1 Affine Gapped Alignment

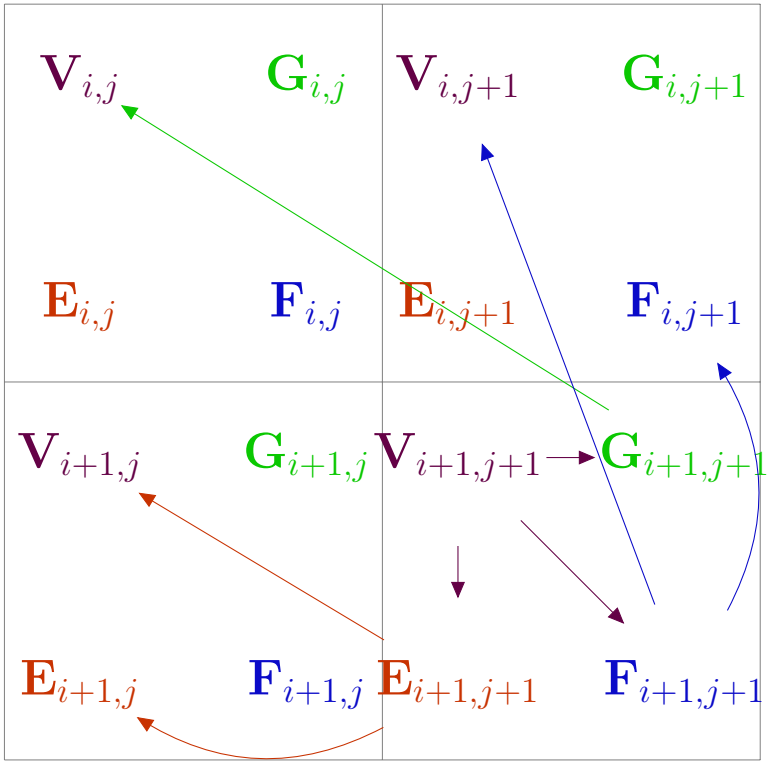
Brown University CS181, Fall 2016, with Professor Sorin Istrail. Document Prepared by Cyrus Cousins

In this document, we show how the recurrence relationship for affine gapped alignment is translated to a dynamic programming solution. First, we review the significance of each term the recurrence relationship:

$$\mathbf{V}_{i,j} \doteq \max\left(\mathbf{E}_{i,j}, \mathbf{F}_{i,j}, \mathbf{G}_{i,j}\right) \qquad \mathbf{G}_{i,j} \doteq \begin{cases} x_i = y_j & : \ \mathbf{V}_{i-1,j-1} + \alpha \\ x_i \neq y_j & : \ \mathbf{V}_{i-1,j-1} - \beta \end{cases}$$

$$\mathbf{E}_{i,j} \doteq \max\left(\mathbf{E}_{i,j-1} - \tau, \mathbf{V}_{i,j-1} - \gamma - \tau\right) \quad \mathbf{F}_{i,j} \doteq \max\left(\mathbf{F}_{i-1,j} - \tau, \mathbf{V}_{i-1,j} - \tau - \gamma\right)$$

When aligning strings of lengths n and m, rather than write out four separate matrices, it's easier to visualize as a single $(n+1) \times (m+1)$ matrix with 4 entries per cell. In this document, the position and color of an entry determines whether it corresponds to **V**, **G**, **E**, or **F**.



In the diagram to the left, each of the arrows corresponds to one of the dependencies in the recurrence. The arrow emanating from **G** represents a match or mismatch, depending on whether the characters at a given position match. The arrows from **E** and **F** correspond to opening and extendig gaps. Finally the arrows from **V** simply take the maximum over the other three matrices within each cell.

2 On Initialization

$$\mathbf{V}_{i,j} \doteq \max\left(\mathbf{E}_{i,j}, \mathbf{F}_{i,j}, \mathbf{G}_{i,j}\right) \qquad \mathbf{G}_{i,j} \doteq \begin{cases} x_i = y_j & : \mathbf{V}_{i-1,j-1} + \alpha \\ x_i \neq y_j & : \mathbf{V}_{i-1,j-1} - \beta \end{cases}$$

$$\mathbf{E}_{i,j} \doteq \max\left(\mathbf{E}_{i,j-1} - \tau, \mathbf{V}_{i,j-1} - \gamma - \tau\right) \quad \mathbf{F}_{i,j} \doteq \max\left(\mathbf{F}_{i-1,j} - \tau, \mathbf{V}_{i-1,j} - \tau - \gamma\right)$$

The above recurrence covers the general case of affine gapped alignment, but we need to be careful about how we initialize our matrix. Specifically, it is not possible to extend a gap that does not exist, so we need to perform some special initialization. The base cases look like this:

$$\mathbf{V}_{0,0} \leftarrow 0 \qquad \mathbf{G}_{i,0}, \mathbf{G}_{0,j} \leftarrow -\infty \qquad \mathbf{F}_{0,j}, \mathbf{E}_{i,0} \leftarrow -\infty$$

The use of $-\infty$ is a bit of an implementation trick: by initializing like this, we ensure that whenever maxima are taken, the $-\infty$ are never maximal. This is equivalently to removing all edges that terminate at a $-\infty$ from the edit graph. In this manner, $-\infty$ is used to vastly simplify an implementation or specification, as we only need to initialize properly to remove all the impossible scenarios from consideration.

In this document, $-\infty$ values are simply represented as blank spaces, and edges to them not drawn. This is primarily done to reduce visual noise, buut it also plays in to the edit graph interpretation, where neither these "impossible" nodes nor edges pointing to them exist.

We're almost ready to align some sequences now! All we need is a scoring function and some sequences.

Take $\alpha = \beta = \gamma = \tau = 1$. Recall that α is the similarity score of a match, β is the cost of a mismatch, γ is the cost of a gap opening, and τ is the cost of extending a gap.

Now, we will align the following sequences:

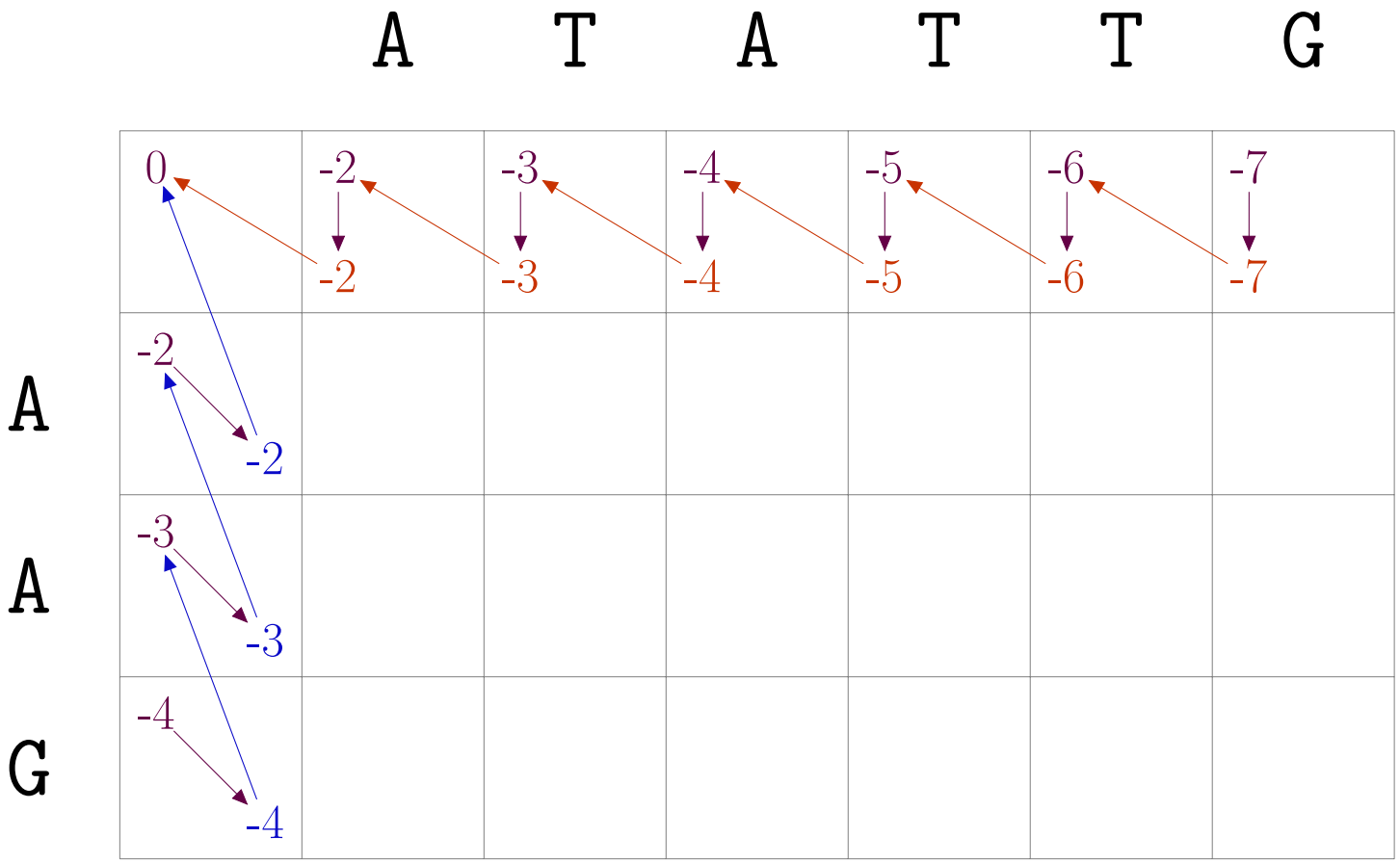
$$x \doteq \texttt{AAG} \quad y \doteq \texttt{ATATTG}$$

3 Initialization

$$\mathbf{V}_{i,j} \doteq \max\left(\mathbf{E}_{i,j}, \mathbf{F}_{i,j}, \mathbf{G}_{i,j}\right) \qquad \mathbf{G}_{i,j} \doteq \begin{cases} x_i = y_j \ : \ \mathbf{V}_{i-1,j-1} + \alpha \\ x_i \neq y_j \ : \ \mathbf{V}_{i-1,j-1} - \beta \end{cases}$$

$$\mathbf{E}_{i,j} \doteq \max\left(\mathbf{E}_{i,j-1} - \tau, \mathbf{V}_{i,j-1} - \gamma - \tau\right) \quad \mathbf{F}_{i,j} \doteq \max\left(\mathbf{F}_{i-1,j} - \tau, \mathbf{V}_{i-1,j} - \tau - \gamma\right)$$

$$\alpha = \beta = \gamma = \tau = 1$$



Here the first row and column are populated, and backtrace arrows are given.

4 Filling Cell 1-1

$$\mathbf{V}_{i,j} \doteq \max\left(\mathbf{E}_{i,j}, \mathbf{F}_{i,j}, \mathbf{G}_{i,j}\right) \qquad \mathbf{G}_{i,j} \doteq \begin{cases} x_i = y_j & : \mathbf{V}_{i-1,j-1} + \alpha \\ x_i \neq y_j & : \mathbf{V}_{i-1,j-1} - \beta \end{cases}$$

$$\mathbf{E}_{i,j} \doteq \max\left(\mathbf{E}_{i,j-1} - \tau, \mathbf{V}_{i,j-1} - \gamma - \tau\right) \quad \mathbf{F}_{i,j} \doteq \max\left(\mathbf{F}_{i-1,j} - \tau, \mathbf{V}_{i-1,j} - \tau - \gamma\right)$$

$$\alpha = \beta = \gamma = \tau = 1$$

		A	T	A	T	T	G
	0	-2 -2	-3 -3	-4 -4	-5 -5	-6 -6	-7 -7
A	-2 -2	1 → 1 -4 -4					
A	-3 -3						
G	-4 -4						

Things begin to get interesting here. Gap extensions are still not possible ($-\infty$ values), but new gaps can now open in either direction.

5 Filling Out Row and Column 1

$$\mathbf{V}_{i,j} \doteq \max\left(\mathbf{E}_{i,j}, \mathbf{F}_{i,j}, \mathbf{G}_{i,j}\right) \qquad \mathbf{G}_{i,j} \doteq \begin{cases} x_i = y_j \ : \ \mathbf{V}_{i-1,j-1} + \alpha \\ x_i \neq y_j \ : \ \mathbf{V}_{i-1,j-1} - \beta \end{cases}$$

$$\mathbf{E}_{i,j} \doteq \max\left(\mathbf{E}_{i,j-1} - \tau, \mathbf{V}_{i,j-1} - \gamma - \tau\right) \quad \mathbf{F}_{i,j} \doteq \max\left(\mathbf{F}_{i-1,j} - \tau, \mathbf{V}_{i-1,j} - \tau - \gamma\right)$$

$$\alpha = \beta = \gamma = \tau = 1$$

		A	T	A	T	T	G
	0	-2	-3	-4	-5	-6	-7
	-2	-2	-3	-4	-5	-6	-7
A	-2	1 1	-1 -3	-2 -2	-3 -5	-4 -6	-4 -7
	-2	-4 -4	-1 -5	-2 -6	-3 -7	-4 -8	-5 -9
A	-3	-1 -1					
	-3	-5 -1					
G	-4	-2 -4					
	-4	-6 -2					

The remainder of row and column 1 proceed as does cell 1-1.

6 Filling Cell 2-2

$$\mathbf{V}_{i,j} \doteq \max\left(\mathbf{E}_{i,j}, \mathbf{F}_{i,j}, \mathbf{G}_{i,j}\right) \qquad \mathbf{G}_{i,j} \doteq \begin{cases} x_i = y_j & : \mathbf{V}_{i-1,j-1} + \alpha \\ x_i \neq y_j & : \mathbf{V}_{i-1,j-1} - \beta \end{cases}$$

$$\mathbf{E}_{i,j} \doteq \max\left(\mathbf{E}_{i,j-1} - \tau, \mathbf{V}_{i,j-1} - \gamma - \tau\right) \quad \mathbf{F}_{i,j} \doteq \max\left(\mathbf{F}_{i-1,j} - \tau, \mathbf{V}_{i-1,j} - \tau - \gamma\right)$$

$$\alpha = \beta = \gamma = \tau = 1$$

		A	T	A	T	T	G
	0	-2	-3	-4	-5	-6	-7
	-2	1 / -4	-1 / -5	-2 / -6	-3 / -7	-4 / -8	-4 / -5
A	-2	1 / -4	-1 / -1	-2 / -2	-3 / -3	-4 / -4	-5 / -5
	-3	-1 / -5	0 / -3				
A	-3	-1 / -5	0 / -3				
	-4	-2 / -6					
G	-4	-2 / -6					

Here we can either open new gaps or extend existing ones: all initialization and $-\infty$ have now been handled.

7 Completing the Matrix

$$\mathbf{V}_{i,j} \doteq \max\left(\mathbf{E}_{i,j}, \mathbf{F}_{i,j}, \mathbf{G}_{i,j}\right) \qquad \mathbf{G}_{i,j} \doteq \begin{cases} x_i = y_j \ : \ \mathbf{V}_{i-1,j-1} + \alpha \\ x_i \neq y_j \ : \ \mathbf{V}_{i-1,j-1} - \beta \end{cases}$$

$$\mathbf{E}_{i,j} \doteq \max\left(\mathbf{E}_{i,j-1} - \tau, \mathbf{V}_{i,j-1} - \gamma - \tau\right) \quad \mathbf{F}_{i,j} \doteq \max\left(\mathbf{F}_{i-1,j} - \tau, \mathbf{V}_{i-1,j} - \tau - \gamma\right)$$

$$\alpha = \beta = \gamma = \tau = 1$$

		A		T		A		T		T		G	
	0		-2		-3		-4		-5		-6		-7	
			-2		-3		-4		-5		-6		-7	
A	-2		1	1	-1	-3	-2	-2	-3	-5	-4	-6	-4	-7
		-2	-4	-4	-1	-5	-2	-6	-3	-7	-4	-8	-5	-9
A	-3		-1	-1	0	0	0	0	-2	-3	-3	-4	-4	-5
		-3	-5	-1	-3	-3	-2	-4	-2	-5	-3	-6	-4	-6
G	-4		-2	-4	-2	-2	-1	-1	-1	-1	-3	-3	-2	-2
		-4	-6	-2	-4	-2	-4	-2	-3	-4	-3	-5	-4	-6

The rest of the matrix proceeds in the same manner.

8 Tracing Back the Alignment

$$\mathbf{V}_{i,j} \doteq \max\left(\mathbf{E}_{i,j}, \mathbf{F}_{i,j}, \mathbf{G}_{i,j}\right) \qquad \mathbf{G}_{i,j} \doteq \begin{cases} x_i = y_j \ : \ \mathbf{V}_{i-1,j-1} + \alpha \\ x_i \neq y_j \ : \ \mathbf{V}_{i-1,j-1} - \beta \end{cases}$$

$$\mathbf{E}_{i,j} \doteq \max\left(\mathbf{E}_{i,j-1} - \tau, \mathbf{V}_{i,j-1} - \gamma - \tau\right) \quad \mathbf{F}_{i,j} \doteq \max\left(\mathbf{F}_{i-1,j} - \tau, \mathbf{V}_{i-1,j} - \tau - \gamma\right)$$

$$\alpha = \beta = \gamma = \tau = 1$$

		A	T	A	T	T	G
	0	-2	-3	-4	-5	-6	-7
		-2	-3	-4	-5	-6	-7
A	-2	1 → 1	-1	-3	-2	-2	-3	-5	-4	-6	-4	-7
	-2	-4	-4	-1	-5	-2	-6	-3	-7	-4	-8	-5	-9
A	-3	-1	-1	0	0	0 → 0	-2	-3	-3	-4	-4	-5
	-3	-5	-1	-3	-3	-2	-4	-2	-5	-3	-6	-4	-6
G	-4	-2	-4	-2	-2	-1	-1	-1	-1	-3	-3	-2 → -2
	-4	-6	-2	-4	-2	-4	-2	-3	-4	-3	-5	-4	-6

```
ATATTG
A_A__G
```

9 Significance of Gapped Alignment

We saw that the optimal alignment had cost -2, and contained 2 gap clusters. We could have gotten this alignment with linear gap penalties: note that we *require* at least 3 gaps, so this alignment is the also the optimal linear gapped alignment[1].

However, consider what would happen if we had a higher gap open cost: take $\gamma = 3$. Then, we would have the following alignment score:

$$3\alpha - 2\gamma - 3\tau = -6$$

On the other hand, this alignment would have a higher alignment score:

```
ATATTG
AA___G
```

$$2\alpha - \beta - \gamma - 3\tau = -5$$

In this case, using affine gap penalties, we obtain an alignment that would be *impossible* with linear gapped alignment. So we see that affine gap penalties can be used to produce alignments that favor a smaller number of gap clusters, even at the cost of additional mismatches.

[1]With score $3\alpha - 3\tau$, we can see that this is in fact optimal: no alignment may have fewer than 3 gaps, and no alignment may have more than 3 matches.