Zarovnávanie sekvencií s použitím metód klasifikácie (rozšírený abstrakt)

Michal Hozza*

Školiteľ: Tomáš Vinař[1†], Michal Nánási[2‡]

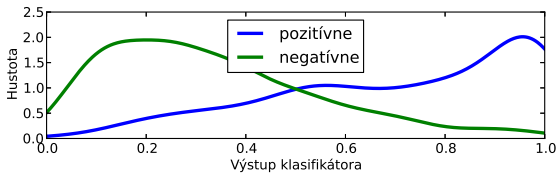
[1] Katedra aplikovanej informatiky, FMFI UK, Mlynská Dolina 842 48 Bratislava
[2] Katedra informatiky, FMFI UK, Mlynská Dolina 842 48 Bratislava

Zarovnávanie dvoch DNA sekvencií je jedným zo základných bioinformatických problémov. Správne zarovnanie identifikuje časti sekvencie, ktoré vznikli z toho istého predka (zarovnané bázy), ako aj inzercie a delécie v priebehu evolúcie (medzery v zarovnaní). Obvykle takéto zarovnanie hľadáme pomocou jednoduchých párových skrytých Markovovských modelov (pHMM) [Durbin et al., 1998]. V tejto práci sa zaoberáme možnosťami použitia prídavnej informácie o funkcii vstupných sekvencií (tzv. anotácie) na zlepšenie kvality takýchto zarovnaní.

Klasifikácia na základe lokálnej informácie. Informácie sme zakomponovali pomocou klasifikátorov, ktoré rozhodujú či dané pozície majú byť zarovnané k sebe alebo nie. Ako klasifikátor sme použili *RandomForest* [Breiman, 2001], pretože aktuálne patrí medzi najlepšie klasifikátory.

Vstupné dáta pre klasifikátor sú okná veľkosti w, v ktorom sa nachádza w dvojíc báz v okolí daných pozícií a ich anotácie (napr. či ide o gén alebo nie). Výstup je hodnota z intervalu $\langle 0,1 \rangle$, ktorá označuje istotu klasifikátora, že dané dve pozície majú byť zarovnané k sebe (v insert stave, že daná pozícia má byť zarovnaná k medzere).

Ukázalo sa, že klasifikátor sa dokáže naučiť, ktoré okná majú byť zarovnané k sebe a ktoré nie (Obr. 1).



Obr. 1: Distribúcia výstupu klasifikátora pre pozitívne a negatívne príklady. Pozitívne príklady sú tie, ktoré majú byť zarovnané k sebe, negatívne príklady sú tie, ktoré k sebe zarovnané byť nemajú.

*Michal.Hozza@ksp.sk
†vinar@fmph.uniba.sk
‡mic@compbio.fmph.uniba.sk

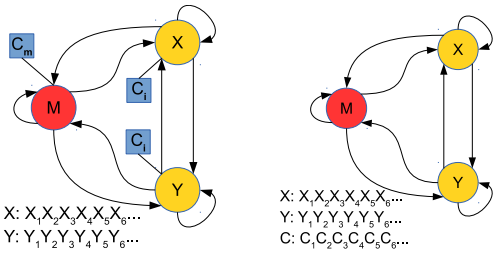
Zakomponovanie výsledkov klasifikácie do pHMM. Vyvinuli sme 2 modely pre zarovnanie sekvencií s anotáciami za pomoci klasifikátora, ktoré sú založené na párových skrytých Markovovských modeloch.

Model s klasifikátorom ako emisiou: (Obr. 2a) V tomto modeli sme nahradili emisné tabuľky stavov výstupom z klasifikátora. Model však nie je korektný pravdepodobnostný model, pretože pravdepodobnosti emisií nesčitujú do 1.

Emisné pravdepodobnosti sme prevzali priamo z klasifikátora, ktorý trénujeme zvlášť. Prechodové pravdepodobnosti sme natrénovali zo zarovnaní z trénovacej vzorky.

Model s klasifikátorovou páskou: (Obr. 2b) Aby sme vyriešili problém s korektnosťou predošlého modelu, navrhli sme alternatívu, ktorá navyše modeluje aj výstup z klasifikátora. Nemodelujeme teda len dvojicu sekvencií, ale aj sekvenciu výstupov klasifikátora.

V tomto modeli sme trénovali všetky parametre na trénovacej vzorke zarovnaní obohatenej o pásku s výstupmi z klasifikátora.



Obr. 2: Modely s klasifikátorom

Literatúra

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

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