Dylan Han

CS 466

December 19, 2021

Nussinov Implementation in Python

Introduction

Nussinov algorithm is a $O(n^3)$ runtime algorithm which may be used to predict the secondary structure of nucleic acids. The algorithm works by trying to maximize the number of base pairs with the assumption that the most stable structure being the most likely. Below is the recurrence for filling in the matrix of size n*n where n is the length of the input sequence.

$$s[i,j] = \max \begin{cases} 0, & \text{if } i \geq j, \\ s[i+1, j-1] + 1, & \text{if } i < j \text{ and } (v_i, v_j) \in \Gamma, \quad (1) \\ s[i+1, j-1], & \text{if } i < j \text{ and } (v_i, v_j) \notin \Gamma, \quad (1^*) \\ s[i+1, j], & \text{if } i < j, \quad (2) \\ s[i, j-1], & \text{if } i < j, \quad (3) \\ \max_{i<k<j}\{s[i,k] + s[k+1, j]\}, & \text{if } i < j, \quad (4) \end{cases}$$

While the nussinov algorithm is a rather quick way of trying to predict for secondary structure for shorter and simpler sequences, it has its limitations. The most prominent setback is the fact that the algorithm is unable to construct for pseudoknots which are very prevalent for many viral RNAs. The presence of pseudoknots makes standard dynamic programming based algorithms

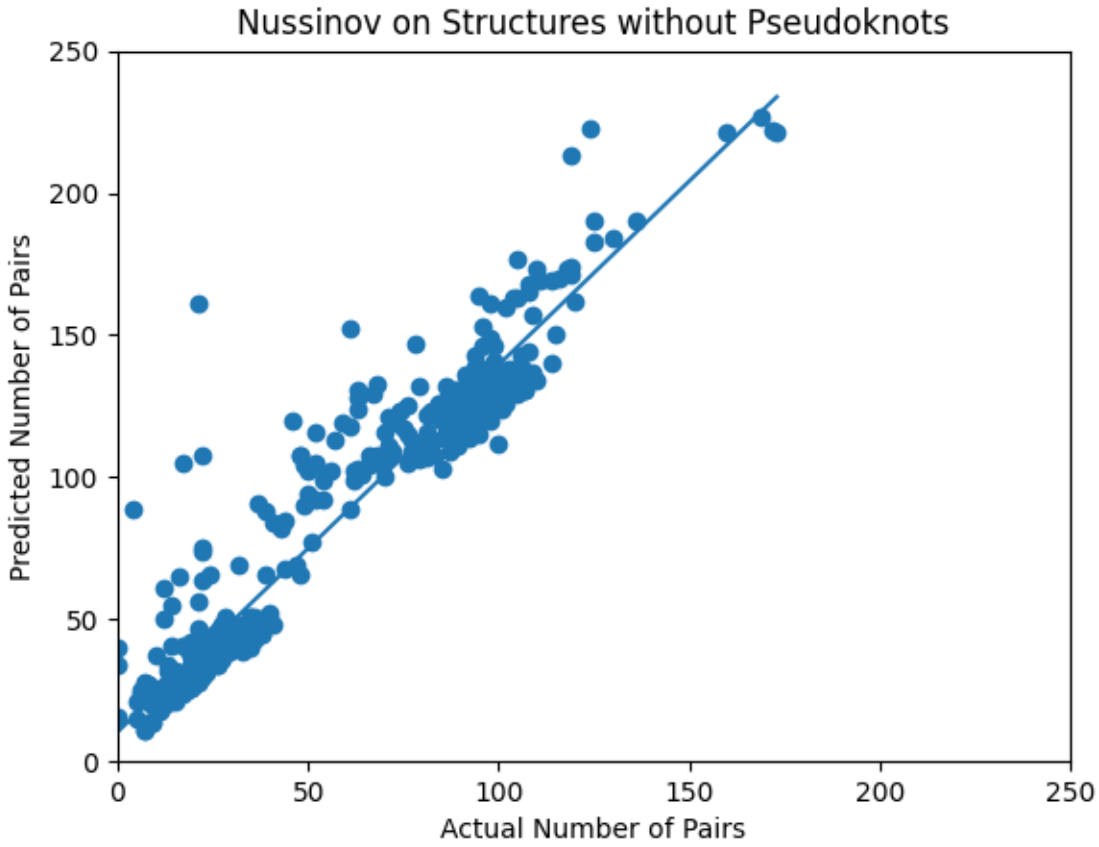
like nussinov much less practical in predicting RNA secondary structures. However, nussinov algorithm provides a good baseline of predicting simple RNA structure with good speed.

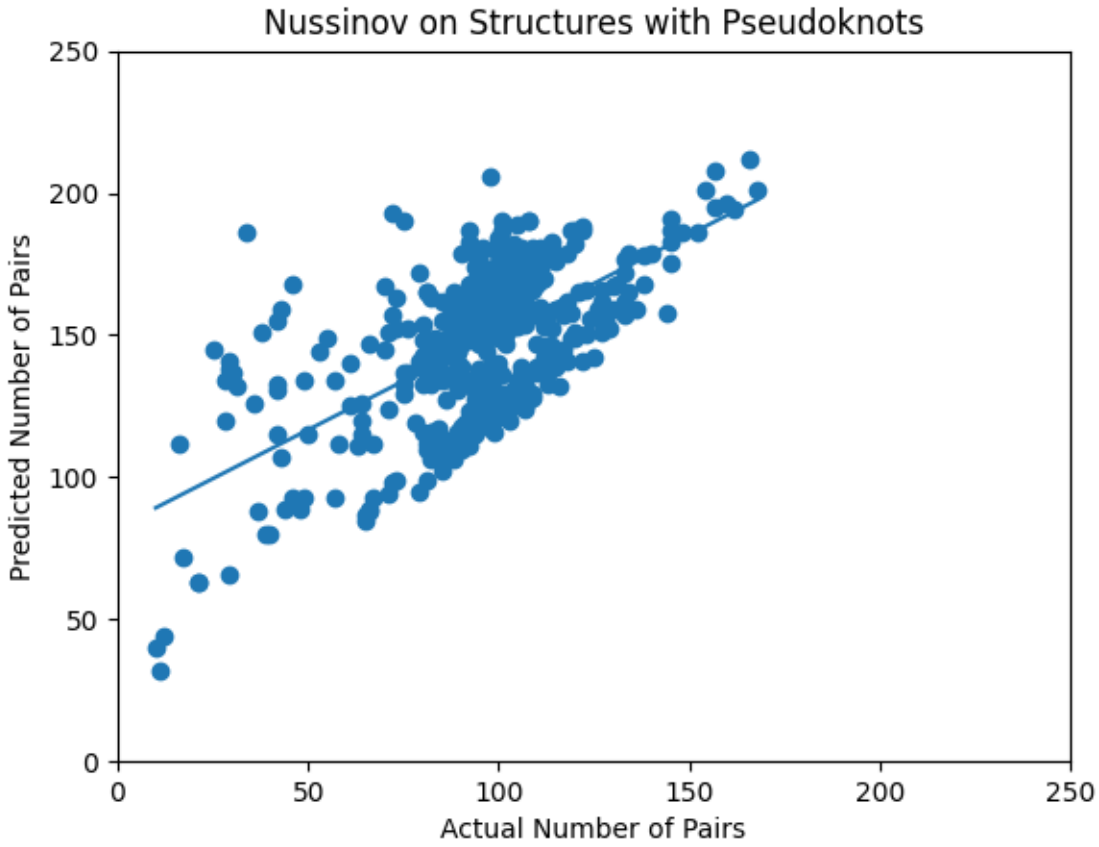
Methods

I implemented Nussinov algorithm and used it to run through documented RNA sequences and their respective structures found at http://bprna.cgrb.oregonstate.edu/about.php. Comparison between the predicted structure and actual structure itself it not of real relevance as I am already aware of the drawbacks of nussinov's algorithm. I considered trying to compare the two and look for local alignments, but due to the large disparity of pairs calculated vs observed I instead opted to simply look at the number of pairs itself.

Note that some of the sequences have structures which contain pseudoknots. Pseudoknots are unaccounted for in the basic version of Nussinov and so to compare likeliness of results, I count for the number of pairs formed versus the number of pairs that actually occur in the real structure.

Validation


Nussinov on Structures without Pseudoknots
Predicted Number of Pairs
Actual Number of Pairs
0
50
100
150
200
250



Results

I also calculated the percentage difference between the actual vs predicted and they are respectively median of 152% for structures without pseudoknots and 158% for structures with pseudoknots present.

As expected, nussinov algorithm overestimates the number of pairs present and in the case of pseudoknots, the number is even greater.

Conclusion + future work

After running nussinov's algorithm on these thousands of sequences of actual RNA strands, it is clear that the algorithm itself is merely good for a baseline in the simplest of structures. There is further work to be done which would include implementing an additional method to include the addition of simple H pseudoknots in the structure. Though the implementation to check for simple pseudoknots would increase the runtime to $O(n^4)$ or possibly higher for even more complex structures, the accuracy would be greatly improved as we are no longer trying to maximize the number of base pairs. In every case, Nussinov with never underpredict and the addition of more restraints would be good for improvement of accuracy at the cost of speed. Although the main advantage of nussinov algorithm is its quick runtime and relative space efficiency $O(n^2)$, it is undeniable from the above testing and evidence that the results are far from reality.

https://github.com/drinkingtea2223/nussinov