CRYO-EM GUIDED DE NOVO PROTEIN FOLDING

By

Steffen Lindert

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

May, 2011

Nashville, Tennessee

Approved:

Professor Al Beth

Professor Phoebe Stewart

Professor Chuck Sanders

Professor Michael Stone

ACKNOWLEDGEMENTS

As I submit this dissertation, it is clear to me that many people have contributed to its creation. I would like to thank all those people who have helped and supported me during the past few years and made this dissertation possible.

My deepest gratitude is to my two advisors, Dr. Jens Meiler and Dr. Phoebe Stewart. Joining their labs when starting my graduate studies at Vanderbilt University has been one of the best decisions that I ever made in my life. They supported me on every step of the way, teaching me how to address scientific problems, how to ask questions whose answers will move my projects along and how to write scientific literature. Jens and Phoebe gave me the freedom to explore the scientific questions that I found most interesting and also to find solutions to problems that I encountered, but were always there when I needed help or guidance. They supported me to attend numerous national meetings, allowing me to gain invaluable practice of presenting my work to my scientific peers. Both Jens and Phoebe fostered lab camaraderie through many lab social events, retreats or lab lunches. As a result of this I established lasting friendships with many of my colleagues.

The members of my dissertation committee, Dr. Al Beth, Dr. Chuck Sanders and Dr. Mike Stone, were a great source of support and guidance for my graduate work. Their insightful comments and constructive criticism kept me and my research on track and allowed me to graduate in such a short time.

None of my projects could have been completed without the help of countless colleagues and friends in the Meiler and Stewart labs. I want to particularly thank Dr. René Staritzbichler, Nathan Alexander, Nils Wötzel and Mert Karakaş for all their help with developing code for the project. Dr. Mariena Silvestry and Dr. Dewight Williams played a pivotal role in teaching me how to use the electron microscope and helped collect countless micrographs with me. I am also thankful to the computing support staff at the CSB and ACCRE, whose work allowed me to concentrate on my actual research and never have to worry about the details of running such a large computer system.

Many friends here in Nashville have made these past 4 years unforgettable. They helped me get through personal and scientific setbacks and were always there to do things outside of work. I made some of my very best friends here and hope that these friendships will last a lifetime even after I leave Vanderbilt.

Finally and most importantly I want to express my deepest gratitude to my family. Without their constant love, support and care I could have never achieved what I have achieved. My parents and grandparents have supported me through every endeavor of mine. And though I know that the large spatial separation puts an incredible strain on my family, they never stopped believing in me. I would like to thank them for everything that they have ever done for me.

TABLE OF CONTENTS

Page

Appendix

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

cryoEM – cryo-electron microscopy

SSE – secondary structure element

BCL – bio-chemical library

ACCRE – Advanced Computing Center for Research & Education

CSB – Center for Structural Biology

SUMMARY

This thesis describes the development of a software package called EM-Fold that combines de-novo protein structure prediction and medium resolution cryoEM density maps. An introduction to the application of computational algorithms to cryoEM data is given in Chapter I. The detailed methodology of the algorithm, its initial results on a benchmark set and an experimental density map are described in Chapter II. Several improvements and additions to the initial implementation of EM-Fold are presented in Chapter III. Here the results of applying the program to a larger benchmark set in combination with using a new version of Rosetta are introduced. Chapter IV summarizes a project that was pursued in the Stewart laboratory. It dealt with the determination of a medium resolution density map of the Adenovirus-Integrin complex. Chapter V is a summary and discussion of the obtained results, tying together the individual angles of research. It also gives an outlook onto future work that might be necessary to improve EM-Fold further. The appendix finally provides detailed background to many aspects that could not be discussed in the chapters. Primarily it serves as an overview over the protocol that was run to obtain the results presented in the chapters. This includes a collection of the scripts and commandlines that were used to generate the protein models presented in this thesis. It also gives a description of how several key classes in the BCL interconnect to perform protein folding into density maps. Additionally, the appendix describes several side projects or ideas that were pursued in the course of the last years that have not been published (yet) however. These include density bump profiles, preliminary results for modeling of DNA-PKcs and results for the cryoEM modeling challenge. The core part of the appendix is contained in this document. More detailed information can be found on a data DVD provided with the thesis.

Chapter I is largely based on the publication S. Lindert, P. L. Stewart, J. Meiler, *Curr Opin Struct Biol* **19**, 218 (Apr, 2009) titled "Hybrid approaches: applying computational methods in cryo-

electron microscopy". The discussion of the work aimed at including density maps into Rosetta was added. Chapter II is based on the publication Lindert *et al.*, *Structure* **17**, 990 (Jul 15, 2009) titled "EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps". Chapter III was written for this thesis and will be the basis of a publication describing the improvements to EM-Fold and its performance in combination with the Rosetta density functionality in terms of achieving atomic resolution models. Chapter IV again is based on the publication S. Lindert, M. Silvestry, T. M. Mullen, G. R. Nemerow, P. L. Stewart, *J Virol* **83**, 11491 (Nov, 2009) titled "Cryo-electron microscopy structure of an adenovirus-integrin complex indicates conformational changes in both penton base and integrin". Chapter V finally was written for the purpose of serving as a discussion chapter for this thesis.

CHAPTER I

INTRODUCTION

This chapter is based on publication (*15*).

Cryo electron microscopy (cryoEM) can provide important structural information about proteins of unknown fold and relative arrangement of proteins of known folds within large macromolecular assemblies such as viruses (*6, 16, 17*). Medium resolution (5-10 Å) cryoEM density maps reveal positions of α-helices (*6*), while near atomic resolution (3.8-4.5 Å) resolution maps can reveal β-sheets and large, aromatic side chains in space (*17*). In addition, near atomic resolution cryoEM density maps can allow tracing of the protein backbone chain and provide restraints for computational atomic-detail refinement techniques (*4, 17*). There is a plethora of computational methods that have been applied to interpret cryoEM density maps and seek to organize newly emerging methodologies with respect to the specific research tasks they address. In the context of cryoEM guided computational protein structure prediction, there are three main computational components: residue-based secondary structure prediction and identification of secondary structure elements, determination of the protein fold, and atomic-detail structure refinement.

Secondary structure prediction algorithms use machine learning techniques like artificial neural networks (ANNs) or hidden Markov models (HMMs) to predict secondary structure, usually as three-state probability (helix, strand, coil) for every residue in the primary sequence of the protein. State of the art techniques like jufo (*18, 19*), psipred (*20*) and sam (*21, 22*) have been demonstrated to achieve accuracies of up to 80%. Determination of secondary structure elements

(SSEs: α-helices and β-strands) from the residue-based predictions is an important aspect in interpreting cryoEM density, however it receives only modest attention in the computational structure prediction field. In order to compensate for prediction inaccuracies from any one method, a consensus prediction protocol in which residue-based predictions are combined and averaged over a sequence window have been developed in the course of this thesis (*5*).

Protein fold or topology prediction algorithms determine the three dimensional arrangement of the amino acids from the sequence of the protein. Two major approaches have to be considered: comparative modeling of a template structure and *de novo* protein structure prediction in the absence of a template. Since 1994, a community-wide blindfold experiment CASP has been carried out bi-yearly to allow these algorithms to be tested on proteins whose structure was already solved but not yet published (*23*). Over the course of the last experiments the program ROSETTA has been identified as one of the most successful *de novo* protein structure prediction algorithms (*24*). However, even if the correct topology is identified, the prediction will still be off by about 5 Å RMSD to the native structure. *De novo* computational structure prediction methods can be applied to soluble proteins smaller than about 180 amino acids in size with success rates of about 50% (*24, 25*). Successful comparative modeling programs include ROSETTA (*26, 27*) and MODELLER (*28-30*) among others and achieve models of 2-5 Å RMSD depending on the similarity between template and target structure.

Atomic detail refinement techniques are used to add side-chain coordinates (e.g. using SQWRL (*31*)) and improve the medium resolution structures produced by comparative modeling or *de novo* prediction to less than 2 Å RMSD to native in favorable cases. These algorithms use higher-resolution energy functions and finer grained sampling techniques. For these algorithms to be successful, the starting structure has to be sufficiently close to the native conformation. Typically side-chain and backbone degrees of freedom need to be optimized in an iterative cycle of rapid side-chain repacking, larger scale backbone perturbations, and gradient minimization (*32*). It has been demonstrated that in a few favorable cases computational algorithms can refine *de novo* models of small proteins (size 40-90 amino acids) to atomic detail (< 2 Å RMSD) (*33*). Atomic-detail comparative models can be built for much larger proteins if a suitable template structure exists.

Figure 1 gives an overview of the most notable hybrid approaches between these computational protein structure prediction methods and cryoEM (*15*).

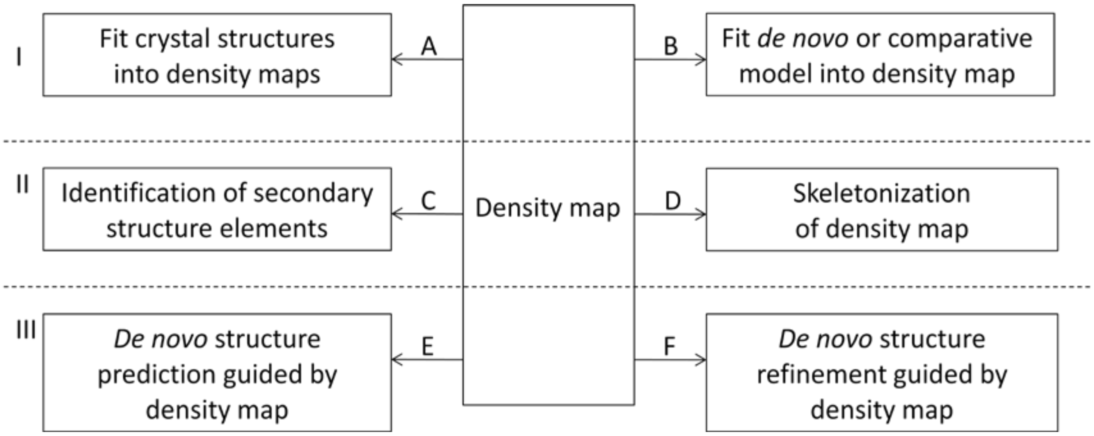
I
Fit crystal structures into density maps
A
B
Fit *de novo* or comparative model into density map
II
Identification of secondary structure elements
C
Density map
D
Skeletonization of density map
III
De novo structure prediction guided by density map
E
F
De novo structure refinement guided by density map

Figure 1. Overview of current hybrid approaches between cryoEM and computational protein structure prediction algorithms.

Computational algorithms that work on cryoEM density maps can be divided into three main classes: I) The first class of algorithms fits structures into cryoEM density maps. These may be A) high resolution experimental structures (X-ray crystallography, NMR) or B) computationally built models (*de novo* models from e.g. ROSETTA or comparative models created with e.g. MODELLER). II) Algorithms that analyze the density map itself. C) SSEHUNTER and HELIXTRACER both attempt to identify regions of the cryoEM density map that correspond to secondary structure elements. D) The skeletonization algorithm described in (*8*) builds skeletons of the density map and can be used to trace the backbone of high resolution density maps. III) The third class of software uses cryoEM density maps as experimental restraints in *de novo* protein structure E) prediction (EM-Fold) and F) refinement (ICM). Reproduced with permission from Elsevier.

Fitting of crystal structures and computational models into cryoEM density maps

In the presence of a crystal structure or a complete model of the target protein a direct fit into cryoEM density maps is possible. The most frequently used fitting methods employ a six dimensional search (three translational and three rotational degrees of freedom) of the rigid-body model in the density map (*34*). Use of a Fast Fourier Transformation (FFT) accelerated translational search is implemented in state-of-the-art algorithms such as COLORES (*35*). Recently

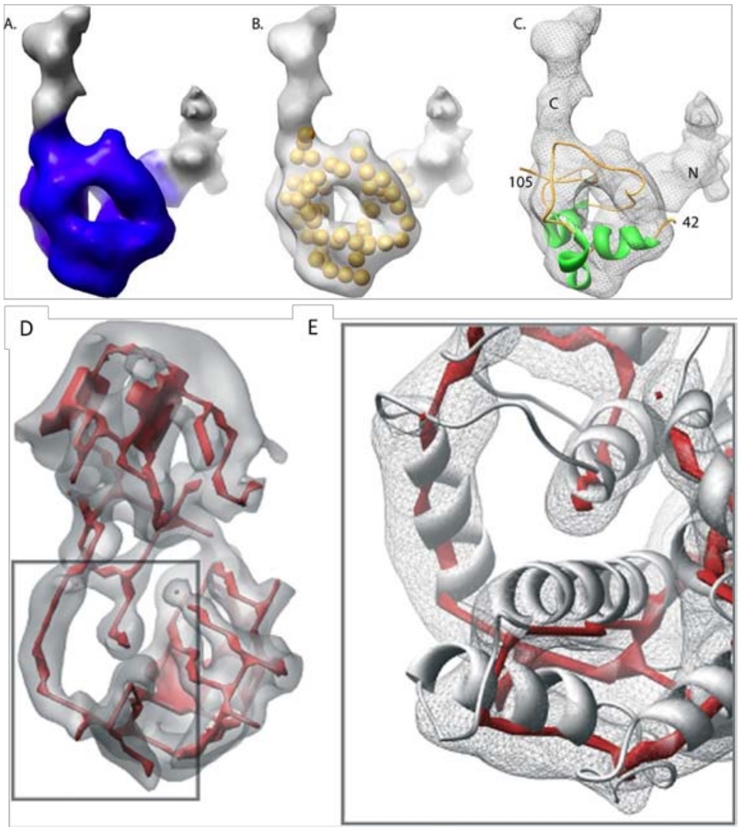

Figure 2. Two different approaches of using cryoEM density maps in conjunction with computational algorithms.
Panels A through C show how an computational model of VP26 is superimposed with the segmented experimental density (*9*). Here the cryoEM density map is used as a filter for *de novo* protein models. D) Example from (*13*) of segmented cryoEM density (gray) and the skeleton that SSEhunter built for the density (red). E) In certain cases the skeleton can approximate the backbone trace of a protein.

(Wötzel et al., unpublished) reported a fitting method based on a geometric hashing algorithm that proved to be faster than traditional fitting methods. In addition even with shorter computational times, the hashing procedure identified more symmetry related positions and independent repeating units within an experimental cryoEM density map. Agreement of a positioned model and density map is determined by a cross correlation coefficient (*36*). Frequently flexible fitting algorithms are used to fit and adjust high resolution structures to optimally fit into EM density maps. Programs such as S-FLEXFIT (*37*, *38*) are suited for medium resolution density maps, while algorithms using normal mode analysis (NMA) (*39*) are tailored towards low resolution density maps. Even though the structures themselves are perturbed during the fitting, the main focus of these algorithms is the optimal fit with the density map.

An example of using fold recognition in cryoEM is SPI-EM (*40*) that identifies the superfamily that a protein belongs to from its density map using CATH. Topf et al. demonstrated that comparative models may be ranked in terms of their accuracy by fitting them into a cryoEM density map (*41*). The authors then went on to develop an iterative protocol that improves comparative models by optimizing their agreement with the cryoEM density maps (*42*). These methods require the presence of a comparative model but have the advantage that no SSEs or even backbone trace have to be identified from the density map.

The authors in (*9*) compared ROSETTA *de novo* predicted structures with the 8.5 Å resolution cryoEM density map of the herpes simplex type 1 capsid. They were able to rank the agreement of the model with the map by using a two-way distance measure. The model for the virus structural protein VP26 that agreed best with the density exhibited a new fold (see Figure 2, panels A, B and C). This approach eliminates the need for a comparative model, which in many cases is not readily available. One drawback is however that the density map is only used as a filter of *de novo* models and the density does not guide the folding step. Therefore this method relies on

Rosetta to fold the protein correctly *de novo* which works in favourable cases for proteins with up to 180 amino acids (*43*).

Computational algorithms to identify secondary structure elements in medium resolution density maps

Density maps begin to reveal α-helices at about 10 Å resolution, β-sheets at about 5-7 Å resolution, and large side chains at about 3.0-4.5 Å resolution (*17*, *44*). Several programs provide an alternative to manual identification of these SSEs. The HelixHunter program was initially developed in 2001 and uses segmentation and feature extraction to identify α-helix positions, orientations and lengths (*45*). HelixHunter has been successfully applied to identify nine α-helices in the 6.8 Å resolution density map of rice dwarf virus outer capsid shell protein P8 (*46, 47*). In order to identify α-helices, EMatch (*48*, *49*) employs a method very similar to HelixHunter exploiting the fact that α-helices generally are observed as continuous, long, thin and highly dense cylindrical regions. A third available algorithm that focuses on the reliable identification of α-helices in medium resolution density maps is called HelixTracer and utilizes gradient analysis to recognize and classify volumes in density maps (*50*). Dal Palu et al. noted significant improvements in recognition and precision over the HelixHunter software.

Tools such as SheetMiner (*51*) and SheetTracer (*52*) have been developed for detecting β-sheets in density maps. The desire to have a single tool capable of identifying both α-helical and β-sheet regions led to the development of the program SSEHunter (*13*). This algorithm uses density skeletonization (see Figure 2, panels D and E), local geometry calculations and a template-based search to identify SSEs in medium resolution density maps.

Skeletonization algorithms help to trace the backbone in higher resolution cryoEM density maps

Density maps at medium resolution (5-10 Å) do not contain sufficient information to unambiguously trace the backbone of the protein from the map. On the other hand, high resolution density maps (<3 Å) can contain enough detail to trace the backbone, as is routinely done in X-ray crystallography. In the intermediate resolution range (4-7 Å) a density map may contain valuable information in the loop regions that can guide model building. The connections between identified SSEs may be clear in some areas of the density map and not evident in other areas. The skeletonization algorithm in SSEHUNTER (*13*) computes skeletons of volumetric data

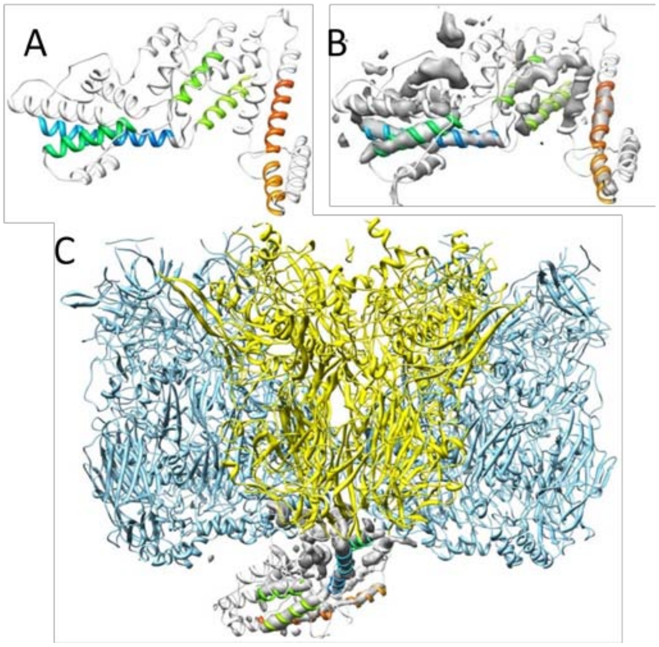

Figure 3. Computational *de novo* protein structure prediction with the cryoEM density map as a folding restraint.
EM-Fold was used to build computational models into an experimental cryoEM density map of human adenovirus protein IIIa at ~6 Å resolution (*5*). A) A reduced model of protein IIIa where only helices that have been placed with at least 60% confidence are colored in rainbow. B) Same as in A, but shown in density. C) Side view of reduced model of protein IIIa (rainbow) in contact with penton base (yellow) and two peripentonal hexons (light blue). Reproduced with permission from Elsevier.

by alternation between a thinning and a skeleton pruning routine (*8*). The authors in (*53*) used a combination of SSEHUNTER and this skeletonization algorithm to trace the backbone of a ~4 Å resolution density map of GroEL.

Using α-helix positions in the density map to build models of proteins promises best results

Membrane proteins are frequently only resolved to medium resolution. Based on the surface charge and evolutional variability of their lipid-exposed faces, Fleishman et al. developed an algorithm (*54*) that can correctly orient transmembrane helices within the density rod. Already in 1979, Cohen et al. investigated the relative placement of helices in general (*55*). In their work, which was not guided by experimental density, they were able to derive simple rules for the assembly of helices into tertiary structure using myoglobin as an example case.

A central part of this thesis was the development of an approach called EM-Fold that uses a Monte Carlo sampling strategy to build and refine protein topologies into intermediate resolution cryoEM density maps (of soluble proteins) where α-helices are resolved as density rods (*5*). The first step is to identify density rods that are likely to be α-helices in a medium resolution density map (5-10 Å). Then a pool of predicted α-helices is used as input to a *de novo* folding algorithm. A novel feature of EM-Fold is that α-helices are only placed in positions where density rods were identified, thus constraining and guiding the *de novo* model building process. The density map is used as a restraint during the initial assembly stage and not just as a post-sampling filter as in other approaches. This decreases the conformational space that has to be sampled considerably and ensures that the final models agree with the density map. Missing loop regions are added and final models are refined using ROSETTA. In a benchmark with 10 proteins of size 250 to 350 amino acids the algorithm identified the correct topology in 70% of the cases and showed that the limiting factor was incorrect secondary structure prediction. A partial model of human adenovirus protein IIIa was built by assembling predicted helices into the experimental density rods (see Figure 3).

High resolution refinement guided by EM density maps

High resolution refinement techniques can also be guided by EM density maps. This was impressively shown in (*4*), where the prediction of atomic-detail structures of helical proteins was aided by simulated EM maps (see Figure 4). The authors assume that a medium resolution density map of a transmembrane helical bundle is available and that helical segments are known (see Figure 4). By a three stage process that included 1) flexible fitting of helices into density rods, 2) optimization of side chains and 3) further refinement of lowest-scoring conformations, Kovacs et al were able to achieve RMSD values between 0.9 and 1.9 Å for their test cases GpA, KcsA and McsL. Even more recently a density map functionality was added to ROSETTA (*56*). This allows using virtually all ROSETTA protocols with a density map as restraint. The authors found that the method can achieve close to an atomic resolution model based on density maps at 4-6 Å resolution. As part of this thesis EM-Fold was improved to yield lower RMSD models after the refinement step. A combined approach using the improved EM-Fold and ROSETTA was used to refine several proteins to atomic resolution starting from medium resolution density maps at 7 Å resolution.

Other notable examples for high resolution refinement in EM density maps are real space refinement algorithms such as RSREF (*57*), FLEX-EM (*58*) and MDFF (*59*). These methods use molecular dynamics to refine the models and use the density map to guide this refinement.

Structure of integrin bound to Adenovirus

A somewhat separate project that was worked on during the course of the thesis dealt with the determination of an experimental cryoEM structure for an Adenovirus-integrin complex. More specifically, a structure of adenovirus type 12 (HAdV12) complexed with a soluble form of integrin αvβ5 was determined by cryoelectron microscopy (cryoEM) image reconstruction. Subnanometer resolution (8 Å) was achieved for the icosahedral capsid with moderate resolution (27Å) for integrin density above each penton base. Modeling with αvβ3 and $\alpha_{IIb}\beta3$ crystal

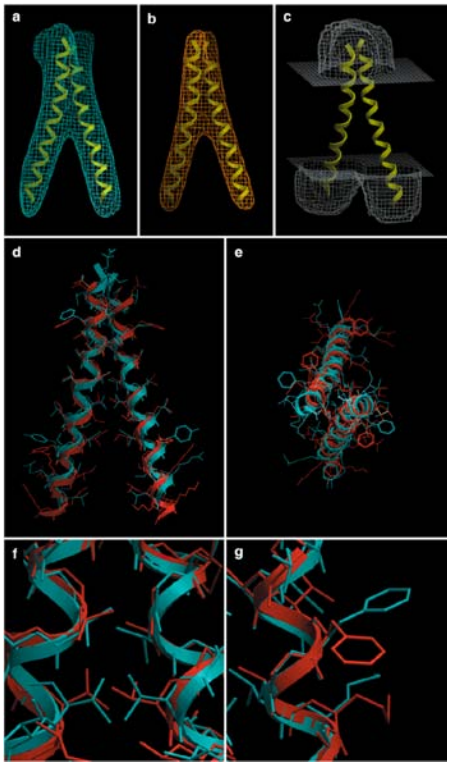

Figure 4. High resolution *de novo* protein structure refinement guided by EM density maps. Benchmark of the high resolution refinement protocol in ICM (*4*). A) A density map was simulated from the NMR structure of Glycophorin A (GpA). B) A tethering map was derived from the EM map and serves to restrain the α-helices. C) A solvent-accessibility map is also calculated. D) and E) Side and top views of the NMR structure (blue) and predicted structure (red). F) Closeup view showing the good agreement of the predicted and experimental structures in a helix packing region. G) Closeup view of a region that faces the lipid where packing constraints are not present and some of the side chain conformations are not recovered. Reproduced with permission from Elsevier.

structures indicates that a maximum of four integrins fit over the pentameric penton base. The close spacing (~60Å) of the RGD protrusions on penton base precludes integrin binding in the same orientation to neighboring RGD sites. Flexible penton base RGD-loops and incoherent averaging of bound integrin molecules explain the moderate resolution observed for the integrin density. A model with four integrins bound to penton base suggests that integrin might extend one RGD-loop in the direction that could induce a conformational change in the penton base involving clockwise untwisting of the pentamer. A global conformational change in penton base could be one step on the way to the release of Ad vertex proteins during cell entry. Comparison of the cryoEM structure with bent and extended models for the integrin ectodomain reveals that integrin adopts an extended conformation when bound to the Ad penton base, a multivalent viral ligand. These findings shed further light on the structural basis of integrin binding to biologically relevant ligands, as well as the molecular events leading to HAdV cell entry.

CHAPTER II

FOLDING PROTEINS INTO MEDIUM RESOLUTION DENSITY MAPS

This chapter is based on publication (*5*).

Introduction

Since the first subnanometer (<10 Å) resolution cryoEM single particle reconstructions, determined for the hepatitis B virus capsid in 1997 (*60, 61*), there have been an increasing number of structures determined by cryoEM in the 6-10 Å resolution range (*6, 62-67*). For example Saban et al. determined a 6.9 Å resolution structure of adenovirus, Booth et al. reached 9 Å resolution for cytoplasmic polyhedrosis virus and Zhang et al. elucidated a 7.6 Å resolution structure of Reovirus. As only a fraction of the viral proteins are amenable to structure elucidation by X-ray crystallography, these experiments yield images of viral proteins of previously unknown structure. CryoEM can also elucidate the structures of large macromolecular complexes such as blue copper protein hemocyanin (Martin et al., 10 Å resolution), elongation factor Tu – ribosome complex (Villa et al., 6.7 Å resolution) and Tetraspanin uroplakins (Min et al., 6 Å resolution). In these cases the density map revealed previously unknown crucial interfaces between subunits of the macromolecular complex. CryoEM has also been used to elucidate subnanometer structures of membrane proteins such as the skeletal muscle Ca^{2+} release channel (Serysheva et al., 9.6 Å resolution). Several near-atomic resolution structures (<5 Å resolution) have been determined recently using cryoEM (*53, 68-70*). While near-atomic resolution maps show details such as β-sheets and large side chains (*17*), these features cannot be identified reliably at intermediate resolution. However α-helices are resolved as density rods at intermediate resolution (*71*).

One of the biggest challenges for the interpretation of medium resolution density regions remains the building of a correct topological model. It is impossible to "thread" the primary sequence through the density map for regions that are assigned to a protein of unknown structure because the connectivity between the density rods cannot be discerned at intermediate resolution. Thus it is not possible to assign particular density rods to specific α-helical regions of the sequence. Even if this obstacle could be overcome, missing loop regions and side chain coordinates need to be built to arrive at an accurate atomic model.

Several computational tools are available that help in the analysis of cryoEM density maps. If a high-resolution structure for the map or parts of the map is available fitting techniques are frequently employed (*34, 37-39, 58, 59, 72-75*). If no high-resolution structures are available for fitting, medium resolution density maps can be interpreted in terms of the α-helices that can be seen in the map. α-helical regions can be identified either manually as rods within the density map, or automatically by methods using segmentation and feature extraction (*45, 50*). The skeletonization algorithm in (*13*) identifies secondary structure elements and suggests a possible secondary structure topology by connecting density rods based on increased density in short loop connections. A protocol that iteratively improves comparative models by fitting these models into cryoEM density maps is reported (*42*). This method requires the presence of a comparative model but is independent of the identification of α-helical regions in the density map. Models built with the *de novo* protein structure prediction software ROSETTA were ranked with respect to their agreement with the cryoEM density maps using a 2-way distance measure (*9*). This approach eliminates the need for an initial comparative model, however it has the drawback that the ROSETTA calculation is not driven by the experimental density map. Therefore the approach only works if ROSETTA is capable of folding the protein correctly *de novo*, which is possible for proteins with up to 150 amino acids (*43*).

De novo protein structure prediction algorithms have experienced considerable improvements during the last ten years. The software ROSETTA has been demonstrated to correctly predict the fold of proteins with up to 150 amino acids (*24, 43, 76-78*). Structurally variable loop regions up to 12 residues long can be modeled routinely with ROSETTA (*79*). More recently iterative side-chain repacking and backbone reconstruction protocols within ROSETTA have been shown to refine initial *de novo* and comparative models to atomic-detail accuracy (*27, 33, 80, 81*). For instance, with a benchmark of 16 small proteins (49-88 residues) Bradley et al. demonstrated that accurate atomic-detail models (<1.5 Å) could be reached from initial *de novo* models for five proteins.

It has been demonstrated that guiding the *de novo* protein structure prediction technique ROSETTA with low resolution or sparse experimental data yields structural models with accurate atomic-detail. Inclusion of NMR data within ROSETTANMR has improved the quality of created atomic models (*32, 82-85*). Similarly EPR data has been combined with Rosetta for enhanced model building (*86, 87*).

The approach presented in this chapter combines computational structure prediction methods with experimental cryoEM density maps to build topological models for large proteins without an atomic resolution structure or an available comparative model. The algorithm first identifies α-helical regions in the density map and in the protein's primary sequence, utilizing a consensus secondary structure prediction protocol. The predicted α-helices are placed into specific α-helical density rods of the density map using a novel Monte Carlo assembly algorithm. Then loop regions and side chain coordinates are added using ROSETTA's iterative side-chain repacking and backbone reconstruction protocols to arrive at a model with atomic detail present.

The initial design of EM-Fold is tailored towards α-helical proteins as β-strands are typically not well resolved in medium resolution density maps. β-strands become visible at 5-7 Å resolution

(*71*). Chapter III will deal with a second development stage of EM-Fold that simultaneously assembles α-helices and β-strands.

Here we present the results of EM-Fold with ten mainly α-helical benchmark proteins and simulated cryoEM density, as well as with experimental cryoEM density maps of bovine metarhodopsin and adenovirus protein IIIa. In the case of metarhodopsin, the EM-Fold models are compared with the atomic resolution structure of rhodopsin.

Table 1. Overview over the ten proteins used in the assembly benchmark.

PDB-ID	Description	residues	α-helices[a]	helical residues[b]	resolution [Å]	contact order[c]
1IE9	Vitamin D3 receptor	259	7	145	1.4	46
1N83	Nuclear receptor ROR-alpha	270	7	161	1.6	44
1OUV	Conserved hypothetical secreted protein	273	14	207	2.0	15
1QKM	Estrogen receptor beta	255	8	168	1.8	43
1TBF	cGMP-specific 3',5'-cyclic phosphodiesterase	347	13	220	1.3	40
1V9M	V-type ATP synthase subunit C	323	8	158	1.9	46
1XQO	8-oxoguanine DNA glycosylase	256	10	151	1.0	43
1Z1L	cGMP-dependent 3',5'-cyclic phosphodiesterase	345	12	201	1.7	41
2AX6	Androgen receptor	256	6	145	1.5	40
2CWC	ADP-ribosylglycohydrolase	303	9	152	1.7	59
1GZM	Bovine Rhodopsin	349	8	197	2.7	82

[a] All α-helices with at least 12 residues are shown
[b] The total helical content per protein varies from 60 to 68%
[c] Contact order is defined as the average sequence separation of all residues that are in contact within the protein. The higher the contact order, the more complex the fold is. Values above 40 are considered to reflect very complex folds.

Results & Discussion

Benchmark database of ten α-helical proteins with 250 to 350 residues

To test the reliability as well as to optimize the parameters of the proposed assembly algorithm EM-Fold, it has been benchmarked on ten proteins of known structure following the protocol outlined in Figure 5. The proteins were chosen to be mostly α-helical (60-68%) and of substantial size (255 to 347 residues, Table 1). Except for one protein (1OUV) all the benchmark cases possess contact orders of 40 or higher. Thus these proteins constitute complex folds, making *de novo* computational structure prediction challenging (*88*). In order to mimic cryoEM density maps, simulated density maps at 6.9 and 9.0 Å resolution were generated for each of the ten proteins. The positions and lengths of the density rods are virtually indistinguishable at both resolutions. The maps however differ by the information they contain in loop regions as well as in delineation of the density rods. The benchmark was performed in two stages depending on the type of secondary structure information used, either the correct secondary structure derived from the atomic resolution structure or a realistic prediction of secondary structure, which can deviate from the true structure.

100% success rate for the perfect secondary structure prediction benchmark

In a first test 20,000 models were built for each of the ten benchmark proteins using the correct secondary structure. The Monte Carlo simulation was run until a total of 2,000 subsequent steps were rejected with no improvement in the overall score. The agreement with the density, which is simulated for the benchmark proteins, is assessed by a combined occupancy score, a loop score, and a connectivity score. A predicted fold is considered correct if all α-helices have been placed in the appropriate simulated density rods with the correct orientation of the α-helical axis. A high rank for the correct fold among the 20,000 models generated indicates success of the protocol.

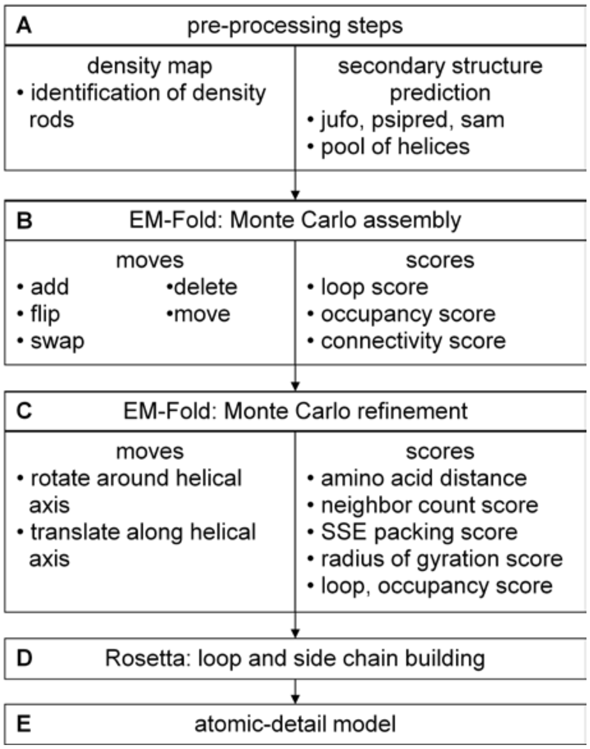

Figure 5. Flowchart of the folding protocol.
A) Density rods are identified in a medium resolution density map. A pool of α-helices is built using secondary structure prediction algorithms. B) The assembly step of EM-Fold places α-helices from the pool into density rods. C) An EM-Fold refinement step improves the placement of α-helices within the density rods. D) Loops and side chains are built in Rosetta for the best of the refined EM-Fold models. E) One of the final full atom models is likely to be very close in RMSD to the native structure. Reproduced with permission from Elsevier.

Table 2. Overview of the benchmark with ten α-helical proteins.

protein	rank assembly [a]	rmsd assembly [Å] [b]	rank refinement [c]	rmsd refinement [Å] [d]	rank loop [e]	rmsd loop [Å] [f]	Helices in final partial model [g]
1IE9	1 (1)	3.7 (3.3)	5 (1)	3.7 (2.6)	1 (1)	5.9 (7.8)	4 [4]
1N83	1 (1)	6.2 (3.2)	2 (1)	5.9 (2.4)	1 (7)	7.1 (3.7)	5 [5]
1OUV	6 (10)	3.0 (3.1)	4 (6)	2.9 (2.3)	1 (1)	4.3 (4.8)	9 [9]
1QKM	16 (1)	3.6 (3.1)	2 (1)	2.7 (3.3)	2 (7)	3.9 (4.2)	5 [5]
1TBF	100 (8)	3.1 (3.2)	20 (17)	2.8 (2.7)	1 (3)	4.1 (4.2)	12 [11] [h]
1V9M	- (1)	- (3.3)	- (1)	- (2.0)	- (2)	- (6.7)	7 [4]
1XQO	- (2)	- (3.3)	- (7)	- (2.1)	- (1)	- (5.0)	6 [2]
1Z1L	150 (3)	3.1 (3.4)	72 (13)	3.2 (2.5)	1 (1)	5.9 (5.5)	9 [9]
2AX6	1 (1)	4.0 (3.4)	5 (1)	3.2 (3.4)	3 (8)	6.6 (9.2)	5 [5]
2CWC	- (2)	- (2.9)	- (8)	- (2.4)	- (2)	- (7.1)	3 [0]
rhodopsin	2	3.4	1	3.1	1	7.9	-

Results are shown for both realistic secondary structure prediction, as well as for perfect secondary structure prediction in parentheses.

[a] rank of true model after assembly step

[b] rmsd of backbone atoms in helices of true model after assembly step (compared to PDB coordinates)

[c] rank of true model after refinement step

[d] rmsd of backbone atoms in helices of true model after refinement step

[e] rank of true model after loop building step

[f] rmsd of all atoms in true model after loop building step

[g] Number of helices in final partial model based on 50% consensus placement; the number of correctly placed helices in these partial models is shown in square brackets.

[h] The one helix in the partial model of 1TBF that has not been correctly placed has been placed into the correct density rod, however with antiparallel orientation

The true model is found among the best 10 scoring models for all the benchmark cases (Table 2). In 50% of the cases the true model is ranked first. In the cases where the true model is not ranked first, the better ranking models are similar in topology to the true model and frequently only have a single α-helix or a pair of α-helices in an incorrect orientation. This demonstrates that the assembly step can clearly distinguish native-like from non-native models if the correct secondary structure is used as input. The RMSDs of the correct topology models range between 2.9 Å and 3.4 Å over the α-helical residues (Table 2).

For each of the ten proteins, the 50 best scoring models from the assembly step were refined. In this process a wider variety of types of scores (described in Materials and Methods) is used to evaluate the models. After refinement the RMSDs of the best scoring correct topology model range between 2.0 Å and 3.4 Å, again considering only the α-helical residues, and the true model is found among the best 17 scoring models (Table 2). These rankings are within the accuracy limit of the scoring functions. ROSETTA was used to build loops for the 20 best scoring models after the refinement run. The RMSD of the true model after loop building ranges between 3.7 and 9.2 Å (Table 2), which is an excellent level of agreement for *de novo* models considering the large size of the proteins. After the loop building step, all of the true models are ranked within the best 8 scoring topologies according to the ROSETTA score. Thus, EM-Fold is able to identify the true topology within the top ten best scoring models built, given completely correct secondary structure information.

EM-Fold selects the best α-helices from a consensus pool generated from state-of-the art secondary structure predictions

A combination of three state-of-the art secondary structure prediction programs jufo (*18, 19*), psipred (*20*) and sam (*21*, *22*) was used to simulate a realistic prediction scenario. The utilization of different programs avoids usage of incorrect secondary structure if one of the methods fails. Wherever an α-helix is predicted with a probability of higher than 0.5 for more than nine subsequent residues, this α-helix is inserted into the pool of considered secondary structure elements. Smaller α-helices are ignored as these cannot be confidently identified in intermediate resolution density maps. Further, a consensus prediction (average of all three methods) and a consensus prediction where α-helices longer than 21 residues are broken into two smaller α-helices are included. Within the ten benchmark proteins there are 93 α-helices that have at least 12 residues. Each of these α-helices is identified by at least one secondary structure prediction technique, although the predicted lengths and confidence levels differ.

Table 3. Evaluation of different secondary structure prediction pools.

	Pool A [a]		Pool B [b]		Pool C [c]	
protein	deviation [d] (residues)	number helices	deviation [d] (residues)	number helices	deviation [d] (residues)	number helices
1IE9	1.3	35	0.7	36	0.4	143
1N83	1.1	38	0.6	40	0.1	156
1OUV	0.9	67	0.4	67	0.1	268
1QKM	1.1	39	0.9	39	0.3	156
1TBF	1.7	52	0.3	52	0.1	206
1V9M	1.6	39	1.6	41	0.9	160
1XQO	2.1	33	1.1	34	0.9	135
1Z1L	1.6	53	0.8	53	0.4	212
2AX6	1.8	35	1.2	35	0.8	140
2CWC	1.8	56	1.1	57	0.7	225
average	1.5	45	0.8	45	0.4	180
prediction	number helices		number helices		number helices	
shorter	255		130		586	
longer	56		153		308	
equal	43		71		168	

[a] Pool A contains predictions from jufo, psipred, sam; a consensus of the three; and a consensus where long helices are broken into smaller pieces.
[b] Pool B replaces all of the helices in pool A with copies that have one residue added to both the N-terminal and C-terminal ends of each helix.
[c] Pool C contains all of the helices in pool A and additional longer copies with one residue added to the N-terminal end, one residue added to the C-terminal end, and one residue added to both the N-terminal and C-terminal ends of each helix.
[d] Average deviation (in residues) of helix lengths between the PDB structure and the best predicted helix in the pool

Secondary structure predictions tend to yield α-helices that are too short, thus three different pools (A, B and C) of secondary structure elements were tested including lengthened α-helices in pools B and C (see Materials and Methods). The best results for the assembly step are obtained with the most diverse pool of secondary structure elements (pool C), where the average deviation between predicted and correct α-helix length is only 0.4 residues per α-helix (Table 3). This finding stresses two points: 1) The more accurate the secondary structure prediction is, the better the results of the assembly algorithm will be - a finding that is also supported by the benchmark test using the correct secondary structure information. 2) A larger pool, which includes many inaccurate secondary structure elements, does not negatively influence the success of the assembly protocol. In other words, the assembly protocol identifies and uses the best possible secondary structure elements available in the pool. Only pool C was used for the realistic secondary structure benchmark as it has been demonstrated to most accurately represent the secondary structure of the proteins.

De novo folding of α-helical benchmark proteins with realistic secondary structure predictions

In the initial assembly step (see Figure 5B) 60,000 models were built for each protein using the most diverse secondary structure pool (pool C). Building one model takes approximately 60s on a single JS20 IBM 2.2GHz PowerPC. The models were ranked by score (Table 2). Our results indicate that despite the inaccuracies of secondary structure prediction, after the assembly step the true model is found among the best 150 scoring models for seven out of the ten proteins. In particular for four of the benchmark proteins the true model is found among the best ten scoring models, and the average rank of the seven correct models is 39. The RMSD of the correct model after the assembly step ranges from 3.0 to 6.2 Å (Table 2). The best 150 models by score enter the refinement protocol without manual analysis.

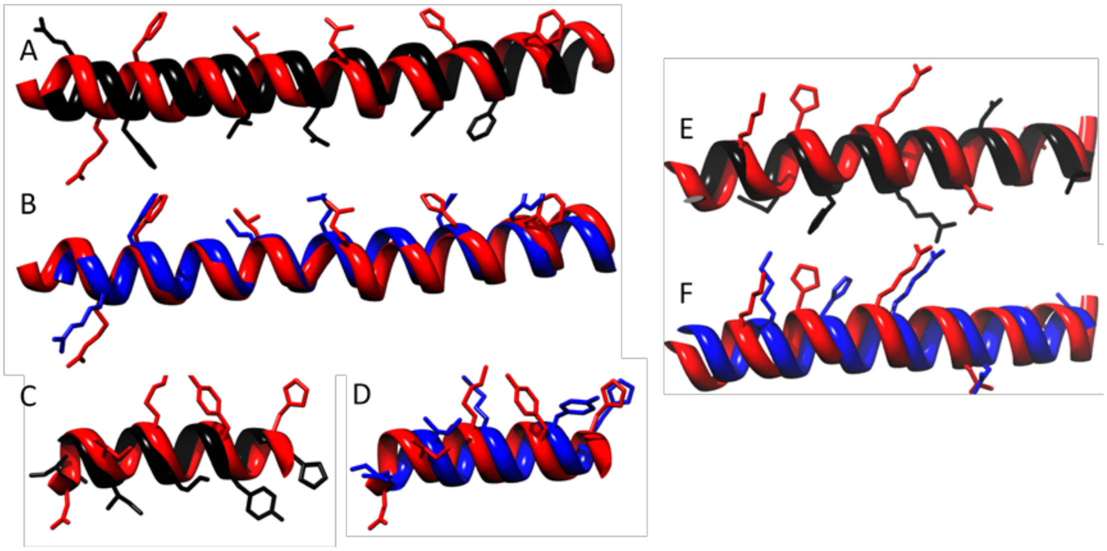

Figure 6. Three examples of improved α-helical orientations after the EM-Fold refinement protocol. Black: model before refinement step. Red: native α-helix from PDB. Blue: model after refinement step. A, C, E) The α-helix in the model before refinement is turned by approximately 180° around the α-helical axis with respect to the native structure. B, D, F) The refinement was able to correctly turn the α-helix with respect to the native structure. Generally, only a slight shift along the α-helical axis remains between the model after refinement and the native α-helix. Reproduced with permission from Elsevier.

After refinement (see Figure 5C) the ranking of the correct model improves to at least rank 72, for five of the benchmark cases it even improves to rank 5 or better. Further, the quality of the true model, as assessed by the RMSD, improves for five out of seven cases with a range over all seven proteins of 2.7 to 5.9 Å (Table 2). Figure 6 illustrates the improvement of α-helix orientations during the refinement step for three examples. The best 75 models by score enter the loop building protocol without manual analysis.

Loops are built for the best 75 scoring models after refinement. For each of the 75 refined models 100 loop models are built using ROSETTA. After ranking of these 7,500 models according to their ROSETTA score, the true model is within the best three scoring models for all seven proteins (see Table 2). Even though the average rank of the correct model after the assembly step was only 39, the user only needs to consider the top three scoring models after loop building. The accuracy of these models is in the range of 3.9 to 7.1 Å (Table 2). This RMSD range is comparable to those built with correct secondary structure elements and acceptable considering the large size of the proteins. Superimpositions of the final ROSETTA model with the native structure are shown for all seven proteins (Figure 7).

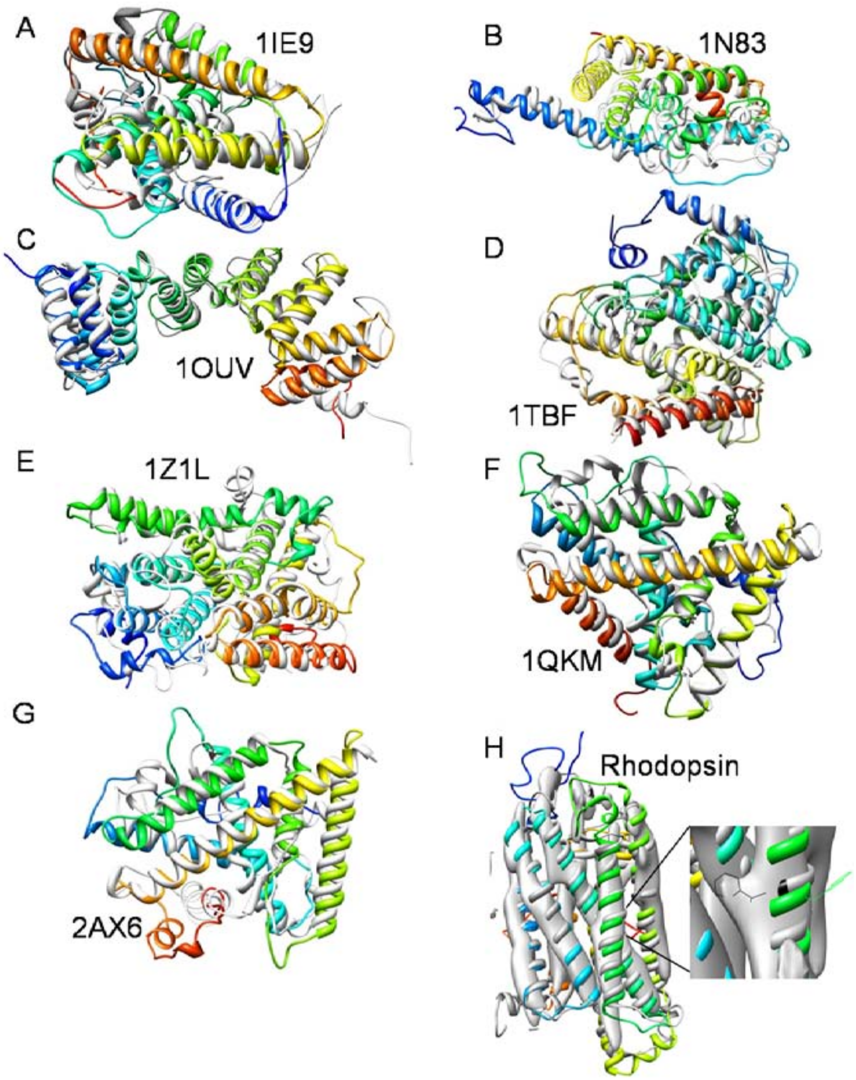

Figure 7. Superimposition of the final models with native structures.

Superimposition of the final models (colored in rainbow) of 1IE9 (A), 1N83 (B), 1OUV (C), 1TBF (D), 1Z1L (E), 1QKM (F) and 2AX6 (G) with the original PDB structures (grey). These proteins range in size from 255 to 345 residues. The displayed models have RMSDs ranging from 3.9 Å to 7.1 Å compared to the PDB structure. Regions that are only seen in the models (such as the N-terminus of 1TBF) correspond to parts of the protein that are missing in the PDB file. Panel H shows the model of rhodopsin after the loop building step (rainbow) in the experimental density model. The crystal structure of rhodopsin is shown in grey for comparison. The model and crystal structure have an RMSD of 7.9 Å. A blow-up of one Trp side chain and its corresponding density bump is shown. The Trp side chain of the crystal structure is shown in black for comparison. It is apparent that the Trp in the model was placed in the correct height of the density rod. The rotation of the α-helix in the model is off by about 150° however. This is not unexpected and could be corrected by a subsequent refinement protocol. Reproduced with permission from Elsevier.

Consensus placement of α-helices correlates with correct positioning and can be used as a measure of confidence

In order to develop a measure that is independent of the score and that can evaluate the correctness of a particular model, the consensus placement of α-helices into specific density rods was analyzed. Models after the assembly step and after loop construction were evaluated. In both cases the benchmarks indicate that if a specific α-helix is found repeatedly in the same density rod within the set of best scoring models it was placed correctly. Receiver Operator Characteristic

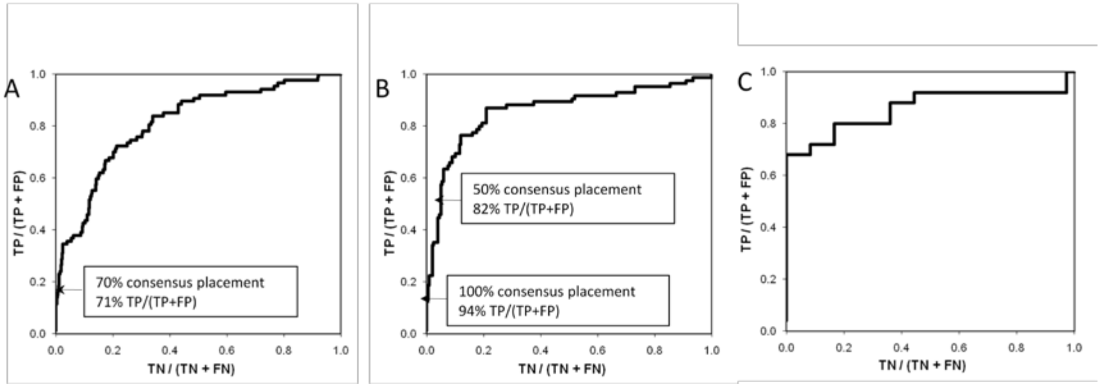

Figure 8. ROC curves for the confidence in repeated placements as well as the performance of the connectivity score.
A) ROC curve of the confidence in placements of single α-helices into density rods based on repeated placements after the assembly step. The fraction of correct placements (true positives / (true positives + false positives)) over the fraction of wrong placements (true negatives / (true negatives + false negatives)) is plotted. The connection between repetition rate and placement confidence has been added to the ROC curve. For example, a placement of a particular α-helix into a specific density rod that is found in 50% of the top scoring models after the assembly step has a 62% confidence of being correct. The area under the curve is 0.81 where 0.5 represents a random measure. B) ROC curve of the confidence in placements of single α-helices into density rods based on repeated placements after the loop building step. The fraction of correct placements (true positives / (true positives + false positives)) over the fraction of wrong placements (true negatives / (true negatives + false negatives)) is plotted. The connection between repetition rate and placement confidence has been added to the ROC curve. The steep increase at the beginning demonstrates that when the same α-helix is placed in one specific density rod in at least 60% of the cases, this placement is virtually always correct. The area under the curve is 0.93 where 0.5 represents a random measure. C) ROC curve of the connectivity score. The fraction of correct connections (true positives / (true positives + false positives)) over the fraction of wrong connections (true negatives / (true negatives + false negatives)) is plotted. The steep increase at the beginning demonstrates that the strongest correct connections score all better than any of the wrong connections. The area under the curve is 0.86 where 0.5 represents a random measure. Reproduced with permission from Elsevier.

(ROC) curve representations for placement confidence after the assembly and loop building steps are shown in panels A and B of Figure 8. The total areas under the curve are 0.81 and 0.86 respectively, indicating strong correlations between frequent placement and correct positioning. For example, a placement of a particular α-helix into a specific density rod that is found in 70% of the top scoring models after the assembly step has a 71% confidence level of being correct. The results for models after the loop building step are even better, corroborating the ability of the algorithm to enrich for true-topology models. For example, a placement of a particular α-helix into a specific density rod that is found in 50% of the top scoring models after the loop building step has an 82% confidence level of being correct.

It would be desirable if the confidence measure allowed distinction between successful and unsuccessful cases in the benchmark. Partial models containing only the α-helices placed with a >50% repetition rate were built for all 10 benchmark proteins. A 50% cutoff ensures that no other placement into that density rod can occur more frequently. We evaluated the overall confidence in a model where k α-helices have been placed confidently out of a total of n α-helices by calculating the number of possibilities to place k α-helices into a total of n density rods (2^k*n!/(n-k)!). This equation explicitly takes into account the number of confidently placed α-helices (k) and the total number of α-helices in the protein (n), and implicitly the fraction of confidently placed α-helices. It also accounts for the fact that placing a specific fraction of α-helices confidently in a large protein is considerably less likely than placing the same fraction of α-helices confidently in a smaller protein. The results of this analysis are plotted in Figure 9. The overall confidence scores for the 10 benchmark proteins fall into two regions within this plot. Some proteins have a low number (3-7) and others have a high number (10-14) on this scale (separated by the dashed line in Figure 9). Proteins below the dashed line contain both successful and unsuccessful cases indicating that there is ambiguity for partial models in this range.

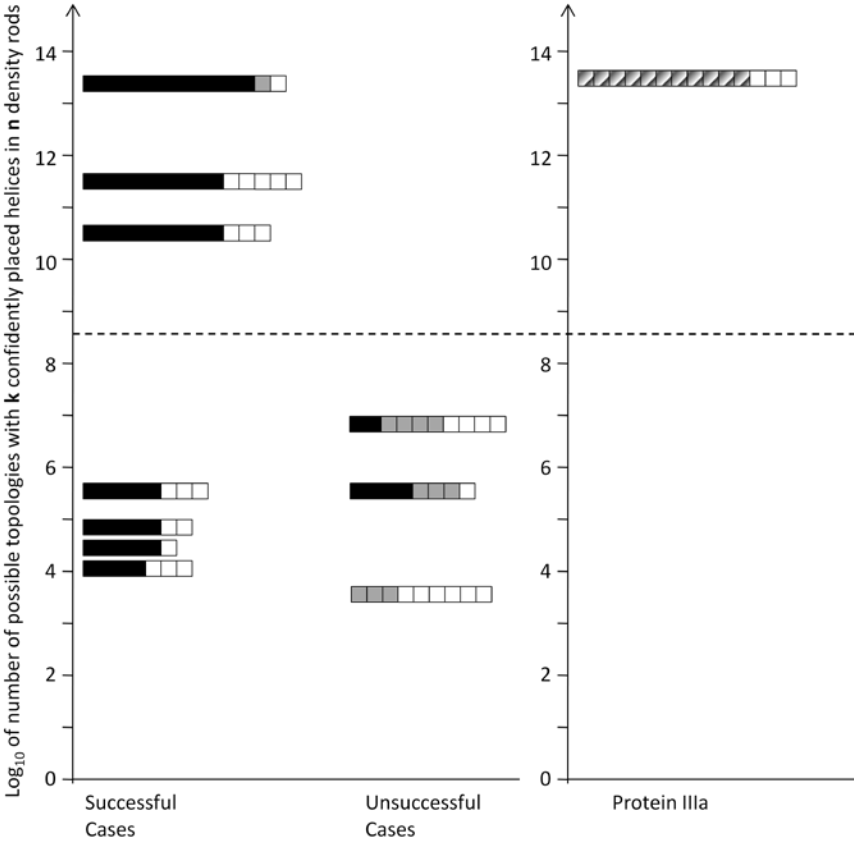

Figure 9. EM-Fold results for the 10 benchmark proteins and adenovirus protein IIIa.
Results are evaluated on the basis of the number of confidently placed α-helices and the total number of α-helices in the protein. The y-axis represents the log base 10 of the number of possible topologies with k confidently placed α-helices in n density rods using the following equation (2^k*n!/(n-k)!). The length of each bar in the plot corresponds to the total number of α-helices in a protein (n). The sum of the black and gray squares within a bar represents the number of α-helices that were confidently placed by EM-Fold (i.e. with >50% repetition rate) (k). Within the subset of confidently placed α-helices, the correctly placed α-helices are in black. The ten benchmark proteins split into two groups as indicated by the dashed line: those with a low number (3-7) and those with a high number (10-14) on this scale. A high number indicates a low probability of confidently placing these α-helices by chance. While there are both successful and unsuccessful benchmark cases below the dashed line, only successful cases are found above the line. For adenovirus protein IIIa 11 out of 14 α-helices are confidently placed by EM-Fold (diagonal pattern, k) and the y-axis number is well above the dashed line. Reproduced with permission from Elsevier.

However, proteins in the upper region (above the dashed line) contain only successful benchmark cases, suggesting that a high value on this overall confidence scale identifies correct topologies. Interestingly, the partial model that we built for adenovirus protein IIIa (discussed below) clusters with the benchmark proteins in the high range of this scale. This gives credence to the protein IIIa model in the absence of an atomic resolution structure.

Poor secondary structure prediction leads to poor assembly results

The three proteins that were not successfully assembled have the poorest secondary structure prediction with an average deviation of 0.8 residues per α-helix in Pool C, compared to an average deviation of 0.3 residues per α-helix for the remaining seven proteins (Table 3). This underscores the fact that failure to find the true solution is not a shortcoming of the assembly algorithm but rather a result of sub-optimal secondary structure prediction. The correct solution of 2AX6 is found despite its poor secondary structure prediction (average deviation of 0.8 residues per α-helix in Pool C) because this protein is small with only six α-helices. In this case the assembly algorithm has to probe a considerably smaller search space and thus can overcome the limitation of poor secondary structure information.

ROSETTA iterative high resolution refinement achieves accurate atomic-detail in parts of the protein models

One of the main challenges of computational protein structure prediction is recovering accurate atomic detail of interfaces within proteins. The top ten scoring loop models of all the seven proteins where the correct topology was identified after loop building were subjected to an iterative ROSETTA refinement protocol (see Experimental Procedures). The objective of this protocol was to test the ability of the method to build accurate atomic-detail structural models at least in part of these proteins. Further it was investigated whether it is possible to uniquely identify the correct topology by the ROSETTA energy score.

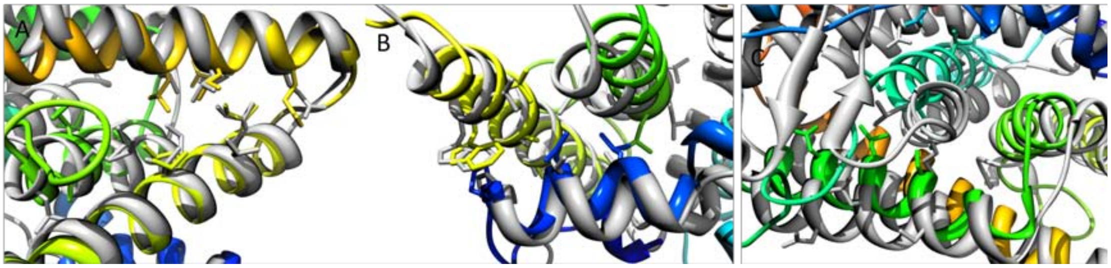

Figure 10. Helical interfaces in best model after ROSETTA iterative high resolution refinement. A-helix-helix interfaces within the best scoring, correct topology, full atom model of protein 1QKM after ROSETTA iterative high resolution refinement. The full atom model is shown in rainbow colors, while the native PDB is depicted in grey. Panels A and B show examples of near-native interfaces in the final model. The α-helix orientations and positions have been correctly identified and the side chain conformations are generally close to the native PDB. Panel C shows an example of a α-helix-helix interface that could not be recovered. Reproduced with permission from Elsevier.

Figure 10 shows close-up views of three α-helix-helix interfaces in the best scoring correct topology model for 1QKM after iterative high resolution refinement. The protocol was able to recover native side chain packing in some of the α-helical interfaces (Figure 10, A and B). However, even in the best scoring model there are still interfaces that are not recovered (Figure 10, C). While generally low RMSD models cannot be identified solely by energy, in six out of seven cases the correct topology can be identified by its enrichment in the 10% model with lowest energy (7.6 for 1Z1L, 4.0 for 1IE9, 3.8 for 1OUV, 2.6 for 1QKM, 1.6 for 1TBF and 1.2 for 1N83). We hypothesize that these enrichments are due to lower energy (higher quality) of the fraction of α-helical interfaces that were built accurately at atomic detail. At the same time non-native α-helix interfaces introduce a background noise that make the energy of models with correct topology often comparable to those of incorrect topology.

For all seven proteins, the native structure obtained from the PDB was minimized in the refinement protocol as well. Its energy is clearly lower than the energy of any of the models built. Thus the absence of models that have accurate atomic detail throughout the entire protein chain is a sampling rather than a scoring problem. This is expected for de novo protein models of 250 and more residues. The size of these systems far exceeds the 90 residue practical limit for de novo high-resolution structure prediction (*33*). However, our finding of native-like α-helix interfaces in portions of these models is an encouraging result that suggests that all-atom accurate atomic-detail models can be achieved as cryoEM reaches higher resolution, and as computational techniques improve. Chapter III deals with a combined approach using improved version of both EM-Fold and ROSETTA to actually successfully build atomic detail models of large proteins.

Comparison of EM-Fold with a computational prediction method for α-helical membrane proteins

In 2007, Kovacs et al. introduced a protocol for predicting atomic resolution details for α-helical membrane proteins guided by EM density maps (*89*). This method uses scripts within the internal coordinate mechanics (ICM) software environment. The ICM-based approach was demonstrated with simulated EM density maps at intermediate resolution for three membrane proteins (GpA, KcsA, MscL). ICM-based flexible fitting of α-helices, optimization of sidechain conformations, and refinement of atomic models resulted in impressive final RMSDs between 0.9 and 1.9 Å for the three test membrane proteins.

While the general idea of guiding protein structure prediction by α-helical density rods observed in intermediate resolution EM density maps is the same for the ICM-based method (*89*) and EM-Fold, there are substantial differences between the methods. In the demonstration of the ICM approach perfect secondary structure prediction was assumed. We have tested EM-Fold with both perfect and realistic secondary structure prediction information including variations in α-helix lengths. Secondly, the test proteins used in the ICM demonstration are sufficiently small (with 1 or 2 α-helices per monomer), and have α-helices of differing lengths (in the case of 2 α-helices per monomer), so that the assignment of α-helices into specific density rods is trivial. The centerpiece of the EM-Fold protocol is the assembly step (Figure 5B), which is designed to identify the topology of a protein from its α-helical secondary structure prediction and the positions of density rods in the density map. Subsequent steps (Figure 5C and D) refine the model. The ICM-based algorithm does not have an assembly step, while the refinement steps in both protocols follow similar principles. In their current setups these algorithms are complementary and it is conceivable that models derived from EM-Fold could be input into ICM for further refinement.

Benchmark of EM-Fold on experimental bovine metarhodopsin density map

To demonstrate EM-Fold's ability to work reliably in conjunction with experimental data, we built a model for bovine metarhodopsin based on the 5.5 Å resolution cryoEM density map obtained from the EMDB Database (*90*). The crystal structure of bovine rhodopsin (PDBID 1GZM, (*91*)) was used to evaluate the results. The crystal structure is in a different conformational state than the cryoEM structure. The overall fold of the protein is the same however as the authors note that the meta I formation involves no large movements or rotations of α-helices from their ground state (*90*). So while there may be structural differences in the loop regions, the α-helical regions that are modeled in the protocol are well described by the crystal structure. Interestingly the authors report density bumps for several Trp side chains in the 5.5 Å resolution cryoEM density map. Bovine rhodopsin is mostly α-helical (63%) and slightly larger than the largest of the 10 benchmark proteins (349 residues, Table 1).

The same protocol that was used for the ten benchmark proteins was applied to bovine metarhodopsin. The results are summarized in Table 2. The correct topology is ranked second after the assembly step and is ranked first after the refinement step. After the loop building step the correct topology is the best scoring model. This model has an RMSD of 7.9 Å to the crystal structure. If the crystal structure was not available, we could evaluate the EM-Fold results on the basis of the overlap between Trp sidechains and Trp density bumps on rods. Only a single good scoring model has all of the Trp containing α-helices in density rods with Trp density bumps. This model corresponds to the correct topology. These results demonstrate the ability of EM-Fold to work accurately in combination with experimental density maps. The rather large RMSD value is in part caused by the conformational change between crystal and cryoEM structure, particularly in the loop regions. The RMSD over α-helical residues is only 3.1 Å, making this an excellent model for a protein of this size.

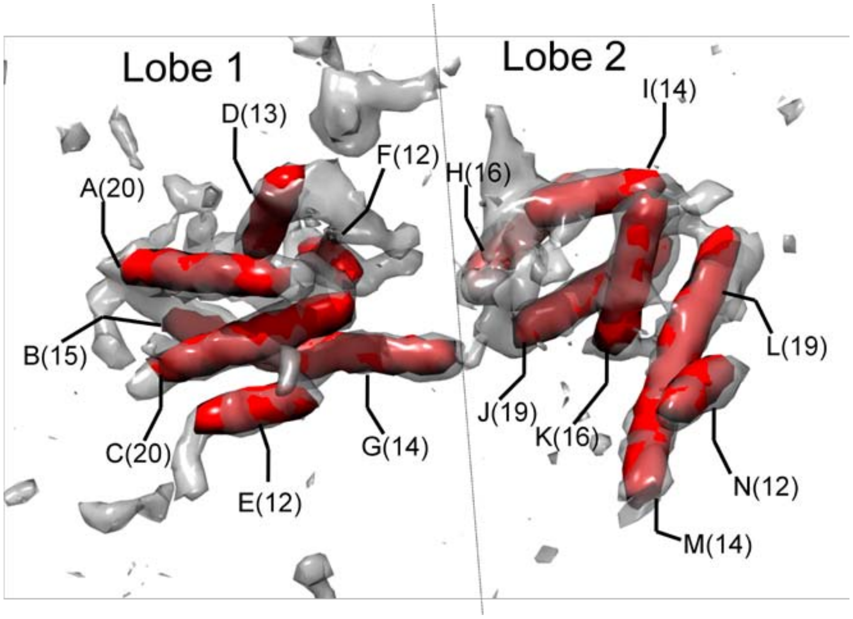

Figure 11. Cryoelectron microscopy density map of adenovirus protein IIIa.
Experimental cryoelectron microscopy (cryoEM) density map of adenovirus protein IIIa (grey) shown segmented from an adenovirus reconstruction at 6.9 Å resolution (FSC 0.5 threshold) (*6*). 14 rods of minimum length 18 Å have been identified asα -helical regions (red). Each rod is labelled with a letter and the number of α-helical residues corresponding to its length. The EM-Fold assembly step involves placing α-helices from the secondary structure prediction pool into the 14 identified density rods. The protein IIIa density has a two-lobe topology, with Lobe 1 comprised of rods A-G and Lobe 2 comprised of rods H-N. In the adenovirus capsid, Lobe 1 is closer to the penton base. Reproduced with permission from Elsevier.

Evaluation of adenovirus protein IIIa folds by EM-Fold

We have also applied EM-Fold to the medium resolution cryoEM density assigned to protein IIIa in the adenovirus capsid (*6*). In this case we do not have an atomic resolution structure for protein IIIa. This is a challenging case for EM-Fold because the α-helical region of protein IIIa is larger than any of the benchmark proteins and it has a two-lobe topology (Figure 11). This two-lobe density region contains 14 manually identified density rods and is assigned to the N-terminal 400 residues of protein IIIa, which are predicted to be highly α-helical. As none of the 10 benchmark

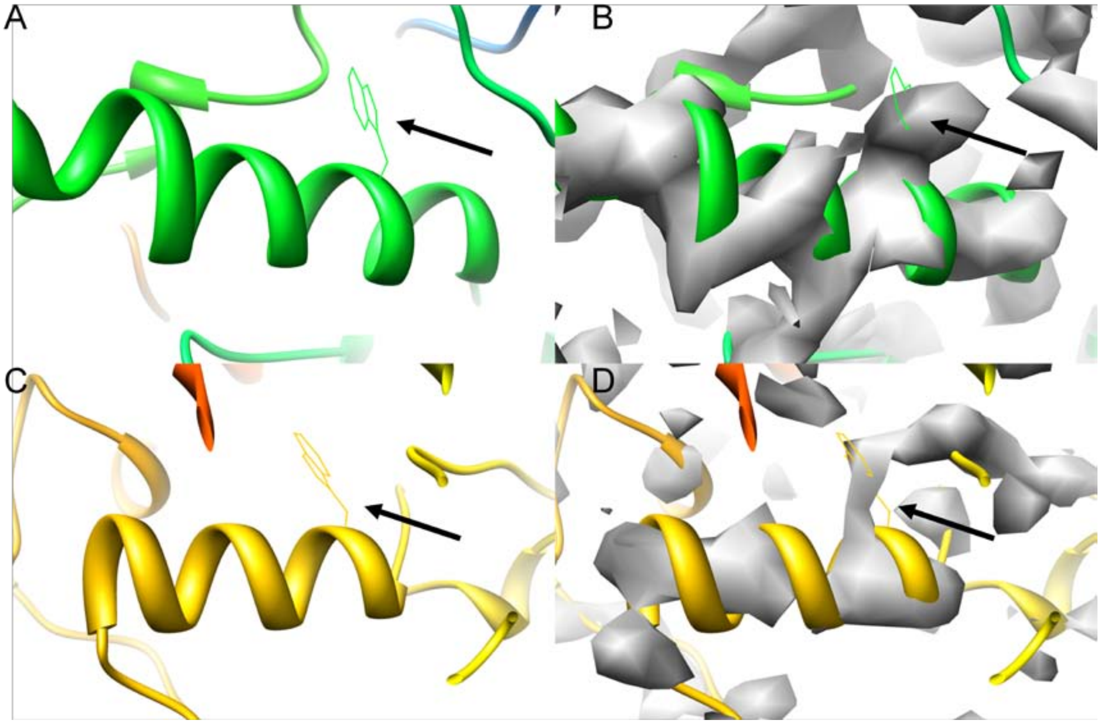

Figure 12. Density bumps in hexon.
Density bumps for the only two Trp sidechains in helices 3 (A, B) and 6 (C, D) of hexon. Panels A and C show the hexon crystal structure for residues 462-480 (helix 3, panel A) and residues 759-773 (helix 6, panel C) with the Trp side chain coordinates. Panels B and D show the crystal structure for the same helices overlaid with the refined Ad35F density map. For both Trp side chains clear density bumps are visible. The black arrows represent the position of the Trp side chain. Reproduced with permission from Elsevier.

proteins or rhodopsin have a two-lobe topology, this complication has not been tested in EM-Fold. Therefore we used experimental information to assign the two lobes and also to filter the models produced by EM-Fold.

In order to extend the resolution of the Ad35F cryoEM structure, we increased the dataset size to a total of 7133 particle images and performed several additional rounds of Frealign refinement. The final Ad35F structure is based on 3040 particle images and has a resolution of 6.8 Å at the FSC 0.5 threshold (and 5.8 Å at the FSC 0.3, and 5.2 Å at the FSC 0.143 thresholds). The crystal structure of the Ad5 hexon reveals that there are two α-helices of ten or more residues that have a Trp (*92*). We observe prominent bumps for the Trp side chains on each of these two α-helices in the 6.8Å cryoEM density map (see Figure 12).

Using the criteria developed to identify Trp in hexon, three possible positions (in rods E, K, L) were identified in protein IIIa that might correspond to a Trp side chain. The side chain bump in rod E is at the end of the rod, while the bumps in rods K and L are both in the middle of the rods and in fact form a connection between these two rods. Analysis of the protein IIIa sequence indicates that there is only one Trp in a predicted α-helix (residue 27) and that it corresponds to the first or second residue in the predicted α-helix. This excludes rods K and L, as corresponding to the α-helix with a Trp, since the observed bumps are in the middle of these rods. We hypothesize that the observed bumps in rods K and L belong to two aromatic side chains that are in contact. After analyzing the cryoEM density, we conclude that the most likely rod to contain the predicted α-helix with a Trp (amino acids 27-39) is rod E. This lobe assignment for protein IIIa is in agreement with the N-terminal tagging experiment recently published (*1*). The protein IIIa peptide tag study localizes the N-terminus of protein IIIa to the inner capsid surface close to the interface between penton base and the peripentonal hexons. Specifically the difference density attributed to an N-terminal FLAG tag on protein IIIa is observed in the vicinity of what we refer to as rod E in lobe 1 of protein IIIa (Figure 11). Therefore both the analysis of the sidechain density and the protein IIIa N-terminal tagging information indicate that lobe 1 should be assigned to the most N-terminal portion of protein IIIa.

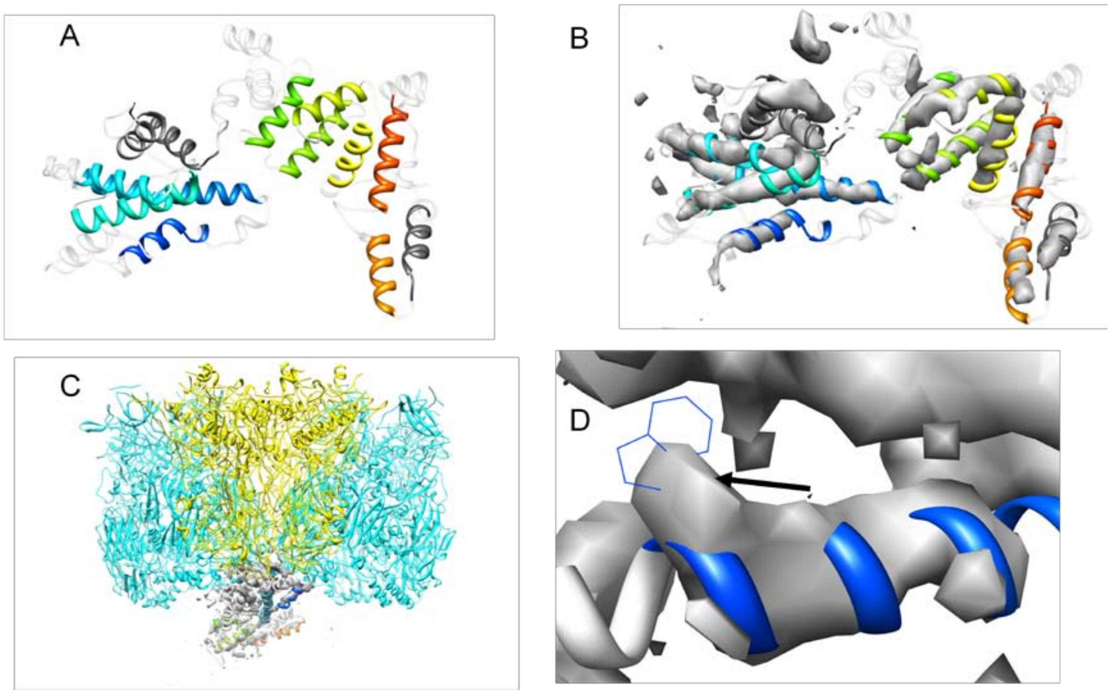

Figure 13. Model of protein IIIa.
A) A reduced model of protein IIIa where only α-helices that have been placed with at least 50% repetition rate are colored in rainbow. This topology agrees with the San Martin et al. (*1*) results. 11 out of 14 α-helices can be placed with a confidence of at least 82%. The remaining three α-helices have been colored in grey while the loop regions are shown in white. B) Same as in A, but shown but shown in density. E) Side view of partial model of protein IIIa (rainbow) in contact with penton base (yellow) and two peripentonal hexons (light blue). F) Density bump in rod E of the refined Ad35F density of protein IIIa that has been assigned to Trp27. The arrow marks the position of the side chain. Reproduced with permission from Elsevier.

After applying the same EM-Fold protocol used for the ten benchmark proteins and rhodopsin, we analyzed the top 100 models for protein IIIa and found that 33 of these have the N-terminal ~200 residues of protein IIIa positioned into lobe 1. A detailed analysis of this subset of models indicates that 14 models have the predicted α-helix for residues 27-39, which includes the Trp at position 27, placed into rod E. We consider these 14 selected models the most likely models for protein IIIa. Within these 14 models, we note that four α-helices (corresponding to residues 50-

60, 70-83, 230-242 and 251-264) are placed into specific rods (G, B, H and J, respectively) in all of the cases. Therefore we assign these α-helices, as well as the Trp-containing α-helix (rod E), as having a very high (>94%) confidence level. An additional six α-helices are placed with >50% repetition rate and thus are assigned a high (>82%) confidence level as labeled in the ROC curve in Figure 8B. A partial model of protein IIIa that contains these 11 confidently placed α-helices is shown in rainbow in Figure 13, panels A and B. The remaining three α-helices are shown in grey and the loop regions are shown in white indicating that their positioning within the density is more ambiguous. The number of confidently placed α-helices puts this partial model into the confident region in Figure 9, further increasing the probability that it is correct. The proposed 50% confidence protein IIIa model is shown in context with penton base and two nearby peripentonal hexons (Figure 13C). Also, the agreement of the Trp (residue 27) side chain with the bump in rod E is shown in Figure 13, panel D. Interestingly, one of the α-helices placed with a high confidence level (rod L) contains a Tyr residue (Y369) in the middle of the α-helix that corresponds to the density connection observed between rods K and L. On top of this another confidently placed α-helix places Y299 in the middle of the connected density rod (rod K). This confidence assignment agrees perfectly with the observed density connection between rods K and L and gives further credence to our model. We anticipate that higher resolution cryoEM density revealing more of the side chains, combined with additional computational modeling, would resolve the remaining ambiguities in the protein IIIa fold model.

Conclusions

EM-Fold is a novel computational protein folding algorithm that assembles α-helical proteins guided by medium resolution density maps. Chapter III deals with the extension of EM-Fold to include β-strands in the assembly algorithm if the cryoEM density maps allow an unambiguous identification of β-strands. A benchmark on ten proteins shows a 100% success rate for the assembly of α-helices when the correct secondary structure information is assumed. When predicted secondary structure information is used, which includes some incorrect information, the success rate drops to seven out of ten. Our results demonstrate that the 30% failure rate is linked to incorrect secondary structure prediction information and future developments will include improving the secondary structure prediction input. This might be done by either improving the secondary structure prediction algorithms themselves or – as demonstrated here – by including more diverse predictions into a more complex pool of α-helices prior to assembly. The final models generated by EM-Fold display RMSDs in the range of 3.9 Å to 7.1 Å for the benchmark proteins. A complete model for rhodopsin with 7.9 Å RMSD could be built based on an experimental density map. These results demonstrate that de novo protein structure prediction can be extended to proteins well beyond 150 amino acids if the search is guided by medium resolution density maps.

The iterative ROSETTA refinement protocol did not completely succeed in refining the models to accurate atomic detail. Given the large size of the proteins this is not entirely surprising. However, portions of the final models, including specific α-helix-helix interfaces, do have correct atomic resolution detail. These partial native-like arrangements lead to an enrichment of correct topology models by energy. Chapter III demonstrates that an improved iterative sampling protocol that includes the density map as an experimental restraint allows refinement to atomic detail accuracy for complete models.

EM-Fold has been applied to build a model of adenovirus protein IIIa, a protein for which we have a medium resolution cryoEM density map but no atomic resolution structure. Based on the experimental constraints provided by N-terminal tagging (*1*) as well as observed side chain density in a refined cryoEM density map, we were able to assign the lobe topology of the protein. We also used this experimental information as a filter to select the most likely fold models for protein IIIa produced by EM-Fold. We present a fold model for protein IIIa with 11 of the 14 α-helices placed with a high level of confidence.

Experimental Procedures

Overall protocol

The flowchart of the full assembly process is shown in Figure 5. The generation of a pool and identification of density rods is followed by the main assembly step in EM-Fold, a refinement step within EM-Fold, and loop and side chain building in ROSETTA. The assembly step builds α-helices from the pool into the density rods. Three sequence-independent, computationally inexpensive, and therefore low resolution scores are used to build a large number of initial models. The best scoring models from the assembly step are refined using sequence-dependent, medium-resolution scores and leaving the overall fold of the protein unchanged. The last step of the assembly protocol uses the existing ROSETTA software (*79, 93*) to model loops for the best-scoring models that emerged from the refinement step. Side chains are constructed using ROSETTA relaxation and repacking strategies (*33*). This is the computationally most expensive and highest resolution step of the model building process and is thus only applied to a handful of final models.

Secondary structure prediction pool

To minimize secondary structure prediction inaccuracies, three different secondary structure pools (A, B, C) were investigated. Pool A uses the secondary structure prediction programs jufo (*18, 19*), psipred (*20*) and sam (*21, 22*) to get three state predictions of the secondary structure of the benchmark cases. Sequences of more than nine amino acids predicted to be α-helical were considered to be a likely position of a non-short α-helix and were added to a “pool” of possible secondary structure elements. In addition to the individual predictions, a consensus secondary structure prediction was calculated by averaging jufo, sam, and psipred. Also α-helices longer than 21 residues were split into two, further expanding this pool.

In pool B copies of the α-helices from pool A were replaced with copies that are extended by one amino acid on both sides. Thus pool B has the same size as pool A, but all of the α-helices are two residues longer. This procedure eliminated the bias in pool A towards α-helices that are too short and reduces the per α-helix deviation from the correct secondary structure from 1.5 residues in pool A to 0.8 residues in pool B.

Pool C combines pools A and B and adds further versions of α-helices extended by one amino acid either on the N-terminus or the C-terminus. As a result the secondary structure element pool C has four versions of each α-helix with different lengths available for assembly. The per α-helix deviation from the correct secondary structure in pool C is 0.4 residues. The length deviations of the elements that are closest in length and have maximal sequence overlap with the true α-helices are reported in Table 3 for all three versions of the prediction pool.

EM-Fold scoring function

Three sequence-independent scores are used during the assembly of the fold: a loop, an occupancy, and a connectivity score. The loop score is a knowledge-based score that evaluates the likeliness of a certain C_α-C_α distance between terminal residues in an α-helix being bridged by a specific number of residues. It has a preference for short EUCLIDEAN distances between beginning and end of a loop (data not shown).

The occupancy score evaluates the length agreement of a density with an α-helix that is placed in it, with unfilled densities getting the maximum unfavorable score. Thus the occupancy score drives the algorithm toward filling the density map completely.

The connectivity score is based on the assumption that for short loops a medium resolution density map contains valuable information in the form of stronger density in the loop regions between density rods. The connectivity score employs a skeletonization algorithm (*8*) to find the highest intensity connection between all pairs of termini of density rods that are closer than 10 Å in space. This information is converted into a score that assesses whether the connection is a strong or a weak one.

The connectivity score has been tested on the ten benchmark proteins. Within the ten proteins there are 65 pairs of density rods whose ends are closer than 10 Å. 25 of these pairs correspond to connected density rods. Figure 8C shows a ROC curve based on the strength of the connection. The area under the curve is 0.86, clearly showing the ability of the connectivity score to enrich for native connections. Out of 14 connections whose strength is more than one standard deviation above the average connection strength, 12 correspond to true connections.

EM-Fold assembly step

The sampling of conformational space is performed in a Monte Carlo algorithm in conjunction with the Metropolis criterion. When placing an α-helix from the pool into a density two physical constraints are checked: First, whether the length of the α-helix fits the density within a deviation of three residues (corresponding to a maximum length deviation of 4.5 Å). This "length-tolerance-check" accounts for inaccuracies both in secondary structure prediction and in length determination of density rods. Secondly, it is checked whether the residues between the α-helix and all previously placed α-helices are sufficient to fill the gaps between α-helices. The maximum loop length was set to 3.0 Å per amino acid plus an additional 6.0 Å per loop. If one of the constraints is violated, the move will be rejected because the resulting model would not agree with the density map. All placements that do not violate these constraints are evaluated by the three sequence-independent scores discussed above. Assuming that *x* density rods have been identified in the density map and the pool contains y α-helices, there is a total of N_{pos} number of possibilities to place the α-helices into the density rods.

$$N_{\text{pos}} = \binom{n}{k} k!\, 2^k \;=\; \frac{n!}{(n-k)!}\, 2^k\,,$$

with n=max(x;y) and k=min(x;y). This same equation is also used to calculate an overall confidence score for a partial model built by EM-Fold by reassigning n to the total number of α-helices and k to the number of confidently placed α-helices (with >50% repetition rate).

The Monte Carlo moves (see Figure 14) that are used in the assembly step are: B) adding an α-helix from the pool to the model, C) deleting an α-helix from the model, D) flipping the orientation of an α-helix in the model, E) swapping the positions of two α-helices within the model, F) swapping an α-helix from the model with one from the pool, and G) moving an α-helix from the model to an empty density rod. The orientation of an α-helix after any move that results in placement of a new α-helix (moves B, E, F, and G) is arbitrary. A simulated annealing Monte Carlo Metropolis search is used where the temperature is decreased linearly from 0.25 to 0.08 over 2000 rejected steps. The weights of the scores are 1.0 (loop), 0.4 (occupancy) and 0.8 (connectivity). The final total scores range from -4.2 (2AX6) to -22.1 (1OUV). It is important to note that the temperature values are somewhat arbitrary and do not correspond to physiologically relevant temperatures.

EM-Fold refinement step

The lowest scoring models are used in a second medium resolution Monte Carlo refinement search. This refinement step uses different moves and scores than the previous assembly step. The moves constitute small perturbations of the model – shifts along the α-helical axis and rotations around the α-helical axis. A set of knowledge-based scores is used including an amino-acid-distance score, a neighbor-count score, a secondary-structure-element-packing score, a compactness-measure in form of a radius-of-gyration score, and the loop and occupancy scores already used in the previous step. These scores are described in detail in the supplementary data. The occupancy score avoids α-helices sliding out of their density rods. This refinement step maintains the fold of the model but identifies correct α-helix-helix-interfaces. A simulated annealing Monte Carlo Metropolis search is used where the temperature is decreased linearly from 0.25 to 0.03 over 2000 rejected steps. The weights of the scores are 10 (loop), 4 (occupancy), 0.2 (aadist), 0.2 (neighbor count), 0.14 (radius of gyration) and 2 (ssepack). The final total scores range from -139 (2AX6) to -367 (1TBF).

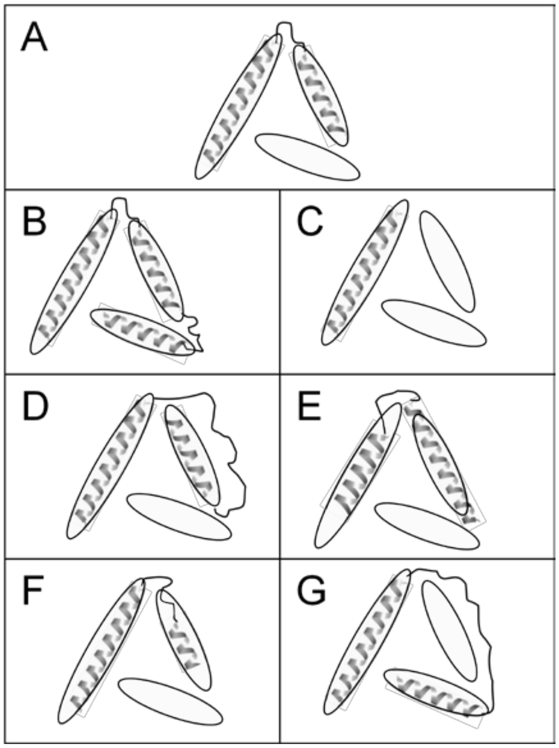

Figure 14. Schematic representation of the moves used in the assembly step of the protocol.
Panel A shows the state of the model before the move. The add move (B) adds an α-helix from the pool into an empty density rod. The delete move (C) removes an α-helix from a density rod and returns it to the pool. The flip move (D) rotates one α-helix within a density rod by 180° perpendicular to its long axis. The swap move (E) exchanges two α-helices within density rods. The swap with pool move (F) exchanges an α-helix within a density rod with one from the pool. Move G removes an α-helix from its density rod and places it into another empty density rod. Reproduced with permission from Elsevier.

ROSETTA loop and side chain building step

For identification of the correct fold as well as for building a full atom model of the protein the ROSETTA software (*33, 79, 93*) was used. The backbone atoms of the residues that are missing in the EM-Fold models are built using the ROSETTA cyclic coordinate descent loop building protocol (*79*). The resulting models with loops are scored in the ROSETTA force field and sorted according to their score. This score can discriminate the correct from non-native topologies as demonstrated in the benchmark. For the seven successful benchmark proteins the ten best scoring topologies according to the ROSETTA score were chosen and underwent an extensive refinement protocol within ROSETTA. This protocol included building 1,000 EM-Fold-refined models per topology (10,000 models total). For each of the 10,000 refined models, 5 loop models were built in ROSETTA (50,000 models total).

Eight rounds of iterative side chain repacking and backbone relaxation in ROSETTA followed (*33*). All 50,000 models undergo round one. Only models that stay within 2.5 Å of the starting structure and are within the best 10 % scoring models according to the ROSETTA full atom energy are run through rounds 2-8. After the eighth round the best 10 % scoring models are analyzed according to their enrichment for the correct topology. The enrichment is computed as the ratio of relative frequency of correct topology models within the best 10 % scoring models to relative frequency of correct topology models within all models.

Benchmark on simulated density maps

The proposed EM-Fold search algorithm was benchmarked on ten proteins that were chosen to be mainly α-helical, exhibit non-redundant folds, possess 250 to 350 residues and form 6 to 14 α-helices of at least 12 residues in length (Table 1). Electron density maps for all ten benchmark cases were created from the coordinates. PDB2VOL of the SITUS package (*35*) was used to simulate density maps with 6.9 Å resolution, a voxel spacing of 1.5 Å and Gaussian flattening.

Positions and lengths of the density rods were identified manually since available α-helix identification algorithms did not perform satisfactorily for either the simulated densities of the benchmark proteins or for the protein IIIa density. Errors in manual identification of α-helix length can be compensated by the length tolerance that is used in the assembly step. To test the influence of the resolution of the simulated medium resolution density map, maps at 9.0 Å resolution were also simulated. Positions and lengths of the density rods were identified manually for the 9.0 Å resolution maps as well.

Furthermore, it should be stressed that, independent of whether the density rods are identified manually or using automated software, there is always the possibility that density regions in medium resolution density maps that do not correspond to α-helices are identified as α-helical regions. An example for this is a β-hairpin of at least 4 residues in each strand. Likewise it is possible that an α-helical region in the protein is not identified as a density rod in the map (in the case of a more flexible α-helix for instance). In both cases EM-Fold is still capable of finding the correct topology, as the algorithm neither requires all identified rods to be filled with α-helices, nor all predicted α-helices to be placed in identified rods.

Benchmark on experimental density map

EM-Fold was also benchmarked on the experimental cryoEM density map of bovine metarhodopsin (EMDB Entry EMD-1079). The density map is reported to have a resolution of 5.5 Å and has a voxel size of (0.4 Å, 0.5 Å, 1.7 Å). A single subunit of the protein was segmented from the density map. Bovine rhodopsin has 349 residues and is highly α-helical (63% α-helical) with 8 α-helices of 12 or more residues.

Protein IIIa structure elucidation

The adenovirus vector Ad35F was used in previous cryoEM structural studies (*6*) and has been refined further with more data (7133 particle images) with the program Frealign (*94*). A negative temperature factor of 450Å^2 was applied to the final map and the structure was filtered at 5.1 Å using a filter with a cosine-shaped cut-off and a width of ~20 Fourier pixels. Ad35F is composed of the Ad5 capsid and the Ad35 fiber. The density for one copy of protein IIIa was segmented from an Ad35F reconstruction. The Ad5 protein IIIa has 585 residues. The 400 N-terminal residues are predicted to be mainly α-helical, while the remaining C-terminal residues are not predicted to have many secondary structural elements. In the density map 14 density rods of at least 18 Å in length and 6-7 Å diameter (corresponding to α -helices of at least 12 residues) were identified manually (Figure 11). A secondary structure element pool with a total of 257 α-helices was built using the protocol described for pool (C). 100,000 models were built for protein IIIa according to the assembly procedure established for the benchmark set of proteins. The models were ranked by score. 100 refined models were constructed for each of the top 150 models produced by the assembly step. The resulting 15,000 models were sorted by score and the top scoring model of each of the 150 topologies was selected for loop construction. A topological model that was built using EM-Fold is presented for the first 400 residues of protein IIIa.

CHAPTER III

BUILDING PROTEIN MODELS ACCURATE AT ATOMIC RESOLUTION FROM MEDIUM RESOLUTION DENSITY MAPS

Introduction

The results published in (*5*) (see also Table 2) showed that the models built by EM-Fold and refined by Rosetta generally had RMSDs between 4 and 7 Å over the full length of the protein. These results demonstrate that it is possible to use computational methods to generate models for proteins that have the correct topology from a medium resolution density map (and the primary sequence) alone. However, the models were accurate at atomic detail only for limited regions despite extensive refinement in Rosetta. Parts of some final models showed side chain recovery in helical interfaces (Figure 10, panels A and B), but in general side chain conformations were not predicted correctly. Often predicted helices lacked bends or differed in length from helices in the experimental structure (Figure 10, panel C). Furthermore, even though the true topology could be enriched by score, it could not be identified by virtue of score alone. On the other hand, the high-resolution experimental structures had considerably better scores than any of the models built using the EM-Fold protocol. It was concluded that model refinement to atomic detail accuracy failed due to insufficient sampling: starting from EM-Fold models Rosetta refinement does not construct models sufficiently close to the native structure to stand out by score.

In short, EM-Fold is capable to use the experimental data at its level of resolution which defines the topology of the fold. However, the EM-Fold protocol failed in adding detail not visible in the experimental data. The latter aspect however was demonstrated to be a strength of the Rosetta algorithm when combined with NMR and EPR experimental data (*32, 82-87*). It was speculated that this shortcoming of EM-Fold may be caused by the following three reasons. Firstly, the models coming out of the EM-Fold refinement step have too large inaccuracies in the backbone. The moves performed in the Rosetta refinement protocol are unable to alter the model sufficiently to resolve these deviations. Secondly, the Rosetta refinement protocol is not guided by the density map. Therefore larger scale deviations such as length or bending of SSEs cannot be corrected. However, accurate construction of loop regions and side chains depends on models with very high agreement of backbone coordinates within secondary structure elements. Lastly, it might be that the size of the proteins was too large. Rosetta has been demonstrated to work well for proteins up to 120-180 amino acids. Even with the perfect setup it might not be possible to refine proteins of 250 to 350 residues. We expect a combination of these three reasons contributing to failure. The present work overcomes these limitations and demonstrates atomic-detail accurate construction of protein structure from medium resolution density maps.

Improve EM-Fold refinement step to build models with lower RMSDs for Rosetta refinement

Specifically, EM-Fold was modified to allow bending, translation and dynamic length modification of secondary structure elements during protein folding and refinement. For the models to accurately reflect such detail the scoring function was adapted to enable direct comparison of the models with the density map. The objective of these modifications was to construct models with higher accuracy in protein backbone coordinates.

Improved version of Rosetta that uses density map as restraint will help refinement

After the EM-Fold algorithm was published (*5*), the capabilities of the Rosetta software were extended to use density maps as restraints in structure refinement (*56*). These modifications allow construction of loops and side chains guided by the density map addressing the second shortcoming of the original algorithm. We expect the most important effect is in the construction of loop regions that connect the secondary structure elements. Accurate construction of the protein backbone in these regions is crucial for successful refinement. While density maps contain information about conformation of short loops and placement of short secondary structure elements, this information was largely ignored in the original implementation of EM-Fold. Using the density map as a restraint will preferentially place the short secondary structure elements and loop regions in the strongest density region between two secondary structure elements. Further, these features also allow the targeted reconstruction of regions that agree least with the density map (*56*).

Rosetta refinement will likely work better with proteins of smaller size

Testing the influence of protein size on the refinement results can be done by benchmarking the established EM-Fold / Rosetta protocol on a set of proteins of smaller size. A mixture of α-helical, β-strand and α/β-proteins with 150 – 250 amino acids were chosen to test the algorithm.

Results & Discussion

Refinement of true topology models with the new Rosetta density functionality

The best true topology model after the EM-Fold refinement step for each of the seven successful cases from the benchmark in (*5*) was subjected to Rosetta loop building and refinement using the new density restraint functionality. This should be considered more as a proof of principle rather than part of a generalizable folding protocol. It is an ideal opportunity to test hypothesis number 2 separately: i.e. is Rosetta (using the density map as restraint) able to build better models than Rosetta without using the density map as a restraint? No change to the EM-Fold refinement protocol as well as to the size of the protein will be done while testing the new refinement on the old benchmark set. It should be stressed that only the true topology model will undergo loop building and refinement in Rosetta. The runs in Rosetta for proteins of this size are so time intensive that it was not feasible to refine a number of top scoring topologies produced by the EM-Fold refinement step. So even if the protocol was successful in producing atomic detail models, it would only be of limited practical value because there is no guarantee that in a blind test the best scoring model after EM-Fold refinement actually corresponded to the true topology.

Table 4. Results of Rosetta refinement on seven successful benchmark proteins

protein	RMSD start model [Å]	best RMSD model after round 1 [Å]	best RMSD after round 2 [Å]	best RMSD after round 3 [Å]	DireX fitted pdb after round 3 [Å]
1IE9	(2.22)	3.88 (2.25)	3.05 (1.90)	2.55 (1.93)	2.15 (1.89)
1N83	(4.68)	5.21 (4.27)	4.31 (3.18)	3.41 (2.63)	3.01 (2.42)
1OUV	(2.21)	2.37 (2.01)	2.05 (1.80)	2.00 (1.79)	1.89 (1.50)
1QKM	(2.95)	3.72 (2.93)	2.87 (3.02)	2.82 (2.33)	2.79 (2.35)
1TBF	(1.93)	3.32 (2.26)	2.86 (2.19)	2.37 (1.96)	2.25 (1.93)
1Z1L	(2.70)	3.94 (3.05)	3.62 (3.12)	3.24 (2.66)	2.97 (2.48)
2AX6	(2.26)	4.22 (2.48)	3.61 (2.73)	2.99 (2.36)	2.81 (1.91)

RMSD values were determined over the backbone atoms N, C_{α}, C and O. Values in parentheses refer to RMSDs over secondary structure elements only.

The best models that were produced in the EM-Fold refinement step for 1IE9 (2.22 Å RMSD), 1N83 (4.68 Å RMSD), 1OUV (2.21 Å RMSD), 1QKM (2.95 Å RMSD), 1TBF (1.93 Å RMSD), 1Z1L (2.70 Å RMSD) and 2AX6 (2.26 Å RMSD) were taken as start models for the Rosetta loop building. The loop building (round 1) was followed by two more rounds of identifying the regions of the models that agree least with the density map and then rebuilding these regions and relaxing the entire protein (*56*). After every step the lowest RMSD model is identified (even if it is not the best scoring model) and used as the starting point for the next round of refinement. After the last round of Rosetta refinement, DireX (*95*) is used to obtain the best possible agreement with the density map. Table 4 summarizes the results of this test. After three rounds of Rosetta refinement and DireX flexible fitting, the RMSDs of the final models range from 1.9 Å to 3.0 Å over the full length of the proteins and between 1.5 Å and 2.5 Å over the helical residues. These results are considerably better than the results obtained when refining with the version of Rosetta that does not use the density map as a restraint, where the RMSDs of the best RMSD models ranged from 3.9 Å to 7.1 Å over the full length of the protein (see Table 2). Figure 15 shows the final models overlaid with the native structure for four of the seven benchmark cases. In fact, achieving 2.3 Å for a protein of 347 residues is very impressive. For proteins of size 250 to 350 residues, RMSD values below 2.5 Å generally mean that side chain conformations at least within the core of the protein are correctly recovered. The agreement of side chain conformations with the native structure is shown in Figure 16. Despite the clear improvement of results compared to (*5*), there

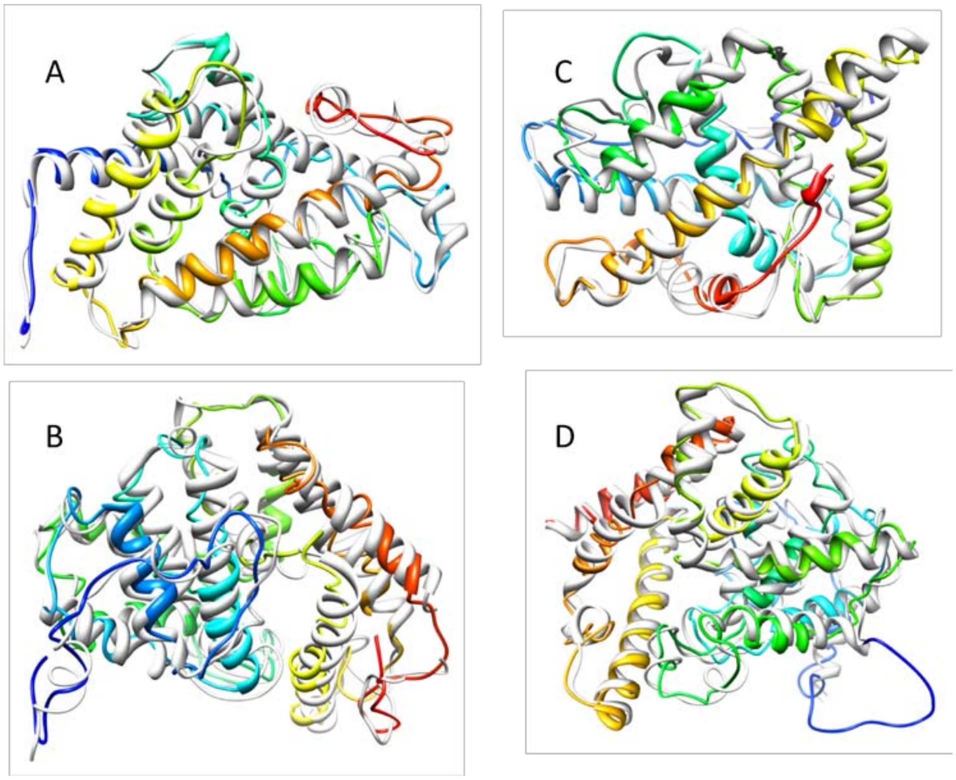

Figure 15. Superimposition of final models after Rosetta refinement with native structures. Superimposition of the final models (colored in rainbow) of 1IE9 (A), 1Z1L (B), 2AX6 (C) and 1TBF (D) with the original PDB structures (grey). (A) 1IE9 has 259 residues. The model shown has a RMSD of 2.15 Å over the full length of the protein and 1.89 Å over the helical residues. (B) 1Z1L has 345 residues. The model shown has a RMSD of 2.97 Å over the full length of the protein and 2.48 Å over the helical residues. (C) 2AX6 has 256 residues. The model shown has a RMSD of 2.81 Å over the full length of the protein and 1.91 Å over the helical residues. (D) 1TBF has 347 residues. The model shown has a RMSD of 2.25 Å over the full length of the protein and 1.93 Å over the helical residues.

remains one caveat to these results. For time limitations only the true topology model was refined and only the best RMSD models were picked after each step to determine the structure that advances into the next round. This does not constitute a realistic benchmark as the lowest scoring topology after EM-Fold refinement is not necessarily the true topology and the very best scoring model after each Rosetta refinement step is not necessarily the lowest RMSD model. This point was addressed in a benchmark that contained 20 α-helical and seven β-sheet proteins with 150 to 250 residues. The rationale behind this choice was that the smaller size of the proteins will make it possible for Rosetta to work on multiple top scoring topologies from the EM-Fold refinement step in a reasonable time. The reduced size might also limit the conformational search space sufficiently for Rosetta to build atomic detail model for at least some of the proteins.

Benchmark database of twenty α-helical and seven β-sheet proteins with 150 to 250 residues

In order to address some of the shortcomings of EM-Fold in its original version and to test the performance of the entire folding protocol with smaller proteins, a benchmark set that contained a total of 27 proteins was assembled. The benchmark set focuses on helical proteins as these represent the majority of the application cases as α-helices are observed readily at medium resolution. However, performance was tested on seven proteins with β-sheets. Density maps at 6.9 Å resolution were simulated for the 20 α-helical proteins, while maps at 5.0 Å resolution were simulated for the seven β-sheet containing proteins. A map of 5 Å resolution or better is needed to identify positions of β-strands.

Results of EM-Fold assembly step with perfect and realistic secondary structure prediction are similar to previous success rates

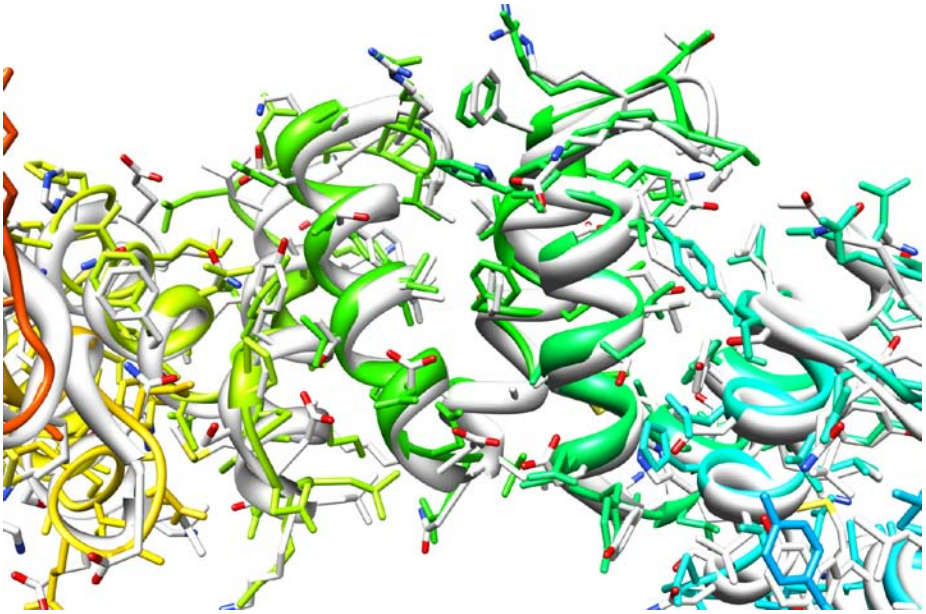

Figure 16. Detailed superimposition of final model after Rosetta refinement with native structure. Superimposition of the final model (colored in rainbow) of 1OUV with the original PDB structures (grey). Side chain conformations are shown. The model shown has a RMSD of 1.89 Å over the full length of the protein and 1.50 Å over the helical residues. Most side chains in helical-helical interfaces are covered correctly at this resolution.

The benchmark of the assembly step in EM-Fold was performed in two stages – using perfect and realistic secondary structure prediction respectively. The only difference to the protocol published in (*5*) is that a dynamic growing and shrinking of secondary structure elements was implemented in the assembly step. In addition to the moves described in Figure 14, the assembly step allows for deletion and addition of residues at the ends of SSEs. This move is accompanied by scoring the agreement of the model's secondary structure with predicted secondary structure to avoid formation of unlikely secondary structure. Growth and shrinkage of SSEs has the potential to compensate for incorrect secondary structure prediction in a more dynamic way than the SSE pool used before. Inaccurate secondary structure prediction was one of the major reasons for failure of EM-Fold. Apart from the new SSE resize move and score, the assembly step has stayed the same as published in (*5*). Table 5 shows the results of the assembly runs of all 27 proteins both with realistic secondary structure prediction where RMSD100 values are calculated over all backbone atoms. The true topology is successfully identified for 15 of the 20 α-helical proteins and for 4 of the 7 β-sheet proteins. Success is defined by having the correct topology among the top 150 scoring topologies after the assembly step. The overall success rate of 70% is comparable to the 70% reported in (*5*). It is more difficult to predict the correct topology of β-sheet containing proteins. It should also be noted that for some of the proteins (1Z3Y, 2FQ4, 2IU1, 2NR7) the correct topology was still identified in the assembly step but is not scored among the best 150 topologies.

Improved EM-Fold refinement protocol improves RMSDs of models

The top 150 scoring models after the assembly step were refined including bending moves in addition to rotations and translations of the individual SSEs as described in (*5*). The agreement of the model with the density map is scored using a cross correlation score. Table 5 shows the improvement in RMSD100 from the assembly step to the

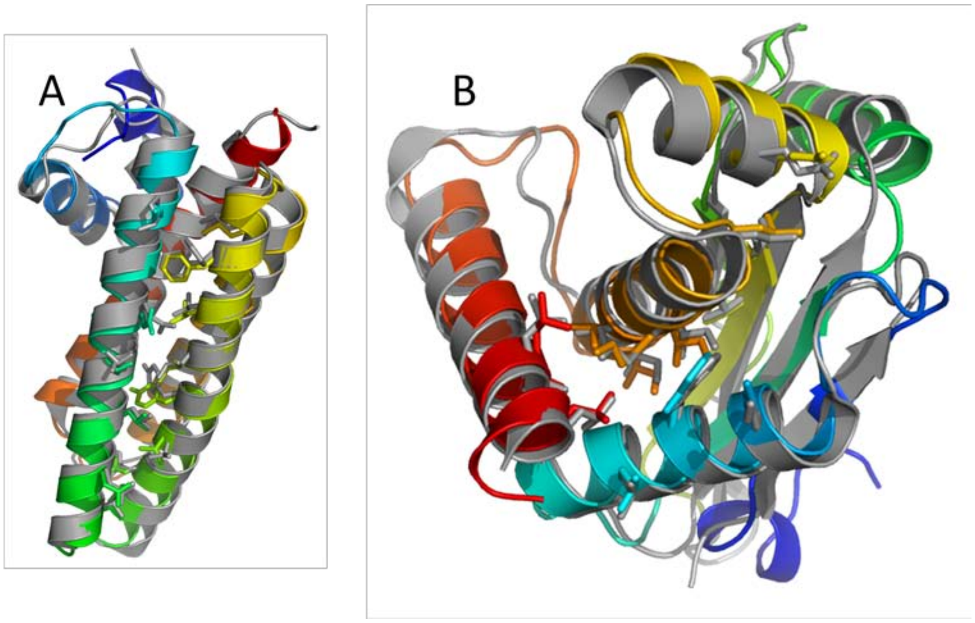

Figure 17. Superimposition of final models after Rosetta refinement with medium sized native structures. Superimposition of the final models (colored in rainbow) of 1X91 (A) and 1OZ9 (B) with the original PDB structures (grey). (A) 1X91 has 153 residues. The model shown has a RMSD of 1.30 Å over the full length of the protein and 0.76 Å over the helical residues. (B) 1OZ9 has 150 residues. The model shown has a RMSD of 1.63 Å over the full length of the protein and 1.19 Å over the helical residues.

refinement step. RMSD100s are calculated over all backbone atoms. For all but one of the 19 successful cases the refinement step generates models that are better in RMSD100 than the model after the assembly step. The maximal improvement was 2.4 Å for 1DVO and the average improvement was 1.0 Å. For the best scoring model after refinement the average improvement in RMSD100 is 0.2 Å, while the best improvement is 2.0 Å (for 1X91). For all but one protein (1CHD, rank 87) the models of the correct topologies are among the top 50 scoring models after the refinement step. Particularly models with score rank after the assembly step worse than 25 are improving their ranking considerably through the refinement. The 50 top scoring topologies after the refinement step are used in the first round of the Rosetta refinement protocol.

Rosetta refinement improves models further and reaches atomic resolution for some of the benchmark cases

An iterative refinement protocol was implemented for the continued refinement in Rosetta. The first round builds loops and side chains for the 50 top scoring topologies from the EM-Fold refinement step and relaxed the resulting models in the Rosetta force field. Regions that agree least with the density map in the best scoring 15 topologies from round 1 were identified. These regions were rebuilt in the second round of the Rosetta refinement followed by a relaxation of the models. Finally the regions with the largest discrepancies to the density map in the top 5 scoring models after round 2 were rebuilt in round 3. Table 5 summarizes the results after each of the three rounds of Rosetta refinement. 14 of the 19 final best scoring models correspond to the correct topology. In the remaining cases the true topology is ranked second in three cases and fourth in the worst case. Rosetta is thus able to identify the correct topology by score whenever a model with a RMSD100 smaller than 2.8 Å was built. This was true for 14 of 19 cases. The RMSD100s of the correct models after the third round of refinement range from 1.3 to 6.9 Å over the full length of the proteins and from 0.8 to 3.8 Å over the secondary structure elements. The average RMSD100 is 3.0 Å over the full length of the protein and 2.2 Å over the SSEs. Thirteen of the proteins have backbone atom RMSD100s of less than 3.0 Å over all residues indicating correct atomic detail. Figure 17 shows two models for 1X91 and 1OZ9 superimposed with the native structure as an example of which model quality is achievable in favorable cases. Side chain conformations in interfaces are shown for both the model and the native structure. At the precision of the models (RMSDs of 1.30 Å and 1.63 Å over all residues respectively) the side chain conformations in interfaces are correctly recovered.

EM-Fold constructs models that display atomic detail not present in the experimental data

Changes to both EM-Fold and Rosetta have been implemented. EM-Fold’s assembly step can dynamically grow and shrink SSEs to offset some of the problems introduced by incorrectly predicted SSEs in the pool. The refinement step can bend secondary structure elements and scores the agreement of the model with the density map directly. More generally, EM-Fold can also assemble and refine proteins containing β-strands if a map of sufficient resolution is available. An electron density agreement score was added to Rosetta. Similar to the density agreement score used in the EM-Fold refinement step, this score uses a cross correlation between the density map and a map generated from the model to evaluate their agreement. Using this score, which proved exceptionally useful when building loops and refining the EM-Fold models, Rosetta scores the correct topology of all but four cases best. This allows identifying native-like models by score in most cases. Almost all of the RMSD vs. score plots show a clear funnel-shape indicating that the Rosetta score correlates with model quality. A combination of the improved EM-Fold and Rosetta refinement yields models that are accurate at atomic detail in 60 to 70% of cases where the correct topology was identified. In summary it can be said that the combination of EM-Fold and Rosetta has become a powerful tool to de novo fold proteins into medium resolution density maps.

Table 5. Results of benchmark on set of 27 α, α/β and β proteins

		Rank / RMSD100 [Å]		Rank / RMSD100 [Å] (RMSD100 SSEs [Å])		
protein	size	EM-Fold assembly	EM-Fold refinement	Rosetta 1	Rosetta 2	Rosetta 3
α-proteins						
1DVO	152, 4, 0	**3 / 4.51**	**19 / 2.72**	1 / 2.49 (1.3)	1 / 2.23 (1.33)	**1 / 2.07 (1.36)**
1GS9	165, 4, 0	**31 / 4.79**	**15 / 5.02**	3 / 3.76 (3.59)	5 / 3.87 (3.69)	4 / 3.96 (3.7)
1IAP	211, 7, 0	**2 / 3.53**	**8 / 2.79**	1 / 2.51 (1.31)	1 / 2.12 (1.27)	**1 / 2.43 (1.28)**
1ILK	151, 5, 0	**73 / 3.72**	**23 / 3.82**	1 / 2.78 (2.6)	1 / 2.75 (2.6)	**1 / 2.64 (2.53)**
1NIG	152, 4, 0	**66 / 7.22**	**22 / 7.77**	4 / 6.04 (4.44)	3 / 5.98 (4.47)	2 / 5.92 (4.37)
1OXJ	173, 4, 0	**71 / 4.08**	**39 / 3.34**	1 / 5.89 (1.5)	1 / 4.34 (1.6)	1 / 4.14 (1.68)
1X91	153, 5, 0	**1 / 4.00**	**1 / 1.61**	1 / 1.37 (0.87)	1 / 1.32 (0.92)	**1 / 1.29 (0.78)**
1Z3Y	238, 7, 0	621 / -	--	-	-	-
2A6B	234, 6, 0	**131 / 4.04**	**24 / 3.64**	1 / 3.9 (1.76)	1 / 3.12 (1.74)	**1 / 2.22 (1.8)**
2FD5	180, 6, 0	**1 / 3.73**	**37 / 2.71**	1 / 1.76 (1.17)	1 / 1.68 (1.11)	**1 / 2.19 (1.64)**
2FM9	215, 9, 0	**126 / 4.27**	**1 / 3.29**	1 / 2.63 (2.29)	1 / 2.29 (2.11)	**1 / 2.29 (2.02)**
2FQ4	192, 7, 0	260 / -	--	-	-	-
2G7S	194, 6, 0	**1 / 3.29**	**29 / 3.35**	1 / 2.04 (1.67)	1 / 2 (1.7)	**1 / 2.01 (1.74)**
2GEN	197, 7, 0	**2 / 3.62**	**18 / 3.70**	4 / 2.67 (2.25)	1 / 2.27 (2.05)	**1 / 2.35 (2.08)**
2IGC	164, 4, 0	**23 / 5.63**	**41 / 7.77**	1 / 7.1 (4.19)	5 / 6.91 (3.81)	2 / 6.93 (3.78)
2IOS	150, 6, 0	**60 / 4.97**	**14 / 3.22**	1 / 3.87 (3.03)	1 / 3.48 (3.18)	1 / 3.31 (3.04)
2IU1	208, 5, 0	849 / -	--	-	-	-
2NR7	195, 5, 0	386 / -	--	-	-	-
2O8P	227, 9, 0	**15 / 3.98**	**18 / 3.90**	2 / 2.35 (2.25)	2 / 2.05 (2.01)	**1 / 2.18 (2.13)**
2QK1	249, 9, 0	- / -	--	-	-	-
α-β-proteins						
1BJ7	156, 1, 8	- / -	--	-	-	-
1CHD	203, 1, 8	**24 / 2.27**	87 / 2.24	2 / 15.76 (1.5)	4 / 15.44 (1.49)	4 / 15.42 (1.5)
1ICX	155, 1, 7	**131 / 2.92**	**47 / 4.68**	1 / 2.51 (2.08)	1 / 2.35 (1.89)	**1 / 2.17 (1.76)**
1JL1	155, 3, 5	**32 / 3.78**	**13 / 4.34**	1 / 3.38 (2.85)	1 / 2.84 (2.03)	**2 / 2.91 (2.3)**
1OZ9	150, 5, 4	**1 / 2.73**	**9 / 2.01**	1 / 1.88 (1.21)	1 / 1.89 (1.41)	**1 / 2.19 (1.85)**
β-proteins						
1WBA	175, 0, 10	- / -	--	-	-	-
2QVK	192, 0, 7	- / -	--	-	-	-
Average RMSD100s		3.19	2.98	3.27	2.97	2.96

This table summarizes the results of the 27 protein benchmark. The first two columns show the pdb ID of the protein and protein size information. Column 2 lists the number of amino acids, number of α-helices with at least 12 residues and the number of β-strands with at least 5 residues. Columns 3 and 4 contain the results of the EM-Fold assembly and refinement step, respectively. The rank of the correct topology model within all scored models as well as the RMSD100 of the correct topology model are given. Columns 5 through 7 show the results of the three rounds of Rosetta refinement. Each of these columns lists the rank of the correct topology model within all scored models, the RMSD100 of the correct topology model as well as the RMSD100 of the correct topology model over residues in secondary structure elements (numbers in paretheses). The last column lists the precent of rotamers in the protein core that were recovered. All RMSD100 values are determined over the backbone atoms N, C_α, C and O. The proteins from the benchmark set that are considered a success after the EM-Fold assembly and refinement steps as well as after the third round of Rosetta refinement are shown in bold. The criteria for the individual success assignments were: correct topology within the top 150 scoring models after the EM-Fold assembly step, correct topology within the top 50 scoring models after the EM-Fold refinement step and correct topology being the top scoring models with an RMSD100 of less than 3Å after the third round of Rosetta refinement.

CHAPTER IV

STRUCTURE OF THE ADENOVIRUS-INTEGRIN-COMPLEX

This chapter is based on publication (*96*).

Introduction

A growing number of viruses have been identified as using one of the 24 types of integrin heterodimers as a receptor for cell entry (*97*). Integrins are cell surface molecules involved in the regulation of adhesion, migration, growth, and differentiation (*98*). The large multi-domained extracellular segments of α and β integrin subunits bind a variety of ligands, including viral ligands, while the smaller intracellular domains interact with cytoskeletal proteins (Figure 18A). These extracellular and intracellular interactions facilitate bidirectional signaling, with the initiating events occurring either outside of the cell (outside-in signaling), or within the cell (inside-out signaling) (*99*). Integrin clustering has been established as having an important role in outside-in signaling (*100-103*). Clustering results in the formation of focal adhesions, which are organized intracellular complexes, that facilitate downstream signaling cascades within the cell (*99*).

Studies of adenovirus (Ad) interactions with αv integrins provided some of the first evidence of the viral induced signaling events (*104, 105*). The Ad penton base capsid protein, which sits at the twelve vertices of the icosahedral capsid, has five prominent Arg-Gly-Asp (RGD) containing loops that are flexible and protrude from the viral surface (*10, 106*). Receptor mediated endocytosis of Ad is stimulated by interaction of the RGD-containing penton base with αvβ3 and αvβ5 integrins (*107*). This interaction leads to receptor clustering followed by tyrosine phosphorylation/activation of focal adhesion kinase (FAK), as well as activation of p130CAS, PI3K and the Rho family of small GTPases, and subsequent actin polymerization and Ad internalization (*97*). Integrin signaling events also lead to production of proinflammatory cytokines (*108*) and may result in increased survival of certain host cells through subsequent

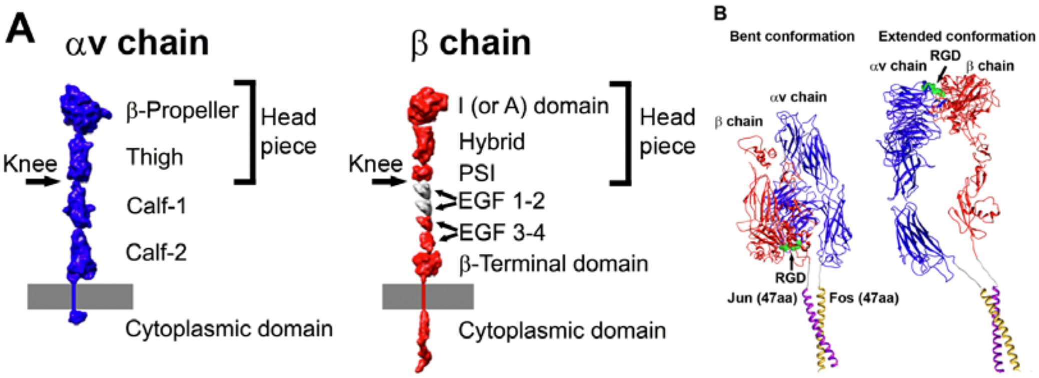

Figure 18. Integrin domains and conformations.
(A) Structural domains of integrin αv and β chains, including the extracellular domains, transmembrane spanning regions, and small cytoplasmic domains, shown in extended schematic forms. The domains are represented as 10Å-resolution density maps based on crystallographic coordinates. The membrane is represented by a gray bar. Modified from Trends in Microbiology, P.L. Stewart and G.R. Nemerow, "Cell Integrins: Commonly Used Receptors for Diverse Viral Pathogens", 15:500-507 (2007) and reprinted with permission from Elsevier. (B) Models for soluble αvβ5 integrin with Fos/Jun dimerization domains. Each chain has a six residue Glycine-rich linker between the ectodomain and the Fos or Jun dimerization domain. The model of a bent integrin conformation (left) was built as a composite of αvβ3 integrin crystal structures, PDB-IDs 1L5G and 1U8C (*2, 3*), and the crystal structure of c-Fos/c-Jun bound to DNA, PDB-ID 1FOS (*12*). The model of an extended integrin conformation (right) is similar to the extended model docked into the HAdV12/αvβ5 cryo structure. Reproduced with permission from ASM.

signaling to protein kinase B (AKT) (*109*).

Multiple studies indicate that after interaction with an RGD-containing ligand a straightening of the integrin extracellular domains occurs, leading to the “extension” or “switchblade” model for integrin activation (*110*, *111*). In the extension model the headpiece domains, which are closest to the RGD interaction site, have a “closed” conformation in the low affinity, unliganded state. This state is characterized by close proximity of the α and β subunits at the “knees” or midpoints of the extracellular segments. In contrast, the high affinity, ligand-bound state in the extension model is distinguished by an “open” headpiece conformation with separation at the knees of the extracellular segments. The location of the RGD binding site between the α-subunit β-propellor and the β-subunit I-domain was first visualized in the crystal structure of the αvβ3 extracellular segment with a bound RGD peptide (*2*). In this structure the RGD site is folded back toward the membrane and the integrin is in a closed conformation. The closed conformation has also been observed in crystal structures of the αvβ3 ectodomain without an RGD peptide (*112*) and the $\alpha_{IIb}\beta3$ ectodomain (*113*).

The open integrin conformation has been characterized as having a large separation of up to ~70Å between the knees of α and β subunits (*110*). Four slightly different open headpiece conformations were observed in crystal structures of the $\alpha_{IIb}\beta3$ headpiece with bound fibrinogen-mimetic therapeutics (*11*). These structures show that the change from a closed to an open headpiece conformation is accompanied by a piston-like motion of helix α7 in the β-chain I-domain, a large swing of the β-chain hybrid domain of up to 69°, as well as extension and separation of the two integrin chains. Comparison of the available αvβ3 and $\alpha_{IIb}\beta3$ crystal structures is providing information on the interdomain angle variation and flexibility between domains (*113*).

One aspect of the extension model is that separation of the C-terminal, intracellular portions of the α and β subunits leads to inside-out activation. This concept is supported by NMR structures of the cytoplasmic tails of $\alpha_{IIb}\beta3$ showing that the membrane-proximal helices engage in a weak interaction that can be disrupted by constitutively activating mutations or by talin, a protein found in high concentrations in focal adhesions (*114*). The concept that the integrin α and β-subunits must also separate during outside-in signaling is supported by a study involving a disulfide-bonded mutant of $\alpha_{IIb}\beta3$ integrin (*115*). When the α and β subunits are linked in the vicinity of the transmembrane helices the mutant $\alpha_{IIb}\beta3$ is still able to bind ligand, mediate adhesion, and undergo antibody-induced clustering. However, the disulfide-bonded mutant exhibits defects in focal adhesion formation and FAK activation. Reduction of the disulfide bond or single cysteine mutants rescues signaling.

A competing model for integrin activation, called the "deadbolt" model, proposes only small conformational changes in the integrin β-chain I-domain upon RGD binding (*116*). This model is based on crystal structures of the αvβ3 ectodomain with and without an RGD peptide (*2, 112*). Both of these αvβ3 structures reveal a bent integrin conformation with a closed headpiece conformation. However, the RGD peptide was soaked into a preformed crystal of αvβ3 and crystal contacts may have prevented conformational changes.

There are relatively few and only moderate resolution structures of virus-integrin complexes. A moderate resolution cryoEM structure has been determined for the Picornavirus echovirus-1 (EV1) in complex with the I-domain of the α2 integrin subunit (*117*). Docking of crystal structures of EV1 and the α2 I-domain into the cryoEM density indicates that the I-domain binds within a canyon on the surface of EV1 and that five integrins could potentially bind at one vertex of the icosahedral capsid. Confocal fluorescence microscopy experiments indicated that EV1 causes integrin clustering on human osteosarcoma cells stably transfected with α2 integrin.

However, it could not be determined whether the bound integrins were in the inactive (bent) or active (extended) conformation.

Moderate resolution (~21Å) cryoEM structures of Ad type 2 (HAdV2) and HAdV12 in complex with a soluble form of αvβ5 integrin revealed a ring of integrin density over each penton base capsid protein (*118*). Better defined integrin density was observed in the HAdV12/integrin complex, supporting the idea suggested from sequence alignments that the RGD loop of the HAdV12 penton base is shorter and less flexible than that of HAdV2. This study also suggested that the precise spatial arrangement of the five Arg-Gly-Asp (RGD) protrusions on the penton base might promote integrin clustering, which may lead to the intracellular signaling events required for virus internalization into a host cell. A similar spacing of RGD-containing integrin-binding sites around the fivefold axis of icosahedral virions has been noted for Ad, foot-and-mouth disease virus (FMDV), and coxsackievirus A9 (CAV9) (*97*).

Here we present a significantly higher resolution cryoEM structure of HAdV12 complexed with soluble αvβ5 that provides insight into the Ad/integrin interaction. The resolution of the icosahedral capsid portion of the Ad/integrin complex was improved to 8Å and the capsid shows clearly resolved α-helices, which allows accurate docking of the penton base crystal structure within the cryoEM density. The resolution of the integrin density is more moderate due to flexibility of the RGD-containing surface loop of penton base and incoherent averaging of integrin heterodimers. Nevertheless, modeling studies with available integrin crystal structures have enabled us to distinguish between a bent or extended conformation (Figure 18B) when αvβ5 binds to the multivalent ligand presented by the Ad penton base. The cryoEM structural analysis also indicates that integrin induces a conformational change in penton base.

Materials and Methods

Preparation and isolation of HAdV12 and soluble αvβ5 integrin

The integrin used for the cryoEM study was a soluble form of recombinant αvβ5 produced in insect cells as previously described (*119*). This recombinant form of αvβ5 has an αv chain formed by the complete αv ectodomain followed by a glycine linker and the Fos dimerization domain. The β5 chain is formed by the entire β5 ectodomain followed by a glycine linker and the Jun dimerization domain.

HAd12V (ATCC) was propagated in 293 cells and purified by two cycles of cesium chloride density gradient ultracentrifugation as previously described (*120*). The banded virus was then dialyzed against Tris buffer at pH 8.0 (50mM Tris, 130mM NaCl, and 3mM KCl).

Cryoelectron microscopy

The HAdV12 sample (250 μg/ml) and a soluble recombinant form of αvβ5 (*119*) (309 μg/ml) were incubated for two hours in a Tris buffer (pH 8.1, 1mM Ca^{2+}, 1mM Mg^{2+}). The sample was prepared to contain about five integrin molecules per RGD binding site. CryoEM grids were prepared as described by Saban *et al.* (*121*) applying 6 µl of sample to each Quantifoil R2/4 holey carbon grid (Quantifoil Micro Tools GmbH). The Vitrobot cryo-fixation device (FEI Company) was used for flash freezing of the grids in a cryogen. An FEI Polara (300 kV; FEG) transmission cryo-electron microscope operated at liquid nitrogen temperature and 300 kV with a Gatan UltraScan 4kx4k charge-coupled device (CCD) camera was used for data acquisition with SAM, a semiautomatic data collection routine (*122*). The absolute magnification for the data was ~398kX.

Image processing

The automatic selection program VIRUS (*123*) was used to extract particle images from the cryoelectron micrographs. The initial box size of the particle images was chosen to include ~500Å beyond the edge of the HAdV12 capsid in order to visualize the bound integrin. An in-house script was used to bin the particle images to three sizes for image processing: 320^2 pixels (4.8Å pixels), 640^2 pixels (3.2Å pixels), and 960^2 pixels (1.6Å pixels). The program CTFFIND3 (*124*) was used to determine initial estimates for the microscope defocus and astigmatism parameters. The cryoEM structure of Ad35f (*7*) filtered to 12Å resolution was used as the starting model for refinement with FREALIGN (*125*). A total dataset of 2,499 particle images was refined with first the coarsest pixel size (4.8Å), then with the intermediate pixel size (3.2Å), and lastly with the finest pixel size (1.6Å). In the initial refinement rounds a modified version of FREALIGN was used to allow the input of externally determined particle centers (*7*). The orientational parameters (two translational parameters and three Euler angles), as well as the defocus, astigmatism, and magnification parameters, were refined on a per particle basis. The final reconstruction included 1,141 particle images, selected on the basis of their FREALIGN phase residual parameter. The resolution of the final reconstruction was assessed using the Fourier shell correlation (FSC) curve calculated by FREALIGN. Maps were calculated for three radial shells: 300-463 Å, corresponding to the icosahedral capsid; 463-515 Å for the proximal integrin density; and 515-700 Å for the distal integrin density.

A temperature factor (B= -300Å^2 or B= -500Å^2 as specified) and cosine edge filtering was applied to the final density maps with the BFACTOR program (http://emlab.rose2.brandeis.edu/grigorieff/download_b.html). The graphics figures were produced with UCSF Chimera (*126*).

The CoLoRes routine from SITUS (*127*) was used to refine the absolute pixel size of the cryoEM density map to within 0.05Å. The absolute pixel size was taken as the value that gave the maximal cross correlation coefficient between the x-ray crystal structure of penton base (PDB-ID 1X9T) (*10*) and a selected portion of the cryoEM density map including one penton base pentamer.

De novo model building for penton base RGD loop

Alignment of penton base sequences from multiple human Ad serotypes and comparison with the HAdV2 penton base crystal structure (*10*) indicates that the HAdV12 penton base is likely to have a flexible and extended RGD-containing loop of 15 amino acids. Within this loop there are 6 amino acids on either side of the RGD residues. The crystal structure of HAdV2 penton base (PDB-ID 1X9T) (*10*) does not contain coordinates for the RGD loop. A Rosetta loop building protocol was used (*14*) to build models for the RGD loop in HAdV12 (residues 296-312). A total of 25 loop models were built. Ten of the 25 models were selected for a subset that is representative of the maximum variation within the set of 25 models.

Integrin docking within the HAdV12/αvβ5 integrin cryoEM density

Given the high homology between the β3 and β5 integrin chains (identity 55%, similarity 70%, gaps 4%), the crystal structures of αvβ3 (*3, 112*) and the αvβ3/RGD peptide complex (*2*) are reasonable for comparison with the integrin density in the HAdV12/αvβ5 complex. There is also significant homology between the αv and α_{IIb} integrin chains (identity 30%, similarity 41%, gaps 4%). Therefore we have also compared the cryoEM density with the platelet integrin $\alpha_{IIb}\beta3$ crystal structures complexed with fibrinogen-mimetic therapeutics, which display open headpiece conformations (*11*).

A density section encompassing the vertex region of the HAdV12/αvβ5 integrin cryoEM structure was extracted for docking with atomic resolution integrin structures. Two integrin domains (the β-propellor of the αv chain, and the I-domain of the β3 chain) together with RGD peptide were selected from the crystal structure of the extracellular segment of integrin αvβ3 in complex with an RGD ligand (PDB-ID 1L5G) (*2*) for initial docking experiments. Interactive docking was performed with UCSF Chimera (*126*). The two αvβ3 integrin domains and RGD peptide were docked as a rigid body into the cryoEM integrin density.

In a second phase of docking experiments four conformations of the β3 integrin chain in crystal structures of $\alpha_{IIb}\beta3$ complexes with fibrinogen-mimetic therapeutics (PDB-IDs 2VDK and 1TYE) (*11*) were evaluated in the context of the HAdV12/αvβ5 cryoEM integrin density. Each of the four β3 conformations has a different angle (57°, 59°, 61°, or 69°) between the β chain I and hybrid domains. The I-domain of each of these conformers was aligned with the previously docked I-domain from the αvβ3/RGD crystal structure using the UCSF Chimera Matchmaker tool.

A model for the complete αvβ5 integrin ectodomain in a bent conformation was built as a composite of crystal structures of αvβ3 (PDB-ID 18UC) (*3*) and the αvβ3/RGD peptide complex (PDB-ID 1L5G) (*2*). A model for the complete integrin ectodomain in an extended conformation was built as a composite of crystal structures of the αvβ3/RGD peptide complex and an $\alpha_{IIb}\beta3$/ligand mimetic complex (PDB-ID 2VDK) (*11*). The α-chain domains calf-1 and calf-2 and the β-chain domains EGF1-EGF4 and terminal domain in the extended model were taken from an αvβ3 crystal structure (PDB-ID 1U8C) and roughly positioned into the cryoEM density.

Results

CryoEM structure of HAdV12/αvβ5 complex indicates disorder for the bound integrin

HAdV12/αvβ5 integrin complexes were formed with a soluble, recombinant form of αvβ5 (*119*). This form of αvβ5 retains the ability to recognize the Ad penton base as well as vitronectin, an

RGD-containing extracellular matrix protein. A kinetic analysis of the interaction between HAdV2 virions with this soluble form of αvβ5 indicated ~4.2 integrin molecules bound per penton base at close to saturation (*118*). The first cryoEM study of the HAdV12/αvβ5 complex reached only moderate (~21Å) resolution (*118*), and at that time there were no atomic resolution structures of αv integrins available for docking into the cryoEM density. We have performed additional cryoEM analyses of HAdV12/αvβ5, pushed the resolution beyond 10Å for the icosahedral capsid, and analyzed the resulting cryoEM density with the available αvβ3 and $\alpha_{IIb}\beta3$ crystal structures.

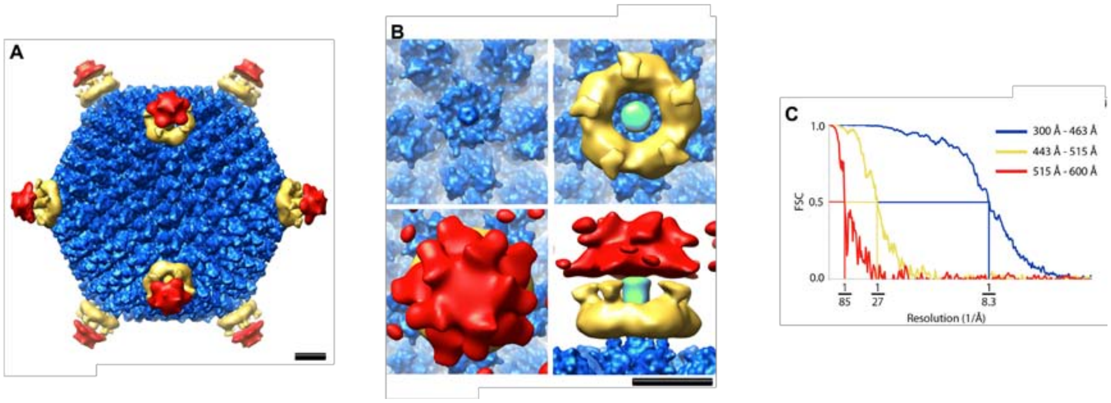

Figure 19. CryoEM structure of the HAdV12/αvβ5 integrin complex.
(A) Full structure viewed along a 2-fold icosahedral axis and shown as three radial shells: icosahedral capsid (300-463Å, blue); integrin ring and fiber (443-515Å, gold); and diffuse integrin density (515-600Å, red). The icosahedral capsid shell is shown filtered to 8Å with an applied B-factor of -300Å^2; the integrin ring and fiber filtered to 19Å with an applied B-factor of -500Å^2; and the diffuse integrin density filtered to 33Å with an applied B-factor of -500Å^2. (B) Enlarged top and side views of the vertex region colored as in (A) with the fiber shaft in green. The diffuse integrin density (red) is shown with a lower isosurface contour level in panel B (1.2σ *vs.* 1.7σ in panel A) to display a fuller extent of this density. Scale bars, 100Å. (C) An FSC plot indicating a resolution range for the HAdV12/αvβ5 icosahedral capsid (radial shell 300-463Å) of 8.3Å to 6.9Å (8.3Å FSC 0.5; 7.5Å FSC 0.3; 6.9Å FSC 0.143) (blue); the HAdV12/αvβ5 radial shell (443-515Å) including the integrin ring and fiber above the penton base of 27Å to 19Å (27Å FSC 0.5; 24Å FSC 0.3; 19Å FSC 0.143) (gold); and the HAdV12/αvβ5 radial shell (515-600Å) with diffuse integrin density of 85Å to 33Å (85Å FSC 0.5; 50Å FSC 0.3; 33Å FSC 0.143) (red). Reproduced with permission from ASM.

A cryoEM dataset with 2,499 particle images of HAdV12/αvβ5 complexes was acquired on an FEI Polara (300kV, FEG) microscope, which allowed significant improvement in the resolution. The final cryoEM structure presented in Figure 19 is based on a subset of ~45% of the dataset (1,141 particle images) selected to yield the highest resolution for the icosahedral capsid. The resolution was assessed in three radial shells corresponding to the icosahedral capsid (radii 300-463Å), the proximal integrin density (radii 443-515Å), and the more distal diffuse integrin density extending away from the virion (radii 515-600Å). Overlapping radial shells were selected for the capsid and proximal integrin density, since the integrin extends over the Arg-Gly-Asp (RGD) containing protrusions of the penton base.

The resolution of the icosahedral HAdV12 capsid is 8.3Å (FSC 0.5 threshold) and, as expected for this resolution, α-helices within the capsid proteins including penton base and protein llla are clearly resolved as density rods (Figure 20B). A ring of integrin density is observed over each penton base with well defined connector regions stretching away from the capsid surface and surrounding the fiber shaft (Figure 19B, gold density). The resolution of the radial shell with the integrin ring is 27Å (FSC 0.5), which is significantly worse than that of the capsid indicating static or dynamic disorder for the bound integrin. Although the integrin ring displays 5-fold symmetry and gives the impression that there are five copies of integrin bound to the penton base, icosahedral symmetry has been imposed during the reconstruction process. In the outermost radial shell the resolution is 85Å (FSC 0.5). This shell has weak and diffuse density over the vertices (Figure 19B, red density) indicating that the disorder of the bound integrin increases with greater distance from the capsid.

When the resolution is assessed over all radial shells (radii 0-700Å) the overall resolution of the HAdV12/αvβ5 structure is 10Å (FSC 0.5). This is significantly better than the resolution previously reported for the same complex (*118*).

Steric hindrance limits the number of integrins bound per penton base

The crystal structure the HAdV2 penton base (*10*) fits well within the cryoEM density for the HAdV12 penton base in the HAdV12/αvβ5 complex (Figure 20A and B). The HAdV2 penton base has a high degree of homology (65% identity, 71% similarity) with the HAdV12 penton base and thus serves as a reasonable model for analysis of the cryoEM density. The biggest difference between the two sequences is that the HAdV12 RGD loop is much shorter, with just 15aa in the HAdV12 loop corresponding to 78aa in the HAdV2 penton base RGD loop. The longer HAdV2 RGD loop is missing from the crystal structure due to disorder. Presumably the HAdV12 RGD loop is also flexible. In the absence of a crystal structure for the HAdV12 penton base, we used the Rosetta loop building protocol to model the HAdV12 RGD loop. Rosetta was first developed for *de novo* protein structure prediction and has been extended to model loops and protein segments. In a benchmark study it was shown that segments of size 13-18aa can be modeled with accuracies of 1 to 7Å root-mean-square deviation (RMSD) (*14*). Rosetta loop building produced 25 models for the HAdV12 RGD loop all with approximately the same score and with varying conformations (Figure 20C and D). We interpret this result to indicate that it is possible for the HAdV12 RGD loop to extend in a variety of directions from the top of the penton base.

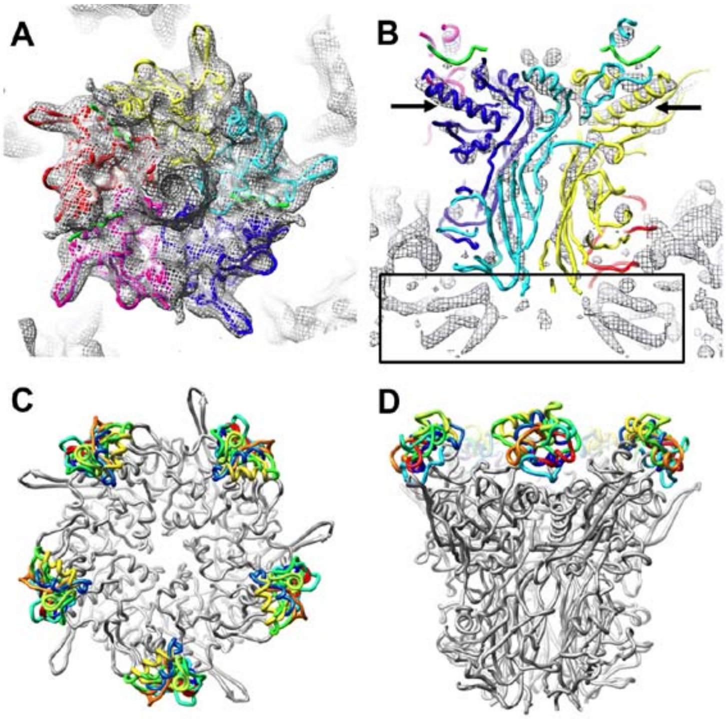

Figure 20. Penton base in the HAdV12/αvβ5 integrin complex and models for the integrin-binding RGD loop.
(A) Top view of the HAdV12 penton base cryoEM density (mesh) with the docked crystal structure of the HAdV2 penton base pentamer and fiber peptide (PDB-ID 1X9T) (*10*). (B) Side view of the same but with the cryoEM density shown at a higher isosurface contour level to reveal α-helical density rods. The arrows indicate the longest α-helix in the penton base monomer. The rectangle indicates the α-helical density below the penton base assigned to protein IIIa (*7*). Each penton base monomer is in a different color with the fiber peptide at the top of the penton base in green. (C and D) Top and side views of the HAdV2 penton base crystal structure (gray) with ten different models for the HAdV12 RGD loop (aa 296-312) built with Rosetta (*14*). Reproduced with permission from ASM.

The crystal structure of the αvβ3 extracellular segment with a bound RGD peptide reveals the RGD binding site between the αv-chain β-propellor domain and the β-chain I-domain (*2*). Manual docking of these two integrin domains within the integrin ring above the penton base shows a basic level of agreement between the size of these domains and the dimensions of the ring in the HAdV12/αvβ5 structure (Figure 19B and Figure 21). A comparison of the interdomain angles in the various integrin crystal structures indicates little variability (~3°) between the α-chain β-propellor and β-chain I-domain (*113*). Therefore this 2-domain integrin unit (β-propellor/I-domain/RGD-peptide) from the αvβ3/RGD crystal structure (*2*) was docked as a rigid body within the cryoEM density. The HAdV2 penton base crystal structure was simultaneously docked in the capsid density. During the integrin docking performed with UCSF Chimera (*126*), distances were monitored between the residues on either side of the RGD-loop gap in the penton base crystal structure with the Arg and Asp residues of the RGD peptide within the αvβ3/RGD crystal structure. There are 7 residues on either side of the central G residue in the RGD sequence of the HAdV12 loop. Our goal was to keep the monitored distances less than 25Å, corresponding to the maximum span of 7 fully extended amino acid residues (*128*).

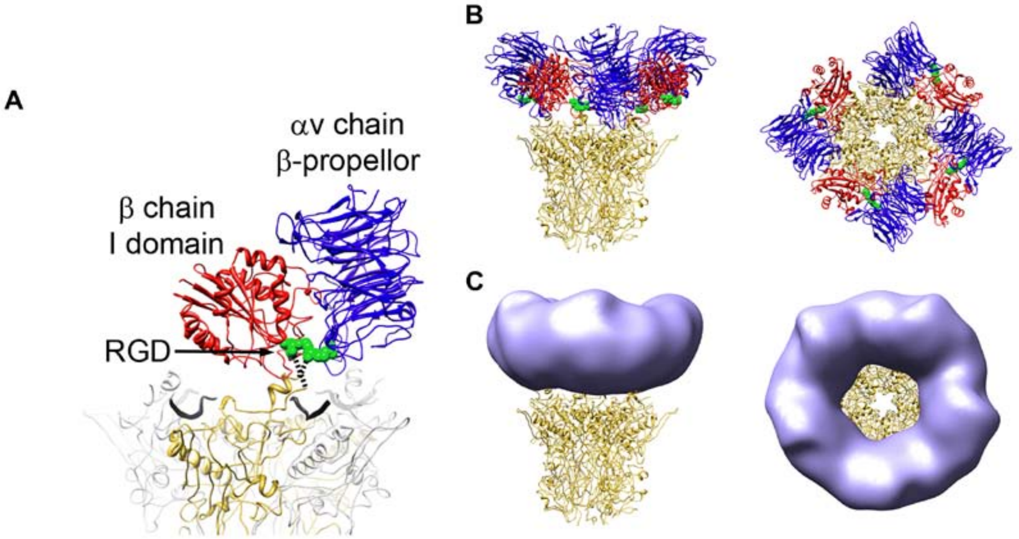

Figure 21. The RGD-binding integrin domains form the ring of density over the penton base in the HAdV12/αvβ5 structure.
(A) The two RGD-binding integrin domains, αv chain β-propellor (blue) and β chain I domain (red), together with the RGD residues (green) from the αvβ3/RGD crystal structure (PDB-ID 1L5G) are shown modeled over the HAdV2 penton base crystal structure (PDB-ID 1X9T). One monomer of penton base is shown in gold, the rest in gray, and the fiber peptides as wide black ribbons. The missing residues in the penton base RGD loop are represented by dashed lines. (B) The penton base is shown with four docked copies of the 2-domain integrin unit. (C) The penton base is shown with simulated density (light blue) generated from the integrin models in panel B with 5-fold averaging and filtered to 27Å resolution. Reproduced with permission from ASM.

Although previous cryoEM analyses of Ad-integrin complexes at lower resolution suggested the possibility that up to five integrins could bind to each penton base, our present high resolution analysis reveals that only four β-propellor/I-domain/RGD-peptide units can be docked without clashes between neighboring integrins (Figure 21B). Three of the four integrin units can be docked without exceeding the RGD-loop distance constraint, and the fourth integrin unit only seems to fit with an apparent violation in the RGD-loop constraint (with distances of ~30Å). If one assumes that small conformational changes at the top of the penton base monomer are possible, a fourth integrin could be bound at the top of the penton base. After docking four integrin β-propellor/I-domain/RGD-peptide units over the penton base, the fifth RGD protrusion is effectively shielded by the bound integrin.

The cryoEM density for the integrin ring appears to have 5-fold symmetry, but this may only be due to the fact that icosahedral symmetry has been imposed on the cryoEM structure. We tried to saturate the RGD binding sites by adding a 5-fold excess of integrin molecules when preparing the HAdV12/αvβ5 complex, but there may be penton bases with fewer than the maximum number of integrin molecules bound. The cryoEM structure represents an icosahedrally averaged representation of the particle images included in the final structure. Therefore the cryoEM density alone doesn't indicate the number of integrin molecules bound per penton base. However, docking of integrin crystal structures into the cryoEM density indicates that a maximum of four integrins can bind to one HAdV12 penton base. This is consistent with the previous Biacore measurement of ~4.2 integrin molecules bound per HAdV2 penton base at close to saturation (*118*).

Integrin molecules bind in various orientations with respect to the penton base

During the docking procedure it became apparent that the close spacing of the RGD protrusions on the penton base relative to the size of the 2-domain integrin unit precludes the possibility of integrin binding in the same orientation to each of the neighboring RGD sites. All four docked integrin units are approximately contained within the cryoEM envelope for the integrin ring, but they are each in a different orientation relative to underlying penton base monomer. Incoherent averaging of different integrin orientations relative to the icosahedral HAdV12 capsid, would explain the markedly different resolutions (8.3Å *vs.* 27Å) between the capsid and the radial shell with the integrin ring. The integrin ring is only ~40Å above the capsid surface at its midpoint, and thus it must be bound in an asymmetric or highly flexible manner in order to explain the dramatic fall off in resolution from the icosahedral capsid. To test the effect of incoherent averaging, coordinates were saved for the four docked 2-domain integrin units and converted to a 5-fold averaged density map (Figure 21C). This resulted in a density ring that strongly resembles that observed above the penton base in the cryoEM structure (Figure 19B).

The adjacent integrin domains, αv thigh and β-chain hybrid and PSI, were included in the analysis in an attempt to model the connector regions extending radially outward above the integrin ring in the cryoEM structure (Figure 22). There is much greater variability observed in crystal structures for the interdomain angle between the β-chain I and hybrid domains (0° to 70°) than between the α-chain β-propellor and thigh domains (up to 10°) (*113*). For the docking analysis we built composite integrin models from the αvβ3/RGD crystal structure (*2*) and the $\alpha_{IIb}\beta3$/ligand crystal structures with various swing angles for the β-chain hybrid domain (*11*). Although the individual integrin domains are not resolved in the cryoEM HAdV12/αvβ5 structure, composite models with an angle between the β-chain I and hybrid domains of 57° to 69° fit the best. With the adjacent integrin domains included with a 69° I/hybrid domain angle the 5-fold averaged density map shows connectors similar to those in the cryoEM structure (Figure 19B and Figure 22C).

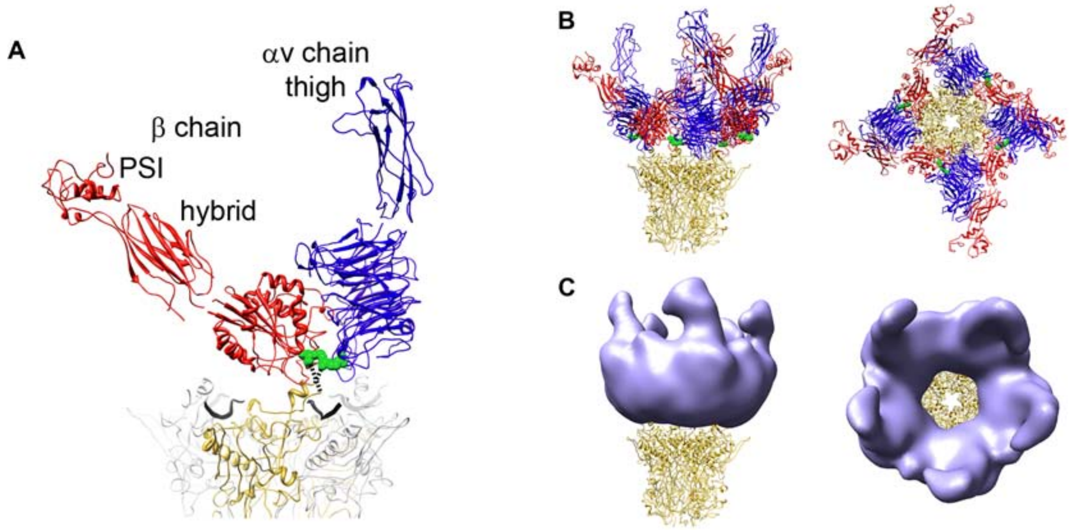

Figure 22. The integrin headpiece forms the ring and connector regions stretching away from the penton base in the HAdV12/αvβ5 structure.
(A) The additional integrin headpiece domains are labeled. The RGD-binding integrins domains are modeled above the penton base as in Figure 4. This is a composite integrin headpiece model built from PDB-IDs 1L5G and chain B of 2VDK with a 69° angle between the β-chain I and hybrid domains (*11*). (B) The penton base is shown with four docked copies of the integrin headpiece. (C) The penton base is shown with simulated density (light blue) generated from the integrin models in panel B with 5-fold averaging and filtered to 27Å resolution. Reproduced with permission from ASM.

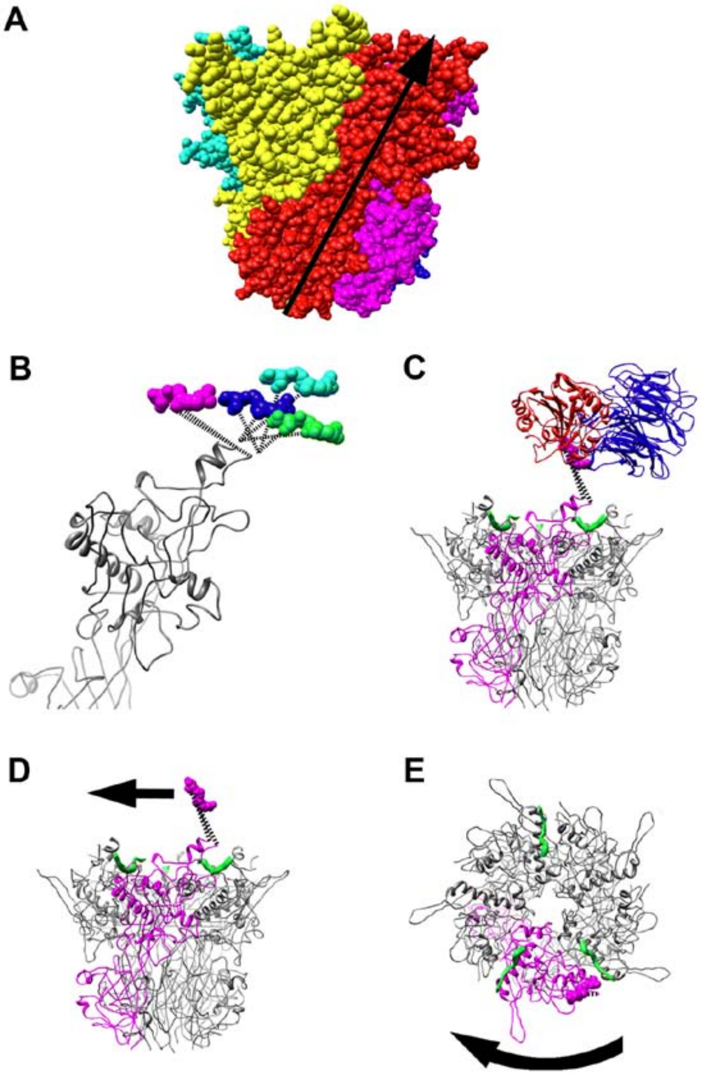

Figure 23. Natural twist of penton base and possible untwisting by integrin.
(A) Space filling representation of the penton base pentamer (PDB-ID 1X9T) with each subunit in a different color. The natural twist of one subunit from the bottom of the pentamer to the top solvent accessible surface in the virion is represented by the arrow. (B) Side view of four superimposed penton base monomers with four copies of the RGD residues (magenta, cyan, blue and green) as modeled in Figures 4 and 5. Note that the magenta copy of the RGD residues extends the RGD loop counter to the natural twist of the penton base monomer. (B) Side view of the penton base pentamer shown with one monomer in magenta, fiber peptides in green, and the modeled RGD-binding integrin domains (blue and red) with the magenta copy of the RGD residues. (C) The integrin domains are removed and an arrow indicates the direction that an integrin bound in the manner of panel B would tend to extend the RGD loop of penton base and lead to untwisting. (D) Top view of the penton base pentamer with a curved arrow indicating clockwise untwisting of the penton base. Reproduced with permission from ASM.

In order to estimate the degree of incoherent averaging that would be caused by four integrins bound as in our integrin/penton base model; we aligned the penton base monomers and measured the distances between all pairs of RGD peptides (Figure 23B). These distances vary from 10Å to 29Å. This large variation would cause a significant reduction, presumably on the order of 10Å, in the resolution of the integrin ring compared to the icosahedral capsid.

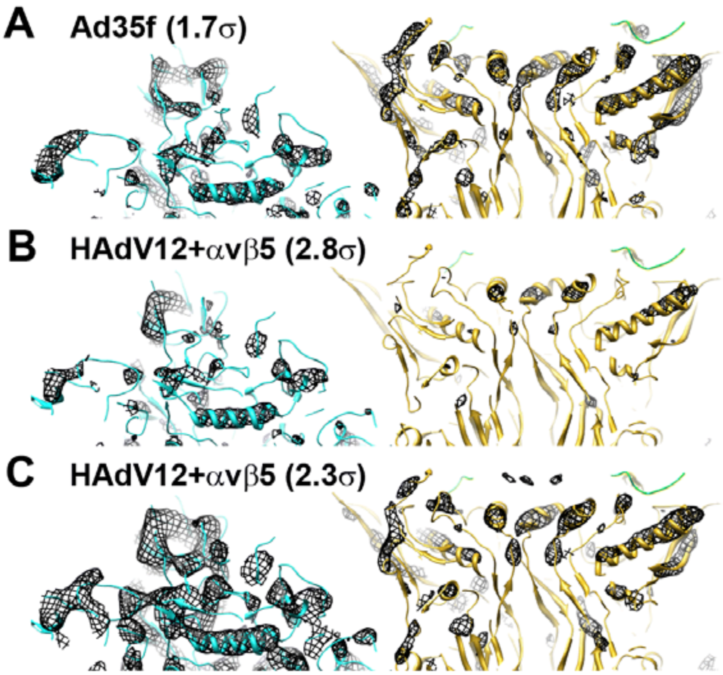

Figure 24. The penton base has low partial occupancy in the HAdV12/αvβ5 cryoEM structure.
(A) A thin (~20Å) slab of the Ad35f cryoEM structure (*7*) with docked crystal structures of hexon (cyan) (PDB-ID 1P30) and penton base (gold) with the N-terminal fiber peptide (green). The cryoEM density (mesh) is isocontoured at 1.7 σ and shows rods for α-helices in both hexon and penton base. (B) A thin slab of the HAdV12/αvβ5 cryoEM structure with the same docked atomic resolution structures. The cryoEM density is isocontoured at 2.8 σ and shows rods for α-helices in hexon and almost no density for the penton base. (C) The same as panel B but with the HAdV12/αvβ5 cryoEM density isocontoured at 2.3 σ showing rods for α-helices in penton base similar to those shown for Ad35f in panel A, but with significantly more hexon density than in panel A. Both maps are shown filtered to 7.5Å with the same applied B-factor (-300Å^2). Reproduced with permission from ASM.

Evidence of conformational change in penton base

One of the docked integrin positions effectively pulls the RGD loop in a direction that is counter to the natural twist of the pentamer (Figure 23). The HAdV2 penton base crystal structure shows that each monomer has two domains, with a right-handed twist between the lower and upper domains (Figure 23A) (*10*). When viewed from the top this appears as a counter-clockwise rotation from the lower to the upper domain. We propose based on our model of the penton base/integrin interaction that when 4 integrins are bound a clockwise rotation is induced in one penton base monomer (Figure 23D and E). This rotation direction is counter to the natural twist of the pentamer. A conformational change of this sort could lead to a reduced number of intermolecular contacts between the penton base and the neighboring hexons in the capsid and to

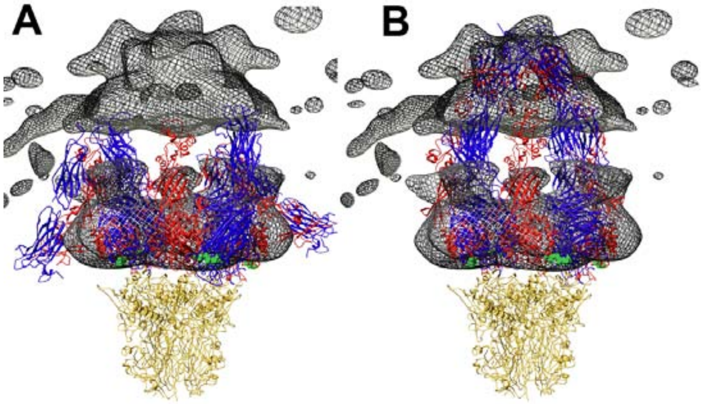

Figure 25. Comparison of HAdV12/αvβ5 integrin density with bent and extended integrin models. (A) The cryoEM integrin density (mesh) is shown with four copies of the integrin ectodomain in a bent conformation. This is a composite integrin model built from PDB-IDs 1L5G (*2*) and 18UC (*3*). (B) The penton base is shown with four copies of the integrin ectodomain in an extended conformation. This is a composite integrin model built from the domains shown in Figure 5 with the remaining domains modeled to approximate the cryoEM density in the outermost radial shell (515-600Å) of the HAdV12/αvβ5 structure. In both panels the αv chains are in blue, the β chains in red, the bound RGD peptides in green, and the penton base in gold. Note that the gap between the upper and lower cryoEM integrin density regions is an artifact from calculating the density in separate radial shells. Reproduced with permission from ASM.

release of the penton base from the capsid. The penton base is known to be released along with fiber and protein IIIa in heat denaturation assays designed to mimic cell entry (*129*).

Close examination of the capsid density in the cryoEM HAdV12/αvβ5 structure and comparison with the cryoEM structure of the Ad35f vector (*7*) suggests that a small percentage of the penton bases and the associated copies of fiber and protein IIIa have been lost in the HAdV12/αvβ5 sample. When the Ad35f cryoEM structure is displayed with an isocontour level of 1.7σ, α-helices within both hexon and penton base are observed as density rods (Figure 24A). When the HAdV12/αvβ5 structure is isocontoured to display the hexon α-helices as rods (at 2.8σ), the penton base density is weak (Figure 24B). Alternatively, when the HAdV12/αvβ5 structure is isocontoured to display penton base α-helices as rods (at 2.3σ), the hexon density includes significantly more than just α-helical density rods. These results indicate that some percentage of the penton bases are missing in the HAdV12/αvβ5 particle images that contribute to the cryoEM structure. The hexon density in the HAdV12/αvβ5 structure also appears stronger than the protein IIIa density underneath the penton base. We speculate that integrin binding to Ad might predispose the penton base and protein IIIa to release from the capsid, even in the absence of the host cell membrane and the low pH environment of the endosome. The loss of penton base and protein llla at the vertex region is consistent with our model of induced conformational changes by integrin engagement.

CryoEM density is consistent with an extended integrin conformation

The cryoEM structure of the HAdV12/αvβ5 integrin complex also provides information on the overall conformation of the extracellular portion of αvβ5 integrin in complex with a multivalent ligand, the Ad penton base. Keeping the RGD-binding domains as modeled in Figure 21, we added the remaining αv (blue) and β (red) extracellular domains to form either a bent conformation or an extended integrin conformation (Figure 25). The αvβ3/RGD crystal structure represents a bent conformation for an almost complete ectodomain just lacking the β-chain PSI domain (*2*). The PSI domain was added from another αvβ3 crystal structure of the complete ectodomain in a bent conformation (*3*). There is currently no available atomic resolution structure for a complete ectodomain in the extended conformation. Therefore we built a composite model starting with the headpiece domains as modeled in Figure 22 and adding the remaining αv and β chain domains to roughly approximate the diffuse integrin density in the outer radial shell (515-600Å) of the cryoEM structure.

Comparison of the bent integrin model with the cryoEM density shows that a significant portion of the integrin model folds back toward the penton base rather than filling the diffuse integrin density in the outer radial shell (Figure 25A). Overall the bent integrin conformation is inconsistent with the shape of the cryoEM density. Instead we find that the integrin density in the HAdV12/αvαβ5 structure is more consistent with an extended conformation with the integrin α and β chains extending away from the penton base (Figure 25B). In our cryoEM-based model for extended integrins bound to penton base, the C-terminal ends of the integrin ectodomains are all clustered around the Ad fiber and are within ~100Å of one another.

The soluble form of αvβ5 integrin used in this cryoEM study has the C-terminal ends of the αv and β5 chains tethered together with 7 amino acid Gly-rich linkers and Fos/Jun dimerization domains (*119*). The distance between the ends of the Fos/Jun domains when they are dimerized is 12Å, as measured in a crystal structure of c-Fos/c-Jun bound to DNA (PDB-ID 1FOS) (*12*). The distance between the C-terminal ends of the αv and β3 integrin chains in the bent conformation in the αvβ3/RGD crystal structure is surprisingly similar at 13Å (Figure 18B). Therefore the Fos/Jun dimerization domains plus Gly-rich linkers should not impede the formation of a bent integrin conformation as observed in the αvB3/RGD crystal structure. The soluble αvβ5 integrin has one additional residue at the C-terminal end of both the αv and β5 chains compared to the αvβ3/RGD crystal structure. So we have modeled the effect of the linkers and Fos/Jun dimerization domains on the extended integrin conformation assuming 8aa linkers in each integrin chain. Our calculations indicate that the Fos/Jun dimerization domains plus two 8aa linkers would allow the C-terminal ends of the αv and β5 chains to spread apart by up to 70Å(assuming an extended length of 3.6Å per residue in the two linkers (*128*)). This maximum distance is consistent with our cryoEM-based model for four extended integrin ectodomains bound to penton base (Figure 25B). Although our model has the αv and β5 C-terminal pairs within ~15Å of each other (Figure 18B), the diffuse nature and low resolution (85Å) of the integrin density in the outermost radial shell (Figure 19B) would suggest that the integrin chains in this region are highly disordered and may well spread farther apart.

Discussion

Symmetry or asymmetry of receptor binding

CryoEM structures of various picornavirus/receptor complexes, including rhinovirus and coxsackievirus with ICAM-1 (*130, 131*) and poliovirus with the poliovirus receptor CD155 (*132-134*), have revealed one receptor per asymmetric unit or 60 receptor molecules per virion. This is in contrast to what we have observed for the interaction of adenovirus with its integrin receptor. In the case of both ICAM-1 and the poliovirus receptor, elongated density was observed for 60 discrete receptor binding sites on the viral capsid after imposed icosahedral symmetry. The highest resolution of these cryoEM studies is an 8.0Å (FSC 0.5) resolution structure of Coxsackievirus A21 (CVA21) complexed with an ICAM-1 variant (*131*). This resolution assessment was made for the radial shells corresponding to the viral protein shell and the first domain (D1) of ICAM-1. CryoEM density was also observed for 4 additional ICAM-1 domains (D2-D5) extending outward from the viral surface. However, the resolution was worse and the density strength was progressively lower for each successive domain suggesting flexibility at each elbow between domains. Although the strength of the cryoEM density for the D1 domain of ICAM-1 suggested only 80% occupancy of ICAM-1 binding sites, docking with crystal structures for both CVA21 and ICAM-1 indicated no reason why 60 receptors molecules might not bind to one virion. Our cryoEM structure of HAdV12/αvβ5 is similar to CAV21/ICAM-1 in that more moderate resolution and weaker density is observed at progressively greater distances from the viral capsid. However, the HAdV12/αvβ5 structure differs in that discrete density is not observed at 60 receptor binding sites. In fact, docking with crystal structures of integrin indicates that 60 integrin molecules are unlikely to bind to one HAdV12 virion simultaneously.

The HAdV12/αvβ5 cryoEM structure provides a model for the penton base/integrin interaction

HAdV12 was chosen for this cryoEM structural study since its penton base has the shortest integrin-binding RGD loop. We reasoned that the short HAdV12 RGD-loop would be the least flexible among the various Ad types and thus the HAdV12/αvβ5 complex would yield the highest resolution for the bound integrin. Here we present a cryoEM structure of the HAdV12/αvβ5 complex with a resolution of 8.3Å (FSC 0.5) for the icosahedral capsid. This allowed visualization of α-helices in the capsid and facilitated accurate docking of the penton base crystal structure within the cryoEM density. The Rosetta *de novo* protein modeling software was used to build atomic models for the RGD-containing loop of the HAdV12 penton base. These models suggest that the RGD loop is quite flexible and might extend in various directions from the top of the penton base.

The HAdV12/αvβ5 cryoEM structure displayed significantly lower resolution for the integrin density than that observed for the capsid. The structure showed a ring of integrin density on top of the penton base RGD-protrusions and surrounding the Ad fibers. The resolution of this radial shell of integrin density is 27Å (FSC 0.5). Beyond this integrin ring, the cryoEM structure showed additional diffuse integrin density with an even lower resolution of 85Å (FSC 0.5) extending out approximately 160Å from the top of the penton base. Rigid body modeling with the RGD-binding domains of the homologous αvβ3 integrin indicates that four integrin molecules can be modeled above the HAdV12 penton base. However, because of the close spacing of the penton base RGD protrusions (~60Å) our model indicates that each of the four bound integrins must adopt a different orientation relative to the underlying penton base monomer. The significant variability in the modeled integrin positions with respect to the penton base RGD protrusions is estimated to be in the range of 10Å to 29Å after alignment with the penton base monomers. This positional variability for bound integrin may explain the lower resolution observed for the integrin density compared to the icosahedral capsid in the cryoEM structure of the complex.

This penton base/integrin modeling analysis also showed that with four integrins bound to the HAdV12 penton base the fifth RGD protrusion is effectively shielded from binding integrin. An earlier Biacore study with HAdV2 virions indirectly attached to a sensor chip via an anti-fiber monoclonal antibody, indicated that at close to saturation ~4.2 soluble αvβ5 integrin molecules could bind to each penton base (*118*). The HAdV2 penton base has a significantly longer RGD-loop than HAdV12 (78 *vs.* 15 amino acids) and thus the HAdV2 RGD-loop may provide enough additional flexibility that either four or five integrin molecules may bind to one penton base.

The HAdV12/αvβ5 cryoEM structure is consistent with the extension model for integrin activation

A variety of experiments including the addition of disulfide bonds to lock integrins in the bent conformation have shown that integrin extension is required for ligand binding during integrin inside-out signaling (*111*). In the extension model, the open headpiece conformation, observed in a crystal structure of αIIbβ3 and a therapeutic antagonist (*11*), represents the high affinity, ligand bound state. The open headpiece conformation has the β-chain hybrid domain swung by up to 70° from its position in the bent integrin conformation (*113*). This large swing of the hybrid domain presumably leads to separation of the two integrin cytoplasmic tails. Separation of integrin transmembrane domains has also been shown to be an important component of integrin outside-in signal transduction (*135*). Unfortunately there is no atomic resolution structure of a complete integrin ectodomain with a bound RGD ligand, except for the αvβ3/RGD crystal structure, which involved soaking an RGD peptide into a preformed αvβ3 crystal. Given the flexible nature of the multi-domained integrin heterodimer and the flexibility of integrin-binding RGD loops, it may be difficult to obtain an atomic resolution structure of an integrin/RGD ligand complex with a complete integrin ectodomain in an extended conformation. Our cryoEM structure of HAdV12 complexed with a soluble form of αvβ5 complex, albeit at moderate resolution for the bound integrin, offers support for the idea that αv integrins adopt an extended conformation upon binding to a multivalent RGD ligand.

Integrin binding may induce a conformational change in penton base and initiate the process of vertex protein release

The cryoEM structure of the HAdV12/αvβ5 complex also indicates that integrin binding may extend the penton base RGD loop in a direction counter to the natural twist of the pentamer and thus induce an untwisting of penton base. This is supported by a finding of significantly weaker penton base density in the HAdV12/αvβ5 complex than in the Ad35f vector. Since the N-terminal fiber peptide binds at the top of the penton base at the interface between monomers (*10*), an untwisting of the penton base multimer would likely disrupt the fiber binding site. This would offer a possible explanation for the finding that fiber is released at the cell surface early in the Ad cell entry process, perhaps after interaction with integrin (*136*).

In conclusion, this considerably improved cryoEM structure of the HAdV12/αvβ5 complex together with crystal structures of penton base, αvβ3, and $\alpha_{IIb}\beta3$ have enabled a more accurate and revealing molecular analysis of the adenovirus integrin interaction. The cryoEM structure indicates that αvβ5 integrin adopts an extended conformation when bound αvβ5 to the multivalent viral ligand and that the viral penton base may be conformationally affected by the binding of multiple integrin molecules to one vertex. The close spacing (~60Å) of the integrin-binding RGD sites on the penton base may promote integrin clustering, lead to the intracellular signaling events required for virus internalization, and prime the adenovirus capsid for programmed disassembly (*137*). The similar spacing of integrin-binding RGD loops noted for foot-and-mouth disease virus (FMDV) and coxsackievirus A9 (CAV9) (*97*) suggests that these viruses might undergo analogous conformational changes after interaction with integrin at the host cell membrane.

CHAPTER V

DISCUSSION

The main goal of this thesis work was to develop a computational algorithm that allows folding proteins into medium resolution cryoEM density maps. Two objectives were pursued. First, is it possible to determine the correct fold of a protein from a medium resolution density map? And secondly, following successful determination of the correct protein fold is refinement to atomic detail accuracy possible? This will answer the question of whether computation can create detail that is not seen in the experiment. A secondary goal was the determination of an experimental cryoEM density map of the Adenovirus-Integrin-Complex. The present chapter will summarize the achievements that were made with respect to each of the goals, their impact on the field as well as an outlook of further scientific developments which could follow the work described in this thesis.

Achievements

EM-Fold, a program that assembles and refines predicted secondary structure elements into medium resolution density maps was developed. Based only on the primary sequence and the positions of density regions that have been identified as corresponding to α-helices and β-strands, a model for the protein of interest is built. In the refinement step secondary structure elements are bent leading to even better models of the protein. Loops and side chains are added to the model using Rosetta. No comparable software for the restrained folding existed at the time when the project was started, thus a large amount of method development was required. The developed program, EM-Fold, is the only program available that can fold large proteins guided by a density map. It is part of a larger software library (BCL) but it can be used as a standalone tool as well. It can be used with maps of any resolution as long as secondary structure elements are visible in the map. If the resolution of the map is high enough (5 Å or better), EM-Fold can be used to build models of β-sheet containing proteins. The software integrates well with existing software that deals with related problems. Helixhunter / SSEHunter can be used to identify secondary structure elements in the map if manual identification is deemed infeasible. Models generated by EM-Fold are ideal input for Rosetta to employ loop and side chain building as well as refinement procedures. Within the scope of this thesis a workflow has been developed that combines EM-Fold and Rosetta with the aim of obtaining models of medium sized benchmark proteins that are accurate at atomic detail. It was shown that a combination of EM-Fold and Rosetta was able to refine several large proteins to a level where they were accurate at atomic detail. This is remarkable as the experimental density map on which all the folding and refinement are based does not contain any atomic detail. Yet computation is able to recover at least some of that non-visible detail. EM-Fold is freely available to the scientific community (available to commercial users for a small fee) and can be downloaded from the BCL commons webpage at http://bclcommons.vueinnovations.com/licensing. Benchmarks of EM-Fold showed higher than average success rates of building a model that is sufficiently close to the native conformation to be considered a success (*33*). Beyond demonstrating EM-Fold's success rate in several benchmarks on both simulated and experimental density maps, it was applied to build a model for protein IIIa of Adenovirus. Building a model for this important capsid protein of Adenovirus can help design experiments to test predictions made by EM-Fold. Finally, it was demonstrated that EM-Fold can be a very integral part in yielding models with atomic resolution detail that are not accessible to computational structure prediction otherwise.

With respect to the second goal, a sub-nanometer structure (8.3 Å resolution) of the Adenovirus-Integrin-Complex has been determined. The resolution in the radial shell containing the integrin density was sufficient to make clear statements about the structure of integrin bound to its ligand (the RGD sequence of penton base). A large conformational change occurs – favoring the switchblade model of integrin adopting an extended conformation when binding a multivalent ligand.

Impact on the field

EM-Fold was the first program that used a cryoEM density map as a folding restraint in de novo protein folding. Several different experimental restraints (such as EPR or NMR distance restraints data) had been shown to help the accuracy and success rate of de novo protein folding algorithms (*32*, *82*-*86*). None of the existing programs used the general position of secondary structure elements in space to help build better models. It was demonstrated that the general position of helices and strands is sufficient to effectively guide the folding process. Using the map as a restraint, proteins as big as 400 amino acids can be folded. This confirms the applicability of cryoEM data as a restraint for protein folding. From the vantage point of the experimentalist, EM-Fold is a unique tool that can help extract more information from a medium resolution density map. Before the release of EM-Fold, researchers often placed predicted secondary structure elements into the map by hand and checked manually whether the model made sense by ensuring that the loops could be closed (*6*). This approach cannot sample the conformational space of possible placements of SSEs into the density as thoroughly as the present algorithm can. The development of EM-Fold has important implications for the cryoEM field as fold information can now be gathered from maps that were previously uninterpretable. This information can at least help to design experiments to test the hypotheses generated by the predictions of EM-Fold. In summary EM-Fold can be considered as a helpful tool in an ever growing toolbox available to electron microscopists to interpret maps that don't show atomic resolution information.

In addition to helping interpret medium resolution density maps in terms of the protein fold, EM-Fold can, in favorable cases, build models that are accurate at atomic detail. This is remarkable since the experimental data does not contain any information at the atomic level. Also computational methods alone will not be able to accurately predict detail at the atomic level. Combining computation and medium resolution experimental data will be able to restore non-visible information. Computationally adding detail not present in the experimental data had been reported for other methods such as NMR, X-ray crystallography and even EPR (*32*, *82*-*86*). The ability to extract non-visible information from cryoEM density maps was demonstrated for the first time in this work.

The determination of a cryoEM structure of integrin bound to a multivalent RGD ligand provided vital evidence to corroborate one of the two competing models of integrin bound to the RGD peptide. It is apparent that in the case of integrin binding to the RGD sequence on the penton base of Adenovirus, a large conformational change in integrin structure occurs. The transition from a bent to an extended conformation has implications for signaling mechanisms that trigger the uptake of the virus into the cell. Understanding the mechanism of Adenovirus cell entry is of importance in light of Adenovirus' use as gene delivery vector.

Outlook / Future Directions

Judging by the impressive growth in the number of cryoEM maps determined (both moderate resolution and sub-nanometer resolution maps) over the past ten years, it can be expected that many more high to medium resolution maps will become available within the near future. As alluded to in length in Chapters I through III, computational methods are required to effectively interpret medium resolution density maps since tracing the backbone directly in the map remains impossible. Another type of density map that EM-Fold could be applied to is a “low resolution” X-ray crystallography electron density map. Crystallographers sometimes only obtain diffraction data out to 5-7 Å, especially when working on membrane proteins which are difficult to crystallize or yield only mediocre diffraction data after being crystallized. Often times this data is not even published because it is deemed useless. EM-Fold could help to build models for these cases, extracting as much information from the maps as possible.

Besides the prospect of an increased number of target maps that the algorithm can be applied to, future improvements to EM-Fold will make it more applicable and accurate in predicting protein models from density maps. One aspect that future work will focus on is using so called “density bumps” on the density rods to score models. Density bumps are protrusions of density to one specific side on the surface of a density rod. As mentioned previously they often represent density reconstructed for large, non-flexible side chains. Maps at different resolution and quality will exhibit different levels of density bump information. For the evaluation of the protein IIIa density map, the analysis of locations of density bumps was done manually. There are two reasons why it would be desirable to automate this analysis. First, any process involving user judgment adds a level of subjectivity. Ideally there should be minimal need for user interaction. Secondly, automated identification and evaluation would allow implementation of these steps into the folding protocol. Then the agreement of the model that is being built with the experimental density map could be evaluated after every step of the Monte Carlo protocol. This would drive the folding to models that agree better with the density profile instead of simply filtering the models that were built using the identified density bumps. Initial work on automating the identification of density bumps in a map was carried out in the course of this thesis. A detailed discussion of developed methods and preliminary results can be found in the Appendix.

The problem that EM-Fold cannot find the correct topology in more than 60-75% of the cases has been identified to be a scoring problem, rather than a sampling problem. This is interesting for several reasons. First, it is generally known that the main bottleneck in computational protein structure prediction is sampling and not scoring. In Rosetta for instance, the native model almost always scores better than any model that was built, indicating that if one could only sample the correct conformation the model would be favored by score. So in that sense, the EM-Fold assembly step deviates from known patterns. Secondly, this means that the main focus of EM-Fold development will be on developing better scores or finding sets of existing scores that yield better results. Within this context, it is important to keep in mind that incorrect secondary structure prediction contributes considerably to the problem. It is anticipated that future work targeted at improving the assembly step success rate will involve both development of more accurate secondary structure prediction techniques as well as identification of computationally inexpensive scores that are less dependent on having a secondary structure element with the correct length in place. This could include scores that look at pair-wise amino acid distances (without taking into account the orientation of the amino acids around the axis of the SSE) or orientation independent contact scores. This area of research will include a large amount of method development. The third conclusion that can be drawn from the fact that the assembly step has a scoring rather than sampling problem is that EM-Fold has not yet been tested at its maximum protein size limit. It is hard to estimate where the limit would be with the current setup because every added secondary structure element increases the folding complexity by tenfold. It would be interesting to perform a benchmark to find out where the limit is.

Besides the problem of not identifying the correct topology for about 30% of the benchmark cases, there were some proteins in the benchmark for which the correct topology was identified and also scored pretty well but whose models still exhibit somewhat larger RMSDs than most other proteins (e.g. 1GS9, 1NIG, 2IGC, 1CHD). Two reasons for these higher RMSDs could be identified: imprecise secondary structure prediction (1GS9, 1NIG, 2IGC) or large loop regions for which no good loop building can be performed in Rosetta (1CHD). The first case happens when one particular SSE is predicted with about the correct length but not in its correct sequence position. For example, there is a helix in the protein from residues 130-145, but the closest prediction yields a helix from residues 135-150. This predicted helix will still be placed in the observed density rod resulting in models with a relatively high RMSD. Currently there is only the SSE resize move in the assembly step that could correct for a situation like this but it is also very unlikely that this will happen assuming region 130-135 is strongly predicted to be coil. One possible improvement could be the implementation of a sequence shift move that shifts the sequence of a SSE without changing its length. The second case only occurred once in the benchmark sets but more frequently during the cryoEM modeling challenge. Whenever there are large parts of the sequence that do not contain any secondary structure elements (or only very short SSEs for which no density can be identified) these regions are left as loop regions for Rosetta to build de novo guided by the density map. As is well known, loop building only works reliably up to loop sizes of 12 amino acids (*79*). Even when guided by the density map it is not possible to build a realistic model of long loops. It has to be honestly stated that the benchmark sets contained an above average amount of secondary structure and that in cases with less secondary structure certain regions will not be predictable to the same accuracy as the rest of the protein.

The medium resolution cryoEM density map (*138*) and the low resolution crystal structure (*139*) of DNAPKcs have been identified as a possible future application of EM-Fold. In both maps the resolution is sufficient to see secondary structure elements in at least parts of the map without seeing connections between SSEs. With more than 4000 residues, the protein is of a challengingly large size for the folding protocol. Preliminary data for the project has been compiled in the Appendix.

So far EM-Fold has been tested on the structure of one membrane protein – rhodopsin (*5*). Future applications of the algorithm might include building models into low resolution X-ray electron density maps. Membrane proteins are very hard to crystallize and crystals frequently only exhibit suboptimal diffraction. So it is conceivable that several of these low resolution X-ray maps will be from membrane proteins. Many of the knowledge based scores used in the EM-Fold assembly and refinement protocol might be different for soluble and membrane proteins. Membrane proteins might reveal other statistics with respect to loop lengths versus number of amino acids in the loop for instance. It might therefore be interesting to contemplate developing a specialized version of EM-Fold (e.g. Membrane-EM-Fold) that is tailored to work with membrane proteins. This will also depend on the number of non-homologous membrane protein structures deposited in the protein data bank. It will only be possible to derive energy potentials once there are a sufficient number of structures available.

Yet another future development of the protein structure prediction field will be the integration of more than one sparse experimental data set into the same protein folding simulation. The BCL is written to accommodate several experimental restraints at once. So EM-Fold will be a potential building block in combining two or more experimental data sets to guide a de novo protein folding run.

Lastly there will hopefully be two external developments that are not directly related to EM-Fold but whose success will have a great influence on the method. The first one, improved secondary structure prediction methods, has been mentioned before. It would be preferable to obtain programs that can more accurately predict secondary structure instead of developing new methods to deal with incorrect secondary structure prediction. The success of major progress in the field is however unclear. The second anticipated development will be improved programs to interpret density maps. We had to resort to manual identification of density rods because existing software gave results inferior to visual inspection. It is hypothesized however that scientists will be more likely to use EM-Fold if it is as easily integrated with software such as SSEHunter as it already is with Rosetta now.

In summary, a program EM-Fold that combines medium resolution cryoEM density maps and de novo protein structure prediction has been successfully developed. It was demonstrated that EM-Fold has about a 70% success rate of folding a model with the correct topology and that in favorable cases it is possible to refine the final models to be accurate at atomic detail. Additionally, the medium resolution cryoEM structure of the Adenovirus-Integrin complex was able to elucidate the conformation of integrin bound to a multivalent RGD ligand. The aims laid out in the Research Proposal were all completed and additional work was performed. EM-Fold is freely available and will hopefully prove useful to many researchers in the years to come.

APPENDIX

Commandlines and scripts to perform an entire folding run

This section of the appendix describes how the folding results for protein 1X91 were obtained. The used scripts and commandlines are provided. The actual source code of scripts is given on the data DVD along with detailed results for 1X91 (see folder scripts_commandlines). The three used executables are provided in a separate folder called executables. The DVD also contains the pdb files for the correct topologies for all the benchmark proteins that were generated during the benchmark (see folder results_benchmark).

The first step in running EM-Fold consists of a few pre-processing steps (see Figure 5). Starting from the fasta file (1X91A.fasta), secondary structure prediction files are generated (1X91A.jufo, 1X91A.psipred_ss2, 1X91A.rdb6) by using the CSB script runs. A pool file is generated using the CreateSSEPool application of the BCL (bcl.exe CreateSSEPool -prefix 1X91 -ssmethods JUFO9D PSIPRED PROFphd -use_consensus -chop_sses -min_sse_sizes_pool 10 4 -sse_threshold 0.4 0.2). The other crucial pre-processing step is the identification of density rods. For the benchmark this was done using the script create_constraint_files.pl. It takes the native pdb as starting point and copies the SEQRES and SSE definitions and the ATOM lines corresponding to atoms in SSEs into files 1X91_10.cst_body and 1X91_12.cst_body, depending on whether 10 or 12 residues are chosen to be the minimum length of identifiable density rods. With these files in hand, the assembly part of the EM-Fold protocol can be started. To distribute the computations over several cores on the cluster, the script generate_pbs.pl is used. The pbs scripts for the assembly run can be found in the pbs/ subdirectory. They are submitted to the cluster using the script submit.pl. The individual commandline is apps_release.exe Fold -nmodels 1 -fasta 1X91A.fasta -pool 1X91_10.pool -mc_number_iterations 2000 500 -mc_temperature_fraction 0.25 0.05 -prefix /tmp/ -em_fold 1X91_12.cst_body 3.5 3.5 4.8 4.8 -1.0 -

em_fold_write_body_assignment -random_seed -message_level Critical -score_weightset_read score_weights_assembly_resize.score -score_density_connectivity 1X91.mrc -print_tracker_history -write_minimization improved -sspred JUFO PSIPRED SAM -sspred_path_prefix proteins/1X91/ 1X91". When the runs are finished the output is written to files 1X91_XX.output in the results/ subdirectory. The script sortMcRuns.pl is used to extract the results from the assembly step into human readable format and sort the models by score. When executed the script creates a file unique_sorted1X91.txt that has all the different unique topologies listed in order of their score. By setup of the script, the correct topology corresponds to model A-B-C-D-E-. In this case A-B-C-D-E- is ranked first among all the topologies that were built during the folding run. This file also contains the seed number and model name needed to reproduce the particular file. The top 150 scoring topologies from the unique_sorted1X91.txt file are chosen to be refined in the EM-Fold refinement step. This is done using the script generate_pbs_refinement.pl. For execution on the cluster, pbs scripts are generated that first reproduce the model (same command line as in the assembly step, this time the random seed is specified). After the starting models for the refinement have been generated (A-B-C-D-E-.pdb for the correct topology), their RMSDs to the native model are calculated. After that 500 refinement runs are performed for all the top 150 scoring models generated by the assembly step. A sample command line for this is "apps_release.exe Fold -nmodels 1 -native 1X91_shifted.pdb -start_model A-B-C-D-E-.pdb -score_density_agreement 1X91.mrc -fasta 1X91A.fasta -pool 1X91_10.pool -mc_number_iterations 2000 400 -mc_temperature_fraction 0.25 0.05 -prefix A-B-C-D-E-/A-B-C-D-E- -random_seed -message_level Critical -score_weightset_read test_score_weightset.score -em_fold_refinement -file_compression GZ -no_superimposition". The RMSDs of all the refined models to the native model are calculated as soon as the refined model is generated. Once the refinement step has finished and all the 75000 refined models have been generated, the script sort_results_refinement.pl extracts the scores and RMSDs for all the models and sorts them. The final results of the refinement step are summarized in the file sorted_results_bestscoreonly.txt in the high_res_refinement/results/ subdirectory. For the 1X91 case, the correct topology is scored best out of all the refined topologies. The RMSD to the native pdb of the model that was generated by the assembly step (A-B-C-D-E-.pdb) is 4.0 Å. The best scored refined model (A-B-C-D-E-_353.pdb) has a RMSD of 1.6 Å to the native structure. The refinement step was clearly able to considerably improve the model from the assembly step. The next parts of the folding protocol are three rounds of Rosetta refinement. As described in Chapter III, the top 50 scoring topologies after refinement are refined in the first round of Rosetta refinement. In the first round, Rosetta is used to build coordinates for all residues that are missing in the models. These are residues in loop regions, as the EM-Fold models only contain amino acids in secondary structure elements. As with all the calculations, the Rosetta runs are distributed over several processors on the ACCRE cluster. The script generate_pbs_rosetta_refinement.pl is used to generate the pbs files used for the Rosetta refinement runs. The pbs files can be found in the atomic_res_refinement/pbs/ subdirectory. In round 1 the loopfiles only contain residues in the loop regions. A sample command line for a round 1 Rosetta run is "loopmodel.linuxgccrelease -database minirosetta_database/ -out::nstruct 50 -loops::input_pdb A-B-C-D-E-_353_cleaned.pdb -in::path atomic_res_refinement/start_models/ -edensity::mapfile 1X91.mrc -edensity::sliding_window 9 -edensity::mapreso 6.9 -edensity::grid_spacing 4.0 -edensity::whole_structure_allatom_wt 0.05 -loops:loop_file atomic_res_refinement/start_models/A-B-C-D-E-_353_cleaned.loopfile -loops::frag_sizes 9 3 1 -loops::frag_files aa1X91A09_05.200_v1_3 aa1X91A03_05.200_v1_3 none -loops::remodel quick_ccd -loops::intermedrelax no -loops::refine no -loops::relax fastrelax -relax::fastrelax_repeats 4 -psipred_ss2 1X91A.psipred_ss2 -score:weights score13_env_hb -out::pdb -out::path atomic_res_refinement/pdbs -out::prefix round1_A-B-C-D-E-_353_1_aa_ -out::overwrite". Each run generates 50 models. Thus for each of the 50 topologies, 500 refined models are built. Once round 1 refinement is finished, the script calculate_rmsds_rosetta.pl can be used to calculate all the RMSDs of the round 1 models with respect to the native model. This script outputs the file scores_sorted_final_models_round1.txt in the atomic_res_refinement/ subdirectory. It also generates a plot of the Rosetta scores vs. RMSD to native for all the models (both over the full length of the protein and over helical residues only): rmsd_vs_score_round1.pdf. Round 2 and 3 of Rosetta refinement are very similar. They both identify regions in the top scoring models from the previous round that agree least with the density map and target these regions for rebuilding of the protein backbone. The only difference between the two rounds is the number of top scoring models from the previous round that are considered for refinement. Round 2 refines the top 15 scoring topologies of round 1. Round 3 refines the top 5 scoring topologies from round 2.

Description of code structure and important objects

EM-Fold is part of the BCL, a collection of classes written in C++ code. This part of the appendix will give an overview over the important objects in the BCL and how they work they are structured to perform the folding of a protein. A copy of all the important header and source files from the BCL is given on the DVD (see folder BCL_code_structure).

EM-Fold itself really is only a specialized version of the general BCL fold protocol (BCL::Fold). Thus it runs the same application (bcl_app_fold.cpp) as BCL::Fold, using certain flags to indicate that folding is restricted to regions of observed density. The main feature of the bcl_app_fold.cpp application is the Monte Carlo Minimizer. It keeps track of the moves performed, how these mutates affect the score of the protein model, whether a move will be accepted and when to end the minimization. The scores used are generally read in from a score file (scores are initialized using fold::Scores). The moves are defaulted in Fold::Mutates. A different set of moves is used specifically for the fold-into-density protocol (chosen by using the -em_fold flag in the bcl_app_fold.cpp application). Fold::Setup serves as the setup class that contains all the flags and data necessary for initializing mutates, scores and various other data. The start models for the minimization runs are chosen depending on the protocol that is used. For the EM::Fold assembly step, a completely empty start model is chosen, while the EM::Fold refinement step reads in a starting pdb to start the minimization from the model generated by the assembly step. The restraints that are read in with the constraint file are used to initialize certain moves and scores. These so called “BodyRestraints” themselves are not part of the protein model but a member of the scores. Special restrained move are needed for the “add” and the “move” move. The specialized version of the adding (MutateProteinModelSSEAdd) is constructed with a shared pointer to a PlacementInterface (PlacementSSEIntoBody). The specialized “move” move is the MutateProteinModelSSESwapBody and also uses the BodyRestraint. In the assembly step, there are two scores that use the BodyRestraints. The function GetScoreDensityAgreement() in bcl_fold_scores.h returns the score that scores the length agreement between the density rods and the secondary structure elements that are placed into them. The function GetScoreBodyConnectivityDensity() in bcl_fold_scores.h returns the score the scores the placement of short loops into strong connecting density between close ends of density rods. Based on the placements of SSEs the class PrinterProteinModelBodyAssignment (bcl_assemble_printer_protein_model_body_assignment.h) determines the assignment information for a particular protein model and body restraint. The assignment information consists of which body (density rod) is occupied by which SSE from the pool as well as the orientation of the SSE in the density rod (this is compared to an arbitrary orientation defined in the constraint file). This assignment information is written to the pdbs produced during the folding run as well as to the standard output. The script sortMcRuns.pl (see Appendix) uses exactly this output to extract the results from the assembly step into human readable format and sort the models by score.

Density profiles of cryoEM density rods

CryoEM density maps at different resolutions show different levels of detail. It is generally accepted that maps start to show helices as density rods starting at around 10 Å resolution. At about 5 Å resolution β-strands become visible and at some point between 3 and 4 Å resolution it will be possible to trace the backbone of the protein trough the map. While generally atomic resolution is considered necessary to clearly identify density corresponding to side chain coordinates, it has been shown that some side chains will exhibit strong density even at resolutions around 7 Å (*5*). It is most likely to see large aromatic side chains in these medium resolution density maps. The side chains will be visible as so called density bumps on secondary structure elements. No clear detail can be seen within the density bump, making it impossible to distinguish between different residue types based on the density. At a resolution of around 7 Å, not every large side chains in SSEs has to correspond to a density bump and conversely not every density bump observed on the density rods needs to correspond to a large side chain. In (*5*) we were able to use the position of a large density bump to manually filter models for protein IIIa. We are hypothesizing that there is enough correspondence between large bumps and side chains to be able to use this information in an automated fashion during the folding protocol. For this to be possible two criteria have to be met. First, an automated way of identifying density bumps on density rods is needed. This will help to speed up the process of evaluating the density map and largely eliminate the subjective user interaction of judging whether a particular part of the density is considered a bump or not. Second, a score that scores the agreement of a protein model with the calculated density bump profile has to be developed. Within the scope of this thesis first steps have been taken to implement density bump evaluation into the folding protocol.

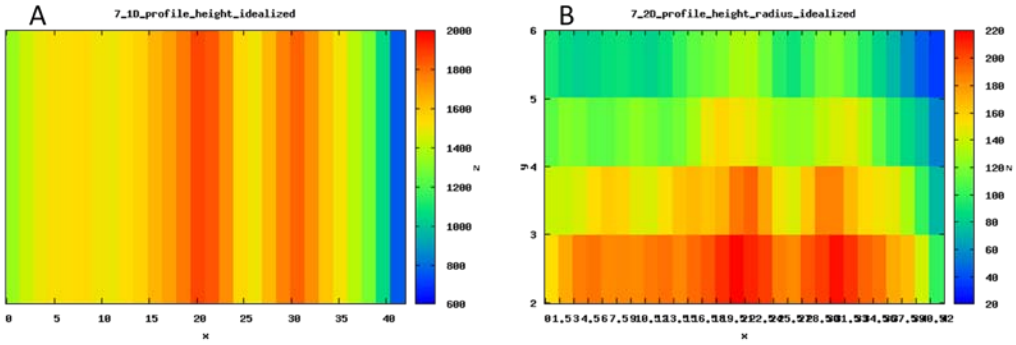

Figure 26. 1D-Height and 2D-Height/Radius profiles for idealized benchmark helix
Panel A shows the 1D-Height profile of the rod. The x-axis corresponds to the main helical axis. All voxels in the radial and angular dimension have been summed up. Red colors indicate strong density. Panel B shows the 2D-Height/Radius profile of the helix. The x-axis corresponds to the main helical axis while the y-axis shows the radial extension away from the main helical axis. All voxels in the angular dimension have been summed up. Red colors indicate strong density.

Work so far has concentrated on the implementation of a density bump profile into the BCL. For this purpose a cylindrical density map was developed (include/density/bcl_density_map_cylindrical.h). This class is similar to the regular density map in that it contains the map as a math::Tensor, but is based on a different coordinate system. The cylindrical density map is not based on a Euclidian coordinate system but on a cylindrical coordinate system. The main axis of the density rod is chosen to be the reference axis of the system. The voxels are divided along the length of the axis, radially away from the axis and angularly around the axis. This way every density rod in the density map can be represented by a separate cylindrical density map. Based on these maps it is possible to calculate 2D and 1D density profiles that describe the distribution of density along a density rod. Six profiles can be calculated: 1D-Angle, 1D-Height, 1D-Radius, 2D-Angle/Height, 2D-Angle/Radius, 2D-Height/Radius. Of these the 1D-Height and 2D-Height/Radius profiles are probably the most useful ones. Figure 26 shows an example of these two profiles for an idealized benchmark helix.

The 1D-Height profile (panel A) sums up all the voxels in cylindrical rings around a section of the helical axis. High (red) values in a part of the rod indicate that the density there is either stronger or extents out further possibly describing a density bump. This profile does not capture the orientation of the bump around the density rod nor how far the bump extents out from the axis. The 2D-Height/Radius (panel B) plots the extension of the density rod on the x axis and the radial extension around the rod on the y axis. Within each of these segments all the angles are summed up. Strong density areas (red) can be seen for certain parts of the rod. It is also possible to see how far out the density extents. This profile still does not capture the orientation of the bump around the density rod.

When using these profiles for scoring whether a helix placement into a density rod agrees with the profile of the rod it is important to use to appropriate density profile – sampling combination. For instance, the 1D-Height and 2D-Height/Radius profiles should be used in the EM-Fold assembly step which is not sampling orientations of helices around their main axis. When using a profile in the EM-Fold refinement step, it should have an angular component as the sampling of rotations around the main axis of the helices is an important move in the refinement.

Very initial work has been done with respect to a score implementation that uses the information contained in the profiles to check against the placement of specific SSEs into the rods. For this a Pearson product moment correlation coefficient was introduced. It checks the correlation between the experimental profile and a profile derived from an individual secondary structure element in the model. Future work will concentrate on which profile to use in which of the steps of the folding protocol. One option that has not been pursued so far is to use the full 3D profile (i.e. the cylindrical density map) for the scoring. Data for this project can be found on the DVD (see folder density_profiles).

DNA-PKcs results

EM-Fold was applied to build a model for a region of DNA protein kinase catalytic subunit (DNAPKcs). Both a medium resolution cryoEM density map (*138*) and a crystal structure (*139*) of the molecule have been determined recently. Neither were at sufficient resolution at trace the backbone of the molecule. The catalytic subunit contains 4128 residues and has about 135 helices predicted. This is much too big to use EM-Fold. However one of the domains in the map exhibits a clear heat repeat structure which would be predictable with EM-Fold. To identify the sequence that corresponds to this part of the map, the entire sequence was submitted to Pfam. Pfam found 4 Pfam-A matches to the search sequence (all significant) but did not find any Pfam-B matches. The matches are NUC194 domain (alignment to residues 1815 - 2210), FAT domain (alignment to residues 3023 - 3470), Phosphatidylinositol 3- and 4-kinase (alignment to residues 3748 - 4014) and FATC domain (alignment to residues 4097 - 4128). Closer inspection of the results revealed that the FAT domain is a member of the clan TPR. Several members of the Tetratrico peptide repeat superfamily (TPR) are heat repeats. Also a visual inspection of the secondary structure prediction for the entire DNA-PKcs showed a highly helical region between residues 2700 and 3500. Consequently these residues (2700 - ~3500) were submitted to Phyre. The results show a region of very high helical content between residues 200 - 840 (corresponding to 2900 - 3540 in the original structure). Several of the structures determined by fold recognition have low E-values and are heat repeats very close in structure to what we see in the density map. Good examples are scope codes d1qbkb and d1qgra. These results corroborate that region 2900 - 3540 is likely to be the region that corresponds to the clearly resolved heat repeat density in the map. A secondary structure element pool for the residues 2900 – 3540 was generated. and folding these predicted helices into the clearly resolved density rods in the cryoEM density map. The programs jufo, psipred and profPhD were used to predict secondary structure for the heat repeat domain. The predictions among those methods agree very well. Looking at the lengths of predicted helices, the following length pattern can be found (no splitting of long helices performed; if there is a splitting then one method predicted long helix, the other method predicted two short helices): 54 - 14 - 20 - 15 - 13 - (27 / 12 - 12) - 13 - 11 - 13 - 16 - (30 / 20 - 10) - (26 / 12) - (24 / 11 - 11) - 12 - 19 - 30 - 12 - 21 - 34 - 13 - 17 - 13 - (19 / 13 / 11) - 10 - 18 - 12 - 27 - 23 - 13 - 12 - 11 A total of about 31 helices of 10 or more residues was predicted. The density region of the clearly resolved density rods was investigated manually. The following pattern was observed (lengths of density rods in number of helical residues, order is based on the assignment in the crystal structure, therefore somewhat arbitrary but spatially close): 16 - 14 - 11 - 9 - 18 - 12 - 14 - 18 - 12 - 14 - 20 - 14 - 19 - 19 - 14 - 12 - 17 - 12 - 12 - 19 - 10 - 15 - 13 - 10 - 10 A total of 25 density rods of 9 or more residues was observed. The mismatch between the number of predicted helices and clearly observed density rods is expected. The clearly observed density rod region likely only encapsulates a subset of residues 2900 – 3540 leading to a smaller number of density rods than there are helices in that part of the sequence. One challenge of the model building process will be to identify the sequence that corresponds to the observed density rods. Ideally the results for the EM-Fold assembly step will point to a preferential placement of helices from a subsection of the sequence. If this is not the case, it might become necessary to manually filter the models. During the folding the best possible density map should be used. There was no density map published alongside the crystal structure. We were able to calculate a density map from the structure factors used to build the model published in (*139*). The problem at the current point is that the map calculated using the structure factors does not have perpendicular axes (90, 105, 90 vs. 90, 90, 90). There seems to be no software that can reformat the map to a perpendicular axial system. The BCL currently cannot read in non-perpendicular axes correctly. So all initial EM-Fold runs were done without density capabilities, i.e. without density connectivity score for the assembly step and without the density agreement score in the refinement step. The next step will be to implement to conversion of non-perpendicular axes to perpendicular ones in the BCL. After this is completed the folding runs will be repeated. Preliminary data for this project can be found on the DVD (see folder DNAPKcs).

Results of cryoEM modeling challenge

Every two years the computational community participates in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment. This community wide experiment is aimed at providing a true blind fold test ground for comparison of computational methods dealing with question of protein structure prediction. Experimentalists hold back submission of structures to the protein data bank for the duration of the experiment and allow computational groups to predict models before the structures are published. This setup has proven useful to the community as it tests the available methods in a truly blindfold manner. When the leading groups that develop computational methods to interpret cryoEM density maps met for the "Modeling of CryoEM Map Workshop" in Houston in January 2010, it was decided that a similar experiment for the cryoEM modeling community would be desirable. The idea for the "CryoEM Modeling Challenge 2010" was born. The challenge, running from July to December 2010, consists of six categories: protein segmentation, secondary structure annotation, backbone tracing, rigid body modeling, flexible modeling and ab-initio modeling. The experiment was set up to be a challenge rather than a competition for several reasons. First, there are not enough maps available where the results for the experiment in question are known to one group only. Thus, it will be hard to compare methods because the experiment was not blindfold. And secondly, there still is a limited number of methods for modeling of cryoEM maps and the existing methods do not overlap too much. So since every method has another objective, it would be hard to compare results and determine a "winner". The Meiler lab decided to participate in the ab-initio modeling category. The organizers released 13 maps with resolutions ranging from 3.0 to 23.5 Å. Many of the maps contain density for several separate protein chains. Our approach was to assume that the maps were either segmented already or that tools existed that could segment density corresponding to a specific protein chain from the entire map. We decided to build models for six different proteins (2C7DU, 3FICD, 3FICG, 3FINF, 3FINR, 3FINU) corresponding to two different maps (GroEL at 7.7 Å resolution and the 70S ribosome at 6.4 Å resolution). The folding protocol (EM-Fold and Rosetta) established in the benchmark was applied to the proteins. The results are summarized in Table 6. The backbone RMSD values for the correct topology model after EM-Fold assembly, EM-Fold refinement and the third round of Rosetta refinement are shown. RMSDs after EM-Fold refinement range from 2.3 to 3.5 Å which is similar to the results seen in the last benchmark (see Table 5). Also the RMSDs over secondary structure elements after Rosetta refinement are below 5 Å for all but one protein (3FINF). The surprising part about the results though was the RMSDs of the full length models after Rosetta refinement. For five out of six models they range between 11 and 20 Å for the lowest scoring model with the correct topology. Only a single protein exhibits RMSDs below 6 Å (3FINU). Inspection of the models revealed that the large deviations between the models and the native structure came from imprecise loop building. The loops in the six proteins from the challenge are longer than the loops of the benchmark proteins which had been chosen to contain a high amount of secondary structure. So while the core of the protein (the SSEs) is still predicted correctly, the prediction for the connecting regions might be wrong. This demonstrates another current limitation of the folding protocol. If loops become too large, it will be unlikely to predict accurate coordinates for that part of the structure even with using the density map to guide the loop building. It might be best not to predict regions that exceed a certain loop length. The main results for the modeling challenge can be found on the DVD (see folder modeling_challenge).

Table 6. Results of cryoEM modeling challenge

protein	Number of residues	map	RMSD assembly step [Å]	RMSD refinement step [Å]	RMSD Rosetta round 3[Å]
2C7DU	97	GroEL, 7.7 Å	(2.6)	(2.4)	11.4; [10.9]; (3.5)
3FICD	208	Ribosome, 6.4 Å	(3.8)	(3.1)	15.9; [15.6]; (4.8)
3FICG	155	Ribosome, 6.4 Å	(5.3)	(3.5)	13.7; [7.2]; (4.0)
3FINF	208	Ribosome, 6.4 Å	(2.7)	(2.9)	19.6; [18.5]; (13.5)
3FINR	117	Ribosome, 6.4 Å	(3.7)	(2.3)	11.5; [11.4]; (2.3)
3FINU	117	Ribosome, 6.4 Å	(4.3)	(3.3)	5.8; [4.4]; (2.9)

All RMSD values are determined over the backbone atoms N, C_{α}, C and O. Values in parentheses refer to RMSDs over secondary structure elements only. Values in square brackets refer to the lowest RMSD model. All other RMSDs refer to the lowest score model. For the Rosetta refinement the RMSDs are reported over the full length of the protein and over secondary structure elements only.

REFERENCES

1. C. San Martin *et al.*, *J Mol Biol* **383**, 923 (Nov 21, 2008).

2. J. P. Xiong *et al.*, *Science (New York, N.Y* **296**, 151 (Apr 5, 2002).

3. J. P. Xiong, T. Stehle, S. L. Goodman, M. A. Arnaout, *J. Biol. Chem.* **279**, 40252 (Sep 24, 2004).

4. J. A. Kovacs, M. Yeager, R. Abagyan, *Biophys J*, (May 11, 2007).

5. S. Lindert *et al.*, *Structure* **17**, 990 (Jul 15, 2009).

6. S. D. Saban, M. Silvestry, G. R. Nemerow, P. L. Stewart, *J Virol* **80**, 12049 (Dec, 2006).

7. S. D. Saban, M. Silvestry, G. R. Nemerow, P. L. Stewart, *J. Virol.* **80**, 12049 (Dec, 2006).

8. T. Ju, M. L. Baker, W. Chiu, *Comput Aided Des* **39**, 352 (May, 2007).

9. M. L. Baker *et al.*, *PLoS Comput Biol* **2**, e146 (Oct 27, 2006).

10. C. Zubieta, G. Schoehn, J. Chroboczek, S. Cusack, *Mol. Cell* **17**, 121 (Jan 7, 2005).

11. T. Xiao, J. Takagi, B. S. Coller, J. H. Wang, T. A. Springer, *Nature* **432**, 59 (Nov 4, 2004).

12. J. N. Glover, S. C. Harrison, *Nature* **373**, 257 (Jan 19, 1995).

13. M. L. Baker, T. Ju, W. Chiu, *Structure* **15**, 7 (Jan, 2007).

14. C. A. Rohl, C. E. Strauss, D. Chivian, D. Baker, *Proteins* **55**, 656 (2004).

15. S. Lindert, P. L. Stewart, J. Meiler, *Curr Opin Struct Biol* **19**, 218 (Apr, 2009).

16. R. Henderson, *Q Rev Biophys* **37**, 3 (Feb, 2004).

17. Z. H. Zhou, *Curr Opin Struct Biol* **18**, 218 (Apr, 2008).

18. J. Meiler, M. Muller, A. Zeidler, F. Schmaschke, *Journal of Molecular Modeling* **7**, 360 (2001).

19. J. Meiler, D. Baker, *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12105 (Oct 14, 2003).

20. D. T. Jones, *Journal of Molecular Biology* **292**, 195 (Sep 17, 1999).

21. K. Karplus *et al.*, *Proteins-Structure Function and Genetics*, 134 (1997).

22. J. M. Chandonia, M. Karplus, *Proteins* **35**, 293 (May 15, 1999).

23. J. Moult, *Philos Trans R Soc Lond B Biol Sci* **361**, 453 (Mar 29, 2006).

24. J. Moult, *Curr Opin Struct Biol* **15**, 285 (Jun, 2005).

25. A. Kryshtafovych, C. Venclovas, K. Fidelis, J. Moult, *Proteins* **61 Suppl 7**, 225 (2005).

26. D. Chivian, D. Baker, *Nucleic Acids Res* **34**, e112 (2006).

27. K. M. Misura, D. Chivian, C. A. Rohl, D. E. Kim, D. Baker, *Proc Natl Acad Sci U S A* **103**, 5361 (Apr 4, 2006).

28. N. Eswar *et al.*, *Curr Protoc Protein Sci* **Chapter 2**, Unit 2 9 (Nov, 2007).

29. M. A. Marti-Renom *et al.*, *Annu Rev Biophys Biomol Struct* **29**, 291 (2000).

30. A. Sali, T. L. Blundell, *J Mol Biol* **234**, 779 (Dec 5, 1993).

31. A. A. Canutescu, A. A. Shelenkov, R. L. Dunbrack, Jr., *Protein Sci* **12**, 2001 (Sep, 2003).

32. B. Qian *et al.*, *Nature* **450**, 259 (Nov 8, 2007).

33. P. Bradley, K. M. Misura, D. Baker, *Science* **309**, 1868 (Sep 16, 2005).

34. W. Wriggers, R. A. Milligan, J. A. McCammon, *J Struct Biol* **125**, 185 (Apr-May, 1999).

35. W. Wriggers, S. Birmanns, *J Struct Biol* **133**, 193 (Feb-Mar, 2001).

36. A. M. Roseman, *Acta Crystallogr D Biol Crystallogr* **56**, 1332 (Oct, 2000).

37. J. A. Velazquez-Muriel, J. M. Carazo, *J Struct Biol* **158**, 165 (May, 2007).

38. J. A. Velazquez-Muriel, M. Valle, A. Santamaria-Pang, I. A. Kakadiaris, J. M. Carazo, *Structure* **14**, 1115 (Jul, 2006).

39. F. Tama, O. Miyashita, C. L. Brooks, 3rd, *J Struct Biol* **147**, 315 (Sep, 2004).

40. J. A. Velazquez-Muriel, C. O. Sorzano, S. H. Scheres, J. M. Carazo, *J Mol Biol* **345**, 759 (Jan 28, 2005).

41. M. Topf, M. L. Baker, B. John, W. Chiu, A. Sali, *J Struct Biol* **149**, 191 (Feb, 2005).

42. M. Topf, M. L. Baker, M. A. Marti-Renom, W. Chiu, A. Sali, *J Mol Biol* **357**, 1655 (Apr 14, 2006).

43. R. Bonneau *et al.*, *J Mol Biol* **322**, 65 (Sep 6, 2002).

44. W. Jiang, S. J. Ludtke, *Curr Opin Struct Biol* **15**, 571 (Oct, 2005).

45. W. Jiang, M. L. Baker, S. J. Ludtke, W. Chiu, *Journal of Molecular Biology* **308**, 1033 (May 18, 2001).

46. A. Nakagawa *et al.*, *Structure* **11**, 1227 (Oct, 2003).

47. Z. H. Zhou *et al.*, *Nat Struct Biol* **8**, 868 (Oct, 2001).

48. O. Dror, K. Lasker, R. Nussinov, H. Wolfson, *Acta Crystallographica Section D-Biological Crystallography* **63**, 42 (Jan, 2007).

49. K. Lasker, O. Dror, M. Shatsky, R. Nussinov, H. J. Wolfson, *Ieee-Acm Transactions on Computational Biology and Bioinformatics* **4**, 28 (Jan-Mar, 2007).

50. A. Dal Palu, J. He, E. Pontelli, Y. Lu, *Comput Syst Bioinformatics Conf*, 89 (2006).

51. Y. Kong, J. Ma, *J Mol Biol* **332**, 399 (Sep 12, 2003).

52. Y. Kong, X. Zhang, T. S. Baker, J. Ma, *J Mol Biol* **339**, 117 (May 21, 2004).

53. S. J. Ludtke *et al.*, *Structure* **16**, 441 (Mar, 2008).

54. S. J. Fleishman, S. Harrington, R. A. Friesner, B. Honig, N. Ben-Tal, *Biophys J* **87**, 3448 (Nov, 2004).

55. F. E. Cohen, T. J. Richmond, F. M. Richards, *J Mol Biol* **132**, 275 (Aug 15, 1979).

56. F. DiMaio, M. D. Tyka, M. L. Baker, W. Chiu, D. Baker, *J Mol Biol* **392**, 181 (Sep 11, 2009).

57. A. Korostelev, R. Bertram, M. S. Chapman, *Acta Crystallogr D Biol Crystallogr* **58**, 761 (May, 2002).

58. M. Topf *et al.*, *Structure* **16**, 295 (Feb, 2008).

59. L. G. Trabuco, E. Villa, K. Mitra, J. Frank, K. Schulten, *Structure* **16**, 673 (May, 2008).

60. B. Bottcher, S. A. Wynne, R. A. Crowther, *Nature* **386**, 88 (Mar 6, 1997).

61. J. F. Conway *et al.*, *Nature* **386**, 91 (Mar 6, 1997).

62. C. R. Booth *et al.*, *Journal of Structural Biology* **147**, 116 (Aug, 2004).

63. A. G. Martin *et al.*, *Journal of Molecular Biology* **366**, 1332 (Mar 2, 2007).

64. G. W. Min, H. B. Wang, T. T. Sun, X. P. Kong, *Journal of Cell Biology* **173**, 975 (Jun 19, 2006).

65. X. Zhang, S. B. Walker, P. R. Chipman, M. L. Nibert, T. S. Baker, *Nature Structural Biology* **10**, 1011 (Dec, 2003).

66. Serysheva, II *et al.*, *Proc Natl Acad Sci U S A* **105**, 9610 (Jul 15, 2008).

67. E. Villa *et al.*, *Proc Natl Acad Sci U S A* **106**, 1063 (Jan 27, 2009).

68. W. Jiang *et al.*, *Nature* **451**, 1130 (Feb 28, 2008).

69. X. Yu, L. Jin, Z. H. Zhou, *Nature* **453**, 415 (May 15, 2008).

70. X. Zhang *et al.*, *Proc Natl Acad Sci U S A* **105**, 1867 (Feb 12, 2008).

71. S. Lindert, P. L. Stewart, J. Meiler, *Curr Opin Struct Biol*, (Mar 30, 2009).

72. M. G. Rossmann, *Acta Crystallographica Section D-Biological Crystallography* **56**, 1341 (Oct, 2000).

73. F. Tama, O. Miyashita, C. L. Brooks, 3rd, *J Mol Biol* **337**, 985 (Apr 2, 2004).

74. M. Topf, A. Sali, *Curr Opin Struct Biol* **15**, 578 (Oct, 2005).

75. N. Volkmann, D. Hanein, *J Struct Biol* **125**, 176 (Apr-May, 1999).

76. C. A. Rohl, C. E. Strauss, K. M. Misura, D. Baker, *Methods Enzymol* **383**, 66 (2004).

77. K. T. Simons *et al.*, *Proteins* **34**, 82 (Jan 1, 1999).

78. K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *J Mol Biol* **268**, 209 (Apr 25, 1997).

79. C. A. Rohl, C. E. Strauss, D. Chivian, D. Baker, *Proteins* **55**, 656 (May 15, 2004).

80. K. M. Misura, D. Baker, *Proteins* **59**, 15 (Apr 1, 2005).

81. O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, D. Baker, *Science* **310**, 638 (Oct 28, 2005).

82. P. M. Bowers, C. E. Strauss, D. Baker, *J Biomol NMR* **18**, 311 (Dec, 2000).

83. J. Meiler, D. Baker, *Proc Natl Acad Sci U S A* **100**, 15404 (Dec 23, 2003).

84. J. Meiler, D. Baker, *J Magn Reson* **173**, 310 (Apr, 2005).

85. C. A. Rohl, D. Baker, *J Am Chem Soc* **124**, 2723 (Mar 20, 2002).

86. N. Alexander, M. Bortolus, A. Al-Mestarihi, H. McHaourab, J. Meiler, *Structure* **16**, 181 (Feb, 2008).

87. S. M. Hanson *et al.*, *Structure* **16**, 924 (Jun, 2008).

88. R. Bonneau, I. Ruczinski, J. Tsai, D. Baker, *Protein Sci* **11**, 1937 (Aug, 2002).

89. J. A. Kovacs, M. Yeager, R. Abagyan, *Biophys J* **93**, 1950 (Sep 15, 2007).

90. J. J. Ruprecht, T. Mielke, R. Vogel, C. Villa, G. F. Schertler, *EMBO J* **23**, 3609 (Sep 15, 2004).

91. J. Li, P. C. Edwards, M. Burghammer, C. Villa, G. F. Schertler, *J Mol Biol* **343**, 1409 (Nov 5, 2004).

92. J. J. Rux, P. R. Kuser, R. M. Burnett, *J Virol* **77**, 9553 (Sep, 2003).

93. V. D. Sood, D. Baker, *J Mol Biol* **357**, 917 (Mar 31, 2006).

94. N. Grigorieff, *Journal of Structural Biology* **157**, 117 (Jan, 2007).

95. G. F. Schroder, A. T. Brunger, M. Levitt, *Structure* **15**, 1630 (Dec, 2007).

96. S. Lindert, M. Silvestry, T. M. Mullen, G. R. Nemerow, P. L. Stewart, *J Virol* **83**, 11491 (Nov, 2009).

97. P. L. Stewart, G. R. Nemerow, *Trends in microbiology* **15**, 500 (Nov, 2007).

98. R. O. Hynes, *Cell* **48**, 549 (Feb 27, 1987).

99. J. Qin, O. Vinogradova, E. F. Plow, *PLoS Biol.* **2**, e169 (Jun, 2004).

100. T. Hato, N. Pampori, S. J. Shattil, *J. Cell Biol.* **141**, 1685 (Jun 29, 1998).

101. S. Miyamoto, S. K. Akiyama, K. M. Yamada, *Science (New York, N.Y* **267**, 883 (Feb 10, 1995).

102. S. Miyamoto *et al.*, *J. Cell Biol.* **131**, 791 (Nov, 1995).

103. K. M. Yamada, S. Miyamoto, *Current opinion in cell biology* **7**, 681 (Oct, 1995).

104. E. Li, D. Stupack, G. M. Bokoch, G. R. Nemerow, *J. Virol.* **72**, 8806 (Nov, 1998).

105. E. Li, D. Stupack, R. Klemke, D. A. Cheresh, G. R. Nemerow, *J. Virol.* **72**, 2055 (Mar, 1998).

106. P. L. Stewart *et al.*, *EMBO J.* **16**, 1189 (Mar 17, 1997).

107. T. J. Wickham, P. Mathias, D. A. Cheresh, G. R. Nemerow, *Cell* **73**, 309 (Apr 23, 1993).

108. N. J. Philpott, M. Nociari, K. B. Elkon, E. Falck-Pedersen, *Proc. Natl. Acad. Sci. USA* **101**, 6200 (Apr 20, 2004).

109. M. S. Rajala, R. V. Rajala, R. A. Astley, A. L. Butt, J. Chodosh, *J. Virol.* **79**, 12332 (Oct, 2005).

110. B. H. Luo, T. A. Springer, *Current opinion in cell biology* **18**, 579 (Oct, 2006).

111. J. Zhu, B. Boylan, B. H. Luo, P. J. Newman, T. A. Springer, *J. Biol. Chem.* **282**, 11914 (Apr 20, 2007).

112. J. P. Xiong *et al.*, *Science (New York, N.Y* **294**, 339 (Oct 12, 2001).

113. J. Zhu *et al.*, *Mol. Cell.* **32**, 849 (Dec 26, 2008).

114. O. Vinogradova *et al.*, *Cell* **110**, 587 (Sep 6, 2002).

115. J. Zhu *et al.*, *Blood* **110**, 2475 (Oct 1, 2007).

116. M. A. Arnaout, B. Mahalingam, J. P. Xiong, *Annual review of cell and developmental biology* **21**, 381 (2005).

117. L. Xing *et al.*, *J. Biol. Chem.* **279**, 11632 (Mar 19, 2004).

118. C. Y. Chiu, P. Mathias, G. R. Nemerow, P. L. Stewart, *J. Virol.* **73**, 6759 (Aug, 1999).

119. P. Mathias, M. Galleno, G. R. Nemerow, *J. Virol.* **72**, 8669 (Nov, 1998).

120. E. Wu *et al.*, *J. Virol.* **78**, 3897 (Apr, 2004).

121. S. D. Saban, R. R. Nepomuceno, L. D. Gritton, G. R. Nemerow, P. L. Stewart, *J. Mol. Biol.* **349**, 526 (Jun 10, 2005).

122. J. Shi, D. R. Williams, P. L. Stewart, *J. Struct. Biol.* **164**, 166 (Oct, 2008).

123. U. Adiga *et al.*, *J. Struct. Biol.* **152**, 211 (Dec, 2005).

124. J. A. Mindell, N. Grigorieff, *J. Struct. Biol.* **142**, 334 (Jun, 2003).

125. N. Grigorieff, *J. Struct. Biol.* **157**, 117 (Jan, 2007).

126. E. F. Pettersen *et al.*, *J. Comput. Chem.* **25**, 1605 (Oct, 2004).

127. P. Chacon, W. Wriggers, *J. Mol. Biol.* **317**, 375 (Mar 29, 2002).

128. G. N. Ramachandran, A. S. Kolaskar, C. Ramakrishnan, V. Sasisekharan, *Biochim. Biophys. Acta* **359** 298 (1974).

129. C. M. Wiethoff, H. Wodrich, L. Gerace, G. R. Nemerow, *J. Virol.* **79**, 1992 (Feb, 2005).

130. P. R. Kolatkar *et al.*, *EMBO J.* **18**, 6249 (Nov 15, 1999).

131. C. Xiao *et al.*, *Structure* **13**, 1019 (Jul, 2005).

132. D. M. Belnap *et al.*, *Proc. Natl. Acad. Sci. USA* **97**, 73 (Jan 4, 2000).

133. Y. He *et al.*, *Proc. Natl. Acad. Sci. USA* **97**, 79 (Jan 4, 2000).

134. L. Xing *et al.*, *The EMBO journal* **19**, 1207 (Mar 15, 2000).

135. B. H. Luo, C. V. Carman, T. A. Springer, *Annu. Rev. Immunol.* **25**, 619 (2007).

136. M. Y. Nakano, K. Boucke, M. Suomalainen, R. P. Stidwill, U. F. Greber, *Journal of virology* **74**, 7085 (Aug, 2000).

137. U. F. Greber, M. Willetts, P. Webster, A. Helenius, *Cell* **75**, 477 (Nov 5, 1993).

138. D. R. Williams, K. J. Lee, J. Shi, D. J. Chen, P. L. Stewart, *Structure* **16**, 468 (Mar, 2008).

139. B. L. Sibanda, D. Y. Chirgadze, T. L. Blundell, *Nature* **463**, 118 (Jan 7, 2010).