Computational Biology and Chemistry 42 (2013) 40–48

Contents lists available at SciVerse ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Understanding the general packing rearrangements required for successful template based modeling of protein structure from a CASP experiment

Ryan Day[a], Hyun Joo[b], Archana C. Chavan[b], Kristin P. Lennox[a], Y. Ann Chen[c], David B. Dahl[d], Marina Vannucci[e], Jerry W. Tsai[b,*]

[a] Lawrence Livermore National Labs, Livermore, CA 94550, United States
[b] Department of Chemistry, University of the Pacific, Stockton, CA 95211, United States
[c] Department of Biostatistics, Moffitt Cancer Center, Tampa, FL 33612, United States
[d] Department of Statistics, Brigham Young University, Provo, UT 84602, United States
[e] Department of Statistics, Rice University, Houston, TX 77251, United States

ARTICLE INFO

Article history:
Received 18 May 2012
Received in revised form 30 October 2012
Accepted 31 October 2012

Keywords:
Protein packing
Loop modeling
Template-based protein structure prediction
Protein statistical function

ABSTRACT

As an alternative to the common template based protein structure prediction methods based on main-chain position, a novel side-chain centric approach has been developed. Together with a Bayesian loop modeling procedure and a combination scoring function, the Stone Soup algorithm was applied to the CASP9 set of template based modeling targets. Although the method did not generate as large of perturbations to the template structures as necessary, the analysis of the results gives unique insights into the differences in packing between the target structures and their templates. Considerable variation in packing is found between target and template structures even when the structures are close, and this variation is found due to 2 and 3 body packing interactions. Outside the inherent restrictions in packing representation of the PDB, the first steps in correctly defining those regions of variable packing have been mapped primarily to local interactions, as the packing at the secondary and tertiary structure are largely conserved. Of the scoring functions used, a loop scoring function based on water structure exhibited some promise for discrimination. These results present a clear structural path for further development of a side-chain centered approach to template based modeling.

1. Introduction

Template based protein structure prediction methods (TBM) have the potential to rapidly 'solve' the structures of many gene sequences of unknown structure. The explicit aim of the structural genomics initiative is to solve the structures of sequences without obvious homology to known structures, increasing the number of templates available for TBM. Despite these efforts, the Protein Data Bank (Berman et al., 2000) has seen virtually no growth in the number of protein folds in the last several years, suggesting that most soluble, globular protein folds have been discovered. As the Critical Assessment of Protein Structure Prediction (CASP) experiments have shown, the ability of predictors to significantly improve these templates' similarity to the target structure remains unimproved (Read and Chavali, 2007; Moult, 2005; Mariani et al., 2011; Cozzetto et al., 2009; Keedy et al., 2009; Kryshtafovych et al., 2011). In this work, we characterize the packing rearrangements that need to be modeled to move a template closer to the native structure.

Current template based methods rely on variations of backbone based chain assembly methods (Eswar et al., 2006; Joo et al., 2007; Zhang et al., 2005; Zhang, 2008; Krieger et al., 2009). In the constraint based approach (Eswar et al., 2006; Joo et al., 2007), template structures are used to define short and long range distances that act as constraints on atom positions. A simulated annealing procedure is then used to generate backbone models, which can be scored or clustered and scored, before rebuilding and packing the side-chains. In the fragment-based approaches (Zhang et al., 2005; Zhang, 2008; Krieger et al., 2009), a threading procedure is used to identify template structure, and these templates are then used to identify short peptide fragments. These fragments are sampled and reassembled to search the backbone conformational space. What is common between these methods and the majority of protein structure prediction approaches is the sampling of conformational space based solely on the protein backbone and without direct influence of the amino acid side-chains (Moult, 2005; Fiser, 2010; Qu et al., 2009). Information about the location of the side-chain centers of mass is retained indirectly in the backbone fragments, but the move space is fundamentally defined by the backbone fragments. This approach seems at odds with the basic idea behind template based structure modeling: because the proteins share the same fold, it

* Corresponding author. Tel.: +1 209 946 2298; fax: +1 209 946 2607.
E-mail address: jtsai@pacific.edu (J.W. Tsai).

1476-9271/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2012.10.008

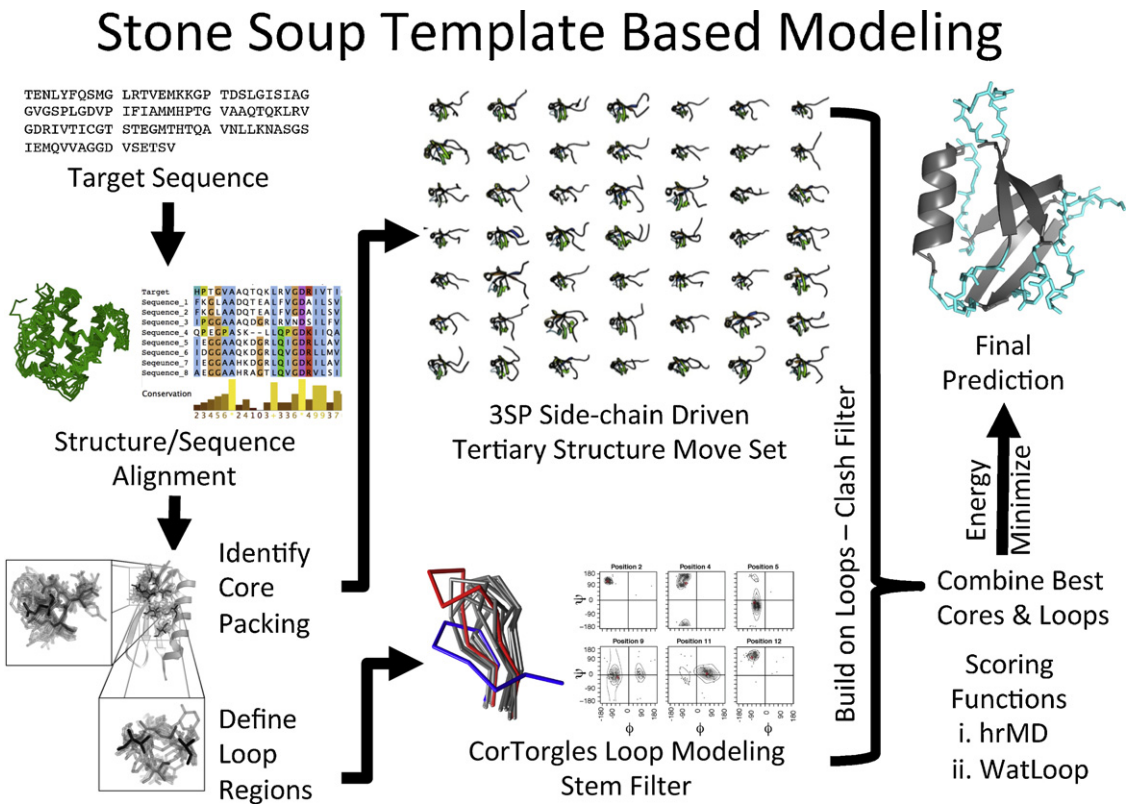

Fig. 1. Stone Soup flowchart. Individual steps are detailed in the methods and results sections.

is the changes in sequence and therefore the contributions from amino acid side-chains that determines the differences between template and native protein structures. Sampling protein conformational space of packing using backbone move sets may account for the current lack of progress in template-based modeling.

In contrast to these backbone based-methods, our group has been actively seeking approaches to capture the side-chain influences on the protein backbone (Holmes and Tsai, 2005; Dahl et al., 2008; Day et al., 2010; Lennox et al., 2009, 2010). In particular, we have shown that cliques in the contact graphs of proteins (i.e. sets of residues where all residues contact all other residues) can be geometrically clustered within and between protein families (Lennox et al., 2009). These clusters represent tertiary packing motifs that are common to all protein structures. For the set of structures from the 9th CASP9 experiment (Moult et al., 2011), we combined this side-chain approach with other methods such as loop modeling and conformational scoring (Joo et al., 2010) developed in the group to make template based structures predictions. Our approach seeks to make moves directly in the packing of the templates, and allow the perturbations in the side-chain packing to define the position of the backbone. The tertiary motifs also allow us to precisely analyze the results of our prediction method and exactly characterize the side-chain perturbations necessary to move a template towards the native protein structure.

2. Methods

2.1. Stone Soup template based structure prediction procedure

The Stone Soup template based structure prediction procedure is an agglomeration of a number of novel methods that approach different parts of the template based structure prediction problem (Fig. 1). The workflow is described generally in the following paragraph and more detail to each individual method is given below. At the start, the sequence provided by the CASP9 organizers was compared to all known protein structures using PSI-BLAST (Altschul et al., 1997). To identify templates, a cutoff of 20% sequence identity was used with a coverage cutoff of 90%. If multiple templates were identified, they were aligned to one another using MUSTANG (Konagurthu et al., 2006). This alignment was then used in a profile–profile alignment to the target using MUSCLE (Edgar, 2004). The aligned structures were averaged to produce a starting template structure, which was then analyzed by our novel tertiary structure prediction (3SP) method that defined regions for core refinement and the remaining for loop modeling: essentially regions with few or no 3SP constraints. For the core of the starting averaged template structure, 3SP was used to identify cliques in the templates, find similar cliques from known protein structures, and statistically model those set of cliques. Side-chain driven backbone samples were drawn from the 3SP distribution of residue cliques and substituted into an average template structure. Since this perturbation of the templates broke the backbone connectivity, all-atom models were generated using Pulchra (Rotkiewicz and Skolnick, 2008). Models were scored with the high resolution molecular dynamics (hrMD) derived volume (see below) and torsion angle score for every second structure, and the top scoring structures were identified to combine with modeled loops. Ranging from 56 to 368, the number of top scoring structures selected depended primarily on available computational resources since the number and size of targets as well as their loops varied over the course of prediction season. For the loops, the new approach of CorTorgles (Joo et al., 2011) uses template data to model backbone ϕ,ψ distributions in unstructured regions of the proteins. Samples from these distributions were converted to loops in Cartesian coordinates using the SNerf algorithm (Parsons et al., 2005). Loops were filtered out if their C-terminal stem α-carbons were greater than 2 Å of the template C-terminal stem when the N-terminal stem

α-carbons were aligned (non-closure) or if any loop α-carbon were within 3.76 Å of any protein α-carbons (backbone overlap, class score). The remaining loops were built onto the best scoring structures from the core refinement. The loops were further filtered according to the bridging water score WatLoop, a new water path based scoring function described below. Because completion of the WatLoop score did not occur until halfway through the prediction season, this scoring step was only applied to the latter half of the targets. Combining the best core and loop structures, complete all-atom models were again generated using Pulchra (Rotkiewicz and Skolnick, 2008) and scored using hrMD. Complete structures were built by selecting the best scoring set of loops for each 3SP structure identified in step 3. All these loops were then combined on each 3SP structure. Each 3SP structure was considered independently, so different base structures could have different sets of best scoring loops. All-atom models of the complete structures were once again generated using Pulchra (Rotkiewicz and Skolnick, 2008), then steepest descent minimized for 1000 steps using the OPLS force field (Kaminski et al., 2001) in Gromacs (Hess et al., 2008) and subjected to a final scoring using hrMD.

2.1.1. 3SP: side-chain driven backbone refinement

The underlying concept of 3SP is to drive backbone perturbations based on the interactions of side-chains. This is accomplished by creating a move-set library that relates side-chain packing variations in Cartesian space to the ϕ,ψ torsion angle space of the backbone main-chain. This library is generated by clustering the maximal contact cliques (Bron and Kerbosch, 1973) computed from the 95% sequence unique ASTRAL (Chandonia et al., 2004) set of known protein structures (hereafter referred to as move-set cliques) based on the relative positions of their Cα atoms and side-chain centers of mass (centroids) (Day et al., 2010). These move-set cliques represent the maximally self-interacting clusters of residues (all residues in the set are in contact with all other residues in the set). For these clustered packing cliques, the distributions of Cα and centroids at each residue position are modeled using a kernel density estimation approach (Day et al., 2010). The distribution of a given centroid position for a packing clique is a mixture of trivariate normal distributions centered on the centroid locations of known cliques. The model also permits straightforward conditional sampling, allowing perturbations at a single clique position to be propagated to other positions. To properly model the residue cliques, this statistical modeling is applied in 2 steps: first the side-chain centroid positions are modeled with respect to each other and then individual side-chain centroid positions are modeled to their respective residue's Cα position as well as backbone ϕ,ψ torsion angles.

To select a specific set of 3SP moves for a particular target, the residues for core refinement need to be identified. From the averaged template structure, maximal contact cliques (Bron and Kerbosch, 1973) are first computed. These template cliques are compared to pre-calculated library of clustered move-set cliques. The move-set cliques that are within 1.2 Å RMSD of the template cliques are pulled for modeling and are further filtered according to the distances between the Cα atoms and centroid for individual residues to ensure that the modeled positions are consistent with the target sequence. For each selected move-set clique, the modeled 3SP distribution of 1000 side-chain positions to the backbone Cα position and torsion angles constitutes the sampling of core repacking. The set of these distributions represents the overall set of moves from which draws are taken during the template-based modeling. A 3SP move consists of making draws first from the centroid distributions, and then obtaining the respective Cα position conditioned on the centroid draws. In this way, the selected side-chain positions inform changes in the backbone structure. These positions were used to build up the model structure as described above.

During the template based modeling, a model's clique is selected at random and the positions of its α-carbon and centroid atoms are changed to those of a randomly selected draw from the 3SP distribution for that clique. The move is accepted if it results in Cα–Cα distances less than 4.4 Å for consecutive residues (the maximum distance observed in the PDB) and if there are no overlaps between centroid atoms (as determined by the minimum observed distance between pairs of centroids of the 20 amino acid types in the PDB). For the next move, a new clique that shares at least two residues with the previously moved clique is selected and moved as described. If no overlapping cliques that were not moved in the previous two steps are found, a new starting clique is chosen at random and moved as described. A single run consisted of 5000 steps.

2.1.2. CorTorgles: correlated torsion angle loop modeling

Contiguous segments of two or more residues with no modeled cliques were modeled with our loop modeling algorithm CorTorgles (Joo et al., 2011), which applies a statistical estimation of continuous backbone ϕ,ψ distributions (Lennox et al., 2010). The ϕ,ψ angles for the loop region to be modeled plus two flanking residues on each side are calculated from all the templates (total residues = $n + 4$). As described in detail (Lennox et al., 2010; Joo et al., 2011), these are used to fit the parameters of a Dirichlet process mixture of bivariate von Mises distributions centered on a hidden Markov model that describes a continuous distribution in ϕ,ψ space. This unique centering distribution allowed us to develop informative template based conformation distributions even at alignment positions with little or no observed data, which allowed us to cope with sparse data and effectively extend a homology modeling approach to loop regions. Samples from this distribution are converted into Cartesian coordinates by building from the template backbone N, Cα, and carboxyl C atom positions of the first residue in the modeled segment (N-terminal stem) (Parsons et al., 2005). The resulting positions of the Cα atoms of the last two residues in the segment (C-terminal stem) are then compared to their positions in the template. If the average of these distances is greater than 2 Å, loop closure is not satisfied and the loop is rejected. Accepted loops are built onto low scoring 3SP structures by aligning the four stem residues. Loops with backbone clashes are eliminated by requiring at least 3.76 Å between non-sequential Cα atoms.

2.1.3. hrMD: high resolution MD scoring function

The hrMD scoring functions is based on an extensive set of molecular dynamics (MD) simulations or dynameome of conformations around the native state ensemble (Joo et al., 2010) and compares main-chain torsion angles, side-chain volume, and side-chain torsion angles of individual residues to values observed in the aforementioned molecular dynamics simulations. Volumes were calculated using Voronoi polyhedra (Tsai and Gerstein, 2002). In order to obtain the volumes of surface residues, the protein was inserted in an equilibrated water box in 10 randomly selected orientations and the resulting residue volumes were averaged. The hrMD score is a unified probabilistic scoring function incorporating (a) distribution of exposed polar groups (*eSA*), (b) backbone dependent residue volumes (v) and (c) backbone dependent χ_1 angles (abbreviated as χ below). These were calculated from the native state dynameome described above. The unified score S_u is proportional to how plausible each candidate structure (denoted by Γ) given the sequence information. With the assumptions of independence between individual amino acids (denoted by aa_i), and the nature of research question that we are comparing the candidate structures for the same sequence, the scoring function is reduced

to the following form:

$$\propto \prod_{i=1}^{n} p(\phi_i, \psi_i, eSA_i, \chi_i, v_i, aa_i)$$

$$= \prod_{i=1}^{n} p(v_i|\phi_i, \psi_i, aa_i)p(\chi_i|\phi_i, \psi_i, aa_i)p(eSA_i|aa_i) \quad (1)$$

The proposed unified score S_u is in the log scale of this estimated probability due to the sparsity of the knowledge space,

$$S_u = \sum_{i=1}^{n} [\log p(v_i|\phi_i, \psi_i, a) + \log p(\chi_i|\phi_i, \psi_i, aa_i) + \log p(eSA_i|aa_i)] \quad (2)$$

One issue that does bear consideration is the computation time involved in calculating the volume portion of the hrMD score. The calculation of Voronoi polyhedra is a much faster algorithm than other volume calculation algorithms, but the overall calculation is still quite slow, especially since a small solvation shell is added to the protein. The computational expense of this calculation could be minimized by introducing a clustering scheme and only scoring cluster centers or by only using the hrMD score in the later stages of prediction.

2.1.4. WatLoop: water path estimation of solvation score

The population density of the distances between polar groups connected through hydrogen bonded water network on the loop region obtained from the dynameome data was used as reference data (Joo et al., 2010). The 0.1 Å bin was used for distances ranging from 1.7 to 12.0 Å to give total 104 bins. For each distance bin, the path was counted as one if it connected from loop to loop, and as a half if it connected loop to helix or sheet. The frequencies were normalized, transformed so the highest peak has 1.0 and lowest valley has −1.0 as score. The candidates are put in the water box obtained from an MD simulation, and then the waters within 1.4 Å from protein surface are removed. Hydrogen bonded water paths between two polar groups in the loop region are searched and counted. Water paths connected through up to six waters are considered to be consistent with the reference data. This polar group distribution was compared with the dynameome data and scored. Since the water structure around the protein is not equilibrated, we rotated the candidate structures three times along each *x*, *y*, and *z* axes and searched the water paths for each rotation. The ten scores are averaged and this averaged score was used to select the good loop structures on the protein after adding the loop to the protein core structures.

2.2. Packing analysis of results

Analysis of differences in packing between templates and the native target structure was performed using the contact order defined packing cliques as described previously (Day et al., 2010). For each template structure, a sequence/structural alignment was performed using MUSTANG (Konagurthu et al., 2006). At the simplest level, the number of residues in packing cliques were compared for equivalent positions. Then, the packing clique class based on the contact order classification system was compared between equivalent sites for the template and native target structure. For example, a template packing clique of 3 + 1 (3 local residues packed against a non-local residue) would not be the same as a native packing clique of 2 + 2 (2 local residues packed against 2 non-local residues), and this would be considered a change. Lastly, cliques were compared based on position of residues in space in a similar fashion to what was done with 3SP to define the move set using a 4.4 Å RMSD cutoff between Cα atoms.

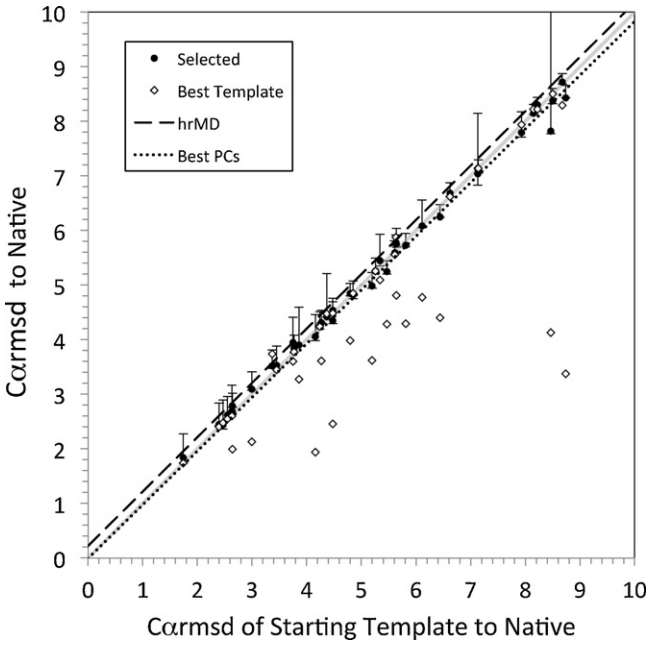

Fig. 2. 3SP scores and final scoring results. The CαRMSD of the starting template to native is always plotted along the *x*-axis, while the CαRMSD to native of the comparison set is plotted on the *y*-axis. The unity line is shown in grey across the diagonal and represents a border for good versus sets. Simply, anything below the line indicate structures closer to the native structure than the template. The best template is plotted with open diamonds. For each target, the selected model's CαRMSD to native is shown by a filled circle, while the distribution of models is shown by the whiskers from the filled circles. The fitted line of the structures built using the best packing cliques (PCs) is shown by the dotted line. The fitted line for the models selected by the hrMD scoring function is shown by the long dashed line.

3. Results/discussion

3.1. Stone Soup performance

The Stone Soup template based structure prediction algorithm was used on 59 CASP9 targets, of which native structures were released for 45. A breakdown of the Stone Soup results is shown in Table 1. Predicted targets ranged in amino acid length from 79 to 611 residues. These targets had from 1 to 67 templates in the PDB. These templates had between 61% and 100% coverage by our packing cliques. With 100% coverage, there were targets with no loops, but there were also targets with up to 21 loops. Template CαRMSD ranged from 1.74 to 19.32 Å, while final CαRMSD values from the closest of the 5 Stone Soup predictions is from 2.13 to 19.32 Å. As shown by the open diamonds in Fig. 2, the averaged template structure in general moved the starting template structure away from the native structure. While this was a major source of error and affected overall performance, it did not significantly impact the sampling and selection capabilities of our approach. Therefore, the discussion will focus on the particular results from the 3SP and CorTorgles components.

3.1.1. 3SP core packing

For each target 56–368 minimized models with all loops were generated, depending on the number of processors available (Table 1). Yet, the diversity of structures generated in 3SP was generally low. The short vertical bars around the unity line in Fig. 2 indicate that 3SP sampled conformational space only around the starting template structure. The skew in the distributions above

Table 1
CASP9 target summary.

Target	Length	Class	templates	nLoops	nProc	Score	coverage	template	start	final	pdbid
T0520	189	H	14	3	352	hrMD	1.00	3.60	3.75	3.92	3mr7
T0547	611	H	7	16	368	hrMD	0.76	8.29	2.68	8.89	3nzp
T0549	84	H+S	20	1	368	hrMD	1.00	7.14	7.13	7.26	2kzv
T0563	279	H+S	3	11	176	hrMD	0.99	5.87	5.63	5.86	3on7
T0565	326	H+S	2	15	184	hrMD	0.95	8.22	8.15	8.61	3npf
T0570	258	H+S	11	6	120	hrMD	0.96	3.74	3.37	4.04	3no3
T0573	311	H+S	4	15	120	hrMD	0.99	5.09	5.34	5.38	3oox
T0584	352	H	8	7	72	hrMD	0.99	4.28	5.47	5.23	3nf2
T0585	234	H+S	1	8	72	hrMD	0.76	4.85	4.85	4.88	3ne8
T0586	125	H	11	3	80	hrMD	0.73	1.94	4.16	4.43	3neu
T0589	465	H+S	11	8	96	hrMD	0.98	4.40	6.44	6.33	3net
T0591	406	H+S	57	3	88	hrMD	1.00	3.61	4.27	4.36	3nra
T0592	144	H	14	5	88	hrMD	0.98	4.13	8.74	7.91	3nhv
T0593	208	H+S	1	12	96	hrMD	0.81	8.50	8.50	8.43	3ngw
T0594	140	H	1	8	96	hrMD	0.94	2.55	2.55	3.07	3ni8
T0597	429	H+S	16	8	88	hrMD	0.90	3.37	8.74	8.47	3nie
T0599	399	H+S	12	10	88	hrMD	0.99	2.46	4.48	4.19	3os6
T0602	123	H	2	4	88	hrMD	0.93	7.94	7.93	7.90	3nkz
T0603	305	H+S	1	16	88	hrMD	0.89	13.49	13.49	12.62	3nkd
T0607	471	H+S	6	21	88	hrMD	0.98	4.47	4.37	4.50	3pfe
T0609	340	H+S	17	14	56	D+Wat	1.00	4.77	6.11	6.10	3os7
T0611	227	H+S	30	2	56	D+Wat	0.99	4.29	5.81	5.75	3nnr
T0613	287	H+S	11	8	56	D+Wat	0.99	1.99	2.64	2.98	3obi
T0615	179	H+S	1	12	56	D+Wat	0.97	15.89	15.89	15.82	3nqw
T0617	148	H+S	39	0	56	D+Wat	0.99	4.81	5.64	5.79	3nrv
T0620	312	H+S	3	19	56	D+Wat	0.98	5.56	5.61	5.92	3nr8
T0623	220	H+S	3	8	56	D+Wat	0.82	3.98	4.80	4.90	3nkh
T0625	233	H	1	17	56	D+Wat	0.98	13.15	13.15	13.04	3oru
T0626	283	H+S	17	6	56	D+Wat	1.00	2.13	3.00	3.16	3o1l
T0632	168	H+S	27	8	56	D+Wat	1.00	2.05	13.05	13.07	3nwz
T0636	336	H+S	67	0	56	D+Wat	1.00	3.27	3.86	3.89	3p1t
T0638	269	H+S	2	13	56	D+Wat	0.84	19.32	19.32	19.32	3nxh
T0640	250	H+S	56	7	56	D+Wat	1.00	3.62	5.20	4.19	3nyw
T0641	296	H+S	10	10	56	D+Wat	1.00	2.60	2.63	2.84	2nyi
TR530	115	R	NA[5]	2	96	hrMD	0.70	2.47	2.47	2.48	2npp
TR557	145	R	NA	7	56	D+Wat	0.86	4.48	4.48	4.63	2kyy
TR567	145	R	NA	7	56	D+Wat	0.98	3.77	3.77	3.95	3n70
TR568	158	R	NA	6	56	D+Wat	0.61	8.22	8.22	8.15	3n6y
TR569	79	R	NA	3	56	D+Wat	1.00	3.45	3.45	3.31	2kyw
TR574	126	R	NA	8	56	D+Wat	0.81	4.24	4.24	4.38	3nrf
TR576	172	R	NA	8	56	D+Wat	0.80	7.14	7.14	7.26	3na2
TR592	144	R	NA	6	56	D+Wat	0.73	1.74	1.74	2.13	3nhv
TR594	140	R	NA	8	56	D+Wat	1.00	2.40	2.40	2.75	3ni8
TR606	169	R	NA	10	56	D+Wat	0.73	5.26	5.26	5.34	3noh
TR622	138	R	NA	4	56	D+Wat	0.88	6.62	6.62	6.82	3nkl

Target is the target number used in CASP9. *Length* refers to the number the amino acid residues for a target. *Class* indicates classification type during CASP9, where H is human, S is server, and R is refinement. The *templates* column shows the number of templates used in our approach, where the NA for all refinement targets indicates that the only structure used was supplied by CASP9. The *nLoops* refers to the number of loops that were modeled by WatLoop. The *nProc* label is the number of structures generated at each step and is proportional to the number of processors used. *Score* indicates which scoring functions were used, where the hrMD scoring function was used for all targets and the water distance based filter was applied to 'D+Wat' targets. Coverage is the percentage coverage of the template to the native structure. The *template*, *start*, and *final* columns are Cα/side-chain center of mass RMSD in Å. The last column *pdbid* is the PDB identification of the native structure.

the unity line shows the sampling was more away from the native structure than towards it. This is partly attributable to our sampling scheme, which was essentially a random walk, with no scoring function to allow us to keep "good" moves and reject "bad" moves. However, the more fundamental issue is that our method is too conservative in its move-set since it relies heavily on information from the template structure. Because the 3SP repacks the conserved residues found in the template core, the approach does not contain those new clique conformations that make the difference between that template and native structure. Even if we used the closest template, our sampling of packing space remains close to the starting structure. As shown in Fig. 2, some targets began with that closest native template, but this did not improve our sampling. To test that limitation of move-set library, model structures for each target were built up by selecting the cliques from our library that were closest in CαRMSD to the template cliques. The dotted line in Fig. 2 shows the average improvement of about 0.2 Å CαRMSD to native, which lies just below the unity line. The best improvement found was just under 0.5 Å CαRMSD and our worst was an increase in CαRMSD from native of 1.2 Å. This increase was due to small lever arm effects in a region where the native had strained backbone torsion angles nearer to disallowed regions. This result shows that the limit of this 3SP approach is not a significant improvement over the starting structure.

3.1.2. CorTorgles loop modeling

For all but two targets, there were regions of the protein that could not be modeled by 3SP. This could be due to lack of coverage in the templates, unpacked residues, or rare packing arrangements that were not well represented in the PDB. We consider all of these cases as candidates for loop modeling and modeled the ϕ,ψ distributions for these residues using the DPM-HMM method detailed in Lennox et al. (2010). Once the ϕ,ψ distributions are calculated, making draws and building putative structures is very fast, allowing us to generate 1,000,000 models for each loop region. However, the stem filter, which enforces loop closure, removes the vast majority of these loops and only a small fraction could be grafted back onto

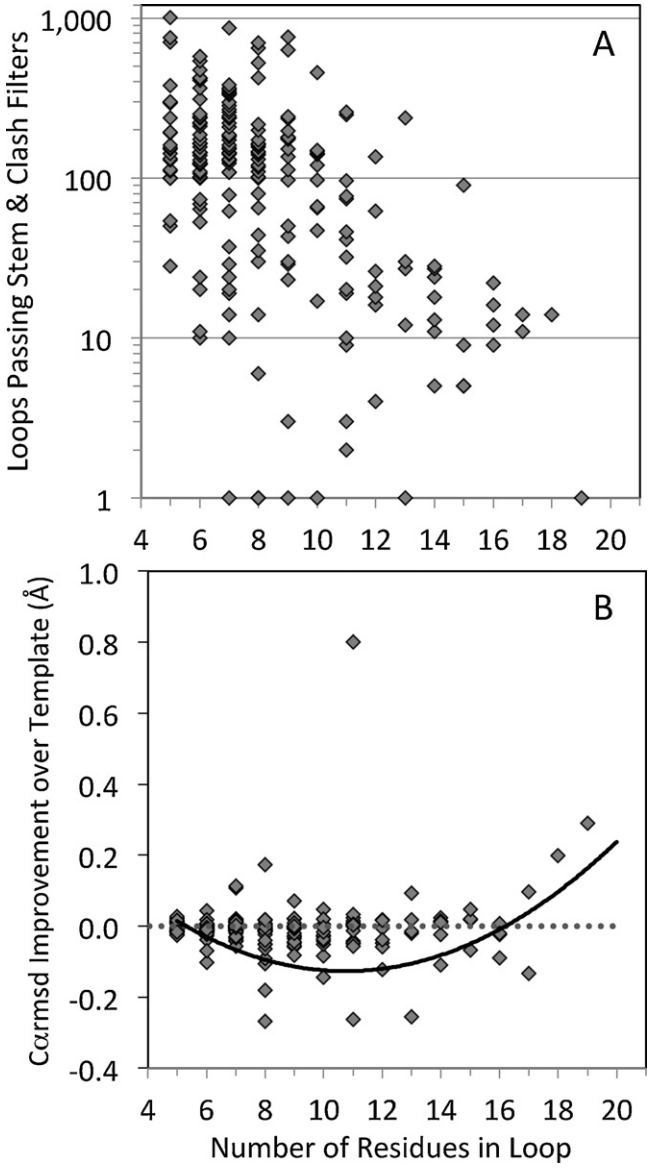

Fig. 3. Loop modeling results. (A) Number of loops that pass the loop closure (stem) filter and clash filter as a function of loop length. In order to save processor time, the clash filter was stopped after 1000 loops had passed. We initially generated 1 million draws for each loop. (B) Whole template Cα/side-chain center of mass RMSD improvement when best loop is built onto template as a function of loop length. The line represents the fitted line of lowest energy loops selected using the WatLoop scoring function.

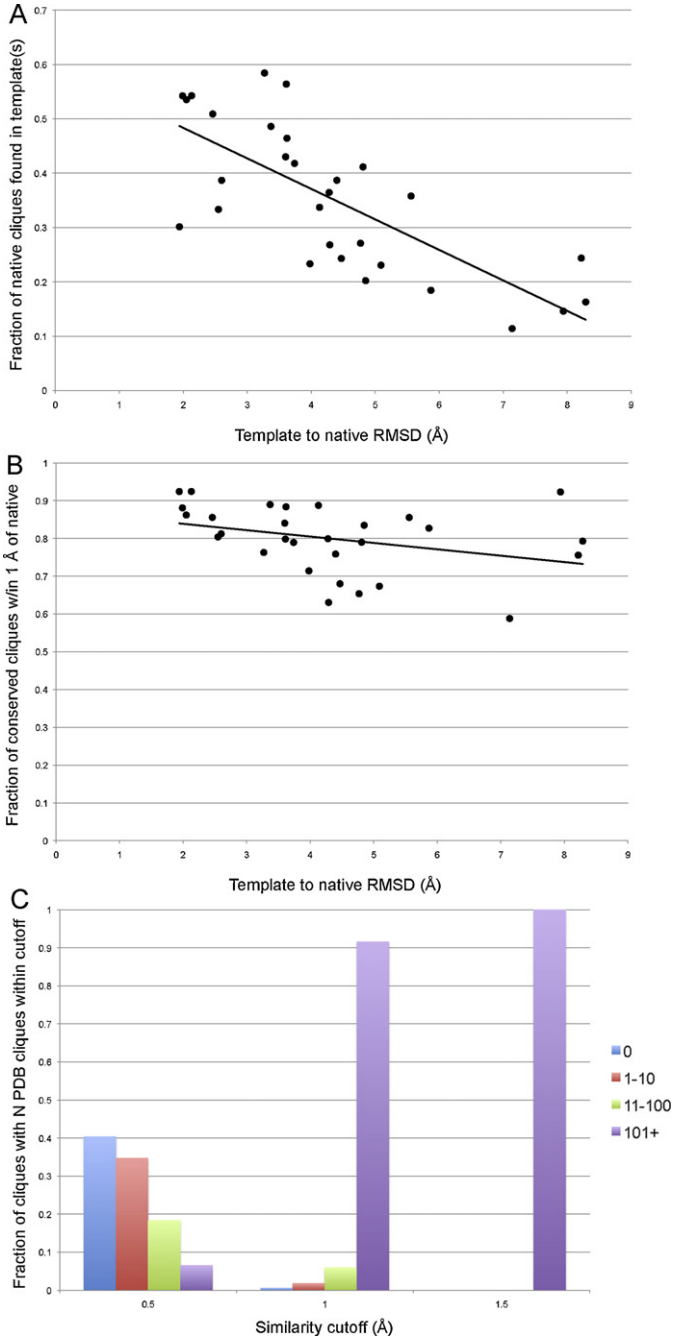

Fig. 4. Similarity between target cliques and template cliques. (A) Fraction of target cliques that are found in any template as a function of the Cα/side-chain center of mass RMSD between the target and best template. (B) Fraction of template cliques that are present in the target and within 1 Å RMSD of the target clique as a function of the RMSD between the target and best template. (C) Similarity between target cliques and PDB cliques. The colors represent the fraction of target cliques with N cliques found in the PDB within the distance cutoff on the *x*-axis. Thus, for ~40% of target cliques there are no cliques in our PDB set within 0.5 Å RMSD, but for ~90% of target cliques there are more than 100 cliques in our PDB set within a 1.0 Å RMSD cutoff. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

the template structure (Fig. 3A). The median number of remaining loops after applying the stem filter is 572. These loops are then built onto the best scoring 3SP models and a backbone clash filter is applied, leading to a further reduction in the number of structures that must be scored. As Fig. 3B shows, it was difficult to improve on loops of shorter length from 5 to 7 residues. In this regime, the starting loops usually began very close to native leaving little room for improvement. At the other end of the spectrum, longer loops of 16 and up were not sampled well by CorTorgles, and increases in CαRMSD to native is seen. CorTorgles exhibits its best performance from 8 to 15 residues, where improvements to the overall template by the loops pushed 0.3 Å. As the template put constraints on the starting and end points of the loop as well as the path it takes to connect those points, this improvement is a significant contribution to building better models for loops in the 8–13 residue length range.

3.2. Scoring functions

Addressing the effectiveness of the scoring function used in this study is difficult as we generally did not move the templates significantly closer to or further from the target (Fig. 2). Using just the hrMD scoring function to select from the distribution of model structures on average selected structures on average that were just

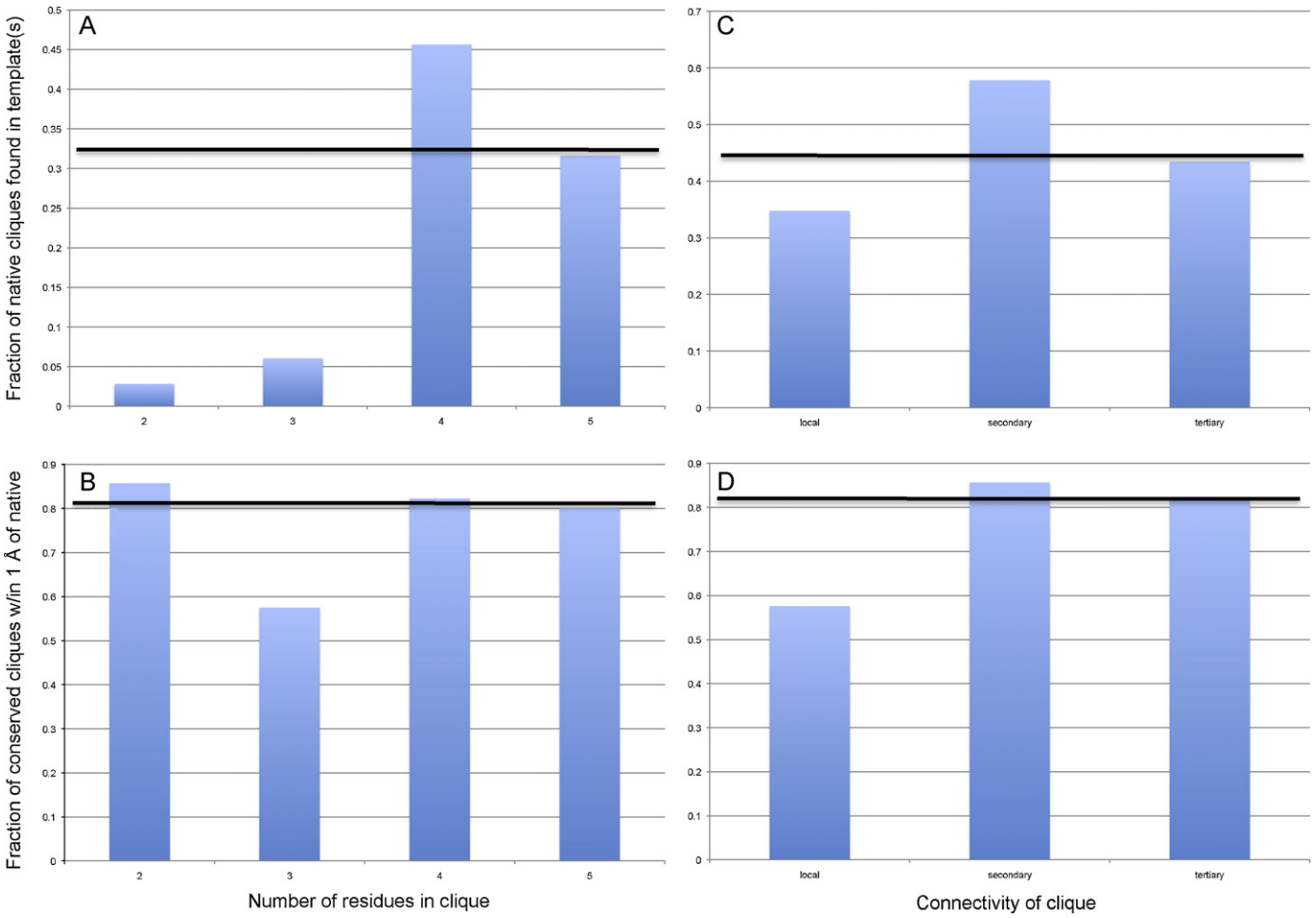

Fig. 5. Clique conservation and similarity by clique type. In A and B, the cliques are grouped according to their size (i.e. the number of residues in each clique). In C and D, the 4 residue cliques are classified according to whether the residues in the clique are local in sequence (local), are within the same element of h-bonded secondary structure (secondary), or connect multiple secondary structure elements (tertiary). A and C show clique conservation, i.e. the fraction of target cliques for which there is a clique in one or more templates formed by residues in the same alignment positions. B and D show the fraction of conserved template cliques that are within 1.0 Å RMSD of the target clique. In all figures, the horizontal black line indicates the average value for all cliques.

slightly worse than the starting template structure. The long dashed line in Fig. 2 shows this average and indicates that hrMD performs consistently regardless of how close the template structure starts to the native. The hope was that using the molecular dynamics data would be able to discriminate structures that were closer to native, which is not the trend shown by the line. As scoring function based on physical principles of a protein structure, these results are consistent and suggest the limitation of a score like hrMD. Since the hrMD scores structures on their physical reasonableness, deviations that unfold or perturb a fold would not be allowed. For structures far from native that require large rearrangement, hrMD would score movement away from the template structure poorly. In a similar manner, the hrMD only selects close structures with templates that are nearer to native. Therefore, the hrMD scoring function is good at keeping the structure stable, but inappropriate for sampling across conformational space.

Adding the water path distance filter (WatLoop) displayed a small improvement over hrMD alone, so the improvement of the template that WatLoop provided was investigated for the loops. Fig. 2B shows the fitted line to the average improvement of loops selected by WatLoop. Consistent with the ability of the CorTorgles to make loops, the WatLoop was able to generally find the better candidates. In 39% of the 213 loops modeled, WatLoop selected loops that were farther from native. In the remaining 61%, WatLoop was able to select loops that moved the structure towards the native structure. In 3 instances, WatLoop was able to select the best loop made by CorTorgles. In each of these, the starting structure was below 3 Å CαRMSD to the native. Since the WatLoop scoring function relies on the network of waters around the loop and its respective structure, the WatLoop scoring function is promising for refinement in template based structure prediction when the template structure is close to the native structure. Alternatively, these results suggest that WatLoop should be used in the final steps of model sampling when the model structure is hopefully closer to the native structure to allow the WatLoop discrimination.

3.3. *Packing rearrangements*

Overall, the templates possessed a certain amount of variation in their packing from the samples and representation in the PDB as explored by Fig. 4. Even for templates that are geometrically similar to the target (Cα/side-chain center of mass CαRMSD < 4 Å), the fraction of cliques that are identical in target and templates is always less than 60% and may be as low as 30% (Fig. 4A). As expected, clique conservation is even lower for poorer templates. In contrast, the structural similarity between cliques that are conserved is high regardless of the similarity between target and template (Fig. 4B). Thus, if regions of low clique conservation could be predicted, we could be confident that the remaining regions provide a good template for packing in the template. Another issue that may affect a

packing centered template based modeling approach is the completeness of the PDB in describing different packing arrangements. This issue is investigated by considering the number of representatives in the filtered PDB set for each target clique in all CASP9 targets. The three Cα/side-chain center of mass RMSD cutoffs were considered for defining representatives. At the shortest cutoff, 0.5 Å, the PDB set appears to be quite incomplete. Three in four target cliques have fewer than 10 representatives in the PDB that are <0.5 Å RMSD. The PDB appears to be much more complete when a 1.0 Å RMSD cutoff is used, with less than 5% of cliques having fewer than 10 representatives. At a 1.5 Å RMSD cutoff, the PDB is essentially complete. While the incompleteness of the PDB at the 0.5 Å RMSD cutoff suggests a lower limit on the resolution of packing based approaches to TBM, the clique conservation issues discussed above represent a much larger practical challenge.

A simple analysis of the types of cliques that are conserved provides some insight into what regions of the protein are more likely to be different in the target and template (Fig. 5). Cliques formed by four residues are by far the most common in the PDB. These also appear to be the most likely to be conserved between target and template. Two and three residue cliques are poorly conserved in the CASP9 targets (Fig. 5A) by a significant margin. Even when they are conserved, three residue cliques are less likely than larger cliques to be geometrically similar to the template clique (Fig. 5B). Taken together, these observations at first suggest that regions of the protein with many two or three residue cliques are poorly packed and more sensitive to changes in sequence. Based on a new analysis (Joo et al., 2012), isolated changes in two and three residue cliques occurs in less than 9% of the cases for two and three body cliques. The 91% majority of changes in two and three body cliques results from repacking and rearrangements of three and four body clique packing. Cliques can also be classified according to the backbone connectivity of the residues in the clique.

In Fig. 4C and D, we classify four residue cliques as local if all residues are near each other in primary sequence, secondary if they all come from the same element of hydrogen bonded secondary structure (i.e. consecutive turns of a helix or neighboring strands in a sheet), and tertiary if one or more residues is neither local nor hydrogen bonded to the other members of the clique. We find that 4 residue local cliques are the least conserved and the least structurally similar. At 82%, most of these local cliques are found in regions classified as loops. The remaining 18% are those local cliques that start in defined secondary structure and extend into loop regions. Secondary cliques defined by relatively rigid secondary structural elements are more likely to be conserved than tertiary cliques, which are more sensitive to changes in the detailed orientation of different secondary structure elements.

4. Conclusion

Stone Soup, a novel side-chain based packing algorithm coupled with a new loop modeling protocol, was tested against the CASP9 set of template based homology modeling targets (Mariani et al., 2011; Kinch et al., 2011). An analysis of Stone Soup's performance indicated that the scoring functions are limited and the approach's move set is overly conservative. As a physical scoring function, hrMD restricts a model structure from sampling into physically unreasonable regions of conformational space. The WatLoop function shows promise, but requires the core structure to be close to the native (<3 Å) to perform well. For the move set, larger perturbations from the template structure need to be included, because template and target side-chain packing contacts tend to differ significantly even if the sequences are close homologs. Therefore, improvement requires the prediction of regions packing rearrangements between template and native structure, which has been a core problem in template based structure prediction.

To this end, a recent development in our group has the characterization of a basic principle underlying protein packing of knobs into sockets that allows us to identify the amino acid code for protein structure (Joo et al., 2012). The knob-socket construct allows us to identify the exact changes in packing between the template and native structure that need to be modeled. In effect, sockets identify regions of protein structure that will form or not form interactions with other parts of the protein. If there is an interaction, the socket packs with a socket. So, perturbations in structure that either create or remove tertiary packing can be identified and then move sets can be sampled using 3SP. Inclusion of the knob-sockets and the corresponding amino acid code has great potential and should improve our predictions.

Acknowledgment

This work is funded by the National Institutes of Health grants R01GM81631 and R01GM104972.

References

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25 (17), 3389–3402.
Berman, H.M., et al., 2000. The Protein Data Bank. Nucleic Acids Research 28 (1), 235–242.
Bron, C., Kerbosch, J., 1973. Finding all cliques of an undirected graph. Communications of the ACM 16 (9), 575–577.
Chandonia, J.M., et al., 2004. The ASTRAL Compendium in 2004. Nucleic Acids Research 32 (Database issue), D189–D192.
Cozzetto, D., et al., 2009. Evaluation of template-based models in CASP8 with standard measures. Proteins 77 (Suppl. 9), 18–28.
Dahl, D.B., Bohannan, Z., Mo, Q., Vannucci, M., Tsai, J., 2008. Assessing side-chain perturbations of the protein backbone: a knowledge-based classification of residue Ramachandran space. Journal of Molecular Biology 378 (3), 749–758.
Day, R., Lennox, K.P., Dahl, D.B., Vannucci, M., Tsai, J.W., 2010. Characterizing the regularity of tetrahedral packing motifs in protein tertiary structure. Bioinformatics 26 (24), 3059–3066.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32 (5), 1792–1797.
Eswar, N., et al., 2006. Comparative protein structure modeling using Modeller. Current Protocols in Bioinformatics 6 (Chapter 5, Unit 5).
Fiser, A., 2010. Template-based protein structure modeling. Methods in Molecular Biology 673, 73–94.
Hess, B., Kutzner, C., van der Spoel, D., Lindahl, E., 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. Journal of Chemical Theory and Computation 4 (3), 435–447.
Holmes, J.B., Tsai, J., 2005. Characterizing conserved structural contacts by pair-wise relative contacts and relative packing groups. Journal of Molecular Biology 354 (3), 706–721.
Joo, K., Lee, J., Lee, S., Seo, J.H., Lee, S.J., 2007. High accuracy template based modeling by global optimization. Proteins 69 (Suppl. 8), 83–89.
Joo, H., Qu, X.T., Swanson, R., McCallum, C.M., Tsai, J., 2010. Fine grained sampling of residue characteristics using molecular dynamics simulation. Computational Biology and Chemistry 34 (3), 172–183.
Joo, H., et al., 2011. Near-native protein loop sampling using nonparametric density estimation accommodating sparsity. PLoS Computational Biology 7 (10), e1002234.
Joo, H., Chavan, A.G., Phan, J., Day, R., Tsai, J., 2012. An amino acid packing code for alpha-helical structure and protein design. Journal of Molecular Biology 419 (3–4), 234–254.
Kaminski, G.A., Friesner, R.A., Tirado-Rives, J., Jorgensen, W.L., 2001. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. Journal of Physical Chemistry B 105 (28), 6474–6487.
Keedy, D.A., et al., 2009. The other 90% of the protein: assessment beyond the Calphas for CASP8 template-based and high-accuracy models. Proteins 77 (Suppl. 9), 29–49.
Kinch, L.N., et al., 2011. CASP9 target classification. Proteins 79 (Suppl. 10), 21–36.
Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., Lesk, A.M., 2006. MUSTANG: a multiple structural alignment algorithm. Proteins 64 (3), 559–574.
Krieger, E., et al., 2009. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in CASP8. Proteins 77 (Suppl. 9), 114–122.
Kryshtafovych, A., Fidelis, K., Moult, J., 2011. CASP9 results compared to those of previous CASP experiments. Proteins 79 (Suppl. 10), 196–207.

Lennox, K.P., Dahl, D.B., Vannucci, M., Tsai, J., 2009. Density estimation for protein conformational angles using a bivariate von Mises distribution and Bayesian nonparametrics. Journal of the American Statistical Society 104, 586–596.
Lennox, K.P., Dahl, D.B., Vannucci, M., Day, R., Tsai, J.W., 2010. A Dirichlet process mixture of hidden Markov models for protein structure prediction. Annals of Applied Statistics 4 (2), 916–942.
Mariani, V., Kiefer, F., Schmidt, T., Haas, J., Schwede, T., 2011. Assessment of template based protein structure predictions in CASP9. Proteins 79 (Suppl. 10), 37–58.
Moult, J., 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Current Opinion in Structural Biology 15 (3), 285–289.
Moult, J., Fidelis, K., Kryshtafovych, A., Tramontano, A., 2011. Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins 79 (Suppl. 10), 1–5.
Parsons, J., Holmes, J.B., Rojas, J.M., Tsai, J., Strauss, C.E., 2005. Practical conversion from torsion space to Cartesian space for in silico protein synthesis. Journal of Computational Chemistry 26 (10), 1063–1068.
Qu, X., Swanson, R., Day, R., Tsai, J., 2009. A guide to template based structure prediction. Current Protein and Peptide Science 10 (3), 270–285.
Read, R.J., Chavali, G., 2007. Assessment of CASP7 predictions in the high accuracy template-based modeling category. Proteins 69 (Suppl. 8), 27–37.
Rotkiewicz, P., Skolnick, J., 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. Journal of Computational Chemistry 29 (9), 1460–1465.
Tsai, J., Gerstein, M., 2002. Calculations of protein volumes: sensitivity analysis and parameter database. Bioinformatics 18 (7), 985–995.
Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9, 40.
Zhang, Y., Arakaki, A.K., Skolnick, J., 2005. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 61 (Suppl. 7), 91–98.