PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Challenging the state of the art in protein structure prediction: Highlights of experimental target structures for the 10th Critical Assessment of Techniques for Protein Structure Prediction Experiment CASP10

Andriy Kryshtafovych,[1] John Moult,[2,3] Patrick Bales,[3] J. Fernando Bazan,[4,5] Marco Biasini,[6,7] Alex Burgin,[8] Chen Chen,[3] Frank V. Cochran,[9] Timothy K. Craig,[10] Rhiju Das,[9,11] Deborah Fass,[12] Carmela Garcia-Doval,[13] Osnat Herzberg,[3,14] Donald Lorimer,[10] Hartmut Luecke,[15,16,19,20] Xiaolei Ma,[4,5] Daniel C. Nelson,[3,17] Mark J. van Raaij,[13] Forest Rohwer,[18] Anca Segall,[18] Victor Seguritan,[18] Kornelius Zeth,[19,20] and Torsten Schwede[6,7]*

[1] Genome Center, University of California, Davis, California 95616
[2] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, Maryland 20742
[3] Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Maryland 20850
[4] Department of Protein Engineering, Genentech, South San Francisco, California 94080
[5] Department of Structural Biology, Genentech, South San Francisco, California 94080
[6] Biozentrum, University of Basel, Basel 4056, Switzerland
[7] SIB Swiss Institute of Bioinformatics, Basel 4056, Switzerland
[8] Broad Institute, Cambridge, Massachusetts 02142
[9] Department of Biochemistry, Stanford University, Stanford, California 94305
[10] Emerald Bio, Bainbridge Isle, Washington 98110
[11] Department of Physics, Stanford University, Stanford, California 94305
[12] Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel
[13] Centro Nacional de Biotecnologia (CNB-CSIC), Madrid E-28049, Spain
[14] Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742
[15] Department of Biochemistry and Biophysics, Center for Biomembrane Systems, University of California, Irvine, California 92697-3900
[16] Department of Computer Science, Center for Biomembrane Systems, University of California, Irvine, California 92697-3900
[17] Department of Veterinary Medicine, University of Maryland, College Park, Maryland 20742
[18] Department of Biology, San Diego State University, San Diego, California 92182
[19] Unidad de Biofisica (CSIC-UPV/EHU), Bizkaia Spain
[20] IKERBASQUE, Basque Foundation for Science, Bilbao Spain

ABSTRACT

For the last two decades, CASP has assessed the state of the art in techniques for protein structure prediction and identified areas which required further development. CASP would not have been possible without the prediction targets provided by

Additional Supporting Information may be found in the online version of this article.
Abbreviations: CASP, community-wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction; gp, gene product; LPS, lipopolysaccharide

Grant sponsor: US National Institutes of Health (NIH); Grant number: R01AI78000; Grant sponsor: Spanish Ministry of Economy and Competitiveness; Grant number: BFU2011–24843; Grant sponsor: US National Institutes of General Medical Sciences (NIGMS/NIH); Grant number: R01GM100482; Grant sponsors: UC Irvine Center for Biomembrane Systems, Spanish Ministry of Education.
J. Fernando Bazan's current address is 44th & Aspen Life Sciences, 924 4th St. N., Stillwater, Minnesota 55082.
Xiaolei Ma's current address is Novartis Institutes for Biomedical Research, 4560 Horton St., Emeryville, California 94608.
*Correspondence to: Torsten Schwede, Biozentrum, University of Basel, Klingelbergstrasse 50, Basel 4056, Switzerland. E-mail: Torsten.Schwede@unibas.ch
Received 8 June 2013; Revised 1 November 2013; Accepted 9 November 2013
Published online 8 December 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24489

the experimental structural biology community. In the latest experiment, CASP10, more than 100 structures were suggested as prediction targets, some of which appeared to be extraordinarily difficult for modeling. In this article, authors of some of the most challenging targets discuss which specific scientific question motivated the experimental structure determination of the target protein, which structural features were especially interesting from a structural or functional perspective, and to what extent these features were correctly reproduced in the predictions submitted to CASP10. Specifically, the following targets will be presented: the acid-gated urea channel, a difficult to predict transmembrane protein from the important human pathogen *Helicobacter pylori*; the structure of human interleukin (IL)−34, a recently discovered helical cytokine; the structure of a functionally uncharacterized enzyme OrfY from *Thermoproteus tenax* formed by a gene duplication and a novel fold; an ORFan domain of *mimivirus* sulfhydryl oxidase R596; the fiber protein gene product 17 from bacteriophage T7; the bacteriophage CBA-120 tailspike protein; a virus coat protein from metagenomic samples of the marine environment; and finally, an unprecedented class of structure prediction targets based on engineered disulfide-rich small proteins.

Proteins 2014; 82(Suppl 2):26–42.

Key words: X-ray crystallography; NMR; protein structure prediction; critical assessment; CASP; model quality.

INTRODUCTION

For the last two decades, the community-wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP) has assessed the state of the art in protein structure prediction, documented the progress, and identified areas which require further development of improved methods. The experiment is based on "blind prediction," that is, at the time of modeling the experimental structure has not yet been published. Thereby, CASP depends on the experimental structural biology community to suggest protein sequences as prediction targets, for which an experimental structure will become available in the near future. Over the last 20 years, the experimental structural biology community has contributed more than 700 protein structures as prediction targets. CASP would not have been possible without this fruitful collaboration. For CASP10, more than 130 sequences were suggested by the experimental community, and 114 were selected by the CASP organizers as prediction targets during the prediction season. In addition to this, 18 targets have been submitted to CASP_Roll, a newly introduced version of the CASP experiment aiming at assessing predictions for challenging remote homology/de novo targets all year round.[1] We hope that this will motivate the development of advanced methods in this area.

Selection of targets for experimental structure determination is typically motivated by a specific research question. In this article, the experimentalists providing targets for CASP10 present selected target highlights and discuss which aspects of the targets were specifically interesting and to what extent they were correctly reproduced in the predictions. We hope that this type of manuscript, which was introduced in CASP9,[2] will help the structure prediction community to better understand which features of a structure are important from the point of view of crystallographers and NMR spectroscopists and how these features should be taken into account to develop better prediction tools. The article can also be of interest for future CASP assessors to decide which additional aspects of a structure require special attention in the assessment.

The article reflects the views of the contributing authors on selected challenging CASP10 targets. Specifically, the following proteins will be discussed: the acid-gated urea channel, a difficult to predict transmembrane protein from the important human pathogen *Helicobacter pylori*; the structure of human interleukin (IL)−34, a helical cytokine in the twilight zone; the structure of a functionally uncharacterized enzyme OrfY from *Thermoproteus tenax* formed by a gene duplication and a novel fold; an ORFan domain of mimivirus sulfhydryl oxidase R596; the fiber protein gene product (gp) 17 from bacteriophage T7; the bacteriophage CBA-120 tailspike protein; a phage coat protein from the marine environment isolated by metagenomics; and finally, an unprecedented class of structure prediction targets based on engineered disulfide-rich small proteins.

For each target protein, the prediction center website provides a numerical analysis of the submitted models (http://www.predictioncenter.org) using standard measures such as GDT,[3] local Distance Difference Test (lDDT),[4] Dali,[5] SphereGrinder,[6] CAD,[7] or RPF[8] scores. The results of the detailed evaluation by the human assessors in the FM[9] and TBM[8] categories are discussed in dedicated manuscripts elsewhere in this issue.

THE ACID-GATED UREA CHANNEL FROM *H. PYLORI* (T0666, PDB: 3UX4; HARTMUT LUECKE)

Approximately 50% of the world's population is chronically infected with the neutralophilic pathogen *H. pylori*.[10] Infection with this bacterium produces gastric inflammation and predisposes infected individuals to the development of both peptic ulcer disease (10–20% of infected

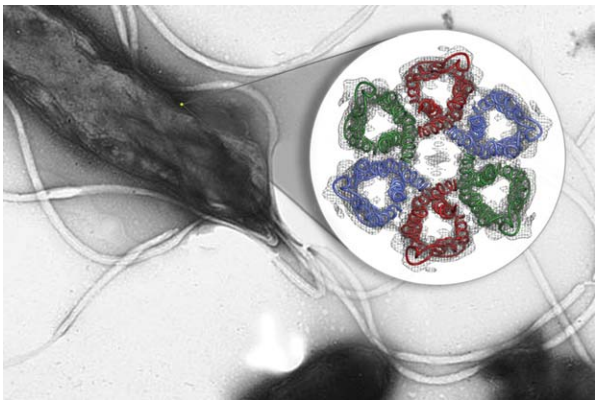

Figure 1

This composite shows the enlarged membrane-embedded hexameric ring of urea channels next to an electron micrograph of a *H. pylori* cell. Urea passes through the center of each of the six channel molecules (two green, two red, and two blue molecules). The center of the ring is filled with a lipid bilayer plug. [Credit: Hartmut Luecke (UC, Irvine) and Andy Freeberg (SLAC National Accelerator Laboratory)].

individuals) and gastric adenocarcinoma (up to 1% of infected individuals).[11]

In recent years, triple and even quadruple therapy with broadband antibiotics has been suffering increasingly from resistance (up to 30% of eradication regimens fail). The pathogen's proton-gated urea channel, HpUreI, is essential for gastric infection, making it a target for specific eradication.[12] HpUreI allows rapid urea entry from the gastric juice into the cytoplasm where urease generates NH_3 and CO_2 that buffer the periplasmic space to pH $\sim$ 6.1 even at a medium pH of $<$2.5.

It is well known that membrane proteins are notoriously difficult to crystallize: to date, the atomic structures of just over 1300 membrane proteins are known (vs. over 95,000 soluble protein structures). Crystallization optimization and structure determination of this membrane protein of 195 residues were particularly challenging and required a multilaboratory effort of over 5 years.[13] The structure reveals a novel fold that assembles into a hexameric ring of protomers surrounding a central lipid bilayer plug. Each protomer forms an hourglass-shaped channel within a twisted bundle of six transmembrane helices (TMHs) in a novel fold, a two-helix hairpin motif repeated three times. The urea pathway is defined entirely by side chains that are predominantly hydrophobic with several tryptophans in key positions. The side chains belong to TMHs 1, 3, and 5 and are highly conserved in the AmiS/UreI superfamily of channels (Fig. 1). Constrictions above and below conserved Glu177 represent the proton rejection and urea selectivity filters.[13–15] A major component of the gating machinery resides in periplasmic loop 2, shown on the periphery of the hexamer in Figure 1.

Perhaps not surprisingly, the prediction target with the number 666 proved to be devilishly difficult (see: "Number of the Beast," http://en.wikipedia.org/wiki/Number_of_the_Beast). The best predictions for this target correctly feature six TMHs. The only prediction closely resembling the native structure of a loosely packed twisted bundle of six helices with the urea channel through its center (Fig. 1) is TS079_1 (TASSER). Even though this is the best prediction according to the FM assessment,[9] only less than one-third of its residues are in proximity to the corresponding residues in the crystal structure[3] (GDT_TS = 31.1%). Two other main groups of predictions either display a bundle of five helices packed around a central helix (e.g., TS035_1) or a two-layer structure with three helices in each layer (e.g., TS237_1). All predictions on this target quite loosely align with the experimental structure, with the best 25 predictions (according to the GDT_TS score) showing all-atom RMSD of 8.5–9.0 Å and GDT_TS of 32.6–33.8.

One reason for the failure to predict the correct fold of the HpUreI protomer may be the hexameric arrangement of protomers observed in the crystal structure, which was shown to predominate in solution as well.[13] Unrestrained molecular dynamic (MD) simulations with the entire HpUreI hexamer in an explicit lipid bilayer show this arrangement to be stable for more than 1000 ns, whereas equivalent MD simulations of a single protomer show it collapsing in a few 100 ns.[15] Thus, it seems likely that a single protomer of HpUreI does not possess a stable fold, suggesting a structural role for the hexameric scaffold perhaps by holding the loosely packed helices of the six-helix bundle of each protomer in the correct relative positions to form a urea channel that is impermeable to protons. Another function of the hexamer might be cooperativity in gating.

IL-34: A HELICAL CYTOKINE IN THE TWILIGHT ZONE (R0007, PDB:4DKC, 4DKF; XIAOLEI MA AND J. FERNANDO BAZAN)

The signaling functions of a remarkably diverse group of helical cytokines appear to be inextricably linked to a unique superfamily of hemopoietic receptors; these latter molecules share a binding scaffold evolutionarily designed to engage the conserved helical ligand fold.[16] Still, exceptions arise! A small clan of helical cytokines has escaped the tyranny of the hemopoietic receptors, and instead signal through three Class III receptor tyrosine kinases (RTKs), Fms/CSF-1R, Kit, and Flt3, relatives of the Class V RTKs for Cys-knot growth factors PDGF and VEGF.[17] The three rogue helical cytokines, CSF-1, SCF, and Flt3L, are distinctively membrane-bound and form head-to-head homodimers.[18] Recently, a novel secreted ligand for CSF-1R was isolated from a large-

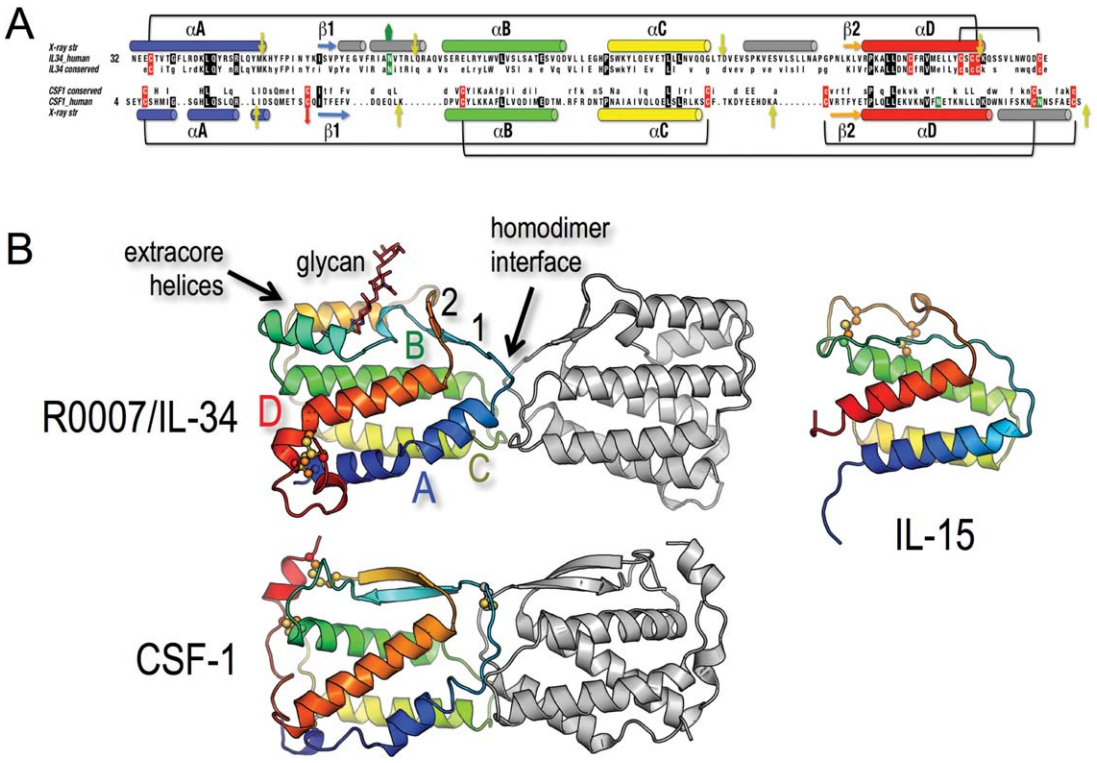

Figure 2

Helical cytokine fold of human IL-34. **A**: Comparison of IL-34 (PDB: 4DKC) and "best template" CSF-1 (PDB: 3UF2) chains by SSM superposition,[25] with scant identities boxed in reverse lettering. PSIPRED-defined conserved sequences (with capital letters indicating nearly invariant residues) are adjacent to the X-ray chains. Cys residues are boxed in red, and disulfide links are noted; the CSF-1 Cys involved in an intermolecular disulfide bridge is marked with a red arrow. The N-gly site in IL-34 is boxed in green and marked by a topmost green pentagon. Exon junctions in the corresponding IL-34 and CSF-1 genes are mapped to the chains with yellow arrows, guiding their alignment.[18,22] The four core helices are color ramped blue to red (and labeled A–D), β-strands are named β1 and β2, and extracore helices noted in gray. **B**: Human IL-34 (target R0007) and its best template, CSF-1, are aligned; the chains color-ramped blue to red, and secondary structure labeled as in panel A. Disulfide bridges and the IL-34 glycan chain are highlighted. The compact IL-15 (PDB: 2Z3Q) is the closest short-chain helical cytokine[23] match to IL-34 by SSM search (2.3 Å RMSD). The structures were drawn by Pymol (www.pymol.org).

scale functional screen and is designated as IL-34.[19] Bearing no discernible sequence similarity to CSF-1, the mechanism of receptor sharing by IL-34 sparked our interest in pursuing the new complex structure and was compared with the known CSF-1/CSF-1R assembly.[20] We further intuited that IL-34 likely shared a helical fold with CSF-1, as the register of PSIPRED-located helices[21] fit the previously noted watermark of exon-encoded structural elements in CSF-1, SCF, and Flt3L,[18,22] and might critically shed light—as the outlier, secreted member of the clan—on their ancient divergence from the short-chain helical cytokine group.[23] In this vein, IL-34 as target R0007 in CASP10-ROLL proved to be one of the more challenging structures to predict.

The antiparallel four-helix bundle structure of dimeric human IL-34 was captured in solution (at 1.85 Å resolution; PDB: 4DKC), in complex with the three N-terminal Ig domains of CSF-1R (3 Å resolution; PDB: 4DKF) and also bound to two therapeutic antibody FAB fragments (7); the homologous mouse IL-34 complex structure is architecturally similar.[24] As assessed by Secondary Structure Matching (SSM),[25] the monomer structure of IL-34 was most closely similar to SCF (2.75 Å RMSD) with 7% chain identity over 103 superposed residues, and next matched CSF-1 (3.1 Å RMSD, 14% ID over 108 aligned residues) and Flt3L (3.6 Å RMSD, 9% ID over 92 residues). As the R0007 target sequence would have perhaps drawn attention to IL-34 as a cytokine and to CSF-1 as its competitor for CSF-1R binding, we chose to capture CSF-1 as the most likely sought template for an IL-34 model [Fig. 2(A)] and thereby highlight a few hurdles for the predictors. (a) Compounding the sparse degree of chain identity, (b) the disulfide bridge pattern in CSF-1 (notably shared with SCF and Flt3L) is distinct from the deduced links in IL-34; two Cys residues remain unpaired in IL-34, whereas CSF-1 uses a free Cys for an intermolecular disulfide link.[20,22] (c) Unexpectedly, the N-linked glycan attached to Asn76 serves as an integral part of the IL-34 structure (by interacting with residues in the β1–αB loop) and is essential for proper folding.[22]

(d) The constitutive homodimer interface in IL-34 conserves hydrophobic contacts from two top-mounted loops (αA–β1 and αB–αC), which do not otherwise contribute to the core monomer fold.[22]

IL-34 is resolutely related to short-chain helical cytokine folds [Fig. 2(B)][22]; however, the greater length of the globular chain (160 residues when compared with the typical 120–130 amino acids) is absorbed by three extracore helices in the long loops connecting αA–αB and αC–αD, which pack against the core helix bundle, and perhaps overshadow the two short β-strands β1 and β2, which form a compact antiparallel sheet between loops.[22] (e) As a result, secondary structure predictions of the IL-34 chain will find seven consecutive helices separated by short hairpin loops and mistakenly point the fold recognition algorithms toward larger antiparallel helical arrays. (f) The plastic engagement of CSF-1R by IL-34 and CSF-1 involves only the αA faces of the dimer helical scaffolds, and thus, the αD side of the short-chain-like cytokine fold is unusually degenerated in the rogue cytokines.[16] However, IL-34 is a surprising exception, as the receptor-free face is highly conserved (and targeted by a nonblocking antibody) for a yet undiscovered functional purpose that is not shared with CSF-1 [Fig. 2(B)].[22]

Among the CASP10 predictions, models by five groups appear numerically better than the rest (GDT > 45). The best prediction according to GDT_TS (49.07) was submitted by the group "BAKER" and shows an all-atom RMSD of 6.3 Å. This model correctly predicts the overall topology of the cytokine helix bundle; however, the relative orientation of the secondary structure elements is not always preserved resulting in significant misalignments: according to the sequence-independent LGA, overall, only 20% of the residues in the model correctly align with the reference structure in a superposition generated with a 4 Å distance cutoff. The model fails to reproduce some specific characteristics of IL-24 such as the unusual disulfide connectivity and the arrangement as homodimer. Three groups (FOLDIT, Anthropic_-Dreams, and Void_Crushers) aimed at predicting the oligomeric state and submitted five models as dimers. However, in most cases, the arrangement does not reflect the correct quaternary structure. One model (group Anthropic_Dreams, Model 5) at least partially traced the overall orientation and interaction interface (residues 51–64 and 106–116) of the native structure.

Cytokine folds are notoriously plastic to sequence divergence, and receptors painstakingly accommodate these recognition and specificity challenges by various mechanisms.[22] As noted in the case of R0007, the prediction challenge begins with the negligible sequence similarity to other folds and is then compounded by a host of other issues (e.g., divergent disulfide patterns, extracore structures, and variant conservation patterns). Successful mapping of IL-34 to helical cytokine folds best followed a functional approach that winnowed down the candidates to other cytokines and then applied family-specific constraints, such as delineation of the core helices by their genetic watermark,[18,22] to find the best rogue cytokine template.

GENE DUPLICATION AND A NOVEL FOLD FORM THE STRUCTURAL BASIS OF A FUNCTIONALLY UNKNOWN ENZYME ORFY FROM *T. TENAX* (T0734, PDB: 3ZPW; KORNELIUS ZETH)

OrfY is a protein of unknown function derived from the hyperthermophilic archaeum *T. tenax*. This protein was studied due to its co-occurrence with the treS/P protein in an operon structure regulating the synthesis of trehalose and possible implications in trehalose metabolism. The treS/P conducts unidirectional glycosyl transferring synthase activity causing the formation of trehalose from UDP (ADP)-glucose and glucose.[26] Because of the co-occurrence of the protein and according to the sequence-based searches that we performed, OrfY was predicted to belong to a class of bacterial transcription factors possibly activated by trehalose derivatives. Representatives of this class of proteins have previously not been characterized by structural biology methods according to sequence-based searches using HHpred.[27]

OrfY contains 216 residues yielding a molecular weight of 24 kDa. Using HHrep for the detection of repetitive sequences, a motif of ∼100 residues was observed in OrfY.[28] We crystallized OrfY to study this protein by means of X-ray crystallography for subsequent cocrystallization with sugar derivatives, to perform *in silico* docking approaches with substrate molecules, and to approach its function based on the structure. Structure determination by MIR techniques and subsequent model building resulted in the refined structure with R/R_{free}-factors of 0.18/0.22 at a resolution of 2.7 Å.

In agreement with the repeat prediction, the monomeric structure of OrfY displays two domains of similar fold and an internal pseudo-twofold axis giving a structure deviation of 2.2 Å for 100 aligned residues. The domain structure consists of an N-terminally located and extended β-strand of 5 nm length [β1 comprising residues 7–28; see Fig. 3(A)], which together with the second and sequence-related β2-strand of the homologous domain (residues 120–138) forms an unusually extended protein architecture of a DNA-like "antiparallel β-helix motif." This motif is stabilized through a continuous H-bond mediated backbone structure and allows the enwinding around the pseudo-helical axis of ∼360° [see Fig. 3(B)]. The C-terminal part of both domains is

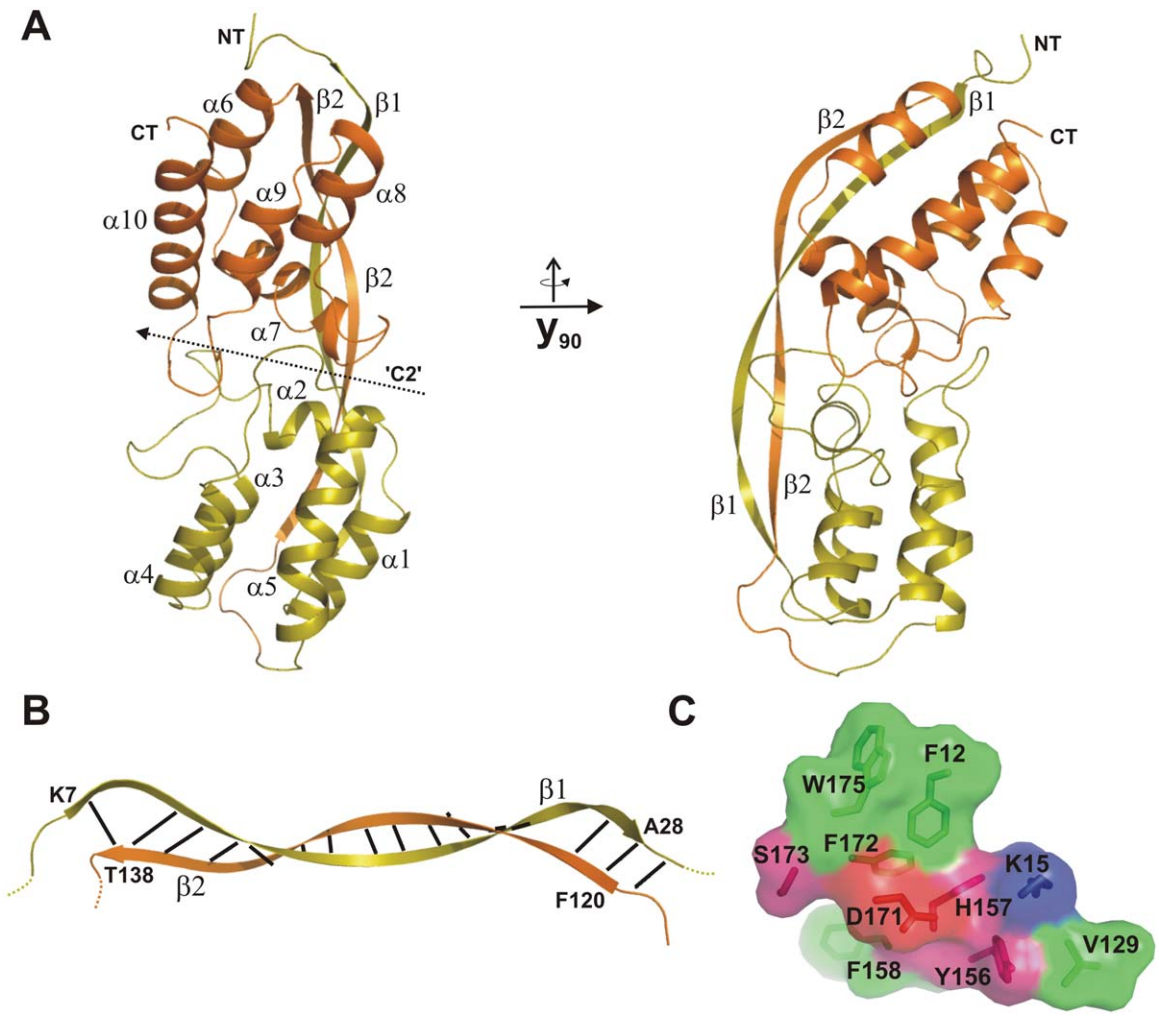

Figure 3

The structure of OrfY from *T. tenax*. **A**: Structure of the monomeric protein in cartoon representation with the two homologous domains color coded in orange and splitpea green. The secondary structure elements are depicted and assigned with β1–β2 for the strands and α1–α10 for the helices. The structure is shown in two different orientations related to each other by a rotation of 90° around the *Y*-axis. The pseudo-twofold symmetric axis is indicated by a dashed arrow and C2. **B**: The extended antiparallel β-helix motif is shown formed by the β1 and β2 strands and the H-bond pattern is indicated. **C**: Conserved residues forming the presumptive sugar-binding site and mapped on the structure in surface representation.

formed by a globular and entirely helical structure comprising five α-helices (α1–α5 and α6–α10). The structure in the asymmetric unit is a tetramer; however, the biological unit is only represented by a dimer that is stabilized by a large number of salt bridges and H-bonds yielding an interface of 1500 Å^2 (12% of entire protein surface). Using a multiple alignment of sequences, we identified 10 conserved residues clustering at the interface between two molecules. These residues are predominantly of aromatic nature and surround a conserved Asp and Lys residue [see Fig. 3(C)]. Interestingly, the conservation of residues in the sequence has been observed only once despite the duplication of the sequence and the pseudo-twofold symmetry of the protein monomer. Using the structure data, we mapped these conserved residues onto the accessible surface whereby the hypothetical binding area of sugar (trehalose) molecules was identified. Residues of both domains contribute to the formation of this cavity including two N-terminal residues (Phe12 and Lys15) of the β1-strand and eight residues located in the globular fold (mostly on α7 and α8 helices). Cocrystallization with glucose yielded a positive electron density in the putative binding groove (unpublished data), and docking analysis using monosaccharide and disaccharide sugar structures including trehalose further confirmed this cavity to resemble the major binding site of the putative transcription factor. In summary, we learned that OrfY represents a repetitive structure of two novel domains that evolutionarily diverged to form one binding site for monosaccharides and disaccharides. Fold

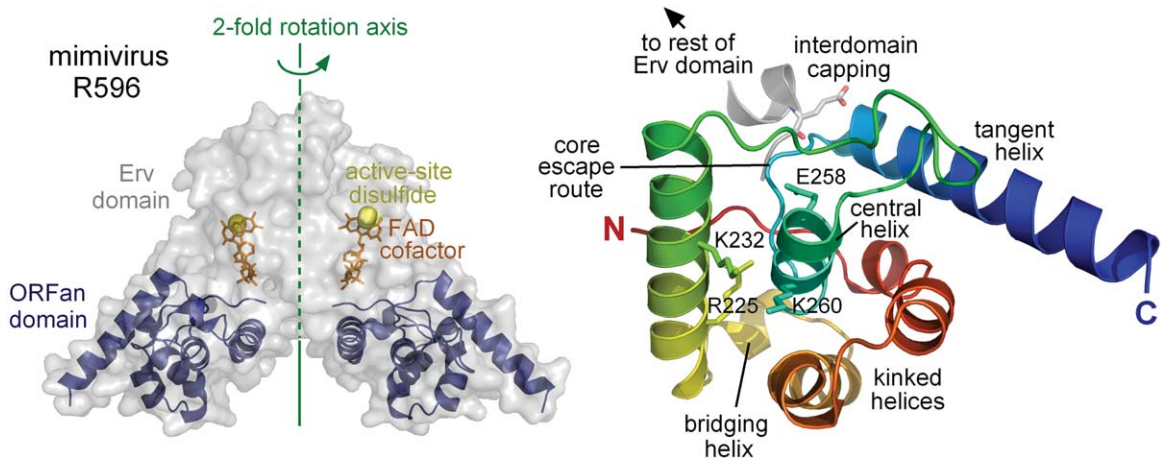

Figure 4

Hard-knock helices in the ORFan domain of mimivirus sulfhydryl oxidase R596. Left, a molecular surface representation of the mimivirus disulfide catalyst R596 dimer is shown with the ORFan domains highlighted as dark blue ribbons. The flavin adenine dinucleotide (FAD) cofactor is in orange sticks, and the sulfur atoms in the FAD-proximal, redox-active disulfide bonds are shown as yellow spheres. Right, the features of the ORFan domain described in the text are labeled on the domain structure. Charged side chains arising from buried positions in ORFan domain helices are shown in stick representation and labeled. A small fragment of the adjacent Erv domain is shown in gray, displaying the glutamate residue that caps the ORFan terminal helix.

analysis of the structure using the DALI program did not indicate any related 3D folds and confirms our initial HHpred results. Missing structural homologs in the PDB database also had implications for the prediction of the target by the CASP10 participating groups. None of the CASP10 predictions resemble the target structure—the best predictions reached only GDT of 20 and all-atom RMSD around 20 Å. Prediction of OrfY by the CASP groups has presumably failed for two reasons: the protein size that hampers a reliable *de novo* prediction and missing structural homologs for proper homology modeling.

HARD-KNOCK HELICES IN THE ORFAN DOMAIN OF MIMIVIRUS SULFHYDRYL OXIDASE R596 (T0737, PDB:3TD7; DEBORAH FASS)

The Erv fold—a five-helix bundle with an embedded flavin adenine dinucleotide cofactor and a redox-active dicysteine motif—is a compact module dedicated to catalyzing the formation of disulfide bonds.[29] The module is called into service for oxidative protein folding or assembly in a variety of subcellular and extracellular settings in eukaryotes. Erv family enzymes are also encoded by large double-stranded DNA viruses, such as poxviruses, and expressed in the cytosol of infected cells.[30–32]

The giant mimivirus encodes two distinct Erv fold enzymes. The precise role of the two mimivirus disulfide catalysts in viral replication or assembly is unknown, but they are distinguished from one another by their carboxy-terminal regions. One has a membrane anchor, whereas the other, designated R596, has a domain-sized extension fused to the Erv module. This extension was not recognized as a known fold and has no apparent sequence similarity to any protein outside of closely related viruses. This phenomenon is quite general in mimivirus: in addition to a large number of ORFan proteins,[33] mimivirus encodes many proteins that clearly fall into recognized fold and functional classes but are decorated by substantial stretches of additional, unrecognizable sequence. These segments may tune the functions of their host proteins or they may simply be the residue of a massive and messy replication history. An argument has been made in favor of a functional role for mimivirus ORFan proteins in general.[33] Supporting positive selection for retention of the R596 ORFan domain in particular, this domain is fused to the Erv disulfide catalyst of related giant viruses isolated from remote regions of the planet.[34] Experimental determination of the structure of the mimivirus R596 protein addressed the relationship of this unclassifiable extension to the known catalytic components of the enzyme[35] and also addressed the more general question of whether mimivirus ORFan proteins and domains are structured (Fig. 4).

CASP10 predictions for T0737 were in general poor and did not provide a realistic representation of the R596 ORFan domain. The best model, submitted by "Zhang_Ab_Initio," reached a GDT_TS of 40 (all-atom RMSD 8.5 Å), in which only about one-third of the residues can be partially aligned to the reference structure (based on a sequence-independent superposition generated at 4 Å distance cutoff). Two striking features of the

R596 ORFan domain, revealed by experimental structure determination, likely contributed to the difficulty in predicting the domain structure computationally. One feature is the unusual topology of the ORFan domain helical bundle. The central helix in the tertiary structure is far off-center in primary structure, being the second-to-last in sequence. Furthermore, rather than crossing the full length of the domain, this core, three-turn helix spans only about half the domain. At that point, the helix breaks, and the amino acid chain escapes out the side to form the final, peripheral helix, which projects tangentially from the rest of the domain. In general, secondary and tertiary structures, as well as hydrophobic and polar residue patterning, appear to be frustrated in the R596 ORFan domain. Two of the helices that pack against the truncated central helix have disrupted hydrogen bonding patterns and corresponding kinks. A single-turn helix serves as a bridge between two others. The core helix itself contains two charged residues, and another of the surrounding helices uses the aliphatic portions of basic side chains to form significant parts of its interaction surface with the hydrophobic core. This basic helix and to a lesser extent the long, straight, peripheral helix at the carboxy-terminus of the domain were predicted with moderate success by CASP competitors to be helical. However, this information did not contribute to success in tertiary structure prediction because of the general inability to pack the core of the protein or to assign the relative positions of the surrounding helices. Only in very few cases, the short central helix was placed by the predictors at the core of the domain.

The second feature that made prediction challenging is that the R596 ORFan domain is not entirely structurally independent from its host Erv domain. For example, the final helix of the ORFan domain is capped at the amino terminus by a glutamate side chain and a backbone carbonyl from the Erv domain. The two domains are intimately associated with one another, and the ORFan domain contributes to interesting shape and surface electrostatic properties in the overall quaternary structural assembly of the R596 enzyme. Specifically, the dimeric Erv and accompanying ORFan domains together generate a concave face lined with a striking collection of basic amino acid residues. The width of the cavity, about 26 Å, suggests a possible role in nucleic acid binding. However, the concave face does not contain other structural features expected of specific DNA binding proteins, and experimental evidence for such a role is currently lacking. Together, the awkward topology and the capping dependence suggest that the ORFan domain has not achieved the status of an independent folding unit, although attempts to produce the domain in isolation and to falsify this hypothesis have not yet been made. A successful structure prediction would presumably have required taking into consideration the relationship between the ORFan domain and the known structure of its neighboring Erv domain. It is not clear whether any of the prediction strategies had this capability. Given the prevalent scenario of protein segments with unknown structures juxtaposed to domains with known or readily templated structures, it seems that improved predictions of interdomain interactions would be exceptionally valuable to future structure prediction efforts.

Despite the ungainliness of the R596 ORFan domain, it nevertheless does clearly fold, at least in the context of the neighboring Erv module, to yield a crystallizable protein. One point scored for mimivirus ORFan foldability. However, the R596 structure raises another unanticipated question. Might new folds in general be born in this manner? Secondary structure accretion against a generous surface of a well-folded domain, coupled here and there with a subtle interdomain helix capping solution, seems like a reasonable way to diffuse the combinatorial problem of bringing a full-fledged new domain into the world. It is interesting to consider that mimivirus, with its enormous genome, rapid replication rate, large population numbers, and potential for engaging in horizontal gene transfer,[36] may use this mechanism to contribute to the expansion of protein fold space.

BACTERIOPHAGE T7 FIBER PROTEIN GP17 (R0001, PDB: 4A0U; CARMELA GARCIA-DOVAL AND MARK J. VAN RAAIJ)

Bacterial viruses, or bacteriophages, are important predators of bacteria. A majority of bacteriophages belong to the *Caudovirales* order and have a tail attached to a special vertex of their DNA-containing capsid.[37] The tail is involved in specific host recognition and subsequent DNA ejection across the bacterial membrane. The tailed phages can be divided into three families: *Myoviridae* with a long contractile tail, *Siphoviridae* with a long, flexible, noncontractile tail, and *Podoviridae* with a short noncontractile tail. Apart from their biological importance, bacteriophages have been studied as model systems for nucleic acid metabolism, protein assembly, DNA packaging, and other biochemical processes essential for life, leading to seminal discoveries such as that of messenger RNA.[38] Bacteriophages have also been used in applications such as phage therapy, phage display, and proposed as gene delivery vectors.[39]

Bacteriophage T7 is a member of the *Podoviridae* family.[40] It is composed of an icosahedral capsid, formed by the protein gp10, and a short noncontractile tail [Fig. 5(A,B)]. The capsid contains 40 kb of linear dsDNA and the core complex, formed by gp14, gp15, and gp16. This complex is attached to the tail via the connector, a dodecamer of gp8. The tail is formed by the proteins gp11 and gp12. Gp13, gp6.7, and gp7.3 have also been

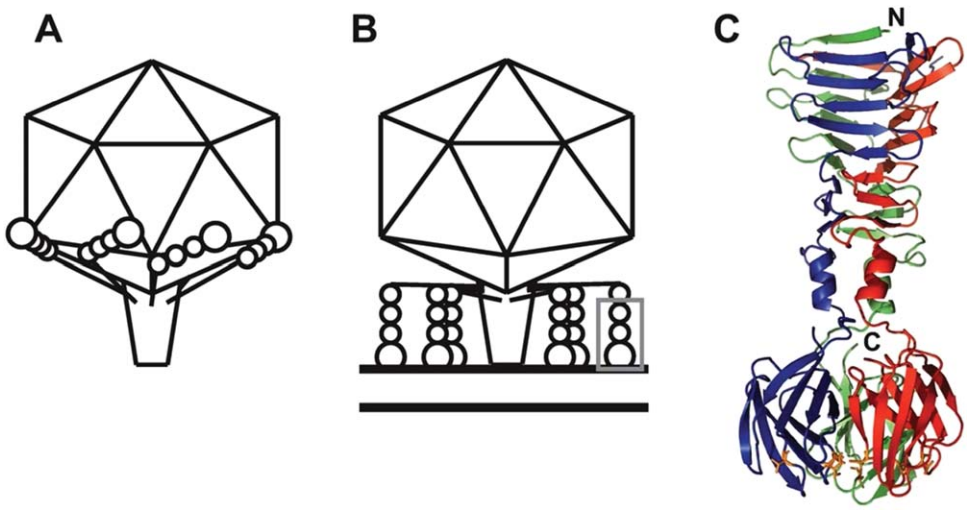

Figure 5

Bacteriophage T7 and its fiber protein gp17. **A** and **B**: Schematic diagram of bacteriophage T7 in the free state (A) and when bound to its host *E. coli* (B). The crystallized fragment is shown in a gray box. C: Cartoon representation of the structure of the C-terminal part of the gp17 trimer containing amino acids 371 to 553 (PDB: 4a0u). The N- and C-terminal ends are indicated. Residues that are thought to be important for host attachment and host range determination (518, 520, and 544) are shown as sticks. (This figure was prepared using the PyMOL Molecular Graphics System, Version 1.4.1 Schrödinger LLC.)

described as being important for the infection; however, their location is still unknown. To the top part of the tail, six protein fibers are attached [Fig. 5(A,B)]. Each one is a homotrimer of gp17, a protein of 553 amino acids. The fibers and the tail together are responsible for efficient host cell recognition.[41] Presumably, the phage diffuses through the medium with the tail fibers retracted (i.e., bound to the capsid), although they transiently detach and may encounter a suitable host cell receptor. The phage may then diffuse two-dimensionally along the bacterial surface, attaching and detaching the individual reversible fiber–lipopolysaccharide (LPS) interactions until a suitable location for DNA injection is found. A second, irreversible receptor interaction may be involved in this. The fibers bind to the *Escherichia coli* LPS core heptose region[42]; however, a putative second receptor has not yet been described.

Electron microscopy on fibers combined with image averaging and mass measurements on unstained fibers has led to a low-resolution model of the organization of the T7 fiber.[43] The fiber contains an N-terminal phage-binding domain consisting of residues 1–149 followed by a thin rod-like domain, which may contain a triple α-helical coiled coil, containing residues 151–260. After this proximal rod-like domain, there is a kink separating it from the distal part, consisting of residues 268–553. This distal part consists of four nodules as observed by electron microscopy. Based on the mass measurements, the nodules were assigned as containing residues 268–365, 366–432, 432–465, and 466–553, respectively.

The bacteriophage T7 fiber does not need phage-encoded chaperones for soluble expression in *E. coli* and correct trimerization. However, crystallization trials of the full-length protein were not successful.[44] Expression vectors for four different N-terminal deletion mutants of gp17 were constructed, each one starting more or less where one of the four nodules start, while taking care not to interrupt predicted secondary structure elements. N-terminal purification tags encoding six consecutive histidine residues were included. The fragment gp17 (371–553) was crystallized, and its structure could be resolved at 1.9 Å resolution [Fig. 5(C)].[44,45] The structure can be divided into two parts [Fig. 5(C)]: a tip domain (residues 465–553) and a pyramid domain (residues 371–464). The tip domain is globular, made up of β-strands, and corresponds to the last nodule. Each monomer contains a β-sandwich of two antiparallel four-stranded β-sheets, with a topology that has not been observed before (but see below). The pyramid domain corresponds to the second and third nodules and is mainly β-structured. The pyramid domain is composed of three nine-stranded β-sheets, to which each of the monomers contributes strands. The pyramid and tip domains are connected to each other by three short α-helices. The tip domain is likely responsible for receptor interaction, as reported mutations that change the host range of bacteriophage T7 are located in amino acids found in loops at the top of the tip domain [Fig. 5(C)].[46,47] When the sequence of the T7 tail fiber is compared

with other related tail fibers, most of the differences are located in the tip domain, which is also consistent with the fact that this tip domain is more variable, responsible for cell attachment, and can be adapted to different host bacteria.

The predictions submitted for T7 fiber gp17 within the CASP10 experiment were overall not very accurate (GDT_TS less than 20) and did not realistically reflect the biology of the protein. The structure of the phage T7 fiber gp17 was difficult to predict due to the lack of sequence homology with any protein of known structure. It appears that the trimeric nature of the protein was not taken into account in the prediction, whereas this fact was provided with the target sequence. In general, the β-stranded nature of the structure is found in the predictions, and in some cases, the residues that are in α-helical conformation in the crystal structure also have this conformation in the predicted structures. However, the topology of the predicted structures does not resemble the solved crystal structures exactly, which means that predictions based on threading the new sequence on an existing structural backbone are bound to at least partially fail.

The predictions that were guided by a limited amount of given long-rang contacts (amino acids spaced apart in the protein sequence but close together in the three-dimensional structure) do not appear to have fared significantly better than those without. In some of the predictions, the residues known to be important for receptor recognition [Fig. 5(C)] are located far away from each other, whereas in others, they are close together, as they are in the crystal structure. The information regarding colocalization of putative receptor-binding residues would have been useful in the absence of the current crystal structure for designing site-directed mutagenesis experiments, either to validate important amino acids or to attempt to modulate receptor-binding properties.

The structure of the pyramid domain did show structural homology to the needle domains of bacteriophage P2 and Phi92,[48] and the tip domain has a similar topology to the globular domain of the bacteriophage Sf6 needle.[49] If it had been possible to predict this despite the lack of sequence homology, these structures could have been used for more successful structure predictions. If the fact that the protein forms a parallel homotrimer would have been taken into account, predictions might also have been more accurate. Furthermore, the low-resolution mask of the protein as determined by Steven *et al.*[43] could have been used to guide structure predictions. Now that the structure of this part of the T7 fiber is known, it should be possible to make reliable structure predictions for homologous domains of the fibers of *E. coli* phages T3 and 13a, the *Yersinia* phage PhiA1122, and perhaps other phages.

CRYSTAL STRUCTURE OF BACTERIOPHAGE CBA-120 TAILSPIKE (T0739; CHEN CHEN, PATRICK BALES, DANIEL NELSON, AND OSNAT HERZBERG)

Similar to the tail fiber described in the previous section, bacteriophage tailspikes are trimeric proteins involved in recognition of the bacterial host, usually through reversible binding to the repeating glycan units of the LPS or capsular polysaccharides. However, in contrast to the tail fibers that lack catalytic activity, tailspike proteins act as endoglycosidases. This enzymatic activity is thought to assist the phage in penetration of the capsule and outer LPS matrix to reach a secondary receptor that is irreversibly bound by the phage for subsequent DNA ejection.[40] Bacterial phage CBA-120 belongs to the *Myoviridae* family and specifically infects *E. coli* strain O157, an important food-borne pathogen.[50] The genome of this phage encodes four putative tailspike proteins, TSP1–4 (ORFs 210–213), which is unique for *Myoviridae* as tailspike proteins are usually associated with the *Podoviridae* family.[51] Recently, we have discovered that the CBA-120 TSP1 (770 amino acid residues) degrades biofilm polysaccharides formed by a variety of bacterial species (Daniel Nelson, unpublished data). Thus, CBA-120 TSP1 is a candidate for combination therapy for biofilm-forming bacterial infections, where the biofilm mediates resistance to antibiotics by encapsulating persister cells maintained in stationary phase inaccessible to the drugs.

The structures of tailspike proteins from a number of *Podoviridae* phages have been determined by X-ray crystallography at high resolution, including those from HK620, P22, ϕ29, det7, and SF6.[52–56] Although they have limited sequence homology, they share common overall fold.[40] All tailspike proteins are elongated in shape and assemble into homotrimers. They are composed of two functional domains: the N-terminal head-binding domain that binds the virion particle and the C-terminal receptor binding domain that binds and degrades the extracellular polysaccharide. The receptor-binding domain consists of primarily a right-handed parallel β-helix structure, spanning approximately two-third of the full-length protein, followed by an intervening fragment and an ensuing C-terminal domain exhibiting different β-structure folds in different tailspikes. The receptor-binding sites among the tailspikes of known structures are located at the β-helix region, either along the interface between trimer subunits or within a single subunit. As with most endoglycosidases, such as lysozyme, the catalytic machinery is thought to consist of two carboxylate groups acting as general acid and base. The substrate specificity and location of the CBA-120 TSP1 active site are currently unknown. The crystallographic studies of TPS1 aim at examining its relationship

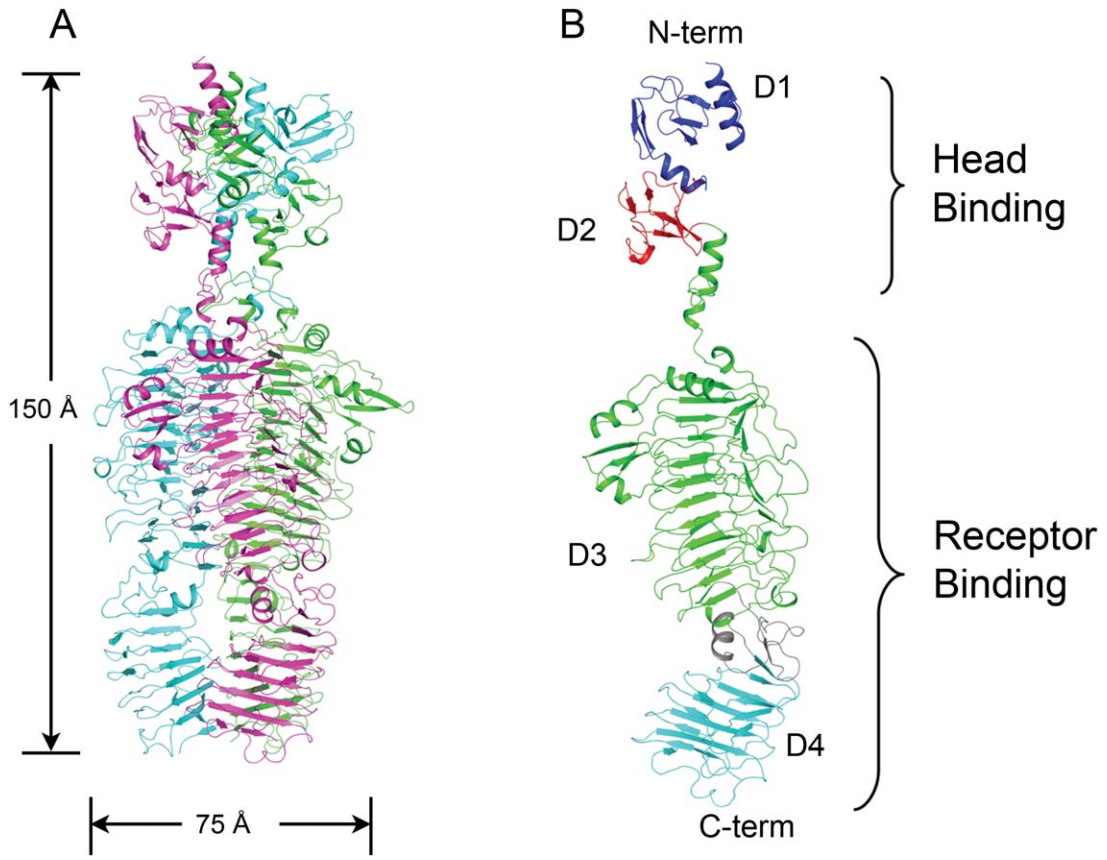

Figure 6

Crystal structure of bacteriophage CBA-120 tailspike. **A**: The overall structure of Tailspike TSP1 homotrimer. **B**: The structure of a TSP1 monomer. In CASP, the structures were assessed in full length and parsed into four structural domains (D1–D4). The N-terminal head-binding domain includes D1 (residues 12–96, colored blue) and D2 (residues 97–154, colored red). The ligand-binding domain includes D3 (residues 198–580, colored green) and D4 (residues 581–796, colored gray and cyan). The ligand-binding domain assumes primarily right-handed parallel β-helical structure; however, the helical axis is bent by an intervening fragment (residues 581–623, colored gray).

to Podoviridae tailspike structures at atomic level and at providing insight into the enzyme catalytic mechanism and specificity.

The structure of the full-length CBA-120 TSP1 was determined by Se-MAD phasing because molecular replacement using known tailspike structures as search models failed, consistent with the lack of significant sequence homology. As expected, TSP1 exhibits a rod-shaped homotrimer [Fig. 6(A)] consisting of the putative N-terminal head-binding domain and a C-terminal receptor-binding domain [Fig. 6(B)]. The latter domain contains primarily a three-stranded right-handed parallel β-helical structure. In CASP10, the predicted full-length structures were assessed and also parsed into four structural subdomains (D1–D4), among which the head-binding domain consists of D1 and D2 and the receptor-binding domain consists of D3 and D4, as outlined in Figure 6(B). Because D4 contains ∼40 amino acids that intervene between the D3 and D4 β-helix regions, we analyzed two D4 subdomains separately, the intervening region (amino acid residues 581–623) and the ensuing β-helix (amino acids 624–766).

For the N-terminal head-binding domain, a Dali[57] search revealed no significant structure analog of the subdomain D1. The C-terminal subdomain D2 (residues 97–154) folds similarly to the NMR structure of the chitin-binding domain of Chitinase from *Bacillus circulans* (PDB: 1ED7) with RMSD of 2.1 Å over 38 paired Cα atoms.

A second Dali search with the receptor-binding domain as the query revealed that the D3 subdomain (residues 198–580) of the receptor binding domain is quite similar to a number of known tailspike structures, with the closest structural homolog being phage Sf6 tailspike (PDB: 2VBE). The RMSD value for 300 paired Cα atoms is 2.5 Å. The region that breaks the D3 and D4 β-helices does not resemble any other bacteriophage tailspikes. The D4 β-helical region has no structural homologs in other tailspikes with known structures. Most D4 domains contain β-sheets at their C-termini but these

are not folded into β-helix. The tailspike from bacteriophage ϕ29 is an exception, as both its D3 and D4 subdomains are β-helical (with large translational shift between the two domains). However, these are unusual two-stranded β-helices, quite different from the CBA-120 TSP1 β-helices.

The lack of significant amino acid sequence homology between TSP1 and tailspike proteins of known structure implies that this is a difficult structure to predict accurately even though the sequence database provides the functional annotation and secondary structure prediction programs show a strong β-strand signal. Issues that complicate structure prediction include out-of-register secondary structure alignment; difficulties to predict the structures of regions connecting the subdomains and hence the orientations between the subdomains; and problems with prediction of the loop regions that lack templates. These loop regions are particularly important because some loops of known tailspike structures delineate the active site. As expected, the structures of D1 and D4 regions that lack structure homologs in the PDB were not predicted correctly in CASP10. The D3 β-helix region was identified by most predictors, although the models exhibit high RMSD values. The fold of the D2 domain, which resembles that of the chitin binding domain of Chitinase, was not identified.

A METAGENOMIC PHAGE COAT PROTEIN FROM THE MARINE ENVIRONMENT (R0009, PDB: 4DMI; TIMOTHY K. CRAIG, ALEX BURGIN, DONALD LORIMER, FOREST ROHWER, ANCA SEAGAL, AND VICTOR SEGURITAN)

An estimated 4×10^{30} viruses[58] in the oceans represent one of the largest reservoirs of genetic diversity on earth. Many of the genes in this reservoir encode phage structural proteins such as capsid or tail proteins; however, they also encode genes that confer evolutionary advantages to their hosts including antibiotic resistance and acceleration of bacterial evolution through horizontal gene transfer.[59] By studying phage metagenomic sequences, we aim to uncover new enzymes with novel functions that could be exploited for various biotechnological purposes, including diagnostics as well as vaccine development.

This protein sequence was identified from a metagenomic pool of sequences isolated from marine environmental samples. The metagenomic sequences were then analyzed with an artificial neural network to identify protein-coding regions.[60] Highly pure protein was obtained for an expression construct, #5936. Crystallization trials were carried out at 16°C, and large, well-diffracting crystals were obtained. A native dataset was collected to 1.5 Å. Unfortunately, the amino acid sequence of this protein has extremely low sequence identity (7%) to any previously solved structures currently deposited in the PDB, which is one of the main reasons we believed that the structure would be an excellent candidate for CASP10. As no molecular replacement models were available, a second dataset was generated for SAD phasing using iodide ions.[61] After SAD phasing, we used the C-terminal domain as a model for molecular replacement with the 1.5 Å native data.

To our surprise, the resulting pentameric structure clearly shows a two-domain architecture for each monomer [Fig. 7(A)]. The N-terminal domain is a fold similar to the six-stranded β-barrel of the cowpea chlorotic mosaic virus,[62,63] except in the structure of construct 5936 only five strands form the β-barrel. This β-barrel is surrounded by two antiparallel β-sheets and a single helix from each monomer. The helices form a pentagonal outline sitting atop the β-barrel [Fig. 7(B)]. The linkage between the N-terminal domain and the C-terminal domain has weak density and high temperature-factors, suggesting that it may be somewhat flexible, perhaps due to the absence of a binding partner. The C-terminal domain consists of a "jelly-roll" fold formed from two sets of four antiparallel β-sheets, which is similar in structure to other viral coat proteins.[64] A large, water-filled internal cavity is formed in the center of the full pentamer, lined with extensive networks of structured water molecules in our high-resolution structure.

Because of the large size and complex interactions at the monomer interfaces, the CASP predictions for this protein failed to produce a model similar to the crystal structure. The pentameric nature of the protein was not predicted by any of the models, and of the monomer structures predicted, none showed a fold similar to the monomers in the crystal structure even at a qualitative level. With low sequence identity to any known structures, we expected that this would be a challenging target for CASP. Robust protein predictions are of high utility for proteins generated from metagenomic samples, where sequence identity with known proteins is low or nonexistent.

ENGINEERED DISULFIDE-RICH SMALL PROTEINS: AN UNPRECEDENTED CLASS OF STRUCTURE PREDICTION TARGETS (T0711, PDB:2M7T; FRANK V. COCHRAN AND RHIJU DAS)

Small proteins containing multiple disulfide bonds are increasingly being developed for a variety of biomedical applications.[65] These short sequences (30–50 residues) adopt well-defined and highly stable 3D structures

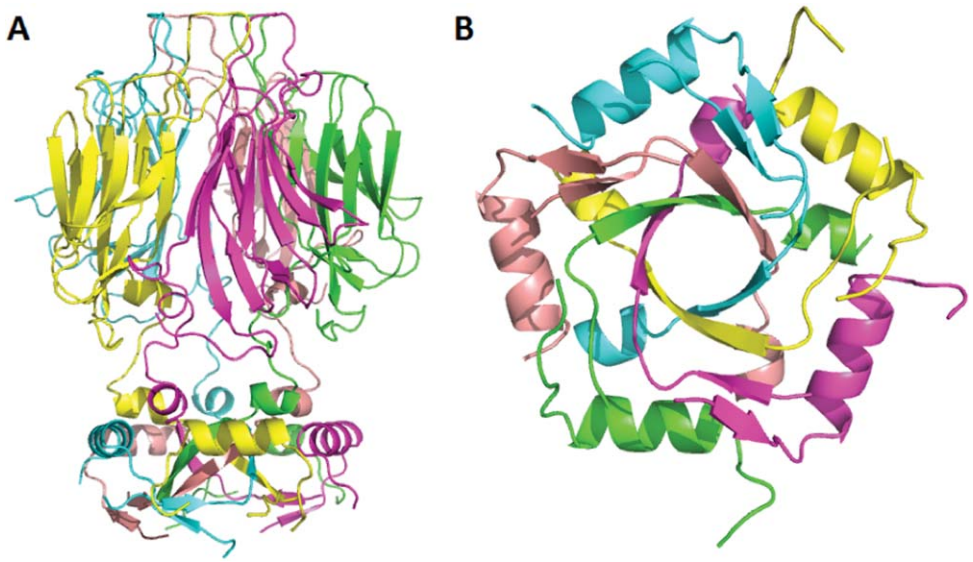

Figure 7

Cartoon representation of 4DMI. This putative coat protein contains a semiflexible linker between highly structured N- and C-terminal domains (**A**). Each of the chains is colored differently to highlight the interconnectedness of the N-terminal domain (**B**).

composed largely of irregular, yet often rigid, loop conformations. The natural sequence diversity in these loops results in a wide variety of biological activities and allows novel functions to be introduced through molecular engineering. In a recent example, *Ecballium elaterium* trypsin inhibitor II (EETI-II), a member of the cystine knot family, was engineered for high affinity binding to tumor-associated integrin receptors by yeast surface display.[66] Libraries were prepared in which the Arg-Gly-Asp integrin-binding motif, flanked by randomized positions, were grafted in place of the trypsin-binding loop. High-throughput screening identified several variants that bound integrin receptors with low-nanomolar affinity. Evaluation as molecular imaging agents in living animals demonstrated high tumor localization, low accumulation in non-target tissue and organs, such as kidney and liver, and high proteolytic and metabolic stability.[67] The combination of high-affinity receptor binding and exceptional *in vivo* performance has established these engineered integrin-binding proteins as leading candidates for further clinical development.

NMR studies with ^{15}N and ^{13}C double-labeled samples were undertaken to begin characterizing the structural basis for binding in addition to exploring the opportunities presented by small systems for advancing computational modeling.[68] The 33-amino-acid sequence of 2.5D was released as target T0711 in CASP10. The 3D structure (PDB: 2M7T; Fig. 8) showed that the cystine knot fold was preserved and that the conformer ensemble agreed well with the published X-ray structure of EETI-II[70] in regions outside the engineered loop. Minor backbone and side-chain differences when compared with the wild type occurred proximal to the engineered loop and were further validated by solving the structure of EETI-II with the same NMR methods. The entire 11-amino-acid engineered loop in T0711 was sufficiently well defined to be included in the assessment unit based

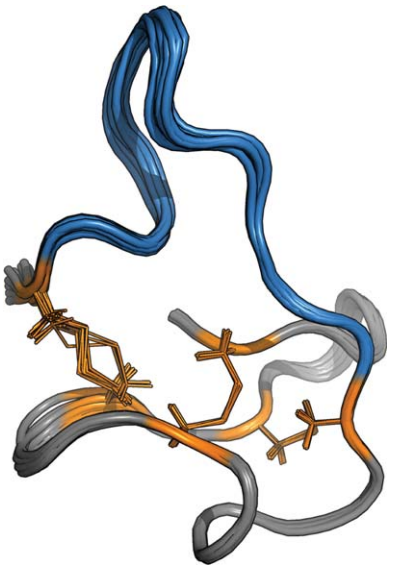

Figure 8

Solution NMR structure of 2.5D (PDB ID 2M7T) represented as 20 lowest energy conformers superimposed using the THESEUS maximum likelihood method.[69] The engineered integrin-binding loop is rendered in blue, and the disulfide bonds are shown in orange. This figure was prepared with PyMol (www.pymol.org).

on the specific structural criteria chosen for CASP10.[71] Two other related integrin-binding cystine knot proteins were also released as prediction targets (T0709 and R0003 in CASP ROLL, both 33 amino acids in length); however, multiple residues were omitted from the corresponding assessment units. Refined structures using restraints from more recently acquired NMR data will be discussed in a future report that discusses these results in the context of receptor-binding specificity.

As the sequence of T0711 differs from that of EETI-II only in the engineered loop and a K15S mutation for site-specific chemical conjugation at the N-terminal amine, standard database searches readily identified structural templates. As a result, T0711 was assigned as a TBM target.[71] However, the global scoring metrics in this category are not the most informative measures of predictive performance for the engineered loop, which is clearly the most interesting component as database searches are less likely to provide closely similar structural templates. This is emphasized by the reliance on global whole-model measures to determine how targets were used for prediction rankings.[3,8] Given that loops play important roles in proteins, T0711 highlights the need for amended CASP protocols that capture a more complete picture of our understanding of protein structure.

SCORING OF LOOP PREDICTIONS

Evaluating the accuracy of loop conformation prediction within the CASP experiment has been challenging for several reasons: in a typical CASP prediction situation, the effect of selecting different templates, using different target-template alignments, and applying different loop prediction techniques are highly convolved in the final models. In addition, loop conformations in target structures are frequently influenced by crystal contacts in the reference structures. In the CASP10 experiment, a new type of prediction target such as T0711 became available, where specifically one loop was engineered within a mini-protein framework for which robust templates were available. Furthermore, these structures were solved by NMR, that is, the loop conformation was not influenced by crystal contacts.

Local scores such as the lDDT[4] allow the evaluation of loops in the context of the rest of the model using an NMR ensemble as reference. The lDDT measures the fraction of interatomic distances that were correctly reproduced in the model at certain accuracy thresholds, deriving the expected distance intervals from the NMR ensemble. Segments, which show large deviations within the ensemble of reference structures, will be characterized by large distance tolerance intervals, and therefore, a wider range of models will be considered as locally correct prediction for this region.

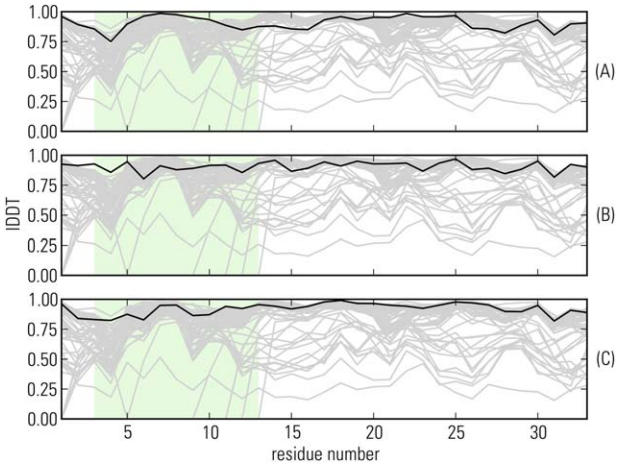

Figure 9

Local accuracy assessment of an engineered loop in target T0711. The per-residue accuracy of predictions was evaluated using all-atom lDDT in multireference mode against the NMR ensemble (cutoff radius = 10 Å, with a sequence separation of zero). The engineered loop region (residues 3–13) is shaded. The results by all groups are shown in gray with the best loop predictions highlighted in bold. Predictions by (**A**) BAKER-ROSETTASERVER and PconsM, (**B**) BhageerathH, and (**C**) MULTICOM-REFINE are shown. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

For example, the assessment of the engineered loop of T0711 (Fig. 9) indicates that the best all-atom prediction was provided by "BAKER-ROSETTASERVER" and picked-up by the meta-method "PconsM" (see Supporting Information for assessment of T0709, T0711, and R0003). One has to keep in mind, however, that the assessment of a small number of loops has no statistical significance which would allow a meaningful ranking of the methods. This analysis should be interpreted as proof of principle for the kind of analysis which would be possible in the future, if a larger number of targets of this type would become available.

SUMMARY AND OUTLOOK

Structure modeling and experimental structural biology complement each other in making structure information available to researchers addressing important problems in life science. Especially, template-based modeling has matured as a field and has become main stream in life science research today.[72] However, despite the significant progress over the last two decades, structure prediction is still far from being perfect, and more research is required to improve accuracy of models, especially in the remote homology/*de novo* modeling area. Together with the improvement of structure prediction methods themselves, techniques for their assessment also have to evolve to better reflect the issues important for structural

biology research.[8,9,73,74] The examples presented in this article highlight a series of reoccurring themes that appear to be challenging for current methods and might deserve development of specific methods in the future. (a) The accuracy of *de novo* predictions of new folds or regions for which templates cannot be detected is not satisfying, especially for longer protein sequences. (b) The oligomeric state of the target protein has often been shown to be relevant for structural integrity, and the oligomeric structure may also assist predictions of the interdomain orientation in multidomain proteins. However, it was not taken into account for the predictions, even in cases where information about the correct oligomeric state was provided. (c) Post-translational modifications that form an integral part of a structure are not considered in structure predictions. (d) Domain-sized extensions fused to known structures are often not recognized and modeled correctly; this is especially true for fast evolving viral proteins. (e) The specific structure of individual loops is often a key for the functional understanding of a protein; however, global evaluation measures in CASP do not capture small differences in the structures. Therefore, detailed assessments of loop modeling accuracy might be of interest in future exercises.

We hope that this study will guide future CASP assessors in emphasizing relevant criteria in the assessment and also inspire the developers of new improved techniques for structure prediction.

AUTHOR CONTRIBUTIONS

The parts of the manuscript on target T0711 were contributed by F.V.C. and R.D.; target T0737 by D.F.; R0007 by X.M. and J.F.B.; target T0666 by H.L.; target T0734 by K.Z.; target R0001 by C.G.D. and M.J. van Raaij; target T0739 by C.C., P.B., D.N., and O.H.; target R0009 by T.K.C., A.B., D.L., F.R., A.S., and V.S.; loop assessment on T0709, T0711, and R0003 by M.B.; and concept, editing, introduction, discussion, and coordination by A.K., J.M., and T.S.

REFERENCES

1. Moult J, Fidelis K, Kryshtafovych, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)–round X. Proteins 2014;82(Suppl 2):1–6.
2. Kryshtafovych A, Moult J, Bartual SG, Bazan JF, Berman H, Casteel DE, Christodoulou E, Everett JK, Hausmann J, Heidebrecht T, Hills T, Hui R, Hunt JF, Seetharaman J, Joachimiak A, Kennedy MA, Kim C, Lingel A, Michalska K, Montelione GT, Otero JM, Perrakis A, Pizarro JC, van Raaij MJ, Ramelot TA, Rousseau F, Tong L, Wernimont AK, Young J, Schwede T. Target highlights in CASP9: experimental target structures for the critical assessment of techniques for protein structure prediction. Proteins 2011;79 (Suppl 10): 6–20.
3. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
4. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 2013; 29(21): 2722–2728.
5. Holm L, Kaariainen S, Wilton C, Plewczynski D. Using Dali for structural comparison of proteins. Curr Protoc Bioinformatics 2006; Chapter 5: Unit 5.5.
6. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82(Suppl 2):7–13.
7. Olechnovic K, Kulberkyte E, Venclovas C. CAD-score: a new contact area difference-based function for evaluation of protein structural models. Proteins 2013;81:149–162.
8. Mao B, Huang YJ, Aramini JM, Montelione GT. Assessment of template based protein structure predictions in CASP10. Proteins 2014; 82(Suppl 2):43–57.
9. Tai CH, Bai H, Taylor TJ, Lee BK. Assessment of template free modeling in CASP10 and ROLL. Proteins 2013; • • •; • • •.
10. Pounder RE, Ng D. The prevalence of Helicobacter pylori infection in different countries. Aliment Pharmacol Ther 1995;9 (Suppl 2): 33–39.
11. Peek RM, Jr, Blaser MJ. Helicobacter pylori and gastrointestinal tract adenocarcinomas. Nat Rev Cancer 2002;2:28–37.
12. Weeks DL, Eskandari S, Scott DR, Sachs G. A H+-gated urea channel: the link between Helicobacter pylori urease and gastric colonization. Science 2000;287:482–485.
13. Strugatsky D, McNulty R, Munson K, Chen CK, Soltis SM, Sachs G, Luecke H. Structure of the proton-gated urea channel from the gastric pathogen Helicobacter pylori. Nature 2013;493:255–258.
14. Luecke H. Neutralizing a pathogen. International innovation, Vol. 20. Bristol: Research Media; 2013. pp 104–106.
15. McNulty R, Ulmschneider JP, Luecke H, Ulmschneider MB. Mechanisms of molecular transport through the urea channel of Helicobacter pylori. Nature Communications 2013;4:2900.
16. Bazan JF. Haemopoietic receptors and helical cytokines. Immunol Today 1990;11:350–354.
17. Verstraete K, Savvides SN. Extracellular assembly and activation principles of oncogenic class III receptor tyrosine kinases. Nat Rev Cancer 2012;12:753–766.
18. Bazan JF. Genetic and structural homology of stem cell factor and macrophage colony-stimulating factor. Cell 1991;65:9–10.
19. Lin H, Lee E, Hestir K, Leo C, Huang M, Bosch E, Halenbeck R, Wu G, Zhou A, Behrens D, Hollenbaugh D, Linnemann T, Qin M, Wong J, Chu K, Doberstein SK, Williams LT. Discovery of a cytokine and its receptor by functional screening of the extracellular proteome. Science 2008;320:807–811.
20. Chen X, Liu H, Focia PJ, Shim AH, He X. Structure of macrophage colony stimulating factor bound to FMS: diverse signaling assemblies of class III receptor tyrosine kinases. Proc Natl Acad Sci USA 2008;105:18267–18272.
21. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
22. Ma X, Lin WY, Chen Y, Stawicki S, Mukhyala K, Wu Y, Martin F, Bazan JF, Starovasnik MA. Structural basis for the dual recognition of helical cytokines IL-34 and CSF-1 by CSF-1R. Structure 2012;20: 676–687.
23. Rozwarski DA, Gronenborn AM, Clore GM, Bazan JF, Bohm A, Wlodawer A, Hatada M, Karplus PA. Structural comparisons among the short-chain helical cytokines. Structure 1994;2:159–173.
24. Liu H, Leo C, Chen X, Wong BR, Williams LT, Lin H, He X. The mechanism of shared but distinct CSF-1R signaling by the non-homologous cytokines IL-34 and CSF-1. Biochim Biophys Acta 2012;1824:938–945.
25. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr Sect D: Biol Crystallogr 2004;60 (Part 12; Part 1):2256–2268.

26. Kouril T, Zaparty M, Marrero J, Brinkmann H, Siebers B. A novel trehalose synthesizing pathway in the hyperthermophilic Crenarchaeon Thermoproteus tenax: the unidirectional TreT pathway. Arch Microbiol 2008;190:355–369.
27. Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure prediction with HHpred. Proteins 2009;77 (Suppl 9):128–132.
28. Soding J, Remmert M, Biegert A. HHrep: de novo protein repeat detection and the origin of TIM barrels. Nucleic Acids Res 2006;34 (Web Server issue):W137–W142.
29. Fass D. The Erv family of sulfhydryl oxidases. Biochim Biophys Acta 2008;1783:557–566.
30. Senkevich TG, White CL, Koonin EV, Moss B. A viral member of the ERV1/ALR protein family participates in a cytoplasmic pathway of disulfide bond formation. Proc Natl Acad Sci USA 2000;97: 12068–12073.
31. Rodriguez I, Redrejo-Rodriguez M, Rodriguez JM, Alejo A, Salas J, Salas ML. African swine fever virus pB119L protein is a flavin adenine dinucleotide-linked sulfhydryl oxidase. J Virol 2006;80:3157–3166.
32. Long CM, Rohrmann GF, Merrill GF. The conserved baculovirus protein p33 (Ac92) is a flavin adenine dinucleotide-linked sulfhydryl oxidase. Virology 2009;388:231–235.
33. Ogata H, Claverie JM. Unique genes in giant viruses: regular substitution pattern and anomalously short size. Genome Res 2007;17: 1353–1361.
34. Yoosuf N, Yutin N, Colson P, Shabalina SA, Pagnier I, Robert C, Azza S, Klose T, Wong J, Rossmann MG, La Scola B, Raoult D, Koonin EV. Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the megavirus lineage. Genome Biol Evol 2012;4:1324–1330.
35. Hakim M, Ezerina D, Alon A, Vonshak O, Fass D. Exploring ORFan domains in giant viruses: structure of mimivirus sulfhydryl oxidase R596. PLoS One 2012;7:e50649.
36. Yoshida T, Claverie JM, Ogata H. Mimivirus reveals Mre11/Rad50 fusion proteins with a sporadic distribution in eukaryotes, bacteria, viruses and plasmids. Virol J 2011;8:427.
37. Ackermann HW. Tailed bacteriophages: the order caudovirales. Adv Virus Res 1998;51:135–201.
38. Volkin E, Astrachan L, Countryman JL. Metabolism of RNA phosphorus in Escherichia coli infected with bacteriophage T7. Virology 1958;6:545–555.
39. Haq IU, Chaudhry WN, Akhtar MN, Andleeb S, Qadrir I. Bacteriophages and their implications on future biotechnology: a review. Virol J 2012;9:9.
40. Casjens SR, Molineux IJ. Short noncontractile tail machines: adsorption and DNA delivery by podoviruses. Adv Exp Med Biol 2012;726:143–179.
41. Hu B, Margolin W, Molineux IJ, Liu J. The bacteriophage t7 virion undergoes extensive structural remodeling during infection. Science 2013;339:576–579.
42. Qimron U, Marintcheva B, Tabor S, Richardson CC. Genomewide screens for Escherichia coli genes affecting growth of T7 bacteriophage. Proc Natl Acad Sci USA 2006;103:19039–19044.
43. Steven AC, Trus BL, Maizel JV, Unser M, Parry DA, Wall JS, Hainfeld JF, Studier FW. Molecular substructure of a viral receptor-recognition protein. The gp17 tail-fiber of bacteriophage T7. J Mol Biol 1988;200:351–365.
44. Garcia-Doval C, van Raaij MJ. Crystallization of the C-terminal domain of the bacteriophage T7 fibre protein gp17. Acta Crystallogr Sect F: Struct Biol Cryst Commun 2012;68 (Part 2):166–171.
45. Garcia-Doval C, van Raaij MJ. Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers. Proc Natl Acad Sci USA 2012;109:9390–9395.
46. Heineman RH, Springman R, Bull JJ. Optimal foraging by bacteriophages through host avoidance. Am Nat 2008;171:E149–E157.
47. Garcia E, Elliott JM, Ramanculov E, Chain PS, Chu MC, Molineux IJ. The genome sequence of Yersinia pestis bacteriophage phiA1122 reveals an intimate history with the coliphage T3 and T7 genomes. J Bacteriol 2003;185:5248–5262.
48. Browning C, Shneider MM, Bowman VD, Schwarzer D, Leiman PG. Phage pierces the host cell membrane with the iron-loaded spike. Structure 2012;20:326–339.
49. Bhardwaj A, Molineux IJ, Casjens SR, Cingolani G. Atomic structure of bacteriophage Sf6 tail needle knob. J Biol Chem 2011;286: 30867–30877.
50. Kutter EM, Skutt-Kakaria K, Blasdel B, El-Shibiny A, Castano A, Bryan D, Kropinski AM, Villegas A, Ackermann HW, Toribio AL, Pickard D, Anany H, Callaway T, Brabban AD. Characterization of a ViI-like phage specific to Escherichia coli O157:H7. Virol J 2011;8: 430.
51. Adriaenssens EM, Ackermann HW, Anany H, Blasdel B, Connerton IF, Goulding D, Griffiths MW, Hooton SP, Kutter EM, Kropinski AM, Lee JH, Maes M, Pickard D, Ryu S, Sepehrizadeh Z, Shahrbabak SS, Toribio AL, Lavigne R. A suggested new bacteriophage genus: "Viunalikevirus". Arch Virol 2012;157:2035–2046.
52. Barbirz S, Muller JJ, Uetrecht C, Clark AJ, Heinemann U, Seckler R. Crystal structure of Escherichia coli phage HK620 tailspike: podoviral tailspike endoglycosidase modules are evolutionarily related. Mol Microbiol 2008;69:303–316.
53. Muller JJ, Barbirz S, Heinle K, Freiberg A, Seckler R, Heinemann U. An intersubunit active site between supercoiled parallel β-helices in the trimeric tailspike endorhamnosidase of Shigella flexneri phage Sf6. Structure 2008;16:766–775.
54. Steinbacher S, Baxa U, Miller S, Weintraub A, Seckler R, Huber R. Crystal structure of phage P22 tailspike protein complexed with Salmonella sp. O-antigen receptors. Proc Natl Acad Sci USA 1996;93: 10584–10588.
55. Xiang Y, Leiman PG, Li L, Grimes S, Anderson DL, Rossmann MG. Crystallographic insights into the autocatalytic assembly mechanism of a bacteriophage tail spike. Mol Cell 2009;34:375–386.
56. Walter M, Fiedler C, Grassl R, Biebl M, Rachel R, Hermo-Parrado XL, Llamas-Saiz AL, Seckler R, Miller S, van Raaij MJ. Structure of the receptor-binding protein of bacteriophage det7: a podoviral tail spike in a myovirus. J Virol 2008;82:2265–2273.
57. Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. Nucleic Acids Res 2010;38 (Web Server Issue):W545–W549.
58. Suttle CA. Viruses in the sea. Nature 2005;437:356–361.
59. Suttle CA. Marine viruses—major players in the global ecosystem. Nat Rev Microbiol 2007;5:801–812.
60. Seguritan V, Alves N, Jr, Arnoult M, Raymond A, Lorimer D, Burgin AB, Jr, Salamon P, Segall AM. Artificial neural networks trained to detect viral and phage structural proteins. PLoS Comput Biol 2012;8:e1002657.
61. Abendroth J, Gardberg AS, Robinson JI, Christensen JS, Staker BL, Myler PJ, Stewart LJ, Edwards TE. SAD phasing using iodide ions in a high-throughput structural genomics environment. J Struct Funct Genomics 2011;12:83–95.
62. Speir JA, Munshi S, Wang G, Baker TS, Johnson JE. Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by X-ray crystallography and cryo-electron microscopy. Structure 1995;3:63–78.
63. Willits D, Zhao X, Olson N, Baker TS, Zlotnick A, Johnson JE, Douglas T, Young MJ. Effects of the cowpea chlorotic mottle bromovirus β-hexamer structure on virion assembly. Virology 2003; 306:280–288.
64. Rossmann MG, Johnson JE. Icosahedral RNA virus structure. Annu Rev Biochem 1989;58:533–573.
65. Kolmar H. Natural and engineered cystine knot miniproteins for diagnostic and therapeutic applications. Curr Pharm Des 2011;17: 4329–4336.

66. Kimura RH, Levin AM, Cochran FV, Cochran JR. Engineered cystine knot peptides that bind alphavbeta3, alphavbeta5, and alpha5beta1 integrins with low-nanomolar affinity. Proteins 2009;77:359–369.
67. Kimura RH, Cheng Z, Gambhir SS, Cochran JR. Engineered knottin peptides: a new class of agents for imaging integrin expression in living subjects. Cancer Res 2009;69:2435–2442.
68. Das R. Four small puzzles that Rosetta doesn't solve. PLoS One 2011;6:e20044.
69. Theobald DL, Wuttke DS. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics 2006;22:2171–2172.
70. Kratzner R, Debreczeni JE, Pape T, Schneider TR, Wentzel A, Kolmar H, Sheldrick GM, Uson I. Structure of Ecballium elaterium trypsin inhibitor II (EETI-II): a rigid molecular scaffold. Acta Crystallogr Sect D: Biol Crystallogr 2005;61 (Part 9):1255–1262.
71. Taylor TJ, Tai CH, Huang YJ, Block J, Bai H, Kryshtafovych A, Montelione GT, Lee BK. Definition and Classification of Evaluation Units for CASP10. Proteins 2013;This volume(This issue).
72. Schwede T, Sali A, Honig B, Levitt M, Berman HM, Jones D, Brenner SE, Burley SK, Das R, Dokholyan NV, Dunbrack RL, Jr, Fidelis K, Fiser A, Godzik A, Huang YJ, Humblet C, Jacobson MP, Joachimiak A, Krystek SR, Jr, Kortemme T, Kryshtafovych A, Montelione GT, Moult J, Murray D, Sanchez R, Sosnick TR, Standley DM, Stouch T, Vajda S, Vasquez M, Westbrook JD, Wilson IA. Outcome of a workshop on applications of protein models in biomedical research. Structure 2009;17:151–159.
73. Gallo Cassarino T, Bordoli L, Schwede T. Assessment of ligand binding predictions in CASP10. Proteins 2014;82(Suppl 2):154–163.
74. Nugent T, Cozzetto, Jones DT. Evaluation of predictions in the CASP10 model refinement category. Proteins 2014;82(Suppl 2):98–111.