MC²hem a . o . . t . . m . h . . . p

Math/Chem/Comp 2022

33th MC2 Conference

NOVEL STRUCTURAL ATTRIBUTES BASED ON THE AVERAGE DISTANCES OF HYDROPHOBIC AMINO ACIDS IMPROVE MODELS FOR PREDICTING PROTEIN FOLDING RATES

Jadranko Batista[a], Antonija Kraljević[b,c], Bono Lučić[d]

[a]*University of Mostar, Faculty of Science and Education, Mostar, Bosnia and Herzegovina*

[b]*University of Mostar, Faculty of Mechanical Engineering, Computing and Electrical Engineering, Mostar, Bosnia and Herzegovina* (*antonija.kraljevic@fsre.sum.ba*)

[c]*University of Split, Postgraduate university study in biophysics, Split, Croatia*

[d]*Ruđer Bošković Institute, NMR Centre, Zagreb, Croatia*

INTRODUCTION

Protein folding is a very important problem in the biosciences that is studied experimentally but also by modelling (theoretical and simulation analyses) [1]. For the past 20 years, attempts have been made to model the dependences of the protein folding constants k_f (s^{-1}), which are equal to 1/(time needed for protein folding) from its primary, secondary and 3D structure.

It was observed in the very beginning that this is a size-dependent problem, which means that $\ln(k_f)$ significantly depends on the length of the protein sequence, *i.e.* on the number of amino acid residues in the protein [1]. Then, the dependence on the number of amino acids having regular secondary structure alpha or beta, on the topology of 3D structure, was also observed [2].

Initial studies indicated the possible importance of the physicochemical properties (like polarity/non-polarity) of segments of amino acids in the protein chain [3]. However, later studies rarely mentioned this as an important factor/descriptor in correlation with protein folding rates.

Most studies suggest that protein length (N) is the dominant factor determining protein folding rate [1,4,5]. Therefore, later this strong relationship between folding constants and length (N) was sought to improve **by the introduction** of the contribution of (1) protein secondary structure [4], (2) protein structural class and (3) the number of contacts of atoms in protein 3D structure within the sphere of defined radius. Details are summarized below.

(1) the **secondary structure of a protein** to compute effective protein length L_{eff} [4]:

$$L_{eff} = L - L_H + l_1 \times N_H \qquad \log(k_f) \sim const - (L_{eff})^P$$

where L_H is the number of residues in alpha-helical conformation, N_H is the number of alpha-helices, and l_1 means that the whole block (a helix) as l_1 chain residues was considered ($l_1 < 4$ residues, and it should be optimised in comparison with experimental data). The protein folding rate is then proportional to the power P of effective protein length L_{eff} [4], where P is determined by fitting to experimental $\log(k_f)$ data. We see that $\log(k_f)$ is linearly proportional to $P \cdot \log(L_{eff})$. On the set of 64 proteins [4], several topic models gave correlation coefficients between R =-0.79 and R = -0.82.

(2) the class of secondary structure to which a particular protein belongs, and

(3) the number of contacts of C_α atoms of amino acids observed in a sphere of defined radius (e.g. 6, 8, 10, .. angstroms) taking into account the topology of contacts [5]. This parameter named N_α gave better correlation (R = -0.83) with folding rates $\ln(k_f, s^{-1})$ of 80 proteins.

METHODS

In this paper, the protein folding rates, $\boldsymbol{k_f} = \mathbf{1}/\boldsymbol{t_f}$ (in s^{-1}), where $\boldsymbol{t_f}$ is the time needed to complete protein folding, are correlated (on the logarithmic scale) by protein-structure descriptors calculated from:

(**A**) the sum of distances of single amino acids (D-des): 20 distance descriptors - for 20 amino acids. (If amino acid A is found at positions 5, 12, 14, then the value of descriptor is 18.)

(**B**) the sum of distances of single amino acids from N- (N-des)and C-terminus (C-des) and the sum of products of distances (NC-des) from N- with the distance from C-terminus. And finally P-des is defined as the sum of the products of the distance of the amino acid from the middle of the protein sequence (20 descriptors in each group).

(If a protein has 50 amino acids, and if amino acid (AA) is found at positions 5, 12, 14, then the value of descriptor summing the distances from N-end is N-des = 4+11+13, from C-end is C-des = 45+38+36, the sum of products is simply NC-des = 4*45+11*38+13*36, and P-des = 15+13+11)

(**C**) the combinations of distances of single amino acids described in (A) and (B) when two o r three amino acids were combined into one descriptor, respectively.

One (+1) is added to these descriptor values due to the definition of the logarithmic function in cases where there are no amino acids in the sequence.

Scripts were written in Python to perform calculations on a set of 80 proteins [5] and another set with 95 proteins [6].

Protein sequence

VELSKKVTGKLDKTTPGIQIWRIENMEMVPVPTKSYGNFYEGDCYVLLSTRKTGSGFSYVNIHYWLGK
NSSQDEQGAAAIYTTQMDEYLGSVAVQ

AA_i AA_j

$$\text{D-des(AA)} = \ln\left\{\left[\sum_i \sum_{j>i} (AA_j - AA_i)\right] + 1\right\}$$

$$\text{N-des(AA)} = \ln\left\{\left[\sum_i (AA_i - 1)\right] + 1\right\}$$

$$\text{C-des(AA)} = \ln\left\{\left[\sum_i (N - AA_i)\right] + 1\right\}$$

$$\text{P-des(AA)} = \ln\left\{\left|\sum_i \left(\frac{N}{2} - AA_i\right)\right| + 1\right\}$$

$$\text{NC-des(AA)} = \ln\left\{\left[\sum_i (AA_i - 1) \cdot (N - AA_i)\right] + 1\right\}$$

Acknowledgement: This research is supported in part by the Croatian Ministry of Science and Education through basic grants given to their institutions and by the Croatian Government and the European Union through the European Regional Development Fund – the Competitiveness and Cohesion Operational Programme (KK.01.1.1.01) The Scientific Centre of Excellence for Marine Bioprospecting – BioProCro.

References

[1] D.N. Ivankov, A.V. Finkelstein, *Biomolecules* 10 (2020) 250.
[2] S. Lifson, C. Sander, *Nature* 282 (1979) 109–111.
[3] M. Levitt, *Biochemistry* 17 (1978) 4277–4285.
[4] D.N. Ivankov, A.V. Finkelstein, *PNAS,. U. S. A.* 101 (2004) 8942–8944.
[5] Z. Ouyang, J. Liang, *Protein Sci.* 17 (2008) 1256–1263.
[6] L. Censoni, L. Martinez, *Bioinformatics* 34 (2018) 4034–4038.

MOTIVATION

In this study, we investigated the correlation between $\ln(k_f, s^{-1})$ and the distance of:

(1) Single amino acids to each other along the protein sequence,

(2) The distribution of Single amino acids from the N- and C-terminus of sequences and from the middle of the protein chain, and

(3) Combinations under (1) and (2) when two and three amino acids were combined into one descriptor, respectively.

RESULTS

In modelling folding rates (protein folding constants) of 80 protein chains taken from [5] we correlated several descriptors derived from protein structure with $\ln(k_f)$.

A) The best correlations are obtained with the descriptor D-des for amino acid valine (V) being R = -0.833 for set1 (80 proteins) and R = -0.625 for set2 (95 proteins). These correlations are higher than with the the logarithm of the total number of valines ln(nV+1) R = -0.797 for set1 and R = -0.586 for set2 – Table 1.

Table 1. The best correlations obtained between ln(kf, s-1) and

Descriptors for a set of 80 proteins with 1 amino acids	Correlation of descriptors with ln (kf) for a set of 80 proteins	Descriptors for a set of 95 proteins with 1 amino acids	Correlation of descriptors with ln (kf) for a set of 95 proteins
Correlation with sequence lengths (N)	-0.727	Correlation with sequence lengths (N)	-0.560
Total number of valines *ln(nV+1)*	-0.797	Total number of valines *ln(nV+1)*	-0.586
D-des(V)	-0,833	D-des(V)	-0,625
P-des(V)	-0,795	P-des(G)	-0,618
N-des(V)	-0,787	C-des(G)	-0,606
C-des(V)	-0,782	C-des(V)	-0,603
NC-des(T)	-0,780	D-des(G)	-0,593

B) The best correlations for combinations of three amino acids are obtained with the descriptors in which they are taken: Valine (V) and Proline (P), with Tyrosine (Y), Threonine (T) and Serine (S) –Table 2.

Table 2. The best correlations obtained between ln(kf, s-1) and distance descriptors when three amino acids are taken together (PTV: distances between V, P and T are considered)

Descriptors for a set of 80 proteins with 3 amino acids	Correlation of descriptors with ln (kf) for a set of 80 proteins	Descriptors for a set of 95 proteins with 3 amino acids	Correlation of descriptors with ln (kf) for a set of 95 proteins
N-des(PTV)	-0,859	C-des(PVY)	-0,704
N-des(STV)	-0,858	C-des(VWY)	-0,702
D-des(STV)	-0,851	D-des(PVY)	-0,698
N-des(GTV)	-0,850	D-des(SVY)	-0,697
D-des(PTV)	-0,849	C-des(SVY)	-0,695

C) Example of a scattering diagram for a C-des (PVY) descriptor with three amino acids in a set of 95 proteins (Figure 1).

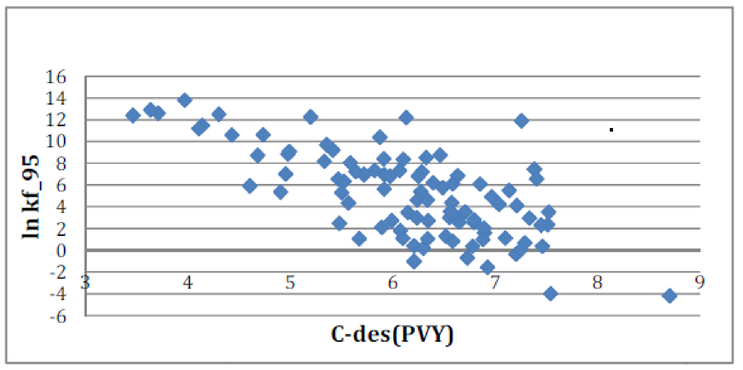

Figure 1. Scatter plot for C-descriptor for P,V,Y amino acids in a set of 95 proteins

The best obtained correlations (Tables 1 and 2) between $\ln(k_f, s^{-1})$ and distance-based descriptors are better than those with the protein length (the number of single amino acids in proteins) (Table 1), and are also better than the best correlations from literature [5] (R = -0.83, 80 proteins) and [6] (R = -0.64 and R = -0.69, 95 proteins).

CONCLUSION

Using simple sequence-based distance descriptors for single amino acids and their combinations (combinations of two and combinations of three amino acids) we obtained improved correlations with the protein folding rates ln(comparing to known literature data.

In addition, the resulting descriptors are: (1) significantly simpler than the often calculated contact order distance or average or reduced topological information used as descriptor, and (2) presented new descriptors are calculated only from the primary protein structure.

We will continue this preliminary research by calculation of optimal combination of two or three descriptors described above. We will also calculate combinations distances of a more than 3 (4 or 5) amino acids - whose distances will be summed into one descriptor.