J. Chem. Sci. Vol. 124, No. 1, January 2012, pp. 83–91. © Indian Academy of Sciences.

Bhageerath—Targeting the near impossible: Pushing the frontiers of atomic models for protein tertiary structure prediction[#]

B JAYARAM[a,b,c,*], PRIYANKA DHINGRA[a,b], BHARAT LAKHANI[b]
and SHASHANK SHEKHAR[b]

[a]Department of Chemistry, [b]Supercomputing Facility for Bioinformatics and Computational Biology,
[c]School of Biological Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India
e-mail: bjayaram@chemistry.iitd.ac.in

Abstract. Protein folding, considered to be the holy grail of molecular biology, remains intractable even after six decades since the report of the first crystal structure. Over 70,000 X-ray and NMR structures are now available in protein structural repositories and no physico-chemical solution is in sight. Molecular simulation methodologies have evolved to a stage to provide a computational solution to the tertiary structures of small proteins. Knowledge base driven methodologies are maturing in predicting the tertiary structures of query sequences which share high similarities with sequences of known structures in the databases. The void region thus seems to be medium (>100 amino acid residues) to large proteins with no sequence homologs in the databases and hence which has become a fertile ground for the genesis of hybrid models which exploit local similarities together with *ab initio* models to arrive at reasonable predictions. We describe here the development of *Bhageerath* an *ab initio* model and *Bhageerath*-H a hybrid model and present a critique on the current status of prediction of protein tertiary structures.

Keywords. *Ab initio*; protein folding; molecular dynamics simulation; protein structure prediction; *Bhageerath*; critical assessment of protein structure prediction (CASP).

1. Introduction

Protein folding, considered to be a challenging task[1–3] remains unsolved for the last six decades. It is classified as an NP complete or NP hard problem.[4,5] This notwithstanding, the dire need for tertiary structures of proteins in drug discovery and other areas[6–8] has propelled the development of a multitude of computational recipes. In this article, we focus on *ab initio*/*de novo* strategies, *Bhageerath* in particular, for protein tertiary structure prediction.

The *ab initio* term in the context of protein structure prediction is used to signify the usage of physics based all atom molecular mechanics potentials for predicting the three-dimensional structure of a protein. Molecular dynamics, Monte Carlo simulations and their variants are pooled under this category. Molecular dynamics simulations, in particular, have provided extremely high resolution spatial and temporal data, enhancing our knowledge and understanding of the protein folding mechanism. Simulations of biologically relevant processes, with atomistic accuracy on timescales beyond microsecond are now possible due to advances in software and hardware.[9]

In the year 1975, Levitt and Warshel simulated the folding of bovine pancreatic trypsin inhibitor (BPTI) using a simple representation of protein conformation, energy minimization and thermalization. They succeeded in 'renaturing' the protein from an open fully extended conformation to a folded native like conformation.[10] Later in 1977, McCammon, Gelin and Karplus studied for the first time the dynamics of folded BPTI in vacuum at a molecular level over a period of 9.2 picoseconds.[11] Since then extensive research has been carried out in the area of protein folding, unfolding, dynamics and structure. Table 1, summarizes some major milestones capturing the advancements. In the year 1995, Li and Daggett studied the structure and dynamics of native chymotrypsin inhibitor 2 with explicit water from a 5.3 ns simulation.[12] Insights were gained from 550 ps unfolding simulations of reduced BPTI at high temperature.[13] Folding simulation of small peptide fragments such as *beta*-turn in a short linear peptide and β-heptapeptides in aqueous solution were performed for 20 ns and 50 ns.[18,19] Using a Cray T3E, a massively parallel supercomputer consisting hundreds of CPUs, Duan and Kollman

[#]Dedicated to Prof. N Sathyamurthy on his 60th birthday
[*]For correspondence

Table 1. Increasing length of simulations with advances in computing resources.

Sl. No.	System	Length of the simulation	Year
1	Bovine Pancreatic Trypsin Inhibitor (BPTI)[11]	9.2 ps	1977
2	Bovine Pancreatic Trypsin Inhibitor (BPTI)[14,15]	60 ps,132 ps	1983
3	Bovine Pancreatic Trypsin Inhibitor (BPTI)[13]	550 ps	1992
4	Apomyoglobin[16]	350 ps,500 ps	1993
5	Chymotrypsin inhinitor 2 (CI2)[12]	5.3 ns	1995
6	Staphylococcal protein[17]	>9 ns	1995
7	Pentapeptide *cis*-AYPYD[18]	20 ns	1997
8	$\beta-$Heptapeptide[19]	50 ns	1998
9	Villin headpiece[20]	1 μs	1998
10	Engrailed Homeodomain[21]	40 ns,70 ns	2000
11	Protein G[22]	38 µs	2001
12	Trp cage[23]	50 ns	2002
13	Trp cage[24]	$\sim$100 µs	2002
14	Human Pin1 WW domain mutant, FiP35[25,26]	10 µs	2009
15	NTL9[27]	1.52 ms	2010
16	FiP35, Villin headpiece[28] Bovine Pancreatic Trypsin Inhibitor (BPTI)	100 µs 1 ms	2010

in 1998 reported one of the longest simulations of that time for a protein in water. They simulated villin headpiece subdomain (HP-36) for $\sim$1 µs with $\sim$3000 water molecules.[20] The growing track record of protein folding simulations in a high performance computing environment stimulated IBM to announce in December 1999, a five year effort to build a massively parallel computer 'Blue Gene' to study biomolecular phenomena.[29] In 2000, protein unfolding simulations of Engrailed Homeodomain (En-HD) from *Drosophila melanogaster* a 61 residue, mainly α-helical protein was carried out for a maximum time scale of 70 ns.[21] The introduction of distributed computing paved the way towards achieving longer time scales in the simulations of the dynamics of biomolecules at atomic level.[30] The Folding@home project used distributed computing techniques and a super cluster of thousands of computer processors to simulate 38 µs of folding time. Folding of the C-terminal β-hairpin from protein G in atomistic detail using the GB/SA implicit solvent model at 300 K was reported.[22] Later in 2001 all atom protein structure prediction using *ab initio* protein folding simulation was carried out for trpcage TC5b with extended initial conformation for a time period of 50 ns at 300 K.[23] Simulations for Trp-cage for an aggregate time of $\sim$100 µs were performed by the Folding@home project to capture the rapid relaxation from an extended starting state to a relaxed unfolded state.[24] In 2003, Langevin dynamics was applied to the physics based united residue (UNRES) force field to generate trajectories for seven proteins with an average folding time of the order of nanoseconds. Folding with Langevin dynamics helped in exploring thousands of folding pathways and also enabled predicting not only the native structure but also the folding scenario of the protein.[31] The improved performance of molecular dynamics softwares and computing resources made it possible to perform multiple microsecond simulations in explicit solvent environment. Using the high performance computing machines Ensign *et al.* in 2007 presented large folding trajectories for villin mutant.[32] Freddolino *et al.* reported in 2008, a 10 µs trajectory of the fast folding human Pin1 WW domain mutant Fip35.[25,26] A recent initiative by Folding@home distributed computing platform was successful in performing a large array of distributed implicit solvent folding simulations of a 39 residue protein NTL9(1–39) using Amber ff96 force field and accelerated version of GROMACS for GPU processors for an aggregate time scale of 1.52 ms.[27] Crossing the barriers of computational resources, Shaw *et al.* developed a special purpose machine christened 'Anton', which has greatly accelerated the execution of simulations and generation of trajectories of 1 ms length. Such massively parallel specialized machines have allowed all-atom molecular dynamics simulations of proteins in an explicit solvent environment at a much faster rate and 100 times longer time scales.[28] Taking the protein folding in a new direction, researchers at the University of Washington developed a protein folding video game Foldit that uses human visual problem solving and strategy development capabilities with traditional computing algorithms.[33]

Despite the significant advantages and power of molecular simulations, the field of *ab initio* protein folding still faces serious challenges. The huge amount of sampling space and the deficiencies in potential functions restrict the use of simulations to smaller proteins and refinement of models produced by low-resolution methods.[34] Increasing availability of experimentally determined protein structures has inspired the development of knowledge based methods for structure prediction. With the exception of 'pure' physico-chemical approaches,[35] these methods rely on searching the protein structure databases and using the available structural information to predict the tertiary structure of new sequences. The bi-annual community wide Critical Assessment of Protein Structure Prediction (CASP) experiments[36,37] classify such methods under the category of Template based modelling. The term *ab initio* is used in much broader sense in CASP and includes methods that compare fragments (short stretches) of query sequence of unknown structure with sequences in protein structure database (RCSB)[38] and assemble atomic models for the whole protein with varying strategies for dealing with regions with missing matches. The primary obstacle to these template based methods is database dependency especially where a related structural homolog is not available or where the query sequence presents a totally new fold. Providing an understanding of the forces driving the protein structure formation is obviously beyond the purview of these methods.[39]

2. Methodology

In a modification of the *ab intio* procedures delineated above (called *de novo* methods which use partial information from structural databases) as for instance using database searches for secondary structures only and no use of database information for the tertiary structure prediction, we describe here an improved and computationally robust version of *Bhageerath*[40] an energy based software suite for narrowing down the search space of tertiary structures of small globular proteins. The protocol comprises eight different computational modules that form an automated pipeline. Proceeding from the input amino acid sequence, the software first predicts the secondary structure information (helix/strand/loop) along the entire length of the protein. The second module creates an atomic-level extended structure using the secondary structure information. The third module generates a large number of trial structures with a systematic sampling of the conformational space of loop dihedrals. The number of trial structures generated is $128^{(n-1)}$ where 'n' is the number of secondary structural elements and '$n - 1$' is the number of loops/junctions between the secondary structural units. These structures are generated by choosing seven dihedrals from each of the loops (three at both ends and one dihedral from the middle of the loop) and sampling two conformational states for each dihedral. The generated trial structures are screened in the fourth module through persistence length, radius of

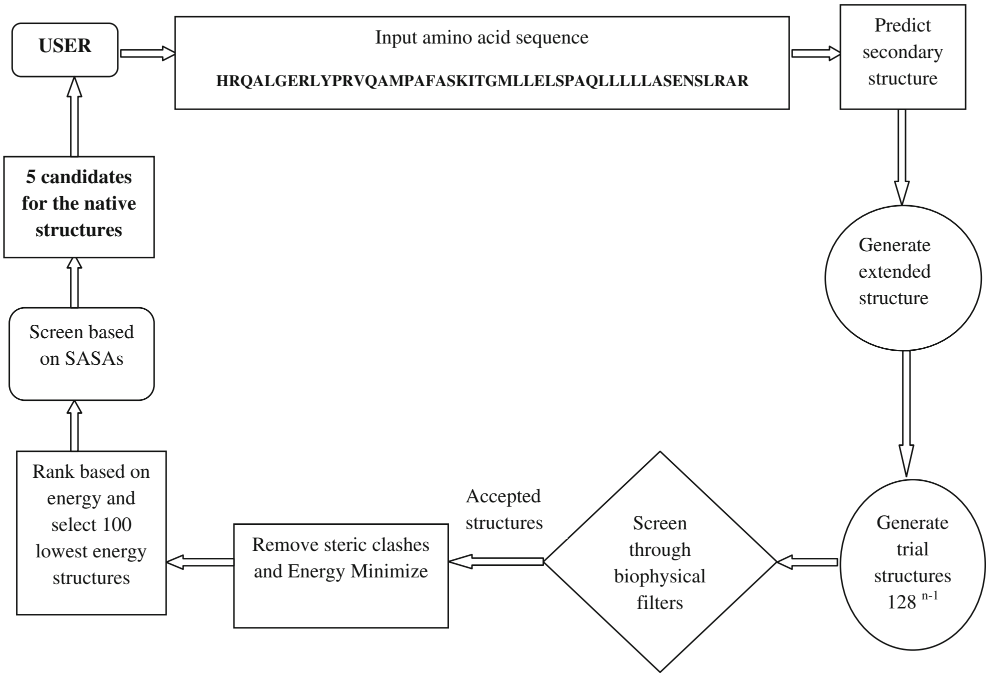

USER
Input amino acid sequence
HRQALGERLYPRVQAMPAFASKITGMLLELSPAQLLLLLASENSLRAR
Predict secondary structure
Generate extended structure
Generate trial structures 128 n-1
Screen through biophysical filters
Accepted structures
Remove steric clashes and Energy Minimize
Rank based on energy and select 100 lowest energy structures
Screen based on SASAs
5 candidates for the native structures

Figure 1. The flow of information in *Bhageerath* web server, starting with the input sequence from the user to the final prediction of five candidate structures for the native.

gyration, topological distinctness of generated structures, inter-atomic distance and C_α loop distance filters,[41] developed for the purpose of reducing the number of improbable candidates. The resultant structures are refined to the fifth module by a Monte Carlo sampling in dihedral space to remove steric clashes and overlaps involving atoms of main chain and side chains. In module six, the structures are energy minimized to further optimize the side chains. Module seven ranks the structures using an all atom energy based empirical scoring function[42] and selects 100 lowest energy structures. Module eight reduces the structures selected in the previous module to five using solvent accessible surface areas (SASA).[43] Short molecular dynamics simulations with explicit solvent for further refinement of the five structures is an optional last step. The protocol has been web-enabled and is freely accessible at http://www.scfbio-iitd.res.in/bhageerath.[40] The flow chart diagram of *Bhageerath* is depicted in figure 1.

3. Results and discussion

The *Bhageerath* methodology has been validated on 80 small globular proteins (<100 amino acids) consisting of up to five helices and strands with known tertiary structures. The results obtained for the 80 small globular proteins with the web server are shown in table 2. For each of these proteins, a structure within 3–7Å RMSD (root mean square deviation) of the native has been obtained in the five lowest energy structures. Figure 2 shows a superimposition of the lowest RMSD predicted structure with the native structure for the 80 test proteins.

All the eight modules of the protocol are currently incorporated on a dedicated 280 AMD Opteron 2.4 GHz processor cluster. In contrast to typical short return times (ranging from 1 to 10 min) for receiving results from comparative (homology based) modelling servers, the expected prediction time with *Bhageerath* web server for small systems (≤3 helices) is ∼ 10 min. The prediction times in case of *Bhageerath* depend on the length of the sequence, number of secondary structural elements, number of trial structures generated and the accepted number of trial structures after biophysical filters which undergo all atom energy processing.

The earlier version of *Bhageerath* had a limitation of predicting structures of proteins with not more than 100 amino acids and 2–3 secondary structures. Pushing the frontiers of the atomic models, we have now developed a new methodology which can handle proteins with more than 100 amino acids. The protocol uses a 'divide and conquer' strategy based on the number of secondary structure elements, wherein the query sequence is divided into overlapping fragments and five structures are generated for each fragment with *Bhageerath* methodology as described above, followed by a patching of all the fragments and scoring them with a final selection of five structures for the overall sequence. This methodology is under rigorous validation.

4. Development of *Bhageerath*-H

The major obstacle in the computational structure prediction based on first principles is the conformational sampling. Sampling the entire conformational space of the polypeptide starting from a fully extended conformation and selecting a conformation which uniquely has a lower energy than the non-native conformations is a daunting task. The *Bhageerath* can predict structures of small proteins with an RMSD < 7–10 Å from the native almost routinely now in an automated mode, but for larger proteins we envisage development of an *ab initio*–homology hybrid methodology christened *Bhageerath*-H for tertiary structure prediction for improved accuracy. In *Bhageerath*-H methodology, we identify regions of polypeptide chain where local sequence similarities are realized, create 3D-structural fragments using conventional bioinformatics tools, use the *ab initio* method for regions with no matches in structural databases, and patch

Table 2. Validation of *Bhageerath* Protocol on 80 small globular proteins.

Sl. No.	PDBID	No. of amino acids	No. of secondary structural elements	Lowest RMSD (Å)	Energy rank of lowest structure in top 5 structures
1	1E0Q	17	2E	**2.5**	2
2	1B03	18	2E	**4.4**	2
3	1WQC	26	2H	**2.5**	3
4	1RJU	36	2H	**5.9**	4
5	1EDM	39	2E	**3.5**	2
6	1AB1	46	2H	**4.2**	5
7	1BX7	51	2E	**3.2**	4

Table 2. (continued)

Sl. No.	PDBID	No. of amino acids	No. of secondary structural elements	Lowest RMSD (Å)	Energy rank of lowest structure in top 5 structures
8	1FME	28	1H,2E	**3.7**	5
9	1ACW	29	1H,2E	**5.3**	3
10	1AIL	70	3H	**4.4**	3
11	1B6Q	56	2H	**3.8**	5
12	1ROP	56	2H	**4.3**	2
13	1NKD	59	2H	**3.9**	1
14	1RPO	61	2H	**3.8**	2
15	1QR8	68	2H	**3.9**	4
16	1YRF	35	3H	**4.8**	4
17	1YRI	35	3H	**4.6**	3
18	2ERL	40	3H	**4**	3
19	1RES	43	3H	**4.2**	2
20	1GVD	52	3H	**5.1**	4
21	1DFN	30	3E	**5**	1
22	1Q2K	31	1H,2E	**4.8**	4
23	1SCY	31	1H,2E	**3.1**	5
24	1XRX	34	1E,2H	**5.6**	1
25	1ROO	35	3H	**2.8**	5
26	1MBH	52	3H	**4**	4
27	1HDD	57	3H	**5.5**	4
28	1BDC	60	3H	**4.8**	5
29	1DF5	68	3H	**3.4**	1
30	1QR9	68	3H	**3.8**	2
31	1VII	36	3H	**3.7**	2
32	1BGK	37	3H	**4.1**	3
33	1BHI	38	1H,2E	**5.3**	2
34	1OVX	38	1H,2E	**4**	1
35	1I6C	39	3E	**5.1**	2
36	2G7O	68	4H	**5.8**	2
37	2OCH	66	4H	**6.6**	3
38	1WR7	41	3E,1H	**5.2**	2
39	2B7E	59	4H	**6.8**	4
40	1FAF	79	4H	**6.4**	4
41	2CPG	43	1E,2H	**5.3**	2
42	1DV0	45	3H	**5.1**	4
43	1IRQ	48	1E,2H	**5.5**	3
44	1GUU	50	3H	**4.6**	4
45	1GV5	52	3H	**4.1**	2
46	1PRB	53	4H	**6.9**	4
47	1DOQ	69	5H	**6.8**	3
48	1I2T	61	4H	**5.4**	4
49	2CMP	56	4H	**5.6**	1
50	1X4P	66	4H	**5.2**	3
51	1GAB	53	3H	**4.9**	1
52	1MOF	53	3H	**2.9**	5
53	1ENH	54	3H	**4.6**	3
54	1IDY	54	3H	**3.6**	5
55	1PRV	56	3H	**5**	5
56	1BW6	56	4H	**4.2**	1
57	2K2A	70	4H	**6.1**	1
58	1TGR	52	4H	**6.8**	2
59	2V75	90	5H	**7**	3
60	1HNR	47	2E,2H	**5.2**	2
61	1I5X	61	3H	**3.6**	3
62	1I5Y	61	3H	**3.4**	5
63	1KU3	61	3H	**5.5**	4

Table 2. (continued)

Sl. No.	PDBID	No. of amino acids	No. of secondary structural elements	Lowest RMSD (Å)	Energy rank of lowest structure in top 5 structures
64	1YIB	61	3H	**3.5**	5
65	1AHO	64	1H,2E	**4.5**	4
66	2KJF	60	4H	**5**	4
67	1RIK	29	2E,2H	**4.4**	4
68	1JEI	53	4H	**5.8**	5
69	2HOA	68	4H	**6.3**	4
70	2DT6	62	4H	**5.9**	3
71	2L37	43	2H	**3**	1
72	2PMR	76	3H	**6.8**	2
73	1I2T	61	4H	**5.7**	2
74	2PM1	30	3E	**4.6**	4
75	2CJJ	63	3H	**5.2**	1
76	1WY3	35	3H	**5**	4
77	1P9I	31	1H	**1.7**	1
78	3NMD	53	1H	**1.9**	1
79	2J15	15	2E	**2.6**	5
80	3E21	40	3H	**5.5**	5

(E=Strand; H=Helix; RMSD=root mean square deviation from the crystal structure i.e., native reported in RCSB)

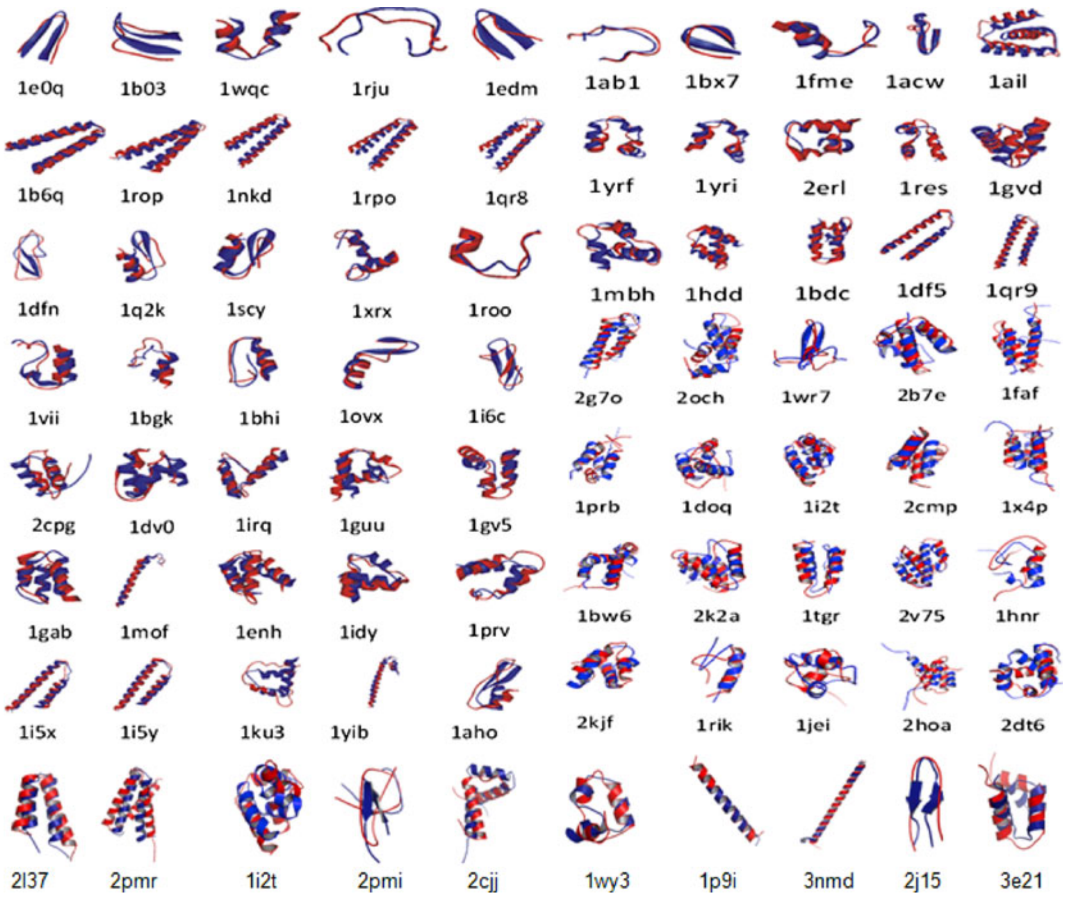


Figure 2. The superimposed lowest RMSD structure for the 80 small globular proteins. The PDB ID's are shown underneath each structure. The predicted structure is in red colour and the native (experimentally determined structure) is in blue.

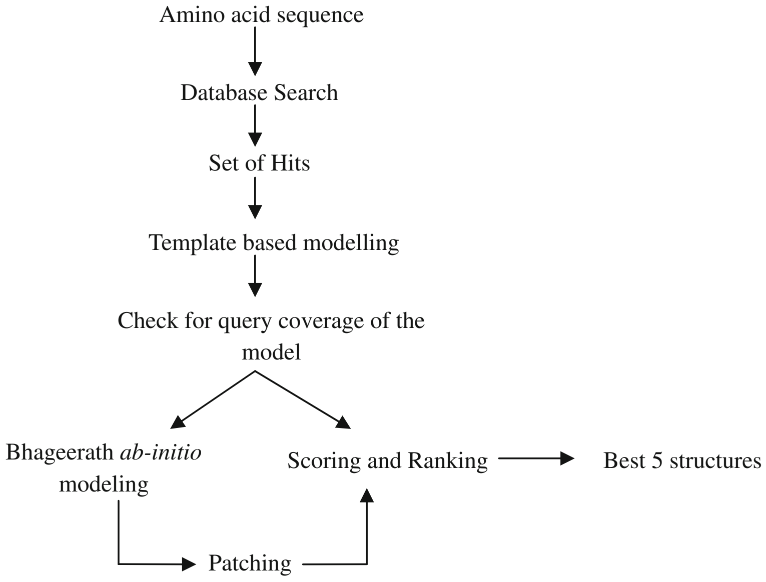


Figure 3. The flow chart of *Bhageerath*-H.

them to put together a complete structure for the proteins. Figure 3 shows a flow chart of *Bhageerath*-H methodology (http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp).

The protocol has been tested on various CASP9 targets of medium to large size (with more than 200 residues). Table 3 shows the root mean square deviation from the native of the best structures predicted by *Bhageerath*-H for five CASP9 targets.

A comparison of the structure predictions by *Bhageerath*, *Bhageerath*-H, Phyre2,[44] Zhang-Server,[45,46] Baker–Rosetta[47] and HHPredA[48] for five CASP9[49] targets was carried out (table 4). The table shows the RMSDs of the predictions by *Bhageerath* and *Bhageerath*-H in CASP9 and post-CASP9 and the RMSDs of the predictions submitted by four other servers in CASP9. In four of the five cases (T0538-D1, T0602-D1, T0605-D1, T0559-D1) *Bhageerath* was able to predict a structure within 10 Å RMSD from the native with target T0643-D1 as an exception. Post CASP9, we have incorporated a new version of structure generator and scoring function in *Bhageerath-H*, which has improved the prediction accuracy of the software (as seen in column 5). Excluding the templates with more than 30% similarity, the latest version of *Bhageerath*-H is able to predict a structure within 7Å RMSD from the native for all the 5 targets. In each of the five illustrative cases, *Bhageerath* and *Bhageerath*-H are able to predict structures with RMSDs comparable to those obtained by some popular servers such as Phyre2, Zhang-server, Baker-Rosetta and HHPredA.

Thus, for sequences with known sequence homologs, *Bhageerath*-H has the potential to predict a structure with higher resolution, accuracy in less time. This clearly demonstrates the advantage of hybrid methods over *ab initio* methods when a close homolog is available in the database, but for new sequences with no available sequence homologs, *ab initio/de novo* servers such as *Bhageerath* are the only alternative. Figure 4 shows a comparison of the structure prediction time and accuracies of *Bhageerath* and *Bhageerath-H* software suites for small globular proteins.

Table 3. Validation of the *Bhageerath*-H protocol on 5 CASP9 targets.

Sl. No.	Target name	No. of residues	Lowest RMSD (Å)
1	T0515	365	2.7
2	T0518	288	2.5
3	T0524	325	4.3
4	T0597	429	1.9
5	T0607	471	3.8

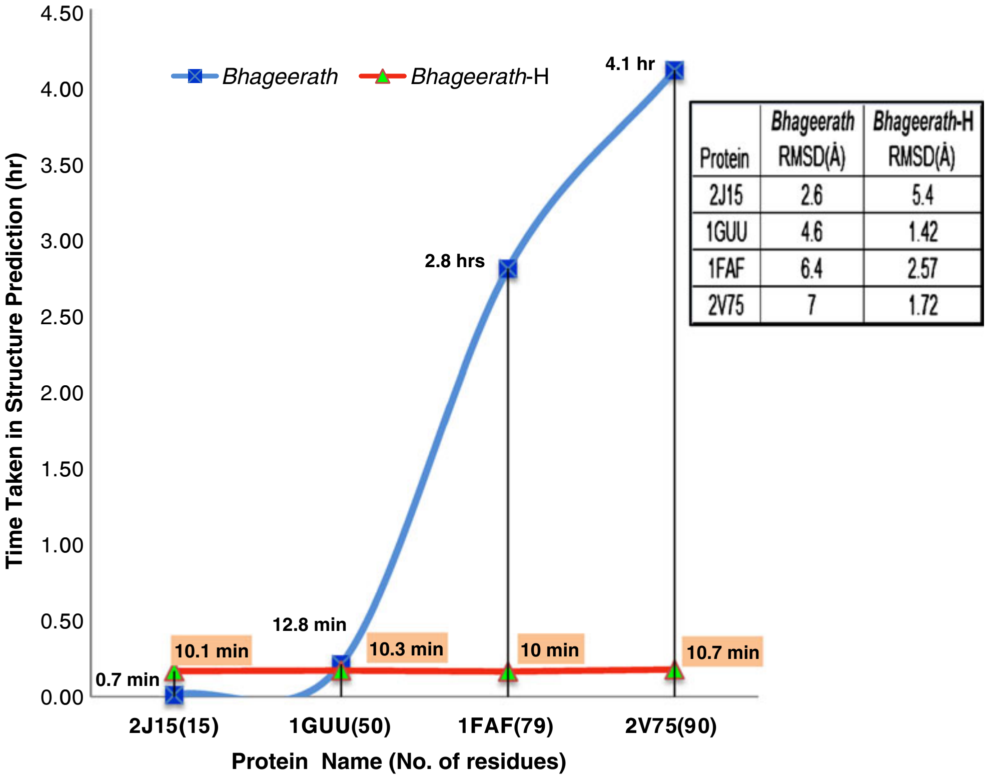
5. Conclusions

We have described here an all atom energy based computational methodology, *Bhageerath* for tertiary structure prediction of small soluble proteins. Results on 80 globular proteins show that *Bhageerath* web server predicts one or more candidate structures within an RMSD of 7Å from the native for proteins with less

Table 4. A comparison of protein tertiary structure prediction accuracies of *Bhageerath* and *Bhageerath*-H with Phyre2, Baker–Rosetta, Zhang-Server and HHpredA software for 5 CASP9 (3 May to 17 July, 2010) targets.

Sl. No.	PDBID	Residues	CASP9 ID	Lowest RMSD (Å) prediction by *Bhageerath* post CASP9 (in CASP9)	Lowest RMSD (Å) prediction by *Bhageerath*-H post CASP9 (in CASP9)	Lowest RMSD (Å) prediction by Phyre2 in CASP9	Lowest RMSD (Å) prediction by Baker-Rosetta Server in CASP9	Lowest RMSD (Å) prediction by Zhang-Server in CASP9	Lowest RMSD (Å) prediction by HHpredA in CASP9
1	2L09	53	T0538-D1	6.88 (9.02)	1.87#	2.11	2.17	1.39	2.15
2	3NKZ	55	T0602-D1	3.43 (4.71)	2.90 (2.90)	2.14	1.61	1.49	2.14
3	3NMD	49	T0605-D1	1.97 (2.48)	3.80 (16.49)	2.84	1.84	1.97	1.62
4	3NZL	73	T0643-D1	4.81 (10.31)	3.49 (4.59)	5.19	5.68	4.30	8.9
5	2L01	67	T0559-D1	8.33 (8.15)	5.59#	2.88	0.98	1.66	2.19

#These targets were fielded only for server prediction category and not for human expert group in CASP9.



Protein	*Bhageerath* RMSD(Å)	*Bhageerath*-H RMSD(Å)
2J15	2.6	5.4
1GUU	4.6	1.42
1FAF	6.4	2.57
2V75	7	1.72

Figure 4. A comparison of the structure prediction time and accuracy of *Bhageerath* and *Bhageerath*-H software suites for four small globular proteins with <100 amino acids.

than five secondary structural elements. CASP9 results reflect the potential of the protocol to predict a structure within 10Å RMSD from the native for sequences with no known sequence homologs. For proteins with local sequence and structure matches involving short fragments, it is expedient to use a hybrid method such as *Bhageerath*-H for tertiary structure prediction. In a nutshell, for small proteins the structure prediction problem is under control with *ab initio* methods, and for larger proteins computational protocols involving hybrid models are getting better and better.

Acknowledgements

Funding from the Department of Information Technology and programme support from the Department of Biotechnology, Government of India to Supercomputing Facility for Bioinformatics are gratefully acknowledged.

References

1. Creighton T E 1990 *Biochem. J.* **270** 1
2. Dobson C M 2003 *Nature* **426** 884
3. Editorial 2005 *Science* **309** 78
4. Unger R and Moult J 1993 *Bull. Math. Biol.* **55** 1183
5. Fraenkel A S 1993 *Bull. Math. Biol.* **55** 1199
6. Baker D 2000 *Nature* **405** 39
7. Klepeis J L and Floudas C A 2004 *SIAM News* **37** 1
8. Venkatraman J, Shankaramma S C and Balaram P 2001 *Chem. Rev.* **101** 3131
9. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis J L, Dror R O and Shaw D E 2010 *Proteins* **78** 1950
10. Levitt M and Warshel A 1975 *Nature* **253** 694
11. McCammon J A, Gelin B R and Karplus M 1977 *Nature* **267** 585
12. Li A and Daggett V 1995 *Protein Eng.* **8** 1117
13. Daggett V and Levitt M 1992 *Proc. Natl. Acad. Sci. USA* **89** 5142
14. Levitt M 1983 *J. Mol. Biol.* **168** 595
15. Levitt M 1983 *J. Mol. Biol.* **168** 621
16. Tirado-Rives J and Jorgensen W L 1993 *Biochemistry* **32** 4175
17. Boczko E M and Brooks C L III 1995 *Science* **269** 393
18. Demchuk E, Bashford D and Case D A 1997 *Fold. Des.* **2** 35
19. Daura X, Jaun B, Seebach D, van Gunsteren W F and Mark A E 1998 *J. Mol. Biol.* **280** 925
20. Duan Y and Kollman P A 1998 *Science* **282** 740
21. Mayor U, Johnson C M, Daggett V and Fersht A R 2000 *Proc. Natl. Acad. Sci. USA* **97** 13518
22. Zagrovic B, Sorin E J and Pande V S 2001 *J. Mol. Biol.* **313** 151
23. Simmerling C, Strockbine B and Roitberg A E 2002 *J. Am. Chem. Soc.* **124** 11258
24. Snow C D, Zagrovic B and Pande V S 2002 *J. Am. Chem. Soc.* **124** 14548
25. Freddolino P L, Liu F, Gruebele M and Schulten K 2008 *Biophys. J.* **94** L75
26. Freddolino P L, Park S, Roux B and Schulten K 2009 *Biophys. J.* **96** 3772
27. Voelz V A, Bowman G R, Beauchamp K and Pande V S 2010 *J. Am. Chem. Soc.* **132** 1526
28. Shaw D E, Maragakis P, Lindorff-Larsen K, Piana S, Dror R O, Eastwood M P, Bank J A, Jumper J M, Salmon J K, Shan Y and Wriggers W 2010 *Science* **330** 341
29. Allen F *et al.* 2001 *IBM Syst. J.* **40** 310
30. Shirts M and Pande V S 2000 *Science* **290** 1903
31. Liwo A, Khalili M and Scheraga H A 2005 *Proc. Natl. Acad. Sci. USA* **102** 2362
32. Ensign D L, Kasson P M and Pande V S 2007 *J. Mol. Biol.* **374** 806
33. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z and Players F 2010 *Nature* **466** 756
34. Bonneau R and Baker D 2001 *Annu. Rev. Biophys. Biomol. Struct.* **30** 173
35. Petrey D and Honig B 2005 *Mol. Cell* **20** 811
36. Moult J, Pedersen J T, Judson R and Fidelis K 1995 *Proteins* **23** ii
37. Moult J, Fidelis K, Kryshtafovych A, Rost B and Tramontano A 2009 *Proteins* **77** 1
38. Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N, Bourne P E 2000 *Nucleic Acids Res.* **28** 235
39. Floudas C A, Fung H K, McAllister S R, Monnigmann M and Rajgaria R 2006 *Chem. Eng. Sci.* **61** 966
40. Jayaram B, Bhushan K, Shenoy S R, Narang P, Bose S, Agrawal P, Sahu D and Pandey V 2006 *Nucleic Acids Res.* **34** 6195
41. Narang P, Bhushan K, Bose S and Jayaram B 2005 *Phys. Chem. Chem. Phys.* **7** 2364
42. Narang P, Bhushan K, Bose S and Jayaram B 2006 *J. Biomol. Struct. Dyn.* **23** 385
43. Hubbard S J and Thornton J M 1993 *NACCESS Computer Program* (London: Department of Biochemistry and Molecular Biology, University College London)
44. Kelley L A and Sternberg M J E 2009 *Nature Protocols* **4** 363
45. Roy A, Kucukural A and Zhang Y 2010 *Nature Protocols* **5** 725
46. Zhang Y 2008 *BMC Bioinformatics* **9** 1
47. Kim D E, Chivian D and Baker D 2004 *Nucleic Acids Res.* **32** W526
48. Soding J, Biegert A and Lupas A N 2005 *Nucleic Acids Res.* **33** W244
49. http://predictioncenter.org/casp9/