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

10. júna 2014

Obsah

Cieľ

- 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

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)

Definícia (Globálne zarovnanie)

Vstupom sú dve sekvencie $X = x_1 x_2 \ldots x_n$ a $Y = y_1 y_2 \ldots y_m$
Výstupom je zarovnanie celých sekvencií X a Y.

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)

Definícia (Globálne zarovnanie)

Vstupom sú dve sekvencie $X = x_1 x_2 \dots x_n$ a $Y = y_1 y_2 \dots y_m$
Výstupom je zarovnanie celých sekvencií X a Y.

Príklad:

```
GTGGACCGTT------CCTTCCGGCAATCACGACAAAAGCCACGT
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)

Definícia (Globálne zarovnanie)

Vstupom sú dve sekvencie $X = x_1 x_2 \ldots x_n$ a $Y = y_1 y_2 \ldots y_m$
Výstupom je zarovnanie celých sekvencií X a Y.

Príklad:

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

Skryté Markovovské modely

- pravdepodobnostný model inšpirovaný konečnými automatmi
- pozostáva z 3 distribúcií
  - distribúcia začiatočných stavov (π_i)
  - distribúcia prechodov ($a_{i,j}$)
  - distribúcia emisií ($e_{i,x}$)
- pravdepodobnosť, že model vygeneruje sekvenciu x dĺžky n s anotáciou s je súčin pravdepodobností prechodov a emisií.

$$P(X = x|S = s) = \pi_{s_1} \left(\prod_{i=1}^{n-1} e_{s_i,x_i} a_{s_i,s_{i+1}} \right) e_{s_n,x_n}$$

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

Skryté Markovovské modely

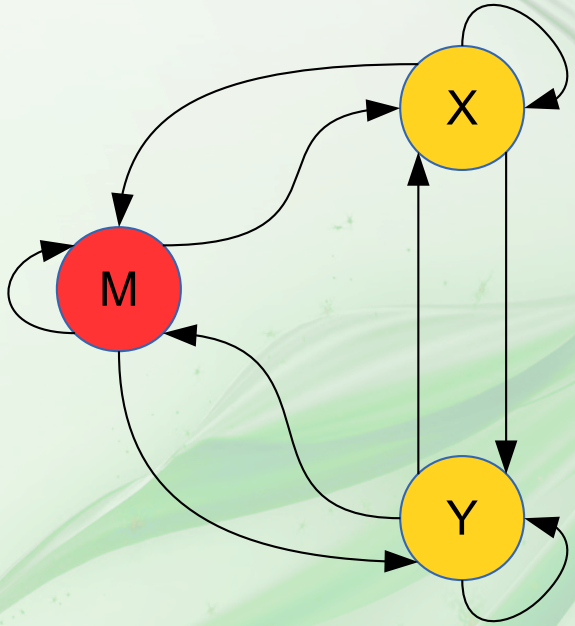
- pravdepodobnostný model inšpirovaný konečnými automatmi
- pozostáva z 3 distribúcií
  - distribúcia začiatočných stavov (π_i)
  - distribúcia prechodov ($a_{i,j}$)
  - distribúcia emisií ($e_{i,x}$)
- pravdepodobnosť, že model vygeneruje sekvenciu x dĺžky n s anotáciou s je súčin pravdepodobností prechodov a emisií.

$$P(X = x|S = s) = \pi_{s_1} \left(\prod_{i=1}^{n-1} e_{s_i,x_i} a_{s_i,s_{i+1}} \right) e_{s_n,x_n}$$

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

Zarovnávanie sekvencií pomocou pHMM

- 3 stavový pHMM
  - Match – emituje dvojice: $AA, AC, AG, \ldots, TT$
  - 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



Existujúce metódy zarovnávania sekvencií s dodatočnou informáciou

Definícia (Problém inverzného zarovnania)

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

- doterajšie publikácie boli o zarovnávaní proteínových sekvencií
- doterajší výskum sa zaoberal havne riešením problému IZ
- v [Yu et al., 2007] využili SVM na nájdenie parametrov a použili anotáciu o štruktúre sekvencií
- podobný prístup zvolili v [Do et al., 2006], kde na nájdenie parametrov využili CRF
- obe riešenia ťažia z výhod diskriminatívneho učenia, kde hľadáme $P(Z|W)$ (namiesto $P(Z, W)$, ktoré hľadáme pri generatívnych modeloch)

Existujúce metódy zarovnávania sekvencií s dodatočnou informáciou

Definícia (Problém inverzného zarovnania)

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

- doterajšie publikácie boli o zarovnávaní proteínových sekvencií
- doterajší výskum sa zaoberal havne riešením problému IZ
- v [Yu et al., 2007] využili SVM na nájdenie parametrov a použili anotáciu o štruktúre sekvencií
- podobný prístup zvolili v [Do et al., 2006], kde na nájdenie parametrov využili CRF
- obe riešenia ťažia z výhod diskriminatívneho učenia, kde hľadáme $P(Z|W)$ (namiesto $P(Z, W)$, ktoré hľadáme pri generatívnych modeloch)

Existujúce metódy zarovnávania sekvencií s dodatočnou informáciou

Definícia (Problém inverzného zarovnania)

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

- doterajšie publikácie boli o zarovnávaní proteínových sekvencií
- doterajší výskum sa zaoberal havne riešením problému IZ
- v [Yu et al., 2007] využili SVM na nájdenie parametrov a použili anotáciu o štruktúre sekvencií
- podobný prístup zvolili v [Do et al., 2006], kde na nájdenie parametrov využili CRF
- obe riešenia ťažia z výhod diskriminatívneho učenia, kde hľadáme $P(Z|W)$ (namiesto $P(Z, W)$, ktoré hľadáme pri generatívnych modeloch)

Naše riešenie

- zarovnávanie s dodatočnou informáciou
- dodatočnú informáciu sme zakomponovali pomocou klasifikátorov
  - klasifikátory na základe lokálnej informácie rozhodujú, či dané pozície majú byť zarovnané k sebe
  - natrénujeme ich na existujúcich zarovnaniach
- klasifikátory sme zakomponovali do pHMM pre zarovnávanie

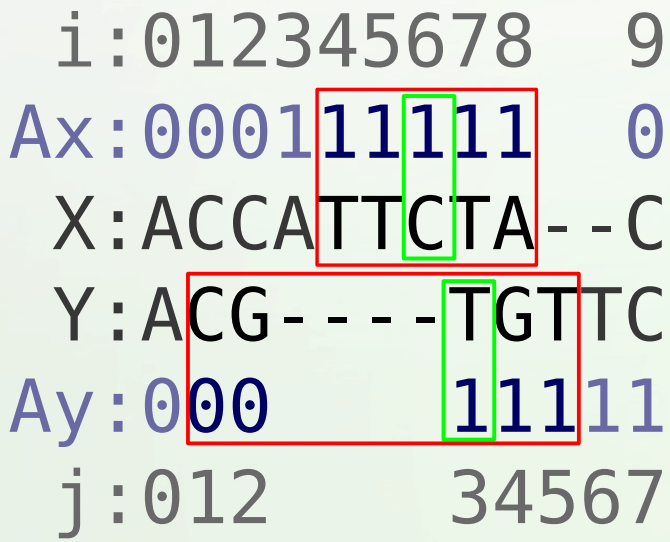
Naše riešenie – odlišnosti

- zaoberáme sa DNA sekvenciami, nie proteínmi
- nemáme k dispozícii sekundárnu štruktúru, ale iný typ anotácií
- naše modely sú kombináciou generatívneho a diskriminačného prístupu
- naše modely sú založené na pHMM
- používame dva rôzne klasifikátory
- architektúry našich modelov abstrahujú od klasifikátora
- ako klasifikátor sme použili *náhodný les (angl. Random forest)* [Breiman, 2001]

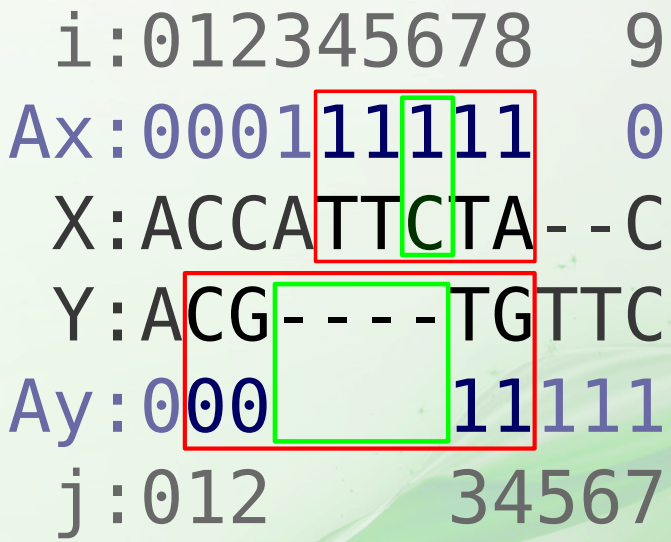
Klasifikácia na základe lokálnej informácie

- 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 insert stave, že daná pozícia má byť zarovnaná k medzere)
- atribúty sú okná veľkosti w

Okno pre klasifikátor



(a) Match klasifikátor



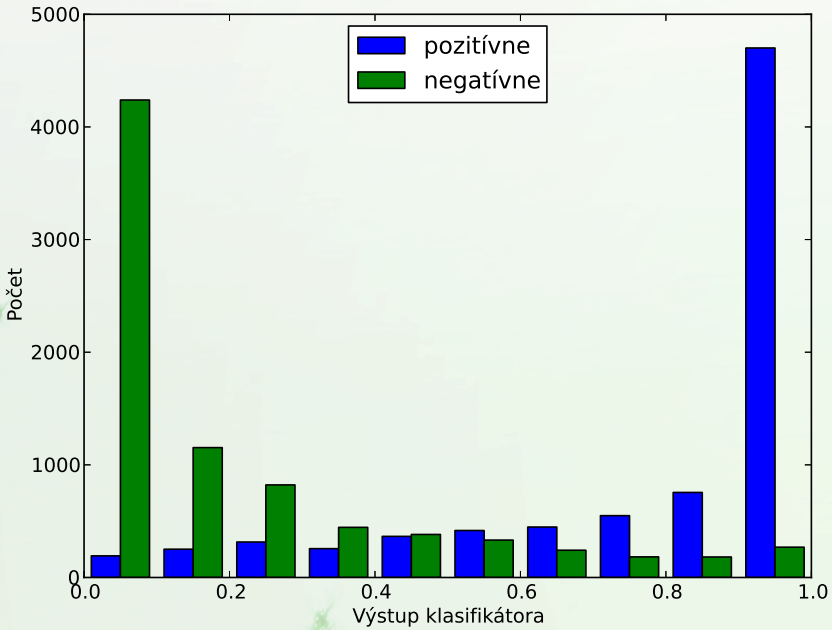
(b) InDel klasifikátor

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

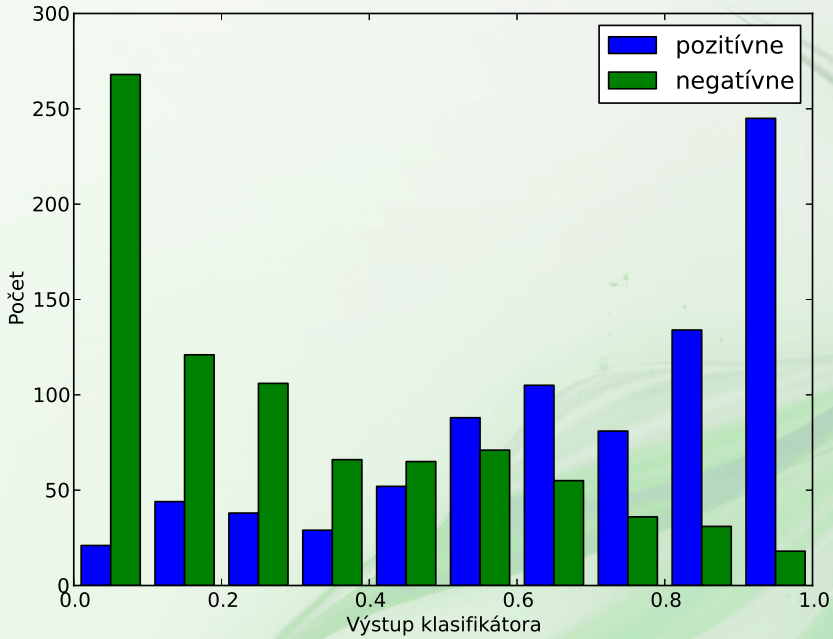
Klasifikácia na základe lokálnej informácie

- 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



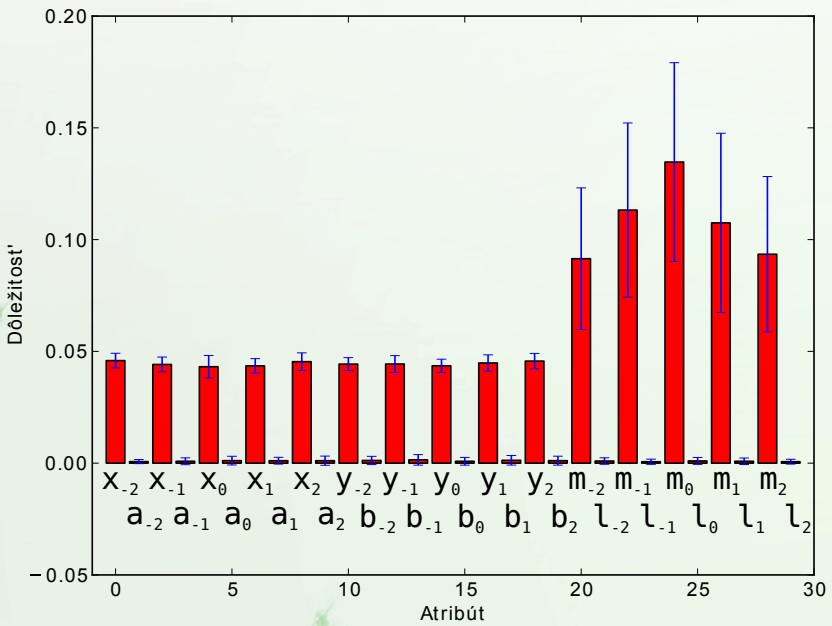
(a) Match klasifikátor



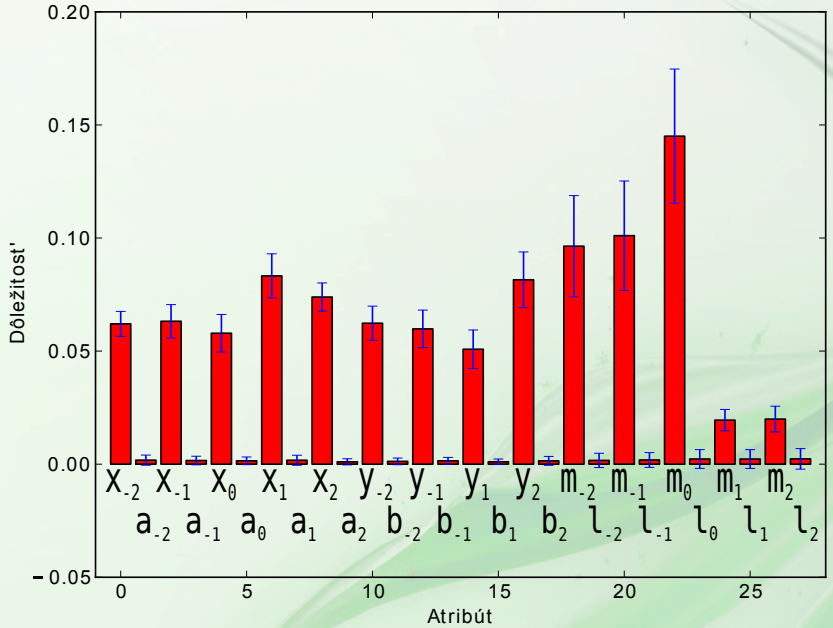
(b) InDel klasifikátor

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

Dôležitosť atribútov



(a) Match klasifikátor



(b) InDel klasifikátor

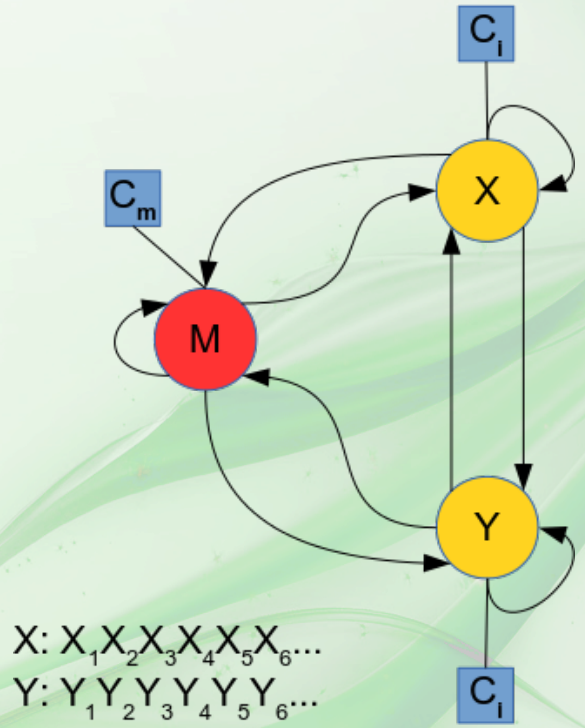
Obr. : Dôležitosť atribútov v klasifikátore.

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.

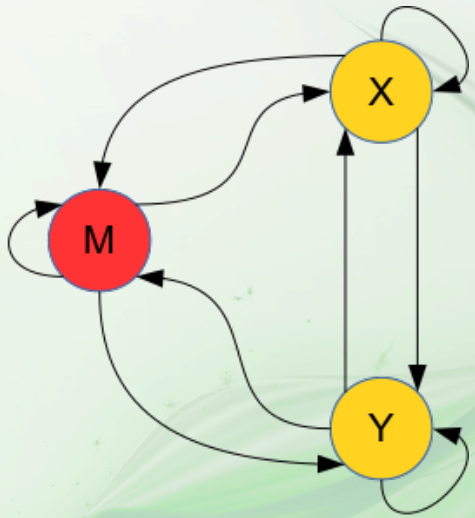
Model s klasifikátorom ako emisiou (Model A)

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



Model s klasifikátorovou páskou (Model B)

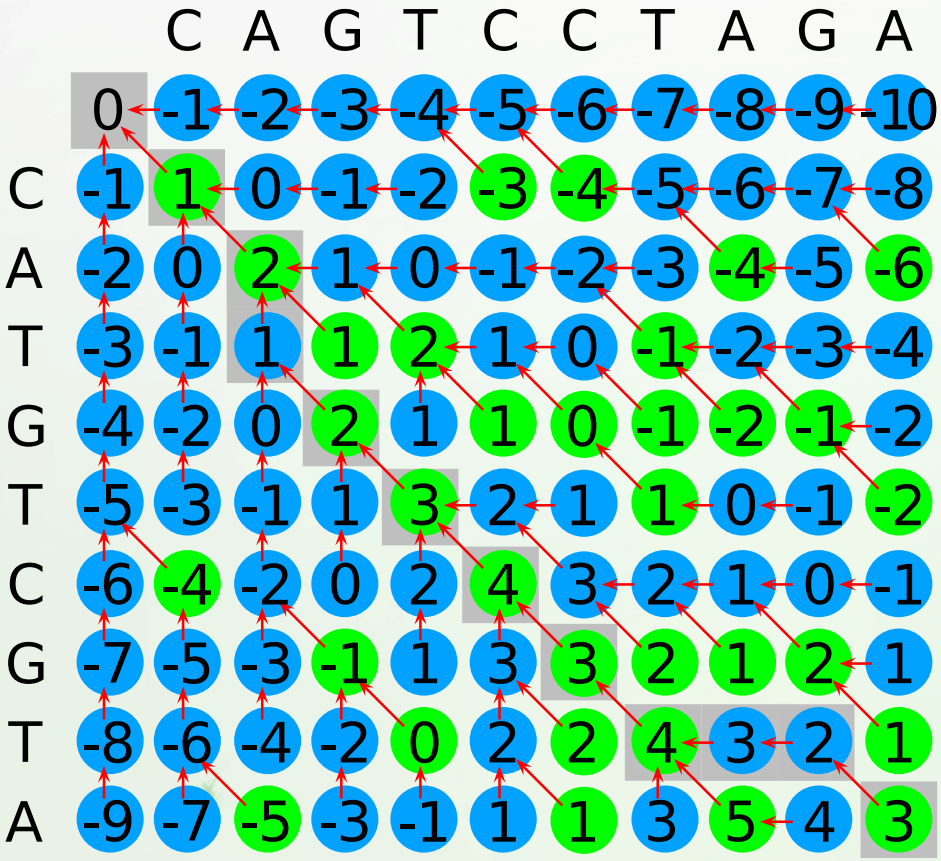
- modelujeme navyše sekvenciu výstupov klasifikátora vo forme pásky
- trénujeme všetky parametre na trénovacej vzorke zarovnaní obohatenej o pásku s výstupmi z klasifikátora



- 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




```
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.

Ďakujem za pozornosť!

https://github.com/mhozza/realigner

Literatúra I


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.


Do, C., Gross, S., and Batzoglou, S. (2006).
Contralign: Discriminative training for protein sequence alignment.
In Apostolico, A., Guerra, C., Istrail, S., Pevzner, P., and Waterman, M., editors, *Research in Computational Molecular*

Literatúra II

Biology, volume 3909 of *Lecture Notes in Computer Science*, pages 160–174. Springer Berlin / Heidelberg. 10.1007/11732990_15.

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

Yu, C.-N., Joachims, T., Elber, R., and Pillardy, J. (2007). Support vector training of protein alignment models. In Speed, T. and Huang, H., editors, *Research in Computational Molecular Biology*, volume 4453 of *Lecture Notes in Computer Science*, pages 253–267. Springer Berlin / Heidelberg. 10.1007/978-3-540-71681-5_18.