PROTEINS:
Structure, Function, and Bioinformatics

Target highlights from the first post-PSI CASP experiment (CASP12, May-August 2016)

Journal:	*PROTEINS: Structure, Function, and Bioinformatics*
Manuscript ID	Prot-00198-2017
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	06-Jul-2017
Complete List of Authors:	Kryshtafovych, Andriy; University of California, Davis, Genome Center Albrecht, Reinhard; Max Planck Institute for Developmental Biology, Department of Protein Evolution Basle, Arnaud; University of Newcastle, Institute for Cell and Molecular Biosciences Bule, Pedro; Universidade de Lisboa Faculdade de Medicina Veterinaria, CIISA Caputo, Alessandro; University of Oxford, Oxford Glycobiology Institute, Department of Biochemistry Carvalho, Ana Luisa; Universidade Nova de Lisboa Instituto de Tecnologia Quimica e Biologica, Departamento de Química, Faculdade de Ciências e Tecnologia Chao, Kinlin; University of Maryland, Institute for Bioscience and Biotechnology Research Diskin, Ron; Weizmann Institute of Science, Department of Structural Biology Fidelis, Krzysztof; University of California, Genome Center Fontes, Carlos; Universidade de Lisboa Faculdade de Medicina Veterinaria, CIISA Fredslund, Folmer; University of Copenhagen, Department of Chemistry Gilbert, Harry J.; University of Newcastle, Institute for Cell and Molecular Biosciences Goulding, Celia; University of California, Irvine, Department of Molecular Biology & Biochemistry, Pharmaceutical Sciences Hartmann, Marcus; Max Planck Institute for Developmental Biology, Department of Protein Evolution Hayes, Christopher; University of California, Santa Barbara , Department of Molecular, Cellular and Developmental Biology, Biomolecular Science and Engineering Program Herzberg, Oznat; University of Maryland, Institute for Bioscience and Biotechnology Research; University of Maryland, Department of Chemistry and Biochemistry Hill, Johan; University of Oxford, Oxford Glycobiology Institute, Department of Biochemistry Joachimiak, Andrzej; Argonne National Laboratory, Structural Biology Center; University of Chicago, Department of Biochemistry and Molecular Biology Kohring, Gert-Wieland; Saarland University, Microbiology

	Koning, Roman; Leiden University , Netherlands Centre for Electron Nanoscopy, Institute of Biology Leiden; Leiden University Medical Center, Department of Molecular Cell Biology Lo Leggio, Leila; University of Copenhagen, Department of Chemistry Mangiagalli, Marco; University of Milano-Bicocca, Department of Biotechnology and Biosciences Michalska, Karolina; Argonne National Laboratory, Biosciences Division Moult, John; University of Maryland , Institute for Bioscience and Biotechnology Research; University of Maryland, Department of Cell Biology and Molecular Genetics Najmudin, Shabir; Universidade de Lisboa Faculdade de Medicina Veterinaria, CIISA Nardini, Marco; University of Milano, Department of Biosciences Nardone, Valentina; University of Milano, Department of Biosciences Ndeh, Didier; University of Newcastle, Institute for Cell and Molecular Biosciences Nguyen, Thanh; Centro Nactional de Biotecnologia, CNB-CSIC Pintacuda, Guido; Université de Lyon, Institut des Sciences Analytiques Postel, Sandra; University of Maryland School of Medicine, Institute of Human Virology van Raaij, Mark; Centro Nactional de Biotecnologia, CNB-CSIC Roversi, Pietro; University of Oxford, Oxford Glycobiology Institute, Department of Biochemistry Shimon, Amir; Weizmann Institute of Science, Department of Structural Biology Sundberg, Eric; University of Maryland School of Medicine, Institute of Human Virology Tars, Kaspars; Latvian Biomedical Research and Study Center, Latvian Biomedical Research and Study Center; University of Latvia, Department of Molecular Biology Zitzmann, Nicole; University of Oxford, Oxford Glycobiology Institute, Department of Biochemistry Schwede, Torsten; University of Basel, Biozentrum; SIB Swiss Institute of Bioinformatics, Protein Structure Bioinformatics group
Key Words:	X-ray Crystallography; NMR; CASP, Protein Structure Prediction

Target highlights from the first post-PSI CASP experiment (CASP12, May-August 2016)

Running title: CASP12 target highlights

Andriy Kryshtafovych, Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616, USA

Reinhard Albrecht, Department of Protein Evolution, Max Planck Institute for Developmental Biology, Spemannstraße 35, 72076 Tübingen, Germany

Arnaud Baslé, Institute for Cell and Molecular Biosciences, University of Newcastle, Newcastle upon Tyne NE2 4HH, UK

Pedro Bule, CIISA - Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

Alessandro T. Caputo, Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, England, United Kingdom

Ana Luisa Carvalho, UCIBIO, REQUIMTE, Departamento de Química, Faculdade de Cieñcias e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Kinlin L. Chao, Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD 20850

Ron Diskin, Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel

Krzysztof Fidelis, Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616, USA

Carlos M.G.A. Fontes, CIISA - Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

Folmer Fredslund, Department of Chemistry, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

Harry J. Gilbert, Institute for Cell and Molecular Biosciences, University of Newcastle, Newcastle upon Tyne NE2 4HH, UK

Celia W. Goulding, Department of Molecular Biology & Biochemistry /Pharmaceutical Sciences, University of California Irvine, Irvine, CA 92697, USA

Marcus D. Hartmann, Department of Protein Evolution, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany

Christopher S. Hayes, Department of Molecular, Cellular and Developmental Biology /Biomolecular Science and Engineering Program, University of California, Santa Barbara, Santa Barbara, CA 93106, USA

Osnat Herzberg, Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD 20850; Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742

Johan C. Hill, Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, England, United Kingdom

Andrzej Joachimiak, Midwest Center for Structural Genomics /Structural Biology Center, Biosciences Division, Argonne National Laboratory, USA; Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637, USA

Gert-Wieland Kohring, Microbiology, Saarland University, Campus Building A1.5, Saarbrücken, D-66123 Saarland, Germany

Roman I. Koning, Netherlands Centre for Electron Nanoscopy, Institute of Biology Leiden, Leiden University Einsteinweg 55, 2333 CC Leiden, the Netherlands; Department of Molecular Cell Biology, Leiden University Medical Center, P.O.Box 9600, 2300 RC, Leiden, The Netherlands

Leila Lo Leggio, Department of Chemistry, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen Ø, Denmark

Marco Mangiagalli, Department of Biotechnology and Biosciences, University of Milano-Bicocca, Piazza della Scienza 2, 20126, Milano, Italy

Karolina Michalska, Midwest Center for Structural Genomics /Structural Biology Center, Biosciences Division, Argonne National Laboratory, USA

John Moult, Institute for Bioscience and Biotechnology Research, Department of Cell Biology and Molecular genetics, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850, USA

Shabir Najmudin, CIISA - Faculdade de Medicina Veterinária, Universidade de Lisboa, Avenida da Universidade Técnica, 1300-477 Lisboa, Portugal

Marco Nardini, Department of Biosciences, University of Milano, Via Celoria 26, 20133 Milano, Italy

Valentina Nardone, Department of Biosciences, University of Milano, Via Celoria 26, 20133 Milano, Italy

Didier Ndeh, Institute for Cell and Molecular Biosciences, University of Newcastle, Newcastle upon Tyne NE2 4HH, UK

Thanh H. Nguyen, Department of Macromolecular Structures, Centro Nacional de Biotecnologia (CSIC), calle Darwin 3, 28049 Madrid, Spain

Guido Pintacuda, Université de Lyon, Centre de RMN à Très Hauts Champs, Institut des Sciences Analytiques (UMR 5280 - CNRS, ENS Lyon, UCB Lyon 1), 69100 Villeurbanne, France

Sandra Postel, Institute of Human Virology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Mark J. van Raaij, Department of Macromolecular Structures, Centro Nactional de Biotecnologia (CNB-CSIC), calle Darwin 3, 28049 Madrid, Spain

Pietro Roversi, Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, England, United Kingdom

Amir Shimon, Department of Structural Biology, Weizmann Institute of Science, Rehovot, Israel

Eric J. Sundberg, Institute of Human Virology, Department of Medicine and Department of Microbiology and Immunology, University of Maryland School of Medicine, Baltimore, MD 21201, USA

Kaspars Tars, Latvian Biomedical Research and Study Center, Rātsupītes 1, LV1067, Riga, Latvia; Faculty of Biology, Department of Molecular Biology, University of Latvia, Jelgavas 1, LV-1004 Riga, Latvia

Nicole Zitzmann, Oxford Glycobiology Institute, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, England, United Kingdom

Torsten Schwede, Biozentrum /SIB Swiss Institute of Bioinformatics, Klingelbergstrasse 50, 4056 Basel, Switzerland

Keywords: X-ray Crystallography; NMR; CASP, Protein Structure Prediction.

Abbreviations:

CASP, community wide experiment on the Critical Assessment of Techniques for Protein Structure Prediction; **VLP**, virus-like particle; **TfR1**, Transferrin Receptor 1; **WWAV**, Whitewater Arroyo Virus; **GP1**, glycoprotein 1; **RG-II**, Rhamnogalacturonan-II; **HGM,** Human gut microbiota; **GH**, Glycoside hydrolases (GH); **IBP**, ice binding protein; **TH**, thermal hysteresis; **IRI**, ice recrystallization inhibition.

Author contributions:

Names of the authors contributing to specific sections are provided in the sections' titles; concept, abstract, introduction, editing and coordination - by AK, KF, JM and TS.

Abstract

The functional and biological significance of the selected CASP12 targets are described by the authors of the structures. The crystallographers discuss the most interesting structural features of the target proteins and assess whether these features were correctly reproduced in the predictions submitted to the CASP12 experiment.

Introduction

Integrity of the CASP experiment rests on the blind prediction principle requesting models to be built on proteins of unknown structures. To get a supply of modeling targets, the CASP organization relies on the help of the experimental structural biology community. In the latest seven experiments (2002-2014), the vast majority (>80%) of CASP targets came from structural genomics centers participating in the Protein Structure Initiative (PSI) program. With the disintegration of the PSI in 2015, CASP faced a challenging task of replenishing the target supply normally provided by the PSI Centers. Dealing with this problem required diversification of target sources and going beyond the existing network of the recurring CASP target providers. Soliciting for targets, the organizers directly approached a wider set of structure determination groups, and also worked out a better protocol for obtaining and analyzing information about the structures placed on hold with the PDB. These efforts bore fruits, and 82 targets were secured for the CASP12 experiment. This number is quite impressive (considered that targets were collected in a short 3-month span of time) and is only somewhat smaller than the number of targets in a typical PSI-era CASP experiment (cf. 100 targets in the most recent CASP11 experiment). It is also worth mentioning that CASP12 targets came from 33 different protein crystallography groups stationed in 17 countries worldwide. Because of this variety, CASP12 targets exhibited wide diversity of sizes (from 75 to 670 residues), difficulties (from high accuracy modeling targets to new folds), quaternary structure composition (from single-domain targets to hetero-complexes), organisms (from rare extremophilic archaea from the depths of the Red Sea to *Homo Sapiens*), and protein types (from globular to viral and membrane). Such diversity is vital for comprehensive testing of prediction methods. CASP organizers, who are co-authors of this paper, want to thank every experimentalist who contributed to CASP12 and this way helped developing more effective protein structure prediction methods. The list of all

crystallographers who contributed targets for the CASP12 experiment is provided in Table 1 of the Supplementary material.

This manuscript is the fourth in a series of CASP target highlight papers[1-3]. The chapters of the paper reflect the views of the contributing authors on twelve CASP12 targets: 1) the flagellar cap protein from *Pseudomonas aeruginosa* – **T0886**; 2) bacteriophage AP205 coat protein – **T0859**; 3) toxin-immunity protein complex from the contact-dependent growth inhibition system of *Cupriavidus taiwanensi* – **T0884/T0885**; 4) sorbitol dehydrogenase from *Bradyrhizobium japonicum* – **T0889**; 5) C-terminal domain of human gasdermin-B – **T0948**; 6) receptor-binding domain of the whitewater arroyo virus glycoprotein – **T0877**; 7) glycoside hydrolase family 141 founding member BT1002 – **T0912**; 8) an DNA-binding protein from *Aedes aegypti* – **T0890**; 9) snake adenovirus-1 LH3 hexon-interlacing protein – **T0909**; 10) an ice-binding protein from Antarctica – **T0883;** 11) a domain of UDP-glucose glycoprotein glucosyltransferase from *Chaetomium thermophilum* – **T0892;** and 12) a cohesin from *Ruminococcus flavefaciens* scaffoldin protein complexed with a dockerin – **T0921/T0922**. The results of the comprehensive numerical evaluation of CASP12 models are available at the Prediction Center website (http://www.predictioncenter.org). The detailed assessment of the models by the assessors is provided elsewhere in this issue.

1. FliD, the flagellar cap protein from *Pseudomonas aeruginosa* PAO1 (CASP: T0886, Ts886, PDB: 5FHY) – provided by Sandra Postel and Eric J. Sundberg.

Bacterial flagella are long helical cell appendages that are important for bacterial motility and pathogenicity [4]. These extracellular hollow filaments are formed by thousands of copies of FliC (flagellin) molecules and connected via a hook to the flagellar rotary motor

anchored in the bacterial membrane [5]. The motor drives the propeller like motion of the filament that confers swimming motility to the bacteria [6]. An important structural and functional component of bacterial flagella is the flagellar capping protein, FliD, that is located at the distal end of the flagellar filament [7]. Unfolded FliC molecules are translocated from the cell cytoplasm through the hollow filament pore to the tip of the growing flagellum where FliD regulates flagellar assembly by chaperoning and sorting FliC proteins. An absence of FliD leads to improperly constructed filaments and, consequently, impaired bacterial motility and infectivity [8]. In the most commonly studied organism for flagella, *Salmonella serovar Typhimurium*, FliD is known to form a homopentameric complex on the tip of the flagellum, as shown in a low-resolution cryo-EM structure [7,9,10]. Until recently, these data provided the only available structural insight to FliD. Our crystal structure of a large fragment of FliD, $FliD_{78-405}$, from *Pseudomonas aeruginosa* PAO1 was the first high-resolution structure of any FliD from any bacterium, providing novel details concerning FliD function [11].

In our crystal structure [11], the *Pseudomonas* $FliD_{78-405}$ monomer exhibits an L-shaped structure **(Figure 1A),** which can be divided into two globular domains and a helical region. Domain D3 is a loop insertion into domain D2 and both domains have structural similarity to other flagellar proteins. Residues 309 to 405 of $FliD_{78-405}$ are highly flexible as revealed by hydrogen/deuterium exchange (HDX) and, therefore, we were unable to model those residues in our structure. Full-length *Pseudomonas* $FliD_{1-474}$ encodes predicted N- (residues 1 to 77) and C-terminal (residue 406 to 474) coiled coil domains that prohibited crystallization in our hands.

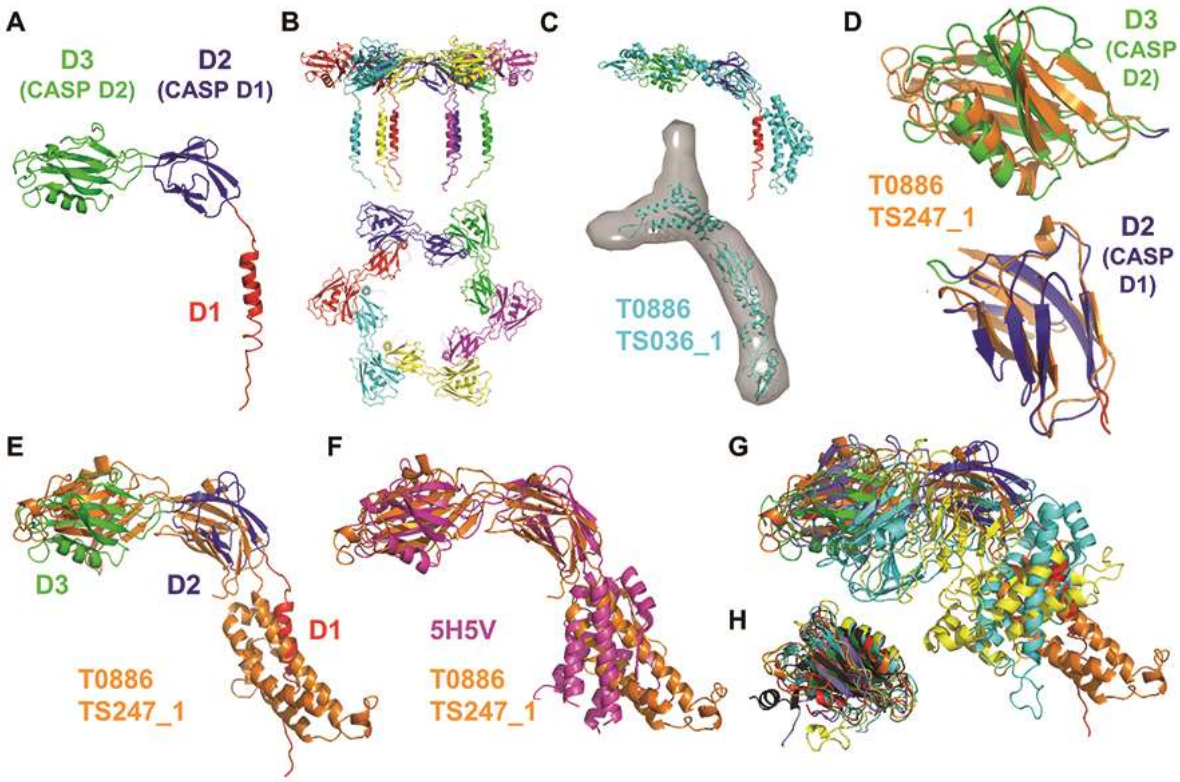

Figure 1.

In contrast to the *Salmonella* FliD that forms a pentamer, *Pseudomonas* FliD adopts a hexameric oligomeric state in the crystal structure (Figure 1B), as well as in solution and functions as a hexamer *in vivo* [11]. The number of protofilaments that comprise the flagellar filament upon which FliD oligomers reside varies between bacteria [12], suggesting that FliD oligomer stoichiometries also vary between bacteria, which is supported by our results. More recently, the crystal structure of FliD from *E. coli* that includes all residues except the N- and C-terminal coiled coils showed that this FliD protein also forms a hexamer [13].

Pseudomonas FliD was included in CASP12 as a regular target and small-angle X-ray scattering (SAXS)-assisted target. SAXS data of the monomeric full-length protein, $FliD_{1-474}$, for which no crystal structure yet exists, was collected and the data provided to the modelers

to aid the structure prediction process of the shorter construct that we had crystallized. All the SAXS-assisted target models exhibit low similarity to the FliD crystal structure as shown in an overlay of the best model T0886TS036_1 with our crystal structure in Figure 1C, but do fit well into the SAXS envelope (Figure 1C).

The models obtained during the regular prediction round without using the SAXS envelopes to assist model-building vary greatly. The highest ranked model T0886TS247_1 closely resembles the crystal structure of *Pseudomonas* $FliD_{78-405}$ on the individual domain level (Figure 1D). However, the connection between domain D2 (CASP domain D1) and domain D3 (CASP domain D2) diverges resulting in a relative positioning of these two domains that is different than in the crystal structure (Figure 1E). The multi-domain-like SAXS molecular envelope of $FliD_{1-474}$ may have made it difficult to predict the exact positioning of the individual domains (Figure 1C). Residues 309 to 405 of $FliD_{78-405}$, which we could not model in the crystal structure due to poor or missing electron density, were in general modeled as helical bundles in T0886TS247_1. A superposition with the recently solved crystal structure of *E. coli* $FliD_{43-416}$ (PDB 5H5V [13]) shows the correct prediction of helical bundles in those regions, but also places those bundles in a different orientation relative to the D2 and D3 domains, as well as differences in the placement of individual helices (Figure 1F). These discrepancies between model and structure may be due to the high flexibility in the linker region and in the helical regions that we detected by HDX[11].

Compared to T0886TS247_1, all of the other models exhibit substantially less similarity to the $FliD_{78-405}$ crystal structure (Figure 1G). Models of domain D3 (CASP domain D2) alone, however, exhibited greater likenesses to the crystal structure with secondary structural elements generally predicted properly (Figure 1H). This might be related to the lower flexibility (as shown by HDX) of domain D3 in comparison to the rest of the FliD

molecule. Overall, FliD seemed to be a difficult target to model, despite the SAXS data provided, and only domain D3 appeared to yield more accurate models by multiple modeling groups.

2. Structure of bacteriophage AP205 coat protein (CASP: T0859; PDB: 5FS4, 5JZR, 5LQP) - provided by Kaspars Tars, Roman I. Koning and Guido Pintacuda.

Virus-like particles (VLPs) are empty, non-infectious shells of viruses, devoid of genomic nucleic acid, but morphologically similar to the corresponding viruses. VLPs have several applications, the best known of which is vaccine development. For example, VLPs of Hepatitis B virus have been used as successful vaccines for a few decades[14]. VLPs can be used not only as vaccines against the disease, caused by the virus of VLP origin, but also as a powerful platform to induce strong immune response against virtually any antigen[15]. In this case, multiple copies of antigen of interest should be attached to the surface of VLP. The immune system recognizes patterns of regularly repeating antigens on VLP surface as a potential threat to organism, inducing highly elevated titres of antibodies and stronger T-cell responses[16]. To avoid pre-existing immune responses, non-human pathogens are preferable as carriers of antigens. For this purpose, VLPs of ssRNA phages like MS2, Qβ and AP205 have been widely used [17]. ssRNA phages are among the simplest known viruses, used for decades as simple models to study various problems in molecular biology. Capsid of ssRNA phages contains 178 copies of coat protein (CP) and a single copy of maturation protein, responsible for attachment of phage particles to bacterial receptor[18]. When produced in bacteria, recombinant CP of ssRNA phages spontaneously assembles in VLPs, containing 180 copies of CP. Due to strong interactions between two adjacent CP monomers, VLPs can be regarded as built from 90 CP dimers.

For creation of vaccine candidate, the antigen of choice can be attached to VLPs either by chemical coupling or genetic fusion of CP and antigen genes. Genetic fusion is technologically more efficient, since production of vaccine candidate requires only a single protein expression and purification without a need for a chemical coupling step. Since antigens must be presented on the surface of VLPs, the knowledge of the exact three-dimensional structure of VLP provides essential information about suitable sites of insertion of antigens in coat protein sequences. Due to folding problems, large insertions are often tolerated only at either N- or C-termini of CP, but this is possible only if the terminal end of CP is well exposed on the VLP surface. However, in VLPs of ssRNA phages studied so far, like MS2 [19], Qβ [20], GA [21], PP7 [22], PRR1 [23] and Cb5 [24] both terminal ends are poorly exposed on the surface. Instead, a so-called AB loop is well exposed, but only relatively short amino acid sequences can be inserted in it without compromising VLP stability. In contrast, AP205 VLPs have been known before to tolerate significantly longer insertions at both C- and N-termini[25], but the structural reason for this remained unknown. Since we failed to obtain high resolution crystals of recombinant AP205 VLPs, we constructed and crystallized an assembly-deficient AP205 CP mutant, capable to form dimers, but not VLPs. The obtained crystal structure was further fitted into a medium resolution cryo-EM map of native recombinant AP205 VLPs. Additionally, a solid-state NMR structure of AP205 coat protein was obtained from labelled AP205 VLPs. The obtained results revealed that compared to related ssRNA phages, structure AP205 CP is circularly permutated [26], meaning that about 20 N-terminal residues including the first beta strand are found at the C-terminal part instead. This feature is made possible due to the close proximity of N- and C-terminal parts of two monomers within the dimer (Figure2ab). The result is that in AP205 VLPs both N- and C- termini are found in

the same position as AB loops in other phages (Figure 2cd). This provides a structural basis for construction of vaccine candidates using AP205 VLPs.

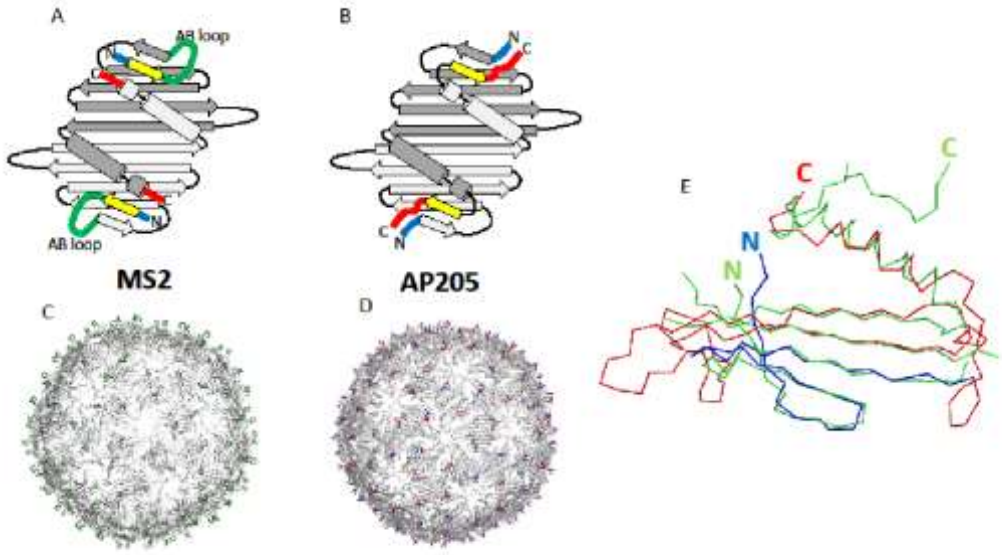

Figure 2.

Out of 499 submitted CASP models, only one had a reasonably accurate overall structure (Figure 2e, red and blue). Model T0859TS001, made by researchers at Francis Crick institute included almost all of the actual secondary structure elements apart from the C-terminal beta strand, which is unique for AP205, compared to other similar phages. About one third of the protein, comprising approximately 40 N-terminal residues was placed fairly accurately in respect to sequence, as compared to the crystal structure. This means that researchers have correctly deduced that the first beta strand is missing in AP205. After residue 40, progressively increasing out-of-register errors occur in the model. At the C-terminal part the out-of-register shift is about 20 residues. Due to this shift, the C-terminal residues are

modelled as alpha-helix although in crystal structure they form the extra (C-terminal) beta strand, not observed in similar phages. Therefore, C-terminal part, is not modelled correctly and does not suggest the placement of C-termini on the surface of VLP, close to AB loops in related phages. Even though the overall precision of the model is somewhat limited, the model correctly suggests that N-terminal part is indeed well-exposed on the surface of VLP and occupies the position of AB loops in related phages. Therefore, in the absence of experimental data, the model T0859TS001 would provide significant biologically relevant information for construction of VLP based vaccines.

3. Structure of the toxin-immunity protein complex from the contact-dependent growth inhibition system of *Cupriavidus taiwanensis* (CASP: T0884/T0885, PDB: 5T87) – provided by Karolina Michalska, Christopher S. Hayes, Celia W. Goulding and Andrzej Joachimiak.

Contact-dependent growth inhibition (CDI) is an important mechanism of inter-cellular competition found in Gram-negative bacteria. CDI^+ cells use CdiB-CdiA two-partner secretion systems to deliver protein toxins directly into neighboring bacteria [27,28]. CdiB is an outer membrane transport protein exporting the CdiA effector onto the cell surface. CdiA recognizes specific receptors on susceptible bacteria and translocates its C-terminal toxin domain (CdiA-CT) into the target cell [29-31]. CdiA proteins carry a variety of toxin domains, most commonly exhibiting nuclease or pore-forming activities [32-35]. To protect against self-inhibition, CDI^+ bacteria produce CdiI immunity proteins, which bind and neutralize cognate CdiA-CT toxins. The variable CdiA-CT toxin region is usually demarcated by a conserved peptide motif, such as the VENN sequence found in enterobacterial CdiAs [33]. Different CdiA-

CTs can be fused to heterologous CdiA proteins at the VENN motif to generate novel chimeric effectors [28,33,34].

We have selected the CdiA-CT/CdiI^{Ctai} complex from *Cupriavidus taiwanensis* LMG 19424 for structural analysis. PSI-BLAST searches for CdiA-CT^{Ctai} homologs recover several predicted S-type pyocins from *Pseudomonas* species and MafB toxins from *Neisseria* species[36]. Other hits include CdiA-CT domains from *Rhizobium leguminosarum* and *Achromobacter* strains, and Rhs peptide-repeat proteins from *Streptomyces* species. All of these homologs are predicted to mediate inter-bacterial competition [37,38], though none have been validated experimentally. An HHpred-based search identified the C-terminal domain of 16S rRNA-cleaving colicin E3 [39,40] as a possible structural homolog having 9% sequence identity to CdiA-CT^{Ctai}. The CdiI^{Ctai} immunity protein is less conserved than CdiA-CT^{Ctai}, with homologs sharing ~30-40% sequence identity. An HHpred analysis recovered proteins with α-helical hairpin repeats, with the armadillo-like γ-COP coatomer (13% sequence identity with CdiI^{Ctai}) being the closest match.

The 2.40 Å resolution crystal structure of the CdiA-CT/CdiI^{Ctai} complex (Figure 3A) shows that the toxin putative catalytic domain (75 residues) consists of a central four-stranded antiparallel β-sheet, sandwiched by two N- and C-terminal α-helices and one 3_{10} helix. The immunity protein (116 residues) is composed of three consecutive α-hairpins creating an armadillo-like structure. The N-terminal β-strand of CdiI^{Ctai} protrudes from the helical body to complement the CdiA-CT^{Ctai} β-sheet, potentially influencing toxin conformation. This arrangement also suggests that the N-terminal segment of CdiI^{Ctai} is likely disordered in the free CdiI^{Ctai}. A Dali server search for CdiA-CT^{Ctai} homologs identified only low-similarity matches: inorganic triphosphatase (Z-score 3.7, rmsd 3.3 Å, PDB:3TYP) (Figure 3B) and WW domain of human transcription elongation regulator 1 (Z-score 3.5, rmsd 2.9 Å,

PDB:2DK7). More distant hits include *E. coli* ParE toxin (Z-score 3.0, rmsd 2.4 Å, PDB:3KXE) (Figure 3C), which belongs to the barnase/EndoU/colicin E5-D/RelE (BECR) family (PMID:22731697). Although structurally related, these toxins display different activities: ParE family poison DNA gyrase [41], RelE is a ribosome-dependent mRNase [42], and colicins D/E5 cleave the anticodon loops of specific tRNAs [43]. Therefore, the exact biochemical function of CdiA-CT^{Ctai} cannot be predicted easily and may include RNase or DNase activity. The CdiI^{Ctai} fold is well-represented in the PDB and is a popular scaffold for designer proteins. The closest match corresponds to human deoxyhypusine hydroxylase (Z-score 12.3, rmsd 2.0 Å, PDB:4D4Z), followed by protein phosphatase 2 (Z-score 12.3, rmsd 2.5 Å, PDB:2IE3) and other proteins with virtually no sequence similarity to CdiI^{Ctai}. Though many of the homologs engage in protein-protein interactions, none are annotated as an immunity protein.

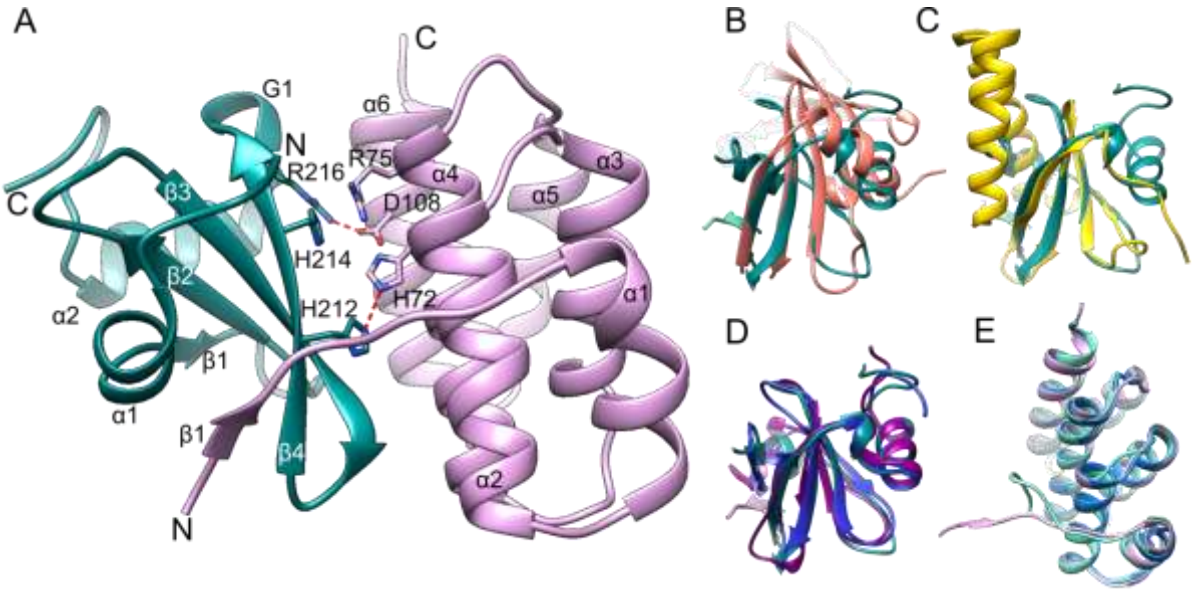

Figure 3.

Antitoxin proteins often bind over nuclease toxin active sites to prevent substrate access. Typically, nuclease toxins are highly electropositive and the cognate immunity proteins carry complementary acidic residues to promote electrostatic interactions. CdiA-CT^{Ctai} contains several basic residues, including conserved His212, His214 and Arg216 (Figure 3A), which may be key catalytic residues. CdiI^{Ctai} is more electrostatically neutral than previously characterized immunity proteins. It directly interacts with the toxin's putative active site using conserved His72, Arg75 and Asp108 residues that form a hydrogen bond, stacking interaction and salt-bridge, respectively. As outlined above, β1 of CdiI^{Ctai} complements the toxin fold.

For the CASP12 competition, CdiA-CT^{Ctai} and CdiI^{Ctai} were modeled as monomers. Out of 43 total predictions, the best model of CdiA-CT^{Ctai} (T0884TS183_1-D1) was generated by QUARK, which uses *ab initio* algorithms with no global template information. This model scored 66 GDT_TS points (% residues under distance cutoff ≤ 4Å), 10 points higher than the next model T0884TS236_1-D1 generated by MULTICOM-CONSTRUCT and T0884TS287_1-D1 from MULTICOM-CLUSTER. The original model was further improved to GDT_TS of 76 by PKUSZ_Wu_group (TR884TS118_1).

T0884TS183_1-D1 closely resembles the crystal structure, though helix α1 is misoriented and the β3-β4 hairpin is distorted (Figure 3D). However, we note that toxin helix α1 is constrained by the immunity protein in the CdiA-CT/CdiI^{Ctai} complex. Therefore, it is possible that the free toxin domain adopts the conformation predicted by the computational model. Toxin residues that interact with the immunity protein are generally located in proper positions,

though a more accurate prediction of β4 would provide better agreement for conserved His212 and His214.

CdiI^{Ctai} (T0885) is a more straightforward structure prediction target, with fewer discrepancies between the 43 predicted models. The best model, T0885TS005_2-D1 (Figure 3E), was generated by BAKER-ROSETTASERVER with 88 GDT_TS points (or 92 without 10 N-terminal residues). This score is 4 and 7 points higher than the subsequent structures T0885TS405_1-D1 generated by IntFold4, and T0885TS183_1-D1 produced by QUARK. 11 more models scored within 15 points of the best scoring models. As we found with CdiA-CT^{Ctai}, the major misalignments were observed for peripheral elements (β1 and the C-terminus of helix α6) involved in protein-protein interactions. The N-terminally truncated variant of the protein achieved 95 GDT_TS in the refinement (TR885TS247_1-D1).

This example shows that computational prediction can yield models with correct folds, and when combined with sequence conservation analysis, can inform rational mutagenesis and biochemical analyses. Important questions remain on how to identify the best computational model in the absence of the experimental data. In addition, though *in silico* approaches often provide insights into protein-protein interactions, such models for the CdiA-CT-CdiIC^{tai} complex failed to properly predict protein-protein interface, leaving the putative active site fully exposed. Thus crystal structures are still required to confidently determine conformational states important for function and catalysis.

4. Sorbitol dehydrogenase (BjSDH) from *Bradyrhizobium japonicum* (CASP: T0889; PDB: 5JO9) - provided by Leila Lo Leggio, Folmer Fredslund and Gert-Wieland Kohring.

Rare sugars are monosaccharides and their derivatives which are rare in nature and they have attracted interest for potential medical and food applications [44]. Consequently, enzymes able to produce and interconvert rare sugars have also attracted attention. We initiated structural studies of *Bj*SDH as part of a collaborative EU project devoted to the development of an electro-enzymatic flow-cell device for the production of rare sugars [45]. Several enzymes were investigated in the study, and *Bj*SDH was selected for structure determination due to some favorable properties. First of all, while *Bj*SDH preferentially catalyses the oxidation of D-glucitol (a synonym for D-sorbitol) to D-fructose, it also can catalyse the oxidation of L-glucitol to the rare sugar D-sorbose with enzymatic cofactor regeneration and high D-sorbose yield [46] (Figure 4a). Sorbitol dehydrogenases are additionally of particular interest in biosensor technology, since D-sorbitol is a marker for onset of diabetes as well as a food ingredient [47]. Furthermore, it is a thermostable enzyme with T_m of 62 ◦C [46], which is a desirable property for potential industrial use and biosensor technology, as thermostability often correlates with general stability. *Bj*SDH is a Zn-independent short chain dehydrogenase using NAD^+/NADH as non-covalently bound cofactor.

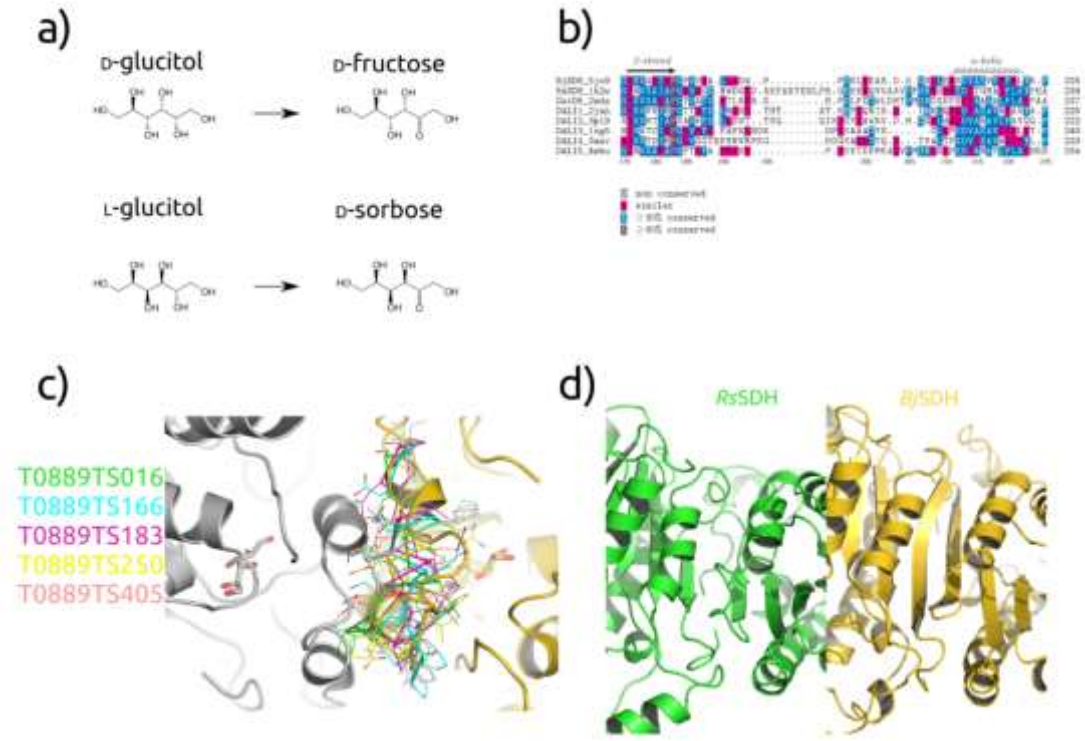

Figure 4.

Structure determination [48] was not straightforward since the resolution was limited. The resolution could be estimated to 2.9Å according to $CC_{1/2}$ of about 50% in the outer resolution shell [49], but closer to 3.2Å with more conventional evaluation of resolution limit at $I/\sigma(I)$ around 2. Furthermore, the Molecular Replacement model chosen (PDB code 4NBU [50]) was only 29 % sequence identical to target (after structure-based alignment). All the closest structural relatives identified with DALI after structure determination (reported in Fredslund et al [48]), have only around 30% sequence identity, and while most are dehydrogenases, none are denoted as sorbitol dehydrogenases. *Bj*SDH was co-crystallized with NAD^+ and D-glucitol. D-glucitol could be modelled in the electron density map and phosphate is clearly bound, mimicking part of the cofactor, however a full co-factor molecule could not be modelled. This is probably due to presence of 1.4 M NaH_2PO_4/K_2HPO_4 in the crystallization

conditions, competing with the cofactor. Although there is only one molecule in the asymmetric unit, the enzyme forms a tetramer in the crystal structure due to crystallographic symmetry, and this is also assumed to be the predominant form in solution [48].

To see if structural features were correctly predicted by the top models in CASP12, we selected the 5 top scoring hits (based on GDT_TS) for compariosn. These 5 top models were based solely, or in part on PDB entry 2JAH (or the related 2JAP), clavulanic acid dehydrogenase from *Streptomyces clavuligerus* [51], which was also the top DALI hit.

The structure showed that the catalytic tetrad (Asn112, Ser140, Tyr153 and Lys157 in *Bj*SDH) present in short chain dehydrogenases is highly conserved structurally in *Bj*SDH compared to similar dehydrogenases. The 5 top hits (based on GDT_TS) from CASP12 all, unsurprisingly, predict correctly the positioning of the catalytic residues.

In contrast the length, sequence and conformation of the loop lid covering the active site is poorly conserved (Figure 4b), even in enzymes with relatively similar specificity like *R. sphaeroides* sorbitol dehydrogenase *Rs*SDH [52]. This loop is different in the 5 top scoring hits from CASP12 and the crystal structure, and indeed also in the model used for molecular replacement. Since the resolution of the crystal structure is limited, and this loop in particular was difficult to trace, there might be errors in the crystallographic model too, but the conformation of the loop from several CASP12 models is definitely incompatible with crystal packing (Figure 4c) and cannot accurately represent the conformation it assumes in the crystal. On the other hand, packing could have affected the conformation and furthermore, the loop is involved in ligand binding, which would not be taken into account explicitly by the modelling programs and could also affect its conformation.

One of the most important features of *Bj*SDH was its thermostability[46], as the knowledge of its structural determinants may help stabilizing related enzymes by protein

engineering. In particular, we compared the structure to the sorbitol dehydrogenase *Rs*SDH, for which the melting temperature by CD spectroscopy was also measured and found to be much lower than for *Bj*SDH under similar conditions (T_m of 47 ◦C vs 62◦C). One of the striking features in *Bj*SDH is a much higher Proline/Glycine ratio compared to *Rs*SDH, a feature which is obvious from the sequence and does not require knowledge of the 3D structure. An additional feature which is likely to affect stability becomes obvious only through analysis of the quaternary structure. As previously mentioned *Bj*SDH is a tetramer in the structure and in solution, as are many members of the short chain dehydrogenase family, and probably also *Rs*SDH[52]. In *Bj*SDH, two monomers of the tetramer make strong interactions so that a continuous β-sheet is formed between the two monomers, while this is not the case in *Rs*SDH, indicating a less stable tetramer in the latter (Figure 4d). As the top CASP12 models for *Bj*SDH were all based on the clavulanic acid dehydrogenase structure, which shares tetrameric formation, including the continuous β-sheet between two monomers, the resulting top models of *Bj*SDH all received this motif from the template and thus are modelled consistent with an intersubunit β-sheet formation, despite the fact that the models are monomeric. This is not surprising, since the determining factor in producing an accurate model of the interaction are the backbone atoms, and are thus easily transferred to a homology model.

In conclusion, the top CASP12 models reproduce correctly some but not all biologically and biotechnologically interesting features of SDH, for example they cannot predict the lid loop conformation, as this loop is poorly conserved.

5. Crystal Structure of the C-terminal Domain of Human Gasdermin-B (CASP: T0948; PDB: 5TJ4, 5TJ2, 5TIB) - provided by Kinlin L. Chao and Osnat Herzberg.

Biological Significance of Gasdermin-B. The human genome encodes four gasdermins (GSDMA-D) that are expressed in epithelial cells of the gastrointestinal tract and skin, regulating the maintenance of the epithelial cell barrier, normal cell proliferation, and differentiation processes via the lytic and non-lytic forms of programmed cell-death [53,54]. Based on the different protein levels in cancers, human GSDMA, GSDMC and GSDMD are considered tumor suppressors and GSDMB (CASP12 target T0948), a tumor promoter. *GSDMB* amplification and GSDMB overexpression indicate poor response to HER2-targeted therapy in HER2-positive breast cancer [55]. The N-terminal domain of gasdermins possesses membrane-binding activity, whereas the C-terminal domain autoregulates the lipid binding function. Multiple genome-wide association studies (GWAS) revealed a correlation between single nucleotide polymorphisms (SNPs) in the protein coding and transcriptional regulatory regions of the neighboring *GSDMA, GSDMB* and *ORDML3* genes in the 17q12.2.1 loci with susceptibility to asthma [56], type 1 diabetes [57,58], Crohn's disease, ulcerative colitis [58,59] and rheumatoid arteritis [58,60]. Pal and Moult identified 2 *GSDMB* SNPs (dbSNP:rs2305479 and dbSNP:rs2305480) in linkage disequilibrium with a marker of disease risk [58]. They correspond to a Gly299 → Arg299 change (rs230549), and a Pro306 → Ser306 change (rs2305480) in the C-terminal domain of GSDMB (GSDMB_C) (numbering scheme according to Uniprot isoform Q8TAX9-1). Analyses of the 1000 Genomes Project Consortium data [61] showed co-occurrence of the 2 SNPs (Gly299:Pro306 or Arg299:Ser306) with ~50% occurrence of each combination in the general population (Pal and Moult, unpublished). Unlike monogenic diseases which are caused by high penetrance SNPs in single

genes, complex-trait diseases are associated with multiple low penetrance SNPs in multiple genes. Because of linkage disequilibrium, most of the SNPs present in a genome are actually not disease causative and the challenge for the large-scale genome sequencing is to reveal the disease causative SNPs. The structural studies of GSDMB_C provided insights into the possible mechanisms that the SNPs may contribute to disease risk [62].

Key features of Gasdermin-B C-terminal domain. The structure of mouse Gsdma3 (PDB 5B5R, an orthologue of GSDMA) revealed a 2-domain protein connected by a long flexible linker. The N-terminal lipid-binding domain folds into an α+β structure and the C-terminal inhibitory domain adopts an α-helical fold comprising 8 helices [63]. The 7-helix bundle topology of GSDMB_C (α5-α11 in PDB 5TJ4, 5TJ2, 5TIB) is the same as that of Gsdma3, except that it lacks a Gsdma3 subdomain comprising an α-helix and a 3-stranded β-sheet between the last two α-helices (Fig 5A-C).

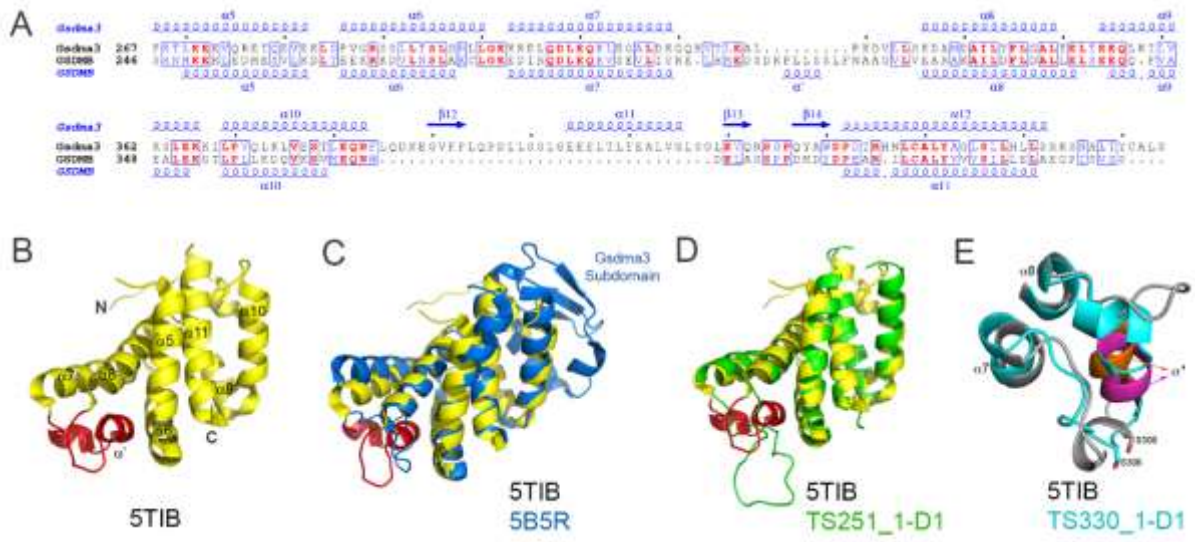

Figure 5.

We determined three crystal structures of the GSDMB_C containing (1) the Arg299:Ser306 pair corresponding to individuals with increased disease risk, (2) the Gly299:Pro306 present in healthy individuals, and (3) the Gly299:Ser306 combination, one from each allele [62]. The SNP residues at positions 299 and 306 are located on a loop connecting the α7 and α8 helices of GSDMB (Figure 1A&B). Three GSDMB_C structures provide 16 independently determined molecules in their asymmetric units: 6 with Ser at position 306 and 10 molecules with Pro at that position. All 16 versions of this loop contain a 5-residue α-helix (α', Pro309-Ser313) (Figure 5A&B). However, the loops with Ser306 adopt an additional well-ordered 4-residue helical turn (Met303-Ser306) between the α7 and α' helices (Figure 5B). By contrast, the loops with a Pro306 do not form this helical turn and each assumes different backbone conformations [62]. In addition, a Gly299→ Arg299 alters the charge distribution on the protein surface. Examination of the structures shows that, unlike a more flexible Ser306 side chain, Pro306 cannot be accommodated at the end of the helical turn because its side chain would clash with main chain carbonyl atoms of the preceding residues. One or both of these changes may contribute to the susceptibility of individuals to develop diseases by possibly modulating the selectivity and binding affinity of its N-terminal domain to lipids or the association with partner proteins, for example HSP90β or fatty acid synthase [64].

CASP12 predictions for the functionally important regions of GSDMB_C (T0948). The 166-residue GSDMB_C CASP12 target sequence (T0948) contained the Arg299:Ser306 pair found in individuals with increased disease risk (PDB 5TIB). The publication of the full-length Gsdma3 structure shortly prior to the CASP12 prediction deadline provided a homologous template for T0948 (PDB 5B5R [63]). T0948 and the 198 residues Gsdma3 C-terminal domain share 34.5% sequence identity and super positioning yields a RMSD of 2.3

Å for 113 shared Cα positions (Figure 5C). However, a 33 amino acid Gsdma3 subdomain between α10 and the last helix (Gsdma3 α12 or GSDMB α11) corresponds to a disordered loop in GSDMB that is too short to form an analogous subdomain (Met366–Tyr382) [62], and therefore could not be predicted. This Gsdma3 region is functionally important because it interacts with a segment on the N-terminal domain that is involved in membrane disruption [63].

A total of 422 predictions for T0948 were deposited in CASP12, and 150 of them had GDT_TS scores > 70. The Gsdma3-based models for T0948 were quite accurate for the well-aligned core 7-helix bundle region, but not for the functionally important polymorphism loop. The superposed structures of GSDMB_C and the highest GDT_TS scored model from group 251 (myprotein-me server, Skwark and colleagues) illustrate the similarity within the core 7-helix bundle (Figure 5D). However, the predictions for the polymorphism loop conformation (i.e. residues Arg299–Val322 of GSDMB corresponding to Arg54-Val77 in T0948) were poor, presumably because the GSDMB loop is 8 residues longer than that of Gsdma3 and lacks significant sequence homology (Figure 5A [62]). Encouragingly, many top models (although not TS251_1-D1, Figure 5D) predicted the α' helix (Pro309-Ser313) in the polymorphism loop. However, its length was overestimated and its orientation was wrong in all cases. Large differences exist even for the position-specific alignment of the polymorphism loop closest to the crystal structure (e.g., group 330, Laufer_seed, Perez and colleagues - Figure 5E). No group reproduced in their prediction the 4-residue helical turn preceding Ser306, a key structural difference that distinguishes the GSDMB produced by IBD and asthma patients from that of healthy individuals. Thus, the GSDMB example shows that prediction of the conformations of large loops that deviate substantially from their template structures has not yet achieved the level of accuracy required for drawing conclusions about structure-function relationships.

6. Receptor-binding domain of the Whitewater Arroyo Virus glycoprotein: studying pathogenicity from a structural point of view (CASP: T0877; PDB: 5NSJ) – provided by Amir Shimon and Ron Diskin.

A subgroup of the enveloped RNA viruses that are known as arenaviruses attach to Transferrin Receptor 1 (TfR1), a highly conserved housekeeping protein, and use it as their cellular receptor for cell entry. For binding to TfR1 and for subsequently catalyzing fusion between the viral and the host membranes, arenaviruses use a class-I trimeric spike complex of which the GP1 portion is serving for receptor binding. Several viruses from this group are pathogenic and can cause disease in humans, due to their ability to utilize the human-TfR1 (hTfR1) in addition to TfR1 from rodents, which are the natural reservoir of these viruses.

Since both pathogenic and non-pathogenic arenaviruses use similar rodent-TfR1 receptors but only the pathogenic viruses can utilize hTfR1, we wanted to understand what the structural barriers are that prevent non-pathogenic viruses from doing so. This information is important if we want to understand the molecular mechanisms that may allow non-pathogenic viruses to emerge into the human population as novel pathogens. Our model system for studying such structural barriers is the non-pathogenic Whitewater Arroyo virus (WWAV) [65,66]. Critical structural information that allows us to perform this study is the crystal structure of a GP1 from the pathogenic Machupo arenavirus in complex with hTfR1 that was solved by the Harrison group [67]. Based on the crystal structure of WWAV-GP1 that we have solved and using the structural information of the Machupo-GP1/hTfR1, we were able to model a putative complex of WWAV-GP1 / hTfR1 (Figure 6A). Relevant to this effort is the observation that Machupo-GP1 fully adopts a TfR1-compatible conformation when it is in the unbound state[68].

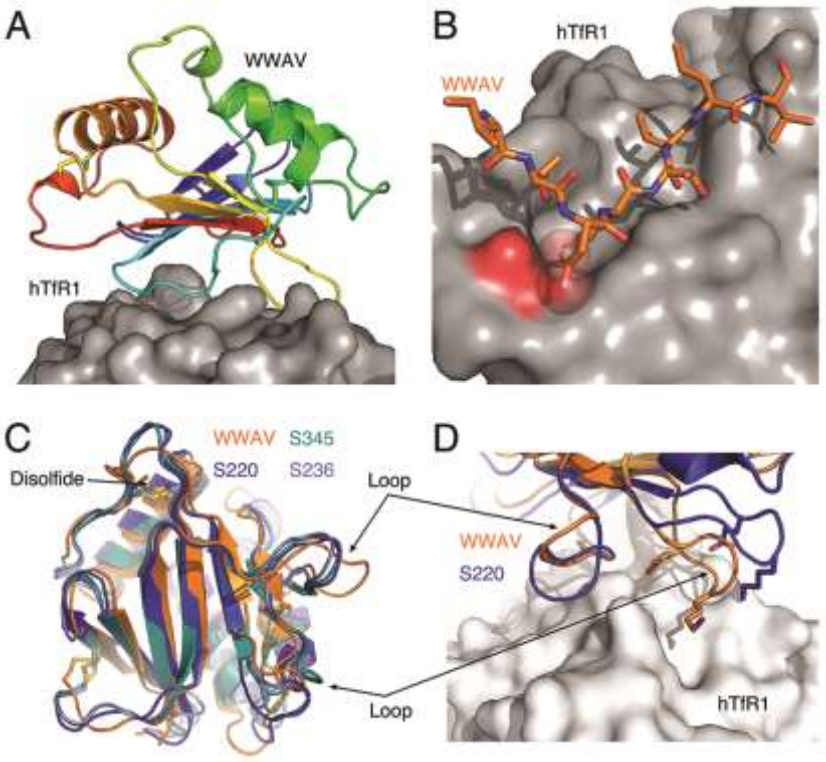

Figure 6.

After modeling the WWAV-GP1 / hTfR1 complex we conducted structural analysis to identify potential barriers that may account for the incompatibility of these two proteins. This analysis indicated that there are two prominent structural barriers on hTfR1 that interfere with the binding of WWAV-GP1 (Figure 6B), compared to binding of TfR1-orthologs from rodents. Interestingly, the same incompatibilities equally affect the binding of GP1 from pathogenic viruses. The reason that pathogenic viruses can engage with hTfR1 is a more elaborate set of weak interactions that they evolved to form with TfR1 that allows them to overcome the energetic cost associated with the structural incompatibilities. Although there are some conserved interactions that GP1 from different viruses form with TfR1, many of the interactions are virus-specific leading to an overall altered TfR1 binding sites. Thus, targeting

the receptor-binding site of these viruses using classical immunotherapy approach or designing GP1-based vaccines may fail to provide broad response that would be effective against different members of this family of viruses.

To be able to construct and analyze a putative complex of WWAV-GP1/hTfR1, we had to have an accurate structure of WWAV-GP1. Sequence conservation of viral glycoproteins like the GP1 domains from TfR1-tropic viruses is generally very low, due to rapid evolution under strong immunological pressure. Thus, a modeling approach may not fully reveal the fine details that are needed for such an analysis. In CASP12, the GP1 domain from WWAV was designated as a target for automated servers. Most of the predictors were able to provide models that faithfully represent the overall structure of this domain with GDT_TS > 50. We compared the top three models to the crystal structure of WWAV-GP1 (Figure 6C). 'MULTICOM-CONSTRUCT', 'MULTICOM-NOVEL', and 'GOAL' achieved the best overall ranking with GDT_TS of 67.78, 68.66, and 70.25, respectively. The central β-sheet and the α-helices were modeled correctly along the primary structure but slightly deviate from their real positions (Figure 6C). Interestingly, a disulfide bond that WWAV has but is not shared by GP1 domains for which structural information was previously available, was not modeled although the cysteine residues were placed in their correct orientations (Figure 6C). Since this bond influences the local geometry of a near-by loop, the modelers were unable to accurately model its conformation (Figure 6C). In general, the conformations of the loops from the various predictors cluster together, but in conformations that deviate from the real structure of WWAV-GP1. Considering the goal of our study, this is a major drawback since some of the important contacts that GP1 makes with TfR1 are mediated through these loops (Figure 6B). Thus, modeling loops is a challenging task and since loops

are often involved in protein-protein interactions bona fide structural information would be preferred for the type of analysis that we have performed.

7. Structure features and biological significance of a new glycoside hydrolase family 141 founding member BT1002 (CASP: T0912; PDB: 5MPQ) - provided by Didier Ndeh, Arnaud Baslé and Harry J. Gilbert.

Rhamnogalacturonan II (RG-II) is a primary cell wall pectin of plants present in fruits, vegetables, wine and chocolate. It is the most complex carbohydrate known and despite its remarkable structural complexity, it is highly conserved across the plant kingdom [69,70]. RGII is a 10 kda acidic polysaccharide and the structure has recently been revised [69,71]. It consists of 12 different sugars held together in the main structure by at least 21 glycosidic linkages. The basic structure is a linear backbone of α-(1-4)-linked D-galacturonic acid decorated stochastically at O2 with two highly complex octa- and enneasaccharide side chains (chains A and B respectively) and at O3 with two disaccharides chains (chains C and D) and monosaccharide side chains (chains E and F). To elucidate how the human gut microbiota (HGM) has evolved to utilise complex glycans in the diet we investigated the RG-II degradome of the prominent gut microbe *Bacteroides thetaiotaomicron*. The organism is capable of metabolising RGII in in-vitro growth experiments, and combined transcriptomic and biochemical data revealed that at least 23 enzymes induced in culture conditions with RGII as the sole carbon source are directly involved in its metabolism [71,72]. The organism is capable of cleaving 20 out of the 21 unique glycosidic linkages in RG-II and biochemical evidence suggests that the CASP12 target T0912 (BT1002) is one of 7 novel enzymes recruited by *B. thetaiotaomicron* to achieve this purpose[71].

BT1002 is a novel α-L-fucosidase and founding member of the new glycoside hydrolase family 141 (GH141) [73]. BT1002 targets the complex tetrasaccharide structure; 2-O-methyl-α-D-xylose-1,3-α-L-fucose-1,4-α-L-rhamnose-1,3'-β-apiose (mXFRA) that spans the length of RGII side chain-A. The products of mXFRA hydrolysis are the two disaccharides 2-O-methyl-α-D-xylose-1,3-α-L-fucose (-1 subsite) and α-L-rhamnose-1,3'-β-apiose (+1 subsite). As the -1 subsite fucose in mXFRA is substituted at O3 with 2-O-methyl-α-D-xylose, the enzyme must have an extended pocket to accommodate this substitution. The importance of BT1002 in RGII metabolism is exemplified by the fact that genetic mutants lacking the enzyme are unable to metabolise mXFRA during in-vitro growth on RGII, leading to accumulation mXFRA in the growth medium. This implies that the enzyme is unique and indispensable for the breakdown of its target in RGII.

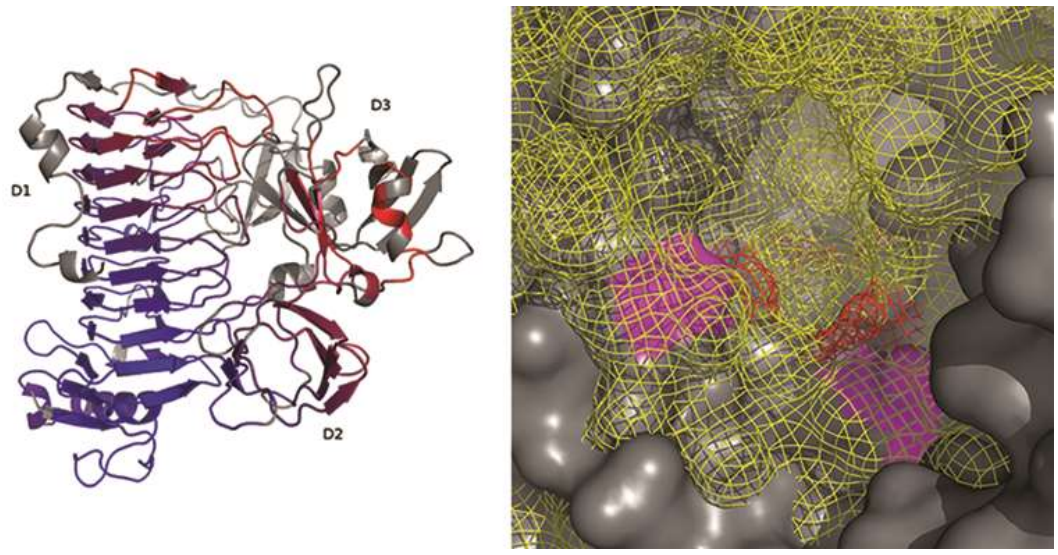

Figure 7.

We solved BT1002 phase problem using selenomethionine single-wavelength anomalous diffraction. The crystallized construct diffracted up to a resolution of 2Å. It comprises 624 amino acids of which 605 are visible in the PDB model (5MQP). BT1002

contains 12 alpha helices and 50 beta-strands forming 6 sheets. The catalytic domain is made of the C-terminal and N-terminal ends of the protein (residues 19-113 and 300-618 respectively) that fold into a beta-parallel helix. An extended loop of the catalytic domain comprising residues 323 to 370 mediates contacts between the beta-parallel helix and the beta-sandwich domains (D1 and D2) made of residues 114 to 299. Additionally the domain D3 is flanked by two alpha-helices (Figure 7, panel A). While efforts to identify specific active site interactions between BT1002 and its tetrasaccharide target are ongoing, we identified two aspartates (Asp523 and Asp564) as potential catalytic residues through site directed mutagenesis[71]. The residues are 6.1 Å apart in a pocket suggesting an acid-base assisted double displacement mechanism. The closest structural homolog we found using a DALI search with the catalytic domain was a GH-120 beta-xylosidase (PDB code 3VSU) with a root mean square deviation of 2.7 Å. While the active site pockets were conserved their primary sequence (20% identity), the catalytic centre and specificity are very different. BT1002 as a CASP12 target was evaluated in full as well as the catalytic domain D1 (residues 24-113 and 299-622). Additionally domains D2 (114-154 and 258-299) and D3 (155-257) were evaluated as targets but their biological significance is not clear. Prediction for the full target was successful overall with 54 models having a GDT_TS score above 30. The model with the highest GDT_TS score is T0912TS303_1 from the wfMESHI-TIGRESS group. To illustrate how well different regions of the protein are predicted, we aligned the BT1002 crystal structure with a mid-range model (T0912TS349_1, HHPred1, GDT_TS 40.78). The result is presented in figure 7 (panel A) where colder colors indicate a close match and hotter colors a higher RMSD (residues in grey were not used). The catalytic domain D1 backbone was very well predicted with the 11 parallel beta-strand stacks of the beta-helix correctly identified (66 predicted models have a GDT_TS above 30 in the T0912-D1 category). This is

not surprising as such a domain had been solved structurally 24 years ago and is well described with multiple examples in the PDB data bank. Side chain positioning is more distant to the crystal protein structure. For instance the catalytic residues Asp564 and Asp523 are distant by about 9 Å in the best D1 model rather than 6.1 Å in the crystal structure. The domain D2 was less correctly modelled overall (58 predicted model with a GDT_TS above 30). Finally the third domain was poorly predicted (only one model with a GDT_TS above 30). The best D3 model T0912TS247_1-D3 from the BAKER group correctly predicted the beta-strands and the beta sandwich but with a registry error. As a consequence, the flanking alpha helices were missed. The overall fold prediction accuracy is essential for this target. Indeed the binding pocket important for ligand recognition and binding, is not only constituted of the surface of the catalytic domain D1 and its extended loop but also the surface of domain D3. Therefore we had to consider only the full target predictions. Figure 7 (panel B) shows an overlay of the best predicted model (T0912TS303_1) and the PDB model (5MQP). The PDB model surface represented as a yellow mesh is clearly smaller than the predicted model surface in dark grey. Additionally the putative catalytic residues are more distant in the predicted model (magenta surface) than in the PDB model (red mesh).

In summary, the BT1002 structure prediction results are very encouraging but shows the challenges facing the community in order to elucidate complex biological functions.

8. A cryptic DNA-binding protein from *Aedes aegypti* (CASP: T0890; PDB: N/A) - provided by Reinhard Albrecht and Marcus D. Hartmann.

During their development, pupating insects (holometabola) may accumulate uracil in the DNA of larval tissues. The protein UDE has been implicated in the development of

holometabola in the late larval stages as a uracil-DNA degrading factor. At the time of its experimental identification in *Drosophila* larval extracts, homologs were only found in holometabola[74]. Its sequence revealed a domain organization with a tandem sequence repeat in the N-terminal half, and several conserved motifs in the C-terminal half of the protein. In some holometabola, only one copy of the N-terminal tandem repeat is found, and it was shown for UDE from *Drosophila melanogaster* (*Dm*UDE), that the first copy of the tandem repeat may be functionally dispensable[75]. Now, however, with more genomes sequenced, sequence searches result in a more diverse picture, including UDE proteins with a more complex domain arrangement in holometabola, but also homologs in plant-pathogenic fungi.

With its developmental implications and due to the absence of sequence matches to domains of known structure, UDE potentially posed an attractive target for the development of insecticides specific to holometabola, or fungicides specific to certain plant pathogens. Initially, UDE caught our attention as we just had identified a novel uracil-recognition mode in the protein cereblon, which we thought could be linked to the recognition of uracil in DNA, and which was in competition to the binding of the drug thalidomide [76,77]. Inspired by the topicality of the Zika virus at that time, we decided to tackle the UDE protein from the yellow fever mosquito *Aedes aegypti* (*Aa*UDE; AAEL003864), a major virus vector.

*Aa*UDE is a canonical UDE protein with the N-terminal tandem repeat and a length of 306 residues; *In vitro*, it showed DNA binding properties similar to *Dm*UDE. While full-length *Aa*UDE withstood crystallization attempts, a recombinant protein corresponding to a proteolytic fragment encompassing residues 87-277, thus omitting the first copy of the tandem repeat and the potentially flexible C-terminal end, yielded well-diffracting crystals. The structure, which we solved via SAD phasing using a platinum derivative, shows an all-helical two-domain protein. The N-terminal domain corresponds to the second copy of the tandem

repeat and forms a three-helix bundle, while the C-terminal half is folded into a compact domain consisting of six helices (Figure 8A).

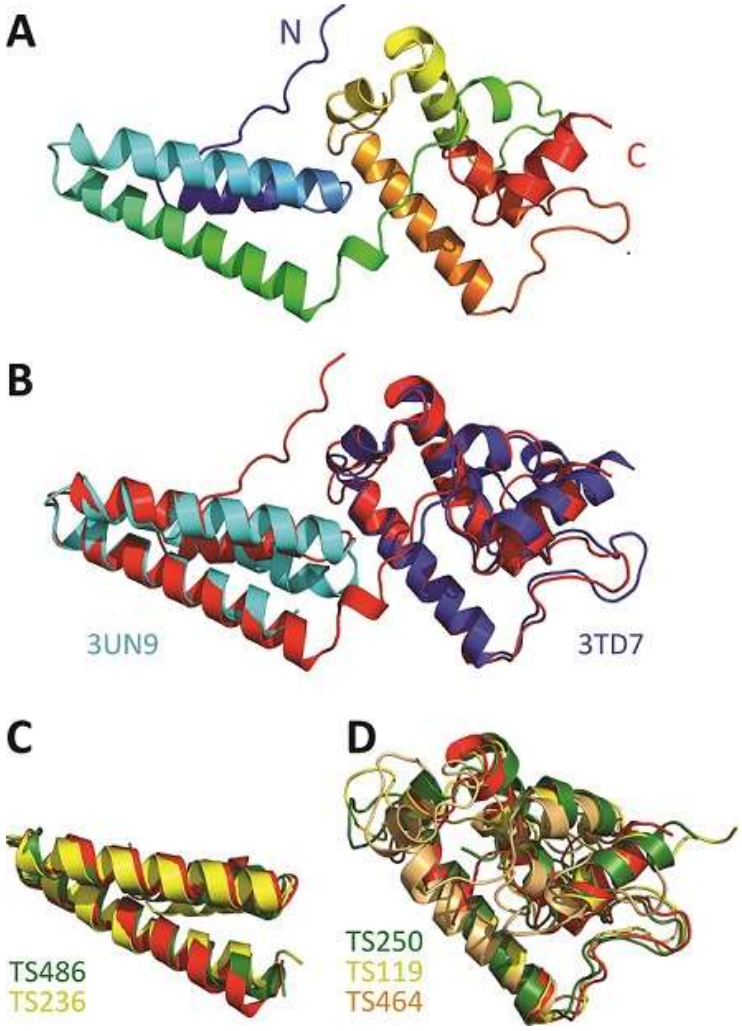

Figure 8.

A DALI search with the full structure yields countless hits for the N-terminal domain with Z-scores of up to 7.5. It had previously been predicted to be a three-helix bundle and had been implicated in DNA binding[75]. This notion is supported by our crystal structure, as this domain presents longer stretches of positively charged residues along its helices. However, the highest-scoring DALI hit is a single - and the only - hit for the C-terminal domain. With a

Z-score of 10.1 it matches a non-conserved additional C-terminal domain of the mimivirus sulfhydryl oxidase R596, which had previously been described as an ORFan domain of novel fold, and which is functionally not understood[78] (Figure 8B).

For the CASP predictors, *Aa*UDE posed a tough but not intractable target. There were many good predictions for the simpler N-terminal domain, and a few good predictions for the C-terminal domain. Curiously, none of the groups could predict both domains. The five best overall models, ranging between a GDT_TS of 44.7 and 33.4 (submitted by the Seok-server, HHGG, HHPred1, HHPred0 and tsspred2) owe their accuracy to the correctly identified similarity of the C-terminal domain to the aforementioned mimivirus ORFan domain. They do, however, not reasonably predict the N-terminal domain. The overall models from rank 6 on mostly contain fair-to-good predictions of the N-terminal but not the C-terminal domain, as they missed the link to the mimivirus protein. The best-matching predictions for the individual domains are depicted in Figure 8C and 8D.

9. The snake adenovirus 1 LH3 hexon-interlacing protein (CASP: T0909; PDB: 5G5N and 5G5O) – provided by Thanh H. Nguyen and Mark J van Raaij.

Adenoviruses are non-enveloped double-stranded DNA viruses that infect vertebrates. Five genera of adenoviruses are known: Mastadenoviruses (infecting mammals), Aviadenoviruses (infecting birds), Ichtadenoviruses (infecting fish), Siadenoviruses (infecting certain birds and amphibian species) and Atadenoviruses (infecting birds, snakes, lizards, ruminants and possums). From the vertices of the icosahedral adenovirus particles (diameter around 100 nm), fiber proteins protrude that are responsible for primary host cell recognition[79]. Internalization in human adenoviruses is mediated by the penton base protein,

but some other adenoviruses lack the necessary integrin-binding sequence. The LH3 gene is a genus-specific Atadenovirus gene found at the left end of the genome, near to the p32K gene. Both LH3 and p32K gene products are believed to be involved in stabilization of the viral capsid[80,81]. The LH3 protein forms trimeric protrusions on the faces of the Atadenovirus particle[81], wedged in between three H_3 hexons and between the H_2, H_3, and H_4 hexons. In total, four LH3 trimers are present on each of the faces, and 80 in the entire Atadenovirus particle.

The Snake Atadenovirus 1 LH3 protein was expressed in *E. coli*, crystallized, the structure was solved using SAD from a mercury derivative crystal, and the structure was refined using native data of a different crystal form at 2.0 Å resolution. Evidence of proteolysis was observed and is consistent with the first 25 residues missing from the experimentally determined structure. The structure revealed a compact, knob-like trimer of right-handed beta-helices, as predicted by the BetaWrap server [82]. The missing part was evident when fitting structure into an overall knob-like shape resulted by SAXS data (Figure 9) and in an overall 11 Å averaged cryo-EM map of SnAdV-1.

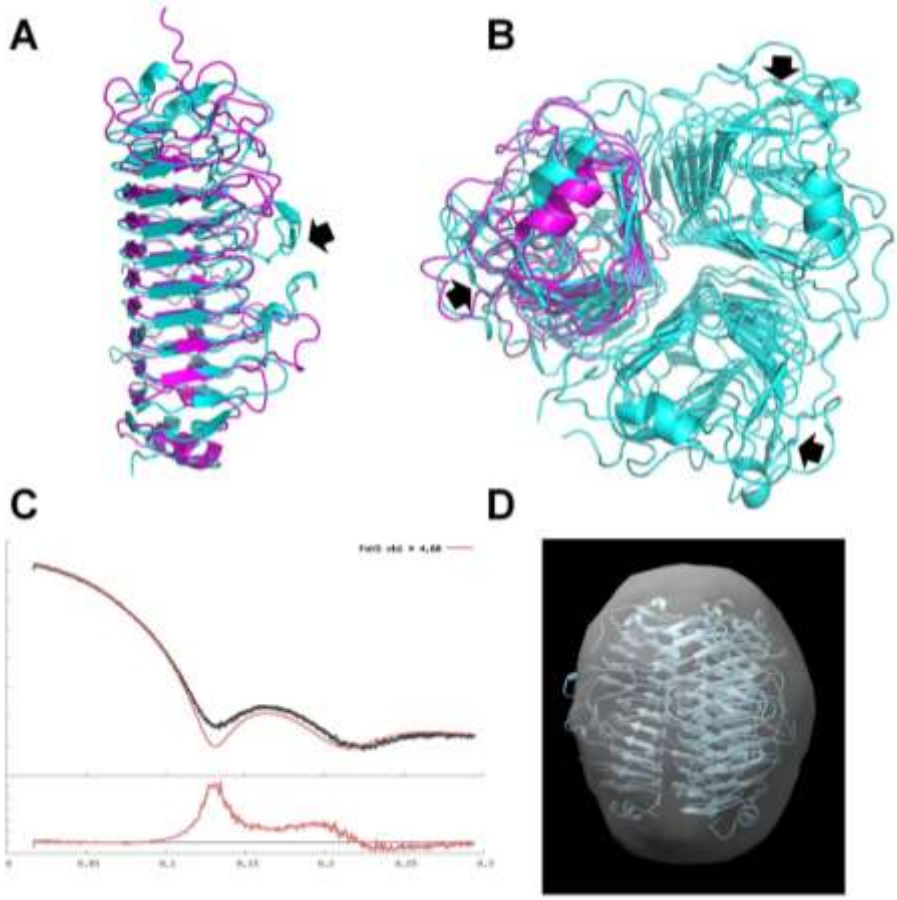

Figure 9.

Each LH3 monomer contains of eleven beta-helical rungs stacked on top of each other. Each beta-helical rung consists of three beta-strands that form long anti-parallel beta-sheets with their counterparts from the other rungs. The beta-sheets are named PB1, PB2 and PB3, following the nomenclature proposed by Mayans [83]. Turns between beta-strands are named T1 (between PB1 and PB2), T2 (between PB2 and PB3), and T3 (between PB3 and PB1). PB1 connects to PB2 mainly by short beta-turns, at the trimer interface, while PB2 connects to PB3 and PB3 to PB1 by longer loops.

Amino acid ladders are observed in the structure of the LH3 protein, as is common for beta-helical structures[83,84]. Asparagine-, isoleucine- and phenylalanine-ladders are found in

the core of each monomer, stabilizing the basic beta-helical architecture of the monomer. The asparagine ladder (residues 193, 214, 248 and 291) is located right at the T1 turn, while the isoleucine (residues 68, 98, 134, 167, 311, 357) and phenylalanine (residues 103, 139, 172, 195) ladders are found in the PB1 and PB2 sheets, respectively. A mixed isoleucine/leucine ladder (Ile84, Leu125, Ile147 and Ile179) in the PB3 sheet. It is possible that the hydrogen bonds in the asparagine ladder helps to avoid out-of-register interactions when the beta-helix folds.

A structural homology search using the DALI server[85] showed the best matches for tailspikes from *Bacillus* phage phi29 [86], *Shigella* phage Sf6 [87] and *Salmonella* phage P22 [88]. Structure superposition between SnAdV-1 LH3 and Sf6 TSP with its ligands revealed a strikingly similar beta-helix topology, despite the low sequence identity (about 13%). It should be noted that the *Shigella* phage SF6 tailspike has endorhamnosidase activity. At the binding site, loops from T2 and T3 turns were identified to involve in the interaction with the lipopolysaccharide substrate. Superimposition between two structures did not show conservation at the loop conformations, however, it is possible to form a potential inter-subunit binding groove in the structure of SnAdV-1 LH3 or on the surface of a single monomer like in the phage P22 tailspike [88]. Evidence for non-conserved binding sites among bacteriophage tailspike proteins was discussed previously [89]. The structural similarity with bacteriophage tailspikes and its location on the viral cell surface suggested the LH3 protein may be involved in binding a (carbohydrate) ligand. However, we have not been able to demonstrate this or a role for the LH3 protein in host interaction.

Structural superimposition between the crystal structure and the best CASP12 model (T0909TS303_5) showed a very similar beta-helical fold between the two. The beta-helix motif was predicted correctly. The best model, with a GDT_TS score greater than 60,

suggested a model comprising three anti-parallel beta-sheets PB1, PB2 and PB3 connected by beta-turns T1, T2 and T3, as observed in the experimentally determined structure. The length and orientation of beta-strands are represented quite accurately, although there are some mismatches or beta-strands missing in the best predicted model. Surface loop conformations are, as expected, predicted less reliably. Structure superimposition of different predicted models also showed that the main beta-helix is present generally but loop conformations are very distinct. Careful inspection of the predicted models shows that most of the beta-strands are well represented in the predicted models in terms of length and location, given that there is low sequence identity of SnAdV-1 LH3 protein with known structures (less than 15%).

It should be kept in mind that SnAdV-1 LH3 protein is a homotrimer. However, the predictions did not take this given feature into account. If they would have, it is likely an overall correct trimeric model could have been derived, which in turn, might have allowed us to solve the structure by molecular replacement without having to resort to a heavy atom derivative. Availability of a SAXS envelope might also have helped to derive an accurate trimeric model computationally.

10. Crystal structure of an ice binding protein from an Antarctic Biological Consortium (CASP:T0883; PDB:N/A) – provided by Valentina Nardone, Marco Mangiagalli and Marco Nardini).

Organisms exposed to permanent subzero temperatures or seasonal temperature dropping are protected from freezing damage by producing Ice Binding Proteins (IBPs) which adsorb to the ice surface and stop ice crystal growth in a non-colligative manner[90]. A measurable effect of ice binding is that IBPs decrease the water freezing temperature, thereby

creating a thermal hysteresis (TH) gap between the melting and the freezing temperature[91]. TH has been explained by the fact that IBP induces a micro curvature on the ice surface. In this way, ice growth is restricted in between the adsorbed IBP and the curved surface. This makes the association of other water molecules thermodynamically unfavorable, causing the decrease of water freezing temperature. The second activity of IBPs is the ice recrystallization inhibition (IRI), that consists in the growth of large ice crystals at the expenses of smaller ones. Ice recrystallization is involved in dehydration and cellular damage and it is very injurious for biological matter[92]. Because of these properties, in recent years the potential application of IBPs has been recognized in several different fields in which materials and substances have to be preserved from freezing, including food processing, cryopreservation, cryosurgery, fishery and agricultural industries, and anti-icing materials development[90,93].

IBPs have been isolated in different species, including fishes, insects, plants, algae, fungi, yeasts and bacteria. Proteins from different sources share the ability to bind ice crystals, but they can exhibit very diverse 3D structures, including small globular proteins, single α-helices, four helix bundles, polyproline type II helix bundles and β-solenoids. Overall, these observations suggest that IBP distribution originates from independent and combined evolutionary events, such as convergent evolution and horizontal gene transfer [90].

Structural studies may contribute to better delineate the "natural history" and the function of IBPs. For instance, many IBPs share threonine-rich repeats, such as Thr-X-Thr or Thr-X-Asx, forming large surfaces complementary to ice crystals. The comparisons of threonine repeats forming the ice-binding sites of structurally very diverse IBPs could help to recognize the core elements involved in ice binding and to understand its mechanism [90].

Among different IBPs, we focused our attention on *Efc*IBP, a bacterial IBP identified by metagenomic analysis of the Antarctic ciliate *Euplotes focardii* and the associated bacterial

consortium. Tested for its effects on ice, recombinant *Efc*IBP shows atypical combination of TH and IRI activities not reported in other bacterial IBPs. Its TH activity was 0.53 °C at 50 μM, but it presented high IRI activity with an effective concentration in the nanomolar range. This value is one of the best described to date. As a result, *Efc*IBP effectively protected purified proteins and bacterial cells from damages deriving freezing. Furthermore, the presence in the *Efc*IBP sequence of a secretion signal seems to indicate that *Efc*IBP might be either concentrated around cells or anchored at the outer cell surface, permitting the entire consortium to thrive/survive at challenging temperature [94]. To shed light on the antifreeze properties of *Efc*IBP at the molecular level it is crucial to elucidate its ice-binding mechanism through a combination of structural and molecular biology studies. Therefore, we decided to solve the *Efc*IBP structure by means of X-ray crystallography.

*Efc*IBP crystals diffracted to atomic resolution (up to 0.84 Å), and the *Efc*IBP structure was solved by molecular replacement with the crystal structure of the IBP from the antarctic bacteria *Colwellia sp.* (PDB-code 3WP9; DALI Z-score of 32.3, residue identity of 38%) as a search model [95]. The overall structure of *Efc*IBP consists of a right-handed β-helix with a triangular cross-section formed by three faces made by parallel β-sheets, and by an additional single 5-turn α-helix, aligned along the axis of the β-helix. The first face of the β-helix (9 β-strands) is screened from the solvent region by the long α-helix and by the N-terminal region. This protein surface is, therefore, not suited for the interaction with ice crystals. The second face (8 β-strands) is flat and regular, while the third (8 β-strands) is only partly flat, with two β-strands which markedly diverge towards the exterior of the protein body. The latter two faces are fully exposed to the solvent region and, therefore, potentially suited for the interaction with ice crystals.

Overall, the CASP12 results indicate that right-handed β-helix can be predicted extremely well. All β-strands of the three faces of the *Efc*IBP structure are correctly positioned as well as the 5-turn α-helix, aligned along the β-helix axis. It should be noted, however, that the β-strand located immediately after the α-helix is correctly placed within the β-helix fold, but is shifted of two residues in sequence, meaning that the loop before this β-strand is two-residue longer and the loop after is two-residue shorter. The top ten ranked models (CASP GDT_TS score >89.0) are characterized by an RMSD of ~1.4 Å for the core of the protein (181 Cα pairs over 207 residues), devoid of the first 9 N-terminal residues. The structure of this terminal segment is not predicted correctly partly because this region is shorter in the homologous proteins used as templates, partly because its conformation might be selected by crystal contacts and, therefore, difficult to predict.

CASP12 is also able to model correctly a deletion at the top of the right-handed β-helix, where in homologous proteins is present a small cap subdomain (about 12 residues). In this region, however, the Gly-Pro-Pro sequence at the closure of the deletion does not superimpose well with the corresponding *Efc*IBP crystal structure.

Interestingly, the overall quality of the CASP12 prediction does not seem to improve significantly when multiple protein templates are used for modeling instead of a single template. This is probably due to the high structural conservation and rigidity of the β-helix scaffold which is reproduced similarly in all protein templates.

11. The TRXL1 domain of *Chaetomium thermophilum* UGGT (CASP: T0892; PDB: 5MU1, 5MZO, 5N2J and 5NV4) – provided by Pietro Roversi, Alessandro T. Caputo, Johan C. Hill and Nicole Zitzmann.

One of the last unsolved mysteries of the eukaryotic endoplasmic reticulum glycoprotein folding quality control (ERQC) machinery is its single checkpoint enzyme, the ER UDP-glucose glycoprotein glucosyltransferase (UGGT). Once monoglucosylated by this enzyme, glycoproteins are retained in the ER bound to the lectins calnexin and/or calreticulin and the associated chaperones and foldases that assist their folding[96]. The mechanism by which UGGT recognizes and glucosylates a large variety of misfolded glycoprotein substrates remains unknown.

The N-terminal ~1200 residues of UGGT harbor the enzyme's misfold sensing activity [97,98]. The lack of any obvious sequence homology of this portion of UGGT with proteins of known fold led to the creation of a UGGT-specific protein fold family (Pfam family PF06427) which gathers all known eukaryotic UGGT N-terminal sequences. The most recent secondary structure and domain boundary predictions for UGGT detected three thioredoxin-like (TRXL) domains in this region[99,100]. The canonical TRXL fold (Pfam family PF13848) comprises a thioredoxin fold (a four-stranded beta sheet sandwiched between three alpha helices, TRX=βαβ−αββα Pfam family PF00085, red in Figure 10), modified by the insertion of a 4-helix subdomain (TRXL=βαβ−αααα−αββα blue in Figure 10)[101,102].

To aid our understanding of UGGT structure and function, we determined four distinct crystal structures of *Chaetomium thermophilum* UGGT, aka *Ct*UGGT. An unexpected structural feature of the UGGT molecule is the unusual subdomain structure of the first thioredoxin-like domain (TRXL1), encoded by residues 43-216 in *Ct*UGGT. The published sequence–based secondary structure predictions in this region was rather accurate, with most helices and sheets correctly predicted from sequence – but the UGGT TRXL1 domain boundaries were not well predicted [101,102].

Indeed, the UGGT TRXL1 domain folds with sequential pairing of a helical subdomain with a thioredoxin subdomain (blue and red in Figure 10), while all other known TRXL domains present a helical subdomain as an insertion within the thioredoxin subdomain (see for example in Figure 10B the closest structural homologue of *Ct*UGGT TRXL1, *Staphylococcus aureus* DsbA, PDB ID 3BD2). The *Ct*UGGT crystal structures also reveal that the *Ct*UGGT TRXL1 domain harbors a disulfide bridge between Cys138 and Cys150 (represented as spheres in Figure 10A).

We submitted the *Ct*UGGT TRXL1 sequence to CASP12 (target T0892) in order to test prediction methods for their ability to model i) its non-canonical subdomain structure, in which an N-terminal α–helical subdomain is followed by a C-terminal thioredoxin subdomain and ii) the presence of a disulfide bridge between *Ct*UGGT TRXL1 C138 and C150.

We compare here the top 10 CASP12 T0892 models (as ranked by the GDT_TS score on the CASP12 results server) to the coordinates of the TRXL1 domain in the 2.8 Å *Ct*UGGT crystal structure (PDB ID 5NV4), residues 43-216. The overall $RMSD_{C\alpha}$ across the ensemble of the top ten T0892 models is 10.7 Å over 174 C_αs[103]. All these CASP12 T0892 models predict an N-terminal 4-helix subdomain followed by a C-terminal subdomain which resembles to various degrees a TRX fold. None of the top T0892 CASP12 models predicts the *Ct*UGGT TRXL1 C138-C150 disulfide bond.

If one restricts the analysis to the *Ct*UGGT TRXL1 N-terminal helical subdomain (residues 43-110) and the first α-helix (residues 111-126) of the C-terminal thioredoxin subdomain, the top ten T0892 models align rather well with each other and with the crystal structure. The overall $RMSD_{C\alpha}$ for the ten structures over these 84 C_αs is 1.7 Å. The major differences between the CASP12 T0892 models in the 43-126 portion arise at the hinge

(*Ct*UGGT residues 108-111, denoted by a black star in Figure 10C) between the helical subdomain and the first α-helix of the thioredoxin subdomain. The two top-ranked CASP12 models (T0892TS011_1 and T0892TS011_2, green and cyan in Figures 10C-D) show a different hinge region from the rest. As a result of these differences, in the same top-ranking two models, the relative angle between the N-terminal helical subdomain and the first helix of the thioredoxin subdomain also differs from the crystal structure and the rest of the T0892 CASP12 ensemble of models. The *Ct*UGGT 111-126 α-helix is marked by a dotted circle in Figure 10C.

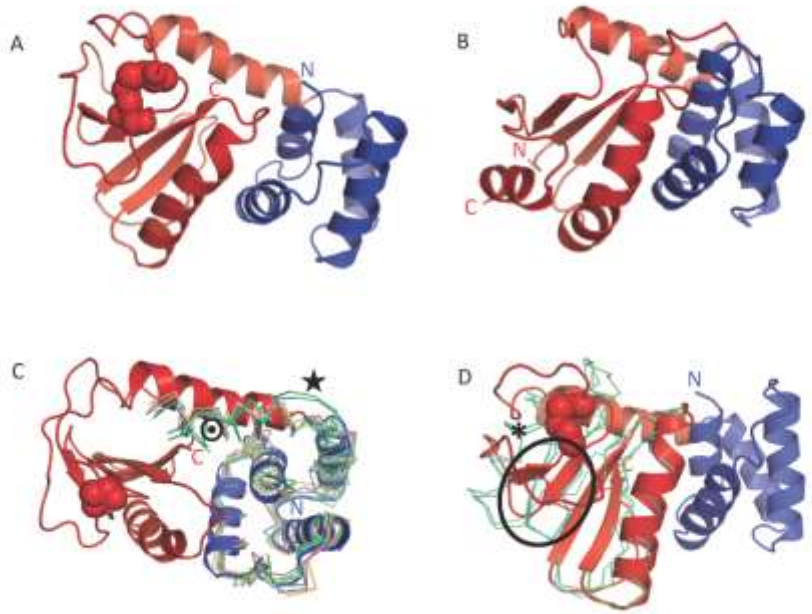

Figure 10.

In the C-terminal thioredoxin subdomain (residues 111-216), the top ten CASP12 T0892 models align poorly with each other and with the crystal structure of the target. The overall $RMSD_{C\alpha}$ for the ten models over these 84 C_αs is 9.5 Å[103]. Only the two top ranking CASP12 T0892 models (T0892TS011_1 and T0892TS011_2, green and cyan in Figures 10C-D) correctly contain a 4-stranded beta-sheet at the center of the TRXL1 thioredoxin subdomain. Even restricting attention to these two models only, across residues 127-216 the $RMSD_{C\alpha}$ between the models and the crystal structure is still as high as 6.5 Å over 90 C_αs[103] (see Figure 10D). In particular, the first two β-strands of the thioredoxin subdomain β-sheet in the models do not superimpose well on the same β-strands in the crystal structure (circled in Figure 10D). Moreover, in both models, the stretch of sequence 151-164 – which immediately follows those strands - is wrongly predicted to fold as an α-helix (marked by an asterisk in Figure 10D) which is not present in the crystal structure.

Overall, none of the models predict the *Ct*UGGT TRXL1 C138-C150 disulfide bond, and the 128-181 region between the first TRX helix, and the third TRX strand is not well defined in any of the models. On the other hand, the best CASP12 T0892 models are successful in predicting the structure of the N-terminal 4-helix subdomain, and the two top-scoring ones also manage to correctly predict that the domain is a linear fusion of an N-terminal 4-helix subdomain and a C-terminal subdomain of TRX fold. In summary, as far as this target was concerned, the CASP12 predictors did well, but did not put us out of our job just yet.

12. Structural characterization of the third cohesin from *Ruminococcus flavefaciens* scaffoldin protein, ScaB (*Rf*CohScaB3) complexed with a group 1a

dockerin (*Rf*Doc1a) (CASP: T0921/T0922; PDB: 5AOZ (*Rf*CohScaB3), 5M2O (*Rf*CohScaB3-Doc1a complex) – provided by Pedro Bule, Ana Luisa Carvalho, Carlos M.G.A. Fontes and Shabir Najmudin.

The plant cell wall represents a major untapped global source of carbon and energy. It is especially important for herbivores, as they are able to utilize this energy source thanks to the presence of cellulolytic bacteria in their gastrointestinal tract. *Ruminococcus flavefaciens*, a Gram-positive Firmicutes, is a major symbiont in the rumen. It possesses multi-enzyme complexes termed cellulosomes, which comprise a range of cellulases and hemicellulases that degrade the structural polysaccharides in a highly efficient and concerted way. The assembly of cellulosomes occurs via highly ordered protein–protein interactions between cohesins (Cohs), which are located in a macromolecular scaffold (scaffoldin), and dockerins (Docs), which are found in the enzymes or on the scaffoldins themselves [104,105]. Strain FD-1 of *R. flavefaciens* produces one of the most intricate and potentially versatile cellulosomes described to date. The *R. flavefaciens* FD-1 genome encodes 223 dockerin-bearing proteins, with the majority of them being carbohydrate-active enzymes [106]. In this highly elaborate cellulosome, scaffoldin B (ScaB) acts as the backbone to which other components attach. It comprises 9 cohesins of 2 distinct types. Cohesins 1 to 4 are similar to the 2 cohesins on ScaA, while cohesins 5 to 9 are closer to the ones found in ScaB of *R. flavefaciens* strain 17. It also has a dockerin with an X-module that binds to the cohesin on ScaE attached to the bacterial cell wall. The ScaA dockerin binds to the second type of ScaB cohesins allowing more carbohydrate active modules to bind to the complex. ScaC acts as an adaptor that binds to both ScaA and the first type ScaB cohesins, thus serving to increase the repertoire of proteins that can be in the complex. In *Clostridium* species studied so far, enzyme-borne Docs

interact with their cognate Cohs through a dual-binding mode [105]. They can bind in either of two orientations resulting in two different Coh-Doc conformations, related by a ~180° rotation. This dual-binding mode results from the characteristic internal symmetry of the Doc sequence and is believed to add flexibility to the cellulosomal macromolecular organization. Recent studies have shown that groups 3 and 6 *R. flavefaciens* Docs display a single-binding mode for their target Cohs [107]. Group 1 Docs also do not seem to possess the internal sequence symmetry required to support the dual-binding mode. Thus, it would be interesting to see if modelling studies would be able to predict the correct binding mode between various types of Coh-Doc complexes and if they could predict which amino acid residues act as molecular specificity determinants.

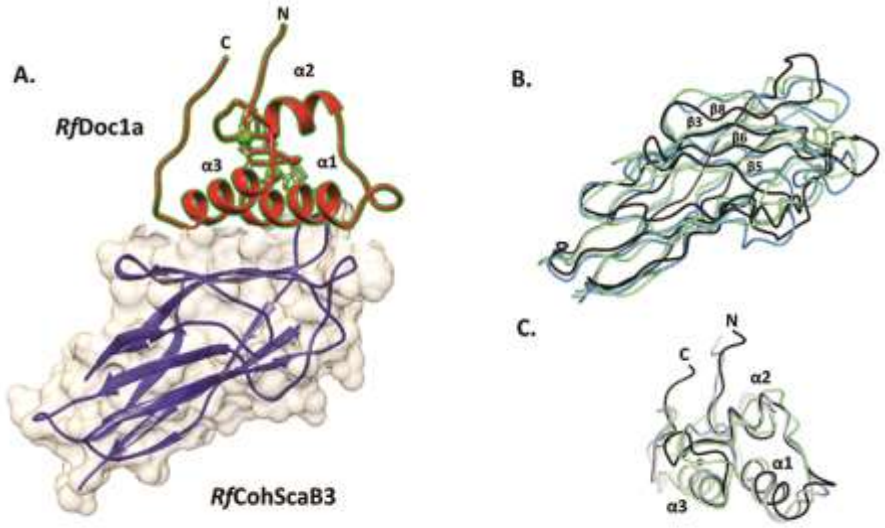

Figure 11.

X-ray crystal structures of the third *R. flavefaciens* cohesin from ScaB (*Rf*CohScaB3) and group 1 Doc (*Rf*Doc1a) in complex with *Rf*CohScaB3 (Figure 11a) were recently solved,

and characterized by comprehensive biochemical analyses [108]. *Rf*ScaBCoh3 forms an elongated nine-stranded β-sandwich in a classical jelly-roll topology. One face of the molecule is formed by anti-parallel β-strands 8, 3, 6, 5, and the other by β-strands 9, 1, 2, 7, 4, with the jelly roll completed by β-strands 1and 9 aligning parallel to each other. The flat surface formed by β-strands 8, 3, 6, 5 constitutes the Doc-interacting surface. Despite the very low sequence similarity, the overall *Rf*ScaBCoh3 structure is similar to other enzyme-borne Doc binding Cohs, with major differences in the Doc binding interface. The loops between β-strands 6 and 7, and 8 and 9 of *Rf*CohScaB3 are slightly raised to form a rim at one edge of the flat surface. *Rf*Doc1a structure in complex with *Rf*CohScaB3 comprises two α-helices (helix-1 and -3) arranged in antiparallel orientation connected by two loops either side of a seven-residue α-helix (helix-2). The overall tertiary structure of *Rf*Doc1a is also very similar to other enzyme-borne Docs and contains two Ca^{2+} ions coordinated by several amino-acid residues, similar to the canonical EF-hand loop motif described in all other Docs [105]. The whole of helix-1 makes predominantly hydrophobic interactions with the Coh, while helix-3 interacts mainly through its C-terminus. Ile-39 and Val-43 on helix-1 of the *Rf*Doc1a and Ala-38 and Leu-79 on the binding platform of *Rf*CohScaB3 were shown to be the key specificity determinants by substituting these residues in ITC experiments. Thus, the X-ray crystal structure of *R. flavefaciens* group 1 Doc (*Rf*Doc1a) in complex with a ScaB (*Rf*CohScaB3) together with comprehensive biochemical analyses suggest that integration of a large repertoire of enzymes into the *R. flavefaciens* cellulosome operates through a single-binding mode unlike in the simpler *Clostridia* cellulosomes [108].

How do these experimental observations match with the modelling studies of CASP12? Predictions for both the *Rf*CohScaB3 and *Rf*Doc1a were very successful, with 147 models for the former and 143 out of 186 for the latter having GDT_TS scores greater than

50. The top model for each target and a slightly poorer model scoring ~10 GDT_TS below the top model were chosen for comparative purposes. For *Rf*CohScaB3, these were models T0921TS220 from the GOAL group (GDT_TS of 70.65) and T0921TS452 from the Zhou-Sparks-X group (GDT_TS of 60.69). Superpositions of these models using SSM onto the X-ray structure gave r.m.s.d. of 2.11 Å for 127 Cα atoms and 2.37 Å for 120 Cα atoms, respectively (Figure 11b). It can be seen that though the core structure matches really well, there are major differences in the β6-7 and β8-9 loops and in the β8 strand on the dockerin binding interface. Ala 38 is generally in the correct position, but there is considerable variation in the Leu 79 position. For *Rf*Doc1a, we chose T0922TS005 from the Baker-Rosetta group (the top scorer with GDT_TS of 83.78) and T0922TS077 from the Falcon_Topo group (GDT_TS of 73.65). Superpositions of these models using SSM onto the X-ray structure gave r.m.s.d. of 1.36 Å for 69 Cα atoms and 1.63 Å for 63 Cα atoms, respectively (Figure 11c). Generally, the α-helices 1 and 3 are well modelled and consequently so are the key specificity residues, like Leu 39 and Val 43, with differences mainly in the loop regions and N- and C-termini. With the modelling of the *Rf*CohScaB3-Doc1a heterocomplex, it was a different story, with only three models out of 325 (TS203_4 from the Seok group, TS188_1 from the Chuo_U group and TS208_3 from the SVMQA group) correctly modeling half or more of the intermolecular surface contacts compared to the crystal structure. One reason for this could be incorrect modelling of the loops in the binding surface of the cohesins. In these three predicted complexes the cohesins have similar or less prominent loops between β-strands 6 & 7, and 8 & 9 compared to the crystal structure (cf. Figure 11b), thus avoiding steric clashes when complexing with the cognate dockerin models in the single-binding mode.

Acknowledgements

CASP experiment and open access fees for this manuscript are supported by the US National Institute of General Medical Sciences (NIGMS/NIH), grant number GM100482.

T0859: Grant sponsor: the Latvian Council of Sciences, grant number: 12.094; Grant sponsor: the European Regional Development Fund, grant number: 2010/0314/2DP/2.1.1.1.0/10/APIA/VIAA/052); Grant sponsor: Biostruct-X and the Latvian-French cooperation program Osmosis, grant number: 7869.

T0884/T0885: Grant sponsor: National Institutes of Health, grant number: GM102318 (CWG, CSH & subcontract to AJ); Grant sponsor: National Institutes of Health, grant number: GM094585 (to AJ); Grant sponsor: National Institutes of Health, grant number: GM115586 (to AJ); Grant sponsor: U. S. Department of Energy, Office of Biological and Environmental Research, contract number: DE-AC02-06CH11357 (to AJ)

T0889: Initial funding for structure determination was from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement No. NMP3-SL-2008-213487. Thanks to Harm Otten and Jens-Christian N. Poulsen for their contributions to structure determination of BjSDH.

T0948: Grant sponsor: National Institutes of Health (NIH), grant number: R01GM102810 (to OH and JM).

T0877: Grant sponsor: Israel Science Foundation (ISF), grant number 682/16 to RD.

T0892: ATC and JCH were funded by Wellcome Trust 4-year Studentships 097300/Z/11/Z and 106272/Z/14/Z, respectively; NZ is a Fellow of Merton College, Oxford.

T0909: Grant sponsor: Spanish Ministry of Economy, Industry and Competitiveness, grant number BFU2014-53425-P (to MJvR).

T0921/T0922: Grant sponsor: Fundação para a Ciência e a Tecnologia (Lisbon, Portugal), grant numbers PTDC/BIA-MIC/5947/2014 and RECI/BBB-BEP/0124/2012, and SFRH/BD/86821/2012 to PB.

References

1. 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.
2. Kryshtafovych A, Moult J, Bales P, Bazan JF, Biasini M, Burgin A, Chen C, Cochran FV, Craig TK, Das R, Fass D, Garcia-Doval C, Herzberg O, Lorimer D, Luecke H, Ma X, Nelson DC, van Raaij MJ, Rohwer F, Segall A, Seguritan V, Zeth K, Schwede T. 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. Proteins 2014;82 Suppl 2:26-42.
3. Kryshtafovych A, Moult J, Basle A, Burgin A, Craig TK, Edwards RA, Fass D, Hartmann MD, Korycinski M, Lewis RJ, Lorimer D, Lupas AN, Newman J, Peat TS, Piepenbrink KH, Prahlad J, van Raaij MJ, Rohwer F, Segall AM, Seguritan V, Sundberg EJ, Singh AK, Wilson MA, Schwede T. Some of the most interesting CASP11 targets through the eyes of their authors. Proteins 2016;84 Suppl 1:34-50.
4. Duan Q, Zhou M, Zhu L, Zhu G. Flagella and bacterial pathogenicity. J Basic Microbiol 2013;53(1):1-8.
5. Arora SK, Ritchings BW, Almira EC, Lory S, Ramphal R. The Pseudomonas aeruginosa flagellar cap protein, FliD, is responsible for mucin adhesion. Infect Immun 1998;66(3):1000-1007.
6. Berg HC. The rotary motor of bacterial flagella. Annu Rev Biochem 2003;72:19-54.
7. Yonekura K, Maki S, Morgan DG, DeRosier DJ, Vonderviszt F, Imada K, Namba K. The bacterial flagellar cap as the rotary promoter of flagellin self-assembly. Science 2000;290(5499):2148-2152.
8. Kim JS, Chang JH, Chung SI, Yum JS. Molecular cloning and characterization of the Helicobacter pylori fliD gene, an essential factor in flagellar structure and motility. J Bacteriol 1999;181(22):6969-6976.
9. Maki-Yonekura S, Yonekura K, Namba K. Domain movements of HAP2 in the cap-filament complex formation and growth process of the bacterial flagellum. Proc Natl Acad Sci U S A 2003;100(26):15528-15533.
10. Yonekura K, Maki-Yonekura S, Namba K. Complete atomic model of the bacterial flagellar filament by electron cryomicroscopy. Nature 2003;424(6949):643-650.
11. Postel S, Deredge D, Bonsor DA, Yu X, Diederichs K, Helmsing S, Vromen A, Friedler A, Hust M, Egelman EH, Beckett D, Wintrode PL, Sundberg EJ. Bacterial flagellar capping proteins adopt diverse oligomeric states. Elife 2016;5.
12. Galkin VE, Yu X, Bielnicki J, Heuser J, Ewing CP, Guerry P, Egelman EH. Divergence of quaternary structures among bacterial flagellar filaments. Science 2008;320(5874):382-385.

13. Song WS, Cho SY, Hong HJ, Park SC, Yoon SI. Self-Oligomerizing Structure of the Flagellar Cap Protein FliD and Its Implication in Filament Assembly. J Mol Biol 2017;429(6):847-857.
14. Hepatitis B vaccines: WHO position paper--recommendations. Vaccine 2010;28(3):589-590.
15. Jennings GT, Bachmann MF. The coming of age of virus-like particle vaccines. Biol Chem 2008;389(5):521-536.
16. Bachmann MF, Rohrer UH, Kundig TM, Burki K, Hengartner H, Zinkernagel RM. The influence of antigen organization on B cell responsiveness. Science 1993;262(5138):1448-1451.
17. Pumpens P, Renhofa R, Dishlers A, Kozlovska T, Ose V, Pushko P, Tars K, Grens E, Bachmann MF. The True Story and Advantages of RNA Phage Capsids as Nanotools. Intervirology 2016;59(2):74-110.
18. Koning RI, Gomez-Blanco J, Akopjana I, Vargas J, Kazaks A, Tars K, Carazo JM, Koster AJ. Asymmetric cryo-EM reconstruction of phage MS2 reveals genome structure in situ. Nat Commun 2016;7:12524.
19. Valegard K, Liljas L, Fridborg K, Unge T. The three-dimensional structure of the bacterial virus MS2. Nature 1990;345(6270):36-41.
20. Golmohammadi R, Fridborg K, Bundule M, Valegard K, Liljas L. The crystal structure of bacteriophage Q beta at 3.5 A resolution. Structure 1996;4(5):543-554.
21. Tars K, Bundule M, Fridborg K, Liljas L. The crystal structure of bacteriophage GA and a comparison of bacteriophages belonging to the major groups of Escherichia coli leviviruses. J Mol Biol 1997;271(5):759-773.
22. Tars K, Fridborg K, Bundule M, Liljas L. The three-dimensional structure of bacteriophage PP7 from Pseudomonas aeruginosa at 3.7-A resolution. Virology 2000;272(2):331-337.
23. Persson M, Tars K, Liljas L. PRR1 coat protein binding to its RNA translational operator. Acta Crystallogr D Biol Crystallogr 2013;69(Pt 3):367-372.
24. Plevka P, Kazaks A, Voronkova T, Kotelovica S, Dishlers A, Liljas L, Tars K. The structure of bacteriophage phiCb5 reveals a role of the RNA genome and metal ions in particle stability and assembly. J Mol Biol 2009;391(3):635-647.
25. Tissot AC, Renhofa R, Schmitz N, Cielens I, Meijerink E, Ose V, Jennings GT, Saudan P, Pumpens P, Bachmann MF. Versatile virus-like particle carrier for epitope based vaccines. PLoS One 2010;5(3):e9809.
26. Shishovs M, Rumnieks J, Diebolder C, Jaudzems K, Andreas LB, Stanek J, Kazaks A, Kotelovica S, Akopjana I, Pintacuda G, Koning RI, Tars K. Structure of AP205 Coat Protein Reveals Circular Permutation in ssRNA Bacteriophages. J Mol Biol 2016;428(21):4267-4279.
27. Ruhe ZC, Low DA, Hayes CS. Bacterial contact-dependent growth inhibition. Trends Microbiol 2013;21(5):230-237.
28. Willett JL, Ruhe ZC, Goulding CW, Low DA, Hayes CS. Contact-Dependent Growth Inhibition (CDI) and CdiB/CdiA Two-Partner Secretion Proteins. J Mol Biol 2015;427(23):3754-3765.
29. Aoki SK, Malinverni JC, Jacoby K, Thomas B, Pamma R, Trinh BN, Remers S, Webb J, Braaten BA, Silhavy TJ, Low DA. Contact-dependent growth inhibition requires the essential outer membrane protein BamA (YaeT) as the receptor and the inner membrane transport protein AcrB. Mol Microbiol 2008;70(2):323-340.
30. Ruhe ZC, Nguyen JY, Xiong J, Koskiniemi S, Beck CM, Perkins BR, Low DA, Hayes CS. CdiA Effectors Use Modular Receptor-Binding Domains To Recognize Target Bacteria. MBio 2017;8(2).
31. Ruhe ZC, Wallace AB, Low DA, Hayes CS. Receptor polymorphism restricts contact-dependent growth inhibition to members of the same species. MBio 2013;4(4).
32. Morse RP, Nikolakakis KC, Willett JL, Gerrick E, Low DA, Hayes CS, Goulding CW. Structural basis of toxicity and immunity in contact-dependent growth inhibition (CDI) systems. Proc Natl Acad Sci U S A 2012;109(52):21480-21485.

33. Aoki SK, Diner EJ, de Roodenbeke CT, Burgess BR, Poole SJ, Braaten BA, Jones AM, Webb JS, Hayes CS, Cotter PA, Low DA. A widespread family of polymorphic contact-dependent toxin delivery systems in bacteria. Nature 2010;468(7322):439-442.
34. Nikolakakis K, Amber S, Wilbur JS, Diner EJ, Aoki SK, Poole SJ, Tuanyok A, Keim PS, Peacock S, Hayes CS, Low DA. The toxin/immunity network of Burkholderia pseudomallei contact-dependent growth inhibition (CDI) systems. Mol Microbiol 2012;84(3):516-529.
35. Aoki SK, Webb JS, Braaten BA, Low DA. Contact-dependent growth inhibition causes reversible metabolic downregulation in Escherichia coli. J Bacteriol 2009;191(6):1777-1786.
36. Jamet A, Jousset AB, Euphrasie D, Mukorako P, Boucharlat A, Ducousso A, Charbit A, Nassif X. A new family of secreted toxins in pathogenic Neisseria species. PLoS Pathog 2015;11(1):e1004592.
37. Zhang D, de Souza RF, Anantharaman V, Iyer LM, Aravind L. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. Biol Direct 2012;7:18.
38. Zhang D, Iyer LM, Aravind L. A novel immunity system for bacterial nucleic acid degrading toxins and its recruitment in various eukaryotic and DNA viral systems. Nucleic Acids Res 2011;39(11):4532-4552.
39. Carr S, Walker D, James R, Kleanthous C, Hemmings AM. Inhibition of a ribosome-inactivating ribonuclease: the crystal structure of the cytotoxic domain of colicin E3 in complex with its immunity protein. Structure 2000;8(9):949-960.
40. Ng CL, Lang K, Meenan NA, Sharma A, Kelley AC, Kleanthous C, Ramakrishnan V. Structural basis for 16S ribosomal RNA cleavage by the cytotoxic domain of colicin E3. Nat Struct Mol Biol 2010;17(10):1241-1246.
41. Jiang Y, Pogliano J, Helinski DR, Konieczny I. ParE toxin encoded by the broad-host-range plasmid RK2 is an inhibitor of Escherichia coli gyrase. Mol Microbiol 2002;44(4):971-979.
42. Pedersen K, Zavialov AV, Pavlov MY, Elf J, Gerdes K, Ehrenberg M. The bacterial toxin RelE displays codon-specific cleavage of mRNAs in the ribosomal A site. Cell 2003;112(1):131-140.
43. Masaki H, Ogawa T. The modes of action of colicins E5 and D, and related cytotoxic tRNases. Biochimie 2002;84(5-6):433-438.
44. Li Z, Gao Y, Nakanishi H, Gao X, Cai L. Biosynthesis of rare hexoses using microorganisms and related enzymes. Beilstein J Org Chem 2013;9:2434-2445.
45. Wang Z, Etienne M, Quiles F, Kohring GW, Walcarius A. Durable cofactor immobilization in sol-gel bio-composite thin films for reagentless biosensors and bioreactors using dehydrogenases. Biosens Bioelectron 2012;32(1):111-117.
46. Gauer S, Wang Z, Otten H, Etienne M, Bjerrum MJ, Lo Leggio L, Walcarius A, Giffhorn F, Kohring GW. An L-glucitol oxidizing dehydrogenase from Bradyrhizobium japonicum USDA 110 for production of D-sorbose with enzymatic or electrochemical cofactor regeneration. Appl Microbiol Biotechnol 2014;98(7):3023-3032.
47. Kant R, Tabassum R, Gupta BD. A highly sensitive and distinctly selective D-sorbitol biosensor using SDH enzyme entrapped Ta2O5 nanoflowers assembly coupled with fiber optic SPR. Sensor Actuat B-Chem 2017;242:810-817.
48. Fredslund F, Otten H, Gemperlein S, Poulsen JC, Carius Y, Kohring GW, Lo Leggio L. Structural characterization of the thermostable Bradyrhizobium japonicumD-sorbitol dehydrogenase. Acta Crystallogr F Struct Biol Commun 2016;72(Pt 11):846-852.
49. Karplus PA, Diederichs K. Linking crystallographic model and data quality. Science 2012;336(6084):1030-1033.
50. Javidpour P, Pereira JH, Goh EB, McAndrew RP, Ma SM, Friedland GD, Keasling JD, Chhabra SR, Adams PD, Beller HR. Biochemical and structural studies of NADH-dependent FabG used to increase the bacterial production of fatty acids under anaerobic conditions. Appl Environ Microbiol 2014;80(2):497-505.

51. MacKenzie AK, Kershaw NJ, Hernandez H, Robinson CV, Schofield CJ, Andersson I. Clavulanic acid dehydrogenase: structural and biochemical analysis of the final step in the biosynthesis of the beta-lactamase inhibitor clavulanic acid. Biochemistry 2007;46(6):1523-1533.
52. Philippsen A, Schirmer T, Stein MA, Giffhorn F, Stetefeld J. Structure of zinc-independent sorbitol dehydrogenase from Rhodobacter sphaeroides at 2.4 A resolution. Acta Crystallogr D Biol Crystallogr 2005;61(Pt 4):374-379.
53. Tamura M, Tanaka S, Fujii T, Aoki A, Komiyama H, Ezawa K, Sumiyama K, Sagai T, Shiroishi T. Members of a novel gene family, Gsdm, are expressed exclusively in the epithelium of the skin and gastrointestinal tract in a highly tissue-specific manner. Genomics 2007;89(5):618-629.
54. Carl-McGrath S, Schneider-Stock R, Ebert M, Rocken C. Differential expression and localisation of gasdermin-like (GSDML), a novel member of the cancer-associated GSDMDC protein family, in neoplastic and non-neoplastic gastric, hepatic, and colon tissues. Pathology 2008;40(1):13-24.
55. Hergueta-Redondo M, Sarrio D, Molina-Crespo A, Vicario R, Bernado-Morales C, Martinez L, Rojo-Sebastian A, Serra-Musach J, Mota A, Martinez-Ramirez A, Castilla MA, Gonzalez-Martin A, Pernas S, Cano A, Cortes J, Nuciforo PG, Peg V, Palacios J, Pujana MA, Arribas J, Moreno-Bueno G. Gasdermin B expression predicts poor clinical outcome in HER2-positive breast cancer. Oncotarget 2016;7(35):56295-56308.
56. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SA, Wong KC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WO. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007;448(7152):470-473.
57. Saleh NM, Raj SM, Smyth DJ, Wallace C, Howson JM, Bell L, Walker NM, Stevens HE, Todd JA. Genetic association analyses of atopic illness and proinflammatory cytokine genes with type 1 diabetes. Diabetes Metab Res Rev 2011;27(8):838-843.
58. Pal LR, Moult J. Genetic Basis of Common Human Disease: Insight into the Role of Missense SNPs from Genome-Wide Association Studies. J Mol Biol 2015;427(13):2271-2289.
59. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, Lee JC, Schumm LP, Sharma Y, Anderson CA, Essers J, Mitrovic M, Ning K, Cleynen I, Theatre E, Spain SL, Raychaudhuri S, Goyette P, Wei Z, Abraham C, Achkar JP, Ahmad T, Amininejad L, Ananthakrishnan AN, Andersen V, Andrews JM, Baidoo L, Balschun T, Bampton PA, Bitton A, Boucher G, Brand S, Buning C, Cohain A, Cichon S, D'Amato M, De Jong D, Devaney KL, Dubinsky M, Edwards C, Ellinghaus D, Ferguson LR, Franchimont D, Fransen K, Gearry R, Georges M, Gieger C, Glas J, Haritunians T, Hart A, Hawkey C, Hedl M, Hu X, Karlsen TH, Kupcinskas L, Kugathasan S, Latiano A, Laukens D, Lawrance IC, Lees CW, Louis E, Mahy G, Mansfield J, Morgan AR, Mowat C, Newman W, Palmieri O, Ponsioen CY, Potocnik U, Prescott NJ, Regueiro M, Rotter JI, Russell RK, Sanderson JD, Sans M, Satsangi J, Schreiber S, Simms LA, Sventoraityte J, Targan SR, Taylor KD, Tremelling M, Verspaget HW, De Vos M, Wijmenga C, Wilson DC, Winkelmann J, Xavier RJ, Zeissig S, Zhang B, Zhang CK, Zhao H, International IBDGC, Silverberg MS, Annese V, Hakonarson H, Brant SR, Radford-Smith G, Mathew CG, Rioux JD, Schadt EE, Daly MJ, Franke A, Parkes M, Vermeire S, Barrett JC, Cho JH. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491(7422):119-124.
60. Stahl EA, Raychaudhuri S, Remmers EF, Xie G, Eyre S, Thomson BP, Li Y, Kurreeman FA, Zhernakova A, Hinks A, Guiducci C, Chen R, Alfredsson L, Amos CI, Ardlie KG, Consortium B, Barton A, Bowes J, Brouwer E, Burtt NP, Catanese JJ, Coblyn J, Coenen MJ, Costenbader KH, Criswell LA, Crusius JB, Cui J, de Bakker PI, De Jager PL, Ding B, Emery P, Flynn E, Harrison P, Hocking LJ, Huizinga TW, Kastner DL, Ke X, Lee AT, Liu X, Martin P, Morgan AW, Padyukov L,

Posthumus MD, Radstake TR, Reid DM, Seielstad M, Seldin MF, Shadick NA, Steer S, Tak PP, Thomson W, van der Helm-van Mil AH, van der Horst-Bruinsma IE, van der Schoot CE, van Riel PL, Weinblatt ME, Wilson AG, Wolbink GJ, Wordsworth BP, Consortium Y, Wijmenga C, Karlson EW, Toes RE, de Vries N, Begovich AB, Worthington J, Siminovitch KA, Gregersen PK, Klareskog L, Plenge RM. Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. Nat Genet 2010;42(6):508-514.

61. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491(7422):56-65.
62. Chao KL, Kulakova L, Herzberg O. Gene polymorphism linked to increased asthma and IBD risk alters gasdermin-B structure, a sulfatide and phosphoinositide binding protein. Proc Natl Acad Sci U S A 2017;114(7):E1128-E1137.
63. Ding J, Wang K, Liu W, She Y, Sun Q, Shi J, Sun H, Wang DC, Shao F. Pore-forming activity and structural autoinhibition of the gasdermin family. Nature 2016;535(7610):111-116.
64. Hergueta-Redondo M, Sarrio D, Molina-Crespo A, Megias D, Mota A, Rojo-Sebastian A, Garcia-Sanz P, Morales S, Abril S, Cano A, Peinado H, Moreno-Bueno G. Gasdermin-B promotes invasion and metastasis in breast cancer cells. PLoS One 2014;9(3):e90099.
65. Zong M, Fofana I, Choe H. Human and host species transferrin receptor 1 use by North American arenaviruses. J Virol 2014;88(16):9418-9428.
66. Fulhorst CF, Bowen MD, Ksiazek TG, Rollin PE, Nichol ST, Kosoy MY, Peters CJ. Isolation and characterization of Whitewater Arroyo virus, a novel North American arenavirus. Virology 1996;224(1):114-120.
67. Blokland JA, Vossepoel AM, Bakker AR, Pauwels EK. Automatic assignment of elliptical ROIs: first results in planar scintigrams of the left ventricle. Eur J Nucl Med 1989;15(2):87-92.
68. Frankel AI, Chapman JC, Cook B. The testicular response to hemicastration in the male rat cannot be maintained in vitro. J Endocrinol 1989;121(1):43-48.
69. O'Neill MA, Ishii T, Albersheim P, Darvill AG. Rhamnogalacturonan II: structure and function of a borate cross-linked cell wall pectic polysaccharide. Annu Rev Plant Biol 2004;55:109-139.
70. Matsunaga T, Ishii T, Matsumoto S, Higuchi M, Darvill A, Albersheim P, O'Neill MA. Occurrence of the primary cell wall polysaccharide rhamnogalacturonan II in pteridophytes, lycophytes, and bryophytes. Implications for the evolution of vascular plants. Plant Physiol 2004;134(1):339-351.
71. Ndeh D, Rogowski A, Cartmell A, Luis AS, Basle A, Gray J, Venditto I, Briggs J, Zhang X, Labourel A, Terrapon N, Buffetto F, Nepogodiev S, Xiao Y, Field RA, Zhu Y, O'Neill MA, Urbanowicz BR, York WS, Davies GJ, Abbott DW, Ralet MC, Martens EC, Henrissat B, Gilbert HJ. Complex pectin metabolism by gut bacteria reveals novel catalytic functions. Nature 2017;544(7648):65-70.
72. Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, Abbott DW, Henrissat B, Gilbert HJ, Bolam DN, Gordon JI. Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. PLoS Biol 2011;9(12):e1001221.
73. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res 2014;42(Database issue):D490-495.
74. Bekesi A, Pukancsik M, Muha V, Zagyva I, Leveles I, Hunyadi-Gulyas E, Klement E, Medzihradszky KF, Kele Z, Erdei A, Felfoldi F, Konya E, Vertessy BG. A novel fruitfly protein under developmental control degrades uracil-DNA. Biochem Biophys Res Commun 2007;355(3):643-648.
75. Pukancsik M, Bekesi A, Klement E, Hunyadi-Gulyas E, Medzihradszky KF, Kosinski J, Bujnicki JM, Alfonso C, Rivas G, Vertessy BG. Physiological truncation and domain organization of a novel uracil-DNA-degrading factor. Febs J 2010;277(5):1245-1259.

76. Hartmann MD, Boichenko I, Coles M, Zanini F, Lupas AN, Hernandez Alvarez B. Thalidomide mimics uridine binding to an aromatic cage in cereblon. J Struct Biol 2014;188(3):225-232.
77. Hartmann MD, Boichenko I, Coles M, Lupas AN, Hernandez Alvarez B. Structural dynamics of the cereblon ligand binding domain. PLoS ONE 2015;10(5):e0128342.
78. 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(11):e50649.
79. Singh AK, Menendez-Conejero R, San Martin C, van Raaij MJ. Crystal structure of the fibre head domain of the Atadenovirus Snake Adenovirus 1. PLoS ONE 2014;9(12):e114373.
80. Gorman JJ, Wallis TP, Whelan DA, Shaw J, Both GW. LH3, a "homologue" of the mastadenoviral E1B 55-kDa protein is a structural protein of atadenoviruses. Virology 2005;342(1):159-166.
81. Pantelic RS, Lockett LJ, Rothnagel R, Hankamer B, Both GW. Cryoelectron microscopy map of Atadenovirus reveals cross-genus structural differences from human adenovirus. J Virol 2008;82(15):7346-7356.
82. Bradley P, Cowen L, Menke M, King J, Berger B. BETAWRAP: successful prediction of parallel beta -helices from primary sequence reveals an association with many microbial pathogens. Proc Natl Acad Sci U S A 2001;98(26):14819-14824.
83. Mayans O, Scott M, Connerton I, Gravesen T, Benen J, Visser J, Pickersgill R, Jenkins J. Two crystal structures of pectin lyase A from Aspergillus reveal a pH driven conformational change and striking divergence in the substrate-binding clefts of pectin and pectate lyases. Structure 1997;5(5):677-689.
84. Garnham CP, Campbell RL, Walker VK, Davies PL. Novel dimeric beta-helical model of an ice nucleation protein with bridged active sites. BMC Struct Biol 2011;11:36.
85. Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. Nucleic Acids Res 2010;38(Web Server issue):W545-549.
86. 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(3):375-386.
87. Muller JJ, Barbirz S, Heinle K, Freiberg A, Seckler R, Heinemann U. An intersubunit active site between supercoiled parallel beta helices in the trimeric tailspike endorhamnosidase of Shigella flexneri Phage Sf6. Structure 2008;16(5):766-775.
88. Steinbacher S, Miller S, Baxa U, Budisa N, Weintraub A, Seckler R, Huber R. Phage P22 tailspike protein: crystal structure of the head-binding domain at 2.3 A, fully refined structure of the endorhamnosidase at 1.56 A resolution, and the molecular basis of O-antigen recognition and cleavage. J Mol Biol 1997;267(4):865-880.
89. Leiman PG, Molineux IJ. Evolution of a new enzyme activity from the same motif fold. Mol Microbiol 2008;69(2):287-290.
90. Bar Dolev M, Braslavsky I, Davies PL. Ice-Binding Proteins and Their Function. Annu Rev Biochem 2016;85:515-542.
91. Raymond JA, DeVries AL. Adsorption inhibition as a mechanism of freezing resistance in polar fishes. Proc Natl Acad Sci U S A 1977;74(6):2589-2593.
92. Yu SO, Brown A, Middleton AJ, Tomczak MM, Walker VK, Davies PL. Ice restructuring inhibition activities in antifreeze proteins with distinct differences in thermal hysteresis. Cryobiology 2010;61(3):327-334.
93. Cid FP, Rilling JI, Graether SP, Bravo LA, Mora Mde L, Jorquera MA. Properties and biotechnological applications of ice-binding proteins in bacteria. FEMS Microbiol Lett 2016;363(11).
94. Mangiagalli M, Bar-Dolev M, Tedesco P, Natalello A, Kaleda A, Brocca S, de Pascale D, Pucciarelli S, Miceli C, Bravslavsky I, Lotti M. Cryo-protective effect of an ice-binding protein derived from Antarctic bacteria. Febs J 2017;284(1):163-177.

95. Hanada Y, Nishimiya Y, Miura A, Tsuda S, Kondo H. Hyperactive antifreeze protein from an Antarctic sea ice bacterium Colwellia sp. has a compound ice-binding site without repetitive sequences. Febs J 2014;281(16):3576-3590.
96. Michalak M, Corbett EF, Mesaeli N, Nakamura K, Opas M. Calreticulin: one protein, one gene, many functions. Biochem J 1999;344 Pt 2:281-292.
97. Arnold SM, Kaufman RJ. The noncatalytic portion of human UDP-glucose: glycoprotein glucosyltransferase I confers UDP-glucose binding and transferase function to the catalytic domain. J Biol Chem 2003;278(44):43320-43328.
98. Guerin M, Parodi AJ. The UDP-glucose:glycoprotein glucosyltransferase is organized in at least two tightly bound domains from yeast to mammals. J Biol Chem 2003;278(23):20540-20546.
99. Zhu T, Satoh T, Kato K. Structural insight into substrate recognition by the endoplasmic reticulum folding-sensor enzyme: crystal structure of third thioredoxin-like domain of UDP-glucose:glycoprotein glucosyltransferase. Sci Rep 2014;4:7322.
100. Calles-Garcia D, Yang M, Soya N, Melero R, Menade M, Ito Y, Vargas J, Lukacs GL, Kollman JM, Kozlov G, Gehring K. Single-particle electron microscopy structure of UDP-glucose:glycoprotein glucosyltransferase suggests a selectivity mechanism for misfolded proteins. J Biol Chem 2017.
101. Ferrari DM, Soling HD. The protein disulphide-isomerase family: unravelling a string of folds. Biochem J 1999;339 (Pt 1):1-10.
102. Kozlov G, Maattanen P, Thomas DY, Gehring K. A structural overview of the PDI family of proteins. Febs J 2010;277(19):3924-3936.
103. Theobald DL, Steindel PA. Optimal simultaneous superpositioning of multiple structures with missing data. Bioinformatics 2012;28(15):1972-1979.
104. Bayer EA, Belaich JP, Shoham Y, Lamed R. The cellulosomes: multienzyme machines for degradation of plant cell wall polysaccharides. Annu Rev Microbiol 2004;58:521-554.
105. Fontes CM, Gilbert HJ. Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. Annu Rev Biochem 2010;79:655-681.
106. Dassa B, Borovok I, Ruimy-Israeli V, Lamed R, Flint HJ, Duncan SH, Henrissat B, Coutinho P, Morrison M, Mosoni P, Yeoman CJ, White BA, Bayer EA. Rumen cellulosomics: divergent fiber-degrading strategies revealed by comparative genome-wide analysis of six ruminococcal strains. PLoS ONE 2014;9(7):e99221.
107. Bule P, Alves VD, Leitao A, Ferreira LM, Bayer EA, Smith SP, Gilbert HJ, Najmudin S, Fontes CM. Single Binding Mode Integration of Hemicellulose-degrading Enzymes via Adaptor Scaffoldins in Ruminococcus flavefaciens Cellulosome. J Biol Chem 2016;291(52):26658-26669.
108. Bule P, Alves VD, Israeli-Ruimy V, Carvalho AL, Ferreira LM, Smith SP, Gilbert HJ, Najmudin S, Bayer EA, Fontes CM. Assembly of Ruminococcus flavefaciens cellulosome revealed by structures of two cohesin-dockerin complexes. Sci Rep 2017;7(1):759.
109. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302(1):205-217.

FIGURE CAPTIONS

Figure 1. (A) Crystal structure of the *Pseudomonas* FliD$_{78-405}$ monomer subunit in which the domain D3 (CASP domain D2, green), domain D2 (CASP domain D1, blue) and the helical region (red), which belongs to domain D1 (not evaluated in CASP), are indicated. **(B)** Side view (top panel) and top view (bottom panel) showing cartoon representations of the hexameric FliD$_{78-405}$ crystal structure. Each monomer subunit is colored distinctly. **(C)** SAXS-generated molecular envelope of the monomeric FliD$_{1-474}$ with the CASP prediction model T0886TS036_1 (cyan). **(D)** Superposition of CASP prediction models T0886TS247_1_D1 (orange) and T0886TS247_1_D2 (orange) with D2 (CASP domain D1, blue) and D3 (CASP domain D2, green) of the FliD$_{78-405}$ monomer crystal structure. **(E)** Superposition of CASP prediction model T0886TS247_1 (orange) with the FliD$_{78-405}$ monomer crystal structure (domain coloring as in Panel A). **(F)** Superposition of CASP prediction model T0886TS247_1 (orange) with the *E. coli* FliD$_{43-416}$ crystal structure 5H5V (magenta). **(G)** Superposition of CASP prediction models T0886TS247_1 (orange), T0886TS011_1 (cyan), T0886TS064_1_1 (light blue), T0886TS411_1 (yellow) with the FliD$_{78-405}$ monomer crystal structure (domain coloring as in Panel A). **(H)** Superposition of CASP prediction models T0886TS247_1-D2 (orange), T0886TS064_1_1-D2 (light blue), T0886TS011_1-D2 (cyan), T0886TS411_1-D2 (yellow), T0886TS456_1-D2 (dark grey), T0886TS173_1_1-D2 (red) with D3 of the FliD$_{78-405}$ monomer crystal structure (green).

Figure 2. Structural features of bacteriophage AP205 coat protein. Coat protein in AP205 and related phages, such as MS2, builds very stable dimers, both monomers are shown in different shades of grey (panels **A** and **B**). Notice the close proximity of N- (blue) and C- (red) termini in dimers. 90 dimers further assemble into VLPs (panels C and D). In MS2, AB loop (green) is the most exposed structure on the surface of VLPs. Compared to MS2, in

AP205 the first beta strand (yellow) is shifted to the C-terminus, although it remains in the same position in 3D. As a result, in AP205, C-and N- termini are the most exposed features on VLPs. In panel E, crystal structure of AP205 monomer (green) is superimposed with the modelled structure (blue and red). The overall fold of model is approximately correct, except that it lacks C-terminal beta strand. Residues 1-39 (blue) are correctly placed in respect to the sequence, corresponding to the first four beta strands. For the rest of model (red) residues are placed incorrectly according to the sequence and out-of-register errors occur. Notice also that position of N-terminus is relatively well predicted, while C-terminus is in a very different position.

Figure 3. The CdiA-CT/CdiI^{Ctai} complex. (**A**) Experimental structure with the most conserved residues and their interactions shown in stick representation. The CdiA-CT^{Ctai} toxin domain is shown in teal and the CdiI^{Ctai} immunity protein in pink. Hydrogen bonds are depicted as red broken lines. Superposition of CdiA-CT^{Ctai} with (**B**) the closest PDB homolog, inorganic triphosphatase (coral, PDB:3TYP), (**C**) with ParE toxin from *E. coli* (yellow, PDB:3KXE) and (**D**) with T0884TS183_1-D1 (purple) and refined TR884TS118_1 model (blue). β1 from CdiI^{Ctai} is shown for reference. (**E**) Superposition of CdiI^{Ctai} with T0885TS005_2-D1 (cyan) and refined TR885TS247_1-D1 model (blue).

Figure 4. (**a)** Products of reaction catalyzed by BjSDH with D-glucitol and L-glucitol as substrates; (**b)** Structure based sequence alignment of region around loop 193-203 covering the active site of BjSDH. Sequences of GatDH, RsSDH and top 5 DALI hits searching with the BjSDH structure are shown; **c)** BjSDH structure shown as cartoon (gold) and symmetry

related molecule packing against is (grey). Ligands in the structure are shown as sticks, while loop 193-203 in top 5 models from CASP12 are shown as lines; **d)** Continuous β-sheet between two monomers in BjSDH crystal structure, and same region in the RsSDH crystal structure.

Figure 5. (A) Structure-based sequence alignment of the GSDMB (T0948 comprises GSDMB's C-terminal domain) and mouse Gsdma3 C-terminal domains with secondary structure elements shown above or below the respective sequences. Identical and conservatively replaced residues are colored in red and blue. The alignment was performed using the programs Clustal Omega [109] and ESPript 3 (espript.ibcp.fr/Espript/). **(B)** Ribbon diagram of the GSDMB_C fold (PDB 5TIB). The α7–α8 GSDMB loop containing the polymorphism residues is colored in red. **(C)** Superposition of the experimental GSDMB_C structure (colored yellow) and the corresponding Gsdma3 domain that served as a modeling template (blue, 5B5R), **(D)** Superposition of the experimental GSDMB_C structure (colored yellow) and the best GTD_TS CASP12 scored model of group 251 (green). **(E)** Superposition of the polymorphism loop of the experimental structure (colored gray with α' highlighted in orange) with the corresponding loop assessed as the closest (Group 330) based on the position specific criterion (colored cyan with α' highlighted in magenta).

Figure 6. The structure of WWAV-GP1 compared to the top three models. (A): Ribbon diagrams of the WWAV-GP1 colored in rainbow and shown in a putative complex with hTfR1 (surface representation) (PDB ID: 3KAS). **(B)**: A potential charge-repulsion between two negatively charged groups on WWAV and hTfR1 that was identified using this analysis.

(C): Comparison of the top three models from 'MULTICOM-CONSTRUCT', 'MULTICOM-NOVEL', and 'GOAL' (designated S236, S345, and S220, respectively) with WWAV-GP1. **(D)**: A close-up view comparing the loops of WWAV-GP1 that interact with hTfR1 to the top model. Structures were rendered using PyMOL (www.pymol.org).

Figure 7. Panel A: **Cartoon representation of BT1002** (5MPQ, chain A) aligned with T0912TS349_1 using align in pymol (sequence alignment followed by structural superposition with c-alpha atoms only). Residues are colored by a RMSD gradient (dark blue is a good alignment and red are higher deviations). Residues not used are colored grey. The domain are labelled D1 to D3. **Panel B**: Binding pocket surface representation. The predicted model (T0912TS303_1) surface is represented in solid dark grey and the PDB model surface in yellow mesh. The putative catalytic residues in the predicted model are colored magenta and red in the PDB model.

Figure 8. The crystal structure of *Aa*UDE(87-277) in comparison to the best DALI matches and CASP predictions. **(A)** The full crystal structure in cartoon representation. **(B)** The crystal structure (red) superimposed with the best DALI matches for the N-terminal (PDB: 3UN9; DALI Z-score 7.5) and the C-terminal domain (PDB: 3TD7; DALI Z-score 10.1). **(C)** The two best CASP predictions for the N-terminal domain (D1), models T0890TS236_1 (MULTICOM-CONSTRUCT) and T0890TS486_1 (TASSER), yielded a GDT_TS of 68.0 and 67.7 for D1 and of 30.0 and 31.8 for the whole structure. **(D)** The best CASP predictions for the C-terminal domain (D2). T0890TS250_1 (Seok-server) yielded a GDT_TS of 74.8 for D2 and 44.7 for the whole structure. T0890TS119_1 represents the three

almost identical models T0890TS119_1 (HHPred0), T0890TS349_1 (HHPred1) and T0890TS313_1 (HHGG), which yielded a GDT_TS of 69.8, 69.8 and 70.5 for D2 and of 40.8, 40.8 and 41.0 for the whole structure. T0890TS464_1 (tsspred2) yielded a GDT_TS of 59.2 for D2 and 33.4 for the whole structure.

Figure 9. Crystal structure of SnAdV-1 LH3 in comparison with CASP12 models. (**A**) Superimposition between a monomer of the experimentally determined structure (cyan) and best predicted model (magenta). The missing loop at the rung 7 is indicated with the black arrow. (**B**) Predicted monomer was superposed in a trimeric structure obtained by X-ray crystallography. Models were colored as in the A and black arrow indicates the missing loop in the model. (**C**) and (**D**).

Figure **10. The TRXL1 domain of *Ct*UGGT. A.** In blue, the *Ct*UGGT TRXL1 N-terminal α-helical subdomain (residues 43-110). In red, the TRXL1 thioredoxin subdomain (residues 111-216). The disulphide bridge C138-C150 is represented as spheres. **B.** The structure of the closest structural homologue to *Ct*UGGT TRXL1, *Staphylococcus aureus* DsbA, with the α-helical insertion subdomain (residues 63-129) in blue and the thioredoxin subdomain (residues 14-62 and 130-177) in red. In (**A)** and (**B)** N- and C-termini are denoted by the letters "N" and "C", respectively. (**C).** The superposition of the top ten CASP12 T0892 models, overlayed on the *Ct*UGGT TRXL1 crystal structure in the region of the N-terminal helical subdomain and the first helix of the thioredoxin subdomain. The *Ct*UGGT TRXL1 crystal structure is colored and represented as in panel A. The top ten CASP12 T0892 models are in ribbon representation and colored as follows: T0892TS011_1:green; T0892TS011_2:

cyan; T0892TS017_1: magenta; T0892TS017_2: yellow; T0892TS017_5: grey; T0892TS411_2; T0892TS017_3: salmon pink; T0892TS079_5: violet; T0892TS479_3: steel blue; T0892TS320_4: orange. A black star marks the hinge between the helical subdomain and the thioredoxin subdomain. A dotted circle marks the first helix in the thioredoxin subdomain. **(D).** The superposition of the top two CASP12 T0892 models (T0892TS011_1 and T0892TS011_2, in green and cyan respectively, in ribbon representation), overlayed on the *Ct*UGGT TRXL1 crystal structure in the region of the C-terminal thioredoxin subdomain, without its first α-helix. The *Ct*UGGT TRXL1 crystal structure is colored and represented as in panel A. The wrongly predicted first two strands of the thioredoxin subdomain are circled, and an asterisk marks the incorrectly predicted α-helix for the stretch of residues 151-164 of *Ct*UGGT TRXL1.

Figure 11. Structure of the *Rf*CohScaB3-Doc1a complex. **(A)** Structure of *Rf*CohScaB3-Doc1a complex with the dockerin in red and the cohesin in blue. The dockerin N- and C-terminus and the α-helices are labeled, and a transparent gray molecular surface of the cohesin is shown. **(B)** Superposition of CASP12 prediction models T0921TS220_2_D1 (light blue) and T0921TS166_1_D1 (light green) with *Rf*CohScaB3 crystal structure (black). **(C)** Superposition of CASP12 prediction models T0922TS005_3_D1 (light blue) and T0922TS077_4_D1 (light green) with the *Rf*Doc1a crystal structure (black). Ca^{2+} ions are depicted as pink spheres.