Zarovnávanie sekvencií s použitím metód klasifikácie

Bc. Michal Hozza

Školiteľ: Mgr. Tomáš Vinař PhD.
Konzultant: Mgr. Michal Nanási

Cieľ

- Cieľom práce je vytvoriť nové metóty na korekciu zarovnaní biologických sekvencií na základe prídavnej informácie.
- Integrácia tejto informácie bude zabezpečená pomocou techník využívaných na klasifikáciu v strojovom učení.

Zarovnávanie sekvencií

- Kľúčové problémy:
  - Aké typy zarovnávania by sme mali uvažovať
  - Skórovací systém, ktorý použijeme na ohodnotenie zarovnania a trénovanie
  - Algoritmus, ktorý použijeme na hľadanie optimálneho alebo dobrého zarovnania podľa skórovacieho systému
  - Štatistická významnosť zarovnania.

Generatívny vs. diskriminačný model

- Generatívny:
  - sa snaží modelovať proces, ktorý generuje dáta ako pravdepodobnosť P (X, Y, Z)
  - rozložíme ju pomocou nezávislých predpokladov na procese → obmedzujúce
- Diskriminačný
  - priamo odhaduje P(Z|X,Y) alebo prislúchajúcu diskriminačnú funkciu, a preto sa zamerá na podstatnú časť problému odhadu
  - Nepotrebuje nezávislosť → silnejšie

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 Rnadom Fields (CRF)
  - Neumožnuje trénovať pomocou rôznych účových funkcií

Odlišnosti nášho riešenia

- Rôzne metódy trénovania
- Možnosť učenia bez učiteľa
- Iný klasifikátor (možno viac rôznych klasifikátorov, prípadne abstrakcia od klasifikátora)

Random Forest

- Klasifikátor
- Zložený z klasifikačných (rozhodovacích) stromov, ktoré hlasujú

Simulátor

- Program, ktorý simuluje evolúciu
- Model určený na prvotné experimenty