CSCI1950L: Algorithmic Foundations of Computational Biology February 26, 2014

Supplementary Hirschberg Notes

Lecturer: Sorin Istrail *Scribe: Derek Aguiar*

1 Hirschberg Algorithm Intuition

The section from Gusfield's book handed out in class give a detailed explanation of this algorithm (available on the class webpage). Below, we give a brief introduction to the ideas used in the algorithm. The general space saving idea is we only need to store the $i-1$ row in order to compute the i^{th} row. The algorithm relies on the idea that one can find an index k^* which the optimal alignment passes through by computing the normal alignment DP for the first $\frac{n}{2}$ iterations and then for the reverse strings corresponding to the indices n to $\frac{n}{2}$. The reason why we are able to do this is essentially due to the following lemma:

Lemma 1 $V(n,m) = MAX_{0\leq k\leq m}[V(\frac{N}{2},k) + V^r(\frac{N}{2}, m-k)]$

The subpath defined by the traceback pointers for the forward and reverse DP define the path the optimal alignment takes. This is stored and then the two areas of the DP matrix still possibly containing the optimal alignment are recursed on.

2 Hirschberg Algorithm

Let us consider an $n\times m$ DP matrix. For alignment, the computation of a cell only requires the $(i-1,j-1)$, $(i-1,j)$, $(i,j-1)$ cells thus we can compute the score of any cell in linear space by only storing the previous row and current row. This may be reduced to one row with clever bookkeeping. The selection of row or column can be made such that you iterate row by row if $n >> m$ and column by column if $m >> n$.

The first iteration will compute the alignment from row 1 to the $\frac{n}{2}$ row and back from n to the $\frac{n}{2}$ row. Consider the $\frac{n}{2}$ row.

```
n/2-1      --------------------
                        \
n/2        --------------------
                         |
n/2+1      --------------------
```

Complexity. $\sum_{k=1}^{\infty}\alpha\delta^k = \frac{a}{1-\delta}$
$\frac{1}{2}\sum_{k=0}^{\infty}\frac{1}{2}^k = \frac{1}{2}\frac{1}{1-\frac{1}{2}} = 1$
We then have 2 subproblems each of size $\frac{N}{2}M$ so $cNM + c\frac{N}{2}M + c\frac{N}{4}M ... = \sum_{i=1}^{log(N)} c\frac{N}{2^{i-1}}M \leq 2CMN$

3 Readings

1. A great book on string algorithms with many applications to computational biology[1]. Only the chapter covering the Hirschberg algorithm is posted online.

References

[1] Carlos Setubal and Joao Meidanis, *Introduction to computational molecular biology*, PWS Publishing, January 1997.