

Zarovňávanie DNA sekvencií s použitím metód klasifikácie

Bc. Michal Hozza

Školiteľ: Mgr. Tomáš Vinař, PhD., Mgr. Michal Nánási

Študentská vedecká konferencia
Fakulta matematiky, fyziky a informatiky, Univerzita Komenského, Bratislava

23.4.2014

Obsah

Zarovňávanie sekvencií

Klasifikácia na základe lokálnej informácie

Zakomponovanie výsledkov klasifikácie do pHMM

Výsledky

Zarovňavanie 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).

Definícia (Globálne zarovnanie)

Vstupom sú dve sekvencie $X = x_1x_2 \dots x_n$ a $Y = y_1y_2 \dots y_m$
 Výstupom je zarovnanie celých sekvencií X a Y .

Príklad:

```
GTGGACCGTT-----CCTTCCGGCAATCACGAGAAAAGCCACGT
GTCGACCGTTTTCAGTGAAGCAATCAGG---AACACCACCT
```

Skryté Markovovské modely

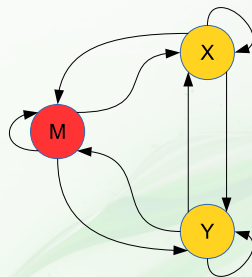
- ▶ pravdepodobnostný model inšpirovaný konečnými automatmi
- ▶ pozostáva z 3 distribúcií
 - ▶ distribúcia začiatočných stavov
 - ▶ distribúcia prechodov
 - ▶ distribúcia emisií
- ▶ pravdepodobnosť, že model vygeneruje sekvenciu x dĺžky n s anotáciou s je súčin pravdepodobností prechodov a emisií.

$$P[X = x | S = s] = \pi_{s_1} e_{s_1, x_1} a_{s_1, s_2} e_{s_2, x_2} a_{s_2, s_3} e_{s_3, x_3} \dots a_{s_{n-1}, s_n} e_{s_n, x_n}$$

[Brejová and Vinař, 2011, Durbin et al., 1998]

Zarovňávanie sekvencií pomocou pHMM

- ▶ 3 stavový pHMM
 - ▶ Match – emituje dvojice:
 AA, AC, AG, \dots, TT
 - ▶ Insert X, Insert Y – emitujú jednotlivé bázy A, C, G, T
- ▶ Prechodové a emisné pravdepodobnosti sa trénujú pomocou frekvenčnej tabuľky
- ▶ Najpravdepodobnejšie zarovnanie nájdeme pomocou Viterbiho algoritmu



Inverzné zarovnanie

Definícia (Problém inverzného zarovnaní)

Vstupom sú dve sekvencie X , Y a ich zarovnanie Z . Výstupom sú parametre, podľa ktorých je toto zarovnanie optimálne.

- pod parametrami rozumieme skórovací systém – napr. skórovaciu maticu alebo skrytý markvovský model

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
- ▶ Použili sme *náhodný les* (angl. *Random forest*) [Breiman, 2001]
- ▶ Výstup $\in \langle 0, 1 \rangle$ – istota 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

Okno pre klasifikátor

```

i:012345678 9
Ax:000111111 0
X:ACCATTCCTA--C
Y:ACG-----TGTTC
Ay:000111111
j:012 34567

```

(a) Match klasifikátor

```

i:012345678 9
Ax:000111111 0
X:ACCATTCCTA--C
Y:ACG-----TGTTC
Ay:000111111
j:012 34567

```

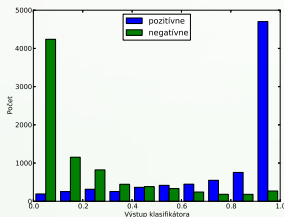
(b) InDel klasifikátor

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

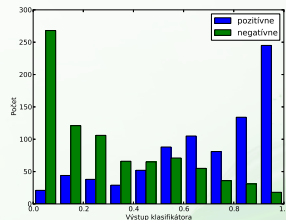
Klasifikácia na základe lokálnej informácie

- ▶ K dátam sme pridali informácie o zhodách na zodpovedajúcich pozíciách, čím sa nám podarilo vylepšiť úspešnosť klasifikátora
- ▶ Úspešnosť Match klasifikátora: **84,32%**
- ▶ Úspešnosť Indel klasifikátora: **76,46%**
- ▶ Klasifikátor sa dokáže naučiť, ktoré okná majú byť zarovnané k sebe a ktoré nie

Úspešnosť klasifikátora



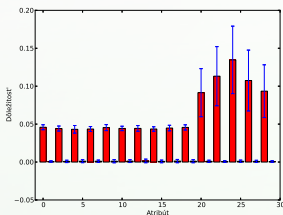
(a) Match klasifikátor



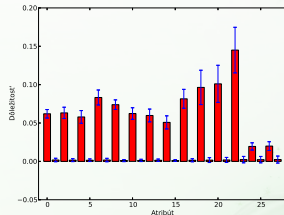
(b) InDel klasifikátor

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

Dôležitosť atribútov



(a) Match klasifikátor



(b) InDel klasifikátor

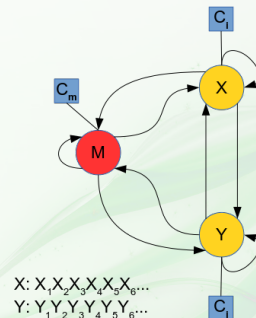
Obr. : Dôležitosť atribútov v klasifikátore. Na párnych pozíciach sú bázy, na nepárnych anotácia. Prvých 10 atribútov zodpovedá oknu v prvej sekvencii, druhých 10 (resp. 8 v InDel klasifikátore) zodpovedá oknu v druhej sekvencii a posledných 10 zodpovedá zhodám na príslušných pozíciách (v InDel klasifikátore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

Zakomponovanie výsledkov klasifikácie do pHMM

- ▶ Vymysleli 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.

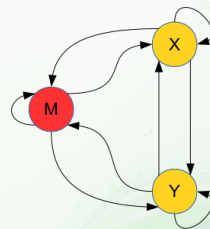
Model s klasifikátorom ako emisiou

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



Model s klasifikátorovou páskou

- ▶ 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
- ▶ Zhoduje sa s cestou zarovnanania
- ▶ Ak sa pohneme horizontálne, alebo vertikálne, používame Indel klasifikátor a ak sa pohneme diagonálne, tak použijeme Match klasifikátor



X: $X_1 X_2 X_3 X_4 X_5 X_6 \dots$

Y: $Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 \dots$

C: $C_1 C_2 C_3 C_4 C_5 C_6 \dots$

Výsledky

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 : Porovnanie úspešností modelov. Zhoda je počítaná ako percentuálna zhoda originálneho a nového zarovňania. Tranzitivita je počítaná z troch zarovnaní AB , BC a AC ako percentuálna zhoda medzi $AB \circ BC$ a AC .

Ďakujem za pozornosť!

Literatúra



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



Brejová, B. and Vinař, T. (2011).
Metódy v bioinformatike [Methods in Bioinformatics].
Knižničné a edičné centrum FMFI UK.
Lecture notes.



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