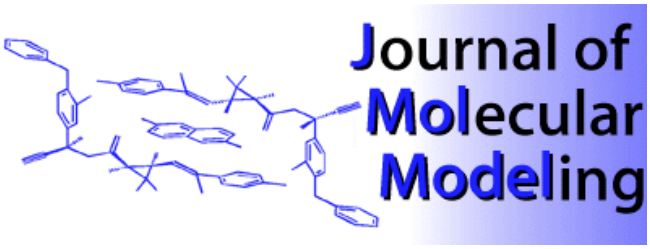
J. Mol. Model. **1998**, *4*, 294 – 309



FULL PAPER

Multiple Model Approach: Exploring the Limits of Comparative Modeling

Łukasz Jaroszewski[1], Krzysztof PawŁowski[2], and Adam Godzik[2]

[1]Department of Chemistry, University of Warsaw, Warszawa, Poland

[2]Department of Molecular Biology TPC-5, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037 USA. Tel: (619) 784 -8823; Fax (619) 784-8895. E-mail: adam@scripps.edu

Received: 10 July 1998 / Accepted: 18 September 1998 / Published: 1 October 1998

Abstract One of the biggest problems in modeling distantly related proteins is the quality of the target-template alignment. This problem often results in low quality models that do not utilize all the information available in the template structure. The divergence of alignments at a low sequence identity level, which is a hindrance in most modeling attempts, is used here as a basis for a new technique of Multiple Model Approach (MMA). Alternative alignments prepared here using different mutation matrices and gap penalties, combined with automated model building, are used to create a set of models that explore a range of possible conformations for the target protein. Models are evaluated using different techniques to identify the best model. In the set of examples studied here, the correct target structure is known, which allows the evaluation of various alignment and evaluation strategies.
For a randomly selected group of distantly homologous protein pairs representing all structural classes and various fold types, it is shown that a threading score based on simplified statistical potentials of mean force can identify the best models and, consequently, the most reliable alignment. In cases where the difference between target and template structures is significant, the threading score shows clearly that all models are wrong, therefore disqualifying the template.

Keywords Comparative modeling, Distantly related proteins, Alignment ambiguity

Running title Multiple Model Approach

Introduction

Comparative modeling, the most reliable protein structure prediction method in existence, is based on the remarkably simple but far-reaching observation that proteins with similar sequences fold to similar structures [1]. To identify this similarity, the well-developed tools of protein sequence analysis are used. They yield a quantitative measure of sequence similarity (score) and a position-by-position equivalence between the two proteins, i.e., the sequence alignment [2]. Such an alignment could be used subsequently to build a three-dimensional model. This paradigm forms the basis of the very active comparative modeling field [3-9]. Building the three-dimensional model of a new protein when only the sequence is known (we call this protein a prediction **target**) can be divided into three logical steps:

- identifying an appropriate protein **template** (a homologous protein whose structure has already been solved)

Correspondence to: A. Godzik

- calculating the alignment of the target sequence with this template
- building the three dimensional model.

Comparative protein modeling is a method of very practical importance with many algorithms available as commercial [10-12] packages or on Web sites [13]. Also, some commercial programs are available for free for the academic community [14, 15].

The first stage of a modeling process is identifying a template, for instance, by a standard sequence similarity search in a database containing sequences of all proteins with known three-dimensional structures. The step is often considered trivial. Indeed, if a closely related template protein is present in the database of known structures, any of the large number of publicly available programs can be used to find it. However, if such template does not exist, problem becomes difficult. In recent years, a large effort was put into developing algorithms that are able to recognize protein homologies and structural similarities beyond the level of easily identifiable homology [16-19]. A success of the fold prediction algorithms greatly increased the range of application of the comparative modeling technique but, at the same time, because the structural similarity between distant homologues is smaller, increased the level of difficulty on the two subsequent steps.

In the next step, which is often merged with the template recognition step, residues of the target sequence are assigned to the residues of the template, i.e. the target-template alignment is prepared. In the third step, a model of the target protein is built on the scaffold of the template structure by side-chain repacking and linking the fragments of aligned structure [9, 20]. This step is usually done with the use of specialized modeling software [10, 12, 21]. At this final stage, it is usually impossible to correct errors that result from mistakes made in the earlier steps [22]. Consequently, problems with the alignment step remain the main reason for errors in the comparative modeling and often result in comparative modeling not taking full advantage of the extent of structural similarity between the target and the template [23].

The inherent problems with the alignment preparation are often unappreciated. This is probably because in most applications, only closely homologous proteins are being compared where the alignment is unique and easy to find. In recent years, there have been very few serious attempts to address the problem of alignment quality and those have concentrated on the problem of optimal alignment parameters [24] or identification of the reliable fragments of the alignment [25, 26]. Standard algorithms can efficiently calculate the score of the best alignment between two proteins and provide one alignment (from many) having this score. What they don't tell is how many different alignments have scores identical to or very close to the optimal one, and how different these alignments are. The information about suboptimal alignments [27, 28] gives some hints about how unique the alignment is. It could be used also to identify the regions of both molecules where the similarity is strongest, which sometimes coincides with the most reliable fragments of the alignment [26]. Unfortunately, the information about the suboptimal alignments is not available from any of the popular "canned" alignment programs, which may also contribute to the lack of appreciation of the "alignment ambiguity problem".

The alignment protocol based on the dynamic programming algorithm [29, 30] requires two types of parameters: a substitution matrix and the penalties for introducing gaps into the alignment. The first group of parameters yields information about the probability of one amino acid being replaced by another via mutation. The derivation of mutation matrices can rely on various information sources, such as structural equivalence of numerous protein sequences, genetic code similarity, chemical similarity of amino acids, hydrophobicity index, physical property indices, main chain folding angles, contact potential and neighborhood selectivity [31]. Over forty various mutation matrices have been derived by various groups and most were recently collected, systematically analyzed and made publicly available on the Internet [31]. The second type of parameter is the gap introduction and extension penalty [2], which have no clear physical interpretation and are usually determined by trial and error.

The alignment methods and parameters can be tested for protein pairs where both structures are known. For instance, the structural alignment can be treated as the standard of truth [24, 25] and various sequence alignment strategies can be evaluated by their ability to reproduce it. Argos and coworkers used this approach in a comprehensive test of different mutation matrices and gap penalties [24]. They tested as many as 41 mutation matrices, optimizing the gap penalty parameters for each matrix. It was shown that the quality of the alignment dramatically depends on the type of mutation matrix and the values of the gap penalties used, which must be optimized separately for each mutation matrix. Unfortunately, as shown later in this paper, the distribution of alignment accuracy for different alignment methods can be very large. A method of alignment calculation, which is optimal in the statistical sense, could actually produce very wrong alignments in a number of specific cases.

The alignment is usually trivial if the sequence identity of the target and template is higher than 50% of identical amino acids. With protein sequence identity ranging from 30 to 50%, significant shifts between different alignments emerge in some regions, mostly loops. When sequence identity drops below 30%, the alignments become very unstable, changing dramatically with scoring matrices and gap penalties [24] and they become essentially random for structurally similar proteins with undetectable sequence similarity [32-34]. It is important to point out that the increasing alignment ambiguity with decreasing sequence similarity is not a technical problem that could be overcome by improving alignment algorithms. It reflects two partly independent problems, both of which are related to the fundamental features of the sequence-structure relationship in proteins. The first is that the extent of sequence similarity decreases with increasing evolutionary distance between proteins, and at a certain point detectable sequence similarity is confined only to a small part of both molecules. This problem could be addressed by changing the definition of sequence similarity and using some structural information in the scoring function, as it is done by

Table 1 *Alignment and model characteristics for test cases of comparative modeling. Values calculated using the correct target structure are highlighted. Such values require the correct answer to be known and are used here only to evaluate various modeling strategies*

case		**The alignments**					**The models**			
target template	mutation matrix	z-score	sequence identity	contact map overlap [a]	Cα RMSD [a]	contact map overlap [b]	Cα RMSD [b]	threading energy	3D profiles [c]	-ln(pdf) Modeller [d]
1.	benner 74 +	-7.1	16.3	54	3.07	63	2.87	-0.109	52.06	594
2hhdA	blossum62 +	-6.4	16.3	50	3.53	56	3.30	-0.053	43.73	605
2gdm	gonnet +	-7.4	16.3	54	3.07	63	2.87	-0.109	52.06	594
	nwsgappep	-4.9	17.0	49	4.16	57	3.93	-0.025	34.50	648
	str +	-6.7	17.0	49	4.16	57	3.93	-0.025	34.50	648
2.	benner 74 +	-4.4	27.4	33	6.66	50	5.93	-0.119	33.49	550
256bA	blossum62 +	-5.8	31.1	40	6.03	53	4.90	-0.219	34.21	551
1bbhA	gonnet +	-4.5	27.4	33	6.66	50	5.93	-0.119	33.49	551
	nwsgappep	-3.7	23.6	27	7.46	34	7.09	0.086	24.24	442
	str +	-5.1	29.2	35	6.54	51	5.48	-0.236	32.44	699
3.	benner 74 +	.0	27.6	29	8.17	30	8.92	0.002	23.28	759
1aaj	blossum62 +	-4.4	34.3	25	10.09	37	11.58	0.044	23.73	634
1paz	gonnet +	-.2	29.5	25	9.69	29	10.56	0.016	22.50	781
	nwsgappep	-6.7	27.6	30	8.13	32	9.16	0.034	24.57	803
	str +	-.8	30.5	31	8.58	37	9.53	-0.007	26.11	740
4.	benner 74 +	-5.4	25.3	48	5.93	57	5.62	-0.118	38.95	708
1plc	blossum62 +	-6.5	26.3	48	5.53	55	5.37	-0.129	39.57	704
1aaj	gonnet +	-5.5	24.2	54	5.43	58	4.68	-0.145	33.49	707
	nwsgappep	-8.4	20.2	49	4.33	50	5.65	-0.122	37.95	817
	str +	-4.2	18.2	22	12.96	24	13.63	0.047	40.60	782
5.	benner 74 +	-14.7	37.6	58	2.88	68	2.46	-0.144	84.73	2429
2sga	blossum62 +	-14.6	38.7	59	2.92	66	2.54	-0.144	87.53	2452
2lrpa	gonnet +	-14.9	38.1	58	3.02	66	2.49	-0.122	81.80	2410
	nwsgappep	-11.7	34.8	61	2.14	70	2.02	-0.121	87.47	1484
	str +	-16.9	37.0	65	1.98	73	1.73	-0.162	88.74	1628
6.	benner 74 +	-9.1	26.3	62	1.63	73	1.85	-0.216	22.98	365
1aboA	blossum62 +	-7.8	28.1	51	1.87	62	2.03	-0.167	19.23	381
1ckaA	gonnet +	-8.8	28.1	53	1.82	57	2.01	-0.185	17.83	359
	nwsgappep	-8.7	22.8	49	1.94	50	2.12	-0.180	14.91	264
	str +	-9.4	22.8	49	1.94	50	2.12	-0.180	14.91	264
7.	benner 74 +	-5.1	28.3	39	7.47	44	5.75	-0.085	49.03	1152
4fxn	blossum62 +	-7.4	31.2	34	7.39	38	6.40	-0.019	42.30	1592
1rcf	gonnet +	-5.2	28.3	39	7.47	44	5.75	-0.085	49.04	1152
	nwsgappep	-6.6	30.4	40	6.03	40	4.96	-0.050	50.33	1454
	str +	-6.9	29.0	39	5.68	45	5.40	0.003	39.74	1782
8.	benner 74 +	-14.5	28.6	70	1.91	74	1.91	-0.184	64.31	859
5dfr	blossum62 +	-12.9	29.9	64	2.47	71	2.35	-0.103	55.39	1032
3dfr	gonnet +	-13.7	28.6	70	1.91	74	1.91	-0.184	64.31	859
	nwsgappep	-14.0	28.6	61	2.21	68	2.23	-0.104	55.05	987
	str +	-14.6	31.2	57	2.63	64	2.59	-0.160	59.97	1285

Table 1 (continued)

case		The alignments					The models			
target template	mutation matrix	z-score	sequence identity	contact map overlap [a]	Cα RMSD [a]	contact map overlap [b]	Cα RMSD [b]	threading energy	3D profiles [c]	-ln(pdf) Modeller [d]
9.	benner 74 +	-1.5	19.5	16	10.14	17	9.96	0.146	13.48	659
1aba	blossum62 +	-2.0	28.7	34	6.62	39	6.39	-0.041	17.95	439
1kte	gonnet +	-1.6	23.0	15	9.97	13	9.65	0.156	12.31	596
	nwsgappep	-1.7	24.1	32	4.77	40	5.08	0.077	22.49	468
	str +	-1.9	13.8	15	9.60	12	9.51	0.015	11.49	606
10.	benner 74 +	-6.1	21.8	20	14.75	20	14.39	0.029	106.64	3372
1qorA	blossum62 +	-6.9	25.5	28	12.51	32	12.18	-0.010	90.33	4851
2ohxA	gonnet +	-6.1	22.4	20	14.80	21	14.39	0.036	103.71	3454
	nwsgappep	-5.3	20.2	18	15.30	20	15.68	0.020	103.42	4737
	str +	-5.5	19.3	18	14.96	19	14.64	0.075	90.80	3344
11.	benner 74 +	-18.5	34.4	55	3.88	56	4.48	-0.111	112.79	2634
1thm	blossum62 +	-21.5	36.9	53	4.34	59	4.37	-0.177	120.48	2655
2prk	gonnet +	-18.5	34.4	52	4.13	54	4.14	-0.133	114.13	2157
	nwsgappep	-22.1	33.7	50	3.93	50	4.49	-0.105	113.48	2830
	str +	-22.3	34.4	58	3.79	61	4.38	-0.148	124.39	2704
12.	benner 74 +	-12.4	22.9	40	4.26	50	4.78	-0.155	91.77	1945
2dri	blossum62 +	-11.0	24.4	41	4.38	51	4.25	-0.156	90.34	1771
8abp	gonnet +	-12.0	23.2	42	4.10	51	3.75	-0.098	81.25	1699
	nwsgappep	-9.4	22.5	40	5.29	50	5.01	-0.036	68.77	1763
	str +	-10.2	23.2	44	3.75	54	3.66	-0.168	99.89	1785
13.	benner 74 +	-7.3	26.4	44	6.16	48	5.44	-0.112	57.20	759
1acf	blossum62 +	-7.7	27.2	44	5.90	49	5.49	-0.070	57.82	779
1pne	gonnet +	-7.3	26.4	42	6.20	50	5.43	-0.119	58.26	834
	nwsgappep	-4.8	26.4	49	3.53	56	2.61	-0.186	53.83	771
	str +	-5.2	25.6	42	5.93	49	5.49	-0.130	53.61	616
14.	benner 74 +	-7.6	30.9	43	3.83	43	3.70	0.018	11.05	249
1fca	blossum62 +	-6.4	34.5	44	3.44	48	3.18	0.035	11.64	249
1fxd	gonnet +	-7.5	32.7	47	3.28	48	2.85	-0.104	14.27	253
	nwsgappep	-5.5	32.7	39	3.96	39	3.87	0.073	9.89	257
	str +	-7.3	32.7	39	3.96	39	3.87	0.073	9.89	257
15.	benner 74 +	-13.2	35.6	62	3.29	68	3.15	-0.261	47.68	689
9rnt	blossum62 +	-13.0	37.5	62	3.42	63	3.15	-0.244	42.76	590
1rtu	gonnet +	-13.4	35.6	61	3.30	67	3.17	-0.244	43.35	717
	nwsgappep	-10.7	35.6	56	3.66	56	3.65	-0.227	39.76	831
	str +	-11.7	36.5	59	3.30	64	3.15	-0.225	47.00	716
16.	benner 74 +	-1.9	20.0	26	4.42	28	3.99	-0.195	20.66	408
1bunB	blossum62 +	-1.6	25.0	27	4.36	26	4.04	-0.165	17.76	457
1tcp	gonnet +	-1.9	21.7	13	4.78	20	4.44	-0.148	13.89	533
	nwsgappep	-2.4	13.3	23	8.24	21	9.94	0.060	15.36	967
	str +	-1.9	20.0	16	4.52	15	4.65	0.005	17.33	472

[a] Target and template structures are compared
[b] Models are compared to their respective target structures
[c] The highest value indicates the best model
[d] The lowest value indicates the best model

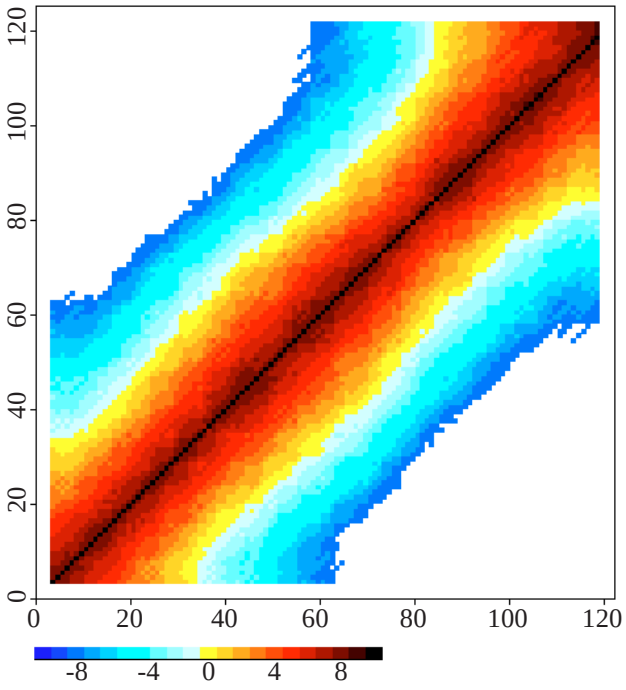


Figure 1a *Plot of density of suboptimal alignments for pairs of proteins with different level of sequence similarity. Pseudoazurin from* Alcaligenes faecalis *(1paz)), pseudoazurin from* Achromobacter cycoclastes *with 68% sequence identity*

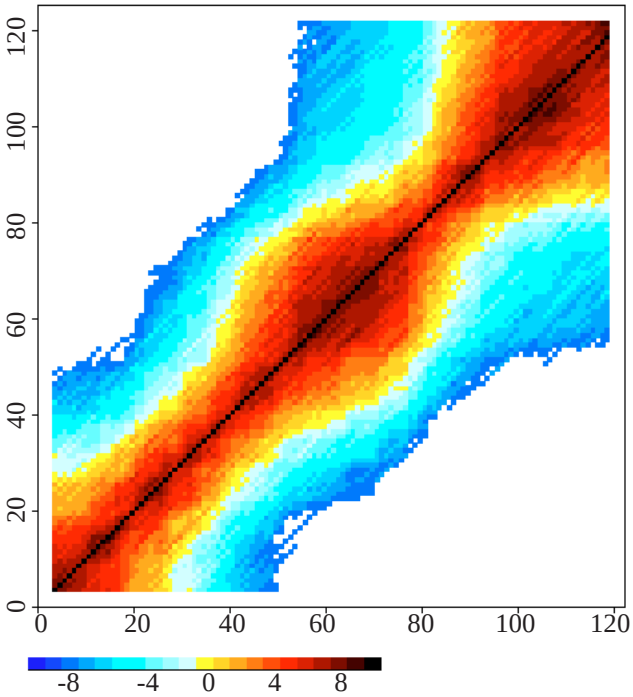


Figure 1b *Plot of density of suboptimal alignments for pairs of proteins with different level of sequence similarity. Pseudoazurin from* Alcaligenes faecalis *(1paz) with pseudoazurin from* Methylobacterium extorquens *(PDB code 1pmy) with 45% sequence identity*

threading algorithms. Surprisingly, although such methods are able to recognize many similarities that are impossible to detect on the sequence level, so far it has not resulted in better quality alignments. The second problem behind sequence alignment ambiguities is that with increasing evolutionary distance between proteins, even the structural alignments become ambiguous [35, 36]. Position-by-position equivalence between two protein structures may depend on how we define the similarity between them [36]; thus, in many cases, sequence alignment may correctly represent reality by being ambiguous.

In such cases, comparative modeling could still work properly because the structural similarity is sufficient. There are, however, two problems to be solved: how to obtain the best possible alignment without the benefit of using both structures to evaluate it and how to estimate the distance between the model and (unknown) target structure. If the modeling algorithm fails for any reason, we should at least know that the resulting models are inaccurate.

A standard approach, suggested by modeling package manuals, is to prepare the alignment using one of the well-known substitution matrices and to assess the quality of the single resulting model. It is suggested that the alignment should be “corrected” if there is evidence that the model is of low accuracy. In this contribution, we present a semi-automatic extension of this approach, which could be implemented using widely available software packages. As shown in the paper, on several test cases, this approach gives consistently good results.

To produce a high quality model of a target protein, which would be as close to the real structure as the actual similarity between template and target would allow, we propose the ***M**ultiple **M**odel **A**pproach*, consisting of the following steps:

1. Exploration of the alignment diversity by using various alignment parameters.
2. Elimination of alignments incompatible with some conserved features of target and template sequences (some features might be known prior to target structure determination - active and binding sites, disulfide bridges, etc.).
3. Construction of full-atom models using all the remaining alignments.
4. Selection of the best model using the value of the threading energy as a criterion of model accuracy.
5. Assessment of the model accuracy using the value of threading energy.

Results

A set of 16 protein pairs, presented in detail in the Methods section, was chosen to represent weakly homologous pro-

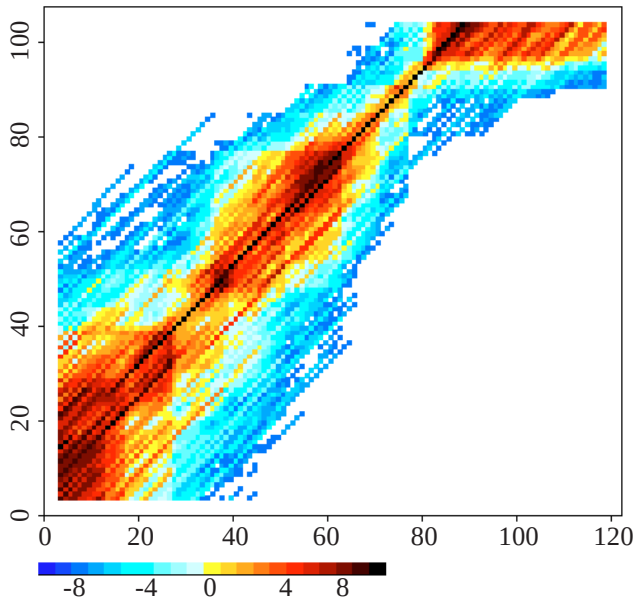


Figure 1c *Plot of density of suboptimal alignments for pairs of proteins with different level of sequence similarity. Amicyanin (PDBcode 1aaj) aligned with pseudoazurin from* Alcaligenes faecalis *(1paz) with 29% sequence identity*

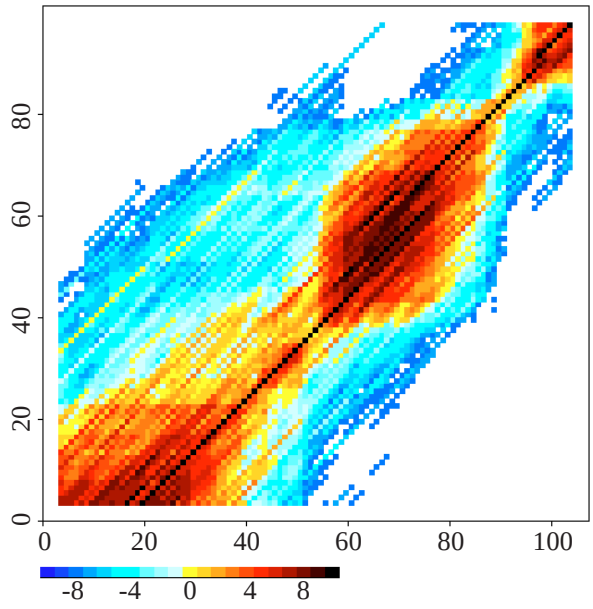


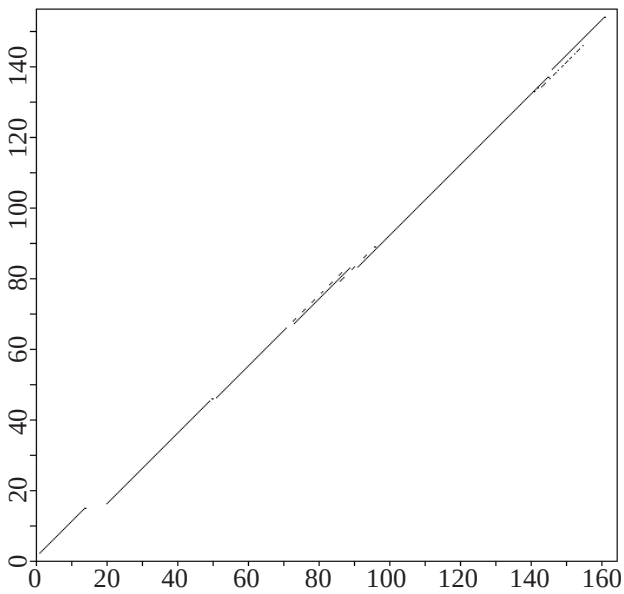
Figure 1d *Plot of density of suboptimal alignments for pairs of proteins with different level of sequence similarity. Amicyanin (PDBcode 1aaj) aligned with plastocyanin from poplar leaves (PDB code 1plc) with 24% sequence identity*

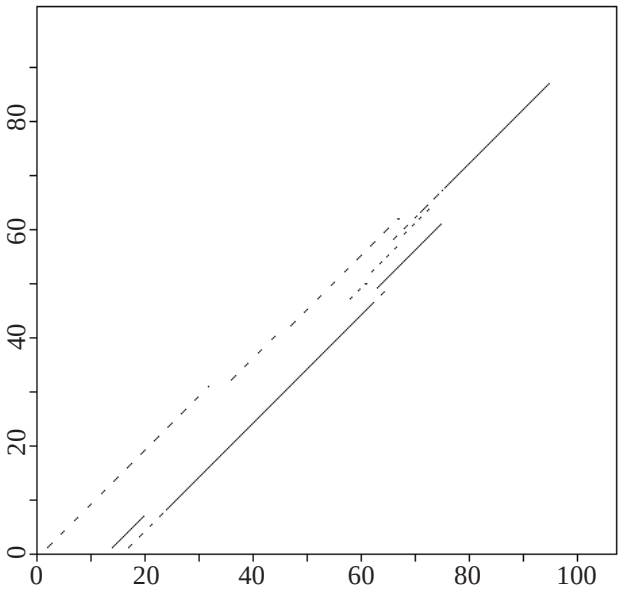
teins of different structural classes. Several pairs represent actual examples where published sequence alignments, prepared before the structural determination of the target, were wrong. For these proteins, the MMA approach, as presented above, is applied and described in detail below. The results and their analyses are presented in Table 1. Structures of all proteins in this set are currently known; therefore, the results shown are not genuine predictions but are presented merely to illustrate the MMA procedure. To stress this point, in Table 1, values obtained using the actual target structure are printed on gray background. These values were used only "outside" the actual MMA procedure to evaluate and discuss its results.

The alignments

As mentioned in the Introduction, the alignment between two proteins becomes increasingly unstable with the decreasing sequence similarity between them. This could be illustrated by looking at the plot of suboptimal alignments for protein pairs with a different level of sequence similarity. Every alignment can be thought of as a position-by-position equivalence between the two proteins. There are many possible alignments, and the one that is usually reported maximizes the similarity score (or minimizes the distance) between two protein sequences [2]. Every alignment can be visualized as a series of points on a plane, where a presence of the point (i,j) means that this particular alignment included the equivalence between position i from one sequence with position j in the second. On a plane, we can present all possible alignments between two proteins, using a third dimension to represent their score. In plots shown in Figure 1, only alignments whose scores differ from the optimal one by 10% or less are shown and only the best score is displayed for every point. Therefore, the single line denotes a unique alignment within the top 10% of the top score. If the line gets wider, it means that other alternative alignments with similar scores are present. If large alignment variations are possible within small score differences, this strongly suggests that the alignment is not well defined in a given region. The most interesting situation arises where the line splits into several parallel lines, meaning that few non-overlapping alignments with similar scores are possible.

For test pairs No. 3 (amicyanin (1aaj) vs. pseudoazurin from *Alcaligenes faecalis* (1paz)), two other proteins, pseudoazurin from *Achromobacter cycoclastes* with 68% sequence identity and pseudoazurin from *Methylobacterium extorquens* (PDB code 1pmy) with 45% sequence identity to 1paz, were chosen. Finally, amicyanin is compared with plastocyanin from poplar leaves (1plc), the fourth pair from Table 1. The density of suboptimal alignments is displayed for all four pairs (see Fig.1). As can be seen in Figure 1, the picture changes qualitatively between pairs with sequence similarity above 30% and for pairs below that level. For closely related proteins, the alignment is essentially unique, as represented by a single black line on Figures 1a and 1b. With lower sequence similarity, large areas of ill-defined alignments appear (compare Figures 1a and 1b) and, finally, for the two





a) 5dfr/3dfr
gonnet +

```
MISLIAALAVDRVIG     PWNLPADLAWFKRNTLDKPVI
 TAFLWAQNRNGLIGKDGHLPWHLPDDLHYFRAQTVGKIMV

-EEEEEEE———      ——HHHHHHHHHHH——EE
 -EEEEEE——EEE———HHHHHHHHH——EEE

MGRHTWESI GRPLPGRKNIILSSQPGTDDR VTWVKSVDE
VGRRTYESFPKRPLPERTNVVLTHQEDYQAQGAVVVHDVAA

EEHHHHHHH ———EEEE——— -EEE—HHH
EEHHHHHH———EEEEE———EEE—HHH

AIAACGDV  PEIMVIGGGRVYEQFLPKAQKLYLTHIDAEV
VFAYAKQHLDQELVIAGGAQIFTAFKDDVDTLLVTRLAGSF

HHHH——  —EEEEE—HHHHHH—EEEEEEEEE——
HHHHHHH——EEE—HHHHHH———EEEEEEE——

EGDTHFPDYEPDDWESVFSEFHDADAQNSHSYCFEILERR
EGDTKMIPLNWDDFTKVSSRTVE DTNPALTHTYEVWQKKA

-———EEEEEEEEE———EEEEEEEE-
---EE———EEEEEEEEE- ——EEEEEEEE—
```

str +

```
MISLIAALAVDRVIG     PWNLPADLAWFKRNTLDKPVI
 TAFLWAQNRNGLIGKDGHLPWHLPDDLHYFRAQTVGKIMV

-EEEEEEE———      ——HHHHHHHHHHH——EE
 -EEEEEE——EEE———HHHHHHHHH——EEE

MGRHTWESIG RPLPGRKNIILSSQPGTDDR VTWVKSVDE
VGRRTYESFPKRPLPERTNVVLTHQEDYQAQGAVVVHDVAA

EEHHHHHHH- ——EEEE——— -EEE—HHH
EEHHHHHH———EEEEE———EEE—HHH

AIA ACGDVPEIMVI GGGRVYEQFLPKAQKLYLTHIDAEV
VFAYAKQHLDQELVIAGGAQIFTAFKDDVDTLLVTRLAGSF

HHH H———EEEE E—HHHHHH—EEEEEEEEE——
HHHHHHH——EEE—HHHHHH———EEEEEEE——

EGDTHFPDYEPDDWESVFS EFHDADAQNSHSYCFEILERR
EGDTKMIPLNWDDFTKVSSRTVEDTNPALTHTY  EVWQKKA

———EEEEEE EEE———EEEEEEEE-
—EE———EEEEEEEEE———EEE  EEEEE—
```

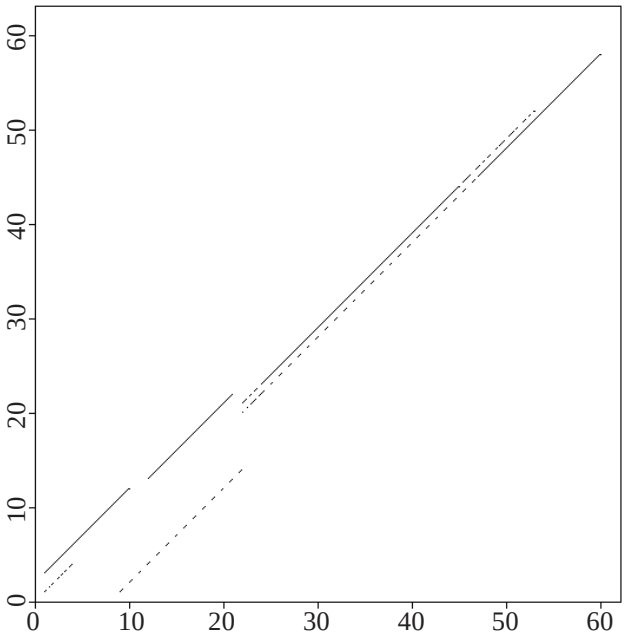
a) 1plc/1aaj
gonnet+

```
             IDVLLGA   DDGSLAFVPSEFSISPGEKIVF
DKATIPSESPFAAAEVADGAIVVDIAKMKYETPELHVKVGDTVTW

             -EEEE—   ———EE—EEEE——EEEE
—EE———EE———EEEEEE—EE—EEEE——EEEE

KNNAGFPHNIVFDEDSIPSGVDASKISMSEEDLLN
INREAMPHNVHF     VAGV LGEAAL KGPMMK

EE———EE———HHHH———
EE———EE     —— ——E E——

AKGETFEVALSNKGEYSFYCSPHQGAGMVGKVTVN
 KEQAYSLTFTEAGTYDYHCTPH  PFMRGKVVVE

——EEEEE——EEEEEE———EEEEEE-
 —EEEEEEE—EEEEEE——  —EEEEEE-
```

str+

```
 IDVLLGADDGSLAFVPSEFSISPGEKIVFKN   NAGFPHNIVF D
DKATIPSESPFAAAEVADGAIVVDIAKMKYETPELHVKVGDTVTWIN

 -EEEE———EE—EEEE——EEEEEE   ———EE -
—EE———EE———EEEEEE—EE—EEEE——EEEEEE

EDSIPSGVDASKISMSEEDL   LNAKGETFEV
REAMPHNVHFVAGVLGEAALKGPMMKKEQAYSL

———HHHH———   ———EEEE
———EE———EE———EEEEE

ALSNKGEYSFYCSPHQGAGMVGKVTVN
TFTEAGTYDYHCTPH  PFMRGKVVVE

E——EEEEEE———EEEEEE-
EE—EEEEEE——  —EEEEEE-
```

Figure 2 *caption see page 301*

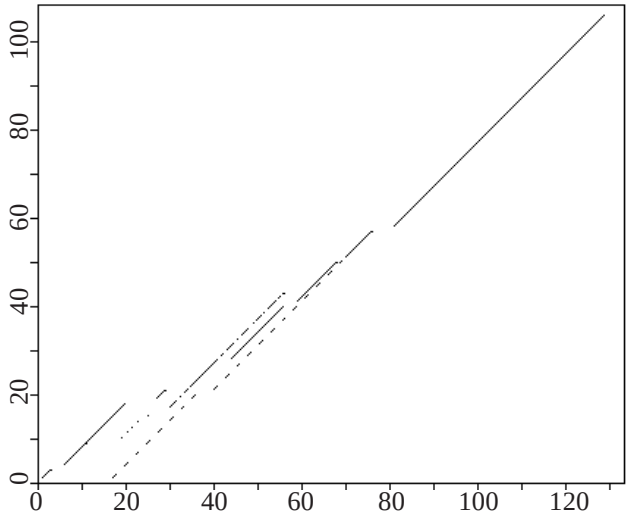


c) 1bunB/1tcp
benner74+

```
RKRHPDCDKPPD TKICQTVVRA  FYYKPSA
  YNRLCIKPRDWIDECDSNEGGERAYFRNGK

_______ ___EEE  EEEE--
  _________HHHH-EEEEE--

KRCVQFRYGGCNGNG NHFKSDHLCRCECLEYR
GGCDSFWICPEDHTGADYYSSYRDCFNACI

-EEEEEEE___ __HHHHHHHH__
-EEEEE_______HHHHHHH-
```

nwsgappep

```
         RKRHPDCDKPPDTKICQTVVRAFYY
YNRLCIKPRDWIDECDSNEGGE      RAYFR

          __________EEEEEE
_________HHHH-      EEEEE

KPSAKRCVQFRYGGCNGNGNHFKSDHLCRCECLEYR
NGKGGCDSFWICPEDHTGADYYSSYRDCFNACI

E__EEEEEEE_______HHHHHHHH__
__EEEEE_________HHHHHHH-
```



d) 256bA/1bbhA
str +

```
ADL  EDNMETLNDNLKV            IEKADNAAQVKDAL
AGLSPEEQIETRQAGYEFMGWNMGKIKANLEGEYNAAQVEAAA

-H  HHHHHHHHHHHHH             HHH-HHHHHHHH
__HHHHHHHHHHHHHHHHHHHHHHHHHH__HHHHHHHH

TKMRAAALDAQKA     TPPKLED KSPDSPE
NVIAAIANSGMGALYGPGTDKNVGDVKTRVKPEFFQN

HHHHHHHHHH-     ___ __HH
HHHHHHH_________________

MKDFRHGFDILVGQIDDALKLANEGKVKEAQAAAEQLKTTRNAYHQKYR
MEDVGKIAREFVGAANTLAEVAATGEAEAVKTAFGDVGAACKSCHEKYRAK

HHHHHHHHHHHHHHHHHHHHHHH-HHHHHHHH-HHHHHHHHHHHH-
HHHHHHHHHHHHHHHHHHHHHHH__HHHHHHHHHHHHHHHHHHHHH-
```

nwsgappep

```
                ADLEDNMETLNDNLKVIEKA   DNAA
AGLSPEEQIETRQAGYEFMGWNMGKIKANLEGEYNAAQVEAAA

                -HHHHHHHHHHHHHHHH-   -HH
__HHHHHHHHHHHHHHHHHHHHHHHHHH__HHHHHHHH

QVKDALTKMRAAALDAQKATPPKLEDKSPDSPE
NVIAAIANSGMGALYGPGTDKNVGDVKTRVKPEFFQN

HHHHHHHHHHHHHHHH_______HH
HHHHHHH_________________

MKDFRHGFDILVGQIDDALKLANEGKVKEAQAAAEQLKTTRNAYHQKYR
MEDVGKIAREFVGAANTLAEVAATGEAEAVKTAFGDVGAACKSCHEKYRAK

HHHHHHHHHHHHHHHHHHHHHHH-HHHHHHHH-HHHHHHHHHHHH-
HHHHHHHHHHHHHHHHHHHHHHH__HHHHHHHHHHHHHHHHHHHHH--
```

Figure 2 *Four examples of the comparison of alignments obtained with different mutation matrices in a dot-matrix representations of alignments. Abbreviations for mutation matrices are identified in the Methods section. Best alignments are shown as continuous lines. Other alignments, where different from the best ones, are shown as dashed lines. The best (first) and the worst (second) sequence alignments are shown in detail, with an indication of the mutation matrices used. Positions of secondary structure elements, defined by their DSSP assignments, are shown below the sequence alignment*

most distant pairs, a number of alternative alignments are present, seen as parallel lines on Figure 1c or 1d.

As mentioned in the Introduction, none of the widely available alignment packages allows one to explore the suboptimal alignments. Therefore, we explore the diversity of alignment using a different method. For a given target-template pair, we align them using several mutation matrix-penalty gap combinations, optimized independently to produce reasonable alignments [24]. The connection between the width

of distribution of suboptimal alignments and the differences between optimal alignments obtained with different scoring matrices and gap penalties is not established. However, both explore the same effect and both undergo the same qualitative change at the same level of sequence similarity. For instance, one can compare Figures 1d and 2b, which present a picture of alignment diversity for the same protein pair of amicyanin and plastocyanin. Alignment ambiguity at the N-terminus and around position 60-70 is captured in the same way in both methods.

For each test pair, several different alignments were obtained (see Fig. 2 and Table 1) following the procedure described in the Methods section. The differences between the alignments were quite large even for the test pairs where sequence identity is above 30%. We can quantify these differences using structural information about both proteins, which is not a part of the MMA procedure since it requires knowledge of the target structure. The percent of conserved inter-residue contacts can change by a factor of 2, depending on the alignment procedure used. The differences in alignments are not limited to outside loops, but also extend to the core regions (see Fig. 2). The global measure of structural similarity, C_α RMSD (root mean square deviation) [37, 38], also can show a large variation for a single test pair from acceptable values of 2-4 Å to a practically meaningless 7-15 Å. The Ca RMSD can be calculated for alignments, comparing equivalenced residues from target and template structure, and for final models, comparing model and target structures. Both values are reported in Table 1.

In several cases, there is a difference in sensitivity between the two measures of alignment accuracy: contact map overlap and C_a RMSD. The value of C_a RMSD changes much more rapidly with distance from the optimal alignment. This reflects the fact that this measure of model quality is more global - it is very sensitive to regions of weak structural similarity and changes of relative positions of otherwise similar protein segments.

Unfortunately, in a real life situation where the target structure is not known, we are not able to distinguish between different alignments. Each of the mutation matrices used here performs very well on a large set of protein pairs [24]. At the same time, for the set studied here, each matrix in at least one case produces the worst alignment (and in at least one case the best). The benner74 matrix performs best in the largest number of cases, but overall has only 40% correct hits. It is often observed that for a given test pair, one of the mutation matrices works much worse than the other, demonstrating that reliance on one alignment protocol is dangerous.

It is noteworthy that the z-score measuring significance of a given alignment in terms of its matrix score, as described in the Methods section, does not distinguish correct from incorrect alignments. In some cases, alignments with the best z-scores are actually the worst from the structural point of view. Another interesting observation arising from this study is the fact that sequence identity is a very poorly defined quantity for protein pairs of low homology. As can be seen in Table 2, different mutation matrices can yield sequence identity values differing more than 10% (in absolute values).

Table 2 *The performance of different threading energy terms in distinguishing the best model*

threading energy term	**number of correct models distinguished using only this term**
local energy	7
burial energy	5
contact energy	5
sum of all terms	11

The differences between alignments can be of various scopes and consequences. In some cases, differences are minor, involving mostly shifts in loop regions and on edges of secondary structure elements. For some cases, e.g., 5dfr/3dfr alignment (see Fig. 2A), small gaps are also observed in one of the helices and one of the β-strands. These apparently minor shifts in alignment can result in substantial deviations in models. It is widely accepted that gaps in alignments should be placed mostly in loops outside the continuous secondary structure elements. Nevertheless, in several cases a more correct model is obtained if the alignment procedure introduces a gap in one of the secondary structure elements. It happens, for instance, with the 256bA/1bbhA pair (see Fig. 2D) where an alignment with a large gap in the middle of the N-terminal helix results in a much better model than an alignment with no big gaps in the central parts of the helices. The resulting models have 5.48 and 7.1 Å RMSD, respectively, relative to the target structure. A similar case is the 1bunB/1tcp pair (see Fig. 2C), where the better alignment has a gap in the middle of a β-strand. What is even more puzzling, the best model for this protein was obtained when an α-helix was partially matched to a β-strand. In this case, better alignment can be easily selected based on matching 6 cysteine residues that form 3 disulfide bridges present in the template and target proteins. The alignments obtained using benner74 and nwsgappep mutation matrices correctly match 5 and 2 cysteine residues, respectively.

In other cases, alignments differ in a more significant way. For example, in the 1plc/1aaj case, (see Fig. 2B) one of the alignment methods does not recognize that an N-terminal segment in 1aaj does not have a corresponding segment in 1plc. This results in a shift of one β-strand in the alignment and the RMSD between model and target structure rises from 4.7 Å (for the better alignment) to a meaningless 13.7 Å for the worst alignment.

The models

As discussed in the previous paragraph, for all 16 cases, alignments prepared with different mutation matrices show significant divergence and, using only information contained in the sequence we are not able to recognize the best alignments. Therefore, in this section, models built based on vari-

Figure 3a *Ribbon diagrams of target and template structures compared with two model structures, the best one and the worst one. Abbreviations for mutation matrices are identified in the Methods section. 1plc vs. 1aaj. Positions of secondary structure elements are calculated using the target structure,* not *the model structures*

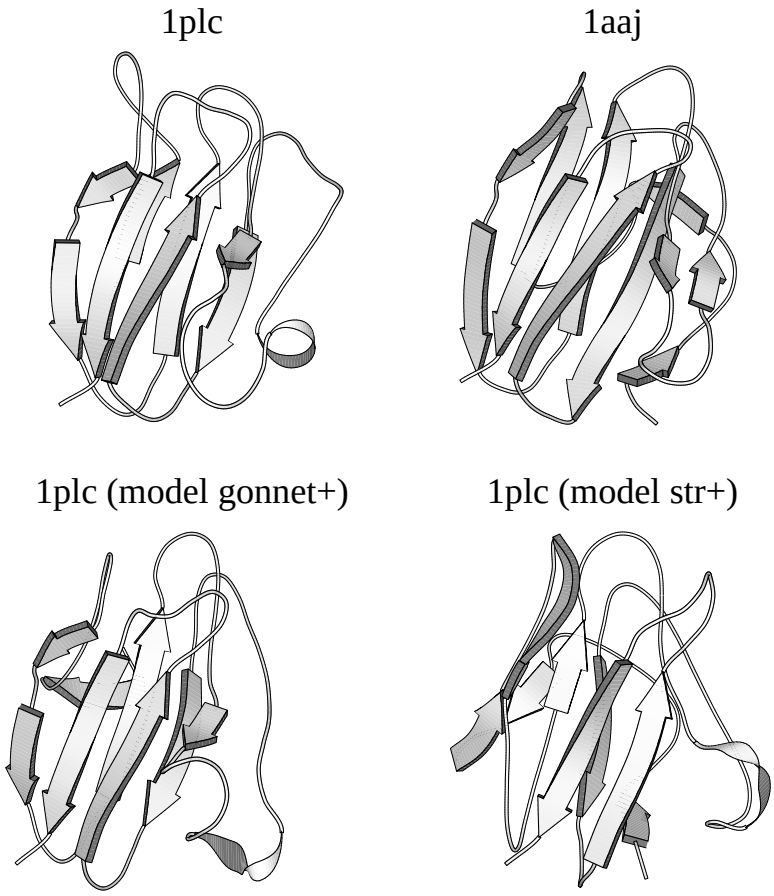


Figure 3A: The structures of 1plc and 1aaj and the best and the worst model of 1plc.

ous alignments will be analyzed. We want to stress that all scores discussed here were calculated without reference to the target structure; therefore, they could also be calculated for *bona fide* predictions.

As seen in Table 2, threading approaches do a very good job at recognizing the best model. In 11 of 16 test cases, the models yielding the lowest value of threading energy were closest to the native target structure (see Table 2). In 9 cases, the models judged to be worst by this criterion are those that differ most from the native structure. The identification of the more accurate models is much better in cases when at least one model is close to the native structure. It can be easily rationalized, since small differences between very wrong models are probably meaningless. In other words, a 2 Å difference between a 2.5 Å and 4.5 Å model is important, but the same difference between a 9 Å and 11 Å model is probably not. The threading energy calculated with the topology fingerprint threading algorithm [39] gives very similar results to the 3D profiles score [40]. Both methods choose the best alignment in the same number of cases and, in one case, they are both wrong in the same way. However, in the cases where the 3D profiles are wrong, they are often completely wrong, picking the worst model such as in the case of 1plc/1aaj where a nonsensical 13.61 Å RMSD alignment is chosen over a 4.7 Å one. The value of an alignment z-score selects the best model in only 5 cases. An internal measure of model quality supplied by the MODELLER program [41] is able to distinguish the best model in 4 cases only. It is not surprising though, since this function does not describe the physical quality of the model, only the degree of satisfaction of the imposed restraints. These numbers can be compared to the random choice of models, which should give 3 correct hits.

The cases where the threading score did not pick up the model closest to the native structure are special because the target and template structures have relatively little similarity. In one of 5 such cases (1aaj/1paz), only the central parts of the sequences have a common fold. The N-terminal part of the target has no corresponding fragment in the template structure, which results in an unacceptable C_α RMSD above 8 Å. This value simply tells us that the template chosen is inappropriate for modeling the given target. This is in agreement with the high values of the threading score for all the models for this target/template pair. In the other four cases (256bA/1bbhA, 4fxn/1rcf, 1aba/1kte and 1thm/2prk), the deviation of best models from the native structure is in the range 4-6 Å. It should be noted that in three of those four cases the models picked up by the criterion of a threading

Figure 3b *Ribbon diagrams of target and template structures compared with two model structures, the best one and the worst one. Abbreviations for mutation matrices are identified in the Methods section. 256b vs. 1bbh. Positions of secondary structure elements are calculated using the target structure,* not *the model structures*

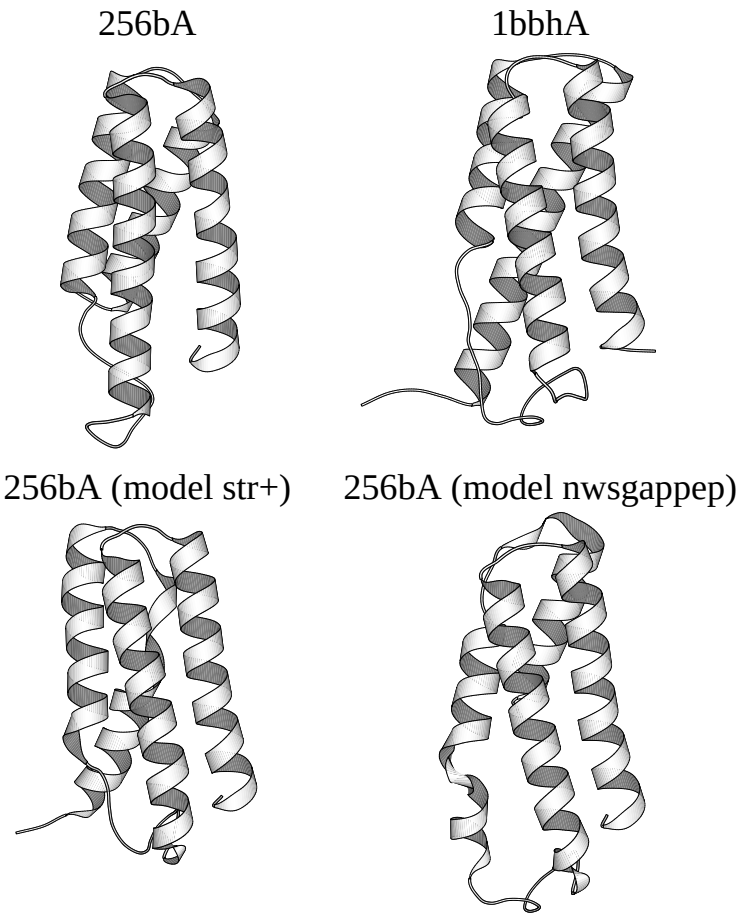


Figure 3B: The structures of 256bA and 1bbhA and two models of 256bA.

score have the second lowest deviations from the native structure.

Figs. 2B, 2D, 3A and 3B illustrate the influence of the alignment and the template structure on the model. In the case of the 1plc/1aaj pair, the N-terminal part of the alignment is unstable (Fig. 2B). Consequently, a very inaccurate model is obtained using the str+ mutation matrix. However, even in the best alignment (calculated using gonnet+ mutation matrix), the N-terminal β-strand of 1plc is incorrectly assigned to a coil segment in the 1aaj structure despite the existence of a corresponding strand. Therefore, the N-terminal fragment of this model has a wrong backbone direction (see Fig. 3A). This confirms the known fact that significant errors occurring at the stage of alignment calculation cannot be corrected at the stage of model building. The most probable reason of alignment instability in this case is that the N-terminal part of the template structure has no corresponding part in the target. Thus, the N-terminal fragments of the target sequence have many possible places to be assigned to in the "dangling" N-terminal fragment of the 1aaj structure.

The case of the 256bA/1bbhA pair is an example of erroneous matching of secondary structure elements. As can be easily observed (see Figs. 2D and 3B), the gaps introduced in the secondary structure elements of the target are less harmful for the model than the assignment of the target's helices to coil regions of the template structure. In this case, the model selected by the criterion of the lowest value of threading score criterion is not the most accurate one. However, the selected model and corresponding alignment are still much more accurate than the worst one (see Fig. 3B).

Another interesting case is the 1qorA/2ohxA pair (sequence identity in the range 19-25%), where the target and template sequences are longer than 300 residues. Various alignment parameters produce a series of models with C_α RMSD in the range 12-15 Å and a threading score in the range of -0.1 to +0.08, which indicates that all models are of very low accuracy. However, local similarity of the native structure and the models is much higher - up to 32% of conserved interresidue contacts for the best model. It is obvious that for larger protein structures, the discrepancy between the local and global measure of structural similarity is higher. In this case, the most accurate model has the best threading score.

Based on the present results (see Table 1), the following rule of thumb can be formulated. If the threading score-per-residue is below -0.1, one can be confident that the method chooses the best model and the model is correct. Values of -0.1-0.0 correspond to models that should be used with cau-

Table 3 *Test pairs of similar proteins used as the test cases for the Multiple Model Approach*

	Target		Template		Structural class	Fold type (SCOP)	Sequence identity (GCG)
	PDB code	length (aa)	PDB code	length (aa)			
1.	2hhdA	141	2gdm	153	all alpha	Globin-like	17.1
2.	256bA	106	1bbhA	131	all alpha	Four-helical up-and-down bundle	23.6
3.	1aaj	105	1paz	120	all beta	Cupredoxins	27.6
4.	1plc	99	1aaj	105	all beta	Cupredoxins	20.2
5.	2sga	181	2lprA	198	all beta	Trypsin-like serine proteases	34.8
6.	1aboA	58	1ckaA	57	all beta	SH3-like barrel	22.8
7.	4fxn	138	1rcf	169	alpha/beta	Flavodoxin-like	30.4
8.	5dfr	154	3dfr	162	alpha/beta	Dihydrofolate reductases	28.9
9.	1aba	87	1kte	105	alpha/beta	Thioredoxin-like	24.1
10.	1qorA	326	2ohxA	374	alpha/beta	NAD(P)-binding Rossmann-fold domain	20.2
11.	1thm	279	2prk	279	alpha/beta	Subtilases	22.1
12.	2dri	271	8abp	305	alpha/beta	Periplasmic binding protein-like I	22.5
13.	1acf	125	1pne	139	alpha+beta	Profilin-like	26.4
14.	1fca	55	1fxd	58	alpha+beta	Ferredoxin-like	32.7
15.	9rnt	104	1rtu	113	alpha+beta	Microbal ribonucleases	35.6
16.	1bunB	61	1tcp	60	small proteins	BPTI-like	13.3

tion, carefully analyzed and perhaps rebuilt while variations in alignments should be explored (i.e., different gap parameters). Values above 0.0 indicate templates inadequate for given target structures. In the latter case, all the models can be regarded as being incorrect.

An analysis of energy of the models

The terms of the threading energy of the models were calculated separately and analyzed to answer the following question. Are all energy terms equally important for distinguishing the best model?

As illustrated in Table 2, all energy terms take part in selecting the best model. The sum of all energy terms works much better than each energy term alone. In two cases, no single energy term was able to distinguish the best model, but the sum of all energy terms indicates which one is the closest to the native structure.

These data exclude the possibility that distinguishing between models of different quality is only the result of strain in the protein backbone caused by an incorrect alignment between the target sequence and the template. This strain is probably the most important and easily detectable result of incorrect alignment as indicated by the number of correctly selected models when using the local energy term only. The significance of burial and contact energy for optimal model selection is lower than the importance of the local energy term. It is probably because the requirement for the conservation of burial status is implicitly encoded in the mutation matrices. It was shown that the matching of side chain hydrophobicity in aligned sequences is the most important feature of the majority of mutation matrices [31]. Thus, the conservation of burial status is at least partially imposed during the calculation of the alignment. On the other hand, side chain hydrophobicity is the dominant component in two-body interaction term [49]. This is in contrast to local propensities information, which depends on pairs of consecutive amino acids [42, 43] and cannot be entirely included in standard mutation matrices.

Discussion

In this work, we introduce a Multiple Model Approach (MMA) to comparative modeling. MMA uses alignment ambiguities, which are one of the most serious problems in comparative modeling, to explore the space of target-template similarity. Series of models are prepared and evaluated using threading algorithms. On a large number of examples, it was shown that in the majority of cases, MMA correctly identifies the best alignment, which is not possible using only sequence information. In addition, it provides a reliable measure of model quality. MMA is easy to implement and all software tools necessary for its application are easily available.

The importance of alignment accuracy can be easily seen in the test models obtained in this study. Models built based on erroneous alignments, although prepared using state-of-the art alignment methods that statistically perform well,

contain errors as severe as completely incorrect matching of disulfide bridges or completely misaligned elements of secondary structure. MMA allows one to worry less about the quality of the single alignment. This approach can be extended to using different structural templates, if the choice of one that is the most appropriate for modeling is not obvious, as presented on the example of S100A1 structure prediction [44].

The importance of sequence alignment is stressed in the manuals of comparative modeling programs, but is not widely recognized as a significant problem despite numerous examples where comparative modeling failed, even when using a correct template [23]. The author of the MODELLER package, which was used in our contribution, suggests an iterative approach, i.e., the evaluation of a model produced using an initial alignment, the correction of the alignment and producing an improved model [45]. The approach presented here can be viewed as a practical realization of these suggestions. Calculation of many alignments and models can be done simultaneously, which allows an automation of the process. It was shown that the use of different substitution matrices is the easy way of generating alternative alignments.

One may wonder if it would be possible to solve the alignment ambiguity problem by developing an alignment strategy that would work in every case. This seems unlikely for a number of reasons. The amino acid properties that are most important for conservation of structural features can depend on the size and structural class of a given fold. The evolutionary distance is another important parameter that differs significantly between test pairs, which probably makes the construction of one optimal alignment matrix and calculation of one set of gap penalties impossible. We have shown that instead of searching for one optimal alignment protocol, one may construct a set of alternative alignments corresponding to various possible structural and evolutionary relationships between sequences and defer the selection of the correct alignment to the stage of model evaluation.

Why is it possible to recognize the most accurate model when the same is very difficult for the alignment? The construction of the model imposes additional restraints on the alignment and significantly improves the sensitivity of some energy terms. At the stage of alignment construction, separate fragments of the target sequence are equivalenced to fragments of the template structure. The model building stage can be regarded as a test of the geometric feasibility of the connectivity between those fragments and of the consistency of the interactions between them. It also provides a realistic description of those parts of the target structure that do not have an equivalent part in a template structure. All this information is not included in one-dimensional alignment algorithms.

An analysis of the energy terms leads to the conclusion that the local energy term (which describes local propensities of amino acids towards different types of secondary structure) plays the most important role in the identification of the best model. This stresses the importance of the correct alignment of secondary structure elements and of the global geometrical feasibility of the model. Random sequence similarities may push the alignment into unphysical regions (e.g., too short loops), which results in a high strain of the protein backbone in the model.

Model building starting from a template of relatively low sequence homology to the target can be quite successful or very disappointing. It should be stressed that MMA provides a reliable measure of model significance. As pointed out in the Results section, the threading score per residue allows discrimination between two modeling cases. In the first case, the sequence similarity corresponds to an important structural similarity, and MMA produces a reasonable model that is consistently selected from those less reasonable. In the second case, the structural similarity resulting from sequence similarity is too low for any model to be useful. No criterion using information from the sequences alone, such as the numerical value of sequence identity or the statistical significance of sequence alignment, allows one to distinguish between these two cases. The present results allow one to expect that MMA may lead to an extension of applicability of comparative modeling to cases of low sequence similarity (the range of 15-30% sequence identity) in the future.

The approach described here attempts to explore the alternative alignments and conformational space more widely than standard modeling techniques. It was shown that the most accurate model could be much more easily identified than the most accurate alignment. This is a result of additional geometrical constraints that are imposed on the alignment at the model building stage. MMA, which can be summarized as the process of construction and evaluation of several models based on a set of alternative alignments, leads to a significant improvement in the accuracy of comparative modeling.

Materials and methods

Protein pairs tested

Test pairs of sequence identity below 35% were selected from the cases of unstable alignments described in the literature and from a database of structurally similar proteins [46]. We selected the pairs representing different structural classes and fold types. Protein pairs forming the testing set are presented in Table 3.

Calculation of the alignments

Four substitution matrices were selected from the best matrices in ranking published by Argos and co-workers [24]. The modified Dayhoff substitution matrix used in the popular GCG package [47] was also included and identified as nwsgappep in the tables. The mutation matrices were downloaded from http://www.genome.ad.jp/ [31]. They were made positive as recommended in [24], which is indicated by the plus sign. Gap penalty parameters optimized for the substitution matri-

ces were adapted from [24]. The alignments were calculated using the GAP program from the GCG package [47], as suggested by Argos and coworkers [24]. The differences in the alignments were surprisingly large (see Fig. 2). Cases where two mutation matrices yielded the same alignment were rare and didn't necessarily coincide with the alignment being correct.

To measure the significance of a given alignment on the sequence level, the Z-scores were calculated as a difference between the score obtained from the alignment algorithm for a given pair of sequences and the average score obtained after randomization of the target sequence (50 randomized sequences were used).

As an additional test, the alignments were evaluated structurally based on our knowledge about target structure. Two measures of alignment accuracy were calculated. The global measure of alignment accuracy was calculated as the value C_α RMSD after optimal superposition of the target and the template, where equivalent pairs of C_α atoms were defined by the given sequence alignment. Contact map overlap between the target and the template, according to each alignment, was used as the local measure of alignment quality. It is important to note that this strategy of alignment evaluation is different from the one used by most authors, where the alignment accuracy is calculated by comparing it position-by-position with the "standard of truth" structural alignment [24, 26]. It was recently shown [35, 36] that structural alignments may be ill-defined in themselves and, in most cases, there are many possible alternative alignments of comparable quality. In a complete parallel to the situation in sequence alignments, the density of suboptimal alignments can be calculated for structural alignments [48] and the width of such distribution becomes quite large for distant homologues such as the test pairs used in this work. Therefore, the structural score of the sequence-based alignment is a much more meaningful measure of its quality.

Model building and evaluation of the models

The models were built using the method of "satisfaction of spatial restraints" implemented in the MODELLER program [41]. A standard MODELLER routine 'model' was applied.

The models were evaluated with the score of threading calculated with the use of energy parameters developed for the topology fingerprint threading method [39] and, independently, by 3D profiles score [40]. We often use the term "energy" instead of "score", which does not mean that those numbers can be rigorously treated as real physical energies. The units of "energy" roughly correspond to the value of one kT at room temperature [49]. To allow comparison of scores for proteins differing in length, we report the score per residue. The same approach, i.e., using the threading energy of a full atom protein model, was used to analyze interactions in a family of structurally divergent homologous proteins [50].

Finally, the models were compared to the true structures of the target proteins. Again, two measures of structural similarity are used: a global measure of the root mean square deviation between C_α positions (RMSD) and the local one as defined by the contact map overlap [36, 48] (see Table 1).

For comparison, the internal measure of restraint satisfaction provided by the modeling program (-ln(pdf)) is shown. This measure does not directly focus on model energy, but describes the degree to which structural restraints imposed on the target by the template were satisfied.

It should be stressed that the MMA method, if used as a prediction tool, does not require any knowledge about the target three-dimensional structure. All the alignments and models discussed in this study were prepared without any knowledge of target structures. Comparisons of the target structures with model and template structures were done only to explore the possibilities and applicability of the MMA.

The local measure of structural similarity

Contact map overlap can be used as the local measure of structural similarity [36, 48]. Here, it was used to evaluate the accuracy of the alignments and protein models. Two contact maps are equivalenced according to the alignment and overlapping contacts are counted. This is described by the following formula:

$$S_c(AB) = \frac{1}{N_c} \sum_{i>j} \sum_{j} C^{A}_{AB(i)AB(j)} C^{B}_{AB(i)AB(j)}$$

where C^{A}_{kl} and C^{B}_{kl} denote residue-residue contacts in proteins A and B, respectively. $C^{B}_{AB(i)AB(j)}$ is the contact map of protein B renumbered according to the alignment AB. The normalization factor N_c is the maximal possible value of overlap, equal to the number of contacts in the smaller of the two proteins. The same formula is used to calculate the value of contact map overlap between the native structure and the protein model. In this case, there is a simple correspondence between two contact maps that are of equal size. Thus no renumbering of contacts is necessary and the overlapping contacts are simply counted.

Threading energy function

The threading energy of a protein is a function of a subset of structural parameters of the protein. This subset of parameters is a simplified description of protein structure (topology fingerprint of the protein [39]). This simplified description makes statistical derivation of potentials of mean force feasible. All the potentials of mean force used in this contribution were derived from the database of high quality protein structures using the inverse Boltzmann law [42, 49]. Threading energy is a function of the following parameters of protein structure:

1. The local conformation of the protein backbone described by "chiral" squared distances between C_α atoms of residues (i-1) and (i+2). This parameter is closely related to

the secondary structure. The corresponding potential of mean force is related to the secondary structure propensity.

2. The burial status of the side chains in the sequence (the side chain is classified as buried if at least 70% of its surface is screened from the solvent). The corresponding potential of mean force describes the energy resulting from exposing a given side chain to the solvent or burying it in the protein interior.

3. The contacts between side chains (the side chains i and j are classified as interacting if the distance between any pair of heavy atoms in the side chains is less than 5 Å). The contact information defines the contact map of the protein. The corresponding potential of mean force describes the effective attraction or repulsion of the side chains.

The threading energy estimate of the protein is the function of parameters 1-3. It is described by the following formula:

$$E = \sum \varepsilon\left(A_i, A_{i+1}, r^{2}_{i-1,i+2}{}^{*}\right) + \sum_i \Gamma_i^A E_i\left(A_i\right) + \\ + \sum_i \sum_{i>j} C_{ij} E_2\left(A_i, A_j\right)$$

where: i, j are numbers of positions along the sequence;
A_i, A_j are residue types found at these positions;
$r^{2}_{i-1,i+2}{}^{*}$ is the value of the "chiral" squared distance between $C_{a\,i-1}$ and $C_{a\,i+2}$ atoms;
Γ_i^A is the burial status of a side chain i;
C_{ij} describes a contact between side chains i and j (C_{ij} =1 if i and j are in contact, otherwise C_{ij} =0).

The functions ε, E_1 and E_2, describe local, burial and contact interactions, respectively.

Acknowledgements The authors are grateful for many stimulating discussions with Drs. A. Bierzynski, A. Kolinski, J. Skolnick and J. Fetrow. We also thank Dr. A. Šali for help and advice in using his MODELLER software package, Dr. K. Olszewski (MSI) for help with the 3D Profile program and Dr. B. Zhang for preparation of Figure 1. Figure 3 was prepared using the MOLSCRIPT program kindly provided by Dr P. Kraulis. Support by the following grants is acknowledged: NSF grant MCB-9506278 (AG), and Howard Hughes Medical Institute grant 75195-543402 and BST-34/97 (LJ).

References

1. Sali, A.; Overington, J.P.; Johnson, M.S.; and Blundell, T.L. *TIBS* **1990**, *15*, 235-240.
2. Waterman, M.S., *Introduction to computational biology: Maps, sequences and genomes (Interdisciplinary Statistics)*. 1995: Chapman & Hall.
3. Svensonn, B.; Vass, I.; Cerdegren, E.; and Strying, S. *EMBO J.* **1990**, *9*, 2051-2060.
4. Siezen, R.J.; de Vos, W.M.; Leunissen, J.A.; and Dijkstra, B.W. *Protein Engineering* **1991**, *4*, 719-737.
5. Sali, A.; Matsumoto, R.; Neil, H.P.; Karplus, M.; and Stevens, R.L. *J.Biol.Chem.* **1993**, *268*, 9023-9034.
6. Schiffer, C.A.; Caldwell, J.W.; Kollman, P.A.; and Stroud, R.M. *Proteins* **1990**, *8*, 30-43.
7. Hilbert, M.; Bohm, G.; and Jaenicke, R. *Proteins* **1993**, *17*, 138-151.
8. Ring, C.S. and Cohen F.E., *FASEB Journal* **1993**, 783-790.
9. Johnson, M.S.; Srinivasan, N.; Sowdhamini, R.; and Blundell, T.L. *Crit. Rev. Biochem. & Mol.Biol.* **1994**, *29*, 1-68.
10. SYBYL 6.0, 1993, Tripos Inc., St. Louis, MO
11. Topham, C.M.; Thomas, P.; Overington, J.P.; Johnson, M.S.; Eisenmerger, F.;, and Blundell, T.L. *Biochem. Soc. Symp.* **1990**, *57*, 1-9.
12. INSIGHT 2.1, 1993, Molecular Simulations Inc., San Diego, CA
13. SWISS-MODEL, *http://www.expasy.ch/swissmod/SWISS-MODEL.html.*
14. Sali, A. and Blundell, T.L. *J.Mol.Biol.* **1993**, 234.
15. NAOMI, 2.4, 1996, *http://www.psynix.co.uk/products/naomi*, Psynix Information Systems Ltd.,
16. Lemer, C.; Rooman, M.; and Wodak, S. *Proteins* **1995**, *23*, 337-355.
17. Marchler Bauer, A.; Levitt, M.;, and Bryant, S.H. *Proteins* **1997**, *Suppl. 1*, 83-92.
18. Jaroszewski, L.; Rychlewski, L.; Zhang, B.; and Godzik, A. *Protein Science* **1998**, *7*, 1431-1440.
19. Rychlewski, L.; Zhang, B.; and Godzik, A. *Folding & Design* **1998**, *3*, 229-238.
20. Thornton, J.M. and Swindells, M.B. in *Molecular Structures in Biology*, R. Diamond, Editor; Oxford Univerity, Oxford: 1993.
21. Sali, A. *MODELLER 3.0, 1994.*
22. Sali, A.; Potterton, L.; Yuan, F.; van Vlijmen, H.; and Karplus, M. *Proteins* **1995**, *23*, 318-326.
23. Mossimann, S.; Meleshko, R.; and James, M.N. *Proteins* **1995**, *23*, 301-317.
24. Vogt, G.; Etzold, T.; and Argos, P. *J. Mol. Biol.* **1995**, *249*, 816-831.
25. Vingron, M. and Argos, P. *Prot. Engineer.* **1990**, *3*, 565-569.
26. Mevissen, H.T. and Vingron, M. *Prot. Engineer.* **1996**, *9*, 127-132.
27. Zuker, M. *J. Mol. Biol.* **1991**, *221*, 403-420.
28. Naor, D. and Brutlag, D.L. *J. Comp. Biology* **1994**, *1*, 349-366.
29. Needelman, S.B. and Wunsch, C.D. *J. Mol. Biol.* **1970**, *48*, 443-453.
30. Smith, T.F. and Waterman, M.S. *Adv. Appl. Math.* **1981**, *46*, 473-500.
31. Tomii, K. and Kanehisa, M. *Prot. Engineer.* **1996**, *9*, 27-36.
32. Pascarella, S. and Argos, P. *Prot. Engineer.* **1992**, *5*, 121-137.
33. Holm, L.; Ouzounis, C.; Sander, C.; Tuparev, G.; and Vriend, G. *Prot. Sci.* **1992**, *1*, 1691-1698.

34. Orengo, C.A.; Flores, T.P.; Jones, D.T.; Taylor, W.R.; and Thornton, J.M. *Curr. Biol.* **1993**, *3*, 131-139.
35. Zu-Kang, F. and Sippl, M. *Folding & Design* **1996**, *1*, 123-132.
36. Godzik, A. *Protein Science* **1996**, *5*, 1325-1338.
37. Kabsch, W. *Acta. Cryst.* **1978**, *A34*, 827-828.
38. Maiorov, V.N. and Crippen, G.M. *J. Mol. Biol.* **1994**, *235*, 625-634.
39. Godzik, A.; Skolnick, J.; and Kolinski, A. *J. Mol. Biol* **1992**, 227, 227-238.
40. Bowie, J.U.; Luethy, R.; and Eisenberg, D. *Science* **1991**, *253*, 164-170.
41. Sali, A. and Overington, J.P. *Protein Sci.* **1994**, 3.
42. Kolinski, A. and Skolnick, J. *Lattice Models of Protein Folding, Dynamics and Thermodynamics*. 1996, Austin,TX: R. G. Landes Company.
43. Rychlewski, L. and Godzik, A. *Prot. Eng.* **1997**, *10*, 1143-1153.
44. Pawlowski, K.; Jaroszewski, L.; Bierzynski, A.; and Godzik, A. in *Pacific Symposium on Biocomputing, 97*, R.B. Altman, *et al.*, Editors. 1997, World Scientific: Singapore. p. 328-339.
45. Sanchez, R. and Sali, A. *Proteins* **1997**, *Suppl. 1*, 50-58.
46. Holm, L. and Sander, C. *J. Mol. Biol.* **1993**, *233*, 123-138.
47. Wisconsin Package 9.1, 1997,, Genetics Computer Group, Inc., Madison, WI
48. Godzik, A.; Skolnick, J.; and Kolinski, A. *Prot. Engineering* **1993**, *6*, 801-810.
49. Godzik, A.; Kolinski, A.; and Skolnick, J. *Protein Science* **1995**, *4*, 2107-2117.
50. Pawlowski, K.; Bierzynski, A.; and Godzik, A. *J.Mol.Biol.* **1996**, *258*, 349-366.