Confronting the Problem of Interconnected Structural Changes in the Comparative Modeling of Proteins.

June 25, 1997

Ram Samudrala[†‡], Jan T Pedersen[†], Huai-bei Zhou[†], Rui Luo[†¶], Krzysztof Fidelis[§],

and John Moult[†*]

[†]Center for Advanced Research in Biotechnology
University of Maryland Biotechnology Institute
9600, Gudelsky Drive, Rockville, MD 20850

[‡]Molecular and Cell Biology Program and [¶]Department of Chemistry
University of Maryland at College Park
College Park, MD 20742

[§]Biology and Biotechnology Research Program
Lawrence Livermore National Laboratory,
Livermore, CA 94551

Running title: Interconnectedness in Comparative Modelling

Keywords: correlated structural changes, eosinophil derived neurotoxin, cellular retinoic acid-binding protein, histidine-containing phosphocarrier protein.

* Corresponding author
Phone: (301) 738-6272
FAX: (301) 738-6255
E-mail: jmoult@indigo5.carb.nist.gov

ABSTRACT

Comparative models of three proteins have been built using a variety of computational methods, heavily supplemented by visual inspection. We consider the accuracy obtained to be worse than expected. A careful analysis of the models shows that a major reason for the poor results is the interconnectedness of the structural differences between the target proteins and the template structures they were modeled from. Side chain conformations are often determined by details of the structure remote in the sequence, and can be influenced by relatively small main chain changes. Almost all of the regions of substantial main chain conformational change interact with at least one other such region, so that they often cannot be modeled independently. Visual inspection is sometimes effective in correcting errors in sequence alignment and in spotting when an alternative template structure is more appropriate. We expect some improvements in the near future through the development of structure based sequence alignment tools, side chain interconnectedness rotamer choice algorithms, and a better understanding of the context sensitivity of conformational features.

INTRODUCTION

Our objective in this work was to test the usefulness of as many of the available computational techniques for comparative modeling as possible, and to try to see where improvements can be made. To this end, models of three of the target proteins, the Histidine-containing Phosphocarrier (HPr) protein from *Mycolasma capricolum* (McHPr; 89 residues [1]), the Mouse Cellular Retinoic Acid-Binding Protein I (CRABPI; 137 residues[2] [2]), and the Eosinophil Derived Neurotoxin (EDN; 134 residues [3]), were built. We divide the modeling into three main stages: (*i*) an alignment mapping the sequence of the target protein on to a template structure, (*ii*) procedures for assigning side chain positions (rotamers) in the context of the surrounding model, and (*iii*) procedures for building regions of main chain. For each stage we indicate what methods were used, what went right, what went wrong, and why (if we think we know). In the last section we discuss what we learned and what type of next generation algorithms may lead to improved model accuracy.

[1]Abbreviations: AMPS, Alignment of Multiple Protein Sequences; C_α, alpha-carbon; CRABPI, Cellular Retinoic Acid-Binding Protein I; EDN, Eosinophil Derived Neurotoxin; HPr, Histidine-containing Phosphocarrier; McHPr, HPr from *Mycoplasma capricolum*; MP, Minimum Perturbation; PDB, Protein Data Bank; RMSD, Root Mean Squared Deviation; SCD, Self Consistent Domain; SCOP, Structural Classification of Proteins; 3D, 3-dimensional

[2]We constructed two models of CRABPI; we only consider the model with the lower RMSD to the experimental structure in this paper. The numbering of the residues in the PDB file for CRABPI differs from the numbering we have used. The model structure begins at M1 whereas the experimental structure begins one residue later, at P1. The first Methionine is probably not present in the protein expressed in *E. coli*.

Structure (PDB code)	Source	Function	Sequence Identity (%)	Resolution (Å)
McHPr (a)				
2hpr	*B. subtilis*	phosphotransferase	40.9	2.0
1ptf	*S. faecalis*	phosphotransferase	40.2	1.6
1poh	*E. coli*	phosphotransferase	34.1	2.0
CRABPI (b)				
2hmb	*H. sapiens*	heart fatty acid-binding	42.7	2.1
1opa	*R. rattus*	retinol transport	36.6	1.9
1lie	*M. musculus*	adipocyte lipid-binding	34.6	1.6
2ifb	*R. rattus*	intestinal fatty acid-binding	29.0	2.0
1mdc	*M. sexta*	fatty acid-binding	23.8	1.6
EDN (c)				
7rsa	*B. taurus*	pancreatic ribonuclease	33.9	1.3
1bsr-A	*B. taurus*	seminal ribonuclease	31.4	1.9
1onc	*R. pipiens*	pancreatic ribonuclease	29.4	1.6

Table 1: Percentage sequence identity between the target sequence and other homologous sequences with known structures as determined by AMPS pairwise alignments: a - Histidine-containing Phosphocarrier from *Mycoplasma capricolum* (McHPr); b - Cellular Retinoic Acid-Binding Protein I (CRABPI); c - Eosinophil Derived Neurotoxin (EDN).

METHODS AND RESULTS

Sequence search

Target protein sequences were obtained from the National Center of Biotechnology Information (NCBI) protein and nucleotide sequence database ENTREZ [4]. A FASTA search [5] was performed on the OWL [4] database to obtain sequences that were related to the target protein. The Structural Classification of Proteins (SCOP) [6] database was used to find the PDB identifiers for the known structures that belonged to the same family as the target sequence. High resolution structures obtained using x-ray crystallography were used as template structures for the modeling. Table 1 shows the structures that were selected for each family and the percentage identity to the target protein.

Sequence and structure alignment

A multiple sequence alignment was generated with the AMPS package [7, 8]. The AMPS-derived alignment was used to identify regions of variability within the target sequence family. AMPS pairwise alignments were also used to determine the degree of homology

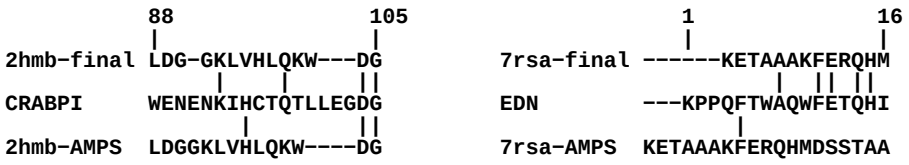

Figure 1: Differences between the final correct sequence alignments and those generated with AMPS. Correct alignments were produced by visual inspection of the sequences and preliminary models.

between the target sequences and the other sequences of known structure (see Table 1). The default PAM250 mutation matrix and a length independent gap penalty of 8.0 were used. Structural alignments between the template structures were generated using the G program[9, 10]. These alignments were used to examine the structural variation at a given position and to assess the correctness of the multiple sequence alignment.

Visual inspection of the initial AMPS alignments revealed two regions where the alignment was dubious (see Figure 1). One of the regions is in CRABPI (insertion at residue 90 which is not seen in the AMPS alignment), and the other is in EDN (the FEQTH sequence (residues 11-15) is aligned incorrectly). The wrong alignment in CRABPI results in K93 being buried, which seemed electrostatically intolerable. In the case of EDN, inspection of the alignment suggested a better alternative. Both alignments were adjusted manually. The final alignments for all proteins agree with those produced by structural superposition of the target experimental structures with the respective primary templates. Figure 2 shows the results of the sequence alignment for CRABPI after correction using structural information.

Side chain replacement and model generation

Following the sequence alignment, an initial model was generated by mutating the residues of the template structure with the highest identity to the target sequence. This was done using a minimum perturbation (MP) technique implemented by the program MUTATE (R. Read). The MP method changes a given amino acid to the target amino acid preserving the equivalent χ angles, as determined by an equivalence table, between the two side chains. The χ angles not present in the model are constructed using a standard library based on the residue type. A careful environment analysis was performed by visual inspection of the intial model using interactive computer graphics. If residue A in a template structure was changed to residue B in the model, then the environments (the contacting residues, their

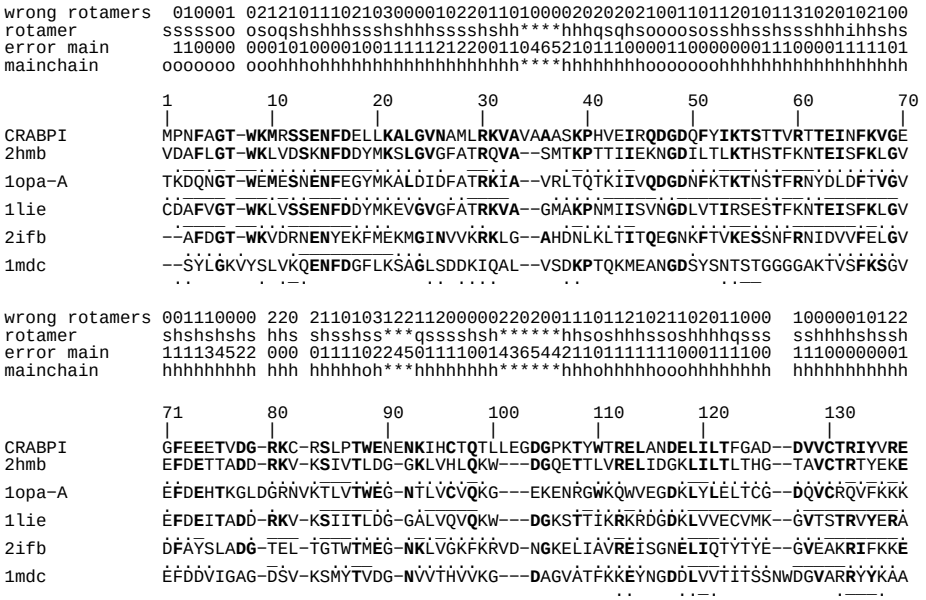

Figure 2: Final alignments of the CRABPI target sequence to other sequences in the family that have known structures. The first line indicates the accuracy of the predicted rotamer by listing the number of χ angles that deviated more than 30° from the experimental structure for each residue. The second line is the list of rotamer choices that were used to generate the final model—for each residue, the rotamer was selected using one of the following methods: s - standard library; i - Insight; q - Quanta; or by selecting from a template structure: h - 2hmb; o - 1opa-A. The third line lists the C_α deviation between the target experimental structure and the model (0: 0-1Å; 1: 1-2Å; . . .). The fourth line indicates the parent structure from which the main chain was taken: o - 1opa-A; h - 2hmb. An '*' indicates that the main chain and/or side chain was generated using loop building techniques. In the multiple sequence alignment, conserved residues are indicated by bold letters. For each amino acid in all the sequences aligned to the target, the C_α distance between the target experimental structure and each related structure after structural alignment is given: a solid line under the one letter code indicates that the C_α distance was within 1.0Å, a dotted line indicates that the C_α distance was within 2.0Å, and a blank indicates that the C_α distance was greater than 2.0Å

locations, and conservation) of residue A and residue B were compared. The rules used to consider plausibility were packing (whether there was too much or too little space left after any change), favorable and unfavorable electrostatic interactions (hydrogen bonding, salt bridges) of side chains and main chain, and burial or exposure of a residue. The confidence of the model at a given position was rated qualitatively using these criteria. Alternate side chain rotamer choices were considered for regions of low confidence.

Rotamer Origin	McHPr	CRABPI	EDN	McHPr	CRABPI	EDN
Library	50.0% (60)	48.0% (100)	50.5% (95)	25.0% (8)	50.0% (6)	37.5% (24)
Identity	34.7% (46)	38.0% (84)	25.0% (48)	26.6% (15)	37.5%(8)	24.2% (33)
Loops	80.0% (5)	66.6% (15)	66.6% (75)	50.0% (2)	00.0% (0)	73.6% (19)
Manual	65.5% (29)	41.1% (34)	00.0% (5)	50.0% (2)	33.3% (3)	00.0% (5)
All	48.5% (142)	45.4% (233)	49.3% (223)	29.6% (27)	41.1% (17)	38.2% (81)

Table 2: Percentage of model χ angles that deviate more than 30° from the experimental structure, considering rotamers that were constructed using a standard library (row 1), identities (row 2), loop builders (row 3), and by other methods (row 4; see Figure 2). The overall percentages are given in the last line. The right hand side omits residues that have contacts closer than 4.0Å to a neighboring protein molecule and χ angles where one or more atoms have a temperature factor greater than 25.0Å^2. The numbers in parenthesis show the total number of χ angles that were included.

Two other methods using different χ libraries were employed in order to generate possible alternative rotamers. These were from the Insight [11] and Quanta [12] packages. In addition, a preliminary version of a self-consistent domain (SCD) method [13] was used. This method iteratively adjusts side chain conformations within a neighborhood to find the electrostatically most favorable clash free set, and checks for consistency with adjacent and overlapping neighborhoods.

An electrostatic energy analysis using point charge electrostatics with an intergroup cutoff distance of 5.0Å was performed on the model using the Eneana program [14]. Residues with unfavorable electrostatic interactions were corrected by examining alternative residue conformations and selecting an energetically favorable one. Residues with unlikely burial were identified by checking the probability of observing that particular burial in an experimental protein structure and similarly corrected.

The percentage of model χ angles that deviated more than 30° from those in the experimental structures is given in the left hand side of Table 2. A number of χ values may be affected because of high temperature factors or contacts with neighboring molecules. For the purpose of evaluating the methods used, it is desirable to eliminate these effects and produce a "no excuse" set of χ angles. We thus calculated additional statistics, excluding residues that have atomic contacts of less than 4.0Å to a neighboring molecule and χ angles where one or more atoms had a temperature factor greater than 25.0Å^2. The right hand side of Table 2 shows these results. Errors are significantly lower in this set, but still surprisingly large, even for cases where the residues in the models and template structures are identical (row 2).

Changes in the position of conserved side chains between related structures must be because of changes in other parts of the structure. To obtain more insight into these correlation effects and others, we examined the seven cases (three Library, three Identity, and one Manual) in

Residue	$\angle$	$\Delta\chi$ (°)	Effect of Using the Rotamer in the Experimental Structure
I53	χ_1	66	Clash with I64 and R112 (see Figure 3).
R112	χ_3	74	Incompatible with experimental solvent structure.
I120	χ_1	34	No clashes.
I120	χ_2	125	No clashes.
F123	χ_2	45	Clash with L121 and V77.
			V77 is in an incorrectly modeled loop.
I133	χ_1	149	Clash with R11 and S12.
			These residues have high temperature factors.
V135	χ_1	66	No clashes.

Table 3: Correlation of individual χ angle errors with other errors in the CRABPI model. Data are for the incorrect angles in the right hand side of Table 2. For each χ listed, the conformation of the corresponding residue in the experimental structure was changed to adopt the model χ value and the resulting environment inspected for inconsistencies.

the "no excuse" set of the CRABPI model where the χ values are wrong. This was done by introducing the model rotamers into the experimental target structure and inspecting the resulting environment. Table 3 shows the results of this analysis. For three of the seven rotamers, the model rotamers were not acceptable in the experimental structure because of clashes that are not present in the model. For two of these, the clashes are directly attributable to main chain differences between the experimental structure and model, so better side chain positioning algorithms would not help. Figure 3 illustrates one of these main chain effects for I53. Here, a difference in the main chain in the target structure relative to the template of the neighboring I64 results in the model rotamer being unacceptable. The side chain conformation of the conserved I64 in the model is similar to that seen in the experimental structure. There is also a side chain clash between the model conformation of I53 and the experimental conformation of R112. Similarly, the side chain conformation of F123 is determined by the conformation of a loop region that was incorrectly modeled. The side chain conformation of I133 is dependent on the conformation of the side chain of R11, which forms a salt bridge with E118 in the experimental structure, but not in the model. The experimental conformation of R112 appears to interact better with solvent than the model conformation. For the other three cases, our criteria could not distinguish between the experimental and model rotamers.

Building insertions and deletions

Insertions and regions flanking the deletion in the target sequences relative to the templates were rebuilt using one of four different methods (Table 4). All of these regions have final conformations with C_α RMSDs greater than 4.0Å.

In McHPr, a lengthening of C terminal region compared with the primary template appeared to enable the formation of an additional short anti-parallel beta strand to pair with the N

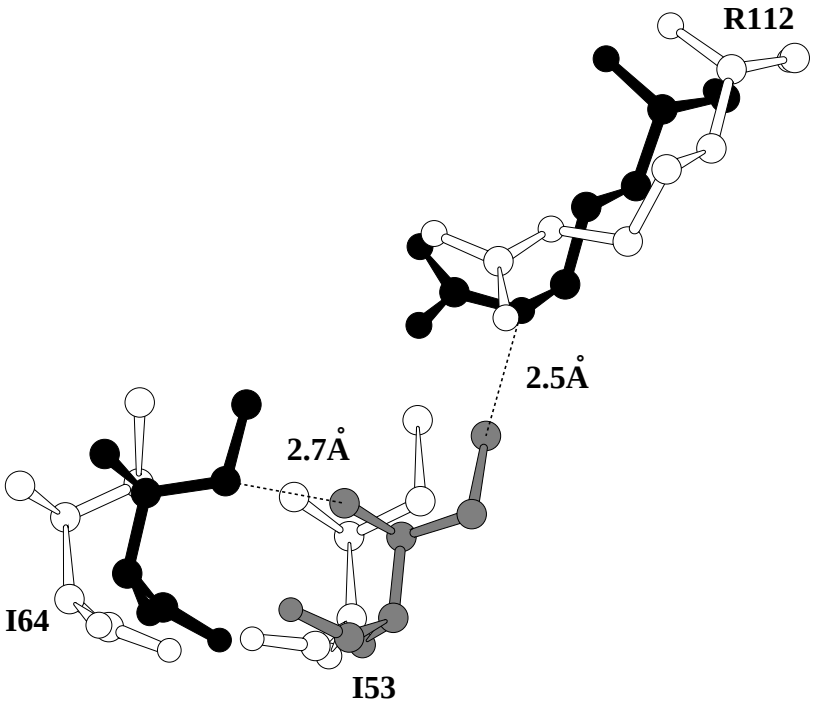

Figure 3: An example of a wrong model rotamer in CRABPI that is unacceptable given the context of the experimental structure. The model structure is white, the experimental structure is black, and the model side chain of I53 placed in the experimental structure is grey. In the model, I64 is further away because of a main chain shift, so the principal clash excluding the I53 model side chain conformation is not present.

terminal strand. These residues (87-89) were rebuilt manually. However, in the experimental structure, this region turns away from the protein surface with the last two residues involved in an intermolecular contact. Thus this conformation could be the result of a crystal packing effect.

In CRABPI, the main chain for residues 34-37 was manually adjusted to extend the C

Region	Structure	RMSD (Å)	Max Root Error (Å)	Method	Intermolecular Contacts	<B> (Å^2)
87-89	McHPr	5.5	0.6	manual	88-89	29.6
34-37	CRABPI	5.0	2.8	manual	37	32.7
90-92	CRABPI	4.2	2.0	pattern matching	-	39.0
101-106	CRABPI	5.3	2.2	[15]	-	80.0
1-5	EDN	9.7	3.3	*ab initio* [16]	3-5	13.3
18-22	EDN	5.3	3.4	manual & [17, 18]	19,21	8.7
62-70	EDN	3.1	1.2	[19]	66-67,69	22.9
89-96	EDN	5.2	6.1	[19]	90-91,95	37.1
112-126	EDN	9.9	7.1	[19]	113-114,116-117 122,124-125	15.9

Table 4: C_α RMSDs between the experimental structure and the model for insertions and residues flanking the single deletion (EDN: 18-22). The larger of the two root C_α atom errors is given in column 4. For each region, the list of residues with at least one atom in the side chain having intermolecular contacts less than 4.0Å is given in column 6. Column 7 lists the average temperature factor for the C_α atoms.

terminus of the α_2 helix. This was a correct guess, but since the adjustment was manual, the shape of the helix is far from ideal. For residues 90-92 in CRABPI, loops that had the same structural pattern as the region of uncertainty (two strands with a three residue loop between them with Glutamate as the center residue of the loop) were obtained from a database of structures. A manual inspection of these loops was used to select the most appropriate one, which are residues 320-322 in 2ach-A. Residues 101-106 in CRABPI were built using the SCS loop building program [15]. This method systematically generates a large set of possible main chain and side chain conformations. In this instance, too many main chain possibilities were generated so a subset had to be chosen by manual inspection. Experimental errors could be a problem in some cases since all the loops in the CRABPI structure have atoms with large temperature factors (Table 4).

Residues 1-5 in EDN were built using *ab initio* methods described in [16] which predicted this set of residues to be partly helical, whereas the correct conformation in this region resembles a turn. 18-22 represents a deletion in EDN with respect to the 7rsa template. It was constructed manually using 1onc as a template (which results in a deletion of only 2 residues, as opposed to a deletion of 6 residues when 7rsa is used) and further refined using Congen [17, 18]. In this procedure, each side chain in the loop and its surroundings is spun in turn to find the lowest energy conformation. The process is iterated until the total energy has converged. For the other three loops in EDN (residues 62-70, 89-96, and 112-126), distance constraints from the parent structure were used to search a database of loops [19] for matching regions. The matching loops were positioned in the model structure using the method of Martin, *et. al.* [19]. Side chains were then rebuilt as described above using Congen. Table 4 shows that all the loops in EDN have contacts with neighboring

Region	Structure	RMSD to Primary Template (Å)	RMSD to Model (Å)	Intermolecular Contacts	<B> (Å^2)
39	McHPr	1.9	1.8	39	23.6
14-17	McHPr	1.5	1.5	14,17	19.1
51-55	McHPr	0.7	1.3	51-52,54-55	17.9
70-83	McHPr	0.5	1.0	71-72,75-76,78-79	18.2
1-10	CRABPI	1.4	0.8	9-10	40.7
46-52	CRABPI	4.1	1.1	46,49	39.3
75-80	CRABPI	3.2	3.1	-	37.8
116-118	CRABPI	2.3	1.4	-	53.1
30-34	EDN	5.1	5.1	33,34	10.8
58	EDN	4.1	3.9	58	11.37

Table 5: C_α RMSDs for other regions of main chain variation. The RMSDs to the primary template shows how much that main chain differs from the experimental structure and the RMSDs to the model shows how accurately the variation was predicted. Three regions in CRABPI were predicted well. The list of residues with at least one atom having intermolecular contacts less than 4.0Å is given in column 5. Column 6 lists the average temperature factor for the C_α atoms.

protein molecules. This factor cannot be taken in account in the modeling.

The errors in the positions of the root residues shown in Table 4 are large—up to 7.0Å, and indicates one reason as to why the loop conformations are so poor. In such cases, the region rebuilt was not large enough and therefore no low RMSD loops could possibly be generated.

Other regions of main chain variation

Comparison of the experimental target structures with the primary templates used in the modeling shows other regions where the main chain conformations are significantly different. We list those regions that have a C_α RMSD greater than 3.0Å in CRABPI and EDN and 1.0Å in McHPr, or regions where we explicitly changed the main chain from the primary template.

Three such regions in CRABPI, residues 1-10, 46-52 and 116-118, were predicted with acceptable accuracy by using 1opa-A as a template rather than 2hmb. The changes in the conformation between CRABPI and 2hmb of the N terminus and the hairpin around residue 49 are correlated (Figure 4), and appear to be the consequence of a set of side chain differences: Two residues (F51L, W88L) are more bulky in CRABPI and there is a loss of a salt bridge between residues 2 and 46. For the third region, there is a Glycine in 2hmb at position 117 with ϕ/ψ values not allowed for other residue types. In CRABPI and 1opa-A, there is an Aspartate here. These and other side chains of the core of CRABPI are more similar to 1opa-A than those in 2hmb, even though the overall sequence identity is significantly lower.

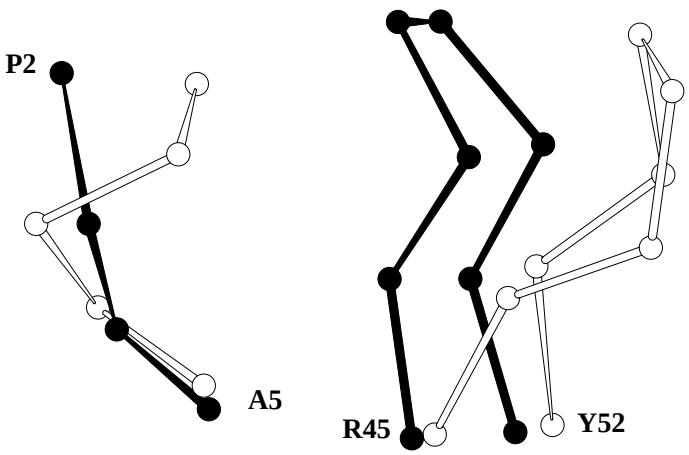

Figure 4: Correlated changes between the N terminus and the loop around residues 46-52 in CRABPI (black) compared with 2hmb (white). In 1opa-A, the conformation is similar to that of CRABPI, providing a better template than the primary one of 2hmb. Correlated changes of this type are common, and such regions of main chain often cannot be modeled independently.

All other regions of main chain variability that do not involve insertions and deletions were not identified and thus incorrectly modeled. With the wisdom of hindsight, some of these can be understood: In McHPr, the region around 14-17 appears to shift from the template because of the presence of a salt bridge between H15 and D10. E39 has a strained ϕ/ψ pair (93,3), and such conformational strain outside of functional regions is rare [20]. In this case it is almost certainly due to the many contacts with a neighboring molecule in the crystal [1]. In CRABPI, the loop around 75-80 appears to move relative to 2hmb because of the V77A and L77Y changes, a main chain shift of about 1.0Å at positions 20-25, and the loss of a salt bridge between residues R59K and D78G. In EDN, the change around residues 30-34 relative to 7rsa may be due to Y33T causing a clash with the conserved Y98. There is a

shift at position 58 caused by the introduction of a Proline, which requires a ϕ angle change.

Two changes in main chain were wrongly introduced. In one of these, residues 51-55 in McHPr, we incorrectly supposed that side chain volume changes would cause a main chain shift seen in one of the other templates. In the other case, the last helix in McHpr (residues 70-83) shifts as a consequence of energy minimization done to accommodate the incorrectly built C terminus.

Model refinement

Once the final side chain rotamers and loop conformations were selected from the variety of choices available, the models were energy minimized for 100 steps using the steepest descent method and either the CHARMM or Discover potentials without electrostatics. This procedure was intended to remove steric clashes and to produce acceptable bond lengths and angles rather than change the conformation significantly. Total movements were small, but increased the C_α Root Mean Square Deviation (RMSD) between the model and experimental structure slightly. For McHPr, the increase of the C_α RMSD between the minimized and the unminimized model with respect to the experimental structure is 0.070Å, for CRABPI it is 0.014Å, and for EDN it is 0.021Å.

Final RMSDs between the model and experimental structures

Throughout the paper, the RMSD between two structures with n equivalent positions is defined as

$$\sqrt{\frac{\sum_{i=1}^{n} dx_i^2 + dy_i^2 + dz_i^2}{n}},$$

where dx_i, dy_i and dz_i are distances in Cartesian space between two structures at position i. RMSDs were computed using the program G [9] and represent global RMSDs (i.e., RMSDs listed for specific regions are calculated after optimally superimposing the complete molecules [10]). Table 6 lists the RMSDs for all residues for the three models.

	McHPr	CRABPI	EDN
C_α	1.18	2.01	4.55
main chain	1.22	2.00	4.44
all atoms	1.76	2.62	5.50

Table 6: Root Mean Square Deviation (RMSD) in Å between the complete final models and the experimental structures for all residues.

DISCUSSION

The accuracy of the models is very unsatisfactory but the modeling experiment has been educational. Three common themes have emerged: The first is the usefulness of visual inspection rather than a reliance on numerical algorithms. The second is the extraordinary interconnectedness of changes between different homologous proteins. The third is the possibility, in some cases, of devising automatic procedures that may significantly improve accuracy.

Alignment

It has been known for some time that alignment of sequences with less than 40% identity tends to produce frequent errors in the mapping of a sequence on to a template structure [21, 22]. We encountered two cases of that (see Figure 1). In one, inspection of the alignment at the amino acid sequence level suggested a better solution. In the other, inspection of the structural implications of the alignment allowed a correction. With these adjustments, the sequences of all three models were correctly aligned with the available template structures. It should be possible to develop algorithms that examine the structural implications of alternative alignments.

Selecting side chain rotamers

Inspection of the structural implications of default rotamer choices did lead to a small improvement in accuracy, but the error level is still very high, even for those residues not likely to be affected by crystal packing or high crystallographic temperature factors. Better methods based on consideration of interacting sets of side chains are clearly needed. Such algorithms have been published, with reported high accuracy in core regions [23, 24, 25]. However, from the analysis of the CRABPI errors (Table 3), it is clear that these algorithms will be seriously affected by the main chain inaccuracies present in real models.

Insertions and deletions

Several algorithms [15, 17, 18, 19] have been shown to produce usefully accurate structures of short stretches of chain in the context of the surrounding protein. There are four obvious explanations as to why they did not work here, all related to the difference between real modeling versus algorithm development tests. The first and in the long run most difficult to address is the interconnectedness of the differences between related protein structures. An example of this is the interaction between the N terminal region of EDN relative to ribonuclease A and the long insertion at residues 112-126. These two regions pack against each other in the experimental structure, so that predicting one in isolation from the other is likely to be very problematic (see Figure 5). Spotting these correlated changes can some

McHPr	14-17	39	51-55	70-83	87-89
14-17			3		
39					
51-55	3				
70-83					2
87-89				2	

CRABPI	1-10	34-37	46-52	75-80	90-92	101-106	116-118
1-10		2	4		3		2
34-37	2						
46-52	4						
75-80						4	
90-92	3						
101-106				4			
116-118	2						

EDN	1-5	18-22	30-34	58	62-70	89-96	112-126
1-5			3				3
18-22							
30-34	3					2	
58							
62-70							
89-96			2				
112-126	3						

Table 7: The interconnectedness of the insertions and deletions and other regions of main chain variation . The number of residue pairs that that have one more atomic contacts less than 4.0Å is given.

times provide the key to modeling, as in the case involving the N terminus of CRABPI and the conformation of the loop around residues 46-52 (see Figure 4). More often than not, they simply render any automatic loop builder useless. Table 7 shows the striking extent of the interconnectedness between the variable regions in the experimental structures.

A second and related problem is the one of the size of variable region that must be included. Both systematic and database searches are severely limited in the size of region they can consider [26]. Effective maximum loop sizes are probably currently about seven residues, ignoring any changes in the surroundings. The short rebuilt regions that we used resulted in large errors of the root residues which led inevitably to high loop RMSDs (Table 4). It is apparent that insertions and deletions often cause significant main chain adjustment in the adjacent residues even where sequence conservation is high. The third problem is the need for reliable and affordable energy functions with which to screen possible conformations. In no case were we able to do this because of time and computing limitations. The fourth problem

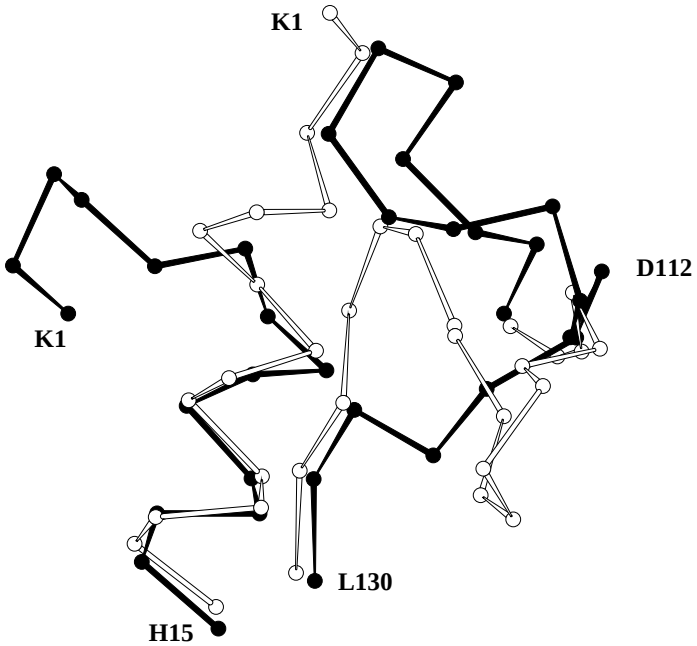

Figure 5: An example of an error in the building of one main chain region excluding the selection of the correct conformation of another region. Experimental structure of EDN is black, model is white. The incorrect structure of the model N terminus occupies space needed for the loop 113-129 (shown in black).

is knowing when to believe the reported experimental structure is relevant to the modeling. In the case of McHPr, we saw one possible case of the crystal environment affecting the conformation around the C terminus. For CRAPBI, all three loop regions have very large temperature factors. In EDN, the temperature factors are more reasonable, but all the loops are involved in intermolecular interactions in the crystal (Table 4). While it is very unlikely that in all cases the high RMSD between the experimental structure and the model are the result of the crystal effects, it does make it difficult to assess the individual predictions.

Identification of regions of main chain variability

When there are insertions and deletions in the sequence alignment, it is obvious that the local main chain conformation is unknown . But there are additional regions of main chain variability that are less easy to identify (Table 5).

Examination of the structural variation within the family may be useful for identifying such regions. For example, positions where the RMSD was greater than the mean RMSD within the CRABPI family (2hmb, 1lie, 1opa-A, 2ifb, and 1mdc) were found to be residues 1-6, 37-40, 47-50, 58-64, 74-83, 89-91, 99-107, 116-118, 127, and 137. This would identify all regions of structural variation listed in Table 5 but also would identify 3 additional regions that are conserved. This analysis, together with consideration of two other factors, changes involving Glycine and Proline residues and the level of local sequence similarity, may help in identifying main chain changes.

As in the case of insertions and deletions, all the regions that vary extensively in main chain conformation have high temperature factors or form intermolecular contacts (see Table 5). For example, the conformation around residue E39 in McHPr is clearly determined by crystal packing.

Choice of alternate templates

In the case of CRABPI, we were able to significantly improve the model by recognizing 1opa-A as a better choice for the backbone in three regions. Inspection of Figure 2 reveals other regions where the prediction could have been improved by choosing main chains from other related structures. For example, the main chain around 108-115 and 128-137 in CRABPI is better modeled by using the main chain from 1opa-A. These choices depend on structural details and are difficult to automate. The usefulness of a "mix and match" approach to template selection is well known [27].

Long term hopes

We can see the way ahead for improvements in sequence alignment, rotamer choice and identification of main chain changes. Loop building is the most glaring and seemingly intractable problem in these results. Its successful treatment requires the development of methods for handling the interconnectedness of features in protein structures. One partial solution may be to consider pieces of chain have their conformation determined essentially independently from the rest of the protein structure [28]. An example of the relevance of that approach is the interaction between the N terminus of EDN and the region 133-129. Analysis suggests that the N terminus has its conformation determined by local sequence effects [16], so it should be built first and then the long loop added.

A complete solution to the comparative modeling problem, *i.e.*, methods rivalling experiment in accuracy, requires the development of radically new approaches that handle the interconnectedness of the structural changes between related protein structures.

ACKNOWLEDGEMENTS

This work was supported in part by NIH GM41034 and by a Life Technologies Fellowship to Ram Samudrala. Work performed at the Lawrence Livermore National Laboratory was under the auspices of the U.S. Department of Energy and supported by Contract W-7405-ENG-48 and Laboratory Directed Research and Development Award 93-DI-003.

References

[1] Pieper, U., Kapadia, G., Zhu, P., Peterkofsky, A., Herzberg, O. Structural evidence for the evolutionary divergence of Mycoplasma from gram-positive bacteria: the histidine-containing phosphocarrier protein. Structure 3:781–790, 1995.

[2] Kleywegt, G., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K., Jones, T. Crystal structure of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid. Structure 2:1241–1258, 1994.

[3] Mosimann, S., Meleshko, R., James, M. A critical assessment of comparative molecular modeling of tertiary structures in proteins. Proteins: Struct., Funct., Genet. 23:301–317, 1995.

[4] Bleasby, A., Wootton, J. Construction of validated, non-redundant composite protein sequence databases. Protein Eng. 3:153–159, 1990.

[5] Lipman, D., Pearson, W. Rapid and sensitive protein similarity searches. Science 227:1435–1441, 1985.

[6] Murzin, A., Brenner, S. E., Hubbard, T. J. P., Chothia, C. SCOP: Structural Classification of Proteins. <`http://www.bio.cam.ac.uk/scop/`>, 1997.

[7] Barton, G. J. Protein multiple sequence alignment and flexible pattern matching. Methods Enzymol 183:403–428, 1990.

[8] Barton, G. J., Sternberg, M. J. E. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. J. Mol. Biol. 198:327–337, 1987.

[9] Pedersen, J. T. G, a molecular modelling program available upon request to the author.

[10] McLachlan, A. Gene duplication in the structural evolution of Chymotrypsin. J. Mol. Biol. 128:49–79, 1979.

[11] MSI. DISCOVER and INSIGHT are trademarks of Biosym Technologies, San diego, California, USA.

[12] MSI. QUANTA is a trademark of MSI Technologies.

[13] Fidelis, K., Moult, J. Unpublished.

[14] Toner, M., Moult, J. Unpublished.

[15] Moult, J., James, M. N. G. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins: Struct., Funct., Genet. 2:146–163, 1986.

[16] Pedersen, J. T., Moult, J. Ab initio structure Prediction for small polypeptides and protein fragments using genetic algorithms. Proteins: Struct., Funct., Genet. 23:454–460, 1995.

[17] Bruccoleri, R. E., Karplus, M. Chain closure with bond angle variation. Macromolecules 18:2767–2773, 1985.

[18] Bruccoleri, R. E., Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.

[19] Martin, A., Cheetham, J., Rees, A. Modelling antibody hypervariable loops: A combined algorithm. Proc. Natl. Acad. Sci. USA 86:9268–9272, 1989.

[20] Herzberg, O., Moult, J. Analysis of steric strain in the polypeptide backbone of protein molecules. Proteins: Struct., Funct., Genet. 11:223–229, 1991.

[21] Risler, J., Delorme, M., Delacroix, H., Mevarech, M. Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J. Mol. Biol. 204:1019–1029, 1988.

[22] Read, J., Brayer, G., Jurášek, L., James, M. Critical evaluation of comparative model building of *Streptomyces griseus* trypsin. Biochemistry 23:6570–6575, 1984.

[23] Holm, L., Sander, C. Database algorithm for generating protein backbone and side-chain co-ordinates from a C_α trace: application to model building and detection of co-ordinate errors. J. Mol. Biol. 218:183–194, 1991.

[24] Ponder, J., Richards, F. Tertiary templates for proteins: use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.

[25] Wilson, C., Gregoret, L., Agard, D. Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. J. Mol. Biol. 229:996–1006, 1993.

[26] Fidelis, K., Stern, P., Bacon, D., Moult, J. Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng. 7:953–960, 1994.

[27] Greer, J. Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins: Struct., Funct., Genet. 7:317–334, 1990.

[28] Unger, R., Moult, J. An analysis of protein folding pathways. Biochemistry 30:3816–3823, 1991.