An attempt on linguistic complexity analysis of five widely-known brain proteins of *homo sapiens*

Onur DEMİREZEN

Abstract

Described by Troyanskaya et al. (2002), linguistic complexity is the ratio of the number of subwords present in the string of interest to the maximum number of possible subwords for a string of the same length over the same alphabet. By this method, we can see possibly important regions in a biological sequence. I attempted to implement a similar algorithm using a sliding-window approach to look for regions of interest in protein sequences.

Algorithm

The algorithm takes a window size, and slides the window one amino-acid at time, calculating linguistic complexity (LC) of the regions then returns an array of LC values. Two methods are used when counting distinct subwords in a given string: suffix tree and linear string traversal.

The algorithm first tries the calculation with the suffix tree approach. In case of a stack overflow, usually meaning the windows size is too large to handle with this particular suffix tree implementation, linear string traversal is used.

Implementation

Python 3 is used in implementation with *Biopyhon* library for fetching FASTA sequences. *numpy* and *matplotlib* are used for plotting the LC data. To run the algorithm in default mode, type into the terminal:

```
python3 calculate_lc <REFSEQ_number>
-w <window_size>
```

Program fetches the FASTA using the given RefSeq number and runs the algorithm with given window size, then draws the complexity-region plot.

If -w is not used, the program just outputs the linguistic complexity of the whole strand.

For the sake of this report, I have implemented a plotting mechanism that takes 3 window sizes and draws the plot accordingly but to reach it, the user must change the main method of the code.

Properties of the selected proteins

The 5 proteins selected for analysis are some of the most commonly found proteins in the human brain. While their functions are out of the scope of this report, some information is given in Table 1.

Name	*RefSeq*	*RefSeq Version*	*Definition*	*Size (aa)*
Reelin	NP_005036	2	reelin isoform a precursor [Homo sapiens]	3460
Tau	NP_001116538	2	microtubule-associated protein tau isoform 6 [Homo sapiens].	776
Amyloid-β precursor	NP_000475	1	amyloid-beta precursor protein isoform a precursor [Homo sapiens]	770
BDNF	NP_001137277	1	brain-derived neurotrophic factor isoform a preproprotein [Homo sapiens]	247
NGF precursor	NP_002497	2	beta-nerve growth factor precursor [Homo sapiens]	241

Table 1 *Protein properties from NCBI RefSeq database.*

Results

It seems that LC analysis on protein strands differs from DNA sequences mainly because of the size. Protein sequences are far shorter than genome sequences. Titin, the largest protein known in the human body has only around 30 000 amino acids. Even when converted to mRNA, it would have a length of approximately 90 000 nucleotides. While obviously more than some organisms, it is not even close to *H. influenzae* virus genome. Therefore, possible window sizes do not vary widely for protein LC analysis.

I tried many different window sizes for the proteins and included the most "clean" figures of complexity analysis in this report. When looked into the *Reelin (Fig. 2)* for a window size of 15, we can see some relatively major fluctuations but complexity almost always stays above 0.9. For a window size of 25, data points seems to be in harmony with the green points that represent the window size of 15. When window size is increased to 50, differences in complexity are present but as minor as at most 0.03 change.

Most of the things said for Reelin can also be said for *Tau (Fig 2)*. Even though they significantly differ in length, they show similar characteristics when the same window sizes are used. Tau has more significant spikes in some areas, indicating possible regions of interest. Complexity seems low relative to Reelin.

Amyloid-β precursor shows interesting results. The same window sizes as Tau are applied to it. They are also very similar in length. While the general profile is similar to Tau, Amyloid-β precursor shows significant

spikes between 200th and 300th amino-acids, for all three window sizes.
Smaller window sizes, 10, 15 and 25 are used for our two little proteins. While *BDNF (Fig 4)* shows only one spike around the 100th amino-acid, *BNGF* has significantly more spikes and seems like a less-complex protein than BDNF.

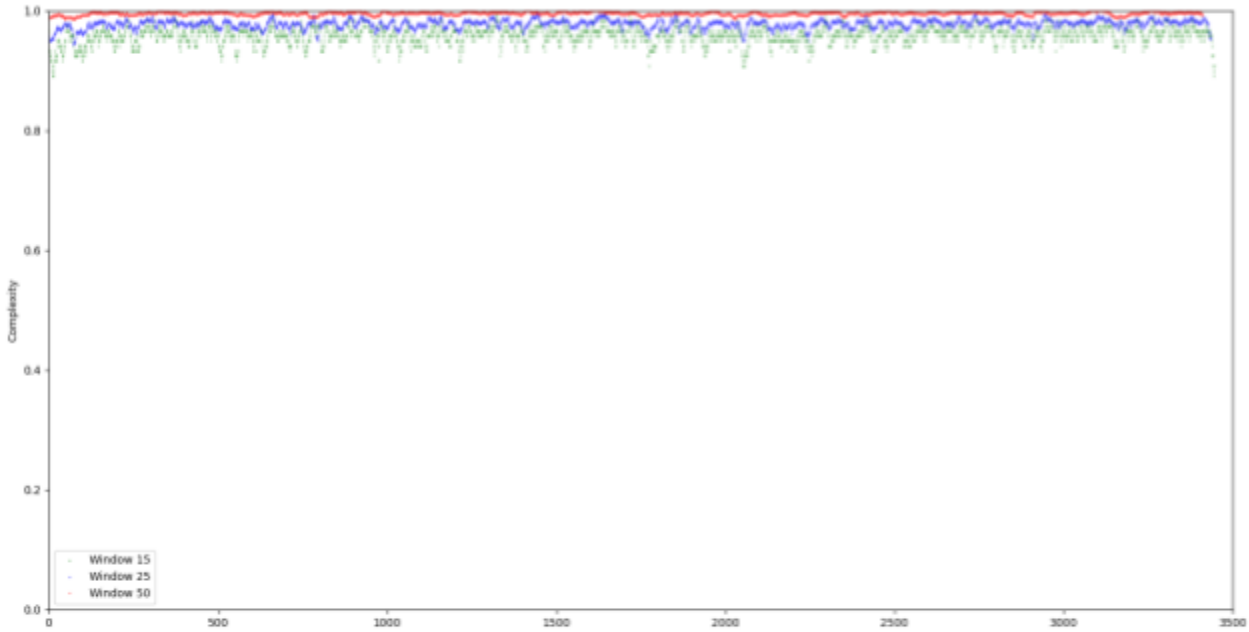

Fig 1. *Reelin* complexity profile for window sizes 15, 25 and 50

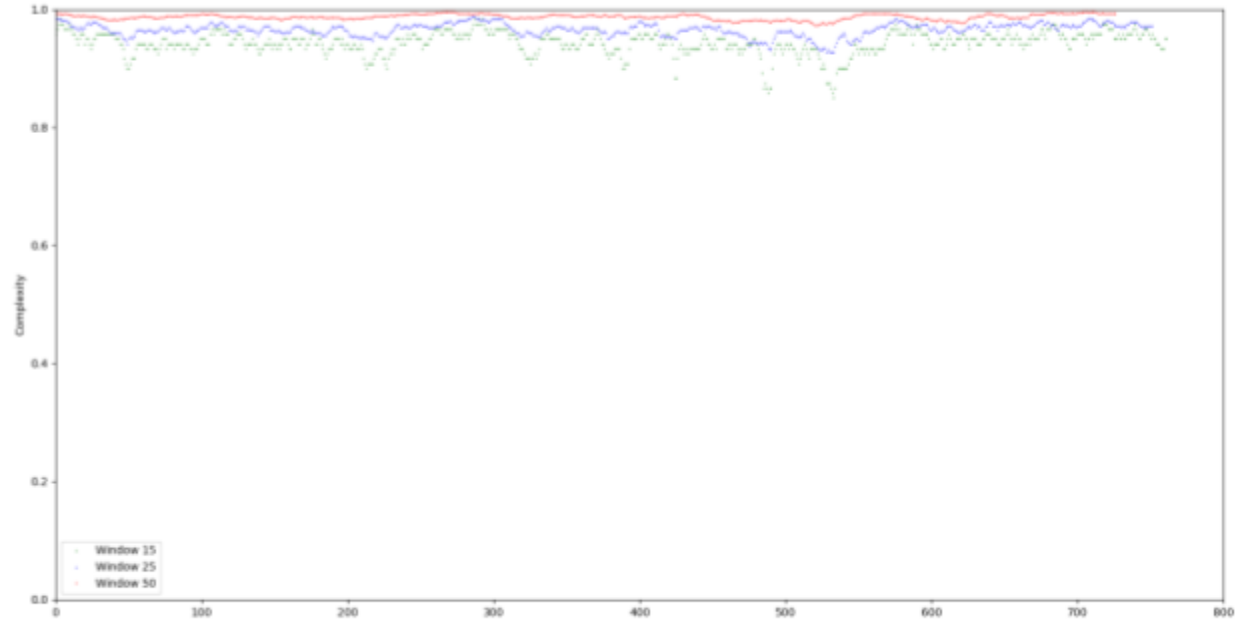

Fig. 2. *Tau* complexity profile for window sizes 15, 25 and 50

Fig. 3. *Amyloid-B* complexity profile for window sizes 15, 25 and 50

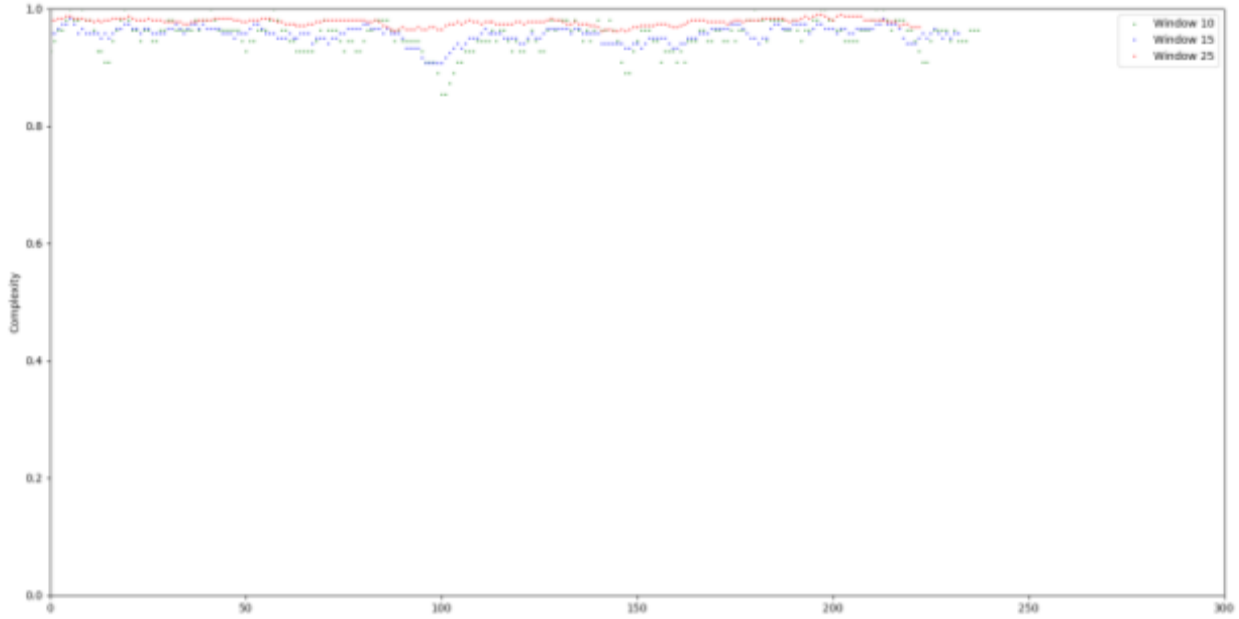

Fig. 4. *BDNF* complexity profile for window sizes 10, 15 and 25

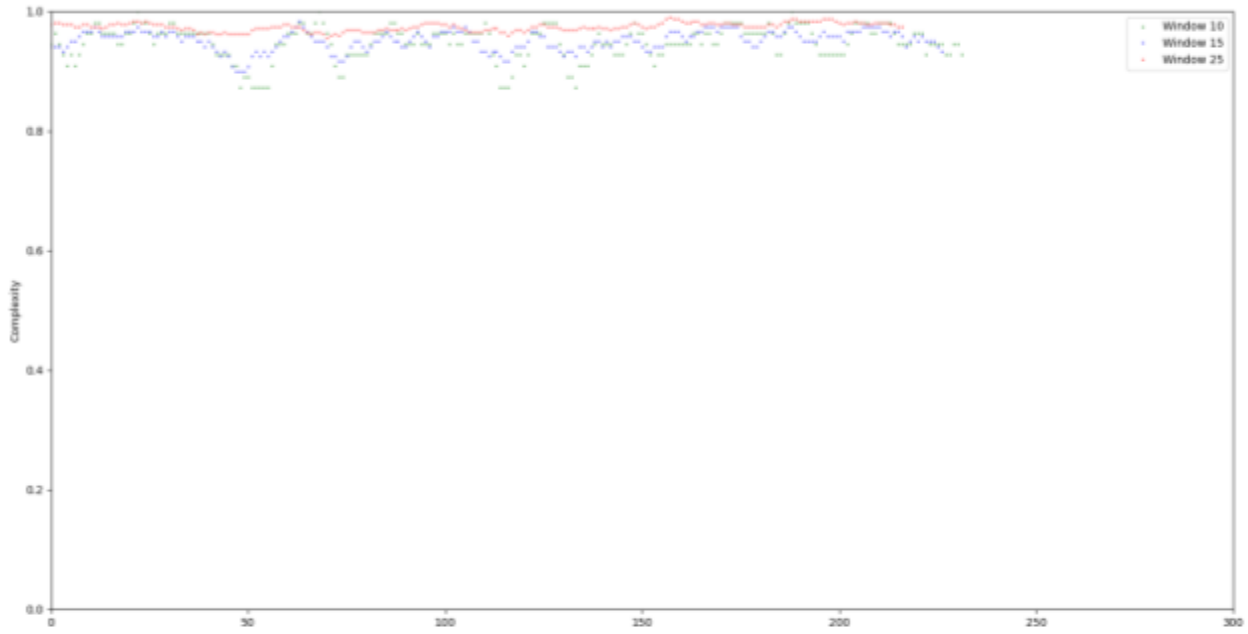

Fig. 5. *BNGF* complexity profile for window sizes 10, 15 and 25

References

- Troyanskaya, Olga G., et al. ”Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity.” Bioinformatics 18.5 (2002): 679-688.
- *https://www.ncbi.nlm.nih.gov/protein/NP_005036.2*
- *https://www.ncbi.nlm.nih.gov/protein/NP_001116538.2*
- *https://www.ncbi.nlm.nih.gov/protein/NP_000475.1*
- *https://www.ncbi.nlm.nih.gov/protein/NP_001137277.1*
- *https://www.ncbi.nlm.nih.gov/protein/NP_002497.2*