FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

BIOINFORMATIKA

Određivanje poravnanja parova sljedova korištenjem HMM

Autori
Tomislav Božurić
Martin Pisačić
Krešimir Topolovec

Zadatak
doc. dr. sc. Mirjana Domazet-Lošo

Siječanj, 2019

1 Opis algoritma i vizualizacija

U ovome radu korišten je modificirani Viterbijem algoritam pomoću kojeg korištnjem dinamičkog programiranja možemo pronaći najvjerojatniju sekvencu skrivenih stanja koja ujedno predstavlja optimalno poravnanje. Da bismo klasičan Viterbijem algoritam tranformirali u HMM, moramo napraviti nekoliko izmjena. Prvo moramo odrediti vjerojatnosti za emitiranje simbola iz stanja i vjerojatnosti tranzicija između pojedinih stanja. Npr. stanje M (match) ima vjerojatnosnu distribuciju

Algoritam: Viterbijev algoritam za HMM

Inicijalizacija:

$$v^M(0,0) = 1$$
$$v^\bullet(i,0) = v^\bullet(0,j) = 0$$

Korak:

za svaki $i = 1,...,n$, $j = 1,...,m$

$$v^M(i,j) = p_{x_i y_i} \max \begin{cases} (1-2\delta-\tau)v^M(i-1,j-1) \\ (1-\epsilon-\tau)v^X(i-1,j-1) \\ (1-\epsilon-\tau)v^Y(i-1,j-1) \end{cases} \tag{1}$$

$$v^X(i,j) = q_{x_i} \max \begin{cases} \delta v^M(i-1,j) \\ \epsilon v^X(i-1,j) \end{cases} \tag{2}$$

$$v^Y(i,j) = q_{y_j} \max \begin{cases} \delta v^M(i,j-1) \\ \epsilon v^Y(i,j-1) \end{cases} \tag{3}$$

Uvjet zaustavljanja: $v^E = \max(v^M(n,m), v^X(n,m), v^Y(n,m))$

Algoritam: optimalno poravnanje logaritamskih kvota

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):

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

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

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

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

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

$e = -\log \frac{\epsilon}{1-\eta}$
$c = \log(1 - 2\delta - \tau) - \log(1 - \epsilon - \tau)$

2 Analiza točnosti, vremena izvođenja i utroška memorije

3 Testiranje

3.1 Testiranje na sintetskim podatcima

3.2 Testiranje na stvarnim podatcima

References

[1] 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.

[2] Jun Xie. *Pairwise alignment using HMM*. http://www.stat.purdue.edu/~junxie/topic4.pdf, Purdue University.