Novel Use of a Genetic Algorithm for Protein Structure Prediction: Searching Template and Sequence Alignment Space

Bruno Contreras-Moreira, Paul W. Fitzjohn, Marc Offman, Graham R. Smith, and Paul A. Bates*
Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, Lincoln's Inn Fields Laboratories, London, United Kingdom

ABSTRACT **A novel genetic algorithm was applied to all CASP5 targets. The algorithm simultaneously searches template and alignment space. Results show that the current implementation of the method is perhaps most useful in recognizing and refining remote homology targets. This new method is briefly described and results are analyzed. Strengths and weaknesses of the current implementation of the algorithm are discussed. Proteins 2003;53:424–429.** © 2003 Wiley-Liss, Inc.

Key words: fold recognition; comparative modeling; genetic algorithms; template selection; alignment errors

INTRODUCTION

Comparative Modeling (CM) and Fold Recognition (FR) methods rely on finding one or more Protein Data Bank (PDB) structures, used as templates, whose structural similarity to the query is significant. The underlying assumption is that similar amino acid sequences have the same fold, as supported by empirical observation, so the folding process can be overlooked. An important property of these methods is that errors incorporated into the models are a function of the differences in sequence between query and template(s).[1] Evaluation experiments, such as EVA,[2] show that if the sequence identity ranges from 35% to 100%, models show average deviation values to the experimental structures in the interval [0.5 Å,6 Å]; if the sequence identity is less, the upper limit of the interval reaches values of 15 to 20 Å.[3]

In previous CASPs, the FR category was defined as the set of targets whose modeling templates could not be found with PSI-BLAST,[4] being labeled as CM otherwise. Apart from this difference, the rest of the modeling procedure for these two categories is essentially the same. Indeed, the assessors for these categories in CASP4 pointed out that template selection and sequence alignment errors remained the main (eternal) problems affecting the quality of models.[5,6] For these reasons, we decided to use the same tools and strategies for all CASP5 targets. In our hands, FR and CM are the same problem; only the sequence similarities involved are of different magnitude.

In building models using CM and FR techniques, template selection and sequence alignment must be optimized. We assume that a combination of alignment methods should be better than any individual method and that there is currently no way to confidently identify the best template and, therefore, several templates should be used and combined. Together with the fact that proteins are linear molecules, these features suggested to us that genetic algorithms, widely used in many different protein structure applications,[7–11] could be used to solve these two problems. In fact, the approach used here tackles both problems simultaneously. The idea is that different templates for a given target are just different possible structures for the same sequence. All templates are assumed to be homologous proteins, synthesized from homologous genes, that can undergo genetic recombination or mutation. Recombination is the natural way to exchange linear DNA segments between homologous chromosomes; in this context, it is a mechanism to swap protein fragments around a crossover point. In three dimensions, mutation is needed to create novel conformations, allowing models to have segments different from any of the templates used.

Because a model can be considered as an alignment in three dimensions, models for alternative alignments to the same template can be used. This simple principle was implemented and applied to CASP5 targets as a way to select the optimal templates and the best alignments in a single step. In this artificial evolutionary process, fitness for each model within a population is defined by its folding free energy. Resulting populations of models are optimized in terms of our free energy measure. However, because the correlation between structural agreement and our free energy function is not perfect, this does not guarantee optimized deviations between model and experimental structure.

MATERIALS AND METHODS

Because it is a computational implementation of genetic recombination at the protein level, the procedure used here is named in silico protein recombination. For each CASP 5 target, a population of models was generated from a variety of templates, sequence alignments, and methods. The algorithm can be outlined as follows:

*Correspondence to: Paul A. Bates, Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, Lincoln's Inn Fields Laboratories, 44 Lincoln's Inn Fields, WC2A 3PX London, UK. E-mail: Paul.Bates@cancer.org.uk

Received 4 February 2003; Accepted 22 April 2003

```
initial population of models
↓
(1) grow population: $r$ recombination + (1-$r$) mutation
↓
(2) select best proportion according to fitness
↓
(3) converged? stop: otherwise back to (1)
```

This is a genetic algorithm with two genetic operators (recombination and mutation) and a fitness function acting as an artificial selection agent. Recombination and mutation events are mutually exclusive, occurring with frequencies *r* and 1-*r*. We now briefly describe each step in the protocol. Because of space limitations, a more detailed description has been published elsewhere.[12]

Initial Population of Models

Initially, the Web server DomainFishing[13] was used to define protein domains within each target sequence and to find suitable modeling templates. Resulting alignments (based on sequence profiles and predicted secondary structure agreement) were inspected and corrected if suspected to be incorrect. When found, different alignments to the same template were added to the pool. In several cases (i.e., T0130) annotations from the templates or their corresponding PFAM[14] families were used to check the correctness of the alignment in active/binding sites. In cases in which DomainFishing returned no templates, alignments were generated by using a profile–profile search against a nonredundant PDB[15] library (here profiles are position-specific scoring matrices—pssm—as calculated by PSI-BLAST). Up to seven different alignments for each library member are calculated by using different similarity matrices (pssm1, pssm2, pssm1 and pssm2) and secondary structure predictions. Models from these alignments were built by using the Web server 3D-JIGSAW[16] and the interactive mode to edit the alignments. To gain extra variability in sequence alignments, templates, and alternate loop conformations, models were also taken from different CAFASP 3 servers that return full atomic coordinates. These were FAMS,[17] EsyPred,[18] Arby, Alax, Robetta,[19] and Pmodeller.[20] In cases in which the fold of the target was not clear, models built by using the most popular templates from the most popular SCOP[21] superfamilies were preferred.

Models were inspected and missing parts, typically loops, were added by using in-house software before going to the next step. In essence, this software explores phi/psi space to allow a peptide (the missing loop) to connect a gap in a protein fold. Models were often energy minimized at this stage to smooth their phi/psi geometry and to permit unbiased energy calculations at later stages.

Growing the Population by Recombination and Mutation

Initial populations were grown by randomly selecting pairs of protein models and applying one of the two possible operators. In recombination (with frequency r = 0.95), the models were superimposed on the basis of their sequence alignment and a crossover point drawn. Crossover was not permitted inside secondary structure elements. Resulting recombinant models inherit the N-terminus from one parent and the C-terminus from the other. In mutation events, a new protein model was obtained by simply averaging its parents' coordinates after superposition. Sometimes this process resulted in distorted side-chain conformations, but no attempt was made to correct them in the current implementation. These distorted mutants are filtered out at the next stage, because their free energy estimates tend to be highly disadvantageous. A more rational mutation operator would be needed for further development of the algorithm.

Selecting the Best Proportion—the Fitness Function

The justification for this algorithm is that it should be possible to obtain optimized mosaic models by shuffling them in a rational way, but this requires an appropriate fitness function to evaluate and combine models. After benchmarking, the fitness function was selected to be a free energy estimate based on two terms: protein contact pair-potentials and side-chain solvation energies, estimated from their solvent accessible area. This function E(p) = contacts(p) + solvation(p) seems to provide a consistent measure of protein structural quality while keeping the calculation time within practical limits.

Protein contact pair-potentials are calculated by using a simplified residue representation and summing the all-against-all energies according to a soft Lennard-Jones type potential as published by Robson and Osguthorpe.[22] Solvation terms are calculated as the sum of side-chain solvent-exposed area multiplied by tabulated residue solvation free energies.[23]

When a population reaches the upper limit (between 2 and 4 times its initial size, 30–200 models in our simulations), members are ranked according to their fitness. To ensure that quality models are not lost prematurely, only the worst 25% of the population is discarded at this stage.

Convergence Criterion and Final Refinements

When all members of the population have converged to a similar energy, there is no room for further generation of variability and the evolution process stops. In most cases, this final population consists of several representatives of the same protein conformation with average backbone deviations in the order of 0.1 Å, but sometimes alternative conformations can be obtained.

One of these representatives is then taken as the final model, which is carefully inspected to detect unfavorable peptide conformations, and a final energy minimization using the CHARMM22[24] force field is performed. This procedure is able to fix distorted side-chains generated by mutation. At this point, we have a CASP5 unrefined model.

For targets T0134, T0165, T0177, and T0185, a further refinement step was performed. This consisted of running an all-atom, molecular dynamics (MD) simulation inside a water box, with neutral total charge for around 0.5 ns. For

TABLE I. Performance of Protein Recombination in the CM/FR, FR(Homology), FR(Analogy), and FR/NF Categories

	AL_4				GDT_TS			
	3D-Jigsaw	Pmodeller	Others	Rec	3D-Jigsaw	Pmodeller	Others	Rec
CM/FR								
T0130	61–60[2]			63	43.2–40[2]			37.3
T0132	66.4–56.8[3]	84.9–56.8[2]		82.2	42.3–39.4[3]	60.4–44.3[2]		61.6
T0159_1		53.3–18.6[10]	26.9–13.2[3]	40.7		36.9–16.2[10]	17–12.6[3]*	25.4
T0159_2		53.5–37.3[10]	44.4–32.4[3]	52.8		34.3–23.4[10]	27.8–23.4[3]*	33.1
T0168_1	58.8–49.4[4]	65.9–43.5[10]		53.5	40.1–34.8[4]	42.8–30.4[10]		35.7
T0168_2	26.2–17.7[4]	31.2–16.3[10]		16.3	22.1–19.1[4]	24.2–18.4[10]		19.7
FR(H)								
T0134_1	67.5[1]	72.2–32.5[7]		69.8	40.7[1]	43.8–20.4[7]		39.1
T0134_2	89.6[1]	87.7–70.7[7]		82.1	58.5[1]	66–42.7[7]		63.4
T0138	78.5–15.6[6]	83.7–60.7[10]		66	43.5–12.4[6]	58.3–47.9[10]		48.7
T0157		80.8–30.8[8]	41.7–10.8[4]	74.2		56.4–25[8]	56.4–22.9[4]?	52.5
T0174_1		15.2[2]		16.7		14.2[2]		14.5
T0174_2		23.9[2]		26.4		23.7[2]		23.7
FR(A)								
T0135(w)			25.5[1]				17.4[1]!	
T0147		22.6–14.5[5]	27.8–20.5[2]	43.6		32.9–23.9[5]	29.6–27.1[2]+	27.7
T0148_1	5.6[1]	23.9–5.6[5]		64.8	27.5[1]	45.1–26.8[5]		45.8
T0148_2	13.2[1]	13.2–6.6[5]		27.5	24.7[1]	35.7–28[5]		29.7
T0187_2(w)			15–8.8[2]	17.1			11.8–10.6[2]#	11.9
T0191_1(w)	15.1[1]	49.6–12.2[8]	21.6–12.2[5]	15.8	14.9[1]	34.9–15.3[8]	18–14.9[5]#	16.4
T0191_2	80.4[1]	81.8–61.5[8]	83.9–60.1[5]	80.4	51.6[1]	56.3–40[8]	52.8–43.4[5]#	52.6
FR/NF								
T0170		63.8–13[10]		47.8		49.6–31.9[10]		37.7
T0172_2		36.6–17.8[4]	26.7–14.8[11]	17.8		24.7–19.8[4]	20.5–17[11]$	18.1
T0173		18.1–14.6[3]	19.9[1]	18.1		13–10.1[3]	15.1[1]#	13
T0186_3	36.1–30.6[3]	50–30.6[10]	44.4–33.3[5]	38.9	29.2–27.8[3]	36.8–30.6[10]	29.9–28.5[5]#	29.9
T0187_1			17.6–16[2]	18.2			17.5–16.6[2]#	18.2

The first column states the target name (w for targets modeled by using templates with incorrect folds). The left side of the table shows AL_4 scores for the initial models fed into the recombination algorithm. These models were obtained from different Web servers (3D-JIGSAW, Pmodeller, and Others). Ranges show the best and the worst scored models, with the total number of models in square brackets. The fifth column shows the AL_4 score for the recombinant models. The right side of the table shows the analysis of the same data, using GDT_TS scores. See the main text for the definitions of these scores. "Others" are servers participating in CAFASP 3, where * indicates servers {Fams,Alax,Robetta}, ? Robetta, + {Robetta,Arby}, # Fams and $ {Fams,Alax}. Finally, ! indicates an experimental FR method by secondary structure pattern matching, developed by P.W. Fitzjohn.

these simulations, we used the GROMACS[25] simulation package and the OPLS-AA[26] force field. Snapshots taken from the trajectory were clustered according to average backbone deviations, and one conformation from the most populated cluster was selected. A few more rounds of CHARMM22 energy minimization were performed, and then this was submitted as a refined model. Insufficient computer resources prevented us from refining all targets by MD simulations.

RESULTS AND DISCUSSION

All 67 CASP5 targets were modeled by using the protocol in silico protein recombination. This population approach was used as an attempt to optimize template-based models obtained from different sources. The analysis of the results shows that, in general, recombined models are not significantly different from the best initial model, if that could have been identified at the time of submission. Only in a handful of cases did recombination yield slightly better models. With a similar frequency, the algorithm yields slightly worse models than the best initial, particularly when all the initial models are poor.

The performance of the method is similar across all CASP5 targets, but here only remote homology targets, down to the New Fold (NF) category, are discussed, because alignment errors and incorrect selection of templates become more frequent for these targets. In relative terms, our method appears to be more competitive in this range. Table I shows our analysis for the results of these 24 domains, after comparing our models with the targets for which the experimental structure is available. As described in Materials and Methods, a set of template-based models was constructed for each target to seed the initial population for a recombination experiment. The final model submitted was selected from those in the last generation of models, after convergence. This table shows how different the final recombinant models (Rec) are with respect to the initial models, constructed with the servers stated on the top of each column. To compare models, two standard CASP scores were computed (AL_4 and GDT_TS) by using the program LGA.[27] AL_4 is defined as the percentage of residues in a model for which corresponding residues in the target are within ±4 residues of the correct location (when superposed independently of sequence), and the distance between corresponding residues is <4 Å. GDT_TS is the average percentage of residues under a series of distance cutoffs (1, 2, 4, and 8 Å) after a sequence-dependent superposition. We discuss some cases in the light of these comparisons.

T0132 (HI0827, *Haemophilus influenzae*)

This CM/FR target was identified as a thioesterase by DomainFishing. By using profile–profile searches (see Materials and Methods), the template 1BVQ, a CoA-thioesterase from *Pseudomonas* sp., was confidently found. However, the alignment was not trivial, so three different alignments were used to build models with 3D-JIGSAW, and two more models were taken from Pmodeller, with one of them using a different template, 1C8U, another bacterial CoA-related enzyme. Recombination built a model that incorporated fragments from both templates but eventually had a very similar score to the best initial model, a Pmodeller model based on an alignment generated by INBGU.[28] We now analyze in more detail the major difficulties of the model, the phasing of strands 2 and 5 of the core β-sheet. For strand 2, our initial set of five alignments contained only segments shifted one or two positions with respect to the correct alignment. The resulting recombinant alignment is shifted one position at this point. However, for strand 5, there were two initial correct alignments (the remaining alignments were shifted by one and two positions), and they were incorporated into the final recombinant model. These results show how important it is to properly sample segments of ambiguous alignment, because the algorithm cannot generate alignments omitted from the initial population.

T0157 (yqgF, *Escherichia coli*)

This target was classified as FR(Homology) by the CASP5 assessors and was related to DNA-binding proteins according to the homologous sequences found by PSI-BLAST in the NCBI nr database. We could not find any confident template(s), so we took models from the CAFASP 3 results page. In particular, models from Robetta and Pmodeller were selected because they used the most popular templates (1KCF and 1HJR, *E. coli* and yeast endonucleases). Different alignments were found for each of them, and a recombination experiment was set to select the best. The recombinant model is comparable, although slightly worse than, the best initial one (based on an alignment generated by FUGUE[29] using 1HJR) but incorporating two different loops and a differently phased α-helix. The main difficulty of the target, an α-helix with a different angle to equivalent helices on the templates, was not resolved.

T0147 (*ycdX, E.coli*)

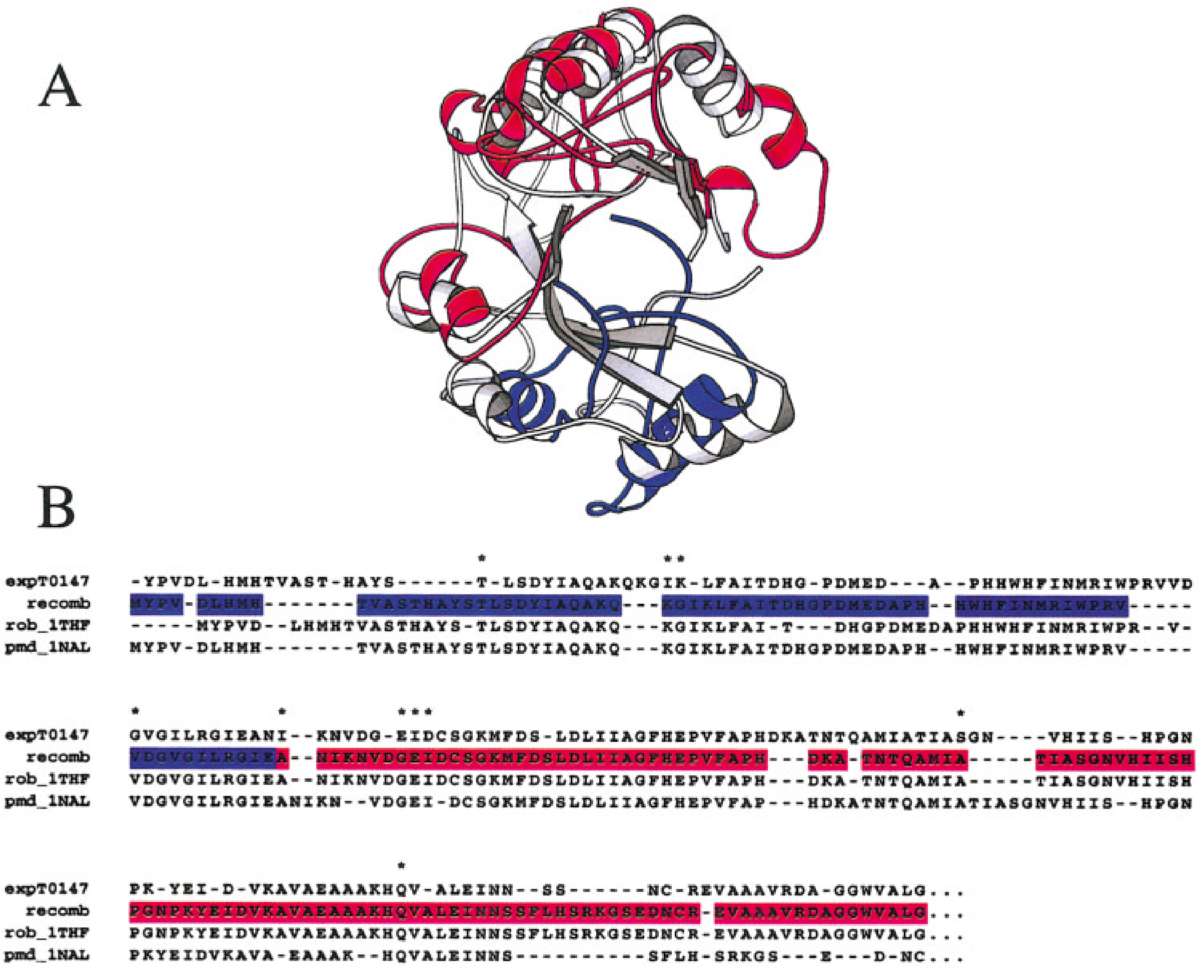
This FR(Analogy) target was identified as a PHP domain by DomainFishing, but no template could be found with our own set of tools, so once again, models for the most popular templates found by the CAFASP servers were downloaded. All these templates (1DHP, 1H5Y, 1QO2, 1THF, and 1NAL) were TIM barrels, with eight β-strands, whereas the target sequence had only seven β-strands predicted. No conclusive functional hint was found to help in selecting templates, so a set of seven models from Pmodeller, Robetta, and Arby was recombined by using the genetic algorithm. Like the initial models, the final recombinant model selected by our fitness function has a poor GDT_TS score. But as shown in Table I, the AL_4 score is considerably better than any of these. This example is shown in Figure 1 and is a good illustration of how the protein recombination algorithm works. In this case, the recombinant model includes two fragments from two models built from two different templates, obtaining a final composite model that can be better equivalenced to the experimental (in AL_4 terms). The algorithm took the better sections from each of the two models to build an improved, hybrid, model. Figure 1(B) shows the set of possible crossover points between these two initial models (marked as *). This points out one important limitation of this technique: useful crossovers between models are only possible if they can be reasonably superimposed in three dimensions, keeping together fragments with the same sequence.

T0170 (FF Domain of HYPA/FBP11, *Homo sapiens*)

Confident modeling template(s) could not be found by using our standard sequence similarity tools (a FR/NF target), and because of its helical secondary structure, many helical folds were nonspecifically recognized in our profile–profile template search. Thus, we decided to take all 10 models provided by Pmodeller and recombine them. Post-CASP analysis shows that the best initial model, based on the homeodomain 1LFB and aligned by 3D-PSSM,[30] is much better than the final recombinant model, suggesting that the algorithm tested may not perform very well with small helical proteins. However, repeating the recombination with the latest version of in silico protein recombination, which calculates free energies per residue, allowing comparison of proteins of different length, provides a recombinant model scoring 58 AL_4 and 46.7 GDT_TS, comparable to the best initial model.

It is interesting to compare the ability of the algorithm to produce recombinant models for CM targets. Despite the simplicity of the potential energy function, in most cases, the algorithm presented here selected the best possible alignments and templates from the initial available ensemble. In some cases, our recombinant models were significantly worse than those constructed by the best predictors. Analysis of some of these results (targets T0137, T0153, T0177, T0178, T0182, and T0192) shows that the quality of the initial models used in the recombination experiments is the main reason. Particularly, we believe that loop conformations were not successfully sampled for each initial model. We also noted that recombination can sometimes improve alignments but at the cost of making GDT_TS scores worse—possibly due to accumulation of errors during the evolutionary procedure.

Finally, we comment on the MD simulations ran for targets T0134, T0165, T0177, and T0185. In T0134_2, 0.5 ns of MD moved the protein considerably, diminishing the



```
                                      *             **
expT0147   -YPVDL-HMHTVAST-HAYS------T-LSDYIAQAKQKGIK-LFAITDHG-PDMED---A--PHHWHFINMRIWPRVVD
  recomb   MYPV-DLHMH-------TVASTHAYSTLSDYIAQAKQ---KGIKLFAITDHGPDMEDAPH--HWHFINMRIWPRV-----
rob_1THF   -----MYPVD--LHMHTVASTHAYS-TLSDYIAQAKQ---KGIKLFAI-T---DHGPDMEDAPHHWHFINMRIWPR--V-
pmd_1NAL   MYPV-DLHMH-------TVASTHAYSTLSDYIAQAKQ---KGIKLFAITDHGPDMEDAPH--HWHFINMRIWPRV-----

           *          *        ***                                        *
expT0147   GVGILRGIEANI--KNVDG-EIDCSGKMFDS-LDLIIAGFHEPVFAPHDKATNTQAMIATIASGN----VHIIS--HPGN
  recomb   VDGVGILRGIEA--NIKNVDGEIDCSGKMFDSLDLIIAGFHEPVFAPH---DKA-TNTQAMIA-----TIASGNVHIISH
rob_1THF   VDGVGILRGIEA--NIKNVDGEIDCSGKMFDSLDLIIAGFHEPVFAPH---DKA-TNTQAMIA-----TIASGNVHIISH
pmd_1NAL   VDGVGILRGIEANIKN--VDGEI-DCSGKMFDSLDLIIAGFHEPVFAP---HDKATNTQAMIATIASGNVHIIS--HPGN

                               *
expT0147   PK-YEI-D-VKAVAEAAAKHQV-ALEINN--SS------NC-REVAAAVRDA-GGWVALG...
  recomb   PGNPKYEIDVKAVAEAAAKHQVALEINNSSFLHSRKGSEDNCR-EVAAAVRDAGGWVALG...
rob_1THF   PGNPKYEIDVKAVAEAAAKHQVALEINNSSFLHSRKGSEDNCR-EVAAAVRDAGGWVALG...
pmd_1NAL   PKYEIDVKAVA-EAAAK--HQVALEINNS----------SFLH-SRKGS---E---D-NC...
```

Fig. 1. **A:** Cartoon showing the superposition of the experimental structure for the first 190 residues of T0147 (white) and the recombinant model submitted by our group (N-terminus in blue, C-terminus in red). **B:** Corresponding structural alignment of the experimental structure of T0147, the recombinant model, and two template-based models generated by Robetta (rob_1THF) and Pmodeller (pmd_1NAL). The N-terminus of the recombinant model was taken from pmd_1NAL, based on that PDB template and highlighted in blue, whereas the C-terminus is derived from rob_1THF, in red. In terms of AL_4 score, the recombinant model is significantly better than both pmd_1NAL and rob_1THF. In terms of GDT_TS, the recombinant model is comparable to them (see text for scores definitions). Asterisks (*) mark possible crossover points between the two initial models after a sequence-based superposition, all within loop regions.

GDT_TS score but increasing the AL_4. For the rest of the targets, MD did not have an important effect over the structure, with final GDT_TS and AL_4 values very close, but slightly worse, than the unrefined model.

CONCLUSIONS

The genetic algorithm tested in CASP5 tends to produce recombinant models that are comparable to the best initial model, if we had identified it. This result suggests that our simple fitness function correctly identifies good models, making it a good candidate to filter and rank models from FR servers as well as models built in-house, or indeed a combination of both. In addition, the method has been shown to be able to improve alignments by recombining well-aligned regions from individual models. Unfortunately, the quality of the models used to seed the first generation seems to be the upper limit for the quality of the final model, showing that the current implementation of the algorithm is not adding much beyond this baseline. Finally, because good global superpositions are required for useful crossover, the current implementation of in silico protein recombination cannot recombine efficiently proteins that are totally different or have different domain orientations. This suggests that local superpositions may be required.

ACKNOWLEDGMENTS

We thank Cancer Research UK for supporting our work and the experimentalists for participating in CASP.

REFERENCES

URL for Web servers mentioned in this article:
3D-JIGSAW::http://www.bmm.icnet.uk/~3djigsaw
3D-PSSM: http://www.sbg.bio.ic.ac.uk/~3dpssm
Alax: http://alax.bio.nagoya-u.ac.jp/
Arby: http://www.gmd.de/SCAl
DomainFishing: http://www.bmm.icnet.uk/~3djigsaw/dom_fish

EsyPred3D: http://www.fundp.ac.be/urbm/bioinfo/esypred
FAMS: http://physchem.pharm.kitasatou.ac.jp
FUGUE: http://www.cryst.bioc.cam.ac.uk/~fugue
INBGU: http://www.cs.bgu.ac.il/~bioinbgu
LGA: http://PredictionCenter.llnl.gov/local/lga
Pmodeller: http://www.sbc.su.se/~arne/pcons

1. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. Embo J 1986;5:823–826.
2. Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: Continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001;17:1242–1243.
3. Contreras-Moreira B, Fitzjohn PW, Bates PA. Comparative modelling: an essential methodology for protein structure prediction in the post-genomic era. Appl Bioinform 2002;1:177–190.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
5. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001;Suppl: 5:22–38.
6. Sippl MJ, Lackner P, Domingues FS, Prlic A, Malik R, Andreeva A, Wiederstein M. Assessment of the CASP4 fold recognition category. Proteins 2001;Suppl:5:55–67.
7. Unger R, Moult J. Genetic algorithms for protein folding simulations. J Mol Biol 1993;231:75–81.
8. May AC, Johnson MS. Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions. Protein Eng 1995;8:873–882.
9. Morris GM, Goodsell DS, Huey R, Olson AJ. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. J Comput Aided Mol Des 1996;10:293–304.
10. Rabow AA, Scheraga HA. Improved genetic algorithm for the protein folding problem by use of a Cartesian combination operator. Protein Sci 1996;5:1800–1815.
11. Xia Y, Levitt M. Roles of mutation and recombination in the evolution of protein thermodynamics. Proc Natl Acad Sci USA 2002;99:10382–10387.
12. 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.
13. Contreras-Moreira B, Bates PA. Domain Fishing: a first step in protein comparative modelling. Bioinformatics 2002;18:1141–1142.
14. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 1998;26:320–322.
15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
16. Bates PA, Sternberg MJ. Model building by comparison at CASP3: using expert knowledge and computer automation. Proteins 1999; 37:47–54.
17. Ogata K, Umeyama H. An automatic homology modeling method consisting of database searches and simulated annealing. J Mol Graph Model 2000;18:258–272, 305–256.
18. Lambert C, Leonard N, De Bolle X, Depiereux E. ESyPred3D: prediction of proteins 3D structures. Bioinformatics 2002;18:1250–1256.
19. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
20. Lundstrom J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network-based consensus predictor that improves fold recognition. Protein Sci 2001;10:2354–2362.
21. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
22. Robson B, Osguthorpe DJ. Refined models for computer simulation of protein folding. Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor. J Mol Biol 1979;132:19–51.
23. Eisenberg D, McLachlan AD. Solvation energy in protein folding and binding. Nature 1986;319:199–203.
24. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization and dynamics calculation. J Comp Chem 1983;4:187–217.
25. Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 2001;7: 306–317.
26. Damm W, Frontera A, Tirado-Rives J, Jorgensen WL. OPLS all-atom force field for carbohydrates. J Comput Chem 1997;18: 1955–1970.
27. Zemla A. LGA program: a method for finding 3-D similarities in protein structures. http://PredictionCenter.llnl.gov/local/lga 2002.
28. Fischer D. Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput 2000:119–130.
29. 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.
30. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.