doi:10.1016/j.jmb.2006.01.062

J. Mol. Biol. (2006) **357**, 1655–1668

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

Refinement of Protein Structures by Iterative Comparative Modeling and CryoEM Density Fitting

**Maya Topf[1], Matthew L. Baker[2], Marc A. Marti-Renom[1]
Wah Chiu[2] and Andrej Sali[1]***

[1] *Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry and California Institute for Quantitative Biomedical Research, QB3, 1700 4th Street Suite 503B, University of California at San Francisco San Francisco, CA 94143-2552 USA*

[2] *National Center for Macromolecular Imaging Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030 USA*

**Corresponding author*

We developed a method for structure characterization of assembly components by iterative comparative protein structure modeling and fitting into cryo-electron microscopy (cryoEM) density maps. Specifically, we calculate a comparative model of a given component by considering many alternative alignments between the target sequence and a related template structure while optimizing the fit of a model into the corresponding density map. The method relies on the previously developed *Moulder* protocol that iterates over alignment, model building, and model assessment. The protocol was benchmarked using 20 varied target–template pairs of known structures with less than 30% sequence identity and corresponding simulated density maps at resolutions from 5 Å to 25 Å. Relative to the models based on the best existing sequence profile alignment methods, the percentage of C^{α} atoms that are within 5 Å of the corresponding C^{α} atoms in the superposed native structure increases on average from 52% to 66%, which is half-way between the starting models and the models from the best possible alignments (82%). The test also reveals that despite the improvements in the accuracy of the fitness function, this function is still the bottleneck in reducing the remaining errors. To demonstrate the usefulness of the protocol, we applied it to the upper domain of the P8 capsid protein of rice dwarf virus that has been studied by cryoEM at 6.8 Å. The C^{α} root-mean-square deviation of the model based on the remotely related template, bluetongue virus VP7, improved from 8.7 Å to 6.0 Å, while the best possible model has a C^{α} RMSD value of 5.3 Å. Moreover, the resulting model fits better into the cryoEM density map than the initial template structure. The method is being implemented in our program *MODELLER* for protein structure modeling by satisfaction of spatial restraints and will be applicable to the rapidly increasing number of cryoEM density maps of macromolecular assemblies.

Keywords: protein structure prediction; comparative modeling; homology modeling; cryo-electron-microscopy; density fitting

Introduction

With the successes of whole genome sequencing, structural biology, and proteomics, there is a growing opportunity and need to study the structure and function of macromolecular assemblies.[1–3] Given the difficulties of applying X-ray crystallography and NMR spectroscopy to large assemblies, a key role is played by cryo-electron microscopy (cryoEM). While currently limited to intermediate resolutions (6–15 Å) for single particle reconstructions, cryoEM offers a number of advantages. Individual particles can be imaged in near-native conditions as well as in different functional states, enabling a structural analysis that can reveal salient features of the assembly and its components.[4–7] Additionally, fitting of high-resolution structures of assembly components into cryoEM density maps may provide pseudo-atomic models for the whole assembly,[8,9] and thus significant insights into

Abbreviations used: cryoEM, cryo-electron microscopy; RDV, rice dwarf virus; BDV, bluetongue virus; RMSD, root-mean-square deviation.

E-mail address of the corresponding author: sali@salilab.org

0022-2836/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.

the structure, function, and dynamics of single proteins and their complexes.[10–18]

Unfortunately, experimentally determined atomic-resolution structures of assembly components are frequently unavailable. Even when they are available, induced fit may severely limit their usefulness in the construction of the complex. In such cases, it may be possible to get useful models of the components with comparative protein structure modeling (homology modeling).[10,19–24,65] In comparative modeling, the structure of a target protein sequence is predicted by: (i) finding one or more related proteins of known structures (i.e. templates); (ii) aligning the target sequence to the template structures; (iii) building a model based primarily on the alignment from the previous step; and (iv) assessing the model. Currently, $\sim 1.1 \times 10^6$ of the $\sim 1.9 \times 10^6$ known proteins sequences[25] have at least one domain that can be modeled based on its similarity to one or more of the ∼32,000 known protein structures.[26] Thus, the number of models with an accuracy that is at least as high as that of the intermediate resolution cryoEM structures is almost two orders of magnitude greater than the number of experimentally determined atomic structures. Moreover, comparative protein structure prediction is becoming increasingly applicable and accurate not only due to the structural genomics initiatives, but also to the availability of faster computers and improved prediction methods.[20,22–24,26,27]

Despite these advances, incorrect fold assignments and/or target–template alignments are the primary sources of errors in comparative models, especially in models of sequences that are only remotely related to their templates (i.e. at less than 30% sequence identity). Unfortunately, most pairs of detectably related protein sequences and structures fall within this category, with correspondingly large alignment errors (Theory).[28,29] Other errors in comparative modeling include distortions and shifts of the backbone and sidechains.[21]

One way of minimizing the alignment errors is to explore multiple alternative alignments, generate the corresponding models, and identify the best ones in the ensemble of structures.[30] An example is the "moulding" method implemented in the program *Moulder* that iterates over alignment, model building, and model assessment.[31] During this iterative process: (i) new alignments are constructed by application of five different genetic algorithm operators, such as alignment mutations and crossovers; (ii) comparative models corresponding to these alignments are built by satisfaction of spatial restraints as implemented in the program *MODELLER*;[19] and (iii) the models are evaluated by a composite model assessment score.[32] Using initial alignments generated from established methods, such as PSI-BLAST[33] and SAM (version 3.3.1),[34] moulding has been shown to improve the alignments by 15%–25%. However, the accuracy of moulding is currently limited by the ability of the fitness function to identify the most accurate model.[31]

Recently, we demonstrated that alternative comparative models can be ranked by fitting them into a cryoEM density map.[35] Following this work, we focus here on model building and refinement that depends on both the cryoEM density and comparative modeling considerations (*Moulder-EM*) (Theory). The method was tested against a benchmark consisting of 20 cases of target–template pairs of known structures that are related at less than 30% sequence identity (Results). To illustrate the method with a practical example, we applied it to a very difficult case, the modeling of the upper domain of the P8 capsid protein of rice dwarf virus. Finally, we discuss the implications of the results for comparative protein structure modeling and for improving the interpretation of cryoEM density maps of whole macromolecular assemblies (Discussion).

Theory

Errors in comparative modeling

The primary requirement for reliable comparative modeling is a detectable similarity between the sequence of interest (target sequence) and a known structure (template).[36,37] The detected similarity between the target and the template sequences is usually quantified in terms of sequence identity or statistical measures such as *E*-value or *Z*-score, depending on the method used. Sequence–structure relationships are coarsely classified into three different regimes in the sequence similarity spectrum: (i) the easily detected relationships characterized by >30% sequence identity; (ii) the "twilight zone" corresponding to relationships with statistically significant sequence similarity in the 10%–30% range; and (iii) the "midnight zone" corresponding to statistically insignificant sequence similarity.[38]

For closely related protein sequences with identities higher than 30%–40%, the alignments produced by all methods are almost always largely correct. The sensitivity of the search and accuracy of the alignment become progressively difficult as the relationships move into the twilight-zone.[38,39] In the twilight-zone, profile-sequence alignment methods[40] (e.g. PSI-BLAST,[33] SAM,[34] HMMER,[41] and *profile.build* in *MODELLER*-8[26]) are more sensitive in detecting related structures than the pairwise sequence-based methods, resulting in approximately 45% of residues in the 0–40% sequence identity range aligned correctly.[29,42] However, profile–profile alignment methods (e.g. FFAS,[43] SP3,[44] HHpred,[48] SALIGN,[29] and the *profile.scan* command in *MODELLER*-8[29] have proven to include the most sensitive and accurate fold assignment and alignment protocols to date.[28,29,45,46] These methods detect ∼28% more relationships at the superfamily level and improve the alignment accuracy by 15%–20% compared to profile-sequence alignment methods.[29,44]

As the sequence identity drops below the threshold of the twilight zone, there is usually insufficient signal in the sequences or in their profiles for the sequence-based methods to detect true relationships and produce accurate alignments.[47] Sequence–structure threading methods (e.g. GenTHREADER,[49] 3D-PSSM,[50] FUGUE,[51] SP3,[44] and SAM-T02 multi-track HMM[52,53]) are most useful in this regime as they can sometimes recognize common folds even in the absence of any statistically significant sequence similarity.[54]

And finally, alignment errors can also be minimized by iterating over the process of calculating alignments, building models, and evaluating models (i.e. moulding).[30,31,55–57]

The *Moulder-EM* protocol

The previously developed program *Moulder* uses a genetic algorithm that iterates over alignment, model building, and model assessment (Figure 1).[31] Briefly, the protocol begins by calculating the initial alignments of each target–template pair that are then submitted to 25 iterations of alignment, model building, and model assessment. Approximately 300 child alignments are produced in each iteration and the corresponding molecular models are built using the *automodel* class of the *MODELLER*-8 Python interface. Alignments for the 25 models with the best fitness function scores are selected as the parent alignments for the next iteration. The final model is the best scoring model from any of the iterations. In this study, we improved the initial alignments, the fitness function that guides the genetic algorithm, as well as the final model assessment; the latter two improvements rely on the cryoEM density maps (*Moulder-EM*). The protocol was implemented on a cluster of computers running Linux. For a 150 residue target sequence, the protocol currently requires ~15 h of CPU time on 50 nodes with dual 1.5 GHz Pentium III CPUs; the CPU time scales approximately linearly with the length of the sequence.

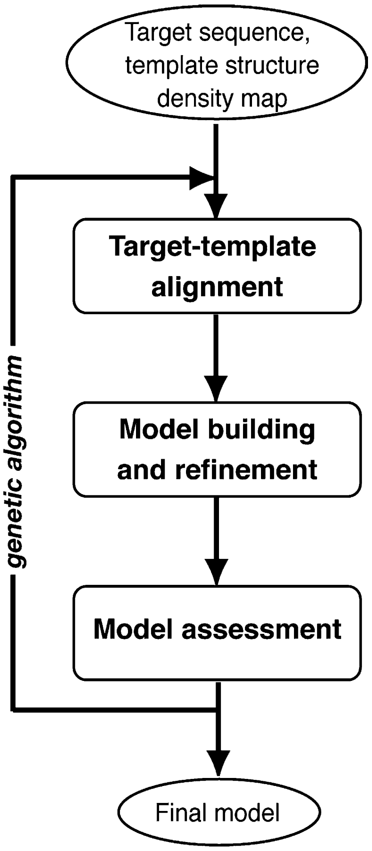

Target sequence, template structure density map
Target-template alignment
Model building and refinement
Model assessment
genetic algorithm
Final model

Figure 1. The *Moulder-EM* protocol for improving the accuracy of molecular models determined by comparative protein structure modeling and fitting into cryoEM density maps. This protocol uses genetic algorithm operators[31] to evolve the alignments and the corresponding models according to a fitness function consisting of both a statistical potential Z-score[32] and a density-fitting Z-score between a model and a cryoEM density map (Mod-EM).[35]

Preparation of initial target–template alignments

First, a multiple sequence alignment (profile) was built for each of the target and template sequences using the *profile.build* command in *MODELLER*-8[26] Next, a profile–profile alignment was calculated between the target and template profiles using the *profile.scan* command of *MODELLER*-8.[29] *profile.scan* was used with changing gap penalties (opening gap penalty ranging from −5000 to 0 in steps of 500, extension gap penalty ranging from −250 to 0 in steps of 50) to get an initial population of 5 to 15 different alignments after redundancy removal.

Fitness function of *Moulder-EM*

The fitness function is a linear combination of a statistical potential Z-score (Z_s) and a density fitting Z-score (Z_c): $F = w_1 Z_s - w_2 Z_c$, where w_1 and w_2 are the weights of the two Z-scores. We used Z-scores instead of the original scores (Z-score = (score − μ)/σ) to ensure good performance of a single set of weights w_1 and w_2 across a full spectrum of applications. The statistical potential Z-scores were calculated using the mean, μ, and standard deviation σ of 200 statistical potential scores obtained from threading random sequences of the same composition as the target sequence onto the structure of the assessed model.[32] The statistical potential score of a model is the sum of the solvent-accessibility terms for all C^{α} atoms and distance-dependent terms for all pairs of C^{α} and C^{β} atoms. The solvent-accessibility term for a C^{β} atom depends on its residue type and the number of other C^{β} atoms within 10 Å; the non-bonded terms depend on the atom and residue types spanning the distance, the distance itself, and the number of residues separating the distance-spanning atoms in the sequence. These potential terms reflect the statistical preferences observed in 760 non-redundant proteins of known structure.

The density-fitting Z-score is a normalized density-fitting score. The density-fitting score is the maximized cross-correlation coefficient between the cryoEM density map and the probe (model) density calculated with Mod-EM (i.e. the *density.grid_search* command in *MODELLER*-8).[35] The normalization relies on the mean and standard deviation obtained from a population of ~7500 alignments constructed in 25 iterations of the *Moulder* program with the original fitness function that depends only on the statistical potential (i.e. independent of the density-fitting score). When the fit is good, the density-fitting *Z*-score is positive; it usually ranges from −10 to 10.

Five protocols of *Moulder-EM* were tested, corresponding to different weights ($[w_1,w_2]$) of [1,0], [1,1], [1,2], [1,8], and [0,1] for the statistical potential *Z*-score and the density-fitting *Z*-score in the fitness function, respectively.

Final model selection

As in the original *Moulder* program, after 25 iterations of alignment, model building, and model assessment, the final alignment is selected by relying on a composite score that is based on the mean and standard deviation of the entire population of alignments and corresponding models (thousands). The use of this composite criterion instead of the fitness function for the selection of the final alignment is based on the benchmarks of *Moulder*.[31] Previously, the composite score was a linear combination of five *Z*-scores[31]: pair statistical potential (P_p), surface statistical potential (P_s), structural compactness (S_c), harmonic average distance (H_a), and an alignment score (A_s). In the *Moulder-EM* protocol, the composite model assessment score (Z') is a linear combination of the original composite score (Z) and the density-fitting *Z*-score (Z_c) using the same weights as in the calculation of the fitness function *F*: $Z'=w_1\cdot Z-w_2\cdot Z_c$. Here, Z_c is calculated by normalizing the density-fitting *Z*-scores using as a reference the models created during *Moulder-EM* iterations.

Benchmark

The benchmark for testing the new moulding protocol consists of 20 pairs of proteins of known structure sharing between 10% and 31% sequence identity (17% on average), including target–template pairs from the two original studies[31,35] as well as several new pairs (Table 1). These proteins range in size from 81 to 388 residues (203 on average) and represent all major fold classes (i.e. α, β, α+β, and α/β).[58] For each of the native structures of the 20 target proteins, a density map was simulated at 10 Å resolution using the PDB2MRC command in the *EMAN* package,[59] an achievable resolution for single-particle cryoEM. For three proteins in the benchmark, additional density maps were simulated at 5 Å , 15 Å , 20 Å , and 25 Å resolution.

Modeling the upper domain of the rice dwarf virus P8 protein

To illustrate the *Moulder-EM* protocol to a practical problem where even fold assignment is ambiguous, we applied it to the upper domain of the P8 protein of rice dwarf virus (RDV) (EMDB code: 1060; PDB code: 1uf2),[60,61] corresponding to the middle part of the protein (residues 143–252). The lower domain was not modeled because the detected template structure[60] has significant secondary structure segment shifts, a problem that we are not addressing in this study. The corresponding density map was segmented from the original 6.8 Å resolution cryoEM density map.[60] The VP7 protein of bluetongue virus (BTV) (PDB code 1bvp) was selected as the template structure. Although there is no statistically significant sequence similarity between RDV P8 and BTV VP7 (3% sequence identity), they share a common fold. Thus, this case study is more difficult than any of the proteins in the benchmark. In constructing the initial alignments, it was impossible to detect any sequences related to both RDV P8 and BTV VP7 by using the *profile.build* command in *MODELLER*-8. Instead, two programs that take into consideration the secondary structure when determining the gap penalty at each position were used: the *alignment. align2d* command in *MODELLER*-8[62] and Fugue.[51] The optimally scoring alignments from each program were refined by 30 *Moulder-EM* iterations using a fitness function with the previously determined optimal weights $[w_1,w_2]$ of [1,2]. We performed 30 iterations instead of the default 25 to make sure the protocol converged.

Measures of alignment accuracy

The accuracy of a target–template alignment was measured by superposing the corresponding target model to its native structure; the target model was calculated with the *automodel* class of *MODELLER*-8. Two criteria were used to assess the geometrical accuracy of the models of each protein. First, the root-mean-square deviation (RMSD) between the corresponding C^α atoms in the model and the native structure was calculated using rigid-body least-squares superposition of all the C^α atoms, as implemented in the *model. superpose* command of *MODELLER*-8. Second, the "native overlap" (NO) was defined as the percentage of the C^α atoms in the model that are within 5 Å of the corresponding atoms in the superposed native structure. For reference, the accuracy of the target–template structure-based alignment calculated with the program CE[63] was assessed in the same way; the CE structure-based alignment optimizes the number of C^α positions superposed within 4 Å of each other while minimizing the RMSD between the two sets of superposed positions.

For assessing the statistical significance of a difference between the accuracies of two alignment methods, the parametric Student's *t*-test (at the 95%

Table 1. The 20 target–template pairs of proteins with known structures included in the benchmark

						Final alignment ($[w_1,w_2]$)									
Target–template		Structural alignment		Initial alignment		[1,0]		[1,1]		[1,2]		[1,8]		[0,1]	
PDB codes	Seq. id.	NO (%)	CR (Å)	NO (%)	CR (Å)	NO (%)	CR (Å)	NO (%)	CR (Å)	NO (%)	CR (Å)	NO (%)	CR (Å)	NO (%)	CR (Å)
A. *Difficult set*															
1atnA(4-354)-1atr(3-382)	13.4	79.2	6.8	45.6	9.3	47.3	9.5	48.4	9.4	46.4	9.4	55.8	8.8	49.0	9.1
1cauB(246-423)-1cauA(48-224)	18.9	86.0	3.5	51.1	8.9	55.6	8.7	65.2	8.9	69.1	4.4	59.6	9.0	65.2	8.1
1eaf(433-633)-4cla(31-218)	19.7	86.1	4.0	49.3	7.5	62.7	6.3	68.2	5.2	64.2	5.5	67.7	5.2	64.6	6.7
1gky(1-186)-3adk(8-194)	19.0	74.2	5.9	68.3	6.6	70.4	6.2	71.0	6.8	73.1	6.5	60.8	7.2	50.5	7.8
1ltsD(17-102)-1bovA(2-69)	4.4	81.4	3.8	36.0	13.1	16.3	13.1	43.0	10.0	59.3	9.4	82.6	3.8	54.7	7.3
1mup(10-161)-1rbp(11-175)	15.4	87.5	3.6	65.8	6.2	67.8	7.0	66.4	6.7	67.1	6.2	67.8	4.6	65.8	5.0
1ten(803-891)-3hhrB(131-233)	18.4	89.9	4.0	65.6	5.1	55.6	5.6	87.8	3.7	92.2	3.7	94.4	3.7	93.3	3.4
2afnA(36-324)-1aozA(1-300)	18.5	87.9	3.4	61.9	7.7	65.4	5.7	64.0	6.2	64.0	5.5	64.4	6.4	59.9	6.0
2omf(10-340)-2por(1-301)	13.2	72.5	5.2	40.2	8.1	35.3	8.9	61.9	6.4	56.8	6.9	58.9	6.7	50.8	6.7
2sar(7A-91A)-9rnt(2-104)	12.5	74.1	4.9	54.1	5.6	56.5	5.6	67.1	5.0	74.1	5.0	63.5	5.3	65.9	4.9
3hlaB(4-98)-2rhe(3-108)	2.4	82.1	4.3	47.4	8.1	47.4	7.3	50.5	6.2	52.6	7.0	35.8	6.3	37.9	6.4
8i1b(7-151)-4fgf(20-143)	14.1	87.5	3.2	36.1	9.1	48.6	8.6	43.1	8.2	69.4	5.6	61.1	5.7	61.1	6.5
Average	14.2	82.4	4.4	51.8	8.0	52.4	7.7	61.4	6.9	65.7	6.3	64.4	6.1	59.9	6.5
B. *Easy set*															
1et0A(7-289)-1daaA(2-273)	21.8	87.4	3.9	87.8	4.7	81.1	7.2	83.9	4.5	87.0	4.5	83.9	4.5	84.3	5.4
1occA(1-188)-1fftC(20:203)	14.7	80.9	3.4	81.9	3.8	84.0	3.4	79.3	4.4	86.2	3.2	76.6	3.8	84.0	3.4
1ivg(82-469)-1nsbA(76-465)	26.8	84.3	4.3	88.1	3.2	86.1	3.3	85.3	3.5	86.9	3.5	93.3	2.7	84.0	3.6
1lgaA(13-291)-2cyp(15-289)	18.0	92.1	2.9	82.1	4.2	84.6	3.8	84.2	4	81.7	3.8	80.6	4.3	77.1	4.8
2 cmd(1-310)-6ldh(21-328)	21.8	91.0	3.8	80.0	4.9	79.5	5.5	81.0	4.2	80.5	4.2	78.6	5.0	73.8	4.1
2fbjL(1-210)-8fabB(1-221)	23.4	87.4	3.9	87.8	4.7	82.7	7.7	83.9	4.5	87.0	4.5	83.9	4.5	84.3	5.4
2mtaC(45-125)-1ycc(2-103)	13.0	93.8	2.5	70.4	3.9	77.8	3.7	81.5	3.4	85.2	3.0	79.0	3.7	82.7	3.3
2pna(20-119)-1shaA(3-104)	31.2	90.0	3.2	86.0	3.5	85.0	3.6	91.0	3.3	86.0	3.4	83.0	4.3	81.0	4.5
Average	21.8	89.2	3.4	81.8	4.1	82.0	4.4	83.1	3.9	84.7	3.6	81.8	4.0	80.2	4.2
Total average	17.0	85.1	4.0	63.8	6.4	64.3	6.4	70.1	5.7	73.1	5.2	71.3	5.2	68.1	5.6

CR is the C^α RMSD value between a model and the corresponding native structure and NO is the native overlap within 5 Å cutoff (see Theory). Target–template sequence identity (Seq. id.) is calculated from the CE structure-based alignment.[63] The initial alignment column refers to the model with the highest statistical potential Z-score (Z_s) among the models built from the initial profile–profile alignments. The final alignment column refers to the model from the whole population that has the highest composite model assessment score (Z'). The ($[w_1,w_2]$) column gives the weights for the statistical potential (Z_s) and the density-fitting Z-score (Z_c), which define the fitness function (F) for all the iterations and for the composite assessment score (Z') in the final model selection.

confidence value) was applied. The accuracy of a method was measured independently by the average C$^{\alpha}$ RMSD value and the average native overlap of all the models in the benchmark. The compared methods included the CE structure-based alignment, the initial alignment for *Moulder-EM* by *profile.scan*, and the five *Moulder-EM* protocols, where ([w_1,w_2] is [1,0], [1,1], [1,2], [1,8], and [0,1]).

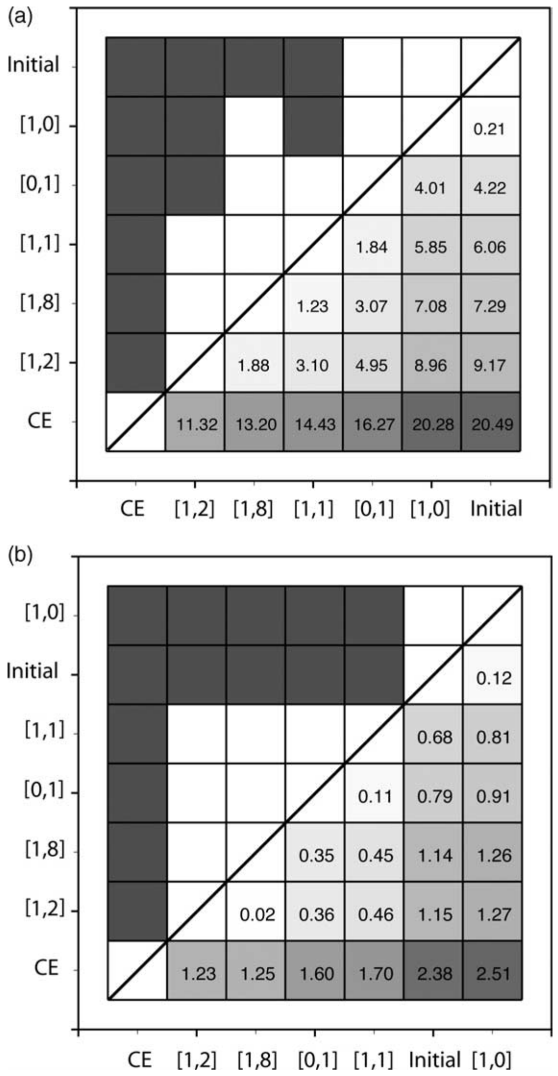
Results

Accuracy of models refined by *Moulder-EM*

To test the *Moulder-EM* protocol, we created a benchmark of 20 target–template protein pairs, related at less than 31% sequence identity that vary in fold and size. The initial alignments for each of these pairs were obtained with the profile–profile alignment method. They were refined by five *Moulder-EM* protocols with different weights (w_1 and w_2) for the statistical potential *Z*-score (Z$_s$) and the density-fitting *Z*-score (Z$_c$) in the fitness function. The five weight sets were chosen to sample reasonable combinations of the statistical potential and density-fitting scores. The accuracies of the highest-scoring initial models and the refined models were quantified by the C$^{\alpha}$ RMSD value and native overlap (Table 1; Figures 2 and 3). The Student's *t*-test shows that all *Moulder-EM* protocols guided by a combination of the statistical potential and density-fitting scores (i.e. [w_1,w_2] is [1,1], [1,2], and [1,8]) resulted in models with significantly better C$^{\alpha}$ RMSD and native overlap compared to the initial models and to the final models based on the fitness function dependent only on the statistical potential (i.e. [w_1,w_2] is [1,0]) (Figure 2). Even moulding with the fitness function dependent only on the density-fitting score ([w_1,w_2] is [0,1]) resulted in a significant improvement in C$^{\alpha}$ RMSD relative to the initial models and to the final models based on the fitness function dependent only on the statistical potential.

Based on the initial native overlap and C$^{\alpha}$ RMSD value, the 20 targets in the benchmark were divided into two groups, "difficult" and "easy". The difficult group includes the targets with an initial native overlap ≤70% and C$^{\alpha}$ RMSD ≥5 Å, while the easy group includes targets with native overlap >70% and C$^{\alpha}$ RMSD <5 Å; there are no cases with native overlap ≤70% and C$^{\alpha}$ RMSD <5 Å (Table 1). Both the difficult and easy targets exhibited improvements in the native overlap as well as the C$^{\alpha}$ RMSD values for all tested weights of the statistical potential and density-fitting *Z*-scores; however, due to poor initial alignments, the improvements upon refinement are larger for the difficult group than for the easy group. On average, modeling of the difficult group based on the fitness function dependent only on the statistical potential did not result in any significant improvement in the native overlap and the C$^{\alpha}$ RMSD (51.8% to 52.4% and 8.0 Å to 7.7 Å, respectively). In contrast, modeling based on the fitness function dependent only on density fitting improved the native overlap from 51.8% to 59.9% and the C$^{\alpha}$ RMSD value from 8.0 Å to 6.5 Å.

In both the difficult and easy groups, the *Moulder-EM* protocol guided by a fitness function consisting of the statistical potential *Z*-score and the density-fitting *Z*-score with weights of 1 and 2, respectively, performed best on average among all five tested protocols (Table 1; Figure 2). The native overlap and the C$^{\alpha}$ RMSD of the final models in the difficult



(a)

	CE	[1,2]	[1,8]	[1,1]	[0,1]	[1,0]	Initial
Initial							
[1,0]							0.21
[0,1]						4.01	4.22
[1,1]					1.84	5.85	6.06
[1,8]				1.23	3.07	7.08	7.29
[1,2]			1.88	3.10	4.95	8.96	9.17
CE		11.32	13.20	14.43	16.27	20.28	20.49

(b)

	CE	[1,2]	[1,8]	[0,1]	[1,1]	Initial	[1,0]
[1,0]							
Initial							0.12
[1,1]						0.68	0.81
[0,1]					0.11	0.79	0.91
[1,8]				0.35	0.45	1.14	1.26
[1,2]			0.02	0.36	0.46	1.15	1.27
CE		1.23	1.25	1.60	1.70	2.38	2.51

Figure 2. Comparison of the alignment accuracies achieved by the following methods: profile–profile sequence alignment by the *profile.scan* command of *MODELLER*-8 (initial); structure-based alignment (CE), *Moulder-EM* with five different weight combinations ([w_1,w_2]) defining the fitness function (Table 1). (a) The difference between the average native overlaps (measured in %). (b) The difference between the average C$^{\alpha}$ RMSDs (measured in Å). Upper diagonal: gray and white squares indicate pairs of protocols whose performances are and are not statistically significantly different at the confidence level of 95%, respectively. Lower diagonal: the intensity of gray is proportional to the accuracy difference between the compared methods.

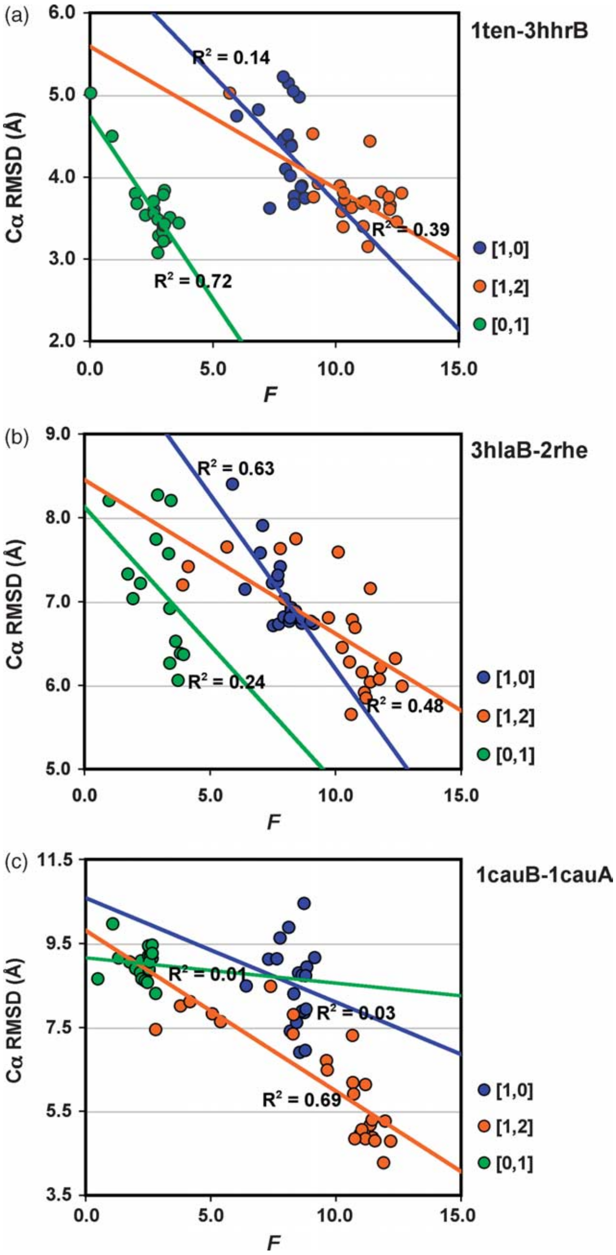


Figure 3. C^{α} RMSD of the highest-ranking model at each iteration of *Moulder-EM versus* its fitness function (*F*). The results are presented for three sample target–template pairs ((a)–(c)). Each symbol represents a fitness function with different weights ($[w_1,w_2]$) for the statistical potential *Z*-score and the density-fitting *Z*-score, as indicated. *R* is the Pearson's correlation coefficient.

group were improved from 51.8% to 65.7% and from 8.0 Å to 6.3 Å, respectively. In contrast, the improvements for the easy targets were smaller (from 81.8% to 84.7% in the native overlap and from 4.1% to 3.6% in the C^{α} RMSD) because the initial alignments were already close to the CE structure-based alignments. In fact, in some cases, the final models are better than the models calculated using the structure-based alignment, because of more accurate loop modeling. For instance, for 1ten modeling based on 3hhrB, the final model has 92.2% native overlap and 3.7 Å C^{α} RMSD, while the model based on the CE structure-based alignment has only 89.9% native overlap and 4.0 Å C^{α} RMSD (Table 1). In modeling 1ivg based on 1nsbA, the final model and the model from the CE structure-based alignment have 86.9% and 84.3% native overlap and 3.5 Å and 4.3 Å C^{α} RMSD, respectively. While none of the model assessment criteria used in this work allow us to predict whether a given target-template pair will be easy or difficult, only one initial alignment (livg-lnsbA, in the easy group) becomes slightly worse as a result of the refinement by *Moulder-EM* with the optimal weights for the statistical potential and density-fitting scores in the fitness function (i.e. $[w_1,w_2]$ is [1,2]).

Sample model optimization

To illustrate the *Moulder-EM* optimization protocol, we describe the improvement in the model of 8i1b based on its 14.1% sequence identity to 4fgf (Table 1; Figures 4 and 5). Using the optimal *Moulder-EM* protocol [*w1*, *w2*] is [1,2], the models generally become more accurate during the iterative process (Figure 4): while the highest-ranking initial model (iteration 0) has a native overlap of 36.1% and a C^{α} RMSD of 9.1 Å, the highest-ranking model in the last iteration (iteration 25) has a native overlap of 68.1% and a C^{α} RMSD of 5.2 Å. The corresponding density-fitting *Z*-scores of these models are −1.6 and 2.8, respectively. Unfortunately, due to limitations of the fitness function, it is generally not possible to select the most accurate model at any iteration. This problem is ameliorated by the composite model assessment score (Z') used in the final model selection (see Theory). The composite score Z' selects a model from the entire population of the models from all iterations that is generally more accurate than the model selected by the fitness function in the last iteration, in agreement with previous experience.[31] For example, in the modeling of 8i1b based on 4fgf, the most accurate model in the population has a native overlap of 59.0%, 72.2%, and 75.0% for the weights of [1,0], [1,2], and [0,1], respectively. Although the model with the best fitness function score in the last iteration has a native overlap of only 43.1%, 68.1%, and 61.1%, the final model based on the composite score Z' has a native overlap of 48.6%, 69.4%, and 61.1%, respectively (Figure 4; Table 1).

The effect of map resolution on model accuracy

In cryoEM, the resolution of the density map often dictates what structural features can be accurately inferred from it. To understand how the resolution affects the *Moulder-EM* refinement protocol, we looked at the accuracy of the final models achieved by the fitness function (*F*) and the composite model assessment score (Z')

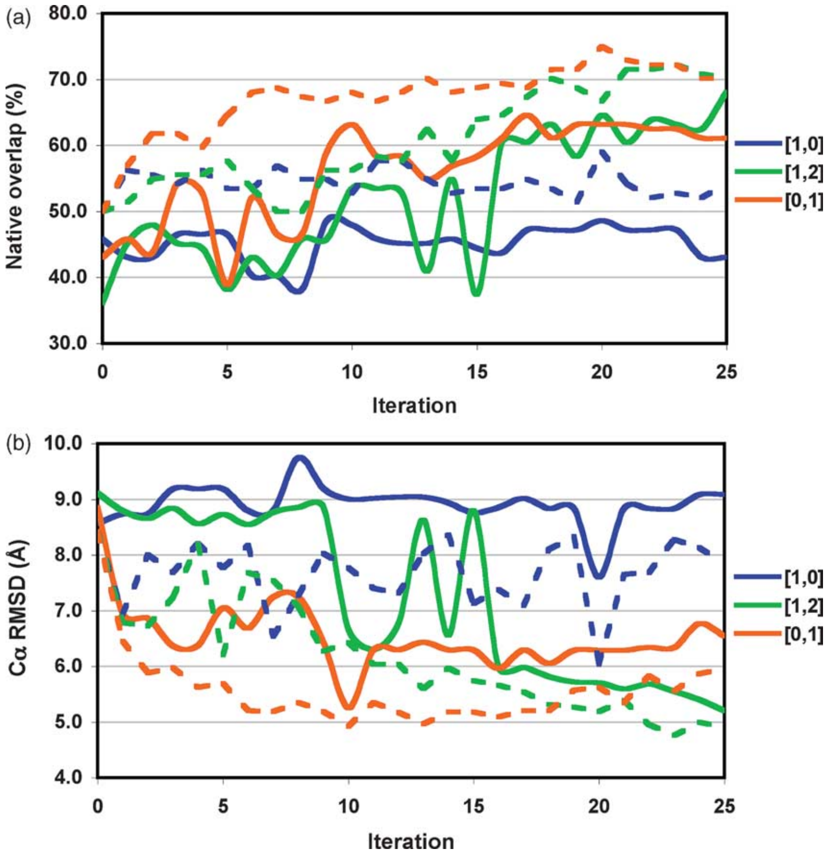


Figure 4. (a) Native overlap and (b) C^{α} RMSD of the highest-ranking models at each iteration of *Moulder-EM* applied to the refinement of the 8i1b target model based on the 4fgf template. The continuous lines represent the model with the best fitness function score, while the broken lines represent the most accurate model based on (a) the native overlap and (b) C^{α} RMSD. Each color represents a fitness function with different weights ($[w_1,w_2]$) for the statistical potential *Z*-score and the density-fitting *Z*-score, as indicated.

dependent on the density-fitting *Z*-score alone (Z_c) ($[w_1,w_2]$ is [0,1]), using density maps at different resolutions (5, 10, 15, 20, and 25 Å) (Figure 6). We also compared these results to the final model achieved by the fitness function (*F*) and by the composite model assessment score (Z') dependent on the statistical potential *Z*-score alone ($[w_1,w_2]$ is [1,0]), which of course has no dependence on the resolution of the cryoEM map.

In the three tested cases, if the resolution is better than 15 Å, the C^{α} RMSDs (except for 2omf at 5 Å resolution) and the native overlaps of the final models based on the density-fitting *Z*-score are better or equal to those based on the statistical potential *Z*-score. At 20–25 Å resolution, the accuracy of the refined models is generally comparable to that based on a statistical potential alone. For the easy target 2mta (where the best initial alignment already corresponds to a model with 70.4% native overlap), improvements over the initial model are noticeable when using density maps at 5 Å and 10 Å resolution. Such improvements are

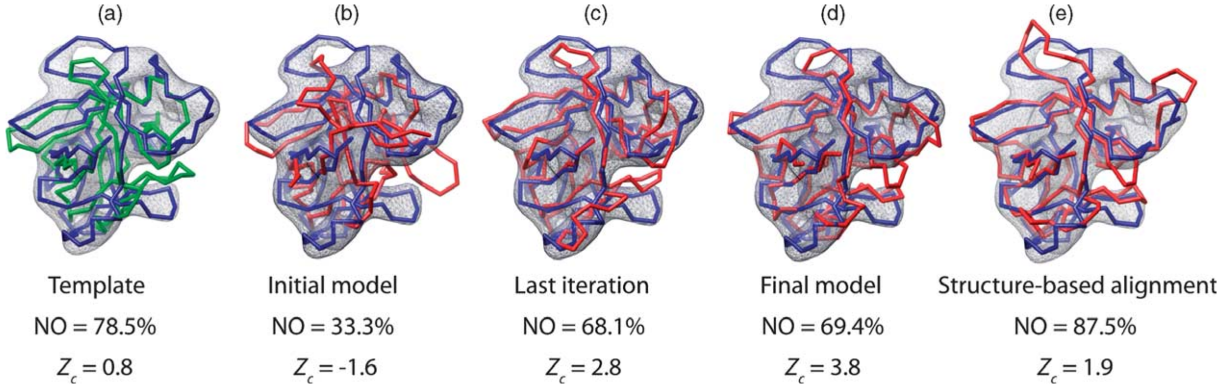


Figure 5. Optimal fitting of different structures into the simulated 10 Å resolution density map of 8i1b. The native structure (8i1b) is shown in blue ((a)–(e)), the template (4fgf) in green (a), and the models are shown in red ((b)–(e)). The model with the highest fitness function score (with weights of [1,2] for the statistical potential *Z*-score and the density-fitting *Z*-score) in the first iteration (initial model) and the last iteration is shown in (b) and (c), respectively. The final model with the best composite score is shown in (d) and the model from the CE structure-based alignment in (e). NO is the native overlap of the template and the models, and Z_c is their density-fitting *Z*-score. The Z_c of the native structure is 9.4. The image was created with the molecular graphics program Chimera.[66]

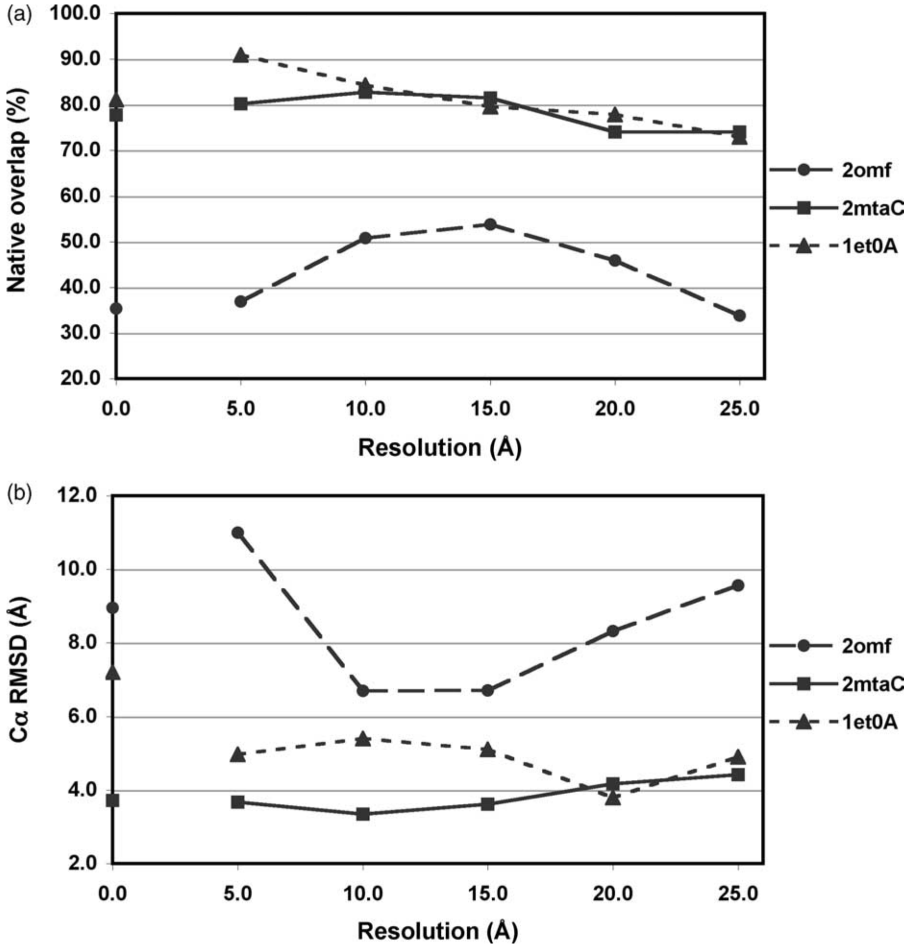


Figure 6. (a) Native overlap and (b) C^{α} RMSD of the final model with the best composite score as a function of the resolution. The model at "zero" resolution was selected from a population evolved *via* a fitness function that was dependent only on the statistical potential *Z*-score ([w_1,w_2] is [1,0]), while the remaining models at different resolutions were selected from a population that was evolved by a fitness function that is dependent on the density-fitting *Z*-score alone ([w_1,w_2] is [0,1]).

less noticeable at lower resolution (worse than 20 Å). For the easy target 1et0, fitting into the 5 Å resolution map improved the native overlap even beyond that of the model from the CE structure-based alignment (90.1% *versus* 87.4%). Conversely, the final model for the difficult target 2omf improved using the lower-resolution maps, while it deteriorated using the highest-resolution map (the initial model has 40.2% native overlap and 8.1 C^{α} RMSD). At 15 Å resolution, the native overlap and C^{α} RMSD improved to 53.8% and 6.8 Å, respectively, while at 5 Å resolution, they deteriorated to 36.9% and 11.0 Å, respectively. We note in passing that for the standard fitness function ([1,2]), the native overlap improved to 51.7% and the C^{α} RMSD to 7.5 Å even with the 5 Å resolution map.

Modeling of RDV P8 upper domain using an experimentally determined cryoEM density map

Moulder-EM was applied to the upper domain of the RDV P8 capsid protein (46 kDa). The density map of the whole virus at 6.8 Å resolution (EMDB code: 1060)[60] as well as the crystal structure at 3.5 Å resolution (PDB code: 1uf2)[61] are available. While no statistically significant sequence similarity can be found between P8 and other proteins of known structure, a structural homolog from a related virus, BTV VP7, is known (PDB code: 1bvp).[64] In particular, the upper domains of P8 and VP7 share a β sandwich fold. The highest-ranking initial model of P8 based on VP7 had a native overlap of 36.7% and a C^{α} RMSD of 8.7 Å. Using the fitness function and the composite model assessment score that performed best in the benchmark [w_1,w_2] is [1,2], the native overlap improved to 53.3% and the C^{α} RMSD to 6.0 Å. The *Moulder-EM* improvement over the initial model is substantial, in particular in the positions of the secondary structure elements (Figure 7(a)). Nevertheless, there is still room for further improvement: the best possible model corresponding to the CE alignment of the two structures has 84.2% native overlap and 5.3 Å C^{α} RMSD. Although some of the loops do not fit the density well (primarily due to the errors in the target–template alignment), the final model fits the cryoEM density map better than

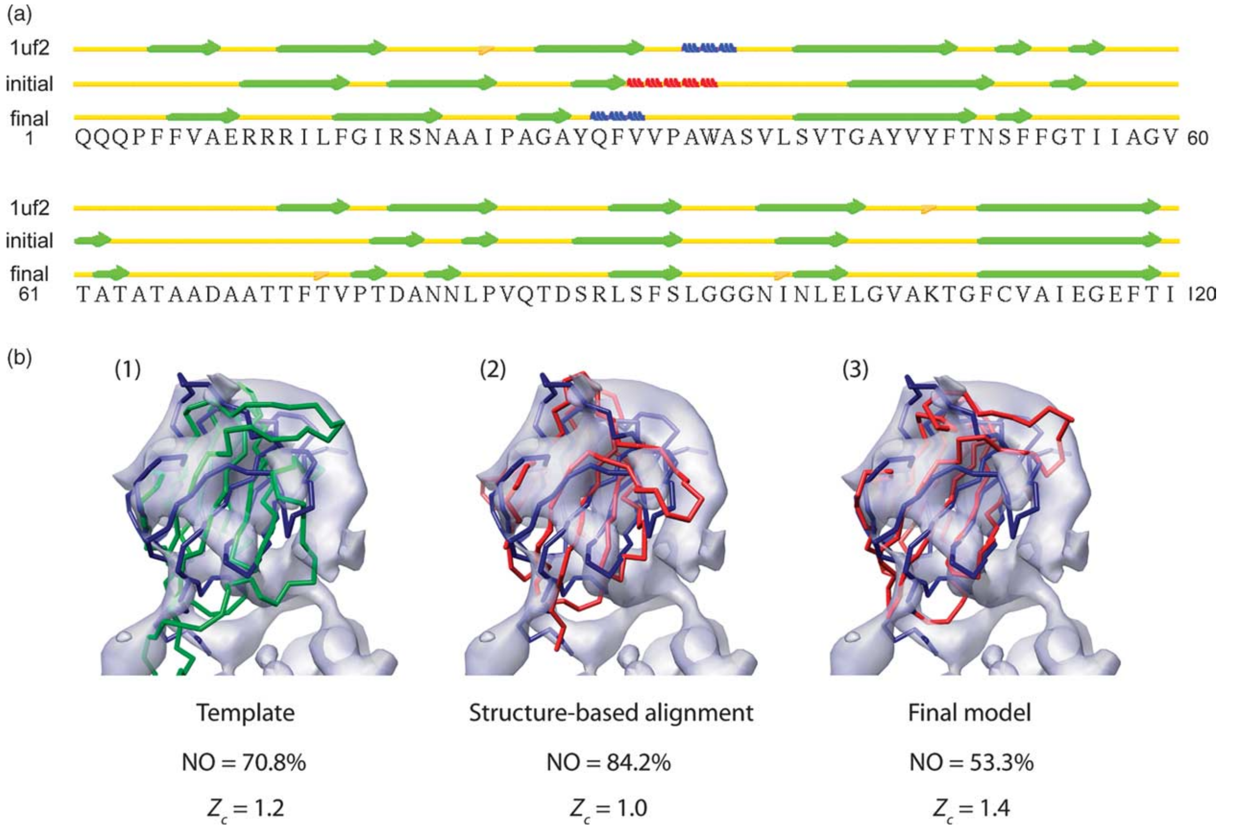

(a)
1uf2
initial
final
1
QQQPFFVAERRRILFGIRSNAAIPAGAYQFVVPAWASVLSVTGAYVVYFTNSFFGTIIAGV
60
1uf2
initial
final
61
TATATAADAATTFTVPTDANNLPVQTDSRLSFSLGGGNINLELGVAKTGFCVAIEGEFTI
120
(b)
(1)
(2)
(3)
Template
NO = 70.8%
$Z_c = 1.2$
Structure-based alignment
NO = 84.2%
$Z_c = 1.0$
Final model
NO = 53.3%
$Z_c = 1.4$

Figure 7. Application of *Moulder-EM* to the experimentally determined cryoEM density map of RDV P8 upper domain. (a) The secondary structure segments in the native structure, the initial model, and the final model are indicated in red (α helix), blue (3-10 helix), and green (β-strand) as assigned by the Stride web server.[67] (b) The template (1bvp, green) and the native structure (1uf2, blue) fitted into the 6.8 Å resolution cryoEM density map of P8 are shown in (1). The model from the CE structure-based alignment (red) is shown in (2). The final model (red) is shown in (3). NO is the native overlap of the model and Z_c is the density-fitting Z-score. The Z_c of the native structure is 5.0. The image was created with the molecular graphics program Chimera.[66]

either the template structure or the model from the CE structure-based alignment (the density-fitting Z-scores are 1.4, 1.2, and 1.0, respectively) (Figure 7(b)).

Discussion

"Moulding" and density fitting of comparative models

We recently showed that the density-fitting score at intermediate resolution is highly correlated with the accuracy of the model; that is, one of the most accurate models can usually be identified based on the quality of its fit into the density map.[35] Here, we took the next step from ranking models by fitting into a cryoEM density map to refining an initial structure by iterative modeling and fitting into a cryoEM density map. Such a refinement is particularly needed in comparative modeling of remotely related proteins where the sequence-structure alignment between the target sequence and the template structure represents a major source of errors in the target model; the alignment errors result in approximately half of the total error in comparative models, the rest of the error originating from our inability to model inserted segments (e.g. loops) and smaller distortions in the correctly aligned segments. The alignment errors are also frequent. More than half of the detectably related sequence-structure pairs share less than 30% sequence identity. At 25% sequence identity, ∼20% of residues are already misaligned.[28,29] Here, we used the "information" encoded in the cryoEM density to improve the accuracy of models refined by an iterative alignment, model building, and model assessment protocol (moulding) guided by a fitness function that is dependent on the cryoEM density map as well as a statistical potential (*Moulder-EM*).

Moulder-EM was tested against a benchmark set of 20 target–template pairs of known protein structures that share less than 32% sequence identity (14% on average) (Table 1). This benchmark contained proteins varying in size, fold type, and the accuracy of the initial alignment. For the benchmark targets, the cryoEM density maps were simulated at 10 Å resolution, which is achievable in single particle cryoEM. Relative to the models based on the best existing sequence profile alignment methods, the percentage of C^α

atoms that are within 5 Å of the corresponding C^{α} atoms in the superposed native structure increases on average from 52% to 66% (difficult cases; Table 1 and Figure 2), which is half-way between the starting models and the models from the best possible alignments (82%). Even cryoEM maps at resolution as low as 20 Å are often more useful than the statistical potentials derived from known atomic structures (Figure 6).

Scoring function and sampling

As for any optimization method, the accuracy of *Moulder-EM* is in principle, limited both by the accuracy of the scoring function (i.e. its ability to pick the most accurate structure among the alternatives) and the thoroughness of the optimizer (i.e. its ability to generate an ensemble of models that includes accurate solutions). Next, we discuss both of these potential bottlenecks.

It was previously shown that moulding compares favorably with two of the most successful sequence alignment methods, PSIBLAST[33] and SAM (version 3.3.1).[34] However, due to the limitations in the ranking of models by the original fitness function consisting of a statistical potential,[31] the most accurate models created during the iterative search could not be identified. By adding the density-fitting score to the fitness function (F) and to the composite model assessment score (Z') of the final model selection, *Moulder-EM* improved upon this limitation, resulting in significantly more accurate alignments and models than the original *Moulder* program (Table 1; Figure 2). For the tested proteins with poor initial alignments, *Moulder-EM* was able to reach up to 92% improvement in the native overlap (from 36.1% to 69.4% for 8i1b-4fgf) and 50% in the C^{α} RMSD (from 8.9 to 4.4 Å for 1cauB-1cauA). Furthermore, *Moulder-EM* was able to improve on targets with very accurate initial alignments, approaching and sometimes even surpassing the structure-based alignments (e.g. 1occA-1fftC).

Even when one of the two Z-scores in the fitness function performed significantly better than the other (e.g. 1ten-3hhr), the combination of both Z-scores performed as well as the best individual Z-scores (Figure 3). This fact is consistent with our previous observation that the density fitting and the statistical potential score generally capture different features of the structure.[35] The density fitting score is likely to perform better when the alternative models have different shapes (e.g. different loop conformations and lengths), whereas the statistical potential score is expected to perform well when the best-assessed model is accurate and its assessment does not suffer from the lack of neighboring subunits. Therefore, the utility of the two Z-scores in the fitness function varies depending on the accuracy of the initial alignment and the target structure. While it is difficult to predict in advance which combination of the scores will produce the most accurate models, the benchmarking shows that the sum of the statistical potential and density fitting Z-scores with weights of 1 and 2, respectively, generally works best. This optimal fitness function generally improves the final model relative to the starting model even when the density fitting or statistical potential scores on their own result in lower accuracy (Table 1).

Although the fitness function of *Moulder-EM* is more accurate than that of *Moulder* (Figure 4), it is far from perfect. The final model selected by the composite score is generally still not the most accurate model in the sample of generated models. It is necessary to improve *Moulder-EM* by increasing the accuracy of the statistical potential score and adding new model assessment scores (D. Eramian *et al.*, unpublished results).

The sampling protocol of *Moulder-EM* is identical with that of *Moulder* (Theory). It typically samples approximately 7500 unique alignments over 25 iterations. The possibility that more thorough sampling would result in more accurate models cannot be excluded and is likely for calculations with moderately high-resolution maps (5 Å) starting with inaccurate alignments (below). However, at 10 Å, the present optimization protocol always produced models that scored better than the models from the CE structure-based alignments, highlighting again that *Moulder-EM* is limited primarily by the accuracy of its scoring function, not the completeness of its sampling.

Currently, the *Moulder-EM* protocol is a relatively slow process, taking several hours on a multiprocessor Linux cluster. When compared to the time and cost it takes to produce a subnanometer resolution structure of a macromolecular complex with cryoEM, however, this time is negligible. *Moulder-EM* will be available in the *MODELLER*−9 release.

Resolution of the cryoEM map

To analyze the dependence of the model accuracy on the map resolution, we modeled three of the targets with maps at resolutions from 5 Å to 25 Å (Figure 6). It is difficult to set the upper limit on the resolution of a cryoEM map that can still be helpful in the refinement of a comparative model by *Moulder-EM* based on our limited tests. However, it appears that cryoEM density maps at 25 Å resolution do not provide any significant advantage over the statistical potential score alone, although they do not deteriorate model accuracy either. If the initial alignment is inaccurate, even the 20 Å resolution density map will improve the model because the large alignment errors can be easily detected by the mismatch of the shape densities. Using the higher resolution map (i.e. 5 Å), the density-fitting score could improve the model accuracy beyond the accuracy of the model from the structure-based alignment because the detailed structural features in the density may contain more potent information than the biased template structure.

When the initial models are relatively accurate, the higher-resolution density maps (5–10 Å) result in more accurate models than the lower-resolution maps, because a higher-resolution map contains fine features that allow the optimization to find a more accurate model (scoring-limited regime). In contrast, when the initial models are inaccurate, the lower-resolution density maps (10–20 Å) result in more accurate models than the higher-resolution maps, because it is easier to find the model with the best fit to a lower-resolution map than a higher-resolution map (optimizer-limited regime). This result reflects the different types of errors that can be minimized with the aid of density maps at different resolutions. With higher resolution maps, finer structural detail can be seen and thus *Moulder-EM* reduces smaller alignment errors. Lower resolution maps (10–20 Å), on the other hand, capture more global features of the structure, and thus *Moulder-EM* can reduce only larger alignment errors that result in large rigid-body shifts.

Rice dwarf virus P8 upper domain

To test *Moulder-EM* with an experimentally determined cryoEM map, we modeled the upper domain of the P8 capsid protein from the 6.8 Å resolution structure of RDV (Figure 7). P8 is a fairly typical example of the type of a structure obtained from cryoEM maps of macromolecular assemblies. These assemblies likely contain multiple protein components, some of which are also determined in isolation at atomic resolution or at least have a known homolog of defined structure. However, as in the case of RDV, this homolog is often only remotely related by sequence, making it difficult to calculate an accurate comparative model. The *Moulder-EM* protocol improves upon this limitation by iterating through simultaneous model building and fitting into the cryoEM density map, resulting in increasingly refined alignments and corresponding models.

Conclusion

In summary, we showed that adding the cryoEM density-fitting score to the fitness function of an iterative alignment and comparative model building process generally decreases alignment errors dramatically, even with density maps at ∼15 Å resolution. Moreover, the resulting models are more suitable for constructing the pseudo-atomic models of whole macromolecular assemblies than the experimentally determined structures of the homologs. Given the increasing number of macromolecular machines under investigation using cryoEM and the rising number of comparative models that can be constructed with useful accuracy, the scope for the application of integrated comparative modeling and cryoEM is large.[65] In the future, combined comparative modeling and cryoEM density fitting is likely to benefit from a more accurate fitness function, more complete sampling of alignments, as well as exploring the conformations of loops and relative orientations of smaller segments of structure in addition to the target-template alignment.

Acknowledgements

We thank Drs Bino John, Eswar Narayanan, Frank Alber, Damien Devos, Friedrich Forster and Wen Jiang for very helpful discussions. We also acknowledge support from NSF (EIA-032645), NIH (P41RR02250, P20RR020647, R01 GM54762, 2 PN2 EY016525-02), Human Frontier Science Program, The Sandler Family Supporting Foundation, SUN, IBM, and Intel.

References

1. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. (2003). From words to literature in structural proteomics. *Nature*, **422**, 216–225.
2. Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M. *et al.* (2004). A structural perspective on protein-protein interactions. *Curr. Opin. Struct. Biol.* **14**, 313–324.
3. Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.
4. Saibil, H. R. (2000). Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallog. sect. D*, **56**, 1215–1222.
5. Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 303–319.
6. Henderson, R. (2004). Realizing the potential of electron cryo-microscopy. *Quart. Rev. Biophys.* **37**, 3–13.
7. Chiu, W., Baker, M. L., Jiang, W., Dougherty, M. & Schmid, M. F. (2005). Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure (Camb)*, **13**, 363–372.
8. Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001). Bridging the information gap: computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* **308**, 1033–1044.
9. Fabiola, F. & Chapman, M. S. (2005). Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure (Camb)*, **13**, 389–400.
10. Beckmann, R., Spahn, C. M., Eswar, N., Helmers, J., Penczek, P. A., Sali, A. *et al.* (2001). Architecture of the protein-conducting channel associated with the translating 80S ribosome. *Cell*, **107**, 361–372.
11. Davis, J. A., Takagi, Y., Kornberg, R. D. & Asturias, F. A. (2002). Structure of the yeast RNA polymerase II holoenzyme: mediator conformation and polymerase interaction. *Mol. Cell.* **10**, 409–415.
12. Holmes, K. C., Angert, I., Kull, F. J., Jahn, W. & Schroder, R. R. (2003). Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. *Nature*, **425**, 423–427.

13. Golas, M. M., Sander, B., Will, C. L., Luhrmann, R. & Stark, H. (2003). Molecular architecture of the multiprotein splicing factor SF3b. *Science*, **300**, 980–984.
14. Kostyuchenko, V. A., Leiman, P. G., Chipman, P. R., Kanamaru, S., van Raaij, M. J., Arisaka, F. *et al*. (2003). Three-dimensional structure of bacteriophage T4 baseplate. *Nature Struct. Biol.* **10**, 688–693.
15. Shin, D. S., Pellegrini, L., Daniels, D. S., Yelent, B., Craig, L., Bates, D. *et al*. (2003). Full-length archaeal Rad51 structure and mutants: mechanisms for RAD51 assembly and control by BRCA2. *EMBO J.* **22**, 4566–4576.
16. Ming, D., Kong, Y., Lambert, M. A., Huang, Z. & Ma, J. (2002). How to describe protein motion without amino acid sequence and atomic coordinates. *Proc. Natl Acad. Sci. USA*, **99**, 8620–8625.
17. Tama, F., Wriggers, W. & Brooks, C. L., III (2002). Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *J. Mol. Biol.* **321**, 297–305.
18. Schmid, M. F., Sherman, M. B., Matsudaira, P. & Chiu, W. (2004). Structure of the acrosomal bundle. *Nature*, **431**, 104–107.
19. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
20. Baker, D. & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
21. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
22. Madhusudhan, M. S., Marti-Renom, M. A., Eswar, N., John, B., Pieper, U., Karchin, R. *et al*. (2005). Comparative protein structure modeling. In *The Proteomics Protocols Handbook* (Walker, J. M., ed.), pp. 831–860, Humana Press Inc., Totowa, NJ.
23. Ginalski, K., Grishin, N. V., Godzik, A. & Rychlewski, L. (2005). Practical lessons from protein structure prediction. *Nucl. Acids Res.* **33**, 1874–1891.
24. Wallner, B. & Elofsson, A. (2005). All are not equal: a benchmark of different homology modeling programs. *Protein Sci.* **14**, 1315–1327.
25. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S. *et al*. (2005). The Universal Protein Resource (UniProt). *Nucl. Acids Res.* **33**, D154–D159.
26. Pieper, U., Eswar, N., Davis, F., Braberg, H., Madhusudhan, M. S. P., Rossi, A. *et al*. (2006). MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucl. Acids Res.* **34**, D291–D295.
27. Sali, A. & Kuriyan, J. (1999). Challenges at the frontiers of structural biology. *Trends Cell. Biol.* **9**, M20–M24.
28. Wang, G. & Dunbrack, R. L., Jr (2004). Scoring profile-to-profile sequence alignments. *Protein Sci.* **13**, 1612–1626.
29. Marti-Renom, M. A., Madhusudhan, M. S. & Sali, A. (2004). Alignment of protein sequences by their profiles. *Protein Sci.* **13**, 1071–1087.
30. Sanchez, R. & Sali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins: Struct. Funct. Genet. Suppl*, **1**, 50–58.
31. John, B. & Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucl. Acids Res.* **31**, 3982–3992.
32. Melo, F., Sanchez, R. & Sali, A. (2002). Statistical potentials for fold assessment. *Protein Sci.* **11**, 430–448.
33. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
34. Karplus, K., Barrett, C. & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
35. Topf, M., Baker, M. L., John, B., Chiu, W. & Sali, A. (2005). Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *J. Struct. Biol.* **149**, 191–203.
36. Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
37. Sanchez, R. & Sali, A. (1998). Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
38. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
39. Saqi, M. A., Russell, R. B. & Sternberg, M. J. (1998). Misleading local sequence alignments: implications for comparative protein modelling. *Protein Eng.* **11**, 627–630.
40. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
41. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
42. Sauder, J. M., Arthur, J. W. & Dunbrack, R. L., Jr (2000). Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins: Struct. Funct. Genet.* **40**, 6–22.
43. Jaroszewski, L., Rychlewski, L., Li, Z., Li, W. & Godzik, A. (2005). FFAS03: a server for profile–profile sequence alignments. *Nucl. Acids Res.* **33**, W284–W288.
44. Zhou, H. & Zhou, Y. (2005). Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins: Struct. Funct. Genet.* **58**, 321–328.
45. Edgar, R. C. & Sjolander, K. (2004). A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, **20**, 1301–1308.
46. Ohlson, T., Wallner, B. & Elofsson, A. (2004). Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. *Proteins: Struct. Funct. Genet.* **57**, 188–197.
47. Lindahl, E. & Elofsson, A. (2000). Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625.
48. Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.
49. McGuffin, L. J. & Jones, D. T. (2003). Improvement of the GenTHREADER method for genomic fold recognition. *Bioinformatics*, **19**, 874–881.
50. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
51. Shi, J., Blundell, T. L. & Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243–257.

52. Karchin, R., Cline, M., Mandel-Gutfreund, Y. & Karplus, K. (2003). Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins: Struct. Funct. Genet.* **51**, 504–514.
53. Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M. & Hughey, R. (2003). Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins: Struct. Funct. Genet.* **53**(Suppl. 6), 491–496.
54. Godzik, A. (2003). Fold recognition methods. *Methods Biochem. Anal.* **44**, 525–546.
55. Saqi, M. A. & Sternberg, M. J. (1991). A simple method to generate non-trivial alternate alignments of protein sequences. *J. Mol. Biol.* **219**, 727–732.
56. Saqi, M. A., Bates, P. A. & Sternberg, M. J. (1992). Towards an automatic method of predicting protein structure by homology: an evaluation of suboptimal sequence alignments. *Protein Eng.* **5**, 305–311.
57. Contreras-Moreira, B., Fitzjohn, P. W. & Bates, P. A. (2003). *In silico* protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. *J. Mol. Biol.* **328**, 593–608.
58. Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, **261**, 552–558.
59. Ludtke, S. J., Baldwin, P. R. & Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97.
60. Zhou, Z. H., Baker, M. L., Jiang, W., Dougherty, M., Jakana, J., Dong, G. *et al*. (2001). Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nature Struct. Biol.* **8**, 868–873.
61. Nakagawa, A., Miyazaki, N., Taka, J., Naitow, H., Ogawa, A., Fujimoto, Z. *et al*. (2003). The atomic structure of rice dwarf virus reveals the self-assembly mechanism of component proteins. *Structure (Camb)*, **11**, 1227–1238.
62. Madhusudhan, M. S., Marti-Renom, M. A., Sanchez, R. & Sali, A. (2006). Variable gap penalty for protein sequence-structure alignment. *Protein Eng. Design Select.* In the press.
63. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747.
64. Grimes, J., Basak, A. K., Roy, P. & Stuart, D. (1995). The crystal structure of bluetongue virus VP7. *Nature*, **373**, 167–170.
65. Topf, M. & Sali, A. (2005). Combining electron microscopy and comparative protein structure modeling. *Curr. Opin. Struct. Biol.* **15**, 578–585.
66. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612.
67. Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Struct. Funct. Genet.* **23**, 566–579.

Edited by W. Baumeister

(Received 13 October 2005; received in revised form 6 January 2006; accepted 17 January 2006)
Available online 2 February 2006