Automated prediction of CASP-5 structures using the ROBETTA server

Dylan Chivian[1], David E. Kim[1], Lars Malmstrom[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, WA, USA
2- Los Alamos National Laboratory, Los Alamos, NM, USA
3- Institute for Systems Biology, Seattle, WA, USA
4- University of California, Santa Cruz, CA, USA

*- dabaker@u.washington.edu
University of Washington
Dept. of Biochemistry and HHMI
Box 357350
Seattle, WA 98195
Phone: (206) 543-1295
Fax: (206) 685-1792

Keywords: Fully automated protein structure prediction server, CASP, CAFASP, Rosetta, fragment assembly, *de novo* modeling, template-based modeling, domain parsing, sequence alignment.

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 some reasonable predictions for targets in the recent CASP-5 and CAFASP-3 experiments.

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], andPcons2 [5][and described in this volume], followed by cutting the sequence into putative domains based on matches to known families and structures, 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 final models from the decoys clusters or from amongst other low energy decoys.

The template-based modeling protocol first requires an alignment to the parent. This is accomplished with 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. From this alignment a template is generated, and variable regions are then modeled in its context with a version of the Rosetta *de novo* method that allows conformational sampling for variable regions in the context of a fixed template (C.A.R., submitted). 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 Angstroms). The last step consists of repacking the side-chains using a backbone-dependent rotamer library [6] with a Monte Carlo conformational search procedure [7].

RESULTS

The protocol used 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, "□": 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. In all, models were often quite reasonable predictions, occasionally on par with the best models produced by human groups.

What went right

Encouragingly, the method performed quite well for some targets. For the targets with the closest structural homologs (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). The automated protocol was considerably more conservative than the human assisted protocol in terms of straying from the template structure and this is likely (unfortunately!) to be the explanation for its superior performance with respect to our human predictions.

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 used (Figure 3a). 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.

For targets with distant structural homologs (the Fold Recognition category) that were modeled with the homology modeling protocol, results were more mixed. Overall, 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 most cases, the automated method did occasionally manage to produce a prediction where the model was equivalent or superior to our human first 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 to handle) and the location of the linker. Interestingly for multidomain all alpha helical proteins and a subset of alpha-beta proteins, Rosetta simulations often separate the chain into distinct domains that correspond roughly to the actual domain boundaries 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 models for T129 had the carbonyl oxygens misplaced. The model for T140 suffered from a collection of errors, which led to an exploded prediction. Alignments for the homology modeled targets, while more complete than those from our human predictions for the targets with parents that are near sequence homologs, were also much less accurate for the more distant targets (Figure 3b). For example, the successful modeling of T186 domain 3 by our human group was never a 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 alignment method employed during CAFASP-3, while fairly decent on average, often does not give the best possible alignment for a given parent, and remains a continuing area of research.

For the *de novo* modeled targets, an obsolete version of the Rosetta code was accidentally used, and 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 [8] 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.

What we learned

Excitingly, the quite good performance of Robetta and other servers in CASP-5 [see assessors' reports in this volume] suggests that automated structure prediction is approaching the accuracy of human experts, continuing a trend that was first noted in CASP-4 [9, 10]. 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, we often made better predictions than the Robetta server for targets that were modeled based on distant structural homologs because of superior alignments. These 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 should be possible to automate. 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 rendered 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 were sometimes good approximations to the native structure, possessing revealing features that may guide further modeling.

ACKNOWLEDGEMENTS

The authors would like to thank the structural biologists for allowing their structures to be used in the 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, Leszek Rychlewski for help integrating Robetta with the BioInfo Meta server and for helpful discussions, and most especially all server developers. 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 HHMI.

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 [11] 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 volume]. 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 volume] 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). Once 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 [12] search with HMMER [13] (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), and 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 [14], the least occupied positions in the PSI-BLAST multiple sequence alignment, and distance from the nearest region of domain confidence. Additionally, a fourth term that boosts the likelihood of a domain boundary in regions of the PSI-BLAST MSA with sequence homolog termini has been 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 appear to be obligate to the fold from the FSSP server [15] in a dynamic programming approach [16] to produce a single alignment. Aligned regions used to generate templates following the mapping defined by the alignment, and borders of unaligned regions ("stems") were trimmed back by 2 residues (or as many as necessary to make the loop at least 5 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 replacement and 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, 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 [7].

***De novo* protocol**

Robetta employs a *de novo* protocol quite similar to that described previously [3, 17]. 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 each sequence homolog decoys that pass contact order filters and strand topology filters are clustered, with the top four cluster centers returned as the four top-ranked models. The model possessing the lowest side-chain centroid 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 Angstroms), 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 [7] with a backbone-dependent side-chain rotamer library [6].

Versions and Parameters

BLAST and PSI-BLAST parent detections were done using PSI-BLASTv 2.2.2 [4, 18] starting from BLOSUM62 [19] against the pdb_seqres.txt [11] 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 better.

Pcons2 uses the following servers as input: PDB-BLAST [20], mGenTHREADER [21], FUGUE [22], Sam-T99 [23, 24], 3D-PSSM [25], BIOINBGU [26], and FFAS [27]. During CAFASP-3, detections were used if they were longer than 30 residues and had Pcons2 consensus confidence of 1.5 or better.

Ginzu uses PSIPREDv2.1 [14] and 5 rounds against the nr with PSI-BLASTv2.2.2 starting from BLOSUM62, e-value for inclusion and reporting .001 or better.

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 PSIPREDv2.1, and structural alignment of the parent with structural homologs from the FSSP server [15] (Z >= 7.0).

FIGURES

Figure 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 render a full-chain all-atom complete model.

Figure 2. Ginzu Domain Parsing Illustration: MutS
MutS (T116 from CASP-4, PDB id 1ewq) was an 811 residue multi-domain target with three domains that can be modeled based on a structural homolog, and another domain that is a new fold. This target was inspirational to the Robetta multi-protocol domain modeling methodology, and well illustrates the concept of the Ginzu parent identification and domain parsing process. In (a) one can see the multiple sequence alignment that results from several rounds of PSI-BLAST after removal of redundant sequences and clustering based on region of query coverage. Also shown are the domain boundaries identified by the assessors, with the additional cuts in domain III indicating which portions are associated with region a and region b. In (b) below, are the domain boundaries that might be predicted by the Ginzu method in the following illustration. First, suppose that residues 29-98 of the query might be associated with 1a79 in a PSI-BLAST search (PSI-BLAST does not in fact find this relationship at an e-value cutoff of .001), as labeled by "A". Next, the remaining sequence 99-811 would be submitted to the Meta server, and Pcons2 might detect the relationship of MutS residues 550-739 to 1b0u with sufficient confidence to be utilized, as labeled by "B". These regions would then be "masked-out". Sequence clusters that do not overlap the masked-out regions and fall within a length range appropriate for modeling by the Rosetta *de novo* method are then assigned in order of cluster size, such that they too are non-overlapping. In this illustration, only the cluster labeled "C" fits these criteria. At this point, cut points between masked-out regions A, B, and C, and the termini of the query must be determined. The cut preference function is shown in (c), and is largely a combination of the terminal transition density in the multiple sequence alignment (d) and the predicted loop density (e). Other weaker terms in the cut preference contribute preference for positions of reduced occupancy in the multiple sequence alignment (not shown), and a positional bias (not shown). If a masked out region is sufficiently close to the N or C terminus of the query, the domain boundary is merely assigned there. If the unmasked region is excessively long for a linker (able to accommodate an entire domain), such as between regions A and C, and between C and B, then more than one cut is performed. This winds up being the correct behavior between A and C, well defining domain II, but not between C and B (although domain III is difficult to predict given the inserted nature of the III□ domain within III□).

Table 1. Robetta Modeling Protocol and Parent Detection Source
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 "y" were based on parents detected by PSI_BLAST, and those labeled with "pc" were based on parents detected by Pcons2. The categories are 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 treatment of this region of the query as merely 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.

Figure 3. Template-based Modeling Alignment Quality Comparison
The alignment quality measures (a) "completeness" and (b) "accuracy" for all domains that Robetta modeled with the template-based protocol. The "gold standard" for the alignment to a given parent was determined by a sequence independent fit with a 4 Angstrom cutoff by the LGA server. Completeness is defined as the percentage of the gold standard alignment achieved by an alignment, accuracy as the

percentage of the alignment that is in agreement with the gold standard. Our human group alignments are indicated with black circles, the Robetta alignments with open squares. The same parent was usually, but not always, used for the same target.

Figure 4. Robetta Model Highlights
(a) The native and our model for 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 entire model was fit to the native by the LGA server with a 4 Angstrom cutoff. 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. Dark blue and red show those residues that are within 4 Angstroms, light blue and yellow deviate less than 8 Angstroms, and ice blue and white are more than 8 Angstroms away from one another in the fit. The domain boundary is denoted by "*". (b) The native and our model for T148 domain 1 and 2 (HI1034 from *Haemophilus influenzae*). Residues are colored according to their role as secondary structure elements in the Native. The domain linker is denoted by "*".

Figure 5. *De novo* Modeling Protocol Revisions: Rerun vs. 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 y-axis represents a distance cutoff under which to fit the model to the native, and ranges from 0-10 Angstroms. The x-axis represents the percentage of the target that will fit below the distance cutoff, and ranges from 0-100 percent.

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. 1997;268:209-25
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. 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. 2001;Suppl 5:119-26
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. 1997;25:3389-402
5. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. 2001;10:2354-62
6. Dunbrack RL, Jr., Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. 1997;6:1661-81
7. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. 2000;97:10383-8
8. Zemla A, Venclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. 1999;Suppl 3:22-9
9. Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. 2001;Suppl 5:55-67
10. 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. 2001;Suppl 5:171-83
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. 2000;28:235-42
12. 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. 2002;30:276-80
13. Eddy SR. Profile hidden Markov models. 1998;14:755-63
14. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. 1999;292:195-202
15. Holm L, Sander C. Mapping the protein universe. 1996;273:595-603
16. Smith TF, Waterman MS. Identification of common molecular subsequences. 1981;147:195-7
17. 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. 2002;322:65-78
18. 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. 2001;29:2994-3005
19. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. 1992;89:10915-9
20. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. Structure prediction meta server. 2001;17:750-1
21. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. 1999;287:797-815
22. Shi J, Blundell TL, Mizuguchi K. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. 2001;310:243-57
23. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. 1998;14:846-56
24. 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? 2001;Suppl 5:86-91
25. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. 2000;299:499-520
26. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. 2000;119-30

27. Rychlewski L, Jaroszewski L, Li W, Godzik A. Comparison of sequence profiles. Strategies for structural predictions using sequence information. 2000;9:232-41

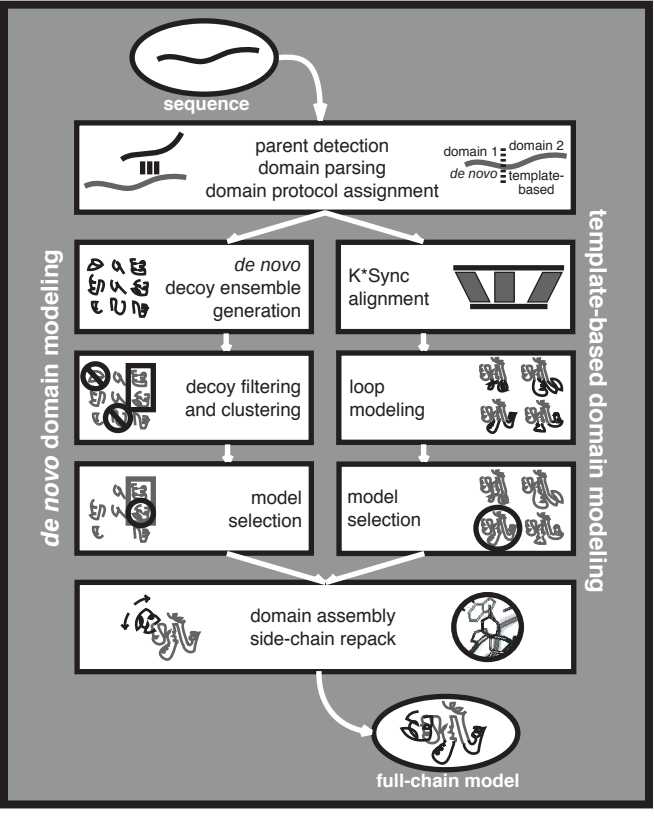

sequence
parent detection
domain parsing
domain protocol assignment
domain 1
domain 2
de novo
template-based
de novo domain modeling
de novo
decoy ensemble
generation
decoy filtering
and clustering
model
selection
template-based domain modeling
K*Sync
alignment
loop
modeling
model
selection
domain assembly
side-chain repack
full-chain model

Figure 1

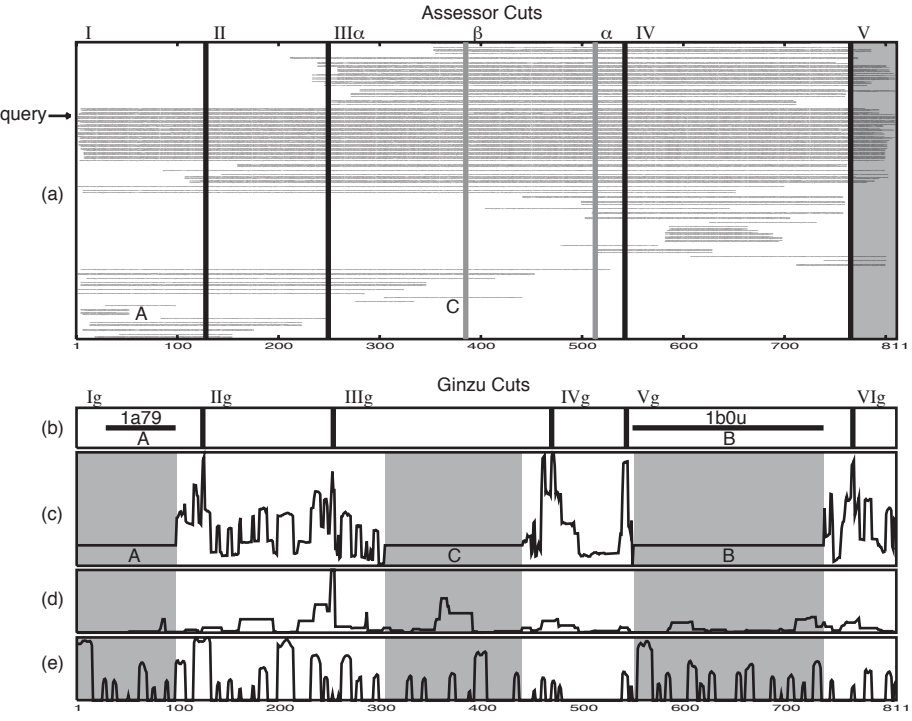
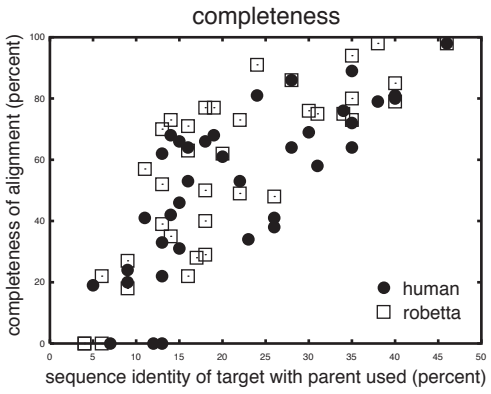

Assessor Cuts
I
II
IIIα
β
α
IV
V
query
(a)
A
C
1
100
200
300
400
500
600
700
811
Ginzu Cuts
Ig
IIg
IIIg
IVg
Vg
VIg
(b)
1a79
A
1b0u
B
(c)
A
C
B
(d)
(e)
1
100
200
300
400
500
600
700
811

Figure 2

(a)


completeness
completeness of alignment (percent)
sequence identity of target with parent used (percent)
human
robetta

(b)

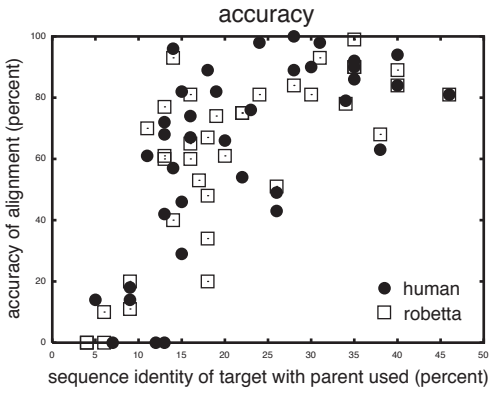

accuracy
accuracy of alignment (percent)
sequence identity of target with parent used (percent)
human
robetta

Figure 3

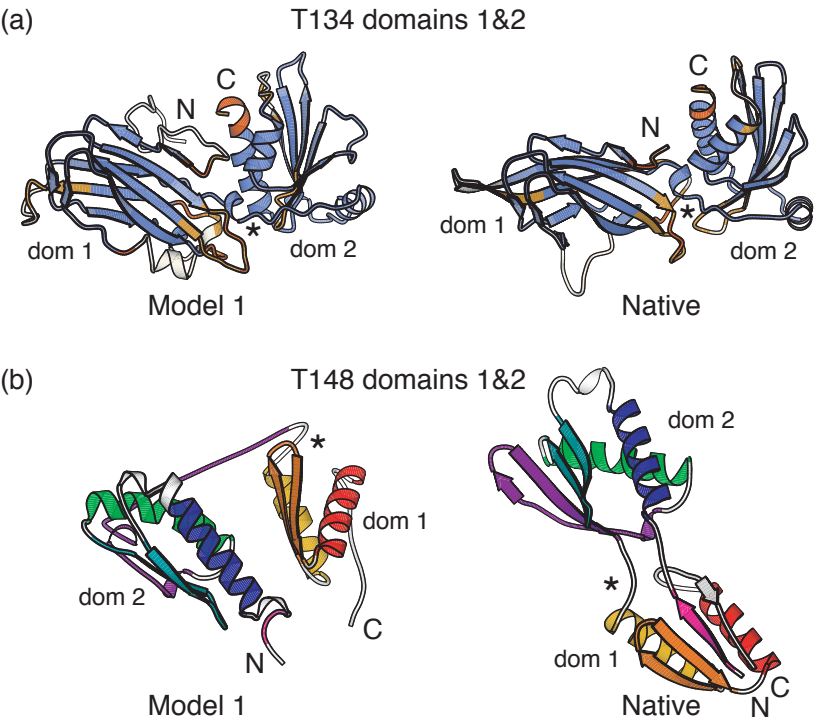

(a)
T134 domains 1&2
C
N
dom 1
dom 2
Model 1
C
N
dom 1
dom 2
Native
(b)
T148 domains 1&2
dom 1
dom 2
C
N
Model 1
dom 2
dom 1
C
N
Native

Figure 4

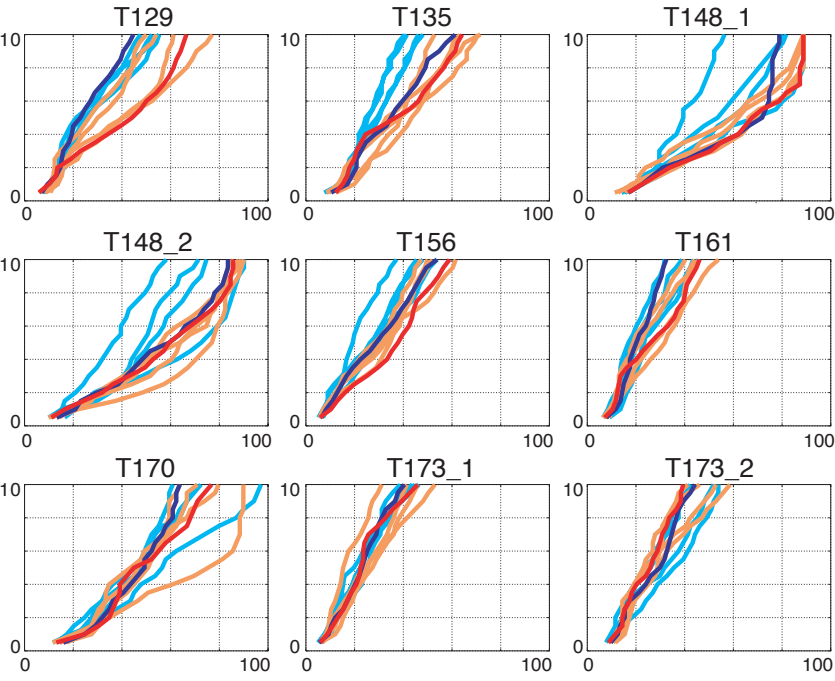


Figure 5