Turkish Journal of Chemistry

http://journals.tubitak.gov.tr/chem/

Research Article

Turk J Chem
(2019) 43: 394 – 403

doi:10.3906/kim-1808-1

Why protein conformers in molecular dynamics simulations differ from their crystal structures: a thermodynamic insight

Filippo PULLARA[1,*], **Mao WENZHI**[1], **Mert GÜR**[1,2,*,**]

[1]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA

[2]Department of Mechanical Engineering, Faculty of Mechanical Engineering, İstanbul Technical University, İstanbul, Turkey

Received: 01.08.2018 • **Accepted/Published Online:** 14.01.2019 • **Final Version:** 03.04.2019

Abstract: Conformers generally deviate structurally from their starting X-ray crystal structures early in molecular dynamics (MD) simulations. Studies have recognized such structural differences and attempted to provide an explanation for and justify the necessity of MD equilibrations. However, a detailed explanation based on fundamental physics and validation on a large ensemble of protein structures is still missing. Here we provide the first thermodynamic insights into the radically different thermodynamic conditions of crystallization solutions and conventional MD simulations. Crystallization solution conditions can lead to nonphysiologically high ion concentrations, low temperatures, and crystal packing with strong specific protein–protein interactions, not present under physiological conditions. These differences affect protein conformations and functions, and MD structures equilibrated or simulated under physiological conditions are usually expected to differ from their X-ray structures at a local scale, while the global fold is usually maintained. To quantify this property, we performed conventional MD simulations for over 70 different proteins spanning a broad range of molecular size and structural and functional families. Our analysis shows that crystal structures are good starting points; however, they do not represent structures in their physiological environment. This fact has to be taken into consideration when computational methods dependent on atomic coordinates, such as substrate/ligand docking, are used to guide experimental analyses.

Key words: X-ray, molecular dynamics, thermodynamics, thermodynamic conditions, crystal contacts, structure

1. Introduction

Protein conformations are affected by several factors. In fact, the same amino acid linear sequence can, in principle, fold into different tertiary structures. Among these tertiary structures of a protein, the native structure is the protein spatial organization in its functional condition. Environment conditions that allow the protein to stay in its native state are called physiological conditions and the native state corresponds, by definition, to the free energy minimum of the system under those given thermodynamic conditions (e.g., temperature, pressure, solutes, type of solvent and solvent conditions,[1−4] and protein concentrations). The energy landscape of a protein is dynamic, such that its shape and the probabilities of the substate populations are dynamically influenced by several factors.[1−5] It has been proven that even small changes in thermodynamic conditions, such as variation of just a few degrees in temperature and/or an increase in salt concentrations,[5] or different solution

* These two authors contributed equally to this work
*Correspondence: gurme@itu.edu.tr

thermodynamic conditions can lead to different native protein structures, not to mention the structural changes that are triggered upon ligand binding.[5,6] or during allosteric interactions.

The relevance of solution conditions is clearly shown when a phase diagram for protein solution is plotted. In the schematic phase diagram shown in Figure 1, we display in black the spinodal line for the model system. Spinodal lines separate the phase diagrams into two regions of thermodynamic stability of the protein solution. In the schematic phase diagram displayed, under the thermodynamic conditions of the region lying above the spinodal line the free energy minimum of the system (protein solution) corresponds to a single phase, homogeneous protein solution. Conversely, under thermodynamic conditions corresponding to a point below the spinodal line, the free energy minimum is a demixed solution with regions of high protein concentration and low protein concentration elsewhere. This phenomenon, usually known as liquid–liquid demixing (LLD), occurs for a variety of systems including hard spheres or polymers.[7]

The spinodal represents the boundary of phase transition. On approaching the spinodal from the stability region of the phase diagram (above), the system experiences spontaneous fluctuations in the local concentration of protein with anomalous amplitudes and lifetimes. The amplitude and lifetime of such anomalous fluctuations are governed by only one variable, ε, which is defined as follows:

$$\varepsilon = \frac{T - T_S}{T_S}, \tag{1}$$

where T is the system temperature and T_S is the corresponding spinodal temperature. In other words, ε is a normalized distance in temperature from the instability region of the phase diagram.[8,9] A simple mathematical derivation of the laws governing anomalous fluctuation in the case of protein crystallization is given in Pullara et al.[8] Because of the fact that functional proteins are stable in solution, protein physiological conditions have to lie in the regions of the phase diagram where the homogeneous protein solution condition is thermodynamically stable. This case corresponds in Figure 1 to the green cloud in the upper-left region, not too close to the thermodynamic instability region. Under those conditions, the free energy minimum corresponds to the native structure of the protein. In contrast, crystallization conditions usually lie much closer to the spinodal (or instability region of the phase diagram), as schematically shown by the red cloud in Figure 1. Those thermodynamic conditions drive the precipitation of proteins, otherwise stable in solution, to eventually form crystals/fibers/aggregates.

Before precipitation into crystalline forms, proteins generally undergo a broad range of conformational rearrangements (from minor up to very relevant ones) and those initial variations in the protein structure are required for facilitating the formation of crystal nuclei and their subsequent growth. In addition, in protein crystals, the occurrence of the so-called "crystal contacts" involves strong protein–protein interaction fields that restrict the motions of the parts of the proteins involved in crystal contacts. These constraints may prevent proteins from exploring the full functional conformational space that is otherwise accessible under physiological conditions. From the native structure of proteins to their structure in crystals two different types of conformational changes occur: (i) the first is thermodynamically driven, renders proteins unstable in solution, and is suitable for crystal nuclei formation; (ii) the second is local and facilitated by intermolecular crystal contacts. These are physical contacts between proteins in the crystal arrangement, which induce constraints, if not alterations in interfacial conformations. Such conformational changes depend on the very local chemical and structural features of the proteins where the crystal contacts take place.

In order to quantify and investigate to what extent the above-discussed formation of crystal contacts and

nonphysiological conditions during crystallization process alter the protein structure, we performed molecular dynamics (MD) simulations under physiological conditions for a large set of 70 randomly selected different protein crystal structures. After 10 ns of MD simulations the overall fold of the proteins was conserved, but a significantly large average root-mean-square deviation (RMSD) of ~ 2 Å from their crystal structures was observed.

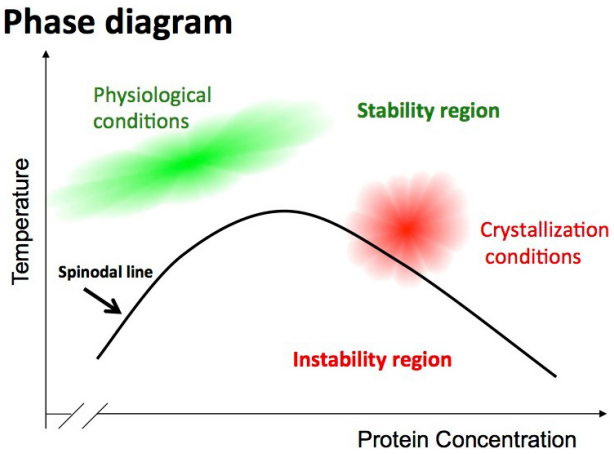

Figure 1. Schematic phase diagram for protein in solution as a function of temperature and protein concentration. The black line represents the spinodal line for the system, which is the boundary for a phase transition. The region above the spinodal line is the thermodynamically stable region for homogeneous protein solution (here only one phase is present for the sake of simplicity); below the spinodal line is the instability region for protein solution. In this region we observe liquid–liquid demixing and formation of subregions that have higher protein concentrations than the remaining; two different phases coexist. The phase transition illustrated in the figure is called "entropy driven".[10] Green and red clouds schematically represent physiological and crystallization conditions, respectively.

2. Results and discussion

Structures in the Protein Data Bank (PDB)[11] come mainly from X-ray crystallographic studies (88.3%, see http://www.rcsb.org for details) resolved with protein crystallization procedures under extreme environmental conditions that are likely to affect structure on a local scale at least, as discussed in the introduction section. Those extreme conditions generally involve salt concentrations often above a few molar (physiological conditions, on the other hand, are ~ 0.15 M), and protein concentrations two or more times above the physiological condition, the presence of cosolutes, low temperatures, cryoprotectant like PEG or glycerol, and sometimes cocrystallization with other proteins/ligands. Protein structures are determined after being immobilized in these extreme thermodynamic conditions and, as a matter of fact, it is a common observation in structures resolved under different conditions to observe different B-factors in residue positions. It should be noted that different crystallization conditions may also lead to different crystal packing and, hence, different crystal contacts.

Two such examples are shown in Figure 2: (i) lysozyme from *Gallus gallus* of ~ 16 kD (left panel, A) and (ii) fatty-acid amide hydrolase (FAAH) 1 from *Rattus norvegicus* of ~ 63 kD (FAAH, right panel, B). Lysozyme structures having PDB IDs 3A3R, 3ULR, and 3A6B were selected for comparison of their B-factors. For 3A3R, crystallization conditions were 10%–15% NaCl (w/v) in sodium-acetate buffer (pH 4.7) at 20 °C. For 3ULR protein crystals were grown in 1.2 M sodium citrate at pH 6.0. For PDB ID 3A6B, on the other hand, crystals were grown in 0.1 M HEPES buffer pH 7.6–7.8, 9%–11% w/v polyethylene glycol (PEG) 6000, and 7%–9% w/v 2-methyl-2,4-pentanediol. Space groups for 3A3R, 3ULR, and 3A6B are P 4_3 2_1 2, P 2_1 2_1 2_1, and P 4_1 2_1 2,

respectively. As can be seen, these three structures were crystallized under significantly different pH values and different temperatures. The B-factors (Figure 2A) obtained for these three different crystallization conditions differ in both magnitude and number/locations of the peaks. Among the 3ULR's five highest peaks only two are also shared by 3A6B and 3A3R. Both 3A6B and 3A3R show additional peaks not present in 3ULR. For FAAH, structures having PDB IDs 4HBP and 3QJ8 were selected for comparison. 4HBP crystals were grown with vapor diffusion in drops of 50 nL of mother liquor (ML) and 50 nL of protein solution. Compositions are 40% PEG 400 and 0.1 M sodium acetate buffer pH 4.5 for ML and 20 mM HEPES (pH 7.5), 150 mM NaCl, 1 mM EDTA, 0.05% LDAO, 5 mM DTT, and 10% glycerol for protein solution. 3QJ8 crystals, on the other hand, were grown in 20 mM HEPES buffer pH 7.9, 1 mM EDTA, 200 mM NaCl, 1 mM DTT, 10% glycerol, and 0.1% LDAO. Space groups are P 3_2 2 1 and P 2_1 2_1 2_1 for 4HBP and 3QJ8, respectively. The two crystallization buffers differ significantly: 40% PEG 400, 0.1 M sodium acetate pH 4.5 for the first and 20 mM HEPES, pH 7.9, 1 mM EDTA, 200 mM NaCl, 1 mM DTT, 10% glycerol, and 0.1% lauryldimethylamine oxide for the second. The crystals for the 4HBP were harvested after 7 days and those for the 3QJ8 were harvested after just 3 days after an exchange of buffer to 0.1 M 2-(N-morpholino) ethanesulfonic acid, pH 5.5, 4%–14% PEG 3350, and 50 mM ammonium fluoride. As can be seen in Figure 2B, the B-factor curve of 4HBP has a much higher number of peaks than 3QJ8. The differences in the features of those two curves are amplified in the region of 300–380.

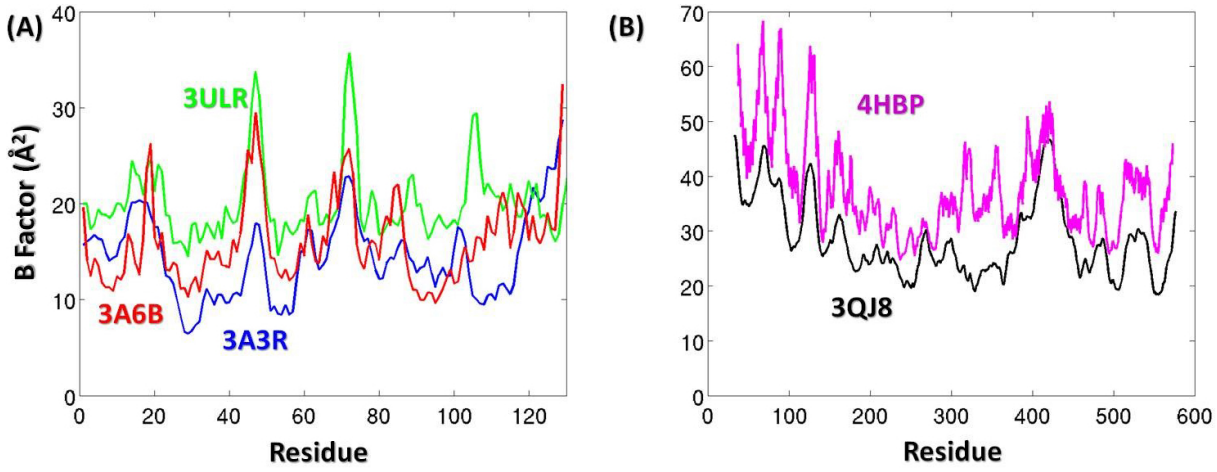

Figure 2. B-factors of structures crystallized under different conditions. (A) B-factors of three lysozyme structures. For PDB ID 3A3R (blue), 3ULR (green), 3A6B (red). (B) B-factors are shown for two FAAH structures; in blue, PDB ID 4HBP and in black, PDB ID 3QJ8.

As was explained earlier, crystallization conditions are generally very different from physiological conditions. Moreover, differences in crystallization conditions have a significant effect on residue flexibility and dynamics (Figure 2). MD simulations are commonly performed under physiological conditions. In order to quantify the effect of the differences in crystallization and physiological conditions on protein conformations and functions, MD simulations were performed for 43 monomeric and 26 multimeric proteins. A list of these proteins is provided in the Supplementary section (Table S1). RMSDs of the newly generated MD conformers from their starting crystal structures were evaluated. To this aim, first all conformers in the MD simulations were aligned with respect to the α-carbon positions of their crystal structures. Once aligned, the C^{α} RMSDs from their crystal structure coordinates were evaluated for the kth MD step as $RMSD_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|\vec{v}_{PDB_i} - \vec{v}_{k_i}\|^2}$.

Here N is the total number of C^{α} atoms, and $\vec{v}_{PDB_i}$ and $\vec{v}_{k_i}$ are the position vectors of the ith C^{α} atom in the crystal structure and MD conformer.

Figure 3A shows the time evolution of these RMSD values averaged over three different categories: (i) all monomers, (ii) all multimers, and (iii) all chains of the multimers treated separately. The drastic increase in RMSDs (up to ~ 0.4 Å) observed during the first 0.1 ns corresponds to the minimization part of MD. In the subsequent 10 ns of conventional MD (CMD) simulation the change in RMSDs became more subtle with time (gradual increase in the departure from the original structure, as the simulation duration increases), and eventually converged. Strikingly, multimers exhibited significantly larger average RMSDs compared to monomeric proteins.

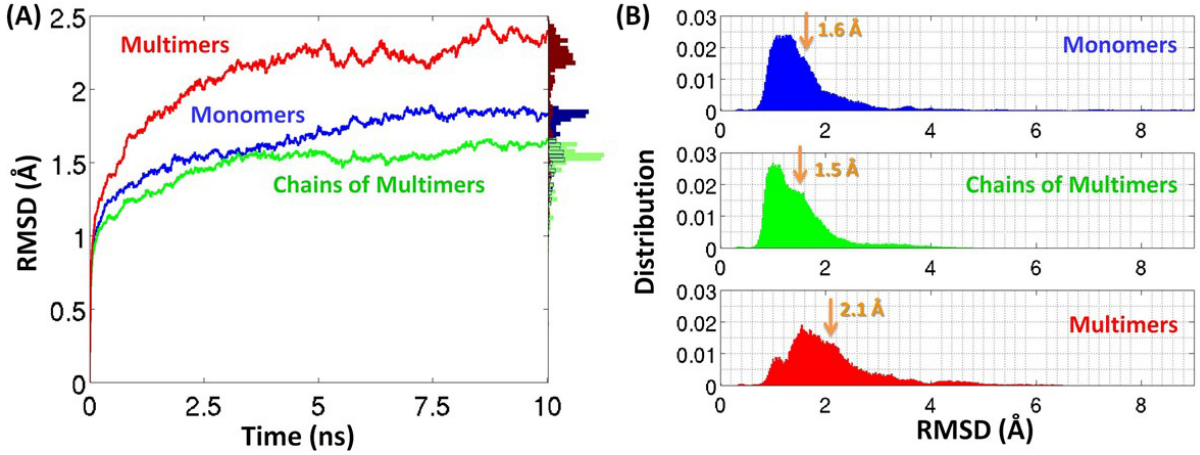

Figure 3. Time evolutions and distributions of RMSDs from the crystal structures. (A) The average of the RMSDs (from the crystal structures) for the 43 monomers (blue), the 26 multimers (red), and the 60 chains of the 26 multimeric proteins (green) are depicted. For the latter each chain was superimposed individually with its counterpart in the multimeric crystal structure prior to RMSD calculations. The distributions of the average RMSD values over the complete simulation length are shown horizontally on the y axis. (B) The distributions of the RMSDs from their crystal for all sampled MD conformers. The distribution was normalized and divided into 360 bins. Coloring is the same as in (A). The orange arrows show the trajectory averages of all RMSDs.

The histograms of the RMSDs collected during the complete MD lengths are presented in Figure 3B, showing that (i) broad ranges of RMSDs were sampled during the minimization and the subsequent 10 ns of CMD simulations, and that (ii) multimers exhibited broader RMSD distributions than monomers. The trajectory average of multimers was ~ 0.5 Å larger than that of single chains. Interestingly, the internal deformations observed for each chain in the multimers (set iii) compared well (scaled) with those observed for monomeric proteins (see Figure 3), clearly demonstrating that the fluctuation amplitudes of the individual chains are intrinsic properties that are closely maintained in the multimeric structures. Larger RMSDs in multichain proteins essentially arise from interchain movements, while intrachain fluctuations remain practically unaffected by multimerization. This observation also draws attention to the possible sensitivity of multichain/multidomain structures to crystallographic conditions, and the possibility of observing different rearrangements of subunits depending on crystallization conditions.

The crystal structures of proteins correspond to the minima of their free energy surfaces under crystallization conditions but are heavily affected by the presence of many crystal contacts. If MD simulations were to be performed under these crystallization conditions different from physiological conditions, the resulting

protein conformer corresponding to the free energy minima would deviate from their crystal structures. This is because, on the one hand, MD has difficulty taking into account crystal contacts and extreme temperatures, and, on the other hand, CMD force fields are optimized around physiological conditions. In the extreme, under crystallization conditions, CMD will not provide an accurate description of the time evolution of protein structures. Interestingly, it happens that changes in RMSDs during the first 2 orders of magnitude in MD simulation time, from 10^{-2} ns to 10^{0} ns, are qualitatively comparable with those happening in the following 2 orders of magnitude in simulation time, from 10^{0} ns to 10^{2} ns (see Figure 4). This behavior suggests that the RMSDs in the first 10 ns of simulation are not simply a result of the equilibrium fluctuations. They may rather be a result of the different thermodynamic conditions of the crystal solution and MD, i.e. the protein is shifting from one free energy surface to another.

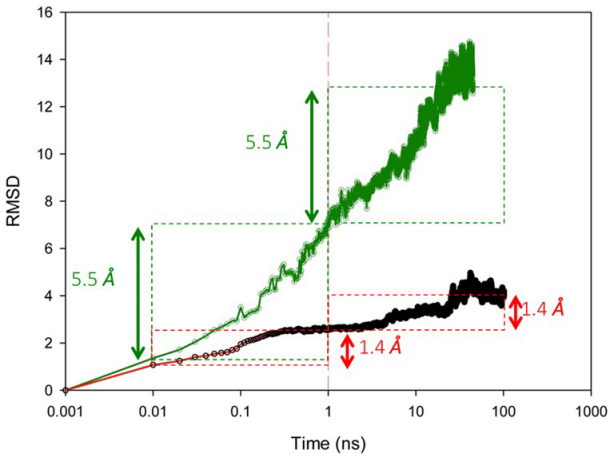

Figure 4. Time evolution of RMSD during MD simulations of complement control protein C3 and RNA Δ47Polymerase-II. In the semilog plot, the green and black symbols refer to complement control protein C3 (PDB ID 1G40) and to RNA Δ47PolymeraseII (PDB ID 1I3Q), respectively. The first is a single chain protein of 28 kDa (crystallized as a dimer); the second is a ten subunit complex and it weighs ~ 450 kDa. Conformers were aligned and RMSDs were calculated.

Crystal contacts affect the protein structure in two ways: (i) they introduce protein–protein interactions, which are not present under physiological protein concentrations, and (ii) they limit the protein–solvent interactions by decreasing the protein surface exposed to the solvent. In contrast to the crystals, the protein surfaces in our MD simulations were completely exposed to the solvent, hence contributing to the structural deviation from the crystal in the MD simulations. In the literature, extensive experimental and computational studies were performed.[2,3,12–15] on the effect of solvent on protein dynamics and structure.

The solvent-exposed surface was identified by the number of close contact water molecules: water molecules within 2 Å of protein. In Figure 5 we present the correlation between the RMSDs collected at 10 ns and the water-exposed surface area. There are 5 outliers (see the crossed data points in Figure 3), two of them (1C44 and 1QVE) falling above the line and 3 of them (1MSP, 1CUN, and 1A4Y) falling below the line. 1C44 is the Sterol Carrier Protein 2 (Scp2) from rabbit and 1QVE is the crystal structure of the truncated K122-4 pilin from *Pseudomonas aeruginosa*. Both of these structures have unstructured residue stretches on both the N and C terminus, which naturally contributes to the RMSD strongly. 1MSP is the crystal structure of the major sperm protein, alpha isoform (recombinant), while 1CUN is the crystal structure of repeats 16 and 17 of chicken brain alpha spectrin. All of these 3 structures share a common feature: they have rather

extended structures that allow them to have a large water contact surface. Except these 5 outliers, the RMSDs from the crystal structures exhibited an approximately linear correlation (Equation $y = ax$ fits the data with a residual of 0.79) with the water-exposed surface, hence strongly supporting the effect of crystal contacts on protein structure.

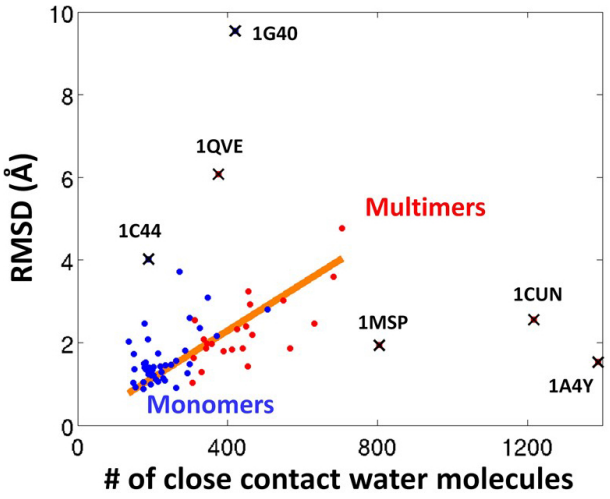

Figure 5. Correlation between solvent exposed surface area and RMSD. The RMSDs from crystals, collected at 10 ns, are shown against the water-exposed surface area of the proteins, which were calculated by taking the average of the number of water molecules within 2 Å of the proteins during last 0.25 ns of the protein backbone fixed (in their crystal positions) simulations. The fit of equation $= ax$, shown in orange, was performed by discarding the data points that are crossed out in the figure.

In addition to the 2 Å cutoff used earlier, the solvent-exposed protein surface was also assessed by the number of close contact water molecules within 5 Å of the protein. As shown in Figures 6A and 6B, the number of close contact water molecules within 2 Å and 5 Å both exhibited a linear correlation with the protein size, which was determined by the number of C^{α} atoms. The only data point not following this linear correlation was the ribonuclease inhibitor (PDB ID: 1A4Y). The ribonuclease inhibitor stands out among all the remaining proteins structurally as it exhibits a very packed structure through repeating alpha helices and beta sheets (Figure 6B). Considering the correlation between the number of close contact water molecules with both the protein size and the RMSD, it can be concluded that a larger RMSD from the crystal structure should be expected in MD simulations as the protein size increases.

3. Experimental

3.1. MD simulations in NAMD

MD simulations were performed for a broad set of 70 proteins ranging from a size of 95 residues up to 1166 residues and a resolution of 1.15–3 Å (Table S1). Each protein was simulated for at least 10 ns, totaling more than 1000 ns of CMD simulations. Structures were solvated in water boxes having at minimum a 10 Å cushion of water in each direction from the exposed atoms. Ions were added to neutralize the systems. Simulations were performed using the NAMD.[16] 2.9 package with CHARMM27 force field.[17] A cutoff distance of 12 Å was adopted for van der Waals interactions, with a switching function starting at 10 Å and reaching zero at 12 Å. The particle-mesh Ewald method[18] was used to compute long-range electrostatic forces. The equilibrated structures

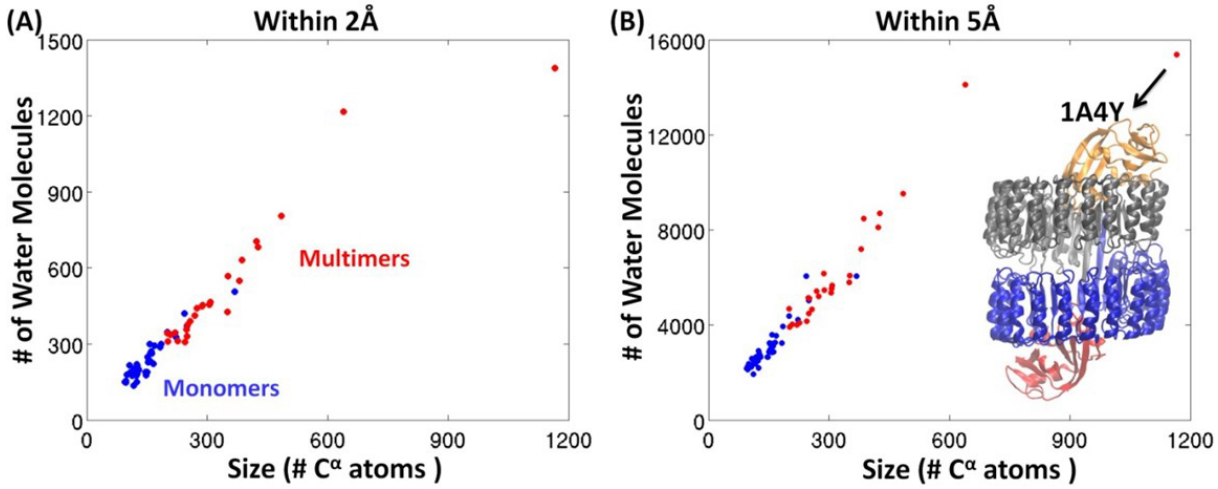

Figure 6. Correlation between the number of close contact water molecules and size of proteins. The numbers of water molecules were determined by counting the number of water molecules that are within 2 Å (A) and 5 Å (B) of the proteins for the last 0.25 ns of the protein fixed MD simulations. In these simulations the proteins are fixed in their crystal coordinates. The sizes of the proteins were identified by the number of their C^{α} atoms. The crystal structure of the ribonuclease inhibitor, which is an outlier to the linear correlation between number of water molecules and protein size, is shown in a new cartoon representation and colored by its chains.

were generated upon two cycles of minimization–equilibration simulations: the first cycle (minimization and subsequent equilibration simulations) at constant temperature (310 K) and pressure (1 atm) (NPT ensemble) with the C^{α} atoms held fixed and the second cycle at constant T and V (canonical ensemble). Each minimization simulation comprised 20,000 steps of minimization. The first CMDs, for which the proteins were held fixed, were performed for 0.5 ns. A damping coefficient of 0.5 ps^{-1} was used to maintain isothermal conditions. NPT simulations were performed with Langevin Nosé-Hoover method to keep the pressure constant.[19,20] Time steps of 1 fs were used in all simulations.

3.2. MD simulations in AMBER

Complement control protein C3 (PDB ID 1G40) and RNAΔ47PolymeraseII (PDB ID 1I3Q) structures were solvated in water boxes having at minimum a 10 Å cushion of water in each direction from the exposed atoms. Ions were added to neutralize the systems. The simulation box was built with explicit water using TLEAP from AMBERTOOLS 12. Simulations were performed using AMBER 12.[21] For both systems the protocol consists of standard four steps: 1) minimization: PDB is checked for possible overlap of loops; 2) heating: 0.1 ns of heating up to 298 K; 3) equilibration: 1 ns at 298 K; 4) production: 100 ns of free simulation. A cutoff distance of 12 Å was adopted for van der Waals interactions, with a switching function starting at 10 Å and reaching zero at 12 Å. The particle-mesh Ewald method was used to compute long-range electrostatic forces.

4. Conclusion

In this study we have described for the first time how different the thermodynamic conditions of the crystallization solutions and those of MD simulations (performed under physiological conditions) are. MD simulations of even short durations showed that the protein structures deviate from their original PDB coordinates, in particular multimers deviating more than monomers. For multimers two types of deviations were observed:

(i) interchain and (ii) intrachain. Interestingly, the interchain deformations appear not to depend on whether the structure is single- or multichain, which indicates that the larger RMSD in multichain/multimeric proteins essentially arises from interchain movements, while intrachain fluctuations remain practically unaffected by multimerization. A possible explanation for the latter is that, in the presence of crystal contacts, monomers act as though they were a part of a multimeric structure, i.e. the crystal packing mimics intrachain interactions of multimeric structures.

The observed structural deviations from crystals in MD simulations are not completely due to the thermodynamic conditional differences between crystal solutions and MD simulations. Several sources of errors in MD also contribute to the structural differences between MD conformers and their crystal structures. One source of error in MD is that Newton's equations of motion are not solved exactly. Instead, numerical methods are applied, which give rather approximate solutions. The force fields used to estimate the potential energy are another source of error. In these force fields mathematical functions are utilized to model interactions, some of these functions being crude approximations. Moreover, the parameters of these functions are derived from quantum mechanics and experimental data, which are susceptible to being sources of errors since they include numerous approximation and different types of experiments.

Crystal packing can also alter the protein dynamics and residue motions.[22–25] Liu et al.[22] showed that the mean square deviations (MSDs) for the X-ray models obtained from two crystal forms of the sugar binding protein LKAMG.[26] exhibit notable differences. Since these two structures are practically identical (having a backbone RMSD of 0.36 Å) the differences in dynamics were attributed to their different crystal arrangements; the crystal contacts with their 14 and 12 neighbors are at different locations on LKAMG. In contrast to the isolated structure, which reproduced only one of the experimental MSDs, the inclusion of the crystal contacts into the Gaussian Network Model (GNM).[27] calculations (a simple physics-based elastic network model) resulted in MSDs that reproduced both experimental MSDs to a high extent, showing unambiguously the effect of crystal packing on the dynamics.

Acknowledgments

We are grateful to Professor Ivet Bahar for the extensive discussions and support. The research reported in this publication was supported by the NIGMS of the NIH under award number P41GM103712 and the NIDA of the NIH under award number P30DA035778.

References

1. Nemethy, G.; Peer, W. J.; Scheraga, H. A. *Annu. Rev. Biophys. Bioeng.* **1981**, *10*, 459-497.
2. Pace, C. N.; Trevino, S.; Prabhakaran, E.; Scholtz, J. M. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **2004**, *359*, 1225-1235.
3. Park, C.; Carlson, M. J.; Goddard, W. A. *J. Phys. Chem. A* **2000**, *104*, 2498-2503.
4. Timasheff, S. N. *Curr. Opin. Struct. Biol.* **1992**, *2*, 35-39.
5. Kumar, S.; Ma, B.; Tsai, C. J.; Sinha, N.; Nussinov, R. *Protein Sci.* **2000**, *9*, 10-19.
6. Okazaki, K. I.; Takada, S. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 11182-11187.
7. Debenedetti, P. G. *Metastable Liquids: Concepts and Principles*; Princeton University Press: Princeton, NJ, USA, 1996.
8. Pullara, F.; Emanuele, A.; Palma-Vittorelli, M. B.; Palma, M. U. *J. Cryst. Growth* **2005**, *274*, 536-544.
9. Pullara, F.; Emanuele, A.; Palma-Vittorelli, M. B.; Palma, M. U. *Faraday Discuss.* **2008**, *139*, 299-308.

10. Vekilov, P. G. *J. Phys. Condens. Matter* **2012**, *24*, 193101.
11. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F. Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *Eur. J. Biochem.* **1977**, *80*, 319-324.
12. Chopra, G.; Summa, C. M.; Levitt, M. *P. Natl. Acad. Sci. USA* **2008**, *105*, 20239-20244.
13. Van Gunsteren, W.; Karplus, M. *Nature* **1981**, *293*, 677-678.
14. Hinsen, K.; Kneller, G. R. *Proteins* **2008**, *70*, 1235-1242.
15. Joti, Y.; Nakagawa, H.; Kataoka, M.; Kitao, A. *Biophys. J.* **2008**, *94*, 4435-4443.
16. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781-1802.
17. Mackerell, A. D. Jr.; Feig, M.; Brooks, C. L. 3rd. *J. Comput. Chem.* **2004**, *25*, 1400-1415.
18. Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089-10092.
19. Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613-4621.
20. Martyna, G. J.; Tobias, D. J.; Klein, M. L. *J. Chem. Phys.* **1994**, *101*, 4177-4189.
21. Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668-1688.
22. Liu, L.; Koharudin, L. M.; Gronenborn, A. M.; Bahar, I. *Proteins* **2009**, *77*, 927-939.
23. Kundu, S.; Melton, J. S.; Sorensen, D. C.; Phillips, G. N. Jr. *Biophys. J.* **2002**, *83*, 723-732.
24. Song, G.; Jernigan, R. L. *J. Mol. Biol.* **2007**, *369*, 880-893.
25. Diamond, R. *Acta Crystallogr.* **1990**, *46*, 425-435.
26. Koharudin, L. M.; Furey, W.; Gronenborn, A. M. *Proteins* **2009**, *77*, 904-915.
27. Bahar, I.; Atilgan, A. R.; Erman, B. *Fold. Des.* **1997**, *2*, 173-181.

Table S1. Dataset of proteins used in this study.[1]

PDB name	Protein name	Resolution, Å	Residue number	Oligomer	Protein family	RMSD, Å (last 0.25 ns)
11BA	Seminal ribonuclease	2.06	124	Dimer	PF00074	1.6096
1A0B	Aerobic respiration control sensor protein ArcB	2.06	117	Monomer	PF01627	1.0553
1A1X	Protein p13 MTCP-1	2	106	Dimer (3 missing residues)	PF01840	1.9287
1A3Z	Rusticyanin	1.9	150	Monomer	PF00127	1.3373
1A4R	Cell division control protein 42 homolog	2.5	190	Dimer	PF00071	3.0964
1A4Y	Ribonuclease inhibitor	2	1166	Tetramer complex	PF00074, PF13516	1.4748
1ACF	Profilin-1B	2	125	Monomer	PF00235	1.3341
1AEP	Apolipophorin-3b	2.7	153	Monomer	No Pfam information	1.0927
1AGI	Angiogenin-1	1.5	125	Monomer	PF00074	1.4196
1ANF	Maltose-binding periplasmic protein	1.67	369	Monomer	PF01547	2.9072
1AQT	ATP synthase epsilon chain	2.3	135	Dimer (1 missing residues)	PF00401, PF02823	1.7234
1AY9	Protein UmuD	3	108	Tetramer (5 missing residues)	PF00717	3.5174

1AZV	Superoxide dismutase [Cu-Zn]	1.9	153	Dimer	PF00080	2.8677
1B1U	Alpha-amylase/trypsin inhibitor	2.2	117	Monomer	PF00234	1.8421
1BP2	Phospholipase A2	1.7	123	Monomer	PF00068	1.5398
1BS4	Peptide deformylase	1.9	168	Monomer	PF01327	1.2956
1BTN	Spectrin beta chain	2	106	Dimer (12 missing	PF15410	1.8930
1BV1	Major pollen allergen Bet v 1-A	2	159	Dimer (10 missing residues)	PF00407	2.2387
1C0E	Low molecular weight phosphotyrosine protein phosphatase	2.2	154	Monomer	PF01451	1.4794
1C44	Sterol carrier protein 2	1.8	123	Monomer	PF02036	4.2381
1CEW	Cystatin	2	108	Monomer	PF00031	1.9788
1CGE	Interstitial collagenase	1.9	161	Monomer	PF00413	3.5973
1CPQ	Cytochrome c'	1.72	129	Monomer	PF01322	1.4728
1CQP	Integrin alpha-L	2.6	182	Monomer	PF00092	1.8064
1CTM	Apocytochrome f	2.3	250	Monomer	PF01333	2.2118
1CUN	Spectrin alpha chain	2	213	Trimer	PF00435	2.3767
1DQE	Pheromone-binding protein	1.8	137	Dimer	PF01395	1.6949
1DY3	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	2	158	Monomer	PF01288	0.9137
1E6K	Chemotaxis protein CheY	2	130	Monomer	PF00072	1.3300
1EDH	Cadherin-1	2	211	Dimer	PF00028	4.8298

1EW4	Protein CyaY	1.4	106	Monomer	PF01491	1.7499
1F21	Ribonuclease HI	1.4	152	Monomer	PF00075	1.4532
1FLM	FMN-binding protein	1.3	122	Dimer	PF01243	1.0932
1G40	Complement control protein C3	2.2	243	Monomer	PF00084	9.0000
1GNU	Gamma-aminobutyric acid receptor-associated protein	1.75	117	Monomer	PF02991	1.2533
1GPR	Glucose permease	1.9	158	Dimer (11 missing residues)	PF00358	3.4936
1GVJ	Protein C-ets-1	1.53	146	Dimer (5 missing residues)	PF00178	2.5914
1H0A	Epsin-1	1.7	158	Monomer	PF01417	1.5411
1H7M	50S ribosomal protein L30e	1.96	97	Monomer	PF01248	1.0026
1HCV	Immunoglobulin G	1.85	112	Monomer	No Pfam information	2.4047
1HUF	Tyrosine phosphatase yopH	2	123	Monomer	PF09013	0.8418
1IAZ	Equinatoxin-2	1.9	175	Dimer	PF06369	2.1222
1IFR	Prelamin-A/C	1.4	113	Monomer	PF00932	1.0644
1ILR	Interleukin-1 receptor antagonist protein	2.1	145	Dimer (1 missing residues)	PF00340	1.3743
1J2A	Peptidyl-prolyl cis-trans isomerase A	1.8	166	Monomer	PF00160	1.2953
1JF4	Globin, monomeric component M-IV	1.4	147	Monomer	PF00042	1.2989
1JHG	Trp operon repressor	1.3	101	Dimer	PF01371	1.5901

1JV4	Major urinary protein 2	1.75	157	Monomer	PF00061	1.4780
1K40	Focal adhesion kinase 1	2.25	126	Monomer	PF03623	1.5464
1KTJ	Mite group 2 allergen Der p 2	2.15	129	Dimer	PF02221	1.7170
1KX8	Chemosensory protein A6	2.8	100	Monomer	PF03392	1.3700
1MB1	Transcription factor MBP1	2.1	98	Monomer	PF04383	1.0494
1MSP	Major sperm protein isoform alpha	2.5	124	Tetramer (12 missing residues)	PF00635	1.7179
1NCO	Neocarzinostatin	1.8	113	Dimer	PF00960	2.2065
1OCV	Steroid delta-isomerase	2	125	Dimer	PF02136	1.3219
1OMR	Recoverin	1.5	201	Monomer	PF00036	2.9329
1PI1	MOB kinase activator 1A	2	185	Monomer	PF03637	2.7661
1PM4	*Yersinia pseudotuberculosis* produced superantigens	1.75	117	Trimer	PF09144	1.7877
1QAU	Nitric oxide synthase, brain	1.25	112	Monomer	PF00595	1.3679
1QJ8	Outer membrane protein X	1.9	148	Monomer	PF13505	1.4211
1QVE	Fimbrial protein	1.54	126	Dimer (1 missing residues)	PF00114	6.1419
1QYM	26S proteasome non-ATPase regulatory subunit 10	2.8	223	Monomer	PF00023	2.1046
1RSY	Synaptotagmin-1	1.9	126	Monomer	PF00168	1.0984
3NYL	Amyloid beta A4 protein	2.8	196	Dimer (6 missing residues)	PF12925	2.2671
1TVQ	Fatty acid-binding protein	2	125	Monomer	PF00061	1.1464

1VC1	Putative anti-sigma factor antagonist TM_1442	2	110	Dimer	PF01740	2.0378
1XMT	Acetyltransferase At1g77540	1.15	95	Monomer	PF14542	1.6281
1XNI	Tumor suppressor p53-binding protein 1	2.8	118	Monomer	PF09038	1.2879
3DFR	Dihydrofolate reductase	1.7	162	Monomer	PF00186	1.0573

Reference

1. Case, D. A.; Darden, T. A.; Cheatham, T. E. 3rd; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M. et al. *AMBER 12*; University of California: San Francisco, CA, USA, 2012.