PROTEINS: Structure, Function, and Genetics 53:524–533 (2003)

Automated Prediction of CASP-5 Structures Using the Robetta Server

Dylan Chivian,[1] David E. Kim,[1] Lars Malmström,[1] Philip Bradley,[1] Timothy Robertson,[1] Paul Murphy,[1] Charles E.M. Strauss,[2] Richard Bonneau,[3] Carol A. Rohl,[4] and David Baker[1]*

[1]*University of Washington, Seattle, Washington*
[2]*Los Alamos National Laboratory, Los Alamos, New Mexico*
[3]*Institute for Systems Biology, Seattle, Washington*
[4]*University of California, Santa Cruz, California*

ABSTRACT **Robetta is a fully automated protein structure prediction server that uses the Rosetta fragment-insertion method. It combines template-based and *de novo* structure prediction methods in an attempt to produce high quality models that cover every residue of a submitted sequence. The first step in the procedure is the automatic detection of the locations of domains and selection of the appropriate modeling protocol for each domain. For domains matched to a homolog with an experimentally characterized structure by PSI-BLAST or Pcons2, Robetta uses a new alignment method, called K*Sync, to align the query sequence onto the parent structure. It then models the variable regions by allowing them to explore conformational space with fragments in fashion similar to the *de novo* protocol, but in the context of the template. When no structural homolog is available, domains are modeled with the Rosetta *de novo* protocol, which allows the full length of the domain to explore conformational space via fragment-insertion, producing a large decoy ensemble from which the final models are selected. The Robetta server produced quite reasonable predictions for targets in the recent CASP-5 and CAFASP-3 experiments, some of which were at the level of the best human predictions. Proteins 2003;53:524–533.**

Key words: automated protein structure prediction server; CASP; CAFASP; rosetta; fragment insertion; fragment assembly; *ab initio* modeling; *de novo* modeling; template-based modeling; domain parsing; homology modeling; comparative modeling; sequence alignment

INTRODUCTION

The best method for predicting the structure of a protein depends on whether it has sequence homology to a protein of known structure. If there is such a similarity, relatively accurate models can be built using the known structure as a template. In the absence of such similarity, models can be built using *de novo* prediction methods, which do not rely on a template structure. In many cases, hybrid template-based/*de novo* methods may be most appropriate: portions of a given target may be modeled based on a template, while it may only be possible to model long variable loops or extra domains or extensions not contained in the template using *de novo* methods.

Full automation of protein structure prediction is a desirable goal as it opens the door to genome-level protein structure modeling and, equally importantly, provides a stringent test of the principles underlying prediction methods unadulterated by the powerful influence of human intuition. The fully automated Robetta structure prediction server attempts to provide the best possible model for the entire length of the protein chain by combining template-based and *de novo* protocols.

PROCESS

Robetta uses the Rosetta fragment-insertion technique[1–3] to build models of protein domains in both template-based and *de novo* modes. Modeling is performed at the domain level based on the assumption that domains are autonomously folding units. Since protein chains are often comprised of more than one domain, it is essential that any server which attempts to model the full length of a query in domain-sized pieces determine the location of putative domains, assign each of those domains to the appropriate template-based or *de novo* protocol, and ideally to restore chain connectivity between the domains by assembling the domain models into a single multi-domain prediction.

An overview of the Robetta process is shown in Figure 1 (for details of the process, see Methods section below). The initial step, called "Ginzu" (see Figure 2), involves screening the query sequence for regions that possess a homolog with an experimentally characterized structure with BLAST, PSI-BLAST,[4] and Pcons2[5] [and described in this volume], followed by cutting the sequence into putative domains based on matches to known families and struc-

---

The Supplementary Materials Referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: Burroughs-Wellcome Fund; Grant sponsor: NIH 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 22 February 2003; Accepted 23 June 2003

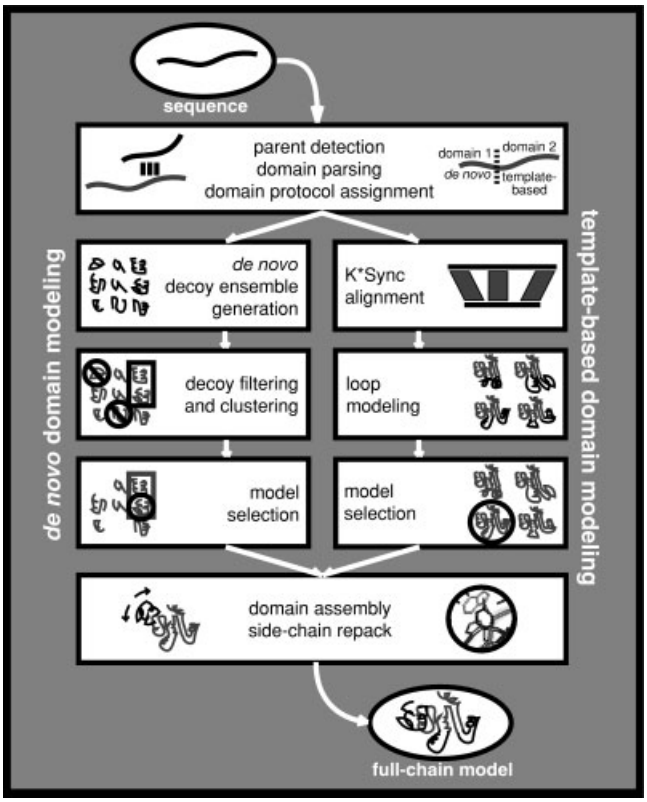

Fig. 1. Robetta process overview. The query sequence is scanned for homologs with experimentally determined structures, domain boundaries are determined, and each domain is modeled separately using either the *de novo* or template-based protocols, assembled into a full-chain model, and side-chains are repacked to produce a full-chain all-atom complete model.

tures, multiple sequence information, and predicted secondary structure information. Any detected parents and the regions of the query with which they are associated are stored and assigned to the template-based modeling protocol. Remaining long unassigned regions are then cut up into sizes amenable to modeling by the Rosetta *de novo* protocol.

After domain parsing, each putative domain then follows its assigned protocol track. For the domains to be modeled *de novo*, an automated version of the CASP-4 Rosetta protocol[3] is used to generate large numbers of alternate "decoy" conformations, and subsequently to filter the decoy ensemble to remove non protein-like conformations and to cluster the remaining structures to identify broad low free energy minima. The final step in the *de novo* domain modeling protocol consists of selection of four final models from the most populated decoy clusters and one model that is the lowest energy decoy remaining that was not in the top clusters.

The template-based modeling protocol first requires an alignment to the parent. Rather than use the PSI-BLAST or Pcons2 alignment, Robetta uses our "K*Sync" alignment program (D.C., manuscript in preparation), which takes into account evolutionary sequence information for both the query and the parent, secondary structure information, and information on regions that are likely to be structurally obligate to the fold (for a further description, see Methods section). From this alignment a template is generated, and variable regions are then modeled with a version of the Rosetta *de novo* method that allows the conformations of variable regions to be sampled in the context of a fixed template.[6] The lowest energy models are selected as the Robetta predictions for the target.

If a target possesses more than one domain, the separate domain models are then combined into one full-length model. This is currently accomplished by fragment- insertion in the putative linker region in order to provide chain connectivity and attempt domain association (unlike CAFASP-3 when domain coordinates were simply spaced by 100 Å). The last step consists of repacking the side-chains using a backbone-dependent rotamer library[7] with a Monte Carlo conformational search procedure.[8]

RESULTS AND DISCUSSION

The protocol used by Robetta for each target is shown in Table I. The targets are separated into columns based on the classification of the assessors, and the method used by Robetta to model the domain is indicated next to the target id: ("*": *de novo*, "bl": parent detected by BLAST, "Ψ": parent detected by PSI-BLAST, "pc": parent detected by Pcons2). As can be seen, Robetta processed the targets in a fashion roughly following the classification of the targets by the assessors, particularly in the extreme categories "Comparative Modeling" and "New Fold". The exception to this was for some of the more challenging Fold Recognition targets, for which a parent was not confidently detected, and were therefore modeled by Robetta's *de novo* protocol rather than utilizing a low-confidence parent. Overall, the models were often quite reasonable predictions, occasionally on par with the best models produced by human groups.

We leave more thorough discussion of Robetta model quality with respect to the field as a whole to the assessors. Instead, to ascertain possibilities for improving our automated modeling protocol, we compare the quality of homology models produced by Robetta to those produced by the human group (a comparative analysis of human vs. automatic *de novo* modeling is in the Baker human group paper in this issue), and to the model produced by the method used to detect the parent structure. BLAST, PSI-BLAST or Pcons2 (for the BLAST category of targets, we compare with PSI-BLAST produced models, as such alignments tend to be longer and therefore of better quality than the corresponding BLAST alignments). We refer to the latter class of models as the "Base" model. Comparison with respect to Base models tells us whether the various stager of Robetta are increasing the quality of the model, either by realigning the query sequence to the parent structure with K*Sync, or by modeling the loops with Rosetta. At the other end of the spectrum, we consider the approximate upper bound to template-only modeling provided by a structure-structure alignment from the LGA server[a] of the target native structure with the parent structure. We refer to the optimal LGA alignment derived model as "LGAopt".

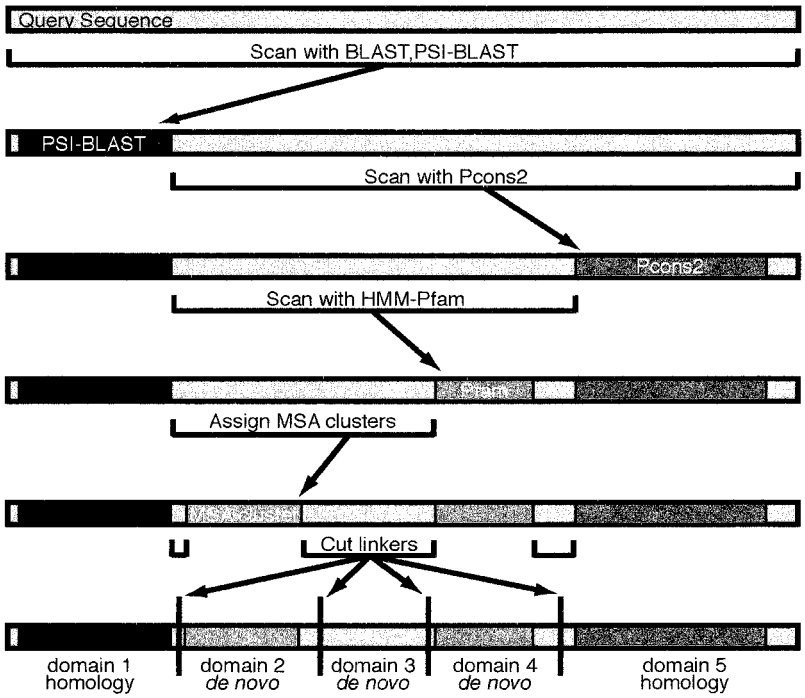

Fig. 2. Ginzu domain parsing illustration: The query sequence is scanned in order to find parent structures and regions of increased domain confidence. Homologous structure searches are performed first, and methods are applied in order of reliability. Remaining uncovered stretches that are long enough to fit a domain are passed forward to the next round. Cuts are applied so that putative query domains without homologous structures are limited in size to lengths accessible to the Rosetta *de novo* protocol.

We examine several stages of homology modeling by Robetta. The model produced by the straightforward mapping of the query sequence onto the parent backbone coordinates based on the default K*Sync alignment we refer to as "RBTdef". The template is then trimmed back into the "stem" regions to allow more flexibility in subsequent loop modeling. We refer to the complete model (which was submitted as Robetta's CASP/CAFASP prediction), with coordinates for all query residues either from the template or from loop modeling, as "RBTcmpl". We also consider the model, called "rebuilt" or "RBTreb", that has coordinates from the complete model, but only includes residues that are defined in the original default template-only model (i.e. new residues added by the loop modeling are left out, leaving only the fixed template and perturbed stem coordinates which have been rebuilt).

There are two types of human-intervention homology models considered as well: human "relaxed" ("HUMrlx") and human "not-relaxed" ("HUMnrlx"). Relaxing allowed conformational sampling along the entire length of the chain of completed models, including template regions, in an effort to make the models more protein-like (e.g. alleviate clashes that resulted from placing the target sequence on a backbone defined by a homologous sequence). None of the Robetta models were relaxed.

What Went Right

Encouragingly, Robetta performed quite well for many targets. For the targets with the closest structural homologs (the BLAST-level targets in the Comparative Modeling category), the automated method was more consistent than our human group in producing first models that were in close agreement with the experimental structure, with even the side-chains quite well rendered (Roland Dunbrack, personal communication and "FORCASP" website posting). Table II, Supplementary Table I, and Figure 3a compare Robetta models to models generated directly from the alignment provided by PSI-BLAST or Pcons2. As can be seen, the default K*Sync alignments are usually as good or better than the alignment provided by the base method (average GDT_TS[9] scores of default K*Sync vs. Base models is BLAST: 74.3 vs. 71.4, PSI-BLAST: 49.0 vs. 47.6, Pcons2: 43.0 vs. 39.0), justifying our decision to generate our own alignment rather than depend on the PSI-BLAST or Pcons2 alignment. The alignment quality obtained for close relatives is on par with that achieved by the base methods, with most of the improvement in alignment quality coming from the more distantly related query-parent pairs, both in terms of detection difficulty and sequence identity (see Figure 3a).

A potential concern in the Robetta method is the trimming back of the alignments in the stem regions to allow for modeling of the loops since the residues that have been removed from the alignment might be better rendered by leaving them untouched as template residues. However this does not appear to be a significant problem, with the quality of the Robetta "rebuilt" models comparable to the Robetta "default" models (average GDT_TS scores of Ro-

TABLE I. Robetta Modeling Protocol and Parent Detection Source

CM-BL	CM-PSI	CM/FR	FR(H)	FR(A)	FR/NF	NF
bl T137	ψ T133	pc T130	pc T134_1	* T135	* T146_1	* T129_1
bl T140	ψ T141	pc T132	pc T134_2	pc T147	* T146_2	* T129_2
bl T142	ψ T149_1	pc T136_1	pc T138	* T148_1	* T146_3	* T139
bl T143	ψ T152	* T136_2	* T156	* T148_2	* T146_4	* T149_2
bl T150	ψ T165	pc T159_1	pc T157	* T162_1	* T170	* T161
bl T151	ψ T169	pc T159_2	* T174_1	* T162_2	* T172_2	* T162_3
bl T153	ψ T171	pc T168_1	* T174_2	pc T187_2	* T173	* T181_1
bl T154_1	ψ T172_1	pc T168_2	pc T193_1	* T191_1	ψ T186_3	* T181_2
bl T154_2	ψ T175	ψ T193_2			* T187_1	
bl T155	ψ T176					
bl T160	bl T184_2					
bl T163	bl T185_1					
bl T167	bl T185_3					
bl T177	ψ T186_1					
bl T178	ψ T186_2					
bl T179	ψ T189					
bl T182	ψ T192A					
bl T183	ψ T192B					
bl T184_1	ψ T195					
bl T185_2						
bl T188						
bl T190						
bl T191_2						

bl - blast; ψ - psi-blast; pc - pcons2; * - de novo

The protocol used for the CASP 5 domains, and the assessors' categorization of the targets. De novo protocol modeled targets are indicated with a "*". All others were modeled following the Rosetta template-based protocol. Targets labeled with "bl" were based on parents detected by BLAST, those labeled with "ψ" were based on parents detected by PSI-BLAST, and those labeled with "pc" were based on parents detected by Pcons2. The categories are as follows: CM-BL: Comparative modeling with BLAST detectable parent; CM-PSI: Comparative Modeling with PSI-BLAST detectable parent; CM/FR: transition category between Comparative Modeling and Fold Recognition (e.g., transitive PSI-BLAST detectable); FR(H): Fold Recognition Homologous; FR(A): Fold Recognition Analogous; FR/NF: the transition category between Fold Recognition and New Fold; and NF: New Fold. The discrepancy between the assessors' categorization and our modeling method for T186_3 results from Robetta's incorrect treatment of this region of the query as part of domain 2, which was detected by PSI-BLAST. Other discrepancies between BLAST and PSI-BLAST categorization likely result from the slightly different results obtained with different PSI-BLAST parameters and sequence databases.

TABLE II. Summary Statistics of Homology Modeled Targets

PARENT SOURCE	N —	LGAopt IDENT	TARGET LEN	BASE LEN	RBTdef LEN	RBTtrim LEN	HUMtrim LEN	LGAopt LEN	BASE GDT_TS	RBTdef GDT_TS	RBTreb GDT_TS	RBTcmpl GDT_TS	HUMnrlx GDT_TS	LGAopt GDT_TS
blast	26	31.5	164.6	152.4	154.7	136.9	126.9	147.1	71.4	74.3	73.6	75.3	75.0	76.4
psibl	14	17.9	205.9	177.4	175.2	143.3	120.1	149.6	47.6	49.0	48.1	50.6	52.2	55.3
pcons2	13	14.2	147.6	125.7	121.9	99.6	82.3	96.1	39.0	43.0	42.4	44.4	44.7	50.7

Average values within each difficulty category (for a target-specific analysis of the data used to generate the summary statistics, see Supplementary Table I). "PARENT SOURCE" is the parent detection method. "N" is the number of targets in the category that were used for the analysis. "LGAopt IDENT" is the sequence identity of the target-parent pair, as determined by the LGA structure-structure alignment. "TARGET LEN" is number of residues in the targets, and in all complete models. "BASE LEN" is the number of residues in the Base model. "RBTdef LEN" is the number of residues in the model built directly from the default K*Sync alignment. "RBTtrim LEN" is the number of residues in the trimmed K*Sync alignment (after stem removal). "HUMtrim LEN", like RBTtrim, is the number of residues in the human group model that were used as fixed template in the complete models. "LGAopt LEN" is the number of residues in the LGA structure-structure alignment. "BASE GDT_TS", "RBTdef GDT_TS", "RBTreb GDT_TS", "RBTcmpl GDT_TS", "HUMnrlx GDT_TS", and "LGAopt GDT_TS" are the GDT_TS scores achieved by the corresponding models.

betta "rebuilt" vs. Robetta "default" models is BLAST: 73.6 vs. 74.3, PSI-BLAST: 48.1 vs. 49.0, Pcons2: 42.4 vs. 43.0). Finally, the Robetta "complete" models, which were submitted as the server's CASP/CAFASP predictions, manage to capture residues not provided by the alignment (average number of Cα atoms < 4 Å of Robetta "rebuilt" vs. Robetta "complete" models is BLAST: 137.9 vs. 141.2, PSI-BLAST: 119.3 vs. 123.3, Pcons2: 69.9 vs. 72.6; see Supplementary Table I), and are often quite good models.

Analysis of the server's performance with respect to our human group's models is complicated by several factors. The automated protocol was considerably more conservative than the human assisted protocol in terms of straying from the template structure and this is likely (unfortu-

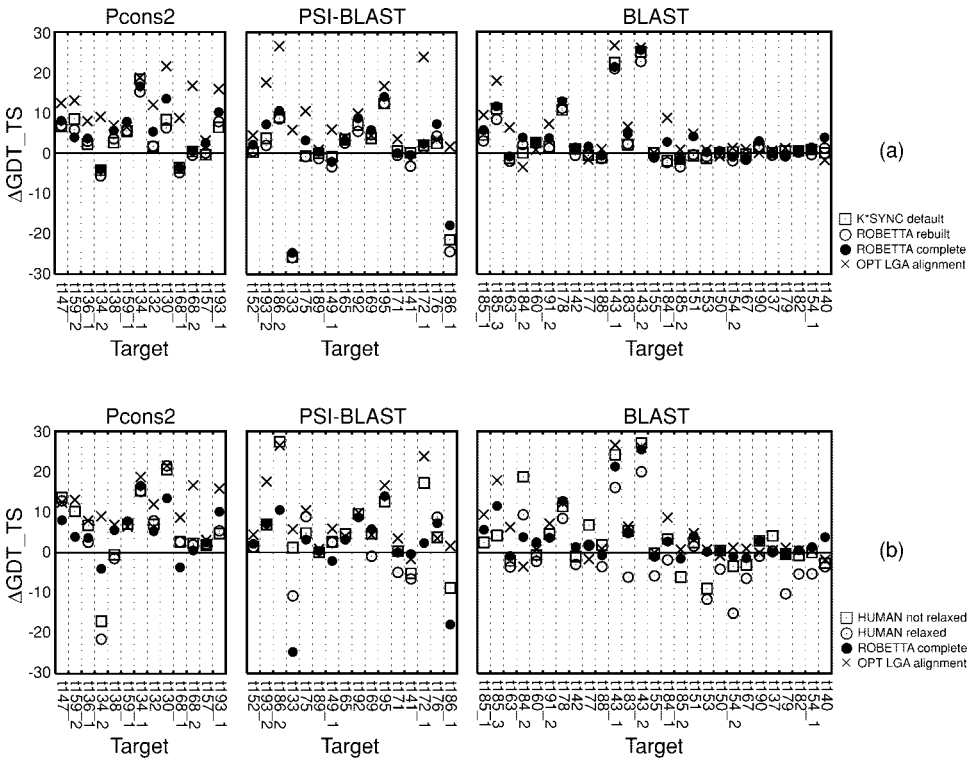

Fig. 3. Robetta and human homology models compared with base method models. (a) Difference in GDT_TS score achieved by Robetta model with respect to base method model. Within the detection category, targets are sorted by sequence identity with higher identity on the right. (b) Difference in GDT_TS score achieved by complete Robetta and Human models with respect to Base method model. (Plots made from data in Supplementary Table I).

nately!) to be the explanation for its superior performance with respect to our human predictions for the easier targets. Considerably larger regions of the query structure were modeled using *de novo* methods in the human assisted protocol and hence less of the template was utilized (average trimmed alignment length for Robetta vs. Human models is BLAST: 136.9 vs. 126.9, PSI-BLAST: 143.3 vs. 120.1, Pcons2: 99.6 vs. 82.3; see Table II and Supplementary Table I). For many targets, our human group also allowed conformational sampling in the template regions in hopes of pushing the model towards the true structure, which sometimes caused the model to move farther away rather than closer to the truth. The fact that our human group's more adventurous attempts to improve on the parent template usually either made no difference or made things worse highlights how far comparative modeling methods still have to go.

While we cannot compare relaxed human models to the Robetta models, other than to state that they are usually worse, we can make a reasonable comparison of the Robetta models with the un-relaxed human model (which, like the Robetta model, did not move the template regions). Such a comparison (Table II and Figure 3b) indicates that human intervention appears to have the greatest opportunity for enhancement of the model for the more challenging targets, in terms of detection difficulty and sequence identity (average GDT_TS scores of Robetta "complete" vs. Human "not-relaxed" models is BLAST: 75.3 vs. 75.0, PSI-BLAST: 50.6 vs. 52.2, Pcons2: 44.4 vs. 44.7; see Table II and Figure 3b). Even so, the server was among the better methods (including humans!) in the Fold Recognition category, both highlighting the quality of the parent detections from Pcons2 and suggesting that the strategy of building a template-based prediction from a confident Pcons2 detection or alternatively a *de novo* model is indeed a sensible approach. Additionally, even though human modeling of Fold Recognition targets led to an improved model in many cases, the automated method did occasionally manage to produce a prediction where the model was equivalent or superior to our human un-relaxed model prediction (e.g. T134_1 and T134_2, see Figure 4a). In this latter example, the automated alignment for domain 2 was not susceptible to second-guessing that our human intervention alignment fell prey to in a failed effort to improve the model quality.

Targets which were predicted by the automated *de novo* protocol were on the whole not close to the native structure, but not particularly worse than many other human groups, and often possessed good features. One reasonable prediction in this set was for T148 (see Figure 4b), for which both Robetta model 1 and model 3 correctly rendered the portion of the topology comprised of the helices and beta-hairpin for both domain 1 and domain 2. Additionally, these models indicated the two-domain nature of the target (it was not parsed into separate domains by Ginzu as it was sufficiently short for the Rosetta *de novo* method

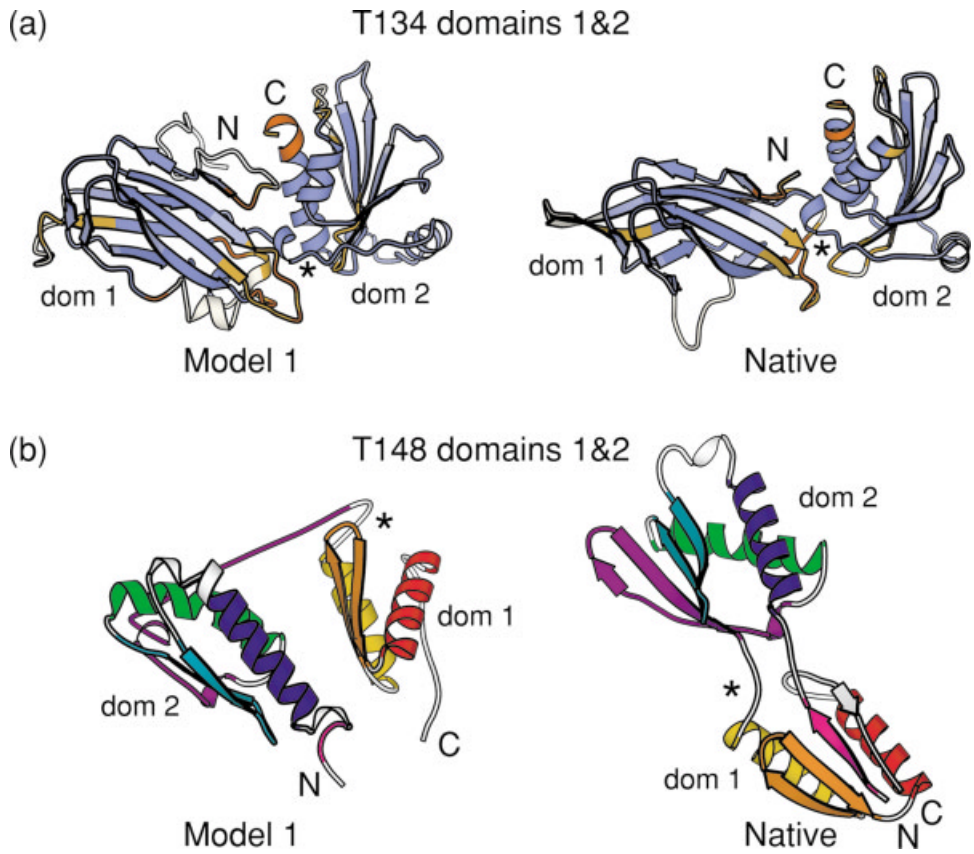

Fig. 4. Robetta model highlights. (a) The correct structure and our model of T134 domain 1 and 2 (delta-adaptin appendage domain from human), built following our template-based protocol from the parent 1qts (Ap-2 Clathrin Adaptor subunit from mouse). The different shades of blue indicate regions that were modeled as template, whereas red, yellow, and white indicate regions that were modeled as loops with our modified *de novo* protocol that takes into account the context of the template. The entire model was fit to the correct structure using the LGA server with a 4 Å cutoff. Dark blue and red, residues within 4 Å, light blue and yellow, residues less than 8 Å, and ice blue and white, residues greater than 8 Å from the corresponding atoms in the correct structure. The domain boundary is denoted by "*". (b) The correct structure and our *de novo* model for T148 domain 1 and 2 (HI1034 from *Haemophilus influenzae*). Residues are colored according to their role as secondary structure elements in the correct structure. The domain linker is denoted by "*".

to attempt) and the location of the linker. Interestingly, Rosetta simulations separate the chain into distinct domains that correspond roughly to the actual domain boundaries about 1/3 of the time even when the conformations of the individual domains are not correctly predicted (D.E.K., unpublished results).

What Went Wrong

Some lapses in model quality were attributable to implementation errors (bugs) that have since been resolved. The *de novo* models for T129 had the carbonyl oxygens misplaced. The alignment for T133 was exceptionally short due to a failure in the trimming logic. The homology model for T140 suffered from a collection of errors, which led to an exploded prediction. Among non-bug-related issues, there still remains considerable room for improvement in alignment quality for the homology-modeled targets. The default K*Sync alignments, while more complete in utilizing available parent coordinates for the template, were often much less accurate than our human predictions for targets with more distant parents. For example, the successful modeling of T186 domain 3 by our human group was never a

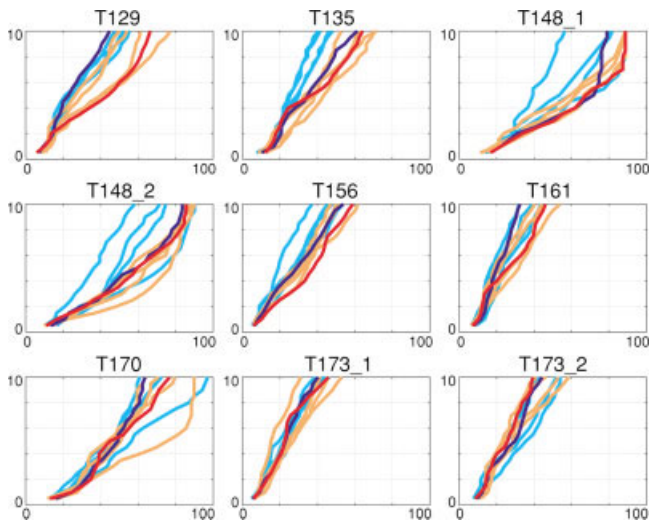

Fig. 5. De novo modeling protocol revisions: rerun versus CAFASP-3 models GDT. Global Distance Test plots for select *de novo* modeled targets show the net improvement after updates to the *de novo* protocol. Models produced during CAFASP-3 are in blue (model 1) and cyan (models 2-5), models produced by the Robetta rerun are in red (model 1) and orange (models 2-5). The x-axis is the percentage of residues that can be superimposed on the correct structure within the distance cutoff (in Å) specified on the y-axis.

possibility for Robetta, which failed to obtain an alignment of sufficient quality for this target's TIM-Barrel domain 2, and therefore didn't model domain 3 as a long loop with the correct template context (many of these residues in fact were treated incorrectly as template by Robetta). The default K*Sync alignment method employed during CAFASP-3, while fairly decent on average, often does not give the best possible alignment for a given parent (as compared to the "optimal" structure-structure alignments and of more immediate concern even compared to the superior human group alignments in Figure 3b), and therefore improving alignment quality remains a continuing area of research.

In the cae of the *de novo* modeled targets, an obsolete version of the Rosetta code was accidentally used, and additionally, the clustering routine used to select the final models did not properly exclude redundant predictions. In an effort to ascertain the expected performance of the current version of the server for targets that were modeled with the *de novo* protocol, we have rerun certain targets. The revised GDT results[9] for those targets are shown in figure 5. While the sample size is small, it does appear that the revised energy function and clustering protocol yields at least comparable results, and in several cases (T129, T135, T148_2, and T170) makes a significant improvement. Still, we can expect for the near term that even with the improvements the server will at best approach the rate of human success for *de novo* targets, producing models that resemble the native structure only some of the time.

What We Learned

Excitingly, the quite good performance of Robetta and other servers in CASP-5 [see assessors' reports in this issue] suggests that automated structure prediction is approaching the accuracy of human experts, continuing a trend that was first noted in CASP-4.[10,11] However, there remains room for improvement of server methods as there was sometimes a gap in quality between the best human predictions and those from servers for some of the more difficult targets.

The Robetta server can potentially be improved by incorporating other methods that capture features of the approaches used by humans. For example, for the more difficult homology modeling targets, the human predictions were often better than the Robetta prediction because the alignments were superior. The human alignments were selected using the Rosetta centroid-based and full-atom energy functions from large ensembles of alternate alignments created by systematically varying the weights on the different terms in the K*Sync alignment scoring function. This process has been included in the current version of the Robetta server. Additionally, the modeling for some of the New Fold category targets by our human group took better advantage of multiple sequence alignment information (e.g. T135 and T173), and it may be possible to generalize and automate some of what was done for these targets.

While the server will continue to undergo improvement as we better understand and attempt to automate what we as humans do to make good predictions, the initial system, tested by CASP-5 and CAFASP-3, performed beyond our expectations. Straightforward Comparative Modeling targets were well modeled with the template-based protocol, and more challenging Fold Recognition targets were often modeled quite reasonably by the template-based or *de novo* protocol. New Fold category predictions in some cases shared regions in common with the native structure, possessing revealing features that may guide further modeling.

METHODS

Domain Assignment and Parent Identification

The first part of the modeling process consists of determination of the locations of putative domains in the query sequence, assignment of domains to the appropriate protocol, and identification of any likely homologs with experimentally characterized structures. These steps are not decoupled, since the ability to assign a region of the target to a known protein structure greatly increases the likelihood that it is at least one protein domain. The approach we have implemented, called "Ginzu" (see Figure 2 for an illustration), consists of scanning the target sequence with successively less confident methods to assign regions that are likely to be domains. Once those regions are identified, cut points in the putative linkers are determined, if possible a single parent PDB chain is associated with each putative domain, and for each putative domain the homology modeling or *de novo* protocol is then initiated.

The initial scan attempts to identify the closest relatives with experimental structures to regions of the query sequence. A straightforward BLAST search[4] against the PDB sequence database[12] detects such relatives. All PDB ids that are detected at this stage are stored. A PSI-BLAST search[4] is then used to detect more distant relatives of the query, as well as provide more complete coverage since such alignments tend to be longer. Non-overlapping regions that possess the best combination of detection confidence and length of coverage are assigned as domains. The associated PDB id and region of the chain matched is retained but not the details of the alignment itself.

Currently, consensus fold recognition methods produce the most reliable fold assignments [see CAFASP-3 and LiveBench-6 results in this issue]. Since the express purpose of our method is to attempt to produce the best-possible model by utilizing the best-possible methodologies, we therefore decided to use Pcons2[5] [and described in this issue] for identification of putative parent PDB ids for any remaining regions of the query that have not already been associated with a parent PDB. Again, as with the PSI-BLAST detected parents, non-overlapping detections are assigned to the query as regions to be modeled as independent domains, and PDB ids and regions are recorded but the alignment discarded.

Any remaining long regions of the query that do not have structural homologs identified are considered suitable for *de novo* modeling, but may require further division into putative domains (For an illustration of how this is accomplished, see Figure 2). After all regions of the query that are likely a contiguous domain are assigned from a

PSI-BLAST or Pcons2 search, or potentially from a Pfam[13] search with HMMER[14] (Pfam search not used in CAFASP-3), any long remaining regions must be further divided into lengths accessible to the Rosetta *de novo* protocol (not much more than about 200 residues). Additionally, potentially excessive "linker" regions between regions of domain confidence must be cut to permit modeling with the domain they are most likely to be structurally associated with. Cut points are selected via a heuristic that considers strongly predicted loop regions by PSIPRED,[15] clusters of sequences in the PSI-BLAST MSA, the least occupied positions in the MSA, and distance from the nearest region of increased domain confidence. A fourth term that boosts the likelihood of a domain boundary in regions of the MSA where the sequences frequently begin or end was added after the CASP experiment.

At this stage, the query has been parsed into putative domains, and parent PDB ids have been associated whenever possible. These domains are passed to either the template-based or *de novo* modeling protocol for structure prediction.

Template-Based Modeling Protocol

The alignment method used by Robetta during CASP-5 and CAFASP-3, called "K*Sync", simultaneously uses residue profile-profile comparison, secondary structure prediction, and information about elements that are obligate to the fold in a dynamic programming approach[16] to produce a single "default" alignment. Pair terms include a PSI-BLAST generated residue substitution profile-profile comparison by inner-product, producing a distribution which is adjusted to possess a mean below zero and a standard deviation of 1.0 in the same fashion as FFAS.[17] Parent residue profiles are adjusted to include counts from the FSSP multiple structural alignment[18] to allow for more distant residue sampling. Secondary structure is added into the pair scores by giving a bonus to matches of PSIPRED predicted query regular secondary structure with DSSP[19] assigned parent regular secondary structure, and penalizing mismatches, weighted by the confidence of the prediction. A novel pair term is then included to attempt to match positions that are obligate to the fold.

Positions that are usually occupied in a multiple alignment are assumed to be obligate to the fold, whereas infrequently aligned positions are likely insertions (or at least conformationally variable) with respect to the core elements. The obligateness of a position in the parent sequence is based on the occupancy of the position in the FSSP multiple structural alignment, and in the query is based on the PSI-BLAST multiple sequence alignment. A bonus is given to matches of obligate positions with each other and a penalty to matches of obligate positions with insertions, with weighting based on the degree of occupancy of the obligate position. Finally, the pair distribution is again adjusted to restore the mean and standard deviation.

Gap penalties are accomplished with position specific gap initiation and gap extension penalties for each of the query and parent. Each position starts with a base value that is appropriate to a sequence-only alignment (again, similar to FFAS), to which are added structurally determined penalties. The values are adjusted to penalize failure to align obligate positions (by increasing the gap extension penalty at such positions) and for inserting a gap between two obligate positions (by increasing the gap initiation penalty at such positions). An additional gap initiation penalty is added to parent positions possessing regular secondary structure. The final distributions of values are not adjusted.

Dynamic programming is then performed to produce a single default alignment (either local-local or local-global in scope, depending on whether the homologous parent region falls within a known domain), which is used to generate a template into the aligned positions. Borders of unaligned regions ("stems") are trimmed back by one or two residues (or as many as necessary to make the loop at least five residues long) to allow more flexibility in the subsequent loop modeling steps.

Loop regions are then modeled in the context of the fixed template using Rosetta fragment assembly. For short and medium loops (< 17 residues), ~300 initial conformations are selected from a database of known structures using similarity of sequence, secondary structure, and stem geometry. The conformations of medium loops (12-16 residues) are then optimized for loop closure and energy using fragment-insertionand random angle perturbations. A gap closure term in the potential in combination with conjugate gradient minimization is used to ensure continuity of the peptide backbone. Optimization of variable regions is accomplished by use of the standard Rosetta potential with a centroid representation of the side-chains. All variable regions are optimized simultaneously starting from a random selection of initial conformations to ensure loop conformations compatible with the stems, the rest of the template, and the other loops. Generally, ~1000 independent optimizations are carried out. The set of loops that produces the lowest energy model is added to the template, and longer loop regions ($>= 17$ residues) are modeled in the context of this revised template. Initial conformations are built up using three and nine residue fragments, as in the full *de novo* protocol, but in the context of the template, followed by closure optimization. About 100 independent simulations are carried out, with a backbone-dependent side-chain rotamer library and a full-atom energy function used to select the lowest energy conformation.[8]

De Novo Protocol

Robetta employs a *de novo* protocol quite similar to that described previously.[3,20] For the purposes of a server, time and space limitations do not permit the generation of an enormous decoy ensemble. During CAFASP-3, Robetta generated 4000 decoys for the query itself and 2000 for each of up to two sequence homologs (since raised to 10000 for the query and 5000 for each of the sequence homologs). Up to 1000 lowest energy query decoys and 500 decoys for each sequence homolog that pass contact-order and strand topology filters are then clustered, with the top four cluster

centers returned as the four top-ranked models. The model possessing the lowest side-chain centroid based energy that is not a member of the clusters represented by the first four models is selected as the fifth model.

Assembly and Side-Chain Repacking

If the query is modeled as more than one domain, the models for individual domains are assembled into a contiguous model. This was not done during CAFASP-3 (multi-domain models were merely placed within the same file spaced by 100 Å), but currently is attempted by the Robetta server by fragment- insertion in the putative linker region(s) to orient the domains in a compact structure. The domain assembler remains under development, and therefore this stage may not do much more than cosmetic enhancement of the model. Finally, side-chains are repacked using a Monte Carlo algorithm[8] with a backbone-dependent side-chain rotamer library.[7]

Versions and Parameters

BLAST and PSI-BLAST parent detections were done using PSI-BLASTv 2.2.2,[4,21] starting from BLOSUM62[22] against the pdb_seqres.txt[12] and using the non-redundant sequence database from the NCBI (nr). The iterative detection was done via automatic restart from a checkpoint file against the pdb_seqres.txt after 5 rounds of profile building against the nr, with an e-value for inclusion of .001 or closer.

Pcons2 uses the following servers as input: PDB-BLAST,[23] mGenTHREADER,[24] FUGUE,[25] Sam-T99,[26,27] 3D-PSSM,[28] BIOINBGU,[29] and FFAS.[17] During CAFASP-3, detections were used if they were longer than 30 residues and had Pcons2 consensus confidence of 1.5 or higher.

Ginzu uses PSIPREDv2.01[14] and 5 rounds against the nr with PSI-BLASTv2.2.2 starting from BLOSUM62, e-value for inclusion and reporting .001 or closer.

K*Sync uses PSI-BLASTv2.2.2 with BLOSUM62 for 2 rounds e-value<=1E-06 against the nr followed by one round e-value<=.001, secondary structure from PSIPREDv2.01, and structural alignment of the parent with structural homologs from the FSSP server[15] (Z >= 7.0).

LGA structure-structure alignments and GDT analysis done with LGA server[9] located at http://predictioncenter.llnl.gov/local/lga-form.cgi. Comparison of models to native structure done with sequence-dependent fit at 4 Å (using the options: "-3 -sda -o1 -d_4.0 -lw_7"). Structure-structure alignment of target native structure to parent structure done with sequence-independent fit at 4 Å (using the options: "-4 -sia -o1 -d_4.0").

ACKNOWLEDGMENTS

The authors would like to thank the structural biologists for allowing their structures to be used in CASP and CAFASP, the CASP organizers and assessors for implementing the CASP-5 experiment, Dani Fischer for running the CAFASP-3 experiment, Arne Elofsson for the use of the Pcons2 server, Liisa Holm for the use of the FSSP server, Adam Zemla for the use of the LGA server, Kevin Karplus for the use of the SAM-T99 software, David Jones for the use of the PSIPRED software, Jens Meiler for the use of the JUFO software, and Leszek Rychlewski for help integrating Robetta with the BioInfo Meta server and for helpful discussions. We would also like to specially thank all server developers, whose collective work is so essential to consensus and consensus-derivative methods such as Robetta. Like the CASP experiment itself, the success of such methods shows that the science of protein structure prediction can best be furthered by our working together as a community. The authors would also like to thank Keith Laidig and Formix for effective and innovative administration and design of the Robetta hardware resources. D.C. is a PMMB fellow, administered by the Florida State University with funding from the Burroughs-Wellcome Fund. This work was also supported by the NIH and the HHMI.

REFERENCES

1. 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.
2. 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.
3. Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, Baker D. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;Suppl 5:119–126.
4. 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.
5. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.
6. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with Rosetta. Proteins 2003.
7. Dunbrack RL, Jr., Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
8. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97: 10383–10388.
9. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins 1999;Suppl 3:22–29.
10. Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001;Suppl 5:55–67.
11. Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL, Jr. CAFASP2: the second critical assessment of fully automated structure prediction methods. Proteins 2001;Suppl 5:171–183.
12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
13. 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.
14. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14: 755–763.
15. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
16. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
17. 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.
18. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
19. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
20. Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D. De novo prediction of three-dimensional structures for major protein families. J Mol Biol 2002;322:65–78.
21. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
22. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
23. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. Structure prediction meta server. Bioinformatics 2001;17:750–751.
24. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.
25. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 2001;310: 243–257.
26. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.
27. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R. What is the value added by human intervention in protein structure prediction? Proteins 2001;Suppl 5:86–91.
28. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
29. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000;119–130.