EVOLUTIONARY CONSERVATION ANALYSIS OF AMINO ACIDS OF HUMAN SERUM ALBUMIN USING CONSURF DATABASE

[1]**Sanghapal S. Kshirasagar**, [2]Vishwas S. Shembekar
[1]Assistant Professor, Dept. of Biotechnology, [2]Professor and Head, Dept. of Zoology and Fishery Science
[1]Department of Biotechnology,
[1]Rajarshi Shahu Mahavidyalaya,
(Autonomous), Latur (413512), (State) Maharashtra, India

Abstract: ConSurf database is used for evolutionary conservation analysis of amino acids of given proteins with known structures in the Protein Data Bank (PDB). In this study we have used PDB ID: 1AO6, which is crystallographic structure of human serum albumin. The evolutionary conservation of each amino acid position in the alignment was calculated by evolutionary profiles of most proteins of known structure. ConSurf-DB was constructed based on the fully automated four steps process. The conservation grades are mapped on the three-dimensional structure of the query protein, which can be viewed using the NGL viewer or FirstGlance in Jmol. Finally, the evolutionary rates are categorized to discrete conservation grades, ranging from 1 to 9.

Keywords: ConSurf database, amino acid conservation score, 1AO6, human serum albumin

I. INTRODUCTION

ConSurf-DB includes pre-calculated estimates of the evolutionary profiles of most proteins of known structure. The evolutionary conservation estimates are highly robust because particularly stringent thresholds were used in constructing the database. A detailed description of regarding use of consurf databes, to study pre calculated evolutionary conservation profiles of protein structure is available in a publication (8).

ConSurf-DB was constructed based on the fully automated four steps process (a) downloading and parsing non-redundant PDB entries, (b) collecting sequence homologues and building the multiple sequence alignment (MSA), (c) calculating evolutionary rates, and (d) formatting the results for presentation in the ConSurf-DB website.

II. MATERIAL AND METHODS

The first step in building ConSurf-DB is retrieving the PDB entries. In this study we have used 1AO6, which is a crystal structure of human serum albumin (2). Each PDB entry can contain one or more protein chains, which are handled separately in ConSurf-DB. The chains are extracted from a PISCES file (17), which contains all non-redundant (unique) chains in the PDB.

The second step is detecting homologues. The sequence homologues are searched in UniRef-90 (14), a clustered version of the UniProt database (16). This is done using one iteration of HMMER (8) with an E-value threshold of 0.0001. Next, CD-HIT (3) removes any redundant homologues with a threshold of 95%. Finally, a multiple sequence alignment (MSA) of the homologues is constructed using the MAFFT-LINSi procedure (4).

The third step is estimating the evolutionary rates. It begins by inferring the best amino acids substitution model based on the MSA (1). Then, a phylogenetic tree is built from the MSA with the Neighbor-Joining method (12), implemented in the Rate4Site program (10). Next, Rate4Site assigns an evolutionary rate to each position in the query sequence, based on the phylogenetic tree and the substitution model, and using an empirical Bayesian methodology (6).

The fourth and final step is formatting and visualizing the results. The conservation grades are mapped on the three-dimensional structure of the query protein, which can be viewed using the NGL viewer (11) or FirstGlance in Jmol (5). The colors are also projected on the query sequence and on the MSA. Moreover, session files, presenting the protein structure colored by the conservation grades, are created using the PyMOL (13) and Chimera (9) programs. A flowchart of the pipeline used to construct ConSurf-DB is shown in figure 2.

III. RESULTS AND DISCUSSION

ConSurf-Database analysis for PDB ID: 1AO6 was done using calculations based on protein 1N5U chain A, which is 100% sequence identical to 1AO6 (17). For all the results following color codes were used, where 1 represents most variable positions and 9 represents most conserved positions as shown in figure 1.

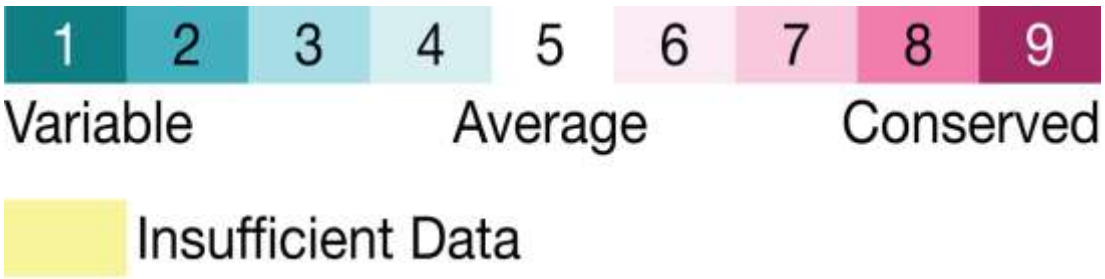

Figure 1: Color scale by conservation score.

NGL 3D viewer was used to analyze ConSurf-DB output and following results were obtained (Figure 3 to figure 7). From these images we can see that chain A and chain B of 1AO6 (human serum albumin) can be visualized by NGL viewer. We can study structural conformations and structural changes in human serum albmin by using different prameters provided by ConSurf-DB.

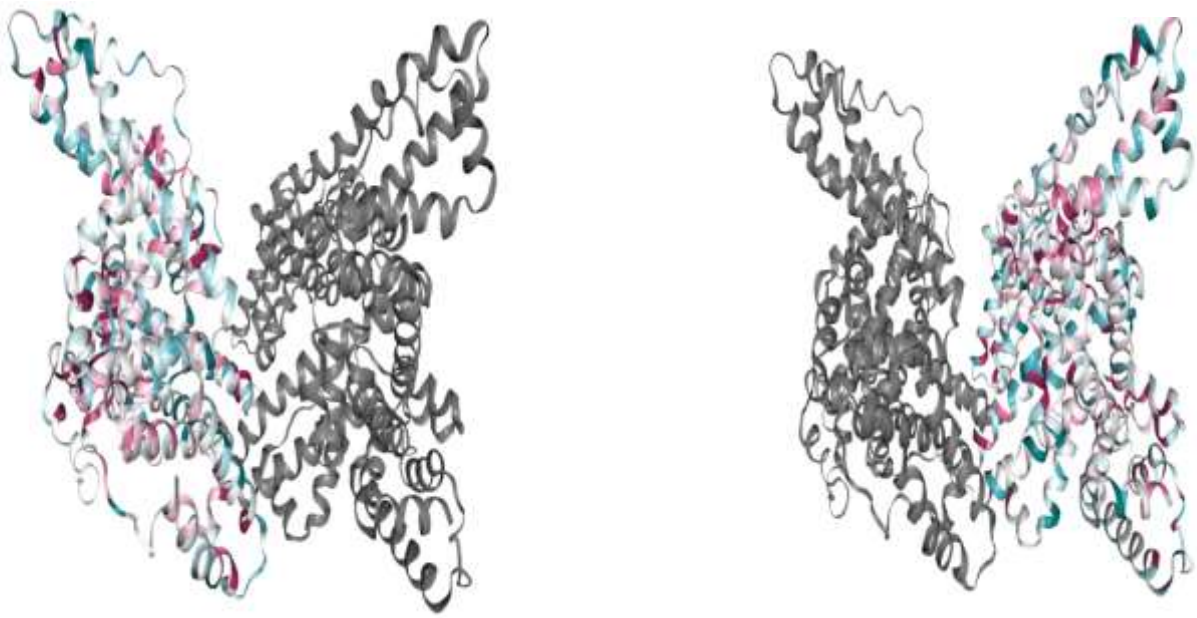

Figure 3: Cartoon style human serum albumin chain A (left) and chain B (right) with NGL 3D viewer

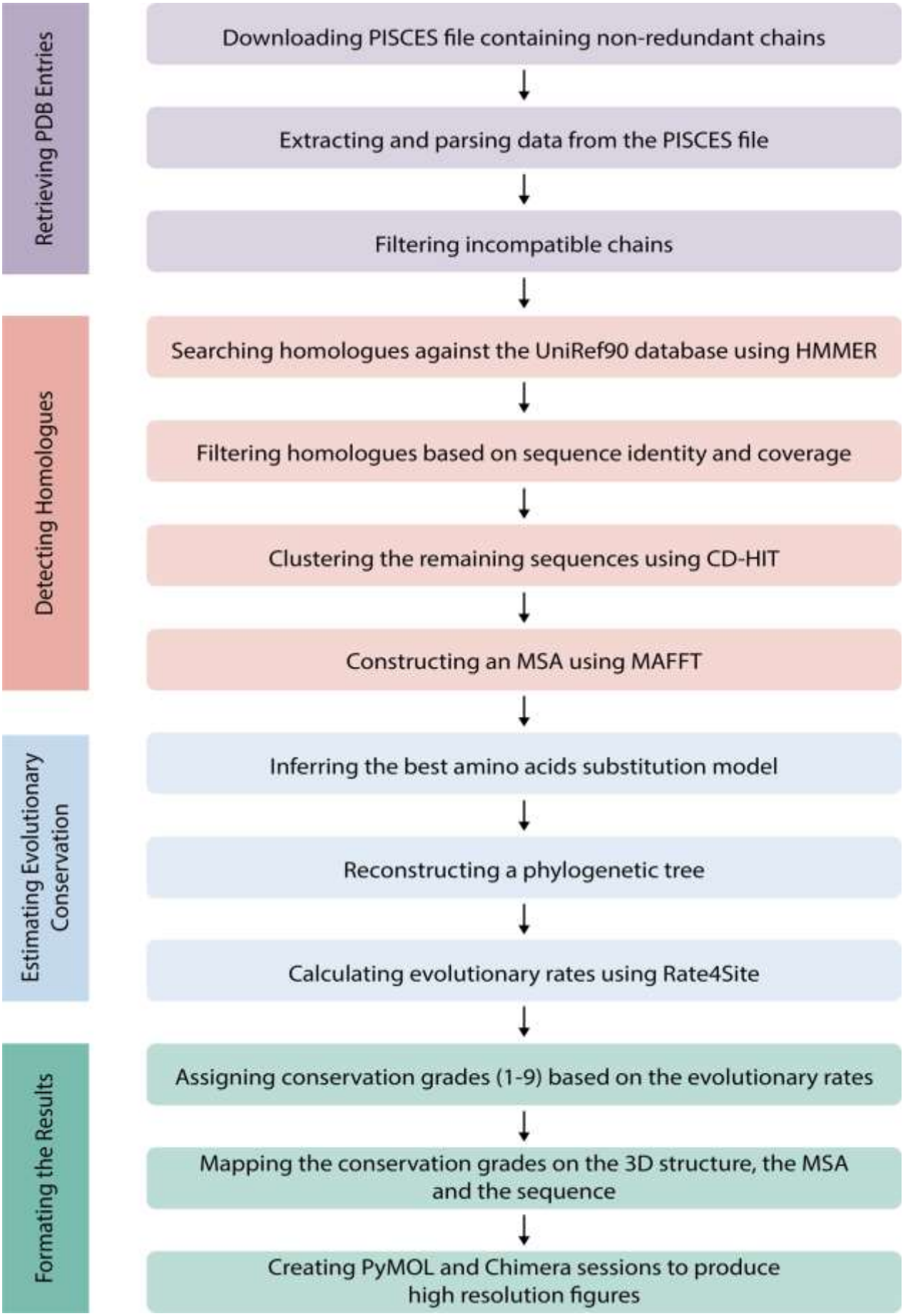

Figure 2: A flowchart of the pipeline used to construct ConSurf-DB

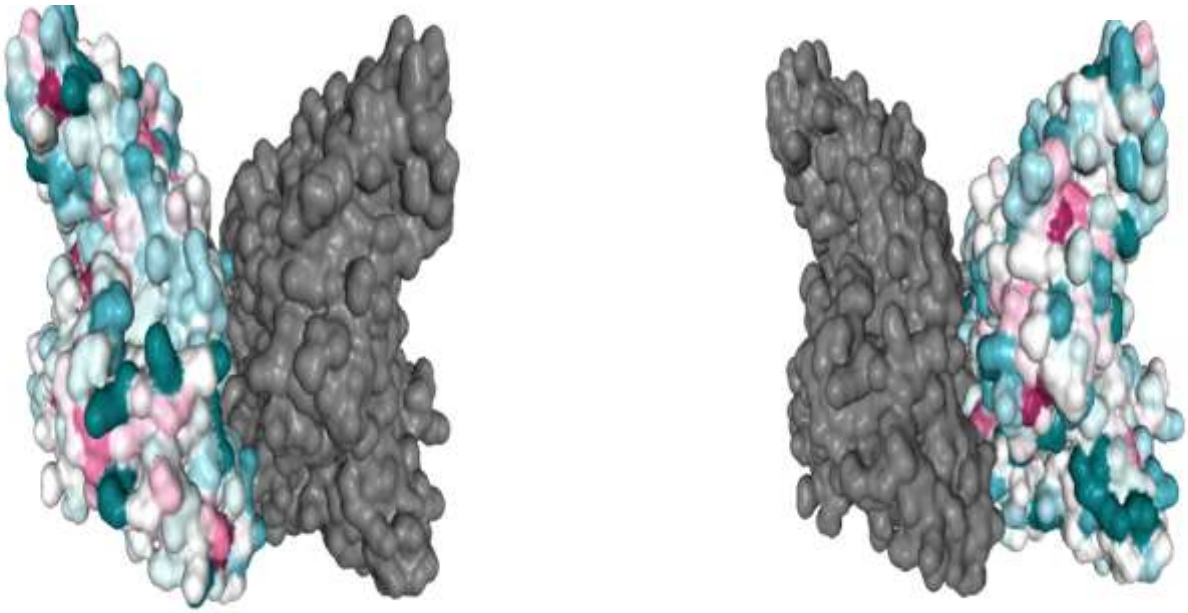

Figure 4: Surface style human serum albumin chain A (left) and chain B (right) with NGL 3D viewer

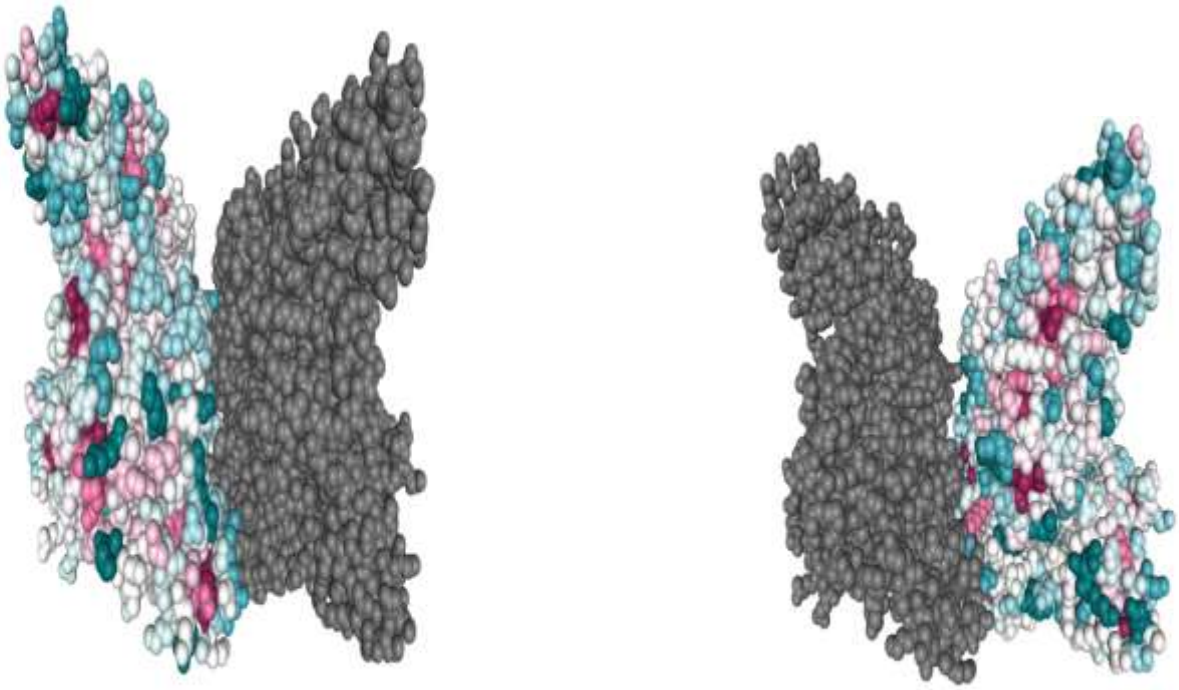

Figure 5: Spacefill style human serum albumin chain A (left) and chain B (right) with NGL 3D viewer

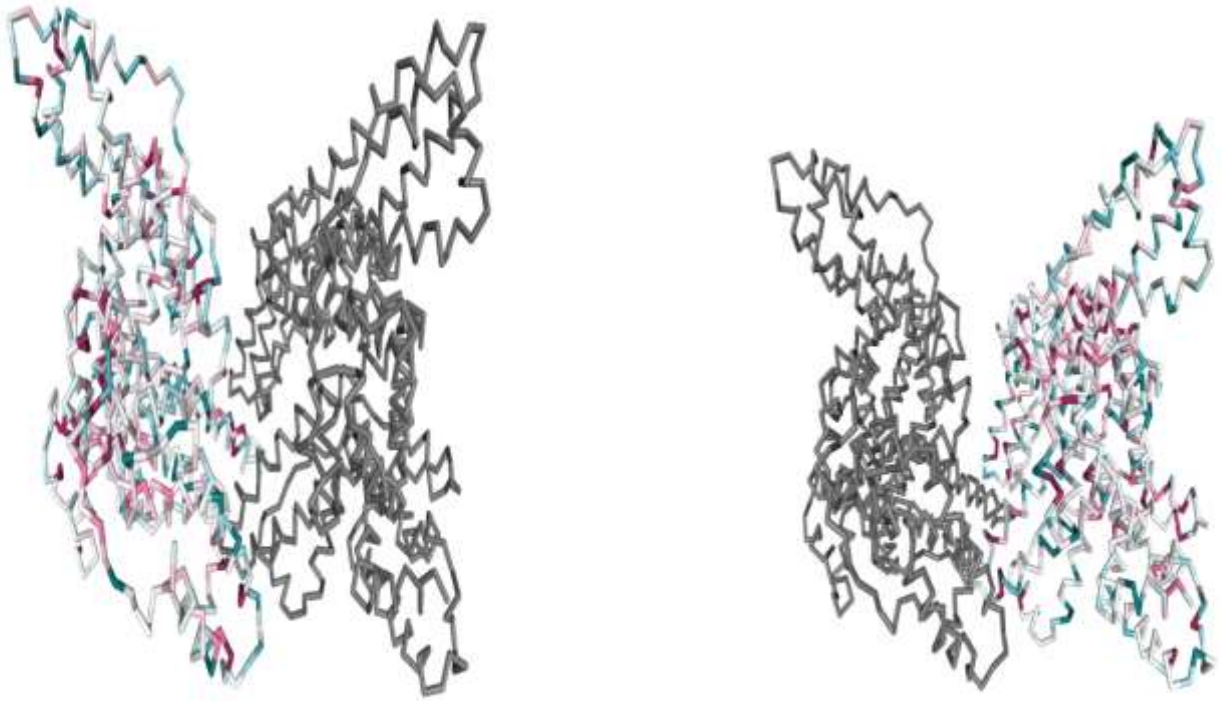

Figure 6: Backbone style human serum albumin chain A (left) and chain B (right) with NGL 3D viewer

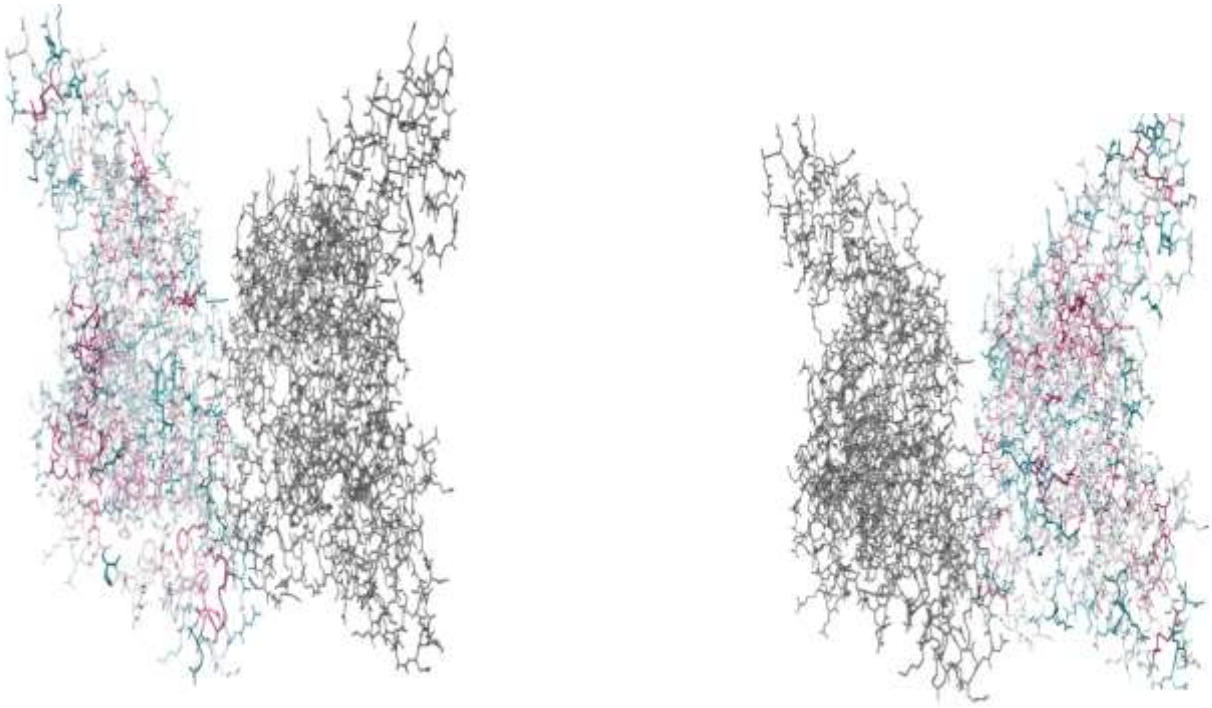

Figure 7: Licorice style human serum albumin chain A (left) and chain B (right) with NGL 3D viewer

```
1           11          21          31          41
DAHKSEVAHR  FKDLGEENFK  ALVLIAFAQY  LQQCPFEDHV  KLVNEVTEFA

51          61          71          81          91
KTCVADESAE  NCDKSLHTLF  GDKLCTVATL  RETYGEMADC  CAKQEPERNE

101         111         121         131         141
CFLQHKDDNP  NLPRLVRPEV  DVMCTAFHDN  EETFLKKYLY  EIARRHPYFY

151         161         171         181         191
APELLFFAKR  YKAAFTECCQ  AADKAACLLP  KLDELRDEGK  ASSAKQRLKC

201         211         221         231         241
ASLQKFGERA  FKAWAVARLS  QRFPKAEFAE  VSKLVTDLTK  VHTECCHGDL

251         261         271         281         291
LECADDRADL  AKYICENQDS  ISSKLKECCE  KPLLEKSHCI  AEVENDEMPA

301         311         321         331         341
DLPSLAADFV  ESKDVCKNYA  EAKDVFLGMF  LYEYARRHPD  YSVVLLLRLA

351         361         371         381         391
KTYETTLEKC  CAAADPHECY  AKVFDEFKPL  VEEPQNLIKQ  NCELFEQLGE

401         411         421         431         441
YKFQNALLVR  YTKKVPQVST  PTLVEVSRNL  GKVGSKCCKH  PEAKRMPCAE

451         461         471         481         491
DYLSVVLNQL  CVLHEKTPVS  DRVTKCCTES  LVNRRPCFSA  LEVDETYVPK

501         511         521         531         541
EFNAETFTFH  ADICTLSEKE  RQIKKQTALV  ELVKHKPKAT  KEQLKAVMDD

551         561         571         581
FAAFVEKCCK  ADDKETCFAE  EGKKLVAASQ  AALGL
```

Figure 8: Sequence of human serum albumin (coloured as per conservation scale)

Table 1: Conservation scores of amino acids with their sequence position and 1 letter codes

Position	Amino Acid	COLOR	Position	Amino Acid	COLOR	Position	Amino Acid	COLOR
1	D	8	41	K	5	81	R	5
2	A	6	42	L	6	82	E	4
3	H	7	43	V	6	83	T	5
4	K	6	44	N	2	84	Y	6
5	S	4	45	E	4	85	G	5
6	E	8	46	V	6	86	E	4
7	V	8	47	T	5	87	M	5
8	A	7	48	E	4	88	A	6
9	H	4	49	F	5	89	D	5
10	R	6	50	A	6	90	C	9
11	F	2	51	K	5	91	C	9
12	K	1	52	T	3	92	A	3
13	D	1	53	C	9	93	K	4
14	L	5	54	V	5	94	Q	3
15	G	6	55	A	5	95	E	6
16	E	5	56	D	3	96	P	4
17	E	3	57	E	1	97	E	6
18	N	5	58	S	5	98	R	9
19	F	6	59	A	3	99	N	6
20	K	2	60	E	5	100	E	3
21	A	3	61	N	3	101	C	9
22	L	4	62	C	9	102	F	8
23	V	5	63	D	2	103	L	7
24	L	5	64	K	6	104	Q	5
25	I	6	65	S	3	105	H	7
26	A	4	66	L	4	106	K	8
27	F	5	67	H	4	107	D	5
28	A	8	68	T	4	108	D	4
29	Q	9	69	L	5	109	N	4
30	Y	6	70	F	7	110	P	4
31	L	6	71	G	7	111	N	2
32	Q	7	72	D	5	112	L	3
33	Q	5	73	K	5	113	P	6
34	C	6	74	L	8	114	R	4
35	P	7	75	C	9	115	L	4
36	F	5	76	T	5	116	V	3
37	E	4	77	V	4	117	R	4
38	D	6	78	A	4	118	P	5
39	H	5	79	T	5	119	E	5
40	V	3	80	L	6	120	V	5

Position	Amino Acid	COLOR	Position	Amino Acid	COLOR	Position	Amino Acid	COLOR
121	D	6	161	Y	7	201	A	5
122	V	5	162	K	5	202	S	6
123	M	4	163	A	5	203	L	2
124	C	9	164	A	6	204	Q	1
125	T	5	165	F	5	205	K	5
126	A	5	166	T	2	206	F	7
127	F	5	167	E	2	207	G	7
128	H	4	168	C	9	208	E	4
129	D	5	169	C	9	209	R	4
130	N	5	170	Q	3	210	A	4
131	E	3	171	A	5	211	F	4
132	E	1	172	A	3	212	K	4
133	T	1	173	D	4	213	A	7
134	F	4	174	K	4	214	W	5
135	L	5	175	A	4	215	A	5
136	K	3	176	A	3	216	V	3
137	K	2	177	C	9	217	A	5
138	Y	5	178	L	7	218	R	4
139	L	4	179	L	3	219	L	3
140	Y	6	180	P	4	220	S	8
141	E	7	181	K	8	221	Q	9
142	I	5	182	L	4	222	R	7
143	A	8	183	D	2	223	F	5
144	R	8	184	E	2	224	P	8
145	R	8	185	L	5	225	K	5
146	H	7	186	R	3	226	A	8
147	P	7	187	D	5	227	E	4
148	Y	5	188	E	2	228	F	7
149	F	6	189	G	2	229	A	2
150	Y	6	190	K	4	230	E	6
151	A	6	191	A	1	231	V	7
152	P	6	192	S	3	232	S	1
153	E	6	193	S	7	233	K	6
154	L	6	194	A	2	234	L	5
155	L	6	195	K	2	235	V	5
156	F	1	196	Q	7	236	T	1
157	F	3	197	R	6	237	D	6
158	A	7	198	L	6	238	L	3
159	K	1	199	K	2	239	T	5
160	R	1	200	C	9	240	K	5

Position	Amino Acid	COLOR	Position	Amino Acid	COLOR	Position	Amino Acid	COLOR
241	V	2	281	K	4	321	E	4
242	H	6	282	P	6	322	A	3
243	T	4	283	L	1	323	K	5
244	E	4	284	L	4	324	D	4
245	C	8	285	E	6	325	V	1
246	C	9	286	K	6	326	F	5
247	H	3	287	S	5	327	L	3
248	G	9	288	H	2	328	G	5
249	D	7	289	C	9	329	M	4
250	L	5	290	I	6	330	F	7
251	L	4	291	A	4	331	L	3
252	E	5	292	E	3	332	Y	6
253	C	8	293	V	4	333	E	8
254	A	6	294	E	5	334	Y	6
255	D	5	295	N	7	335	A	7
256	D	4	296	D	8	336	R	9
257	R	7	297	E	5	337	R	8
258	A	5	298	M	2	338	H	8
259	D	4	299	P	6	339	P	5
260	L	4	300	A	3	340	D	5
261	A	5	301	D	3	341	Y	4
262	K	2	302	L	7	342	S	9
263	Y	6	303	P	7	343	V	4
264	I	4	304	S	2	344	V	5
265	C	9	305	L	2	345	L	6
266	E	5	306	A	5	346	L	7
267	N	4	307	A	3	347	L	7
268	Q	6	308	D	4	348	R	8
269	D	5	309	F	5	349	L	5
270	S	2	310	V	7	350	A	3
271	I	6	311	E	5	351	K	3
272	S	9	312	S	5	352	T	2
273	S	7	313	K	2	353	Y	7
274	K	7	314	D	4	354	E	3
275	L	7	315	V	6	355	T	2
276	K	2	316	C	8	356	T	3
277	E	1	317	K	3	357	L	7
278	C	9	318	N	2	358	E	1
279	C	9	319	Y	3	359	K	5
280	E	3	320	A	4	360	C	9

Position	Amino Acid	COLOR	Position	Amino Acid	COLOR	Position	Amino Acid	COLOR
361	C	8	401	Y	4	441	P	5
362	A	2	402	K	1	442	E	5
363	A	3	403	F	7	443	A	5
364	A	4	404	Q	7	444	K	3
365	D	7	405	N	7	445	R	5
366	P	3	406	A	3	446	M	6
367	H	1	407	L	3	447	P	4
368	E	3	408	L	3	448	C	9
369	C	9	409	V	4	449	A	5
370	Y	4	410	R	4	450	E	8
371	A	1	411	Y	6	451	D	4
372	K	1	412	T	7	452	Y	3
373	V	6	413	K	5	453	L	6
374	F	4	414	K	6	454	S	6
375	D	4	415	V	6	455	V	3
376	E	4	416	P	9	456	V	6
377	F	4	417	Q	7	457	L	6
378	K	3	418	V	6	458	N	7
379	P	4	419	S	6	459	Q	3
380	L	6	420	T	5	460	L	6
381	V	6	421	P	2	461	C	9
382	E	2	422	T	5	462	V	4
383	E	6	423	L	7	463	L	6
384	P	7	424	V	3	464	H	7
385	Q	3	425	E	2	465	E	3
386	N	3	426	V	3	466	K	3
387	L	3	427	S	6	467	T	4
388	I	7	428	R	3	468	P	6
389	K	6	429	N	2	469	V	7
390	Q	3	430	L	6	470	S	9
391	N	6	431	G	6	471	D	4
392	C	9	432	K	1	472	R	4
393	E	3	433	V	6	473	V	8
394	L	3	434	G	7	474	T	4
395	F	4	435	S	4	475	K	5
396	E	4	436	K	4	476	C	9
397	Q	3	437	C	9	477	C	9
398	L	5	438	C	9	478	T	4
399	G	7	439	K	2	479	E	4
400	E	4	440	H	3	480	S	9

Position	Amino Acid	COLOR	Position	Amino Acid	COLOR	Position	Amino Acid	COLOR
481	L	6	521	R	4	561	A	5
482	V	8	522	Q	7	562	D	2
483	N	3	523	I	1	563	D	3
484	R	9	524	K	5	564	K	5
485	R	7	525	K	7	565	E	4
486	P	4	526	Q	6	566	T	4
487	C	9	527	T	4	567	C	9
488	F	8	528	A	6	568	F	8
489	S	6	529	L	8	569	A	2
490	A	4	530	V	7	570	E	1
491	L	7	531	E	5	571	E	8
492	E	2	532	L	7	572	G	3
493	V	2	533	V	8	573	K	4
494	D	9	534	K	8	574	K	5
495	E	4	535	H	6	575	L	7
496	T	6	536	K	7	576	V	7
497	Y	8	537	P	7	577	A	4
498	V	6	538	K	1	578	A	4
499	P	8	539	A	6	579	S	7
500	K	7	540	T	6	580	Q	5
501	E	1	541	K	1	581	A	7
502	F	5	542	E	5	582	A	4
503	N	6	543	Q	7	583	L	8
504	A	5	544	L	6	584	G	7
505	E	4	545	K	3	585	L	7
506	T	5	546	A	4			
507	F	7	547	V	5			
508	T	4	548	M	3			
509	F	6	549	D	1			
510	H	4	550	D	3			
511	A	4	551	F	8			
512	D	6	552	A	3			
513	I	7	553	A	1			
514	C	9	554	F	6			
515	T	6	555	V	6			
516	L	5	556	E	3			
517	S	4	557	K	4			
518	E	4	558	C	9			
519	K	3	559	C	9			
520	E	5	560	K	2			

Hence, with respect to color values, it can be said that amino acids which are in variable regions have high chances of getting interacted with other molecules. It is particularly important because, the amino acids which are in conserved regions are generally present in the interior core of the protein. Similarly, colored Sequence of human serum albumin is obtained by using Consurf databse as shown in figure 8. And conservation scores of amino acids with their sequence position and 1 letter codes are shown in Table 1.

From this it can be said that amino acids in variable region positions are likely to undergo conformation change and are generally present in accessible area for other ligands or small molecules. Determination of conserved and variable regions of protein can be useful to study interaction of different ligands with protein. This is particularly important because each protein has different ligand binding sites. Similarly, to study the possible interactions of amino acids with other small molecules amino acid conservation score can be utilized.

IV. ACKNOWLEDGMENT

Authors are thankful to Principal, Rajarshi Shahu Mahavidyalaya (Autonomous), Latur for providing laboratory facilities.

REFERENCES:

1. Darriba D., Taboada G.L., Doallo R. and Posada D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27:1164-1165

2. Crystal structure of human serum albumin at 2.5 A resolution. Sugio, S., Kashima, A., Mochizuki, S., Noda, M., Kobayashi, K. (1999) Protein Eng. 12: 439-446

3. Fu L., Niu B., Zhu Z., Wu S. and Li W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28:3150-3152

4. Katoh K. and Standley D.M. (2013), MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30:772–780

5. Martz E. (2005). FirstGlance in Jmol.

6. Mayrose I., Graur D., Ben-Tal N. and Pupko T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Molecular Biology and Evolution*, 21:1781-1791

7. Mistry J., Finn R.D., Eddy S.R., Bateman A. and Punta M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, 41:e121

8. Ofir Goldenberg, Elana Erez, Guy Nimrod, and Nir Ben-Tal (2009). The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures Nucleic Acids Res; 37(Database issue): D323–D327.

9. Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C. and Ferrin T.E. (2004). UCSF Chimera a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25:1605-1612

10. Pupko T., Bell R.E., Mayrose I., Glaser F. and Ben-Tal N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18:S71-S77.

11. Rose A.S., Bradley A.R., Valasatava Y., Duarte J.M., Prlic A. and Rose P.W. (2018). NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, 34:3755-3758.

12. Saitou N. and Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406-425.

13. Schrödinger L.L.C. (2015). The PyMOL Molecular Graphics System, Version 2.3.3.

14. Suzek B.E., Wang Y., Huang H., McGarvey P.B., Wu C.H. and UniProt Consortium. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31:926-932

15. The Atomic Structure of Human Methemalbumin at 1.9 A Wardell, M., Wang, Z., Ho, J.X., Robert, J., Ruker, F., Ruble, J., Carter, D.C. (2002) Biochem.Biophys.Res.Commun. 291: 813-819

16. UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47:D506-D515

17. Wang G. and Dunbrack R.L. (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, 33:W94-W98