COARSE-GRAINED SIMULATION OF CALBINDIN D9K

Mahendra Thapa* and Mark Rance**

*Department of Physics, University of Cincinnati (UC), OH, USA.

**Dept. of Molecular Genetics, Biochemistry and Microbiology, UC, OH, USA.

Abstract: The coarse-grained protein modeling tool, Cabs-flex, is feely available online server; it is based on the CABS model in which each residue of a protein has been represented by four points. The server was used for the protein Calbindin D_{9k} in it's doubly calcium loaded state: small and single domain protein of the EF-hand family. Twelve representative structures, in all-atom format corresponding to each cluster, were also downloaded along with trajectories, ready-made plots, images, video, data files of C_α RMSD, atomic fluctuation and GDT_TS. In the present study, simulated C_α atomic fluctuations for residues of the protein was compared with the experimental results and also correlated with the respective C_α RMSD and GDT_TS.

Keyword: Calbindin D_{9k}; Coarse-grained simulation; CABS model; Cabs-flex server.

INTRODUCTION

Identification of flexible regions of protein structures is important for understanding of their biological functions. Local fluctuations (side-chains, loops) occur in picoseconds, while global rearrangements (folding/ unfolding) require typically milliseconds, even for small globular proteins. Nuclear magnetic resonance (NMR) spectroscopy and all-atom molecular dynamics (MD) are the methods of choice for investigation of protein flexibility in solution[2,7,8]. The all-atom MD simulations are considered the gold standard in simulating protein dynamics because they capture the essential physics of protein dynamics. Because of the difficulty of NMR studies and timescale problems in all-atom MD, coarse-grained methods have emerged as an inexpensive and powerful alternative. In many cases, the computational cost in MD is prohibitive for biologically relevant processes due to its limitation to relatively short time sales[2,3]. Much longer time scales can be accessed by properly designed coarse-grained models.

The CABS model [2, 4]is a coarse-grained model in which a residue of a protein chain is represented by up to four atoms (the C_α and C_β atoms and two virtual pseudo-atoms: the center of mass of a side chain and the center of the C_α – C_α virtual bonds). The CABS force field includes knowledge-based statistical potentials (sequence-dependent short conformational preferences, context-dependent potential of pairwise interactions of side chains and a model of the main chain hydrogen bonds) accounting for the solvent effects in an implicit fashion and the sampling is realized by the Monte Carlo method. The CABS dynamics are simulated by a random series of small local moves (controlled by Monte Carlo scheme) whose long-term evaluation describes the protein dynamics well [2-7].

Cabs-flex [1] is an online server for the fast simulation of folded globular protein structure which is based on the CABS model [2,3]: a well-established coarse-grained protein modelling tool which may generate consistent protein dynamics at highly reduced (three orders of magnitude) cost , although with some decrease of the resolution. The authors demonstrated that the consensus view of protein near native dynamics obtained from 10 ns MD simulations (all-atom, explicit water, for all protein meta-folds using the four most popular force fields) is consistent with the

Author for Correspondence: Mahendra Thapa, Department of Physics, University of Cincinnati, OH, USA.
E-mail: mahendrab2007@gmail.com

Received: 2 Jan 2020; First Review: 29 Aug 2020; Second Review: 30 Aug 2020; Accepted: 1 Sep 2020.
Doi: https://doi.org/10.3126/sw.v14i14.34977

CABS dynamics. This server needs only a protein structure in PDB format to start the simulation. The input structure is used as a starting point for the CABS simulation of near-native dynamics. The resulting trajectory is automatically analyzed and processed to provide the useful description of protein dynamics. The CABS-flex pipeline uses the CABS simulation protocol. This server is free and open to all users, and there is no login requirement. After clicking submit button (preceded by filling the input-project name and pdb file data), a web link to the result is provided, which the user can bookmark and access at a later time. Web links to the submitted jobs are displayed on a queue page, unless the option 'Do not show my job on the queue page' is marked. Typically, the computations take about 2 to 3 hours[3].Importantly, the computational cost of obtaining near native dynamics by CABS simulations was proved to be much lower ($\approx 6\times10^3$ times) than that of MD. The CABS design is a compromise between high sampling efficiency and high resolution of protein representation[7].

Calbindin $D_{9k}$is a single domain protein of the EF-hand family with molecular weight 8700 Dalton and 75 amino acids. It consists of a pair of EF-hands connected by a flexible linker (from residue number 36 to 45); the EF-hand is a helix-loop-helix motif. The first EF-hand possesses the following residues: helix (3-15), loop (16-24) and helix (25-35); similarly the second EF-hand possesses helix (46-53), loop(54-62) and helix(63-73). It is found predominantly in tissues involved in the uptake and transport of calcium such as cells of the intestinal brush boarder membrane, also in the kidney and uterus in some mammalian species. It in encoded in humans by the S100G gene. It is one of the vitamin D-dependent calcium-binding proteins. It serves as an attractive model system for computational investigation due to its small size. Its structure in various calcium-loaded states has been extensively characterized experimentally [NMR/X – rays].The apo structure of the protein was determined by high resolution 1H NMR technique and refined with restrained molecular dynamics in vacuo[12,13].

In the present paper, how the data, plots and tables obtained from the computationally cheap means of accessing backbone dynamics than atomic MD but extremely useful online coarse-grained server, CABS flex, were used. The aim was to show how the combination of freely available online server for the protein simulation and the coarse-grained simulation technique could give thermodynamics properties like B-factor of atoms/residues of the protein. The results so obtained were compared with the experimental values. This paper could be a starting point for those who will be planning to use online servers and simulation techniques to address specific research questions.

METHOD

The input file for the Cabs flex server (http://biocomp.chem.uw.edu.pl/CABSflex/[1]) is the protein structure file (PDB code or uploaded by a user). The requirements for the input file are as follows: (i) it should be in PDB format, (ii) it should be a single and continuous (without breaks) protein chain up to 400 amino acids in length, (iii) non-standard amino acids should not be used and (iv) each residue must have backbone atoms N, C_α, C and O; side chain atoms may be missing. If there are multiple conformation (e.g., the structures determined by NMR) in the input file, the server uses only the first one. Also, heteroatoms (e.g., water, ligand) aren't considered by the server. If an input file has breaks in the protein chain, this could be fixed using the software as mentioned in the web-page of the server[1]. In our case, PDB code '4ICB'[14] was used for the server as the input structure. As it has both chain A and chain B for a few residues, so the server might consider the chain A only. It also have two calcium atoms and water molecules, so they might be ignored by the server. After running the server about half an hour, following figures, data, plots and tables can be downloaded:

(i) A PDB file of C_α trajectory with 2000 frames (i.e. 2000 snapshot trajectory in C_α trace representation).

(ii) 12 superimposed PDB files, each of which represented a cluster.

(iii) Residue fluctuation profile in text and EPS format.

(iv) A picture of 12 models of above (ii).

(v) A movie in OGV or MP4 format generated by these 12 predicted models.

(vi) Residue RMSD for each of 12 predicted models in text and EPS format.

(vii) Cluster data, C_α RMSD and GDT_TS to the input structure, C_αRMSD between predicted models and C_αGDT_TS between predicted models.

For the generation of online presented plots, pictures and eps plot files, Gnu plot 4.5 was used. For the generation of online presented pictures and movie showing an ensemble of predicted models in the cartoon representation, open source PyMOL visualization software was used.

RESULTS AND DISCUSSION

Based in C_α RMSD, 2000 models produced by the server for the given input PDB file were grouped into 12 clusters (which were done by k-means methods) as shown in the table 1 and consequently we got 12 corresponding representative models which were called here predicted models. The PDB files corresponding to these 12 predicted models consisted of coordinates of all atoms unlike the total 2000 models which consisted of coordinates of $C_\alpha$atoms only. These predicted models were being superimposed on each other using Theseus[16]software. As shown in the table 1, 340 models with average RMSD 0.8 fell in the cluster number 1, 284 models with average RMSD of 0.8 fell in the cluster number 2 and so on. Even though both clusters had the same average RMSD, they had different cluster density which was defined as cluster size divided by the average RMSD. That is, the clusters were numbered according to the cluster density values, from the most dense (numbered the first) to the least one. Since the cluster 1 being the most dense, the predicted model 1 was considered as the most dominant conformation among these 2000 structures, followed by the predicted model 2 and so on. As shown in the table 1, RMSD values were low which suggested that the simulated structures didn't change much from the initial structure which supported the previous work based on all-atom molecular dynamics simulation of the protein[12].

Table 1. ***Clustering data***

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12
Cluster density	435	349	244	226	193	161	132	115	107	100	84	61
Cluster size	340	284	210	194	175	161	132	123	117	103	92	69
Average Cluster RMSD	0.8	0.8	0.9	0.9	0.9	1.0	1.0	1.1	1.1	1.0	1.1	1.1

Table 2. ***C_α RMSD & GDT_TS to the input structure***

Cluster number	1	2	3	4	5	6	7	8	9	10	11	12
RMSD	2.73	2.77	2.76	3.07	2.61	2.88	2.73	2.84	2.80	2.75	2.85	2.41
GDT_TS	0.69	0.68	0.69	0.65	0.69	0.66	0.68	0.66	0.67	0.69	0.68	0.71

As shown in the table 2, 2.73 was C_α RMSD between the predicted models and input structure for the cluster1and so on; 0.69 was the corresponding global distance test (GDT_TS). GDT_TS metric is considered more accurate measurement than RMSD. The claims made for the table 1 was also applied for the table 2.

Table 3. ***C_α RMSD between predicted models***

#	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1.09	1.72	1.68	0.85	1.38	1.06	1.56	1.75	1.43	1.38	1.57
2	1.09	0	1.53	1.70	1.10	1.37	1.30	1.72	1.91	1.31	1.46	1.60
3	1.72	1.53	0	1.17	1.45	1.74	1.52	1.59	1.96	1.71	1.90	1.81
4	1.68	1.70	1.17	0	1.54	1.75	.51	1.42	2.02	1.90	1.95	2.00
5	0.85	1.10	1.45	1.54	0	1.46	0.99	1.47	1.80	1.48	1.38	1.43
6	1.38	1.37	1.74	1.75	1.46	0	1.56	1.49	1.50	1.45	1.81	1.98
7	1.06	1.30	1.52	1.51	0.99	1.56	0	1.54	1.69	1.51	1.48	1.50
8	1.56	1.72	1.59	1.42	1.47	1.49	1.54	0	1.70	1.82	1.81	2.10
9	1.75	1.91	1.96	2.02	1.80	1.50	1.69	1.70	0	1.54	1.72	1.93
10	1.43	1.31	1.71	1.90	1.48	1.45	1.51	1.82	1.54	0	1.31	1.61
11	1.38	1.46	1.90	1.95	1.38	1.81	1.48	1.81	1.72	1.31	0	1.61
12	1.57	1.60	1.81	2.00	1.43	1.98	1.50	2.10	1.93	1.61	1.61	0

As shown in the table 3, 1.09 was the RMSD between first and second predicted models. All C_α RMSD were within 2.00, so these 12 predicted structures didn't deviate much from the input structure.

Table 4. ***C_α GDT_TS between predicted models***

#	1	2	3	4	5	6	7	8	9	10	11	12
1	1	0.90	0.85	0.88	0.95	0.89	0.91	0.89	0.83	0.87	0.89	0.82
2	0.90	1	0.88	0.83	0.91	0.93	0.89	0.85	0.83	0.93	0.90	0.82
3	0.85	0.88	1	0.89	0.88	0.84	0.87	0.84	0.82	0.86	0.79	0.79
4	0.88	0.83	0.89	1	0.87	0.85	0.90	0.87	0.87	0.82	0.81	0.76
5	0.95	0.91	0.88	0.87	1	0.90	0.92	0.90	0.83	0.88	0.89	0.86
6	0.89	0.93	0.84	0.85	0.90	1	0.88	0.88	0.86	0.92	0.86	0.79
7	0.91	0.89	0.87	0.90	0.92	0.88	1	0.85	0.88	0.86	0.86	0.82
8	0.89	0.85	0.84	0.87	0.90	0.88	0.85	1	0.85	0.82	0.86	0.75
9	0.83	0.83	0.82	0.87	0.83	0.86	0.88	0.85	1	0.87	0.85	0.79
10	0.87	0.93	0.86	0.82	0.88	0.92	0.86	0.82	0.87	1	0.88	0.82
11	0.89	0.90	0.79	0.81	0.89	0.86	0.86	0.86	0.85	0.88	1	0.83
12	0.82	0.82	0.79	0.76	0.86	0.79	0.82	0.75	0.79	0.82	0.83	1

In the table 4, values of C_α GDT_TS between predicted models were shown. These were close to 1, so they further reinforced the results of the table 3.

Fig 1 showed the residue fluctuation profile which gave the relative propensities of protein residues to deviate from an average structure of the trajectories. The fluctuation of the residue i is defined[2] as

$$\langle R_i^2 \rangle = \frac{1}{N}[(p^i_{j,x} - c^i_{j,x})^2 + (p^i_{j,y} - c^i_{j,y})^2 + (p^i_{j,z} - c^i_{j,z})^2]$$

Where j is the trajectory frame, i is the residue index, c is the position of the C_α atom in the average structure and N is the number of trajectory models.<> denotes the average over a whole trajectory for the residue i.

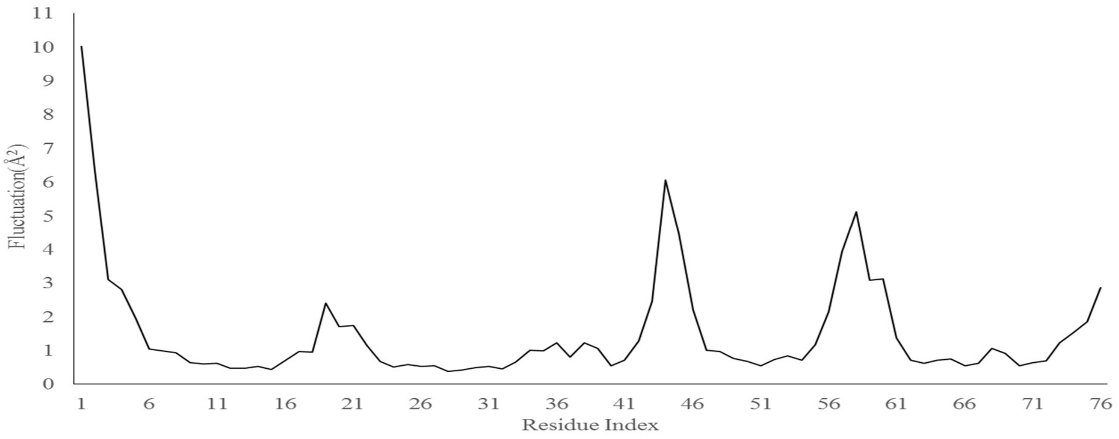

Figure 1: Residue fluctuation profile

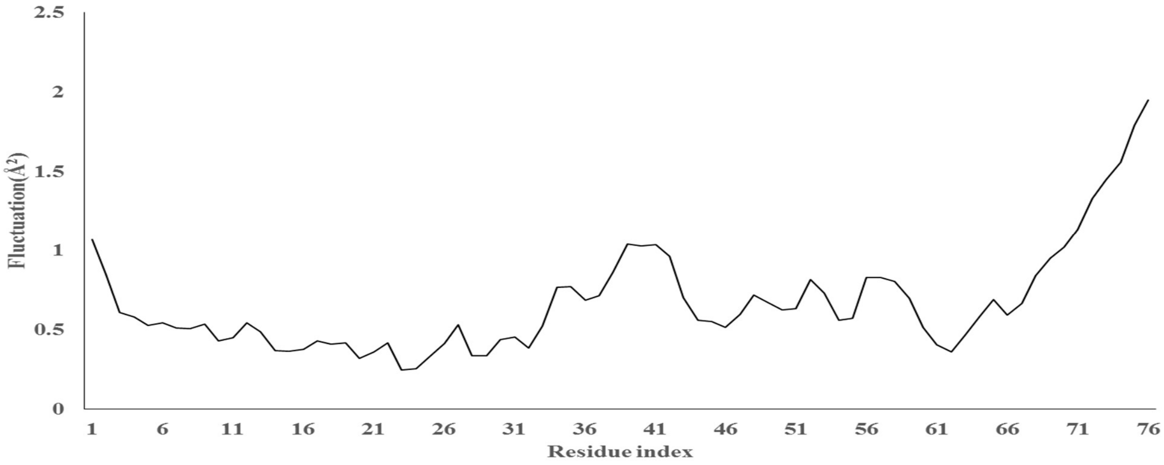

Figure 2: Residue fluctuation profile for C_α from PDB file of 4ICB

Fig 2 showed the plot of fluctuation versus residue number for C_α atom of each residue. B-factor values of these atoms were given in the download pdb file for 4ICB. The relation between the B-factor and atomic fluctuation of the protein is given as[15]

$<\Delta r^2_i> = (3/8\pi^2)\ B_i$, where $<\Delta r^2_i>$ is the average atomic fluctuation for the atom i and B_i is the corresponding B-factor.

[7]The B-factors of residues of the protein reflect protein flexibility and they are influenced by factors such as crystallization conditions, the refinement method (used for the interpretation of X-ray data) and the molecular environment of the crystal structure. The crystal environment has a significant effect on protein flexibility: the spectrum of fluctuations is considerably flattened in crystal as compared with that in solution [9]. It should be noted that X-ray structures have been determined at cryogenic temperatures; at that temperature B-factors have reduced values because of the packing defects and it may result in unrealistically unique non-functional structures. These reasons hence warned that descriptions of protein flexibility derived from X-ray models and B-factors must be approached with caution[7,10,11]. Studies of the plots as shown in Fig 1 and Fig 2 also supported the above mentioned arguments in the following lines: (i) The maximum value of fluctuation was about 1.0 Å (except

last few C-terminus residues) in the experimental case while it was about 6 Å for the simulation case (except first few N-terminus residues). This supported the all atom simulation results[12] that fluctuations were overestimated in the simulation than in the experiment. (ii) Neglecting end residues, fluctuation of residues in simulation is greater than those for the corresponding residues in the experiment. (iii) Since residues from residue number 36 to 45 represent the linker region, which connects a pair of EF hand of the protein, naturally fluctuations of these residues should be higher. The maximum value of fluctuation fell Coarse- grained simulation did not take into account of two calcium atoms bounded by EF-hands of the proteins but in reality the effects of these atoms couldn't be neglected for the cases in which fluctuations were being studied. (v) Because these two calcium binding sites are coupled by a short β-type interaction formed by two backbone-backbone hydrogen bonds between L23 and

in that range of residues for both cases as shown by peaks in Fig 1 and fig 2. (iv) As the calcium atom in the first binding site is bounded by ligands provided by residues 14-27, fluctuations of these residues are expected to be low as seen in the experiment; similar is the case for the calcium atom in the second binding site which is bounded by ligands provide by residues 54-65, Fig 2. In contract, two peaks (representing high fluctuation value) were seen in these regions in simulation, Fig 1. The results were not unusual because the simulation protocol of this

V61, the fluctuations of these residues were low in both simulation and experiment.

The Fig 3 showed the residue RMSD value profile i.e., the RMSD between the input structure and the first predicted model. The nature of the plots for RMSD of the input structure and other predicted models was seen similar.

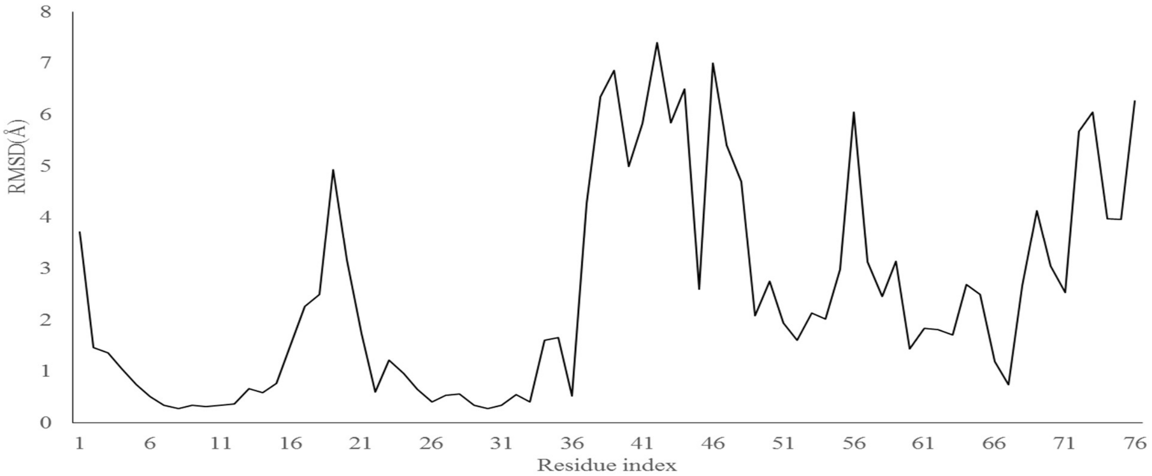

Figure 3: RMSD profile for first cluster

The plot shown in Fig 3 actually repeated the same story as given Fig 1; rmsd of residues (16-24 and 54-62) those binds two calcium atoms possessed peak values as in Fig 1. Similarly, the linker region of the protein (36-45) also showed peak values. Although these rmsd values were high for the residues that provided ligands to bind calcium atoms and linker region residues, residues of helix regions were comparable with those obtained from computationally expansive all-atom simulation of the protein [12].

CONCLUSIONS AND FUTURE DIRECTION

The results of fluctuation of residues of the protein calbindin D_{9k} calculated from coarse-grained simulation resembled with those obtained from the crystallographic experiments taking into accounts on (i) the experimental conditions such as crystallization conditions, the refinement methods and (ii) the simulations conditions such as negligence of calcium atoms, considering only four representative points in each residue instead of all atoms of the protein, simulation time. The simulated structures or trajectories obtained so far from the online Cabs Flex server could be used in another online server (or codes in Fortran/python or any other programs) to find order parameter, say, backbone N-H,for each residue of the protein. Care should be taken that all simulated structures should be aligned before using Henry method for calculating order parameter. If ired approach is used, then there is no necessary to align. The result could be compared with corresponding experimental results for apo state of the calbindin D_{9K} (pdb code 1CLB) instead of the doubly loaded state (pdb code 4ICB) because ,as already said, this Cabs Flex server did not consider calcium atoms in two binding sides of the protein.

REFERENCES

1. http://biocomp.chem.uw.edu.pl/CABSflex/about.php
2. Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field, Jamroz M., Orozco M., Kolinski A., Kmiecik S. 2013. J. *Chem. Theory Comput.* **9 (1)**: 119–125. doi: 10.1021/ct300854w
3. CABS-flex: server for fast simulation of protein structure fluctuations, Michal Jamroz, Andrzej Kolinski and Sebastian Kmiecik, *Nucleic Acids Research, 2013*. **41**: 427–431. doi:10.1093/nar/gkt332
4. Protein modeling and structure prediction with a reduced representation, Kolinski, A. (2004). *Acta Biochim. Polonica*. **51**: 349–371
5. The role of dynamic conformational ensembles in biomolecular recognition, *Nat. Chem. Biol*., Boehr, D.D., Nussinov, R. and Wright, P.E. 2009. **5**: 789–796
6. Fromcoarse-grained to atomic-level characterization of protein dynamics: transition state for the folding of B domain of protein, Kmiecik, S., Gront, D., Kouza, M. and Kolinski, A. 2012 A. J. *Phys. Chem*. B. **116**: 7026–7032
7. CABS-flex predictions of protein flexibility compared with NMR ensembles, Bioinformatics, 2014, pg. 1–5, doi:10.1093/ bioinformatics/btu184
8. Coarse-grained representation of protein flexibility. Foundations, successes, and shortcomings, Orozco M[1], Orellana L, Hospital A, Naganathan AN, Emperador A, Carrillo O, Gelpí JL., *Adv Protein Chem Struct Biol*. 2011. **85**:183-215. doi: 10.1016/B978-0-12-386485-7.00005-3.
9. Protein flexibility in solution and in crystals, Eastman, P. et al. 1999. J. *Chem. Phys*.. **110**: 10141–10152.
10. Accessing protein conformational ensembles using room temperature X-ray crystallography, Fraser, J.S. et al. 2011 Proc. *Natl Acad. Sci. USA*. **108**: 16247–16252
11. Crystalline ribonuclease A loses function below the dynamical transition at 220 K, Rasmussen, B.F. et al. (1992). *Nature*. **357**: 423–424.
12. Molecular Dynamics Study of Calbindin D_{9k} in the Apo and Singly and Doubly Calcium-Loaded States. S.Marchand & Benoit Roux.PROTEINS: *Structure, Function and Genetics. 1998*. **33:**265-284.
13. Computational Prediction of Chemical Shifts of Apo-state of the protein Calbindin D_{9k}. Mahendra Thapa & Mark Rance. *Scientific World.* September 2014. **12 (12),**
14. Proline cis-trans isomers in calbindin D_{9k} observed y X-ray crystallography. Svensson L.A., Thulin E. Forsen. *Phy.J.Mol*. 1992. **223**:601-606.
15. Thermal vibrations in crystallography. Willis B.T.M, Pryor A.W, Cambridge University press. 1975
16. Optimal simultaneous superpositioning of multiple structures with missing data.Theobald, Douglas L. &Steindel, Philip A. (2012) *Bioinformatics*. **28 (15):** 1972-1979.

■