Comp 555 - BioAlgorithms - Spring 2019

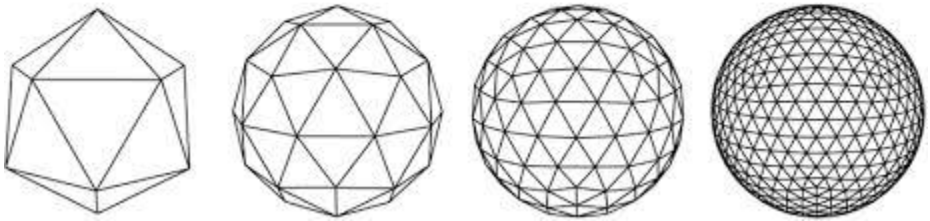
"Really? — my people always say *multiply* and conquer."

- Problem Set #3 is due tonight
- Problem Set #4 is posted

Divide and Conquer Algorithms

The Essence of Divide and Conquer

- Divide problem into sub-problems
- Conquer by solving sub-problems recursively.
  - If the sub-problems are small enough, solve them in brute force fashion
- Combine the solutions of sub-problems into a solution of the original problem
  - This is the tricky part



Divide and Conquer Applied to Sorting

Problem

- Given an unsorted array of items

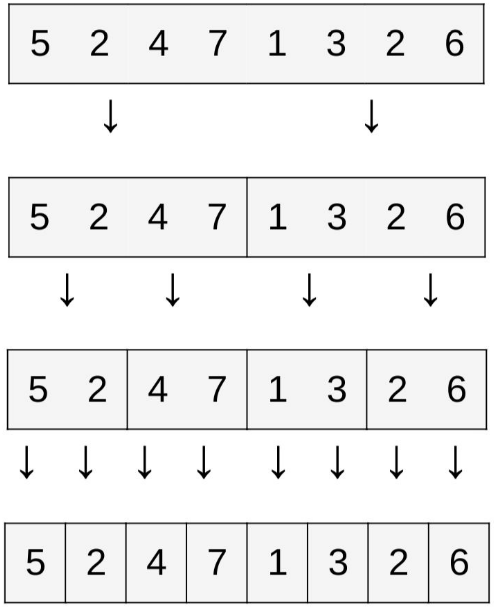
5	2	4	7	1	3	2	6

- Reorder them such that they are in a non-decreasing order

1	2	2	3	4	5	6	7

Merge Sort

Step 1. The Divide Phase

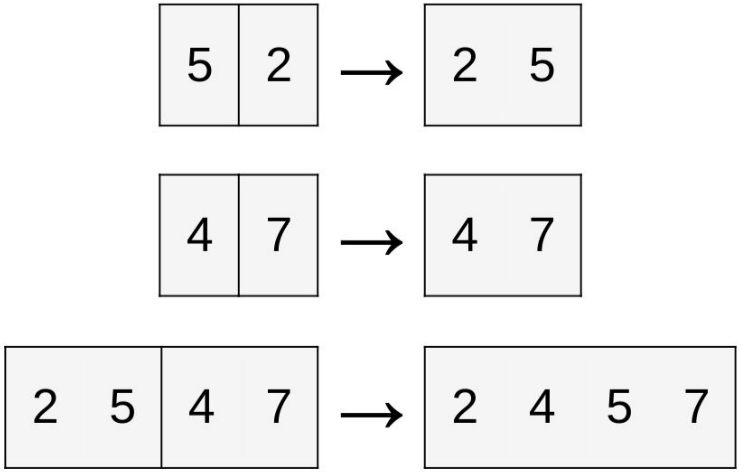


$log_2(n)$ divisions to split an array of size n into single elements

Merge Sort

Merging

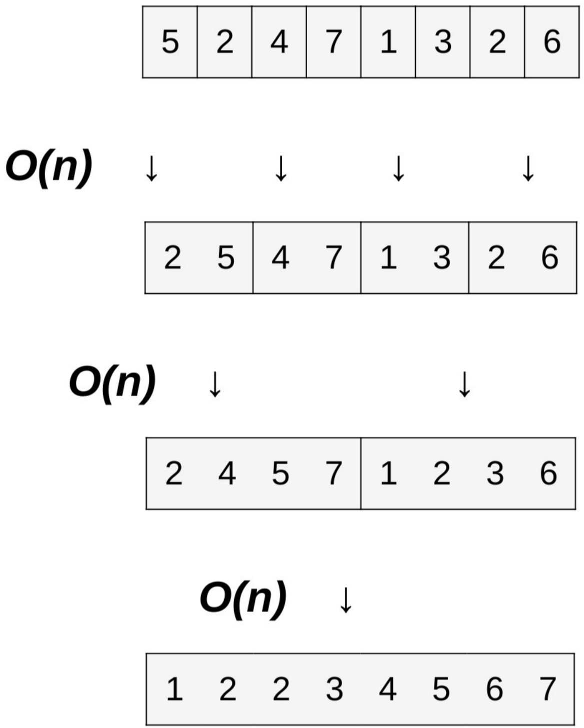
- 2 arrays of size 1 can be easily merged to form a sorted array of size 2



- Move the smaller first value of the two arrays to the next slot in the merged array. Repeat.
- 2 sorted arrays of size p and q can be merged in $O(p+q)$ time to form a sorted array of size p+q

Merge Sort

Step 2. Conquer Phase



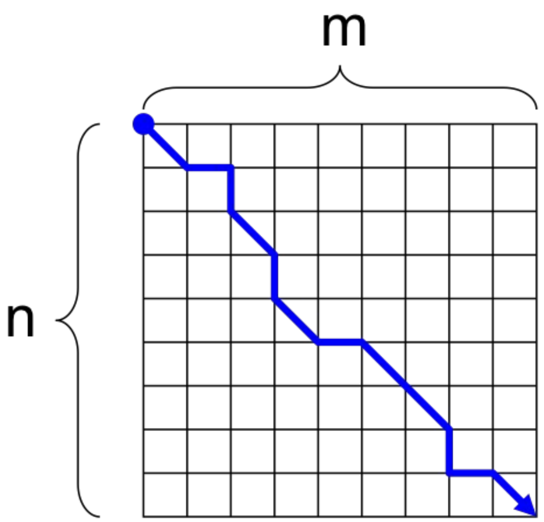
$log_2(n)$ iterations, each iteration takes $O(n)$ time, for a total time $O(n\ log_2(n))$

Now back to Biology

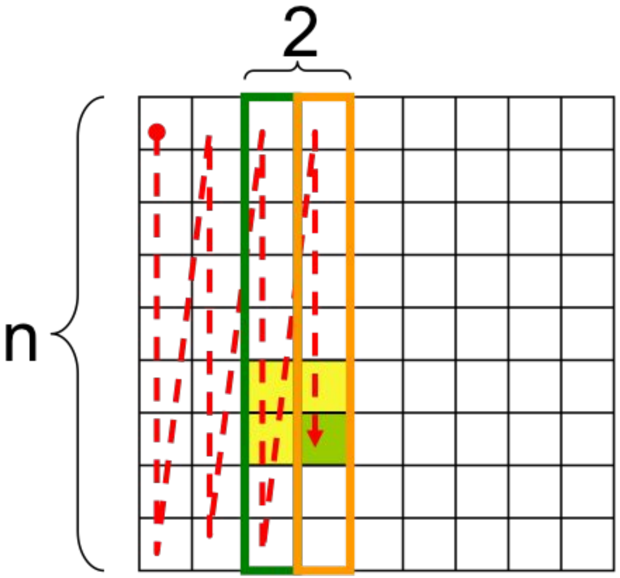
All algorithms for aligning a pair of sequences thus far have required *quadratic memory*

The tables used by the dynamic programming method

- Space complexity for computing alignment path for sequences of length *n* and *m* is *O(nm)*
- We kept a table of all scores and arrival directions in memory to reconstruct the final best path (backtracking)



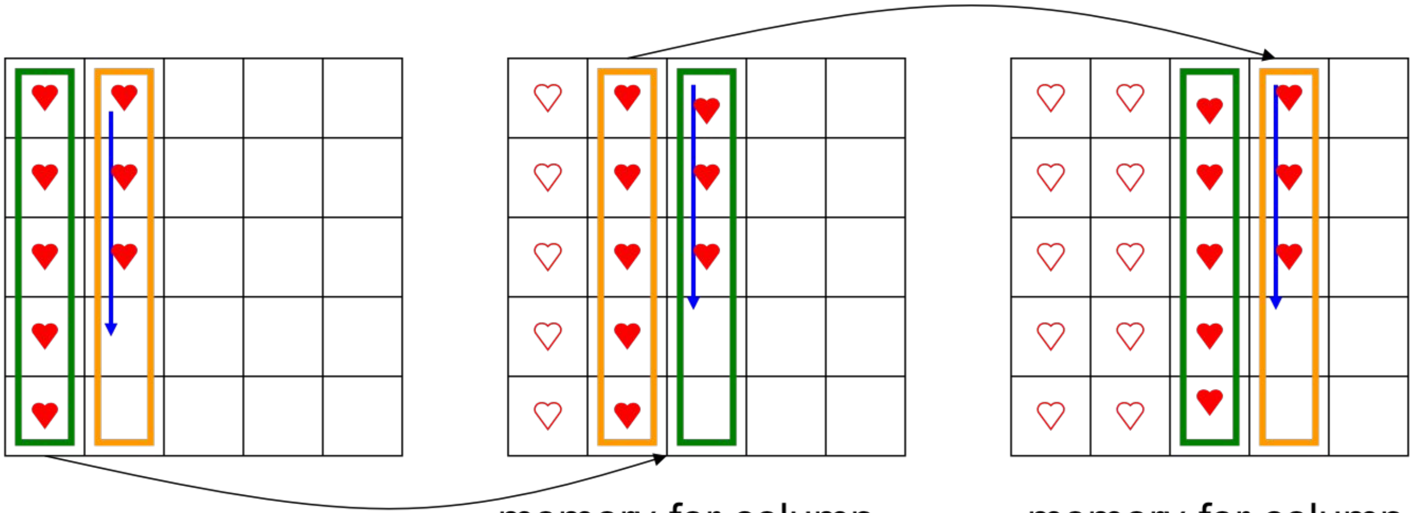
Computing Alignments with Linear Memory



- If appropriately ordered, the space needed to compute ***just the score*** can be reduced to O(n)
- For example, we only need the previous column to calculate the current column, and we can throw away that previous column once we're done using it

Recycling Columns

Only two columns of scores are needed at any given time



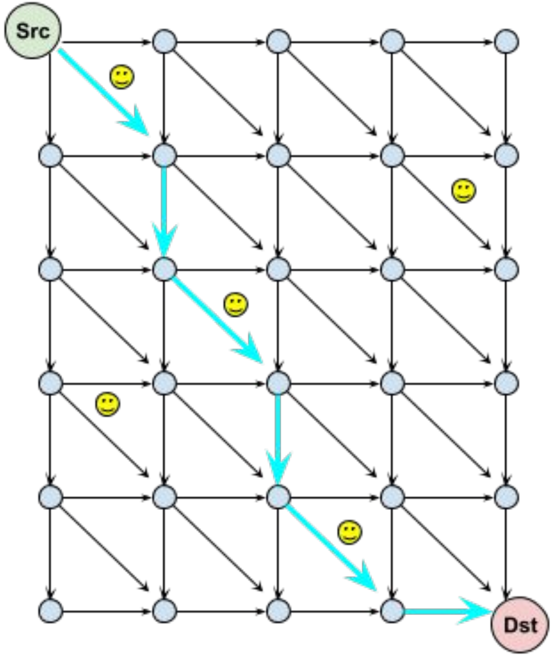
memory for column 1 is used to calculate column 3

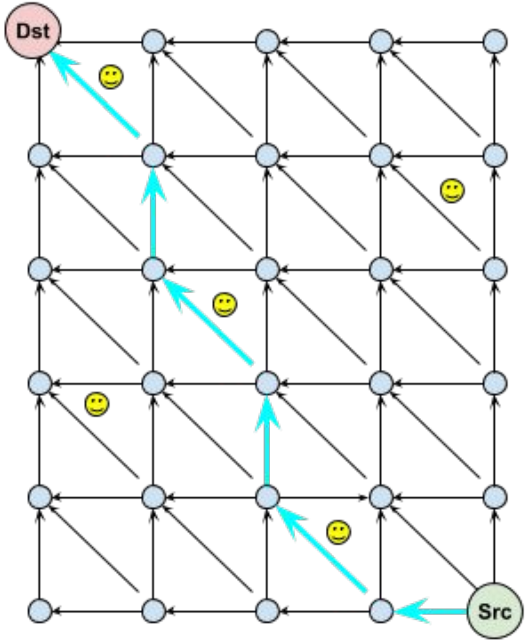
memory for column 2 is used to calculate column 4

An Aside

Suppose that we reverse the source and destination of our Manhattan Tour

- Does the path with the most attractions change?

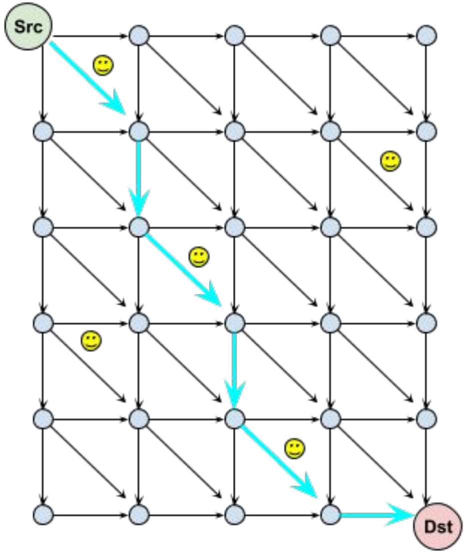


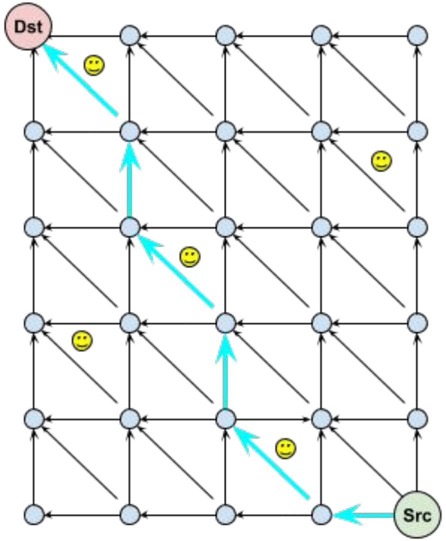


More Aside

Now suppose that we made two tours

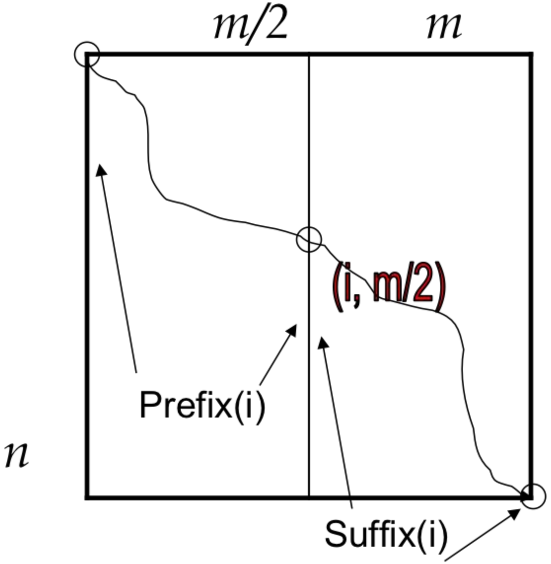
- One from the source towards the destination
- A second from the destination of towards the source
- And we stop both tours at the middle column





- Can we combine these two separate solutions to find the overall best score?

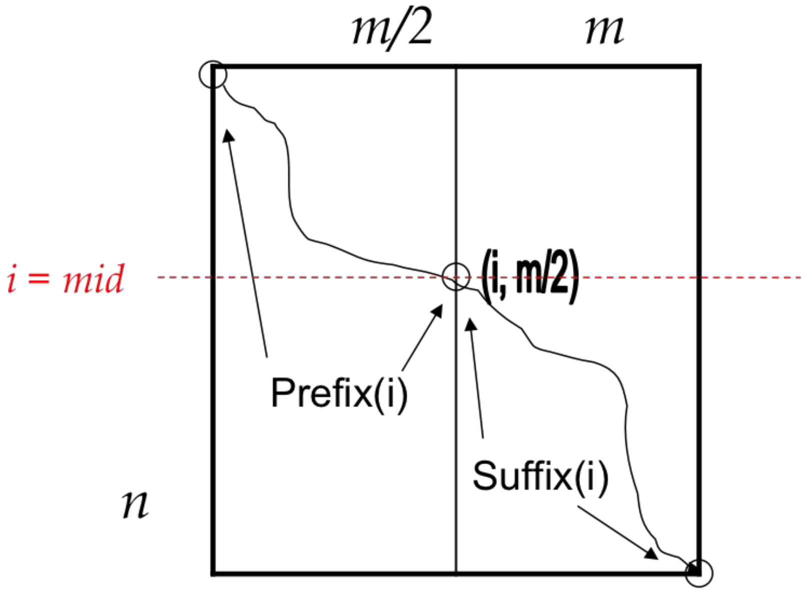
A Divide & Conquer Alignment Approach



- We want to calculate the longest path from (0,0) to (n,m) that passes through (i,m/2) where i ranges from 0 to n and represents the i-th row
- Define Score(i) as the score of the path from (0,0) to (n,m) that passes through vertex (i, m/2)

Finding the Midline

Define (mid,m/2) as the vertex where the best score crosses the middle column.

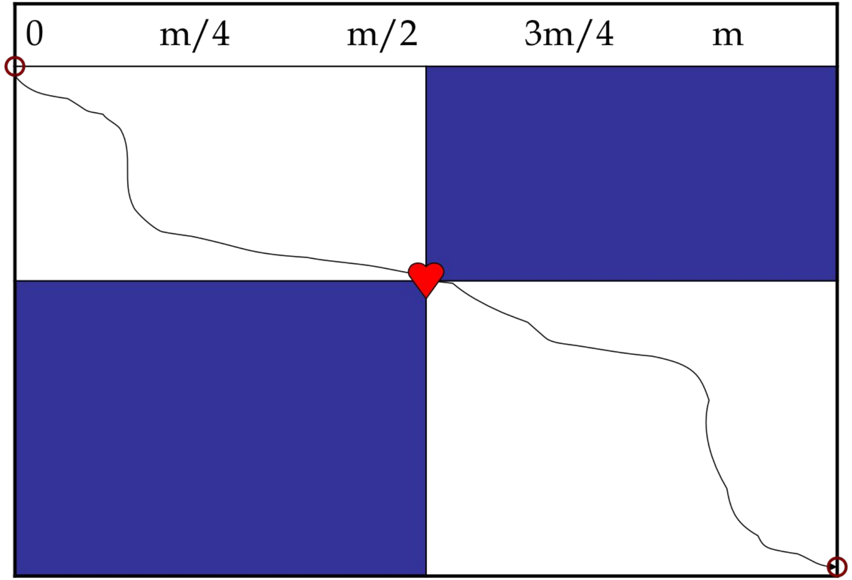


- How hard is the problem compared to the original DP approach?
- What does it lack?

We know the Best Score

How do we find the best path?

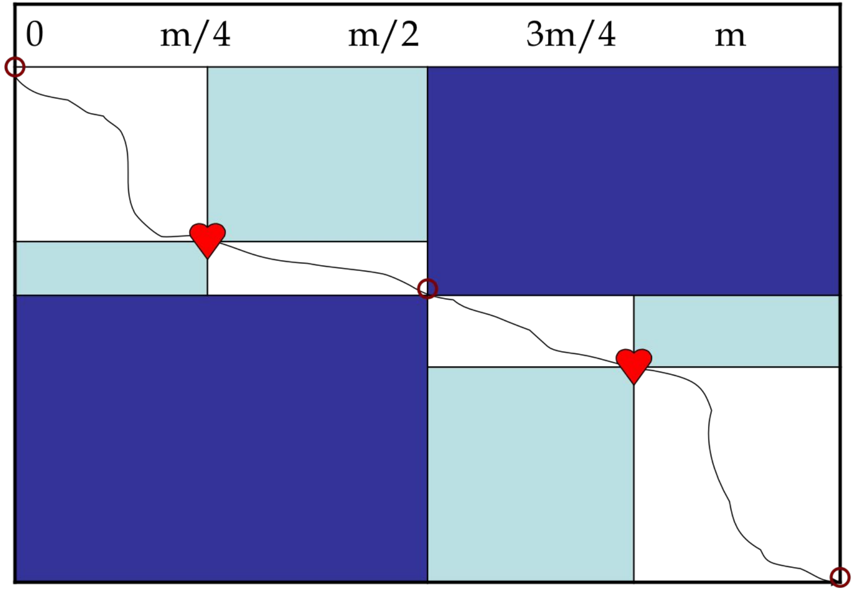
- We actually know one vertex on our path, (m/2, mid).
- How do we find more?



- **Hint:** Knowing *mid* actually constrains where the paths can go

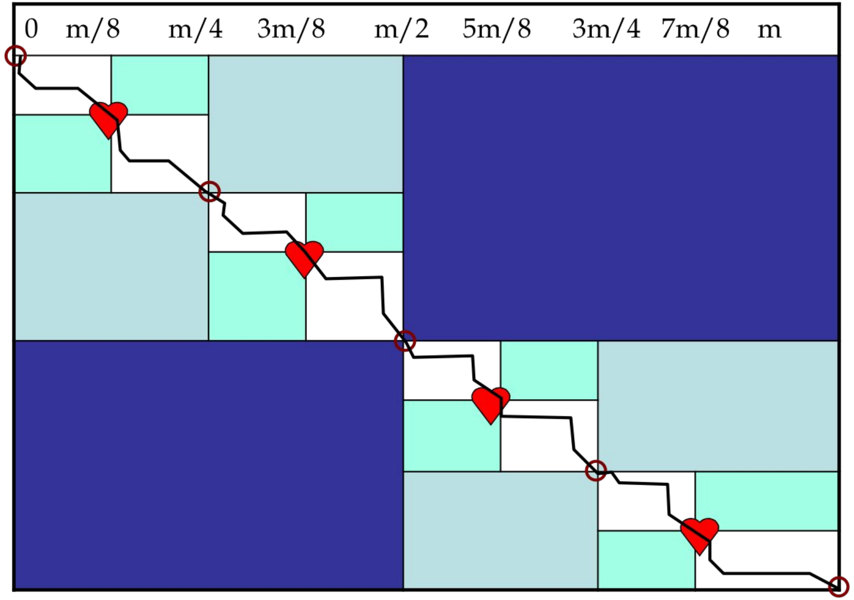
A Mid's Mid

We can now solve for the paths from (0,0) to (m/2, mid) and (m/2, mid) to (m,n)



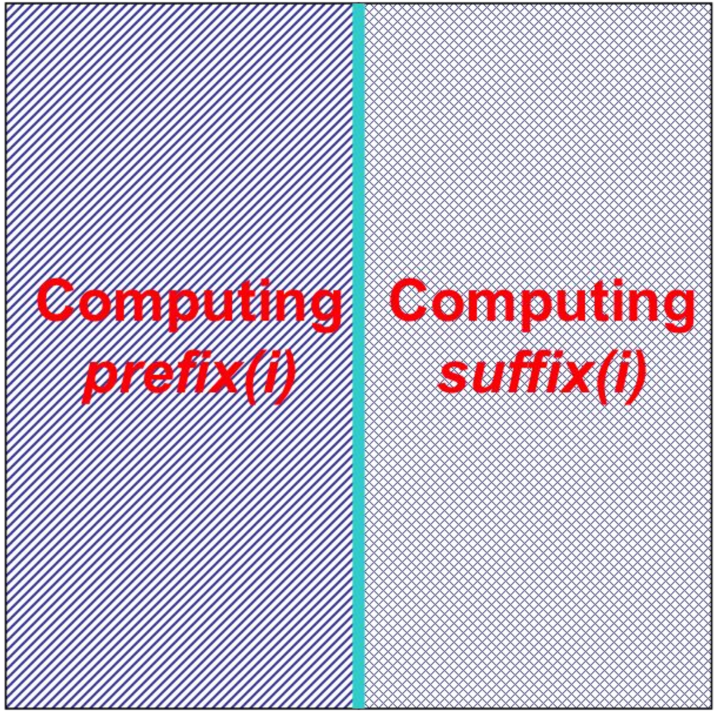
And Mid-Mid's Mids (recursively)

And repeat this process until the path is from (i,j) to (i,j)



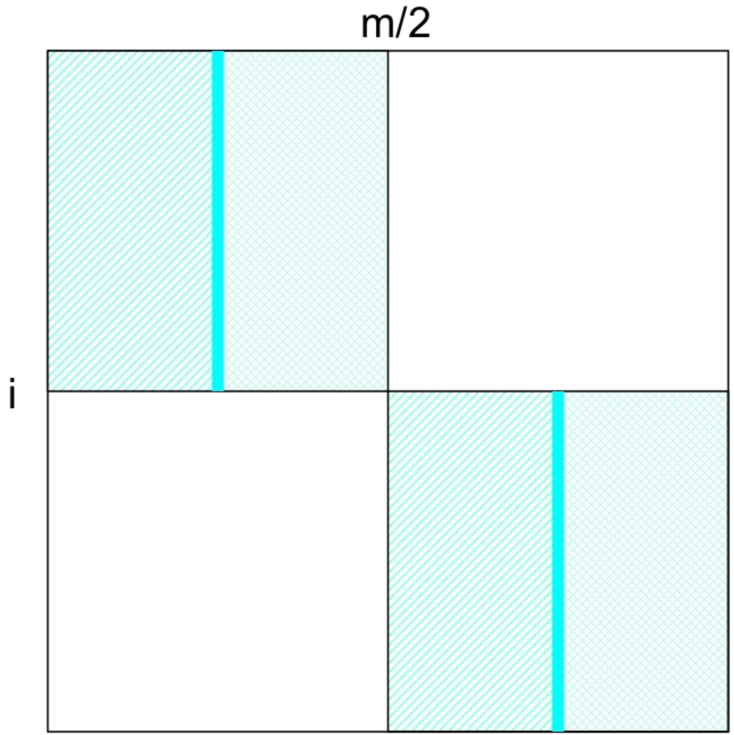
Algorithm's Performance

- On the first level, the algorithm fills every entry in the matrix, thus it does O(nm) work



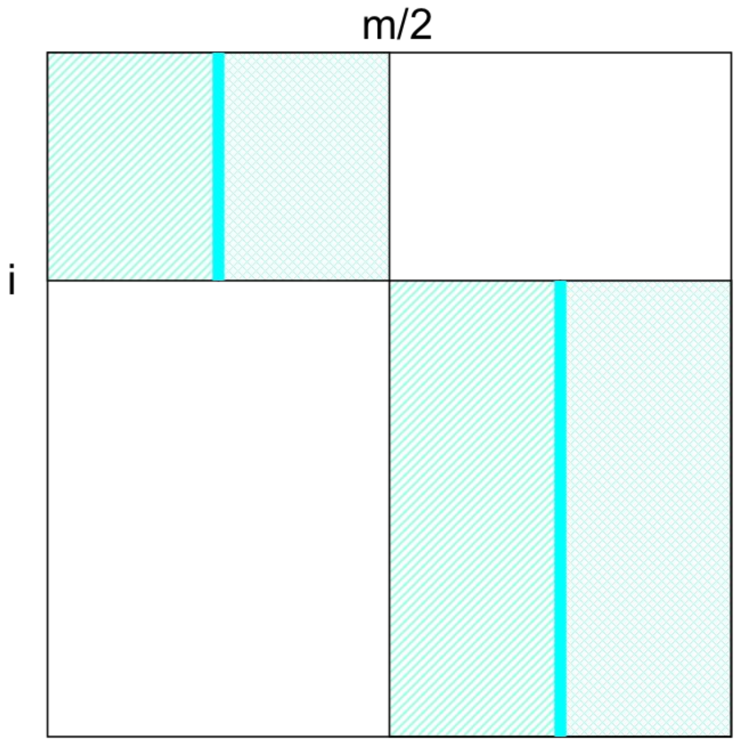
Work done on a second pass

- On second level, the algorithm fills half the entries in the matrix, thus it does O(nm)/2 work



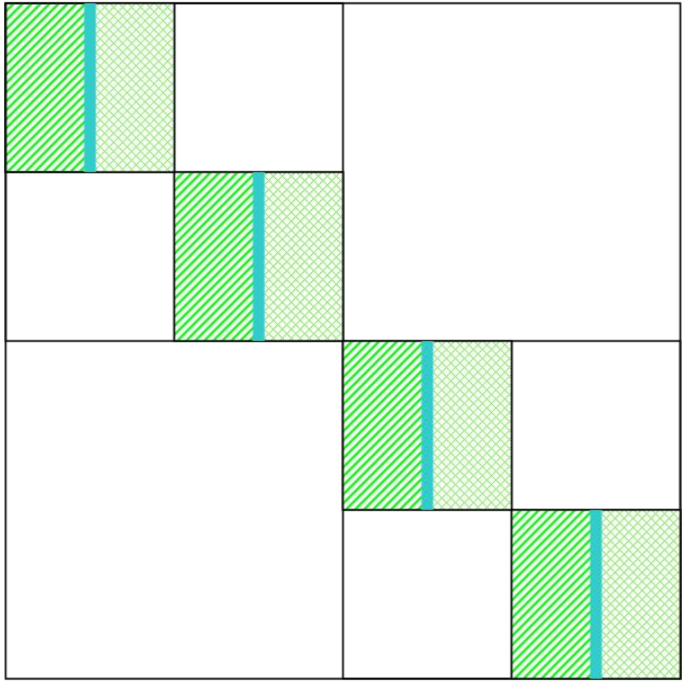
Work done on an Alternate second pass

- This is true regardless of what *mid* is



Work done on a third pass

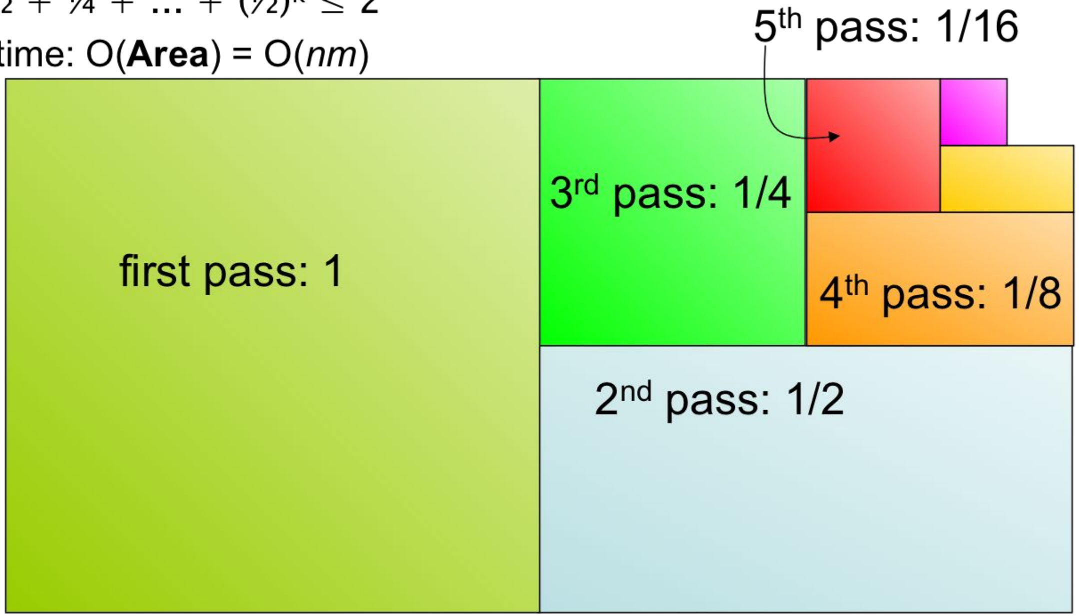
- On the third pass, the algorithm fills a quarter of the entries in the matrix, thus it does O(nm)/4 work



Sum of a Geometric Series

$1 + \frac{1}{2} + \frac{1}{4} + \ldots + (\frac{1}{2})^k \leq 2$

- Runtime: O(**Area**) = O(*nm*)



- Total Space: O(n) for score computation, O(n+m) to store the optimal alignment

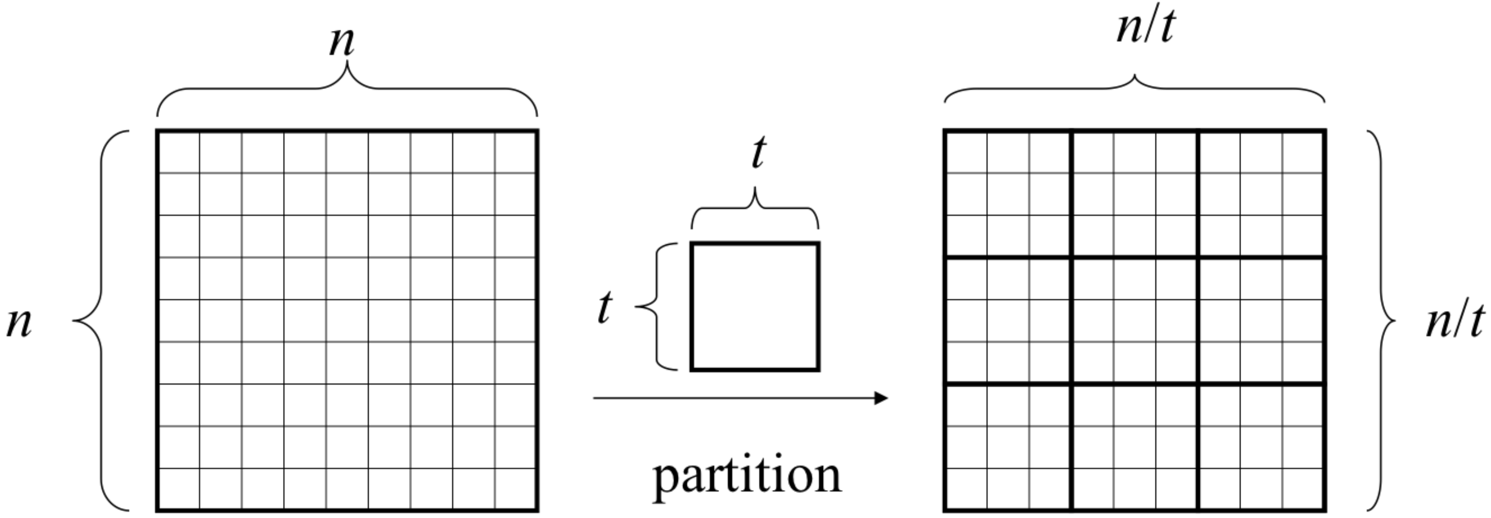
- Time complexity is still O(mn). Actually, we expect it to take about twice as long as the approach using O(mn) space

Can We Do Even Better?

- Align in Subquadratic Time?
- Dynamic Programming takes *O(nm)* for global alignment, which is quadratic assuming n ≈ m
- Yes, using the Four-Russians Speedup

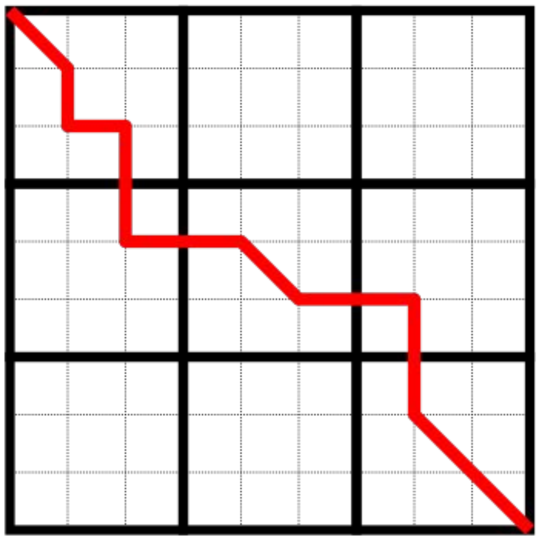
Partitioning the Alignment Grid

Into smaller blocks



Block Logic

- How does a block relate to a correct alignment?
    - the alignment path passes through block
    - the path does not use the block
- The alignment passes through *O(n/t)* total blocks
- Paths enter from the top or left and exit from the right or bottom
- If we know the best score at the boundaries, perhaps we can peice together a solution as we did before.



Recall our Bag of DP Tricks

- A key insight of dynamic programming was to reuse repeated computations by storing them in a tableau
- Are there any repeated computations in Block Alignments?
- Let’s check out some numbers…
    - Lets assume n = m = 4000 and t = 4
    - n/t = 1000, so there are 1,000,000 blocks
    - How many possible many blocks are there?
        - Assume we are aligning DNA with DNA, so there sequences are over an alphabet of {A,C,G,T}
        - Possible sequences are $4^t = 4^4 = 256$,
        - Possible alignments are $4^t \times 4^t = 65536$
- There are fewer possible alignments than blocks, thus we must be frequently revisiting block alignments!

		A	C	A	T
	v	*w*	*x*	*y*	*z*
A	*u*	max(u-1,v+1,w-1)	max(max(u-1,v+1,w-1)-1,w-1,x-1)	f(u,v,w,x,y)	f(u,v,w,x,y,z)
T	*t*	max(t-1,u-1,max(u-1,v+1,w-1)-1)	f(t,u,v,w,x)	f(t,u,v,w,x,y)	f(t,u,v,w,x,y,z)
A	*s*	f(s,t,u,v,w)	f(s,t,u,v,w,x)	f(s,t,u,v,w,x,y)	f(s,t,u,v,w,x,y,z)
G	*r*	f(r,s,t,u,v,w)	f(r,s,t,u,v,w,x)	f(r,s,t,u,v,w,x,y)	f(r,s,t,u,v,w,x,y,z)

- max(max(u-1,v+1,w-1)-1,w-1,x-1) → max(u-2,v,w-2,w-1,x-1) → max(u-2,v,w-1,x-1)
- max(t-1,u-1,max(u-1,v+1,w-1)-1) → max(t-1,u-1,u-2,v,w-2) → max(t-1,u-1,v,w-2)
- All functions are maxs of the 7 block inputs (r,s,t,u,v,w,x,y,z), which can be precomputed.

Next Time

HIdden Markov Models