

# Zarovňávanie 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

5. februára 2014

# Obsah

## Úvod

Cieľ

Zarovnávanie sekvencií

## Modely a Existujúce riešenia

Modely

Príbuzné témy

## Naše riešenia

Odlišnosti nášho riešenia

Simulátor

Modely

## Výsledky

Metódy vyhodnocovania

Simulované dáta

# 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.

# Zarovňavanie sekvencií

Kľúčové problémy:

- ▶ Aké typy zarovňavania by sme mali uvažovať
- ▶ Skórovací systém, ktorý použijeme na ohodnotenie zarovňania a tréňovanie
- ▶ Algoritmus, ktorý použijeme na hľadanie optimálneho alebo dobrého zarovňania podľa skórovacieho systému
- ▶ Štatistická významnosť zarovňania.

# Modely

## Generatívny:

- ▶ sa snaží modelovať proces, ktorý generuje dáta ako pravdepodobnosť  $P(X, Y, Z)$
- ▶ rozložíme ju pomocou nezávislých predpokladov na procese  $\rightarrow$  obmedzujúce

## Diskriminačný

- ▶ priamo odhaduje  $P(Z|X, Y)$  alebo prislúchajúcu diskriminačnú funkciu, a preto sa zamerá na podstatnú časť problému odhadu
- ▶ Nepotrebuje nezávislosť  $\rightarrow$  silnejšie

# Príbuzné témy (existujúce riešenia)

- ▶ Problém inverzného zarovnania
- ▶ Support vector training of protein alignment models
  - ▶ Support Vector Machine (SVM)
  - ▶ Umožňuje trénovať pomocou rôznych účelových funkcií
- ▶ Contralign: Discriminative training for protein sequence alignment.
  - ▶ Conditional Random Fields (CRF)
  - ▶ Neumožňuje trénovať pomocou rôznych účových funkcií

# Odlišnosti nášho riešenia

- ▶ Korekcia existujúcich zarovnaní
  - ▶ Použitie súbežne s existujúcimi zarovnávačmi
- ▶ Rôzne modely využitia klasifikátora
- ▶ Rôzne metódy trénovania
- ▶ Možnosť učenia bez učiteľa
- ▶ Iný klasifikátor
  - ▶ možno porovnanie viac rôznych klasifikátorov
  - ▶ prípadne abstrakcia od klasifikátora

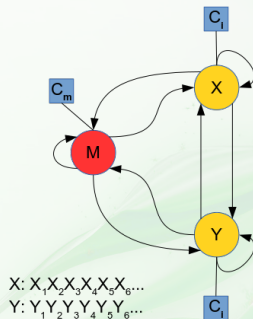
# Simulátor

- ▶ Model určený na prvotné experimenty
- ▶ Program simuluje evolúciu
  - ▶ Generovanie dvojice postupností so správnym zarovnaním
  - ▶ Generovanie dodatočej informácie
  - ▶ Simulácia mutácie a delécie



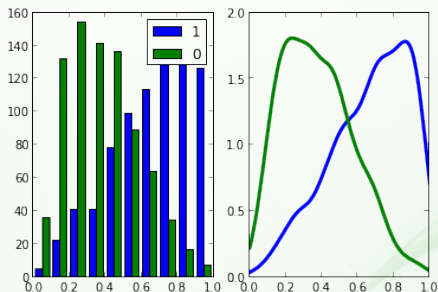
# Základný model

- ▶ 3 stavový HMM
  - ▶ Match
  - ▶ Insert X, Insert Y
- ▶ Klasifikátor vidí okolie báz rozšírené o anotácie
- ▶ V HMM sa trénujú iba tranzície, klasifikátory sa trénujú zvlášť
- ▶ Viterbiho algoritmus
  - ▶ namiesto tabuľky emisných pravdepodobností, máme výstup z natreňovaného klasifikátora
  - ▶ Problém výstupy z klasifikátora nesumujú do 1, čiže model nie je celkom korektný



# Distribúcia výstupu z klasifikátora

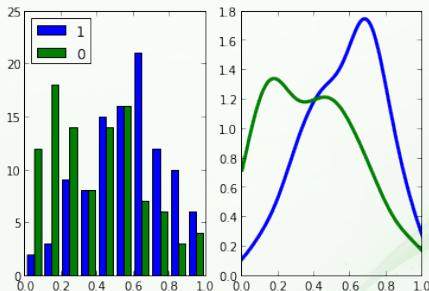
## Match stav



**Obr. :** Distribúcia výstupu klasifikátora pre zarovnané (modrá) a nezarovnané (zelená) pozície. Klasifikátor pre match stav s anotáciou a oknom veľkosti 5 (testovacia množina)

# Distribúcia výstupu z klasifikátora

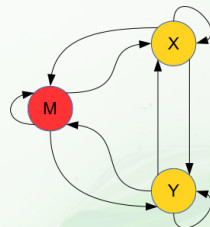
Insert stav



**Obr. :** Distribúcia výstupu klasifikátora pre zarovnané pozície (zelená) a pozície zarovnané k medzere (modrá). Klasifikátor pre insert stav s anotáciou a oknom veľkosti 5 (testovacia množina)

# Model s klasifikátorovou páskou

- ▶ Opäť rovnaký 3 stavový HMM
- ▶ Okrem 2 sekvencií máme ešte pásku s výstupom z klasifikátora
- ▶ Model teda emituje trojicu - dve písmená zo sekvencií (alebo jedno a pomlčku) a výstup z klasifikátora
- ▶ Tento model je narozdiel od predošlého korektný
- ▶ Trénujú sa aj prechodové aj emisné pravdepodobnosti
- ▶ Pre jednoduchosť budeme emisie výstupu klasifikátora aproximovať pomocou normálneho rozdelenia s natrénovanými parametrami



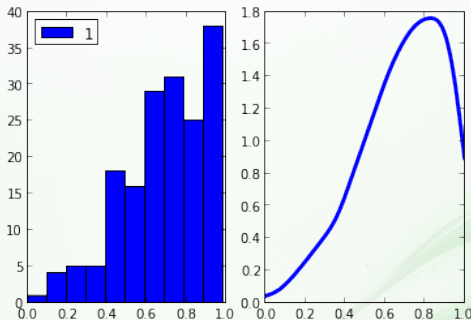
X:  $X_1 X_2 X_3 X_4 X_5 X_6 \dots$

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

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

# Distribúcia výstupu z klasifikátora

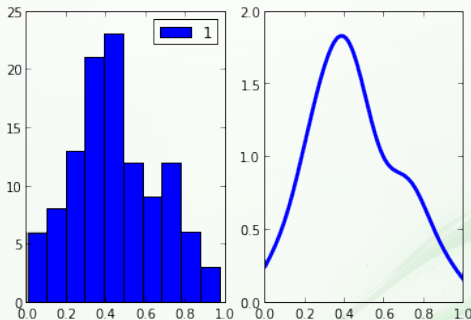
Match stav - AA



**Obr. :** Distribúcia výstupu klasifikátora pre match stav v prípade báz AA zarovnaných k sebe. Klasifikátor s anotáciou a oknom veľkosti 5 (testovacia množina)

# Distribúcia výstupu z klasifikátora

Match stav - AC



**Obr. :** Distribúcia výstupu klasifikátora pre match stav v prípade báz AC zarovnaných k sebe. Klasifikátor s anotáciou a oknom veľkosti 5 (testovacia množina)

# Metódy vyhodnocovania

## Kontrola tranzitivity

Ako základnú mieru úspešnosti nášho algoritmu budeme brať kontrolu tranzitivity

- ▶ Použijeme 3 párové zarovnania 3 sekvencií (každá s každou)
- ▶ Spojíme prvé 2 zarovnania do nového zarovnania
- ▶ Porovnáme percentuálne zhody nového s tretím zarovnaním

# Dodatočné informácie

- ▶ Stopa s informáciou, či daná pozícia je súčasťou génu
- ▶ Ukazuje sa, že dodatočné informácie dokážu pomôcť klasifikátoru v lepšej klasifikácii



# Výsledky

- ▶ Aktuálne len na simulovaných dátach
- ▶ Kontrola tranzitivity:
  - ▶ Referenčný model (3-stavový HMM bez klasifikátora)
    - 40%
  - ▶ Náš základný model (s 1 anotáciou a oknom veľkosti 1)
    - **46%**
  - ▶ Náš základný model (s 1 anotáciou a oknom veľkosti 5)
    - **46%**
  - ▶ Model s klasifikátorovou páskou – ešte nie je dokončený.

# Najbližšie plány

- ▶ Dorobiť ostatné modely
- ▶ Urobiť experimenty na reálnych dátach
- ▶ Experimenty s deravým oknom, poprípade iné
- ▶ V prvom modeli CRF namiesto HMM?

**Ďakujem za pozornosť!**