**Big Data Analysis and Simulation Reveals a Highly Conserved Amino Acid Clustering Trait in Protein Structure**

Francisco Javier Lobo-Cabrera¹\* and José Antonio Prado-Bassas²

(\*) Corresponding author.

1. Departamento de Sistemas Físicos, Químicos y Naturales, Universidad Pablo de Olavide, 41013, Sevilla, Spain. E-mail address: francisco.lobo6@gmail.com. ORCID: 0002-9592-6863.

2. Departamento de Análisis Matemático, Facultad de Matemáticas e IMUS, Universidad de Sevilla, 41090, Sevilla, Spain. E-mail address: bassas@us.es. ORCID: 0002-5670-5400.

**Abstract**

Large-scale analysis of currently available protein structure data reveals the existence of a highly consistent, empirical law (R² = 0.978) linking two variables: i) a specific chemical-geometrical parameter and ii) the number of residues (size) of the protein. The chemical-geometrical parameter, referred here as *Q*, quantifies the degree of spatial clustering of amino acids of the same chemical type (polar, hydrophobic, acidic or basic) in each protein. Once defined *Q*, the analysis involved calculating its value for all X-Ray currently determined structures available in the Protein Data Bank, showing that a specific, remarkably uniform law that relates the obtained *Q* values and protein sizes (n) can be obtained. This equation cannot be explained by the existence of the hydrophobic core, as the property is shown to be independent from protein compactness and surface area, so that proteins of same size but very different level of globularity will display almost identical *Q* value. Novel stochastic simulations were designed to infer possible reasons behind this conserved clustering trait in protein structure, demonstrating that the origin of this conserved equation may be consequence of the presence of amino acid clusters of same chemical type (polar, hydrophobic, acidic or basic) following specific morphology rules. These morphology rules are apparently shared among the great majority of proteins, in what is referred here as the *Q-*mosaic model. Given their apparently almost universal presence in protein structure, the presented clustering trait may effectively play an important role in the biological properties of polypeptides. The *Q*-mosaic model can even be appreciated directly visually using reputed, well-known visualization tools.

**Introduction**

Biomolecules, such as nucleic acids and proteins, exert their biological functions in a 3D environment 1. Therefore, comprehending their spatial characteristics remains a crucial and active area of research 2. Proteins, in particular, are pivotal as they perform most cellular functions, including structural support, catalysis of biochemical reactions, and roles in transport and signaling 3,4,5.

The journey of protein 3D structure analysis began with the resolution of myoglobin in 1958 6. Since then, the Protein Data Bank (PDB) has accumulated over 215,000 protein structures 7,8,9, with X-Ray crystallography accounting for approximately 84.2% and Nuclear Magnetic Resonance for about 6.5% of these structures 8. In addition to experimental determination of structures, simulation techniques that predict tertiary conformation from primary sequences have also developed, including innovative tools like AlphaFold and RoseTTAFold. These AI-driven approaches have shown exceptional accuracy in predicting protein structures, often for proteins without known structural analogs, effectively bridging significant gaps in structural biology 10,11. Further computational techniques, such as Molecular Dynamics and Quantum Mechanics methods can further refine the experimentally obtained or predicted structures under different simulated environmental conditions, enhancing the reliability and utility of the obtained data 12,13.

Nevertheless, even though understanding of protein structure has considerably advanced, it continues to face different aspects that significantly impact both research and practical applications in medical and biological sciences. Key challenges include the complexity of protein-protein interactions, where many proteins function through complex, often transient interactions with other proteins, making it difficult to decipher cellular processes 14,15. Intrinsically Disordered Proteins (IDPs) add another layer of complexity as they lack a fixed or stable three-dimensional structure under physiological conditions and represent a significant fraction of the proteome 16. Also, integration structural data with functional insights to fully understand protein function in is still a developing field of research 17. These are just some examples of factors that underscore the need for continued research in protein structure and the role it plays in different aspects of cell biology.

In this sense, one way to gain new insights 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 18,19 –containing protein secondary structure composition-- or BRENDA 20,21  –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 provide new valuable information about polypeptides. As an illustration, both CATH 22 and SCOP2 23 have used data from the Protein Data Bank to establish structural and evolutionary relationships between protein domains.

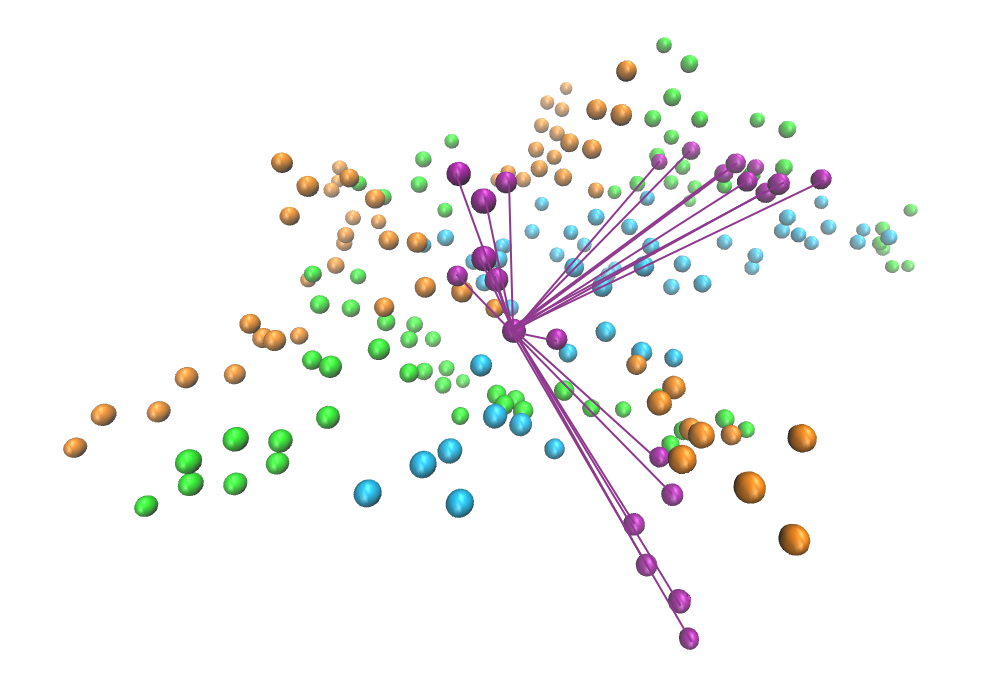
In this work, a specific chemical-geometrical parameter for polypeptides is presented. Referred as *Q*, it measures quantitatively the degree of spatial clustering of amino acids of the same chemical type in the 3D structure. Once defined, an analysis was conducted that included all X-Ray determined entries available in the Protein Data Bank. Remarkably, the *Q* value for a protein structure is highly predictable. In fact, it is only required to take into account the number of amino acids (*n*) that the polypeptide contains, demonstrating the existence of a *Q* related empirical relation that is highly conserved among protein structures. To further understand the origin of this almost constant property, novel stochastic structure simulations were designed that show that such law is consistent with the supra-organization of amino acids in spatial clusters that follow specific size and shape rules. Therefore, the size and shape of such clusters could conceivably be a highly conserved element in polypeptide structure.

**Methods**

*Q* 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, corresponding to the position of the α-carbon atom.

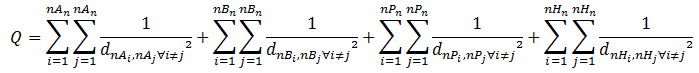


**Fig 1** Representation of the Euclidean distances of one residue with other residues of the same type. Every sphere here depicts the position of the α-carbon atom of a residue in an example protein, and is colored according to its type (acidic, basic, polar or hydrophobic). The lines represent all the Euclidean distances to be calculated for one residue. For the generation of these graphics, the VMD 1.9.3 software was employed 36 .

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 24, whilst Arg, His and Lys as basic 25. Polar residues include Ser, Thr, Asn and Gln 26. Finally, hydrophobic residues comprise Ala, Val, Ile, Leu, Met, Phe, Tyr and Trp 27. The rest of amino acids (Cys, Sec, Gly, Pro) are considered special not included in the analysis. Particularly, cysteine is unique due to its sulfhydryl (-SH) group, which is highly reactive, allowing cysteine residues to form disulfide bonds with each other 28. On the other hand, selenocysteine displays unique chemical properties due to the presence of selenium in its structure 29, whereas glycine is likely to play a unique role due to its flexible dihedral angle and compact size 30 and proline is unusual because its side chain forms a cyclic structure, connecting back to the amino group, which restricts the flexibility of the polypeptide backbone 31 .

Once the alpha-carbon positions and chemical nature for each residue have been established, the calculation of *Q* can proceed. The basic measurement involves the determination of Euclidean distances between residues of the same kind. That is, the minimum distance between the alpha-carbon of residues of the same chemical nature. Fig 1 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 results are added up, rendering *Q*. A formal definition can be seen in Equation 1.



**Equation 1.** The calculation of *Q* involves the computation of the euclidean distance (d) between the position of the alpha carbon of each residue (i) with the alpha carbon (j) of every residue of the same type. 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.

In essence, *Q* constitutes a measurement of spatial clustering of residues of the same chemical type. By measuring the inverse of squared distances, its value becomes higher for increased amino acid clustering. In the calculations, Bio.PDB 32,33 was used to extract the atom coordinates of each residue. An alternative to Equation 1 may also be defined where the addition of squared distances does not occur twice for each residue pair, however, for simplicity in this work Equation 1 was employed.

Stochastic simulations

The stochastic simulations in this work generate simulated structures that follow rules based on basic structural features of proteins. This is useful, as explained in the Results section, to check certain hypothesis on experimental structures. The algorithm for stochastic simulations involves the following steps:

a) Creation of an empty simulation box.For computational reasons it was chosen an 80x80x80 Å³ cubic space as simulation box.

b) Addition of amino acid clusters to the simulation box. This pursues the generation of an imaginary protein conformed by clusters of residues of the same type that coexist in the same protein.

The number of amino acids in each cluster along with the rule for their shape are the only input parameters in the simulations. In this sense, different clusters sizes were tested, ranging from 6 residues to up to 20. For simplicity, in each of the imaginary proteins generated all the clusters have the same size (same number of amino acids per cluster). As for rules for the shape of the clusters, each imaginary protein follows one of these two alternative implementations:

I) In the first implementation, all the residues in each cluster must be at a maximum of 3.8 Å (experimentally found maximum inter-residue distance) from at least another residue in the same cluster. This is named Shape I.

II) In the second implementation, all residues in each cluster must be at a maximum of 3.8 Å from at least two other residues in the same cluster. Logically, this second implementation renders more compact clusters than the first one. This is named Shape II.

For the formation of each amino acid cluster, one residue is added at a time at a random 3D position that is only validated if i) it is at least at a distance >= 2.4 Å from any other already existing amino acid in the structure (to mimic Van der Waals repulsion) and simultaneously ii) it is at a distance < 3.8 Å from at least one or two --depending on the shape implementation-- of other already existing amino acids in the cluster it belongs. Logically, the first residue in each cluster (and also the second for shape II) does not need to follow the latter rule. This iterative process of residue addition is repeated until the cluster achieves the desired size (number of residues) set as input. After the first cluster is completed, the second cluster is formed, and so on. The referred process of generating clusters and adding them to the structure continues until the simulation box reaches a realistic residue density. Such realistic residue density in the simulation box is obtained from a symmetric probabilistic distribution where maximum probability corresponds to the average residue density found for X-ray experimentally obtained structures in the PDB (1.172e⁻03 res Å⁻³) and minimum and maximum allowed values are the quartile Q1 and Q3 experimental values (8.080e⁻04 res Å⁻³ and 1.488e⁻03 res Å⁻³), respectively. The referred experimental residue density statistics were computed as the ratio between the number of residues and the volume of the minimum sphere that encompasses all residues in each X-ray determined PDB entry, normalized by the ratio between the volume of a cube and a sphere (Equation 2).

**Equation 2.** The term d indicates the residue density, while n denotes the number of residues and Vs refers to the volume of the minimum sphere that encompasses all residues in the 3D structure. The value 6/π represents the ratio between the volume of a cube of side length 2r and a sphere of radius r.

c) Creation of voids in the simulation box.At this point, a random percentage (chosen from a uniform distribution in the range [0%, 30%]) of the clusters in the simulation box are eliminated. This is done to create more potential variability in the shape in the imaginary protein to be generated.

d) Selection of an internal 3D subregion of the simulation box.In order to generate even more variability among the imaginary proteins, only the clusters present in an ellipsoid centered in the simulation box are selected. The three ellipsoid semi-axes are chosen each from a uniform distribution in the range [12Å, 60Å).

e) Validation of residue density. After steps c) and d) the selected subregion may not possess a realistic residue density, even though the residue density in the whole simulation box from step b) is realistic. Accordingly, here it is checked whether the chosen subregion still retains a realistic residue density, i.e. in the range (8.080e⁻04 res Å⁻³, 1.488e⁻03 res).

If this validation fails, step d) will be repeated until the chosen subregion possess adequate residue density.

f) Selection of clusters belonging to the hydrophobic core.A random fraction in the range [7%, 15%] of the clusters that are closest to the subregion geometric center are considered the hydrophobic core of the imaginary protein. Hence, these clusters are assigned the hydrophobic type. Importantly, after this it is checked that the number of hydrophobic residues (i.e., those present in the hydrophobic clusters) are in the realistic range of 41.077 +- 5% (both ends inclusive). This is done since the expected frequency of hydrophobic residues in a protein is 41.077% of the total, which is the average empirical percentage found in the X-ray crystallography-resolved proteins in the PDB. If the obtained percentage of residues in hydrophobic clusters is not in the referred realistic range, then steps d), e) and f) are executed over again.

g) Assignment of a type (polar, acidic, basic or special) to each of the non-hydrophobic clusters. The assignment of type to the non-hydrophobic clusters is selected randomly from a distribution with probability values also based on the empirical averages obtained in the PDB X-Ray determined entries. In particular, probability of acidic is 12.199%, probability of basic is 13.257%, probability of polar is 19.980% and probability of special is 13.487%.

After the completion of all the referred steps an imaginary protein will have been created, taking as input only the size of its clusters and the rule for their shape. The process relies on multiple random decisions, such as the 3D coordinates of the residues, the dimensions of the subregion ellipsoid or the probabilistic assignment of the type to each cluster. Nevertheless, the process also incorporates multiple checks to ensure that the stochastic structures obtained follow realistic patterns, such as acceptable inter-residue distances, overall realistic residue density and the existing of a hydrophobic core (all imaginary proteins are assumed to be water-soluble).

Calculation of protein compactness

Protein compactness refers to the arrangement, more or less extended, of the residues in a protein. 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

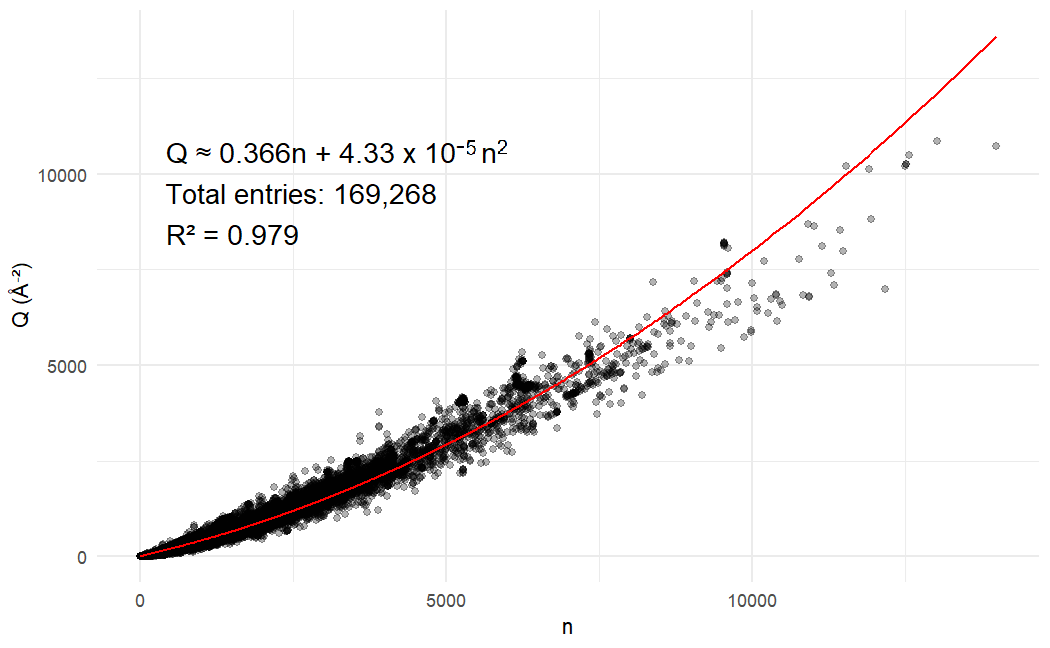
To compute surface areas of proteins, the software 3V 34 was used as a dependency. Probe radius was set to 1.5 A. Subsequently, the data were normalized by the number of residues.

**Results**

An empirical law relating Q and protein size

After the definition of Q (see Methods), the second step in this work involved calculating *Q* for 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 180,000 polypeptides and peptides. Of these, approximately 84.2% 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 an empirical equation for *Q* can also be constructed. In particular, *Q* values are shown to be dependent on the number of residues of proteins (*n*) following a specific linear model equation with very limited variance (R² = 0.978), indicating not a correlation between the two variables (as expected from Equation 1) but rather a distinct equation with limited variance that describes their relation (Fig 2).



**Fig 2** An empirical law highly accurately describes the relationship between the *Q* parameter and the number of residues (*n*). For the calculation, distances between residue alpha-carbon atoms are measured in Ångströms. Those entries where *Q*/*n* > 1.032 Å⁻²; which correspond to < 0.07% of the total, were discarded as outliers.

Delving into the cause of the empirical law

Given the highly structural diversity among proteins, it is surprising that an experimental law for *Q* can be drawn that only depends on the number of amino acids. Alternatively stated, even though according to Equation 1 it is expected that *Q* will *tend* to increase with increasing sizes of the polypeptide structure, the law tells us that any two proteins of equal number of amino acids will display almost the same *Q* value, no matter how different they are in terms of conformation. For example, one protein may be highly globular, and another not globular at all, but both of them shall display in essence the same *Q* value as long as they have the same number of residues.

In addition to the origin of the referred very limited variance found in the experimental law, another question arises: is it trivial that the law displays those particular model coefficients? And if not, why are those the particular coefficients appear in nature and not others?

In this work, we hypothesize that the main reason behind the highly conserved experimental law and its coefficients is the ubiquitous appearance of clusters of amino acids that follow specific size and shape rules, and that the same rules are shared among the great majority of proteins. As explained in detail later, this conclusion was extracted from a series of stochastic simulations that helped us test different controlled conditions.

In this manner, it is suggested here that the 3D structure of polypeptides would essentially be composed of a certain number of these amino acid clusters that follow the same morphology rules in almost all known protein structures. This would presumably be the cause of the empirical law found for *Q*,i.eitsequivalency with Equation 1. The referred clustering organization shared among proteins is designated here as the *Q*-mosaic model. This is compatible with the well-known existence of the hydrophobic core in water soluble proteins 36 . What is stated here is that, aside from the core, the rest of the amino acids also apparently group together in clusters according to their chemical type and that such clustering arrangement follows morphology patterns that are shared among the vast majority of proteins.

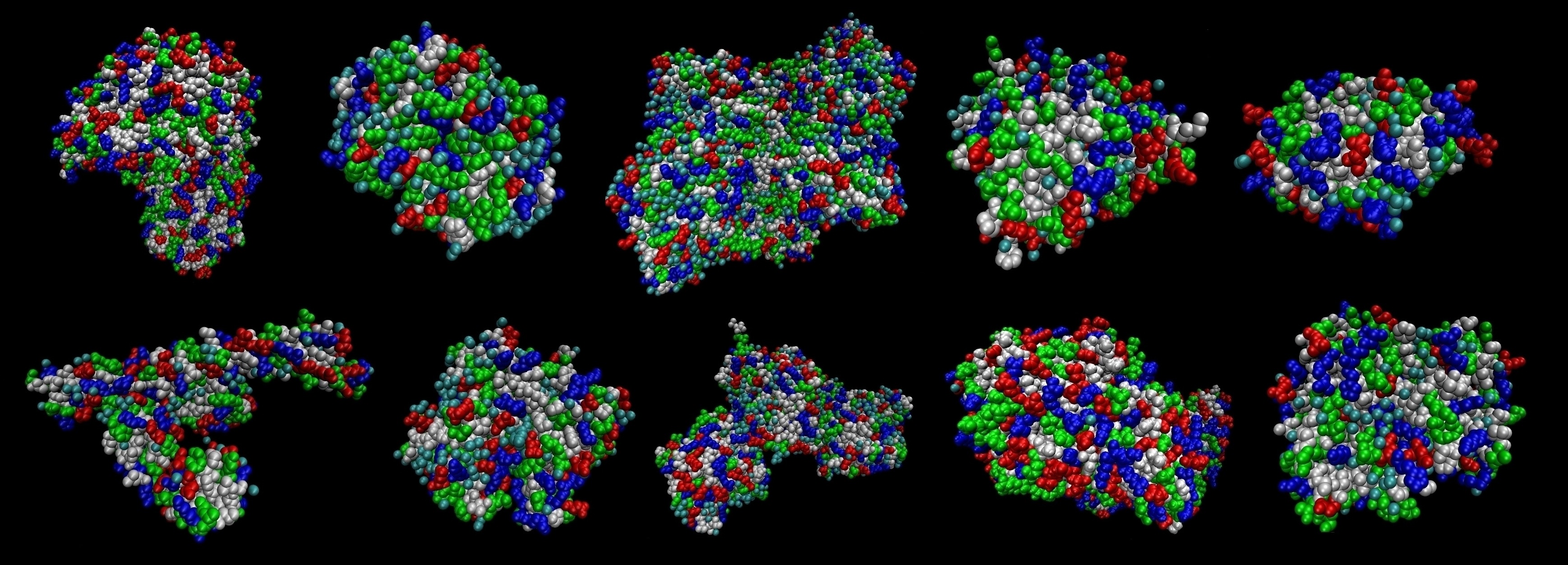
**Fig 3** Top: From left to right 3D structure of PDB entries 2SBL, 3CSJ, 3FG4, 4JG3, 4WDA. Bottom: From left to right 3D structure 5H07, 6JH8, 6ZA5, 7A0A, 8PQK. Every sphere depicts a residue and is colored according to its type (hydrophobic: white, polar: green, basic: blue, acidic: red, special: cyan). It can be seen that residues of the same type are grouped together in groups (clusters). For the generation of these graphics, the VMD 1.9.3 software was employed 37 .

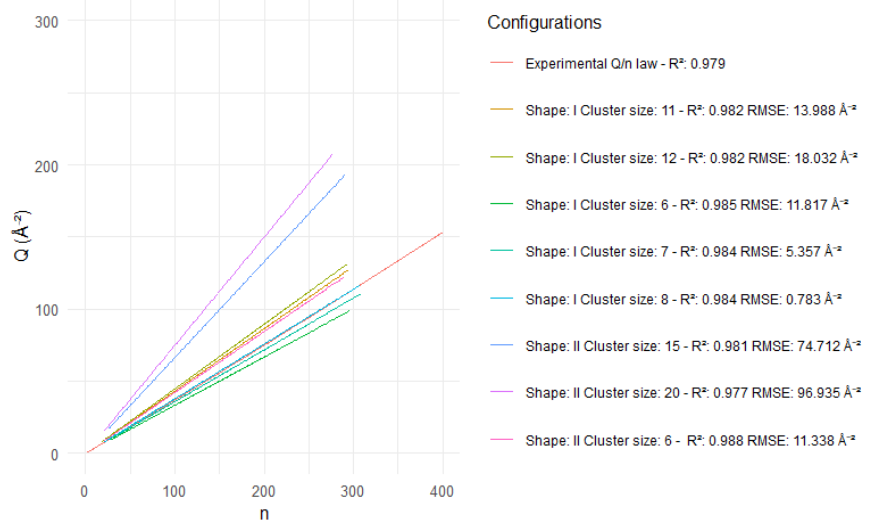
Fig 3 shows the visualization of ten real PDB structures using the VMD 1.9.3 software 36, where residues are colored according to its type. From these images, the referred cluster organization of amino acids seems real (at least in these structures), since residues of a same type are grouped together in clusters of similar size and shape, in an analogous fashion in all the structures. That is, aside from the core, there is no uniform mixture of amino acids according to their type, but rather residues of the same type are present in clusters of similar size and shape pattern in all structures. We encourage the reader to visualize any other PDB structures that they desire coloring residues according to their chemical type to check that the same clustering pattern applies. Apart from these simple visualizations examples, there is evidence in the scientific literature that suggests clustering of residues with similar characteristics does exist, aside from the known existence of hydrophobic/hydrophilic cores. Specifically, a study analyzed changes (differences) in spatially neighboring residues across different proteins, obtaining that hydrophobic residues tend to be interchanged preferably by hydrophobic residues, and charged/polar by other charged/polar residues, and that especially charged neighboring residues tend to be conserved 38. Also, another study was capable of designing an inter-residue distance map generator based on 30 physical-chemical properties of neighboring amino-acids 39. These analyses point that neighboring residues tend indeed to possess similar physical-chemical characteristics, which is consistent with the spatial organization of amino acids in clusters according to their chemical type (hydrophobic, polar, basic, acid) as it is proposed here. However, as explained before, in the present work this clustering pattern is quantitatively described by *Q*, showing the existence of a very well conserved quantitative law for its value. It is also proposed that the reason for that law is that the referred clusters are subject to the same morphology rules in most known polypeptide structures.

But, what exactly would be these specific clustering patterns (size and shape) existing in nature? In this regard, a possibility to discern the size and shape of clusters is to systematically analyze the available experimental structures. However, even performing an analysis on a large scale it would be highly challenging to detect a rule for cluster size and shape of the clusters, or even to determine if such rules exist at all. The reason is that what seems to form one cluster could actually be two or more clusters of the same type that happened to occur closely in the 3D structure. This could be especially the case in the hydrophobic core that most water-soluble proteins display internally, as it encompasses most hydrophobic residues (and in principle most hydrophobic clusters) in the structure. Also, it is hard to determine the real shape of the individual clusters from observation, since highly extended clusters may exist that encompass several of the apparent clusters upon initial observation. Alternatively stated, direct analysis on empirical data may not be suitable in principle for characterizing the existence of possible clusters rules that we propose may be the reason behind the empirical law.

To overcome these limitations an alternative approach based on simulation was followed. Instead of trying to extract a clustering model from the results, multiple models (cluster size and shape) were generated to test which configuration better fitted the results. More specifically, a series of novel stochastic simulations were designed that aimed to generate multiple artificial (imaginary) polypeptide structures. These imaginary proteins, although not really existing, follow a logical set of realistic rules based on basic structural features of proteins, so that conclusions extracted from them can in principle be regarded as useful for the purposes of our study. In particular, testing different cluster sizes and cluster shape rules, obtaining the *Q* values in each case, and comparing resemblance with the experimental data given by the empirical law for *Q*. The algorithm for the generation of each of the imaginary proteins in the simulations is detailed in the Methods section.

For each set of cluster size/shape input rules, multiple imaginary proteins were generated. The aim was to see if from these imaginary proteins a law analogous to the empirical equation could be extracted, and in that case what input values of size and shape rendered most resemblance to the empirical law.

Fig 4 shows that indeed the imaginary proteins formed under each configuration do exhibit, as experimental proteins, a predictable value of *Q* as a function of the number of residues (*n*). More specifically, for each configuration (pair size/shape) 10,000 imaginary proteins were generated, and a simple linear model (*n* vs *Q*) was fit for each case, rendering in all the configurations high R² values. Also, the slope of these linear models varied depending on the configuration, where higher cluster sizes tend to render higher slopes, and vice versa, lower cluster sizes resulted in lower slopes. At the same time, shape II related to higher slopes as compared to shape I. From the configurations tested, shape I seems a better candidate to accommodate the experimental law. Particularly, configuration Shape: I Cluster size: 8 seems to fit very well the empirical law for *Q*, at least in the range of protein sizes analyzed (from 21 to 308 residues) that the simulation box was capable to generate, which seems reasonable when inspecting the real PDB structures shown in Figure 3. It is important to note that even though the stochastic simulations presented here are simplified model and not entirely realistic, since for example it is assumed that all clusters have exactly the same size, they do show that the characteristic slope of the empirical law is not trivial. That is to say, depending on the cluster morphology (size and shape) rules the slope varies. In this manner, the *slope* (model coefficients) of the experimental law can be attributed to the fact that a specific cluster morphology rule is shared among polypeptides.

****

|  |  |
| --- | --- |
| **Configuration** | **RMSE (Å⁻²)** |
| Shape: I Cluster size: 8 | 0.783 |
| Shape: I Cluster size: 7 | 5.357 |
| Shape: II Cluster size: 6 | 11.338 |
| Shape: I Cluster size: 6 | 11.817 |
| Shape: I Cluster size: 11 | 13.988 |
| Shape: I Cluster size: 12 | 18.032 |
| Shape: II Cluster size: 15 | 74.712 |
| Shape: II Cluster size: 20 | 96.935 |

**Fig 4** Comparison between the empirical equation and analogous equations for each group of imaginary protein configurations. The number of entries analyzed in each configuration of imaginary proteins is 10,000. For each configuration, the value R² of the corresponding linear model is shown, along with the RMSE metric obtained by comparing the empirical law values and the predictions according to each model of imaginary proteins.

Assessing the potential role of hydrophobic/hydrophilic core in *Q* values

It could be argued that the existence of a hydrophobic core (for water-soluble proteins) or alternatively a hydrophilic core (for non water-soluble polypeptides) may explain the uniformity found for the empirical law, i.e that the property is direct consequence of the accumulation of residues of the same type in the inner core of polypeptides. However, the uniformity of the law also means that for any two given polypeptides of the same size, the value of *Q* will be essentially the same, regardless their 3D conformation. This means that if the first of these proteins is highly globular (compact) while the second is elongated they both will have the same *Q* value, although it is clear that in the first polypeptide the inner core will be larger than in the second. Nevertheless, to further a possible role of the core on the value of Q two shape parameters were tested, namely protein compactness and surface area, to gauge their impact on Q value.

For this purpose,

**Discussion**

The present work presents a potentially highly conserved amino acid 3D clustering trait in protein structure. Big data analysis shows that it can be described by a very specific mathematical equation with great reliability, which is surprising given the diversity of protein structures in terms of shape and organism of origin present in the PDB. Any two proteins, as long as they have the same number of residues, will display analogous *Q* value. The property cannot be attributed to the existence simply of the hydrophobic/hydrophilic core, since *Q* values are independent from protein compactness and surface area, so that in principle any two proteins with same size but very different shape will display the same Q quantity. Via novel stochastic simulations, it is demonstrated that shared control of cluster size and shape is sufficient to render this apparently conserved feature in polypeptides. Alternatively stated, the experimental law can be explained by a common set of morphology rules for clusters in the great majority of proteins, and not necessarily by more complex patterns involving the 3D disposition of clusters in respect to each other. Apart from the existence of a hydrophobic core in most crystallized proteins that contains the majority of hydrophobic clusters, the position of the rest of clusters needs not to be controlled in order to generate the expected *Q* value. The results also show that depending on the chosen size and shape rules the value of *Q* value will vary significantly, only matching those find in nature for specific configurations. Of the tested configurations, Shape: I Cluster size: 8 represents the one that closest resembles the one found in nature. Nevertheless, the main conclusion from this work is that most protein structures may display a highly conserved clustering morphology pattern, named here as the *Q*-mosaic model, as evidenced by the demonstrated impact of cluster configuration on *Q* by the simulations and the high predictive value (R²=0.978) of the model constructed from the experimental data (empirical law). Also, current protein visualization software allows to directly visualize examples of the *Q*-mosaic in multiple instances (PDB structures) picked at random, which is in support of the conclusions obtained via big data analysis and simulation described in this work. This apparently conserved trait in protein structure highlights the potentially relevant role of 3D clusters of the same type of amino acid in the biological properties of this group of biomolecules.

**Acknowledgements**

Fernando Govantes, Alejandro Cuetos, Francisco R. Villatoro, Brian Jiménez-García, Juan Neftalí Morillo García, Francisco Javier García Moscoso provided valuable technical support and advice.

**Financial disclosure statement**

No funding was received to assist with the preparation of this manuscript.

**Competing interests**

The authors have no competing interest to declare that are relevant to the content of this article.

**Data availability statement**

The results data and employed software is available at:

<https://github.com/UPO-Sevilla-Fco-Javier-Lobo-Cabrera/clustering_trait_proteins>

**Author contributions**

Francisco Javier Lobo-Cabrera: Conceptualization, Methodology, Formal Analysis, Software, Writing – Original Draft Preparation.

José Antonio Prado-Bassas: Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing.

**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. 2014;82:1-6. https://doi.org/10.1002/prot.24452.

3. Whitford D. Protein Structure & Function. Wiley; 2003.

4. Poulos T. Heme enzyme structure and function. Chem Rev. 2014;114:3919-62. https://doi.org/10.1021/cr400415k.

5. Grecco H, Schmick M, Bastiaens P. Signaling from the living plasma membrane. Cell. 2011;144:897-909. https://doi.org/10.1016/j.cell.2011.01.029.

6. Kendrew J, et al. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature. 1958;181:662-6. https://doi.org/10.1038/181662a0.

7. Berman H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235-42. https://doi.org/10.1093/nar/28.1.235.

8. RCSB PDB: Homepage. Rcsb.org. 2018. Available from: http://www.rcsb.org/.

9. Petrey D, Honig B. Protein structure prediction: inroads to biology. Mol Cell. 2005;20:811-9. https://doi.org/10.1016/j.molcel.2005.12.005.

10. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583-9. https://doi.org/10.1038/s41586-021-03819-2.

11. Baek M, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021;373(6557):871-6. https://doi.org/10.1126/science.abj8754.

12. Shen R, et al. Structural refinement of proteins by restrained molecular dynamics simulations with non-interacting molecular fragments. PLoS Comput Biol. 2015;11:e1004368. https://doi.org/10.1371/journal.pcbi.1004368.

13. Yu N, Yennawar H, Merz K. Refinement of protein crystal structures using energy restraints derived from linear-scaling quantum mechanics. Acta Crystallogr D Biol Crystallogr. 2005;61:322-32. https://doi.org/10.1107/S0907444904033669.

14. Perkins JR, et al. Transient protein-protein interactions: structural, functional, and network properties. Structure. 2010;18(10):1233-43. https://doi.org/10.1016/j.str.2010.08.007.

15. Ghadie MA, Xia Y. Are transient protein-protein interactions more dispensable? PLoS Comput Biol. 2022;18(4):e1010013. https://doi.org/10.1371/journal.pcbi.1010013.

16. Trivedi R, Nagarajaram HA. Intrinsically disordered proteins: an overview. Int J Mol Sci. 2022;23(22):14050. https://doi.org/10.3390/ijms232214050.

17. Wang B, et al. Integrating structural information to study the dynamics of protein-protein interactions in cells. Structure. 2018;26(10):1287-96. https://doi.org/10.1016/j.str.2018.07.010.

18. Joosten R, et al. A series of PDB-related databases for everyday needs. Nucleic Acids Res. 2010;39:D411-9. https://doi.org/10.1093/nar/gkq1105.

19. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22:2577-2637. https://doi.org/10.1002/bip.360221211.

20. Placzek S, et al. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Res. 2016;45:D380-8. https://doi.org/10.1093/nar/gkw385.

21. Enzyme Database - BRENDA. Brenda-enzymes.org. 2018. Available from: http://www.brenda-enzymes.org/.

22. Dawson N, et al. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2016;45:D289-95. https://doi.org/10.1093/nar/gkw1098.

23. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin A. SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res. 2014;42:D310-4. https://doi.org/10.1093/nar/gkt1242.

24. Agius C, Von Tucher S, Poppenberger B, Rozhon W. Quantification of glutamate and aspartate by ultra-high performance liquid chromatography. Molecules. 2018;23(6):1389. https://doi.org/10.3390/molecules23061389.

25. Li C, et al. Molecular selective binding of basic amino acids by a water-soluble pillar[5]arene. Chem Commun. 2013;49:1924-6. https://doi.org/10.1039/c3cc38622h.

26. Brinzei M, Stefaniu A, Iulian O, Ciocirlan O. Density functional theory (DFT) and thermodynamics calculations of amino acids with polar uncharged side chains. Chem Proc. 2021;3(1):56. https://doi.org/10.3390/ecsoc-24-08420.

27. Akbarbaglu Z, Tamjidi F, Sarabandi K, Ayaseh A. Physicochemical characteristics and antioxidant stability of spray-dried soy peptide fractions. Food Sci Nutr. 2023;7:3949-58. https://doi.org/10.1002/fsn3.3381.

28. Bak DW, Bechtel TJ, Falco JA, Weerapana E. Cysteine reactivity across the subcellular universe. Curr Opin Chem Biol. 2019;48:96-105. https://doi.org/10.1016/j.cbpa.2018.11.002.

29. Arnér ESJ. Selenoproteins—What unique properties can arise with selenocysteine in place of cysteine? Exp Cell Res. 2010;316(8):1296-303. https://doi.org/10.1016/j.yexcr.2010.02.032.

30. Imamoto Y, et al. Diverse roles of glycine residues conserved in photoactive yellow proteins. Biophys J. 2008;94(9):3620-8. https://doi.org/10.1529/biophysj.107.123414.

31. Ho BK, et al. The flexibility in the proline ring couples to the protein backbone. Protein Sci. 2005;14(4):1011-8. https://doi.org/10.1110/ps.041156905.

32. Cock P, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422-3. https://doi.org/10.1093/bioinformatics/btp163.

33. Hamelryck T, Manderick B. PDB file parser and structure class implemented in Python. Bioinformatics. 2003;19:2308-10. <https://doi.org/10.1093/bioinformatics/btg299>.

34. **Voss N, Gerstein M.** 3V: cavity, channel and cleft volume calculator and extractor. Nucleic Acids Res. 2010;38(Web Server issue):W555–W562. doi:10.1093/nar/gkq395.

35. Smyth M, Martin J. X-ray crystallography. Mol Pathol. 2000;53:8-14.

36. Rose GD, et al. Hydrophobicity of amino acid residues in globular proteins. Science. 1985;229(4716):834-8. https://doi.org/10.1126/science.4023714.

37. Humphrey W, et al. VMD - Visual Molecular Dynamics. J Mol Graph. 1996;14:33-8. https://doi.org/10.1016/0263-7855(96)00018-5.

38. Azarya-Sprinzak E, et al. Interchanges of spatially neighbouring residues in structurally conserved environments. Protein Eng Des Sel. 1997;10(10):1109-22. https://doi.org/10.1093/protein/10.10.1109.

39. Asencio-Cortés G, Aguilar-Ruiz A. Predicting protein distance maps according to physicochemical properties. J Integr Bioinform. 2011. https://doi.org/10.2390/biecoll-jib-2011-181.