Zarovnávanie sekvencií s použitím metód klasifikácie (rozšírený abstrakt)

Michal Hozza*

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

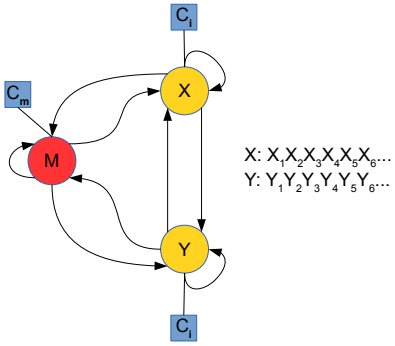
[1] Katedra aplikovanej informatiky, FMFI UK, Mlynská Dolina 842 48 Bratislava
[2] Katedra informatiky, FMFI UK, Mlynská Dolina 842 48 Bratislava

Bežné zarovnávače DNA sekvencií pracujú len s jednotlivými dvojicami báz [Durbin et al., 1998]. Naše zakomponováva dodatočné informácie (poskytnuté formou anotácií k bázam), aby sme vytvorili kvalitnejšie zarovnania.

Na zakomponovanie informácie sme sa rozhodli využiť klasifikátory, ktoré rozhodujú, či dané pozície majú byť zarovnané k sebe alebo nie.

Vyvinuli sme 2 modely pre zarovnanie sekvencií s anotáciami za pomoci klasifikátora, ktoré sú založené na skrytých Markovových modeloch.

Model s klasifikátorom ako emisiou: V tomto modeli sme nahradili emisné tabuľky stavov výstupom z klasifikátora.



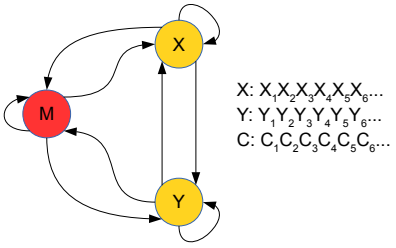
Obr. 1: Model s klasifikátorom ako emisiou

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

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

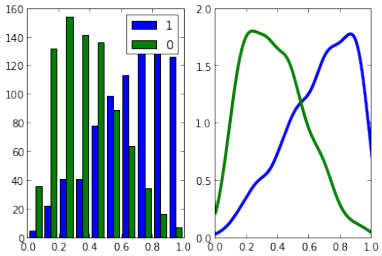
Model s klasifikátorovou páskou: Aby sme vyriešili problém s korektnosťou predošlého modelu, navrhli sme alternatívny model, ktorý navyše modeluje aj výstup z klasifikátora. Nemodelujeme teda len dvojicu sekvencií, ale aj sekvenciu výstupov klasifikátora.

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



Obr. 2: Model s klasifikátorovou páskou

V tomto modeli sme trénovali aj tranzície aj emisie. Výstupy z klasifikátora sme rozdelili do 10 košov rovnomerne na intervale <0, 1>



Obr. 3: Distribúcia výstupu klasifikátora pre pozitívne (modrá) a negatívne (zelená) príklady

Ukázalo sa, že klasifikátor sa dokáže celkom dobre naučiť, ktoré bázy majú byť zarovnané k sebe a ktoré nie. Na obrázku 3 je distribúcia výstupov klasifikátora. Pozitívne príklady sú tie, ktoré majú byť zarovnané k sebe.

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.