SciForum
Mol2Net

# Prot-SSP: A Tool for Amino Acid Pairing Pattern Analysis in Secondary Structures

**Miguel de Sousa [1,*], Cristian R. Munteanu [2] and Alexandre Magalhães [1]**

[1] UCIBIO/REQUIMTE/University of Porto, R. Campo Alegre 687, 4169-007 Porto, Portugal; E-Mail: miguelmsousa@gmail.com (M.S.); almagalh@fc.up.pt (A.M.)
[2] RNASA-IMEDIR group, Computer Science Faculty, University of A Coruna, Campus de Elviña S/N, 15071, A Coruña, Spain (Department of Information and Communication Technologies); E-Mail: crm.publish@gmail.com
* UCIBIO/REQUIMTE/University of Porto, R. Campo Alegre 687, 4169-007 Porto, Portugal; E-Mail: miguelmsousa@gmail.com;
Tel.: +351220402504/+351220402659

*Published: 4 December 2015*

**Abstract:** It is known that individual amino acids can have a decisive role in the stabilization of a protein structure. Moreover, it is likely that specific amino acid combinations also fulfil structural and stabilizing roles in protein structure. We present Prot-SSP, an analytical Python tool designed to gather and parse sequence and structural data from sets of PDB files and determine amino acid residue pairing propensities and correlations in alpha helices and beta strands, in various secondary structure contexts.This versatile and user-friendly bioinformatic tool has proven useful for the analysis of a selected set of protein structures as shown in an illustrative example.

**Keywords:** secondary structure; amino acid pair, alpha helix, beta strand; software

## 1. Introduction

Understanding the formation, stability and function of protein structures requires the characterization of interaction preferences between amino acid residues in secondary structure motifs. It has been established that specific sequences of amino acids may have important roles in protein folding and stability and, more recently, studies show how amino acid patterns, in particular amino acid pairings patterns, may have a stabilizing or destabilizing influence in beta sheets [1], loop sequences [2]

and specific (i, i+4) pairs which stabilize alpha helix structures [3].

Similarly to individual amino acids, whose frequency of occurrence in particular secondary structures varies, it is reasonable to consider that the same principle should also apply to amino acid patterns, with the different propensities of pairs to occur being related to their effect in the formation and stabilization of secondary structures. Likewise, as each residue has an individual part in the stabilization of a secondary structure, the residue distribution is different when considering the different types of positions and contexts (N-terminal, interior and C-terminal) and it is reasonable to consider that this may also apply to amino acid pairing patterns.

To evaluate the preference of a particular amino acid pairing ($X_i$ , $Y_{i+n}$) to occur at the interior of the secondary structure motifs we used a statistic called global propensity ($P^{SS}{}_{X_iY_{i+n}}$), defined as a ratio of the frequency with which that pairing occurs in a given secondary structure and the frequency with which it occurs globally, irrespective of secondary structure:

$$P^{SS}_{X_iY_{i+n}} = \frac{N^{SS}_{X_iY_{i+n}}}{\sum_{A,B} N^{SS}_{A_iB_{i+n}}} \Bigg/ \frac{N^{all}_{X_iY_{i+n}}}{\sum_{A,B} N^{all}_{A_iB_{i+n}}}$$

To determine and make possible the analysis of this statistic, a novel analytical tool conceived to gather and parse sequence and structural data from user-defined sets of PDB files and determine the pairing propensities of amino acid residue pairings, as well as correlation values, in alpha helices and beta sheets, in various possible contexts of secondary structure motifs. This tool, Prot-SSP, is a GUI Python/wxPython application which, for desktop, can be compiled for the Windows XP/Vista/7 operatic systems. Our working version was compiled for Windows 7.

Prot-SSP uses user-defined inputs to cull protein chain sequence data files from the worldwide Protein Data Bank database [4], extracts protein sequence information into local storage and confirms secondary structure assignment using the DSSP algorithm [5]. The resulting output comes in the form of TXT files in CSV-format which can be opened and edited with notepad, worksheet software or, for ease of viewing, using a simple specialized macro-enabled Excel file (.XLSM) supplied with Prot-SSP.

A wide range of parameters, ranging from sample specifications to motif area under study and residue spacing as well as pattern specificity, are contemplated and can be easily and comprehensively set by the user through the program's GUI.

For alpha helices, the minimum size of protein chains to be analysed is six residues and Prot-SSP can analyse pairings up to five residues apart in three different contexts: N-terminal (considering the first three helical residues as not equivalent), interior and C-terminal (considering the last three helical residues as not equivalent).

In beta sheets, pairings are considered up to two residues apart and the minimum size for a chain to be included in analysis is six residues. Again there is a distinction between N-terminal, interior and C-terminal context and the first and last few residues are considered particular cases of study. Additionally, strands are divided into four categories depending on relative strand orientation: terminal, parallel, anti-parallel and mixed.

## 2. Results and Discussion

The original form of Prot-SSP, pre-development into an application with GUI, was used in an analysis of pairing propensities of amino acid residues in the interior of alpha helical structures [6] which reproduced and complemented previous research [7].

In its current form, Prot-SSP has been extensively tested for application in calculations of amino acid pairing propensities in alpha helices and in beta sheets, showing reliability as tool with data gathering and parsing capabilities for calculation of statistical analysis parameters.

**Table 1** presents a typical example of the processed output of a calculation made using Prot-SSP for the particular case of (i, i+2) amino acid residue pairings in beta sheet secondary structures (results for other cases can be supplied upon request).

**Table 1.** Global propensities for amino acid residue pairings in beta sheet protein secondary structure. Parameters: 25% identity; 2.0 A resolution; 0.25 R-value; beta 2S motif; i, i+2 residue spacing; Interior region; all pairs considered; no distinction between types of beta strand (All)

| | P | G | F | V | Y | W | C | A | L | I | M | T | N | S | H | D | Q | R | K | E | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P | 0.1 | 0.3 | 0.3 | 0.5 | 0.6 | 0.3 | 0.2 | 0.3 | 0.3 | 0.5 | 0.7 | 0.5 | 0.2 | 0.2 | 0.3 | 0.2 | 0.3 | 0.4 | 0.3 | 0.2 | 0.0 |
| G | 0.2 | 0.8 | 1.3 | 1.7 | 1.1 | 0.8 | 1.9 | 1.0 | 1.2 | 1.4 | 1.2 | 0.8 | 0.5 | 0.6 | 0.7 | 0.4 | 0.6 | 0.6 | 0.5 | 0.6 | 0.0 |
| F | 0.6 | 1.3 | 2.3 | 2.8 | 1.7 | 1.9 | 2.7 | 1.2 | 1.9 | 2.6 | 2.3 | 1.4 | 1.0 | 1.1 | 1.1 | 0.5 | 0.9 | 0.9 | 0.7 | 0.7 | 0.0 |
| V | 0.8 | 1.5 | 2.9 | 3.3 | 2.4 | 1.9 | 2.3 | 1.7 | 2.1 | 2.7 | 2.5 | 1.9 | 1.0 | 1.5 | 1.5 | 0.9 | 1.3 | 1.3 | 1.0 | 1.3 | 0.0 |
| Y | 0.4 | 1.1 | 2.3 | 2.7 | 2.0 | 1.4 | 4.4 | 1.3 | 1.9 | 2.6 | 2.0 | 1.6 | 0.7 | 1.1 | 1.1 | 0.7 | 1.1 | 1.2 | 1.0 | 1.1 | 0.0 |
| W | 0.2 | 1.0 | 2.0 | 1.8 | 1.9 | 1.2 | 2.2 | 1.1 | 1.5 | 1.7 | 2.0 | 1.2 | 0.9 | 1.0 | 1.0 | 0.6 | 1.0 | 1.1 | 0.4 | 0.8 | 0.0 |
| C | 0.7 | 1.3 | 2.1 | 4.0 | 2.2 | 1.6 | 1.8 | 1.5 | 1.9 | 2.0 | 1.7 | 1.3 | 0.7 | 1.2 | 1.2 | 0.6 | 1.0 | 1.1 | 1.0 | 1.2 | 0.0 |
| A | 0.4 | 0.8 | 1.4 | 1.7 | 1.0 | 0.9 | 1.3 | 0.7 | 1.0 | 1.4 | 1.4 | 0.8 | 0.5 | 0.8 | 0.8 | 0.4 | 0.5 | 0.5 | 0.6 | 0.5 | 0.0 |
| L | 0.4 | 1.1 | 2.0 | 2.6 | 1.8 | 1.4 | 2.9 | 0.9 | 1.6 | 2.2 | 1.6 | 1.3 | 0.6 | 1.0 | 0.8 | 0.4 | 0.7 | 0.8 | 0.7 | 0.6 | 0.0 |
| I | 0.6 | 1.3 | 2.5 | 2.7 | 2.2 | 1.5 | 2.8 | 1.5 | 1.9 | 2.8 | 3.2 | 1.5 | 0.8 | 1.4 | 1.3 | 0.8 | 1.1 | 1.0 | 0.8 | 1.1 | 0.0 |
| M | 0.6 | 1.2 | 2.4 | 2.6 | 2.1 | 1.8 | 2.8 | 1.0 | 2.0 | 2.5 | 1.4 | 1.3 | 0.7 | 1.1 | 1.1 | 0.5 | 0.9 | 0.7 | 0.5 | 0.8 | 0.0 |
| T | 0.3 | 0.8 | 1.4 | 1.8 | 1.6 | 1.4 | 1.4 | 1.0 | 1.2 | 1.5 | 1.5 | 2.0 | 0.8 | 1.3 | 1.5 | 0.7 | 1.2 | 1.3 | 1.1 | 1.0 | 0.0 |
| N | 0.2 | 0.5 | 0.8 | 0.8 | 0.9 | 1.1 | 1.1 | 0.6 | 0.5 | 0.8 | 0.5 | 1.0 | 0.5 | 0.8 | 0.8 | 0.4 | 0.7 | 0.9 | 0.5 | 0.5 | 0.0 |
| S | 0.3 | 0.7 | 1.3 | 1.5 | 1.3 | 1.0 | 1.4 | 0.7 | 0.8 | 1.3 | 1.2 | 1.4 | 0.7 | 1.1 | 1.0 | 0.6 | 0.9 | 1.0 | 0.8 | 0.9 | 0.0 |
| H | 0.4 | 0.6 | 1.4 | 1.6 | 1.2 | 1.5 | 0.5 | 0.9 | 1.0 | 1.4 | 1.6 | 1.3 | 0.7 | 0.8 | 0.7 | 0.6 | 1.0 | 0.9 | 0.9 | 1.0 | 0.0 |
| D | 0.2 | 0.5 | 0.6 | 0.9 | 0.9 | 0.6 | 0.8 | 0.4 | 0.5 | 0.8 | 0.5 | 0.8 | 0.5 | 0.5 | 0.5 | 0.3 | 0.7 | 0.7 | 0.6 | 0.4 | 0.0 |
| Q | 0.4 | 0.4 | 0.9 | 1.2 | 0.9 | 1.2 | 0.5 | 0.6 | 0.6 | 1.1 | 1.1 | 1.5 | 0.7 | 0.8 | 0.8 | 0.5 | 0.6 | 0.9 | 0.7 | 0.7 | 0.0 |
| R | 0.3 | 0.6 | 1.2 | 1.4 | 1.2 | 0.8 | 1.3 | 0.8 | 0.8 | 1.1 | 1.2 | 1.3 | 0.5 | 1.0 | 1.2 | 0.7 | 1.0 | 1.2 | 0.8 | 1.0 | 0.0 |
| K | 0.2 | 0.6 | 0.8 | 1.0 | 0.9 | 1.0 | 0.7 | 0.5 | 0.5 | 0.9 | 0.9 | 1.3 | 0.5 | 0.9 | 1.0 | 0.6 | 0.8 | 1.0 | 0.7 | 0.8 | 0.0 |
| E | 0.2 | 0.5 | 0.7 | 1.0 | 0.7 | 1.0 | 0.7 | 0.5 | 0.6 | 0.7 | 0.8 | 1.3 | 0.6 | 1.0 | 1.0 | 0.6 | 0.7 | 0.9 | 0.8 | 0.9 | 0.0 |
| X | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

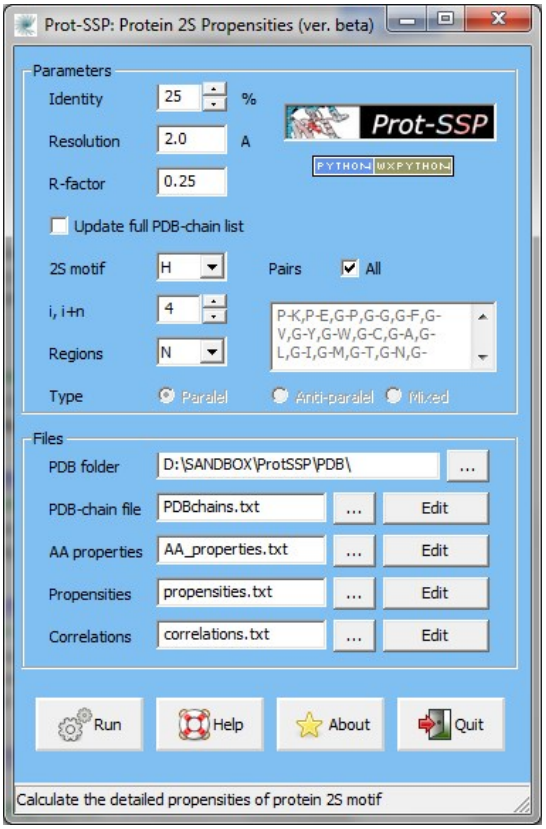


**Figure 1.** Prot-SSP graphic user interface

## 3. Materials and Methods

Through the GUI main window (**Figure 1**), the user can define and adjust parameters for propensity calculations as well as the input/output files. Details of undergoing operations, progress detail, errors and some calculation details are displayed in the console window.

The thresholds defined when culling PDB using the PISCES [8] server are detailed in "Identity", "Resolution" and "R-factor" and the user has the option of checking PDB for more recent versions of those defined in the PDB-chain file input. The user may also specify the secondary structure motif (alpha helix or beta sheet) under study as well as the amino acid pairing separation and the region under study (N-, C-terminal or interior). Specifically for beta sheet structures, the user can define the relative orientation of the strands to be considered for analysis. Specific pairing possibilities can also be defined by the user.

A TXT file containing the chain listing, which can be created manually or culled from the PISCES server, is used by Prot-SSP to download all the relevant structural files from the PDB into local storage and update any present files if judged necessary. During operation, the program cross-checks the locally stored PDB files with DSSP for the attribution of secondary structure to the structural data and culling of relevant sequences and creates an additional text file with lists and details the sequences culled for analysis.

The names and locations of output TXT files may also be defined through the GUI. The visualization and editing of the resulting files, while possible using text editing software can be

easily and comfortably done using a simple macro-enabled Excel XLSM file, supplied with Prot-SSP, to colour-code the entries from the resulting propensity-value tables.

## 4. Conclusions

Prot-SSP was projected and created to address the need for a tool that could simultaneously address the need for data gathering, parsing, calculation and a degree of analysis of amino acid residue patterns in protein secondary structures. A relatively simple tool to use, its application to carefully selected data sets and the results yielded may prove it to have significant potential – both for immediate analysis and for future applications in the field of protein structure prediction.

## Acknowledgments

The authors acknowledge the support provided by the Galician Network of Drugs R+D REGID (Xunta de Galicia R2014/025) and by the "Collaborative Project on Medical Informatics (CIMED)" PI13/00280 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013-2016 and the European Regional Development Fund / GAIN (FEDER - CONECTAPEME - INTERCONECTA). This work was partially supported by the Galician Network for Colorectal Cancer Research (Red Gallega de Cáncer Colorrectal - REGICC, Ref.: CN2014/039), Institute for Biomedical Informatics of A Coruña (INIBIC), and Center for Research of Information and Communication Technologies (CITIC).

## Conflicts of Interest

The authors declare no conflict of interest.

## References

1. Fooks, H.M.; Martin, A.C.R.; Woolfson, D.N.; Sessions, R.B.; Hutchinson, E.G. Amino acid pairing preferences in parallel beta-sheets in proteins. *Journal of Molecular Biology* **2006**, *356*, 32-44.

2. Crasto, C.J.; Feng, J.A. Sequence codes for extended conformation: A neighbor-dependent sequence analysis of loops in proteins. *Proteins-Structure Function and Bioinformatics* **2001**, *42*, 399-413.

3. Andrew, C.D.; Penel, S.; Jones, G.R.; Doig, A.J. Stabilizing nonpolar/polar side-chain interactions in the alpha-helix. *Proteins-Structure Function and Genetics* **2001**, *45*, 449-455.

4. Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.B.; Meyer, E.F.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. Protein data bank - computer-based archival file for macromolecular structures. *Journal of Molecular Biology* **1977**, *112*, 535-542.

5. Kabsch, W.; Sander, C. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577-2637.

6. de Sousa, M.M.; Munteanu, C.R.; Pazos, A.; Fonseca, N.A.; Camacho, R.; Magalhaes, A.L. Amino acid pair- and triplet-wise groupings in the interior of alpha-helical segments in proteins. *J. Theor. Biol.* **2011**, *271*, 136-144.

7. Fonseca, N.A.; Camacho, R.; Magalhaes, A.L. Amino acid pairing at the n- and c-termini of helical segments in proteins. *Proteins-Structure Function and Bioinformatics* **2008**, *70*, 188-196.

8. Wang, G.L.; Dunbrack, R.L. Pisces: Recent improvements to a pdb sequence culling server. *Nucleic Acids Research* **2005**, *33*, W94-W98.