Geometrical Approach to Protein Secondary Structure Assignment $\kappa\tau S^2$ - User's Guide

D. Tamayo, S.A. Quintana, L. Gil, C.A. Arango

May 16, 2022

Contents

List of Figures

1 Introduction

1.1 $\kappa\tau S^2$ algorithm

The $\kappa\tau S^2$ (curvature-torsion secondary structure) is an assignment algorithm for the secondary structure of proteins based on two main geometrical features: curvature and torsion. This tool uses the crystallographic coordinates of a protein as an input to assign secondary structures. $\kappa\tau S^2$ was designed using Wolfram Mathematica.

This document describes a user's guide of using and understanding $\kappa\tau S^2$.

1.2 Acknowledgments

This work has been funded by the Ministry of Science of Colombia by the grant 874-2020 in the Young Undergraduate Researcher, and the Universidad Icesi.

This work has been partially funded by the OMICAS Program: In-silico Multiscale Optimization of Sustainable Agricultural Crops (Infrastructure and validation in Rice and Sugar Cane), sponsored within the Scientific Colombia Ecosystem, made up by the World Bank, Ministry of Science, Technology, and Innovation (Minciencias), Icetex, Ministry of Education and Ministry of Industry and Tourism, Project ID: FP44842-217-2018.

2 Getting Started

2.1 What is needed

Before using $\kappa\tau S^2$, the following are being needed:

- A Wolfram Mathematica active license.
- The PDB code of a protein of interest (note that it must have its crystallographic coordinates in the PDB file.)
- The $\kappa\tau S^2$ assigning tool folder, which can be downloaded in (WEB PAGE).

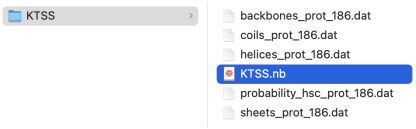

Figure 1: KTSS folder. All the documents provided by this folder are necessary for the well-operation of $\kappa\tau S^2$.

NOTE: Once you have all the requirements shown above, you can start using $\kappa\tau S^2$. In the following sections we are going to use Wolfram Language. Therefore, you must enter the $\kappa\tau S^2$ folder and open the file "KTSS.nb" as it is shown in Figure 1.

2.2 Curvature-torsion pairs

As it was mentioned before, $\kappa\tau S^2$ is based on curvature-torsion pairs. These values are related to a scoring region that has been developing the assigning for secondary structures in proteins. To obtain the curvature-torsion pairs, it must need the crystallographic coordinates of the protein of interest.

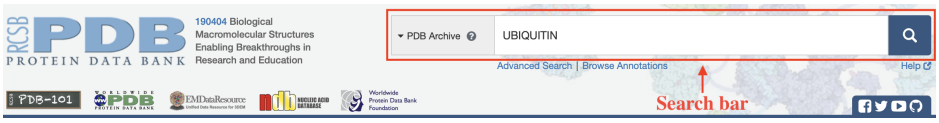

Figure 2: Write in the search bar, on the RCSB PDB web page, a protein of interest. In this example, we are going to use UBIQUITIN.

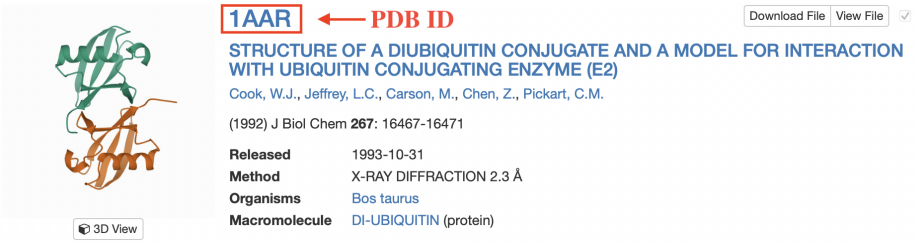

Figure 3: We can see many options for UBIQUITIN. We selected this protein with PDB ID: 1AAR.

Once we identified the PDB ID, open "KTSS.nb". In the first section of the notebook, "*Curvature-torsion pairs*", there is a subsection called "*Reading the necessary information from the PDB repository*".

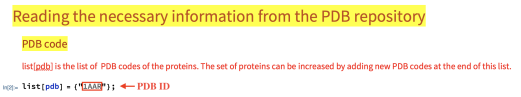

Figure 4: Open the Wolfram Mathematica file. In the first input of the notebook, write the PDB ID of the protein of interest between the quotation marks. In this case, we write "1AAR".

In this subsection, we can see a subsubsection called "**PDB code**" (which is highlighted with yellow). In this, there is one-only input labeled as "list[pdb]". Do type the PDB ID inside this input as seen in Figure 4. This automatically will process the information of the PDB file needed to calculate curvature-torsion pairs (such as protein coordinates, secondary structures, atoms name, and positions.).

2.3 Secondary Structure Assignment

In this section, you are going to visualize all the curvature-torsion pairs within the scoring regions that the algorithm assigns, and the probabilities of belonging to a secondary structure element (QH, QE, and QC).

2.3.1 Protein information

We can find the protein information regarding to the protein's structure and the curvature-torsion pairs as seen in Figure 5.

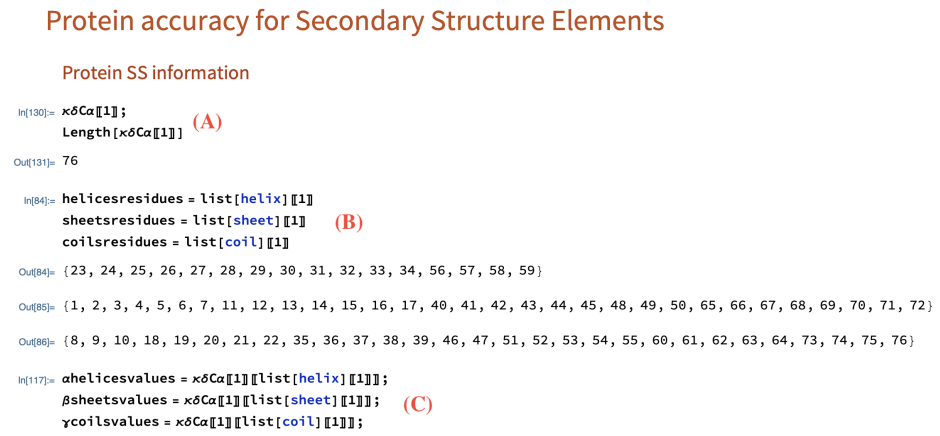

Figure 5: Protein information for 1AAR. A) Shows all the curvature-torsion pairs of the given protein. This must coincide with the total of residues of the protein. B) Shows the list of the positions of a given residue belonging to helices, sheets, or coils. C) Groups the curvature-torsion pairs of the given protein with its secondary structure elements respectively.

2.3.2 Region visualization

The curvature-torsion pairs are distributed in a region plot according to the likelihood of the 186-proteins' curvature-torsion pairs belonging to helices, sheet or coils (REFERENCIAR ARTICULO). Once we obtained the protein information (*section 2.3.1*), it is possible to visualize these values as seen in Figure6. This visualization means that a curvature-torsion value is associated with a secondary structure element depending on its location on the region.

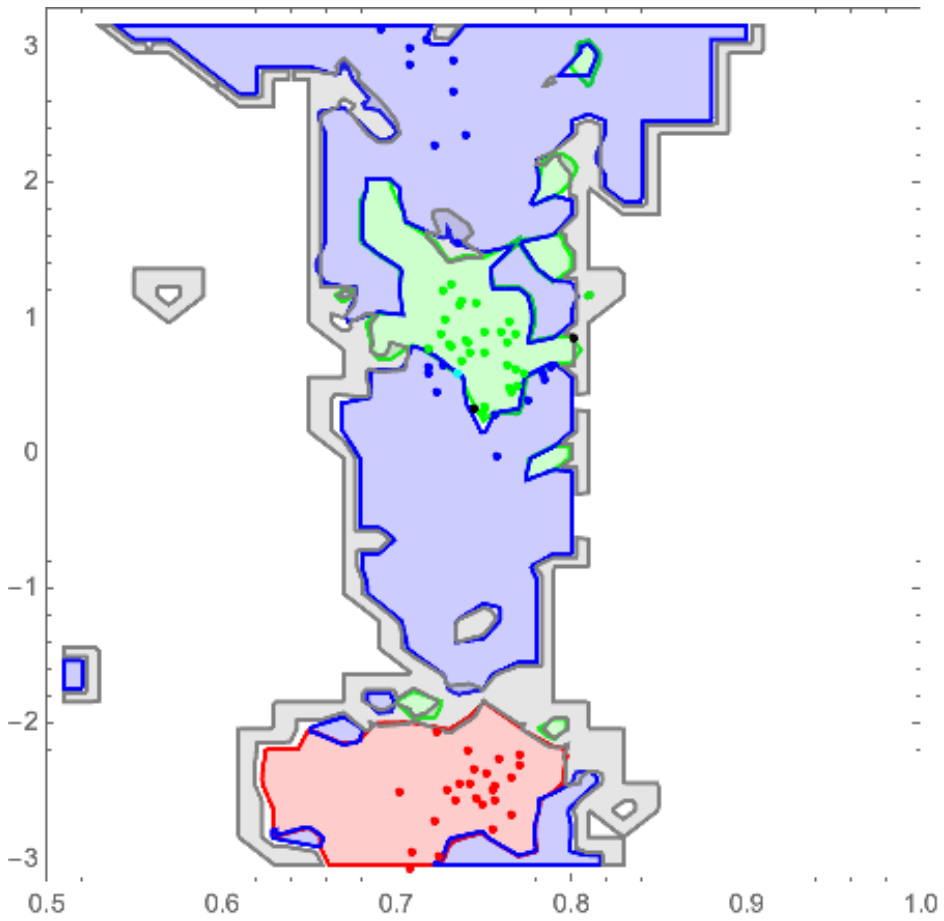

Figure 6: Scoring region for curvature-torsion pairs of 1AAR. Herein, each colour represents a secondary structure element: red for helices, green for sheets, and blue for coils.

2.3.3 Statistics

Figure 7 shows the accuracy values for helices, sheets, and coils (QH, QE, and QC respectively) as well as the standard deviation for each secondary structure element. According to 1AAR, we find that for helices, sheets, and coils, there are an accuracy values of 81.85431%, 71.245%, and 60.7514%, respectively. In terms of standard deviation (σ), we found these values in helices, sheets, and coils of 0.366699, 0.274524, and 0.224403 respectively.

```
valueQH = Mean[1 - allQH]

0.818541

valueQE = Mean[1 - allQE]

0.71245

valueQC = Mean[1 - allQC]

0.607514

σQH = StandardDeviation[probabilitieshelices]

0.366699

σQE = StandardDeviation[probabilitiessheets]

0.274524

σQC = StandardDeviation[probabilitiescoils]

0.224403
```

Figure 7: Statistics for the secondary structure elements in 1AAR.The labels of QH, QE, and QC represent helices, sheets, and coils.

3 Conclusions