Automated Prediction of Domain Boundaries in CASP6 Targets Using Ginzu and RosettaDOM

David E. Kim,† Dylan Chivian,† Lars Malmström, and David Baker*
University of Washington, Seattle, Washington

ABSTRACT **Domain boundary prediction is an important step in both experimental and computational protein structure characterization. We have developed two fully automated domain parsing methods: the first, Ginzu, which we have described previously, utilizes information from homologous sequences and structures, while the second, RosettaDOM, which has not been described previously, uses only information in the query sequence. Ginzu iteratively assigns domains by homology to structures and sequence families using successively less confident methods. RosettaDOM uses the Rosetta de novo structure prediction method to build three-dimensional models, and then applies Taylor's structure based domain assignment method to parse the models into domains. Domain boundaries observed repeatedly in the models are predicted to be domain boundaries for the protein. Interestingly, RosettaDOM produced quite good domain predictions for proteins of a size typically considered to be beyond the reach of de novo structure prediction methods. For remote fold recognition targets and new folds, both Ginzu and RosettaDOM produced promising results, and in some cases where one method failed to detect the correct domain boundary, it was correctly identified by the other method. We describe here the successes and failures using both methods, and address the possibility of incorporating both protocols into an improved hybrid method. Proteins 2005;Suppl 7:193–200.** © 2005 Wiley-Liss, Inc.

Key words: domain prediction; domain parsing; domain identification; CASP; CAFASP; Rosetta; Robetta; protein structure prediction; ab initio modeling; de novo modeling; template-based modeling; comparative modeling; homology modeling

INTRODUCTION

Because many protein chains consist of multiple domains and most structure prediction methods are optimized for single domains, accurate identification of domain boundaries is often an essential step in protein structure prediction. Our lab's development of automated domain prediction methods has been motivated by this application as well as to aid structural genomics efforts by identifying domains that may express and crystallize separately when the full chain of a multidomain protein does not. For CASP5 and CAFASP3, a sequence homology-based domain detection method called Ginzu was incorporated as the first step in our fully automated protein structure prediction server, Robetta.[1,2] Ginzu has been used to identify domains for structure determination by the Structural Genomics of Pathogenic Protozoa (SGPP) consortium, and also to predict the domain content of the *S. cerevisiae* genome (manuscript in preparation).

For CASP6 and CAFASP4, an additional method that does not rely on sequence homology, called RosettaDOM, was developed to predict domain boundaries based on consistencies found in structure based domain assignments of models that were generated using the Rosetta de novo structure prediction method.[3–5] The method is based on the assumption that although Rosetta is unlikely to produce accurate models for large proteins, it may reproduce low-resolution structural features such as the partitioning of the chain into domains. The development of the method was inspired by the quite good domain structure recapitulation of the CASP5 T0148 structure in the models produced by the Robetta server.

Here we describe the methods used by both Ginzu and RosettaDOM to predict domain boundaries, the factors that lead to successful and failed predictions, and the potential for improvement using new structure prediction methods developed in our lab. Because there were cases where domain boundaries were correctly identified by one method but were missed or incorrectly assigned by the other, a hybrid method using both Ginzu and RosettaDOM is also described and evaluated.

MATERIALS AND METHODS

The Ginzu Protocol

The Ginzu domain prediction method used in our automated structure prediction server, Robetta, has been previously described.[1,2] Initially developed to enable structure predictions of full-length protein sequences regardless of domain content, Ginzu consists of a hierarchical

†These authors contributed equally to this article.

Grant sponsor: the NIH Protein Structure Initiative through the SGPP consortium and HHMI.

*Correspondence to: David Baker, Department of Biochemistry and HHMI, University of Washington, Box 357350, Seattle, WA 98195. E-mail: dabaker@u.washington.edu

Received 23 June 2005; Accepted 27 June 2005

Published online 26 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20737

The article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version.

screening procedure that first uses BLAST, PSI-BLAST,[6] FFAS03,[7,8] and 3D-Jury,[9,10] to detect regions in the query sequence that are homologous to experimentally determined structures, and then proceeds with a search against Pfam-A[11] using HMMER[12] to identify putative domains. In each step, unassigned regions are either treated as domain linkers if they are less than 50 residues, or are used as input for the next step. The final detection step utilizes the program "msa2domains," which examines the multiple sequence alignment (MSA) produced by a PSI-BLAST search against the NCBI nonredundant protein sequence database (NR) to find clusters of sequences that may represent domains. Nonoverlapping clusters with the highest number of unique sequences are assigned as regions of increased domain confidence in the remaining uncovered stretches. Last, msa2domains determines where to place the exact cut points in the remaining linkers with a "cut preference" score that employs a heuristic that considers the least occupied positions in the MSA, strongly predicted loop regions by PSIPRED,[13] and the distance from the nearest assigned region. A fourth term boosts the likelihood of a domain boundary in regions of the MSA where sequence clusters frequently begin or end. Each step in Ginzu is ordered by the reliability of the detection method, with de novo predictions (those that utilize Pfam, the MSA clusters from msa2domains, and the cut preference function from msa2domains) applied last.

Because regions that are assigned by homology to an experimentally determined structure may consist of more than one domain, an additional step was developed to assign domain boundaries. In CASP6, these regions were parsed based on the comparative model generated by Robetta. We developed a consensus method that applies Taylor's structure-based domain parsing algorithm[14] to the model as well as to its PSI-BLAST detectable structural homologs, an approach that provides more robust definitions (manuscript in preparation). Taylor's method assigns domains using only α-carbon coordinates by numerically labeling each residue along the protein chain sequentially and then iteratively updating the labels based on the average of neighboring labels within a given radius. The labels of residues surrounded by neighbors that have higher or lower labels on average, are increased or decreased by 1, respectively. Label reassignments are made for each residue until convergence is reached. The end result is the assignment of compact domains that are identified by residues with the same label. This method is fast, and has been shown to be accurate, and is not limited to delineating continuous domains but may also partition complex topologies that consist of domains with discontinuous segments of the protein chain (such as proteins with domain insertions).

The RosettaDOM Protocol

Domain boundaries are currently most accurately assigned from close structural homologs. Therefore, RosettaDOM first uses Ginzu to identify domains that are homologous to known structures in the PDB. If Ginzu assigns a domain using BLAST, PSI-BLAST, or FFAS03, RosettaDOM simply returns the domain boundary predictions provided by Ginzu.

For targets lacking such homology, we developed a de novo domain prediction method that is similar in concept to SnapDRAGON,[15] but uses the Rosetta de novo structure prediction method to produce models. RosettaDOM generates 400 three-dimensional models using Rosetta, and then selects the top 200 scoring models that pass filters that eliminate structures with too many local contacts or unlikely strand topologies. Domain boundaries are then assigned for each of the 200 models using Taylor's structure-based domain identification algorithm described above. Final domain boundary predictions are made based on consistencies found in the domain assignments of these models by taking the sum of boundary assignments at each position along the protein chain, smoothing the values using a center weighted sliding window, and then converting the smoothed boundary distributions to *Z*-scores as described by George et al.[15] Positions with *Z*-scores of 2.5 or greater are treated as potential domain boundaries. Because logic is not applied to assign discontinuous domains and continuous domains are unlikely to be less than 50 residues in length, final domain boundaries are assigned for positions with the highest *Z*-scores that are at least 50 residues apart and are not within 50 residues of the N and C terminus.

Target Classification

Because methods that use structural homology have a clear advantage for targets that have similar sequences in the PDB, it is important to discriminate between predictions that used such structural information (referred to as "template-based" predictions) and those that did not (referred to as "de novo" predictions). The majority of Ginzu predictions were template-based, but many of these predictions were made for remote targets in the FR regime as assigned by the assessors. The template-based Ginzu results were thus separated into "easy" and "hard" categories, depending on whether any domain was detected using BLAST or PSI-BLAST (easy), or FFAS03 or 3D-Jury (hard). Some multidomain targets had domains assigned by detection methods from different categories, and for these cases, the targets were classified by the more confident method. Because RosettaDOM did not use Ginzu template-based results for targets that were assigned by 3D-Jury, a larger fraction of de novo predictions were made compared to Ginzu (43% vs. 18%). The CASP/CAFASP targets that are included in this evaluation are listed in Table I. The targets are grouped by the most confident detection method used by Ginzu, and the classification provided by the assessors is listed next to the target id.

RESULTS AND DISCUSSION

We use three considerations when evaluating our domain prediction results, the accuracy in (1) the domain count (Table II), (2) the accuracy in domain boundary assignments (Table III), and (3) an overlap score similar to the one used by the assessors (Table IV). Successful

TABLE I. CASP/CAFASP Targets Grouped by the Most Confident Ginzu Detection Method

				RosettaDOM *de novo*			
Ginzu, easy BLAST, PSI-BLAST		Ginzu, hard FFAS03		Ginzu, hard 3D-Jury		Ginzu *de novo*	
T0196	CM/hard	T0203	FR/H	T0197	FR/H	T0201	NF
T0200	CM/hard	T0205	CM/hard	T0198	FR/A	T0209_1	FR/A
T0204	CM/easy	T0211	CM/hard	T0199_1	CM/hard	T0209_2	NF
T0206	FR/H	T0228_1[d]	FR/H	T0199_2[d]	FR/H	T0212	FR/A
T0207	*	T0228_2[d]	FR/H	T0199_3	FR/A	T0213	FR/H
T0208	CM/hard	T0255_1	*	T0202_1[d]	FR/H	T0214	FR/H
T0219_1[d]	*	T0255_2	*	T0202_2	FR/H	T0215	FR/A
T0219_2	*	T0255_3	*	T0210	*	T0216_1	NF
T0220_1[d]	*			T0217	*	T0216_2	NF
T0220_2	*			T0224	FR/H	T0227	FR/H
T0222_1	CM/hard			T0226_1[d]	CM/hard	T0239	FR/A
T0222_2	FR/H			T0226_2	CM/hard	T0242	NF
T0223_1	CM/hard			T0230	FR/A	T0243	FR/H
T0223_2	FR/H			T0236	*	T0257	*
T0229_1	CM/easy			T0238	NF	T0272_1	FR/A
T0229_2	CM/easy			T0241_1[d]	NF	T0272_2	FR/A
T0231	CM/easy			T0241_2[d]	NF	T0273	FR/A
T0232_1	CM/hard			T0248_1	FR/A		
T0232_2	CM/hard			T0248_2	NF		
T0233_1	CM/easy			T0248_3	FR/A		
T0233_2	CM/easy			T0249_1	FR/H		
T0234	CM/hard			T0249_2	FR/H		
T0235_1[d]	CM/easy			T0250_1[d]	*		
T0235_2	FR/A			T0250_2	*		
T0237_1	FR/H			T0251	FR/H		
T0237_2	FR/H			T0254	*		
T0237_3	FR/H			T0262_1	FR/A		
T0240	CM/easy			T0262_2	FR/H		
T0244	CM/easy			T0263	FR/H		
T0246	CM/easy			T0281	FR/A		
T0247_1[d]	CM/easy						
T0247_2[d]	CM/easy						
T0247_3	CM/easy						
T0252_1[d]	*						
T0252_2	*						
T0253	*						
T0256_1	*						
T0256_2	*						
T0258	*						
T0260_1	*						
T0260_2	*						
T0264_1	CM/easy						
T0264_2	CM/hard						
T0265	CM/hard						
T0266	CM/easy						
T0267	CM/hard						
T0268_1[d]	CM/easy						
T0268_2	CM/easy						
T0269_1	CM/easy						
T0269_2	CM/hard						
T0271	CM/easy						
T0274	CM/easy						
T0275	CM/easy						
T0276	CM/easy						
T0277	CM/easy						
T0279_1[d]	CM/hard						
T0279_2	CM/hard						
T0280_1[d]	CM/easy						
T0280_2	FR/A						
T0282	CM/easy						

*Indicates targets that were not considered by assessors because of cancellations due to early released or not yet available structures. These targets were included in our assessment because the predictions were made before the structure was released.

[d]Discontinuous domain.

The following classifications are provided by assessors and listed beside the target id: CM/easy, comparative modeling with BLAST detectable parent; CM/hard, comparative modeling with PSI-BLAST detectable parent; FR/H, fold recognition homologous; FR/A, fold recognition analogous; NF, new fold. Discrepancies between the assessors' categorization and the Ginzu detection method are largely due to our classification of all domains of multidomain targets with the most confident detection method.

TABLE II. Accuracy of Domain Count Predictions

	Predicted domain count			
	1	2	3	4
Official domain count				
A. *Ginzu (template-based, easy)*				
1	**19 (86.4%)**	3 (13.6%)	—	—
2	2 (12.5%)	**12 (75%)**	—	2 (12.5%)
3	—	1 (50%)	**1 (50%)**	—
B. *Ginzu (template-based, hard)*				
1	**11 (73.3%)**	4 (26.7%)	—	—
2	1 (14.3%)	**5 (71.4%)**	1 (14.3%)	—
3	1 (33.3%)	2 (66.7%)	**—**	—
C. *Ginzu (de novo)*				
1	**11 (100%)**	—	—	—
2	1 (33.3%)	**1 (33.3%)**	1 (33.3%)	—
3	—	—	**—**	—
D. *RosettaDOM (de novo)*				
1	**21 (91.3%)**	2 (8.7%)	—	—
2	1 (11.1%)	**7 (77.8%)**	1 (11.1%)	—
3	—	2 (100%)	**—**	—

Correct domain count predictions are highlighted. The easy template-based predictions used BLAST or PSI-BLAST to detect at least one domain, whereas the hard template-based predictions used FFAS03 or 3D-Jury. The *de novo* Ginzu predictions used either a search against the Pfam-A database or msa2domains as described in the "The Ginzu Protocol" section. *De novo* RosettaDOM predictions were made for targets that did not have any domains assigned by Ginzu using BLAST, PSI-BLAST, or FFAS03.

TABLE III. Accuracy of Domain Boundary Prediction for Multidomain Targets

	Sensitivity (%)		Specificity (%)	
	Predicted	Control	Predicted	Control
Ginzu				
Easy	66.7	23.3	69.0	24.1
Hard + *de novo*	26.9	11.5	53.9	23.1
RosettaDOM	28.6	9.5	54.6	18.2

A boundary is considered correct if it lies within ±10 residues from the center of a domain linker assigned by the assessors. The sensitivity and specificity were calculated as TP/(TP + FN) and TP/(TP + FP), respectively, where TP is the number of true positives, FN, false negatives, and FP, false positives. The control uses the predicted domain count and just divides the target into equally sized domains.

TABLE IV. Average Domain Overlap Scores for "Hard" and *De Novo* Targets

	Ginzu	Ginzu cutpref	RosettaDOM	Hybrid
Multidomain	64	61	62	69
single + multidomain	81	77	82	82

See assessors' section in this issue for specifics about the overlap score.

predictions are addressed below, with an emphasis toward targets that were predicted using de novo methods. Problems with the current methods are discussed for failed predictions and potential methods for improvement are suggested. Finally, a hybrid method that uses scoring components from Ginzu, msa2domains, and RosettaDOM is described and evaluated.

What Went Right

Template-based predictions for many targets grouped in the "easy" category did well, even for targets with discontinuous domains. De novo predictions made by Ginzu and RosettaDOM were quite encouraging for a number of targets that will be discussed in detail below. Surprisingly, the overall results for the automated Ginzu and RosettaDOM methods were comparable to the best human predictors (see assessor's report in this issue). This was particularly true when considering the subset of targets for which the assessor classified as unambiguously containing multiple domains, where only two human groups (which submitted predictions for only a limited number of targets) exceeded the performance of Ginzu and RosettaDOM. Even when considering all targets, including single-domain targets, only three human groups were superior to these automated methods.

Although the sample is small, encouraging results in domain count accuracy were observed. Table II summarizes the results for the comparison of the "official" domain count with the domain count predictions made for (A) Ginzu "easy," (B) Ginzu "hard," (C) Ginzu de novo targets, and (D) RosettaDOM de novo targets. The domain count was correct for over 70% of two-domain targets predicted using the various detection methods with the exception of the Ginzu de novo set for which there were only three two-domain targets. Domain counts for over 85% of single domain targets were correctly assigned with the exception of Ginzu "hard" targets (73%). Among all sets, the domain count for only one three-domain target was correctly predicted (by Ginzu "easy"), with the majority of three-domain targets predicted as having two domains.

The accuracy of the domain boundary predictions was also encouraging for multidomain targets as shown in Table III. Although the sensitivity of detecting domain boundaries was somewhat low for the Ginzu "hard" + de novo and RosettaDOM sets (below 30%), the specificity

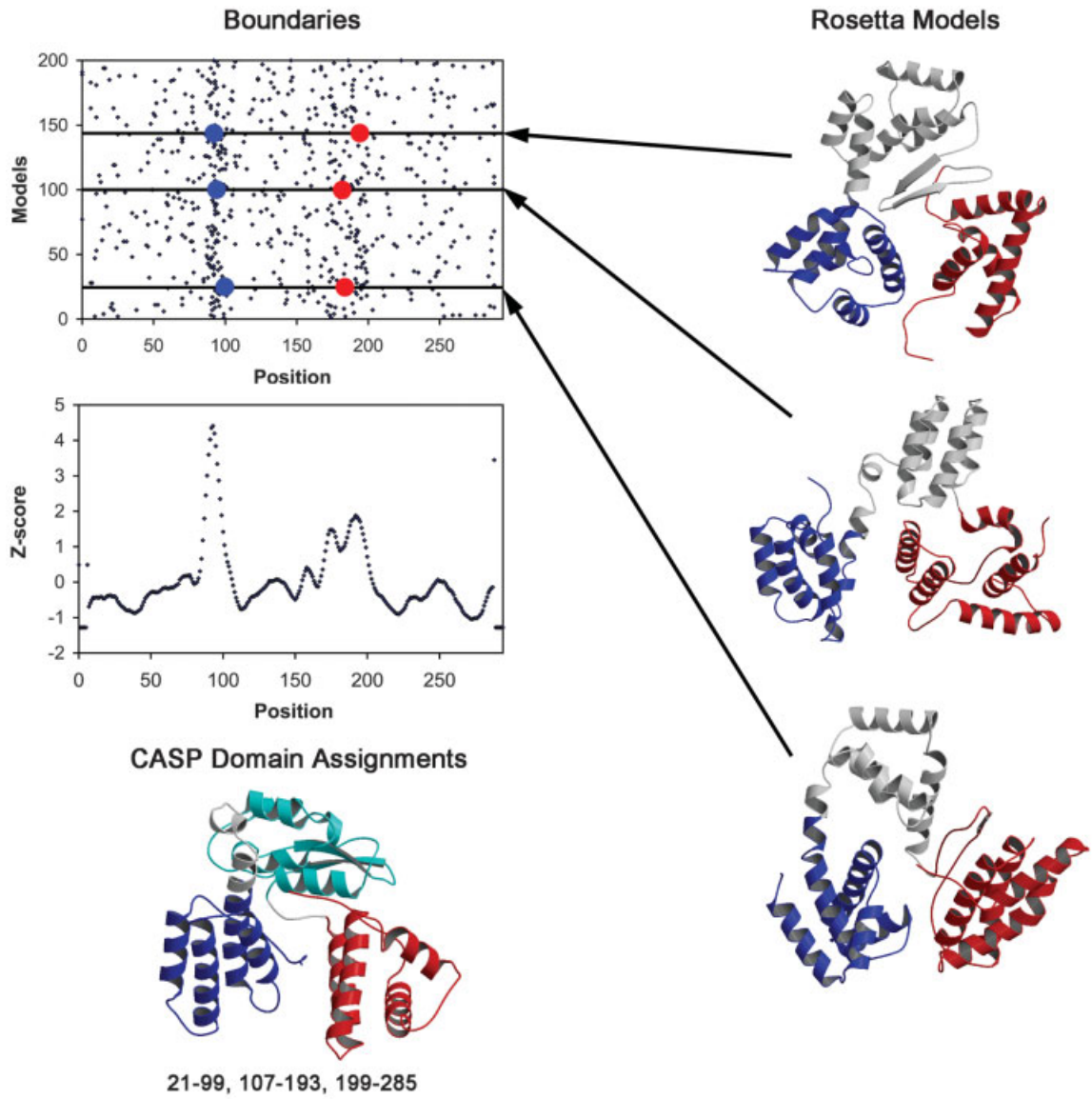

Fig. 1. Domain boundary distribution and models that were made by the Rosetta de novo structure prediction method for T0248. The first plot displays the domain boundaries assigned to models produced by Rosetta and the corresponding models for three examples are shown on the right. The *Z*-scores for each position are shown in the second plot. The CASP domain assignments in the context of the native structure is displayed in the bottom left corner. Interestingly, models with roughly the correct domain boundaries are being produced by Rosetta.

(predicted boundaries were within ±10 residues from the center of a domain linker assigned by the assessors) was greater than 50%.

The most interesting and somewhat surprising results came from predictions made using RosettaDOM. Models for relatively large (>150 aa) proteins generated using the Rosetta fragment insertion method were often partitioned into multiple domains, and for some cases, a significant number of models were partitioned with good correspondence (with the exception of one or two secondary structure elements) to the native structure even though their detailed topologies were incorrect. An example of this is shown for the three-domain target, T0248, in Figure 1. The first domain boundary was correctly predicted (with a *Z*-score of 4.4) based on consistencies in the Rosetta models, and the second was suggested with a *Z*-score of 1.9 at position 192 but not chosen due to being below the threshold of 2.5. The distribution of domain boundaries assigned for the models is broader and slightly bimodal for the second domain boundary, and perhaps may be explained by a beta-hairpin in the native structure that interacts with the neighboring domain. Successful domain boundary predictions for two other targets, T0249 and T0262, are displayed in the context of their native structures in Figure 2(a) and (b), respectively. Interestingly, both targets also display a peak in the boundary distributions beside a long stretch of missing coordinates at the C and N termini, respectively.

It is striking that domain boundary predictions using RosettaDOM can be reasonable when the detailed internal

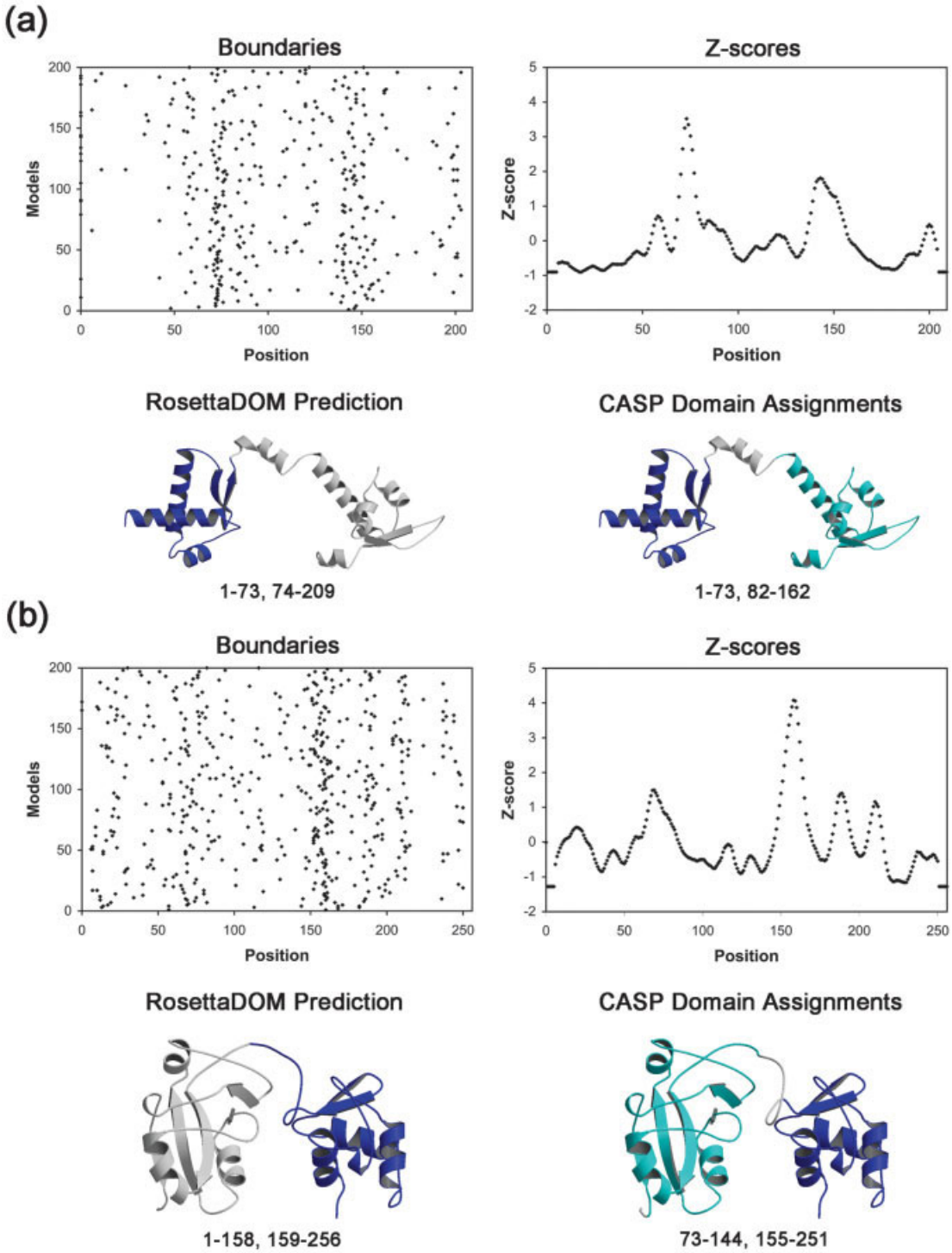

Fig. 2. RosettaDOM domain predictions for (**a**) T0249 and (**b**) T0262. The boundaries assigned to each Rosetta model are shown in the plot on the left for each target. The *Z*-scores for each position are shown in the plot on the right. The prediction made by RosettaDOM in the context of the native structure is shown in the bottom left, and the CASP domain assignments are shown in the bottom right for each target.

structures of the domains are not accurately predicted. The Rosetta de novo structure prediction method may capture basic features of the clustering of hydrophobic residues into domains that represent the best solution to burying hydrophobic residues given the local structural propensities of the protein chain.

For target T0209, the Ginzu de novo method made the correct boundary prediction (Fig. 3), whereas RosettaDOM

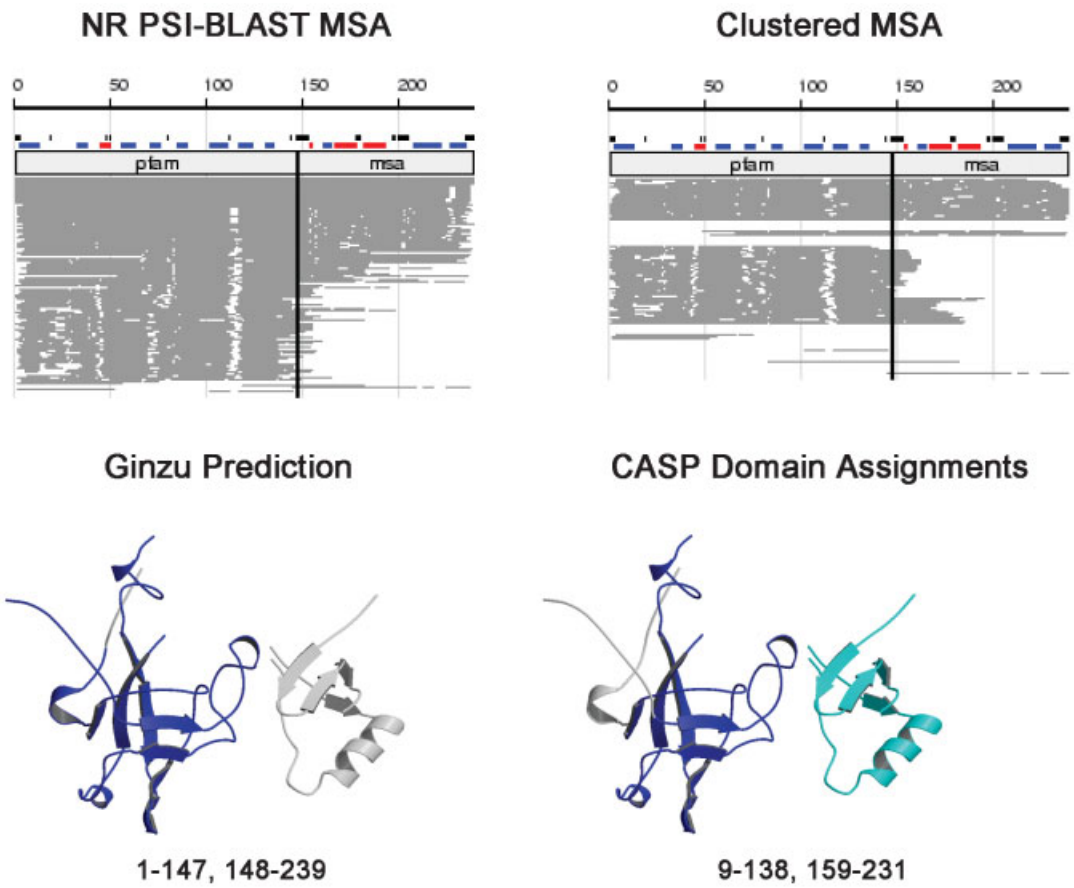

Fig. 3. Ginzu domain prediction for T0209. The MSA from a PSI-BLAST search against the NR is displayed in the top left, and the clustered MSA is shown in the top right. The prediction made by Ginzu in the context of the native structure is shown in the bottom left, and the CASP domain assignments are shown in the bottom right.

failed to detect any boundaries. A large number of sequences homologous to the first domain were found in the NR as shown in the clustered MSA in Figure 3 (top right). Ginzu detected this domain with a search against the Pfam-A database, but likely would have assigned it correctly in the absence of the Pfam search by assigning the domain from the MSA cluster. The boundary between the domains was correctly suggested by the transition from the end of the Pfam-associated cluster to the start of the sequence homolog that matches only the C-terminal domain.

What Went Wrong

Many domain boundaries were not predicted accurately, even for some targets that were categorized as "easy." The difficulties can be attributed to a variety of factors such as the relatively large number of domains that contain discontinuous segments of the protein chain and complex nonlocal topologies, and protein lengths greater than 200 residues that are difficult for de novo modeling. Targets T0197 and T0209 are good examples of difficult topologies for RosettaDOM because they both have beta-barrel-like folds. For the single domain target, T0197, RosettaDOM assigned a domain boundary with a relatively high *Z*-score of 5.4 (most models produced by Rosetta were partitioned near this position). Interestingly, the parse was placed in a strand that pairs nonlocally with another. Such nonlocal strand pairings are difficult for Rosetta to model using the standard fragment insertion protocol and may be a factor that lead to the consistent partitioning of models (it is possible that such "incorrect" parses may be useful in guiding the modeling of nonlocal strand pairs during folding). A new protocol was developed for structure predictions by our human group that uses long-range beta-sheet pairings suggested from models generated by fold-recognition servers for relevant targets, and is explained in the Baker human group article in this issue. The application of this new protocol to T0197, removed the strong bias to incorrectly partition the protein. RosettaDOM assigns the two-domain target, T0209, as a single domain. An incorrect parse just below the *Z*-score threshold is suggested within the beta-barrel-like first domain, which also contains nonlocal strand pairs.

For the "easy" Ginzu target T0235, a discontinuous domain was modeled as two domains due to the detection of one portion of the discontinuous domain using PSI-BLAST and the other using FFAS03. The templates that were used for comparative modeling were in the same

SCOP protein family; thus, logic could have been implemented to use just one template for the discontinuous domain. We will be incorporating such logic into Ginzu in the near future.

What We Learned

The better performance in general of automated domain prediction methods compared to humans (see assessor's report in this issue) is quite promising, particularly for applications at the genome scale. However, there is much room for improvement, specifically in the accurate identification of discontinuous domains, and the sensitivity and specificity of domain boundary predictions. On a positive note, our methods for domain boundary predictions should improve as structure prediction methods get better. The correct reassignment of T0197 using a newly developed structure prediction protocol is an example of this. Also, the surprising capability of Rosetta to generate models for proteins greater than 200 residues in length (outside the range thought to be accessible to automated de novo structure prediction) that are sometimes accurately partitioned into domains gives hope to further development.

Because there were examples of targets that were correctly predicted by one method but missed by the other, and Ginzu "cut preference" and RosettaDOM boundary scores sometimes suggested the correct parse points but were below the threshold, a hybrid method was investigated in an effort to improve results for the "hard" template-based and de novo targets. The results of the hybrid method compared to the individual methods are displayed in Table IV. The hybrid method ("Hybrid") consists of a linear combination of the boundary scores that are obtained by the complete Ginzu method ("Ginzu"), the Ginzu cut preference ("Ginzu cutpref"), and RosettaDOM's boundary frequencies ("RosettaDOM"). The overlap scores in Table IV are a similar formulation to the one used by the assessors (see assessor's report in this issue for details). For the multidomain targets, the hybrid method has an overlap score of 69%, which improves upon the best individual method. These results suggest that the hybrid method may improve the performance of boundary predictions for more difficult proteins. Further research is needed to properly assess the performance of the hybrid method and its component methods on a larger test set.

ACKNOWLEDGMENTS

The authors thank the structural biologists who provided structures, and CASP organizers, and assessors for making the CASP6 experiment possible. We also thank Dani Fischer for running the CAFASP4 experiment, Adam Godzik and Leszek Rychlewski for allowing Ginzu to use information from their servers, William Taylor for the use of his structure based domain assignment algorithm, Sean Eddy for the use of HMMER, David Jones for the use of PSIPRED, Kevin Karplus for the use of the SAM software, and Jens Meiler for the use of JUFO.

REFERENCES

1. Chivian D, Kim DE, Malmstrom L, Bradley P, Robertson T, Murphy P, Strauss CE, Bonneau R, Rohl CA, Baker D. Automated prediction of CASP-5 structures using the Robetta server. Proteins 2003;53(Suppl 6):524–533.
2. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32(Web Server issue):W526–531.
3. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CE, Baker D. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins 2003; 53(Suppl 6):457–468.
4. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
5. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
6. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
7. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile-profile sequence alignments. Nucleic Acids Res 2005;33(Web Server issue):W284–288.
8. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 2000;9:232–241.
9. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
10. Ginalski K, Rychlewski L. Detection of reliable and unexpected protein fold predictions using 3D-Jury. Nucleic Acids Res 2003;31: 3291–3292.
11. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
12. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
13. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
14. Taylor WR. Protein structural domain identification. Protein Eng 1999;12:203–216.
15. George RA, Heringa J. SnapDRAGON: a method to delineate protein structural domains from sequence data. J Mol Biol 2002; 316:839–851.