Relative Packing Groups in Template-Based Structure Prediction: Cooperative Effects of True Positive Constraints

*RYAN DAY,¹ *XIAOTAO QU,² ROSEMARIE SWANSON,³ ZACH BOHANNAN,⁴
ROBERT BLISS,³ and JERRY TSAI¹

ABSTRACT

Most current template-based structure prediction methods concentrate on finding the correct backbone conformation and then packing sidechains within that backbone. Our packing-based method derives distance constraints from conserved relative packing groups (RPGs). In our refinement approach, the RPGs provide a level of resolution that restrains global topology while allowing conformational sampling. In this study, we test our template-based structure prediction method using 51 prediction units from CASP7 experiments. RPG-based constraints are able to substantially improve approximately two-thirds of starting templates. Upon deeper investigation, we find that true positive spatial constraints, especially those non-local in sequence, derived from the RPGs were important to building nearer native models. Surprisingly, the fraction of incorrect or false positive constraints does not strongly influence the quality of the final candidate. This result indicates that our RPG-based true positive constraints sample the self-consistent, cooperative interactions of the native structure. The lack of such reinforcing cooperativity explains the weaker effect of false positive constraints. Generally, these findings are encouraging indications that RPGs will improve template-based structure prediction.

Key words: backbone conformation distance constraints, protein structure prediction, side-chain packing.

1. INTRODUCTION

STRUCTURAL GENOMICS PROJECTS ARE WORKING TO SOLVE a representative of every protein fold (Brenner et al., 1997; Chothia, 1992; Du et al., 2003; Goldsmith-Fischman and Honig, 2003; Grant et al., 2004; Liu et al., 2004; Yan and Jernigan, 2005). This will, in principle, create templates for all protein sequences. When a template is available, template-based prediction methods produce reliable and accurate models of protein

¹Chemistry Department, University of the Pacific, Stockton, California.

²Moffitt Cancer Center, Tampa, Florida.

³Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas.

⁴Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, California.

*These two authors contributed equally to this work.

structure (Baker and Sali, 2001; Ginalski, 2006; Schwede et al., 2009). These predicted structures have been used successfully in a variety of applications, such as studying the effects of mutations, designing site-directed mutagenesis, predicting binding sites, and docking small molecules in structure-based drug discovery. For the past 14 years, template-based structure prediction has evolved steadily during seven CASP experiments, but current methods still struggle to improve on good templates. Current template-based structure prediction focuses on finding the closest backbone conformation first and then fits the side chains into that backbone conformation (Floudas et al., 2006; Qu et al., 2009; Sanchez and Sali, 1997; Xiang, 2006). Unfortunately, while template structures possess backbone conformations that are well packed for the template sequence, these backbone conformations may represent an over-packed environment for side chains of the target sequence (Venclovas and Margelevicius, 2005). Thus, current refining protocols tend to move the starting structure (derived from templates) away from the native conformation (Misura and Baker, 2005; Tramontano, 2003).

A recent study on relative packing groups (RPGs) suggests a new approach to the problem of mapping sequence changes between a template and target to structural changes (Holmes and Tsai, 2005). We define an RPG as a set of residues in which all residues are in contact with one another according to the Delaunay tessellation (Delaunay, 1934) of the atom positions. Whereas radial cutoffs can fail to detect homologous contacts in related proteins if their contact distance is near the cutoff, contacts defined through use of the Delaunay tessellation can be used to accurately and automatically classify folds within a super family (Roach et al., 2005). These contacts correlate linearly up to the twilight zone with the sequence identity (Holmes and Tsai, 2005), and thus provide a direct relationship between the changes in the sequence and the changes in the structure. These contacts are also useful for understanding the structural quality of the sequence or the structure alignments, and for providing the context necessary for calculating a value for structural randomness, both of which are important for properly accessing the quality of a structural alignment.

Spatial constraints have been used extensively in structure prediction to incorporate contact information in a way that is less susceptible to alignment errors (Sali and Blundell, 1993; Wallner and Elofsson, 2005; Wallner et al., 2004; Wang and Dunbrack, 2004; Zhang, 2007; Zhang et al., 2005; Zhou et al., 2007). In this study, we use spatial constraints derived from the conserved RPGs based upon multiple sequence and structure alignment (Holmes and Tsai, 2005). Unlike other methods, constraints derived from RPGs are only between directly packed residues. Although these constraints may be long range in primary sequence, they are always local in space. Using 51 protein domains from the CASP7 experiment, including both structures selected for submission and all the other structures generated during the structure refinement, we were able to test our packing-based structure prediction method in a double-blind test. We find that (1) constraints derived from the conserved RPGs are able to improve the starting structure; (2) correct constraints (true positives), especially non-local constraints, are important for improving predictions; and (3) wrong constraints (false positives) don't affect the quality of the refinement upon the starting structure. With the right constraints, a template structure derived from multiple homologues that is 4 Å from the native conformation by C α RMSD can be improved as much as 2 Å towards the native conformation.

2. METHODS

2.1. Starting structure

The starting structure was generated using the sequence alignment program BLAST (Altschul et al., 1997), the structure alignment program MUSTANG (Konagurthu et al., 2006), and the structure modeling program MODELLER (Sali and Blundell, 1993). First, a target sequence was searched against the PDB database using BLAST, and template structures were selected based on the BLAST E-value. At least one template structure with a significant E-value was selected as the parent structure, and up to 43 templates were used for one target prediction (T0305). All template structural files were downloaded from the PDB and cleaned to remove heteroatoms and alternative conformations. MUSTANG was used to generate a structural alignment in order to calculate the conserved RPGs. Using this alignment as a guideline, up to 20 template structures were selected based on their similarity. These template structures were used to generate constraints as described below and input into MODELLER to generate starting structures using the quick build algorithm. Several possible starting structures were generated, verified and compared to the parent structure by a human expert before being selected as the final starting structure, the starting point of structure refinement.

2.2. Spatial constraints

First, RPGs of every template structure were calculated as described previously (Holmes and Tsai, 2005). The Delaunay tessellation (Delaunay, 1934) was used to determine contacting residues. RPGs are simply cliques on the resulting contact graph (Fig. 1). Those RPGs were then annotated using target sequence numbers based on the multiple sequence alignments, regardless of their residue compositions. If any member of the RPGs is in a non-aligned region, such RPGs were omitted from later calculations. In order to ensure that the packing environment for each residue is conserved as well as the sequence, constraints were defined only for RPGs that were found in at least 50% of template structures. The average distance between each pair of main-chain atoms in an RPG was used to define the spatial constraints and the standard deviation of the distance allowed variation in the constraint length. Similar procedures were used to derive the spatial distance constraints between the $C\beta$ atoms in the RPGs.

For structures with fewer than three homologues, an alternative method was used to derive the average distances and standard deviations for constraints. Each RPG derived from the template structure was compared to clusters of RPGs derived from the PDB. The cluster center with the same number of residues, the same secondary structure composition and the smallest main-chain RMSD value was chosen to represent the RPG. Up to 100 randomly selected cluster members were then used to calculate the average distance and the standard deviation as spatial distance constraints and allowed variations, respectively.

2.3. Candidate structure generation

Candidate structures were built using the NIH-Xplor (Schwieters et al., 2003) package with the spatial constraints described above. Main-chain distance constraints were applied to all targets, while side-chain constraints were only applied to targets with a high sequence identity template. Torsion angle constraints, derived from the template structures, were also used to constrain local structure. The average time for building 100 residues was about 10 hours on an Athlon 1800+ CPU. In order to sample the conformational space more efficiently, multiple XPLOR jobs were run with different random number seeds. An average of 300 candidate structures were generated for each target.

2.4. Analysis

All the structural comparisons between the native structure and structure of interest use the same sets of residues consistent with the official CASP assessment. Residues in the target sequence that didn't correspond to any native structure were omitted. Three-dimensional structure models were viewed using MacPyMol (DeLano, 2002). Figures were generated using R (Becker et al., 1988) and Microsoft Excel[®] programs. C α RMSD values were calculated using the Kabsch and Sander method (Kabsch, 1978). Using the same method, the deviation of each residue was calculated as the C α -C α distance of corresponding residues after superposing two structures. GDT-TS is the average of the percent of C α atoms within 1, 2, 4, and 8 Å of the native structure (Zemla, 2003). Higher numbers represent better structural matches, with a GDT-TS of 100 representing a perfect match between two structures.

3. RESULTS AND DISCUSSION

We based our analysis on single protein domains from the CASP7 experiment. For simplicity, we use “target” to identify each assessment unit, “template” for the homologue structure, “starting structure” for the structure we try to refine, “candidate” for any structure that was refined from the starting structure, and “predicted model” for final structures. There are a total of 51 targets from the template-based modeling (TBM) classification. Nineteen out of the 51 targets are high accuracy targets (HA). In Table 1, C α RMSD value of the starting structure, the best candidate structure, and the difference between these two are shown. Targets are ordered based on their classification and the quality of the template used (“Start RMSD”). This table also shows the relative ranks of the best submissions of each target as compared to all of the submissions to the CASP7 experiment as percentages, the GDT scores of our predicted model, and information about the constraints used in prediction. In general, our method improves template structures both by C α RMSD and GDT by providing a consistent way to sample the conformational space near the native conformation.

FIG. 1. Schematic view of RPG and derived constraints. **(A)** Schematic view of all contacts among residues for protein 1pgb (Gallagher et al., 1994) are shown. Examples of residues forming 4 member RPGs are black, and examples of sets of 4 residues that do not form an RPG are connected by dotted lines. Distances ($\text{\AA}$) of constraints between $C\alpha$ atoms are labeled for one of the RPGs. **(B)** Ribbon diagram of protein G structure (1pgb) with core residues shown.

3.1. Improving template structures

It has been shown that including information from multiple templates provides “added value” to the starting structure, which moves the starting structure closer to its native conformation (Kryshtafovych et al., 2005, 2007). In order to evaluate the ability of our method to further improve the starting structure, we compare our best candidate structures to the starting structure we generated for each target. Since the native structure was not available when the candidate structures were generated, we assumed that our starting structure was the best representation of native conformation at the time of refinement, even though it may not have been the best available template structure identified after the CASP7 experiments.

TABLE 1. C α RMSD VALUE OF THE STARTING STRUCTURE, THE BEST CANDIDATE STRUCTURE, AND THE DIFFERENCE BETWEEN THEM^a

<i>Target</i>	<i>nRes</i>	<i>Class</i>	<i>Start RMSD</i>	<i>Best RMSD</i>	Δ <i>RMSD</i>	<i>Best GDT</i>	<i>% TP</i>	<i>% TP > 8</i>	<i>% FP</i>	<i>Rank</i>
T0346	172	HA	0.51	0.83	0.32	96.2	0.82	0.62	0.28	36.8
T0345	185	HA	1.03	1.10	0.07	95.8	0.89	0.75	0.46	47.7
T0315	257	HA	1.32	1.09	-0.24	93.2	0.73	0.48	0.16	4.9
T0303D1	147	HA	1.95	2.23	0.28	80.1	0.51	0.23	0.24	33.3
T0324D1	142	HA	2.02	2.26	0.24	73.6	0.53	0.20	0.15	54.3
T0324D2	65	HA	2.12	2.91	0.79	69.6	0.53	0.15	0.19	59.8
T0340	90	HA	2.25	2.17	-0.08	96.6	0.84	0.69	0.19	5.5
T0366	106	HA	2.42	2.14	-0.28	86.9	0.70	0.47	0.25	29.8
T0308	165	HA	2.45	2.21	-0.24	83.5	0.73	0.49	0.24	19.2
T0334	530	HA	2.54	3.85	1.31	60.4	0.50	0.12	0.14	55.8
T0339D2	267	HA	2.58	2.31	-0.27	82.9	0.58	0.35	0.24	0.2
T0313	322	HA	2.60	2.67	0.07	78.0	0.70	0.48	0.26	11.8
T0359	97	HA	3.83	2.16	-1.67	80.8	0.64	0.40	0.19	68.1
T0288	93	HA	4.03	2.41	-1.62	84.6	0.72	0.44	0.33	24.1
T0302	132	HA	4.07	1.73	-2.34	82.8	0.80	0.56	0.31	3.9
T0332	159	HA	4.15	2.52	-1.63	79.9	0.60	0.31	0.12	68.5
T0305	297	HA	5.34	2.77	-2.58	77.6	0.75	0.54	0.33	78.7
T0317	163	HA	5.80	3.48	-2.31	76.3	0.65	0.36	0.17	62.7
T0326	304	HA	11.72	11.65	-0.08	48.5	0.77	0.49	0.23	92.8
T0329D1	66	TBM	1.54	1.31	-0.23	83.7	0.72	0.46	0.92	67.8
T0381D2	176	TBM	1.88	1.75	-0.13	89.1	0.70	0.38	0.19	2.9
T0378D2	142	TBM	1.96	2.03	0.07	78.9	0.50	0.24	0.47	59.1
T0381D1	61	TBM	2.25	2.79	0.54	68.0	0.40	0.03	0.63	8.9
T0371D2	121	TBM	2.40	2.72	0.32	64.1	0.63	0.29	0.36	74.7
T0341D2	104	TBM	2.46	2.47	0.01	76.0	0.62	0.26	0.23	8.1
T0318D2	335	TBM	2.50	2.91	0.41	78.8	0.73	0.46	0.33	83.4
T0341D1	148	TBM	2.77	1.86	-0.91	84.1	0.58	0.30	0.30	49.2
T0339D1	136	TBM	2.82	2.59	-0.24	78.7	0.54	0.31	0.16	22.1
T0380	145	TBM	2.89	3.35	0.47	60.7	0.41	0.12	0.14	77.6
T0330D1	153	TBM	3.24	3.36	0.12	71.4	0.44	0.17	0.16	82.9
T0371D1	162	TBM	3.70	3.63	-0.06	74.7	0.57	0.30	0.33	42.7
T0379D1	140	TBM	3.89	3.31	-0.58	65.9	0.58	0.29	0.28	38.6
T0379D2	64	TBM	4.12	3.90	-0.21	60.2	0.58	0.19	0.18	47.4
T0323D2	116	TBM	4.14	4.14	0.00	59.3	0.52	0.08	0.29	62.2
T0362	151	TBM	4.46	3.63	-0.82	68.4	0.40	0.18	0.22	47.3
T0376	321	TBM	5.09	4.32	-0.78	60.1	0.62	0.28	0.22	22.3
T0297	211	TBM	5.97	5.44	-0.54	59.6	0.65	0.33	0.35	63.4
T0375	296	TBM	6.37	5.44	-0.93	51.6	0.48	0.18	0.36	71.2
T0373	147	TBM	6.53	4.52	-2.01	55.0	0.76	0.35	0.26	83.6
T0384	325	TBM	7.06	6.15	-0.91	50.8	0.55	0.25	0.44	73.0
T0303D2	77	TBM	7.44	7.06	-0.38	39.0	0.44	0.05	0.17	75.9
T0330D2	72	TBM	8.00	8.03	0.03	31.3	0.34	0.03	0.43	88.5
T0338D2	113	TBM	8.25	6.74	-1.51	45.1	0.49	0.10	0.15	0.2
T0329D2	92	TBM	8.38	6.43	-1.95	42.9	0.57	0.18	0.32	20.3
T0318D1	154	TBM	8.53	9.58	1.05	29.9	0.41	0.10	0.27	40.5
T0338D1	143	TBM	10.00	9.37	-0.63	57.2	0.53	0.23	0.19	74.7
T0374	160	TBM	10.24	9.33	-0.91	44.4	0.49	0.20	0.34	74.1
T0323D1	101	TBM	11.38	9.23	-2.15	26.7	0.36	0.05	0.40	80.5
T0327	102	TBM	13.13	10.41	-2.72	23.3	0.43	0.04	0.35	11.2
T0382	123	TBM	17.43	14.08	-3.35	17.0	0.46	0.04	0.34	93.1
T0331	149	TBM	18.43	14.80	-3.62	16.0	0.29	0.03	0.19	45.9
T0309	76	FM	14.14	10.94	-3.21	26.2	0.60	0.04	1.15	51.4

^aAlso shown are the relative ranks of the best submissions of each target as compared to all of the submissions to the CASP7 experiment as percentages, the GDT scores of our predicted model, and information about the constraints used in prediction.

FIG. 2. C α RMSD value of best candidate structure and starting structure. C α RMSD to the native structure of the best candidate structure and the starting structure were plotted for each target. The diagonal line represents no improvement from the starting structure to the best candidate. Any data points below this line indicate an improvement on the starting structure. About two thirds of the targets have a candidate structure with smaller C α RMSD value and the majority of the rest of the targets are concentrated in an area where the starting structure has C α RMSD values range between 2 and 3.5 Å.

Figure 2 shows the C α RMSD to the native structure of the starting structure and the best candidate structure for each target. The best candidate structure is the structure generated during the refinement that is most similar to the target, not necessarily the structure selected by a scoring function. In Figure 2, any data point below the diagonal line indicates that our method improved the C α RMSD of the starting structure, making it more native-like. About two-thirds of the targets have a better candidate structure. Targets without any better candidate structures tend to have C α -RMSD less than 3.5Å.

The closer the template structure is to the native conformation, the harder the refinement becomes, and the more the refining procedure tends to move the template away from the native conformation. This difficulty can be thought of in terms of energetic and entropic effects. For starting structures that are near the native state, the volume of conformational space that is closer to the native state is much smaller than the volume of conformational space that is further from the native state, imposing a large entropic penalty on sampling conformations that are closer to the native state (Swanson et al., 2009). Balancing this entropic penalty is an energetic reward provided by satisfying native constraints. This energetic gain is also dependent on the native character of the starting structure. More native-like starting structures have already satisfied or nearly satisfied many of the native constraints, and so receive less of an energetic bonus upon becoming more native-like. In this model, it appears that the entropic effects outweigh the energetic effects for starting structures less than 3.5 Å C α -RMSD to the native structure.

3.2. RPG-derived constraints and cooperativity

The major difference between our method and other methods using spatial constraints is that the constraints used in our method were derived from the conserved RPGs. RPGs are sets of residues that are in direct, physical contact with one another, with no intervening atoms. Thus, the constraints derived from RPGs reflect sets of atoms that interact directly in a protein's tertiary structure. These RPGs are able to characterize the packing environment of the protein tertiary conformation and, when compared within the same folding family, reflect the effects of sequence changes on structure (Holmes and Tsai, 2005). We analyze our results in terms of true positive (TP) and false positive (FP) constraints. True positive (TP) constraints are present in both the template and native structures. False positive (FP) constraints are present in the template structure, but not the native structure. These constraints are reported as a fraction of the constraints found in the experimentally determined native structure; thus, the fraction of false negative constraints (FN) is simply $1 - TP$.

The effects of true and false positive constraints are explored by plotting GDT-TS as a function of TP and FP (Fig. 3). There is a clear correlation ($R = 0.70$) between the fraction of true positive constraints and the GDT-TS scores of the best candidates. No such correlation exists between the GDT-TS scores and the fraction of false positive constraints ($R = 0.02$). This trend is largely independent of the quality of the starting structure. If we limit ourselves to predictions where the starting structure was less than 6 Å C α -RMSD from the native, the correlation coefficients for TP and FP are 0.68 and 0.10, respectively. Additionally, there is surprisingly little correlation between the quality of the starting structure, as measured by C α -RMSD, and TP ($R = -0.45$) or FP ($R = 0.18$). This is probably due to the requirement that constraints are derived from multiple structures rather than the single starting structure.

FIG. 3. Best GDT-TS as a function of different classes of constraints. The GDT-TS (see Methods) is plotted as a function of the fraction of true positive (A) and false positive (B) constraints. Both true positive and false positive constraints are taken relative to the number of constraints present in the true native structure. The lines represent linear fits to the two data sets. The correlation coefficient (R) is 0.70 for the true positives and 0.02 for the false positives. (C) The GDT-TS score of the best candidate structure was plotted against the percentage of the true positive constraints between residues that are not local in sequence (separated by more than 4 residues). The line represents a linear fit with a correlation coefficient (R) of 0.77.

The lack of correlation between false positive constraints and prediction quality is attributable to the cooperative nature of protein folding. All true positive constraints are consistent with a single native conformation and positively reinforce each other to stabilize that conformation. False positive constraints come from a variety of non-native conformations and are not necessarily consistent with one another. Other structure prediction methods have found that false positive contacts have deleterious effects on prediction quality and have made efforts to filter false positives out of their datasets (Liu et al., 2009; Michino and Brooks, 2009). In their case, calculating global constraints between residues regardless of spatial locality includes so many false positives that the true positive are overwhelmed. The specificity of our method in defining constraints based on residue contacts minimizes the number of false positives and thus contributes to our method's lack of sensitivity to false positive constraints. This result suggests a viable strategy for improving prediction quality. Potentially, the specificity of our approach can include poorer sequence matches to improve coverage with more true positive constraints without being overwhelmed by the proportionally larger number of false positive constraints. Since the quality of our predictions limits the

number of false positives, the overall effect of adding these sequences would be to improve prediction quality.

The global topology of a protein is primarily constrained by non-local contacts. As such, we examine the role of constraints between residues that are separated by more than four residues in the primary sequence (Fig. 3C). There is a stronger correlation between non-local true positives and GDT-TS ($R = 0.77$) than there is between all true positives and GDT-TS ($R = 0.70$). Target T0326 is one exception, with a relatively low GDT-TS score, around 50, and a relatively high fraction of non-local true positives of 0.61. This exception is due to the fact that we chose to include residues 1 to 33 in the model for refinement. Since no conserved RPGs were found for these residues, no non-local constraints were used to refine the starting structure and these residues became an extended chain in our submitted structure. This lowered the GDT score even though the rest of the protein is relatively well predicted. In general we find that non-local contacts are disproportionally important in predicting the correct structure.

3.3. Case studies

The target T0302 largeE/H (Graber et al., 1995) is an example of a prediction where candidate structures significantly improve on an already good starting structure (Fig. 4A,B). The starting structure is generated based on 14 templates, including the best template identified by CASP7 assessors. Target T0302 2jm5 (Soundararajan et al., 2008) is an all-alpha protein with 132 residues. The multiple sequence alignment generates a starting structure with all the secondary structure elements correct, but the distance between the three terminal α -helices is clearly different from their native value. This generates most of the structure deviation observed in the starting structure. 426 out of a total of 701 candidate structures had a smaller value of C α RMSD than the starting structure's 4.1 Å. The best candidate has a C α RMSD value of 1.7 Å, which is about a 60% improvement over the starting structure. The largest structure improvement from the starting to the best candidate structure comes from the two terminal regions (Fig. 4B). In this prediction, there were many residues spread evenly throughout the sequence that had a large number of non-local true positive constraints.

Another target, T0303D1 2hsz, is a good example of an unsuccessful prediction (Fig. 4C,D). A very good starting structure was determined (1.95 Å C α -RMSD), but the modeling process moved all candidates further from the native structure. The deviations between the starting and native structure are concentrated in a loop region at the boundary between two domains (residues 95--110), suggesting that the two domains may not be completely independent. Our prediction efforts also moved several other regions closer to the

FIG. 4. Case studies of successful and unsuccessful structure predictions. Case study of target T0302 as an example of successful prediction (A,B) and target T0303D1 as an example of an unsuccessful prediction (C,D). (A,C) Superposition of three structures: native structure (light gray), best candidate structure (black), and starting structure (dark gray). The helices at the end of the starting structure are further away, as compared to the best candidate structure and the native structure. (B,D) The number of correct non-local constraints in relation to the residue's C α -C α distance from the native structure (right y-axis) for both of the starting structure (black) and the best candidate structure (gray). The gray histogram indicates the number of correct non-local constraints (left y-axis). Note the discontinuity in the residue numbering for (D) due to the omission of the independently folding second domain (T0303D2).

C-terminus further from native. In contrast to the successful prediction of T0302, which had many non-local constraints spread evenly through the sequence, there are essentially no non-local constraints in these C-terminal regions. Modeling them becomes a random walk through conformational space that is much more likely to end up further away from native than closer to native. We are currently developing alternative methods of loop prediction based on backbone ϕ, ψ pairs that may prove more effective in regions where there are few or no long range constraints (Lennox et al., 2009).

4. CONCLUSION

We analyzed our packing-oriented structure prediction algorithm's ability to improve the starting templates of 51 targets from the CASP7 experiment. Using spatial constraints derived from the conserved RPGs, we were able to improve the starting structures for a majority of the CASP7 targets. True positive constraints are important to the quality of the prediction, and false positive constraints have little effect on the prediction quality. This is because true positive constraints, as members of a self-consistent set, reinforce one another in pointing toward the true native structure, whereas false positives are not likely to reinforce each other. Among true positive constraints, non-local constraints appear to play a larger role in generating good candidates than local constraints. Taken together, these observations indicate that our methods can be further improved by including more distant homologues to increase the number of co-operative true positive constraints without incurring the penalty for adding too many false positives.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institutes of Health (NIH R01 GM81631-01).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Altschul, S.F., Madden, T.L., Schaffer, A.A., et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Baker, D., and Sali, A. 2001. Protein structure prediction and structural genomics. *Science* 294, 93–96.
- Becker, R.A., Chambers, J.M., and Wilks, A.R. 1988. *The New S Language*. Chapman & Hall, New York.
- Brenner, S.E., Chothia, C., and Hubbard, T.J. 1997. Population statistics of protein structures: lessons from structural classifications. *Curr. Opin. Struct. Biol.* 7, 369–376.
- Chothia, C. 1992. Proteins. One thousand families for the molecular biologist. *Nature* 357, 543–544.
- DeLano, W.L. 2002. *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA.
- Delaunay, B. 1934. Sur la sphere vide. *Izv Akad Nauk SSSR Otdelenie Matematicheskikh Estestvennykh Nauk* 7, 793–800.
- Du, P., Andrec, M., and Levy, R.M. 2003. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Eng.* 16, 407–414.
- Floudas, C.A., Fung, H.K., McAllister, S.R., et al. 2006. Advances in protein structure prediction and de novo protein design: a review. *Chem. Eng. Sci.* 61, 966–988.
- Gallagher, T., Alexander, P., Bryan, P., et al. 1994. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry* 33, 4721–4729.
- Ginalski, K. 2006. Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.* 16, 172–177.
- Goldsmith-Fischman, S., and Honig, B. 2003. Structural genomics: computational methods for structure analysis. *Protein Sci.* 12, 1813–1821.
- Graber, R., Kasper, P., Malashkevich, V.N., et al. 1995. Changing the reaction specificity of a pyridoxal-5'-phosphate-dependent enzyme. *Eur. J. Biochem.* 232, 686–690.
- Grant, A., Lee, D., and Orengo, C. 2004. Progress towards mapping the universe of protein folds. *Genome Biol.* 5, 107.

- Holmes, J.B., and Tsai, J. 2005. Characterizing conserved structural contacts by pair-wise relative contacts and relative packing groups. *J. Mol. Biol.* 354, 706–721.
- Kabsch, W. 1978. Discussion of solution for best rotation to relate 2 sets of vectors. *Acta Crystall. A* 34, 827–828.
- Konagurthu, A.S., Whisstock, J.C., Stuckey, P.J., et al. 2006. MUSTANG: a multiple structural alignment algorithm. *Proteins* 64, 559–574.
- Kryshtafovych, A., Fidelis, K., and Moult, J. 2007. Progress from CASP6 to CASP7. *Proteins* 69, Suppl 8, 194–207.
- Kryshtafovych, A., Venclovas, C., Fidelis, K., et al. 2005. Progress over the first decade of CASP experiments. *Proteins* 61, Suppl 7, 225–236.
- Lennox, K.P., Dahl, D.B., Vannucci, M., et al. 2009. Density estimation for protein conformational angles using a bivariate von Mises distribution and Bayesian nonparametrics. *J. Am. Stat. Soc.* 104, 586–596.
- Liu, T., Horst, J.A., and Samudrala, R. 2009. A novel method for predicting and using distance constraints of high accuracy for refining protein structure prediction. *Proteins* 77, 220–234.
- Liu, X., Fan, K., and Wang, W. 2004. The number of protein folds and their distribution over families in nature. *Proteins* 54, 491–499.
- Michino, M., and Brooks, C.L. 3rd. 2009. Predicting structurally conserved contacts for homologous proteins using sequence conservation filters. *Proteins* 77, 448–453.
- Misura, K.M., and Baker, D. 2005. Progress and challenges in high-resolution refinement of protein structure models. *Proteins* 59, 15–29.
- Qu, X., Swanson, R., Day, R., et al. 2009. A guide to template based structure prediction. *Curr. Protein Pept. Sci.* 10, 270–285.
- Roach, J., Sharma, S., Kapustina, M., et al. 2005. Structure alignment via Delaunay tetrahedralization. *Proteins* 60, 66–81.
- Sali, A., and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815.
- Sanchez, R., and Sali, A. 1997. Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.* 7, 206–214.
- Schwede, T., Sali, A., Honig, B., et al. 2009. Outcome of a workshop on applications of protein models in biomedical research. *Structure* 17, 151–159.
- Schwieters, C.D., Kuszewski, J.J., Tjandra, N., et al. 2003. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* 160, 65–73.
- Soundararajan, M., Willard, F.S., Kimple, A.J., et al. 2008. Structural diversity in the RGS domain and its interaction with heterotrimeric G protein alpha-subunits. *Proc. Natl. Acad. Sci. U. S. A.* 105, 6457–6462.
- Swanson, R., Vannucci, M., and Tsai, J.W. 2009. Information theory provides a comprehensive framework for the evaluation of protein structure predictions. *Proteins* 74, 701–711.
- Tramontano, A. 2003. Of men and machines. *Nat. Struct. Biol.* 10, 87–90.
- Venclovas, C., and Margelevicius, M. 2005. Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. *Proteins* 61, Suppl 7, 99–105.
- Wallner, B., and Elofsson, A. 2005. All are not equal: a benchmark of different homology modeling programs. *Protein Sci.* 14, 1315–1327.
- Wallner, B., Fang, H., Ohlson, T., et al. 2004. Using evolutionary information for the query and target improves fold recognition. *Proteins* 54, 342–350.
- Wang, G., and Dunbrack, R.L., Jr. 2004. Scoring profile-to-profile sequence alignments. *Protein Sci.* 13, 1612–1626.
- Xiang, Z. 2006. Advances in homology protein structure modeling. *Curr. Protein Pept. Sci.* 7, 217–227.
- Yan, A., and Jernigan, R.L. 2005. How do side chains orient globally in protein structures? *Proteins* 61, 513–522.
- Zemla, A. 2003. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* 31, 3370–3374.
- Zhang, Y. 2007. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins* 69, Suppl 8, 108–117.
- Zhang, Y., Arakaki, A.K., and Skolnick, J. 2005. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins* 61, Suppl 7, 91–98.
- Zhou, H., Pandit, S.B., Lee, S.Y., et al. 2007. Analysis of TASSER-based CASP7 protein structure prediction results. *Proteins* 69, Suppl 8, 90–97.

Address correspondence to:

Dr. Jerry Tsai
Chemistry Department
University of the Pacific
3601 Pacific Avenue
Stockton, CA 95211

E-mail: jtsai@pacific.edu