

Dear Author

Here are the proofs of your article.

- You can submit your corrections **online**, via **e-mail** or by **fax**.
- For **online** submission please insert your corrections in the online correction form. Always indicate the line number to which the correction refers.
- You can also insert your corrections in the proof PDF and **email** the annotated PDF.
- For **fax** submission, please ensure that your corrections are clearly legible. Use a fine black pen and write the correction in the margin, not too close to the edge of the page.
- Remember to note the **journal title**, **article number**, and **your name** when sending your response via e-mail or fax.
- **Check** the metadata sheet to make sure that the header information, especially author names and the corresponding affiliations are correctly shown.
- **Check** the questions that may have arisen during copy editing and insert your answers/corrections.
- **Check** that the text is complete and that all figures, tables and their legends are included. Also check the accuracy of special characters, equations, and electronic supplementary material if applicable. If necessary refer to the *Edited manuscript*.
- The publication of inaccurate data such as dosages and units can have serious consequences. Please take particular care that all such details are correct.
- Please **do not** make changes that involve only matters of style. We have generally introduced forms that follow the journal's style.
- Substantial changes in content, e.g., new results, corrected values, title and authorship are not allowed without the approval of the responsible editor. In such a case, please contact the Editorial Office and return his/her consent together with the proof.
- If we do not receive your corrections **within 48 hours**, we will send you a reminder.
- Your article will be published **Online First** approximately one week after receipt of your corrected proofs. This is the **official first publication** citable with the DOI. **Further changes are, therefore, not possible.**
- The **printed version** will follow in a forthcoming issue.

Please note

After online publication, subscribers (personal/institutional) to this journal will have access to the complete article via the DOI using the URL:

<http://dx.doi.org/10.1007/s12039-011-0189-x>

If you would like to know when your article has been published online, take advantage of our free alert service. For registration and further information, go to:

<http://www.springerlink.com>.

Due to the electronic nature of the procedure, the manuscript and the original figures will only be returned to you on special request. When you return your corrections, please inform us, if you would like to have these documents returned.

Metadata of the article that will be visualized in OnlineFirst

1	Article Title	<i>Bhageerath—Targeting the near impossible: Pushing the frontiers of atomic models for protein tertiary structure prediction</i>
2	Article Sub- Title	
3	Article Copyright - Year	Indian Academy of Sciences 2011 (This will be the copyright line in the final PDF)
4	Journal Name	Journal of Chemical Sciences
5		Family Name JAYARAM
6		Particle
7		Given Name B
8		Suffix
9		Organization Indian Institute of Technology
10		Division Department of Chemistry
11	Corresponding Author	Address Hauz Khas, New Delhi 110016, India
12		Organization Indian Institute of Technology
13		Division Supercomputing Facility for Bioinformatics and Computational Biology
14		Address Hauz Khas, New Delhi 110016, India
15		Organization Indian Institute of Technology
16		Division School of Biological Sciences
17		Address Hauz Khas, New Delhi 110016, India
18		e-mail bjayaram@chemistry.iitd.ac.in
19		Family Name DHINGRA
20		Particle
21		Given Name PRIYANKA
22		Suffix
23		Organization Indian Institute of Technology
24	Author	Division Department of Chemistry
25		Address Hauz Khas, New Delhi 110016, India
26		Organization Indian Institute of Technology
27		Division Supercomputing Facility for Bioinformatics and Computational Biology
28		Address Hauz Khas, New Delhi 110016, India
29		e-mail
30		Family Name LAKHANI
31		Particle
32		Given Name BHARAT
33		Suffix

34		Organization	Indian Institute of Technology
35		Division	Supercomputing Facility for Bioinformatics and Computational Biology
36	Author	Address	Hauz Khas, New Delhi 110016, India
37		e-mail	
38		Family Name	SHEKHAR
39		Particle	
40		Given Name	SHASHANK
41		Suffix	
42	Author	Organization	Indian Institute of Technology
43		Division	Supercomputing Facility for Bioinformatics and Computational Biology
44		Address	Hauz Khas, New Delhi 110016, India
45		e-mail	
46		Received	
47	Schedule	Revised	
48		Accepted	
49	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 <i>evolved to a stage to provide</i> a computational solution to the tertiary structures of small proteins. Knowledge-based methodologies all developing rapidly to enable prediction of maturing in 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 <i>ab initio</i> models to arrive at reasonable predictions. We describe here the development of <i>Bhageerath</i> an <i>ab initio</i> model and <i>Bhageerath-H</i> a hybrid model and present a critique on the current status of prediction of protein tertiary structures.	
50	Keywords separated by ' - '	<i>Ab initio</i> - protein folding - molecular dynamics simulation - protein structure prediction - <i>Bhageerath</i> - critical assessment of protein structure prediction (CASP)	
51	Foot note information	Dedicated to Prof. N Sathyamurthy in his 60th birthday	

Q1 ***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

^aDepartment of Chemistry, ^bSupercomputing Facility for Bioinformatics and Computational Biology,
^cSchool of Biological Sciences, Indian Institute of Technology, 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-based methodologies all developing rapidly to enable prediction of maturing in 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 1. Introduction

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

11 The *ab initio* term in the context of protein struc-
12 ture prediction is used to signify the usage of physics
13 based all atom molecular mechanics potentials for pre-
14 dicting the three dimensional structure of a protein.
15 Molecular dynamics, Monte Carlo simulations and their
16 variants are pooled under this category. Molecular
17 dynamics simulations, in particular, have provided

extremely high resolution spatial and temporal data, 18
enhancing our knowledge and understanding of the 19
protein folding mechanism. Simulations of biologi- 20
cally relevant processes, with atomistic accuracy on 21
timescales beyond microsecond are now possible due to 22
advances in software and hardware.⁹ 23

24 In the year 1975, Levitt and Warshel simulated the 25
folding of bovine pancreatic trypsin inhibitor (BPTI) 26
using a simple representation of protein conformation, 27
energy minimization and thermalisation. They suc- 28
ceeded in ‘renaturing’ the protein from an open fully 29
extended conformation to a folded native like con- 30
formation.¹⁰ Later in 1977, McCammon, Gelin and 31
Karplus studied for the first time the dynamics of folded 32
BPTI in vacuum at a molecular level over a period 33
of 9.2 picoseconds.¹¹ Since then extensive research 34
has been carried out in the area of protein folding, 35
unfolding, dynamics and structure. Table 1, summa- 36
rizes some major milestones capturing the advance- 37
ments. In the year 1995, Li and Daggett studied the 38
structure and dynamics of native chymotrypsin inhibi- 39
tor 2 with explicit water from a 5.3 ns simulation.¹² 40
Insights were gained from 550 ps unfolding simula- 41
tions of reduced BPTI at high temperature.¹³ Folding

[#]Dedicated to Prof. N Sathyamurthy in 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 inhibitor 2 (CI2) ¹²	5.3 ns	1995
6	Staphylococcal protein ¹⁷	>9 ns	1995
7	Pentapeptide <i>cis</i> -AYPYD ¹⁸	20 ns	1997
8	β -Heptapeptide ¹⁹	50 ns	1998
9	Villin headpiece ²⁰	1 μ s	1998
10	Engrailed Homeodomain ²¹	40 ns, 70 ns	2000
11	Protein G ²²	38 μ s	2001
12	Trp cage ²³	50 ns	2002
13	Trp cage ²⁴	\sim 100 μ s	2002
14	Human Pin1 WW domain mutant, FiP35 ^{25,26}	10 μ s	2009
15	NTL9 ²⁷	1.52 ms	2010
16	FiP35, Villin headpiece ²⁸	100 μ s	2010
	Bovine Pancreatic Trypsin Inhibitor (BPTI)	1 ms	

42 simulation of small peptide fragments such as *beta*-turn
 43 in a short linear peptide and β -heptapeptides in aque-
 44 ous solution were performed for 20 ns and 50 ns.^{18,19}
 45 Using a Cray T3E, a massively parallel supercom-
 46 puter consisting hundreds of CPUs, Duan and Kollman
 47 in 1998 reported one of the longest simulations of
 48 that time for a protein in water. They simulated villin
 49 headpiece subdomain (HP-36) for \sim 1 μ s with \sim 3000
 50 water molecules.²⁰ The growing track record of pro-
 51 tein folding simulations in a high performance com-
 52 puting environment stimulated IBM to announce in
 53 December 1999, a five year effort to build a massively
 54 parallel computer 'Blue Gene' to study biomolecular
 55 phenomena.²⁹ In 2000, protein unfolding simulations
 56 of Engrailed Homeodomain (En-HD) from *Drosophila*
 57 *melanogaster* a 61 residue, mainly α -helical protein
 58 was carried out for a maximum time scale of 70 ns.²¹
 59 The introduction of distributed computing paved the
 60 way towards achieving longer time scales in the simu-
 61 lations of the dynamics of biomolecules at atomic level.³⁰
 62 The Folding@home project used distributed computing
 63 techniques and a super cluster of thousands of computer
 64 processors to simulate 38 μ s of folding time. Folding
 65 of the C-terminal β -hairpin from protein G in atomistic
 66 detail using the GB/SA implicit solvent model at 300 K
 67 was reported.²² Later in 2001 all atom protein structure
 68 prediction using *ab initio* protein folding simulation
 69 was carried out for trpcage TC5b with extended initial
 70 conformation for a time period of 50 ns at 300 K.²³ Sim-
 71 ulations for Trp-cage for an aggregate time of \sim 100 μ s
 72 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 dynam-
 ics helped in exploring thousands of folding pathways
 and also enabled predicting not only the native struc-
 ture but also the folding scenario of the protein.³¹
 The improved performance of molecular dynamics soft-
 wares and computing resources made it possible to
 perform multiple microsecond simulations in explicit
 solvent environment. Using the high performance com-
 puting machines Ensign *et al.* in 2007 presented large
 folding trajectories for villin mutant.³² Freddolino *et al.*
 reported in 2008, a 10 μ s trajectory of the fast folding
 human Pin1 WW domain mutant Fip35.^{25,26} A recent
 initiative by Folding@home distributed computing plat-
 form was successful in performing a large array of dis-
 tributed 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

104 the protein folding in a new direction, researchers at
 105 the University of Washington developed a protein fold-
 106 ing video game Foldit that uses human visual prob-
 107 lem solving and strategy development capabilities with
 108 traditional computing algorithms.³³

109 Despite the significant advantages and power of
 110 molecular simulations, the field of *ab initio* protein
 111 folding still faces serious challenges. The huge amount
 112 of sampling space and the deficiencies in potential func-
 113 tions restrict the use of simulations to smaller proteins
 114 and refinement of models produced by low-resolution
 115 methods.³⁴ Increasing availability of experimentally
 116 determined protein structures has inspired the develop-
 117 ment of knowledge based methods for structure pre-
 118 diction. With the exception of 'pure' physico-chemical
 119 approaches,³⁵ these methods rely on searching the pro-
 120 tein structure databases and using the available struc-
 121 tural information to predict the tertiary structure of
 122 new sequences. The bi-annual community wide Critical
 123 Assessment of Protein Structure Prediction (CASP)
 124 experiments^{36,37} classify such methods under the cate-
 125 gory of Template based modeling. The term *ab initio*
 126 is used in much broader sense in CASP and includes
 127 methods that compare fragments (short stretches) of
 128 query sequence of unknown structure with sequences
 129 in protein structure database (RCSB)³⁸ and assem-
 130 ble atomic models for the whole protein with vary-
 131 ing strategies for dealing with regions with missing

132 matches. The primary obstacle to these template based
 133 methods is database dependency especially where a
 134 related structural homolog is not available or where the
 135 query sequence presents a totally new fold. Providing
 136 an understanding of the forces driving the protein struc-
 137 ture formation is obviously beyond the purview of these
 138 methods.³⁹

139 2. Methodology

140 In a modification of the *ab initio* procedures delineated
 141 above (called *de novo* methods which use partial infor-
 142 mation from structural databases) as for instance using
 143 database searches for secondary structures only and no
 144 use of database information for the tertiary structure
 145 prediction, we describe here an improved and computa-
 146 tionally robust version of *Bhageerath*⁴⁰ an energy based
 147 software suite for narrowing down the search space of
 148 tertiary structures of small globular proteins. The pro-
 149 tocol comprises eight different computational modules
 150 that form an automated pipeline. Proceeding from the
 151 input amino acid sequence, the software first predicts
 152 the secondary structure information (helix/strand/loop)
 153 along the entire length of the protein. The second mod-
 154 ule creates an atomic-level extended structure using
 155 the secondary structure information. The third mod-
 156 ule generates a large number of trial structures with

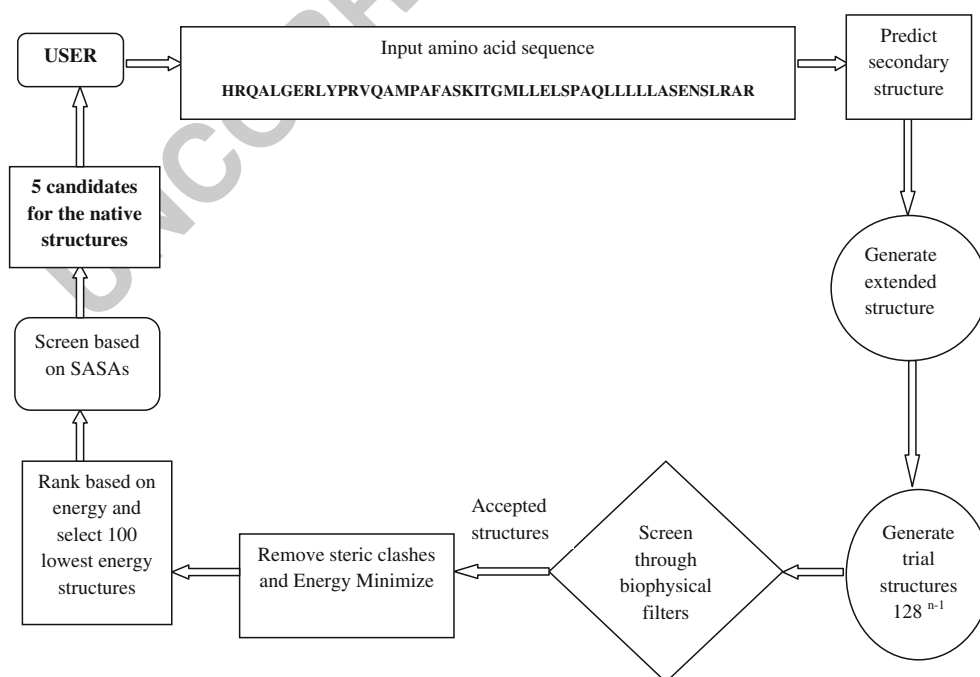


Figure 1. The flow of information in *Bhageerath* web server, starting with the input sequence from the user to the final prediction of five candidate structures for the native.

157 a systematic sampling of the conformational space of
 158 loop dihedrals. The number of trial structures genera-
 159 ted is $128^{(n-1)}$ where 'n' is the number of secondary
 160 structural elements and 'n - 1' is the number of
 161 loops/junctions between the secondary structural units.
 162 These structures are generated by choosing seven dihe-
 163 drals from each of the loops (three at both ends
 164 and one dihedral from the middle of the loop) and
 165 sampling two conformational states for each dihe-
 166 dral. The generated trial structures are screened in the
 167 fourth module through persistence length, radius of
 168 gyration, topological distinctness of generated struc-
 169 tures, inter-atomic distance and C_{α} loop distance fil-
 170 ters,⁴¹ developed for the purpose of reducing the
 171 number of improbable candidates. The resultant struc-
 172 tures are refined to the fifth module by a Monte Carlo
 173 sampling in dihedral space to remove steric clashes and
 174 overlaps involving atoms of main chain and side chains.
 175 In module six, the structures are energy minimized to
 176 further optimize the side chains. Module seven ranks
 177 the structures using an all atom energy based empiri-
 178 cal scoring function⁴² and selects 100 lowest energy
 179 structures. Module eight reduces the structures selected
 180 in the previous module to five using solvent accessible
 181 surface areas (SASA).⁴³ Short molecular dynamics
 182 simulations with explicit solvent for further refinement
 183 of the five structures is an optional last step. The pro-
 184 tocol has been web-enabled and is freely accessible
 185 at <http://www.scfbio-iitd.res.in/bhageerath>.⁴⁰ The flow
 186 chart diagram of *Bhageerath* is depicted in figure 1.

187 3. Results and Discussion

188 The *Bhageerath* methodology has been validated on 80
 189 small globular proteins (<100 amino acids) consisting
 190 of up to five helices and strands with known tertiary
 191 structures. The results obtained for the 80 small glob-
 192 ular proteins with the web server are shown in table 2.

For each of these proteins, a structure within 3–7Å 193
 RMSD (root mean square deviation) of the native has 194
 been obtained in the five lowest energy structures. 195
 Figure 2 shows a superimposition of the lowest RMSD 196
 predicted structure with the native structure for the 80 197
 test proteins. 198

All the eight modules of the protocol are currently 199
 incorporated on a dedicated 280 AMD Opteron 2.4 GHz 200
 processor cluster. In contrast to typical short return 201
 times (ranging from 1 to 10 min) for receiving results 202
 from comparative (homology based) modeling servers, 203
 the expected prediction time with *Bhageerath* web 204
 server for small systems (≤ 3 helices) is ~ 10 min. The 205
 prediction times in case of *Bhageerath* depend on the 206
 length of the sequence, number of secondary structural 207
 elements, number of trial structures generated and the 208
 accepted number of trial structures after biophysical 209
 filters which undergo all atom energy processing. 210

The earlier version of *Bhageerath* had a limitation of 211
 predicting structures of proteins with not more than 100 212
 amino acids and 2–3 secondary structures. Pushing the 213
 frontiers of the atomic models, we have now developed 214
 a new methodology which can handle proteins with 215
 more than 100 amino acids. The protocol uses a 'divide 216
 and conquer' strategy based on the number of sec- 217
 ondary structure elements, wherein the query sequence 218
 is divided into overlapping fragments and five struc- 219
 tures are generated for each fragment with *Bhageerath* 220
 methodology as described above, followed by a patch- 221
 ing of all the fragments and scoring them with a final 222
 selection of five structures for the overall sequence. This 223
 methodology is under rigorous validation. 224

225 4. Development of Bhageerath-H

The major obstacle in the computational structure pre- 226
 diction based on first principles is the conformational 227
 sampling. Sampling the entire conformational space of 228

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

Q3

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)

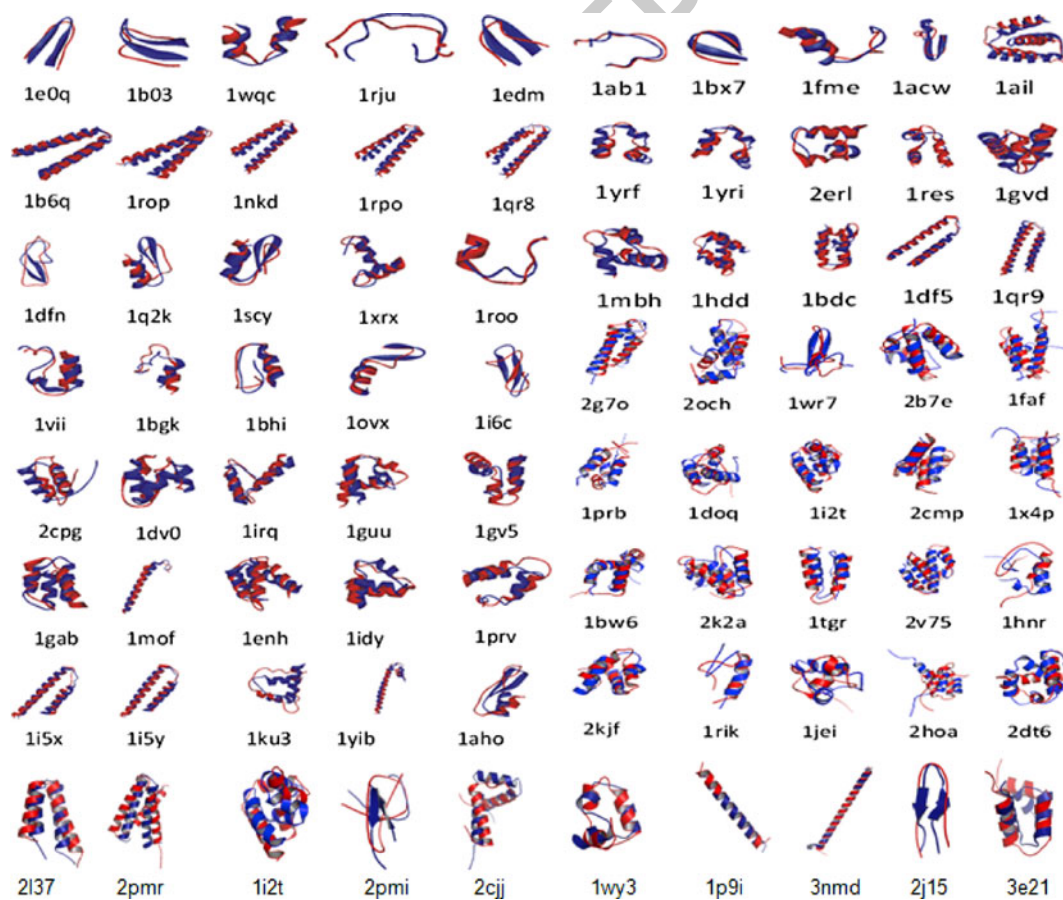


Figure 2. The superimposed lowest RMSD structure for the 80 small globular proteins. The PDB ID's are shown underneath each structure. The predicted structure is in red colour and the native (experimentally determined structure) is in blue.

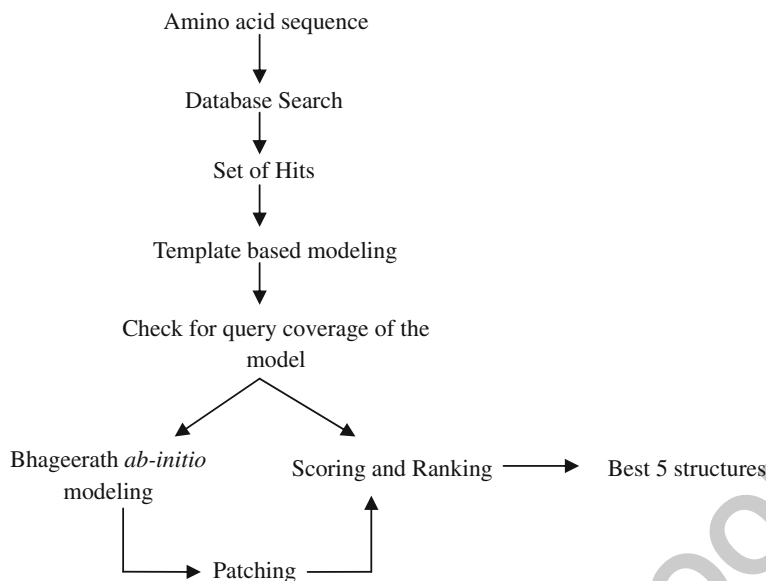


Figure 3. The flow chart of *Bhageerath-H*.

229 the polypeptide starting from a fully extended confor-
 230 mation and selecting a conformation which uniquely
 231 has a lower energy than the non-native conforma-
 232 tions is a daunting task. The *Bhageerath* can pre-
 233 dict structures of small proteins with an RMSD <
 234 7–10Å from the native almost routinely now in an
 235 automated mode, but for larger proteins we envisage
 236 development of an *ab initio*–homology hybrid method-
 237 ology christened *Bhageerath-H* for tertiary structure
 238 prediction for improved accuracy. In *Bhageerath-H*
 239 methodology, we identify regions of polypeptide chain
 240 where local sequence similarities are realized, cre-
 241 ate 3D-structural fragments using conventional bioin-
 242 formatics tools, use the *ab initio* method for regions
 243 with no matches in structural databases, and patch
 244 them to put together a complete structure for the pro-
 245 teins. Figure 3 shows a flow chart of *Bhageerath-H*
 246 methodology ([http://www.scfbio-iitd.res.in/bhageerath/
 247 bhageerath_h.jsp](http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp)).

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

253 A comparison of the structure predictions by
 254 *Bhageerath*, *Bhageerath-H*, Phyre2,⁴⁴ Zhang-
 255 Server,^{45,46} Baker–Rosetta⁴⁷ and HHPredA⁴⁸ for five
 256 CASP9⁴⁹ targets was carried out (table 4). The table
 257 shows the RMSDs of the predictions by *Bhageerath*
 258 and *Bhageerath-H* in CASP9 and post-CASP9 and
 259 the RMSDs of the predictions submitted by four other

260 servers in CASP9. In four of the five cases (T0538-D1, 260
 261 T0602-D1, T0605-D1, T0559-D1) *Bhageerath* was 261
 262 able to predict a structure within 10Å RMSD from 262
 263 the native with target T0643-D1 as an exception. Post 263
 264 CASP9, we have incorporated a new version of struc- 264
 265 ture generator and scoring function in *Bhageerath-H*, 265
 266 which has improved the prediction accuracy of the soft- 266
 267 ware (as seen in column 5). Excluding the templates 267
 268 with more than 30% similarity, the latest version of 268
 269 *Bhageerath-H* is able to predict a structure within 7Å 269
 270 RMSD from the native for all the 5 targets. In each of 270
 271 the five illustrative cases, *Bhageerath* and *Bhageerath-
 272 H* are able to predict structures with RMSDs compar- 272
 273 able to those obtained by some popular servers such as 273
 274 Phyre2, Zhang-server, Baker–Rosetta and HHPredA. 274

275 Thus, for sequences with known sequence homologs, 275
 276 *Bhageerath-H* has the potential to predict a structure 276
 277 with higher resolution, accuracy in less time. This 277
 278 clearly demonstrates the advantage of hybrid methods 278

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

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.

Q3

Sl. No.	PDBID	Residues	CASP9 ID	Lowest RMSD (Å) prediction by <i>Bhageerath</i> post CASP9 (in CASP9)	Lowest RMSD (Å) prediction by <i>Bhageerath-H</i> 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.

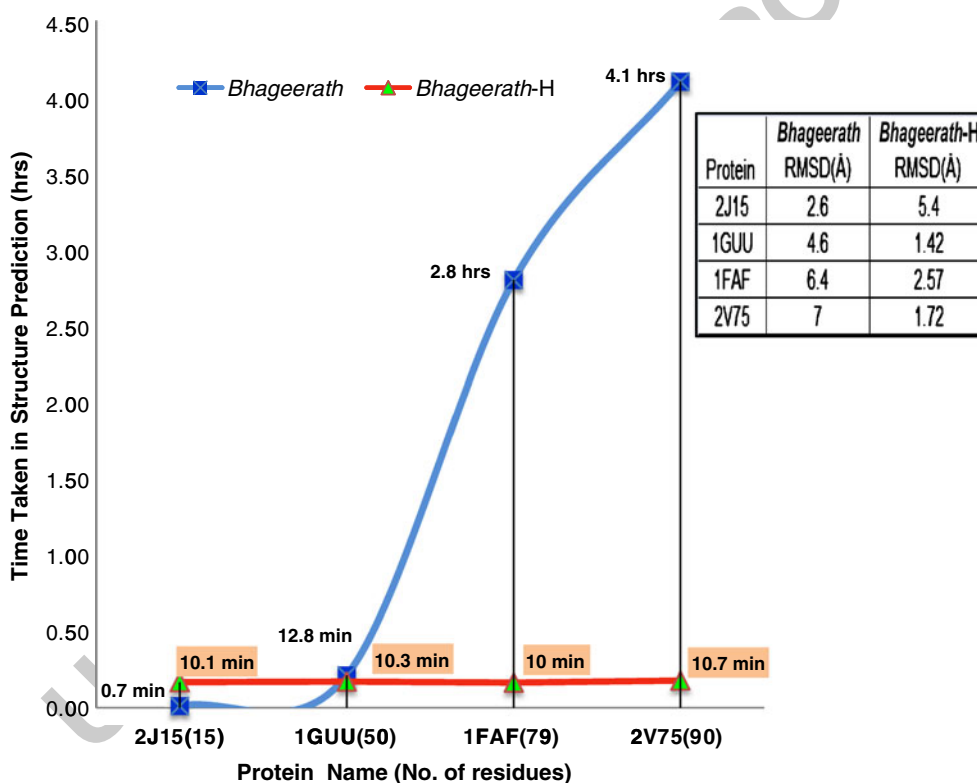


Figure 4. A comparison of the structure prediction time and accuracy of *Bhageerath* and *Bhageerath-H* software suites for four small globular proteins with <100 amino acids.

279 over *ab initio* methods when a close homolog is avail-
 280 able in the database, but for new sequences with no
 281 available sequence homologs, *ab initio/de novo* servers
 282 such as *Bhageerath* are the only alternative. Figure 4
 283 shows a comparison of the structure prediction time and
 284 accuracies of *Bhageerath* and *Bhageerath-H* software
 285 suites for small globular proteins.

5. Conclusions

286

We have described here an all atom energy based com-
 287 putational methodology, *Bhageerath* for tertiary struc-
 288 ture prediction of small soluble proteins. Results on
 289 80 globular proteins show that *Bhageerath* web server
 290 predicts one or more candidate structures within an
 291

292 RMSD of 7Å from the native for proteins with less
 293 than five secondary structural elements. CASP9 results
 294 reflect the potential of the protocol to predict a struc-
 295 ture within 10Å RMSD from the native for sequences
 296 with no known sequence homologs. For proteins with
 297 local sequence and structure matches involving short
 298 fragments, it is expedient to use a hybrid method such
 299 as *Bhageerath-H* for tertiary structure prediction. In
 300 a nutshell, for small proteins the structure prediction
 301 problem is under control with *ab initio* methods, and
 302 for larger proteins computational protocols involving
 303 hybrid models are getting better and better.

304 Acknowledgements

305 Funding from the Department of Information Technol-
 306 ogy and programme support from the Department of
 307 Biotechnology, Government of India to SuperComput-
 308 ing Facility for Bioinformatics is gratefully acknowl-
 309 edged.

310 References

- 311 1. Creighton T E 1990 *Biochem. J.* **270** 1
- 312 2. Dobson C M 2003 *Nature* **426** 884
- 313 3. Editorial 2005 *Science* **309** 78
- 314 4. Unger R and Moulton J 1993 *Bulletin of Mathematical*
 315 *Biology* **55** 1183
- 316 5. Fraenkel A S 1993 *Bulletin of Mathematical Biology* **55**
 317 1199
- 318 6. Baker D 2000 *Nature* **405** 39
- 319 7. Klepeis J L and Floudas C A 2004 *SIAM News* **37** 1
- 320 8. Venkatraman J, Shankaramma S C and Balaram P 2001
 321 *Chem. Rev.* **101** 3131
- 322 9. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P,
 323 Klepeis J L, Dror R O and Shaw D E 2010 *Proteins* **78**
 324 1950
- 325 10. Levitt M and Warshel A 1975 *Nature* **253** 694
- 326 11. McCammon J A, Gelin B R and Karplus M 1977 *Nature*
 327 **267** 585
- 328 12. Li A and Daggett V 1995 *Protein Eng.* **8** 1117
- 329 13. Daggett V and Levitt M 1992 *Proc. Natl. Acad. Sci. USA*
 330 **89** 5142
- 331 14. Levitt M 1983 *J. Mol. Biol.* **168** 595
- 332 15. Levitt M 1983 *J. Mol. Biol.* **168** 621
- 333 16. Tirado-Rives J and Jorgensen W L 1993 *Biochemistry*
 334 **32** 4175
- 335 17. Boczek E M and Brooks C L III 1995 *Science* **269** 393
- 336 18. Demchuk E, Bashford D and Case D A 1997 *Fold. Des.*
 337 **2** 35
- 338 19. Daura X, Jaun B, Seebach D, van Gunsteren W F and
 339 Mark A E 1998 *J. Mol. Biol.* **280** 925
- 340 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 342
22. Zagrovic B, Sorin E J and Pande V S 2001 *J. Mol. Biol.*
313 151 343
23. Simmerling C, Strockbine B and Roitberg A E 2002 *J.*
Am. Chem. Soc. **124** 11258 344
24. Snow C D, Zagrovic B and Pande V S 2002 *J. Am.*
Chem. Soc. **124** 14548 345
25. Freddolino P L, Liu F, Gruebele M and Schulten K 2008
Biophys. J. **94** L75 346
26. Freddolino P L, Park S, Roux B and Schulten K 2009
Biophys. J. **96** 3772 347
27. Voelz V A, Bowman G R, Beauchamp K and Pande V S
2010 *J. Am. Chem. Soc.* **132** 1526 348
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** 341 349
29. Allen F *et al.* 2001 *IBM Systems Journal* **40** 310 350
30. Shirts M and Pande V S 2000 *Science* **290** 1903 351
31. Liwo A, Khalili M and Scheraga H A 2005 *Proc. Natl.*
Acad. Sci. USA **102** 2362 352
32. Ensign D L, Kasson P M and Pande V S 2007 *J. Mol.*
Biol. **374** 806 353
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** 756 354
34. Bonneau R and Baker D 2001 *Annu. Rev. Biophys.*
Biomol. Struct. **30** 173 355
35. Petrey D and Honig B 2005 *Mol. Cell* **20** 811 356
36. Moulton J, Pedersen J T, Judson R and Fidelis K 1995
Proteins **23** ii 357
37. Moulton J, Fidelis K, Kryshtafovych A, Rost B and
Tramontano A 2009 *Proteins* **77** 1 358
38. Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat
T N, Weissig H, Shindyalov I N, Bourne P E 2000 *Nucl.*
Acids Res. **28** 235 359
39. Floudas C A, Fung H K, McAllister S R, Monnigmann
M and Rajgaria R 2006 *Chem. Eng. Sci.* **61** 966 360
40. Jayaram B, Bhushan K, Shenoy S R, Narang P, Bose S,
Agrawal P, Sahu D and Pandey V 2006 *Nucl. Acids Res.*
34 6195 361
41. Narang P, Bhushan K, Bose S and Jayaram B 2005 *Phys.*
Chem. Chem. Phys. **7** 2364 362
42. Narang P, Bhushan K, Bose S and Jayaram B 2006 *J.*
Biomol. Struct. Dyn. **23** 385 363
43. Hubbard S J and Thornton J M 1993 *NACCESS Com-*
puter Program (London: Department of Biochemistry
and Molecular Biology, University College London) 364
44. Kelley L A and Sternberg M J E 2009 *Nature Protocols*
4 363 365
45. Roy A, Kucukural A and Zhang Y 2010 *Nature Proto-*
cols **5** 725 366
46. Zhang Y 2008 *BMC Bioinformatics* **9** 1 367
47. Kim D E, Chivian D and Baker D 2004 *Nucl. Acids Res.*
32 W526 368
48. Soding J, Biegert A and Lupas A N 2005 *Nucl. Acids*
Res. **33** W244 369
49. <http://predictioncenter.org/casp9/> 370

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES.

- Q1. Please provide Graphical Abstract.
- Q2. “developing rapidly to enable prediction of maturing in the tertiary...” Please check changes made if appropriate.
- Q3. Please check tables 1–4 for correctness.
- Q4. Please provide shortened running title.
- Q5. Is et al. usage allowed in reference section?

UNCORRECTED PROOF