**SUPPLEMENTARY MATERIAL**

**Physicochemical Properties**:

A total of 13 physicochemical features, including molecular weight, aromaticity, isoelectric point, secondary\_structure\_fraction\_α, secondary\_structure\_fraction\_β1, secondary\_structure\_fraction\_β2, molar\_extinction\_coefficient\_disulfid\_bridges, instability\_index, three indicators of protein flexibility (mean, maximum, minimum), and gravy, were calculated using the Biopython[1] package (version 1.79). Each of these features reflects a physical property of the protein and contributes to its localization or structural function. For detailed calculations of each feature, please refer to the source code of the Biopython package (<https://biopython.org/>). The ensuing descriptions present the methodologies employed for the computation of these individual features:

Molecular weight:

The molecular weight of a protein reflects the size of the protein molecule and related to the number of amino acids it contains.

Aromaticity:

According to Lobry’s 1994 calculation of the aromatic value of the protein, it is only the relative frequency of .[2]

Where is the relative frequency of amino-acid of kind in the protein and when the amino-acid is aromatic () and otherwise.

For Physical meaning of aromaticity see Burley’s article[3].

Isoelectric point:

Protein are amphidromical ionized in solution. Assuming a protein is present in a solution, When, , the number of positive and negative ions dissociated from the polar group of the protein is equal, and the net charge is 0. At this time, the value of the solution is the value of the protein. The value of a protein is specific, and is related to the protein structure, but not to the environmental . In proteins and peptides, this depends on the dissociation constant () of the seven amino acids at the end of the polypeptide and the charged groups of and groups. For specific calculation method, please refer to the source code of module. The isoelectric point reflects the dissociation level of protein in the electrolytic solution, and is also related to protein localization and function.[4]

secondary\_structure\_fraction\_α, secondary\_structure\_fraction\_β1 and β\_Fold (β2) secondary\_structure\_fraction\_β2:

The original name for these three features.

secondary\_structure\_fraction\_α indicate the Spiral fraction (α Corner).

secondary\_structure\_fraction\_β1 indicate the β Number of corners (β1).

secondary\_structure\_fraction\_β2 indicate the β Number of β\_Fold (β2)

Calculate fraction of helix, turn and sheet.

Returns a list of the fraction of amino acids which tend to be in Helix, Turn or Sheet.

Amino acids in helix（secondary\_structure\_fraction\_a）: include number of V, I, Y, F, W, L. (Note: V etc. include the following capital letters which are abbreviations of the 20 amino acids)

Amino acids in Turn（secondary\_structure\_fraction\_b1）: include number of N, P, G, S.

Amino acids in sheet（secondary\_structure\_fraction\_b2）: include number of E, M, A, L.

Returns a tuple of three floats (Helix, Turn, Sheet).

The physicochemical properties of these three structures see these references[5-8].

molar\_extinction\_coefficient\_disulfid\_bridge:

Calculate the molar extinction coefficient.

Calculates the molar extinction coefficient assuming cysteines (reduced) and cystines residues (Cys-Cys-bond). The specific calculation method can be found in the source code of module and refer to these papers[9, 10].

instability index:

A statistical analysis of 12 unstable and 32 stable proteins by Rogers et al. (1986) showed that there are dipeptides that occur with significantly different frequencies in unstable and stable proteins. So， we chose to look for the occurrence of 400 possible dipeptides in these two types of proteins. Assuming that the components of dipeptides are independent events, the expected (probable) occurrence of dipeptides can be calculated by the following equation:

Where and are the expected and observed incidence of dipeptide xy, respectively; and are the observed incidence of amino acids x and y, respectively, and T is the total number of amino acids in a particular class.

The chi-squared value between the observed and expected incidence of dipeptide (xy) for each class of proteins was calculated by the following equation:

The mean of the cardinal values for each type of protein was calculated using the formula:

Then was used as the confidence limit to select significant dipeptides for each class of proteins, where the significant dipeptide conditions used to distinguish between stable and unstable proteins were:

And

was also calculated from the observed occurrence of dipeptides in unstable and stable classes of proteins, respectively, using the following equation:

where and are the observed incidence of dipeptides in the unstable and stable classes of proteins, respectively. We chose the value of as the third group of dipeptides that satisfy the following conditions:

The potential incidence of each dipeptide in all three groups, , was calculated by the following equation:

The above three groups of dipeptides were further divided into two subgroups according to the significant difference of from unity (>1.5 or <0.64), and the conditions given in **Table II** were formulated according to the corresponding point positions. The dipeptides satisfying each condition given in **Table II** were classified according to the cardinal values and of each class of dipeptides, respectively. The of all the dipeptides (xy) satisfying the condition were summed to obtain the corresponding instability weight values for each condition. The impact factor for the ith condition was thus estimated by the following equation:

where and are the normalized values of the occurrence of dipeptides satisfying the condition in the unstable and stable classes of proteins, respectively. The influence factor of the condition is operated to bring it into the positive range, which is called the instability weight value, and is given by the following equation:

where LIF is the lowest observed influence factor. The contribution of each dipeptide to instability is obtained by summing the instability weight values corresponding to the conditions satisfied by the dipeptide, called dipeptide instability weight values (DIWV), and the DIWVs of all 400 combinations are represented as matrices in **Table III**, and then the protein instability index (II) is calculated using the DIWV with the following equation:

where is a dipeptide, L is the length of the sequence, and 10 is the scale factor.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table II.** Rule-based weightages | | | |
| Conditions | | Instability  Weight value | |
| If the dipeptide satisfies: | | | |
| 1. | Pus ≥ 1.50 and Ps < 1.50 | | +13.34 |
| 2. | Pus < 1.50 and Ps ≥ 1.50 | | - 1.88 |
| 3. | Pus ≤ 0.64 and Ps > 0.64 | | - 6.54 |
| 4. | Pus > 0.64 and Ps ≤ 0.64 | | +24.68 |
| 5. | Pus-s ≥ 1.50 | | +20.26 |
| 6. | Pus-s ≤ 0.64 | | - 7.49 |
| 7. | None of the above conditions | | + 1.0 |

**Table II**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table III.** GRP matrix of condition-based instability values for 400 possible dipeptides | | | | | | | | | | | | | | | | | | | | |
| First amino  acid of  dipeptide | Second amino acid of dipeptide | | | | | | | | | | | | | | | | | | | |
| W | C | M | H | Y | F | Q | N | I | R | D | P | T | K | E | V | S | G | A | L |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| W | 1.0 | 1.0 | 24.68 | 24.68 | 1.0 | 1.0 | 1.0 | 13.34 | 1.0 | 1.0 | 1.0 | 1.0 | -14.03 | 1.0 | 1.0 | -7.49 | 1.0 | -9.37 | -14.03 | 13.34 |
| C | 24.68 | 1.0 | 33.6 | 33.6 | 1.0 | 1.0 | -6.54 | 1.0 | 1.0 | 1.0 | 20.26 | 20.26 | 33.6 | 1.0 | 1.0 | -6.54 | 1.0 | 1.0 | 1.0 | 20.26 |
| M | 1.0 | 1.0 | -1.88 | 58.28 | 24.68 | 1.0 | -6.54 | 1.0 | 1.0 | -6.54 | 1.0 | 44.94 | -1.88 | 1.0 | 1.0 | 1.0 | 44.94 | 1.0 | 13.34 | 1.0 |
| H | -1.88 | 1.0 | 1.0 | 1.0 | 44.94 | -9.37 | 1.0 | 24.68 | 44.94 | 1.0 | 1.0 | -1.88 | -6.54 | 24.68 | 1.0 | 1.0 | 1.0 | -9.37 | 1.0 | 1.0 |
| Y | -9.37 | 1.0 | 44.94 | 13.34 | 13.34 | 1.0 | 1.0 | 1.0 | 1.0 | -15.91 | 24.68 | 13.34 | -7.49 | 1.0 | -6.54 | 1.0 | 1.0 | -7.49 | 24.68 | 1.0 |
| F | 1.0 | 1.0 | 1.0 | 1.0 | 33.6 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 13.34 | 20.26 | 1.0 | -14.03 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| Q | 1.0 | -6.54 | 1.0 | 1.0 | -6.54 | -6.54 | 20.26 | 1.0 | 1.0 | 1.0 | 20.26 | 20.26 | 1.0 | 1.0 | 20.26 | -6.54 | 44.94 | 1.0 | 1.0 | 1.0 |
| N | -9.37 | -1.88 | 1.0 | 1.0 | 1.0 | -14.03 | -6.54 | 1.0 | 44.94 | 1.0 | 1.0 | -1.88 | -7.49 | 24.68 | 1.0 | 1.0 | 1.0 | -14.03 | 1.0 | 1.0 |
| I | 1.0 | 1.0 | 1.0 | 13.34 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | -1.88 | 1.0 | -7.49 | 44.94 | -7.49 | 1.0 | 1.0 | 1.0 | 20.26 |
| R | 58.28 | 1.0 | 1.0 | 20.26 | -6.54 | 1.0 | 20.26 | 13.34 | 1.0 | 58.28 | 1.0 | 20.26 | 1.0 | 1.0 | 1.0 | 1.0 | 44.94 | -7.49 | 1.0 | 1.0 |
| D | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | -6.54 | 1.0 | 1.0 | 1.0 | -6.54 | 1.0 | 1.0 | -14.03 | -7.49 | 1.0 | 1.0 | 20.26 | 1.0 | 1.0 | 1.0 |
| P | -1.88 | -6.54 | -6.54 | 1.0 | 1.0 | 20.26 | 20.26 | 1.0 | 1.0 | -6.54 | -6.54 | 20.26 | 1.0 | 1.0 | 18.38 | 20.26 | 20.26 | 1.0 | 20.26 | 1.0 |
| T | -14.03 | 1.0 | 1.0 | 1.0 | 1.0 | 13.34 | -6.54 | -14.03 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 20.26 | 1.0 | 1.0 | -7.49 | 1.0 | 1.0 |
| K | 1.0 | 1.0 | 33.6 | 1.0 | 1.0 | 1.0 | 24.68 | 1.0 | -7.49 | 33.6 | 1.0 | -6.54 | 1.0 | 1.0 | 1.0 | -7.49 | 1.0 | -7.49 | 1.0 | -7.49 |
| E | -14.03 | 44.94 | 1.0 | -6.54 | 1.0 | 1.0 | 20.26 | 1.0 | 20.26 | 1.0 | 20.26 | 20.26 | 1.0 | 1.0 | 33.6 | 1.0 | 20.26 | 1.0 | 1.0 | 1.0 |
| V | 1.0 | 1.0 | 1.0 | 1.0 | -6.54 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | -14.03 | 20.26 | -7.49 | -1.88 | 1.0 | 1.0 | 1.0 | -7.49 | 1.0 | 1.0 |
| S | 1.0 | 33.6 | 1.0 | 1.0 | 1.0 | 1.0 | 20.26 | 1.0 | 1.0 | 20.26 | 1.0 | 44.94 | 1.0 | 1.0 | 20.26 | 1.0 | 20.26 | 1.0 | 1.0 | 1.0 |
| G | 13.34 | 1.0 | 1.0 | 1.0 | -7.49 | 1.0 | 1.0 | -7.49 | -7.49 | 1.0 | 1.0 | 1.0 | -7.49 | -7.49 | -6.54 | 1.0 | 1.0 | 13.34 | -7.49 | 1.0 |
| A | 1.0 | 44.94 | 1.0 | -7.49 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | -7.49 | 20.26 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| L | 24.68 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 33.6 | 1.0 | 1.0 | 20.26 | 1.0 | 20.26 | 1.0 | -7.49 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

**Table III**

The difference in the II value () for a pair of tripeptides 'axb' and 'ayb' obtained by changing the central residue and keeping the neighbouring residues the same was estimated by the relation:

We have analysed the data on substitutions in the central position of a tripeptide along with their associated II value, the formula that can be used for estimating the change in the II of a protein would be:

Where is the difference in II of a protein of length L before and after the replacement of residue x residue y.

The instability index is a measure of protein stability, indicating the tendency of a protein to denature or unfold under certain conditions. It is calculated based on the difference in free energy between two tripeptides with different central amino acids. A higher instability index indicates greater protein instability and a greater propensity for denaturation or unfolding. Therefore, the instability index is a useful metric for understanding the impact of amino acid substitutions on protein structural stability and function.

Calculate the protein flexibility indices (mean, minimum, maximum):

Calculate the protein flexibility indices first, then calculate mean and select minimum, maximum of protein flexibility indices outcome. For protein flexibility indices calculate detail see below（conference from .

Now that window size nine was found to be optimal in predictions with normalized B-values a new equation was determined:

Where .

Higher flexibility indices indicate that a protein is more flexible and dynamic, and may be more likely to participate in specific biological processes, such as protein interactions, structural transitions, and functional regulation.

gravy:

The specific calculation method can be found in the source code of module.

In addition in Informal reference help materials can reference from Jack Kyte and Russell.F.Doolittle’s paper, *A Simple Method for Displaying the Hydropathic Character of a Protein*, which discusses the hydrophobic GRAVY fraction of various types of membrane proteins，which presented in (https://github.com/yujuan-zhang/ProtLoc-mexl).

**Reference**

1. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.

2. Lobry, J.R. and C. Gautier, *Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes.* Nucleic Acids Res, 1994. **22**(15): p. 3174-80.

3. Burley, S.K. and G.A. Petsko, *Aromatic-aromatic interaction: a mechanism of protein structure stabilization.* Science, 1985. **229**(4708): p. 23-8.

4. Lee, K., et al., *PLPD: reliable protein localization prediction from imbalanced and overlapped datasets.* Nucleic Acids Res, 2006. **34**(17): p. 4655-66.

5. Dobb, M.G., *α-Helix in Fibrous Proteins.* Nature, 1965.

6. Kim, D.I., S.J. Han, and Y.B. Lim, *Unique behaviour of the alpha-helix in bending deformation.* Chem Commun (Camb), 2022. **58**(27): p. 4368-4371.

7. Porter, L.L., et al., *Many dissimilar NusG protein domains switch between alpha-helix and beta-sheet folds.* Nat Commun, 2022. **13**(1): p. 3802.

8. Struthers, M.D., R.P. Cheng, and B. Imperiali, *Design of a monomeric 23-residue polypeptide with defined tertiary structure.* Science, 1996. **271**(5247): p. 342-5.

9. Kendall, E.C. and D.F. Loewen, *The mechanism of oxidation-reduction potential: The oxidation-reduction potential of cysteine and cystine.* Biochem J, 1928. **22**(3): p. 669-82.

10. Spears, R.J., C. McMahon, and V. Chudasama, *Cysteine protecting groups: applications in peptide and protein science.* Chem Soc Rev, 2021. **50**(19): p. 11098-11155.