BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 11(8)

Modelling and Characterization of Glial Fibrillary Acidic Protein

Hemchandra Deka[1, 2], Rajeev Sarmah[2], Ankita Sharma[1], Sagarika Biswas[1]*

[1]CSIR- Institute of Genomics and Integrative Biology, Mall Road, Delhi, India; [2]Centre for Bioinformatics Studies, Dibrugarh University, Assam, India; Sagarika Biswas -- Email: sagarika.biswas@igbi.res.in; Phone: 9818004740; Fax# 011- 27662407; #9818004740; *Corresponding author

Received July 08, 2015; Revised July 30, 2015; Accepted August 01, 2015; Published August 31, 2015

Abstract:
Glial Fibrillary Acidic Protein (GFAP) is an intermediate-filament (IF) protein that maintains the astrocytes of the Central Nervous System in Human. This is differentially expressed during serological studies in inflamed condition such as Rheumatoid Arthritis (RA). Therefore, it is of interest to glean molecular insight using a model of GFAP (49.88 kDa) due to its crystallographic non-availability. The present study has been taken into consideration to construct computational protein model using Modeller 9.11. The structural relevance of the protein was verified using Gromacs 4.5 followed by validation through PROCHECK, Verify 3D, WHAT-IF, ERRAT and PROVE for reliability. The constructed three dimensional (3D) model of GFAP protein had been scrutinized to reveal the associated functions by identifying ligand binding sites and active sites. Molecular level interaction study revealed five possible surface cavities as active sites. The model finds application in further computational analysis towards drug discovery in order to minimize the effect of inflammation.

Keywords: Astrocytes, proteomics, Rheumatoid Arthritis, Modeller 9.11, Gromacs 4.5

Background:
Glial Fibrillary acidic protein (GFAP) is an intermediate filament protein having molecular weight 49.88 kDa and is found to be present in the glial cells of the Central Nervous System (CNS) **[1].** It was first isolated from human multiple sclerosis plaques in 1971 **[2].** This protein is used as a classical marker for astrocytes in the vertebral central nervous system **[3]** and it plays role in the de-differentiation of the axon and maintaining the strength of the astrocytes **[4].** In our previous study we have identified this protein to be significantly up regulated during the chronic inflammation of Rheumatoid Arthritis (RA) **[5].** RA is an autoimmune inflammatory disease that affects the joints in systemic manner. In order to understand the etiology of this disease, the understanding of the insights of protein structure may play an important role. But lack of crystallographic or NMR deduced model structure of GFAP restricted the complete understanding of protein role towards RA. To better understand the function of a protein, 3 dimensional (3D) structure of protein is needed. In case of unavailability of experimentally determined protein structure, we made an attempt to build a model of the protein of interest using *in-silico* methods **[6, 7].** From an earlier study of the protein we came to an assured point of the rod domain of GFAP being conserved **[8].** In the present study, we generated a model based on comparative protein modeling. The model generated has been subjected for its Molecular Dynamic (MD) and protein quality analysis using Gromacs 4.5. This robust check helped us in validating the schematically prepared protein model.

Methodology
The whole experiment had been carried out in Windows7 and Ubuntu 10.1 Operating systems. Comparative modeling has been carried out by Modeller 9.11 package **[9]** in Windows while molecular dynamic simulation was done in Gromacs 4.5 **[10]** in the Ubuntu platform. The models were visualized in Pymol, the active sites were predicted using Molegro Virtual Docker **[11].**

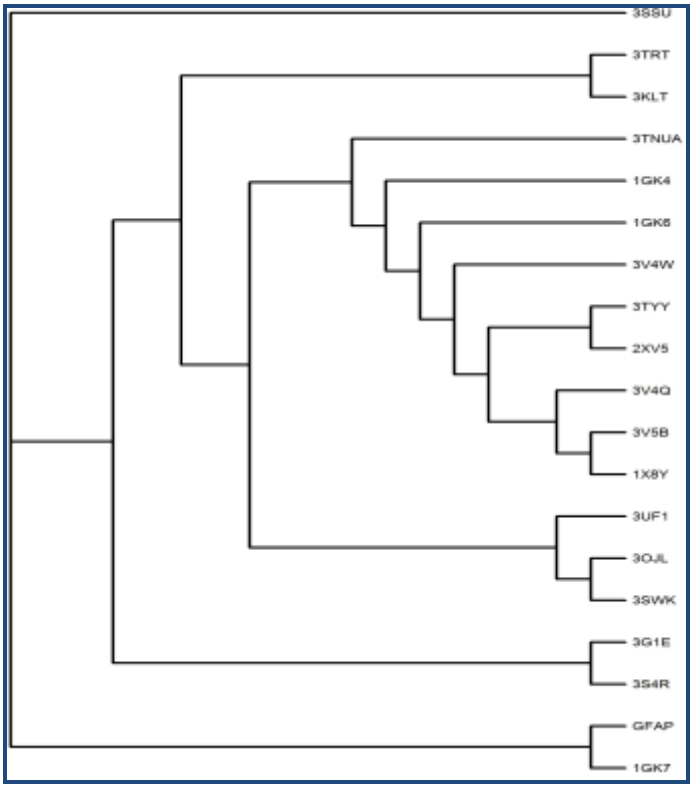

Figure 1: The tree showing relationship among GFAP and the rest templates. Figure is showing close ancestral relationship of GFAP with selected templates such as 3S4R, 3SSU, 3G1E, and 1GK7.

Sequence Retrieval

The sequence of Glial Fibrillary Acidic Protein (GFAP), accession no AAB22581.1 GI: 251802; of *Homo sapiens* was retrieved in FASTA format from NCBI protein database.

Template Identification

The program BLAST-P **[12]** has been used to detect similar crystallographic protein structures of GFAP. The parameters set for the search include all default and maintained the search for highly similar sequences/structure. The BLOSUM 62 **[13]** matrix was used for scoring the similarity. The better identical sequences having low gap percentage and high identity as well as higher number of positives, were chosen for templates. The template structures were then downloaded from the RCSB-PDB. The identification of the templates was based on the results of the query coverage, since there were no homologous proteins available in the PDB database. So the templates were selected on the basis of following criteria: 1) The proteins that share common ancestor with GFAP; 2) Short query coverage must have high identities; 3) Low identities must have high positives and large query coverage. The sequences of the chosen templates including the query were then subjected for multiple sequence alignment and 2D secondary structure alignment for analysis of secondary structural variation among the proteins.

Modeling of GFAP and quality analysis studies

Modeling was performed using Modeller 9.11. This program models protein tertiary structure by satisfaction of spatial restraint using standard parameters sets. The generated three-dimensional model includes all non-hydrogen main-chain and side-chain atoms. Generated model has been refined using energy minimization techniques to optimize stereochemistry and to remove bumps and steric clashes among non-bonded interactions using the commands of Modeller 9.11. Scripts generated five default models of the protein with different DOPE (Discrete Optimized Protein Energy) score. The model with lowest DOPE score was selected due to less steric clashes. This model is further refined by energy minimization by steepest descent method by applying *AMBER03 Force Field* **[14]** and solvating the protein in water. The force of 1000 kj/mol was applied for the purpose. The process of energy minimization was carried out using GROMACS 4.5 **[15]** for 5 nano second for observing the stability, computing the velocity temperature coupling and relaxing the system by pressure temperature coupling. The protein behavior was measured by evaluating Root mean square deviation (RMSD), Root mean square fluctuation (RMSF), and Radius of gyration (Rg). RMSD is the measure of query structure deviation and fluctuation from the crystal structure respectively **[16].** The radius of gyration of a protein is a measure of its compactness. If a protein is stably folded, it is likely to maintain a relatively steady value of Rg. If a protein unfolds, its Rg will change over time. Final structure was validated by Ramachandran plot using Rampage. It is an online tool which validates the dihedral angles present among the amino acids based on the Ramachandran plot **[17].**

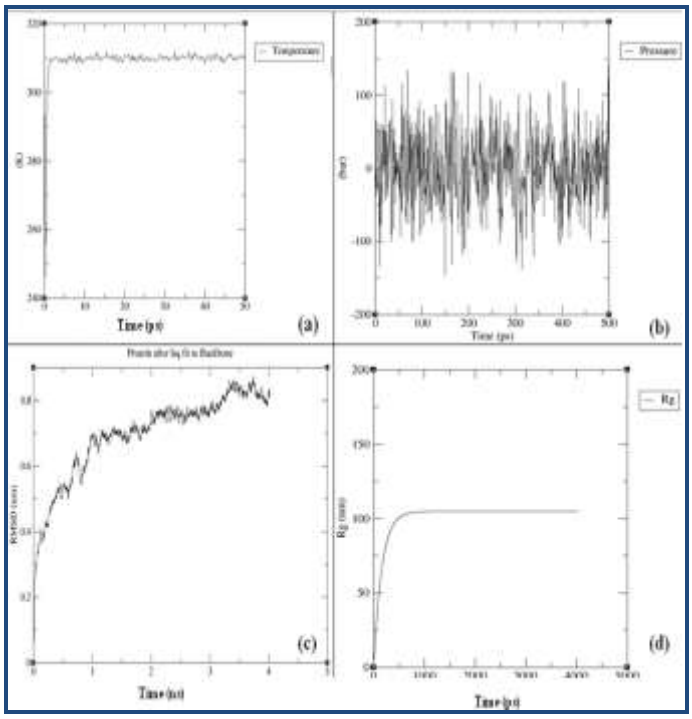

Figure 2: a) Plot representing the NVT ensemble (Temperature (k) vs Time (ps)) Image is showing that system attain 310K temperature during the initial run and remain stable during equilibration process; **b**) Plots depicts the NPT graph (pressure (bar) vs time (ps)) Figure is indicating that the pressure is fluctuating widely during the equilibration phase; **c)** Plot (rmsd (nm) vs time (ns)). The region at 0.8 nm indicating that system tends to be equilibrated in dynamic behaviour and can be used to analyze the molecular features; **d)** Plot showing the radius of Gyration (-Rg (nm) vs Time (ns)). Plot is representing the stability of protein over the course of given time

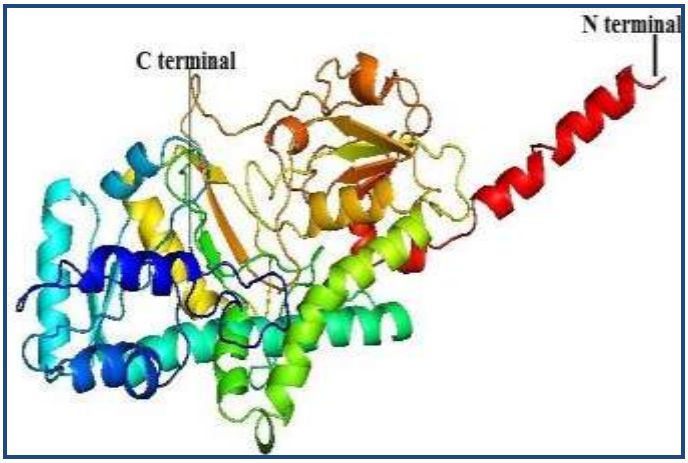

Figure 3: Pymol view of protein model achieved after molecular dynamic simulation. Figure is showing the number of helices and beta sheets after the complete run of simulation.

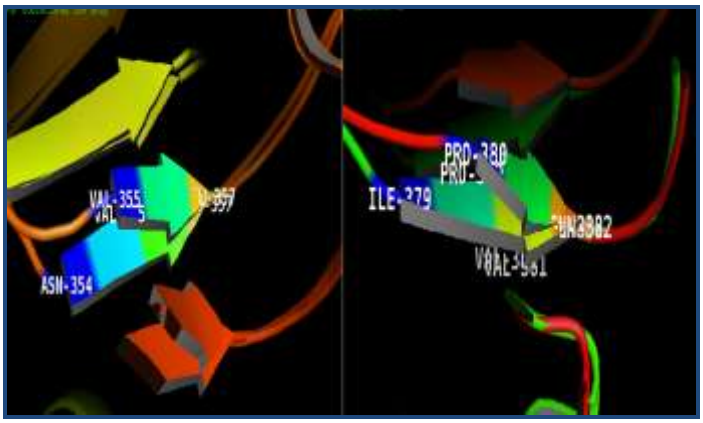

Figure 4: Visual comparison of the GFAP Models before and after Energy Minimization. Figure is showing marked differences after overlapping the structure before and after the energy minimization. Structure was more stable after the energy minimization.

Results & discussion:

In the present communication we have identified the 3D structure of GFAP that can be used to reveal its role in pathogenesis of RA in near future. Due to the absence of crystallographic structure of this protein, it is necessary to predict the 3D structure in order to understand the function of a particular protein. In case of GFAP, there are some already predicted structures in various servers and databases but due to lack of complete sequence coverage we have to predict the complete structure of this protein. Similarity search for the template identification have been shown in **Table 1 (see supplementary material).**

The search resulted that the GFAP protein sequence is similar to vimentin, lamin, keratin 5 and keratin 14 of human and also similar to the *Staphylococcus aueris*, *Francisella tularensis* and *Bacillus subtilis*. Phylogenetic tree was constructed using clustalW for global alignment followed by the proml of phylip which reveals that the vimentin and GFAP were probably sharing common ancestors. This envisages that these two proteins probably arose from genes that are paralogous. Although 3S4R, 3SSU, 3G1E, and 1GK7 share the common ancestors with GFAP, but the length aligned was very poor i.e. 90, 90, 36 and 37 amino acids respectively out of 432 amino acids. But it covers query length with maximum identity percentage of 59%, 60%, 70% and 71% respectively. Therefore, the query coverage of 3S4R, 3SSU, 3G1E, and 1GK7 when aligned to GFAP sequence is 21%, 21%, 8% and 8% respectively which is not sufficient for the purpose of modeling. Hence, 17 suitable templates for the present study have been selected based on the previously discussed criteria (**Figure 1).**

Model generation

A multiple sequence alignment (MSA) generated using a python script 'salign.py'. The MSA was converted into a sequence profile that lists the likelihood of the 20 standard amino acid residue types at every position in a given MSA. Alignments based on sequence profiles rather than single sequences have been shown to be significantly more accurate **[18].** The '*salign*' command of *salign.py* generated a matrix of pairwise alignment, as shown in the **Table 2 (see supplementary material)**, where the raw quality score of the multiple alignment was found to be 17.9. Quality Score (QS) is the average number of structurally equivalent residue pairs. It had RMS cut-off of 1.000. Two residues are considered to be equivalent when they have closer than RMS CUTOFF. There were 136 number of unique protein pairs. The multiple alignment file has been written in protein information resource (PIR) format.

Modeller generated 5 models on the basis of DOPE score that has been considered as the best scoring function **[19].** The structure showing highest stability with least score (-27480.207031) was selected for further analysis. The model generated may have steric hindrance or might possess side-chain bumps that affect the backbone folding. Therefore, stereochemistry was satisfied with energy minimization allowing repositioning of the amino acids in the 3D space solvating the model in water. The solvation was virtually executed in GROMACS by generating a cubic box that places the model at the center of the box (c) and at least 1.0 nm from the box edge (-d 1.0). Solute box specified with distance 1.0 nm means that there are at least 2.0 nm between periodic images of a protein. The whole system was found to possess a net charge of -13e. Since life does not exist at a net charge, we had to add ions to our system. Thus 13 positive Na^+ ions were added to neutralize the system.

The energy minimization was carried out using steepest descent method of Gromacs 4.5 package, applying Amber 03 Force Field **[15]** and the force applied was 1000kj/mol. This process is converged at 2593 steps and the energy was minimized to -3.2263365e+06. The DOPE score of this new model was found to be -32518.039063 indicating the more stable model compared to earlier model. The model was superimposed to determine the structural differences using standard method of computing which apparently resulted to be 0.817 that infer minor changes in positioning of the atoms because of the energy minimization. An attempt was also made to study the stability of the model in the biological environment. The whole process was studied in three steps:

NVT (constant Number of particles, Volume, and Temperature)

The system was equilibrated in NVT ensemble (constant Number of particles, Volume, and Temperature) where

temperature 310 K for 100-ps. The temperature coupling algorithm used for NVT simulation was *Berendsen-thermostat*. Temperature progression analysis can be depicted in graph. From the plot **(Figure 2a),** it is clear that the temperature of the system quickly reaches the target value (310 K), and remains stable during equilibration process.

NPT (constant Number of particles, Volume, and Pressure)

The resulted final model after NVT simulation was then subjected for NPT simulation. Equilibration of pressure is conducted under an NPT ensemble, wherein the Number of particles, Pressure, and Temperature are all constant. The pressure applied for the study was 1-bar**.** The pressure progression computed by Gromacs was stated as a graph plot **(Figure 2b).** When this plot was analyzed we found that the pressure value fluctuated widely over the course of 100-ps equilibration phase. Thus, by measuring the average of the running data we kept the pressure value as 1.05.

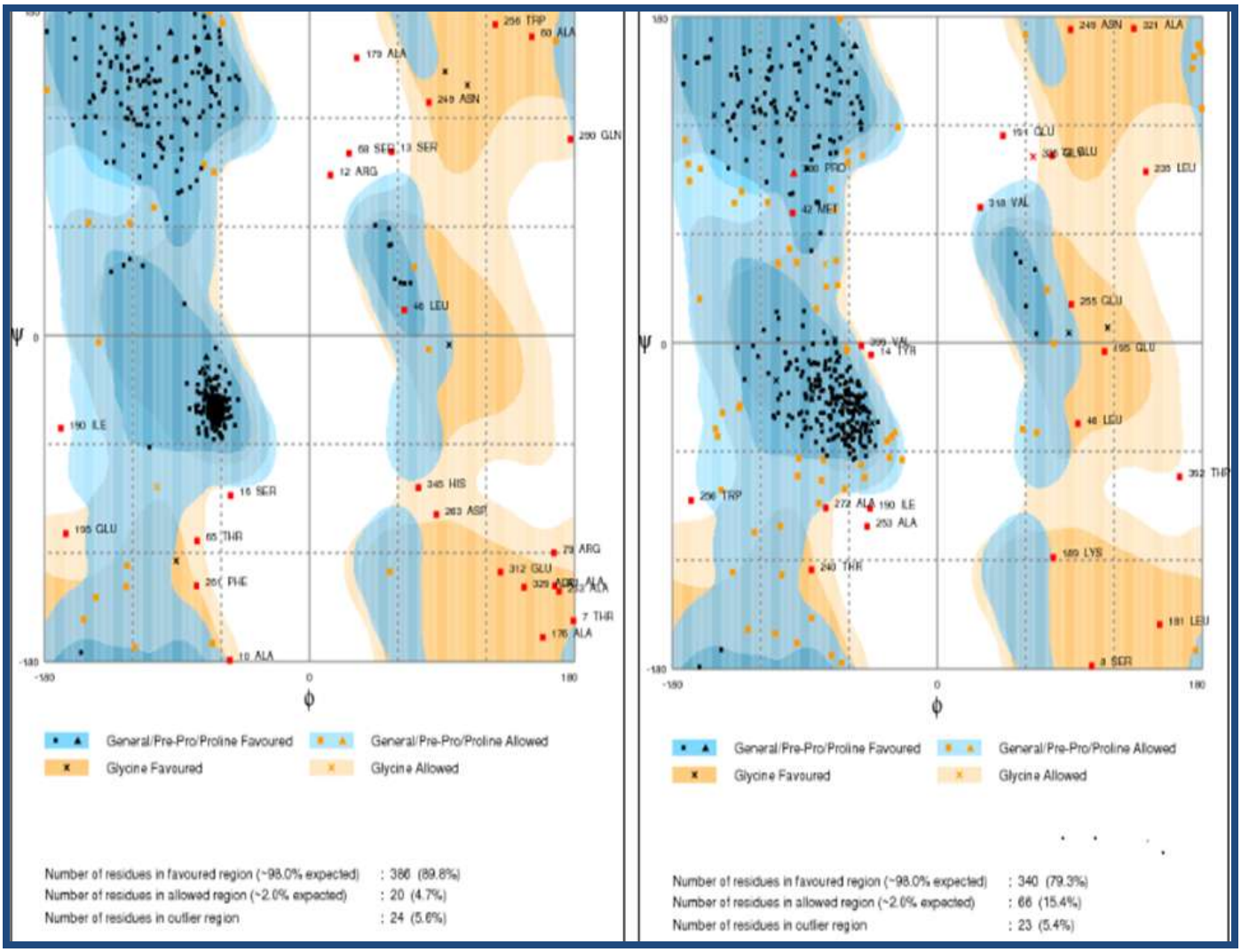

Figure 5: The Ramachandran plot of pre and post moleculardynamic structures respectively.

Production Molecular Dynamics

It is used as a post-processing tool to strip out coordinates, correct for periodicity, or manually alter the trajectory (time units, frame frequency, etc) for both the least squares fit and for RMSD calculation. The output plot shows that the RMSD relative to the structure present in the minimized form and system has been equilibrated. Further, we calculated RMSD relative to the crystal structure. The report generated in both the plots **(Figure 2c)** revealed that RMSD tends to levels off to ~0.8 nm (1 Å). This indicates that the structure tends to stabilize itself in due course of time. Subtle differences between the plots indicate that the structure at t = 0 nano second (ns) is slightly different from this crystal structure. This is to be expected, since it has been energy-minimized, and because the position restraints are not completely perfect. The analysis of the Rg for GFAP in our simulation: we can see from the reasonably invariant R_g values that the protein remains very stable, in its compact (folded) form over the course of 5 ns at 310 K in 1bar pressure **(Figure 2d).** This result was expected, that proves the stability of the model.

The generated models consists of 17 (seventeen) α-Helices (H), 9(nine) β-Sheets (B), 23 Loops (L) and 2(two) very short Turns (T). The N-terminal of the protein consists of a loop of 19 amino acids and the C-terminal is a long tail of alternate helices and loops **(Figure 3).** The overall shape of the protein seems to be

'∞' shape. Here the helix number 7 (seven) is supposed to be stabilizing the two parts of the protein. The movement of the protein is supposed to be controlled by seventh, eighth, ninth and tenth helices. The comparative visualization of the amino acids act on each and every position of superimposed structures was energy minimized **(Figure 4)** and subtle variations are recorded during the production MD for 5ns. These variation show that residues in alpha helices or beta sheets have low stability to remain in the earlier stretch form.

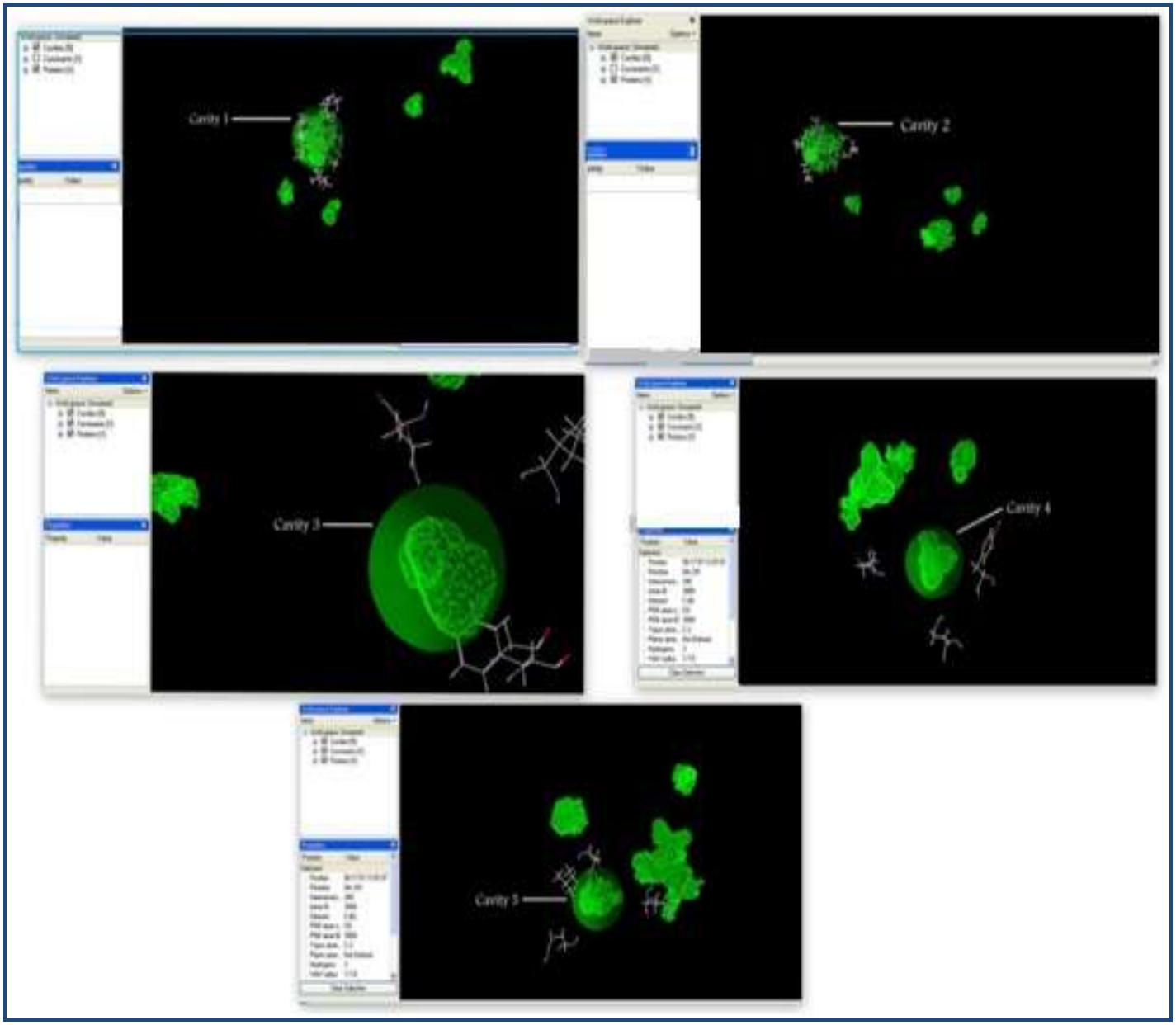

Figure 6: Cavities in the molecular surface of the protein are shown here in the chronological order starting cavity1 from the top-left.

However, Rampage gave a favorable result upon submission of both the protein structures. Rampage results of secondary protein models before and after molecular dynamics. The results obtained from the initial model were 386 residues in favoured region; 20 residues in allowed region and 24 residues in outlier regions. In contrast, the final model consist 340 residues in favored area; 66 residues in allowed region and 23 residues at outlier region **(Figure 5).** It signifies that after the first round molecular dynamic study of the model for 5ns it tend to form a stable state with less steric hindrance compared to the previous model. Thus, upon an increment in the duration, the model might be able to show a permanent stable confirmation. The probable ligand binding pockets were predicted using Molegro Virtual Docker (MVD). There are 5 pockets on the surface of the protein **(Figure 6).** Residues involved in the surface pockets have been listed in **Table 3 (see supplementary material)**. The first big obstacle to model this protein by *in-silico* approach is because of its homologous sequence with close similarity. Although it is similar to human Vimentin and Keratin it covers only to short extent. The instability index of such sequence is 52.74 which is too high. The generated model has no di-sulphide bonds among any residues. This long peptide chain contains only one cystine

residue. This molecular insight may be utilized for revealing the role of this protein in pathogenesis of RA.

Conclusion:

We report a molecular structural model of GFAP with a DOPE score of -27518.980469. Information related to its molecular cavity is documented that would help in further docking studies. The molecular model was subjected to molecular dynamics simulation over 5 ns and trajectories for its molecular properties monitored. The trajectory files towards the end of the time limit shows a tendency of stabilization of the plot. Energy minimization and molecular dynamic simulation was carried out by GROMACS. After the energy minimization, minor changes were observed. MD simulation revealed that protein remains stable in its folded form over the course of 5 ns at 310 K temperature and in 1 bar pressure. These results provide insights in understanding its molecular structural features towards drug discovery.

Acknowledgement:

We acknowledge Indian Council of Medical Research (ICMR) and Council of Scientific and Industrial Research (CSIR), Government of India, New Delhi, India for providing financial support to carry out the research work.

References:

[1] Eng LF, *J Neuroimmunol*. 1985 **8**: 203 [PMID: 2409105]
[2] Eng LF *et al. Brain Res*. 1971 **28**: 351 [PMID: 5113526]
[3] Sofroniew MV & Vinters HV, *Acta Neuropathol*. 2010 **119**: 7 [PMID: 20012068]
[4] Middeldorp J & Hol EM, *Prog Neurobiol*. 2011 **93**: 421 [PMID: 21219963]
[5] Biswas S *et al. PLoS One*. 2013 **8**: e56246 [PMID: 23418544]
[6] Jaroszewski L, *Methods Mol Biol*. 2009 **569**: 129 [PMID: 19623489]
[7] Kopp J & Schwede T, *Pharmacogenomics*. 2004 **5**: 405 [PMID: 15165176]
[8] Biswas S *et al. Scholars Research Library*. 2011 **2**: 40
[9] Eswar N *et al. Curr Protoc Protein Sci*. 2007 **2:** 2 [PMID: 18429317]
[10] Pronk S *et al. Bioinformatics*. 2013 **29:** 845 [PMID: 23407358]
[11] Mishra V & Siva Prasad CV, *Bioinformation* 2011 **7**: 46 [PMID: 21938204]
[12] Altschul SF *et al. Nucleic Acids Res*. 1997 **25:** 3389 [PMID: 9254694]
[13] Henikoff S & Henikoff JG, *Proc Natl Acad Sci U S A*. 1992 **89**: 10915 [PMID: 1438297]
[14] Kini RM & Evans HJ, *J Biomol Struct Dyn*. 1991 **9:** 475 [PMID: 1687724]
[15] Van Der Spoel D *et al. J Comput Chem*. 2005 **26**: 1701 [PMID: 16211538]
[16] Maiorov VN & Crippen GM, *J Mol Biol*. 1994 **235:** 625 [PMID: 8289285]
[17] Lovell SC *et al. Proteins* 2003 **50**: 437 [PMID: 12557186]
[18] Gribskov M *et al. Methods Enzymol*. 1990 **183**: 146 [PMID: 2314273]
[19] Shen MY & Sali A, *Protein Sci*. 2006 **15**: 2507 [PMID: 17075131]

Edited by P Kangueane

Citation: **Deka *et al.*** Bioinformation 11(8): 393-400 (2015)

BIOINFORMATION

open access

Supplementary material:

Table 1: Hit result of BLAST search for similar sequences to GFAP in PDB database

Sub_ID	%idty	aln_len	q. start	q. end	E-value	bit score
1GK4 | A	71.4	84	294	377	1.00E-34	125
3SSU | A	60.4	91	65	155	1.00E-30	114
3UF1 | A	61	105	111	215	1.00E-29	112
3S4R | A	59.3	91	65	155	3.00E-29	110
3KLT | A	62.5	72	229	300	3.00E-27	105
3SWK | A	62.8	86	119	204	5.00E-27	104
3TRT | A	58.7	75	227	301	6.00E-26	101
3TNU | B	42.6	129	245	373	6.00E-24	97.4
3TNU | A	40	130	244	373	3.00E-20	87.4
1GK7 | A	71.1	38	67	104	1.00E-10	58.2
1X8Y | A	42.4	85	293	377	1.00E-10	58.9
1GK6 | A	70.7	41	338	378	6.00E-10	56.6
3V4W | A	47.1	70	307	376	3.00E-09	55.1
3G1E | A	70.3	37	68	104	3.00E-09	54.3
3V58 | A	47.1	70	307	376	4.00E-09	54.7
3V4Q | A	47.1	70	307	376	5.00E-09	54.3
3TYY | A	43.9	82	299	377	6.00E-09	54.7
2XV5 | A	65.7	35	345	379	3.00E-07	48.9
3OJL | A	23.5	200	57	237	2.5	30.4
4F21 | A	19.1	131	67	193	6.7	28.9
1Y23 | A	30	70	177	246	9.6	28.1

Sub_ID = Subject ID, %idty = Percentage of Identity, % +ve = percentage of Positives, aln_len = alignment length, q. start= position of the starting residue of the query, q. end= position of the ending residue of the query

Table 2: Matrix of pairwise equivalences

PDB ID →		1GK4	1GK6	1GK7	1X8Y	2XV5	3G1E	3KLT	3S4R	3SSU	3SWK	3TNU	3TRT	3TYY	3UF1	3V4Q	3V4W	3V58
↓	CHAIN	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
1GK4	A	0	19	27	45	5	32	5	33	31	17	18	12	42	9	42	34	0
1GK6	A		0	9	28	20	23	6	12	14	10	6	6	29	17	24	27	0
1GK7	A			0	14	13	28	21	8	24	27	34	37	22	20	14	14	1
1X8Y	A				0	9	29	7	29	29	12	10	10	39	7	68	66	0
2XV5	A					0	16	9	5	11	4	6	5	9	24	8	6	1
3G1E	A						0	33	23	37	36	36	22	31	24	30	27	1
3KLT	A							0	8	36	8	29	10	5	3	7	9	1
3S4R	A								0	24	9	22	13	22	12	32	30	0
3SSU	A									0	32	39	25	32	9	32	31	0
3SWK	A										0	2	3	17	18	11	10	0
3TNU	A											0	38	14	5	12	13	0
3TRT	A												0	9	3	10	10	1
3TYY	A													0	7	44	44	0
3UF1	A														0	6	6	1
3V4Q	A															0	67	0

3V4W	A	0	0
3V58	A		0

Table 3: Detailed information of 5 molecular Surface Cavities

Cavity Number	Co-ordinates (X,Y,Z)	Volume (Å³)	Radius (Å)	Residues surrounding the cavity
1	59.50, 51.20, 72.90	291.84	8.11	A (145), Q (146), A (149), L (155), R (173), H (221), V (226), K (228), P (229), K (356), V (381), T (383), E (401), L (404)
2	57.98, 81.35, 35.84	115.648	7.11	M (21), G (24), L (31), T (35), L (37), S (53), A (57), K (63), R (66)
3	71.65, 60.72, 83.84	74.752	4.11	Y (349), I (378), Q (382)
4	70.01, 63.56, 46.24	58.368	4.11	Q (93), Y (116), Q (129)
5	60.72, 52.51, 90.12	49.664	4.11	Q (241), I (377), I (379), I (408)

Å= Angstrom, the one letter codes represent the usual amino acids.