Università
degli Studi
di Padova

# Biological datasets for computational physics - Final exam report
# The structural properties of Sororin/Q96FF9 human protein

George Prodan / 2046802

(Dated: June 30, 2022)

Most of the structural properties of the proteins derive from the primary structure. The knowledge of protein structure helps in finding ways of interactions. In this report, I analyzed the physical-chemical properties of Sororin human protein to evaluate the possibility of engagement via different motifs. Also, I present data about potential disease mutations from the structural point of view.

## INTRODUCTION

Proteins are polymers consisting of many amino acids joined together by peptide bonds. There are 20 amino acids with different properties and they can be categorised as hydrophobic, hydrophilic (polar) or charged residues [1]. The primary structure of a protein consists in a sequence of amino acids [2]. From their primary structure, properties such as composition or several functional features (e.g hydrophobicity, transmembrane tendency) can already be predicted. Those properties play an essential role in defining the folding process of a protein [3]. For instance, a high hydrophobicity score may indicate a fingerprint for membrane embedded helices. On the other hand, the loops and turns are mostly polar and are located on the protein surface, as polar amino acids are attracted to the water molecules.

The secondary structure of protein is an intermediary key step to predict the protein folding - the tertiary structure [4]. There are databases such as DSSP [5] which provide the secondary structures found experimentally, and also several tools have been developed for the prediction of such structures.

Also, we can gain some insights about a protein function by identifying sequences of amino acids that are already known to have a specific function. These sequences are called motifs or domains. Usually, a motif is a short sequence between 10 and 20 residues. However, domains are independent units and they are larger (40-700 residues). Both motifs and domains are conserved sequence patterns [6].

In this work, the focus is on analyzing the structural properties of Sororin/Q96FF9 human protein, a member of the Sororin super family. We present the results predicted based on the properties derived from the primary structure of the protein. Then, these results are analyzed from the perspective of the structural variations due to disease mutations.

## METHODS

### Experimental data

The primary structure of Sororin CDCA5 (cell division cycle-associated 5) human protein has been retrieved from the UniProt database [7]. It contains 252 amino acids. Accord-

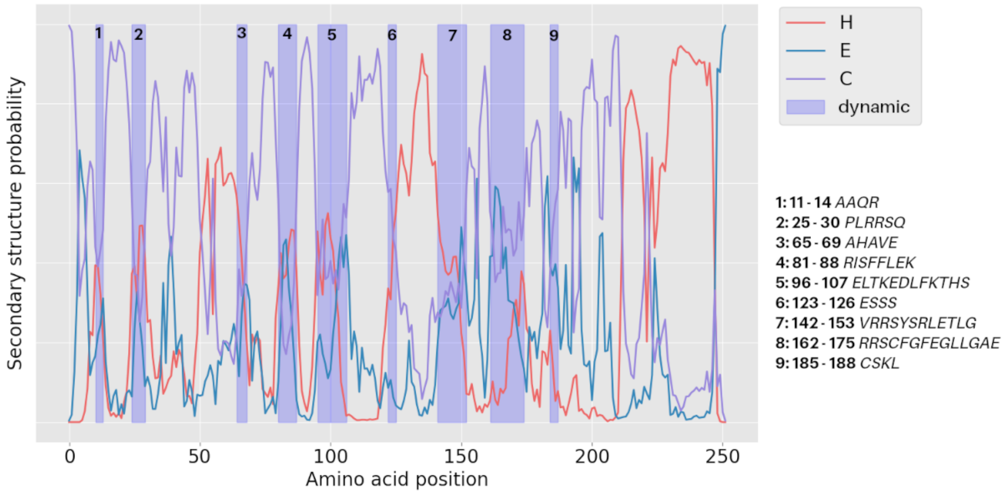


FIG. 1. Predicted probabilities of the secondary structure of Sororin CDCA5 human protein. The regions of comparable probabilities are highlighted in blue and the corresponding sequences are shown on the right side of the plot.

ing to *S. Rankin et al, 2005* [8], this protein can be located inside the nucleus or in the cytoplasm. Data showing the structural variations upon the disease mutations is provided by BioMuta, which is an integrated sequence feature database [9]. It gathers information from sources such as COSMIC, ClinVar, UniProtKB and different publications.

### Analysis methods

To analyze the physical-chemical properties of the protein we consider the hydrophobicity, flexibility. The hydrophobicity profile of the protein is obtained based on the method proposed by Kyte & Doolittle in 1982 [10]. They use a hydropathy scale derived from experimental observations found in the literature. A sliding window of amino acids having a fixed size is considered in order to compute an average hydropathy score for each position of the sequence.

Flexibility along the protein sequence can be characterized by a flexibility score. The method presented in [11] uses a differential equation model to derive this scores taking into account the spatial position and the distribution of the amino acids.

### Statistical analysis of secondary structure

The secondary structure is predicted using GOR4 [12]. An overview of the obtained sequence is presented in Table I.

## RESULTS

### Secondary and tertiary structures

The GOR4 secondary structure prediction of Sororin CDCA5 human protein provides a probabilistic analysis of the resulting sequence. Here, there are three classes: H (alpha helices), E (extended/beta strands) and C (random coils). For each residue, one class has a certain probability to appear in the secondary

| | | | |
|---|---|---|---|
| Alpha helix | Hh | 88 | 34.92% |
| $3_{10}$ helix | Gg | 0 | 0.00% |
| Pi helix | Ii | 0 | 0.00% |
| Beta bridge | Bb | 0 | 0.00% |
| Extended strand | Ee | 17 | 6.75% |
| Beta turn | Tt | 0 | 0.00% |
| Bend region | Ss | 0 | 0.00% |
| Random coil | Cc | 147 | 58.33% |
| Ambiguous states | ? | 0 | 0.00% |
| Other states | | 0 | 0.00% |

TABLE I. A statistical overview of the secondary structure predicted by GOR4

| Positions | Primary | Secondary |
|---|---|---|
| 2 - 10 | GRRTRSGG | cceeeccc |
| 15 - 23 | SGPRAPSP | cccccccc |
| 29 - 37 | SQRKSGSE | cccccccc |
| 75 - 83 | SPRRSPRI | cccccccch |
| 88 - 96 | KENEPPGR | cccccccc |
| 157 - 165 | TSTPGRRS | ccccccce |
| 207 - 215 | GISPPPEK | cccccchh |

TABLE II. Amino acids windows of Sororin/CDCA5 having high flexibility. The corresponding subsequences of primary and secondary structures are shown.

sequence. A plot[1] of those probabilities is shown in Figure 1. One can notice certain regions in which there are comparable probabilities. These regions can be structurally variable. They may initiate aggregations or drive conformational transitions. In the given plot, we spot 9 such regions along the sequence by imposing the conditions that all the probabilities are under 60% and there are at least 4 amino acids involved.

### Hydrophobicity and flexibility

Using Expasy ProtScale [13], properties such as hydrophobicity and flexibility are analyzed. We run the algorithm for a window size of 9 amino acids to get the hydropathy scores. As it can be observed from Figure 2a, Sororin/CDCA5 is more hydrophilic than hydrophobic, thus it is clearly not a transmembrane protein having a very low transmembrane tendency.

In Figure 2b, we can see the average flexibility for a window size of 9 amino acids. Windows that attain a high flexibility score (over 0.50) are highlighted in green. We show these windows in Table II. By spotting these windows we can estimate the best flexible segments of the sequence that could be targeted experimentally. As expected, the segments mostly consist of coils.

### Potential homology models

One can use BLAST [14] in order to retrieve the structure of similar proteins by multiple sequence alignment. I have found homologous structures for 7 human proteins and for another 23 non-human proteins for which BLAST showed a query coverage over 99%. The phylogenetic tree generated by Clustal W [15] for the human proteins is represented in Figure 3. While Sororin/Q96FF9 is a natural protein, the other isoform proteins are generated by insilico approaches (computationally). However, the coverage for this isoform proteins is under 90%. There are similar proteins coming from other species with larger coverage as mentioned before. The full tree created using BLAST is shown in Appendix 1.

The proteins that share the closest ancestor to Sororin human protein are Sororin chimpanzees proteins from species such as Nomascus leucogenys, Hylobates moloch, Gorilla go-

[1] The code used for the plots is available on the following GitHub repository: https://github.com/prodangp/SororinStructuralAnalysis

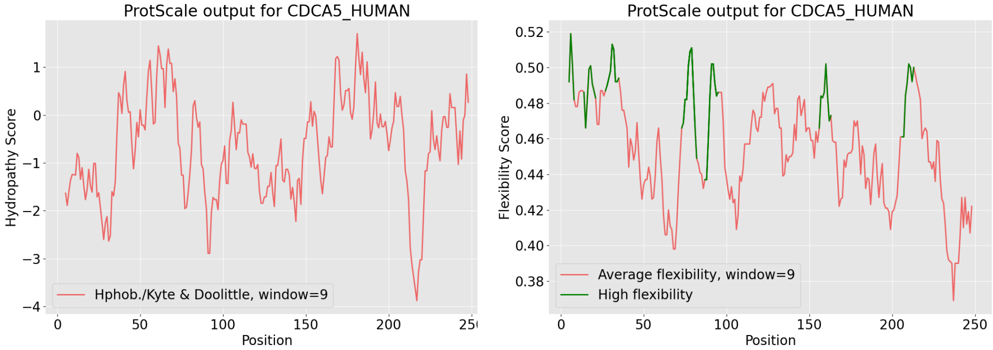


FIG. 2. Plots of hydrophobicity (a) and average flexibility (b) scores as a function of the amino acid position. The data have been retrieved using Expasy ProtScale software.

rilla gorilla, Pan troglodytes and Pan paniscus. The phylogenetic tree in this case can be visualized in Figure 4.

We can notice that the sequence of the human protein is closer to the sequences of Gorilla, Pan troglodytes and Pan paniscus proteins than to the other two proteins and there are strong relationships between Nomascus leucogenys and Hylobates moloch and also between Pan troglodytes and Pan paniscus. The sequence alignment is illustrated in Appendix 2. Regarding the other similar human proteins, we cannot say that one of the isoform proteins we found is closer to the Sororin natural protein than the others.

FIG. 3. The phylogenetic tree of Sororin/CDCA5_HUMAN for similar human proteins.

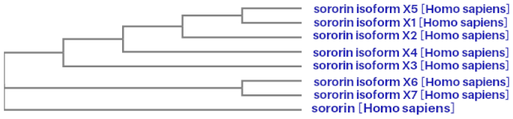
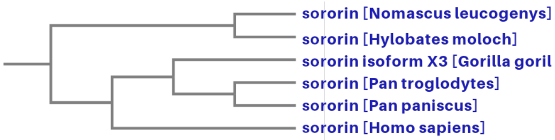

sororin isoform X5 [Homo sapiens]
sororin isoform X1 [Homo sapiens]
sororin isoform X2 [Homo sapiens]
sororin isoform X4 [Homo sapiens]
sororin isoform X3 [Homo sapiens]
sororin isoform X6 [Homo sapiens]
sororin isoform X7 [Homo sapiens]
sororin [Homo sapiens]

FIG. 4. The phylogenetic tree of Sororin/CDCA5_HUMAN for similar Sororin chimpanzee proteins.


sororin [Nomascus leucogenys]
sororin [Hylobates moloch]
sororin isoform X3 [Gorilla goril
sororin [Pan troglodytes]
sororin [Pan paniscus]
sororin [Homo sapiens]

### Structural changes upon disease mutations

Using BioMuta, I have collected information about the possible variations due to mutations for each amino acid. A total of 56 such unique cases are identified after filtering the duplicate variations coming from different sources. If we take into account the localization of the subsequences with high flexibility and, also, of those for which a dynamical behaviour has been indicated in Figure 1, we can spot how many mutations happen in these locations. Thus, 19 mutations are localized in dynamical subsequences (33%), whereas 21 of them happen in high flexibility regions (37.5%). Six of them are spotted in locations in which the two types of regions overlap and 50% of the mutations are outside these regions. A histogram illustrating the variations is shown in Figure 5.

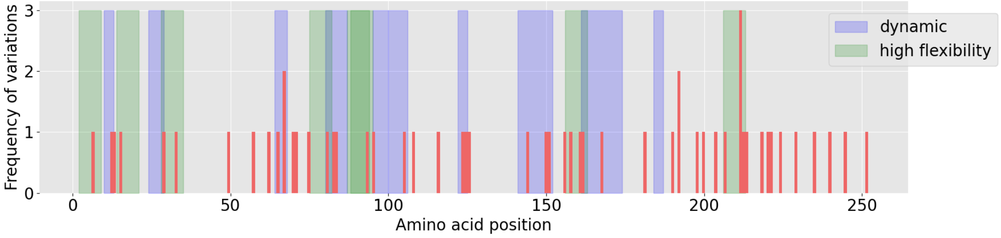


FIG. 5. The frequency of the mutations are shown for each amino acid. High flexibility regions are shadowed in green and the regions covering dynamical subsequences in blue.

## DISCUSSION AND CONCLUSION

It was found experimentally that Sororin interacts with Pds5 protein via a conserved motif FGF (166-168) [16]. Indeed, this supports our results as we have found that this motif is actually a part of the 8th dynamical subsequence shown in Figure 1. One can look for other motifs by using tools such as ELM [17]. For instance, if we refer to the same figure, we can find ELTKEDL phosphothreonine motif (known as a ligand for FHA domains) in the fifth region or the Rev1-Interacting Region RIR motif overlapping with the forth region. The last-mentioned motif is responsible for the interaction between DNA repair proteins and the C-terminal domain, which can be helpful to revert the three mutations located in this region (81, 83, 84).
According to Rankin et al. [8], Sororin is targeted for APC/C-mediated degradation via the KEN box motif, a highly conserved sequence. Referring to the Table II, we can find the KEN motif inside the high flexibility subsequence 88-96. Using ELM again, we identify another possible interacting motif overlapping with the subsequence 75-83: protein phosphatase 1 catalytic subunit (PP1c) - SPRISFF/79-85 . This motif may allow Sororin to be targeted for dephosphorylation.
In conclusion, by analyzing the structural properties derived from the primary structure, one can find efficient procedures to engage the proteins in interactions. Sororin is a very flexible hydrophilic protein and its profile shows a considerable number of disease mutations. Approaches to revert these mutations may be the objective of further studies as we already identified the case of the RIR motif which has a role in DNA repair.

[1] M. Aftabuddin and S. Kundu, Biophysical Journal **93**, 225 (2007).

[2] M. Fuxreiter, Lecture notes. MSc Physics of Data - University of Padua (2021-2022).

[3] V. Daggett and A. R. Fersht, Trends in biochemical sciences **28 1**, 18 (2003).

[4] Ji, Yong-Yun and Li, You-Quan, Eur. Phys. J. E **32**, 103 (2010).

[5] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).

[6] J. Xiong, "Protein motifs and domain prediction," in *Essential Bioinformatics* (Cambridge University Press, 2006) p. 85–94.

[7] T. U. Consortium, Nucleic Acids Research **49**, D480 (2020), https://academic.oup.com/nar/article-pdf/49/D1/D480/35364103/gkaa1100.pdf.

[8] S. Rankin, N. G. Ayad, and M. W. Kirschner, Molecular Cell **18**, 185 (2005).

[9] T.-J. Wu, A. Shamsaddini, Y. Pan, K. Smith, D. J. Crichton, V. Simonyan, and R. Mazumder, Database **2014** (2014), 10.1093/database/bau022, bau022.

[10] J. Kyte and R. F. Doolittle, Journal of Molecular Biology **157**, 105 (1982).

[11] R. BHASKARAN and P. PONNUSWAMY, International Journal of Peptide and Protein Re-

search **32**, 241 (1988).

[12] J. Garnier, J.-F. Gibrat, and B. Robson, in *Computer Methods for Macromolecular Sequence Analysis*, Methods in Enzymology, Vol. 266 (Academic Press, 1996) pp. 540–553.

[13] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. Wilkins, R. D. Appel, and A. M. Bairoch, eng"Protein identification and analysis tools on the expasy server," (Humana Press, Totowa, 2005) pp. 571–607, iD: unige:37793.

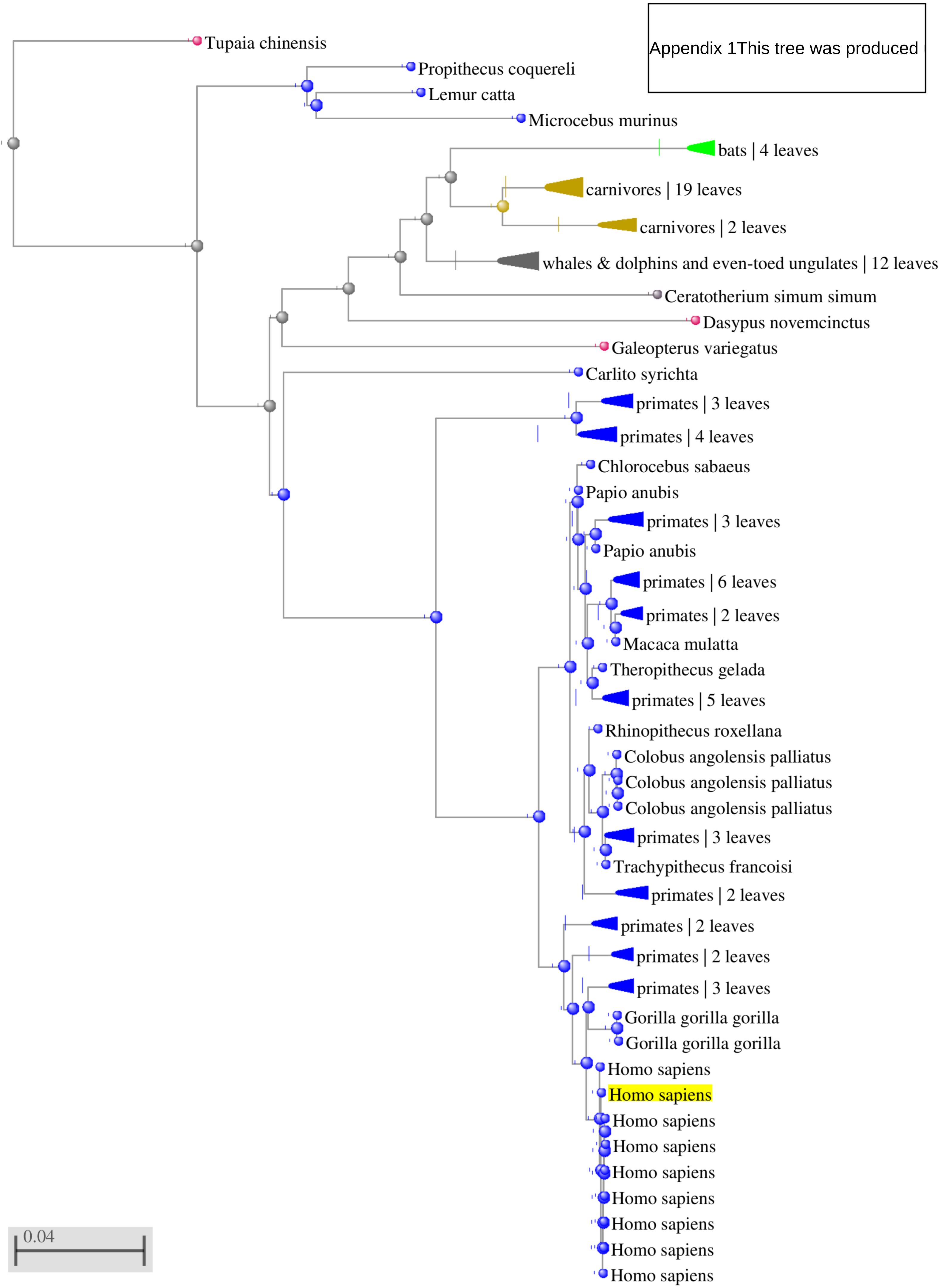
[14] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Nucleic Acids Research **25**, 3389 (1997), https://academic.oup.com/nar/article-pdf/25/17/3389/3639509/25-17-3389.pdf.

[15] J. D. Thompson, D. G. Higgins, and T. J. Gibson, Nucleic Acids Research **22**, 4673 (1994), https://academic.oup.com/nar/article-pdf/22/22/4673/7122285/22-22-4673.pdf.

[16] T. Nishiyama, R. Ladurner, J. Schmitz, E. Kreidl, A. Schleiffer, V. Bhaskara, M. Bando, K. Shirahige, A. A. Hyman, K. Mechtler, and J.-M. Peters, Cell **143**, 737 (2010).

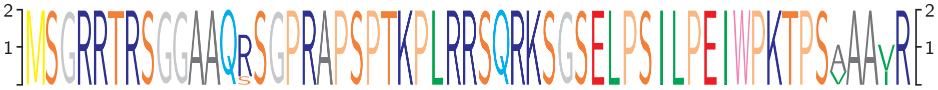
[17] M. Kumar, S. Michael, J. Alvarado-Valverde, B. Mészáros, H. Sámano-Sánchez, A. Zeke, L. Dobson, T. Lazar, M. Örd, A. Nagpal, N. Farahi, M. Käser, R. Kraleti, N. Davey, R. Pancsa, L. Chemes, and T. Gibson, Nucleic Acids Research **50**, D497 (2021), https://academic.oup.com/nar/article-pdf/50/D1/D497/42058167/gkab975.pdf.

Appendix 1This tree was produced

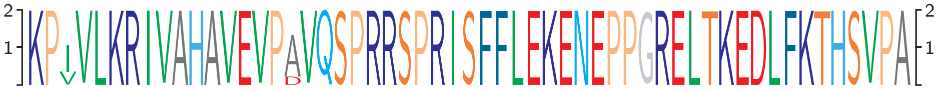


@
@
@

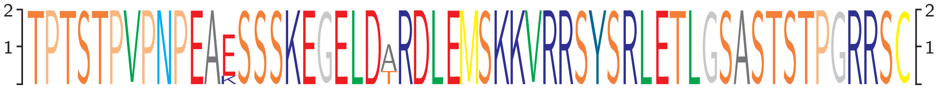
@
@



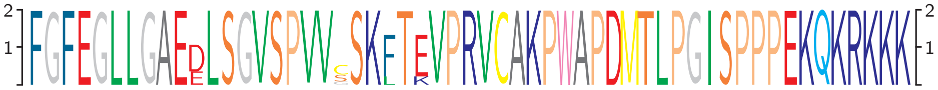
```
Human      MSGRRTRSGGAAQRSGPRAPSPTKPLRRSQRKSGSELPSILPEIWPKTPSAAAVR   55
Pan        MSGRRTRSGGAAQRSGPRAPSPTKPLRRSQRKSGSELPSILPEIWPKTPSVAAVR   55
Gorilla    MSGRRTRSGGAAQRSGPRAPSPTKPLRRSQRKSGSELPSILPEIWPKTPSAAAVR   55
Nomascus   MSGRRTRSGGAAQRSGPRAPSPTKPLRRSQRKSGSELPSILPEIWPKTPSAAAIR   55
Hylobates  MSGRRTRSGGAAQSSGPRAPSPTKPLRRSQRKSGSELPSILPEIWPKTPSAAAIR   55
Pan        MSGRRTRSGGAAQRSGPRAPSPTKPLRRSQRKSGSELPSILPEIWPKTPSVAAVR   55
consensus  !!!!!!!!!!!!!*!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!*!!*!
```



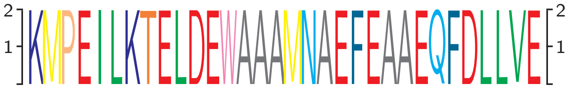
```
Human      KPIVLKRIVAHAVEVPAVQSPRRSPRISFFLEKENEPPGRELTKEDLFKTHSVPA   110
Pan        KPIVLKRIVAHAVEVPAVQSPRRSPRISFFLEKENEPPGRELTKEDLFKTHSVPA   110
Gorilla    KPIVLKRIVAHAVEVPDVQSPRRSPRISFFLEKENEPPGRELTKEDLFKTHSVPA   110
Nomascus   KPVVLKRIVAHAVEVPAVQSPRRSPRISFFLEKENEPPGRELTKEDLFKTHSVPA   110
Hylobates  KPVVLKRIVAHAVEVPAVQSPRRSPRISFFLEKENEPPGRELTKEDLFKTHSVPA   110
Pan        KPIVLKRIVAHAVEVPAVQSPRRSPRISFFLEKENEPPGRELTKEDLFKTHSVPA   110
consensus  !!*!!!!!!!!!!!!!*!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```



```
Human      TPTSTPVPNPEAESSSKEGELDARDLEMSKKVRRSYSRLETLGSASTSTPGRRSC   165
Pan        TPTSTPVPNPEAESSSKEGELDTRDLEMSKKVRRSYSRLETLGSASTSTPGRRSC   165
Gorilla    TPTSTPVPNPEAKSSSKEGELDARDLEMSKKVRRSYSRLETLGSASTSTPGRRSC   165
Nomascus   TPTSTPVPNPEAESSSKEGELDARDLEMSKKVRRSYSRLETLGSASTSTPGRRSC   165
Hylobates  TPTSTPVPNPEAESSSKEGELDARDLEMSKKVRRSYSRLETLGSASTSTPGRRSC   165
Pan        TPTSTPVPNPEAESSSKEGELDTRDLEMSKKVRRSYSRLETLGSASTSTPGRRSC   165
consensus  !!!!!!!!!!!!*!!!!!!!!!*!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```



```
Human      FGFEGLLGAEDLSGVSPVVCSKLTEVPRVCAKPWAPDMTLPGISPPPEKQKRKKK   220
Pan        FGFEGLLGAEDLSGVSPVVCSKFTEVPRVCAKPWAPDMTLPGISPPPEKQKRKKK   220
Gorilla    FGFEGLLGAEDLSGVSPVVCSKFTEVPRVCAKPWAPDMTLPGISPPPEKQKRKKK   220
Nomascus   FGFEGLLGAEELSGVSPVVSSKFTEVPRVCAKPWAPDMTLPGISPPPEKQKRKKK   220
Hylobates  FGFEGLLGAEELSGVSPVVSSKFTEVPRVCAKPWAPDMTLPGISPPPEKQKRKKK   220
Pan        FGFEGLLGAEDLSGVSPVVGSKFTKVPRVCAKPWAPDMTLPGISPPPEKQKRKKK   220
consensus  !!!!!!!!!!*!!!!!!!!*!!*!*!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```



```
Human      KMPEILKTELDEWAAAMNAEFEAAEQFDLLVE   252
Pan        KMPEILKTELDEWAAAMNAEFEAAEQFDLLVE   252
Gorilla    KMPEILKTELDEWAAAMNAEFEAAEQFDLLVE   252
Nomascus   KMPEILKTELDEWAAAMNAEFEAAEQFDLLVE   252
Hylobates  KMPEILKTELDEWAAAMNAEFEAAEQFDLLVE   252
Pan        KMPEILKTELDEWAAAMNAEFEAAEQFDLLVE   252
consensus  !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
```