Zarovnávanie sekvencií s použitím metód klasifikácie

Michal Hozza

Školiteľ: Tomáš Vinař[1], Michal Nánási[2]

[1] Katedra aplikovanej informatiky, FMFI UK, Mlynská dolina, 842 48 Bratislava

[2] Katedra informatiky, FMFI UK, Mlynská dolina, 842 48 Bratislava

Úvod

Zarovnávanie dvoch DNA sekvencií je jedným zo základných bioinformatických problémov. Správne zarovnanie identifikuje časti sekvencie, ktoré vznikli z toho istého predka (zarovnané bázy), ako aj inzercie a delécie v priebehu evolúcie (medzery v zarovnaní). Obvykle takéto zarovnanie hľadáme pomocou jednoduchých párových skrytých Markovovských modelov (pHMM) [Durbin et al., 1998]. V našej práci [Hozza, 2014] sa zaoberáme možnosťami použitia prídavnej informácie o funkcii vstupných sekvencií (tzv. anotácie) na zlepšenie kvality takýchto zarovnaní.

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

Obr. 1: *Zarovnanie dvoch sekvencií. V zarovnaní sa nachádzajú zhody, nezhody a medzery v oboch sekvenciách*

Klasifikácia na základe lokálnej informácie

Anotácie sme zakomponovali pomocou klasifikátorov, ktoré rozhodujú či dané pozície majú byť zarovnané k sebe alebo nie. Ako klasifikátor sme použili *náhodný les (angl. Random forest)* [Breiman, 2001], pretože aktuálne patrí medzi najlepšie klasifikátory.

- Výstup je hodnota z intervalu $\langle 0, 1 \rangle$, ktorá označuje istotu klasifikátora, že dané dve pozície majú byť zarovnané k sebe (v insert stave, že daná pozícia má byť zarovnaná k medzere).
- Atribúty sú okná veľkosti w (obr. 2), v ktorých sa nachádza w dvojíc báz v okolí daných pozícií a ich anotácie (napr. či ide o gén alebo nie).

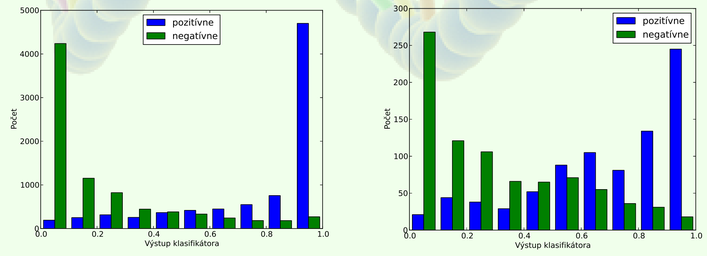
```
 i:012345678  9      i:012345678  9
Ax:0001111111  0    Ax:0001111111  0
 X:ACCATTCTA--C      X:ACCATTCTA--C
 Y:ACG----TGTTC      Y:ACG----TGTTC
Ay:000    11111    Ay:000    11111
 j:012    34567      j:012    34567
```

(a) Match klasifikátor (b) Indel klasifikátor

Obr. 2: *Okno klasifikátora pre pozície $i = 6$ a $j = 3$*

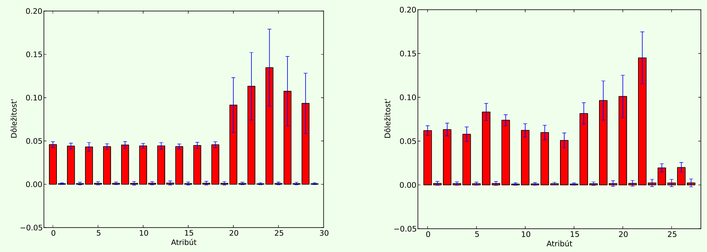
- K dátam z okna sme pridali informácie o zhodách na zodpovedajúcich pozíciách, čím sa nám podarilo vylepšiť úspešnosť klasifikátora.
- Úspešnosť Match klasifikátora: **84,32%**
- Úspešnosť Indel klasifikátora: **76,46%**

Ukázalo sa teda, že klasifikátor sa dokáže naučiť, ktoré okná majú byť zarovnané k sebe a ktoré nie (Obr. 3).



(a) Match klasifikátor (b) Indel klasifikátor

Obr. 3: *Distribúcia výstupu klasifikátora pre pozitívne a negatívne príklady.*



(a) Match klasifikátor (b) Indel klasifikátor

Obr. 4: *Dôležitosť atribútov v klasifikátore. Na párnych pozíciach sú bázy, na nepárnych anotácia. Prvých 10 atribútov zodpovedá oknu v prvej sekvencii, druhých 10 (resp. 8 v Indel klasifikátore) zodpovedá oknu v druhej sekvencii a posledných 10 zodpovedá zhodám na príslušných pozíciách (v Indel klasifikátore sa namiesto medzery zopakuje báze čo je za ňou)*

Zakomponovanie výsledkov klasifikácie do pHMM

Vyvinuli sme dva modely pre zarovnanie sekvencií s anotáciami za pomoci klasifikátora, ktoré sú založené na párových skrytých Markovovských modeloch.

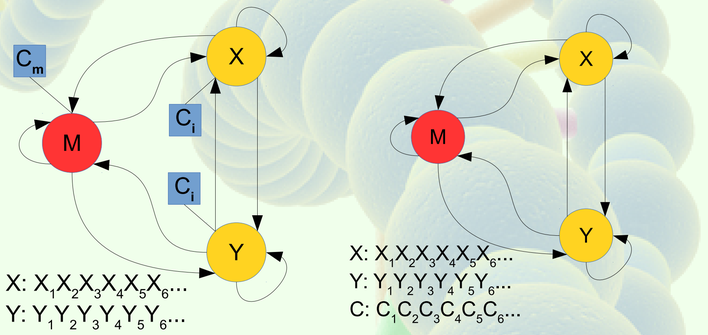
Model s klasifikátorom ako emisiou (Obr. 5(a)):

- Emisné tabuľky stavov nahradíme výstupom z klasifikátora.
- Model nie je korektný pravdepodobnostný model, pretože pravdepodobnosti emisií nesčitujú do 1.
- Prechodové pravdepodobnosti sme natrénovali zo zarovnaní z trénovacej vzorky.

Model s klasifikátorovou páskou (Obr. 5(b)):

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

Keďže stále ide o párový HMM, pásku si musíme predstaviť ako cestu v 2D tabuľke výstupov klasifikátorov, ktorá sa zhoduje s cestou zarovnania. Teda ak sa pohneme horizontálne alebo vertikálne, používame Indel klasifikátor a ak sa pohneme diagonálne, tak použijeme Match klasifikátor.



(a) Model s klasifikátorom ako emisiou (b) Model s klasifikátorovou páskou

Obr. 5: *Modely s klasifikátorom*

Experimenty

Dáta	Simulované 1		Simulované 2	
Model	Zhoda	Tranz.	Zhoda	Tranz.
Ref. model	85,78%	61,03%	6,78%	8,87%
Model A	73,95%	38,10%	72,62%	39,48%
Model A bez an.	76,07%	45,63%	—	—
Model B	81,62%	54,94%	18,57%	10,50%
Model B bez an.	81,22%	54,18%	—	—

Tabuľka 1: *Porovnanie úspešností modelov. Zhora dole - referenčný model (obyčajný pHMM na zarovnávanie DNA sekvencií), model s klasifikátorom ako emisiou s anotáciou a bez nej, model s klasifikátorovou páskou s anotáciou a bez nej. Zhoda je počítaná ako percentuálna zhoda originálneho a nového zarovnania. Tranzitivita je počítaná z troch zarovnaní AB, BC a AC ako percentuálna zhoda medzi $AB \circ BC$ a AC. Simulované dáta 1 sa snažia napodobňovať biologické procesy, Simulované dáta 2 nezodpovedajú biologickým dátam a sú zložitejšie na natrénovanie pre referenčný model.*

Experimenty ukázali, že pri jednoduchších dátach bol obyčajný pHMM postačujúci a dosiahol lepšie skóre ako naše modely, ktoré však nezaostávali príliš. Pri zložitejších dátach už obyčajný pHMM nestačil a ukázala sa sila diskriminačného prístupu v modeli s klasifikátorom ako emisiou. Taktiež si môžme všimnúť, že pri simulovaných dátach 1 anotácia našim modelom vôbec nepomohla. Pri dátach 2 bola anotácia nutná na správne zarovnanie.

Literatúra

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

[Durbin et al., 1998] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

[Hozza, 2014] Hozza, M. (2014). Zarovnávanie sekvencií s použitím metód klasifikácie. Katedra Informatiky, Fakulta matematiky, fyziky a informatiky,Univerzita Komenského, Bratislava.