The Q1n function: A potentially universal characteristic of proteins

Francisco Javier Lobo-Cabrera*.

Affiliation:

- Departamento de Sistemas Físicos, Químicos y Naturales, Universidad Pablo de Olavide, 41013, Sevilla, Spain.

Contact information:

- francisco.lobo6@gmail.com

Abstract

The principles governing protein structure are largely unknown. Here, a structural property universal ($R^2 = 0.978$) among proteins is reported. The characteristic under study --named the Q1n function-- relates a specific chemical-geometrical parameter with the number of residues of the protein. In the analysis, all X-Ray currently determined entries in the Protein Data Bank were studied. The limited variance is shown to be independent from secondary structure composition, compactness or relative surface area. The Q1n function allows an *a priori* protein structure prediction quality-check. Indeed, predictions with unexpected function values correspond to low ranks in the CASP12 experiment. The reason behind a specific, constant rule for protein folding remains unknown.

Introduction

Biomolecules, such as nucleic acids and proteins, exert their biological function in a 3D environment [1]. Therefore, comprehension of their spatial characteristics constitutes an active area of research [2]. For proteins, this is relevant as they carry out most cellular functions, such as structural functions, catalysis of biochemical reactions, transport or signaling [3] [4] [5].

The first protein structure determined was that of myoglobin in 1958 [6]. Since then, over 125,000 protein structures have been determined [7] [8]; mostly by X-Ray crystallography (~90%) and Nuclear Magnetic Resonance (~8%) [7] [8]. Simultaneously, a variety of simulation techniques have been developed that allow prediction of tertiary conformation based on primary sequence [9] [10]. Additionally, Molecular Dynamics and Quantum Mechanics methods enable refinements of the obtained structures [11] [12].

Despite the advancements in the field, prediction of three-dimensional structure remains challenging [13]. For example, it is known that proteins with similar primary sequence may have different 3D arrangements [14]. Also, some proteins change their conformation when interacting with specific substrates [15] or their structure can be modified by interaction with solvent molecules [16]. This emphasizes the need for more knowledge on protein structure and the generation of more sophisticated prediction methods. In turn, this new knowledge on structure will benefit understanding of protein functioning.

One way to make new discoveries on protein structure comprises statistical analysis of already available data. Currently, there are multiple databases containing polypeptide

information. Examples include the Protein Data Bank [7][8] –with experimental structural data, the DSSP database [17][18] –containing protein secondary structure-- or BRENDA [19][20] – an enzyme database. In these sources, the user can simply enter a specific query (e.g protein name or id) and receive the information related to that query as output. The referred databases also allow systematic access by users, so that an automatic protocol can retrieve information for multiple queries. In this fashion, they enable direct analysis of a large number of polypeptides. The power of the aforementioned databases, combined with proper analysis tools, can render new insights into proteins. As an illustration, both CATH [21] and SCOP2 [22] have used data from the Protein Data Bank to establish structural and evolutionary relationships between protein domains.

Databases may also be employed to extract potentially constant features of proteins. Up to know, only a few characteristics have been shown to be general among polypeptides. The 20 standard amino acids, the nature of the peptide bond [23], a limited number of secondary structure conformations [24] or the organization of proteins in domains are among them [25] [26]. These elements have helped gain understanding of general polypeptide structure and functioning. However, not all of them are truly universal. For example, over 120 additional amino acids have been identified as naturally occurring in proteins [27], some proteins do not contain domains [26] and the secondary structure relative composition is highly variable from protein to protein; where some proteins may contain only alpha helices and not beta sheets or vice versa [28]. Noteworthy, these general features have been identified by traditional means and not by statistical screening of databases. Hence, the possibility exists that there are hidden universal traits that only database mining can unveil.

In order to search in databases for unraveled constant characteristics it is necessary first to define those characteristics. In this case, the number of possible designs is enormous; one can measure practically any kind of physical, chemical or biological protein parameters, or even functions or ratios between them. For example, number of residues, compactness, electric charge, aliphatic index... Again, the existing databases provide the necessary input information. There are already tools for the computation of multiple of these parameters, such as ProtParam [29], PDBParam [30] or Vossolvox [31]. Once the characteristic to measure is defined, the polypeptides can then be scored. If the score is the same for all proteins, in principle the characteristic or trait is shown to be constant. Of course, since proteins are a highly diverse group of molecules [32] it is expected that only a few characteristics will be invariant.

In this article, a chemo-spatial parameter for polypeptides is presented. Referred as Q1, it quantifies the presence of spatial clusters of the same type of amino acid. Once defined, an analysis was conducted that included all X-Ray determined entries in the Protein Data Bank. Remarkably, the Q1 value for a protein structure is highly predictable. In fact, it is only required to take into account the number of amino acids the protein contains. This demonstrates the existence of a Q1 related function –the Q1n function, universal apparently among protein structures.

Q1 parameter definition

Given a polypeptide, let's associate each of its residues with two characteristics; i) a position in space and ii) a chemical nature --acidic, basic, polar or hydrophobic.

The position in space is established not as a volume but rather as a single point in space. This point is calculated as the geometrical center of the smallest cubic box encompassing the center of every atom in the residue (Figure 1a).

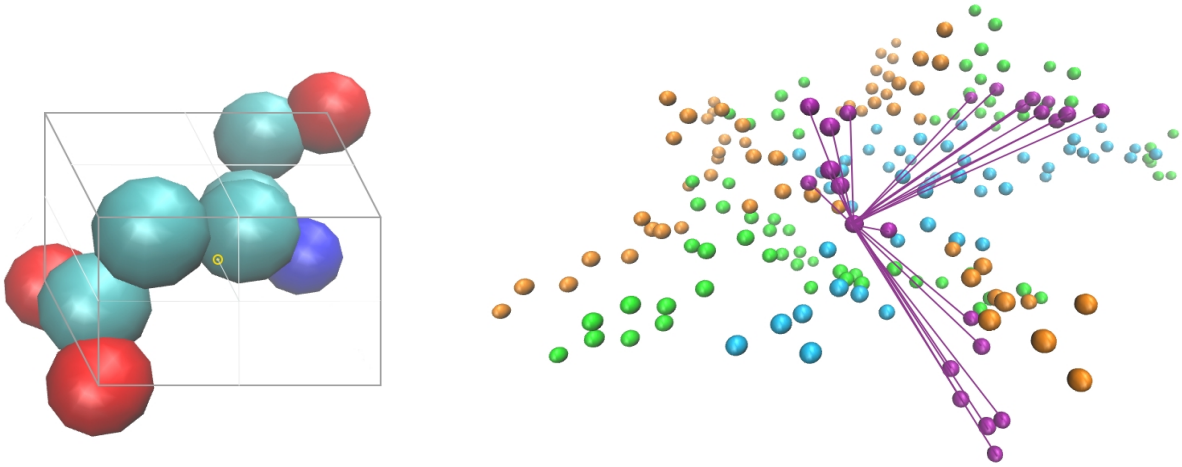

Figure 1. a (left): Determination of the space point associated with a residue. In the image, the spheres depict the different atoms of an amino acid, in this instance a glutamate residue. A cubic box is constructed as small as possible that contains the centers of every atom. Then, the geometrical center of the cubic box (yellow point) is considered as the space point for the residue. **B (right): Representation of the Euclidean distances of one residue with other residues of the same type.** Every sphere here depicts the space point of a residue in an example protein, and is coloured according to its type (acidic, basic, polar or hydrophobic). The lines represent all the Euclidean distances to be calculated for one residue.

On the other hand, the chemical nature of each residue corresponds to that of its R-chain. Specifically, Asp and Glu are designated as acidic, whilst Arg, His and Lys as basic. Polar residues include Ser, Thr, Asn and Gln. Finally, hydrophobic residues comprise Ala, Val, Ile, Leu, Met, Phe, Tyr and Trp. The rest of amino acids (Cys, Sec, Gly, Pro) are not included in

the analysis.

Once the space point and chemical nature for each residue have been established, the calculation of Q1 can proceed. The basic measurement involves the determination of Euclidean distances between residues of the same kind. That is, the minimum distance between the space point of residues of the same chemical nature. Figure 1b shows an illustration of this.

For every residue, the Euclidean distance with every other residue of the same type in the protein is calculated. Then, the inverse of each distance is squared, and the are results added up, rendering Q1. A formal definition can be seen in Equation (1).

$$Q1=\sum_{i=1}^{nA_n}\sum_{j=1}^{nA_n}\frac{1}{d_{nA_i,nA_j\forall i\neq j}^{\ 2}}+\sum_{i=1}^{nB_n}\sum_{j=1}^{nB_n}\frac{1}{d_{nB_i,nB_j\forall i\neq j}^{\ 2}}+\sum_{i=1}^{nP_n}\sum_{j=1}^{nP_n}\frac{1}{d_{nP_i,nP_j\forall i\neq j}^{\ 2}}+\sum_{i=1}^{nH_n}\sum_{j=1}^{nH_n}\frac{1}{d_{nH_i,nH_j\forall i\neq j}^{\ 2}}$$

Equation 1. Calculation of Q1. First, the Euclidean distance (d) of each residue space point (i) with every other residue space point (j) of the same type is calculated. The types are denoted as (A) for acidic, (B) for basic, (P) for polar and (H) for hydrophobic. All the distances are squared and the inverse is added up. The process is repeated for the nA acidic residues, nB basic residues, nP polar residues and nH hydrophobic residues of the protein.

An alternative parameter (Q2) is also presented in this work. Although being analogous to Q1, it does not include chemical nature specification. So, its definition --contained in Equation (2)-- becomes much simpler. As discussed later, Q2 offers less predictive power than Q1.

$$Q2=\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{d_{n_i,n_j\forall i\neq j}^{\ 2}}$$

Equation 2. Calculation of Q2. The Euclidean distance (d) of each residue space point (i) with every other residue space point (j) is calculated. The values are squared and the inverse added up. The process is repeated for all n residues.

Both Q1 and Q2 constitute measurements of residue spatial clustering. By measuring the inverse of squared distances, their value becomes higher for increased amino acid clustering. In the calculations, Bio.PDB [33] [34] was used to extract the atom coordinates of each residue.

Results

The Q1n function**:**

The second step in this work involved calculating Q1 (see prior section) in a large set of proteins. For this purpose, the Protein Data Bank (PDB) archive --specifically those entries determined by X-Ray diffraction-- was employed. Currently, PDB includes structural data of more than 130,000 polypeptides and peptides. Of these, approximately 90% have been resolved by X-Ray diffraction [7][8], a technique that offers very high atomic resolution [35].

Results of the analysis of X-Ray entries from PDB show that a structural function exists, universal ($R^2 = 0.978$) among proteins. This function relates the Q1 value and the number of residues of a protein (n). Specifically, it shows that Q1 increases its value with the number of residues following a specific function (Figure 2).

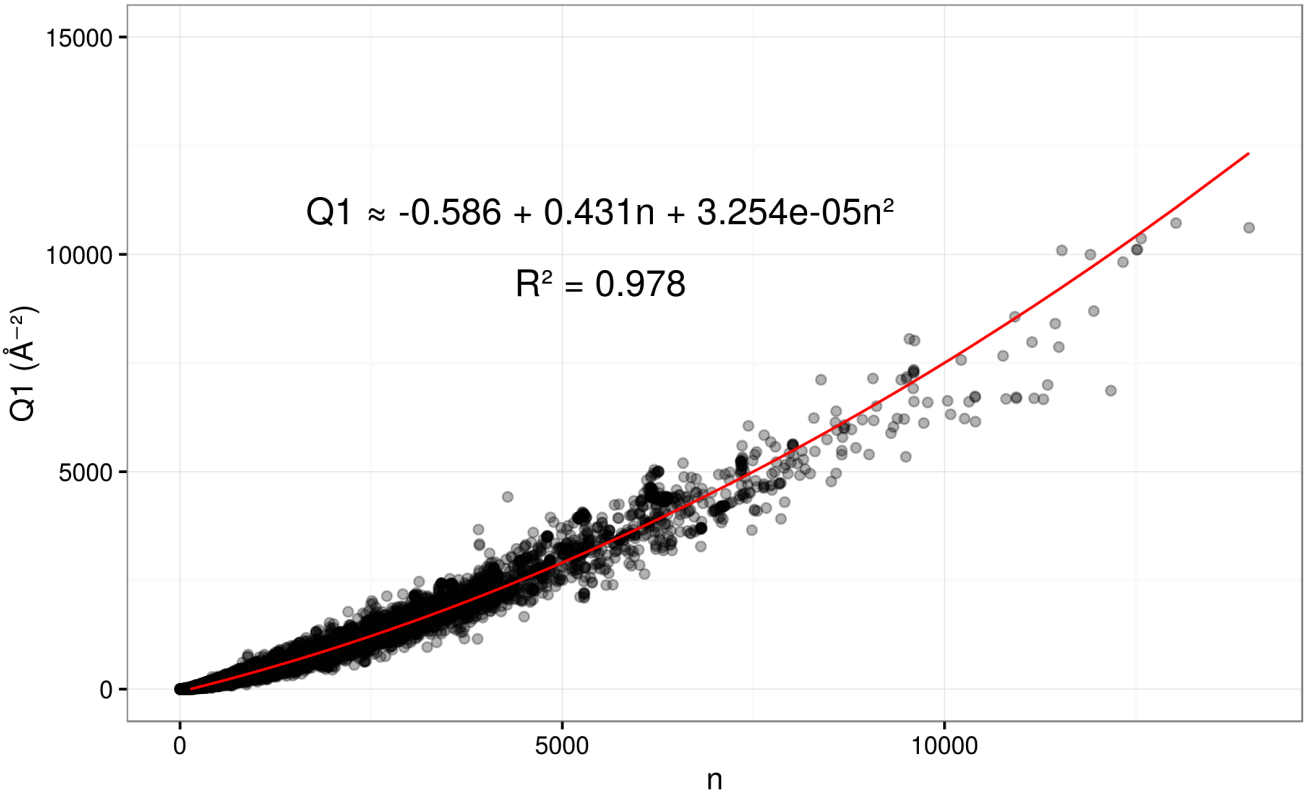

Figure 2. The Q1n function. A second-degree polynomial describes the relationship between Q1 and the number of residues (n). For the calculation, distances between residue space points are measured in Ångströms. Those entries where Q1/n > 1.032 $Å^{-2}$ --i.e less than 0.1% of total entries-- were discarded as outliers.

The other defined parameter --Q2-- was likewise assessed to search for a function. In this case, the model showed lower predictive power correlation (R^2 = 0.898).

Independence from different factors:

Albeit that the Q1n function is shown to be highly conserved, there is still some limited variation from expected Q1 among proteins. Deviations consist of the difference between obtained Q1 values and those predicted by the function. These variations could be simply

stochastic, but several analyses were performed to try and correlate those differences with key structural characteristics. In this manner, the function was evaluated in its relation with secondary structure composition, protein compactness and surface area per residue. The aim was to test the independence between the Q1n function and each characteristic. The details on the calculations performed are described in the Method section.

a) Secondary structure composition:

Proteins display a limited number of secondary structure types, being alpha helix and beta strand the most abundant [36] [37]. To determine whether secondary structure affects Q1 values, the percentage of residues in beta strand of each protein was calculated. Approximately, this percentage becomes inversely proportional to that of residues in alpha helix.

Figure 3a shows absence of correlation between secondary structure and the Q1n function.

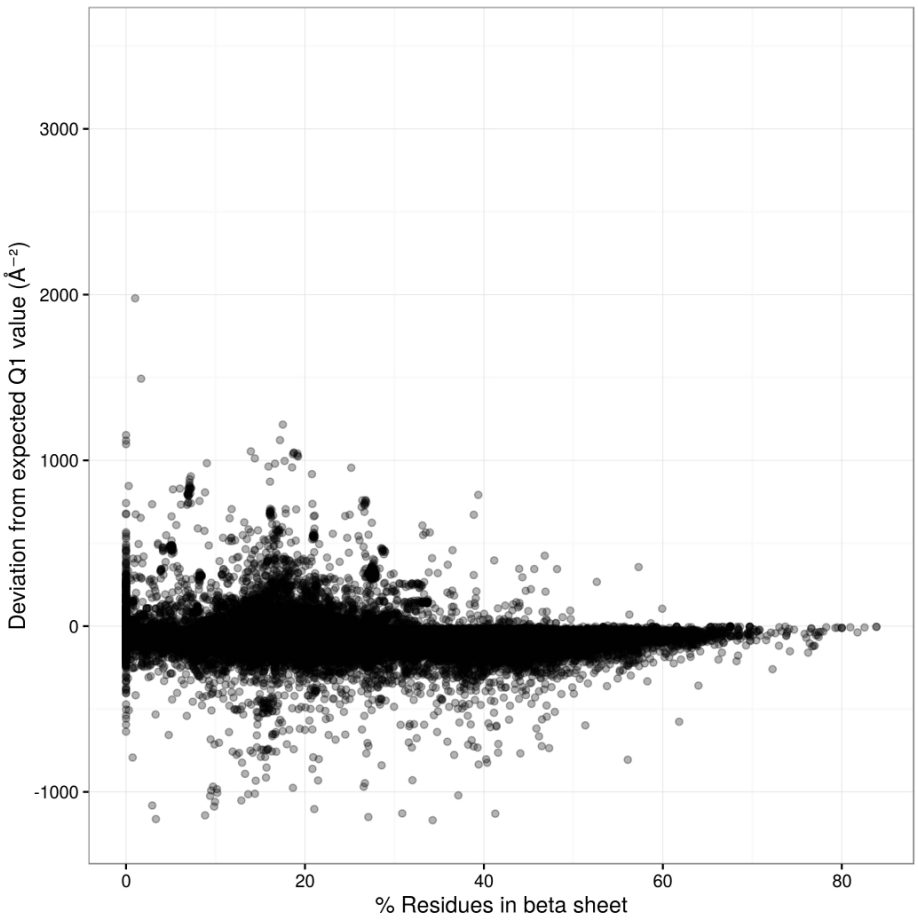

Figure 3a. Independence from secondary structure. Plotting of the percentage of residues in beta sheet and the deviation from the expected Q1 value.

b) Compactness:

Protein compactness refers to the arrangement, more or less extended, of the residues in a protein. As Q1 measures distances between residues, it was hypothesized that it may be dependent on the overall compactness.

As depicted in Figure 3b, there is very low correlation between the two variables, indicating little effect of protein compactness.

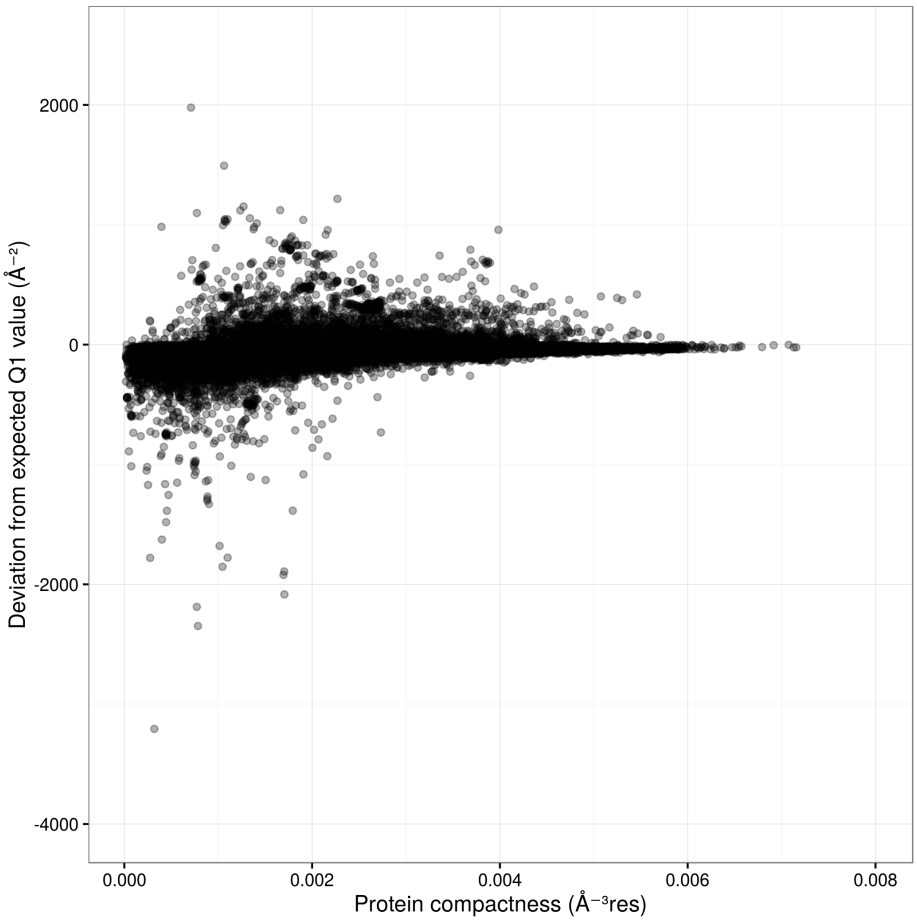

Figure 3b. Independence from protein compactness. Protein compactness here is referred as the residue density in an imaginary sphere of minimum volume containing all the protein atoms. Note: One point whose compactness is > 0.06 $Å^{-3}$ res was not plotted.

c) Surface area per residue:

Another studied structural parameter was surface area per residue. Polypeptides with higher proportion of buried residues; that is, residues not accessible by the solvent, will have lower surface areas per residue.

As shown in Figure 3c, the spatial function under study in this work does not depend apparently on surface area per residue.

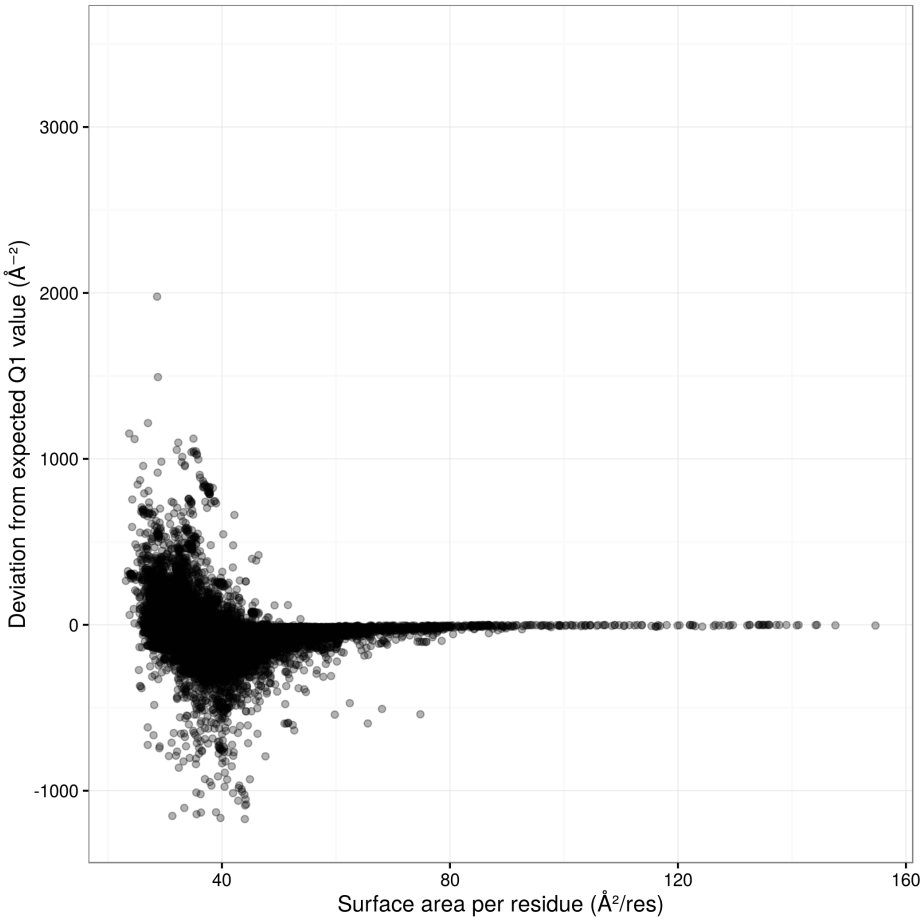

Figure 3c. Independence from surface area per residue. The surface area for each protein was calculated with the 3V software [38], using a probe radius of 1.5 Å.

An *a priori* method for quality-check of structure prediction:

An empirical function enables in principle inference on structure. Therefore, an *a priori* method can be constructed to check the quality of structure predictions, without the need to compare with experimental results.

To explore the use of the Q1 rule as a quality-check, data from the CASP12 experiment [39] was analyzed. CASP12 [39] constituted a recent evaluation of computational methods for

protein structure prediction. It ranked simulations from different research groups in relation to an obtained experimental structure. CASP12 assessed in this way predictions of a total of 53 protein targets. The goal here was to analyze whether there was a correlation between CASP12 rankings and deviations from expected Q1 values.

Results show that i) the Q1n function does not allow to compare between good predictions, but rather to ii) identify poor predictions. This can be explained by the fact that even experimental structures do not have a difference from expected Q1 equal to 0. So, structure predictions do not necessarily need to have Q1 differences equal to 0 to be optimum.

On the other hand, in the PDB X-Ray data set 90% of the entries have absolute difference < 111.6866 $Å^2$ from the expected Q1 value. In this manner, it would be relatively safe to discard (90% probability) those predictions displaying differences larger than 111.687 $Å^2$. This opens the door for *a priori* quality assessment of structure predictions. Accordingly, Figure 4 depicts how in CASP12 predictions surpassing the referred limit correspond to low ranks.

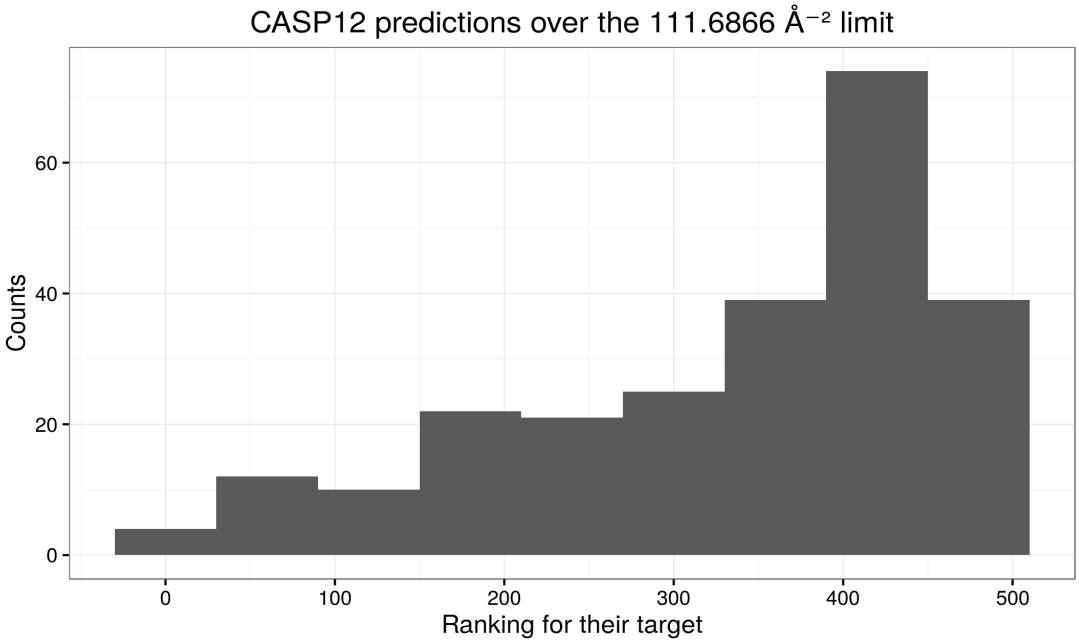

Figure 4. Checking the validity of the test on the CASP12 data. This diagram represents the rankings of those predictions surpassing the 113.2107 Å² difference limit. Instead of a uniform distribution, the diagram shows that low rankings are favored.

Origins of the function Q1n**:**

Some hypotheses about the Q1n function are considered here. These include a thermodynamic and a biological perspective.

a) Thermodynamic approach: a general principle for protein folding strongly suggests free energy stabilization as a key factor. In order to test this theory, a thermodynamic variable was chosen for the variety of proteins in the PDB X-Ray data set. In this case, it was the temperature of the system, meaning the temperature at which the protein is normally found in nature. A thermodynamic basis for the function could mean that system temperature is related with deviations from expected Q1 values.

Figure 5 shows that no real correlation is found between system temperature and the function Q1n. Nevertheless, this does not exclude thermodynamics as a key factor.

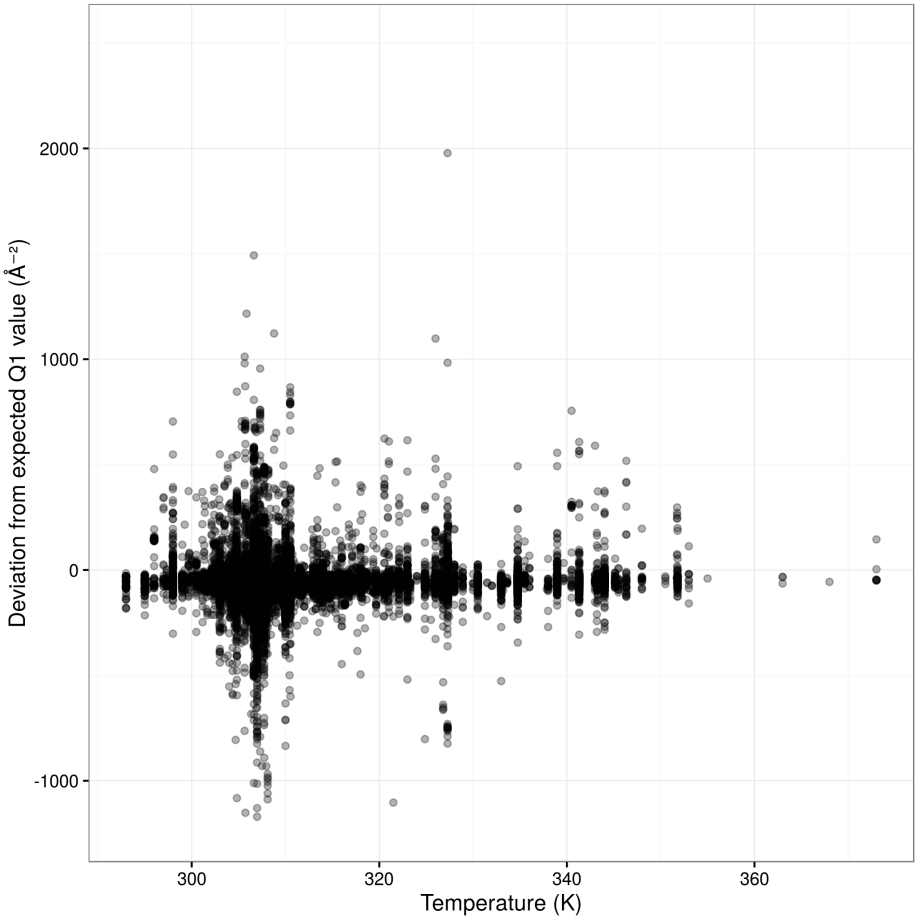

Figure 5. System temperature and the Q1 function. System temperature corresponds to the approximate temperature at which the polypeptide is present in nature. It is calculated as the average optimum temperature of the organism's enzymes included in the BRENDA database.

b) Biological approach: Other explanations regarding this function may be biological. Nonetheless, they would also have to be shared among proteins of all kind.

According to its definition, high values of Q1 imply that the protein contains large islands of

amino acids of one type. This potentially can lead to strong interactions with molecules of similar chemical nature as those amino acids. This is especially relevant since the number of residues is dependent on the protein's mass. Alternatively, low values of Q1 would imply soft interaction with other molecules, as no great areas of amino acids of any type would exist. Life depends on molecular recognition, i.e interaction between molecules. These interactions must not be either too strong (molecules would aggregate unspecifically) or too weak (recognition would never occur). A common, shared rule like Q1 could be the *radio frequency* at which molecular recognition among proteins --or among proteins and other molecules-- must happen.

Simultaneously, deviations from Q1 expected values exist for different proteins of same number of residues. This could be related to fine-tuning molecular interactions, and thus protein function. To check this, a GO-terms based analysis was conducted (see details in Methods). The results show that only one GO-term (p-value < 0.05) is underrepresented in proteins with comparative low Q1 values and overrepresented in proteins with comparative high Q1 values. The referred GO-term is GO:0003824, which corresponds to catalytic activity. This may be consistent with the hypothesis of increasing Q1 values implying stronger molecular recognition.

Methods

Easy-to-use Python code for the calculations is available at https://github.com/UPO-Sevilla-Fco-Javier-Lobo-Cabrera/Q1n . On the other hand, ggplot2 [40] was employed in the

generation of the graphics.

Calculation of secondary structure composition:

The DSSP database and related program mkdssp [17] [18] were employed as a dependency. In this manner, the percentage of residues in beta strand of each X-Ray PDB entry was obtained.

Calculation of protein compactness:

To evaluate protein compactness, the following strategy was followed. First, a sphere of minimum volume was generated for each protein that contained all its atoms. Then, the number of residues per sphere volume unit; i.e residue density in the sphere, was calculated. Since the spherical form is the most compact shape, the calculated residue density constitutes a direct measure of compactness for each entry.

Calculation of surface area per residue:

To compute surface areas of proteins, the software 3V [38] was used as a dependency. Probe radius was set to 1.5 Å. Subsequently, the data were normalized by the number of residues.

Assessment of CASP12 protein predictions using the Q1 function:

Prediction structures and their corresponding rankings were obtained obtained from: http://predictioncenter.org/download_area/CASP12/ .

For all the structures, the theoretical and real Q1 value was calculated. Finally, the deviation from expected and real Q1 was plotted against the structure's ranking.

Calculation of system temperature:

System temperature refers here to the normal temperature in which the protein is found in nature. Calculating this value required making several approximations. Firstly, it was assumed that its value matched the species optimum growth temperature. In this way, all the proteins produced by a species would have the same system temperature. Also, that optimum growth temperature would coincide with the average optimum temperature of the species' enzymes. The BRENDA database [19][20] provided the optimum temperatures of each species' enzymes. These approximations, along with the use of the BRENDA database, enabled the calculation of system temperature for a total of 87,061 PDB entries.

GO-term analysis of Q1 deviations :

First, all X-Ray PDB entries were ordered according to their deviation from expected Q1 values. Then, two groups were selected; a) those entries under percentile 10, and b) those above percentile 90. The GO-terms present in these two groups were extracted, and Fisher

tests were carried out for each GO-term. In this way, a series of overrepresented GO-terms and underrepresented GO-terms (p-value < 0.05) were obtained for groups a) and b).

Discussion

The present work reports a presumably universal characteristic of proteins. Along with the nature of the peptide bond, the presence of 20 standard amino acids or the organization in domains, this could be one of the few constant characteristics in polypeptides. This is very much consistent with the existence of a *protein linguistics*, as predicted by Mohammed AlQuraishi [41]: *"...The end result of all this would be the emergence of something resembling a linguistic structure, a grammar that defines the reusable parts and how these parts can be combined to form larger assemblies. Given that this is biology, it's unlikely to be rigid or minimal. It would be messy and hacky, with many exceptions and ad hoc evolutionary optimizations. But the manifold would be there, potentially discoverable and learnable..."* .

Finally, the Q1n function can be employed to assess protein structure predictions without the need of experimental results. However, this test is shown to identify only possible low quality predictions.

Acknowledgements

I thank Fernando Govantes, Alejandro Cuetos, Brian Jiménez-García, Juan Neftalí Morillo García and Francisco Javier García Moscoso for their advice.

Competing interests

The author declares no competing interests.

Author contributions

FL is the author of this work.

Data availability

The datasets generated during and/or analysed during the current study are available in the Github repository, https://github.com/UPO-Sevilla-Fco-Javier-Lobo-Cabrera/Q1n .

References

1.Durham, A., Gruber, A., Huynh, C. & Portillo, H. *Bioinformatics in tropical disease research*. (U.S. National Library of Medicine, NCBI, 2008).

2.Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP) - round x. Proteins: Structure, Function, and

Bioinformatics 82, 1-6 (2013).

3.Whitford, D. Protein structure & function. (Wiley, 2003).4.Gruenbaum, Y. & Foisner, R. Lamins: Nuclear Intermediate Filament Proteins with Fundamental Functions in Nuclear Mechanics and Genome Regulation. Annual Review of Biochemistry 84, 131-164 (2015).

4.Poulos, T. Heme Enzyme Structure and Function. Chemical Reviews 114, 3919-3962 (2014).

5.Grecco, H., Schmick, M. & Bastiaens, P. Signaling from the Living Plasma Membrane. Cell 144, 897-909 (2011).

6.Kendrew, J. et al. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181, 662-666 (1958).

7.Berman, H. et al. The Protein Data Bank. Nucleic Acids Res 28, 235-242 (2000).

8.RCSB PDB: Homepage. Rcsb.org (2018) at <http://www.rcsb.org/>

9.Petrey, D. & Honig, B. Protein Structure Prediction: Inroads to Biology. Molecular Cell 20, 811-819 (2005).

10.Peng, J. & Xu, J. Raptorx: Exploiting structure information for protein alignment by statistical inference. Proteins: Structure, Function, and Bioinformatics 79, 161-171 (2011).

11.Shen, R. et al. Structural Refinement of Proteins by Restrained Molecular Dynamics Simulations with Non-interacting Molecular Fragments. PLOS Computational Biology 11, e1004368 (2015).

12.Yu, N., Yennawar, H. & Merz, K. Refinement of protein crystal structures using energy

restraints derived from linear-scaling quantum mechanics. Acta Crystallographica Section D Biological Crystallography 61, 322-332 (2005).

13.Kryshtafovych, A. et al. Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10. Proteins: Structure, Function, and Bioinformatics 82, 26-42 (2014).

14.Kosloff, M. & Kolodny, R. Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins: Structure, Function, and Bioinformatics 71, 891-902 (2008).

15.Fink, A. Natively unfolded proteins. Current Opinion in Structural Biology 15, 35-41 (2005).

16.Lai, J., Ambia, J., Wang, Y. & Barth, P. Enhancing Structure Prediction and Design of Soluble and Membrane Proteins with Explicit Solvent-Protein Interactions. Structure 25, 1758-1770.e8 (2017).

17.Joosten, R. et al. A series of PDB related databases for everyday needs. Nucleic Acids Research 39, D411-D419 (2010).

18.Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-2637 (1983).

19.Placzek, S. et al. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Research 45, D380-D388 (2016).

20.Enzyme Database - BRENDA. Brenda-enzymes.org (2018) at <http://www.brenda-enzymes.org/>

21.Dawson, N. et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Research 45, D289-D295 (2016).

22.Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Research 42, D310-D314 (2013).

23.Edison, A. Linus Pauling and the planar peptide bond. Nat Struct Biol 8, 201-202 (2001).

24.Sneha, P. & Priya Doss, C. Advances in Protein Chemistry and Structural Biology. 181-212 (Elsevier, 2016).

25.Jaenicke, R. Protein folding: local structures, domains, subunits, and assemblies. Biochemistry 30, 3147-3161 (1991).

26.Marsh, J. & Teichmann, S. How do proteins gain new domains?. Genome Biology 11, 126 (2010).

27.Ambrogelly, A., Palioura, S. & Söll, D. Natural expansion of the genetic code. Nature Chemical Biology 3, 29-35 (2007).

28.Toll-Riera, M., Rado-Trilla, N., Martys, F. & Alba, M. Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. Molecular Biology and Evolution 29, 883-886 (2011).

29.Walker, J. The Proteomics Protocols Handbook. 571-607 (Springer, 2005).

30.Nagarajan, R. et al. PDBparam: Online Resource for Computing Structural Parameters of Proteins. Bioinformatics and Biology Insights 10, BBI.S38423 (2016).

31.Voss, N., Gerstein, M., Steitz, T. & Moore, P. The Geometry of the Ribosomal Polypeptide

Exit Tunnel. Journal of Molecular Biology 360, 893-906 (2006).

32.Zaretsky, J. & Wreschner, D. Protein Multifunctionality: Principles and Mechanisms. Translational Oncogenomics 1, 99-136 (2008).

33.Cock, P. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422-1423 (2009).

34.Hamelryck, T. & Manderick, B. PDB file parser and structure class implemented in Python. Bioinformatics 19, 2308-2310 (2003).

35.Smyth, M. & Martin, J. x Ray crystallography. Molecular Pathology 53, 8-14 (2000).

36.Haimov, B. & Srebnik, S. A closer look into the α-helix basin. Scientific Reports 6, (2016).

37.Badal, S. & Delgoda, R. Pharmacognosy. 477-494 (2017).

38.Voss, N. & Gerstein, M. 3V: cavity, channel and cleft volume calculator and extractor. Nucleic Acids Research 38, W555-W562 (2010).

39.Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)-Round XII. Proteins: Structure, Function, and Bioinformatics 86, 7-15 (2017).

40.Wickham, H. Ggplot2 - elegant graphics for data analysis. (Springer-Verlag).

41.AlQuraishi, M. Protein Linguistics. *Some Thoughts on a Mysterious Universe* (2019). At <https://moalquraishi.wordpress.com/2018/02/15/protein-linguistics/>.