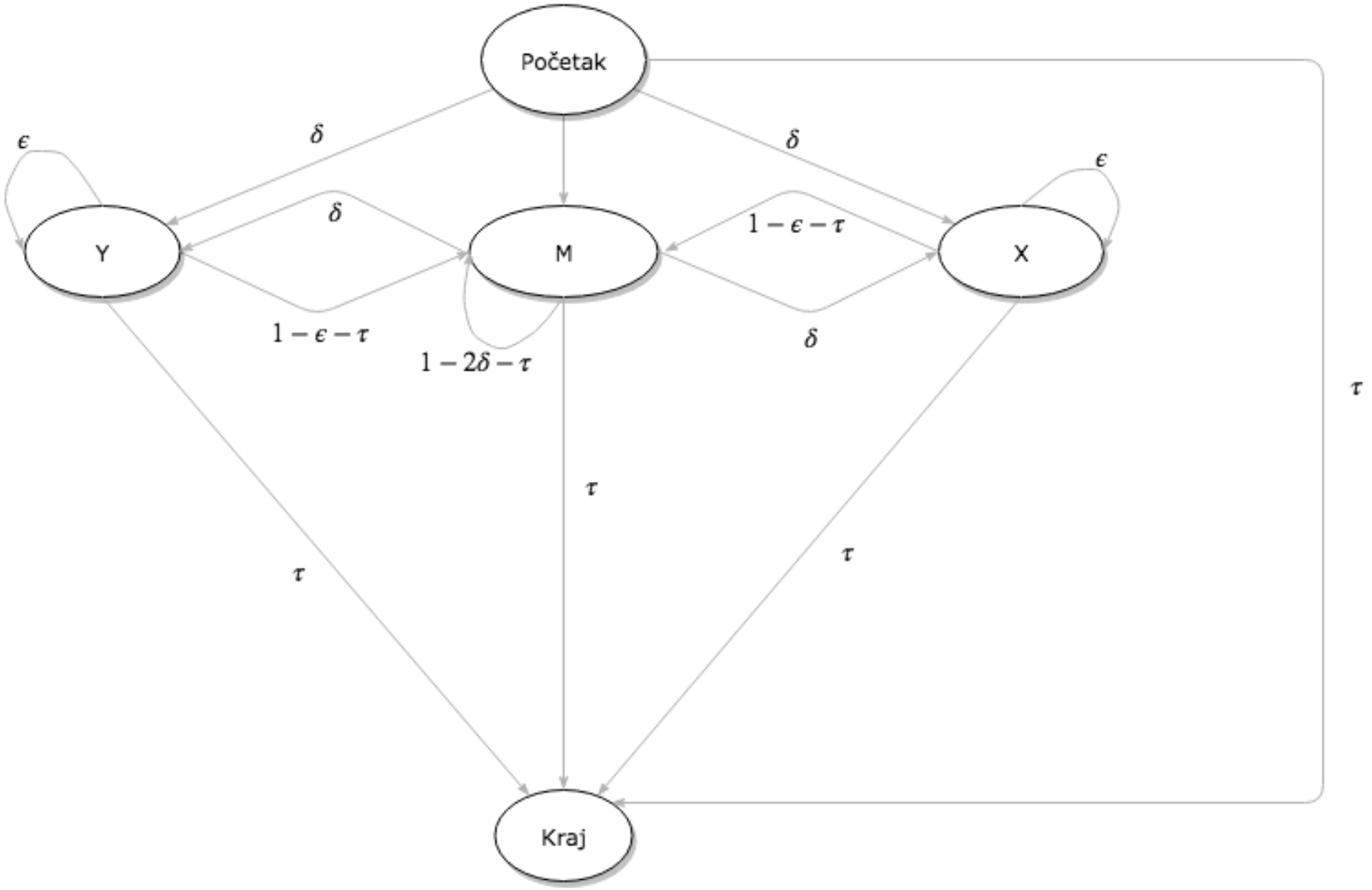
Određivanje poravnanja parova sljedova korištenjem HMM

Tomislav Božurić, Martin Pisačić, Krešimir Topolovec

Hidden Markov Model



Optimalno poravnanje logaritamskim kvotama

Algoritam: optimalno poravnanje logaritamskih kvota [3]

Inicijalizacija:

$$V^M(0,0) = -2log(\eta), V^X(0,0) = V^Y(0,0) = -\infty$$
$$V^\bullet(i,-1) = V^\bullet(-1,j) = -\infty$$

Korak:

za svaki $i = 0,...,n$, $j = 0,...,m$ osim (0,0):

$$V^M(i,j) = s(x_i,y_j) + \max \begin{cases} V^M(i-1,j-1) \\ V^X(i-1,j-1) \\ V^Y(i-1,j-1) \end{cases} \quad (4)$$

$$V^X(i,j) = \max \begin{cases} V^M(i-1,j) - d \\ V^X(i-1,j) - e \end{cases} \quad (5)$$

$$V^Y(i,j) = \max \begin{cases} V^M(i,j-1) - d \\ V^Y(i,j-1) - e \end{cases} \quad (6)$$

Uvjet zaustavljanja: $V = \max(V^M(n,m), V^X(n,m) + c, V^Y(n,m) + c)$

Pri čemu su:

$$s(a,b) = \log \frac{p_{ab}}{q_a q_b} + \log \frac{1-2\delta-\tau}{(1-\eta)^2}$$
$$d = -\log \frac{\delta(1-\epsilon-\tau)}{(1-\eta)(1-2\delta-\tau)}$$
$$e = -\log \frac{\epsilon}{1-\eta}$$
$$c = \log(1-2\delta-\tau) - \log(1-\epsilon-\tau)$$

Procjena parametara

- Maximum Likelihood Estimation
- Smoothing level = 1

$$e(\alpha, \beta) = \frac{E(\alpha,\beta)+smoothing}{\sum_{\alpha,\beta} E(\alpha,\beta)+2 \cdot smoothing}$$

$$t(X, Y) = \frac{T(X,Y)+smoothing}{\sum_{X,Y} T(X,Y)+2 \cdot smoothing}$$

Rezultati modela učenog nad bazom različitih virusa

Sekvenca	*Pairwise HMM score*	*MAFFT score*
HIV:Ref.A1.RW.92.92RW008.AB253421, Ref.A1.UG.92.92UG037.AB253429	10296	49486
AF086833.2 Ebola virus - Mayinga, Zaire, 1976, complete genome, JF828358.1 Lloviu virus strain MS-Liver-86/2003, complete genome	13473	46349
FJ424484.1 Rabies virus red fox, MG996466.1 Rabies lyssavirus	1268	7439
Tropomyiosin : Homo sapiens cDNA, Soares Thymus Mus musculus cDNA	545	1761
Hepatitis B virus isolate G376-A6, complete genome, Hepatitis C virus genotype 1, complete genome	−6504	2691

Tablica 7: Usporedba rezultata poravnanja modela učenog na bazi više virusa poravnatoj s alatom *MAFFT*

Konačni rezultati

Sekvenca	*Pairwise HMM score*	*MAFFT score*
HIV:Ref.A1.RW.92.92RW008.AB253421, Ref.A1.UG.92.92UG037.AB253429	46609	49486
HIV:Ref.A1.RW.92.92RW008.AB253421, Ref.A2.CD.97.97CDKTB48.AF286238	36264	42534
Tropomyiosin : Homo sapiens cDNA, Soares Thymus Mus musculus cDNA	855	1761
Hepatitis B virus isolate G376-A6, complete genome, Hepatitis C virus genotype 1, complete genome	-3273	2691
Nasumične sekvenca duljine 101 i 105 znakova	102	139
Nasumične sekvenca duljine 5097 i 5053 znakova	4300	8751

Tablica 8: Usporedba rezultata poravnanja modela učenog na bazi HIV-a poravnatoj s *ClustalW*-om

Literatura

- Mile Šikić, Mirjana Domazet-Lošo. Bioinformatika. Fakultet elektrotehnike i računarstva, 2013.
- Byung-Jun Yoon. Hidden Markov Models and their Applications in Biological Sequence Analysis. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2766791, US National Library of Medicine, National Institutes of Health, 2009.
- Jun Xie. Pairwise alignment using HMM. http://www.stat.purdue.edu/~junxie/topic4.pdf, Purdue University.
- Luay Nakhleh,. Pairwise HMMs and Sequence Alignment. https://www.cs.rice.edu/~nakhleh/COMP571/Slides-Spring2015/SequenceAlignment-PairwiseHMM.pdf , Rice University, 2015.