Ultimate Secondary Structure Assignment Method

Projet Long

Manon Curaudeau

16th January 2020

https://github.com/mnncrd/UltimateSSAM

Contents

Introduction

Secondary structure assignment is a necessary step in the analysis of 3D protein structures. The most popular software for secondary structure assignment is DSSP (*Define Secondary Structure of Proteins*), a pattern-recognition algorithm of hydrogen-bonded and geometrical features extracted from atomic coordinates, described in Kabsch and Sander (1983). DSSP allows the assignment of helices (α, π, and 3_{10}), β-strands, β-bridges, turns, and bends.

Several other methods have been developed to assign secondary structures. Some focus on other structures than the classic Helices-Strands-Coils defined by DSSP: PolyProline type II helices (Mansiaux *et al.*, 2011), β-turns (Chan *et al.*, 1993; De Brevern, 2016), β-hairpins, and β-buldges (Craveur *et al.*, 2013), such as ProMotif (Hutchinson and Thornton, 1994), or helices axes such as Helanal (Kumar and Bansal, 1998; Bansal *et al.*, 2000).

The goal of this project is to create a tool, UltimateSSAM, combining and centralising the preexisting tools to implement a more generic approach.This will allow to go further in the analyses.

Material and Methods

Programming

UltimateSSAM was divided into several modes, each of them performing a secondary structure assignment method. All scripts were written in Python 3.7. and are available at https://github.com/mnncrd/UltimateSSAM.

Necessary inputs for UltimateSSAM are: the mode to run, an input file, and an output file.

```
$ python3 ssam.py mode -i input -o output
```

The common first step of all the modes was to read input files. Both PDB (`.pdb`) and PDBx/mmCIF (`.cif`) are supported, to account for the fact that the PDB format is no longer extented. Indeed, the PDB format will gradually be replaced by the PDBx/mmCIF format, which became the standard PDB archive format in 2014.

The atomic coordinates and information were stored in a Atom object instance for each atom present in the input file. These Atom object instances were then stored in a Residue object instance. A list containing Residue object instances was created for each chain in the protein, which were stored a list of chains. Following this, this list was passed as input to the different functions of the selected mode.

DSSP mode

As a first approach, a DSSP like method, subsequently referred to as the `dssp` mode, was implemented. It can be accessed using the following command:

```
$ python3 ssam.py dssp -i input -o output
```

The DSSP method focuses solely on hydrogen bonds, which are calculated using the following equation:

$$E = q_1 q_2 (\frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}) * f \quad (1)$$

The `dssp` mode checks for the existence of Elementary Hydrogen Bond Patterns – n-turns and β-bridges – by computing the energy with (1). When this energy is lower than 0.5, a hydrogen bond exists. Depending on the pattern of hydrogen bonds along the molecule backbone, n-turns or β-bridges are created. From the pattern of n-turns and β-bridges, it checks for the existence of Cooperative Hydrogen Bond Patterns: helices and β-ladders (for more information, see Kabsch and Sander, 1983).

Following DSSP, eight different conformational states exist in the `dssp` mode: 3_{10}-helices (G), α-helices (H), and π-helices (I), β-strands (E), β-bridges (B), bends (S), turns (T), and coils (C). Note that coils are represented by an empty string in UltimateSSAM.

DSSP does not use the hydrogen atoms already present in a PDB/mmCIF file, but instead places new hydrogen atoms. On the contrary, the default `dssp` mode uses the hydrogen atoms

present in the input file. This can be overidden by specifying the argument `-hy`, which will result in the placement of hydrogen atoms *à la* DSSP.

```
$ python3 ssam.py dssp -i input -o output -hy
```

DSSPCOMPARE mode

Another mode, the `dsspcompare` mode was implemented in UltimateSSAM. It allows the comparison of the output of the `dssp` mode with the output if DSSP for a given protein. It computes the percentage of matches between the `dssp` mode and DSSP for the eight conformational states, as well as the eight conformational states reduced to three classes: Helices (H), representing 3_{10}-helices (G), α-helices (H), and π-helices (I); Strands (E), representing β-strands (E) and β-bridges (B); Loops (C), representing bends (S), turns (T), and coils (C).

It returns a `.csv` file as output, and thus requires the argument `-oc` followed by a `.csv` file. It can be accessed using the following command:

```
$ python3 ssam.py dsspcompare -i input -o output -oc output-compare
```

SSAM and SSAMCOMPARE modes

Two other modes, the `ssam` and `ssamcompare` modes were implemented in UltimateSSAM, but are not fully developed. As for now, they only contain `dssp` and `dsspcompare` respectively. They could be expanded to include other secondary structure assignment methods, such as a Promotif- and a Helanal-like modes.

The `ssam` mode can be accessed using the following command:

```
$ python3 ssam.py ssam -i input -o output
```

The `ssamcompare` mode can be accessed using the following command:

```
$ python3 ssam.py ssamcompare -i input -o output -oc output-compare
```

Data

Three non-redundant datasets of one hundred proteins each (Table 1) were generated to test the performances of UltimateSSAM, and compare it to those of DSSP, via the `dsspcompare` mode. The first dataset, One chain, is composed of monomers generated with solution Nuclear magnetic resonance (NMR), and thus the proteins contain hydrogen atoms. Similarly, the second dataset, Multiple chains, is composed of polymers generated with solution NMR. Finally, the Random proteins dataset is composed of randomly chosen proteins.

The first two datasets were analysed both with and without the argument `-hy`, to decipher if the differences between UltimateSSAM and DSSP are due to the position of the hydrogen atoms only.

Table 1: Proteins used for each dataset. The One chain and Multiple chains datasets are composed of monomers and polymers, respectively, generated with solution NMR. The Random proteins dataset is composed of randomly chosen proteins. One hundred proteins were used for each dataset.

Dataset	Proteins
One chain	1BTA, 1FKS, 1GO1, 1J2M, 1J56, 1JNT, 1JU8, 1K37, 1K76, 1KM7, 1M3G, 1OPZ, 1OR5, 1OW5, 1P0R, 1PRR, 1PX9, 1Q3T, 1Q59, 1R9K, 1RK9, 1RZW, 1S05, 1S2H, 1SP0, 1SVJ, 1SZV, 1TTG, 1V3A, 1V49, 1VMC, 1W7D, 1WCJ, 1WH1, 1WUG, 1YG0, 1YLB, 1YSG, 1YSN, 1YX7, 1Z1V, 1Z1Z, 1Z3J, 1Z7R, 1ZS5, 2A29, 2ABO, 2AFE, 2AJ1, 2AQF, 2ARW, 2B0G, 2B8F, 2BVB, 2D46, 2E6W, 2ERS, 2EYV, 2EYX, 2FVN, 2FZ5, 2GQL, 2GT6, 2GVP, 2HQ3, 2HRF, 2HSX, 2HZ8, 2IDY, 2J52, 2J76, 2JNQ, 2JQX, 2JT5, 2JU2, 2JVH, 2JYL, 2K0J, 2K51, 2K61, 2K6N, 2KHO, 2L8M, 2LQD, 2M3K, 2MA1, 2MC5, 2MOR, 2MPO, 2MPV, 2N1L, 2ND4, 2O1Y, 2O21, 2PLP, 2RNA, 2WNM, 3TRX, 5UHU, 6OVC
Multiple chains	1BI6, 1BVG, 1BXL, 1CQH, 1DHM, 1DOM, 1E08, 1F2R, 1F95, 1FXT, 1G3F, 1G5J, 1GX7, 1HAA, 1HUM, 1HZE, 1IHV, 1IL8, 1ILQ, 1IO6, 1J4V, 1KA7, 1KLC, 1KLQ, 1L4W, 1L5E, 1MH6, 1MSG, 1N3J, 1N9J, 1NIQ, 1OLG, 1OO9, 1OVX, 1PES, 1PRM, 1Q6A, 1QNZ, 1QTG, 1QWE, 1RGJ, 1RLQ, 1RQU, 1RQV, 1RTO, 1SAK, 1SAL, 1SUY, 1WCR, 1WJA, 1WLP, 1WTU, 1XOX, 1XR0, 1YFB, 1YUR, 1YUT, 1ZZF, 2A7U, 2ADL, 2ADN, 2BBM, 2BTX, 2E8J, 2EZO, 2EZX, 2FFK, 2FXP, 2FYL, 2IXQ, 2JOD, 2JXG, 2JZN, 2JZO, 2K2S, 2K5X, 2K79, 2KRI, 2L7L, 2L9H, 2LE9, 2LGF, 2LV6, 2MBO, 2MBQ, 2MC6, 2MJ5, 2MMA, 2MP0, 2MZ6, 2MZW, 2ODG, 2OI3, 2PE9, 2PEA, 2PJH, 2VER, 3EZA, 3EZE, 6O22
Random proteins	115L, 1A0P, 1A1R, 1A3K, 1A7C, 1AF4, 1ARR, 1BOX, 1BRF, 1C2A, 1CHG, 1CZN, 1D01, 1ETU, 1F2Y, 1GCI, 1GF3, 1HI1, 1HOE, 1ITV, 1K1B, 1KHT, 1KI6, 1L02, 1L16, 1MBN, 1NGP, 1S7M, 1S9H, 1SBT, 1SH8, 1SLD, 1U8S, 1UD4, 1W9K, 1WA9, 1YK4, 1YRT, 2B97, 2JAR, 2LYM, 2LZT, 2MZ6, 2O2T, 2O3B, 2OV0, 2P7M, 2PRB, 2PRK, 2PVB, 2QXW, 2RIV, 2VKN, 2VRA, 2W4P, 2WTC, 2WX0, 2Z5O, 3BCJ, 3DHZ, 3HDF, 3J6P, 3KEE, 3PYP, 3SEA, 3VDF, 3VH9, 3WOB, 3WUY, 4B46, 4CMX, 4D3H, 4D7C, 4FGV, 4KBF, 4LBS, 4LZT, 4NNI, 4NPR, 4NS4, 4O00, 4O32, 4PT1, 4QYN, 4R21, 4R30, 4ROV, 4RSJ, 4RY7, 4UAT, 4UAU, 4ZM7, 5FGN, 5IHH, 5QTL, 5YCE, 6GOD, 6IPB, 6KJD, 6PTI

Results

All three hundred files from the three datasets produced error-free outputs, for both the analyses with and without the argument `-hy`. The outputs for four proteins were looked at in more details, and the assignation of UltimateSSAM seem correct as for the number of secondary structures.

One chain dataset

Example 1: 1BTA

The barstar is a small protein of 89 amino-acids synthesised by *Bacillus amyloliquefaciens*. It has four α-helices and one β-sheet with three parallel β-strands (Figure 1). Its PDB entry is `1bta.pdb` (Lubienski *et al.*, 1994).

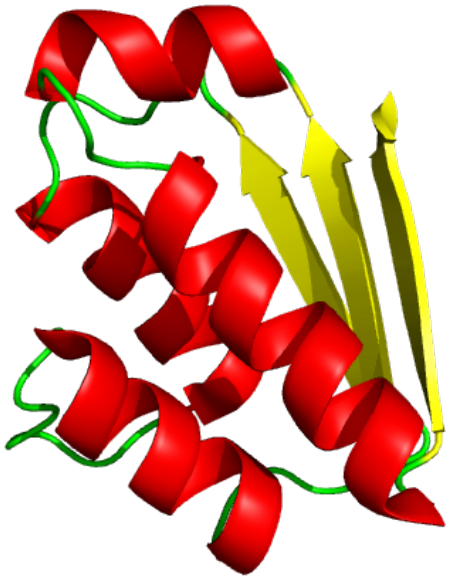

Figure 1: Structure of the barstar synthesised of *Bacillus amyloliquefaciens* (1BTA). Cartoon display coloured by secondary structure.

UltimateSSAM finds four α-helices and three parallel β-strand, forming a β-sheet.

Example 2: 1PX9

The native CnErg1 Ergtoxin is a small protein of 42 amino-acids synthesised by *Centruroides noxius*. It has one α-helice, one 3_{10}-helice, and one β-sheet with two antiparallel β-strands (Figure 2). Its PDB entry is `1px9.pdb` (Frénal *et al.*, 2004).

UltimateSSAM finds one α-helice, one 3_{10}-helice, and one β-sheet with two antiparallel β-strands.

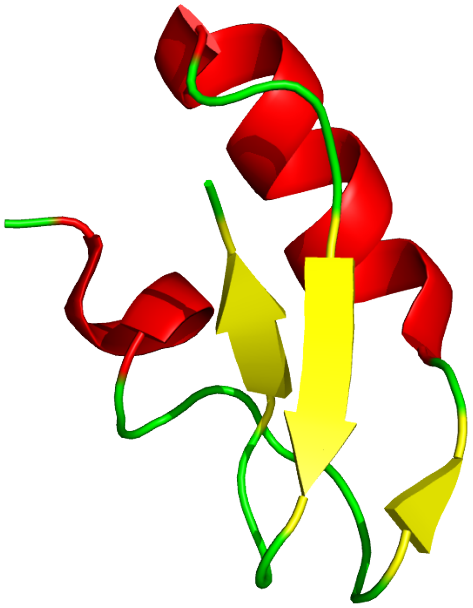

Figure 2: Structure of the CnErg1 Ergtoxin synthesised of *Centruroides noxius* (1PX9). Cartoon display coloured by secondary structure.

Multiple chains dataset

Example 3: 2L7L

The calmodulin-binding domain of calmodulin kinase I is a two-chains protein of 170 amino-acids (148 and 22 respectively) synthesised by *Homo sapiens*. It has nine α-helices and one antiparallel β-strand (Figure 3). Its PDB entry is `2l7l.pdb` (Gifford *et al.*, 2011).

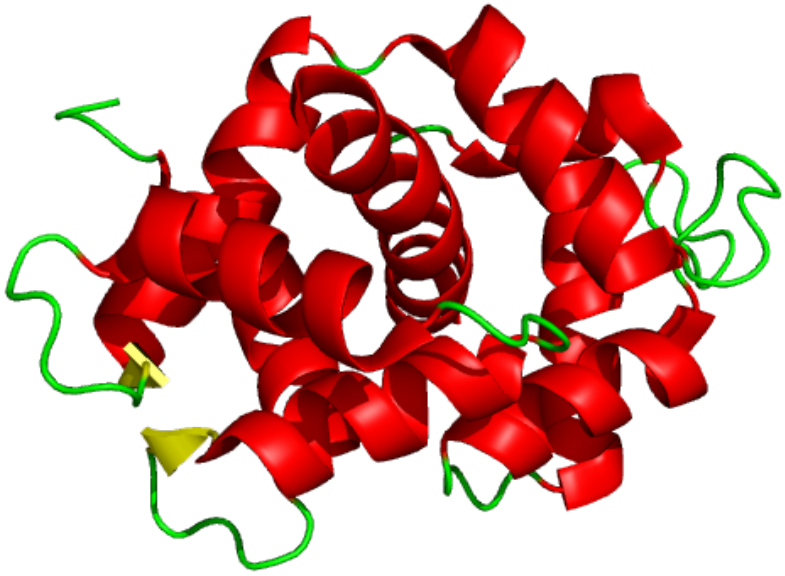

Figure 3: Structure of the calmodulin-binding domain of calmodulin kinase I synthesised of *Homo sapiens* (2L7L). Cartoon display coloured by secondary structure.

UltimateSSAM finds nine α-helices and one antiparallel β-strand.

Random entries

Example 4: 6PTI

The pancreatic trypsin inhibitor is a small protein of 58 amino-acids synthesised by *Bos tarurus*. It has one α-helice, one 3_{10}-helice, and one β-sheet with two antiparallel β-strands (Figure 4). Its PDB entry is `6pti.pdb` (Wlodawer *et al.*, 1987).

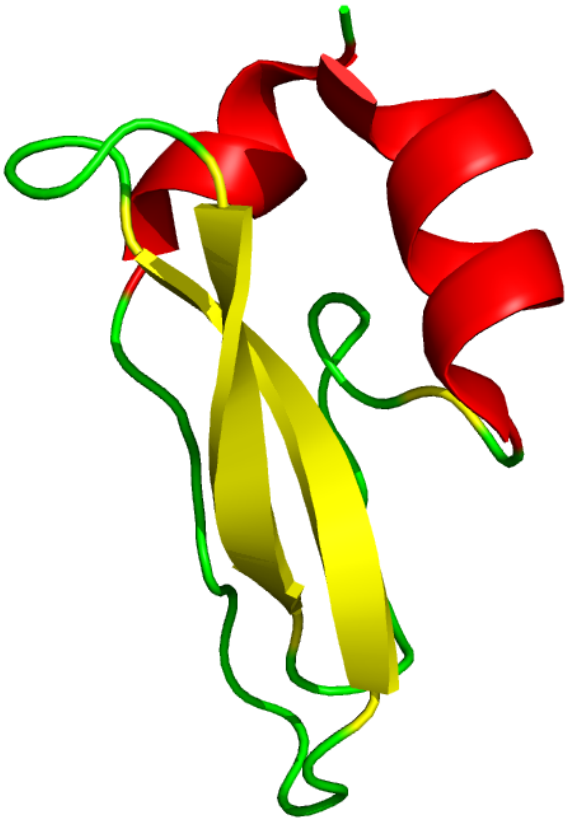

Figure 4: Structure of the pancreatic trypsin inhibitor synthesised of *Bos taurus* (6PTI). Cartoon display coloured by secondary structure

UltimateSSAM finds one α-helice, one 3_{10}-helice, and one β-sheet with two antiparallel β-strands.

Discussion

Overall, UltimateSSAM seems to agree with DSSP on most secondary structure assignment, with more than 96% matches for all the datasets for the eight conformational states (Table 2).

Table 2: Matches between DSSP and UltimateSSAM for the eight conformational states. Accuracy percentages and standard deviation of UltimateSSAM compared to DSSP for infile or added hydrogen atoms.

	Matches between DSSP and UltimateSSAM			
	With infile hydrogen atoms		With added hydrogen atoms	
Dataset	Percentage (%)	Standard deviation	Percentage (%)	Standard deviation
One chain	97.65	2.68	98.82	2.24
Multiple chains	96.09	4.57	96.83	4.41
Random proteins	–	–	97.62	2.96

There is a slight increase in performance with the argument `-hy` for both One chain and Multiple chains datasets, with an increased percentage (from 97.65% to 98.82%, and from 96.09% to 96.83% respectively) and decreased standard deviation (from 2.68 to 2.24, and from 4.57 to 4.41 respectively).

UltimateSSAM agrees more often with DSSP for monomers than polymers, since the One chain dataser has the highest percentage (98.82%) and lowest standard deviation (2.24). Furthermore, the Random proteins is also partly composed of monomers, contrary to the Multiple chains datasets, and also exhibit a higher percentage and standard deviation (97.62% and 2.96 respectively) than the Multiple chains dataset (96.83% and 4.41 respectively).

Regarding the matches between DSSP and UltimateSSAM for the eight conformational states reduced to three classes (Table 3), helices and coils seems to match more than sheets.

The percentage of matches is lower and the standard deviation is higher for every classes for both One chain and Multiple chain datasets (with the exception of standard deviation sheets for Multiple chains, which is greater by 0.1) for hydorgen atoms added with the `-hy`.

Coils are the most accurately assigned structures, followed closely by helices, and then by sheets. Sheets are probably the most affected by the placement of the hydrogen atoms, as they are more remote interactions than helices.

Table 3: Matches between DSSP and UltimateSSAM for the eight conformational states reduced to three classes. Accuracy percentages and standard deviation of UltimateSSAM compared to DSSP for the three classes. Helices (H) represent 3_{10}-helices (G), α-helices (H), and π-helices (I). Strands (E) represent β-strands (E) and β-bridges (B). Loops (C) represent bends (S), turns (T), and coils (C). Note that coils are represented by an empty string in UltimateSSAM.

	Matches between DSSP and UltimateSSAM					
	Helices (H = HGI)		Sheets (E = BE)		Coils (C = STC)	
Dataset	Percentage (%)	Standard deviation	Percentage (%)	Standard deviation	Percentage (%)	Standard deviation
One chain	97.25	12.042	97.59	3.96	99.39	1.41
One chain with added hydrogen atoms	99.95	0.47	97.96	3.18	99.97	0.34
Multiple chains	99.37	2.19	86.68	25.43	99.42	1.35
Multiple chains with added hydrogen atoms	99.94	0.43	87.02	25.53	99.99	0.06
Random proteins	99.80	1.47	94.17	11.31	99.934	0.42

Conclusion

As for now, UltimateSSAM assigns secondary structure based on DSSP (Kabsch and Sander, 1983) using the `dssp` mode, and can also compare its results to the DSSP version implemented by Maarten Hekkelman as part of a series of PDB-related databanks (Touw *et al.*, 2014) using the `dsspcompare` mode. Both the `ssam` and `ssamcompare` modes only perfom DSSP, but could be expanted in the future. The `dsspcompare` and `ssamcompare` modes could also be expanded to include more statistics, such as the specificity, or to give detailed statistics for each conformational state.

DSSP

The implementation of DSSP in UltimateSSAM performs relatively well compared to the actual DSSP (Tables 2 and 3), but some minor adjustments are still necessary to allow the assignment of secondary structures to every entry in the Protein Data Bank (Berman *et al.*, 2000).

The first improvment on UltimateSSAM would be to take into account the fact that residue assignement can be uncertain, and that as a result, for a given residue, multiple amino-acids may be possible and appear in the PDB file. For instance, at the position 36, `1est.pdb` (Sawyer *et al.*, 1978) has three different amino-acids: ARG (36A), SER (36B), and GLY (36C).

In the same fashion, with progresses in Nuclear magnetic resonance (NMR), PDB files can now contain several models for the same protein, with slightly different atomic coordinates. Namely, `2mkz.pdb` (Jiao *et al.*, 2014) contains 20 models. The way DSSP treats these models is to average the coordinates for each atom across all models. UltimateSSAM could be improved by doing the same.

In some PDB files, some residues have negative a negative residue number. Unlike DSSP, UltimateSSAM do not consider these residues to be directly precedent to the first residue.

A standard output is not present and should be added to UltimateSSAM. If this was to be the case, a filename for the DSSP output should be provided. Alternatively, output from the real DSSP could be generated directly as standard output and catched using the `os` package of python, with no intermediary file being created.

Also, SS-bonds are not supported for `.cif` input files, hetero-atoms are never taken into account, and solvent accessibility was not calculated.

All the improvements mentionned above concerns functionnalities that were not implemented in UltimateSSAM. However, a minor bug is present in UltimateSSAM regarding a functionnality that was implemented. In the output histogram, UltimateSSAM counts single bridges not involved in a sheet as a sheet of length 1, as can be seen with `2l7l.pdb` (Gifford *et al.*, 2011). They should be removed from the list containing all sheets after being assigned a sheet label.

ProMotif and Helanal

The different secondary structure assignment method implemented by ProMotif (Hutchinson and Thornton, 1994) and Helanal (Bansal *et al.*, 2000) such as polyproline helices, β-turns, β-hairpins, β-bulges, and helix axes were not implemented, but could be in the future, in order to have a true Ultimate Secondary Structure Assignment Method.

PDBx/mmCIF

PDBx/mmCIF were successfully implemented as input, as they are now the standrad format for PDB entries. However, on Windows, the available DSSP version (2.0) do not run with `.cif` files, which impedes the use of the `compare` modes with `.cif` files as input. Errors generated by this were not handled, but this could be fixed in the future.

References

Bansal, M., Kumart, S. and Velavan, R., 2000. Helanal: a program to characterize helix geometry in proteins. *Journal of Biomolecular Structure and Dynamics*, 17(5), 811–819. doi : 10.1080/07391102.2000.10506570.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242. doi : 10.1093/nar/28.1.235.

Chan, A.E., Hutchinson, E.G., Harris, D. and Thornton, J.M., 1993. Identification, classification, and analysis of beta-bulges in proteins. *Protein Science*, 2(10), 1574–1590. doi : 10.1002/pro.5560021004.

Craveur, P., Joseph, A.P., Rebehmed, J. and Brevern, A.G. de, 2013. B-bulges: extensive structural analyses of β-sheets irregularities. *Protein Science*, 22(10), 1366–1378. doi : 10.1002/pro.2324.

De Brevern, A.G., 2016. Extension of the classical classification of β-turns. *Scientific reports*, 6, 33191. doi : 10.1038/srep33191.

Frénal, K., Xu, C.-Q., Wolff, N., Wecker, K., Gurrola, G.B., Zhu, S.-Y., Chi, C.-W., Possani, L.D., Tytgat, J. and Delepierre, M., 2004. Exploring structural features of the interaction between the scorpion toxincnerg1 and erg k+ channels. *Proteins: Structure, Function, and Bioinformatics*, 56(2), 367–375. doi : 10.1002/prot.20102.

Gifford, J.L., Ishida, H. and Vogel, H.J., 2011. Fast methionine-based solution structure determination of calcium-calmodulin complexes. *Journal of biomolecular NMR*, 50(1), 71–81. doi : 10.1007/s10858-011-9495-3.

Hutchinson, E.G. and Thornton, J.M., 1994. A revised set of potentials for β-turn formation in proteins. *Protein Science*, 3(12), 2207–2216. doi : 10.1002/pro.5560031206.

Jiao, L., Ouyang, S., Shaw, N., Song, G., Feng, Y., Niu, F., Qiu, W., Zhu, H., Hung, L.-W., Zuo, X. *et al.*, 2014. Mechanism of the rpn13-induced activation of uch37. *Protein & cell*, 5(8), 616–630. doi : 10.1007/s13238-014-0046-z.

Kabsch, W. and Sander, C., 1983. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen–Bonded and Geometrical Features. *Biopolymers: Original Research on Biomolecules*, 22(12), 2577–2637.

Kumar, S. and Bansal, M., 1998. Geometrical and sequence characteristics of α-helices in globular proteins. *Biophysical Journal*, 75(4), 1935–1944. doi : 10.1016/S0006-3495(98)77634-9.

Lubienski, M.J., Bycroft, M., Freund, S.M. and Fersht, A.R., 1994. Three-dimensional solution structure and 13c assignments of barstar using nuclear magnetic resonance spectroscopy. *Biochemistry*, 33(30), 8866–8877. doi : 10.1021/bi00196a003.

Mansiaux, Y., Joseph, A.P., Gelly, J.-C. and Brevern, A.G. de, 2011. Assignment of polyproline ii conformation and analysis of sequence–structure relationship. *PloS one*, 6(3), e18401. doi : 10.1371/journal.pone.0018401.

Sawyer, L., Shotton, D., Campbell, J., Wendell, P., Muirhead, H., Watson, H., Diamond, R. and Ladner, R., 1978. The atomic structure of crystalline porcine pancreatic elastase at 2.5 Åresolution: comparisons with the structure of α-chymotrypsin.

Journal of molecular biology, 118(2), 137–208. doi : 10.1016/0022-2836(78)90412-6.

Touw, W.G., Baakman, C., Black, J., Beek, T.A. te, Krieger, E., Joosten, R.P. and Vriend, G., 2014. A series of pdb-related databanks for everyday needs. *Nucleic acids research*, 43(D1), D364–D368. doi : 10.1093/nar/gku1028.

Wlodawer, A., Nachman, J., Gilliland, G.L., Gallagher, W. and Woodward, C., 1987. Structure of form iii crystals of bovine pancreatic trypsin inhibitor. *Journal of molecular biology*, 198(3), 469–480. doi : 10.1016/0022-2836(87)90294-4.