sme sa na začiatok rozhodli použiť jednoduchší model, ktorý neskôr nebude veľmi obtiažne nahradiť zložitejším modelom.

2.2 Pravdepodobnostný model evolúcie sekvencie s duplikáciami

Ak zoberieme do úvahy procesy väčšieho rozsahu, vieme vytvoriť zložitejší pravdepodobnostný model. Sekvencie druhov sa budú meniť nielen mutáciami, ale aj pomocou duplikačných udalostí (duplikácia, reverzná duplikácia a delécia). Pre vetvu v vstupného fylogenetického stromu T a históriu H definujeme pravdepodobnosť

$$P(H, v) = \left(\prod_{i=1}^{k} P(U_i|d_i) \right) \cdot \left(\prod_{e_i \in \vartheta(v)} P(k|e_i) \right) \cdot P_K(t_K) \ , \tag{2.2}$$

kde $U_1, \ldots, U_k$ sú všetky duplikačné udalosti na vetve v a $P(U_i|d_i)$ je pravdepodobnosť udalosti U_i aplikovanej na sekvenciu dĺžky d_i, $\vartheta(v)$ je množina všetkých súvislých úsekov delécií na vetve v a $P(k|e_i)$ je pravdepodobnosť, že úsek e_i obsahuje práve k delécií a $P_K(t_K)$ je pravdepodobnosť, že od poslednej udalosti po koniec vetvy uplynul čas t_K.

Nech fylogenetický strom T má vetvy $v_1, v_2, \ldots$. Následne až na normalizačnú konštantu vieme vyrátať pravdepodobnosť, že vidíme súčasné sekvencie druhov X a históriu H, nasledovným vzťahom

$$P(H, X) \propto \left(\prod_{v_i \in T} P(H, v_i) \right) \cdot \left(\prod_{T_x \in \sigma(H)} P(X|T_x) \right) \ , \tag{2.3}$$

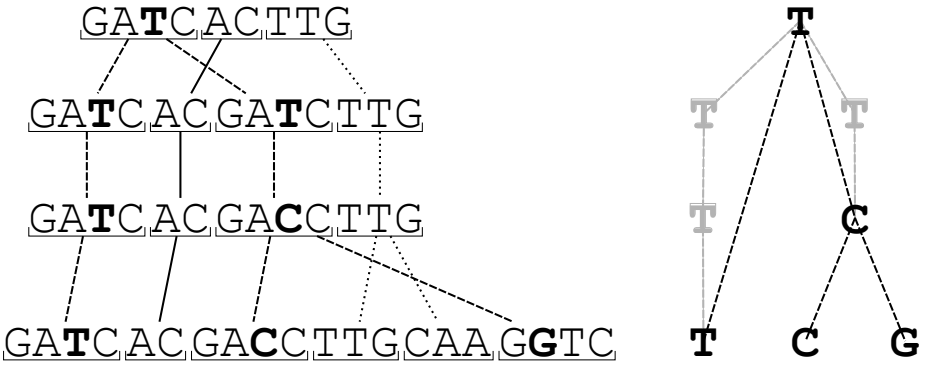
kde $P(H, v_i)$ je vyššie popísaná pravdepodobnosť vetvy v_i, $\sigma(H)$ je množina všetkých mutačných stromov T_x (popíšeme neskôr) a $P(X|T_x)$ je pravdepodobnosť sekvencií X za predpokladu mutačného stromu T_x.

Pozrieme sa teraz na pravdepodobnosť jednotlivých udalostí. Duplikáciu (aj reverznú) tvorí zdrojová a cieľová časť sekvencie a charakterizujeme ju štyrmi súradnicami:

- stredom - súradnica v strede medzi ľavým a pravým koncom zdrojovej časti sekvencie,
- dĺžkou - dĺžka zdrojovej časti sekvencie,
- vzdialenosťou - vzdialenosť medzi zdrojovou a cieľovou častou sekvencie,
- smerom - označuje buď reverznú (s pravdepodobnosťou P_R) alebo priamu (s pravdepodobnosťou $1 - P_R$) duplikáciu.

Stred je vyberaný rovnomerným rozdelením, pretože každá pozícia v sekvencii má rovnakú pravdepodobnosť stať sa stredom udalosti. Dĺžka a vzdialenosť sú vyberané geometrickým rozdelením, lebo pravdepodobnosť, že dĺžka sa rovná presne k je súčin pravdepodobností, že sa nerovná $1, \ldots, (k-1)$ ale sa rovná k. Analogicky to isté platí aj pre vzdialenosť. Treba si uvedomiť, že niektoré kombinácie stredu, dĺžky a vzdialenosti vedú k neplatným duplikáciám. Tieto neplatné kombinácie sú zamietnuté. Podobne, dvomi súradnicami, stredom a dĺžkou, charakterizujeme aj deléciu. Udalosť je buď duplikáciou s pravdepodobnosťou P_D, alebo deléciou s pravdepodobnosťou $1 - P_D$. Pre každú duplikáciu sa uplynutie času od predchádzajúcej duplikácie na tej istej vetve stromu alebo od predchádzajúcej speciácie riadi exponenciálnym rozdelením. Exponenciálne rozdelenie pravdepodobnosti sa totiž bežne používa na vyjadrenie doby čakania na určitú udalosť. Delécie priraďujeme úseku medzi dvomi duplikáciami, prípadne medzi speciáciou a duplikáciou a neurčujeme ich čas, iba počet, ktorý sa riadi poissonovským rozdelením s parametrom λ. Poissonovo rozdelenie pravdepodobnosti vyjadruje počet výskytov málo pravdepodobných javov (v našom prípade delécií) v určitom časovom intervale.

Do tohto modelu ešte potrebujeme zahrnúť mutácie. Pre každý znak x pôvodnej sekvencie vo vrchole stromu T vieme z duplikačnej histórie H zostrojiť strom T_x a to tak, že pokiaľ bol znak v duplikácii alebo speciácii, vytvoril v strome T_x vnútorný vrchol a pokiaľ bol v delécii alebo už je v súčasnej sekvencii, tvorí list stromu T_x. Príklad takto vytvoreného stromu s históriou, akú sme ukázali na obrázku 1.2, môžeme vidieť na obrázku 2.1. Množinu všetkých týchto stromov v histórii H pre všetky



Obr. 2.1: Vľavo duplikačná história a vpravo jeden zo stromov evolúcie mutácií pre vyznačený znak.

znaky x pôvodnej sekvencie vo vrchole stromu T označíme $\sigma(H)$. Pre každý strom $T_x \in \sigma(H)$ zvlášť aplikujeme Jukes-Cantorov model (popísaný v predchádzajúcej podkapitole) a podľa neho vieme vypočítať pravdepodobnosť $P(X|T_x)$.

2.3 Počítanie vierohodnosti histórie

Teraz, keď už poznáme pravdepodobnostný model, nás zaujíma možnosť vedieť efektívne vypočítať vierohodnosť histórie H s množinou sučasných sekvencií X, čiže pravdepodobnosť $P(H, X)$. Túto vieme vypočítať ako súčin pravdepodobností jednotlivých udalostí zvlášť na každej vetve stromu T a pravdepodobností mutácií znakov na jednotlivých mutačných stromoch T_x (výraz 2.3).

Pravdepodobnosť udalosti U za predpokladu dĺžky d celej sekvencie pred aplikovaním udalosti U vieme vyrátať ako

$$P(U|d) = \begin{cases} P_D \cdot P(U|d, e_s) \cdot P(U|d, e_d) \cdot P(U|d, e_v) \cdot & \text{, ak } U \text{ je duplikácia,} \\ \quad \cdot P(U|d, e_r) \cdot P(U|t) & \\ (1 - P_D) \cdot P(U|d, e_s) \cdot P(U|d, e_d) & \text{, ak } U \text{ nie je duplikácia,} \end{cases} \tag{2.4}$$

kde

$P(U|d, e_s)$ je pravdepodobnosť udalosti U pri dĺžke d a stredom udalosti so súradnicou e_s,

$P(U|d, e_d)$ je pravdepodobnosť udalosti U pri dĺžke d a dĺžke udalosti e_d,

$P(U|d, e_v)$ je pravdepodobnosť duplikácie U pri dĺžke d a vzdialenosťou e_v,

$P(U|d, e_r)$ je pravdepodobnosť, ktorá určuje, či je duplikácie reverzná alebo priama,

$P(U|t)$ - pravdepodobnosť duplikácie U, ak uplynul čas t od predchádzajúcej duplikácie alebo speciácie na tej istej vetve stromu.

Keďže sme zadefinovali, že stred udalosti vyberáme rovnomerným rozdelením, pravdepodobnosť udalosti U pri dĺžke d a strede udalosti so súradnicou e_s bude

$$P(U|d, e_s) = \frac{1}{d} \tag{2.5}$$

Dĺžku a vzdialenosť vyberáme geometrickým rozdelením, preto

$$P(U|d, e_d) = (1-p)^{e_d-1} \cdot p \ , \tag{2.6}$$

$$P(U|d, e_v) = (1-p')^{e_v-1} \cdot p' \ , \tag{2.7}$$

kde sme ako parametre p a p' použili odhady Zhang et al. (2008) [ZSV+08] z ľudských génových zhlukov (stredná dĺžka $\frac{1}{p} = 14307$ a stredná vzdialenosť $\frac{1-p'}{p'} = 306718$). Pravdepodobnosť reverznej, respektíve priamej duplikácie sme definovali v modeli ako pravdepodobnosť P_R, respektíve $(1 - P_R)$, preto

$$P(U|d, e_r) = P_R^{(1-e_r)} \cdot (1-P_R)^{e_r} \ ; \quad e_r = \begin{cases} 1 & , \text{ ak } U \text{ je priama duplikácia} \ , \\ 0 & , \text{ ak } U \text{ je reverzná duplikácia} \ . \end{cases} \tag{2.8}$$

Nakoniec pravdepodobnosť duplikácie U, ak uplynul čas t od predchádzajúcej duplikácie respektíve speciácie na tej istej vetve stromu je riadená exponenciálnym rozdelením, preto

$$P(U|t) = \lambda \cdot e^{-\lambda_e \cdot t} \tag{2.9}$$

Parametre λ_e ($\lambda_e = 0,25$), P_D ($P_D = 0,95$) a $P_R = 0,25$ ($P_R = 0,39$) sme použili rovnaké ako Vinař et al. [VBSS09]. Pravdepodobnosť, že od poslednej udalosti U už po koniec vetvy nenastala žiadna ďalšia duplikácia je integrál z exponenciálneho rozdelenia, teda

$$P_K(t) = e^{-\lambda_e \cdot t} \tag{2.10}$$

Ak medzi dvomi duplikáciami nastane úsek delécií, určujeme pravdepodobnosť ich počtu z poissonového rozdelenia. Množinu všetkých úsekov delécií na jednej vetve v budeme označovať $\vartheta(v)$. Pravdepodobnosť, že i-ty úsek s časom t má práve k delécií je

$$P(k|v_i) = \frac{e^{-\lambda_p t}(\lambda_p t)^k}{k!} \tag{2.11}$$

Parameter λ_p sme podobne použili rovnaký ako Vinař et al. [VBSS09] ($\lambda_p = 0,31$).

Každú túto čiastkovú pravdepodobnosť pre udalosť U vieme vypočítať v konštantnom čase, preto aj pravdepodobnosť $P(U|d)$ (výraz 2.4) vieme vypočítať v čase $O(1)$. Výraz 2.2 potom vieme vyhodnotiť v časovej zložitosti

$$O(k) \tag{2.12}$$

a v pamäťovej zložitosti $O(1)$, kde k je počet udalostí na vetve, keďže počet úsekov delécií ϑ je určite menší ako celkový počet udalostí.

Ostáva nám už len vypočítať pravdepodobnosť $P(X|T_x)$ pre každý mutačný strom T_x. To znamená, že pre daný strom T_x s dĺžkami hrán, znakmi v listoch a maticou pravdepodobností S zmien znakov definovanou vzťahom 2.1, chceme vypočítať pravdepodobnosť, že z modelu dostaneme práve takúto kombináciu znakov v listoch.

Na výpočet hľadanej pravdepodobnosti sme použili Felsensteinov algoritmus [Fel81]. Nech X_v je premenná reprezentujúca znak vo vrchole v stromu T_x a x_v nech je konkrétny znak vo vrchole v. Nech listy sú $X_1, \ldots, X_n$ a vnútorné vrcholy $X_{n+1}, \ldots, X_{2n-1}$. Nech dĺžka hrany z vrchola v do rodiča je t_v. Nech $P(b|a,t)$ je pravdepodobnosť zmeny znaku a na znak b za čas t podľa matice S. Nech q_a je pravdepodobnosť znaku a v koreni r stromu T_x (ekvilibrium matice S). Chceme zistiť pravdepodobnosť

$$P(X_1 = x_1, \ldots, X_n = x_n|T_x, R) = \sum_{x_{n+1},\ldots,x_{2n-1}} P(X_1 = x_1, \ldots, X_{2n-1} = x_{2n-1}|T_x, R)$$

Túto pravdepodobnosť vieme rýchlo spočítať dynamickým programovaním. Nech $A[v,a]$ je pravdepodobnosť dát v podstrome v ak $X_v = a$. Hodnoty $A[v,a]$ počítame od listov ku koreňu a v liste je $A[v,a] = [a = x_v]$. Vo vnútornom vrchole s deťmi y a z máme

$$A[v,a] = \left(\sum_{b\in\Sigma} A[y,b] \cdot P(b|a,t_y)\right)\left(\sum_{c\in\Sigma} A[z,c] \cdot P(c|a,t_z)\right)$$

Celková hľadaná pravdepodobnosť pre koreň r je potom

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T_x, R) = \sum_{a \in \Sigma} A[r, a] \cdot q_a \tag{2.13}$$

Časová zložitosť Felsensteinovho algoritmu je $O(n|\Sigma|^2)$, kde n je počet listov stromu T_x. Keďže ale máme konštantný počet znakov v abecede, časová zložitosť je nakoniec $O(n)$. Pamäťová zložitosť je taktiež $O(n)$.

Výpočtom mutačných pravdepodobností sme dokončili celý aparát pre výpočet pravdepodobnosti $P(H, X)$ danej výrazom 2.3. Ukážeme ešte algoritmickú zložitosť tohoto výpočtu a na koniec načrtneme niektoré obmedzenia nášho modelu.

Nech N je počet všetkých listov všetkých stromov z množiny $\sigma(H)$, čo je presne rovné súčtu dĺžok všetkých sučasných sekvencií z množiny X. Nech u je počet všetkých udalostí v historii H. Potom celková časová zložitosť pre vyrátanie vierohodnosti histórie H a množiny súčasných sekvencií X definovaná vzťahom 2.3 je

$$O(N \cdot u) \ , \tag{2.14}$$

pretože sme ukázali, že mutačné pravdepodobnosti pre stromy T_x vieme vypočitať Felsensteinovým algoritmom v čase $O(N)$ a keďže súčet udalostí na každej vetve je rovný počtu všetkých udalostí u v histórii H, potom zo vzťahu 2.12 pre čas na jednej vetve priamo vychádza časová zložitosť pre výpočet pravdepodobnosti celej histórie. Analogicky pamäťová zložitosť je

$$O(N) \tag{2.15}$$

2.4 Obmedzenia modelu

V predstavenom modeli sme ukázali spôsob, ako efektívne počítať vierohodnosť histórie nad týmto modelom. Táto efektívnosť nás ale stála niekoľko obmedzení. Ako sme už spomínali, do nášho modelu sme nezahrnuli operácie, ktoré pracujú na celých genómoch, preto náš model aplikujeme len na zhluky génov. Na rozdiel od modelu, ktorý používal Vinař et al.[VBSS09], používame na modelovanie mutácií jednoduchý Jukes-Cantorov model (namiesto HKY a iných modelov), ktorý napríklad nezachytáva skutočnosť, že v niektorých častiach genómu prebiehajú mutácie pomalšie (zachovalejšie úseky, exóny, ...) ako v iných častiach.

Napriek týmto a viacerým iným obmedzeniam to vyzerá tak, že náš model dáva uspokojivé výsledky a kvôli svojej efektivite skrýva veľa možností na budúce rozširovanie.

Kapitola 3

MCMC vzorkovanie

V takýchto zložitých pravdepodobnostných modeloch je presná inferencia pomerne náročná. Štandardným postupom v týchto prípadoch je použiť tzv. vzorkovanie. Ak H je história a X sekvencia, ktorú analyzujeme, snažíme sa o stochastický algoritmus, ktorý vytvorí požadovaný počet vzoriek histórií $H_1, H_2, \ldots, H_k$ takých, že distribúcia pravdepodobností týchto histórií sa blíži k rozdeleniu $P(H|X)$. Z takýchto vzoriek môžeme buď preskúmať tie, ktoré majú najväčšiu vierohodnosť, alebo počítať odhady očakávaných hodnôt rôznych zaujímavých vlastností (napr. počet duplikácií v evolučnej histórii).

V tejto kapitole ukážeme konkrétnu realizáciu algoritmu pre náš problém rekonštrukcie duplikačných histórií, pričom používame tzv. Metropolis-Hastingsov MCMC algoritmus (Hastings (1970) [Has70], Gilks (1996) [GRS96]).

3.1 MCMC algoritmus

Na vstupe máme množinu súčasných sekvencií X a výsledkom algoritmu by mala byť séria vzoriek histórií z pravdepodobnostného rozdelenia $P(H|X)$. Rozhodli sme sa použiť MCMC algoritmus, ktorý definuje Markovov reťazec náhodných premenných $H_0, H_1, \ldots$ z neznámeho ale statického rozdelenia (angl. stationary distribution), ktoré je naším hľadaným rozdelením. Premenné v Markovovom reťazci sú vyberané navrhovacím rozdelením (angl. proposal distribution), kvôli náročnosti vzorkovania premenných priamo z hľadaného rozdelenia. Platí

$$P(H_t \mid H_0, \ldots, H_{t-1}) = P(H_t|H_{t-1}) \ ,$$

čo znamená, že hodnota náhodnej premennej H_t závisí len od premennej H_{t-1} a nie od ďalších predchádzajúcich premenných.

Algoritmus prebieha nasledovne. Začína s nejakou štartovacou vzorkou H_0. V každej iterácii i vyrobí navrhovací algoritmus kandidátsku vzorku H' podľa navrhovacieho pravdepodobnostného rozdelenia podmieneného vzorkou H_{i-1}. Kandidátsku vzorku prijmeme do Markovovho reťazca s pravdepodobnosťou $\alpha(H_{i-1}, H')$. Ak prijmeme kandidátsku vzorku, potom $H_i = H'$, inak $H_i = H_{i-1}$. Pravdepodobnosť α sa vyráta ako

$$\alpha(H, H') = \min\left(1, \frac{\pi(H') \ \cdot \ q(H|H')}{\pi(H) \ \cdot \ q(H'|H)}\right) \tag{3.1}$$

kde π je vierohodnosť vzorky a $q(H'|H)$ je pravdepodobnosť navrhnutia vzorky H', ak predchádzajúca vzorka bola H a analogicky, $q(H|H')$ je pravdepodobnosť navrhnutia vzorky H, ak predchádzajúca vzorka bola H'. Akceptačná pravdepodobnosť α zabezpečuje, že statické rozdelenie náhodných premenných v Markovovom reťazci bude konvergovať ku hľadanému rozdeleniu (Hastings [Has70]).

Tento algoritmus môžeme priamo aplikovať na duplikačné histórie. Vierohodnosť π histórie H respektíve histórie H' môžeme vypočítať vzťahmi

$$\pi(H) = P(H|X) = \tfrac{P(H,X)}{P(X)} \ ,$$
$$\pi(H') = P(H'|X) = \tfrac{P(H',X)}{P(X)} \ .$$

Po dosadení do vzorca 3.1 dostávame

$$\alpha(H, H') = \min\left(1, \frac{\frac{P(H',X)}{P(X)} \ \cdot \ q(H|H')}{\frac{P(H,X)}{P(X)} \ \cdot \ q(H'|H)}\right) = \min\left(1, \frac{P(H', X) \ \cdot \ q(H|H')}{P(H, X) \ \cdot \ q(H'|H)}\right) \tag{3.2}$$

Pravdepodobnosti $P(H, X)$ respektíve $P(H', X)$ vieme vypočítať vzťahom 2.3 uvedenom v predchádzajúcej kapitole. Ostáva nám už len vypočítať pravdepodobnosti $q(H'|H)$ a $q(H|H')$, ktoré vychádzajú z navrhovacieho rozdelenia. Pravdepodobnosť $q(H'|H)$ vypočítame priamo navrhovacím algoritmom ako vedľajší produkt navrhnutia duplikačnej histórie H' a pravdepodobnosť $q(H|H')$ vieme vypočítať miernou modifikáciou navrhovacieho algoritmu.

Obrázok 3.1 predstavuje celkovú schému MCMC algoritmu aplikovaného na duplikačné histórie. Blok označený ako A reprezentuje navrhovací algoritmus, ktorý na základe predchádzajúcej histórie H_{i-1} navrhne novú históriu H'. Navrhovací algoritmus popíšeme podrobnejšie v podkapitole 3.4. Podobnou schémou sa už zaoberali

Vinař et al. (2009) [VBSS09]. Hlavným príspevkom tejto práce je časť označená *B* - efektívny algoritmus na vzorkovanie jednej duplikácie v kontexte navrhovacieho algoritmu. Tento popíšeme v kapitole 4.

3.2 Atomizácia

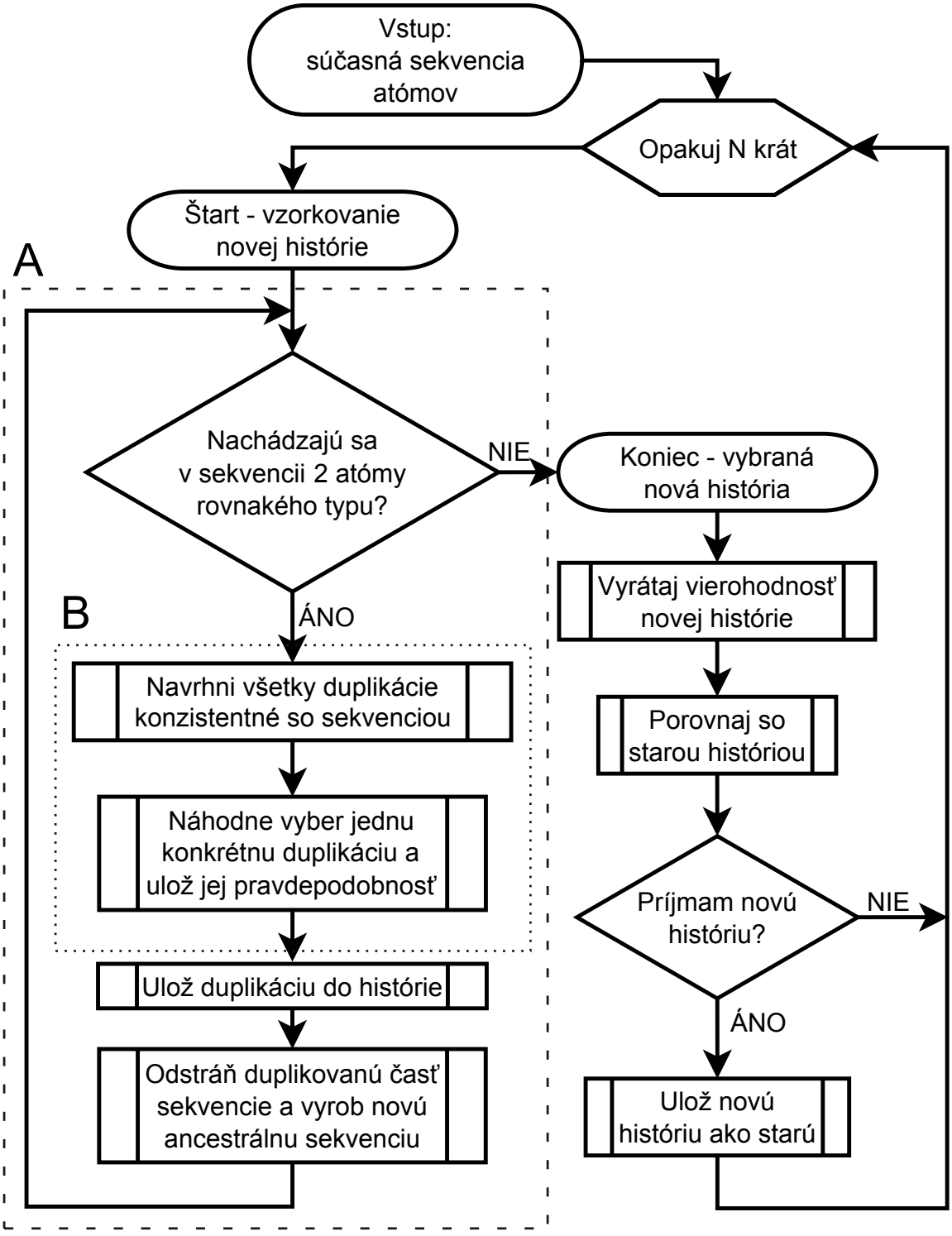
Duplikácie, speciácie a delécie sú udalosti, ktoré sa v histórii vyskytujú pomerne zriedkavo a sú to udalosti väčšieho rozsahu. Ak by sme hľadali duplikačné udalosti len v sekvenciách znakov, bola by táto úloha obtiažna.

Pre ľahšie vyhľadávanie udalostí sme preto použili koncept **atómov**, ktorý ukázali Burger [Bur10] a Brejová et al. [BBV11]. Vychádzajú z toho, že ak sú dve dlhšie sekvencie veľmi podobné (líšia sa iba v niekoľkých znakoch), museli pravdepodobne v minulosti vzniknúť zo spoločnej časti, takže jedna zo sekvencií sa duplikačnou udalosťou skopírovala na miesto, kde sa momentálne nachádza druhá sekvencia. Každý takýto súvislý úsek sekvencie, ktorý už nezahŕňal žiadnu kópiu iného úseku, nazvali atóm.

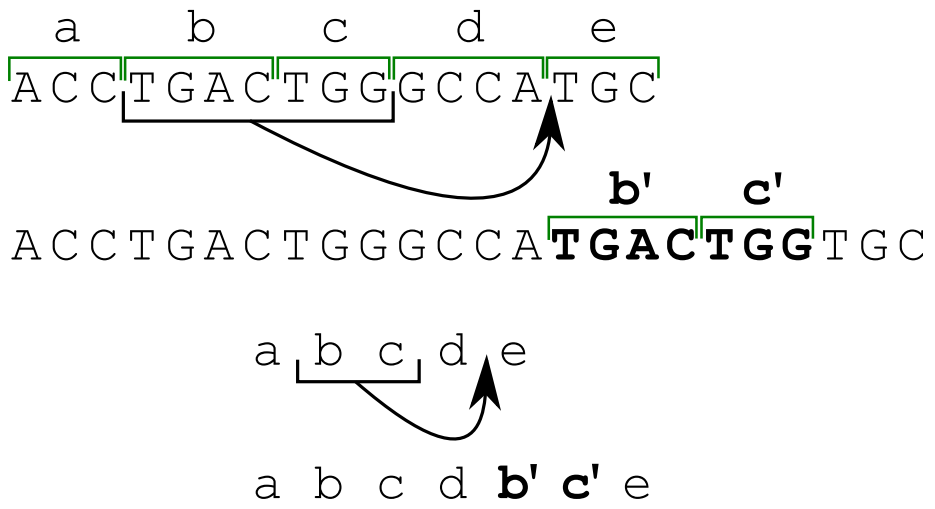
Atómom teda rozumieme dlhšiu súvislú časť sekvencie o ktorej predpokladáme, že v minulosti v rámci nej nenastala žiadna duplikačná udalosť. To znamená, že žiadna časť sekvencie jedného atómu sa v minulosti nemohla kopírovať na iné miesto, ani nemohla byť zmazaná. Mohlo sa tak stať iba s celou sekvenciou atómu.

Atómom, ktorých sekvencie boli veľmi podobné pridelíme rovnaký **typ**. Z toho teda vyplýva, že dva rôzne atómy rovnakého typu museli v minulosti vzniknúť zo spoločného predka duplikáciou alebo speciáciou. Taktiež platí, že spoločný ancestrálny predok všetkých organizmov na vstupe musel mať vo svojej sekvencii z každého typu práve jeden atóm. Tieto vlastnosti sú kľúčové pri hľadaní duplikačných udalostí. Príklad duplikácie z obrázku 1.1 po prerobení sekvencií na atómy vidieť na obrázku 3.2.

Rozloženie sekvencie na atómy nám umožňuje abstrahovať od množstva problémov súvisiacich s vyhľadávaním podobnných sekvencií a tým vytvoriť efektívny algoritmus na vzorkovanie duplikačných histórií.



Obr. 3.1: Celková schéma MCMC algoritmu aplikovaná na duplikačné histórie. Blok označený ako A reprezentuje navrhovací algoritmus. Časť označená ako B reprezentuje rýchly algoritmus na vzorkovanie jednej duplikácie.



Obr. 3.2: Príklad duplikácie sekvencie atómov b c. Na obrázku vidíme len príklad, ako by mohla byť DNA sekvencia rozdelená na atómy. Rovnaké písmená označujú rovnaký typ atómu. V skutočnosti majú atómy obvykle dlhšiu sekvenciu, väčšinou viac ako 500 báz.

3.3 Navrhovacie rozdelenie

Navrhovacie rozdelenie je pravdepodobnostné rozdelenie, ktorým navrhujeme novú históriu a ktoré nejakým spôsobom závisí od predchádzajúcej histórie. História H' sa skladá z jednotlivých udalostí $u'_1, \ldots, u'_n$. Každá udalosť má priradenú pravdepodobnosť $p'_1, \ldots, p'_n$, ktorá závisí napríklad od jej dĺžky a iných atribútov, ale taktiež od toho, či sa udalosť vyskytla aj v predchádzajúcej histórii H. Takto zaistíme, aby sme navrhovali podobné histórie k tým predchádzajúcim. MCMC algoritmus zabezpečí, že Markovov reťazec bude viac zotrvávať v oblastiach, kde má história väčšiu vierohodnosť ako v tých, kde ju má menšiu. Pravdepodobnosť histórie v navrhovacom rozdelení je jednoducho súčin pravdepodobností jednotlivých duplikačných udalostí

$$q(H'|H) = \prod_{i=1}^{n} p'_i \ , \tag{3.3}$$

pričom pravdepodobnosti jednotlivých udalostí popíšeme nižšie.

3.4 Navrhovací algoritmus

Predstavíme navrhovací algoritmus pre výber vzoriek z navrhovacieho rozdelenia. V každom kroku navrhne jednu duplikáciu alebo speciáciu, za ktorými môže nasle-

dovať niekoľko delécií. Tieto udalosti navrhuje odzadu (od najnovších po najstaršie udalosti v histórii). Delécie nezanechávajú pozorovateľné stopy v sekvenciách žijúcich druhov a preto je nemožné ich datovať presne. Namiesto toho v navrhovacom algoritme priradíme delécie k speciácii alebo duplikácii, ktorá sa udiala pred nimi. Uvažujeme iba delécie, ktoré sú kompletne vnútri zdrojovej alebo cieľovej časti sekvencie predchádzajúcej duplikácie alebo speciácie. Speciáciu reprezentujeme ako kopírovanie všetkých atómov jedného druhu do prázdnej sekvencie atómov iného druhu s možnosťou, že okamžite nasleduje niekoľko delécií v oboch druhoch.

Navrhovací algoritmus má na vstupe vyvážený fylogenetický strom T (všetky listy majú rovnakú vzdialenosť od koreňa), súčasné atomizované sekvencie $X_1, \ldots, X_k$ patriace k druhom nachádzajúcich sa v listoch $\ell_1, \ldots, \ell_k$ stromu T a predchádzajúcu duplikačnú históriu H. V každom okamihu pre každý list ℓ_i aktuálneho stromu T existuje d_i možných duplikácií $D_i^{(1)}, \ldots, D_i^{(d_i)}$ konzistentných so sekvenciou X_i. Algoritmus každú z nich ohodnotí pomocou skórovacej funkcie f. Ak sa niektorá z navrhnutých duplikácií vyskytla aj v duplikačnej histórii H, potom dostane v skórovacej funkcii bonus. Taktiež pre najbližšie dva listy ℓ_i a ℓ_j v aktuálnom strome T existuje s možných speciácií $S^{(1)}, \ldots, S^{(s)}$ konzistentných so sekvenciami X_i a X_j, ktorým tiež priradíme skóre skórovacou funkciou f. V prvom rade potrebujeme vypočítať sumu ς všetkých ohodnotených duplikácií a speciácií (udalostí):

$$\varsigma = \left(\sum_{i=1}^{k} \sum_{j=1}^{d_i} f(D_i^{(j)}) \right) + \sum_{i=1}^{s} f(S^{(i)})$$

Stochasticky si zo všetkých možných udalostí vyberieme jednu podľa pomeru jej skóre k sume všetkých udalostí ς. Udalosť u teda vyberieme s pravdepodobnosťou $p' = \frac{f(u)}{\varsigma}$. Ak bola udalosť u duplikáciou na vetve ℓ_i, upravíme sekvenciu X_i tak, ako by vyzerala pred aplikovaním duplikácie u. Ak bola udalosť u speciáciou medzi druhmi v listoch ℓ_i a ℓ_j, vytvoríme ancestrálnu sekvenciu X' sekvencií X_i a X_j, všetky vetvy stromu T prislúchajúce listom $\ell_1, \ldots, \ell_k$ skrátime o dĺžku vetvy prislúchajúcej k listu ℓ_i respektíve ℓ_j, zrušíme tým listy ℓ_i a ℓ_j a najbližší spoločný vrchol listov ℓ_i a ℓ_j sa stane novým listom ℓ', ktorému bude prislúchať sekvencia X'. Prečíslujeme listy a k nim príslúchajúce sekvencie od 1 po $(k-1)$ (zrušili sme dva listy a pridali jeden) a tým vytvorí upravený aktuálny strom T. Iterácie opakujeme, kým vieme navrhnúť aspoň jednu duplikáciu prípadne speciáciu. Efektívny algoritmus, ako vybrať jednu udalosť bez toho, aby sme explicitne vymenovali všetky možné udalosti, ukážeme

v nasledujúcej kapitole 4.

Miernou modifikáciou navrhovacieho algoritmu vieme vypočítať aj pravdepodobnosť $q(H|H')$. Keďže je história H predchádzajúcou históriou, musela byť niekedy predtým vytvorená navrhovacím algoritmom. Nech teda vznikla udalosťami $u_1, \ldots, u_n$ presne v tomto poradí. Modifikovaný algoritmus podobne ako predchádzajúci navrhovací algoritmus tiež v každej iterácii i ohodnotí všetky možné duplikácie a speciácie, ale v skórovacej funkcii zoberie bonus pre udalosti z duplikačnej histórie H'. V bode, kedy si má vybrať jednu konkrétnu udalosť nevyberie stochasticky, ale vyberie udalosť u_i a pravdepodobnosť $p_i = \frac{f(u_i)}{\varsigma}$, kde ς je rovnako súčet všetkých ohodnotených duplikácií a speciácií v iterácii i. Potom

$$q(H|H') = \prod_{i=1}^{n} p_i \tag{3.4}$$

V každej iterácii navrhovacieho algoritmu potrebujeme ohodnotiť všetky konzistentné udalosti a stochasticky vybrať jednu z nich bez nutnosti vymenovania všetkých udalostí. V nasledujúcej kapitole ukážeme náš najväčší prínos v tejto práci, efektívny prístup k riešeniu tohoto problému.

Kapitola 4

Vzorkovanie duplikačnej udalosti

V navrhovacom algoritme, ktorý sme popísali v predchádzajúcej kapitole, je potrebné vytvoriť stochastický algoritmus na navrhnutie konkrétnej duplikačnej udalosti potrebnej na odstránenie duplikovanej časti a získanie ancestrálnej sekvencie, ako aj vypočítať pravdepodobnosť, že danú udalosť algoritmus navrhne.

Máme danú postupnosť atomických segmentov *seq*. Každej možnej duplikácii v tejto postupnosti priradíme skóre, ktoré závisí od jej dĺžky a vyjadruje „vhodnosť“ tejto duplikácie v kontexte danej sekvencie. Zo všetkých možných duplikácií chceme potom vybrať jednu, pričom ak súčet skóre všetkých duplikácií je S a skóre konkrétnej duplikácie x je S_x, tak pravdepodobnosť vybrania duplikácie x bude $P(x|seq) = \frac{S_x}{S}$.

Napríklad pre sekvenciu $\langle\ a\ b\ c\ a'\ b'\ c'\ \rangle$ a skórovaciu funkciu $f(x) = 2^{\ell-1}$, kde ℓ je dĺžka duplikácie x (počet duplikovaných atómov), môžeme vidieť zoznam všetkých jednoduchých duplikácií aj s ich skóre v nasledujúcej tabuľke:

duplikácia	skóre	ancestrálna sekvencia po odvinutí duplikácie
$\langle\ a\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ a'\ \rangle$	1	$\langle\ a\ b\ c\ b'\ c'\ \rangle$
$\langle\ a'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ a\ \rangle$	1	$\langle\ b\ c\ a'\ b'\ c'\ \rangle$
$\langle\ b\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ b'\ \rangle$	1	$\langle\ a\ b\ c\ a'\ c'\ \rangle$
$\langle\ b'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ b\ \rangle$	1	$\langle\ a\ c\ a'\ b'\ c'\ \rangle$
$\langle\ c\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ c'\ \rangle$	1	$\langle\ a\ b\ c\ a'\ b'\ \rangle$
$\langle\ c'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ c\ \rangle$	1	$\langle\ a\ b\ a'\ b'\ c'\ \rangle$
$\langle\ a\ b\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ a'\ b'\ \rangle$	2	$\langle\ a\ b\ c\ c'\ \rangle$
$\langle\ a'\ b'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ a\ b\ \rangle$	2	$\langle\ c\ a'\ b'\ c'\ \rangle$
$\langle\ b\ c\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ b'\ c'\ \rangle$	2	$\langle\ a\ b\ c\ a'\ \rangle$
$\langle\ b'\ c'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ b\ c\ \rangle$	2	$\langle\ a\ a'\ b'\ c'\ \rangle$
$\langle\ a\ b\ c\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ a'\ b'\ c'\ \rangle$	4	$\langle\ a\ b\ c\ \rangle$
$\langle\ a'\ b'\ c'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ a\ b\ c\ \rangle$	4	$\langle\ a'\ b'\ c'\ \rangle$

Súčet skóre všetkých duplikácií $S = 22$. Takže napríklad pravdepodobnosť duplikácie $\langle\ b'\ c'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ b\ c\ \rangle$ je

$$P(\langle\ b'\ c'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ b\ c\ \rangle \mid \langle\ a\ b\ c\ a'\ b'\ c'\ \rangle) = \frac{2}{22}$$

Vinař et al.(2009)[VBSS09] urobili váhovanú skórovaciu schému, ktorá závisela od dĺžky vymazanej časti sekvencie po odvinutí udalosti, predchádzajúcich videných udalostí (udalosť, ktorá bola použitá aj v predchádzajúcej histórií dostala bonus), vzdialeností atómov v ich stromoch (čerešničky stromu boli uprednostňované), podčastí duplikácií (ak bola duplikácia podmnožinou väčšej duplikácie, dostala penalizáciu), penalizovaných delécií a viacerých iných faktorov. Schéma bola príliš zložitá na to, aby sa dal urobiť efektívny algoritmus na vzorkovanie duplikácií. Preto ich prístup bol vymenovať všetky možné duplikácie explicitne (v ich skórovacej schéme to bolo cca $O(n^3)$ duplikácií bez medzier, alebo s jednou medzerou $O(n^4)$), vypočítať pre každú z nich jej skóre (v čase $O(n)$), a potom vybrať jednu z týchto vymenovaných duplikácií náhodne v pomere ku skóre. Výsledná časová zložitosť $O(n^5)$ je značne nepraktická.

Príspevok tejto práce je vytvorenie nových skórovacích schém, ktoré sú uspôsobené na to, aby sa dal vytvoriť efektívny algoritmus a príspevkom je tiež samotný efektívny algoritmus. V ďalších podkapitolách predstavíme niekoľko skóro-

vacích schém (od najjednoduchšej až po pomerne zložitú, biologicky motivovanú) a ukážeme vzorkovací algoritmus, ktorý dokáže vybrať duplikáciu v čase $O(n^2)$. Algoritmus je veľmi podobný na techniky používané pri lokálnych a globálnych zarovnaniach DNA sekvencií (Smith-Waterman [SW81] a Needleman-Wunsch [NW70]), v kombinácii so vzorkovacím algoritmom na skrytých Markovových modeloch pomocou stochastického spätného prechodu (Boomsma [BMT+08] a Cawley [CP03]).

4.1 Základný model

Na vstupe máme súčasnú sekvenciu atómov *seq*. i-ty atóm sekvencie *seq* budeme označovať seq_i. Typ i-teho atómu budeme označovať $seqtype_i$. Vidíme teda súčasnú sekvenciu a chceme vedieť, ako vyzerala ancestrálna sekvencia pred duplikáciou. Povolené budú len jednoduché priame duplikácie (žiadne delécie ani reverzné duplikácie). Skórovacia funkcia duplikácie x s dĺžkou ℓ bude

$$f(x) = 2^{\ell-1}$$

Takúto úlohu vieme reprezentovať pomocou grafu, ktorý má jeden začiatočný vrchol B, jeden koncový vrchol E a maticu vrcholov S. Vrchol $S_{i,j}$ predstavuje všetky duplikácie, v ktorých posledným krokom duplikácie je kopírovanie atómu seq_i na atóm seq_j. Hodnota vrchola $S_{i,j}$ je potom súčet týchto duplikácií. Z tohto jasne vyplýva, že vrcholy s hodnotou rôznou od nuly budú len tie, ktorých indexy i a j ukazujú na dva rozdielne atómy s rovnakým typom a budeme ich nazývať aktívnymi.

$$\forall\ S_{i,j}\ \in\ S:\ S_{i,j} \text{ je aktívny} \Leftrightarrow (seq_i \neq seq_j)\ \wedge\ (seqtype_i = seqtype_j) \tag{4.1}$$

Hrany medzi vrcholmi v matici S prechádzajú len šikmo medzi aktívnymi vrcholmi. Napríklad z vrcholu $S_{i,j}$ môže ísť hrana do vrchola $S_{i+1,j+1}$, ale len v prípade, že sú vrcholy $S_{i,j}$ a $S_{i+1,j+1}$ aktívne. Nech majú hrany medzi vrcholmi v matici S hodnotu z. Z vrchola B vychádzajú hrany len do aktívnych vrcholov v matici S a majú hodnotu b. Podobne do vrchola E vchádzajú hrany len z aktívnych vrcholov a majú hodnotu e. Hodnoty vrcholov v matici S vieme potom počítať nasledujúcim spôsobom:

$$S_{i,j} = \begin{cases} S_{i-1,j-1} \cdot z\ + B \cdot b & ,\ (S_{i,j} \text{ je aktívny})\ \wedge\ (S_{i-1,j-1} \text{ je aktívny}) \\ B \cdot b & ,\ (S_{i,j} \text{ je aktívny})\ \wedge\ \neg(S_{i-1,j-1} \text{ je aktívny}) \\ 0 & ,\ \text{inak} \end{cases} \tag{4.2}$$

Hodnota koncového vrchola je potom:

$$E = \sum_{v \in S \ \wedge \ v \text{ je aktívny}} v \cdot e \tag{4.3}$$

Príklad tohoto základného modelu pre sekvenciu $seq = \langle\ a\ b\ c\ d\ e\ b'\ c'\ d'\ a'\ b''\ e'\ \rangle$ vidieť na obrázku 4.1. Pre našu skórovaciu funkciu f majú parametre modelu hodnoty $z = 2$, $b = 1$, $e = 1$, $B = 1$.

Ukážeme, že každá cesta zo začiatočného vrchola B do koncového vrchola E predstavuje jednu duplikáciu. Nech je to cesta u. Keďže medzi vrcholom B a E neexistuje priama hrana, ak existuje cesta, musí v nej byť aspoň jeden vrchol z matice S. Nás budú zaujímať práve tieto vrcholy S v ceste u pre ktoré platí, že k nim prislúchajúce atómy tvoria zdrojovú a cieľovú časť duplikácie.

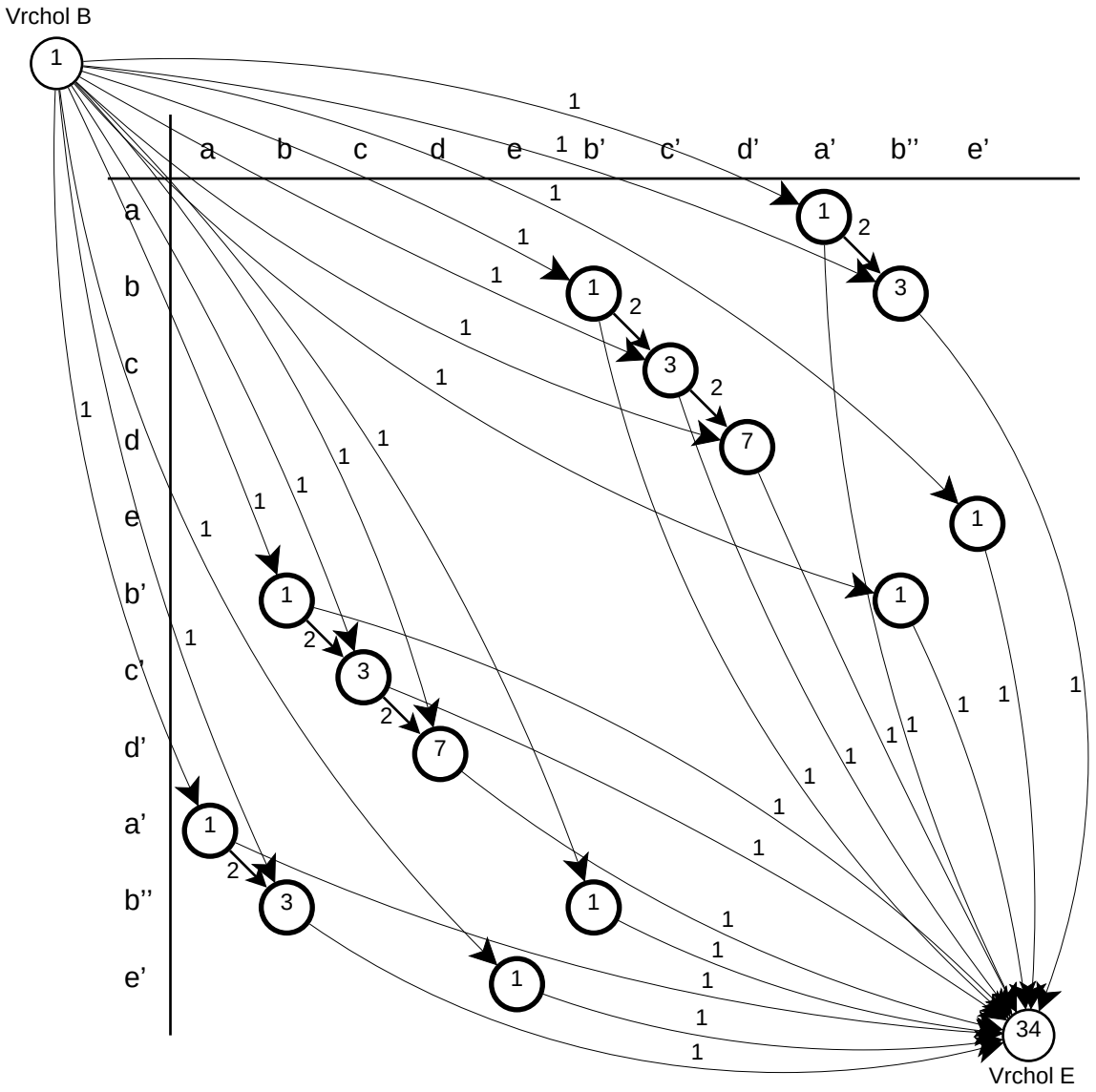
$$\begin{aligned} \forall S_{i,j} \in S : S_{i,j} \in u \ \Rightarrow \quad & (seq_i \in \text{zdrojovej časti duplikácie}) \ \wedge \\ & (seq_j \in \text{kopírovanej časti duplikácie}) \end{aligned} \tag{4.4}$$

Je to preto, lebo aktívne vrcholy v matici S sú len tie, ktoré spájajú rôzne atómy s rovnakým typom, čiže tie, ktoré sa môžu kopírovať. Samotné aktívne vrcholy sú taktiež spojené medzi sebou len tak, aby mohli byť spojené atómy kopírované za sebou a teda aby vytvárali dve súvislé časti sekvencie, zdrojovú a kopírovanú časť duplikácie.

Hodnota cesty tiež spĺňa našu skórovaciu funkciu f, pretože ak chceme predĺžiť nedokončenú cestu v končiacu vo vrchole $S_{i-1,j-1}$ o vrchol $S_{i,j}$, hodnota predĺženej cesty v' sa vyráta ako $S_{i-1,j-1} \cdot z$, kde $z = 2$. Preto ak predtým bola hodnota cesty v rovná 2^ℓ, kde ℓ je dĺžka duplikácie, tak sa v našom prípade dá interpretovať tiež ako počet vrcholov z matice S v ceste v a potom po pridaní ďalšieho vrchola z matice S bude hodnota cesty v' rovná $2 \cdot 2^\ell = 2^{\ell+1}$. V ceste v' sa skutočne nachádza $\ell + 1$ vrcholov z matice S (pretože na ceste v ich bolo ℓ). V koncovom vrchole E sa potom už len nasčítajú všetky hodnoty skóre všetkých povolených duplikácií (bez delécií a reverzných duplikácií), ktoré sa dajú nájsť v sekvencii seq.

Maticu S môžeme efektívne vypočítať dynamickým programovaním v časovej aj pamäťovej zložitosti $O(n^2)$, kde n je dĺžka vstupnej sekvencie seq. Vrchol E vieme vypočítať tiež v čase $O(n^2)$, pretože sa na každý prvok matice S pozrie práve raz.

Ako sme už ukázali, každá cesta zo začiatočného vrchola B do koncového vrchola E predstavuje duplikáciu. My ale chceme nájsť jednu konkrétnu. To dosiahneme pomocou stochastického spätného prechodu. Začneme v koncovom vrchole E a pozrieme



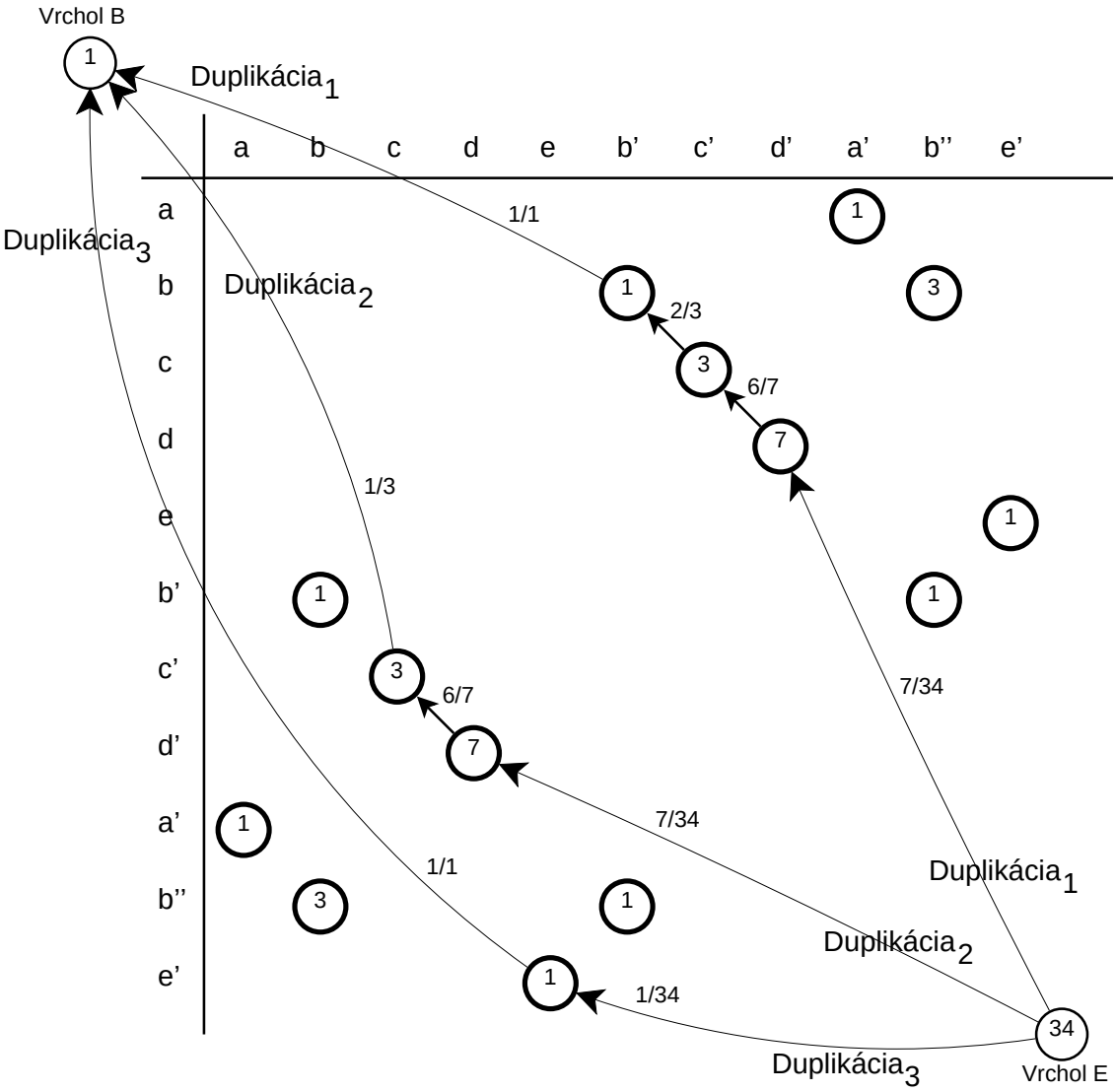
Obr. 4.1: Úlohou bolo nájsť duplikáciu v sekvencii $\langle\ a\ b\ c\ d\ e\ b'\ c'\ d'\ a'\ b''\ e'\ \rangle$. Na obrázku sú zobrazené len aktívne vrcholy matice S. Ostatné vrcholy mali hodnotu 0 a stupeň vrchola tiež 0.

sa, ktoré hrany do neho vchádzajú. Každá hrana vchádzajúca do vrchola E prispieva veľkosťou skóre danou súčtom skór všetkých duplikácií, ktoré cez ňu prechádzajú. Pomerne k týmto hodnotám a súčtu všetkých skór duplikácií obsahujúcich daný vrchol náhodne vyberieme jednu z hrán. To znamená, že sme počet všetkých možných ciest (duplikácií) obmedzili len na tie, ktoré prechádzajú vybranou hranou. Uložíme pomer, ktorým sme vybrali hranu a prejdeme do vrchola, z ktorého hrana vychádzala. Toto spätne opakujeme, až kým sa nedostaneme do začiatočného vrchola B. Súčin všetkých pomerov na tejto ceste nám dáva pomer hodnoty nájdenej duplikácie ku všetkým duplikáciám, ktoré vieme nájsť v sekvencii *seq* a súčasne predstavuje pravdepodobnosť výberu danej duplikácie naším algoritmom.

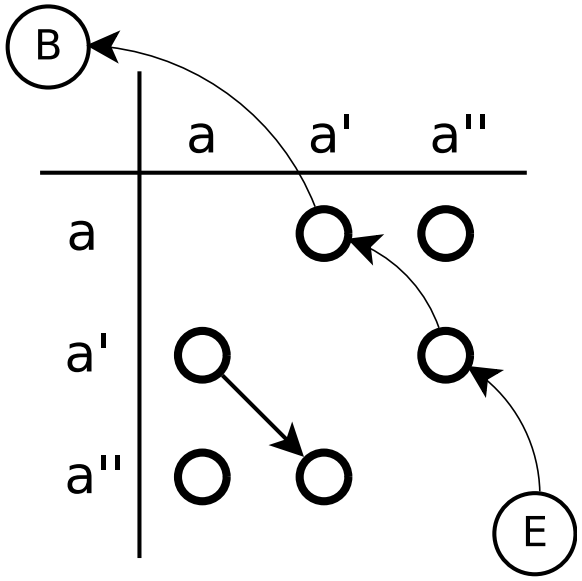
Najlepšie je to vidieť na troch rôznych duplikáciách Duplikácia$_1$, Duplikácia$_2$ a Duplikácia$_3$ zobrazených na obrázku 4.2, ktorý zobrazuje čo by sa stalo, ak by sme na predchádzajúci príklad so sekvenciou $\langle\ a\ b\ c\ d\ e\ b'\ c'\ d'\ a'\ b''\ e'\ \rangle$ na obrázku 4.1 aplikovali spätný stochastický prechod trikrát a vrátil by nám 3 rozličné cesty (duplikácie).

- Duplikácia$_1$ - $\langle\ b\ c\ d\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ b'\ c'\ d'\ \rangle$ s hodnotou cesty (duplikácie) $\frac{7\cdot1}{34}\cdot\frac{3\cdot2}{7}\cdot\frac{2\cdot1}{3}\cdot\frac{1\cdot1}{1}=\frac{2}{17}$. Ancestrálnu sekvenciu potom tvoria atómy $\langle\ a\ b\ c\ d\ e\ a'\ b''\ e'\ \rangle$.
- Duplikácia$_2$ - $\langle\ c'\ d'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ c\ d\ \rangle$ s hodnotou cesty (duplikácie) $\frac{7\cdot1}{34}\cdot\frac{3\cdot2}{7}\cdot\frac{1\cdot1}{3}=\frac{1}{17}$. Ancestrálnu sekvenciu potom tvoria atómy $\langle\ a\ b\ e\ b'\ c'\ d'\ a'\ b''\ e'\ \rangle$.
- Duplikácia$_3$ - $\langle\ e'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ e\ \rangle$ s hodnotou cesty (duplikácie) $\frac{1\cdot1}{34}\cdot\frac{1\cdot1}{1}=\frac{1}{34}$ Ancestrálnu sekvenciu potom tvoria atómy $\langle\ a\ b\ c\ d\ b'\ c'\ d'\ a'\ b''\ e'\ \rangle$.

Môže sa nám ale stať, že nájdená cesta bude ukazovať na neplatnú duplikáciu, ktorá by znamenala kopírovanie časti sekvencie vnútri duplikovanej časti. Napríklad cesta na obrázku 4.3 znázorňuje duplikáciu $\langle\ a\ a'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ a'\ a''\ \rangle$. Takéto duplikácie nechceme, preto ak detegujeme neplatnú duplikáciu, cestu zahodíme a spätným stochastickým prechodom nájdeme ďalšiu cestu. Toto opakujeme až kým nenájdeme platnú duplikáciu. Ak je v matici aspoň jeden aktívny vrchol, musí v nej existovať aspoň jedna platná duplikácia, pretože už samotná cesta z vrchola B do aktívneho vrchola a z neho priamo do vrchola E je platná duplikácia. Detegovať neplatnú duplikáciu vieme pomerne jednoducho. Označme z_{orig} - začiatok a k_{orig} - koniec pôvodnej sekvencie a z_{copy} - začiatok a k_{copy} - koniec kopírovanej sekvencie. Ak platí, že koniec pôvodnej sekvencie je pred začiatkom kopírovanej sekvencie, je duplikácia platná.



Obr. 4.2: Tri rozličné cesty nájdené spätným stochastickým prechodom.



Obr. 4.3: Problematická neplatná duplikácia.

Ak to neplatí, tak musí platiť, že koniec duplikovanej sekvencie je pred začiatkom pôvodnej sekvencie. Ak ani to neplatí, duplikácia nie je platná.

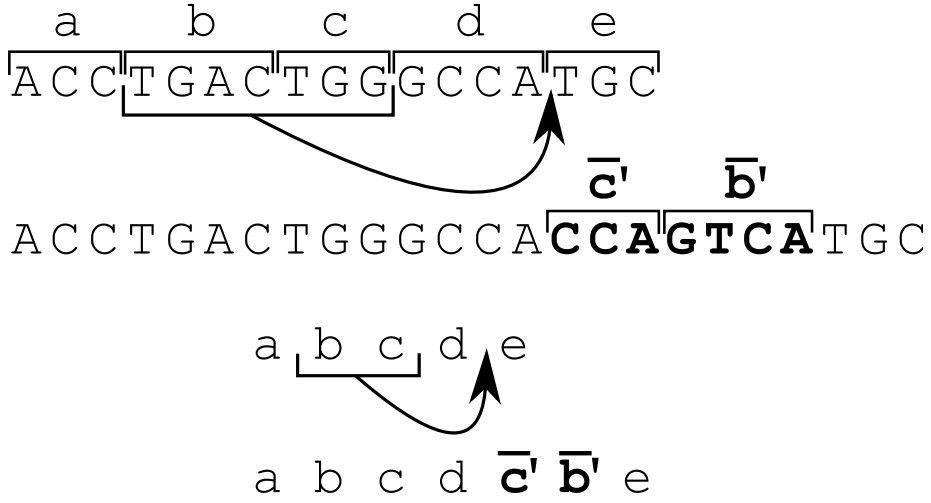
$$\begin{aligned} \text{duplikácia je platná} \ \Leftrightarrow \quad & (k_{orig} < z_{copy}) \vee \\ & (k_{orig} > z_{copy} \ \wedge \ z_{orig} > k_{copy}) \end{aligned} \tag{4.5}$$

V problematickom príklade na obrázku 4.3 môžeme vidieť, že ($z_{orig} = 1$, $k_{orig} = 2$, $z_{copy} = 2$, $k_{copy} = 3$) $\Rightarrow$ duplikácia nespĺňa podmienku pre platnosť.

4.2 Základný model s reverziou

V predchádzajúcom modeli sme ukázali spôsob, ako hľadať jednoduché duplikácie. Z biologického hľadiska sa ale dejú aj iné duplikácie, ktoré by sme radi zahrnuli do nášho modelu, reverzné duplikácie. Reverznú duplikáciu sme popísali v kapitole 1. Príklad reverznej duplikácie na atomizovanej sekvencii môžeme vidieť na obrázku 4.4, ktorý je podobný obrázku 3.2 s tým rozdielom, že teraz nastala reverzná duplikácia.

Zavedieme ďalšie označenia. Pole orientácií atómov nazveme *seqstrand*, pričom i-ty prvok $seqstrand_i$ označuje orientáciu i-teho atómu seq_i a každý prvok môže nadobúdať hodnoty 1 (kladná orientácia) alebo -1 (záporná orientácia). Dĺžku atomizovanej sekvencie *seq* označíme N. Reverznou sekvenciou atómov potom rozumieme



Obr. 4.4: Príklad reverznej duplikácie $\langle\ b\ c\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ \overline{c'}\ \overline{b'}\ \rangle$.

sekvenciu *seqrev*:

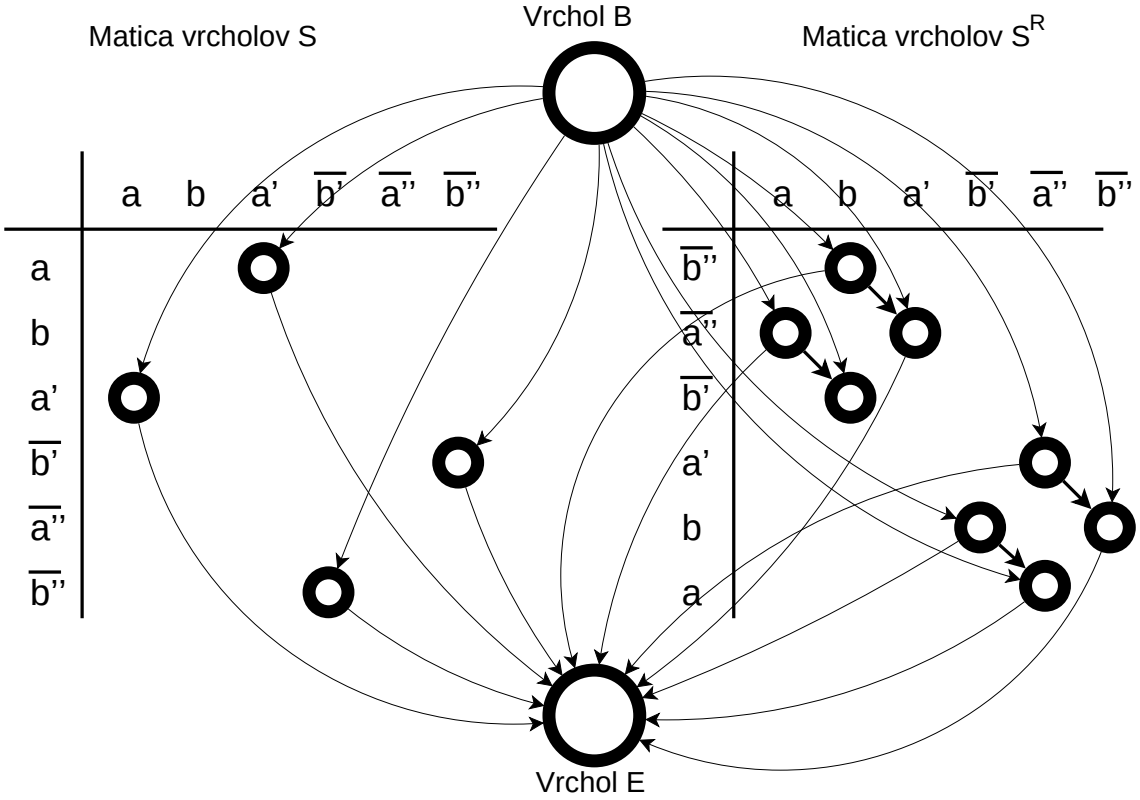
$$\forall_{i=1}^{N} \quad seqrev_i = seq_{N-i+1} \tag{4.6}$$

Skórovacia funkcia zostala rovnaká ako v predchádzajúcom modeli, $f(x) = 2^{\ell-1}$, kde ℓ je dĺžka duplikácie x.

Oproti predchádzajúcemu modelu ale naviac vytvoríme maticu S^R veľmi podobnú matici S s tým rozdielom, že predtým sme v matici S porovnávali prvky dvoch rovnakých sekvencií a v matici S^R budeme porovnávať prvky pôvodnej sekvencie *seq* s jej reverznou sekvenciou *seqrev*. Taktiež upravíme podmienku, kedy je vrchol $S_{i,j}$ respektíve vrchol $S^R_{i,j}$ aktívny. Vrchol $S_{i,j}$ bude aktívny práve vtedy, ak sú atómy seq_i a seq_j rozdielne ale majú rovnaký typ a orientáciu. Vrchol $S^R_{i,j}$ bude aktívny práve vtedy, ak sú atómy seq_i a $seqrev_j$ rozdielne ale majú rovnaký typ a opačnú orientáciu.

$$\begin{aligned} \forall\ S_{i,j}\ \in\ S:\ S_{i,j} \text{ je aktívny} \Leftrightarrow\ & (seq_i \neq seq_j) \wedge \\ & (seqtype_i = seqtype_j) \wedge \\ & (seqstrand_i = seqstrand_j) \end{aligned} \tag{4.7}$$

$$\begin{aligned} \forall\ S^R_{i,j}\ \in\ S^R:\ S^R_{i,j} \text{ je aktívny} \Leftrightarrow\ & (seq_i \neq seq_{N-j+1}) \wedge \\ & (seqtype_i = seqtype_{N-j+1}) \wedge \\ & (seqstrand_i \neq seqstrand_{N-j+1}) \end{aligned} \tag{4.8}$$



Obr. 4.5: Príklad základného modelu duplikácií s reverziou.

Upravíme ešte vrcholy B a E. Z vrchola B budu vychádzať hrany nielen do aktívnych vrcholov v matici S, ale aj do aktívnych vrcholov v matici S^R. Vzťah pre výpočet hodnôt vrcholov v matici S sa nezmenil. Veľmi podobný vzťah zavedieme aj pre vrcholy v matici S^R:

$$S^R_{i,j} = \begin{cases} S^R_{i-1,j-1} \cdot z \ + B \cdot b & , \ (S^R_{i,j} \text{ je aktívny}) \ \wedge \ (S_{i-1,j-1} \text{ je aktívny}) \\ B \cdot b & , \ (S_{i,j} \text{ je aktívny}) \ \wedge \ \neg(S_{i-1,j-1} \text{ je aktívny}) \\ 0 & , \ \text{inak} \end{cases} \tag{4.9}$$

Podobne do vrchola E budu vchádzať hrany z aktívnych vrcholov v oboch maticiach S a S^R. Pravidlo pre vyrátanie hodnoty koncového vrchola bude:

$$E = \left(\sum_{v \in S \ \wedge \ v \text{ je aktívny}} v \cdot e \right) + \left(\sum_{v \in S^R \ \wedge \ v \text{ je aktívny}} v \cdot e \right) \tag{4.10}$$

Obrázok 4.5 znázorňuje príklad modelu s reverziou pre sekvenciu $\langle \ a \ b \ a' \ \overline{b'} \ \overline{a''} \ \overline{b''} \ \rangle$.

Cesty v matici S stále zodpovedajú duplikáciám, podobne ako v predchádzajúcom modeli. V matici S^R cesty zodpovedajú reverzným duplikáciám. Maticu S^R sme totiž

vytvárali analogicky ako v predchádzajúcom modeli maticu S, preto v nej platia rovnaké pravidlá, ako pre maticu S.

Následne získame duplikáciu analogicky ako v predchádzajúcom základnom modeli bez reverzie spätným stochastickým prechodom z koncového vrchola E do začiatočného vrchola B. V prvom kroku spätného stochastického prechodu vyberieme jednu z hrán vchádzajúcich do vrchola E. Tieto hrany vychádzajú z vrchola matice S alebo matice S^R. Keďže medzi maticami S a S^R hrany neprechádzajú, výberom tejto prvej hrany sa rozhodne, či duplikácia bude alebo nebude reverznou.

4.3 Rozšírený model s deléciami

V predchádzajúcich kapitolách sme ukázali modely na nájdenie duplikácie v sekvencii atómov vrátane reverzií. Základný model bez reverzie teraz rozšírime o delécie. V sekvencii budeme hľadať duplikáciu a povolíme, aby bezprostredne po nej mohlo nastať niekoľko delécií. Skórovacia funkcia f bude o niečo zložitejšia. Nech ℓ je dĺžka duplikácie x a nech d je súčet dĺžok všetkých delécií naledujúcich hneď po duplikácií. Potom

$$f(x) = 2^{\ell-d-1} \cdot 0.5^d$$

Predstavený základný model potrebuje len niekoľko málo úprav. V matici S ostane stále rovnaká podmienka pre aktívne vrcholy a stále bude platiť, že z vrchola B vychádzajú hrany len do aktívnych vrcholov v matici S a do vrchola E vchádzajú hrany len z aktívnych vrcholov. Z aktívnych vrcholov môžu vychádzať ešte šikmé hrany do susedných vrcholov s hodnotou $z = 2$. Naviac ale povolíme horizontálne aj vertikálne hrany medzi susednými vrcholmi a každá hrana bude mať hodnotu $g = 0.5$. Tieto hrany nám zabezpečia to, že ak zoberieme nejakú cestu z vrchola B do vrchola E, ktorá predstavuje nejakú duplikáciu a v tejto ceste sa nachádza aj takáto hrana, musela duplikácia po aplikovaní na ancestrálnu sekvenciu skopírovať aj atómy, ktoré sme v súčasnej sekvencii *seq* nemali. Takéto dočasné atómy je potrebné vymazať za pomoci delécie. Mohli by sme každý dočasný atóm vymazávať postupne. Keďže sa ale snažíme minimalizovať počet udalostí, budeme vymazávať celú súvislú časť horizontálnej, respektíve vertikálnej časti cesty. Podmienky na aktívnosť vrchola zostali zachované, taktiež výpočet hodnoty vrchola E zostal rovnaký.

Treba len zmeniť výpočet pre vrcholy v matici S:

$$S_{i,j} = \begin{cases} S_{i-1,j-1} \cdot z \ + S_{i-1,j} \cdot g + S_{i,j-1} \cdot g + B \cdot b & , \ (S_{i-1,j-1} \text{ je aktívny}) \wedge \\ & \quad (S_{i,j} \text{ je aktívny}) \\ S_{i-1,j-1} \cdot z \ + S_{i-1,j} \cdot g + S_{i,j-1} \cdot g & , \ (S_{i-1,j-1} \text{ je aktívny}) \wedge \\ & \quad \neg(S_{i,j} \text{ je aktívny}) \\ S_{i-1,j} \cdot g + S_{i,j-1} \cdot g & , \ \text{inak} \end{cases} \tag{4.11}$$

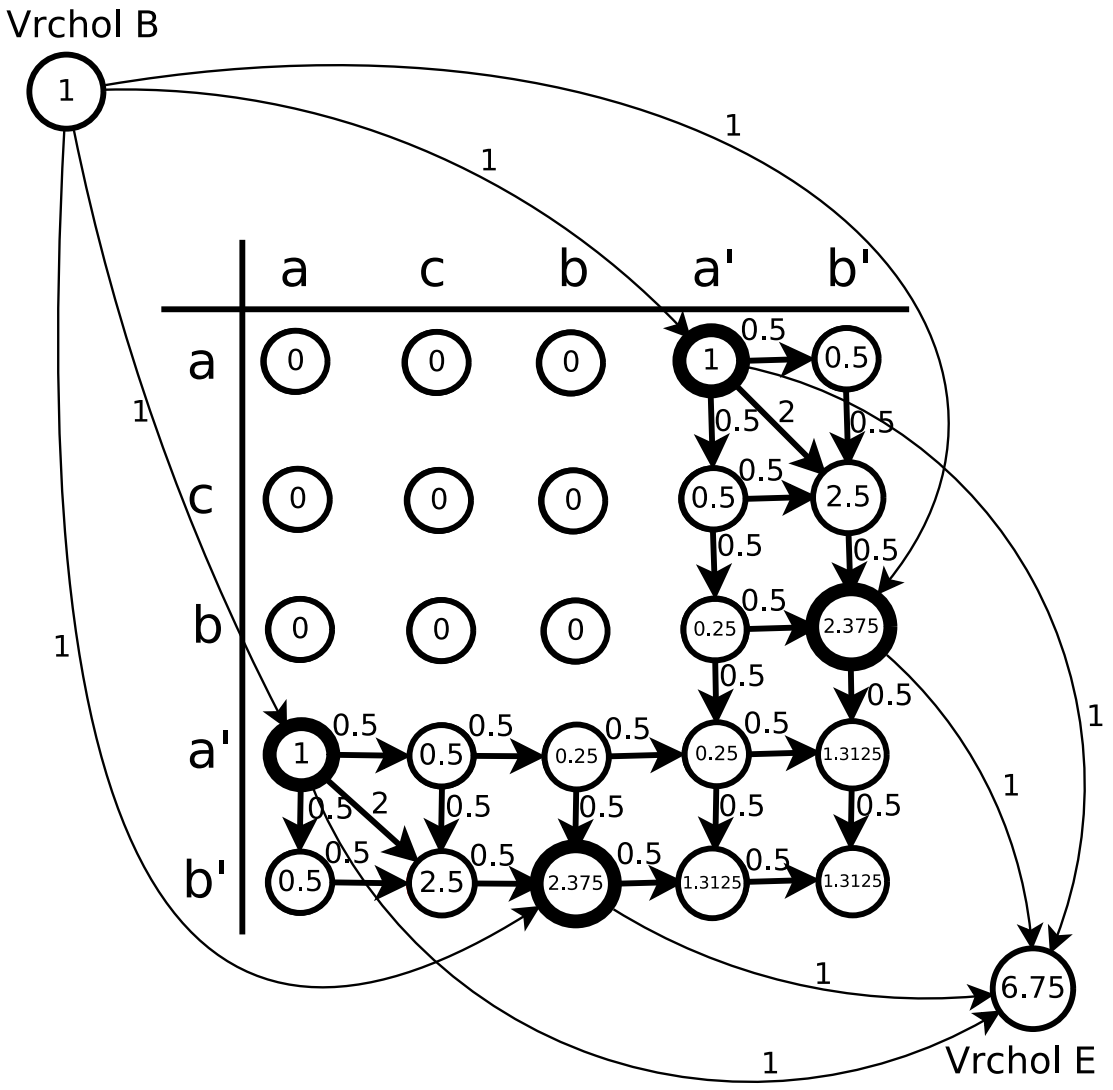
Pred vypĺňaním má každý vrchol matice S hodnotu 0. Po vyplnení zrušíme vrcholom, ktoré majú nulovú hodnotu, všetky hrany. Na obrázku 4.6 vidieť príklad rozšíreného modelu duplikácií s deléciami pre sekvenciu $seq = \langle\ a\ c\ b\ a'\ b'\ \rangle$.

Ukážeme, že skóre duplikácie skutočne zodpovedá hodnota cesty prislúchajúcej danej duplikácii v takejto upravenej matici vrcholov S. Je to preto, že každá zvislá alebo vodorovná hrana v matici prislúcha jednému vymazávanému atómu, preto počet horizontálnych a vertikálnych hrán je súčet dĺžok d všetkých delécií. Tieto hrany majú hodnotu 0.5, preto delécie prispievajú do skóre cesty hodnotou 0.5^d. Každá šikmá hrana zase zmenší počet vnútorných vrcholov v ceste o jeden, keďže sa namiesto jednej horizontálnej a vertikálnej hrany použije šikmá hrana. Počet použitých šikmých hrán k prispieva ku skóre cesty hodnotou 2^k. Počet vnútorných hrán h v ceste u (čiže bez hrany z vrchola B a do vrchola E) zodpovedá priamo počtu vrcholov w matice S v ceste u a to tak, že $w = h + 1$. Dĺžka duplikácie ℓ je rovná počtu vrchlov w, preto

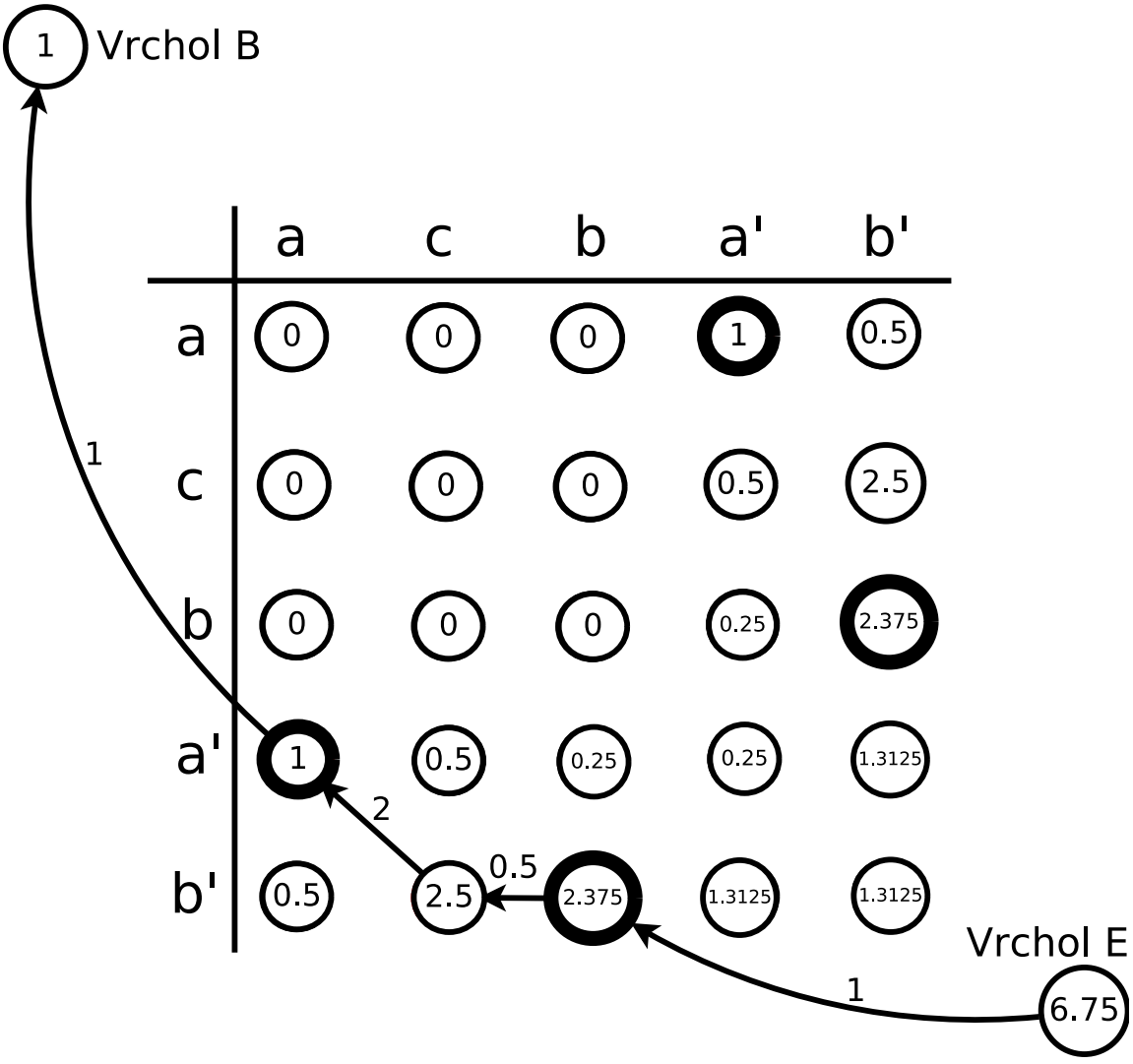
$$\ell = w = h + 1 = k + d + 1 \ \Rightarrow \ k = \ell - d - 1$$

Šikmé hrany teda prispievajú ku skóre cesty hodnotou $2^{\ell-d-1}$, horizontálne a vertikálne prispievajú hodnotou 0.5^d, preto hodnota cesty je $2^{\ell-d-1} \cdot 0.5^d$ čo je požadovaná skórovacia funkcia f.

Podobne ako v predchádzajúcich modeloch nájdeme jednu konkrétnu duplikáciu spätným stochastickým prechodom. Na obrázku 4.7 vidíme príklad jednej konkrétnej duplikácie. Prvou udalosťou je duplikácia $\langle\ a'\ c'\ b'\ \rangle \xrightarrow{\text{sa kopíruje na}} \langle\ a\ c\ b\ \rangle$. Dostaneme ale sekvenciu $\langle\ a\ c\ b\ a'\ c\ b'\ \rangle$, ktorá sa líši od sekvencie seq. Preto hneď po duplikácii nastane druhá udalosť, zmazanie prebytočného atómu $\langle\ c\ \rangle$. Následne dostaneme pôvodnú sekvenciu $\langle\ a\ c\ b\ a'\ b'\ \rangle$. Ancestrálna sekvencia je potom $\langle\ a'\ c\ b'\ \rangle$. Táto duplikácia mala hodnotu $\frac{2.375 \cdot 1}{6.75} \cdot \frac{2.5 \cdot 0.5}{2.375} \cdot \frac{1 \cdot 2}{2.5} \cdot \frac{1 \cdot 1}{1} = \frac{1}{6.75}$.



Obr. 4.6: Príklad rozšíreného modelu duplikácií s deléciami.



Obr. 4.7: Príklad cesty zobrazujúcej duplikáciu s deléciou.

Podobným spôsobom, akým sme rozšírili základný model o reverzie, vieme rozšíriť aj tento model s deléciami. Rovnako vyrobíme druhú maticu S^R, ktorá bude predstavovať reverzné duplikácie a povolíme v nej horizontálne aj vertikálne hrany. Vzťah pre výpočet hodnôt vrcholov v matici S^R bude:

$$S^R_{i,j} = \begin{cases} S^R_{i-1,j-1} \cdot z \ + S^R_{i-1,j} \cdot g + S^R_{i,j-1} \cdot g + B \cdot b & , (S^R_{i-1,j-1} \text{ je aktívny}) \wedge \\ & \quad (S^R_{i,j} \text{ je aktívny}) \\ S^R_{i-1,j-1} \cdot z \ + S^R_{i-1,j} \cdot g + S^R_{i,j-1} \cdot g & , (S^R_{i-1,j-1} \text{ je aktívny}) \wedge \\ & \quad \neg(S^R_{i,j} \text{ je aktívny}) \\ S^R_{i-1,j} \cdot g + S^R_{i,j-1} \cdot g & , \text{inak} \end{cases} \tag{4.12}$$

Doteraz sme ukázali niekoľko modelov na hľadanie duplikácií. Prešli sme postupne od najjednoduchšieho modelu popisujúceho jednoduché duplikácie, až po model zahŕňajúci reverzné duplikácie aj s deléciami. V nasledujúcej podkapitole ukážeme ešte zložitejší model, ktorým sme sa snažili viac priblížiť k biologicky motivovanej tvorbe duplikácií.

4.4 Hlavný model

Z predchádzajúcich podkapitol máme k dispozícii model na hľadanie duplikácií, ktorý v sebe zahŕňa reverzie aj delécie. Tento model ale neobsahuje dve dôležité vlastnosti. Prvou je, že v biológii je už samotná duplikácia zriedkavý jav a k nej pripojené delécie ešte zriedkavejší. Preto sme upravili model tak, aby uprednostňoval dlhšie delécie pred viacerými kratšími. Dosiahli sme to penalizáciou začiatkov delécií. Druhou, ešte dôležitejšou vlastnosťou, ktorú zahrnieme v tomto modeli, je fakt, že chceme navrhovať duplikácie so skórovacou funkciou, ktorá pridáva bonus, ak bola rovnaká udalosť použitá v predchádzajúcej histórii. Budeme sa nateraz venovať len duplikáciám bez reverzií. Všetky nasledujúce postupy sú ale priamo aplikovateľné aj na reverznú časť duplikácií.

Nech duplikácia x má dĺžku ℓ a nech za ňou hneď nasleduje k delécií s dĺžkami $d_1, \ldots, d_k$ a nech H je množina udalostí predchádzajúcej histórie. Potom skórovacia funkcia f má tvar

$$f(x, H) = s^{\ell - \left(\sum_{i=1}^{k} d_i\right) - 1} \ \cdot \ \left(\prod_{i=1}^{k} z_D \ \cdot \ p_D^{d_i - 1} \ \cdot \ k_D \right) \ + \ (b \cdot g(x, H)) \tag{4.13}$$

kde

- z_D - začiatok delécie,
- p_D - pokračovanie delécie,
- k_D - koniec delécie,
- s - duplikácia medzi dvomi rôznymi atómami s rovnakým typom,
- b - bonus, ak sa duplikácia nachádzala aj v predchádzajúcej histórii H,
- $g(x, H) = \begin{cases} 1 & , \text{ak } x \in H \\ 0 & , \text{ak } x \notin H \end{cases}$

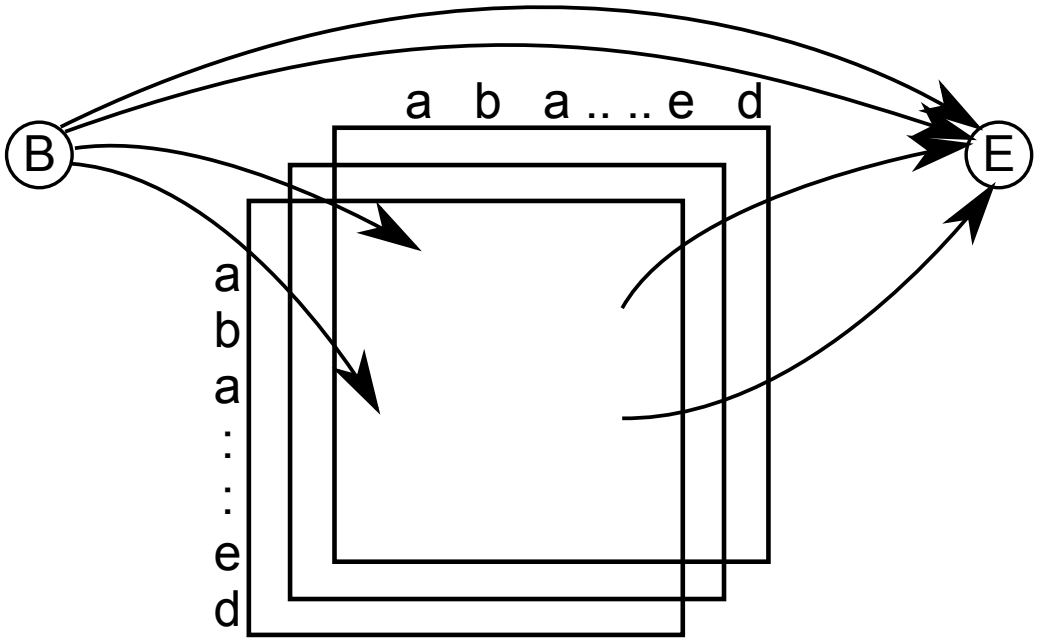
Pôvodnú maticu vrcholov S sme rozdelili na 3 matice vrcholov S, U a D. V novej matici S boli aktívne vrcholy vyberané podľa rovnakého pravidla ako v pôvodnej matici. Vrcholy v maticiach U a D predstavujú atómy, ktoré bude treba po duplikácii odstrániť (delécie). Na začiatku mali všetky vrcholy v maticiach nulovú hodnoty. To, ktoré hrany prechádzajú medzi vrcholmi v maticiach najlepšie vidieť na vzťahoch, ktorými sa rátajú hodnoty týchto vrcholov:

$$S_{i,j} = \begin{cases} U_{i-1,j-1} \cdot k_D + D_{i-1,j-1} \cdot k_D + & , (S_{i,j} \text{ je aktívny}) \wedge \\ \quad + S_{i-1,j-1} \cdot s + B \cdot b & \quad (S_{i-1,j-1} \text{ je aktívny}) \\ U_{i-1,j-1} \cdot k_D + D_{i-1,j-1} \cdot k_D + & , (S_{i,j} \text{ je aktívny}) \wedge \\ \quad + B \cdot b & \quad \neg(S_{i-1,j-1} \text{ je aktívny}) \\ 0 & , \text{inak} \end{cases} \tag{4.14}$$

$$U_{i,j} = \begin{cases} U_{i,j-1} \cdot p_D + D_{i,j-1} \cdot k_D \cdot z_D + S_{i,j-1} \cdot z_D & , S_{i,j-1} \text{ je aktívny} \\ U_{i,j-1} \cdot p_D + D_{i,j-1} \cdot k_D \cdot z_D & , \neg(S_{i,j-1} \text{ je aktívny}) \end{cases} \tag{4.15}$$

$$D_{i,j} = \begin{cases} D_{i-1,j} \cdot r_D + D_{i-1,j} \cdot k_D \cdot z_D + S_{i-1,j} \cdot z_D & , S_{i-1,j} \text{ je aktívny} \\ U_{i-1,j} \cdot r_D + D_{i-1,j} \cdot k_D \cdot z_D & , \neg(S_{i-1,j} \text{ je aktívny}) \end{cases} \tag{4.16}$$

Ešte potrebujeme v grafe reprezentovať bonusy za duplikácie, ktoré sa nachádzajú v predchádzajúcej histórii H. Z vrchola B nebudú vychádzať hrany len do aktívnych vrcholov v matici S, ale pre všetky udalosti $u \in H$, ktoré by sme vedeli aplikovať na sekvenciu seq, budú vychádzať hrany aj do vrchola E. Tieto hrany budú mať



Obr. 4.8: Príklad hlavného modelu duplikácií s deléciami obsahujúci aj bonusové hrany.

hodnotu b. Do vrchola E teda vchádzajú hrany nielen z aktívnych vrcholov, ale aj z vrchola B. Stručný príklad tohoto upraveného grafu môžeme vidieť na obrázku 4.8.

Všetko ostatné funguje presne rovnako ako v predchádzajúcich modeloch. Matice U a D sú len rozšírením pôvodnej matice S do ďalšieho rozmeru, aby sme vedeli ohodnotiť začiatok a koniec delécie. Hrany z vrholov matice S do vrcholov matice U, respektíve D znamenajú začiatok delécie, hrany späť do matice S znamenajú koniec delécie a preto sú aj ohodnotené hodnotami z_D, respektíve k_D. Rozšírenie o reverzie je už pomerne jednoduché. Stačí rozdeliť maticu S^R na tri matice S^R, U^R a D^R podobne, ako sme rozdelili pôvodnú maticu S a ako sme rozširovali modely o reverzie.

Výber jednej konkrétnej duplikácie funguje analogicky ako v predchádzajúcich modeloch až na prípad, keď si algoritmus vyberie hranu h vedúcu z vrchola E priamo do vrchola B (tzv. bonusovú hranu). V tom prípade ešte musí nájsť pravdepodobnosť udalosti prislúchajúcej k vybranej hrane cieľavedomým (nie náhodným) prechodom cez matice S, U a D a pripočítať ju k pravdepodobnosti vybranej hrany h.

Tento model sme následne používali pri hľadaní duplikácií. Algoritmicky sa tieto

vrcholy v maticiach dajú tiež vyplniť dynamickým programovaním v časovej aj pamäťovej zložitosti $O(n^2)$, kde n je veľkosť vstupnej atomizovanej sekvencie seq. Následný výber cesty pomocou stochastického prechodu má časovú zložitosť $O(n)$.

4.5 Speciácie

Jednoduchou modifikáciou tohoto algoritmu môžeme riešiť aj poslednú duplikačnú udalosť, speciáciu. Rozdiel je v tom, že na vstupe máme súčasné sekvencie dvoch organizmov a uvažujeme len speciácie bez reverzií. Sekvencia atómov organizmu A bude seq^A a organizmu B bude seq^B. Skórovaciu funkciu pre speciácie ponecháme rovnakú ako v hlavnom modeli.

Zmeníme ale pravidlo pre aktívnosť vrcholov v matici S:

$$\forall\ S_{i,j}\ \in\ S:\ S_{i,j} \text{ je aktívny} \Leftrightarrow (seq_i^A \neq seq_j^B)\ \wedge\ (seqtype_i^A = seqtype_j^B) \quad (4.17)$$

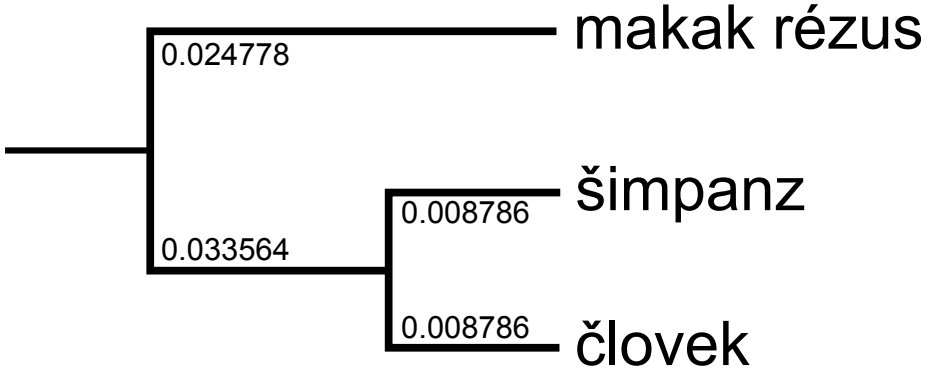
Podmienka $(seq_i^A \neq seq_j^B)$ je v tomto prípade zbytočná, keďže porovnávame dve nezávislé rozdielne sekvencie dvoch organizmov, ale pre analógiu s predchádzajúcim modelom sme ju ponechali.

Do matíc U a D pridáme ešte jeden počiatočný riadok a stĺpec nulových vrcholov. Tie budú predstavovať počiatočné dlhé delécie v jednej zo sekvencií seq^A alebo seq^B. Je tu ešte jedna dôležitá zmena. Z vrchola B budú vychádzať hrany len do vrcholov $U_{0,0}$, $D_{0,0}$ a do vrchola $S_{1,1}$, ak je aktívny. Do vrchola E zase budú vchádzať hrany len z vrcholov U_{N_A,N_B}, D_{N_A,N_B} a z vrchola S_{N_A,N_B}, ak je aktívny, kde N_A je dĺžka sekvencie seq^A a N_B je dĺžka sekvencie seq^B. Analogicky potom bude cesta z vrcholu E do vrchola B nájdená spätným stochastickým prechodom tvoriť jednu možnú speciáciu.

Kapitola 5

Aplikácia na dáta

Skúšobnú implementáciu nášho vyššie popísaného MCMC vzorkovacieho algoritmu sme sa rozhodli otestovať na simulovaných dátach, ktoré sme prevzali z práce Vinař et al. [VBSS09]. Dáta boli simulované na fylogenetickom strome človeka, šimpanza a makaka rézusa. Dĺžky hrán vo fylogenetickom strome (viď obr. 5.1) boli získané porovnaním skutočných sekvencií týchto organizmov získaných z celogenómových zarovnávaní ľudského chromozómu 22, ktoré sú k dispozícii v databáze UCSC Genome browser (Karolchik et al. (2008) [KKB+08]).



Obr. 5.1: Fylogenetický strom troch skúmaných druhov aj s dĺžkami hrán.

Pre naše účely sme použili 20 simulovaných sekvencií zhlukov génov s pomalšou rýchlosťou evolúcie (viď Vinař et al.[VBSS09]). Základné vlastnosti týchto dát môžeme vidieť v tabuľke 5.1. Keďže ide o simulované sekvencie, u ktorých poznáme správnu históriu, použili sme skutočnú správnu atomizáciu, pričom sme z dát vylúčili atómy kratšie ako 500 znakov.

MCMC algoritmus na rekonštrukciu duplikačných histórií sme implementovali

Tabuľka 5.1: Základné vlastnosti 20 simulovaných génových zhlukov s pomalšou evolúciou.

	min	max	priemer
Dĺžky sekvencií (v 1000)	91	295	172
Počet typov atómov	15	53	36
Počet duplikácií	5	24	15
Počet delécií	0	3	0,8

v jazyku *C++*. Nastavovali sme rôzne parametre skórovacej funkcie danej výrazom 4.13 a ukážky niekoľkých rekonštruovaných histórií z dát jednej simulovanej histórie (vybrali sme ôsmu z 20 simulovaných histórií) môžeme vidieť na nasledujúcich obrázkoch.

Obrázok 5.2 predstavuje simulovanú históriu, ku ktorej by sme sa radi priblížili. Na všetkých týchto obrázkoch znázorňuje každá farba jeden typ atómu. Následne sme si zo simulovanej histórie vybrali iba súčasné sekvencie atómov a s nimi sme sa snažili zrekonštruovať duplikačnú históriu. Pozitívnym zistením bol fakt, že v každom prípade sme vedeli vzorkovať 30000 duplikačných histórií za menej ako 6 hodín (počítač 2,27 GHz CPU, 16 GB RAM). Kvôli menším problémom so správou pamäte v našom algoritme sme neboli schopní vzorkovať dlhšie Markovove reťazce duplikačných histórií. Vzorkovanie s podobnými parametrami pomocou pôvodného algoritmu [VBSS09] by trvalo niekoľko týždňov.

Pre rekonštrukciu histórie na obrázku 5.3 sme použili parametre skórovacej funkcie, ktoré viac penalizovali delécie. Vybraná bola 11523. vzorka z Markovovho reťazca.

- $z_D = 0,1$ - začiatok delécie
- $p_D = 0,7$ - pokračovanie delécie
- $k_D = 0,8$ - koniec delécie
- $s = 2$ - duplikácia medzi dvomi rôznymi atómami s rovnakým typom

Vidíme, že na rozdiel od simulovanej histórie, kde väčšina duplikácií nastala hlavne na vetve medzi koreňom stromu a spoločným predkom človeka a šimpanza, nastali

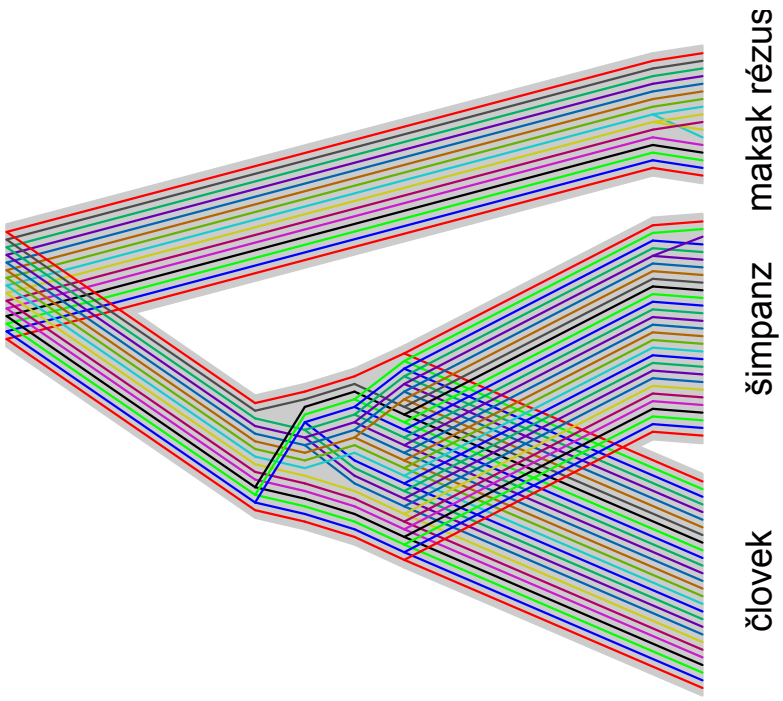
v tejto rekonštruovanej histórii duplikácie najmä na vetvách medzi listami a spoločným predkom. Následná speciácia človeka a šimpanza je v našom príklade oveľa kratšia ako v simulovanej histórii. Dôvodom mohlo byť, že ak obmedzíme skóre za delécie, príspevok speciácie pri náhodnom výbere udalostí nebol dostatočne veľký na to, aby prekonal hodnotu duplikácií. Vidíme totiž, že tesne pred speciáciou sa už v sekvenciách človeka a šimpanza nedali nájsť žiadne duplikácie a tak bola vybraná speciácia. Rozhodli sme sa preto zmenšiť penalizáciu za delécie

Na obrázkoch 5.4 a 5.5 sme použili parametre, ktoré menej penalizovali delécie a naviac na obrázku 5.5 sme viac preferovali speciáciu pred duplikáciou.

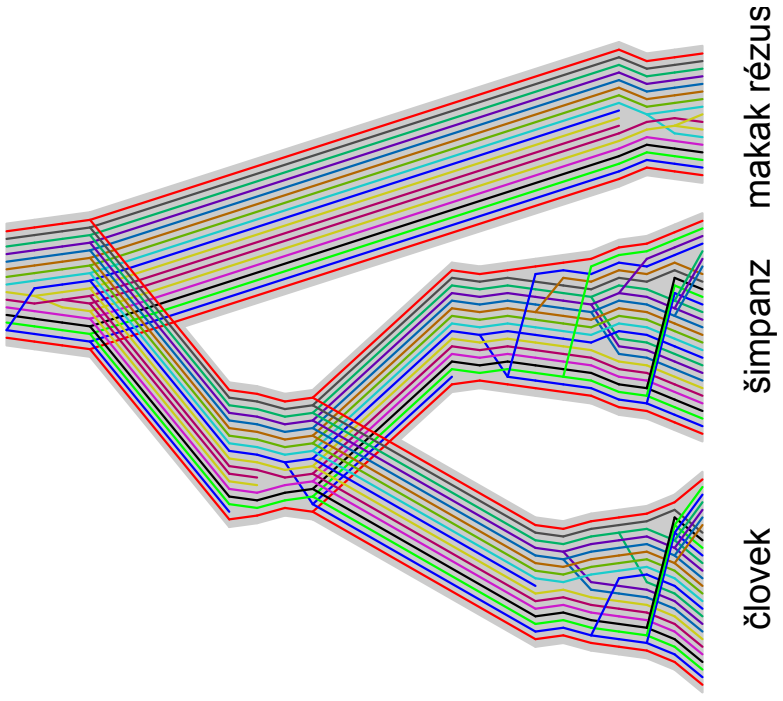
- $z_D = 0,4$ - začiatok delécie
- $p_D = 0,8$ - pokračovanie delécie
- $k_D = 1$ - koniec delécie
- $s = 2$ - duplikácia medzi dvomi rôznymi atómami s rovnakým typom

Na obrázku 5.4 bola vybraná 6805. vzorka a na obrázku 5.5 27215. vzorka. Ako vidíme na obrázku 5.4, zmena penalizácie delécií nepriniesla očakávané výsledky. Preto sme sa rozhodli zvýhodniť speciácie. Výsledok vidieť na obrázku 5.5. Všetky duplikácie sa presunuli až za spoločného predka všetkých troch organizmov.

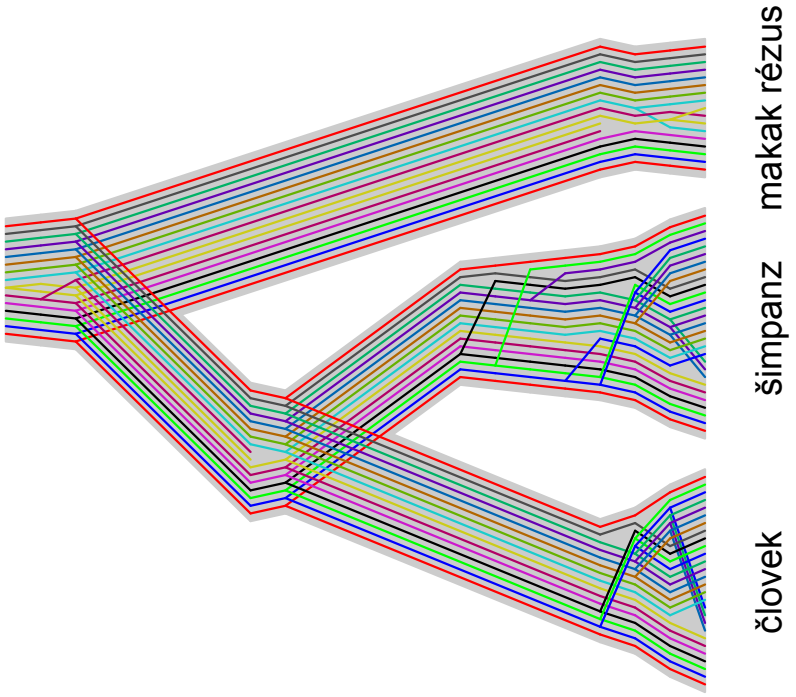
Ako vidíme, nastavenie parametrov skórovacej funkcie je veľmi dôležitým faktorom pri aplikácii MCMC na tento typ dát. V budúcnosti plánujeme preskúmať systematické spôsoby na nastavovanie týchto parametrov napr. pomocou metód strojového učenia, ako aj pridanie nových členov do skórovacej funkcie tak, ako to navrhujú Vinař et al. [VBSS09]. Výsledkom bude algoritmus na rekonštrukciu duplikačných histórií, ktorý bude presný a bude podstatne efektívnejší ako implementácia [VBSS09].



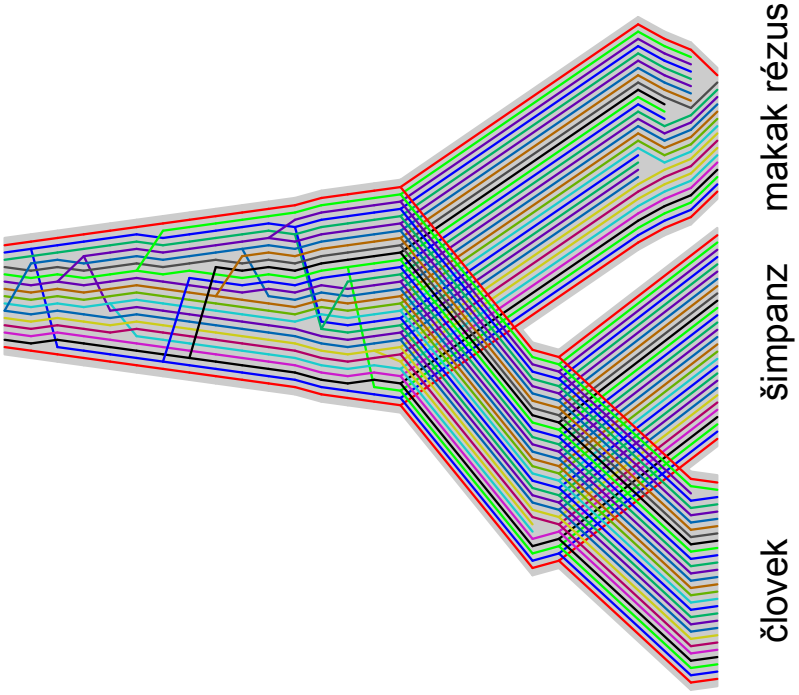
Obr. 5.2: Simulovaná história. Každá farba predstavuje jeden typ atómu.



Obr. 5.3: Rekonštruovaná duplikačná história s viac penalizovanými deléciami.



Obr. 5.4: Rekonštruovaná duplikačná história s menej penalizovanými deléciami.



Obr. 5.5: Rekonštruovaná duplikačná história s menej penalizovanými deléciami a zvýhodnenými speciáciami.

Záver

V tejto práci sme sa venovali skúmaniu bioinformatického problému rekonštrukcie duplikačných histórií. Zaviedli sme pravdepodobnostný model evolúcie, ukázali sme algoritmus na výpočet vierohodnosti duplikačných histórií ale najmä sme vytvorili efektívny pravdepodobnostný algoritmus pre navrhovanie jednotlivých duplikačných udalostí. Na rekonštruovanie duplikačných histórií sme používali inferenciu vzorkovaním za pomoci MCMC algoritmu.

Náš algoritmus má veľký potenciál na zlepšenia. Jednou z možností zlepšenia je určenie parametrov skórovacej funkcie. Tie by sa dali odhadnúť metódami strojového učenia, prípadne štatistickými metódami. Ďalším problémom je zavedenie ďalších členov do skórovacej funkcie pre pravdepodobnostný výber duplikácie, ktoré budú odzrkadľovať informácie obsiahnuté v DNA sekvenciách jednotlivých atómov.

Po vyladení konštánt MCMC algoritmu a doplnenia skórovacej funkcie plánujeme nami navrhnutý algoritmus použiť na analýzu biomedicínsky významných zhlukov génov v ľudskom genóme.

Literatúra

[BBV11] B. Brejová, M. Burger, and T. Vinař. Automated Segmentation of DNA Sequences with Complex Evolutionary History. 2011. Nepublikovaný report.

[BMT+08] W. Boomsma, K.V. Mardia, C.C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932, 2008.

[Bur10] Michal Burger. Atomization of DNA Sequences with Complex Evolutionary History. Master's thesis, Comenius University in Bratislava, 2010. Supervised by Tomáš Vinar.

[CP03] S.L. Cawley and L. Pachter. HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, 19(suppl 2):ii36, 2003.

[Fel81] Joseph Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981. 10.1007/BF01734359.

[GRS96] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996.

[Has70] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[HKY85] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2):160–174, 1985.

[JC69] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. *Mammalian protein metabolism*, page 21–123, 1969.

[KKB+08] D. Karolchik, RM Kuhn, R. Baertsch, GP Barber, H. Clawson, M. Diekhans, B. Giardine, RA Harte, AS Hinrichs, F. Hsu, et al. The UCSC genome browser database: 2008 update. *Nucleic acids research*, 36(suppl 1):D773, 2008.

[MRR+08] J. Ma, A. Ratan, B.J. Raney, B.B. Suh, W. Miller, and D. Haussler. The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences*, 105(38):14254, 2008.

[NW70] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[SW81] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.

[VBSS09] T. Vinař, B. Brejová, G. Song, and A. Siepel. Reconstructing histories of complex gene clusters on a phylogeny. *Comparative Genomics*, pages 150–163, 2009.

[ZSV+08] Y. Zhang, G. Song, T. Vinař, E.D. Green, A. Siepel, and W. Miller. Reconstructing the evolutionary history of complex human gene clusters. In *Proceedings of the 12th annual international conference on Research in computational molecular biology*, pages 29–49. Springer-Verlag, 2008.