Zarovnávanie sekvencií s použitím metód klasifikácie (rozšírený abstrakt)

Michal Hozza*

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

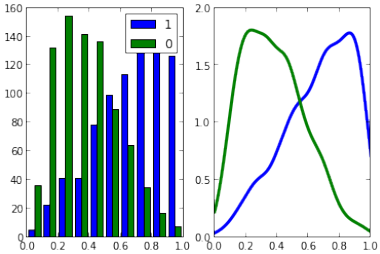
[1] Katedra aplikovanej informatiky, FMFI UK, Mlynská Dolina 842 48 Bratislava
[2] Katedra informatiky, FMFI UK, Mlynská Dolina 842 48 Bratislava

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 zaklade lokálnej informácie Na zakomponovanie informácie sme sa rozhodli využiť klasifikátory, ktoré rozhodujú, či dané pozície majú byť zarovnané k sebe alebo nie. Ako klasifikátor sme sa rozhodli využiť *RandomForest*[Breiman, 2001], pretože aktuálne patrí medzi najlepšie klasifikátory. V našich modeloch sme použili rôzne klasifikátory pre Match stav a Insert stavy.

Vstupné dáta pre klasifikátor sú okná veľkosti w, v ktorom sa nachádzaja 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 2 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. Na obrázku 1 je distribúcia výstupov klasifikátora. Pozitívne príklady sú tie, ktoré majú byť zarovnané k sebe.



Obr. 1: Distribúcia výstupu klasifikátora pre pozitívne (modrá) a negatívne (zelená) príklady. Okno veľkosti 5 a anotácia s génom. Vpravo je spojitá aproximácia ľavého obrázku. Na x-ovej osi je výstup z klasifikátora, na y-ovej je početnosť daného výstupu.

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

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

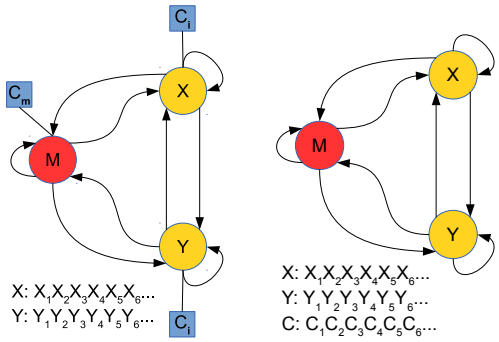
Model s klasifikátorom ako emisiou: (Obr. 2a) V tomto modeli sme nahradili emisné tabuľky stavov výstupom z klasifikátora.

Model však nie je úplne korektný, pretože pravdepodobnosti emisií nesumujú do 1. Avšak ukázalo sa, že model aj napriek tomu funguje dobre.

V tomto modeli sme trénovali iba tranzície, emisie sme mali priamo z natrénovaného klasifikátora.

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

V tomto modeli sme trénovali aj tranzície aj emisie. Výstupy z klasifikátora sme rozdelili do 10 košov rovnomerne na intervale $\langle 0, 1\rangle$



(a) Model s klasifikátorom ako emisiou

(b) Model s klasifikátorovou páskou

Obr. 2: Modely s klasifikátorom

Literatúra

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

[Brejová and Vinař, 2011] Brejová, B. and Vinař, T. (2011). *Metódy v bioinformatike [Methods in Bioinformatics]*. Knižničné a edičné centrum FMFI UK. Lecture notes.

[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.