3982–3992 *Nucleic Acids Research, 2003, Vol. 31, No. 14*
DOI: 10.1093/nar/gkg460

Comparative protein structure modeling by iterative alignment, model building and model assessment

Bino John and Andrej Sali[1,*]

Laboratory of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, New York, NY 10021, USA and [1]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry and California Institute for Quantitative Biomedical Research, University of California at San Francisco, San Francisco, CA 94143, USA

Received April 4, 2003; Revised and Accepted May 8, 2003

ABSTRACT

Comparative or homology protein structure modeling is severely limited by errors in the alignment of a modeled sequence with related proteins of known three-dimensional structure. To ameliorate this problem, we have developed an automated method that optimizes both the alignment and the model implied by it. This task is achieved by a genetic algorithm protocol that starts with a set of initial alignments and then iterates through re-alignment, model building and model assessment to optimize a model assessment score. During this iterative process: (i) new alignments are constructed by application of a number of operators, such as alignment mutations and cross-overs; (ii) comparative models corresponding to these alignments are built by satisfaction of spatial restraints, as implemented in our program MODELLER; (iii) the models are assessed by a variety of criteria, partly depending on an atomic statistical potential. When testing the procedure on a very difficult set of 19 modeling targets sharing only 4–27% sequence identity with their template structures, the average final alignment accuracy increased from 37 to 45% relative to the initial alignment (the alignment accuracy was measured as the percentage of positions in the tested alignment that were identical to the reference structure-based alignment). Correspondingly, the average model accuracy increased from 43 to 54% (the model accuracy was measured as the percentage of the Cα atoms of the model that were within 5 Å of the corresponding Cα atoms in the superposed native structure). The present method also compares favorably with two of the most successful previously described methods, PSI-BLAST and SAM. The accuracy of the final models would be increased further if a better method for ranking of the models were available.

INTRODUCTION

High throughput sequencing of many genomes is yielding a plethora of protein sequences (1,2). The functions of these proteins now need to be described, understood and manipulated. To this end, it is generally useful to know the three-dimensional structures of the proteins. In the absence of an experimentally determined structure, comparative or homology modeling can sometimes provide a useful model of a protein (target) that is related to at least one known protein structure (template) (3,4). As a result, comparative protein structure modeling is relevant to structure-based functional annotation of proteins and thus enhances the impact of genome sequencing, structural genomics and functional genomics on biology and medicine.

The overall accuracy of useful comparative models spans a wide range, from models with only the correct fold to more accurate models that are comparable to structures determined by low resolution X-ray crystallography or medium resolution nuclear magnetic resonance (NMR) spectroscopy (5). In general, errors in comparative models include errors in side chain packing, distortions and shifts of core segments of the fold, errors in modeling of insertions (e.g. loops) and errors resulting from an incorrect alignment and fold assignment. Alignment errors are particularly detrimental because they are frequent and have a large impact on the model accuracy. Unfortunately, no current comparative model building method can generally recover from errors in the input alignment. Consequently, even residues that are misaligned by a single position are most likely modeled with an error larger than the spacing between two consecutive Cα positions (i.e. 3.8 Å). Alignment mistakes not only trigger large errors in a model, but are also frequent. Most pairs of detectably related protein sequences and structures are related at less than 30% sequence identity, where the alignment errors become significant (6,7); at 30% sequence identity, ~20% of residues are misaligned on average (8). In large-scale modeling of all known protein sequences that are detectably related to at least one known structure, the alignment errors on their own are responsible for about half of the grossly mismodeled residues (i.e. residues whose Cα positions are modeled with an error >5 Å) (8). In summary, due to both their impact and frequency, alignment

*To whom correspondence should be addressed at Mission Bay Genentech Hall, Suite N472D, 600 16th Street, University of California at San Francisco, San Francisco, CA 94143-2240, USA. Tel: +1 415 514 4227; Fax: +1 415 514 4231; Email: sali@salilab.org

errors are the most important single limitation on comparative modeling (9).

When correct alignment of the target with the template is difficult, the best comparative model for the target may be obtained by using a number of alternative alignments to build and assess the corresponding models (10–14). A step further in this direction would be to use model assessment techniques in the search for alignments that generate models with good model assessment scores; recovery from alignment errors may be possible if the corresponding errors in the comparative models can be detected. Fortunately, there is a great variety of model assessment methods, which can sometimes detect some errors in the assessed models (15–17).

In this paper, we focus on minimizing the impact of an incorrect input alignment on the final comparative model. In particular, we describe an automated protocol for comparative modeling that is capable of refining the initial target–template alignment in the search for the best model. Thus, the protocol is expected to be less sensitive to errors in the input alignment than the traditional approaches to comparative modeling that rely on one or a few input alignments. The protocol iteratively optimizes both the alignment and the model. The alignments are explored by a genetic algorithm, while a model given an alignment is obtained by our standard comparative modeling procedure implemented in MODELLER (3).

Because the number of all possible alignments of a given target–template pair is enormous (18), enumerating and testing all of the alignments by assessing the corresponding models is computationally prohibitive. Assuming that each alignment and model building take a few minutes of CPU time, it is necessary to use an efficient optimization algorithm that can find a good alignment by testing on the order of 10 000 possibilities. Genetic algorithms (19) have been used for a variety of difficult optimization problems, such as protein folding (20,21), protein docking (22), *de novo* design of protein sequences (23), protein sequence alignment (24–26) and phylogeny estimation (27). Genetic algorithms are inspired by natural selection in evolution. A population of individuals evolves through selection of various mutations of the individuals and recombinations between individuals. The selection is guided by a fitness function. In our implementation, the individual is a target–template alignment, the mutations and recombinations are changes of the alignments and the fitness of an alignment is a composite assessment score for a model implied by the alignment. This iterative approach blurs the boundary between traditional comparative modeling, which calculates a highly refined model for one alignment, and the threading methods (28–30), which predict the optimal alignment by scoring a simple implicit model for each one of the many tested alignments.

We begin by describing the iterative modeling protocol, the benchmarking criteria and the remotely related protein structure pairs used for benchmarking (Methods). In Results, we assess the accuracy of the protocol and illustrate it by a detailed description of the modeling of one protein sequence. We conclude by discussing the implications of the results for comparative protein structure modeling (Discussion).

METHODS

A flow chart of our approach to iterative alignment and modeling is shown in Figure 1. The inputs are: (i) a reliable multiple sequence alignment of the target with its close homologs (target profile); (ii) one or more template structures; (iii) a reliable multiple alignment of the template structures with their close homologs (template profile). The outputs are refined target–template alignments with the corresponding comparative models for the target sequence, ranked by a composite model score. The modeling protocol was implemented on a cluster of computers running the Linux operating system. For a 150 residue target sequence, the protocol currently requires ~1 day of CPU time on 50 nodes with dual 1 GHz Pentium III CPUs; the CPU time scales approximately linearly with the length of the sequence. In the following sections, we describe the individual steps in the flow chart. The steps at the beginning of the section titles refer to the flow chart in Figure 1.

Steps 1 and 2. Generating initial target–template alignments

The target and template profiles were obtained by PSI-BLAST (version 2.0.11) (31), scanning the non-redundant protein sequence database at NCBI (March 2002) with the E-value cut-off of 10^{-4} for up to 20 iterations and retaining only up to 1000 matches with the most significant E-values. The target and template profiles were then aligned by the SALIGN command in MODELLER (M.A.Marti-Renom, M.S.Madhusudhan and A.Sali, in preparation). SALIGN implements global dynamic programming (32) for alignment of two sequence profiles, with a linear gap penalty function. This method is similar to that of FFAS (33) and usually results in alignments that are 5–10% more accurate than those of PSI-BLAST. A single 'initial comparative model' was obtained based on the profile–profile alignment with the optimal initiation and extension gap penalties. If this initial comparative model was assessed to be insufficiently accurate (steps 5 and 6), a variation of the initiation and extension gap penalties on a 5×5 grid centered on the optimal penalties was used to calculate 25 alignments. Each of these alignments was ranked by the alignment score described in step 3 below. The 15 top scoring alignments (i.e. 'initial parent alignments') were subsequently subjected to evolution by the genetic algorithm operators (step 7).

Step 3. Ranking alignments by an alignment score

An alignment between the target and template profiles was scored by the sum of the substitution scores and gap penalties, as implemented in the SALIGN command of MODELLER. The substitution score between a target profile position and a template profile position is defined as 1000 times the correlation coefficient between the estimated relative frequencies of each of the 20 standard residue types at these two positions. Gaps were scored with the linear gap penalty function $u + n\,v$, where u is the gap initiation penalty of –575, v is the gap extension penalty of –35 and n is the length of the gap.

Step 4. Building molecular models

For a given alignment, a single comparative model of the target sequence that contains all non-hydrogen atoms was built

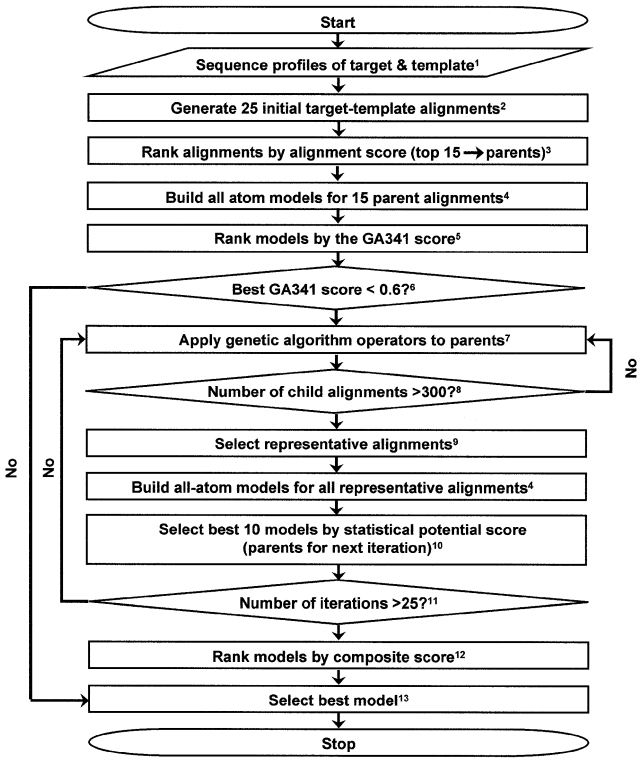

Figure 1. Overview of modeling by iterative alignment, model building and model assessment. An initial set of alignments is generated using sequence profiles of the target and template sequences (steps 1 and 2). Comparative models implied by the alignments are built by MODELLER (step 4) and ranked by the GA341 score (step 5). If the predicted model accuracy is low (GA341 score <0.6), genetic algorithm operators are applied to the selected initial alignments to generate new alignments (step 7). The cycle of alignment (steps 7–9), model building (step 4) and model assessment (step 10) is continued for up to 25 iterations (step 11). A composite model assessment score is used at the end to assess the accuracy of the models corresponding to all of the representative alignments from all 25 iterations (step 13). The top model is selected as the final output from the protocol (step 13). Refer to Methods for a detailed description of the steps.

by MODELLER-6, applying the default model building routine 'model' with fast refinement (3).

Steps 5 and 6. Ranking models by the GA341 score

A molecular model was assessed by a GA341 score that combines a Z-score (Z_s) calculated with a statistical potential function (16), target–template sequence identity (S_i) and a measure of structural compactness (S_c) (6,16). The GA341 score is defined as:

$$GA341 = 1 - [\cos(S_i)]^{(S_i + S_c)/\exp(Z_s)} \quad \mathbf{1}$$

Sequence identity is the fraction of positions with identical residues in the target–template alignment. Structural compactness is the ratio between the sum of the standard volumes of the amino acid residues in the protein and the volume of the sphere with the diameter equal to the largest dimension of the model. The Z-score is calculated for the combined statistical potential energy of a model, using the mean and standard deviation of the statistical potential energy of 200 random sequences with the same composition and structure as the model (16). The combined statistical potential energy of a model is the sum of the solvent accessibility terms for all Cβ atoms and distance-dependent terms for all pairs of Cα and Cβ atoms. The solvent accessibility term for a Cβ atom depends on its residue type and the number of other Cβ atoms within 10 Å; the non-bonded terms depend on the atom and residue types spanning the distance, the distance itself and the number of residues separating the distance-spanning atoms in sequence. These potential terms reflect the statistical preferences observed in 760 non-redundant proteins of known structure. The GA341 scoring function was evolved by a genetic algorithm that explored many combinations of a variety of mathematical functions and model features, to optimize the discrimination between good and bad models in a training set of models. The GA341 score ranges from 0 for models that tend to have an incorrect fold to 1 for models that tend to be comparable to at least low resolution X-ray structures.

If the top ranked model in the initial population had a GA341 score >0.6, the initial alignments were assumed to be of sufficiently high accuracy to avoid refinement by our relatively coarse scheme; in such a case, the initial comparative model (steps 1–3) is the final output of the protocol.

Step 7. Genetic algorithm operators

A new generation of alignments ('child' alignments) evolves through an application of five genetic algorithm operators on a random subset of the current parent alignments. For the very first iteration, the parent alignments are the 15 'initial parent alignments' (steps 1–3). For the subsequent iterations, the parent alignments are the 10 best alignments selected in step 10. The operators are applied iteratively until at least 300 child alignments are accumulated.

The single point crossover operator swaps alignment segments between two parent alignments to generate two child alignments (Fig. 2A). The swapped segment is defined by a randomly chosen alignment position that matches the same pair of residues in both parents. While the first parent is chosen randomly, the second parent is selected by the roulette wheel rule, whereby the probability of selection is proportional to the rank of the parent in the population (19). The rank of an alignment is defined by its alignment score (step 3).

The two point crossover operator also swaps alignment segments between two parent alignments to generate two child alignments (Fig. 2B). The two parent alignments are selected as described for the single point crossover operator. Next, all alignment segments that are different between the two parents are mapped. One of these segments is selected randomly and swapped between the two parents to generate two children.

The gap insertion operator inserts a gap into a parent alignment (Fig. 2C). An alignment segment is selected randomly. Gaps of equal length are inserted at the beginning and the end of the alignment segment in the target and the template sequences, respectively. The length of the inserted gaps is a random number distributed uniformly between 1 and 20.

The gap deletion operator deletes a gap from a parent alignment (Fig. 2D). A gap is selected randomly from all the gaps in the parent alignment. Next, the gap is shortened by a random number of positions, distributed uniformly from 1 to the gap length.

The gap shift operator shifts existing gaps in a parent alignment (Fig. 2E). An alignment position with a gap to the

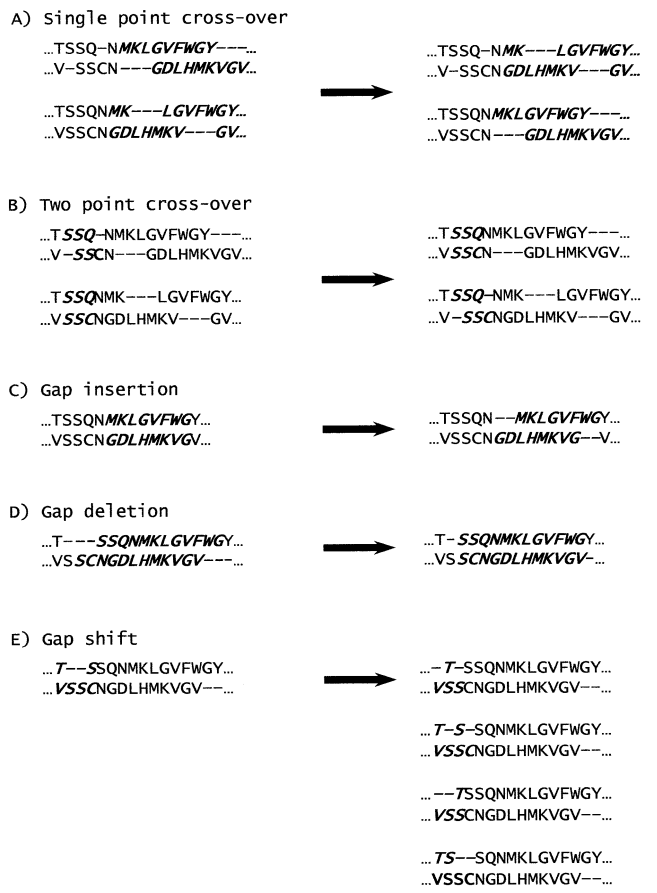

Figure 2. Genetic algorithm operators used in iterative alignment, model building and model assessment. The five genetic algorithm operators that transform parent alignment(s) on the left into child alignment(s) on the right are illustrated in (**A**–**E**). Alignment segments shown in bold italic type are altered by the operation. See Methods for details.

right of it is selected randomly in the parent alignment. Next, the first *m* gap positions are moved before each one of the 20 preceding residues, generating 20 children for each integer *m* from 1 to the length of the gap. Finally, the same procedure is repeated at the C-terminal end of the gap, moving the gapped positions of increasing length after each one of the 20 subsequent residues. This operator creates many similar children, but redundancy is eliminated during the selection of the representative alignments in step 9.

Step 9. Selecting representative alignments

The parent and child alignments from the current iteration are pooled. The redundant alignments from this pool are eliminated, such that the remaining representative alignments share no more than 95% identically aligned positions or have at least five different alignment positions. This filtering typically eliminates 20% of the alignments.

Step 10. Selecting best models by a statistical potential score

The statistical potential Z-score, Z_s (step 5), was used to assess the overall accuracy of the molecular models for each of the representative alignments from the current iteration (16). Alignments for 10 models with the lowest statistical potential Z-scores were selected as the parent alignment for further evolution in the next iteration (steps 7–11).

Steps 11 and 12. Ranking models by a composite model assessment score

After 25 iterations of the target–template alignment and model building, all models calculated based on the representative alignments from all iterations are ranked by a composite model assessment score. The following five scores contribute to the composite score: pair (P_p) and surface (P_s) statistical potential values (step 10) (16); structural compactness (S_c in step 5) (6); harmonic average distance score (H_a) (34); the alignment score (A_s) in step 3. Each one of these five individual scores is transformed into the corresponding Z-score, by relying on the mean μ and standard deviation σ of the score for the assessed models [Z(score) = (score – μ)/σ]. The five Z-scores are then linearly combined to form the composite model assessment score

$$Z = 0.17\ Z(P_p) + 0.02\ Z(P_s) + 0.1\ Z(S_c) + 0.26\ Z(H_a) + 0.45\ Z(A_s) \quad \mathbf{2}$$

The coefficients in this linear combination were obtained by optimizing the performance of the Z-score on the training set of nine template–target pairs described below (B.John and A.Sali, in preparation).

Step 13. Selecting the best model

For a target sequence with a GA341 score of <0.6 in step 6, the best model was selected using the composite score (step 12), whereas for a sequence with a GA341 score >0.6 in step 6, the best model was selected by the alignment score.

Evaluation of the alignment and model accuracy

The accuracy of a model was determined by comparison with the corresponding native structure extracted from the Protein Data Bank (PDB) (35). First, the root mean square deviation (RMSD) between the corresponding Cα atoms in the model and the native structure was calculated upon rigid body least squares superposition of all the Cα atoms, as implemented in the SUPERPOSE command of MODELLER. Second, the percentage of structurally equivalent positions was defined as the percentage of the Cα atoms in the model that are within 5 Å of the corresponding atoms in the superposed native structure ('native overlap'). In addition to the assessment of a model, we evaluated the accuracy of the corresponding alignment through a comparison with the structure-based alignment produced by the CE program (37). The percentage of correctly aligned positions was defined as the percentage of positions in the tested alignment that were identical to those in the CE structure-based alignment ('CE overlap'); residue–gap matches are ignored in this calculation.

The training and testing sets of pairs of related proteins of known structure

We relied on the Fischer set of 68 pairs of remotely related protein structures from 51 to 568 residues in size (36). This set was devised to test fold assignment methods in the most difficult regime of no statistically significant sequence

similarity. According to the structure superpositions by the program CE (37), the sequence identity for these pairs ranges from 2.4 to 31.6%, with an average of 16.3% and a standard deviation of 6.4%. The percentages of the pairs in the α, β, α/β and α + β protein structure classes are 19, 36, 29 and 10%, respectively. Because comparative protein structure modeling requires a degree of structural similarity between the template and target structures, we applied the following two filters to select the target–template pairs for training and testing of our method: (i) target–template pairs had to have >70% of their Cα atoms within 5 Å of each other, upon rigid body least squares superposition using the CE alignment; (ii) target sequences had to have at least 80% of their residues aligned with the template residues in the CE alignment. Application of these two filters to the Fischer set of 68 structure pairs yielded 48 target–template pairs, 29 of which had a GA341 score of >0.6 for the top ranking model in the corresponding initial population of alignments (steps 1 and 2). Nine out of these 29 pairs were randomly selected as the training set to devise the current protocol. The other 20 of the 29 pairs ('difficult' set) and the remaining 19 pairs with a GA341 score of <0.6 ('very difficult' set) were used as the testing sets. For each target sequence, we modeled only the segment spanned by the first and last aligned residues in the CE structure-based alignment.

PSI-BLAST and SAM alignments

For comparison, we also used the difficult and very difficult testing sets to assess alignment accuracy by PSI-BLAST (31,38) and SAM (39). Alignments by PSI-BLAST were obtained by aligning the target and the template profiles (step 2, but using PSI-BLAST 2.2.3 instead of the earlier version 2.0.11) to the template and the target sequences, respectively. Of the two resulting alignments, the alignment with the most significant E-value was used as the final PSI-BLAST alignment. We determined the dependence of the accuracy of the final PSI-BLAST alignments on the number of PSI-BLAST iterations used to construct the profiles in step 2. Using up to 20 iterations results in alignments that have the optimal CE overlap for the very difficult set and are within 2% of the optimal alignments for the difficult set obtained with up to three iterations. Moreover, the average PSI-BLAST alignment accuracy for our PSI-BLAST 2.2.3 protocol with up to 20 iterations (step 2) is within 4% of the average of the highest alignment accuracies obtained by optimizing the number of iterations for each individual target–template pair. Therefore, the PSI-BLAST protocol in step 2 is robust and appropriate for benchmarking PSI-BLAST.

To construct alignments by SAM (version 3.3.1), we applied the following protocol (Rachel Karchin, personal communication). First, the 'w0.5' script in the SAM package was used to build hidden Markov models for the target and template sequences, using their PSI-BLAST 2.2.3 profiles from step 2 above. Next, the program 'hmmscore' in the SAM package was used (sw = 0; select_align = 8; adpstyle = 5) to align the hidden Markov models of the target and the template with the template and the target sequences, respectively, resulting in two generally different template–target alignments. The average CE overlap for these two alignments was defined to be the SAM accuracy for a given template–target pair.

RESULTS

In Methods, we described our iterative alignment, model building and model assessment protocol, the alignment and model accuracy criteria and the training and testing sets of protein structure pairs. In Results, we first validate the performance of the protocol with an ideal fitness function. Next, we quantify the significant improvements in the alignment and model accuracies achieved by our protocol using a realistic fitness function. We also compare the accuracy of our protocol with those of PSI-BLAST and SAM. Finally, we illustrate the method by describing in detail its application to a single protein sequence.

Sampling efficiency of the modeling protocol with an ideal fitness function

As in any other practical problem, it is necessary to attain a useful solution in a reasonable amount of computer time (e.g. 1 day on 100 Intel Pentium III CPUs). Therefore, we first explore the sampling efficiency of our protocol. It is necessary, although not sufficient, for a good protocol to be able to find a substantially correct alignment when an ideal fitness function is used. In particular, we tested the performance of the protocol by assessing models with the CE overlap criterion instead of the statistical potential score (step 10). With such an artificial fitness function, the current implementation of the protocol typically achieves an essentially correct alignment in approximately 30 iterations, both for the best alignment in the population and the population as a whole (Fig. 3).

Accuracy of refined alignments and models for the 'very difficult' testing set

With the assurance that the modeling protocol is at least in principle capable of finding an accurate alignment (Fig. 3), we proceeded with testing the protocol using the realistic fitness function and the 'very difficult' testing set of 19 target–template pairs described in Methods (Table 1). Target–template sequence identities based on the CE structure alignments in this set are <27%. The average sequence identity and coverage (percentage of the modeled residues in the target sequence) are 14 and 85%, respectively.

We contrast the accuracy of the models in steps 3 (profile-based alignments by SALIGN) and 10 (model assessment by a statistical potential score) to assess the utility of the statistical potential score. The Cα RMSD error and native overlap of the highest ranking model in step 3, averaged over the 19 template–target pairs in the test set, are 9.6 Å and 42.8%, respectively (Fig. 4). In comparison, the average accuracy of the highest ranking model is generally higher in step 10. For example, after 25 iterations, the average native overlap increased from 42.8 to 49.2%, while the average Cα RMSD error decreased from 9.6 to 8.7 Å. The differences between the highest ranking model and the average accuracy of the 10 top models in step 10 become small after approximately 15 iterations. The observed increase in accuracy validates our approach to optimizing a model by evolving alignments under the selective pressure of a model assessment score.

To gain insight into the relative importance of the ranking inaccuracies and the incomplete sampling of alignments and models, we compared the accuracy of the highest ranking models in steps 3 and 10 with the best model sampled in steps

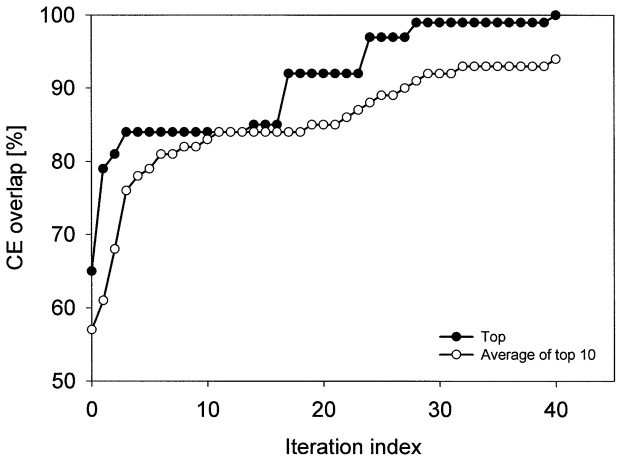

Figure 3. Efficiency of the genetic algorithm protocol relying on an ideal fitness function (Results). CE overlap of the evolving alignments is plotted against the iteration index (step 11 in Fig. 1). The evolution is shown for one of the testing template–target pairs, the 1MOL–1CEW pair; similar results were obtained for the other testing pairs (not shown). Closed circles, the top ranking alignment; open circles, the average of the top 10 alignments.

1–11. In all iterations, the best models are significantly more accurate than the highest ranking models (Fig. 4). For instance, the average native overlaps for the highest ranking model and the best model after 25 iterations are 49.2 and 64.2%, respectively. The difference between these accuracies increases considerably with the number of iterations. The steady increase in the average accuracy of the best sampled models reaches a plateau at approximately 20 iterations. These results indicate that the accuracies of the final alignment and model are severely limited by our ability to correctly rank the models by their accuracy in the absence of knowing their actual structures.

We now describe the total improvement achieved by our entire modeling protocol, which also includes the final model selection based on the composite model criterion in steps 12 and 13. The average native overlap over all 19 template–target pairs for the highest ranking models in step 13 increased from 42.8 to 54.4%, relative to the 'initial comparative model' in step 3 (Table 1). Similarly, the Cα RMSD error and CE overlap improved from 9.6 to 7.7 Å and from 36.7 to 44.8%, respectively. Moreover, nine of the 19 test cases were modeled with <6 Å Cα RMSD error. Improvements in the accuracies of the alignments and models were found to be statistically significant at the 95% confidence level using Student's *t*-test (40). Despite our partial heuristic sampling of the alignment space, typically involving less than 8000 unique alignments, the optimization protocol was generally able to significantly improve the input alignments, even when the sequence identity between the template and target sequences was <20%.

Accuracies of the refined alignments and models for the 'difficult' testing set

The current optimization protocol does not refine the initial alignments for the target–template pairs in the 'difficult' testing set, because they all have an initial GA341 score >0.6 (step 6). These target–template pairs are more similar to each other than those in the 'very difficult' set. The average sequence identity and coverage for this testing set are 20.1 and 88.8%, respectively.

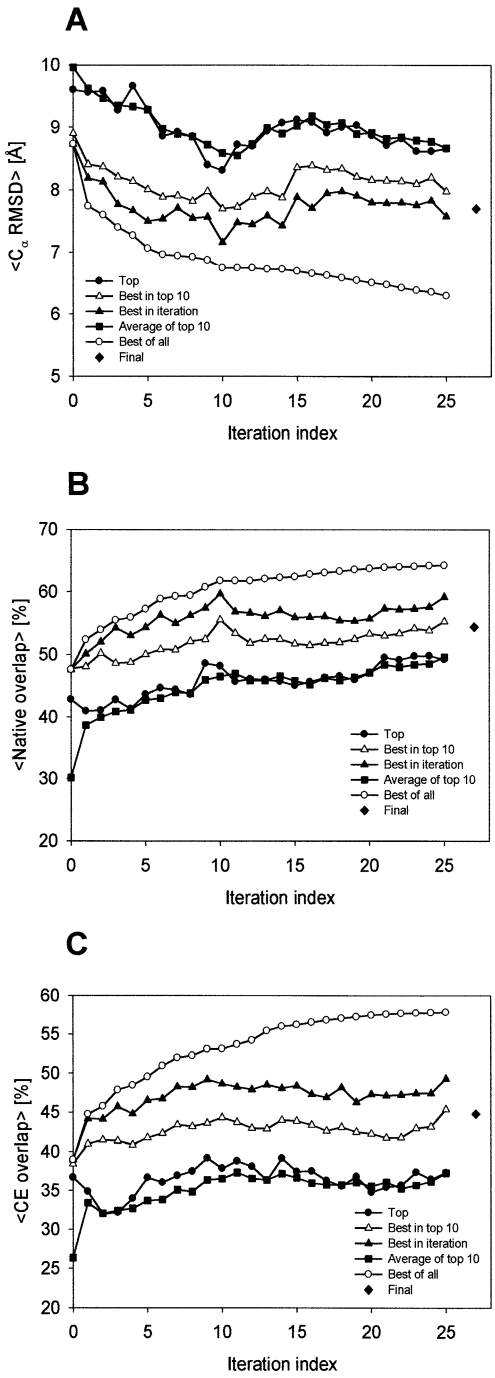

Figure 4. Accuracy of the genetic algorithm protocol as a function of the optimization progress. All three panels show the averages over the 19 target–template pairs in the 'difficult' testing set. The model accuracy is measured both by the Cα RMSD of a model from the native structure (**A**) and by the native overlap (**B**). The alignment accuracy is measured by the CE overlap (**C**). Closed circles, the highest ranking model in step 10; closed triangles, the most accurate model in step 9; closed squares, the average of the 10 highest ranking models in step 10; open triangles, the most accurate model among the 10 highest ranking models in step 10; open circles, the most accurate model generated in any of the steps 1–11 up to the current iteration index; diamonds, the final model (step 13) selected using the composite score in step 12.

Table 1. Accuracies of alignments and models for the 'very difficult' testing set of 19 target–template pairs

Template–target pair[a]	Sid[b]	Cov[c]	Initial prediction			Initial best			Final prediction			Population best			PB	SM
			NO[d]	CR[e]	CE[f]	NO[d]	CR[e]	CE[f]	NO[d]	CR[e]	CE[f]	NO[d]	CR[e]	CE[f]	CE[f]	CE[f]
1ATR(3-382)-1ATN(4:A-354:A)	13.8	94.3	13.1	19.2	20.2	16.8	17.1	24.6	11.7	18.8	20.2	17.7	17.1	24.6	0.0	29.6
1BOV(2:A-69:A)-1LTS(17:D-102:D)	4.4	83.5	44.2	10.1	29.4	51.2	10.1	29.4	84.9	3.6	79.4	91.9	3.1	92.6	0.0	27.2
1CAU(48:A-224:A)-1CAU(246:B-423:B)	18.8	96.7	18.5	11.7	15.6	18.5	11.7	15.6	29.8	10.0	27.4	43.3	7.6	47.4	2.5	69.4
1COL(71:A-187:A)-1CPC(29:L-170:L)	11.2	81.4	37.1	8.6	44.0	37.1	8.6	44.0	53.6	5.6	58.6	70.0	4.8	59.3	0.0	15.1
1LFB(15-86)-1HOM(7-57)	17.6	75.0	100.0	1.2	100.0	100.0	1.2	100.0	100.0	1.2	100.0	100.0	1.1	100.0	27.5	80.4
1NSB(113:A-449:A)-2SIM(35-374)	10.1	89.2	22.4	13.2	20.2	25.3	12.8	26.8	23.5	13.2	20.1	26.2	12.3	26.8	0.0	6.5
1RNH(4-143)-1HRH(437:A-555:A)	26.6	91.2	10.5	13.0	21.2	10.5	13.0	21.2	69.3	4.8	35.4	87.7	3.5	57.5	63.7	57.0
1YCC(2-103)-2MTA(45:C-125:C)	14.5	55.1	81.5	3.4	72.4	81.5	3.4	72.4	71.6	5.3	58.4	87.7	3.1	75.0	51.3	47.3
2AYH(43-214)-1SAC(4:A-163:A)	8.8	78.4	55.0	5.8	33.8	66.2	5.2	42.6	66.2	5.5	48.0	79.4	4.8	64.9	12.2	11.1
2CCY(5:A-128:A)-1BBH(5:A-131:A)	21.3	97.0	77.2	4.1	52.4	77.2	4.1	52.4	88.2	3.1	73.0	96.1	2.6	77.0	68.9	66.4
2PLV(43-275)-1BBT(1-192)	20.2	91.4	61.2	7.3	58.9	61.2	7.3	58.9	62.9	7.3	58.9	66.5	6.2	60.7	55.2	45.4
2POR(1-301)-2OMF(10-340)	13.2	97.3	15.1	18.3	11.3	15.7	16.7	11.3	28.7	11.4	14.7	31.7	10.5	25.9	33.1	21.4
2RHE(6-112)-1CID(1-109)	12.2	61.6	42.2	9.2	33.7	42.2	8.8	33.7	54.1	7.5	51.1	61.5	4.4	71.1	0.0	43.4
2RHE(3-108)-3HLA(4:B-98:B)	2.4	96.0	43.2	8.1	16.5	43.2	8.1	16.5	47.4	7.6	9.4	56.8	6.7	43.5	0.0	10.6
3ADK(8-194)-1GKY(1-186)	19.5	100.0	18.3	13.8	26.6	39.8	9.3	33.8	36.6	11.5	37.7	53.2	7.7	48.1	68.8	44.9
3HHR(131:B-233:B)-1TEN(803-891)	18.4	98.9	65.2	7.3	60.9	65.2	6.7	62.1	70.8	6.0	66.7	80.9	4.9	79.3	71.3	59.8
4FGF(20-143)-8I1B(7-151)	14.1	98.6	29.9	11.3	24.0	52.1	6.2	31.4	41.0	9.3	30.6	52.8	5.4	41.2	24.0	11.9
6XIA(8-239)-3RUB(235:L-429:L)	8.7	44.1	28.2	10.5	14.5	28.2	10.4	14.5	23.1	10.1	11.0	39.0	9.0	34.3	0.0	4.4
9RNT(2-104)-2SAR(7:A-91:A)	13.1	88.5	50.6	5.8	41.7	71.8	5.5	48.8	70.6	5.1	51.2	76.5	4.8	69.0	0.0	25.0
Average	14.2	85.2	42.8	9.6	36.7	47.6	8.7	38.9	54.4	7.7	44.8	64.2	6.3	57.8	25.2	35.6

Initial prediction, the model based on the SALIGN alignment with the optimal gap parameters in step 3; initial best, the most accurate model among those based on the 15 initial parent alignments in step 3; final prediction, the final model selected by the composite model accuracy score in step 13; population best, the most accurate model generated through steps 1–11. PB and SM, PSI-BLAST and SAM alignment accuracy, respectively (Methods).

[a]The segments from the PDB files are indicated by the beginning and ending residue numbers.

[b]Percentage sequence identity based on the CE alignment.

[c]Coverage, measured as the percentage of the number of residues modeled in the complete target sequence.

[d]Native overlap.

[e]C RMSD between the model and the native structure.

[f]CE overlap.

The average Cα RMSD, native overlap and CE overlap of the final models (step 13) of the 20 target sequences, obtained without refinement of their initial alignments, are 6.3 Å, 69.4% and 67.4%, respectively (Table 2). Out of the 20 target sequences, 13 were modeled with <6 Å Cα RMSD error and 15 were modeled with >60% native overlap. Five of the latter 15 targets share <20% sequence identity to their templates. Ten of the 11 target sequences with >20% sequence identity to their templates are predicted with a relatively high accuracy of >60% native overlap.

To investigate the potential gain in the prediction accuracy if a perfect model assessment score were available for the final selection of the best model (steps 12, 13), the sampling protocol for alignment refinement was applied to the 20 target sequences. With perfect model ranking, the average Cα RMSD error would improve from 6.3 to 4.3 Å. Similarly, the average native and CE overlaps would increase from 69.4 to 83.1% and from 67.4 to 78.4%, respectively.

Comparison with the PSI-BLAST and SAM alignments

For direct comparison with our protocol, we assessed the alignment accuracy of PSI-BLAST and SAM for the very difficult (Table 1) and difficult testing sets (Table 2). The average CE overlaps (Tables 1 and 2) of our iterative genetic algorithm protocol are higher than those produced by SAM and PSI-BLAST: 44.8 versus 35.6 and 25.2%, and 67.4 versus 62.9 and 58.8% for the very difficult and difficult testing sets, respectively. Taken together, these improvements in the alignment accuracies are statistically significant at the 95% confidence level using Student's *t*-test (40). Thirteen of the 19 targets in the very difficult set were aligned more accurately by the iterative genetic algorithm protocol than by SAM. Similarly, 15 targets in the very difficult set were aligned more accurately by the genetic algorithm protocol than by PSI-BLAST. Moreover, 11 and 15 of the 20 targets in the difficult set were aligned more accurately by the iterative genetic algorithm protocol than by SAM and PSI-BLAST, respectively.

Similar conclusions are obtained when using the native overlap and Cα RMSD accuracy measures instead of CE overlap (data not shown). This observation is not surprising because of strong correlations among the three accuracy measures.

Sample application to a 'very difficult' modeling case

In the previous sections, we benchmarked the optimization protocol using average improvements for two different testing sets. Here, we illustrate the protocol in detail by describing its application to a single target–template case of 1LTS–1BOV. This example was chosen because it has both the lowest sequence identity (4.4%) and the largest gain in the alignment accuracy (50% gain in the CE overlap between steps 3 and 13) among all of the template–target pairs in the very difficult testing set (Table 1). We contrast the accuracies of the predicted model (step 13) and the initial model based on the SALIGN alignment (step 3). Refinement of the initial parent alignments (step 3), which would have been useless in most applications of comparative modeling, resulted in a reasonably accurate prediction (Table 1). The Cα RMSD decreased from 10.1 Å for the initial comparative model to 3.6 Å for the final refined model. Correspondingly, the native and CE overlaps

Table 2. Accuracies of alignments and models for the 'difficult' testing set of 20 target–template pairs

Template–target pair[a]	Sid[b]	Cov[c]	Initial prediction			Initial best			Population best			PB	SM
			NO[d]	CR[e]	CE[f]	NO[d]	CR[e]	CE[f]	NO[d]	CR[e]	CE[f]	CE[f]	CE[f]
1EGO(1-84)-1ABA(1-87)	16.9	100.0	63.2	5.6	35.1	87.4	4.5	53.2	87.4	4.1	53.2	24.7	22.7
1FRN(35-314)-2PIA(6-225)	14.5	68.5	73.6	5.4	64.4	80.0	4.1	69.2	88.2	3.2	74.3	60.3	50.4
1HBG(2-145)-1DXT(4:B-146:B)	22.4	97.3	88.8	3.3	84.3	91.6	2.6	90.3	93.7	2.5	91.0	84.3	84.7
1MOL(8:A-94:A)-1STF(17:I-119:I)	10.3	85.3	37.0	12.2	30.8	67.9	4.6	53.8	98.8	2.9	72.5	0.0	46.2
1PAZ(3-93)-1AAJ(21-105)	27.5	81.0	84.7	3.4	70.0	84.7	3.4	70.0	95.3	2.1	91.2	63.7	68.8
1YCC(1-102)-1C2R(1:A-115:A)	31.6	99.1	75.7	5.1	83.7	75.7	5.1	83.7	81.7	4.0	88.8	88.8	86.2
2CPP(40-414)-2HPD(27:A-453:A)	15.6	93.4	72.8	6.1	79.2	73.8	6.1	79.2	76.6	5.2	82.5	70.8	47.3
2FB4(1:H-219:H)-1FC1(301:A-444:A)	22.5	69.6	59.7	8.5	53.6	59.7	8.4	53.6	72.9	6.1	73.9	58.0	58.7
2HIP(2:A-71:A)-1HIP(5-85)	23.5	95.3	71.6	4.6	57.4	71.6	4.6	57.4	87.7	3.3	79.4	52.9	82.3
2HIP(7:A-71:A)-1ISU(7:A-62:A)	20.4	90.3	96.4	2.4	66.7	98.2	2.2	75.9	100.0	1.8	94.4	14.8	78.7
2MNR(4-359)-1CHR(1:A-368:A)	17.9	99.5	83.7	4.5	71.1	83.7	4.5	71.7	89.1	3.2	76.9	69.7	78.3
2RHE(6-112)-1TLK(47-135)	20.4	86.4	88.8	3.0	72.7	88.8	3.0	72.7	98.9	2.3	83.1	52.3	48.9
2RHE(8-112)-3CD4(2-100)	21.7	100.0	60.6	7.6	60.9	60.6	7.5	60.9	64.7	6.5	62.0	34.8	58.1
2SCP(1:A-172:A)-2SAS(3-183)	16.5	97.8	75.7	4.3	85.3	75.7	4.3	85.3	82.3	3.9	87.6	76.5	84.1
3GRS(21-478)-1NPX(1-438)	15.9	98.0	30.6	13.4	46.5	54.6	8.6	51.7	59.4	8.1	53.8	63.4	55.4
3HLA(1:B-99:B)-1PFC(342-444)	22.7	91.0	79.2	3.9	89.7	80.2	3.8	89.7	89.1	3.3	93.8	77.3	66.5
3MIN(61:B-504:B)-1MIO(44:C-510:C)	17.2	89.0	21.0	15.1	78.5	27.8	13.5	78.5	48.0	10.4	78.7	71.9	49.9
4CPV(37-108)-1OSA(79-147)	30.4	46.6	97.1	2.4	97.1	97.1	2.4	97.1	98.5	2.1	97.1	95.7	95.0
4ENL(1-399)-2MNR(5-320)	12.6	88.5	50.3	10.8	50.6	57.9	7.7	53.2	67.1	7.0	61.3	43.9	28.8
6LDH(21-328)-2CMD(1-310)	21.7	99.4	78.1	4.4	70.8	78.4	4.4	71.9	82.6	4.2	71.9	71.2	66.5
Average	20.1	88.8	69.4	6.3	67.4	74.8	5.3	71.0	83.1	4.3	78.4	58.8	62.9

See Table 1 for a description of the columns. There are no final prediction columns here because the final prediction corresponds to the initial prediction when the GA341 score of the initial prediction is >0.6.

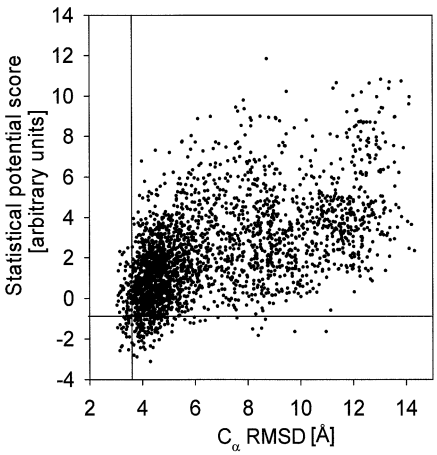

Figure 5. The statistical potential score of a model of 1LTS based on 1BOV as a function of its Cα RMSD error. The model at the crossing of the vertical and horizontal lines corresponds to the best model according to the composite model accuracy criterion.

increased from 44.2 to 84.9% and from 29.4 to 79.4%, respectively.

Although evolution of alignments guided by the statistical potential scores of the models improved the accuracy of the alignments and the models, final model selection based on this model score alone would have resulted in a Cα RMSD error of 4.2 Å (Fig. 5). In contrast, the composite model score in step 12 selected a more accurate model with a Cα RMSD error of 3.6 Å. If perfect ranking of the models were available, the model with a Cα RMSD error of 3.1 Å would have been selected.

Improvement in the model and alignment accuracies is gradual and not attained in a single step of optimization (Fig. 6A–C). It is matched by a significant improvement in the model assessment score over the course of the whole optimization (Fig. 6D). These observations demonstrate the usefulness of 'evolution' in the genetic algorithm protocol and justify the number of optimization steps applied.

DISCUSSION

We have described an automated protocol that refines an initial alignment between a given sequence and structure and increases the accuracy of the corresponding comparative model. These improvements are achieved by a genetic algorithm protocol that starts with a set of initial alignments and then iterates through re-alignment, model building and model assessment to optimize a model assessment score. During this iterative process: (i) new alignments are constructed by application of a number of operators, such as alignment mutations and crossovers; (ii) comparative models corresponding to these alignments are built by satisfaction of spatial restraints, as implemented in our program MODELLER; (iii) the models are assessed by a variety of criteria, partly depending on an atomic statistical potential.

The new method was benchmarked on a test set of 39 comparative modeling cases, divided into the difficult and very difficult testing sets (Tables 1 and 2). When testing the procedure on a very difficult set of 19 modeling targets sharing only 4–27% sequence identity with their template structures, the average final alignment accuracy increased from 37 to 45% relative to the initial alignment (the alignment accuracy was measured as the percentage of positions in the tested alignment that were identical to the reference structure-based alignment). Correspondingly, the average model accuracy increased from 43 to 54% (the model accuracy was measured as the percentage of the Cα atoms of the model that were

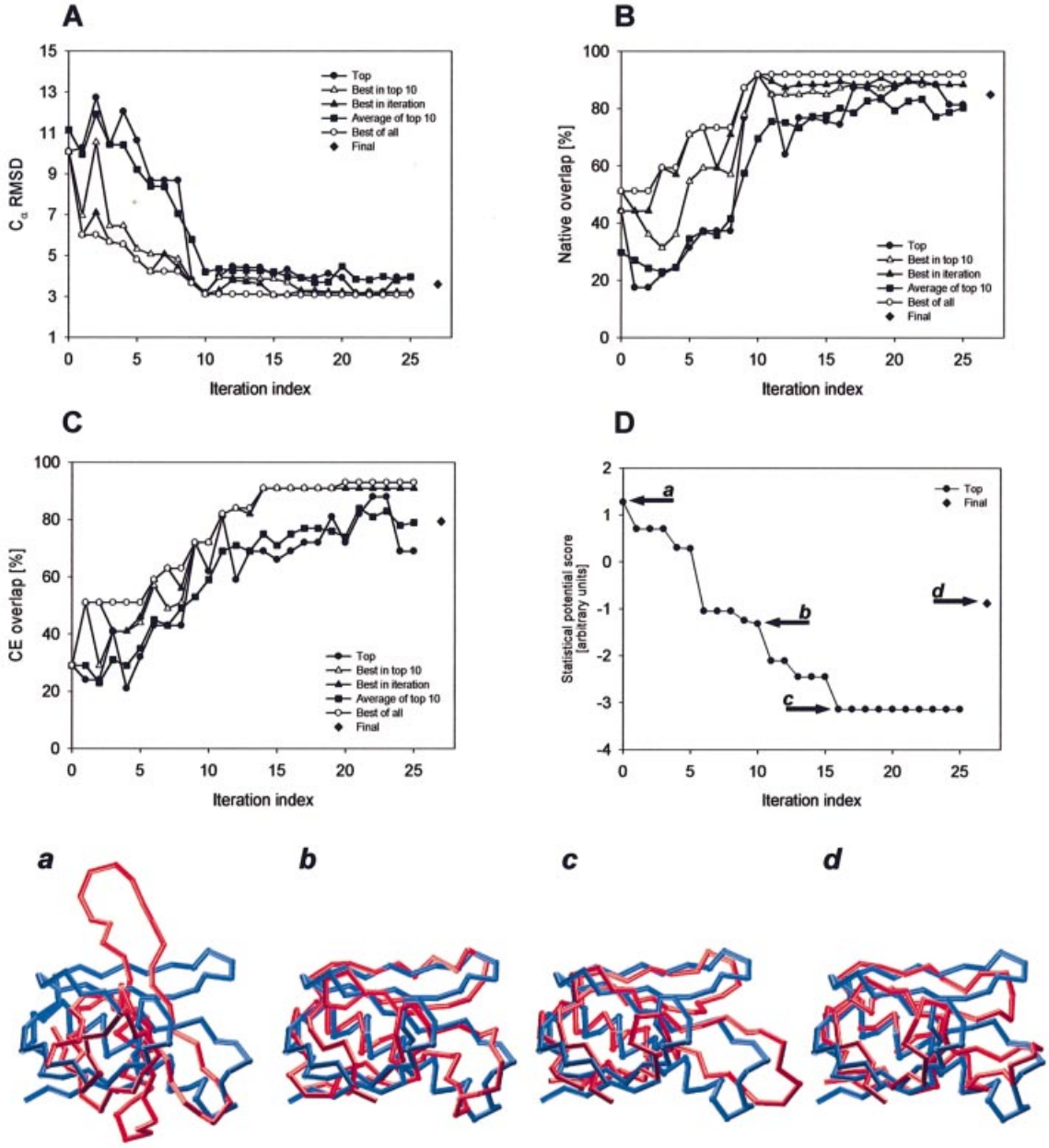

Figure 6. Accuracies of the 1LTS alignment and model as a function of the optimization progress. (**A**–**C**) See Figure 4 for a description of the symbols. (**D**) Statistical potential score. The bottom panels (**a**–**c**) show models (red) of representative iterations superposed on the native structure (blue). (**d**) The final model superposed on the native structure. The Cα RMSD errors for these models are 10.1 (a), 3.8 (b), 4.3 (c) and 3.6 Å (d).

within 5 Å of the corresponding Cα atoms in the superposed native structure).

The accuracy of any new prediction method has to be compared with previous results. There are a great many existing alignment methods and it is not practical to consider all of them. Thus, we chose to compare the present results with only two, but carefully selected, previous studies. We chose the PSI-BLAST (31) and SAM (39) programs because they are easily available, widely used, well documented, automated and accurate (41–43). PSI-BLAST depends on sequence profiles and SAM on hidden Markov models. Their alignment accuracies have been benchmarked previously (41,42). The average CE overlaps (Tables 1 and 2) of our iterative genetic algorithm protocol are higher than those produced by SAM and PSI-BLAST: 44.8 versus 35.6 and 25.2%, and 67.4 versus 62.9 and 58.8% for the very difficult and difficult testing sets, respectively.

Despite errors, even comparative protein structure models based on alignments with only 30% CE overlap can be useful in biology. For example, such models may be sufficiently accurate in some of their parts to allow for interpreting site-directed mutagenesis experiments (44), constructing macromolecular assemblies (45), identifying catalytic residues (46), refining NMR structures (47) and fitting into low resolution electron density maps (48). These models are also useful for assessing the fold of the target sequence (8).

Seven of the 12 targets in the very difficult set (19 targets in total) that had a CE overlap of <30% for the PSI-BLAST alignments were refined to achieve CE overlaps of >30% in the final step (step 13). Thus, ~58% of the targets that were inaccurately modeled based on the PSI-BLAST alignments can be considered modeled with useful accuracy by our protocol (>30% CE overlap). Furthermore, CE overlap of 15 of the 19 targets (80%) in the very difficult set was improved by the current protocol with respect to the PSI-BLAST-based alignments.

These benchmarks indicate that our new protocol is useful for refining inaccurate models in MODBASE (6), a comprehensive database of annotated comparative models for all known protein sequences (49) that are detectably related to at least one known protein structure. Currently, 40% (~167 000) of the models in MODBASE are based on PSI-BLAST alignments with <20% sequence identity to the closest template. An extrapolation from the results presented here indicates that the current protocol might in principle be able to produce useful models for 97 000 additional protein sequences (58% of 167 000). In practice, refining 167 000 models in MODBASE by the current protocol is not feasible because it would require too much computer time. However, a smaller subset of biologically important sequences could be selected for refinement by this protocol.

The genetic algorithm protocol for refining a given sequence–structure alignment can already be selected as an option in MODWEB, a web server for automated comparative modeling that relies on MODPIPE (8), which in turn depends on PSI-BLAST (31), IMPALA (50) and MODELLER (3). However, due to the high demand for CPU time (e.g. 1 day on 100 Intel Pentium III CPUs for one alignment refinement), web access is currently restricted to a small number of selected users. In addition to our current benchmarks, the protocol will also be evaluated by the EVA web server for automated and continuous assessment of protein structure prediction methods (51).

The iterative alignment, modeling and model assessment protocol is currently limited at least by the errors in the assessment of model accuracy. A comparison of the accuracies of the best model generated in steps 1–11 and the final model selected in step 13 (Table 1) reveals that there is considerable scope for further improvement by using more accurate model ranking in steps 3, 10 and 13. Our results indicate that a statistical potential score used in conjunction with other model assessment scores can lead to more accurate predictions than those obtained based on the individual scores (Table 1 and Figs 4–6). While our composite model score is frequently able to select nearly the best model from a large ensemble of structures, it is generally unable to choose the best model (Figs 4–6). For instance, if model ranking were to select the best generated model (steps 1–11), the native overlap would increase from 42.8% in step 3 to 64.2% in step 13 for the very difficult test set. Similar improvements would also be observed for CE overlap (36.7 to 57.8%) and Cα RMSD (9.6 to 6.3 Å). A more accurate scoring scheme would not only afford selection of the most accurate generated model, but would also bias the search towards the more relevant parts of the alignment space and therefore further increase the accuracy and efficiency of the whole protocol.

REFERENCES

1. Cantor,C.R. and Little,D.P. (1998) Massive attack on high-throughput biology. *Nature Genet.*, **20**, 5–6.
2. Grunenfelder,B. and Winzeler,E.A. (2002) Treasures and traps in genome-wide data sets: case examples from yeast. *Nature Rev. Genet.*, **3**, 653–661.
3. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
4. Marti-Renom,M.A., Stuart,A.C., Fiser,A., Sanchez,R., Melo,F. and Sali,A. (2000) Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.
5. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
6. Pieper,U., Eswar,N., Stuart,A.C., Ilyin,V.A. and Sali,A. (2002) MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.*, **30**, 255–259.
7. Rost,B. (1998) Marrying structure and genomics. *Structure*, **6**, 259–263.
8. Sanchez,R. and Sali,A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
9. Venclovas,C., Zemla,A., Fidelis,K. and Moult,J. (2001) Comparison of performance in successive CASP experiments. *Proteins*, **5** (suppl.), 163–170.
10. Saqi,M.A., Bates,P.A. and Sternberg,M.J. (1992) Towards an automatic method of predicting protein structure by homology: an evaluation of suboptimal sequence alignments. *Protein Eng.*, **5**, 305–311.
11. Pawlowski,K., Bierzynski,A. and Godzik,A. (1996) Structural diversity in a family of homologous proteins. *J. Mol. Biol.*, **258**, 349–366.
12. Guenther,B., Onrust,R., Sali,A., O'Donnell,M. and Kuriyan,J. (1997) Crystal structure of the delta′ subunit of the clamp-loader complex of *E. coli* DNA polymerase III. *Cell*, **91**, 335–345.
13. Sanchez,R. and Sali,A. (1997) Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.*, **7**, 206–214.
14. Sanchez,R. and Sali,A. (2000) Comparative protein structure modeling. Introduction and practical examples with modeller. *Methods Mol. Biol.*, **143**, 97–129.
15. Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.
16. Melo,F., Sanchez,R. and Sali,A. (2002) Statistical potentials for fold assessment. *Protein Sci.*, **11**, 430–448.
17. Koehl,P. and Levitt,M. (2002) Improved recognition of native-like protein structures using a family of designed sequences. *Proc. Natl Acad. Sci. USA*, **99**, 691–696.
18. Waterman,M.S. (2000) *Introduction to Computational Biology*. Chapman & Hall/CRC, USA.
19. Goldberg,D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.
20. Pedersen,J.T. and Moult,J. (1997) Protein folding simulations with genetic algorithms and a detailed molecular description. *J. Mol. Biol.*, **269**, 240–259.
21. Sun,Z., Xia,X., Guo,Q. and Xu,D. (1999) Protein structure prediction in a 210-type lattice model: parameter optimization in the genetic algorithm using orthogonal array. *J. Protein Chem.*, **18**, 39–46.
22. Gardiner,E.J., Willett,P. and Artymiuk,P.J. (2001) Protein docking using a genetic algorithm. *Proteins*, **44**, 44–56.
23. Pegg,S.C., Haresco,J.J. and Kuntz,I.D. (2001) A genetic algorithm for structure-based *de novo* design. *J. Comput. Aided Mol. Des.*, **15**, 911–933.
24. Notredame,C. and Higgins,D.G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.*, **24**, 1515–1524.
25. Zhang,C. and Wong,A.K. (1997) A genetic algorithm for multiple molecular sequence alignment. *Comput. Appl. Biosci.*, **13**, 565–581.
26. Szustakowski,J.D. and Weng,Z. (2000) Protein structure alignment using a genetic algorithm. *Proteins*, **38**, 428–440.
27. Lemmon,A.R. and Milinkovitch,M.C. (2002) The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation. *Proc. Natl Acad. Sci. USA*, **99**, 10516–10521.
28. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.
29. Moult,J. (1996) The current state of the art in protein structure prediction. *Curr. Opin. Biotechnol.*, **7**, 422–427.
30. Forster,M.J. (2002) Molecular modelling in structural biology. *Micron*, **33**, 365–384.

31. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
32. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
33. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
34. Pazos,F., Helmer-Citterich,M., Ausiello,G. and Valencia,A. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, **271**, 511–523.
35. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
36. Fischer,D., Elofsson,A., Rice,D. and Eisenberg,D. (1996) Assessing the performance of fold recognition methods by means of a comprehensive benchmark. *Pac. Symp. Biocomput.*, 300–318.
37. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
38. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
39. Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
40. Marti-Renom,M.A., Madhusudhan,M.S., Fiser,A., Rost,B. and Sali,A. (2002) Reliability of assessment of protein structure prediction methods. *Structure*, **10**, 435–440.
41. Friedberg,I., Kaplan,T. and Margalit,H. (2000) Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Sci.*, **9**, 2278–2284.
42. Elofsson,A. (2002) A study on protein sequence alignment quality. *Proteins*, **46**, 330–339.
43. Edgar,R.C. and Sjolander,K. (2003) Simultaneous sequence alignment and tree construction using hidden Markov models. *Pac. Symp. Biocomput.*, 180–191.
44. Matsumoto,R., Sali,A., Ghildyal,N., Karplus,M. and Stevens,R.L. (1995) 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.*, **270**, 19524–19531.
45. Spahn,C.M., Beckmann,R., Eswar,N., Penczek,P.A., Sali,A., Blobel,G. and Frank,J. (2001) Structure of the 80S ribosome from *Saccharomyces cerevisiae*—tRNA-ribosome and subunit-subunit interactions. *Cell*, **107**, 373–386.
46. Navaratnam,N., Fujino,T., Bayliss,J., Jarmuz,A., How,A., Richardson,N., Somasekaram,A., Bhattacharya,S., Carter,C. and Scott,J. (1998) *Escherichia coli* cytidine deaminase provides a molecular model for ApoB RNA editing and a mechanism for RNA substrate recognition. *J. Mol. Biol.*, **275**, 695–714.
47. Nagata,T., Gupta,V., Sorce,D., Kim,W.Y., Sali,A., Chait,B.T., Shigesada,K., Ito,Y. and Werner,M.H. (1999) Immunoglobulin motif DNA recognition and heterodimerization of the PEBP2/CBF Runt domain. *Nature Struct. Biol.*, **6**, 615–619.
48. Palaniyar,N., McCormack,F.X., Possmayer,F. and Harauz,G. (2000) Three-dimensional structure of rat surfactant protein A trimers in association with phospholipid monolayers. *Biochemistry*, **39**, 6310–6316.
49. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
50. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
51. Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Fiser,A., Pazos,F., Valencia,A., Sali,A. and Rost,B. (2001) EVA: continuous automatic evaluation of protein structure prediction servers. *Bioinformatics*, **17**, 1242–1243.