Stochastic Pairwise Alignments and Scoring Methods for Comparative Protein Structure Modeling

Adam C. Marko, Kate Stafford, and Troy Wymore*

Pittsburgh Supercomputing Center, National Resource for Biomedical Supercomputing, 300 South Craig Street, Pittsburgh, Pennsylvania 15213

Received November 1, 2006

Despite recent advances in fold recognition algorithms that identify template structures with distant homology to the target sequence, the quality of the target-template alignment can be a major problem for distantly related proteins in comparative modeling. Here we report for the first time on the use of ensembles of pairwise alignments obtained by stochastic backtracking as a means to improve three-dimensional comparative protein models. In every one of the 35 cases, the ensemble produced by the program probA resulted in alignments that were closer to the structural alignment than those obtained from the optimal alignment. In addition, we examined the lowest energy structure among these ensembles from four different structural assessment methods and compared these with the optimal and structural alignment model. The structural assessment methods consisted of the DFIRE, DOPE, and ProsaII statistical potential energies and the potential energy from the CHARMM protein force field coupled to a Generalized Born implicit solvent model. The results demonstrate that the generation of alignment ensembles through stochastic backtracking using probA combined with one of the statistical potentials for assessing three-dimensional structures can be used to improve comparative models.

INTRODUCTION

One of the driving forces for determining the structure of proteins is to aid in determining their function. Crucial insight into substrate binding, modes of inhibition, enzymatic chemistry, and protein−protein interactions can be obtained through the X-ray or NMR determination of protein structure, though often this is insufficient for a full accounting of a protein's function due to the dynamic nature of proteins. Despite the fact that the number of protein structures with known three-dimensional structures is steadily growing due in part to various Structural Genomics Initiatives,[1,2] the number of known protein sequences is several orders of magnitude larger than will be solved in the near or even distant future. Yet, with continuing advances in algorithms that (1) identify sequences with known structures (the template sequence) that are related to the target sequence, (2) align the template and target sequences, and (3) can explore conformational space to locate native conformations for insertions, there is much enthusiasm that these modeling techniques will provide structures that can assist in more complete functional assays of target proteins.[3−5]

Though the sensitivity of fold recognition algorithms based on sequence and structural information has greatly increased, the alignment between the target and template can still contain significant errors. Sequence alignments are generally correct if the amino acid sequences of the target and template exhibit 50% or more sequence identity. As sequence identity drops below this level, alignment errors in loop regions begin to emerge. More dramatic errors begin to occur when the sequence identity falls to 30% or less, resulting in misalignment of secondary structure elements. As shown by Jaroszewski et al.,[6] these alignment errors span the range from being "correct" to being completely wrong when this 30% threshold is passed, often referred to as the twilight zone.

Aligning two protein sequences can be done efficiently with dynamic programming algorithms[7,8] which require the assignment of a substitution matrix and penalties for introducing gaps into the alignment. Substitution matrices have been developed based on several criteria; structural equivalence of numerous protein sequences, genetic code similarity, chemical similarity of amino acids, hydrophobicity index, and other physical property indices.[9] Often there are two to three types of gap penalties: the open gap penalty, a penalty for extending the gap, and end-gap penalties. The quality of a pairwise sequence alignment can vary dramatically by the particular choice of substitution matrix and gap penalty scheme.[6,10] Furthermore, for sequences of two proteins that have even a small region of variability there may be many alignments of equal or almost equal score.[11] Though some programs may give information on the reliability of aligned pairs over the entire alignment, they often do not output alternative alignments. Alternative alignments could result in the construction of better comparative protein models.

The investigation of suboptimal alignments has been reported by several researchers.[12−14] Recently, the program probA was introduced that extends previous approaches to stochastic pairwise alignments by a probabilistic backtracking procedure to generate ensembles of suboptimal alignments.[15] The algorithm uses a statistical thermodynamic interpretation of the alignment problem where alignment scores are analogous to energy.[16,17] The stochastic pairwise alignment program and standard alignment program differ in the

* Corresponding author phone: (412)268-4960; fax: (412)268-8200; e-mail: wymore@psc.edu.

10.1021/ci600485s CCC: $37.00
Published on Web 03/29/2007

induction phase where the alignment matrix is populated with probabilities instead of scores and pointers. The trace back is then done stochastically instead of with deterministic rules. Repeated application of the stochastic alignment program produces an equilibrium sample of alignments. The program probA was able to obtain correct alignments (as judged by a structural alignment) between two proteins sharing 14% sequence identity with significant probabilities, even though the optimal alignment deviated significantly from the structural alignment. These correct alignments came from an ensemble of *1 million*. It would clearly be of interest to determine if the program probA could assist in improving comparative protein models by constructing a practical number of alternative alignments. This is the first report that examines the use of stochastic pairwise alignments for improving comparative protein structure models.

If a protein modeler is to utilize alternative alignments as part of a comparative modeling strategy, then a method for identifying the best model among the ensemble is essential. A few groups (including our group) in the sixth round of the Critical Assessment of Techniques for Protein Structure Prediction (CASP6) were described using alternative alignments and fold assessment methods as part of their comparative modeling strategy. The Multiple Model Approach described by the Godzik[6] group constructed alternative alignments by varying the substitution matrix after optimizing the gap penalty scheme and utilized a threading energy to pick the structure closest to native. They discovered that the percent of conserved residue contacts can change by a factor of 2 depending on alignment parameters, the alignment differences were not just confined to the loop regions, and that in some cases the best z-score calculated by the sequence alignment program produced the worst structural model. An in silico protein recombination method developed in the Bates[18] group simulated artificial genetic selection on a population of single-template models created from different templates and different sequence alignments per template. The number of alternative alignments was small (in the range of 5−10). A fitness function derived from a free energy estimate based on residue−residue contacts and solvation energies was used to guide the optimization. This method was able to recover from some alignment errors present in the initial population with the limitation that partially correct alignments must be present in the initial population. The method also consistently converged around the optimal template's conformation which is not readily identified through sequence alignment programs. Recently, another method that combines alignments from several sources in variable regions and finds the best combination to improve comparative protein models has been reported.[19]

John and Sali[20] have reported on a very promising method that iteratively optimizes both the alignment and the model. Alternative alignments were constructed by application of a number of operators such as alignment mutations and crossovers in a genetic algorithm. Both sequence similarity and statistical potentials were used to guide the optimization process. A composite score made up of five Z-scores that calculate various aspects of the protein structure was used to select the final model. For all of the 'very difficult' targets, the method was able to increase the overlap with the native structure over their initial PSI-BLAST alignment. For most cases the scoring function was able to identify a better structure among the ensemble created though often not the one closest to native.

Chivian and Baker[21] have demonstrated success in comparative modeling using large scale parametric alignment generation with consensus and energy-based model selection. They found that generation of large ensembles of alignments can improve the alignment quality and that they were able to frequently sample near the best possible alignment (as compared to a structural alignment). Despite this, the researchers still faced challenges in sampling and selection. For example, in the most difficult targets, the near best possible alignments were never sampled even with the large ensemble size.

As shown by the studies mentioned above, a method for fold assessment must be used to distinguish between several different protein models with ideally the one closest to native being identified as such. There are several reported methods for fold assessment; many of which are referenced in an article by Melo, Sanchez, and Sali.[22] These methods can have precision that ranges from ranking structures as coarsely as whether or not the fold is correct in distinguishing similar models from an ensemble. For our purposes, we will be most interested in the latter methods since it is often the case in comparative modeling that a correct fold is first identified through sequence-based methods. Two popular methods for ranking protein structures are statistical potentials, sometimes called knowledge-based potentials, and potentials based on molecular mechanical (MM) force fields. Statistical potentials are derived from known protein structures and quantify the observed preference of the different residue or atom types to be exposed to the solvent or to interact with each other in a pairwise or higher order fashion.[23] MM force fields have been designed by several groups for proteins and are highly parametrized to reproduce vibrational and torsional potential energy surfaces and nonbonded interaction energies of amino acid fragments while also producing stable molecular dynamics trajectories of proteins at room temperature in bulk water.[24,25] Because the surrounding water molecules play a central role in the thermodynamics and structure of proteins yet can represent a large fraction of the computational expense, implicit solvent models have been developed and applied to calculate the solvation energy of proteins.[26,27] The combination of the two terms (molecular mechanics energy plus solvation energy) has been used to distinguish models in comparative modeling[28] and ab initio folding simulations.[29]

Finally, in order to assess the performance of the stochastic alignment protocol and the methods to rank protein structure, a procedure for comparing model structures with those taken from the X-ray or NMR data must be employed. It is common to simply report the overall root-mean-square deviation (rmsd) of all corresponding Cα atoms after a superposition transformation has been performed. Yet, a small deviation in just one part of the protein such as a loop region can create a large rmsd. A measure that has been used extensively by the assessors for CASP has been the Global Distance Test-Total Score (GDT_TS herein after referred to as GDT).[30] GDT values can range from 100 for essentially an exact match of the native structure on the Cα level to very low single digits for models with no recognizable native structure.

EXPERIMENTAL DETAILS

All the target-template pairs that were selected to analyze the performance of probA for comparative modeling applications were either past CASP5 or CASP6 pairs taken from the literature[31,32] or were ones from the paper by John and Sali.[20] The identification of suitable template(s) can be challenging when the sequence identity between target and template is very low, but we do not address this issue here.

Alternative sequence alignments between template and target were obtained from the program probA.[15] This program uses a probabilistic interpretation of the sequence alignment problem. For each target-template pair we constructed 2000 alternative alignments with the following parameters. The "temperature" was set to 1.0. In the limit that $T \rightarrow 0$, the probability is equal to zero for all alignments with a score less than the maximal score. In the limit that $T \rightarrow \infty$, all alignments have the same probability. Therefore, we can interpret T as a measure of our interest in alternative alignments. The program also selected the appropriate scoring matrix by doing an initial alignment of the two sequences and calculating a PAM-distance.[33] The Gonnet series[34,35] of scoring matrices were used (see Table 1).

Perl scripts were written to translate the output of probA into input for the program MODELLER version 6.2,[36] though these scripts have been updated to work with the current release. MODELLER was used to construct all 2000 models. A sample input script is given in the Supporting Information. We set the md_level to 'nothing' which instructs the program to minimize the restraints generated by the sequence alignment and other stereochemical features of proteins but does not perform any extra optimizations or molecular dynamics.

The graphical programs VMD[37] from the Resource for Macromolecular Modeling and Bioinformatics and the UCSF Chimera[38] package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081) were used to visualize the structures and create figures.

The resulting 3D models were assessed with three statistical potentials: ProsaII,[23] DFIRE,[39,40] and DOPE.[41] The result used from the ProsaII energy was a summation of the Cα-Cα, Cβ-Cβ, and surface potential energies. A sample input script used to run ProsaII is given in the Supporting Information. DOPE and DFIRE are all-heavy (not hydrogen) atom statistical potentials. All statistical potential analysis was performed on structures that were not minimized.

We used the all-22 MM potential energy function[24] implemented in CHARMM[42] coupled with a generalized born implicit solvent model[26] to calculate the MM/GB energy. This procedure was facilitated by use of the Multi-scale Modeling Tools for Structural Biology (MMTSB, www.scripps.mmtsb.edu).[43] MMTSB enables the analysis of ensembles of models without the need for additional scripting by the end user. Hydrogen atoms were added to the models using the CHARMM 22 parameters, and the structures then were minimized for 100 steps using a distance dependent dielectric with $\epsilon = 4.0$. In essence, this means that the structures analyzed with the statistical potentials were not in the absolute sense the same as the ones analyzed with the all-atom MM potential. But in practice, this procedure changes the structures in a nonconsequential manner on the level that the structures have been analyzed; that is the Cα positions. However, this process is required for all-atom models where the minor overlap of a few hydrogen atoms can raise the energy substantially despite the accuracy of the overall model.

For five targets, we took the all-atom structures obtained before minimization with a distance dependent dielectric and instead performed a 1000 step steepest-descent energy minimization with the all-atom CHARMM 22 parameters coupled to a Generalized Born implicit solvation model.[26]

Table 1. Information on the Target-Template Pairs

target	residues	template	template residues	sequence identity	Gonnet matrix	target fold class
1ATN	4-354:A	1atr	3-382	13.8	250	α/β
2SIM	35-374	1nsb	113-449:A	10.1	250	β
1T70 (T0200)	1-255:A	1ush	26-550	16.4	160	β
1YK3 (T0152)	10-209:D	1cjw	1-166:A	16.2	250	α+β
1L7A (T0165)	1-318:A	1qfs	432-706:A	15.4	160	α/β
1PV1 (T0195)	1-299:A	1dqz	3-282:A	16.3	250	α+β
1TZ9 (T0208)	1-353:A	1i60	1-278:A	11.4	250	α/β
1IYA (T0141)	1-187:A	1lba	1-146:A	23.9	160	α+β
1CPC	29-170:L	1col	71-187:A	10.2	250	α
1VL7 (T0234)	11-145:A	1nrg	49-261:A	10.6	300	β
2SAR	7-91:A	9rnt	2-104	11.5	300	α+β
1GKY	1-186	3adk	8-194	19.1	250	α/β
1MK4 (T0169)	1-181:A	1b87	1-156:A	17.5	160	α+β
1XFK (T0282)	43-332:A	1pq3	24-329:A	23.3	160	α/β
2PIA	6-225	1frn	35-314	14.5	250	α+β
1O14 (T0189)	1-319:A	1rkd	4-309:A	13.5	250	α/β
2MTA	45-125:C	1ycc	2-103	15.6	250	α
1N91 (T0176)	1-100:A	1jrm	1-104:A	18.9	250	α+β
T0265	1-109	1lj9	2-145:A	16.7	250	α+β
1TVG (T0211)	4-130:A	1jhj	11-162:A	12.8	250	α+β
2SAS	3-183	2scp	1-172:A	16.5	250	α
1XE1 (T0196)	17-108:A	1jny	4-429:A	29.4	160	β
1BBH	5-131:A	2ccy	5-128:A	21.3	160	α
1CHR	1-368:A	2mnr	4-359	17.9	250	α/β
1TEN	803-891	3hhr	131-233:B	18.4	250	β
1VLO (T0247)	32-365:A	1pj5	485-830:A	24.5	160	α+β
1XFK (T0282)	43-332:A	1gq6	9-309:A	22.6	160	α/β
1VM0 (T0205)	19-123:A	1h0x	9-97:A	25.9	160	α+β
1VGG (T0271)	1-161:A	1rlh	1-135:A	41.4	80	α+β
1WK4 (T0267)	1-174:A	1tiq	2-174:A	17.3	250	α+β
1WGB (T0274)	1-159:A	1i0r	1-161:A	22.9	160	α+β
1M3S (T0167)	2-185	1jeo	4-180:A	36.9	120	α/β
1WDV (T0266)	3-152:A	1dbu	1-157:A	24.2	160	α+β
1H7M (T0150)	1-96:A	1ck2	2-105:A	34.0	160	α+β
1VLC (T0246)	1-354:A	1cnz	1-363:A	56.7	80	α/β

The GDT of the protein with the lowest energy at the final step of the minimization was compared to the GDT of the lowest energy from the 100 step minimization performed with a distance dependent dielectric. This extensive minimization was only performed on five targets due to the amount of CPU time required per model.

The analysis of the structural properties including GDT, native contacts, and phi, psi, and chi1 angles was performed using the MMTSB toolset. We calculated the GDT scores using the typical cutoffs of 1.0, 2.0, 4.0, and 8.0 Å cutoffs. We also examined other structural features such as the fraction of native contacts, which is based on two heavy atoms within a residue having a separation less than 4.2 Å, and the percent correct phi, psi, and chi1 angles. These features closely parallel the analysis of GDT values and so are not included in this report.

The structural alignment models were created by aligning the template and target PDB structures using the Combinatorial Extension[44] Web server (cl.sdsc.edu/ce.html) and using the resulting sequence alignment as input to MODELLER. This allows us to compare our modeled structures to the 'structural alignment' model in terms of GDT rather than comparing the percent overlap in sequence alignments. Each of the targets contain some residues that are aligned to gaps in the structural alignment, though MODELLER does not have any information to build these residue's coordinates except for the position of the anchoring residues,[45] i.e. where the alignment continues with matched residues. Depending on the size of the insertion, it is possible to create an incorrect alignment but do so in a way that guides the conformation of the unmatched residues closer to the actual X-ray structure than the structural alignment model. Despite this possibility, we found the 'structural alignment' models to generally exhibit higher GDT and thus serve as a good comparison for what can be achieved when the alignment is correct.

RESULTS

Stochastic Backtracking To Generate Alternative Pairwise Sequence Alignments. For each of 35 target-template (t−t) pairs, 2000 alternative alignments were created that were subsequently used to construct a total of 70 000 three-dimensional protein models. An ensemble of 2000 models represents what can reasonably be constructed on a modern single processor workstation in 30 h. As shown in Figure 1, the 3D model with the highest GDT value from the ensemble (best probA) is always better than the GDT value of the model resulting from the optimal alignment. The data used to generate all the plots are contained in a separate table as Supporting Information (Table 1S). The number of probA alignments (out of 2000) that result in 3D models closer to the native GDT value than those resulting from the optimal alignment ranges from as little as 2 alignments to as many as 1894 alignments. Further details are contained in the Supporting Information (Table 2S).

Alignment sampling produced models that were within 5% of the structural alignment models for 16/35 targets (counting the targets which had slightly higher GDT values for the probA model with the highest GDT, see Methods for further explanation). The alignment sampling also generates many alignments that result in models that are further away from native than the optimal alignment. The optimal alignment

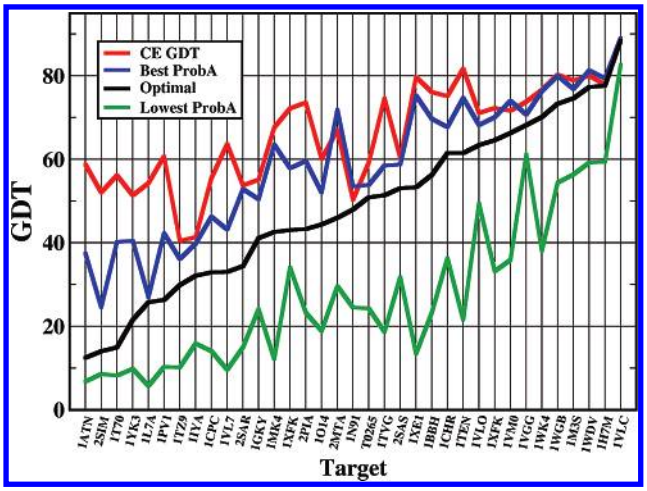

Figure 1. GDT values over all targets obtained from the optimal alignment (black line), highest value from the ensemble (blue line), lowest value from the ensemble (green), and from the structural alignment (red line).

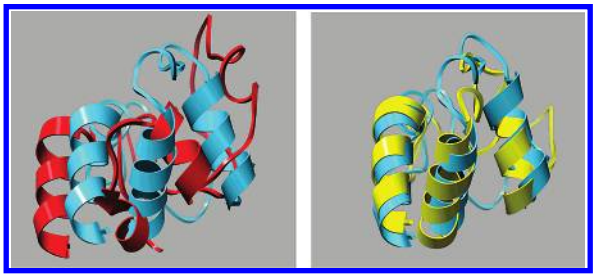

Figure 2. (left) Superposition of X-ray structure (blue) and optimal alignment model (red) for target 2mta. **(right)** Superposition of X-ray structure (blue) and best probA-generated model (yellow) for target 2mta (residues 10−60).

Table 2. Root-Mean-Square Coordinate Deviation (RMSCD) versus Crystal Structure for the Best Models Produced from the Iterative Method of John and Sali with the One Produced by Alignment Sampling with probA

target	iterative method population best	probA population best	target	iterative method population best	probA population best
1ATN	17.1	9.9	2MTA	3.1	3.4
2SIM	12.3	13.3	2SAS	3.9	4.0
1CPC	4.8	5.1	1BBH	2.6	3.1
2SAR	4.8	4.6	1CHR	3.2	3.0
1GKY	7.7	6.3	1TEN	4.9	2.3
2PIA	3.2	4.2			

GDTs span the range of 12.5−88.5, while the GDT of models built from structural alignments span a smaller range of 58.8−88.5. Table 1 shows that the sequence identity between t−t pairs is concentrated in the 10−25% range with two targets in the mid-30s and two at 41 and 57%. Noticeable improvement for targets that have optimal alignment models in the range of 50−60 GDT can be seen in Figure 1 for targets 1TEN, 1BBH, and 1XE1. In addition, two targets (2MTA and 1MK4) that have optimal alignment GDT values in the 40s show a large improvement. The difference is shown graphically for 2MTA in Figure 2. The graphic clearly shows not only that the helices are correctly aligned but also that their relative orientation to each other is greatly improved. Despite these large improvements for some targets, the targets with the six lowest optimal alignment GDTs have a best probA model (the one with the highest GDT values)

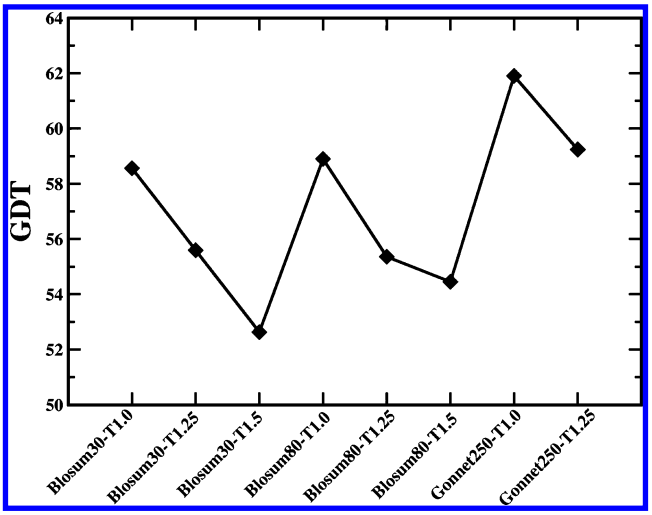

Figure 3. Highest GDT value obtained for target 2PIA among the 10K ensemble produced by the program probA using different scoring matrices and temperature (T) settings.

that deviates from the structural alignment by 21−53%. Four other targets (1VL7, 1XFK, 2PIA, and 1TVG) with higher GDT values for the best probA model deviate from the structural alignment model by 19−32%.

Due to the large deviation in the latter four targets mentioned above, we experimented with changing the scoring matrix, temperature, and number of alignments from 2000 to 10 000 for the target 2PIA. Visual examination of the sequence alignment showed no major insertions or deletions suggesting that a simple increase in sampling would achieve improved alignments. The results shown in Figure 3 demonstrate that increasing the temperature results in a degradation of the best probA model for all scoring matrices used. There is some minor improvement in the GDT (2.3) with increasing the sampling from 2 to 10K when using the default temperature (1.0) and scoring matrix (Gonnet 250) for this target. Finally, we tried increasing the number of alternative alignments to 100K using the Gonnet 250 scoring matrix and temperatures of 1.0, 1.05, and 1.1. The best model from these large ensembles improved the GDT value over the 2K ensemble by only 3.8. The structural alignment model would be an improvement in GDT over the best probA model of 14.0.

Scoring Functions. We examined four structural assessment methods by extracting the lowest energy model from each and comparing its GDT value with the best one from the ensemble. Our assessment of scoring functions differs from more traditional assessments in that the crystal structure is not part of the ensemble or "decoy" set. But as demonstrated by John and Sali[20] and Offman et al.,[46] these scoring functions are being used to guide the refinement of alignments and structures making this type of analysis relevant. The structural assessment methods were DFIRE,[39,40] DOPE,[41] and ProsaII[23] and the all-atom CHARMM 22 force field coupled to a generalized born implicit solvent model (MM-GB). DFIRE identifies models that are within 5% of the best GDT for 11/35 targets (see Figure 4), and it identifies models that are equivalent to or better than the optimal alignment in 27/35 cases. DOPE identifies models that are within 5% of the best GDT in 12/35 cases (see Figure 5), and it identifies models that are equivalent to or better than the optimal alignment in 28/35 cases. ProsaII identifies models

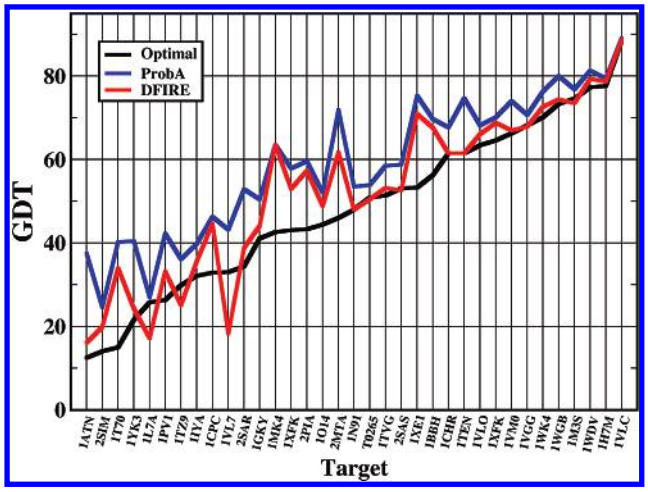

Figure 4. GDT values over all targets obtained from the optimal alignment (black line), highest value from the ensemble (blue line), and for the structure with the lowest DFIRE energy (red line).

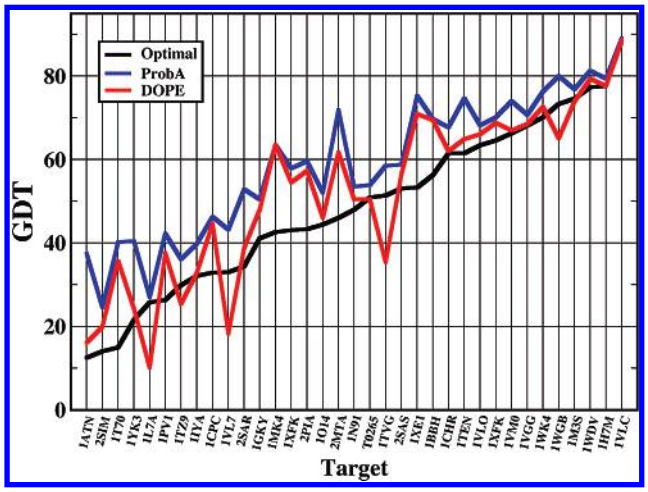

Figure 5. GDT values over all targets obtained from the optimal alignment (black line), highest value from the ensemble (blue line), and for the structure with the lowest DOPE energy (red line).

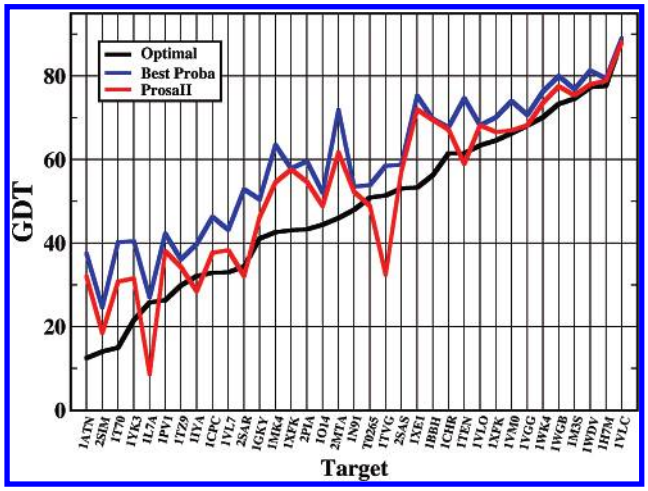

Figure 6. GDT values over all targets obtained from the optimal alignment (black line), highest value from the ensemble (blue line), and for the structure with the lowest ProsaII energy (red line).

within 5% of the best GDT in 15/35 targets, and it identifies models that are equivalent to or better than the optimal alignment in 29/35 cases (see Figure 6). MM-GB identifies models within 5% of the best GDT for 4/35 targets, and it identifies models that are as good as or better than the optimal

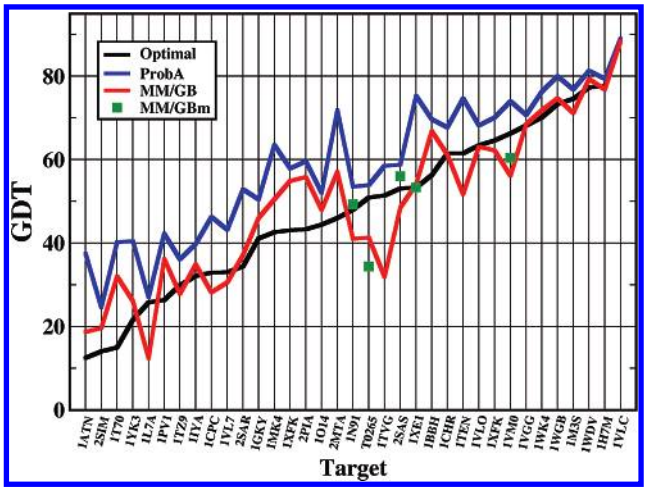

Figure 7. GDT values over all targets obtained from the optimal alignment (black line), highest value from the ensemble (blue line), for the structure with the lowest MM/GB energy (red line), and for the structure with the lowest MM/GB energy from a more extensive minimization procedure (green squares).

alignment in 19/35 cases (see Figure 7). For the five targets where the models were minimized more extensively and in the presence of the GB implicit solvent, the identification of models with a higher GDT value is obtained in 3 of the 5 targets with one target exhibiting insignificant change.

All of the potentials identify a model much lower in GDT than the optimal alignment for target 1L7A. The target is a multiple domain 318-residue protein in which the template had multiple insertions, and the best probA model only had a GDT value of 27.1. The target 1TVG was more noteworthy in that the best probA model had a much higher GDT value of 58.5 though DOPE, ProsaII, and MM/GB had the lowest energy model with a GDT near 35. The lowest energy structure by ProsaII contains significant alignment errors in secondary structure elements and loop regions including one region with a knot. Based on this visual cue, one would easily discount this structure. The lowest energy structure by MM-GB has errors in both secondary structures and loops. This model would be harder to discount based on visual features. The target contained many variable loop regions that likely led to the failure of the potentials to identify a model closer to native.

Secondary Structure Location of Sequence Alignment Errors. We graphically examined the structural superposition of models constructed from a structural alignment and the best probA models with those from the X-ray coordinates to examine where sequence alignment errors were occurring. The most difficult modeling targets all not only contain significant errors in loops but also have secondary structure elements that are totally misaligned (see Figure 8). These errors occur not only in the optimal alignment model but also in the best probA model. The easiest targets contain only minor errors in loop regions and in the N- and C-termini. The other targets in general only contain errors in loop regions, though the best probA model for 1bbh (see Figure 8) contains a gap in the middle of a helix. This structure has the best GDT value due to superior loop structures.

DISCUSSION

We have shown that the program probA can be used to construct a reasonable number of alternative pairwise

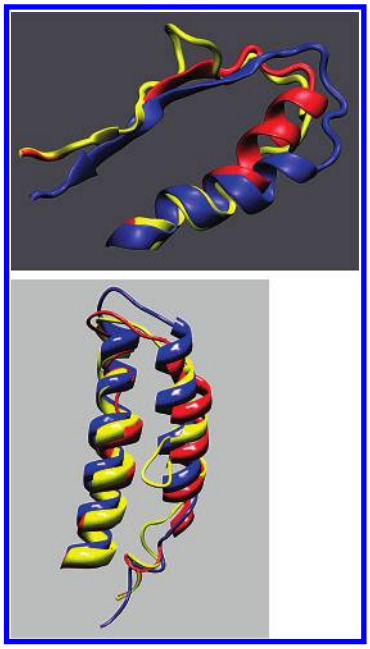

Figure 8. (top) Superposition of X-ray structure (blue), structural alignment model (red), and best probA model (yellow) for target 1xfk (residues 67−96) showing the misaligned β sheet region in the best probA model. **(bottom)** Superposition of X-ray structure (blue), structural alignment model (red), and best probA-generated model (yellow) for target 1bbh (residues 10−60) showing the misaligned helix in the best probA-generated model.

sequence alignments (2000) that result in comparative structural models closer to native than if an "optimal" alignment is used. This is true for all of the 35 cases in this study. Yet, for 19 of the 35 cases the best model generated differed from the structural alignment model by at least 5% and in a few cases by much larger percentages. This further demonstrates what has been reported before: structures are conserved to a greater degree than are sequences.[47]

Comparing the best models for 11 targets produced by probA (number of alternative alignments = 2000, $T = 1.0$ and assigning the scoring matrix through a calculated PAM distance) with the best one produced with the iterative method of John and Sali,[20] we find respectively results of similar accuracy (see Table 2). The computational cost of building the 2000 models with MODELLER with a model length of 150 residues and scoring them with DFIRE or DOPE is approximately 30 h on a 1 GHz Pentium III processor which is at least 80 times less expensive than the iterative method reported by John and Sali.[20]

The theory of probabilistic alignments is derived from a thermodynamic partition function in which the sequence alignment scores are analogous to the energy of the system. Therefore, increasing the sampling of alignment space to increase the likelihood of producing better alignments can be compared to running longer molecular dynamics simulations to increase the likelihood of observing conformational fluctuations. If the energetic barrier is low for converting from one conformation/sequence alignment to another, then these alternatives will be sampled more frequently than if the energetic barrier is higher. For distantly related sequences (t−t pairs), the optimal alignment may be separated from the structural alignment by several barriers of varying magnitude. In other words, the scoring/energetic landscape is rugged.[17]

As mentioned above, one way to navigate this landscape is to simply increase the number of alternative alignments. Unfortunately, the extent of alignment sampling needed in order to achieve the best possible result is difficult to determine a priori. The effect of increasing the number of alternative alignments for the t−t pair 2PIA-1FRN from 2K to 10K and ultimately to 100K was minimal. Another option would be to increase the temperature. While this latter option increases the diversity of sampled alignments, it also leads to increased probabilities of mismatching residues that should be matched. In general, we found that increasing the temperature above the default of 1.0 produces models further away from native.

CONCLUSION

Ultimately, it appears that some sort of iterative strategy will have to be implemented within a stochastic pairwise alignment program to significantly improve comparative protein models for difficult targets in which there is feedback from the scoring methods. For example, common sequence alignment features between models with low energies (either from the ProsaII , DOPE, or DFIRE statistical potential and/or the MM-GB potential) could be used to increase the probabilities of these features in a subsequent round of alternative alignment generation. This would effectively concentrate the diversity of alignments in areas where the best alignment has not yet been generated. Of course, this algorithm will require the protein structure scoring methods not only to identify the structure closest to native among a large ensemble but also to guide the optimization. Because of this, it is crucial to examine the behavior of the scoring methods in assessing models that were further away from native (in the range of 30−50 for GDT values). The DFIRE, DOPE, and ProsaII energy values for most targets exhibited an increase in energy as the structure moved away from native (results not shown). This general trend was less evident when using MM-GB energies of structures that were first minimized with a distance dependent dielectric function. Energy minimization of the ensembles from five targets with the same generalized born implicit solvent model used to score the final structure did lead to the identification of models with higher GDT values though this is much more computationally demanding.

ACKNOWLEDGMENT

This research was supported by funding from NIH-NCRR (RR06009).

Supporting Information Available: Tables of the raw data for Figures 4−7, sample input files representative of the one used to build all three-dimensional models with MODELLER and of how energies were calculated with ProsaII, and a Perl script that takes the output from probA and converts it into input for building comparative models with MODELLER (recently updated to work with version 8). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Chayen, N. E. Turning Protein Crystallization from an Art into a Science. *Curr. Opin. Struct. Biol.* **2004**, *14*, 577−583.
(2) Chandonia, J.-M.; Brenner, S. E. The Impact of Structural Genomics: Expectations and Outcomes. *Science* **2006**, *311*, 347−351.
(3) Baker, D.; Sali, A. Protein Structural Prediction and Structural Genomics. *Science* **2001**, *294*, 93−96.
(4) Ginalski, K.; Grishin, N. V.; Godzik, A.; Rychlewski, L. Practical Lessons from Protein Structure Prediction. *Nucleic Acids Res.* **2005**, *33*, 1874−1891.
(5) Dunbrack, R. L., Jr. Sequence Comparison and Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2006**, *16*, 374−384.
(6) Jaroszewski, L.; Pawlowski, K.; Godzik, A. A Multiple Model Approach: Exploring the Limits of Comparative Modeling. *J. Mol. Model.* **1998**, *4*, 294−309.
(7) Needlemen, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443−53.
(8) Smith, T. F.; Waterman, M. S. Identification of Common Molecular Subsequences. *J. Mol. Biol.* **1981**, *147*, 195−197.
(9) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices. *Adv. Protein Chem.* **2000**, *54*, 73−97.
(10) Vogt, G.; Etzold, T.; Argos, P. An Assessment of Amino Acid Exchange Matrices in Aligning Protein Sequences: The Twilight Zone Revisited. *J. Mol. Biol.* **1995**, *249*, 816−31.
(11) Jaroszewski, L.; Weizhong, L.; Godzik, A. In Search for More Accurate Alignments in the Twilight Zone. *Protein Sci.* **2002**, *11*, 1702−1713.
(12) Vingron, M. Near Optimal Sequence Alignment. *Curr. Opin. Struct. Biol.* **1996**, *6*, 346−352.
(13) Saqi, M. A.; Sternberg, M. J. A Simple Method to Generate Non-trivial Alignments of Protein Sequences. *J. Mol. Biol.* **1991**, *219*, 727−732.
(14) Naor, D.; Brutlag, D. L. On Near-optimal Alignments of Biological Sequences. *J. Comput. Biol.* **1994**, *1*, 349−66.
(15) Mückstein, U.; Holfacker, I. L.; Stadler, P. F. Stochastic Pairwise Alignments. *Bioinformatics* **2002**, *18*, S153−S160.
(16) Miyazawa, S. A Reliable Sequence Alignment Method Based on Probabilities of Residues Correspondences. *Protein Eng.* **1994**, *8*, 999−1009.
(17) Kschischo, M.; Lassig, M. Finite-temperature Sequence Alignment. *Pac. Symp. Biocomputing* **2000**, *1*, 624−635.
(18) Contreras-Moreira, B.; Fitzjohn, P. W.; Bates, P. In Silico Protein Recombination: Enhancing Template and Sequence Alignment Selection for Comparative Protein Modeling. *J. Mol. Biol.* **2003**, *328*, 593−608.
(19) Rai, B. K.; Fiser, A. Multiple Mapping Method: A Novel Approach to the Sequence-to-Structure Alignment Problem in Comparative Protein Structure Modeling. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 644−661.
(20) John, B.; Sali, A. Comparative Protein Structure Modeling by Iterative Alignment, Model Building and Model Assessment. *Nucleic Acids Res.* **2003**, *31*, 3982−3992.
(21) Chivian, D.; Baker, D. Homology Modeling using Parametric Alignment Ensemble Generation with Consensus and Energy-based Model Selection. *Nucleic. Acids Res.* **2006**, *00*(00), e2−18.
(22) Melo, F.; Sanchez, R.; Sali, A. Statistical Potentials for Fold Assessment. *Protein Sci.* **2002**, *11*, 430−448.
(23) Sippl, M. J. Recognition of Three-dimensional Structures of Proteins. *Proteins* **1993**, *17*, 355−362.
(24) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem.* **1998**, *102*, 3586−3616.
(25) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(26) Lee, M. S; Feig, M; Salsbury, F. R., Jr.; Brooks, C. L., III. New Analytic Approximation to the Standard Molecular Volume Definition and its Application to Generalized Born Calculations. *J. Comput. Chem.* **2003**, *24*, 1348−1356.
(27) Tsui, V.; Case, D. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000**, *1222*, 2489−2498.
(28) Fiser, A.; Feig, M.; Brooks, C. L., III; Sali, A. Evolution and Physics in Comparative Protein Structure Modeling. *Acc. Chem. Res.* **2002**, *35*, 413−421.
(29) Bradley, P.; Misura, K. M. S.; Baker, D. Toward High-resolution De Novo Structure Prediction for Small Proteins. *Science* **2005**, *309*, 1868−1871.
(30) Zemla, A. LGA: A Method for Finding 3D Similarities in Protein Structures. *Nucleic Acids Res.* **2003**, *31*, 3370−3374.

(31) Kinch, L. N.; Qi, Y.; Hubbard, T. J. P.; Grishin, N. V. CASP5 Target Classification. *Proteins* **2003**, *53*, Suppl. 6, 340−351.
(32) Tress, M.; Ezkurdia, I.; Grana, O.; Lopez, G.; Valencia, A. Assessment of Predictions Submitted for the CASP6 Comparative Modeling Category. *Proteins* **2005**, *61*, Suppl. 7, 27−45.
(33) Hunt, L. T.; Barker, W. C.; Schwartz, R. M.; Orcutt, B. C.; Young, C. L. In *Atlas of Protein Sequence and Structure*; Dayhoff, M. O., Ed.; National Biomedical Research Foundation: Washington, DC, 1978; Vol. 5, Suppl. 3, pp 345−352.
(34) Gonnet, G. H.; Cohen, M. A.; Benner, S. A. Exhaustive Matching of the Entire Protein Sequence Database. *Science* **1992**, *256*, 1443−1445.
(35) Benner, S. A.; Cohen, M. A.; Gonnet, G. H. Amino Acid Substitution During Functionally Constrained Divergent Evolution of Protein Sequences. *Protein. Eng.* **1994**, *7*, 1323−32.
(36) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.
(37) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics Modell.* **1996**, *14*, 33−38.
(38) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(39) Zhou, H.; Zhou, Y. Distance-scaled, Finite Ideal-gas Reference State Improves Structure-derived Potentials of Mean Force for Structure Selection and Stability Prediction. *Protein Sci.* **2002**, *11*, 2714−2726.
(40) Zhang, C.; Liu, S.; Zhou, H.; Zhou, Y. The Dependence of All-atom Statistical Potentials on Structural Training Database. *Biophys. J.* **2004**, *86*, 3349−3358.
(41) Shen, M. Y.; Sali, A. Statistical Potential for the Assessment and Prediction of Protein Structures. *Protein Sci.* **2006**, *15*, 2507−2524.
(42) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(43) Feig, M; Karanicolas, J.; Brooks, C. L.; 3rd. MMTSB Tool Set: Enhanced Sampling and Multiscale Modeling Methods for Applications in Structural Biology. *J. Mol. Graphics Modell.* **2004**, *22*, 377−95.
(44) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−47.
(45) Marti-Renom, M. A.; Stuart, A.; Fiser, A;, Sanchez, R.; Melo, F.; Sali, A. Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.
(46) Offman, M. N.; Fitzjohn, P. W.; Bates, P. A. Developing a move-set for protein model refinement. *Bioinformatics* **2006**, *22*, 1838−1845.
(47) Chothia, C.; Lesk, A. M. The Relation Between the Divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823−826.

CI600485S