Computational Biology

Lecture 11

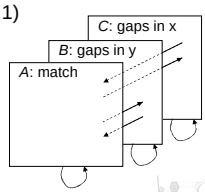
Saad Mneimneh

Alignment
Affine gap model

$$A(i,j) = s(i,j) + \max \begin{cases} A(i-1,j-1) \\ B(i-1,j-1) \\ C(i-1,j-1) \end{cases}$$

$$B(i,j) = \max \begin{cases} A(i-1,j) - e \\ B(i-1,j) - d \end{cases}$$

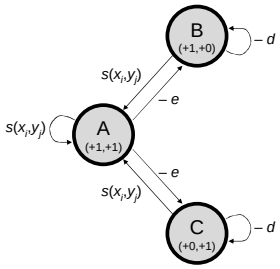
$$C(i,j) = \max \begin{cases} A(i,j-1) - e \\ C(i,j-1) - d \end{cases}$$



Saad Mneimneh

Finite state machine

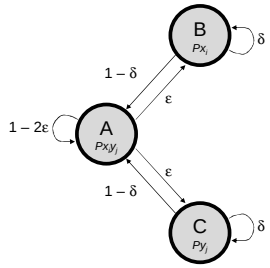
An alignment corresponds to a path in the finite state machine



Saad Mneimneh

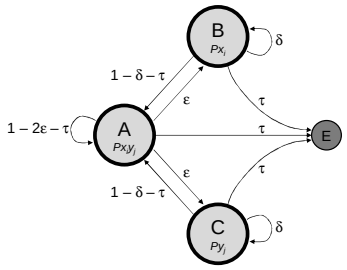
Probabilistic model HMM for related sequences

With transitional probabilities and emission probabilities, generates two sequences simultaneously



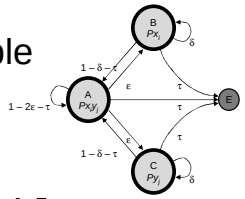
Modeling length

By adding an end state



Example

x = VLSPADK y = HLAESK



Possible alignment:

```
states:   A A A B B A A C A E
            | | | | | | | | |
emissions:  | | | | | | | | |
            | | | | | | | | |
x:          V L S P A D   K
y:          H L     A E S K
```

P = $(1-2\varepsilon-\tau)P_{VH}\cdot$
$(1-2\varepsilon-\tau)P_{LL}\cdot$
$\varepsilon P_S\cdot$
$\delta P_P\cdot$
$(1-\delta-\tau)P_{AA}\cdot$
$(1-2\varepsilon-\tau)P_{DE}\cdot$
$\varepsilon P_S\cdot$
$(1-\delta-\tau)P_{KK}\cdot$
τ

What is the most probable alignment (path in HMM)?

Viterbi algorithm

- All the algorithms we have seen for HMM apply, e.g. Viterbi
- But there is an extra dimension in the search space because of the extra emitted sequence
- Instead of using $v_k(i)$, we will use $v_k(i,j)$ because an observation x_i does not necessarily mean an observation for y_i.
- Imagine we have two clocks, one for x, and one for y, that work differently in different zones.
- Therefore, $v(i,j)$ can advance only in certain ways:
  - In zone A, both i and j advance.
  - In zone B, only i advances.
  - In zone C, only j advances.

Saad Mneimneh

Viterbi

$v_A(0,0) = 1$, $v_A(i,0) = v_A(0,i) = 0$
$v_B(0,0) = 0$, $v_B(i,0) = \varepsilon\delta^{i-1}p_{x1}\ldots p_{xi}$, $v_B(0,j) = 0$
$v_C(0,0) = 0$, $v_C(i,0) = 0$, $v_C(0,j) = \varepsilon\delta^{j-1}p_{y1}\ldots p_{yj}$

for i=1..m, j=1..n

$$v_A(i,j) = Px_iy_j \cdot \max \begin{cases} (1-2\varepsilon-\tau)v_A(i-1,j-1) \\ (1-\delta-\tau)v_B(i-1,j-1) \\ (1-\delta-\tau)v_C(i-1,j-1) \end{cases}$$

$$v_B(i,j) = Px_i \cdot \max \begin{cases} \varepsilon v_A(i-1,j) \\ \delta v_B(i-1,j) \end{cases}$$

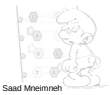
$$v_C(i,j) = Py_j \cdot \max \begin{cases} \varepsilon v_A(i,j-1) \\ \delta v_C(i,j-1) \end{cases}$$

$v_E = \tau.\max(v_A(m,n), v_B(m,n), v_C(m,n))$

Saad Mneimneh

Observation

- Viterbi algorithm looks very similar to Needleman-Wunsch
- Let's investigate more this similarity
- Performing Viterbi in log space will definitely make it look like Needleman-Wunsch (multiplication → additive), let's do even more…
- Let's compute the log-odds ratio of the max probability of an alignment (path) to the probability of obtaining the sequences at random, therefore compute the maximum log-odds ratio.
  - Construct a model for random sequences
  - Perform Viterbi in log space with both models in mind
- We should then see almost the same algorithm



Saad Mneimneh

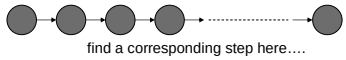
Two HMMs at the same time

alignment HMM path

$p = p_1.p_2.p_3……$

for each step here….

random HMM path

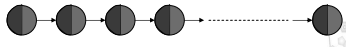


find a corresponding step here….

$p = q_1.q_2.q_3……$

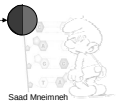
and associate steps together

alignment HMM path



$p = (p_1/q_1) . (p_2/q_2) . (p_3/q_3)……$

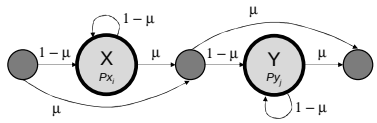
$\log p = \log (p_1/q_1) + \log (p_2/q_2) + \log (p_3/q_3)……$



Saad Mneimneh

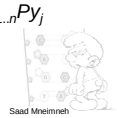
Random model

x and *y* are generated independently according to the individual probabilities Px_i and Py_i



$p(x,y \mid R) = \mu(1-\mu)^m \prod_{i=1..m} Px_i \,.\, \mu(1-\mu)^n \prod_{j=1..n} Py_j$

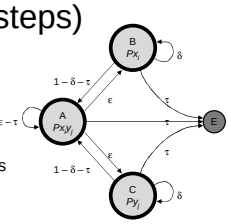
$= \mu^2(1-\mu)^{m+n} \prod_{i=1..m} Px_i \prod_{j=1..n} Py_j$



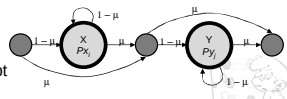
Saad Mneimneh

Aligned v.s. random (identifying steps)

- In the alignment model:
  - Each match step contributes $(1\text{-}2\varepsilon\text{-}\tau)Px_iy_j$ *(…depends)*
  - Each start gap step contributes εPx_i or εPy_j
  - Each additional gap step contributes δPx_i or δPy_j



- In the random model:
  - Each letter step contributes $(1\text{-}\mu)Px_i$ or $(1\text{-}\mu)Py_j$
  - All terms are counted except for a μ^2



Saad Mneimneh

Aligned v.s. random (associating steps)

Alignment model	Random model
match x_i y_j	generate x_i and y_j
start gap in y with x_i	generate x_i
start gap in x with y_j	generate y_j
gap in y with x_i	generate x_i
gap in x with y_j	generate y_j

Saad Mneimneh

Obtain p/q for each step…

match step

$$s(x_i, y_j) = \log \frac{Px_i y_j (1-2\varepsilon-\tau)}{(1-\mu)Px_i (1-\mu)Py_j} \cong \log \frac{Px_i y_j}{Px_i Py_j} + \log \frac{(1-2\varepsilon-\tau)}{(1-\mu)^2} \cong \log \frac{Px_i y_j}{Px_i Py_j}$$

corresponding random steps

gap step

$$d = -\log \frac{\delta Px_i}{(1-\mu)Px_i} = -\log \frac{\delta}{(1-\mu)} \cong -\log \delta$$

corresponding random step

start gap step

$$e = -\log \frac{\varepsilon Px_i}{(1-\mu)Px_i} \frac{1-\delta-\tau}{(1-2\varepsilon-\tau)} \cong -\log \varepsilon$$

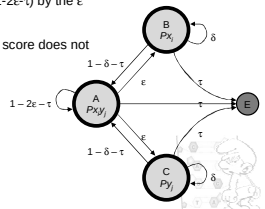
corresponding random step

adjustment factor to correct for $s(x_i,y_j)$

Saad Mneimneh

Start gap adjustment factor

- The contribution of state A assumes that a match always follows a match.
- If state A (match) is reached from state B or state C (a gap), then the term contributed to the probability of the path is $(1-\delta-\tau)\ Px_iPy_j$ and not $(1-2\varepsilon-\tau)Px_iPy_j$
- We correct this before it happens when we first open the gap (go from A to B or C) by multiplying $(1-\delta-\tau)/(1-2\varepsilon-\tau)$ by the ε contribution of the start gap
- But what if we never go back to A? Then the score does not exactly correspond to the log-odd ratio.



Saad Mneimneh

log-odds ratio Viterbi (Needleman-Wunsch)

$v_A(0,0) = -2\log\mu$, $v_A(i,0) = v_A(0,i) = -\infty$
$v_B(0,0) = -\infty$, $v_B(i,0) = -e-(i-1)d$, $v_B(0,j) = -\infty$
$v_C(0,0) = -\infty$, $v_C(i,0) = -\infty$, $v_C(0,j) = -e-(j-1)d$

for i=1..m, j=1..n

$$V_A(i,j) = s(x_i,y_j) + \max \begin{cases} V_A(i-1,j-1) \\ V_B(i-1,j-1) \\ V_C(i-1,j-1) \end{cases}$$

$$V_B(i,j) = \max \begin{cases} V_A(i-1,j) - e \\ V_B(i-1,j) - d \end{cases}$$

$$V_C(i,j) = \max \begin{cases} V_A(i,j-1) - e \\ V_C(i,j-1) - d \end{cases}$$

We don't go back to state A.
The anticipated error did not occur.
Correct the adjustment.

$V_E = \max(V_A(m,n), V_B(m,n) - c, V_C(m,n) - c)$
$c = \log(1-\delta-\tau) - \log(1-2\varepsilon-\tau)$

Saad Mneimneh

What is the meaning of the scores?

- From Viterbi, we obtained Needleman-Wuncsh with scores directly related to the parameters of a probabilistic model.

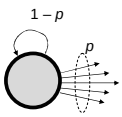
Needleman-Wunsch is just Viterbi with an appropriate transformation!

- We knew from before that the score is related to the log-odds ratio of probabilities. Now Viterbi gives more justification of this (we are actually after the most likely alignment compared to just a random instance!).
- We have a clear relation between the scores (even for gaps now) of Needleman-Wunsch and the parameters of a probabilistic model.
- The question now is: what do the model parameters ε, δ, τ, and μ really represent?

Saad Mneimneh

Duration of a state

Let us look at a single state with probability p of transitioning outside

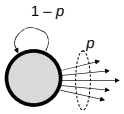


What is the probability of staying in that state for a duration D.

Saad Mneimneh

Duration of a state

$Pr[D=1]=p$
$Pr[D=2]=(1-p)p$
$Pr[D=3]=(1-p)(1-p)p$
.
.
.
$Pr[D=i]=(1-p)^{i-1}p$



i.e. geometric distribution

This is a disadvantage of a Markov Chain: it imposes a geometric distribution on the duration in a state, which might not be true.

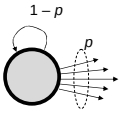
There are ways around that… (see Durbin p. 69)

Saad Mneimneh

Mean duration

What is the mean duration?

$E[D] = 1/p$



Saad Mneimneh

Therefore…

$2\varepsilon = 1/L_u$, L_u = mean length of ungapped regions

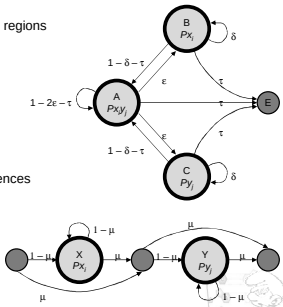
$L_u \nearrow$ $\varepsilon \searrow$ $\log \varepsilon \searrow$ $e \nearrow$

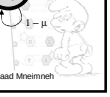
$1 - \delta = 1/L_g$, L_g = mean length of gap

$L_g \nearrow$ $\delta \nearrow$ $\log \delta \nearrow$ $d \searrow$

$\tau = 1/L_r$, L_r = mean length of related sequences (~ 0)

$\mu = 1/L$, L = mean length of a sequence (~ 0)





Saad Mneimneh