Zarovnávanie sekvencií s použitím metód klasifikácie

Diplomová práca

Bc. Michal Hozza

Vedúci práce: Mgr. Tomáš Vinař, PhD.
Konzultant: Mgr. Michal Nánási

Fakulta matematiky, fyziky a informatiky, Univerzita Komenského, Bratislava

17.6.2014

Obsah

Zarovnávanie sekvencií

- jedným zo základných bioinformatických problémov
- identifikuje časti sekvencie, ktoré vznikli z toho istého predka (zarovnané bázy), inzercie a delécie v priebehu evolúcie (medzery v zarovnaní)

Príklad:

```
GTGGACCGTT------CCTTCCGGCAATCACGAGAAAAGCCACGT
GTCGACCGTTTCAGTGACTTGAAGCAATCAGG--AACACCACCT
```

Zarovnávanie sekvencií

- jedným zo základných bioinformatických problémov
- identifikuje časti sekvencie, ktoré vznikli z toho istého predka (zarovnané bázy), inzercie a delécie v priebehu evolúcie (medzery v zarovnaní)

Príklad:

```
GTGGACCGTT------CCTTCCGGCAATCACGAGAAAAGCCACGT
GTCGACCGTTTCAGTGACTTGAAGCAATCAGG---AACACCACCT
```

Zarovnávanie sekvencií ako informatický problém

- Vstupom sú dve sekvencie $X = x_1 x_2 \ldots x_n$ a $Y = y_1 y_2 \ldots y_m$
- Výstupom je zarovnanie sekvencií, ktoré má najvyššie možné skóre podľa danej skórovacej schémy
- Skórovacia schéma môže byť napr. $+1$ za zhodu a -1 za medzeru alebo nezhodu
- V tomto prípade sa dá najlepšie zarovnanie nájsť v čase $O(nm)$ dynamickým programovaním
- V praxi sa používajú zložitejšie schémy, v našej práci používame párové HMM

Párové Skryté Markovovské modely (pHMM)

- pravdepodobnostný model inšpirovaný konečnými automatmi

Definícia (pHMM)

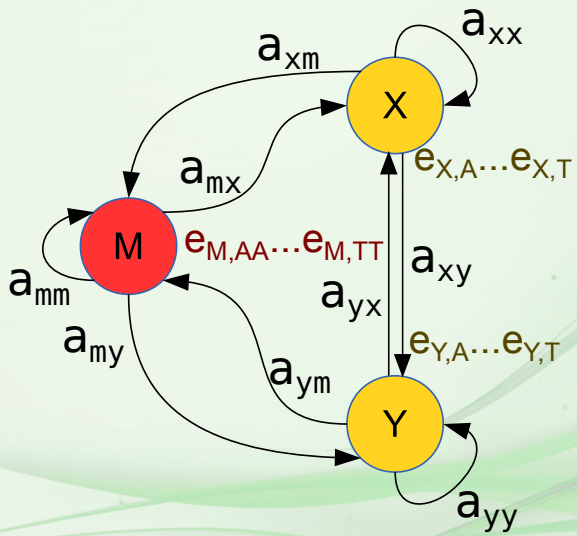
pHMM je 6-ica $(K, \Sigma_X, \Sigma_Y, \pi, a, e)$, kde K je množina stavov, Σ_X je množina symbolov $X - ovej$ sekvencie, Σ_Y je množina symbolov $Y - ovej$ sekvencie, $\pi = \{\pi_i\}_{i \in K}$ je distribúcia začiatočných stavov, $a = \{a_{i,j}\}_{i,j \in K}$ je distribúcia prechodov a $e = \{e_{i,x,y}\}_{i \in K, x \in \Sigma_X \cup \{\varepsilon\}, y \in \Sigma_Y \cup \{\varepsilon\}}$ je distribúcia emisií.

- narozdiel od jednoduchého HMM generuje dve sekvencie, pričom môže mať stavy, ktoré generujú symboly v oboch sekvenciách, alebo len v jednej zo sekvencií

Párové Skryté Markovovské modely (pHMM)

Príklad: pHMM pre zarovnávanie sekvencií

- $K = \{M, X, Y\}$, $\Sigma_X = \Sigma_Y = \{A, C, G, T\}$
- M *(Match)* – emituje dvojice báz: $AA, AC, AG, \ldots, TT$
- X *(Insert X)* a Y *(Insert Y)* – emitujú jednotlivé bázy: A, C, G, T



Párové Skryté Markovovské modely (pHMM)

- Postupnosť stavov definuje konkrétne zarovnanie vstupných sekvencií
- Pri zarovnávaní hľadáme najpravdepodobnejšiu postupnosť stavov (pomocou Viterbiho algoritmu)

Príklad:

- Vstupné sekvencie

  GTGGACCGTTCCTTCCGGCAATCACGAGAAAAGCCACGT
  GTCGACCGTTTCAGTGACTTGAAGCAATCAGGAAGACCACCT

- Jedno z možných zarovnaní spolu so stavmi

  GTGGACCGTT------CCTTCCGGCAATCACGAGAAAAGCCACGT
  MMMMMMMMMMIIIIIIMMMMMMMMMMMMMMMMXXXMMMMMMMMMM
  GTCGACCGTTTCAGTGACTTGAAGCAATCAGG---AACACCACCT

Párové Skryté Markovovské modely (pHMM)

- Postupnosť stavov definuje konkrétne zarovnanie vstupných sekvencií
- Pri zarovnávaní hľadáme najpravdepodobnejšiu postupnosť stavov (pomocou Viterbiho algoritmu)

Príklad:

- Vstupné sekvencie

```
GTGGACCGTTCCTTCCGGCAATCACGAGAAAAGCCACGT
GTCGACCGTTTCAGTGACTTGAAGCAATCAGGAACACCACCT
```

- Jedno z možných zarovnaní spolu so stavmi

```
GTGGACCGTT------CCTTCCGGCAATCACGAGAAAAGCCACGT
MMMMMMMMMMYYYYYYMMMMMMMMMMMMMMMMMXXXMMMMMMMMM
GTCGACCGTTTCAGTGACTTGAAGCAATCAGG---AACACCACCT
```

Zarovnávanie sekvencií s anotáciou

- máme k dispozícii anotácie k sekvenciám (napr. gén/negén)
- anotácia môže byť aj komplexnejšia – pozostávajúca z rôznych indikátorov, pochádzajúcich z biologických experimentov
- to nám môže pomôcť lepšie určiť biologicky korektné zarovnanie

Príklad:

- S anotáciou

```
0000000      1111111111111111111100000000000
GTGGACC------GTTCCTTCCGGCAATCACGAGAAAAGCCACGT
GTCGACCGTTTCAGTGACTTGAAGCAATCAGGAA---CACCACCT
0000000000000111111111111111111111   00000000
```

- Bez anotácie

```
GTGGACCGTT------CCTTCCGGCAATCACGAGAAAAGCCACGT
GTCGACCGTTTCAGTGACTTGAAGCAATCAGG---AACACCACCT
```

Zarovnávanie sekvencií s anotáciou

- máme k dispozícii anotácie k sekvenciám (napr. gén/negén)
- anotácia môže byť aj komplexnejšia – pozostávajúca z rôznych indikátorov, pochádzajúcich z biologických experimentov
- to nám môže pomôcť lepšie určiť biologicky korektné zarovnanie

Príklad:

- S anotáciou

```
0000000       11111111111111111111000000000000
GTGGACC------GTTCCTTCCGGCAATCACGAGAAAAGCCACGT
GTCGACCGTTTCAGTGACTTGAAGCAATCAGGAA---CACCACCT
000000000000011111111111111111111   00000000
```

- Bez anotácie

```
GTGGACCGTT------CCTTCCGGCAATCACGAGAAAAGCCACGT
GTCGACCGTTTCAGTGACTTGAAGCAATCAGG---AACACCACCT
```

Ciele práce

- cieľom práce je vytvoriť nové metódy na korekciu zarovnaní biologických sekvencií na základe prídavnej informácie
- integrácia tejto informácie bude zabezpečená pomocou techník využívaných na klasifikáciu v strojovom učení

- vzhľadom ku komplexnosti pridanej informácie, je veľmi ťažké systematickým spôsobom zakomponovať anotáciu do skórovacích schém založených na pHMM
- modelovanie takejto informácie v rámci pHMM by bolo neefektívne

Ciele práce

- cieľom práce je vytvoriť nové metódy na korekciu zarovnaní biologických sekvencií na základe prídavnej informácie
- integrácia tejto informácie bude zabezpečená pomocou techník využívaných na klasifikáciu v strojovom učení

- vzhľadom ku komplexnosti pridanej informácie, je veľmi ťažké systematickým spôsobom zakomponovať anotáciu do skórovacích schém založených na pHMM
- modelovanie takejto informácie v rámci pHMM by bolo neefektívne

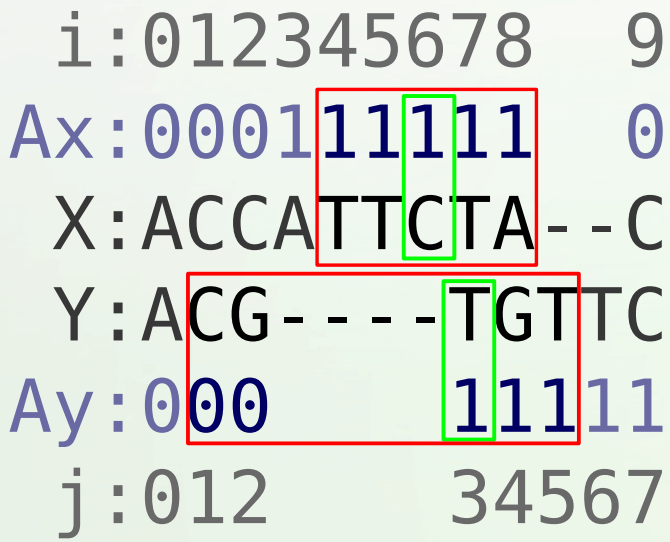
Zhrnutie nášho riešenia

1. informáciu zosumarizujeme do jedného čísla
   - pre pár pozícií (i, j) chceme číslo $S(i, j)$, ktoré na základe lokálnej informácie vyjadruje, ako dobre k sebe príslušné pozície pasujú
   - to je vhodná úloha pre klasifikátory
2. navrhneme systematický spôsob, ako toto číslo zakomponovať do skórovacej schémy pHMM

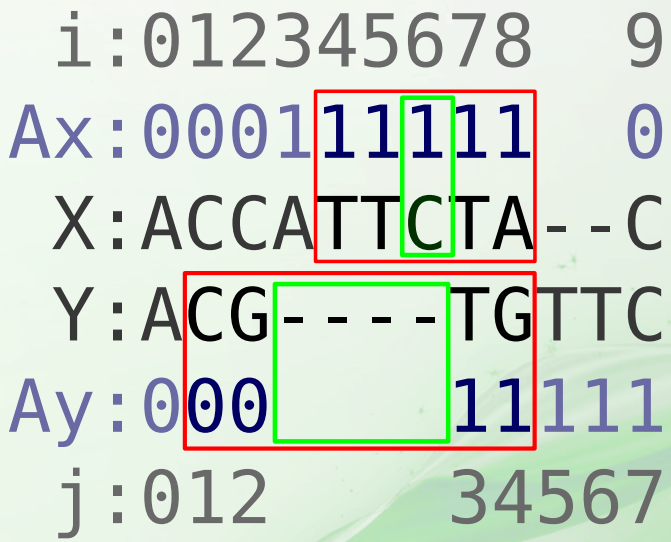
Klasifikácia na základe lokálnej informácie

- anotácie sme zakomponovali pomocou dvoch typov klasifikátorov
  1. Match – rozhoduje či dané pozície majú byť zarovnané k sebe
  2. Indel – rozhoduje či daná pozícia má byť zarovnaná k medzere
- výstup $\in \langle 0, 1 \rangle$ – istota klasifikátora, že dané dve pozície majú byť zarovnané k sebe (v Indel klasifikátore, že daná pozícia má byť zarovnaná k medzere)
- atribúty sú okná veľkosti w

Okno pre klasifikátor



(a) Match klasifikátor



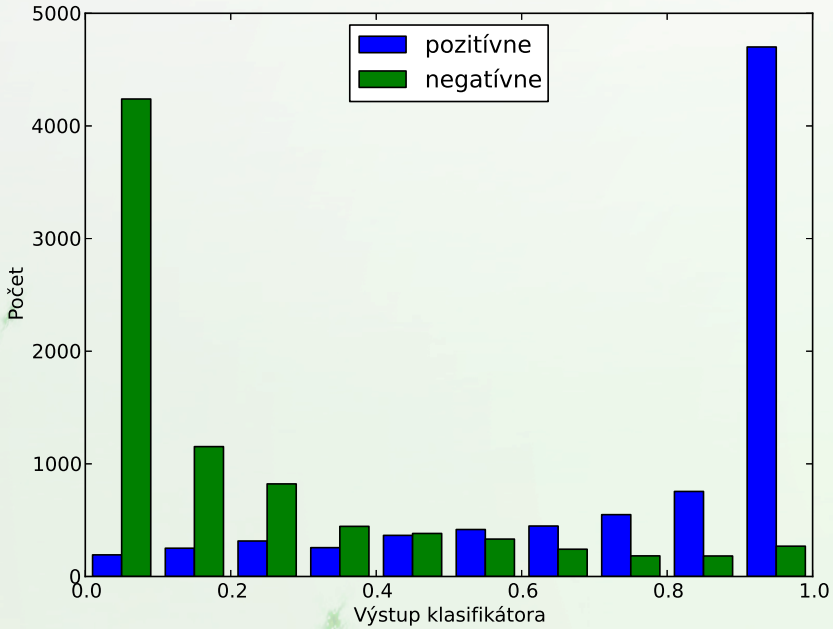
(b) Indel klasifikátor

Obr. : Okno klasifikátora pre pozície $i = 6$ a $j = 3$

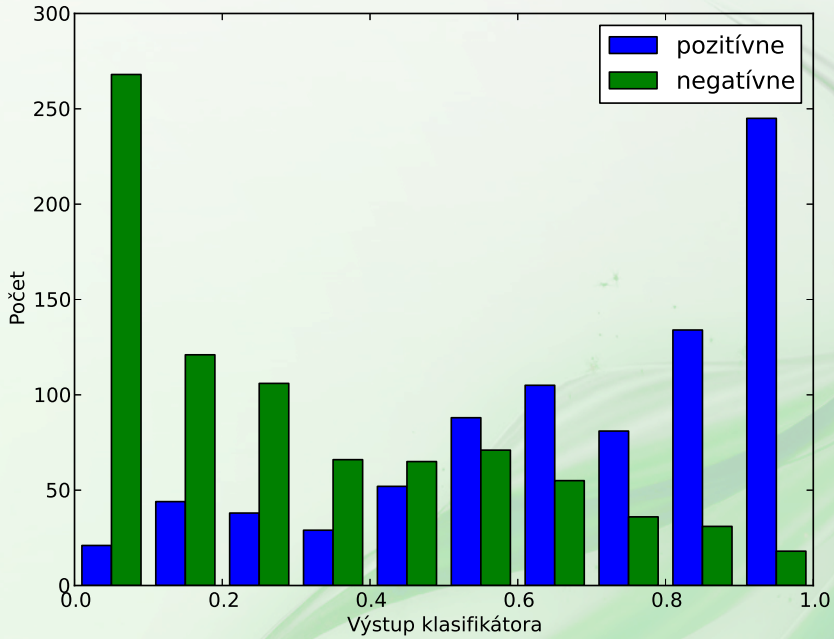
Klasifikácia na základe lokálnej informácie

- k dátam sme pridali informácie o zhodách na zodpovedajúcich pozíciách, čím sa nám podarilo vylepšiť úspešnosť klasifikátora
- úspešnosť Match klasifikátora: **89,87%**
- úspešnosť Indel klasifikátora: **81,78%**
- klasifikátor sa dokáže naučiť, ktoré okná majú byť zarovnané k sebe a ktoré nie

Úspešnosť klasifikátora



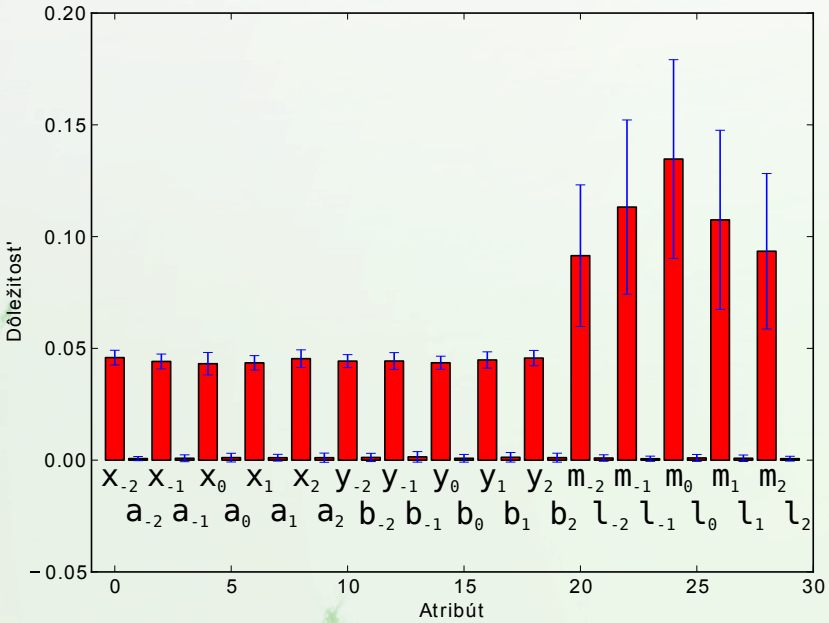
(a) Match klasifikátor



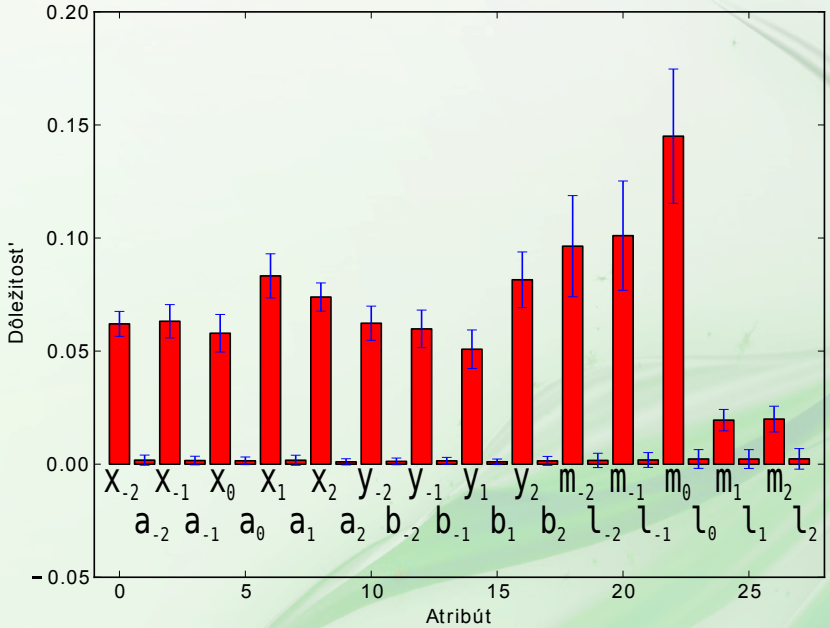
(b) Indel klasifikátor

Obr. : Distribúcia výstupu klasifikátora pre pozitívne a negatívne príklady.

Dôležitosť atribútov



(a) Match klasifikátor



(b) InDel klasifikátor

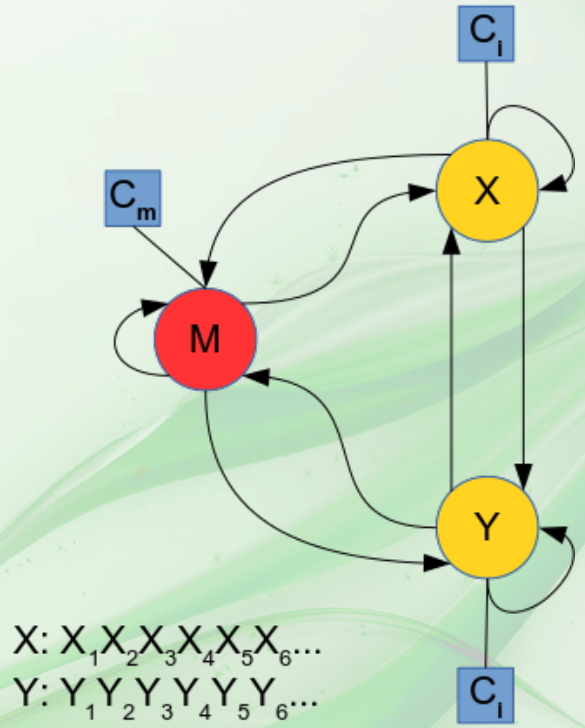
Obr. : Dôležitosť atribútov v klasifikátore.

Zakomponovanie výsledkov klasifikácie do pHMM

- skonštruovali sme dva modely pre zarovnanie sekvencií s anotáciami za pomoci klasifikátora
  1. Model s klasifikátorom ako emisiou
  2. Model s klasifikátorovou páskou
- založené na párových skrytých Markovovských modeloch.

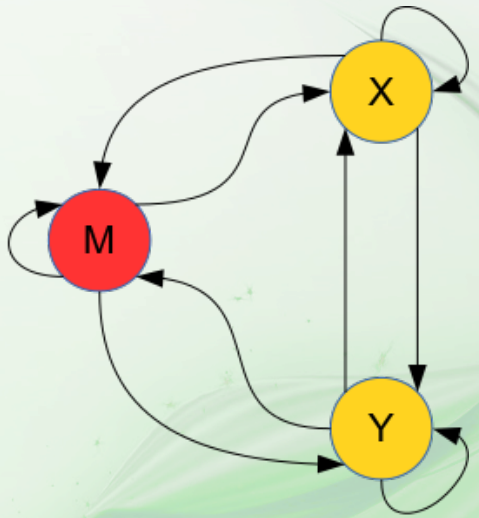
Model s klasifikátorom ako emisiou (Model A)

- emisné tabuľky stavov nahradíme výstupom z klasifikátora
- model nie je korektný pravdepodobnostný model, pretože výstupy klasifikátora nesčitujú do 1
- prechodové pravdepodobnosti sme natrénovali zo zarovnaní z trénovacej vzorky



Model s klasifikátorovou páskou (Model B)

- modelujeme navyše sekvenciu výstupov klasifikátora vo forme pásky
- trénujeme všetky parametre na trénovacej vzorke zarovnaní obohatenej o pásku s výstupmi z klasifikátora

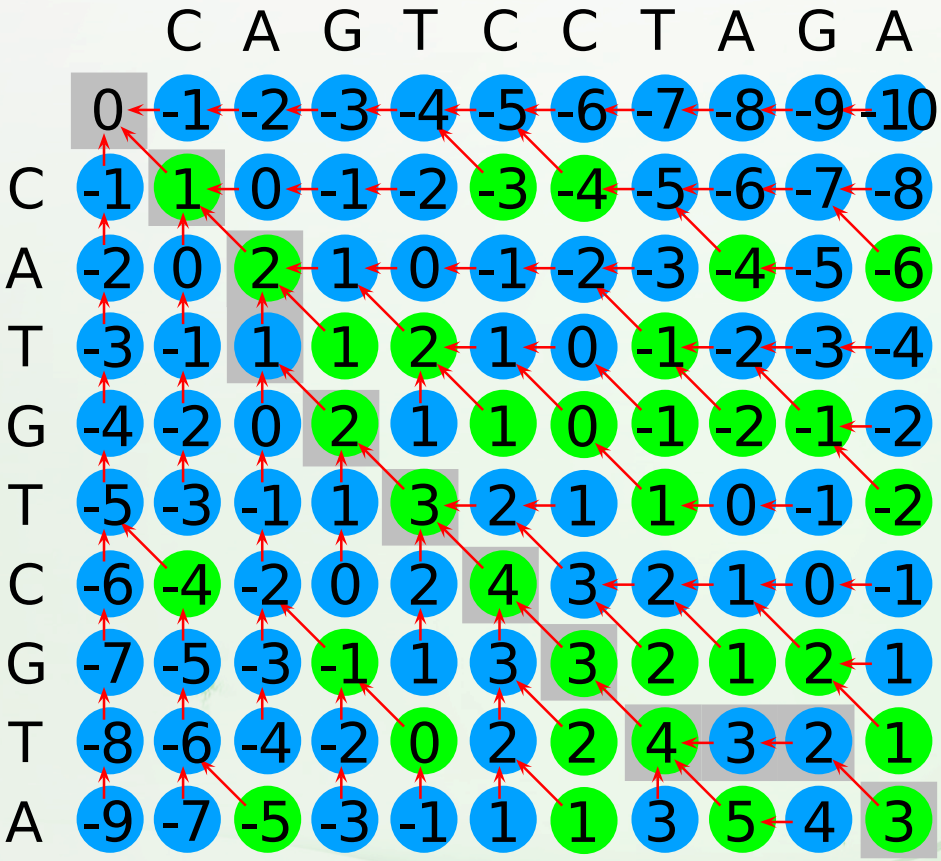


- páska je cesta v 2D tabuľke výstupov klasifikátorov

X: $X_1X_2X_3X_4X_5X_6...$
Y: $Y_1Y_2Y_3Y_4Y_5Y_6...$
C: $C_1C_2C_3C_4C_5C_6...$

- zhoduje sa s cestou zarovnania
- ak sa pohneme horizontálne, alebo vertikálne, používame Indel klasifikátor a ak sa pohneme diagonálne, tak použijeme Match klasifikátor

Klasifikátorová páska




```
CATGTCAT--A
CA-GTCCTAGA
MMIMMMMMIIM
```

Obr. : Použité klasifikátory v klasifikátorovej páske

Výsledky

Dáta	Model A		Model B		Ref. Model		Muscle	
	Zhoda	Tranz.	Zhoda	Tranz.	Zhoda	Tranz.	Zhoda	Tranz.
sim1	79,75%	44,97%	84,35%	56,5%	**85,78%**	**61,03%**	82,72%	58,76%
sim2	70,14%	—	**71,47%**	—	60,38%	—	61,47%	—
bio	**91,40%**	96,63%	91,24%	**96,89%**	91,34%	96,45%	91,28%	95,98%

Tabuľka : Porovnanie našich modelov s referenčným modelom a zarovnávačom muscle.

- Tranzitivitu počítame z troch zarovnaní AB, BC a AC, ako percentuálnu zhodu medzi zložením prvých dvoch zarovnaní $(AB \circ BC)$ a tretieho zarovnania AC.

Záver

- navrhli sme spôsob zakomponovania dodatočnej informácie do zarovnávania sekvencií
- vytvorili sme vhodnú sadu atribútov pre klasifikátory
- vytvorili sme dva modely, ktoré zahŕňajú výstup klasifikátora do zarovnania
- úspešne sme implementovali zarovnávač s dodatočnou informáciou
  - naše riešenie je ľahko rozšíriteľné (umožnuje pridanie nových modelov, sady atribútov, výmenu klasifikátora . . .)
- podarilo sa nám prekonať úspešnosť referenčných zarovnávačov bez anotácie na biologických dátach

Ďakujem za pozornosť!

https://github.com/mhozza/realigner

Čo považujeme za heuristiku?

- Algoritmy, ktoré hľadajú optimálne zarovnanie vzhľadom na nejakú skórovaciu schému a vždy ho nájdu nepovažujeme za heuristické. Takýmto algoritmom je aj Viterbiho algoritmus, ktorý používame v našej práci. Skórovacia schéma je v našom prípade pHMM s klasifikátorom.
- Okrem toho existujú rôzne heuristické algoritmy, ktoré pracujú rýchlejšie za cenu nenájdenia optimálneho zarovnania.

Dôvody pre výber klasifikátora Náhodný les

- V práci sme sa nezaoberali porovnaním rôznych klasifikačných algoritmov
- Namiesto toho sme implementovali zarovnávač, v ktorom je možné jednoducho klasifikátor vymeniť za iný
- Za najlepší tento klasifikátor považujú jeho autori v článku [Breiman, 2001]
- Nás zaujímalo hlavne to, že funguje dostatočne dobre, a že je rýchly
- V rámci pokusov sme vyskúšali aj algoritmus SVM (dosahoval značne horšiu úspešnosť a bol pomalý) a rôzne stromové algoritmy (dosahovali podobnú úspešnosť aj rýchlosť trénovania)
- Z dôvodu časovej náročnosti detailnejších experimentov sme takéto porovnanie nezahrnuli do našej práce

Trénovanie náhodných lesov

- V práci sme si zaviedli dva typy listov – uzavretý a otvorený
- Uzavretý má priradenú triedu, otvorený nie. Jediný list na začiatku je otvorený (pokiaľ neobsahuje len dáta jednej triedy)
- List uzavrieme a priradíme mu triedu, ak všetky dáta v tomto liste patria do tej triedy. Keďže ide učenie s učiteľom, príslušnosť dát do danej triedy poznáme.
- Dátovú sadu, pre každý strom v lese, získame náhodným výberom N vektorov s opakovaním z danej trénovacej sady veľkosti N.
- Druhý prvok náhodnosti spočíva v tom, že v každom vrchole sa vyberie len malá podmnožina atribútov, z ktorých sa vyberá najinformatívnejší atribút
- Časť dát z globálnej sady sa nám v danej množine nevyskytne. Tieto dáta sú tzv. *out of bag* a OOB chyba je chyba počítaná na týchto dátach.

Klasifikátory vs. modely

- v našich modeloch máme tri stavy dvoch typov – jeden Match stav a dva Insert stavy
- Match stav používa Match klasifikátor a Insert stavy používajú Indel klasifikátor

DNA vs. proteíny

- z hľadiska modelov a algoritmov, medzi DNA a proteínmi nie je rozdiel
- rozdiel je vo vlastnostiach sekvencie a v dostupných anotáciách
- časť DNA (gény) kóduje proteíny, preto sa na zarovnávanie proteínov môžme pozerať ako na podproblém zarovnávania DNA.

Literatúra I


Breiman, L. (2001).
Random forests.
Machine learning, 45(1):5–32.