Alignment with dynamic programming

Heng Li

July 17, 2022

1 General notations

Suppose we have two sequences: a *target* sequence and a *query* sequence. The length of the target sequence is ℓ_t with each residue indexed by i. The length of query is ℓ_q with each residue indexed by j. Gaps on the target sequence are *deletions* and gaps on the query are *insertions*. Function $S(i,j)$ gives the score between two residues on the target and the query, respectively. $q>0$ is the gap open/initiation penalty and $e>0$ the gap extension penalty. A gap of length k costs $q+k\cdot e$.

2 Global alignment with affine-gap penalties

2.1 Durbin's formulation

The original Durbin's formulation is:

$$
\begin{aligned}
M_{ij} &= \max\{M_{i-1,j-1},E_{i-1,j-1},F_{i-1,j-1}\}+S(i,j)\\
E_{ij} &= \max\{M_{i-1,j}-q,E_{i-1,j}\}-e\\
F_{ij} &= \max\{M_{i,j-1}-q,F_{i,j-1}\}-e
\end{aligned}
$$

This formulation disallows a deletion immediately followed an insertion, or vice versa. A more general form is:

$$
\begin{aligned}
M_{ij} &= \max\{M_{i-1,j-1},E_{i-1,j-1},F_{i-1,j-1}\}+S(i,j)\\
E_{ij} &= \max\{M_{i-1,j}-q,E_{i-1,j},F_{i-1,j}-q\}-e\\
F_{ij} &= \max\{M_{i,j-1}-q,E_{i,j-1}-q,F_{i,j-1}\}-e
\end{aligned}
$$

2.2 Green's formulation

If we define:

$$H_{ij}=\max\{M_{ij},E_{ij},F_{ij}\}$$

the Durbin's formulation can be transformed to

$$\begin{aligned} E_{ij} &= \max\{H_{i-1,j} - q, E_{i-1,j}\} - e \\ F_{ij} &= \max\{H_{i,j-1} - q, F_{i,j-1}\} - e \\ H_{ij} &= \max\{H_{i-1,j-1} + S(i,j), E_{ij}, F_{ij}\} \end{aligned}$$

I first saw this formulation in Phrap developed by Phil Green, though it may have been used earlier. If we further introduce

$$\begin{aligned} E'_{ij} &= E_{i+1,j} \\ F'_{ij} &= F_{i,j+1} \end{aligned}$$

we have

$$\begin{aligned} H_{ij} &= \max\{H_{i-1,j-1} + S(i,j), E'_{i-1,j}, F'_{i,j-1}\} \\ E'_{ij} &= \max\{H_{ij} - q, E'_{i-1,j}\} - e \\ F'_{ij} &= \max\{H_{ij} - q, F'_{i,j-1}\} - e \end{aligned}$$

In fact, we more often use this set of equations in practical implementations. The initial conditions are

$$\begin{aligned} H_{-1,j} &= \begin{cases} 0 & (j = -1) \\ -q - (j+1)\cdot e & (0 \le j < \ell_q) \end{cases} \\ H_{i,-1} &= \begin{cases} 0 & (i = -1) \\ -q - (i+1)\cdot e & (0 \le i < \ell_t) \end{cases} \\ E'_{-1,j} &= E_{0,j} = H_{-1,j} - q - e = -2q - (j+2)\cdot e \\ F'_{i,-1} &= F_{i,0} = -2q - (i+2)\cdot e \end{aligned}$$

2.3 Suzuki's formulation

2.3.1 Standard coordinate

Now let

$$\begin{aligned} u'_{ij} &= H_{ij} - H_{i-1,j} \\ v'_{ij} &= H_{ij} - H_{i,j-1} \\ x'_{ij} &= E'_{ij} - H_{ij} \\ y'_{ij} &= F'_{ij} - H_{ij} \end{aligned}$$

We have

$$\begin{aligned} x'_{ij} &= \max\{-q, E'_{i-1,j} - H_{i-1,j} + H_{i-1,j} - H_{ij}\} - e \\ &= \max\{-q, x'_{i-1,j} - u'_{ij}\} - e \end{aligned} \tag{1}$$

Similarly

$$y'_{ij} = \max\{-q, y'_{i,j-1} - v'_{ij}\} - e \tag{2}$$

To derive the equation to compute $u'(i,j)$ and $v'(i,j)$, we note that

$$\begin{aligned} H_{ij} - H_{i-1,j-1} &= \max\{S(i,j), E'_{i-1,j} - H_{i-1,j-1}, F'_{i,j-1} - H_{i-1,j-1}\} \\ &= \max\{S(i,j), x'_{i-1,j} + v'_{i-1,j}, y'_{i,j-1} + u'_{i,j-1}\} \end{aligned}$$

and

$$H_{ij} - H_{i-1,j-1} = u'_{ij} + v'_{i-1,j} = v'_{ij} + u'_{i,j-1}$$

We can derive the recursive equation for u'_{ij} and v'_{ij}:

$$\begin{aligned} z'_{ij} &= \max\{S(i,j), x'_{i-1,j} + v'_{i-1,j}, y'_{i,j-1} + u'_{i,j-1}\} \\ u'_{ij} &= z'_{ij} - v'_{i-1,j} \\ v'_{ij} &= z'_{ij} - u'_{i,j-1} \\ x'_{ij} &= \max\{0, x'_{i-1,j} + v'_{i-1,j} - z'_{ij} + q\} - q - e \\ y'_{ij} &= \max\{0, y'_{i,j-1} + u'_{i,j-1} - z'_{ij} + q\} - q - e \end{aligned}$$

From eq. (**??**) we can infer that $x'_{ij} \geq -q-e$ and similarly $y'_{ij} \geq -q-e$. We further have:

$$u'_{ij} = H_{ij} - H_{i-1,j-1} - v'_{i-1,j} \geq x'_{i-1,j} \geq -q - e$$

Therefore, we have a lower bound $-q-e$ for u', v', x' and y'. This motivates us to redefine the four variables as:

$$\begin{aligned} u''_{ij} &= H_{ij} - H_{i-1,j} + q + e \\ v''_{ij} &= H_{ij} - H_{i,j-1} + q + e \\ x''_{ij} &= E'_{ij} - H_{ij} + q + e \\ y''_{ij} &= F'_{ij} - H_{ij} + q + e \end{aligned}$$

The recursion becomes

$$\begin{aligned} z''_{ij} &= \max\{S(i,j) + 2q + 2e, x''_{i-1,j} + v''_{i-1,j}, y''_{i,j-1} + u''_{i,j-1}\} \\ u''_{ij} &= z''_{ij} - v''_{i-1,j} \\ v''_{ij} &= z''_{ij} - u''_{i,j-1} \\ x''_{ij} &= \max\{0, x''_{i-1,j} - u''_{ij} + q\} = \max\{0, x''_{i-1,j} + v''_{i-1,j} - z''_{ij} + q\} \\ y''_{ij} &= \max\{0, y''_{i,j-1} - v''_{ij} + q\} = \max\{0, y''_{i,j-1} + u''_{i,j-1} - z''_{ij} + q\} \end{aligned}$$

Here z_{ij} is a temporary variable. u'', v'', x'' and y'' are all non-negtive.

2.3.2 Rotated coordinate

We let

$$\begin{aligned} r &= i + j \\ t &= i \end{aligned}$$

We have

$$
\begin{aligned}
z_{rt} &= \max\{S(t,r-t)+2q+2e, x_{r-1,t-1}+v_{r-1,t-1}, y_{r-1,t}+u_{r-1,t}\} \\
u_{rt} &= z_{rt}-v_{r-1,t-1} \\
v_{rt} &= z_{rt}-u_{r-1,t} \\
x_{rt} &= \max\{0, x_{r-1,t-1}+v_{r-1,t-1}-z_{rt}+q\} \\
y_{rt} &= \max\{0, y_{r-1,t}+u_{r-1,t}-z_{rt}+q\}
\end{aligned}
$$

Due to the definition of r and t, the following inequation must stand:

$$0 \le r-t \le \ell_q - 1$$

$$0 \le t \le \ell_t - 1$$

where ℓ_t is the length of the sequence indexed by i and ℓ_q the length indexed by j. In case of banded alignment with a fixed diagonal band of size w,

$$-w \le j-i \le w$$

In the (r,t) coordinate, it is:

$$\frac{r-w}{2} \le t \le \frac{r+w}{2}$$

Putting these together:

$$0 \le r \le \ell_q + \ell_t - 2$$

$$\max\left\{0, r-\ell_q+1, \frac{r-w}{2}\right\} \le t \le \min\left\{\ell_t-1, r, \frac{r+w}{2}\right\}$$

2.3.3 Initial conditions

$$x_{r-1,-1} = x''_{-1,r} = E'_{-1,r} - H_{-1,r} + q + e = 0$$

$$y_{r-1,r} = y''_{r,-1} = 0$$

$$v_{r-1,-1} = v''_{-1,r} = H_{-1,r} - H_{-1,r-1} + q + e = \begin{cases} q & (r>0) \\ 0 & (r=0) \end{cases}$$

$$u_{r-1,r} = u''_{r,-1} = H_{r,-1} - H_{r-1,-1} + q + e = \begin{cases} q & (r>0) \\ 0 & (r=0) \end{cases}$$

3 Alignment with dual affine-gap penalties

3.1 Green's formulation

$$
\begin{aligned}
H_{ij} &= \max\{H_{i-1,j-1} + S(i,j), E'_{i-1,j}, F'_{i,j-1}, \tilde{E}'_{i-1,j}, \tilde{F}'_{i,j-1}\} \\
E'_{ij} &= \max\{H_{ij} - q, E'_{i-1,j}\} - e \\
F'_{ij} &= \max\{H_{ij} - q, F'_{i,j-1}\} - e \\
\tilde{E}'_{ij} &= \max\{H_{ij} - \tilde{q}, \tilde{E}'_{i-1,j}\} - \tilde{e} \\
\tilde{F}'_{ij} &= \max\{H_{ij} - \tilde{q}, \tilde{F}'_{i,j-1}\} - \tilde{e}
\end{aligned}
$$

The initial conditions are:

$$
\begin{aligned}
H_{-1,j} &= \begin{cases} 0 & (j = -1) \\ \max\{-q - (j+1)\cdot e, -\tilde{q} - (j+1)\cdot \tilde{e}\} & (0 \le j < \ell_q) \end{cases} \\
H_{i,-1} &= \begin{cases} 0 & (i = -1) \\ \max\{-q - (i+1)\cdot e, -\tilde{q} - (i+1)\cdot \tilde{e}\} & (0 \le i < \ell_t) \end{cases} \\
E'_{-1,j} &= E_{0,j} = H_{-1,j} - q - e \\
F'_{i,-1} &= F_{i,0} = H_{i,-1} - q - e \\
\tilde{E}'_{-1,j} &= \tilde{E}_{0,j} = H_{-1,j} - \tilde{q} - \tilde{e} \\
\tilde{F}'_{i,-1} &= \tilde{F}_{i,0} = H_{i,-1} - \tilde{q} - \tilde{e}
\end{aligned}
$$

3.2 Suzuki's formulation

$$
\begin{aligned}
z'_{ij} &= \max\{S(i,j), x'_{i-1,j} + v'_{i-1,j}, y'_{i,j-1} + u'_{i,j-1}, \\
&\qquad \tilde{x}'_{i-1,j} + v'_{i-1,j}, \tilde{y}'_{i,j-1} + u'_{i,j-1}\} \\
u'_{ij} &= z'_{ij} - v'_{i-1,j} \\
v'_{ij} &= z'_{ij} - u'_{i,j-1} \\
x'_{ij} &= \max\{0, x'_{i-1,j} + v'_{i-1,j} - z'_{ij} + q\} - q - e \\
y'_{ij} &= \max\{0, y'_{i,j-1} + u'_{i,j-1} - z'_{ij} + q\} - q - e \\
\tilde{x}'_{ij} &= \max\{0, \tilde{x}'_{i-1,j} + v'_{i-1,j} - z'_{ij} + \tilde{q}\} - \tilde{q} - \tilde{e} \\
\tilde{y}'_{ij} &= \max\{0, \tilde{y}'_{i,j-1} + u'_{i,j-1} - z'_{ij} + \tilde{q}\} - \tilde{q} - \tilde{e}
\end{aligned}
$$

In the rotated coordinate:

$$
\begin{aligned}
z_{rt} &= \max\{S(t, r-t), x_{r-1,t-1} + v_{r-1,t-1}, y_{r-1,t} + u_{r-1i,t}, \\
&\qquad \tilde{x}_{r-1,t-1} + v_{r-1,t-1}, \tilde{y}_{r-1,t} + u_{r-1,t}\} \\
u_{rt} &= z_{rt} - v_{r-1,t-1} \\
v_{rt} &= z_{rt} - u_{r-1,t} \\
x_{rt} &= \max\{0, x_{r-1,t-1} + v_{r-1,t-1} - z_{rt} + q\} - q - e \\
y_{rt} &= \max\{0, y_{r-1,t} + u_{r-1,t} - z_{rt} + q\} - q - e
\end{aligned}
$$

$$
\begin{aligned}
\tilde{x}_{rt} &= \max\{0, \tilde{x}_{r-1,t-1} + v_{r-1,t-1} - z_{rt} + \tilde{q}\} - \tilde{q} - \tilde{e} \\
\tilde{y}_{rt} &= \max\{0, \tilde{y}_{r-1,t} + u_{r-1,t} - z_{rt} + \tilde{q}\} - \tilde{q} - \tilde{e}
\end{aligned}
$$

By definition, it is easy to see the initial conditions except u and v:

$$x_{r-1,-1} = x'_{-1,r} = E'_{-1,r} - H_{-1,r} = -q - e$$

$$y_{r-1,r} = y'_{r,-1} = F'_{r,-1} - H_{r,-1} = -q - e$$

$$\tilde{x}_{r-1,-1} = -\tilde{q} - \tilde{e}$$

$$\tilde{y}_{r-1,-1} = -\tilde{q} - \tilde{e}$$

$$v_{r-1,-1} = H_{-1,r} - H_{-1,r-1} = \begin{cases} \max\{-q-e, -\tilde{q}-\tilde{e}\} & (r = 0) \\ -e & (r < \lceil \frac{\tilde{q}-q}{e-\tilde{e}} - 1 \rceil) \\ r(e-\tilde{e}) - (\tilde{q}-q) - \tilde{e} & (r = \lceil \frac{\tilde{q}-q}{e-\tilde{e}} - 1 \rceil) \\ -\tilde{e} & (r > \lceil \frac{\tilde{q}-q}{e-\tilde{e}} - 1 \rceil) \end{cases}$$