A Linear Space Local Alignment Algorithm for TTN Gene Sequencing

Hamilton Evans and Bruce Atwood

Middlebury College CSCI 0321 Bioinformatics Algorithms, Spring 2019

Goal

Our goal was to implement a linear space local alignment algorithm for protein sequencing of the TTN gene in humans compared to certain animal species.

Background

An important problem in genomics is the identification of maximally homogenous subsequences of DNA among sets of longer sequences. This allows biologists to identify highly conserved regions, as well as variations among different genomes. By sequencing the TTN gene[3], a gene which helps to control movement of skeletal muscles and heart muscles, we can learn more about humans physical relations to other animals.

The Smith-Waterman Algorithm is a local alignment algorithm which performs in O(nm) time, with O(nm) space complexity, where n and m are the lengths of the sequences[1]. However, when aligning long sequences (100,000+ proteins), this space requirement can become quickly overwhelming.

Our implementation uses Hirschberg’s Algorithm, a linear space global alignment algorithm. We combined both a modified Smith-Waterman Algorithm and Hirschberg's Algorithm to run linear space local alignment, while retaining similar time complexity to that of the Smith-Waterman Algorithm[2]. To test this method, we aligned protein sequences of the TTN gene in humans against that of monkeys, dogs, and cats. Both results from the TTN gene comparison and comparison of our algorithm versus the Smith-Waterman Algorithm were reported.

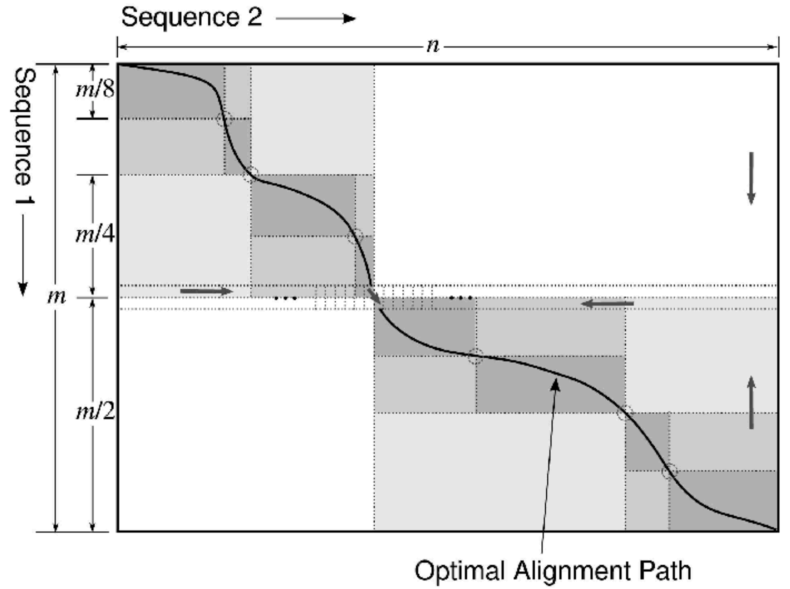


Figure 6. Hirschberg Algorithm Storage Visualized.

Linear Space Local Alignment Algorithm

To run local alignment, we first wanted to find the area of the whole sequence used in the local alignment. To do this, we ran the Smith-Waterman Algorithm with two rows, one for the information of the previous row and the other for the current row, keeping track of the best score seen and the corresponding cell (end points). This was a simple modification from our original code for local alignment. The starting points were more difficult. For the starting points, each cell initially starts with itself. When an alignment path continues, each cell copies the starting point from its preceding cell on the path. The starting point is set to a cell itself whenever a score zero is reached.

With start and endpoints of the portion of the sequence used in the local alignment, the Hirschberg Algorithm (linear space global alignment) can be applied. Scores were given based on the pam250 scoring matrix. Hirschberg Algorithm pseudocode can be seen to the right.

Results

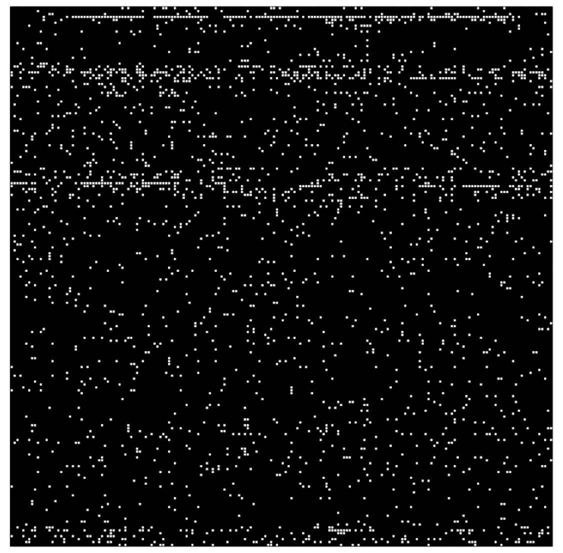


Figure 1 Human vs Dog Protein TTN Alignment

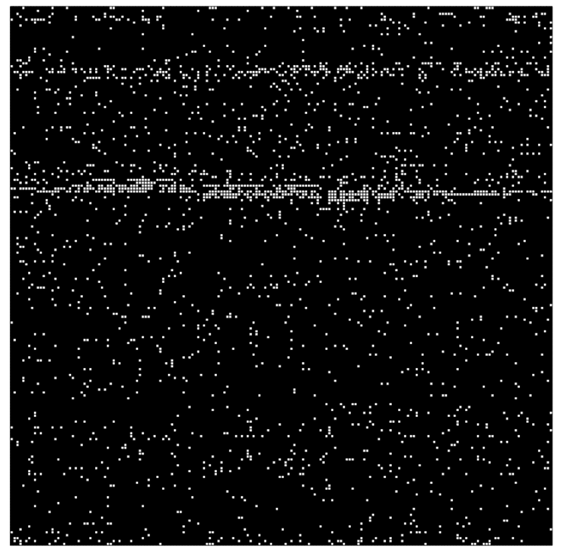


Figure 2. Human vs Cat Protein TTN Alignment

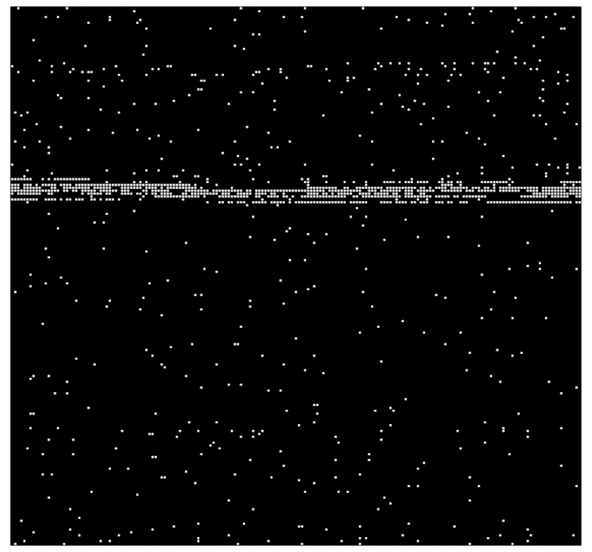


Figure 3. Human vs Monkey Protein TTN Alignment

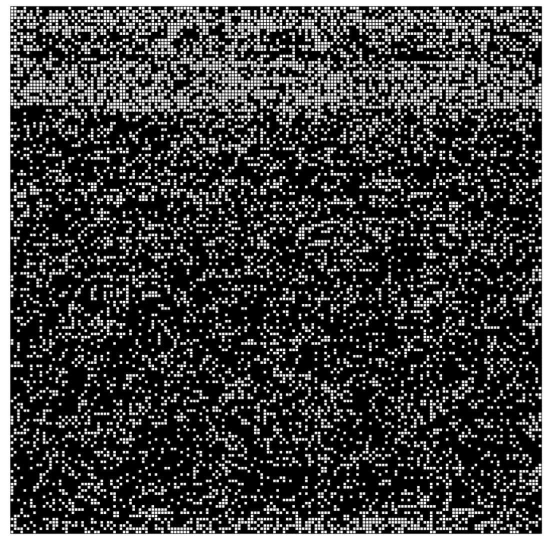
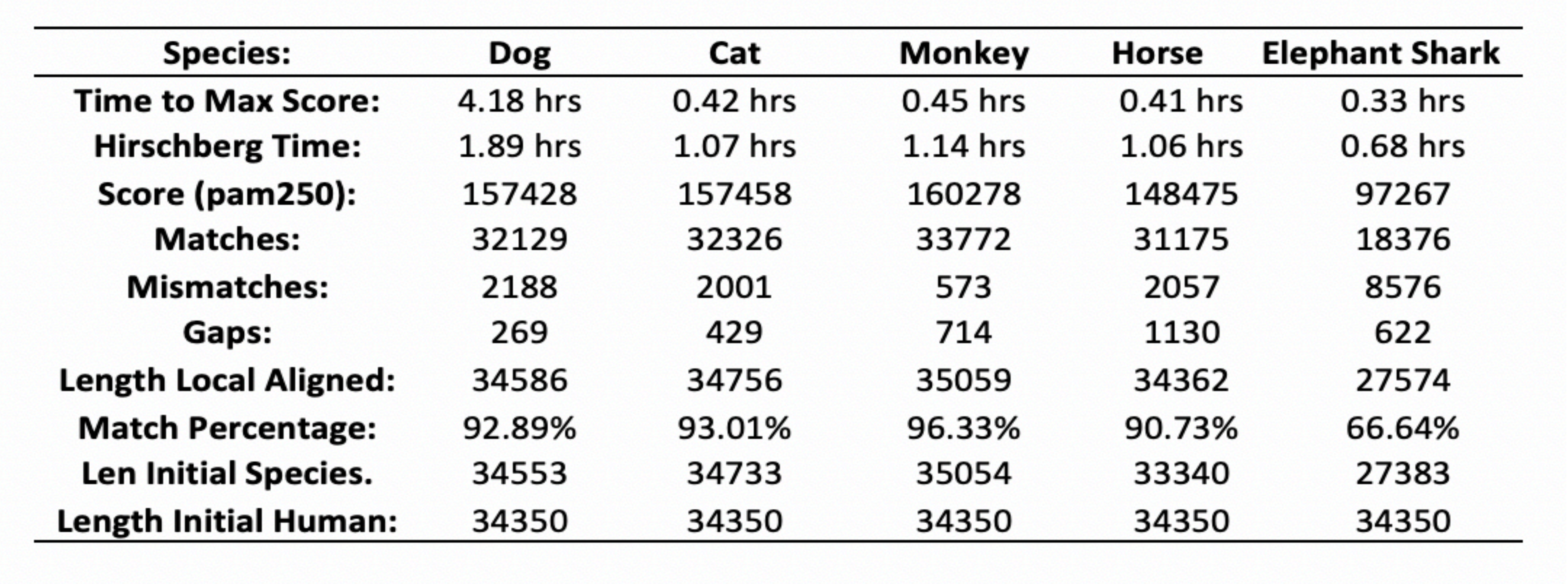


Figure 4 Human vs Elephant Shark Protein TTN Alignment



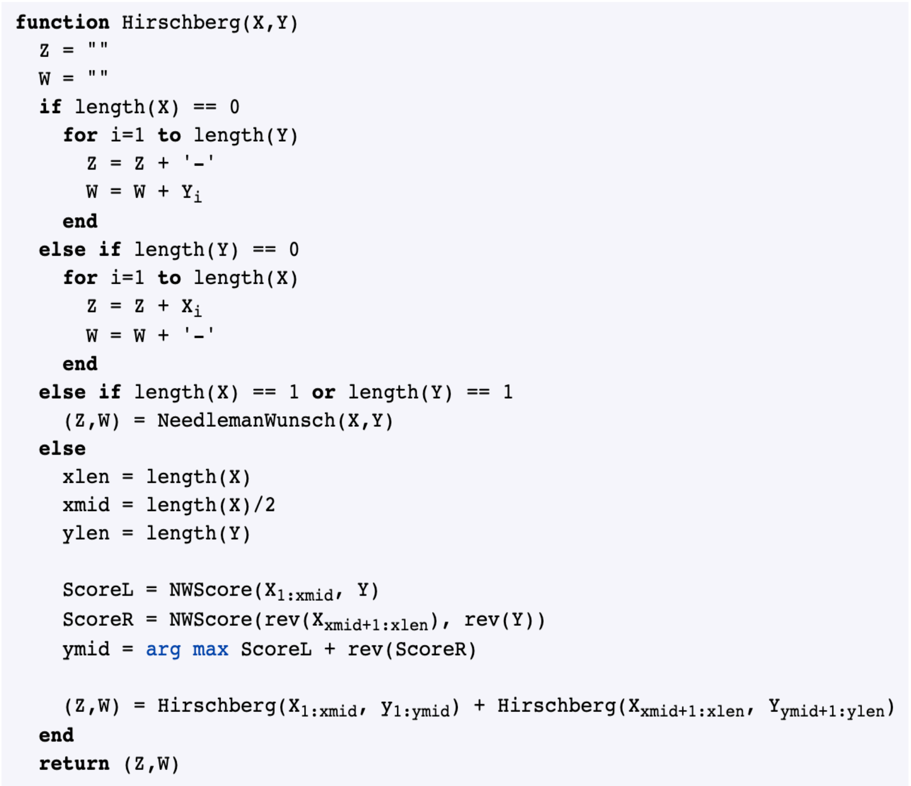
Species:	Dog	Cat	Monkey	Horse	Elephant Shark
Time to Max Score:	4.18 hrs	0.42 hrs	0.45 hrs	0.41 hrs	0.33 hrs
Hirschberg Time:	1.89 hrs	1.07 hrs	1.14 hrs	1.06 hrs	0.68 hrs
Score (pam250):	157428	157458	160278	148475	97267
Matches:	32129	32326	33772	31175	18376
Mismatches:	2188	2001	573	2057	8576
Gaps:	269	429	714	1130	622
Length Local Aligned:	34586	34756	35059	34362	27574
Match Percentage:	92.89%	93.01%	96.33%	90.73%	66.64%
Len Initial Species.	34553	34733	35054	33340	27383
Length Initial Human:	34350	34350	34350	34350	34350

Table 1. Human vs Dog, Cat, Monkey, Horse, and Elephant Shark, Protein TTN Sequence Results

References:

1. *Smith, Temple F. & Waterman, Michael S. (1981). "Identification of Common Molecular Subsequences" (PDF). Journal of Molecular Biology. **147** (1): 195–197. CiteSeerX 10.1.1.63.2897. doi:10.1016/0022-2836(81)90087-5. PMID 7265238.*
2. *Hirschberg, D. S. (1975). "A linear space algorithm for computing maximal common subsequences". Communications of the ACM. **18** (6): 341–343. CiteSeerX 10.1.1.348.4774. doi:10.1145/360825.360861. MR 0375829*
3. Reference, G. H. TTN gene. *Genetics Home Reference* Available at: https://ghr.nlm.nih.gov/gene/TTN.
4. ROSALIND | Find a Highest-Scoring Local Alignment of Two Strings. Available at: http://rosalind.info/problems/ba5f/.
5. Ortholog_gene_7273[group] - Gene - NCBI. Available at: https://www.ncbi.nlm.nih.gov/gene/?Term=ortholog_gene_7273[group].

Hirschberg Algorithm



```
function Hirschberg(X,Y)
  Z = ""
  W = ""
  if length(X) == 0
    for i=1 to length(Y)
      Z = Z + '-'
      W = W + Y_i
    end
  else if length(Y) == 0
    for i=1 to length(X)
      Z = Z + X_i
      W = W + '-'
    end
  else if length(X) == 1 or length(Y) == 1
    (Z,W) = NeedlemanWunsch(X,Y)
  else
    xlen = length(X)
    xmid = length(X)/2
    ylen = length(Y)

    ScoreL = NWScore(X_1:xmid, Y)
    ScoreR = NWScore(rev(X_xmid+1:xlen), rev(Y))
    ymid = arg max ScoreL + rev(ScoreR)

    (Z,W) = Hirschberg(X_1:xmid, y_1:ymid) + Hirschberg(X_xmid+1:xlen, Y_ymid+1:ylen)
  end
  return (Z,W)
```

Figure 4. Hirschberg Algorithm Pseudocode

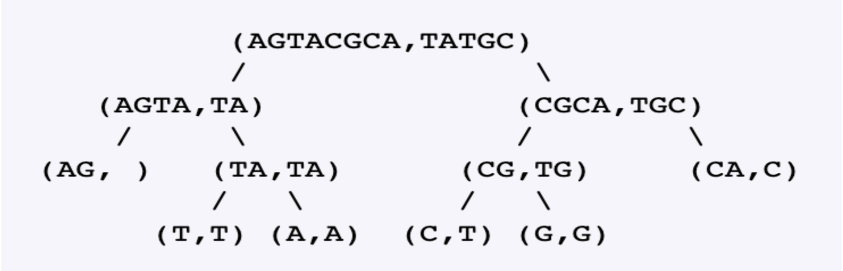


Figure 5. Hirschberg Algorithm Visualization

DISCUSSION

-A space optimized local alignment algorithm was created by using a modified Smith-Waterman Algorithm and the Hirschberg Algorithm

-Our results for both max score and length matched that of the Smith-Waterman algorithm, proving it was correctly finding the Local Alignment while using optimized space.

-Though efforts were made, we were unsuccessful in actually cataloging the space used by the program (to prove it was linear space complexity)

-In looking at the results, as expected, humans TTN gene sequence is closest to that of monkeys. It isn’t far off from dogs, cats, and horses, but is quite different than that of elephant sharks. These can be seen as visualized in Figure 1, Figure2, Figure 3, and Table 1.

-1/3 of the way through, large sequence of gaps and mismatches, implying that is the location of difference between animal species and humans.