# MULTIPLE MODEL APPROACH — DEALING WITH ALIGNMENT AMBIGUITIES IN PROTEIN MODELING

K. PAWŁOWSKI
*Institute of Biochemistry & Biophysics, Polish Academy of Sciences, Warszawa, Poland*

Ł. JAROSZEWSKI
*Department of Chemistry, Warsaw University, Warszawa, Poland*

A. BIERZYŃSKI
*Institute of Biochemistry & Biophysics, Polish Academy of Sciences, Warszawa, Poland*

A. GODZIK
*The Scripps Research Institute, 10666 N.Torrey Pines, La Jolla, CA 92037, USA*

Sequence alignments for distantly homologous proteins are often ambiguous, which creates a weak link in structure prediction by homology. We address this problem by using several plausible alignments in a modeling procedure, obtaining many models of the target. All are subsequently evaluated by a threading algorithm. It is shown that this approach can identify best alignments and produce reasonable models, whose quality is now limited only by the extent of the structural similarity between the known and predicted protein. Using a similar approach structure prediction for the oxidized dimer of S100A1 protein, for which the structure is not known, is presented.

## 1. Introduction

Despite intense efforts we are not able to predict the three dimensional structure of a protein from its amino acid sequence alone. Comparative modeling, the only consistently successful prediction method, is based on a simple observation that proteins with similar sequences fold to similar structures. In this technique, an unknown structure of a new protein (the target) is predicted on the basis of a sequentially related known structure (the template).[1, 2]

The modeling process consists of template identification, creation of an alignment and building of a model. For homology modeling the first step is trivial, or at least much easier than others. The creation of the alignment is the most crucial step. Any errors made at this stage are usually impossible to correct later and lead to significant errors in final models.[3] At the same time, inherent problems with the alignment preparation are often unappreciated.[4] In this contribution we will focus on this stage in the modeling process.

Efficient algorithms for sequence alignments were developed many years ago [5] and are now widely available. They require two groups of parameters: a scoring matrix and penalties for introducing gaps into the alignment. Unfortunately no optimal set of parameters has been identified to work well for different proteins. [4]

If the target and the template proteins have over 50% of identical amino acids, the alignment is trivial. With protein sequence similarity between 50 and 30% significant

differences between different alignments emerge in some regions. But in the protein core the alignment is still well defined and doesn't depend on parameters used. The situation changes again, when the sequence similarity drops below 30%. Alignments become very ill defined, changing dramatically with scoring matrices and gap penalties.[4] Ambiguities are no longer confined to loops, but involve protein core. In such situations, although comparative modeling in principle could work properly, models can be useless due to alignment errors.

For protein pairs where both structures are known, alignment strategies could be investigated with the help of structural alignments.[4, 6] Unfortunately, while some sets of parameters are better than others, this is true only on the average.[4] For any single case we cannot be sure that strategy which is statistically the best would actually produce the „true" alignment. At the same time, the approach suggested in modeling packages' manuals is to prepare the „best possible alignment" and to stake everything on the resulting model. In this contribution we propose a simple method, the multiple model approach, which intends to improve on this strategy. It consists of the following steps:

1. exploration of the alignment diversity, by using various alignment protocols
2. elimination of alignments contradicting our knowledge of the target.
3. construction of full-atom models using all the remaining alignments
4. choice of the best model

In this contribution quality of the models (point 4 above) is evaluated by a threading algorithm.[7] Similar strategy was used by other groups to assess quality of crystallographical models.[8] Self-threading energy of a full atom protein model was used in evaluating alignment strategies for threading alignments [9] and to analyze interactions in a family of structurally divergent homologous proteins.[10] A related approach based on the measurement of core volume and packing pair potential was used to compare suboptimal sequence alignments for difficult to align protein pairs.[11]

## 2 Methods

***The multiple model method.*** The accuracy of sequence alignment depends strongly on the substitution matrix used and vary significantly for different gap penalties.[4] If the real structure of the target is not known, the selection of the correct alignment is not trivial. Several methods have been suggested to identify a part of the alignment one could be certain about,[6] but such methods focus only on a small fraction of the alignment.

Here, to explore the range of possible alignments, we use five different scoring matrices, judged to be among the best in a comprehensive test.[4]

***The test case 1: EETI and PCI.*** The two small disulfide-rich proteins, trypsin inhibitor (EETI, 2ETI), and potato carboxypeptidase inhibitor (PCI, 4CPA) have weak, but detectable homology, although they inhibit different proteases. Although, as

expected, the proteins share similar structural scaffold, the binding sites are located at different positions in the structures. In EETI the trypsin-binding loop consists of residues Gly1- Met7 and the scissile peptide bond is situated between Arg4 and Ile5. In PCI the active site of carboxypeptidase is blocked by the last four residues. Early alignments of this protein pair were partly in error since only two of three disulfide bridges are structurally conserved between the two proteins.[12]

***The test case 2: Calbindin and parvalbumin***. Calbindin (4ICB) and parvalbumin (5CPV) belong to the EF-hand calcium binding protein family.[13] These all-alpha proteins are built of ion binding motifs named EF-hands, usually combined in pairs to form a functional unit. Calbindin contains one such unit, even that the first EF-hand motif differs significantly from other such motifs, with a two residue insertion and a missing aspartic acid (conserved in all other proteins from this family) at the N-terminal end of the calcium binding loop. Prior to calbindin structure determination it was speculated [14] than the organization of both motifs might be significantly different that the one in parvalbumin, which was the first known protein from this family. Parvalbumin has three motifs, because apart from the typical EF-hand unit it contains an additional motif at the N-terminus. This motif is inactive due to several mutations, but significant sequence similarity between it and other motifs is retained. It is widely believed that both proteins evolved by gene duplication, with two copies present in calbindin and three in parvalbumin. Differences between disparate alignment procedures in this case arise from attempts to align the inactive motif 1 of parvalbumin with modified motif 1 of calbindin, instead of the structurally correct alignment: motif 2 of parvalbumin with motif 1 of calbindin.

***Calculation of the alignments.*** Four substitution matrices were selected from the best scoring substitutions matrices recommended by Argos and co-workers.[4] The Dayhoff matrix as modified in the GCG package,[15] was also included, identified as GCG. Notation for the substitution matrices as well as the optimal parameter values were adapted from.[4] The alignments were determined using the GAP program from GCG package.[15]

***Structure prediction: S100A1 protein.*** S100A1 is a homodimer containing a pair of EF-hand motifs in every chain.[13] S100A1 contains one cysteine residue per chain and occurs in oxidized and reduced forms. The three-dimensional structures of only two proteins closely homologous to S100A1 are known so far: calbindin [16] and calcyclin.[17] Neither of these structures can be used as a template for the oxidized S100A1. Calbindin is monomeric while in calbindin the residues corresponding to S100A1 cysteines are far away from each other, what makes formation of a disulfide bridge impossible without much structural change. This is supported by experimental data on cysteine oxidation [18] and ion binding [Goch and Bierzynski, unpublished data]. In such a situation other structural templates from the EF-hand family are sought. In this case alignment ambiguity stems from the possibility of changing the order in which four EF-hand motifs from a S100A1 dimer are threaded into the template structures (see below).

***Structure prediction: modeling templates and sequence - structure alignments*** For modeling of the oxidized S100A1 dimer structures of bovine recoverin (PDB code 1REC) and SCaBP (sarcoplasmic $Ca^{2+}$-binding protein from sandworm; PDB code 2SCP) were chosen as templates. Only these two structures contain four EF-hand motifs (as S100A1 dimer) and at the same time have compact conformations allowing the formation of the interchain disulfide bridge without much structural rearrangement. Other known four motif EF-hand calcium proteins have extended conformations with little or no interactions between units.

Structures of all the four EF-hand units present in the templates are very similar, with $C_\alpha$ RMSD on the order of 2-3 Å. Still, due to different unit packing, the overall structures are completely different.[10] Therefore the two structures could be used as distinct templates for modeling. Only the EF-hand motifs were used as templates, because inter-motif linkers and terminal segments vary in length. In the modeling studies the sequence of the bovine S100A1 [19] was used.

At first, all the possible orders in which four EF-hand motifs from S100A1 can be threaded into the template structures were taken into account. Then only these alignments were retained that allowed EF-hand motifs from one S100A1 chain form a standard EF-hand unit. Given this condition there are four non-equivalent alignments of S100A1 dimer within a four-motif template (see Table 2). In order to account for the fact that S100A1 homodimer should be symmetric, only those alignments were used that allowed for symmetry on the level of motif - motif contacts, if not on the level of contacts between individual residues.

The alignments of S100A1 sequence to the recoverin and SCaBP template structures were done in three steps: **1°** S100A1 chain sequence was aligned with the calbindin sequence using the Dayhoff PAM 250 matrix. Given the high sequence identity, other scoring matrices gave almost identical alignments. **2°** a contact map overlap algorithm [20] was used to produce structural alignments of all the EF-hand motifs from recoverin and SCaBP with both EF-hand motifs of calbindin. **3°** the results of the first two steps were combined to obtain alignments of an S100A1 dimer with the templates' structures.

Since modeling the N- and C- terminal segments outside the EF-hand motifs would be arbitrary, only residues Thr 10 - Cys 85 were modeled. The chain segments linking the N- and C-terminal motifs were allowed to relax the strains caused by different alignments. The alignments used are schematically shown in Table 2.

***Automated modeling, analysis of models' energy and structural similarity*** One of recent automated modeling methods was used, implemented in MODELLER.[21] Standard MODELLER routine 'homol' was used. Oxidized structures were modeled with additional disulfide restraint.

The energy parameters, developed for the topology fingerprint threading algorithm [7] were used. The structures of proteins that were modeled in this study were not used for the derivation of energy parameters. The term „energy" is often used here instead of „score", which does not mean that scores can be rigorously treated as real

physical energies. The „energy" units roughly correspond to the value of one kT in room temperature.[22]

Two measures of structural similarity are used: root mean square deviation between equivalent $C_\alpha$ positions (RMSD) and the contact map overlap.[20, 23]

For pairs of two native protein structures final RMSD was calculated after superposing the two structures by minimization of $C_\alpha$ RMSD whereas equivalent pairs of $C_\alpha$ atoms were defined by the given sequence alignment.

The quality of sequence alignments was measured by calculating the structural similarity (measured as contact map overlap), between the target and the template according to each alignment. This strategy was used, because it was recently discovered [23, 24] that structural alignments are very ill-defined, with very high density of high-scoring, alternative alignments. Therefore, comparing the alignment to a single „standard of truth" alignment would bias the procedure to an arbitrary alignment chosen from the group of high scoring alignments. This problem becomes serious only for very distant homologues, therefore it is unlikely to change the results of the gap parameters optimization [4] which was primarily done on highly homologous proteins.

## 3 Results

### *3. 1 Test cases: alignments*

For both pairs of proteins studied here, several different alignments were obtained (see Figures 1, 2). The differences between the alignments were quite high: sequence identity reported by various methods ranged from 21 to 32 % for the EETI / PCI pair and from 26 to 34 % for calbindin and parvalbumin. Alignment ambiguity is even more striking on the structural level: various sequence alignments imply structural alignments ranging from 19 to 44% contact map overlap for EETI and PCI and from 26 to 39 % for calbindin and parvalbumin (see Table 1). It means that the number of conserved inter-residue contacts can change by the factor of 2 depending on the alignment procedure used. Also the more commonly used measure of structural similarity, $C_\alpha$ RMSD, shows variation (see Table 1) corresponding to a broad range of structural alignments - from acceptable values to practically meaningless 8 Å in the case of calbindin and parvalbumin. The latter value indicates that while some regions might have been aligned correctly, some gross misalignments occurred (see below).

```
A)
EETI residues 1-28
..GCPRIL.MRCKQDSDCLAGCVC....GPNGFCG... b_74, str
..GCPRIL.MRCKQDSDCLAGCVCGP..NG..FCG... b_62
..GCPRIL.MRCKQDSDCLAG..C.VC.GPNGFCG... gonnet
..GCPRIL.MRCKQDSDCLAGCVCGPNGFCG.......          GCG
..G.CPRILMRCKQDSDC.LAG..CVCGPN.GFCG... [12]

PCI residues 1-37
EHADPICN.KPCKTHDDCSGAWFCQACWNSARTCGPYV


B)
calbindin, residues 1-26
.....................................MKSPEELKGIFEKYAAKEGDPNQLSK  b_74,GCG
..................MKSPEEL..KGIFEK..YAAKEGDP..........NQ.....LSK  b_62
..................MKSPEEL..KGIFEK..............YA..AKEGDPNQ.LSK  gonnet
...................MKSPEELKGIFEKYAAKEGDPNQLSK.................  str

parvalbumin, residues 1-60
AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFA..IIDQDKSGFIEE


calbindin, residues 27-76
EELKLLLQ...TEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ       b_74, GCG
EELKLLLQ...TEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ       b_62
EELKLLLQ...TEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ       gonnet
EELKLLLQT.EFPSLLKGPSTL..DELFEELDKNGDGEVSFEEFQVLVKKISQ       str

parvalbumin, residues 61-108
DELKLFLQNFKADARALTDGET..KTFLKAGDSDGDGKIGVDEFTALVKA...
```

Figure 1: Sequence alignments obtained using different substitution matrices.
For both cases the first lines contain the shorter sequence as aligned using the substitution matrix indicated on the right. The template sequence is shown below. A) EETI/PCI; B) calbindin/parvalbumin.

The differences in alignments are not limited to outside loops, but extend to the core regions (see Figure 2). This is well documented by another quantity, the buried contact map overlap (see Table 1). It is calculated as the ordinary contact map overlap, but only contacts between buried residues are included. Different values of buried contact map overlap obtained for various alignments mean that those different fractions of protein cores were aligned correctly.

The benner74 (denoted b_74 in the Figure 1) and str substitution matrices produced identical EETI/PCI alignments, however different from the optimal one. The mere fact that two different substitution matrices give identical alignment does not imply this is the best one. All the alignments of these proteins agree in the N-terminal part of the sequence (residues 1-20; see Figures 2 A, B). In the C-terminal part of the sequence for every position at least three different alignments were obtained. Although these differences didn't involve more than a shift of four residues, they were sufficient to generate erroneous models with visibly higher energy.

```
A)
EETI residues 1-28
..GCPRIL.MRCKQDSDCLAGCVC....GPNGFCG... b_74, str
..GCPRIL.MRCKQDSDCLAGCVCGP..NG..FCG... b_62
..GCPRIL.MRCKQDSDCLAG..C.VC.GPNGFCG... gonnet
..GCPRIL.MRCKQDSDCLAGCVCGPNGFCG.......          GCG
..G.CPRILMRCKQDSDC.LAG..CVCGPN.GFCG... [12]

PCI residues 1-37
EHADPICN.KPCKTHDDCSGAWFCQACWNSARTCGPYV


B)
calbindin, residues 1-26
.................................MKSPEELKGIFEKYAAKEGDPNQLSK  b_74,GCG
.................MKSPEEL..KGIFEK..YAAKEGDP..........NQ.....LSK  b_62
.................MKSPEEL..KGIFEK..............YA..AKEGDPNQ.LSK  gonnet
..................MKSPEELKGIFEKYAAKEGDPNQLSK.................  str

parvalbumin, residues 1-60
AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFA..IIDQDKSGFIEE


calbindin, residues 27-76
EELKLLLQ...TEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ      b_74, GCG
EELKLLLQ...TEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ      b_62
EELKLLLQ...TEFPSLLKGPSTLDELFEELDKNGDGEVSFEEFQVLVKKISQ      gonnet
EELKLLLQT.EFPSLLKGPSTL..DELFEELDKNGDGEVSFEEFQVLVKKISQ      str

parvalbumin, residues 61-108
DELKLFLQNFKADARALTDGET..KTFLKAGDSDGDGKIGVDEFTALVKA...
```

Figure 1: Sequence alignments obtained using different substitution matrices.
For both cases the first lines contain the shorter sequence as aligned using the substitution matrix indicated on the right. The template sequence is shown below. A) EETI/PCI; B) calbindin/parvalbumin.

The differences in alignments are not limited to outside loops, but extend to the core regions (see Figure 2). This is well documented by another quantity, the buried contact map overlap (see Table 1). It is calculated as the ordinary contact map overlap, but only contacts between buried residues are included. Different values of buried contact map overlap obtained for various alignments mean that those different fractions of protein cores were aligned correctly.

The benner74 (denoted b_74 in the Figure 1) and str substitution matrices produced identical EETI/PCI alignments, however different from the optimal one. The mere fact that two different substitution matrices give identical alignment does not imply this is the best one. All the alignments of these proteins agree in the N-terminal part of the sequence (residues 1-20; see Figures 2 A, B). In the C-terminal part of the sequence for every position at least three different alignments were obtained. Although these differences didn't involve more than a shift of four residues, they were sufficient to generate erroneous models with visibly higher energy.

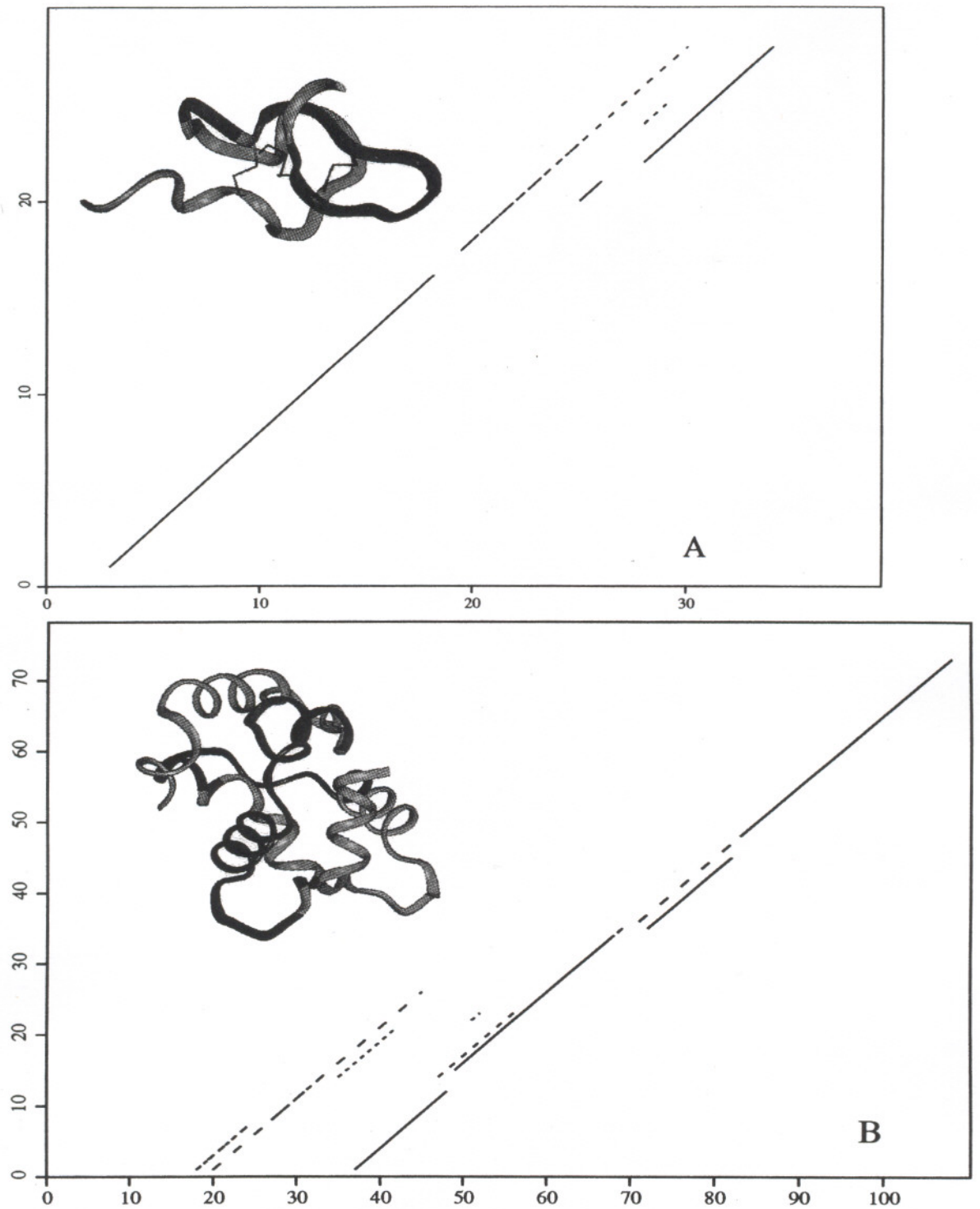


Figure 2: Regions where alignments diverge

Dot-matrix representations of alignments for the EETI/PCI [A] and calbindin/parvalbumin [B] pairs. Best alignments are shown as continuous lines. Other alignments, where different from the best ones, are shown as dashed lines. Ribbon representations of PCI and parvalbumin structures are shown in the insets. The regions where differences in alignments occur are shown in black.

**Table 1**: Model characteristics. t/t: target/template comparison; t/m: target/model comparison For target/template contact map overlap the values in parentheses were calculated for the buried residues only.

| Substitution matrix | sequence identity | model energy | MODE-LLER score | $C_\alpha$ RMSD [Å] | | contact map overlap | |
|---|---|---|---|---|---|---|---|
| | | | | t/t | t/m | t/t | t/m |
| EETI/PCI | | | | | | | |
| benner74 | 28.6 | -0.4 | 244 | 3.6 | 3.2 | 33 (47) | 39 |
| blosum62 | 32.1 | 5.7 | 288 | 4.0 | 3.8 | 27 (47) | 34 |
| gonnet | 32.1 | -2.2 | 200 | 3.1 | 2.9 | 44 (56) | 58 |
| GCG | 21.4 | 11.7 | 197 | 5.3 | 5.3 | 19 (28) | 23 |
| str | 28.6 | -0.4 | 244 | 3.6 | 3.2 | 33 (47) | 39 |
| structural [12] | 28.6 | -2.0 | 164 | 2.6 | 2.5 | 33 (44) | 62 |
| calbindin / parvalbumin | | | | | | | |
| benner74 | 30.4 | -18.5 | 420 | 4.4 | 4.2 | 39 (54) | 51 |
| blosum62 | 33.8 | -16.1 | 595 | 8.9 | 8.2 | 32 (49) | 41 |
| gonnet | 32.4 | -12.4 | 680 | 8.1 | 7.4 | 34 (50) | 49 |
| \|GCG | 30.4 | -18.5 | 420 | 4.4 | 4.2 | 39 (54) | 51 |
| str | 26.0 | -12.3 | 554 | 8.8 | 8.9 | 26 (26) | 36 |

In the case of calbindin and parvalbumin identical alignments were produced using benner74 and GCG substitution matrices. The calbindin / parvalbumin alignments differ, as expected, mostly in the 1-40 region (the inactive EF-hand motif) in parvalbumin sequence (see Figure 2), which is erroneously aligned by most substitution matrices with parts of the first EF-hand motif of calbindin (with a shift of ca. 20 residues relative to the best alignment). This error had to propagate, so large gaps are aligned with the second parvalbumin EF-hand motif (region 40-80; see Figure 1B) and differences end only at the beginning of the third motif. Only the 26 positions at the C-terminus are aligned in the same manner by all the substitution matrices used. This case may seem to be more trivial than the first one since some knowledge on sequence patterns of EF-motifs could allow to discard the wrong alignments. Such knowledge is not possible however without first knowing the target calbindin structure.

*3. 2 Test cases: models*

Protein models obtained starting from alignments as divergent as described above, are not surprisingly very different one from another. It is very encouraging that in both test cases the models picked up by the energy score criterion are those that are closest to the native target structure (see Table 1). Moreover, the models judged to be worst by this criterion are those that differ most from the native structure. It was checked that all the energy terms used here are important for selection of the most correct model.

The differences in energies between the models are largest for those residues that are misaligned relative to the best alignment (data not shown). In „bad" EETI models residues with high energy are found mostly in the C-terminal region, where alignments differ. Interestingly, residues with high energy are found also in the N-terminal part of some bad EETI models. In this region all the alignments are identical, so differences in energy of a residue reflect propagation of alignment errors. Similar observations can be made for calbindin models.

In the two test cases studied, different mutation matrices gave good and bad alignments. In the case of EETI and PCI the gonnet substitution matrix produced the best alignment while the GCG gave the worst. In the case of calbindin and parvalbumin the best alignments were obtained using the benner74 and GCG substitution matrices, and the worst one using the str matrix.

The best model of EETI, built without the knowledge of the native structure, is nearly as accurate as the model based on the previously published alignment that was prepared using the EETI structure [12] (See Table 1).

An internal measure of model quality supplied by the MODELLER program,[21] did not always allow to distinguish the most correct model (see Table 1). In one case (calbindin model) it selected the best one, while in the other case (EETI model) - the worst. It is not surprising though, since this function doesn't describe physical quality of the model, only degree of satisfaction of the restraints imposed.

There is a difference in sensitivity between contact map overlap and $C_\alpha$ RMSD. In case of calbindin models compared to the native structure RMSD varies from 4.4 to 8.9 Å whereas contact map overlap varies only from 36 to 51 %. This reflects the fact that the former measure is more global - it is not sensitive to regions of weak structural similarity and changes of relative positions of otherwise similar protein segments.

*3. 3 Predicted structure of the oxidized S100A1 protein*

Energies of models (see Table 2) suggest that the oxidized S100A1 dimer prefers the recoverin fold. Some of the scores however may be partly affected by the difficulty of modeling the segments linking the EF-hand motifs and the terminal „tails".

The best model of the oxidized S100A1 follows the general fold of recoverin, and the two chains occupy place of recoverin EF-hand motifs. S100A1 chains are linked by the disulfide bridge and stabilized by hydrophobic interactions between several interface residues. The placement of chain segments linking the motifs, which is different than in recoverin, does not distort the recoverin-like structure.

The model proposed here represents S100A1 in the calcium-loaded state. Although calcium ions were not present in the modeling process, they were included *implicite* since the templates were structures with ion-binding sites occupied.

**Table 2:** Energy for the models of S100A1 dimer. Different alignment modes are represented in column 1. EF-hand motifs from templates are numbered from the N-terminal. For the S100A1 dimer N1 denotes N-terminal motif from first S100A1 chain, etc. Empty cells represent models that were not considered due to the symmetry requirement (see the „Methods" section).

| Alignment of EF-hand motifs | | | | | Template | |
|---|---|---|---|---|---|---|
| | | | | | Recoverin | SCaBP |
| S100A1 motifs | N1 | C1 | N2 | C2 | | +4.1 |
| template motifs | 1 | 2 | 3 | 4 | | |
| | N1 | C1 | C2 | N2 | **-17.6** | |
| | C1 | N1 | N2 | C2 | -6.1 | |
| | C1 | N1 | C2 | N2 | | +22.6 |

## 4 Discussion

In this work we proposed a novel approach to the problem of alignment ambiguities which is encountered in the initial stages of comparative protein modeling. The importance of the alignment choice can be seen in some of the wrong models obtained in this study. These models, although prepared using state-of-the art alignment methods that statistically perform well, contain errors as severe as completely incorrect ion binding site in the case of calbindin. It was shown, that in both test cases studied, our criterion of model energy can correctly recognize the best model. Using an automated modeling procedure and analyzing resulting models with the statistical potential allows one to worry less about quality of the alignment. This approach can be extended, as presented on the example of S100A1 structure prediction, to using different structural templates if it is not clear a priori, which is the most appropriate for modeling.

The manuals of comparative modeling programs stress the importance of the sequence alignment and ask the user to prepare it with utmost care. The author of the MODELLER package goes even further and suggests an iterative approach, i. e. careful analysis of a model produced using an initial alignment, reexamination and

correction of the alignment and producing an improved model.[21] The approach presented here can be viewed as a practical realization of these suggestions, but construction of many alignments and models can be done simultaneously what allows for partial automation of the process.

At the stage of alignment construction separated fragments of target sequence are equivalenced to fragments of template structure. The construction of the model can be regarded as a test of geometric feasibility of the connectivity between aligned target sequence fragments. At this point one may wonder, what actual characteristics of the model give raise to good/bad self-threading score and if it would be possible to incorporate them into the threading program itself. The search for such features was an initial objective of the research described here and it remains our ultimate goal. Unfortunately, so far we were not able to identify the high energy fragments of the model in a way which could be utilized on the alignment stage. Thus, model building remains a slow, but necessary ultimate test for the accuracy of the alignment. It should be stressed, that even if a relatively small number of alignments can be tested and compared, each of these alignments come from the well optimized alignment procedure.

The approach described here attempts to explore the conformational space more widely than standard modeling techniques, in recognition of the fact that constructing just one model requires one to be able to formulate very decisive assumptions about the modeled structure. Proposing a number of models for subsequent theoretical evaluation and experimental verification is a step towards more reliable modeling in situations, where substantially homologous structures can be identified, but it is not clear which of them to use and how to use them. Still, this approach is limited to testing and comparing models, which all come from within the space of already known protein topologies.

## Acknowledgments

We thank A.Koliński for stimulating discussions. We thank Dr. W. Chazin for making the calcyclin structure available to us prior to its release. We thank Dr A. Šali for kindly providing us with his MODELLER program. Support by the following grants is acknowledged: NIH GM-48835 (AG), Howard Hughes Medical Institute grant 75195-53402 (£J), Polish Committee for Scientific Research grant KBN-4111789101 (AB and KP).

## References


1. Thornton, J.M. and M.B. Swindells, in *Molecular Structures in Biology*, R. Diamond, Editor. 1993, Oxford Univerity Press: Oxford.
2. Johnson, M.S., *et al. Crit. Rev. Biochem. & Mol. Biol.* **29,** 1 (1994)
3. Karplus, M. and A. Sali. *Current Opinion Struct Biol.* **5,** 58 (1995)
4. Vogt, G., T. Etzold, and P. Argos. *J. Mol. Biol.* **249,** 816 (1995)
5. Needelman, S.B. and C.D. Wunsch. *J. Mol. Biol.* **48,** 443 (1970)
6. Vingron, M. and P. Argos. *Prot. Engineer.* **3,** 565 (1990)
7. Godzik, A., J. Skolnick, and A. Kolinski. *J. Mol. Biol* **227,** 227 (1992)
8. Luethy, R., J.U. Bowie, and D. Eisenberg. *Nature* **356,** 83 (1992)
9. Jaroszewski, L. and A. Godzik. *Prot. Engineer.* submitted (1996)
10. Pawlowski, K., A. Bierzynski, and A. Godzik. *J. Mol. Biol.* **258,** 349 (1996)
11. Saqi, M.A., P.A. Bates, and M.J. Sternberg. *Prot. Engineer.* **5,** 305 (1992)
12. Chiche, L., *et al. Prot. Engineer.* **6,** 675 (1993)
13. Kawasaki, H. and R. Kretsinger. *Protein Profile* **1,** 343 (1994)
14. Kretsinger, R.H. *CRC Crit. Rev. Biochem.* **8,** 119 (1980)
15. Genetic Computer Group. Program Manual for the GCG Package, Version 7 (1991)
16. Svensson, L.A., E. Thulin, and S. Forsen. *J. Mol. Biol.* **223,** 601 (1992)
17. Potts, B.C., *et al. Nature Str. Biol.* **2,** 790 (1995)
18. Mely, Y. and D. Gerard. *J. Neurochem.* **55,** 1100 (1990)
19. Bairoch, A., *SwissProt*, . 1994.
20. Godzik, A., J. Skolnick, and A. Kolinski. *Prot. Engineering* **6,** 801 (1993)
21. Sali, A. and J.P. Overington. *Protein Science.* **3,** 1582 (1994)
22. Godzik, A., A. Kolinski, and J. Skolnick. *Protein Science* **4,** 2107 (1995)
23. Godzik, A. *Protein Science* **5**, 1325 (1996)
24. Zu-Feng, F. and M. Sippl. *Folding & Design* **1,** 123 (1996)