In memory of Professor Marek Cieplak

REVIEW

The Role of Cavities in Biological Structures

Ł. Mioduszewski[a], K. Wołek[b] and M. Chwastyk[b,*]

[a] *Cardinal Stefan Wyszyński University, Dewajtis 5, 01-815 Warsaw, Poland*
[b] *Institute of Physics, Polish Academy of Sciences, al. Lotników 32/46, PL-02668 Warsaw, Poland*

Doi: 10.12693/APhysPolA.145.S51 *e-mail: chwastyk@ifpan.edu.pl

We investigate the significance of cavities within biological structures, ranging from single proteins to large complexes, such as viruses and even protein clusters composed of intrinsically disordered proteins. Utilizing our SPACEBALL algorithm, we detect empty spaces within these structures and quantify their volumes. This enables us to elucidate the impact of cavities on the properties of the given structures. Finally, we discuss how the presence of cavities in protein clusters facilitates the assessment of their hydration levels within a coarse-grained implicit solvent approach. Our discussion aims to demonstrate that the functions of various proteins originate from their specific tertiary structures containing cavities.

topics: cavities in biological structures, molecular dynamics simulations, viruses, simulations of gluten

1. Introduction

In the intricate world of biomolecular science, the diversity of shapes that proteins, their aggregates, and complexes can adopt is a captivating and fundamental phenomenon. These varied tertiary or quaternary structures play a pivotal role in determining the functional capabilities of biomolecules, as they dictate their ability to interact with other molecular entities [1]. Within this realm of structural exploration, a number of techniques are employed to unveil the hidden architectures of biomolecules, each offering a unique perspective. Notable among these methodologies are nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, and electron microscopy, which empower researchers to construct precise atomic models [2]. These models, once derived, serve as the foundation for theoretical analyses, shedding light on the detailed mechanisms of biomolecular systems.

Intriguingly, these investigative techniques occasionally unveil enigmatic and fascinating topological features within biomolecular structures. Such findings include the knotting of the protein's main chain [3, 4], the entanglement of two chains within multi-chain proteins [5–7], or the hidden cavities within a molecule's core [8, 9]. The most recent of these discoveries is widely discussed in the context of pathogenesis-related proteins of class 10 (PR-10), a category of plant proteins that has long been a source of scientific interest. Despite their conspicuous presence and high expression levels, PR-10 proteins continue to confound researchers by defying easy categorization of their functions. Beyond their purported roles, these enigmatic proteins have been found to participate in various biological processes, including the regulation of development and symbiotic interactions with other organisms [10]. Furthermore, they feature a hollow cavity within their molecular core, formed by a relatively short polypeptide chain comprising 154–163 residues. This cavity is surrounded by a seven-stranded antiparallel β-sheet, intersected by an elongated C-terminal α-helix, as illustrated in Fig. 1 (see also [11]). The figure presents the tertiary structure of the yellow lupine LIPR-10.2B protein obtained from the Protein Data Bank (PDB). This structure was extracted from the LIPR-10.2B/zeatin complex, which involves the plant hormone, *trans*-zeatin [12].

These structural elements are supported by a V-shaped framework formed by two additional helices, H1 and H2, as demonstrated in previous studies [13, 14] and depicted in Fig. 1. This distinctive folding pattern, commonly referred to as the PR-10 fold or Bet v1 fold, owes its nomenclature to the elucidated crystal structure.

To gain insights into the roles of these proteins, it becomes imperative to precisely determine the position of the cavity within a protein and describe its unique characteristics. This necessity served as the impetus for research initiated by Professor Marek Cieplak, resulting in the development of an algorithm and the establishment of the public server known as SPACEBALL [15, 16]. This innovative program facilitates the objective identification of cavity positions and the detailed description of their geometrical and chemical attributes. Notably,

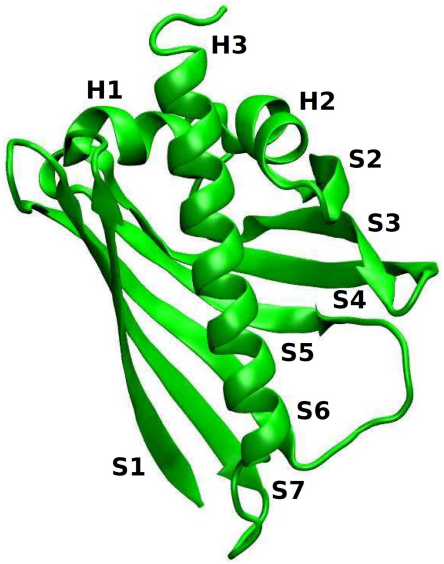

Fig. 1. The native structure of the LIPR-10.2B/zeatin PR-10 protein (PDB: 2QIM) [11]. The β-strands (S) and α-helices (H) are numbered consecutively from the N- to the C-terminus. The three zeatin molecules within the protein's hydrophobic cavity are omitted from the presentation as they were excluded from the analysis.

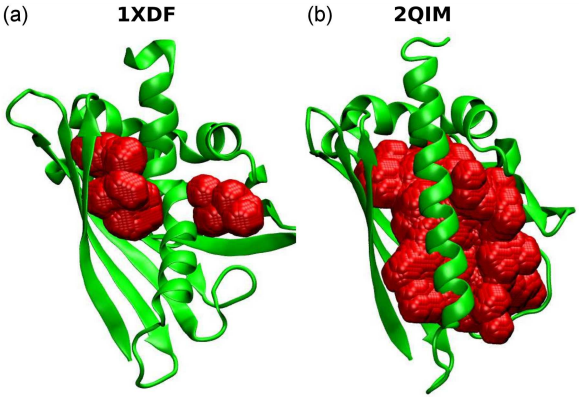

Fig. 2. The structures of two PR-10 protein representatives with PDB codes (a) 1XDF and (b) 2QIM. The protein structure is indicated in green, and the detected cavities are highlighted in red.

SPACEBALL not only enables the characterization of cavities within individual proteins, but also extends its utility to protein aggregates, such as gluten, and complex systems, like virus capsids.

2. Proteins with cavities

In 2013, the group led by Professor Cieplak calculated the volumes of cavities and described the surfaces of eighteen plant pathogenesis-related proteins of class 10. At that time, they characterized the cavities as large, given that the average calculated volume was 326 ± 162 Å^3. Three years later, an updated algorithm was published, and it yielded an average volume of 1309 ± 556 Å^3 for the same set of algorithm parameters. The new version of the algorithm accurately accounted for regions in the immediate proximity of the cavity walls. Since the surface of the cavity interior is often highly irregular, resulting in a large volume, the portion of the cavity volume near the cavity wall contributes significantly to its total volume [16]. Additionally, for calculations using the van der Waals radii proposed by Pauling instead of those proposed by Tsai et al. [17], the average volume was 1494 ± 609 Å^3. The largest cavity, with a volume of 2179 ± 16 Å^3, was detected in LIPR-10.2B/zeatin protein (PDB: 2QIM) [11], while the smallest, 273 ± 10 Å^3, in LIPR-10.2A protein (PDB: 1XDF) [18]. It is important to note that in the case of 1XDF, the protein's interior is composed of three smaller cavities, and the given value represents the volume of the largest one. Nevertheless, even when the volumes of all three are summed, this protein ranks at the lower end of the list of calculated volumes. The structures of these two proteins and the cavity positions are presented in Fig. 2.

The expression of PR-10 proteins increases after viral, bacterial, or fungal infection, as well as due to abiotic factors, such as cold, drought, oxidative stress, or UV radiation [19, 20]. Despite the wide range of factors impacting their expression, no unique function can be attributed to them [20], as mentioned in the introduction. These proteins exhibit considerable uniformity in their behavior, with notable disparities primarily observed in the internal cavity volumes and variations in the optimal folding time [15]. Despite these variations, they demonstrate mechanical robustness and display nearly identical structural rupture patterns when subjected to mechanical forces [15]. This suggests a high stability of the PR-10 fold. Interestingly, this stability is not immediately apparent, given the presence of a large cavity in their structures [15]. Consequently, it is suggested that this protein family may serve as versatile ligand binders, playing diverse roles in small-molecule signaling, transport, or storage. It is essential to highlight that, owing to variations in cavity volumes, shapes, topologies, and internal surface amino acid compositions, individual proteins within this family may offer distinct chemical environments for ligands. This suggests that different proteins possess the capability to host and potentially transport ligands with varying atomic compositions, in line with previous suggestions in the literature [12]. Furthermore, this suggests that such proteins can serve as selectors for ligands.

In 2020, another group led by Professor Cieplak extended the aforementioned analysis to calculate the volumes of cavities within each of the 24 280 single-chain protein structures from the CATH

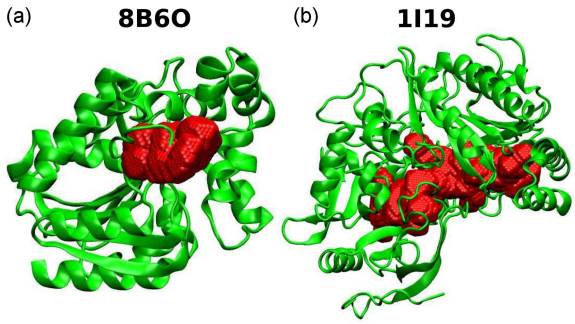

Fig. 3. (a) Tertiary structure (green) and cavity position (red) in haloalkane dehalogenase (PDB: 8B6O). (b) Tertiary structure (green) and cavity position (red) in cholesterol oxidase II (PDB: 1I19).

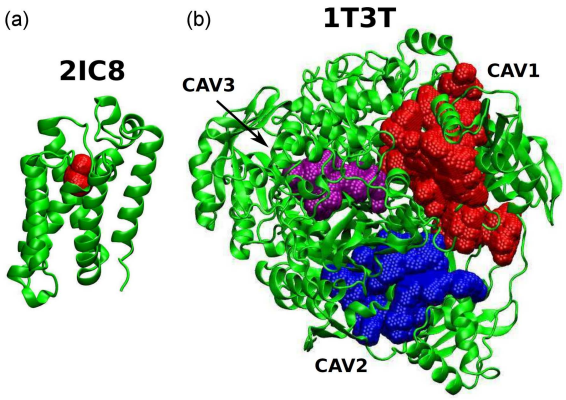

Fig. 4. (a) Structure (green) of the intramembrane protease GlpG from *Escherichia coli* (PDB: 2IC8) [25] with a single cavity (red) detected using the SPACEBALL algorithm. (b) Structure (green) of the StPurL protein (PDB: 1T3T) [26] with three cavities, labeled CAV1 (red), CAV2 (blue), and CAV3 (purple).

database [21]. Their findings demonstrated the existence of cavities with volumes of almost 40 nm^3 (PDB: 1KMP, 1KMP, 1PNZ) and a great number of smaller ones, showcasing the diverse range of cavity sizes within protein structures. It should be emphasized that the volume of the smallest considered cavity is of the order of 12 $Å^3$, which is sufficient to accommodate a single water molecule. This indicates that cavities initially considered to be large were found, upon examination of all available structures, to be relatively small compared to structures with much larger cavities. Moreover, very often the role of these large cavities is much better specified. Beyond ligand binding or small molecule transport, as mentioned in the case of PR-10 proteins, there are several other reasons for the presence of cavities within protein structures. Now, we will discuss the most interesting ones.

Some cavities act as active sites where enzymatic reactions take place. These folds provide a specific microenvironment for the binding and transformation of substrates. One example of structures with an active site buried in a cavity is haloalkane dehalogenases, where the active site is deeply embedded in the predominantly hydrophobic cavity at the interface of the α/β-hydrolase core domain and the helical cap domain [22]. The tertiary structure and the position of the cavity with a volume of 802 ± 34 $Å^3$ are presented in Fig. 3a. Another example of an active site deeply buried within a protein structure is cholesterol oxidase II. The active site in this case consists of a cavity (with a volume of 1977 ± 13 $Å^3$) bounded on one side by the β-pleated sheet in the substrate-binding domain and, on the opposite side, by the isoalloxazine ring of the flavin adenine dinucleotide cofactor covalently attached to the protein [23], as presented in Fig. 3b.

Moreover, cavities can play a role in regulating protein activity. Changes in cavity conformation may control the accessibility of substrates to the active site or modulate the protein's overall function. For example, the protease GlpG of *Escherichia coli* can be inactivated via the selective stabilization of the flexible C subdomain by cavity-filling mutations in this subdomain. On the other hand, cavity-creating mutations might enhance GlpG activity by providing even more flexibility [24]. The structure of GlpG with a cavity of volume 179 ± 19 $Å^3$, calculated using the SPACEBALL algorithm, is presented in Fig. 4a (see also [25]).

Cavities also play a crucial role in allosteric regulation, where binding at one site (allosteric site) influences the activity or conformation of another site in the protein. Surprisingly, the presence of empty spaces in the protein can trigger domain movements that facilitate the activation of the enzyme. In the investigation conducted by Tanwar et al. [26], focusing on the FGAR-AT protein derived from *Salmonella typhimurium* (StPurL), it was elucidated that this protein contains specific hydrophobic cavities that allow for breathing motions. The residues delineating these vacant regions establish a correlation network, interlinking them with the active centers, thereby constituting a functional communication conduit. Additionally, the protein's regions containing cavities, even if lacking a contiguous network with the active center, demonstrate inherent plasticity, rendering them capable of accommodating substantial structural perturbations, including those leading to direct steric conflicts with adjacent neighbors. Here, we show that the mentioned empty spaces are much larger and more extensive than previously detected [26]. Using the SPACEBALL algorithm, we identified three main cavities in the structure of the StPurL protein, as presented in Fig. 4b (see also [26]). Their volumes are: $V_{\mathrm{CAV1}} = 6674 \pm 249$ $Å^3$, $V_{\mathrm{CAV2}} = 3416 \pm 111$ $Å^3$, and $V_{\mathrm{CAV3}} = 1177 \pm 93$ $Å^3$. The positions of these cavities within the protein's structure are marked in red, blue, and purple, respectively.

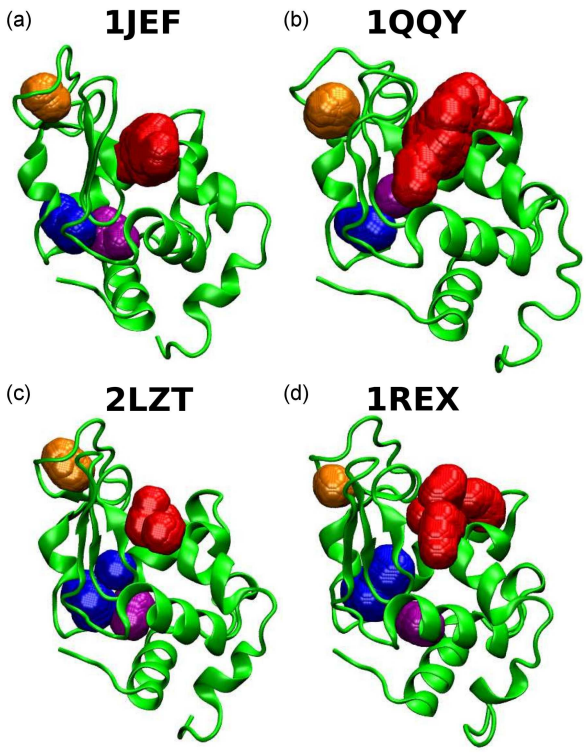

Fig. 5. Lysozyme structures from different biological species: turkey (PDB: 1JEF), dog (PDB: 1QQY), hen (PDB: 2LZT), and human (PDB: 1REX). Specific cavities, detected using the SPACEBALL algorithm, are distinguished by different colors. The largest cavity is marked in red, the smallest one in orange, the second-largest in purple, and the third-largest in blue.

Another intriguing example highlighting the impact of cavities on protein activity is found in the case of lysozymes. According to H. Li and Y.O. Kamatari [27], the location of cavities within structures from different biological species remains the same, despite variations in amino acid sequences. We further investigated whether the volumes of cavities at these specific locations are comparable. To do so, we utilized the SPACEBALL algorithm for structures from various biological species, including turkey (PDB: 1JEF) [28], dog (PDB: 1QQY) [29], hen (PDB: 2LZT) [30], and human (PDB: 1REX) [31]. The detected cavities align with the positions reported in the literature [27]. The volumes of these cavities, calculated using the SPACEBALL algorithm, are recorded in Table I; see also Fig. 5.

Based on the presented data, we observe that, despite a significant difference in the volume of the largest cavities, the volumes of the others, which are more precisely defined, are comparable. The substantial difference in the volumes of the largest cavities arises from their less precisely defined shapes, resembling pockets, whose volumes cannot be determined with high accuracy, yet their positions remain consistent. This observation underscores the

TABLE I

The volumes of cavities detected in lysozyme structures from different biological species. The first column specifies the particular species, the second column indicates its PDB code, and the third through sixth columns provide the volumes of specific cavities, as presented in Fig. 5.

Species	PDB	V_1 [Å^3]	V_2 [Å^3]	V_3 [Å^3]	V_4 [Å^3]
dog	1QQY	397 ± 42	51 ± 12	39 ± 11	29 ± 9
turkey	1JEF	194 ± 24	68 ± 11	49 ± 11	39 ± 12
human	1REX	154 ± 20	70 ± 10	40 ± 8	30 ± 5
hen	2LZT	87 ± 10	75 ± 9	67 ± 7	41 ± 8

potential significance of cavities in the functional attributes of lysozyme, supporting the perspective that these cavities play a pivotal role in the catalytic cycle of lysozymes. Their presence allows for a level of mobility within the active site, maintaining a constant volume available for water molecules. This arrangement is posited to contribute to the hydrolysis of substrate molecules. Furthermore, this outcome supports the notion that cavities are evolutionarily conserved elements essential for protein function [27].

The examples discussed so far illustrate the pivotal role of cavities in proteins, revealing their significance. Thus far, our focus has been on cavities within individual protein chains. Now, we turn our attention to a larger system — the protein complex known as the capsid, which serves as the protective protein coat of a virus.

3. Interior of viral capsid

A virus capsid is an assembly of proteins that shields viral genomes, possessing remarkable mechanical properties that have captured scientific interest. This fascination has led to extensive studies of various capsids to unveil their elastic behavior [32–34]. Computer simulations of nanoindentation experiments [35–38] have revealed that virus capsids, especially those protecting single-stranded RNA, exhibit significant elasticity. The study emphasizes that capsid sturdiness results from a combination of protein mechanical properties and interprotein binding [39]. Here, we explore whether the cavity within the capsid also affects the virus's stability.

Our theoretical research is based on molecular dynamics (MD) simulations, which can be conducted using either all-atom or coarse-grained models. While all-atom simulations offer valuable insights, they are limited by computation time. To address these challenges, we employed a coarse-grained molecular dynamics model to explore the mechanical response of a virus capsid, taking the

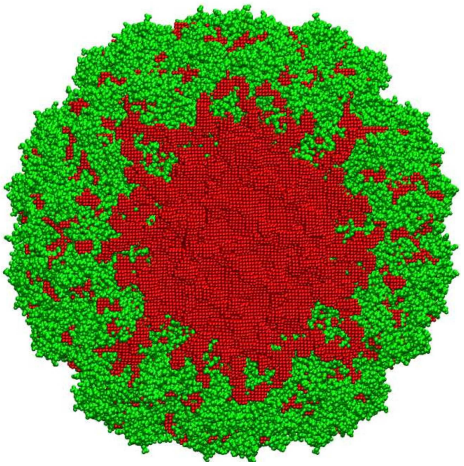

Fig. 6. The capsid of the CCMV virus in its basic native structure (PDB: 1CWP) [40]. The structure is highlighted in green, while the detected cavity is marked in red.

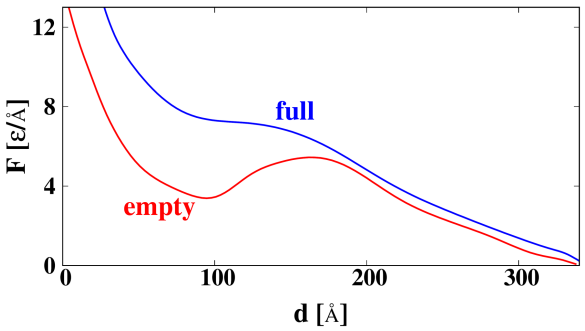

Fig. 7. The force F acting on the opposite walls of the simulation box during the squeezing of the virus capsid as a function of the distance d between the walls, considering both empty and full (with the inclusion of an RNA molecule) CCMV capsids.

cowpea chlorotic mottle virus (CCMV) [40] as an example. The structural analysis includes the identification of a cavity with a volume on the order of 5185 ± 2 nm^3, calculated using the SPACEBALL algorithm, as depicted in Fig. 6 (see also [40]).

The model used for the MD simulations has been developed by Professor Cieplak's group over many years. This is a Gō-like [41] model where each residue interacts with other residues via a pairwise Lennard–Jones potential, while residues in a single chain are connected by harmonic bonds. The model is based on representing each amino acid residue as a single pseudo-atom with an implicit solvent, and the temperature is controlled by a Langevin thermostat [34, 42]. The molecular dynamics within this approach is based on a contact map, i.e., a list of residues in contact, determined from the PDB structure through atomic overlaps [17, 43]. Finally, the native contacts between the Cα atoms i and j at distance r_{ij} are described by the Lennard–Jones potential

$$V(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \qquad (1)$$

where σ_{ij} is calculated as $\sigma_{ij} = 2^{-1/6}\, d_{ij}$ for each ij pair, so that the potential minimum coincides with the native distance d_{ij}, and the binding energy parameter ϵ is of the order 110 pN/Å. The interactions between residues not in the contact map are purely repulsive and are modeled using a truncated Lennard–Jones potential, with a cutoff at the minimum of 4 Å [41, 44–46]. The same criteria for atom dynamics were applied to describe interactions between protein chains forming the fully assembled virus capsid [39, 42].

Within the aforementioned model, we conducted a mechanostability analysis of the virus capsid through nanoindentation studies [39]. These studies involved measuring the force, F, acting on the opposite walls of the simulation box while squeezing the virus capsid as a function of the distance, d, between walls. We considered two structures of the capsid: the empty structure (PDB: 1CWP) and the same one but with the inclusion of an RNA molecule composed of 3171 bases. The RNA is modeled as a chain of beads connected by a harmonic potential with an equilibrium distance of 5.8 Å. RNA beads interact only via the excluded-volume effect, which imposes a repulsive interaction for other RNA beads within a distance closer than 8 Å or amino acids closer than 6 Å [39]. Both examined capsids exhibited a linear elastic response under small deformations, followed by a sudden force drop, signaling irreversible structural changes due to bond rupture within the capsid, as depicted in Fig. 7.

It must be noted that in our implicit solvent model, amino acids receive only random "kicks" from the thermostat and experience friction, but we do not consider the potentially stabilizing role of water within the capsid. However, since the capsid wall is semipermeable, allowing water to flow in and out of the capsid [47], we expect this effect to be negligible.

Figure 7 illustrates markedly different mechanical properties of the examined structures. Inter-protein bonds break much more easily in the empty capsid but are much more stable in the case of the one with RNA. This observation reveals another role of cavities within biological structures, i.e., they serve as activators of a cascade of inter-chain bond ruptures after the initial bond breaks, which can be considered the trigger of this process. Such a situation is not observed in structures filled with an RNA molecule that stabilizes the full structure. This may suggest that a larger ratio of empty space to the space occupied by the genetic material within the virus increases the opportunity for the virus to break and release the genetic material, but further research is needed to test this hypothesis [48].

4. Cavities in protein clusters

Cavities, as discussed thus far, have typically been described within proteins or their complexes with well-defined structures. Now, we will explore a slightly different system where cavities can emerge, namely, within intrinsically disordered protein (IDP) clusters. A notable example of such a system is storage proteins of grains, with gluten made from wheat [49] serving as an interesting case study due to its importance in the elasticity of dough in breadmaking [50]. Gluten can be categorized into two main fractions: shorter, water-soluble gliadins and longer, insoluble glutenins [49]. Glutenins are expected to contribute significantly to the elasticity of gluten [49, 50]. In contrast, storage proteins from maize and rice do not exhibit the same extraordinary elastic properties. Here, we present the results of our simulations involving various systems of interest: gluten, its gliadin and glutenin fractions, as well as proteins from rice and maize, revealing another function of empty regions within biological structures.

The research on gluten, similar to the virus capsid case, was conducted within a coarse-grained model. However, the model described in the previous section can be applied to simulate systems with well-defined structures. In this assembly, intrinsically disordered regions are included, making it challenging to use a classic Gō-like model [51]. To analyze this system, we modified the model so that the contact map is no longer based on the native structures from PDB. Instead, the contact map is constructed dynamically based on the geometry of the chain at any given moment. The geometrical criteria for a contact are derived statistically from a large database of contacts. The contacts are turned on and off quasi-adiabatically, reflecting changes in the chain conformation. The model was validated on a set of IDPs and partially ordered proteins [51], demonstrating its ability to simulate not only IDPs but also large clusters of IDPs and partially structured proteins as the Gō-like contact map can coexist with the dynamic one.

The viscoelastic properties of storage proteins depend on their hydration levels [52]. In a coarse-grained model with an implicit solvent [53], water molecules cannot be represented explicitly. Instead, their presence is visualized by cavities within the simulated system, large enough to accommodate at least one water molecule but small enough to remain part of a separate protein cluster. The analysis of the cavities in the simulated systems was conducted using our SPACEBALL algorithm [16], enabling the identification of the number and volume of the cavities.

Our simulations were conducted in several steps. Initially, the simulation box was compressed to achieve the desired protein concentration of 3.5 residues per cubic nanometer, as detailed in [52]. In the next step, the system was equilibrated in preparation for periodic box deformations. The simulation box was deformed by shearing, as illustrated in Fig. 8a, or by pulling, as shown in Fig. 8b. In both cases, the position of two opposing walls changed periodically, back and forth. The wall displacement as a function of time was a sinusoid with an amplitude of 1 nm and an oscillation period of 40 μs. The residues were attracted to the walls with the Lennard–Jones potential [54]. Following five full oscillation cycles, the system underwent the next equilibration. We also performed control simulations with no periodic deformation. Finally, the simulation box with a well-equilibrated system was stretched in one direction to induce the rupture of the protein network. The same pair of box walls continually attracted protein residues, causing them to adhere to these walls and enabling the stretching of the entire system [54].

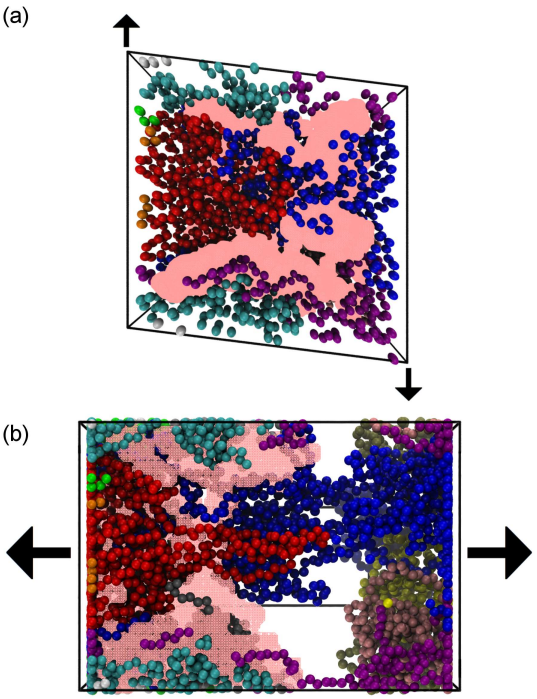

Fig. 8. Examples of gluten systems during shearing (a) and pulling (b) deformations. In both panels, the left and right walls attract residues, while all other walls have periodic boundary conditions. Residues are represented as balls, and cavities are shown in pink.

In our simulations, we observed a reduction in the average volume of the largest cavity during the stretching of all simulated storage proteins [53]. This indicates a decrease in the amount of solvent inside the protein cluster. The result is depicted in Fig. 9 (see also [53]) as the ratio of the volumes of the largest cavity, averaged over the second and first parts of the stretching trajectory. The observation is consistent with experimental findings [55] and may correspond to the stretching-induced release of water molecules [56].

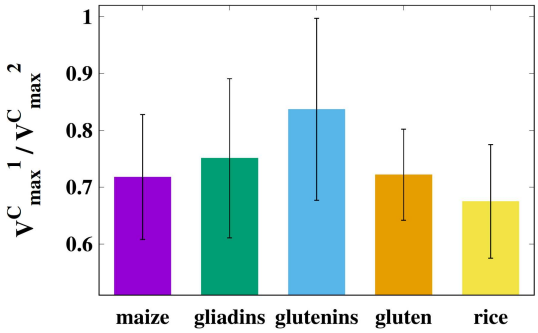

Fig. 9. The ratio of the largest cavity volumes, averaged over the second ($V^{C}_{max}1$) and first ($V^{C}_{max}2$) halves of the stretching trajectory for the five simulated storage protein systems. The data for this graph were obtained from [53]. Both halves of the stretching trajectory were simulated after the periodic mechanical deformation of the sample.

The high uncertainties in Fig. 9 result from volume fluctuations during stretching due to the dynamic nature of this process.

It is important to note that the final stretching is the last stage of the simulation, and the results presented in Fig. 9 only cover the first and second halves of that stage. Other figures display results from an earlier stage of the simulation, namely periodic deformation, where we periodically distorted the simulation box to investigate its effects on the system.

To explore this, we examined the total number of cavities in the systems resulting from either pull or shear periodic deformation. Figure 10a illustrates the ratio of the number of cavities in the systems after and before box deformation in the pulling mode. The decrease in the number of cavities suggests that the pulling mode facilitates the merging of smaller cavities into larger ones during the elongation step. On the other hand, in the shearing mode, as depicted in Fig. 10b, the number of cavities seems to slightly increase only for gluten. However, this change is still within the error bar, assessed at 20%, indicating that the shearing mode does not seem to lead to the merging of smaller cavities into larger ones.

This observation is confirmed by the results pertaining to the volume of the largest cavity. Its average value increases significantly in the pulling mode, while it remains relatively stable in the shearing mode, as presented in Fig. 11. On the other hand, when considering the total volume of cavities in the examined systems, a slight increase is observed in the pulling mode, as depicted in Fig. 12a. This increase is primarily due to the merging of small cavities, too small to contain a water molecule, into larger ones. Consequently, the average total cavity volume is higher. In contrast, the situation remains stable in the case of the shearing mode, as presented in Fig. 12b.

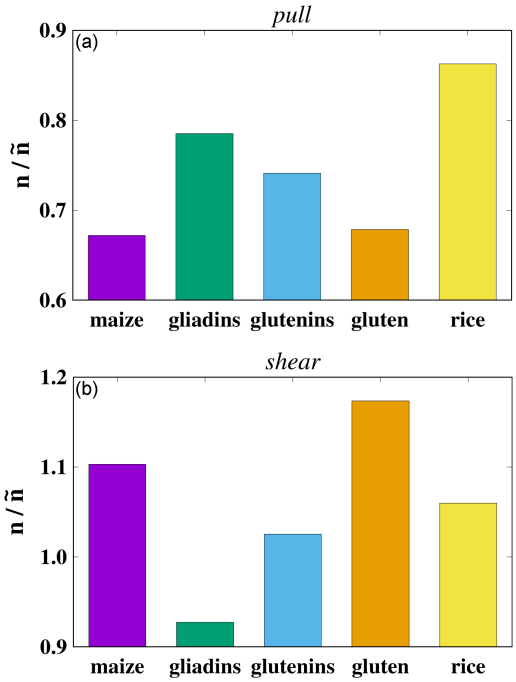

Fig. 10. The ratio of the number of cavities before ($\tilde{n}$) and after (n) periodic deformation for the 5 simulated systems in the pulling mode (a) and the shearing mode (b). The uncertainty is approximately ± 0.1 for pulling and ± 0.2 for shearing.

The presented results show more pronounced changes in cavity properties after periodic deformation in the pulling mode (normal stress) compared to the shearing mode. This aligns with the "loops and trains" theory [57], which predicts that after elongation, proteins form "trains" composed of parallel chains connected by hydrogen bonds, leading to the expulsion of water from the system. In the undeformed state, proteins form loops that are partially solvent-exposed and establish hydrogen bonds with water (all plant storage proteins contain high amounts of hydrophilic residues [49, 58, 59]). Closing these loops by elongating them in one direction acts as a kinetic trap, compelling the proteins to remain in the "train" state even when they are no longer deformed and no stress is applied. This phenomenon explains the lower number of cavities, the higher volume of the largest cavity (due to water expulsion), and minimal change in the total cavity volume — outcomes expected in an explicit solvent simulation but not necessarily in an implicit solvent one. In the shearing mode, "trains" are not formed, and the aforementioned process does not occur, resulting in much smaller changes.

While the differences between the studied systems were smaller than those arising from the deformation mode, it is noteworthy that the number of cavities before and after deformation changed the most in the gluten system (after pulling deformation, it was the lowest of all, and after shearing, it was the highest), making gluten much more

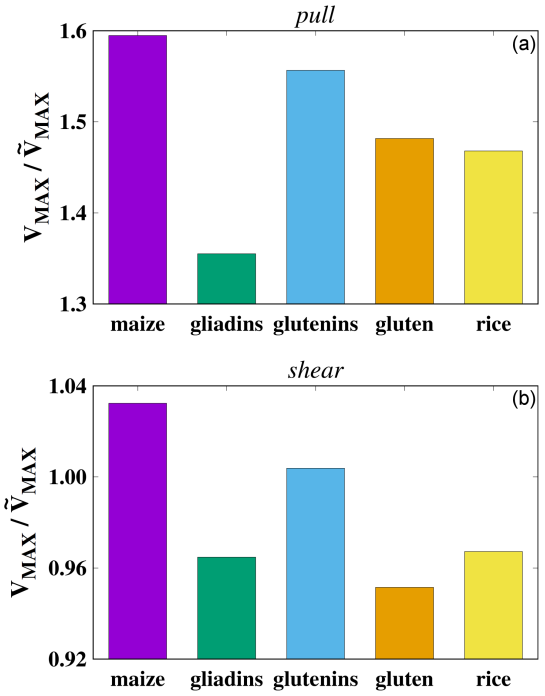

Fig. 11. The ratio of the volumes of the largest cavity after (V_{MAX}) and before ($\tilde{V}_{\mathrm{MAX}}$) periodic deformation for the five simulated storage protein systems. The error bars are approximately ± 0.12 for the pulling mode (a) and ± 0.07 for the shearing mode (b).

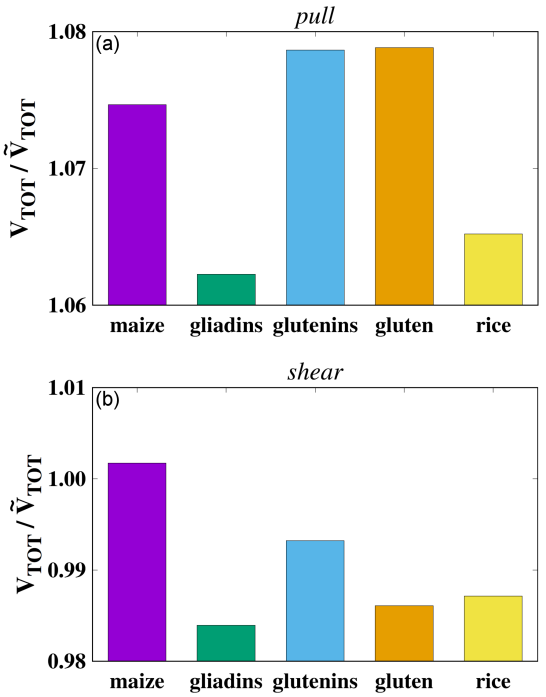

Fig. 12. The ratio of the total volumes of all cavities in the system after (V_{TOT}) and before ($\tilde{V}_{\mathrm{TOT}}$) periodic deformation for the five simulated storage protein systems. The error bars are approximately ± 0.11 for the pulling mode (a) and ± 0.18 for the shearing mode (b).

susceptible to changes than, for example, rice storage proteins. This susceptibility may stem from the fact that the gluten protein network solidifies after deformation, forming more hydrogen bonds [50, 57]. Cavities act as "holes" in that network, allowing adaptation to large deformations [57].

5. Conclusions

We have surveyed various structures containing cavities, detecting and calculating their volumes. With support from literature data, we described the functions of empty spaces within the considered structures. We concluded that, in the case of PR-10 proteins, their interior can serve as a space for ligands specific to a particular protein, consequently allowing such proteins to act as selectors for ligands. Using the examples of haloalkane dehalogenase and cholesterol oxidase II, we demonstrated that cavities provide an environment for enzymatic reactions. Through the example of GlpG, we discussed the role of cavities in regulating protein activity. Examining the StPurL protein, we illustrated that cavities are crucial in allosteric regulation. Finally, based on the analysis of lysozyme structures from different species, we asserted that cavities play a pivotal role in the catalytic cycle of lysozyme. Our calculations supported the notion that cavities are evolutionarily conserved elements in protein structure.

Analyzing the mechanostability and structure of the cowpea chlorotic mottle virus led us to the hypothesis that the large ratio of empty space to the space occupied by the genetic material within the virus increases the opportunity for the virus to break and release the genetic material.

For the simulated plant storage protein systems, we correlated changes in the size and number of cavities after periodic deformation with the "loops and trains" theory [57]. This enabled us to demonstrate that, even in an implicit solvent model, cavities can serve as a measure of the hydration level [52], and processes like solvent expulsion can be simulated using only geometric constraints [54]. The variations in the number of cavities also contributed to confirming the unique nature of the viscoelastic gluten protein network [53].

In all the discussed phenomena, cavities within protein systems play a pivotal role. The research initiated by Professor Cieplak in 2013 allowed us to uncover this role and may shed more light on many other systems in the future.

Acknowledgments

The authors acknowledge Professor Marek Cieplak for encouraging the investigation of structures containing cavities, knots, and various unstructured systems, explored using coarse-grained models.

This research has received support from the National Science Centre (NCN), Poland, under grant No. 2018/31/B/NZ1/00047 and the European H2020 FETOPEN-RIA-2019-01 grant PathoGelTrap No. 899616. The computer resources were supported by the PL-GRID infrastructure.

References

[1] J. Janin, R.P. Bahadur, P. Chakrabarti, *Q. Rev. Biophys.* **41**, 133 (2008).

[2] M. Schneider, X. Fu, A.E. Keating, *Proteins* **77**, 97 (2009).

[3] M. Jamroz, W. Niemyska, E.J. Rawdon, A. Stasiak, K.C. Millett, P. Sulkowski, J.I. Sulkowska, *Nucleic Acids Res.* **43**, D306 (2015).

[4] M. Chwastyk, M. Cieplak, *J. Phys. Chem. B* **124**, 11 (2020).

[5] P. Dabrowski-Tumanski, A.I. Jarmolinska, W. Niemyska, E.J. Rawdon, K.C. Millett, J.I. Sulkowska, *Nucleic Acids Res.* **45**, D243 (2017).

[6] M. Kroger, J.D. Dietz, R.S. Hoy, C. Luap, *Comput. Phys. Commun.* **283**, 108567 (2023).

[7] Y.I. Zhao, M. Chwastyk, M. Cieplak, *J. Chem. Phys.* **146**, 225102 (2017).

[8] M. Chwastyk, E.A. Panek, J. Malinowski, M. Jaskólski, M. Cieplak, *Front. Mol. Biosci.* **7**, (2020).

[9] M. Chwastyk, M. Cieplak, *Front. Mol. Biosci.* **8**, 692230 (2021).

[10] H. Fernandes, K. Michalska, M. Sikorski, M. Jaskolski, *FEBS J.* **280**, 1169 (2013).

[11] H. Fernandes, O. Pasternak, G. Bujacz, A. Bujacz, M.M. Sikorski, M. Jaskolski, *J. Mol. Biol.* **378**, 1040 (2008).

[12] H. Fernandes, A. Bujacz, G. Bujacz, F. Jelen, M. Jasinski, P. Kachlicki, J. Otlewski, M. Sikorski, M. Jaskolski, *FEBS J.* **276**, 1596 (2009).

[13] M. Gajhede, P. Osmark, F.M. Poulsen, H. Ipsen, J.N. Larsen, R.J.J. van Neerven, C. Schou, H. Lowenstein, M.D. Spangfort, *Nat. Struct. Biol.* **3**, 1040 (1996).

[14] J. Biesiadka, G. Bujacz, M.M. Sikorski, M. Jaskolski, *J. Mol. Biol.* **319**, 1223 (2002).

[15] M. Chwastyk, M. Jaskolski, M. Cieplak, *FEBS J.* **281**, 416 (2014).

[16] M. Chwastyk, M. Jaskolski, M. Cieplak, *Proteins* **84**, 1275 (2016).

[17] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, *J. Mol. Biol.* **290**, 253 (1999).

[18] O. Pasternak, J. Biesiadka, R. Dolot, L. Handschuh, G. Bujacz, M.M. Sikorski, M. Jaskolski, *Acta Cryst.* **D61**, 99 (2005).

[19] S. Führer, J. Unterhauser, R. Zeindl, R. Eidelpes, M.L. Fernández-Quintero, K.R. Liedl, M. Tollinger, *Int. J. Mol. Sci.* **23**, 8252 (2022).

[20] H. Fernandes, K. Michalska, M. Sikorski, M. Jaskolski, *FEBS J.* **280**, 1169 (2013).

[21] N.L. Dawson, T.E. Lewis, S. Das, J.G. Lees, D. Lee, P. Ashford, C.A. Orengo, I. Sillitoe, *Nucl. Acids Res.* **45**, D289 (2016).

[22] K.H.G. Verschueren, F. Seljee, H.J. Rozeboom, K.H. Kalk, B.W. Dijkstra, *Nature* **363**, 693 (1993).

[23] R. Coulombe, K.Q. Yue, S. Ghisla, A. Vrielink, *J. Biol. Chem.* **276**, 30435 (2001).

[24] R. Guoa, Z. Cangb, J. Yaoa, M. Kima, E. Deansc, G. Weib, S. Kangd, H. Hong, *PNAS* **117**, 22146 (2020).

[25] Y. Wang, Y. Zhang, Y. Ha, *Nature* **444**, 179 (2006).

[26] A.S. Tanwar, V.D. Goyal, D. Choudhary, S. Panjikar, R. Anand, *PLoS ONE* **8** e77781 (2013).

[27] H. Li, Y.O. Kamatari, *Subcell Biochem.* **72**, 237 (2015).

[28] K. Harata, M. Muraki, *Acta Cryst.* **D53**, 650 (1997).

[29] T. Koshiba, M. Yao, Y. Kobashigawa, M. Demura, A. Nakagawa, I. Tanaka, K. Kuwajima, K. Nitta, *Biochemistry* **39**, 3248 (2000).

[30] M. Ramanadham, L.C. Sieker, L.H. Jensen, *Acta Cryst.* **B46**, 63 (1990).

[31] M. Muraki, K. Harata, N. Sugita, K. Sato, *Biochemistry* **35**, 13562 (1996).

[32] B. Kiss, D. Mudra, G. Török, Z. Martonfalvi, G. Csik, L. Herenyi, M. Kellermayer, *Biophys. Rev.* **12**, 1141 (2020).

[33] A. Zolochevsky, S. Parkhomenko, A. Martynenko, Quantum, molecular and continuum modeling in nonlinear mechanics of viruses. *V.N. Karazin Kharkiv Nat. Univ. Ser. Med.* **44**, (2022).

[34] M. Cieplak, M. Robbins, *J. Chem. Phys.* **132**, 015101 (2010).

[35] M. Cieplak, in: *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*, Ed. A. Liwo, Springer Series on Bio- and Neurosystems, vol 8. Springer, Cham 2019.

[36] M. Bravo, L.J. Gomez, J. Hernandez-Rojas, *Soft Matter* **16**, 3443 (2020).

[37] M. Medrano, A. Valbuena, A. Huete, M. Mateu, *Nanoscale* **11**, 9369 (2019).

[38] M. Aznar, S. Roca-Bonet, D. Reguera, *J. Phys. Condens. Matter* **30**, 264001 (2018).

[39] K. Wolek, M. Cieplak, *J. Condens. Matter Phys.* **29**, 474003 (2017).

[40] R. Konecny, J. Trylska, F. Tama, D. Zhang, N. Baker, C. Brooks III, A. McCammon, *Biopolymers* **82**, 106 (2006).

[41] M. Cieplak, T.X. Hoang, M.O. Robbins, *Proteins* **49**, 114 (2002).

[42] M. Cieplak, M.O. Robbins, *PLOS ONE* **8**, e63640 (2013).

[43] M. Chwastyk, A.P. Bernaola, M. Cieplak, *Phys. Biol.* **12**, 046002 (2015).

[44] M. Chwastyk, A. Galera-Prat, M. Sikora, Á. Gómez-Sicilia, M. Carrión-Vázquez, M. Cieplak, *Proteins* **82**, 717 (2014).

[45] M. Gunnoo, P.-A. Cazade, A. Orlowski, M. Chwastyk, H. Liu, D.T. Ta, M. Cieplak, M. Nashde, D. Thompson, *Phys. Chem. Chem. Phys.* **20**, 22674 (2018).

[46] Y.N. Zhao, M. Chwastyk, M. Cieplak, *Sci. Rep.* **7**, 39851 (2017).

[47] E. Tarasova, I. Korotkin, V. Farafonov, S. Karabasov, D. Nerukh, *J. Mol. Liq.* **245**, 109 (2017).

[48] H.V. Chaudhari, M.M. Inamdar, K. Kondabagil, *iScience* **24**, 102452 (2021).

[49] H. Wieser, Food Microbiol. **24**, 115 (2007).

[50] D.N. Abang Zaidel, N.L. Chin, Y.A. Yusof, *J. Appl. Sci.* **10**, 2478 (2010).

[51] Ł. Mioduszewski, M. Cieplak, *Phys. Chem. Chem. Phys.* **20**, 19057 (2018).

[52] Ł. Mioduszewski, *Eur. Biophys. J.* **52**, 583 (2023).

[53] Ł. Mioduszewski, M. Cieplak, *PLOS Comp. Biol.* **17**, e1008840 (2021).

[54] Ł. Mioduszewski, M. Cieplak, *Tribol. Lett.* **69**, 60 (2021).

[55] P. Belton, I. Colquhoun, A. Grant, N. Wellner, J. Field, P. Shewry, A. Tatham, *Int. J. Biol. Macromol.* **17**, 74 (1995).

[56] S. Liese, M. Gensler, S. Krysiak, R. Schwarzl, A. Achazi, B. Paulus, T. Hugel, J.P. Rabe, R.R. Netz, *ACS Nano.* **11**, 702 (2017).

[57] H. Singh, F. MacRitchie, *J. Cereal Sci.* **33**, 231 (2001).

[58] Y. Wu, J. Messing, *Front. Plant Sci.* **5**, 240 (2014).

[59] P. Chen, Z. Shen, L. Ming et al., *Front. Plant Sci.* **9**, 612 (2018).