

Zarovňávanie sekvencií s použitím metód klasifikácie

Diplomová práca

Bc. Michal Hozza

Vedúci práce: Mgr. Tomáš Vinár, PhD.

Konzultant: Mgr. Michal Nanási

Fakulta matematiky, fyziky a informatiky, Univerzita Komenského, Bratislava

20. marca 2013

Obsah

Úvod

Cieľ

Zarovnávanie sekvencií

Modely a Existujúce riešenia

Modely

Existujúce riešenia

Naše riešenie

Odlišnosti nášho riešenia

Klasifikátor

Simulátor

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:

Zarovňavanie sekvencií

Kľúčové problémy:

- ▶ Aké typy zarovňavania by sme mali uvažovať

Zarovňavanie sekvencií

Kľúčové problémy:

- ▶ Aké typy zarovňavania by sme mali uvažovať
- ▶ Skórovací systém, ktorý použijeme na ohodnotenie zarovňania a tréňovanie

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

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 \longrightarrow obmedzujúce

Modely

Generatívny:

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

Diskriminačný

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

Existujúce riešenia

- Problém inverzného zarovnania

Existujúce riešenia

- ▶ Problém inverzného zarovnania
- ▶ Support vector training of protein alignment models
 - ▶ Support Vector Machine (SVM)
 - ▶ Umožňuje trénovať pomocou rôznych účelových funkcií

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

Odlišnosti nášho riešenia

- ▶ Korekcia existujúcich zarovnaní
 - ▶ Použitie súbežne s existujúcimi zarovnávačmi
- ▶ Rôzne metódy trénovania

Odlišnosti nášho riešenia

- ▶ Korekcia existujúcich zarovnaní
 - ▶ Použitie súbežne s existujúcimi zarovnávačmi
- ▶ Rôzne metódy trénovania
- ▶ Možnosť učenia bez učiteľa

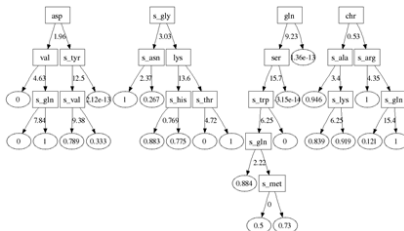
Odlišnosti nášho riešenia

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

Klasifikátor

Random Forest

- Zložený z klasifikačných (rozhodovacích) stromov
- Stromy hlasujú o výsledku ktoré hlasujú



Simulátor

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

Ďakujem za pozornosť!