PROTEINS: Structure, Function, and Bioinformatics 63:644–661 (2006)

Multiple Mapping Method: A Novel Approach to the Sequence-to-Structure Alignment Problem in Comparative Protein Structure Modeling

Brajesh K. Rai and András Fiser*
Department of Biochemistry and Seaver Center for Bioinformatics, Albert Einstein College of Medicine, Bronx, New York

ABSTRACT **A major bottleneck in comparative protein structure modeling is the quality of input alignment between the target sequence and the template structure. A number of alignment methods are available, but none of these techniques produce consistently good solutions for all cases. Alignments produced by alternative methods may be superior in certain segments but inferior in others when compared to each other; therefore, an accurate solution often requires an optimal combination of them. To address this problem, we have developed a new approach, Multiple Mapping Method (MMM). The algorithm first identifies the alternatively aligned regions from a set of input alignments. These alternatively aligned segments are scored using a composite scoring function, which determines their fitness within the structural environment of the template. The best scoring regions from a set of alternative segments are combined with the core part of the alignments to produce the final MMM alignment. The algorithm was tested on a dataset of 1400 protein pairs using 11 combinations of two to four alignment methods. In all cases MMM showed statistically significant improvement by reducing alignment errors in the range of 3 to 17%. MMM also compared favorably over two alignment meta-servers. The algorithm is computationally efficient; therefore, it is a suitable tool for genome scale modeling studies. Proteins 2006;63:644–661.**

Key words: sequence-to-structure alignment; multiple mapping method; comparative protein structure modeling; environment dependent scoring function; structural genomics

INTRODUCTION

Comparative or homology protein structure modeling generates a three-dimensional full-atom model for a protein sequence (target) using experimentally determined structures of one or more proteins (templates).[1] Comparative modeling typically involves the following steps: (1) one or more proteins with experimentally determined structures are identified to serve as template for modeling the target sequence. (2) A sequence to structure alignment between the target and template is calculated to determine the residue equivalencies between the two. (3) Models are built by optimizing the positions of target residues according to template dictated spatial restraints and other constraints. (4) The models are evaluated to reveal incorrectly modeled regions.

Each step of the comparative modeling procedure is important and critical in building accurate models. For example, an error in template identification (such as selecting a template that has a different fold than the target) would result in models that show little relevance to the real structure. Similarly, alignment errors would map target residues to incorrect positions in the template structure, leading to inaccurate models. The errors introduced in these first two steps cannot be corrected in the subsequent steps of the modeling procedure. For instance, an error in template selection cannot be corrected by a good alignment, and currently there is no efficient optimization approach available that could recover a good model from serious alignment errors. A misalignment by only one residue position could result in an error of approximately 4 Å in the model.[2]

Fold-recognition algorithms[3–9] and profile-alignment methods[10–14] allow efficient recognition of remotely related sequences by intruding into the distant homology regions, commonly known as the "twilight zone,"[15] of protein sequence similarity. These methods "outperform" the needs of comparative modeling in a sense that they are able to locate remotely related template-target sequence pairs, that are sometimes identified only by a few short conserved segments, and for which no reliable comparative model can be built. To obtain an all-atom comparative model for the target sequence one needs to establish a correct alignment along the entire length of the target sequence, not just for a few conserved segments. Alignments in the low sequence identity region are, in general, not reliable.[15,16] It has been estimated that, on average, ~20% of positions are misaligned if the overall sequence identity is around 30%.[17] The alignments produced by

Grant sponsor: the Howard Hughes Medical Institute. Grant sponsor: the Seaver Foundation. Grant sponsor: NIH; Grant number: GM62519-04.

*Correspondence to: András Fiser, Department of Biochemistry and Seaver Center for Bioinformatics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461. E-mail: fiser@fiserlab.org

Received 1 July 2005; Revised 30 September 2005; Accepted 5 October 2005

Published online 25 January 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20835

different methods or by different combinations of substitution matrices and gap penalties show large variations, which in general, increase with decreasing sequence similarity. Various studies have compared the performance of commonly used alignment methods, and found that no single method or choice of scoring parameters always produced the best results.[18] The accuracy of alignments in the "twilight zone" is a serious limitation in comparative modeling because most of the structurally related proteins fall into this region.[15,17] It has been projected that a realistic 7% improvement in accuracy for alignments that fall in the "twilight zone" over state-of-the art methods would result in an 11% average improvement in model quality and an estimated ~50% increase in the number of reliable comparative models,[19] with a profound impact on the ongoing Protein Structure Initiative.[20,21]

Several recent studies have used sets of alternative (or suboptimal) alignments to improve the quality of input alignment for comparative modeling.[22] Two approaches have been explored: (1) to identify consensus, reliable regions from a set of alignments,[18,22] and (2) to use a genetic algorithm to correct alignment errors by recombining target models.[19,23]

The motivation for the first approach comes from a general observation that alignments of difficult cases by different methods or parameter sets often agree over certain regions of the alignment, and disagree over others. It has been shown that identically aligned stretches in a large set of suboptimal alignments generally agreed with the "gold standard" alignments derived from structural superpositions.[24–26] Therefore, a general strategy for improving the accuracy of the structure prediction is to include only the reliable regions of alignments, while discarding the segments where alternative alignments differ. However, these improvements come at the expense of an incomplete model. Furthermore, the less reliable, alternatively aligned, and consequently not modeled parts frequently correspond to unique sequence signals in these proteins, which often carry the functional fingerprint of the molecules.[27]

Two recent studies focused on the second approach, and they both used a genetic algorithm to simulate artificial genetic selection on a population of target models built from a set of alignments.[19,23] This approach tries to establish a correct alignment from an initial set of random alignments by iterating through a genetic algorithm protocol guided by a fitness function. A major limitation of this method, as pointed out by both studies, is that it requires large computing time, which limits its practical applicability. However, a more fundamental limitation of this approach lies in the fact that during the course of the computation the genetic algorithm samples large numbers of irrelevant and biologically meaningless input alignments, for each of which a model needs to be built and evaluated by a fitness function. In principle, with an ideal scoring function in hand, this would not be a problem, as incorrect models would be filtered out and only the superior conformations would be kept after each iteration. However, scoring functions such as residue–residue contact potentials are not sensitive enough to discriminate among a large number of models that may differ from each other only marginally, as indicated by the fact that the currently available methods rarely converge to the optimal solution.

In this work, we introduce a new method that seeks to minimize alignment errors by selecting and optimally combining differently aligned fragments from a set of alternative input alignments. This selection is guided by a scoring function that determines the preference of each alternatively aligned fragment of the target sequence in the structural environment of the template. The advantage of the method is that in the sampling step it considers a rather small number of biologically relevant alternatives, that is, the best scoring solutions of any given method and not all mathematically possible alignment solutions. Most of the suboptimal alignments correspond to biologically impossible scenario; therefore, they would burden the scoring function without providing relevant alternatives solutions. Furthermore, the current approach optimizes only those parts of the alignment that are differently aligned by the input methods. Second, the method takes advantage of structural information by directly mapping the alternative solutions on the template; therefore, no time-consuming model building is required. These steps ensure that the scoring function needs to explore only a reasonably small number of relevant choices, and it can do it efficiently, without numerous model-building exercises. Finally, we present an extensive exploration of scoring function terms that can efficiently discriminate among alternative alignment choices. We demonstrate on 11 combinations of a benchmark dataset of 1400 structurally aligned protein pairs that this quick approach efficiently builds alignments that are systematically and statistically significantly more accurate than the accuracy of any input alignments. In another comparison the MMM method improves over two alignment meta-servers tested.

RESULTS

We first describe a carefully designed dataset containing 6635 structurally related pairs of proteins encompassing a wide range of sequence identity relationships that present a challenging task for sequence alignment methods. This dataset of alignments was used to develop and test our new approach, the Multiple Mapping Method (MMM). Next, we illustrate the importance of considering alternative alignments for a given protein pair, to improve accuracy of comparative models. The analysis of comparative models of the target sequences of our dataset, built on various alternative alignments, shows that no single alignment method consistently produces superior models. Next, we illustrate that using root-mean-square difference (RMSD) between the experimental structure and the comparative model as an ideal scoring function, it is possible to combine the best variable regions to produce an alignment, which is systematically more accurate than the average of the input alignments. Once we established the theoretical upper limit for this approach, we present the

actual improvements in optimally combining alternate alignments with MMM, using our newly developed template environment-dependent scoring function. Finally, a detailed performance analysis of the method is provided (1) on a large set of data using 11 combinations of four different alignment inputs on 1400 test cases; (2) compared to two alignment meta serves; and (3) on two specific examples using various input alignment scenarios.

Benchmark Dataset of Structurally Related Protein Pairs

As described in the Materials and Methods section, our benchmark dataset was constructed using protein chains extracted from the Protein Data Bank (PDB).[28] The final dataset iscomprised of 6635 pairs of protein chains with high structural similarity and overlap. Only those structurally related protein pairs were retained for benchmarking, where the RMSD differences between the models built from alternative alignments were at least 0.5 Å. For parameterization and initial testing two methods were used in two different scenarios. In the first scenario CLUSTALW[10] and Align2D[17] were used with the default parameters to produce alternative alignments of a given protein pair. In the second scenario, the same alignment protocol and substitution table were used from CLUSTALW, but with default and modified gap penalty functions. CLUSTALW is one of the most frequently used and publicly available alignment methods, whereas Align2D is the fundamental alignment method in the MODELLER package,[29] which is used widely for comparative modeling. One thousand six hundred forty pairs of proteins were identified from the set of 6635 structurally related protein pairs whose sequences were alternatively aligned by $\text{CLUSTALW}_{\text{def}}$ and Align2D, and the difference between the RMSDs of the corresponding comparative models to the experimental structure, built using the two alternative alignments, was at least 0.5 Å ($\text{CLUSTALW}_{\text{def}}$-Align2D dataset). On the other hand, 750 template-target pairs were intrinsically defined by the fact that the difference in RMSDs of the models to experimental structure, built from the two alternative CLUSTALW alignments, was at least 0.5 Å ($\text{CLUSTALW}_{\text{def}}$-$\text{CLUSTALW}_{\text{mod}}$ dataset). The percent of sequence identity between template and target pairs spans a wide range. It includes examples from the "trivial region" up to 80% pairwise sequence identity, where accurate sequence alignment is thought to be easy to obtain. However, it is not surprising that most of the test cases emerged from the "twilight" and "midnight" zones,[15] all the way down to 6% pairwise sequence identity level, which traditionally represent a major challenge for alignment methods. A notable feature of the two datasets is that they contain a significant proportion of pairs from globin and immunoglobulin superfamilies, whose folds are strongly preserved but their sequences show a large divergence.

The RMSDs of most of the target models from both datasets are in the range of 1.0 to 3.0 Å, with the mean of the distribution at 2.0 Å. This illustrates that even at very low sequence identity levels, an accurate "gold standard" input alignment derived from structural superposition can produce models with high accuracy.

We compared the relative accuracies of alignments obtained from $\text{CLUSTALW}_{\text{def}}$ and Align2D, and those from $\text{CLUSTALW}_{\text{def}}$ and $\text{CLUSTALW}_{\text{mod}}$ through the accuracy of the comparative models that were built using these alternative alignments. A model built from one alignment is considered better (or worse) than the model built from a different alignment if the RMSD differences of their corresponding models to the experimental structure is at least 0.5 Å. The relative ranking of the models are shown in Figure 1. A general observation is that, in approximately three out of four cases the $\text{CLUSTALW}_{\text{def}}$ models are of better quality than the models based on the Align2D alignments [Fig. 1(a)]. In contrast, the models built using $\text{CLUSTALW}_{\text{def}}$ and $\text{CLUSTALW}_{\text{mod}}$ have a roughly similar performance over the entire sequence identity spectrum [Fig. 1(b)]. This illustrates that instead of relying on a single most optimal solution from one method, alternative alignments from different methods, including suboptimal alignments, and those calculated using different parameters (such as mutation matrix and gap penalty) should be considered competitively to build better models. These alternative alignments provided superior results (at least a 0.5 Å RMSD improvements in model quality) in 30–50% of the cases in our experiment (Fig 1).

Alignment Discrepancy and Variable Regions

To accurately monitor the differences among alternative alignments, besides comparing the resulting comparative models, we also compared the alignments obtained from the $\text{CLUSTALW}_{\text{def}}$-Align2D and $\text{CLUSTALW}_{\text{def}}$-$\text{CLUSTALW}_{\text{mod}}$ datasets. We define discrepancy between two alignments as the percentage of differently aligned residues in the shorter of the two alignments. As expected, alignment discrepancies exponentially decrease with increasing sequence identity (not shown).

The differences among the alternative alignments are also described in terms of the ratio of the length of the variable region to the length of the consensus alignment. As shown in Figure 2(a) and (b), the percent length of the variable region rapidly runs up with decreasing template-target sequence identity, indicating the difficulty in obtaining an accurate alignment. Depending on the differences between the input alignments, one or more variable regions, separated by nonvariable parts, may occur in the consensus alignment. Because the performance of the MMM critically depends on its ability to correctly select and combine the variable regions, there are three cases to consider: (1) the consensus alignment contains only one variable region; (2) the consensus alignment contains more than one variable regions, and the better aligned variable regions all come from the same input alignment; (3) the consensus alignment contains more than one variable regions, and the better aligned variable regions come from a combination of both input alignments. In the first two cases, the MMM, in an ideal case, would reproduce

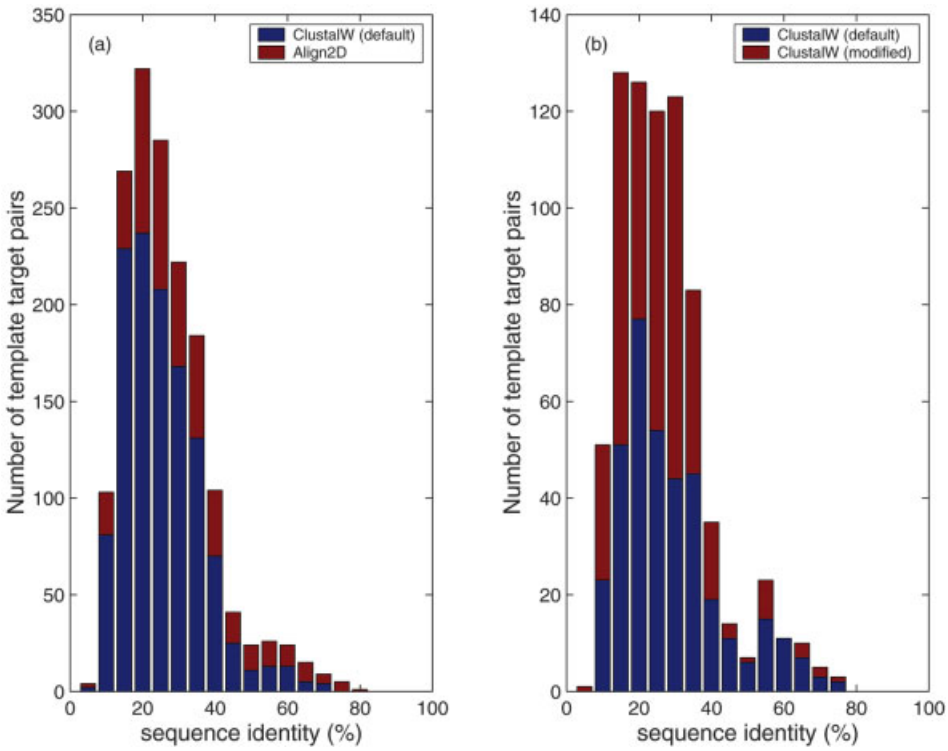

Fig. 1. Ranking of the alternative models from the (**a**) CLUSTALW$_{def}$-Align2D and (**b**) CLUSTALW$_{def}$-CLUSTALW$_{mod}$ datasets. Total number of cases when one alignment produces better model than the other is indicated as a function of the percentage sequence identity from the STAMP structural alignments. Model A is defined as better than model B if the RMSD of the superposition of comparative model A to the experimental structure is at least 0.5 Å less than the corresponding RMSD for model B.

the best of the two input alignments. However, in a scenario when both input alignment have variable regions that are more accurate than those in the other alignment, a correct selection and combination of the variable regions would produce a unique MMM alignment with better accuracy than either of the two input alignments. A histogram of the number of cases, when each of the three cases of variable regions, occur is shown in Figure 2(c) and (d), for the CLUSTALW$_{def}$-Align2D and CLUSTALW$_{def}$-CLUSTALW$_{mod}$ datasets, respectively. Although above 40% sequence identity level scenario 1 is dominant below this level, all three cases occur with roughly comparable chances with scenario 3 being the most frequent one.

Alignment Recombination

The aim of the MMM approach is to build an accurate alignment for the entire length of a target sequence from a set of alternative inputs. Although it is straightforward to include the reliable regions of the consensus alignment in the final alignment, the less reliable regions, which are different among the alternative alignments, should be chosen, and optimally combined, based on a scoring scheme, which should minimize the modeling errors associated with these variable segments. The details of the scoring scheme and how the alternative variable regions are selected are discussed in the next few sections.

Establishing the Theoretical Limit for MMM Using an Ideal Scoring Function

First, we study the performance of MMM using an ideal scoring function to explore the theoretical limits of this approach. In this case, because the experimental structures of the targets are known, RMSD of main-chain atoms between the models and the experimental structure can be used to score the alternatively aligned and modeled target segments. Variable regions with smaller RMSD can then be selected to construct the most optimal alignment between the target sequence and the template structure. This optimal combination of variable regions would not only select the superior alignment in general, but if it combines parts from different alignments, it will produce models with smaller overall RMSDs compared to the models built from any of the input alignments. We test the performance of the ideal scoring function by selecting the best variable regions from the consensus of the input alignment set.

For easier reference we balanced out our benchmark dataset so that it contained equal number of cases where one input alignment method performs better than the other. Therefore, in subsequent benchmarks the expected accuracy of reference input alignments can be associated with the average accuracy of the corresponding comparative models compared to the experimentally determined structure. In the CLUSTALW$_{def}$-Align2D set, there are

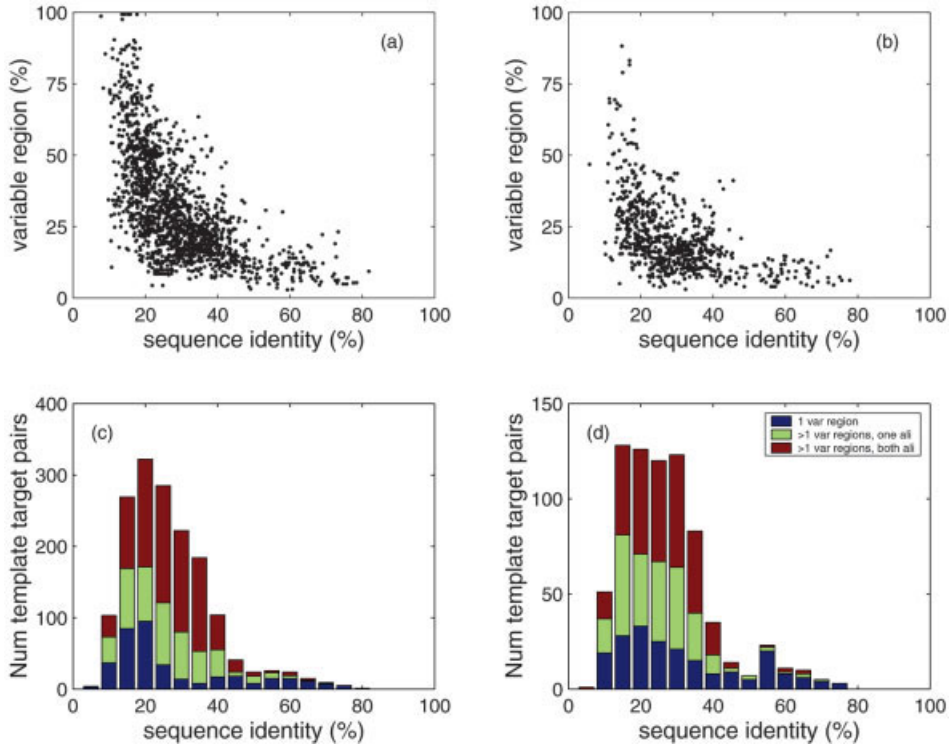

Figure 2.

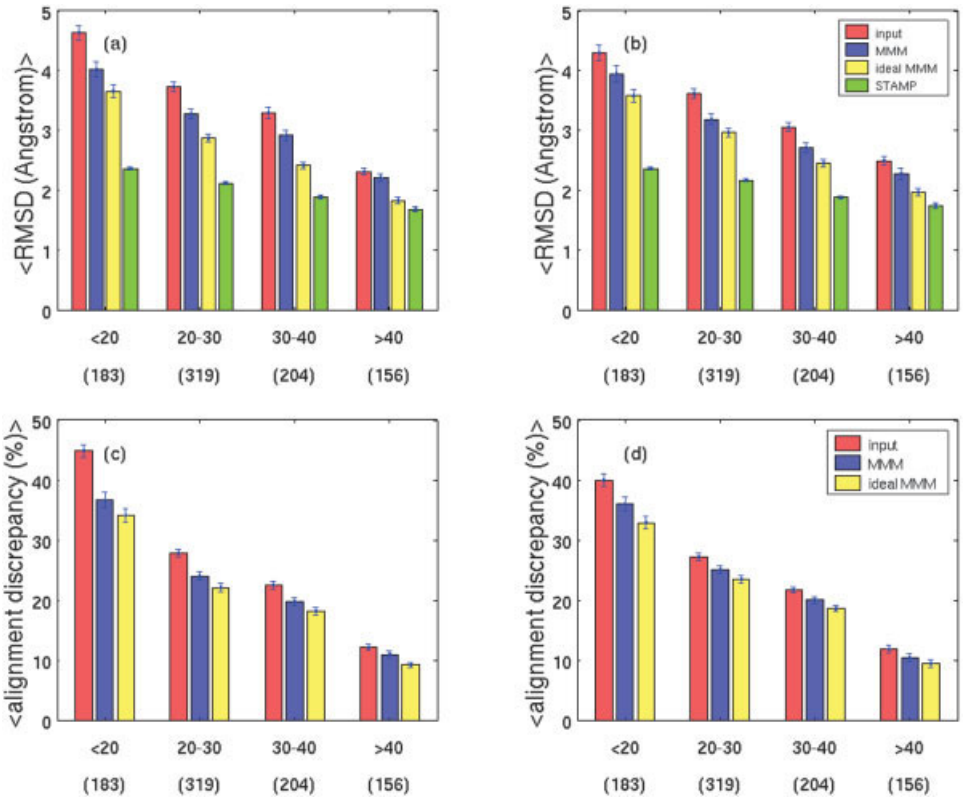

Figure 3.

431 cases where ALIGN2D alignment produces significantly better model than $CLUSTALW_{def}$; thus, we randomly selected and added 431 cases where $CLUSTALW_{def}$ was superior resulting in 862 template-target pairs (out of 1640). In a similar manner, 720 cases were selected (out of 747) from the $CLUSTALW_{def}$-$CLUSTALW_{mod}$ dataset.

Next, we built MMM alignments by combining the more accurate parts of each input alignment. Subsequently, comparative models of each input and the MMM alignment are compared to the experimental structure of the target. The performance of this ideal MMM approach, on $CLUSTALW_{def}$-Align2D and $CLUSTALW_{def}$-$CLUSTALW_{mod}$ datasets, is shown in Figure 3 using RMSD and percent of alignment discrepancies as quality measures. The average RMSDs of the MMM models, along with the mean RMSDs of the average input and the models based on the STAMP structural alignments are shown in Figure 3(a) and (b) as a function of the sequence identities. These results (Fig. 3 and Table I) show that a combination of alternative input alignments using MMM produces models that are significantly more accurate (average improvements of 0.82 and 0.63 Å for $CLUSTALW_{def}$-Align2D and $CLUSTALW_{def}$-$CLUSTALW_{mod}$, respectively) than the average input models. The improvements in the accuracies of the MMM models are even more significant (0.90 and 0.68 Å RMSD) in the low-sequence identity region (below 30%), where differences between alternative alignments are more serious. Similarly, as shown in Figure 3(c) and (d) and Table II, the MMM approach reduces the discrepancy of the alignments compared to the STAMP solution. The average discrepancy between input and STAMP decreases from 27.6 and 26.8% to 21.6 and 22.5% for the $CLUSTALW_{def}$-Align2D and $CLUSTALW_{def}$-$CLUSTALW_{mod}$ datasets, respectively. As in the case of models, the MMM method produces larger improvements (34.0 to 26.5 and 32.7 to 27.5 for $CLUSTALW_{def}$-Align2D and $CLUSTALW_{def}$-$CLUSTALW_{mod}$ datasets, respectively) in the low-sequence identity region of alignments ($<30\%$).

Although the gain in the model accuracies using MMM are significant, these improvements come only about halfway to the maximum possible improvements that is set by the STAMP structural alignments [Fig. 3(a) and (b)]. This observation can be attributed to the sampling of the alignment search space. With a better choice of input alignments one could expect to further increase the expected upper limits of the method.

Scoring Scheme to Select Alternative Variable Regions

The aim of the scoring scheme is to select the best alternatively aligned variable regions based on their compatibility in the structural environment defined by the template protein. We use a composite score, consisting of three rather independent terms, each of which assesses the local structural environment of the template protein. The three different scoring terms are: (1) environment specific substitution matrices from FUGUE;[7] (2) A 3D-1D substitution matrix, H3P2, that scores the matches of predicted secondary structure of the target sequence to the observed secondary structures and accessibility types of the template residues;[30] (3) statistically derived residue–residue contact energy term (MJ), which determines the compatibility of alternative variable segments in the protein environment.[31]

Model Quality Improvements Using the MMM for Combining Input Alignments

The ideal scoring function sets a theoretical upper limit for improvement in our approach. We now assess the performance of our MMM algorithm on the $CLUSTALW_{def}$-Align2D and $CLUSTALW_{def}$-$CLUSTALW_{mod}$ datasets using a composite scoring function. As quality measures we use RMSD of the corresponding comparative models and experimental structures, and the percent of alignment discrepancy. We compare the performance of the method using our composite scoring function. Figure 3(a) and (b) shows a histogram of the average RMSDs of MMM models as function of percent sequence identity for the alignment test cases. The quality (with RMSD as performance measure) of MMM models, built using the environment specific composite scoring function, are consistently better than the average input models, with mean improvements of 0.41 and 0.35 Å for the models from the $CLUSTALW_{def}$-Align2D and $CLUSTALW_{def}$-$CLUSTALW_{mod}$ datasets, respectively. The gain in the MMM performance is even more pronounced (improvements of 0.51 and 0.40 Å) in the low-sequence identity region ($<30\%$).

Fig. 2. General features of the variable regions in the input alignments from (**a**) $CLUSTALW_{def}$-Align2D and (**b**) $CLUSTALW_{def}$-$CLUSTALW_{mod}$ datasets. The number of template-target pairs consisting of only one variable segment (blue color) and multiple variable segments are shown. The pairs with multiple variable regions are further divided into two groups, according to whether the best segments come from the same alternative alignment (green color) or a combination of the two (red color). Ratio of the total length of the variable region(s) with the length of the shorter of the two alignments is shown as a function of the percentage sequence identity.

Fig. 3. MMM performance on the $CLUSTALW_{def}$-Align2D (left panels) and $CLUSTALW_{def}$-$CLUSTALW_{mod}$ (right panels) benchmark datasets. The MMM performance using the composite scoring function is shown as a function of the pairwise sequence identity. The numbers on the second row below the x-axes correspond to the number of template-target pairs for the corresponding bin. (**a**) The performance measure RMSD is calculated from the superposition of the equivalent main-chain atoms between the model and experimental structure. For each bin, the average RMSD of the MMM models are compared to the average of input models, to MMM models using ideal (RMSD) scoring function, and to the models based on the STAMP structural alignments. (**c**) A comparison of the average discrepancy between the test and STAMP structural alignments. The average alignment discrepancy (%) of the MMM alignments are compared to the average input alignments as well as to the ideal MMM alignments. The results presented here are the means $\pm$ standard error of the performance data corresponding to each bin.

TABLE I. Accuracies of the Average Input Models, Models Based on the STAMP Alignments, Ideal MMM Models, and the MMM Models Using a Composite Scoring Function

	CW_{def}-Align2D set	CW_{def}-CW_{mod} set	CW_{def}-Align2D set	CW_{def}-CW_{mod} set
Sequence identity	0–100%	0–100%	<30%	<30%
Number of cases	862	720	502	416
Average input	3.58	3.47	4.06	3.90
STAMP	2.04	2.08	2.22	2.26
Ideal MMM[a]	2.75	2.83	3.16	3.22
MMM[b]	3.16	3.10	3.55	3.50
input-STAMP	1.53	1.38	1.84	1.64
input-Ideal MMM	0.82	0.63	0.90	0.68
input-MMM	0.41	0.35	0.51	0.40

The numbers in the table represent the average RMSD (in Angstroms) between models and corresponding experimental structures. CW_{def} and CW_{mod} refer to the $CLUSTALW_{def}$ and $CLUSTALW_{mod}$ datasets, respectively. The last three rows show the average RMSD improvements compared to the average input models.

[a]Ideal MMM alignments are calculated using RMSD as scoring function.

[b]A composite scoring function combining FUGUE, H3P2, and MJ scores, is used to obtain the MMM alignment.

TABLE II. Accuracies of the Average Input, Ideal MMM, and MMM Alignments Using Percent Alignment Discrepancy from the STAMP Structural Alignments as Performance Measure

	CW_{def}-Align2D set	CW_{def}-CW_{mod} set	CW_{def}-Align2D set	CW_{def}-CW_{mod} set
Sequence identity	0–100%	0–100%	<30%	<30%
Number of cases	862	720	502	416
Average input	27.6	26.8	34.0	32.7
Ideal MMM	21.6	22.5	26.5	27.5
MMM	23.5	24.4	28.6	29.8

Using the percent discrepancy to STAMP structural alignment as the performance measure also shows [Fig. 3(c,d)] that the MMM alignments improved from 27.6 (input) to 23.5% (MMM) for the $CLUSTALW_{def}$-Align2D dataset, and 26.8 (input) to 24.4% for the $CLUSTALW_{def}$-$CLUSTALW_{mod}$ dataset (Table II). Again, the gain in the MMM performance is most significant in the low sequence identity region (below 30%). Although the improvements in the alignment accuracy of the entire $CLUSTALW_{def}$-Align2D and $CLUSTALW_{def}$-$CLUSTALW_{mod}$ datasets are 4.1 and 2.4, respectively, the corresponding improvements for the cases <30% pairwise sequence identity are 5.4 and 2.9.

An additional comparison of the MMM models to the average input models, using a three-state classification scheme (good, bad, and ambiguous) were considered. An MMM model is classified as: good, if $\Delta RMSD_{MMM-input} \leq -\epsilon$; bad, if $\Delta RMSD_{MMM-input} \geq \epsilon$; and ambiguous, if $|\Delta RMSD_{MMM-input}| < \epsilon$; where, ϵ is a variable number for each template-target pair, that indicates the observed RMSD difference between the worst and the best models from the 10 models that were built using the same STAMP structural alignment.

For both datasets, in approximately 68% of the cases, the MMM models are superior to the average of the input models. Meanwhile, it happens only 2% of the cases that MMM produces such an alignment that is inferior to any of the input alignments.

MMM Performance as a Function of Sampling Space

Using the scoring function parameters derived from our benchmark $CLUSTALW_{def}$-Align2D dataset, we assessed the performance of this method as a function of sampling space. We extended the alignment sampling space by using an additional method, MUSCLE,[32] and by exploring all combinations of all four types of input alignments: $CLUSTALW_{def}$, $CLUSTALW_{mod}$, Align2D, and MUSCLE.[32] As shown in Figure 4, we considered all 11 different combinations of these alignments (all six, four, and one combinations of possible pairs, triplets, and the full quartet, respectively) as inputs to the MMM, and compared the performance of the output MMM alignment to each of the input alignments. The percentage alignment discrepancy of the test alignment to the STAMP structural alignment was used as the performance measure. In each case, the performance data were calculated on the same 1400 alignment pairs. Because the MMM attempts minimize alignment errors by selecting the better aligned variable regions from the input alignments, it was neces-

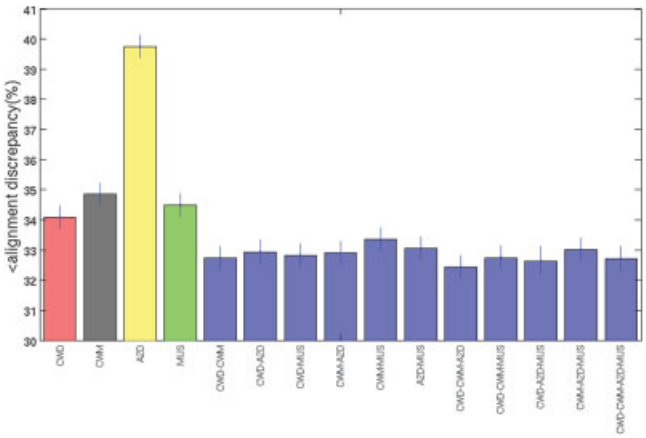

Fig. 4. MMM performance using a combination of doublet, triplet, and quartet alignments from CLUSTALW$_{def}$ (CWD), CLUSTALW$_{mod}$ (CWM), Align2D (A2D), and MUSCLE (MUS) as input to the method. The percent alignment discrepancy between the test alignment and the corresponding STAMP structural alignment are used as performance measure. For each input alignment combination, the average (±standard error) percent discrepancy of the input and MMM alignments to the STAMP structural alignments are shown. (left to right) The six doublet combination correspond to 1400 alignments from: CLUSTALW$_{def}$-CLUSTALW$_{mod}$; CLUSTALW$_{def}$-Align2D; CLUSTALW$_{def}$-MUSCLE; CLUSTALW$_{mod}$-Align2D; CLUSTALW$_{mod}$-MUSCLE; and Align2D-MUSCLE. The four triplet combination correspond to alignments from: CLUSTALW$_{def}$-CLUSTALW$_{mod}$-Align2D; CLUSTALW$_{def}$-CLUSTALW$_{mod}$-MUSCLE; CLUSTALW$_{def}$-Align2D-MUSCLE, CLUSTALW$_{mod}$-Align2D-MUSCLE. The quartet combination corresponds to the four alignments from CLUSTALW$_{def}$-CLUSTALW$_{mod}$-Align2D-MUSCLE.

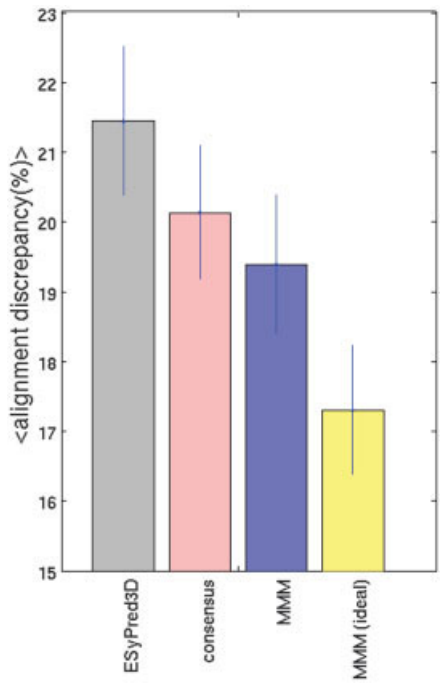

Fig. 5. MMM performance on a dataset of 70 template-target protein pairs, using two sets of alternative input alignments from ESyPred3D and Consensus servers. Percentage alignment discrepancy of alignments to STAMP structural alignment is used as the performance measure. The mean alignment discrepancies of the input alignments are compared to the average alignment discrepancies of MMM based alignments.

sary to consider only those cases where input alignments considerably differed from each other. Therefore, by using a filtering criterion, we selected only those cases, from the dataset of 6635 template-target protein pairs, for which the percent alignment discrepancy between each input alignments pair was greater than 10%. This filtering cutoff at the alignment level is consistent with those imposed in previous section at the input models (i.e., >0.5 Å RMSD).

For each of the 11 sets of performance data shown in Figure 4, the MMM alignments have smaller mean discrepancy to the gold standard STAMP structural alignments compared to the mean discrepancy of any of the input alignments to the STAMP. Furthermore, as shown from the error bars corresponding to each datasets, the gain in the MMM performance are statistically significant. The percentage gain the alignment accuracy spans a wide range, from 3% (MUSCLE to MMM using CLUSTALW$_{mod}$-MUSCLE as input) to 17% (Align2D to MMM using CLUSTALW$_{def}$-Align2D as input). Because the same 1400 test cases were used in all cases in Figure 4, the performance of MMM is directly comparable among each of the 11 scenarios. It is notable that the most accurate MMM solution resulted in from the combination of CLUSTALW$_{def}$-CLUSTALW$_{mod}$-Align2D methods. Align2D, in general, is the least accurate out of the four methods tested, that is, it ranks least frequently as the top solution. Nevertheless, when combinations of alignments are needed, Align2D, presumably due to its unique approach, presents such solutions, which are frequently unique and sometimes superior to all others. The results illustrate that MMM is efficient in selecting these cases. Another notable result in Figure 4 that MMM is not simply selecting the superiorly performing input alignment from a set of inputs, but by optimally "splicing" the more accurate parts, together it delivers a unique solution that is not present in any inputs, and this unique solution is, on average, superior to the best available input.

MMM Performance Using Inputs from Methods That Combine Alignments

A number of methods and meta-servers[18,33–36] that combine alignments and/or models from diverse sources, have been developed. A direct comparison of the MMM performance to these existing programs is not meaningful here because in this work we only introduce the MMM algorithm and present an optimized environment specific scoring function; the sampling component of this method, which would involve derivation of an optimal combination of alignments by an exhaustive exploration of the available alignment programs is a subject of future study. MMM has been shown to improve over input alignments; hence, we present an assessment of the MMM performance vis-à-vis two other methods that combine alignments by using the output from these programs as the sampling space for our method. Two sets of alignments for the same template-target pairs were obtained from Consensus[18,36] and ESyPred3D,[33] both of which are publicly available online servers that produce alignments based on a combination of various alignment techniques. These two servers were

chosen because they satisfied the following two constraints pertaining to this assessment study: (1) it was possible to specify a template structure for a given target sequence; and (2) the servers returned an alignment between target sequence and template (not just the model for target protein).

From a set of 1400 protein pairs selected for the performance assessment in the previous section, we randomly selected 200 template-target pairs. For each of these 200 protein pairs, we submitted the target sequence along with the PDB code for the corresponding template structure to the ESyPred3D and Consensus servers. Out of these 200 pairs the ESyPred3D returned alignments only for 87 pairs and failed for others due to low sequence identity; in contrast, the Consensus server returned alignment for 176 pairs. For many template-target pairs, ESyPred3D and Consensus servers returned partial alignments with mean coverage of 99 and 96%, respectively. The partial alignments were converted to full alignments by adding missing residues and aligning those positions to gaps.

Figure 5 shows the performance of MMM using a dataset of 70 protein pairs and two sets of alternative input alignments from ESyPred3D and Consensus. The MMM based alignments using the composite scoring function have smaller mean alignment discrepancy to STAMP structural alignments compared to the input alignments from ESyPred3D and Consensus (19 vs. 20% and 21%). The theoretical limit on the possible improvements is defined by the optimal combination of the alternative alignments sampled by the two servers. We calculated that an optimal combination of inputs would results in a 17% alignment discrepancy.

Modeling a Target Sequence with Low Sequence Identity

In the previous sections, we benchmarked the MMM technique on large datasets of template-target pairs. Here, we illustrate the application of this method in detail on two template-target pairs, 1a6m–1spgB (chain B) and 1ecn–1myt (the four-letter identifiers represent the PDB code of these proteins). These examples were chosen as they have low template-target sequence identities (1a6m–1spgB, 21%; 1ecn–1myt, 16%), belonging to the "twilight" (1a6m–1spgB) and "midnight" (1ecn–1myt) zones of alignment accuracy.[15] The alignments obtained in "twilight" zone change drastically with different methods and scoring functions. The "midnight zone" of the alignments corresponds to a sequence identity region where sequence identities between pairs of proteins are comparable to the level expected for unrelated sequences.

Sequence Alignment in the Twilight Zone

The consensus of the input alignments of the 1a6m–1spgB pair, from CLUSTALW and Align2D, reveals two variable regions, labeled as variable region I and variable region II in Figure 6(a). Using the MMM protocol, we first rank the alternative alignments generated by joining all possible combinations of the variable regions to the nonvariable segments of the alignment. The two variable regions of the consensus alignment [Fig. 6(a)] can be combined in four different ways to the core part of the alignment, as shown in Figure 6(b). For each of these alternative alignments, a fitness score is calculated, which is the sum of the scores from each individual positions in the variable regions.

The FUGUE and H3P2 scores are calculated as the sum of substitution scores from each aligned position in the variable regions, and are obtained directly from the appropriate substitution tables, whereas the MJ scores are calculated from a summation of the pairwise contact energies of each variable region residues from the target sequence to its local environment. The residues from the alternative alignments are mapped to different positions in the template structure. For example, the CLUSTALW$_{\text{def}}$ and Align2D align the residue D (the first target residue of the variable region I) to template residues E and D, respectively [Fig. 6(a)]. The positions of these two residue positions are indicated as red in the sequence shown in Figure 6(b), and in the template structure shown in Figure 6(c) and (d). The local environment of these residues (with a 10.0 Å cutoff) is shown in blue.

From the four alternative alignments, A1–A4 [Fig. 6(b)], the MMM selects A3 as the best scoring alignment combination. The MMM alignment consists of variable region I from Align2D and II from CLUSTALW$_{\text{def}}$. The model of the target sequence based on the MMM alignment is shown in Figure 6(e), and its accuracy is compared to the two input models in Table III.

As noted from Table III, a correct selection of the variable region fragments results in a model that has an overall smaller RMSD to the experimental structure (1.79 Å) compared to each of the two input models (2.04 and 2.74 Å). In addition, compared to each of the two input alignments, the MMM alignment has smaller discrepancy to the STAMP alignment than the two input alignments (7.83 vs. 12.42% and 20.39%). This result is further illustrated in Figure 6(e) and (f), which show the experimental structure of the target, as well as the MMM, CLUSTALW$_{\text{def}}$, and Align2D models.

Sequence Alignment in the Midnight Zone

We use MMM to build a model of the target sequence 1myt, which has only 16% sequence identity with its template structure 1ecn. The pairwise alignments from CLUSTALW$_{\text{def}}$ and Align2D were used as the input to the MMM. The consensus alignment of the input shows two variable regions (I and II), occurring at the terminal ends of the sequence and separated by a nonvariable region [Fig. 7(a)]. A superposition of the two alternative target models with experimental structure clearly shows the differences of the variable segments of the two models [Fig. 7(b)]. Although the variable region I from the CLUSTALW$_{\text{def}}$ model has a smaller RMSD to the experimental structure (RMSD 3.9 vs. 6.9 Å), the backbone corresponding to the variable segment II of Align2D has a better fit to the experimental structure (RMSD 2.5 vs. 5.4 Å).

A set of alternative alignments is constructed (by combining variable regions I and II to the core part) and scored

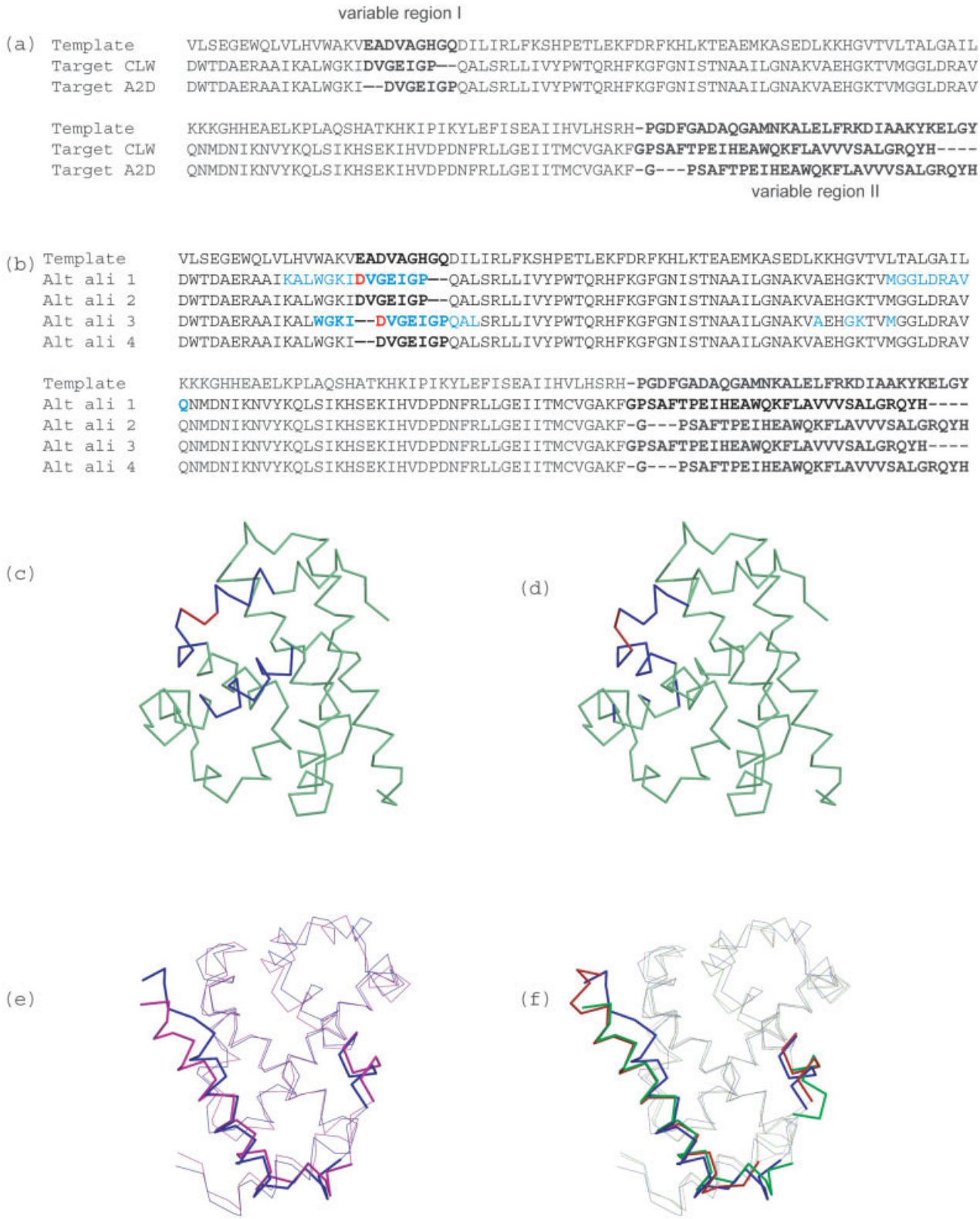

Fig. 6. A detailed illustration of the multiple mapping method using an example from the "Twilight zone" of the alignment (template-target pair 1a6m-1spgB, 21% sequence identity). (**a**) Consensus of the two input alignments from CLUSTALW and Align2D shows two variable regions in bold (I and II). (**b**) The two variable regions are combined to the core part of the alignment in four possible ways. The two alternative mapping of the first variable region residue from the variable region I to the template structure are shown in red for CLUSTALW (**c**) and Align2D (**d**). The local environment of the reference residue, in each case, is represented in blue. (**e**) Superposed structures of the target model, built using the MMM alignment, and the experimental structure. (**f**) The experimental structure superposed with the target models build from the CLUSTALW and Align2D alignment. The backbone corresponding to the variable parts of the consensus alignment are shown in bold, as opposed to a relatively thinner trace for the rest of the protein.

TABLE III. An Example from the "Twilight Zone" of Alignments

	Alignment discrepancy	Full protein RMSD	Variable region (1) RMSD	Variable region (2) RMSD
CLUSTALW	12.4 %	2.0 Å	4.6 Å	2.6 Å
Align2D	20.4 %	2.7 Å	1.1 Å	6.1 Å
MMM	7.8 %	1.8 Å	1.1 Å	2.6 Å
STAMP	—	1.7 Å	—	—

The template and target share only 21% sequence identity according to the STAMP structural alignment. Accuracies for each input, the MMM and STAMP alignment (in RMSD and in percent alignment discrepancy) are shown for 1spg (chain B) alignments and models, using 1a6m as template structure.

using the scoring scheme described previously. The MMM alignment, and the corresponding model, is shown in Figure 7(c) and (d), respectively. MMM approach accurately selects and combines the variable regions that have smaller errors. The MMM model, as shown in Figure 7(d) has smaller RMSD to the experimental structure than any of the two input models (RMSD 2.0 vs. 3.2 Å and 2.4 Å; Table IV). In this particular case, the MMM outperforms the model based on the STAMP structural alignment, showing an improvement (though marginal) of 0.1 Å. Considering the alignment discrepancy as performance measure, the MMM outperforms each of the two input alignments (10.9 vs. 42.1% and 17.1%).

DISCUSSION

The Multiple Mapping Method addresses an important problem in comparative protein structure modeling. From a set of alternative sequence-to-structure alignments with competing quality, which one is more accurate or what parts needs to be combined to obtain the best results? Because none of the existing alignment methods is superior in all scenarios, it is useful to combine several methods instead of relying on just one. The Multiple Mapping Method provides an approach to optimally combine parts of alternative alignments. In this section we discuss the following two main features of the method: (1) alignment search space, and (2) the scoring scheme. A comparison with some previously developed techniques that are similar to this method are also presented. Finally, we comment on possible applications.

Alignment Search Space

MMM generates an alignment by selecting the best scoring combination of the variable region fragments from a discrete combinatorial search space of biologically relevant alternatives, which is defined by the input alignments. Therefore, the selection of input alignments to the MMM is critical to the accuracy of the output of this method. MMM can be applied in those practical cases where at least one alternatively aligned region exists among the input alignments. The method has the capability to select the more accurate variable regions from different input alignments, and not just confined to a single alignment of the input set. In this case, the MMM solution provides a unique alternative, which is better than any of the individual input alignments.

The speed, and at least in part the accuracy, of the MMM approach originates from the fact that the alignment search space is restricted to a small number of biologically relevant alternative variable regions, which differ from each other by the relative positions of gaps in these fragments. For example, the two alternative alignments of the target sequences from variable region I, *DVGEIGP* and DVGEIGP, of Figure 6(a) differ only by the position of gaps but the corresponding RMSD values are 4.6 and 1.1 Å. The task of the MMM is to select one of these two variable fragments based on their compatibility score to the local environment. At the expense of longer computing time, one might consider artificially creating many more variable fragments by placing gaps at all possible locations. All these variable segments, along with those identified by the alternative input alignments, will need to be scored to select the one having best score. In principle, using an ideal scoring function such as RMSD, this exhaustive sampling approach would lead to a better performance of the MMM. However, the generally available scoring functions, such as the ones used in this study, although capture many important features of the sampled systems, are not perfect, and not expected to be sensitive enough to accurately select superior variable segments from very large set of decoys, which are overwhelmed by biologically irrelevant possibilities. We tested the exhaustive sampling strategy on the MMM performance using our composite scoring function. The result (not presented here) was a significant (approximately 10%) drop in the MMM performance compared to the case when only biologically more relevant alternative variable segments, as determined from the input alignments, were used.

Scoring Function

The three components of our scoring scheme provide three rather independent terms for assessing the compatibility of the alternative variable region fragments to the template environment. Although the contribution to the FUGUE and H3P2 terms of the scoring function comes only from the aligned residue pairs of the variable regions, the MJ contact energy scores consider residues that are close in the structure, but may be at distant sequential positions. It is clear from the results presented that this scoring function on average accurately selects correct combinations of variable region fragments, and contributes to improving the alignment accuracy. Nevertheless, a

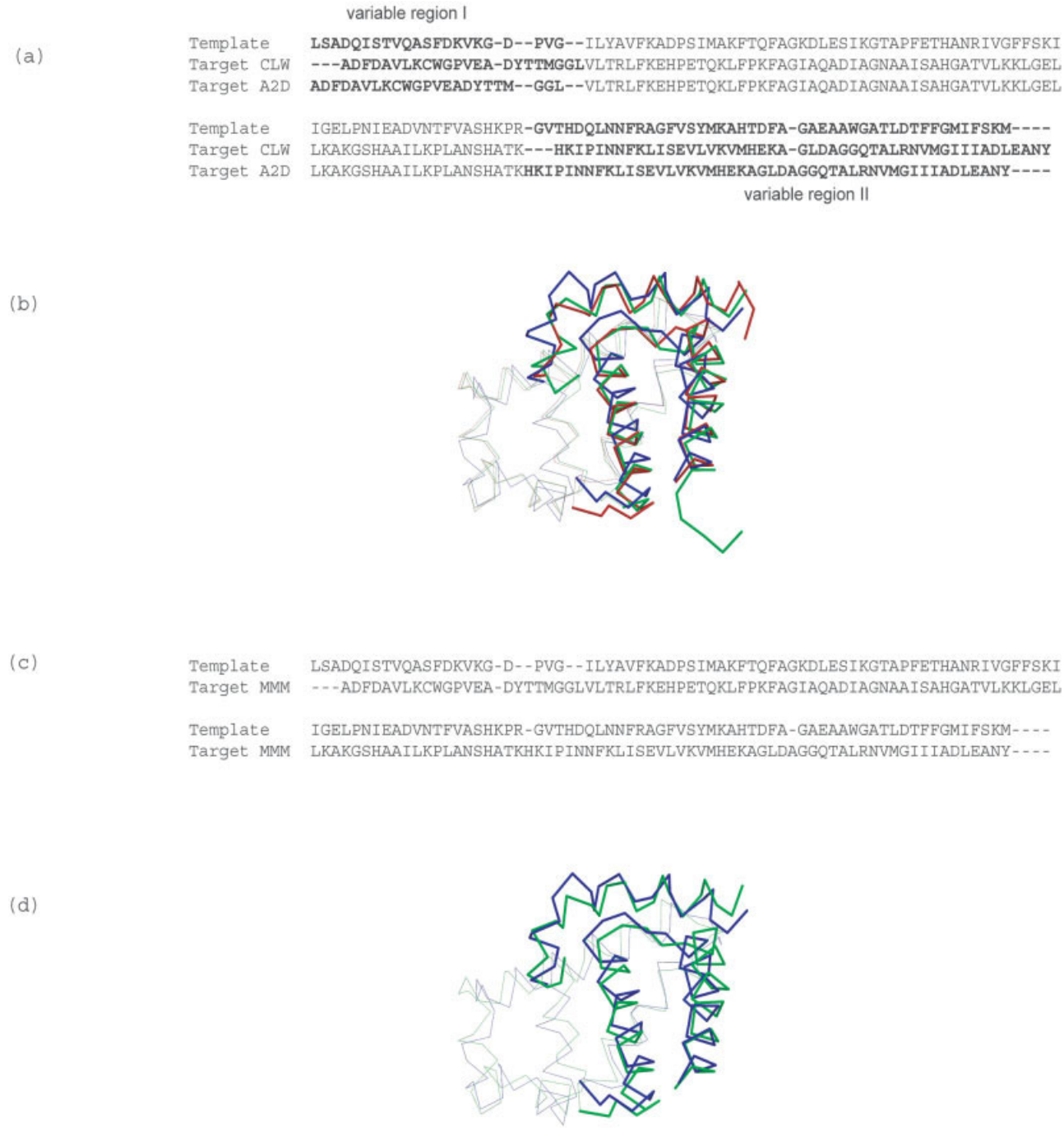

Fig. 7. An example from the "Midnight zone" of the alignment. (**a**) The consensus alignment of template structure 1ecn with target sequence 1myt, having 16% sequence identity. The variable stretches of the consensus alignment are shown in bold. (**b**) The two input models from the CLUSTALW and Align2D alignments are superposed with the experimental structure. To emphasize the differences in the variable parts of the models, the segments corresponding to the variable regions are shown by thicker trace than other parts of the proteins. (**c**) Superposition of the model built from the MMM alignment to the experimental structure.

comparison with the ideal scoring function (RMSD) suggests that further improvements can be made (Fig. 8).

Selecting Alternative Variable Regions: Mapping versus Modeling

Two alternative approaches to calculate fitness score of a variable region segment using MJ residue–residue contact energy score has been described in the Materials and Methods section. The fitness score of the variable region residues can be calculated using (1) the framework of the template structure, or (2) the target model. The results presented here used the former approach by mapping the target residues onto the template structure according to the alignment. This is a very quick approach, but the

TABLE IV. An Example from the "Midnight Zone" of Alignments

	Alignment discrepancy	Full protein RMSD	Variable region (1) RMSD	Variable region (2) RMSD
CLUSTALW	42.1 %	4.2 Å	3.9 Å	5.4 Å
Align2D	17.1 %	2.4 Å	6.9 Å	2.5 Å
MMM	10.9 %	2.0 Å	3.9 Å	2.5 Å
STAMP	—	2.1 Å	—	—

The target and template share only 16% sequence identity according to their STAMP structural alignment. Data types as in Table 3.

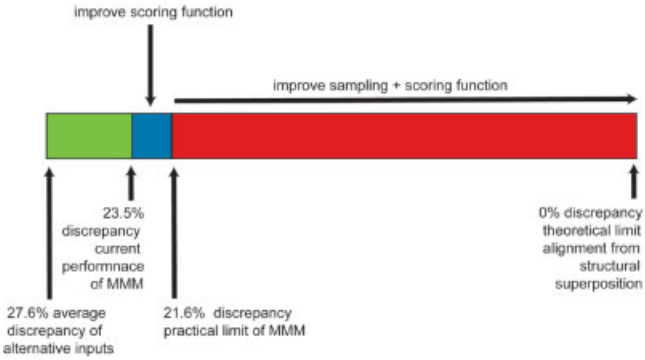

Fig. 8. Schematic overview of MMM performance and possible strategies for improvement. Lengths of bars are realistic. "Average discrepancy of input alignments" is the baseline of the approach. "Practical limit of MMM" is obtained by the optimal combination of more accurate parts of input alignments. "Current performance" reflects actual benchmark results with MMM composite scoring function. "Theoretical limit for improvements" is defined by the "gold standard" alignment derived from a structural superposition using STAMP.

representation of insertions in the mapping is only approximate. A more accurate position for these target residues can be determined only after building a model. We calculated the MMM performance using the second approach, where the MJ component of the scoring function was calculated using the target models that were built for all alternative inputs instead of just using the framework defined by the template structure. Although the second approach is more realistic than the first one, the gain in the MMM performance, compared to the case when first approach is used, is approximately 1%. This statistically insignificant improvement comes at the expense of a many-fold more demanding computation.

MMM versus Genetic Algorithm

The MMM approach is somewhat similar to the genetic algorithm-based methods, which have been applied before to correct alignment errors by recombination and crossover of a set of target models obtained from different alignments of the same template-target pair.[19,37] Both approaches, MMM as well as the genetic algorithm techniques, use a set of alternative alignments (models) in conjunction with a scoring function to generate optimal alignment (models). Differences between the two methods lie in the scoring function and the search algorithm. Although the search space of the MMM is limited to a small number of variable region fragments and their combinations, the genetic algorithm generates a large number of models at each step of iteration by applying recombination and crossover operators at randomly selected protein positions of two randomly selected models of the population set. Compared to the genetic algorithm-based techniques, the MMM approach has two advantages. First, because the MMM does not require to build models and its search space is limited to biologically relevant alignments, the computational time required for this method is one to two orders of magnitude smaller than the genetic algorithm based techniques. Typically, the MMM alignment (model) can be generated in minutes using a one processor Pentium machine, whereas a similar calculation using genetic algorithm may require several hours or even days of computer time. Second, the genetic algorithm builds a large number of models by cut and paste, which often corresponds to biologically unrealistic situations and challenges the scoring functions. In contrast, MMM does not encounter such problems, because it considers a small number of alternative alignments built from existing methods that are expect to produce biological meaningful results, although with varying accuracies. The restricted sampling space is also a limitation for MMM, but it can be systematically extended by combining more alternative inputs (Fig. 8).

Applications

The improved accuracy of MMM alignments directly influences the accuracy of the comparative models. The improvement in the comparative models has a direct bearing on further studies based on those models. Numerous applications of comparative modeling involving site-directed mutagenesis experiments,[38,39] construction of macromolecular assemblies,[40] identification of catalytic residues,[41,42] and refinement of NMR structures,[43,44] low-resolution electron density maps,[45] etc., critically rely on the accuracy of the alignment.

Because the MMM algorithm is computationally efficient, it can be applied to large-scale structure modeling experiments to build database of models for all known protein sequences that have detectable similarity to at least one known protein structure.[17,21]

A Web service has been set up for the current method that is accessible at http://www.fiserlab.org/servers/MMM.

CONCLUSION

We presented a novel approach to improve the accuracy of sequence to structure alignment for comparative modeling by optimally combining two or more input alignments

from existing methods. The gain in the alignment accuracy from the baseline that corresponds to the average accuracy of input alignments is achieved by first identifying the stretches where the input alignments differ and subsequently selecting the better aligned fragments from the input set. The scoring scheme evaluates the compatibility of the alternatively aligned fragments in the structural environment of the template structure. The method provides particular advantage in the low-sequence identity regions of "twilight" and "midnight zones," where different methods or choice of scoring functions produce dramatically different results. Because MMM does not require input models to be built, and the search space explored by this algorithm is rather limited, this method is computationally efficient, making it a suitable tool even for large scale modeling of genomic sequences.

MATERIALS AND METHODS

Dataset of Structurally Related Sequence Pairs

The template and target sequences were obtained from the PDB[46] using a series of strict filtering criteria. As the first step in the preparation of the benchmark dataset, 16321 distinct protein chains were extracted from the PDB using the following constraints: (1) crystallographic resolution of the chains 2.5 Å or better, (2) no missing residues in the chain, (3) length of the chain 70 residues or longer. To avoid the bias in the PDB associated with the presence of multiple entries of identical proteins, or the proteins that differ only by a small number of residues in the chain, further filtering of the data set was done by clustering the protein chains at 90% sequence identity cutoff. The clustering of the protein chains was done using the NRDB90 program by Holm and Sander,[47] producing 3137 clusters containing one or more protein chains within each cluster. For clusters with more than one chain, one representative protein chain with the best atomic resolution was selected.

All-to-All Pairwise Structural Alignment Using STAMP

An all-to-all pairwise structural alignment of the 3137 protein chains, obtained from the previous filtering step, was performed using STAMP.[48] In total, $3137 \times 3136/2 = 4{,}918{,}816$ pairwise structural superpositions were made. The STAMP pairwise alignments indicating structural relationship between template and target proteins were retained. A combined criterion for this filtering step was based on the STAMP score (5.5 or higher) and a minimum 80% structural overlap. With these strict constraints, only 5659 alignments out of 4,918,816 were retained, leading to 2×5659 unique template-target pairs in the data set.

Separating the Alignment and Loop Modeling Problem

Three possible scenarios occur for each residue position in the alignment: (1) match, where a template residue is aligned with a target residue; (2) gap, where a template residue is aligned to a gap in the target; or (3) insertion, where a gap in the template is aligned to a residue in the target. From the model building perspective, cases (1) and (2) pose little difficulty. In contrast, accurate modeling of segments with continuous multiple residue insertions in the template, which often correspond to loop regions of protein, is challenging. The presence of long loop segments can potentially lead to large modeling errors that are not a consequence of alignment quality, but reflect on the accuracy of loop modeling. Similarly, the presence of insertions at either end of the template protein chain can potentially cause large modeling errors in terms of overall RMSD, as the modeling program would not be constrained to place the inserted residues in any particular orientation. Because the aim of the present work is to build improved comparative models by optimizing the alignments, we decided to exclude all cases with more than eight residue-long insertions from our data set because MODELLER[29] is able to provide reasonable conformations for loops upto size 8;[27] therefore, loop errors would not burden the overall RMSD. Further, to minimize the errors caused by the terminal flanking residues, we truncated the target chains to eliminate all inserted residues at both ends of the target protein.

Separating the Alignment Problem from Intrinsic Modeling Errors

We use MODELLER with its default parameters to build a model of target sequence from a given alignment. Using the "model" subroutine of MODELLER a number of models can be generated for the same input alignment, which slightly differ from each other because of the stochastic nature of optimization routine. To assess the extent of variability among the models produced from the same alignment, we built 10 structures using the STAMP alignment as input for each template-target pair in our dataset. The RMSD difference between the worst and the best of the 10 models covers a range. To minimize the uncertainty associated with intrinsic modeling errors, we eliminated all cases from our dataset where the maximum deviation among the models exceeded 0.5 Å; therefore, reproducibility was not granted. The remaining 6635 template-target pairs, with high-resolution X-ray structure and high template-target structural similarity, were used to develop and test the Multiple Mapping Method.

Alternative Input Alignments

The input to the MMM is a set of two or more pairwise alignments of target sequence to the template sequence/structure. For the development and testing of the MMM method, we used two sets, each consisting two different pairwise alignments of the same template-target pair: (1) the first set included alignments by CLUSTALW[10] (using default gap penalty parameters) and Align2D[17] methods; (2) the second set included two alternative alignments both generated by CLUSTALW but using different gap penalty parameters.

The Align2D program (with default parameters) from MODELLER was used to produce alignments between target sequence and template structure. This method implements Needleman-Wunsch[49] global dynamic programming algorithm with a structure-dependent gap pen-

alty function, such that gaps in the buried secondary structure regions get higher penalty compared to those in the loop and solvent accessible parts.

For the second benchmark set we used alternatively parameterized CLUSTALW to generate pairwise global alignments between template and target sequences. For each template-target pair of our dataset, we generated two sets of CLUSTALW alignments; using default gap penalty parameters (gap initiation and extension penalties 10 and 0.2, respectively) in one case, and modified parameters with gap initiation and extension penalties 5 and 0.1, respectively, in the other case. We refer these two types of alignments as $\text{CLUSTALW}_{\text{def}}$ and $\text{CLUSTALW}_{\text{mod}}$. Additional benchmarking of MMM was done using alignments from MUSCLE[32] program along with those obtained from $\text{CLUSTALW}_{\text{def}}$, $\text{CLUSTALW}_{\text{mod}}$, and Align2D as input to the method. The MMM performance was studied using all four input alignments, as well as all possible combinations of input pairs and triplets. We also study the performance of our method using inputs from two publicly available servers (ESyPred3D and Consensus) that combine alignments from diverse sources.

Multiple Mapping Method: Sampling Alignment Space

The method takes as input two or more nonidentical alignments of the target sequence to a template sequence or structure. The input alignments can be generated using different alignment methods, or by using optimal and several suboptimal alignments from the same method. A consensus alignment of the input set is built to locate variable and nonvariable regions of the alignment. A variable region is defined as a segment of the alignment, where one or more residues of the target sequence are aligned differently to the template. In contrast, the nonvariable region is defined as identically aligned template-target residue pair positions in the input alignments. The input alignment sets, depending on their relative differences, may contain one or more stretches of variable regions that are separated by nonvariable segments. The quality of the input alignments is determined by the accuracy of the variable regions.

In the MMM algorithm, an exhaustive set of alternative alignments are built by joining the nonvariable regions with all possible combinations of the variable regions. The algorithm uses a scoring scheme to rank these alternative alignments, and the best scoring alignment is the output of MMM. The output is an alignment of the same template-target pair. This output can be a hybrid alignment, generated by combining spliced fragments from the input alignment set.

Multiple Mapping Method: Scoring Function to Evaluate Sequence–Structure Compatibility

The scoring function determines the preference of each variable region residue (and its predicted secondary structure) of the target sequence in the corresponding position of the template as a function of the template environment. The total score of a given variable region is the sum of scores from each individual positions of this segment. The alternative variable regions are scored using a scoring function consisting of three components, each of which takes into account the 2D or 3D structures of the template protein. The three components of the scoring function are as follows: (1) a set of environment specific substitution matrices (FUGUE);[7] (2) a substitution matrix for secondary structure match (H3P2);[30] (3) a statistical potential, describing residue–residue contact energy by Miyazawa and Jernigan (MJ).[31]

Environment Specific Substitution Scores

The environment specific substitution scores for each alternative alignment are calculated as the linear sum of the score for each variable region residue using data from the FUGUE substitution tables[7]

$$S_{\text{essm}} = \sum_{i=1}^{N_{\text{vr}}} S_{i,\text{essm}}$$

where the summation is over all variable region positions, N_{vr}. Environment-specific scores consist of two components: (1) Environment-specific substitution matrices, and (2) environment-dependent gap penalties. The score of a residue or a gap at the position i is given by

$$S_{i,\text{essm}} = \begin{cases} S(\text{res}_i^{\text{tm}}, \text{res}_i^{\text{tr}} \,|\, E_i) \\ s(\text{res}_i^{\text{tm}}, - \,|\, E_i) \\ s(-, \text{res}_i^{\text{tr}} \,|\, E_i) \end{cases}$$

where $s(\text{res}_i^{\text{tm}}, \text{res}_i^{\text{tr}}|E_i)$ represents the likelihood of substituting res_i^{tm} by res_i^{tr} in the local environment E_i of the template structure. The environment specific gap penalty function corresponding to deletion and insertion in the template structure is given by $g(\text{res}_i^{\text{tm}}, -|E_i)$ and $g(-, \text{res}_i^{\text{tr}}|E_i)$, respectively. The superscripts "tm" and "tr" refer to the template and target sequences, respectively.

The local environment of the template structure, corresponding to the ith position in the alignment, is a function of three groups of structural features: main-chain conformation and secondary structure (ss), solvent accessibility (acc), and hydrogen bonding status (hb).

$$E_i \equiv E_i(ss_i, acc_i, hb_{i,ss}, hb_{i,so}, hb_{i,sn})$$

where ss represents four classes of secondary structures: alpha-helix, beta-strand, irregular structure (coil), and residue with a positive phi main-chain torsion angle. The solvent accessibility (acc) is defined by two classes: buried and accessible. Three groups of hydrogen bonds, corresponding to side-chain to side-chain ($hb_{i,ss}$), side-chain to main-chain CO ($hb_{i,so}$), and side-chain to main-chain NH ($hb_{i,sn}$), were considered. All possible combinations of these structural features defines 64 local structural environments, resulting 64 environment-specific substitution matrices.

An environment-dependent gap penalty scheme, as implemented in FUGUE, was used to penalize insertions/deletions in the variable regions. In this scheme, the gap costs depend upon the local structural features of the template. For example, gaps in the middle of the secondary

structure elements are penalized more than those occurring in the loop regions. Similarly, the introductions of gaps in the buried protein regions cost more than those in the accessible regions.

To calculate environment specific score, we first classify each variable region residue of the template structure according to the three FUGUE structural groups. The classification of the template residues into the three secondary structural groups (alpha, beta, and coil), as well as the accessibility information, were derived from DSSP.[50]

Statistical Potential-Based Scoring Scheme

The second component of the scoring function uses the statistical potentials by Miyazawa and Jernigan.[31] We determine the fitness of alternative variable regions in their local environments using statistically derived residue–residue contact energies. The fitness score for a variable region is the sum of contact energies of each residue from that variable region. We explored two alternative ways for calculating the fitness score of the variable regions.

In one approach, each target residue is mapped to a corresponding position in the template structure. In the case of one or more insertions in the template, we map the target residues at the midpoint of the straight line joining the two template residues between which the target residues are to be placed. It is arguably a highly approximate mapping, especially in the case of several consecutive insertions when multiple target residues are mapped to the same position in the template structure. However, this simplified approach avoids the necessity of model building, which would be the most time-consuming part of the method.

The local environment can also be identified from the models of target sequences, built from the alternative alignments. This approach is more realistic than a direct mapping of target residues onto the template structure.

The procedure for calculating the contact energy is as follows: first, the target residue from the variable region is mapped to a position in the template structure that corresponds to the template residue in the aligned pair. Next, the local environment of the reference residue is determined by selecting the residue positions that are within 10 Å cutoff distance from the reference residue in the template structure. The template residues in the environment are replaced by the corresponding target residues from the alignment; finally, the pairwise contact energy of each residue in the environment with the reference residue is calculated using the statistical potential table.

Secondary Structure-Based Substitution Scores

The H3P2[30] score for an alternative alignment is obtained by summing over the scores from each position in the variable region.

$$S_{\mathrm{H3P2}} = \Sigma s(\mathrm{res}_{c,i}^{tm}, ss_{i,o}^{tm}, acc_{i,o}^{tm}; res_{c,i}^{tr}, ss_{i,p}^{tr})$$

This component of the scoring system uses substitution scores from the comparison of the secondary structure classifications and residue accessibility of the aligned residue pairs. Each residue position in the target sequence of the alignment is defined by one of seven residue classes and three secondary structure classes. Each residue position in the template is defined by seven residue classes, three secondary structure classes, and two residue accessibility classes. The resulting substitution matrix is five dimensional, with three dimensions for the template structure and two dimensions from the target sequence. The prediction of the three-state (helix, strand, coil) secondary structure of the target sequence is done using a profile-based neural network program (PHD).[51] For the template, we use DSSP to determine the three-state (helix, strand, coil) secondary structure classes. The accessibility types of the template residues were calculated from the fractional buried area of residues (obtained using DSSP program) using the same buried/accessible boundary cutoffs for each of the seven residue classes as before.[30]

Multiple Mapping Method: Scoring Strategy

The three scoring function components, as described previously, are used to rank the alternative variable regions.

Composite Score from a Linear Combination of Scoring Function Components

The three components of the scoring scheme (environment specific substitution scores, statistical potential scores, and secondary structure based substitution scores) contribute to a composite score. The individual values are first transformed into the corresponding *Z*-scores from using the mean (μ) and standard deviation (σ) of the scores from random alignments, as $Z(\mathrm{score}) = (\mathrm{score} - \mu)/\sigma$. The procedure for generating the randomized data set for *Z*-score calculation for different components of our scoring scheme varies.

In the case of environment specific substitution scores, the random dataset is produced by replacing each variable region residues of the target sequence by amino acid residues according to the probabilities of their occurrence in native proteins. The local environment of each residue position in the variable region of the template structure is randomly assigned to 1 of the 64 environment classes as described previously. The gap positions in the variable regions of either template or target are preserved, and are not substituted by other residues. However, the local environment of the template structure corresponding to the gap position is randomly changed to one of the sixty four classes, as before.

The *Z*-scores corresponding to the statistical potential based scores are calculated from a dataset of random scores, which are obtained by replacing each residue of the environment by a randomly generated residue according to their probability of occurrence in native proteins. As in the case of the *Z*-scores for the environment-specific substitution matrix, all gap positions in the environment are preserved.

The *Z*-scores for the H3P2 substitution matrix-based scoring component are calculated by randomly shuffling

the residues with the same predicted secondary structure elements of target sequence.

The composite score is a weighted linear combination of these three *Z*-scores, where weights for each component were determined in an iterative fitting using the CLUSTALW$_{def}$-Align2D dataset. For all the results presented in this work, a weight combination of 0.4 for Fugue and H3P2 components and 0.2 for MJ was used, as it was found to give better performance than all other combinations that were explored from all possible discrete combinations of the weights.

Evaluating the Performance of MMM

The performance of the MMM is evaluated on the structural level. The RMSD between the corresponding main-chain atoms (N, Cα, C, O) in the model and the PDB structure was calculated upon rigid body superposition of the main-chain atoms, as implemented in the SUPERPOSE routine of MODELLER.

Models are built for each input alignment, for MMM and for the "gold standard" STAMP alignment. The relative accuracies of the MMM models and the models built on input alignments are determined through a comparison to the structure-based alignments obtained from the STAMP program. The performance of MMM is benchmarked by comparing the accuracy of MMM model with the overall accuracy of models from input alignments. The alignment discrepancy is defined as the percentage of positions in the tested alignment that are identical to those in the STAMP structure-based alignment. The accuracies of the models built from the MMM alignments and those from the input alignments was determined by comparison with the corresponding target structure extracted from the PDB.

ACKNOWLEDGMENTS

We are grateful to Drs. Eduardo Fajardo, Dmitrij Rykunov, and Narcis Fernandez-Fuentes for their comments on the manuscript, and to Mr. Carlos Madrid for setting up the MMM server.

REFERENCES

1. Fiser A. Protein structure modeling in the proteomics era. Expert Rev Proteomics 2004;1:89–102.
2. Fiser A.; Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 2003;374:461–491.
3. Finkelstein AV, Reva BA. A search for the most stable folds of protein chains. Nature 1991;351:497–499.
4. Bowie JU, Luthy R, Eisenberg D. A Method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
5. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287:797–815.
6. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J Mol Biol 2000;297:1003–1013.
7. 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.
8. Blake JD, Cohen FE. Pairwise sequence alignment below the Twilight Zone. J Mol Biol 2001;307:721–735.
9. Przybylski D, Rost B. Improving fold recognition without folds. J Mol Biol 2004;341:255–269.
10. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
11. 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.
12. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
13. 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.
14. Marti-Renom MA, Madhusudhan MS, Sali A. Alignment of protein sequences by their profiles. Protein Sci 2004;13:1071–1087.
15. Rost B. Twilight Zone of protein sequence alignments. Protein Eng 1999;12:85–94.
16. Wang G, Dunbrack RL Jr. Scoring profile-to-profile sequence alignments. Protein Sci 2004;13:1612–1626.
17. Sanchez R, Sali A. Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci USA 1998;95:13597–13602.
18. Prasad JC, Comeau SR, Vajda S, Camacho CJ. Consensus alignment for reliable framework prediction in homology modeling. Bioinformatics 2003;19:1682–1691.
19. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 2003;31:3982–3992.
20. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol 2001;8:559–566.
21. Chance MR, Fiser A, Sali A, Pieper U, Eswar N, Xu G, Fajardo JE, Radhakannan T, Marinkovic N. High-throughput computational and experimental techniques in structural genomics. Genome Res 2004;14:2145–2154.
22. Tress ML, Jones D, Valencia A. Predicting reliable regions in protein alignments from sequence profiles. J Mol Biol 2003;330:705–718.
23. Contreras-Moreira B, Fitzjohn PW, Bates PA. In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. J Mol Biol 2003;328:593–608.
24. Vingron M, Argos P. Determination of reliable regions in protein sequence alignments. Protein Eng 1990;3:565–569.
25. Chao KM, Hardison RC, Miller W. Locating well-conserved regions within a pairwise alignment. Comput Appl Biosci 1993;9:387–396.
26. Mevissen HT, Vingron M. Quantifying the local reliability of a sequence alignment. Protein Eng 1996;9:127–132.
27. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
28. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem 1977;80:319–324.
29. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
30. Rice DW, Eisenberg D. A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
31. Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256:623–644.
32. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC—Bioinformatics 2004;5:113.
33. Lambert C, Leonard N, De BX, Depiereux E. ESyPred3D: prediction of proteins 3D structures. Bioinformatics 2002;18:1250–1256.
34. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.
35. Fischer D, Rychlewski L, Dunbrack RL Jr, Ortiz AR, Elofsson A. CAFASP3: the third critical assessment of fully automated

structure prediction methods. Proteins 2003;53(Suppl 6):503–516.
36. Prasad JC,Vajda S, Camacho CJ. Consensus alignment server for reliable comparative modeling with distant templates. Nucleic Acids Res 2004;32:W50–W54.
37. Contreras-Moreira B, Fitzjohn PW, Offman M, Smith GR, Bates PA. Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. Proteins 2003;53(Suppl 6):424–429.
38. Matsumoto R, Sali A, Ghildyal N, Karplus M, Stevens RL. Packaging of proteases and proteoglycans in the granules of mast cells and other hematopoietic cells. A cluster of histidines on mouse mast cell protease 7 regulates its binding to heparin serglycin proteoglycans. J Biol Chem 1995;270:19524–19531.
39. Wu G, Fiser A, ter Kuile B, Sali A, Muller M. Convergent evolution of trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase. Proc Natl Acad Sci USA 1999;96:6285–6290.
40. Spahn CM, Beckmann R, Eswar N, Penczek PA, Sali A, Blobel G, Frank J. Structure of the 80S ribosome from *Saccharomyces cerevisiae*—TRNA-ribosome and subunit-subunit interactions. Cell 2001;107:373–386.
41. Navaratnam N, Fujino T, Bayliss J, Jarmuz A, How A, Richardson N, Somasekaram A, Bhattacharya S, Carter C, Scott J. *Escherichia coli* cytidine deaminase provides a molecular model for ApoB RNA editing and a mechanism for RNA substrate recognition. J Mol Biol 1998;275:695–714.
42. Fiser A, Filipe SR, Tomasz A. Cell wall branches, penicillin resistance and the secrets of the MurM protein. Trends Microbiol 2003;11:547–553.
43. Nagata T, Gupta V, Sorce D, Kim WY, Sali A, Chait BT, Shigesada K, Ito Y, Werner MH. Immunoglobulin motif DNA recognition and heterodimerization of the PEBP2/CBF runt domain. Nat Struct Biol 1999;6:615–619.
44. Barrientos LG, Campos-Olivas R, Louis JM, Fiser A, Sali A, Gronenborn AM. 1H, 13C, 15N resonance assignments and fold verification of a circular permuted variant of the potent HIV-inactivating protein cyanovirin-N. J Biomol NMR 2001;19:289–290.
45. Palaniyar N, McCormack FX, Possmayer F, Harauz G. Three-dimensional structure of rat surfactant protein A trimers in association with phospholipid monolayers. Biochemistry 2000;39:6310–6316.
46. Bourne PE, Addess KJ, Bluhm WF, Chen L, Deshpande N, Feng Z, Fleri W, Green R, Merino-Ott JC, Townsend-Merino W, Weissig H, Westbrook J, Berman HM. The distribution and query systems of the RCSB Protein Data Bank. Nucleic Acids Res 2004;32(Database issue):D223–D225.
47. Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998;14:423–429.
48. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins 1992;14:309–323.
49. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
50. Kabsch W, Sander C. On the Use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. Proc Natl Acad Sci USA 1984;81:1075–1078.
51. Rost B. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 1996;266:525–539.