Use of secondary structural information and Cα-Cα distance restraints to model protein structures with MODELLER

BOOJALA V B REDDY[1,*] and YIANNIS N KAZNESSIS[2]

[1]*Laboratory of Bioinformatics and In Silico Drug Design, Department of Computer Science, Queens College, CUNY 65-30 Kissena Blvd, Flushing, NY 11367, USA*

[2]*Digital Technology Center and Department of Chemical Engineering and Materials Science, University of Minnesota Minneapolis, MN 55455, USA*

**Corresponding author (Fax, 718-997-3513; Email, breddy@qc.cuny.edu)*

Protein secondary structure predictions and amino acid long range contact map predictions from primary sequence of proteins have been explored to aid in modelling protein tertiary structures. In order to evaluate the usefulness of secondary structure and 3D-residue contact prediction methods to model protein structures we have used the known Q3 (alpha-helix, beta-strands and irregular turns/loops) secondary structure information, along with residue-residue contact information as restraints for MODELLER. We present here results of our modelling studies on 30 best resolved single domain protein structures of varied lengths. The results shows that it is very difficult to obtain useful models even with 100% accurate secondary structure predictions and accurate residue contact predictions for up to 30% of residues in a sequence. The best models that we obtained for proteins of lengths 37, 70, 118, 136 and 193 amino acid residues are of RMSDs 4.17, 5.27, 9.12, 7.89 and 9.69, respectively. The results show that one can obtain better models for the proteins which have high percent of alpha-helix content. This analysis further shows that MODELLER restrain optimization program can be useful only if we have truly homologous structure(s) as a template where it derives numerous restraints, almost identical to the templates used. This analysis also clearly indicates that even if we satisfy several true residue-residue contact distances, up to 30% of their sequence length with fully known secondary structural information, we end up predicting model structures much distant from their corresponding native structures.

[Reddy B V B and Kaznessis Y N 2007 Use of secondary structural information and Cα-Cα distance restraints to model protein structures with MODELLER; *J. Biosci.* **32** 929–936]

1. Introduction

Several analyses of known protein structures are carried out and hundreds of methods have been published to predict secondary structure using amino acid sequence information (Chandonia and Karplus 1996, 1999; Hung and Samudrala 2003; Jiang 2003; Lin *et al* 2004; Petersen *et al* 2000; Wood and Hirst 2004; Wu *et al* 2004). The secondary structure prediction accuracies of several top performing methods have been evaluated and compared. From recent reviews we have recognized that the Q3 (alpha-helix, beta-strands and irregular turns/loops) prediction accuracy of 80% to 82% is achieved by most of the best performing methods (Wu *et al* 2004) with considerable variation that depends on the structural class of the proteins. This area appears to be at a matured state with no significant further improvement. There are also numerous computational methods developed to predict 3D residue contact maps from the sequence information (Fariselli and Casadio 1999; Fariselli *et al* 2001a,b; Hamilton *et al* 2004; MacCallum 2004), stating that such predictions are useful

Keywords. Model evolution; protein modelling; residue contact prediction; secondary structure prediction

Abbreviations used: BRP, Buried residue restraint pairs; CRP, conserved residue restraint pairs; GDT, global distance test; GRP, general restraint pairs; RMSD, root-mean square deviation; SRP, surface residue restraint pairs; TMS, template model score

in protein modelling. However, we do not see any attempts showing quantitatively the usefulness of these prediction results for accurate modelling of protein structures. Here we have attempted to test how best one can model the protein structures using the residue secondary structures and 3D contact information using MODELLER (Fiser and Sali 2003; Sali and Blundell 1993), a popularly used, comparative protein restrain optimization program.

Here we present our modelling studies on 30 well resolved protein structures of different lengths with varied Q3 secondary structure composition. For these 30 structures we have also computed the residue contact information by identifying the residues that have 6.5 Å or less distance between their Cα atoms that are 5 or more residues in distance apart. We have then built random models using the Q3 secondary structure information of these proteins. We then generated final models applying the various combinations of known Cα-Cα distance restraints using MODELLER, a powerful comparative modelling program that generates possible models of protein structures by satisfaction of spatial restraints of amino acids observed in known protein structures. Here we discuss the results obtained from this modelling exercise.

2. Materials and methods

From the Protein Data Bank (PDB) we have selected 1.8 Å or better resolution X-ray defined single domain protein structures of lengths 30 to 200 residues. We have used only the protein structures for which coordinates available for all the residues, and which have at least 8 distant homologous sequences in the NR sequence database at NCBI Blast site (*http://www.ncbi.nlm.nih.gov/BLAST/*). We selected a total of 30 proteins which satisfied these conditions (table 1). Secondary structures are assigned to each residue using the DSSP method (Kabsch and Sander 1983) as helix (H), strand (E) or coil (C) for each residue. We have obtained distant homologues of each of these sequences in the non-redundant set of sequence data (NR) and calculated conservation index of each position (Reddy and Kaznessis 2005). Using the structural coordinates, we have computed the solvent accessible surface area (Lee and Richards 1971) of each residue to identify buried residues and surface residues in the protein structure.

2.1 *Calculation of spatial restraints*

For each of the structure we have identified the interacting pairs of Cα atoms that are 6.5 Å or less distance in structure and grater than 5 residues apart in the sequence. All such pairs of C_αs are stored as general restraint pairs (GRP). Within the GRP we have grouped the pairs of amino acids whose positions are highly conserved depending on their conservation index value (Reddy and Kaznessis 2005) as conserved residue restraint pairs (CRP). The pairs of residues whose inaccessible surface area is greater than 40% are grouped as buried residue restraint pairs (BRP) and the residue pairs which are accessible greater than 40% of their surface area are grouped as surface residue restraint pairs (SRP).

2.2 *Modelling of protein structures*

We have taken the amino acid sequence of each of the protein and the corresponding secondary structure of each residue in Q3 form as helix (H), extended structure (E) or coil (C). A random seed structure was generated for each of the sequences using ProteinShop (Crivelli *et al* 2004) software that builds alpha-helix and extended beta-strand structures for residues defined as H and E respectively, in the input file and gives a random phi- and psi- angles for residues defined as C.

We have used such generated random structure as a seed template for MODELLER program to generate final models. MODELLER that generates possible models of protein structures by satisfaction of spatial restraints of amino acid Cα distances given as input parameters. The restraints in principle can be derived by number of different sources such as related protein structures (as used in comparative protein modelling), NMR experiments, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy, site-directed mutagenesis, intuition, residue-residue and atom-atom potentials of mean force etc. The restraints can operate on distances, angles, dihedral angles and other spatial features defined by atoms and pseudo atoms. For comparative modelling MODELLER automatically derives restraints from the given related structures and their alignment with the target sequence. MODELLER optimizes the molecular probability density function with variable target function procedure in cartesian space that employs methods of conjugate gradients and molecular dynamics with simulated annealing.

We have given seed structure generated by ProteinShop with additional restraints: (i) Not to disturb the residues in the rigid secondary structures in helices and beta-strands but only to move the residues in the coil structures. (ii) Restrain the distances of selected pairs of Cα atoms from the GRP, CRP, SRP and BRP groups separately. The Cα-Cα distances were restrained to be 5.5 ±1.0 Å and generating the model structures.

2.3 *Model evaluation*

In order to evaluate the resultant models, (i) the root-mean square deviation (RMSD) of the Cα atoms of the model structures compared to the native structure, in a globally

optimized superposition of the two structures, as one of the measures of similarity to asses the models. However, improper modelling of a small region of the protein can some times lead to large variation in RMSD. Therefore we have used several other measures introduced by the Critical Assessment for Structure Prediction (CASP) contest evaluators (Moult *et al* 2003) to evaluate these model structures. These are (ii) Global distance test (GDT) (Zemla *et al* 1999), which identifies largest set of residues deviating from the target by no more than a specified C*α* distance cutoff using many different methods of superpositions. (iii) MaxSub (Siew *et al* 2000), is a measure aims at identifying the largest subset of C*α* atoms of a model that superimpose 'well' over the experimental structure, and produces a single normalized score that represents the quality of the model. (iv) Template model score (TMS) is a new scoring function introduced recently by (Zhang and Skolnick 2004) by extending the approaches used in GDT and MaxSub scores. We have used all these four measures to evaluate the models.

3. Results

The proteins we have used are single domain proteins, fall in all major secondary structural classifications namely,

Table 1. Best possible models generated using MODELLER with randomly selected CA-distances and rigid secondary structures as restraints

ID	Length	Lowest RMSD	Best MaxSub	Best GDT	Best TMS
1ajj	37	4.17 (7443)	0.419 (6026)	0.520 (6026)	0.285 (3047)
2fdn	55	7.17 (0257)	0.224 (2109)	0.311 (2093)	0.254 (2342)
2igd	61	4.76 (7461)	0.373 (3119)	0.537 (7442)	0.387 (7400)
1nxb	62	5.97 (6435)	0.324 (4396)	0.415 (3252)	0.339 (5278)
1aho	64	7.49 (6436)	0.289 (3006)	0.386 (7029)	0.300 (3006)
1utg	70	5.27 (7064)	0.398 (5053)	0.539 (5053)	0.457 (5053)
1vcc	77	7.63 (3374)	0.246 (0185)	0.351 (4326)	0.314 (5340)
1bdo	80	9.86 (6450)	0.233 (7046)	0.284 (7064)	0.274 (7046)
1opd	85	6.83 (6081)	0.328 (6005)	0.421 (6005)	0.393 (6005)
1aba	87	8.69 (3262)	0.275 (6047)	0.374 (6047)	0.351 (6047)
1gvp	87	10.67 (7020)	0.207 (2161)	0.267 (7418)	0.270 (7482)
1bm8	99	8.24 (4296)	0.266 (7474)	0.354 (7474)	0.366 (7474)
1plc	99	9.95 (6000)	0.168 (1052)	0.240 (6024)	0.271 (6000)
3vub	101	10.07 (7485)	0.199 (6465)	0.285 (7463)	0.280 (7463)
1bkf	107	10.94 (1074)	0.184 (3044)	0.262 (3346)	0.287 (3044)
2mcm	112	9.90 (7068)	0.185 (2192)	0.259 (6464)	0.282 (7070)
2tgi	118	9.12 (7464)	0.224 (6065)	0.337 (6065)	0.375 (6065)
2mhr	118	9.14 (7063)	0.302 (2217)	0.379 (1232)	0.417 (2217)
2sak	121	11.19 (3281)	0.198 (3281)	0.312 (3281)	0.353 (3281)
1whi	122	11.41 (6033)	0.169 (3145)	0.246 (3145)	0.300 (3145)
1bgf	124	13.14 (4151)	0.329 (2157)	0.355 (4341)	0.380 (4341)
3lzt	129	11.68 (3105)	0.190 (3020)	0.244 (5065)	0.280 (6457)
1c52	131	9.90 (6052)	0.231 (3079)	0.313 (6062)	0.363 (0065)
1eca	136	7.89 (3066)	0.280 (6018)	0.384 (7050)	0.432 (6050)
2hbg	147	13.45 (7475)	0.164 (5242)	0.207 (4282)	0.261 (7407)
1a6m	151	10.85 (3174)	0.287 (5059)	0.358 (5025)	0.441 (5025)
1koe	172	15.60 (5075)	0.119 (5062)	0.198 (4351)	0.277 (4351)
1gbs	185	12.79 (4271)	0.151 (7061)	0.226 (5209)	0.297 (4254)
2pth	193	9.69 (6482)	0.184 (4308)	0.257 (7074)	0.391 (7074)
1iab	200	12.47 (7093)	0.190 (6485)	0.231 (6461)	0.324 (6485)

Various model evaluation measures with reference to the native structure are given. The model number is given in the parenthesis.

Table 2. Average RMSD values for the best 10 models generated using MODELLER with varied number of restraints of different groups of residues

Len	ID	10%A	10%C	10%B	10%S	15%A	15%C	15%B	15%S	20%A	20%C	20%B	20%S	30%A	30%T
37	1ajj	*6.68*	7.85	6.74	6.76	7.30	*6.08*	7.93	7.40	*5.60*	5.95	5.95	7.48	**5.31**	5.68
55	2fdn	8.80	9.75	8.85	**8.00**	10.43	9.60	9.75	*9.51*	10.39	10.29	10.75	*9.92*	*10.31*	10.53
61	2igd	10.33	*5.78*	13.20	6.78	*6.07*	6.34	10.18	6.65	8.33	*5.50*	10.78	7.32	6.15	**5.03**
62	1nxb	10.23	9.76	*9.69*	10.30	9.27	9.25	*7.54*	10.54	10.12	9.10	*6.78*	10.48	9.68	**6.74**
64	1aho	*8.96*	10.98	10.01	10.05	9.43	10.94	*9.10*	9.23	*9.12*	9.96	9.22	9.22	8.67	**8.20**
70	1utg	*5.86*	6.76	7.12	6.60	5.54	*6.72*	7.34	6.82	9.18	6.86	7.59	*6.85*	**5.47**	**5.47**
77	1vcc	10.61	20.16	10.97	*9.14*	9.10	20.34	12.44	*8.34*	10.14	18.84	11.85	**8.04**	*8.89*	11.14
80	1bdo	*11.58*	12.70	13.89	16.20	13.63	*11.62*	13.60	16.46	12.18	*11.04*	13.02	17.23	10.99	**10.86**
85	1opd	*10.38*	10.50	10.66	12.41	*8.86*	10.73	10.32	11.23	11.14	10.80	*10.09*	12.04	8.05	**7.63**
87	1aba	12.11	13.01	11.50	*10.64*	11.23	12.52	*10.41*	12.05	13.61	12.25	**9.71**	11.03	10.59	*10.49*
87	1gvp	15.40	14.68	14.73	*14.57*	14.54	14.94	*12.25*	14.62	14.68	*13.19*	14.44	14.44	**11.23**	12.77
99	1bm8	*8.77*	12.50	12.79	11.35	11.16	12.47	*10.66*	10.75	12.46	12.46	*11.15*	11.59	11.34	**8.48**
99	1plc	14.20	*12.70*	15.59	14.14	11.90	13.08	14.31	*11.41*	11.94	13.80	13.80	*11.54*	**10.52**	11.43
101	3vub	14.49	14.92	15.02	*12.71*	*11.39*	12.95	21.87	12.96	*11.08*	11.42	20.56	13.80	12.89	**10.30**
107	1bkf	**11.67**	16.79	17.86	16.50	*14.11*	16.09	16.48	15.52	*13.30*	15.87	17.25	14.56	13.84	*13.63*
112	2mcm	*12.70*	17.62	15.83	30.74	13.43	15.18	*10.90*	28.86	13.19	14.93	*12.29*	27.99	**9.11**	10.24
118	2mhr	10.44	*10.23*	11.10	10.26	10.56	*10.14*	11.28	10.38	11.40	11.08	12.13	*10.65*	**9.52**	9.85
118	2tgi	11.44	*11.30*	15.13	24.16	11.28	*10.36*	12.73	22.91	*9.82*	10.54	23.41	23.41	10.96	**9.50**
121	2sak	17.67	13.97	*13.05*	19.38	12.62	12.54	*11.94*	17.29	12.93	12.34	**11.61**	16.30	*12.25*	14.37
122	1whi	17.47	*16.01*	17.34	37.75	*14.63*	16.78	15.07	22.64	16.33	16.75	*13.34*	21.21	**11.93**	12.23
124	1bgf	15.82	14.66	*14.36*	15.27	**13.97**	14.26	14.75	15.29	17.74	16.14	16.96	*15.48*	*14.36*	15.24
129	3lzt	16.88	14.13	20.85	*12.22*	15.07	*12.91*	18.07	16.75	13.18	**12.04**	17.10	14.16	12.74	*12.52*
131	1c52	*11.13*	11.78	13.67	11.39	12.40	13.43	13.10	*10.78*	*11.84*	13.38	13.40	12.60	**10.25**	11.45
136	1eca	9.95	10.15	*8.80*	9.34	9.22	*8.63*	8.86	9.25	*8.74*	9.15	9.14	9.56	8.85	**8.54**
147	2hbg	14.55	*14.32*	14.76	14.56	*13.79*	14.21	14.74	14.43	*14.08*	14.22	14.22	14.36	13.80	**13.68**
151	1a6m	14.84	*14.19*	15.31	15.88	*11.57*	12.34	14.07	14.83	14.09	**11.19**	15.01	14.87	*13.73*	14.41
172	1koe	*17.45*	19.78	17.79	17.79	*15.79*	19.83	16.70	17.59	16.98	19.17	*16.27*	17.55	16.64	**15.99**
185	1gbs	16.42	18.26	*15.28*	15.54	13.10	17.43	**13.00**	16.35	14.01	17.61	*13.17*	16.28	15.11	*15.00*
193	2pth	14.02	16.08	13.70	*12.19*	12.16	15.55	14.38	*10.92*	11.43	15.67	13.68	*11.33*	10.26	**10.14**
200	1iab	19.81	*18.38*	19.86	18.49	*15.08*	16.13	15.61	17.46	20.53	20.53	*15.75*	17.62	**13.65**	13.77

The 10%, 15%, 20% or 30% indicates if the protein is of length 100 residues 10%, 15%, 20% or 30/% Cα-Cα distances, respectively, have been restrained to obtain final models using randomly chosen restraints from all residues (A), Conserved residues (C), burried residues (B) or surface residues (S). The lowest values in each row are shown bold and the lowest value in a class of restraints is shown italic-underlined.

Table 3. Average TMS scores for the ten best models generated using MODELLER with different number of restraints of different groups of residues

Len	ID	10%A	10%C	10%B	10%S	15%A	15%C	15%B	15%S	20%A	20%C	20%B	20%S	30%A	30%T
37	1ajj	0.231	0.207	*0.244*	0.239	0.217	0.228	0.227	*0.246*	*0.251*	0.233	0.233	0.228	**0.253**	0.243
55	2fdn	0.213	0.223	0.201	**0.228**	0.195	*0.212*	0.199	*0.212*	*0.206*	0.200	0.198	0.202	*0.212*	0.202
61	2igd	0.311	*0.360*	0.318	0.334	*0.353*	0.339	0.306	0.342	0.306	*0.340*	0.308	0.312	0.349	**0.373**
62	1nxb	0.260	0.261	0.253	*0.281*	0.297	0.243	*0.300*	0.281	0.267	0.234	*0.296*	0.268	0.263	**0.307**
64	1aho	0.269	0.239	0.272	*0.283*	**0.289**	0.247	0.266	0.267	*0.288*	0.252	0.287	0.287	*0.283*	0.282
70	1utg	*0.403*	0.385	0.361	0.379	**0.433**	0.368	0.355	0.361	0.342	*0.366*	0.345	0.340	0.419	*0.422*
77	1vcc	0.255	**0.285**	0.210	0.269	*0.272*	0.276	0.246	*0.290*	0.237	*0.283*	0.236	0.283	*0.261*	0.235
80	1bdo	0.234	*0.249*	0.215	0.189	0.201	*0.237*	*0.237*	0.224	0.227	0.229	*0.235*	0.204	**0.261**	0.249
85	1opd	0.264	0.291	*0.296*	0.272	0.275	*0.295*	0.293	0.276	0.252	0.289	*0.301*	0.280	**0.354**	0.332
87	1aba	0.250	0.269	*0.273*	0.262	0.247	0.255	*0.282*	0.268	0.238	0.255	*0.273*	0.259	**0.319**	0.283
87	1gvp	0.204	*0.233*	0.229	0.198	0.210	0.207	*0.239*	0.203	*0.207*	0.203	0.202	0.202	0.246	**0.255**
99	1bm8	*0.308*	0.249	0.273	0.273	0.274	0.270	**0.333**	0.249	0.250	0.250	*0.329*	0.236	0.316	*0.330*
99	1plc	*0.233*	0.217	0.187	0.232	0.196	0.194	0.185	*0.240*	0.208	0.203	0.203	**0.248**	*0.244*	0.226
101	3vub	*0.231*	0.225	0.197	0.218	*0.257*	0.234	0.215	0.238	*0.259*	0.231	0.228	0.226	0.243	**0.265**
107	1bkf	0.229	0.191	*0.238*	0.200	0.212	0.197	*0.270*	0.213	**0.272**	0.205	0.264	0.251	0.236	*0.252*
112	2mcm	0.211	*0.244*	0.196	0.176	0.224	0.242	*0.257*	0.195	0.237	0.237	*0.246*	0.188	0.268	**0.270**
118	2mhr	0.399	0.395	**0.401**	0.394	0.391	0.380	*0.393*	*0.393*	0.392	0.378	*0.396*	0.385	0.382	*0.387*
118	2tgi	*0.285*	0.263	0.197	0.213	0.249	*0.291*	0.266	0.251	**0.310**	0.252	0.239	0.239	0.288	*0.307*
121	2sak	0.216	*0.248*	0.246	0.237	0.270	0.258	*0.296*	0.210	0.264	0.278	**0.323**	0.228	0.279	*0.283*
122	1whi	*0.208*	0.201	0.206	0.156	*0.255*	0.250	0.214	0.176	0.235	*0.270*	0.248	0.190	0.281	**0.290**
124	1bgf	0.332	0.336	0.300	**0.354**	0.305	0.303	0.299	*0.332*	0.278	0.281	0.289	*0.326*	*0.306*	0.289
129	3lzt	0.229	0.230	*0.246*	0.240	*0.265*	0.253	0.231	0.236	0.260	0.257	*0.268*	0.245	**0.271**	0.265
131	1c52	0.327	*0.337*	0.291	0.295	0.301	0.267	0.243	*0.322*	*0.313*	0.285	0.259	0.277	**0.346**	0.315
136	1eca	*0.382*	0.364	0.372	0.370	*0.403*	0.398	0.401	0.368	0.402	0.405	*0.406*	0.351	**0.422**	0.406
147	2hbg	0.243	*0.251*	0.244	0.223	0.248	**0.253**	0.247	0.221	*0.252*	0.250	0.250	0.219	**0.253**	**0.253**
151	1a6m	0.287	*0.317*	0.284	0.293	**0.432**	0.354	0.305	0.332	0.338	*0.366*	0.316	0.332	*0.344*	0.339
172	1koe	0.221	0.220	*0.232*	0.219	0.225	0.200	0.225	**0.263**	*0.245*	0.205	0.221	0.235	*0.233*	0.225
185	1gbs	0.247	0.233	0.266	*0.269*	0.243	0.213	**0.294**	0.253	0.250	0.221	*0.287*	0.241	0.270	*0.275*
193	2pth	0.284	0.279	0.252	*0.296*	0.285	0.251	0.244	*0.338*	0.309	0.270	0.252	*0.330*	**0.360**	0.353
200	1iab	*0.249*	0.229	0.228	0.228	*0.251*	0.217	0.241	0.211	*0.261*	*0.261*	0.243	0.218	0.281	**0.309**
AVERAGE		0.267	0.267	0.257	0.261	0.276	0.264	0.270	0.267	0.272	0.266	0.273	0.250	0.295	0.294

The 10%, 15%, 20% or 30% indicates if the protein of length 100 residues 10, 15, 20 or 30 Cα-Cα distances, respectively, have been restrained to obtain final models using the randomly chosen restraints from all residues (A), conserved residues (C), burried residues (B) or surface residues (S), . The highest values in each row are shown bold and the highest value in a group of restraints is shown italic-underlined.

all-alpha, alpha/beta and all-beta protein classes. Length of these proteins varies from 37 to 200 amino acids. As described in the method we restrained the major secondary structure elements, the helices and beta-strands, and generated random models of protein structures. In the table 1 we have listed the best possible models that were obtained for each protein according to the four different model evaluation measures described in the methods section. The GDT, MaxSub and TMS values range between 1.0 - 0.0 , 1.0 is given by the perfect model that has 0.0 RMSD compared to the native structure. The TMS measure is said to be comparable among the proteins of different lengths (Zhang and Skolnick 2004). It can be seen from table 1 that the highest TMS of 0.457 is obtained for protein 1utg, which is an alpha-helical protein. In fact all the proteins which have higher TMS value are predominantly alpha-helical proteins. It is clear from the table 1 that we could not obtain a single useful model (lower than 4.0 Å RMSD or better than 0.6 TMS value) in all our modelling exercise. However, here we present the data showing how the models are dependent on the number and the type of Cα-Cα residue distance constraints used.

We have generated 2800 model structures for each protein starting with the random model template given by ProteinShop program. We have used a number of Cα-Cα distance restraints, equivalent to 10%, 15%, 20% and 30% of the length of protein sequence. In other words, a number of 10% restraint pairs for a protein of length 100 residues, we use randomly 10 pairs of Cα atoms each time from the GRP, CRP, SRP or BRP groups of pairs, and restrained their Cα-Cα distances to generate models by MODELLER. We modeled 100 structures for every selected group of Cα-Cα distance restraints. The groups of Cα-Cα distance pairs represent specific set of residues such as in CRP we have pairs of residues that are highly conserved in the family of homologous sequences. The BRP set of pairs represent buried residue pairs in the protein structures which are hydrophobic and shown to be involved in nucleation of the structure in the protein folding process. The SRP set of residues are on the surface region of protein structure.

In table 2 we have given the average RMSD values of ten best models when compared to their experimental structures. The RMSD value for each structure in each group of residue Cα-Cα distance pairs is give. In order to study how the increase in number of distance constraints affects the resultant models, we have chosen number of Cα-Cα pair equivalent to 10%, 15%, 20% and 30% of the length of the modeled protein structure. For example for the protein 1ajj which is of length 37 residues we have randomly taken 4, 6, 8 and 11 Cα-Cα pair for 10%, 15%, 20% and 30% respectively, from the GRP, CRP, BRP and SRP set of pairs separately. We constrained the Cα-Cα distances to 5.5 Å with 1.0 Å as standard deviation. In the case of 30% number of Cα-Cα pairs group we have chosen the constraints from GRP set and a combined set, CRP+BRP+SRP pairs, named as TRP. This is because we do not have sufficient number of Cα-Cα pairs and by grouping all conserved, buried and surface residue pairs we choose restraints from different regions of the structure. It is again clear that using even such a high number of residue-residue distance constraints and Q3 secondary structure information we are unable to model any structure, at least with 4.0 Å RMSD compared to the native structure.

It can be seen from the table 2 that the models constrained using higher number (30%) of Cα-Cα pair distances have given better models as lowest RMSD models are mostly in 30% number constraints. In other words the more the number of distance constraints we use the better we are able to model the structure. However, this is not clearly seen as a trend when we compare the lowest RMSD values (*see* italic-underlined values in table 2) in each constrain-number-group of 10%, 15%, 20% and 30%. For a significant number of structures we see a high RMSD value with higher number of constrains used in modelling. This shows that it is not only number of constraints but also the position of constrained residues in the structure may be important to obtain better models. In each constrain-number-group in table 2 we have given sub groups that pertain to the kind of residues restrained such as general (GRP), conserved (CRP), buried (BRP) or surface (SRP). The RMSD values do not show any preference for specific subgroup, the lowest RMSD values are approximately equal in all the sub groups.

In table 3 we have given average TMS score, which suppose to be length independent measure, for the best ten structures for each of the constrain-number-groups and their sub-groups for all the 30 structures studied. The TMS scores also shows that the models obtained using higher number of restrains give better models. The distribution of the TMS values, in each of the subgroups, do not indicate any preference in restraining the specific type of residues to obtain better models.

4. Discussion

The analysis presented here evaluates the usefulness of comparative protein modelling software MODELLER which is popularly used to model protein structures using homologous structures as template(s). This analysis clearly shows that MODELLER calculates numerous restraints that lead to copying the entire template structure(s) in the process of building model. Though it is claimed to be a restrain-based model it severely fails to model a useful structure with limited number of Cα-Cα distances as restraints. Any selective use of Cα-Cα restraints such as conserved key residues or residues buried in the structures did not improve the models. This shows that the secondary

structure predictions and contact predictions will not help much unless we improve the restraint based modelling programs. The other homology restraint based modelling programs include a method by Brocklehurst and Perham 1993, GENECOMP (Kolinski *et al* 2001), a method by Zhang *et al* (2002) and PERMOL ((Moglich *et al* 2005), not popularly used in the literature but needs to be tested for their predictive capabilities.

5. Conclusions

The analysis indicates that predicting the Q3 secondary structures and residue contact information can not help much to obtain useful models of protein structures. We also need to improve the methods that use the given restraints while obtaining the complete model of the structure. The results shows that it is very difficult to obtain useful model even with 100% accurate secondary structure predictions and accurate residue contact predictions for up to 30% of residues in a sequence. The best models that we obtained for proteins of lengths 37, 70, 118, 136 and 193 amino acid residues are of RMSDs 4.17, 5.27, 9.12, 7.89 and 9.69 respectively and strongly dependent on secondary structure content with better models for the proteins which have high percent of alpha-helix content. This analysis further shows that MODELLER restrain optimization program can be useful only if we have truly homologous structure(s) as a template. This analysis also clearly indicates that even if we satisfy several true residue-residue contact distances with fully known secondary structural information we end up predicting model structures much distant from their corresponding native structures.

References

Brocklehurst S M and Perham R N 1993 Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipoylated H-protein from the pea leaf glycine cleavage system: a new automated method for the prediction of protein tertiary structure; *Protein Sci*. **2** 626–639

Chandonia J M and Karplus M 1996 The importance of larger data sets for protein secondary structure prediction with neural networks; *Protein Sci.* **5** 768–774

Chandonia J M and Karplus M 1999 New methods for accurate prediction of protein secondary structure; *Proteins* **35** 293–306

Crivelli, S, Kreylos O, Hamann B, Max N and Bethel W 2004 ProteinShop: a tool for interactive protein manipulation and steering; *J. Comput. Aided Mol. Des.* **18** 271–285

Fariselli P and Casadio R 1999 A neural network based predictor of residue contacts in proteins; *Protein Eng*. **12** 15–21

Fariselli P, Olmea O, Valencia A and Casadio R 2001a Prediction of contact maps with neural networks and correlated mutations; *Protein Eng.* **14** 835–843

Fariselli P, Olmea O, Valencia A and Casadio R 2001b Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations; *Proteins Suppl*. **5** 157–162

Fiser A and Sali A 2003 Modeller: generation and refinement of homology-based protein structure models; Methods Enzymol. **374** 461–491

Hamilton N, Burrage K, Ragan M A and Huber T 2004 Protein contact prediction using patterns of correlation; *Proteins* **56** 679–684

Hung L H and Samudrala R 2003 Accurate and automated classification of protein secondary structure with PsiCSI; *Protein Sci.* **12** 288–295

Jiang F 2003 Prediction of protein secondary structure with a reliability score estimated by local sequence clustering; *Protein Eng.* 16 651–657

Kabsch W and Sander C 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features; *Biopolymers* **22** 2577–2637

Kolinski A, Betancourt M R, Kihara D, Rotkiewicz P and Skolnick J 2001 Generalized comparative modelling (GENECOMP): a combination of sequence comparison, threading, and lattice modelling for protein structure prediction and refinement; *Proteins* **44** 133–149

Lee B and Richards F M 1971 The interpretation of protein structures: estimation of static accessibility; *J. Mol. Biol*. **55** 379–400

Lin K, Simossis V A, Taylor W R and Heringa J 2004 A simple and fast secondary structure prediction method using hidden neural networks; *Bioinformatics* **21** 152–159

MacCallum R M 2004 Striped sheets and protein contact prediction; *Bioinformatics (Suppl.)* **20** I224–I231

Moglich A, Weinfurtner D, Gronwald W, Maurer T and Kalbitzer H R 2005. PERMOL: restraint-based protein homology modelling using DYANA or CNS; *Bioinformatics* **21** 2110–2111

Moult J, Fidelis K, Zemla A and Hubbard T 2003 Critical assessment of methods of protein structure prediction (CASP)-round V; *Proteins* (*Suppl. 6)* **53** 334–339

Petersen T N, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert G P *et al* 2000 Prediction of protein secondary structure at 80% accuracy; *Proteins* **41** 17–20

Reddy B, and Kaznessis Y 2005 A Quantitative analysis of amino acid position conservation in interface regions of protein-protein hetrocomplexes; *J. Bioinfo. Comput. Bio.* **3** 1137–1150

Sali A and Blundell T L 1993 Comparative protein modelling by satisfaction of spatial restraints; *J. Mol. Biol.* **234** 779–815

Siew N, Elofsson A, Rychlewski L and Fischer D 2000 MaxSub: an automated measure for the assessment of protein structure prediction quality; *Bioinformatics* **16** 776–785

Wood M J and Hirst J D 2004 Predicting protein secondary structure by cascade-correlation neural networks; *Bioinformatics* **20** 419–420

Wu K P, Lin H N, Chang J M, Sung T Y and Hsu W L 2004 HYPROSP: a hybrid protein secondary structure prediction algorithm--a knowledge-based approach; *Nucleic Acids Res.* **32** 5059–5065

Zemla A, Venclovas C, Moult J and K Fidelis K 1999 Processing and analysis of CASP3 protein structure predictions; *Proteins (Suppl.)* **3** 22–29

Zhang C, Hou J and Kim S H 2002 Fold prediction of helical proteins using torsion angle dynamics and predicted restraints; *Proc. Natl. Acad. Sci. USA* **99** 3581–3585

Zhang Y and Skolnick J 2004 Scoring function for automated assessment of protein structure template quality; *Proteins* **57** 702–710

*e*Publication: 21 June 2007