Computational Modeling of Biomolecular Interactions

Marc F. Lensink

December 6^{th}, 2017

Document for the obtention of the HDR
"Habilitation à diriger des recherches"
"Accreditation to supervise research"
University of Lille, France

Discipline:	Life sciences and health
Promotor:	Ralf Blossey
Jury:	Tony Lefebvre Marc Baaden Savvas Savvides Alessandra Carbone Isabelle Landrieu
Laboratory:	Institute for Structural and Functional Glycobiology CNRS UMR8576 UGSF
Doctoral school:	Biology/Health
Order:	42471

CONTENTS

1 *Curriculum Vitae* **3**
1.1 Personal Information . . . 3
1.2 Address . . . 3
1.3 Responsibilities . . . 3
1.4 Grants obtained (as coordinator or partner) . . . 3
1.5 Academic Degree . . . 3
1.6 Career positions . . . 4
1.7 Undergraduate, graduate and post-graduate supervision . . . 5

2 Scientific Research **6**
2.1 Background . . . 6
2.2 Research activities . . . 6
2.3 Citation analysis . . . 6
2.4 Selected publications . . . 6

3 Research activities **10**
3.1 The CAPRI protein docking experiment . . . 10
3.2 Protein-lipid interaction studies . . . 13
3.3 Protein and protein-small molecule modelling . . . 15

4 Selected publications **19**
Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment . . . 20
Identification of specific lipid-binding sites in integral membrane proteins . . . 46
On the ability of molecular dynamics simulation and continuum electrostatics to treat interfacial water molecules in protein-protein complexes . . . 54
Probing the conformation of FhaC with small-angle neutron scattering and molecular modeling 64
Virulence regulation with Venus flytrap domains: Structure and function of the periplasmic moiety of the sensor-kinase BvgS . . . 76

5 Outlook **97**

6 Acknowledgements **98**

7 Jury **98**

8 Cover images **98**

1. *Curriculum Vitae*

1.1 Personal Information

Name:	Marc Ferdinand Lensink, Ph.D.
Nationality:	Dutch
Date of Birth:	20-02-1972
Place of Birth:	Amsterdam
Marital Status:	Married to Emmanuelle Rangeard, French, born in Dole (Jura)
Children:	Téa Elena Marie, born in Ghent, Belgium, July 12th, 2004 Luca Maëli Julian, born in Uccle, Belgium, October 3rd, 2007
URL:	`http://ugsf.univ-lille1.fr/spip.php?article108`
e-Mail:	`marc.lensink@univ-lille1.fr`
Home address:	Rue de l'Aigle 8, B-1480 Tubize, Belgium

1.2 Address

CNRS UMR8576 UGSF – Institute for Structural and Functional Glycobiology
Parc de la Haute Borne
50 Avenue de Halley
F-59658 Villeneuve d'Ascq, France
+33 3 62 53 17 28 (tel)
+33 3 62 53 17 01 (fax)

1.3 Responsibilities

- Team leader UGSF: "Computational and Molecular Systems Biology"
- Member of the CAPRI Management Committee
- Member of the CAPRI Assessment Team
- Member of the "bilille" bureau de gestion
- Member of the "bilille" cellule animation scientifique
- Member of the GENCI national resource expert panel
- Scientific reviewer for (list not complete): Biophys J, Biochim Biophys Acta, Biochemistry, Mol Member Biol, BMC Bioinformatics, BMC Struct Biol, J Chem Phys B, Bioinformatics, J Mol Biol, PLoS Comput Biol, PLoS One, Nucl Acids Res, Proteins, J Biol Macromol, Scientific Reports, ...

1.4 Grants obtained (as coordinator or partner)

- ANR 2017-2020; HICARE "Heterovalent inhibitors of carbohydrate processing enzymes"
- H2020 MSCA-IF-2016 2017-2019; FimH-Mech "The molecular mechanisms of E. coli pathogenicity"
- H2020 ITN 2016-2019; GLYCOVAX "A training network for the rational design of the next generation of well-defined glycoconjugate vaccines"
- Région Nord-Pas-de-Calais, Accueil Jeune Chercheur 2014-2016; "Targeting the Ets-1 oncoprotein"
- ANR 2014-2017; MECA VENUS "Mécanismes moléculaires de la transduction par BvgS, un modèle de la famille de récepteurs kinases bactériens à domaines Vénus Flytrap"
- ANR 2013-2016; FSCF "Fluctuations in structured Coulomb fluids"
- Max Planck Gesellschaft & CNRS Post-doctoral program for systems biology 2010-2011

1.5 Academic Degree

Ph.D. (March 2002)
"Non-adiabatic Proton Transfer in Biomolecular Systems"
Mathematics and Natural Sciences
University of Groningen
The Netherlands

1.6 Career positions

Jan 2015 – present	Team Leader Institute for Structural and Functional Glycobiology, CNRS UMR 8576, France Computational Molecular Systems Biology
Oct 2011 – Dec 2014	CNRS CR1 Research Associate Interdisciplinary Research Institute, CNRS USR 3078, France Computational and Theoretical Molecular Biology
Sep 2010 – Sep 2011	Post-doctoral Fellow Interdisciplinary Research Institute, CNRS USR 3078, France Biological Nanosystems, and Max Planck Institute for Informatics, Saarbrücken, Germany
Project funding:	*"Post-doctoral Programme for Systems Biology"* *CNRS & Max Planck Gesellschaft*
Mar 2006 – Aug 2010	Post-doctoral Fellow Dept. of Chemistry, Université Libre de Bruxelles, Belgium Center for Structural Biology and Bioinformatics Structure and Function of Biological Membranes, and Genome and Network Bioinformatics
Project funding:	*"Development of Anti-Allergy Vaccination based on Immunosomes"* *Région Wallonne of Belgium, Waleo programme contract 515993*
and	*"BioSapiens: A Network of Excellence for Genome Annotation"* *European Union, contract LSHG-CT-2003-503265*
Sep 2004 – Mar 2006	Post-doctoral Fellow Dept. of Molecular Biology, Université Libre de Bruxelles, Belgium Center for Structural Biology and Bioinformatics Structure of Biological Macromolecules and Bioinformatics
Project funding:	*"GeneFun: In-silico Prediction of Gene Function"* *European Union, contract LSHG-CT-2004-503567*
Sep 2002 – Sep 2004	Post-doctoral Fellow Dept. of Biochemistry, University of Ghent, Belgium Dept. of Lipoprotein Chemistry
Project funding:	*"Messenger Proteins: Mechanisms of Action and Biological Significance"* *European Union, contract NPRN-CT-2001-0242*
Oct 2000 – Sep 2002	Post-doctoral Fellow Dept. of Biochemistry & Biocenter Oulu, University of Oulu, Finland Biocomputing Group
Aug 2000 – Oct 2000	UNIX System Manager Dept. of Biophysical Chemistry, University of Groningen, The Netherlands Molecular Dynamics Group
May 1996 – May 2000	Ph.D. Mathematics & Natural Sciences Dept. of Biophysical Chemistry, University of Groningen, The Netherlands Molecular Dynamics Group
Project funding:	*"Non-adiabatic Proton Transfer in Biomolecular Systems"* *NWO, Dutch Scientific Research Institute*
	of which half a year was spent in
Jul 1996 – Nov 1996	Dept. of Chemistry, Boston University, USA Theoretical Chemistry Group
Jan 1995 – May 1996	Junior Research Fellow Centre Européen de Calcul Atomique et Moléculaire (CECAM) Ecole Normale Supérieure de Lyon (ENS-Lyon), France
Sep 1994	M.Sc. Mathematics & Natural Sciences *"Density Matrix Evolution"* Dept. of Biophysical Chemistry Molecular Dynamics Group University of Groningen, The Netherlands

1.7 Undergraduate, graduate and post-graduate supervision

- *Undergraduate supervision*
- M.Sc. thesis, Gerrit Groenhof, “Signal Transduction in the Photoactive Yellow Protein”. As a Ph.D. student I was responsible for the everyday guidance. The project led to two publications and was later continued by Gerrit for his Ph.D. thesis. More importantly, it laid the foundations for the incorporation of quantum-mechanics into the GROMACS molecular dynamics simulations package. Gerrit later became group leader at the Max Planck Institute for Biophysical Chemistry in Göttingen and is now assistent professor in Jyväskylä, Finland.
- M.Sc. thesis, Theo Rispens, “Cosolvent Effects on Neutral Hydrolysis”. As a Ph.D. student I was responsible for the everyday guidance. We analyzed the occurrence of key spatial organizations in molecular dynamics simulations, reproducing the experimentally observed linear dependency on cosolvent concentration as well as estimate the rate constant of the reaction to an order of magnitude correct. Theo has done a Ph.D. thesis and a few years of post-doctoral research and then left the academic world.
- *Graduate supervision*
- Ph.D. thesis, Fernanda Sirota-Leite, “Role of the amino acids sequences in domain swapping of the B1 domain of protein G by computation analysis”. As a post-doc I was responsible for the day-to-day guidance in Brussels, while the promotor had moved to Toronto. Fernanda later became post-doctoral researcher at the Bioinformatics Institute of Singapore.
- Ph.D. thesis, Raul Méndez, “Critical Assessment of Predicted Interactions at Atomic Resolution”. As a post-doc I was responsible for the day-to-day guidance in Brussels, while the promotor had moved to Toronto. This project laid the foundation for my later involvement in the CAPRI protein docking experiment. Raul later become a post-doc at the Universidad Autónoma Madrid in Spain and is currently at The University of North Carolina at Chapel Hill, USA.
- Ph.D. thesis, Benoît Dessailly, “Binding sites in protein structures: characterisation and relation with destabilising regions”. As a post-doc I was responsible for the day-to-day guidance in Brussels, while the promotor had moved to Toronto. The project dealt with the investigation of stability in protein structures and led to two publications and the development of a database of ligand-binding sites in proteins. Benoît later became post-doctoral researcher at University College London, the obtained a JSPS post-doctoral fellowship in Japan and is currently working in industry in Cambridge, UK.
- Ph.D. thesis, Raghvendra Pratap Singh, “Computer modeling of the application of mechanical forces to biomolecules”. I was involved in the molecular modeling aspects of his thesis. This work later resulted in a publication. Raghvendra went on to become a post-doctoral researcher.
- Marco Miele, “Modelling of GPCR’s in bilayers of mixed content. Marco was a six-month visitor in the group in 2011. He successfully defended his Ph.D. thesis, entitled “Comparative structural analysis of human cytokine membrane receptors and modeling of lipid membrane containing the adiponectin receptor, ADIPOR1, by molecular dynamics simulations” in November, 2011.
- *Post-graduate supervision*
- Patricia Urbina, “Specific binding of phospholipids to phospholipases”. Patricia has spent two years as a post-doc in Brussels on a project that we jointly wrote. Patricia performed docking calculations which led to a publication. She then continued on an other post-doctoral contract in Brussels.
- Jérôme de Ruyck, “Targeting the Ets-1 oncoprotein”. Jérôme obtained a two-year fellowship “Accueil de jeunes chercheurs 2013” on a project that we jointly wrote. We have several joint publications from that period. Afterwards, Jérôme became a post-doctoral researcher in Montpellier and will come back to my group as assistant professor (Maître de Conférences) in 2017.
- Eva-Maria Krammer, “Molecular mechanisms of signal transduction in BvgS”. Eva-Maria came as a post-doc on the ANR MECA VENUS project. After a year as post-doctoral researcher in Brussels, she returned to the group last April, on a Marie Skłodowska-Curie fellowship.

2. SCIENTIFIC RESEARCH

2.1 Background

Whereas both sequence and structure space are being rapidly filled, the sampling of the interactome remains sparse. Protein-protein interaction (PPI) is one of the key processes in cellular functioning. Proteins involved in cellular signaling and regulation transiently bind other proteins to propagate a signal, thereby forming an intricate network of PPIs. These processes are initiated at the membrane surface and a significant portion of proteins in the cell are membrane-associated, accounting for as much as 30% of all genomic sequences. But their lipidic environment plays a role that is to-date poorly understood. In addition, the majority of proteins are glycosylated, with glycosylation playing such diverse roles as in cellular adhesion, immunology, and protein folding and quality control.

2.2 Research activities

My main research activities are aimed towards the understanding at the molecular level of three types of fundamental interactions governing these protein complexes, namely *(i)* protein-lipid interaction, *(ii)* protein-carbohydrate interaction and *(iii)* protein-protein interaction.

My main expertise lies in molecular modeling and dynamics; through pragmatic and collaborative research projects, I develop expertise in modeling of molecular recognition processes.

I am the team leader of the “Computational Molecular Systems Biology” group at the “Institute for Structural and Functional Glycobiology” (UGSF). We develop applicable knowledge and technology in the study of molecular recognition and dynamics, using primarily computational and crystallographic techniques. We work at the interface of biology and physics, applying a combination of multi-scale modeling approaches and structural biology techniques to relevant biological questions. With a natural overlap between the various topics, our main axes of research are:

- Computational modeling and dynamics of biomolecular systems
- Structural biology of protein-carbohydrate interactions
- Protein interaction and regulatory networks
- Statistical physics of biomolecular interactions

I am an active member of the Bililleinformatics platform, with responsibilities in the Management Bureau, in the organization of workshops and training courses, and in providing access to high-performance computing resources.

In addition, I am a member of the CAPRI Management Committee, a key member in the organization of prediction rounds, and the person responsible for the assessment of docking models.

2.3 Citation analysis

ResearcherID:	A-1678-2008
Scopus ID:	8637947800
ORCID:	0000-0003-3957-9470
Total publications:	52
h-index:	18
Average citations per item:	25.2
Sum of times cited*:	1 238
Citing articles*:	977

* *without self-citations*

(Data from Web of Science, September 2017)

2.4 Selected publications

- *“Sites for Dynamic Protein-Carbohydrate Interactions of O- and C-Linked Mannosides on the E. coli FimH Adhesin”*, Touaibia M, Krammer EM, Shiao TC, Yamakawa N, Wang Q, Glinschert A, Papadopoulos A, Mousavifar L, Maes E, Oscarson S, Vergoten G, Lensink MF, Roy R, Bouckaert J. **Molecules** 2017;22:E1101.

- *“Mutation of Tyr137 of the universal Escherichia coli fimbrial adhesin FimH relaxes the tyrosine gate prior to mannose binding”*, Rabbani S, Krammer EM, Roos G, Zalewski A, Preston R, Eid S, Zihlmann P, Prvost M, Lensink MF, Thompson A, Ernst B, Bouckaert J. **IUCrJ** 2017;4:7.

- *"FlexPepDock lessons from CAPRI peptide-protein rounds and suggested new criteria for assessment of model quality and utility"*, Marcu O, Dodson EJ, Alam N, Sperber M, Kozakov D, Lensink MF, Schueler-Furman O. **Proteins** 2017;85:445.
- *"On the ability of molecular dynamics simulation and continuum electrostatics to treat interfacial water molecules in protein-protein complexes"*, Copie G, Cleri F, Blossey R, Lensink MF. **Sci Rep** 2016;6:38259.
- *"Modeling protein-protein and protein-peptide complexes: CAPRI 6th edition"*, Lensink MF, Velankar S, Wodak SJ. **Proteins** 2017;85:359.
- *"Introducing a Clustering Step in a Consensus Approach for the Scoring of Protein-Protein Docking Models"*, Chermak E, De Donato R, Lensink MF, Petta A, Serra L, Scarano V, Cavallo L, Oliva R. **PLoS One** 2016;11:e0166460.
- *"Molecular docking as a popular tool in drug design, an in silico travel"*, De Ruyck J, Brysbaert G, Blossey R, Lensink MF. **Adv Appl Bioinform Chem** 2016;9:1.
- *"CLUB-MARTINI: Selecting Favourable Interactions amongst Available Candidates, a Coarse-Grained Simulation Approach to Scoring Docking Decoys"*, Hou Q, Lensink MF, Heringa J, Feenstra KA. **PLoS One** 2016;11:e0155251.
- *"Structures of C-mannosylated anti-adhesives bound to the type 1 fimbrial FimH adhesin"*, De Ruyck J, Lensink MF, Bouckaert J. **IUCrJ** 2016;3:163.
- *"Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment"*, Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman-Duhovny D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath S, Elber R, Grudinin S, Popov P, Neveu E, Lee H, Baek M, Park S, Heo L, Rie Lee G, Seok C, Qin S, Zhou HX, Ritchie DW, Maigret B, Devignes MD, Ghoorah A, Torchala M, Chaleil RA, Bates PA, Ben-Zeev E, Eisenstein M, Negi SS, Weng Z, Vreven T, Pierce BG, Borrman TM, Yu J, Ochsenbein F, Guerois R, Vangone A, Rodrigues JP, van Zundert G, Nellen M, Xue L, Karaca E, Melquiond AS, Visscher K, Kastritis PL, Bonvin AM, Xu X, Qiu L, Yan C, Li J, Ma Z, Cheng J, Zou X, Shen Y, Peterson LX, Kim HR, Roy A, Han X, Esquivel-Rodriguez J, Kihara D, Yu X, Bruce NJ, Fuller JC, Wade RC, Anishchenko I, Kundrotas PJ, Vakser IA, Imai K, Yamada K, Oda T, Nakamura T, Tomii K, Pallara C, Romero-Durana M, Jimnez-Garca B, Moal IH, Frnandez-Recio J, Joung JY, Kim JY, Joo K, Lee J, Kozakov D, Vajda S, Mottarella S, Hall DR, Beglov D, Mamonov A, Xia B, Bohnuud T, Del Carpio CA, Ichiishi E, Marze N, Kuroda D, Roy Burman SS, Gray JJ, Chermak E, Cavallo L, Oliva R, Tovchigrechko A, Wodak SJ. **Proteins** 2016;84 Suppl 1:323.
- *"The Antiadhesive Strategy in Crohn's Disease: Orally Active Mannosides to Decolonize Pathogenic Escherichia coli from the Gut"*, Alvarez Dorta D, Sivignon A, Chalopin T, Dumych TI, Roos G, Bilyy RO, Deniaud D, Krammer EM, de Ruyck J, Lensink MF, Bouckaert J, Barnich N, Gouin SG. **Chembiochem** 2016;17:936.
- *"Balance between Coiled-Coil Stability and Dynamics Regulates Activity of BvgS Sensor Kinase in Bordetella"*, Lesne E, Krammer EM, Dupre E, Locht C, Lensink MF, Antoine R, Jacob-Dubuisson F. **MBio** 2016;7:e02089.
- *"Signal transduction by BvgS sensor kinase. Binding of modulator nicotinate affects the conformation and dynamics of the entire periplasmic moiety"*, Dupré E, Lesne E, Guérin J, Lensink MF, Verger A, de Ruyck J, Brysbaert G, Vezin H, Locht C, Antoine R, Jacob-Dubuisson F. **J Biol Chem** 2015;290:26473. Erratum in J Biol Chem 2015;290:26473.
- *"Virulence regulation with Venus flytrap domains: structure and function of the periplasmic moiety of the sensor-kinase BvgS"*, Dupré E, Herrou J, Lensink MF, Wintjens R, Vagin A, Lebedev A, Crosson S, Villeret V, Locht C, Antoine R, Jacob-Dubuisson F. **PLoS Pathog** 2015;11:e1004700.
- *"Regulatory motifs on ISWI chromatin remodelers: molecular mechanisms and kinetic proofreading"*, Brysbaert G, Lensink MF, Blossey R. **J Phys Condens Matter** 2015;27:064108.
- *"Membrane-associated proteins and peptides"*, Lensink MF. **Methods Mol Biol** 2015;1215:109.

- *"Score_set: a CAPRI benchmark for scoring protein complexes"*, Lensink MF, Wodak SJ. **Proteins** 2014;82:3163.
- *"Probing the conformation of FhaC with small-angle neutron scattering and molecular modeling"*, Gabel F, Lensink MF, Clantin B, Jacob-Dubuisson F, Villeret V, Ebel C. **Biophys J** 2014;107:185.
- *"The structure of the CD3ζζ transmembrane dimer in lipid bilayers"*, Sharma S, Lensink MF, Juffer AH. **Biochim Biophys Acta** 2014;1838:739.
- *"Blind prediction of interfacial water positions in CAPRI"*, Lensink MF, Moal IH, Bates PA, Kastritis PL, Melquiond AS, Karaca E, Schmitz C, van Dijk M, Bonvin AM, Eisenstein M, Jimnez-Garca B, Grosdidier S, Solernou A, Prez-Cano L, Pallara C, Fernndez-Recio J, Xu J, Muthu P, Praneeth Kilambi K, Gray JJ, Grudinin S, Derevyanko G, Mitchell JC, Wieting J, Kanamori E, Tsuchiya Y, Murakami Y, Sarmiento J, Standley DM, Shirota M, Kinoshita K, Nakamura H, Chavent M, Ritchie DW, Park H, Ko J, Lee H, Seok C, Shen Y, Kozakov D, Vajda S, Kundrotas PJ, Vakser IA, Pierce BG, Hwang H, Vreven T, Weng Z, Buch I, Farkash E, Wolfson HJ, Zacharias M, Qin S, Zhou HX, Huang SY, Zou X, Wojdyla JA, Kleanthous C, Wodak SJ. **Proteins** 2014;82:620.
- *"Docking, scoring, and affinity prediction in CAPRI"*, Lensink MF, Wodak SJ. **Proteins** 2013;81:2082.
- *"On the molecular basis of D-bifunctional protein deficiency type III"*, Mehtälä ML, Lensink MF, Pietikäinen LP, Hiltunen JK, Glumoff T. **PLoS One** 2013;8:e53688.
- *"Oligoarginine vectors for intracellular delivery: role of arginine side-chain orientation in chain length-dependent destabilization of lipid membranes"*, Bouchet AM, Lairion F, Ruysschaert JM, Lensink MF. **Chem Phys Lipids** 2012;165:89.
- *"Unexpected wide substrate specificity of C. perfringens -toxin phospholipase C"*, Urbina P, Collado MI, Alonso A, Goñi FM, Flores-Díaz M, Alape-Girón A, Ruysschaert JM, Lensink MF. **Biochim Biophys Acta** 2011;1808:2618.
- *"Blind predictions of protein interfaces by docking calculations in CAPRI"*, Lensink MF, Wodak SJ. **Proteins** 2010;78:3085.
- *"Docking and scoring protein interactions: CAPRI 2009"*, Lensink MF, Wodak SJ. **Proteins** 2010;78:3073.
- *"Identification of specific lipid-binding sites in integral membrane proteins"*, Lensink MF, Govaerts C, Ruysschaert JM. **J Biol Chem** 2010;285:10519.
- *"Fusogenic activity of cationic lipids and lipid shape distribution"*, Lonez C, Lensink MF, Kleiren E, Vanderwinden JM, Ruysschaert JM, Vandenbranden M. **Cell Mol Life Sci** 2010;67:483.
- *"Cationic lipids activate cellular cascades. Which receptors are involved?"*, Lonez C, Lensink MF, Vandenbranden M, Ruysschaert JM. **Biochim Biophys Acta** 2009;1790:425.
- *"Characterization of the cationic DiC(14)-amidine bilayer by mixed DMPC/DiC(14)-amidine molecular dynamics simulations shows an interdigitated nonlamellar bilayer phase"*, Lensink MF, Lonez C, Ruysschaert JM, Vandenbranden M. **Langmuir** 2009;25:5230.
- *"Homology-based modeling of 3D structures of protein-protein complexes using alignments of modified sequence profiles"*, Kundrotas PJ, Lensink MF, Alexov E. **Int J Biol Macromol** 2008;43:198.
- *"Membrane-associated proteins and peptides"*, Lensink MF. **Methods Mol Biol** 2008;443:161.
- *"Recognition-induced conformational changes in protein-protein docking"*, Lensink MF, Méndez R. **Curr Pharm Biotechnol** 2008;9:77.
- *"LigASite–a database of biologically relevant binding sites in proteins with known apo-structures"*, Dessailly BH, Lensink MF, Orengo CA, Wodak SJ. **Nucleic Acids Res** 2008;36:D667.
- *"Docking and scoring protein complexes: CAPRI 3rd Edition"*, Lensink MF, Méndez R, Wodak SJ. **Proteins** 2007;69:704.

- *"Relating destabilizing regions to known functional sites in proteins"*, Dessailly BH, Lensink MF, Wodak SJ. **BMC Bioinformatics** 2007;8:141.
- *"Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures"*, Méndez R, Leplae R, Lensink MF, Wodak SJ. **Proteins** 2005;60:150.
- *"Penetratin-membrane association: W48/R52/W56 shield the peptide from the aqueous phase"*, Lensink MF, Christiaens B, Vandekerckhove J, Prochiantz A, Rosseneu M. **Biophys J** 2005;88:939.
- *"Phosphorylation by protein kinase CK2 modulates the activity of the ATP binding cassette A1 transporter"*, Roosbeek S, Peelman F, Verhee A, Labeur C, Caster H, Lensink MF, Cirulli C, Grooten J, Cochet C, Vandekerckhove J, Amoresano A, Chimini G, Tavernier J, Rosseneu M. **J Biol Chem** 2004;279:37779.
- *"A conserved arginine plays a role in the catalytic cycle of the protein disulphide isomerases"*, Lappi AK, Lensink MF, Alanen HI, Salo KE, Lobell M, Juffer AH, Ruddock LW. **J Mol Biol** 2004;335:283.
- *"Response of SCP-2L domain of human MFE-2 to ligand removal: binding site closure and burial of peroxisomal targeting signal"*, Lensink MF, Haapalainen AM, Hiltunen JK, Glumoff T, Juffer AH. **J Mol Biol** 2002;323:99.
- *"Signal transduction in the photoactive yellow protein. II. Proton transfer initiates conformational changes"*, Groenhof G, Lensink MF, Berendsen HJ, Mark AE. **Proteins** 2002;48:212.
- *"Signal transduction in the photoactive yellow protein. I. Photon absorption and the isomerization of the chromophore"*, Groenhof G, Lensink MF, Berendsen HJ, Snijders JG, Mark AE. **Proteins** 2002;48:202.

3. RESEARCH ACTIVITIES

A wealth of genomic information has recently become available and will continue to do so in the future. The rapid development of high-throughput structural biology techniques will result in an exponential increase of structural information to feed into (structural) systems biology projects. Whereas both sequence and structure space are being rapidly filled, the sampling of the interactome remains sparse. This is the ideal time to invest in molecular modelling techniques in order to further develop computational techniques that are able to provide a *reliable* description of the recognition between interacting entities.

Protein-protein interaction (PPI) is one of the key processes in cellular functioning. Proteins involved in cellular signalling and regulation transiently bind other proteins to propagate a signal, thereby forming an intricate network of PPIs. These networks are the core component of cellular functioning; by transmitting a signal through these networks, the cell is able to respond to various external and internal stimuli. Complex diseases like autoimmunity, diabetes and cancer result from deficiencies in such signal transduction pathways.

PPI networks and most major processes in the cell are initiated at the membrane surface and a significant portion of proteins in the cell are membrane-associated. In fact, membrane proteins account for about 30% of all genomic sequences. Between their synthesis and delivery to the plasma membrane, membrane proteins encounter a host of lipid environments, in all of which they have to adopt a stable structure, but in some of which the protein is not supposed to show any functional activity.

In addition, the majority of proteins are glycosylated, with glycosylation occuring in the endoplasmic reticulum, but also in the cytoplasm and nucleus. Glycosylation adds significant diversity to the proteome, modifying both structure and dynamics. It serves various functions, going from protein folding and quality control to cellular adhesion; yet no molecular (DNA) template exists for glycosylation.

My research activities are aimed at the development and application of advanced molecular modelling and simulation techniques to study the process of molecular recognition, focusing primarily on three types of fundamental interactions:

- **Protein-protein interaction**,
- **Protein-lipid interaction**,
- **Protein-carbohydrate interaction**.

I have experience in all aspects of molecular recognition, including protein-lipid, protein-carbohydrate and protein-small molecule interaction. I am an expert in molecular dynamics simulation techniques and have extensive expertise in the simulation of protein-lipid systems. In addition, I hold a key role in the community-wide CAPRI protein-protein docking experiment.

The remainder of this document will briefly talk about the various modelling projects that I am or have been involved in and finishes with a small outlook section. It would be impractical to include all relevant publications, so I have decided to list the references and include only a few key ones in their entirety. The interested reader is refered to the list of publications on page 6, or my full publication list, which can be found on-line using any of the on-line ID's listed on page 6.

3.1 The CAPRI protein docking experiment

Computational protein-protein docking is the procedure of producing a three-dimensional structure of the complex, starting from the individual structures of the interacting proteins. The CAPRI protein docking experiment is a community-wide effort aimed at the improvement of computational docking methods. It does so by working in close collaboration with experimental scientists, who make their data available to this community, in the fullest confidence and prior to publication. Participants in CAPRI are asked to predict the three-dimensional structure of a protein complex; these predictions are then assessed in a double-blind procedure against an unpublished and confidential (X-ray) target structure. Present-day realistic scenarios often require a step of homology modeling prior to docking. The project is overseen by a Management Committee; its members are listed in Table I. The commitee makes decisions concerning CAPRI targets, the protocol for evaluating models, the planning of CAPRI assessment meetings, and represents CAPRI at various scientific meetings.

The assessment protocol evaluates the prediction quality on the basis of interface accuracy and ligand positioning. When either of these two achieves a certain threshold quality, the prediction is labeled as as being of "acceptable" quality. Higher threshold values lead to the labels "medium" or "high" quality. Two rmsd-based quantities are thus calculated: the ligand rmsd (L-rms) is the backbone rmsd calculated

Alexandre Bonvin	Utrecht University, The Netherlands
Marc Lensink	CNRS, France
Michael Sternberg	Imperial College London, UK
Sandor Vajda	Boston University, USA
Ilya Vakser	University of Kansas, USA
Sameer Velankar	European Bioinformatics Institute, UK
Zhiping Weng	University of Massachusetts Medical School, USA
Shoshana Wodak	University of Toronto & Hospital for Sick Children, Canada

Table I: *Composition of the CAPRI Management Committee, as of April, 2013.*

over the set of common residues after structural superposition of the receptor entities, and the interface rmsd (I-rms) is the backbone rmsd over the common set of interface residues after structural superpositioning of these residues. An important third quantity whereby models are assessed is the fraction of native contacts that is correctly predicted, or $f(nat)$. These quantities together define the quality of a prediction, following Table II.

	Score	$f(nat)$	L-rms		I-rms
$***$	High	≥ 0.5	≤ 1.0	OR	≤ 1.0
$**$	Medium	≥ 0.3	$< 1.0 - 5.0]$	OR	$< 1.0 - 2.0]$
$*$	Acceptable	≥ 0.1	$< 5.0 - 10.0]$	OR	$< 2.0 - 4.0]$
	Incorrect	< 0.1	> 10.0	AND	> 4.0

Table II: *Summary of the requirement for a model to be placed in any of the four CAPRI model quality categories.*

The experiment is illustrated in Figure 1, which depicts the prediction space of a given target. Fig. 1A shows in dark and light blue cartoon representation the structure of target receptor in its unbound and bound form, resp. Every single prediction is overlapped on this structure and the geometric center of the ligand molecule depicted as a coloured sphere. For this target, most predicted models were incorrect, and the colour of the sphere is yellow. A cluster of near-acceptable structures is located near the center of the target ligand (shown as transparent surface), with predictions that are of acceptable quality or better coloured other than yellow (here: shades of blue).

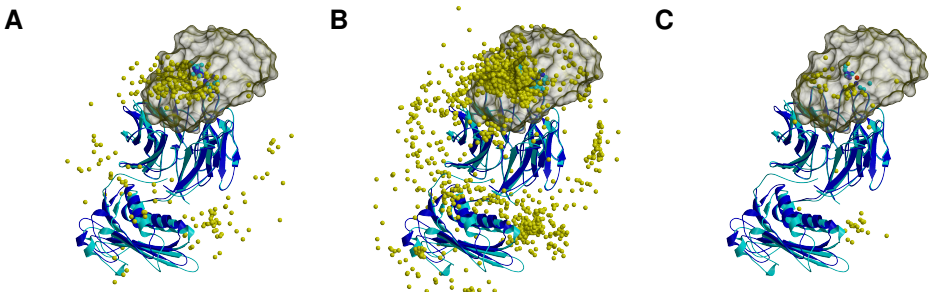


Figure 1: *Representation of the prediction space for CAPRI Target 26. Bound (cyan) and unbound (blue) receptor in cartoon, ligand in surface representation. The dots indicate the ligand geometric centre, colored according to prediction quality.* ***A****: Predictor set,* ***B****: Uploader set,* ***C****: Scorer set. Scorers select decoys from the Uploader set, which is an extension of the Predictor set, enriched in acceptable solutions.*

In addition to the submission of 10 models for evaluation, CAPRI participants are also invited to upload a set of 100 models. Once submissions are closed, the uploaded models are shuffled and made available to the participating groups as part of the CAPRI scoring experiment. The predictor groups, then

called “scorers”, are invited to evaluate the uploaded models using the scoring function of their choice and make a second submission. This submission is evaluated using the exact same criteria as were used for the docking experiment. The prediction space of the scoring experiment is shown in Fig. 1.

CAPRI targets follow the demand of the experimentalists and as such represent a wide variety of biological processes. The experiment now includes, besides docking and scoring of protein-protein and protein-peptide interaction, the prediction of multi-component assemblies, protein-nucleic acid and protein-polysaccharide binding, and the prediction of binding affinities and the positions of interfacial water molecules. The CAPRI project gives a fair assessment of the performance of present-day protein docking methods for the more difficult targets one may encounter and as such it provides an upper limit of what can be expected from docking algorithms.

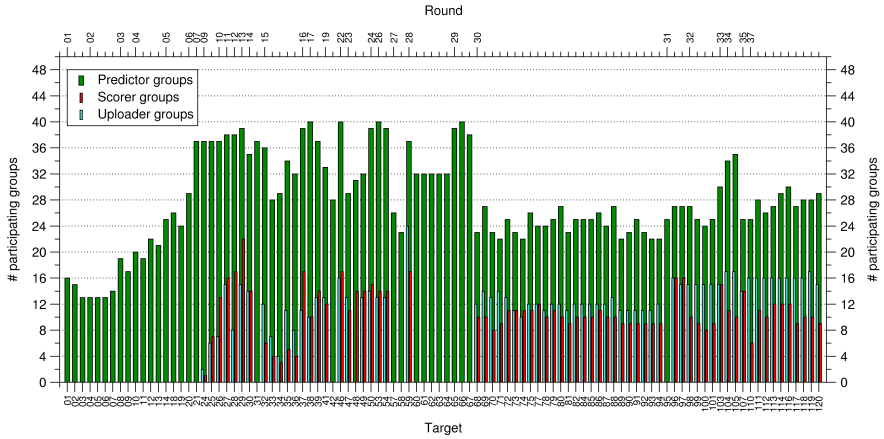


Figure 2: *Number of participating research teams in CAPRI since its inception. Predictor groups (blue bars) participate in the docking experiment, scorer groups (red bars) participate in the scoring experiment, selecting models made available by uploader groups (cyan bars), who form a subset of the predictor groups.*

The number of research teams participating in the experiment lies around 100, of which a subset of 25 to 45 participate in any given target, often involving several members of the same team (See Fig. 2. Participants in CAPRI meet at a regulat basis at Evaluation Meetings; six of these meetings have been held so far, the most recent ones in April 2013 (Utrecht, The Netherlands) and April 2016 (Tel Aviv, Israel). The latest two CASP installments (CASP11 2014 and CASP12 2016) have seen a joint CASP/CAPRI prediction round and such is planned for CASP13 2018 as well. The CASP and CAPRI Evaluation Meetings are accompanied by dedicated issues of the journal Proteins. CAPRI and CASP have catalyzed the development of protein structure prediction algorithms and together define the standard in protein-protein docking and protein structure prediction.

Publications

- Marcu *et al*, **Proteins** 2017;85:445
- Lensink and Wodak, **Proteins** 2017;85:359
- Janin *et al*, in: **Reviews in computational chemistry Vol. 28**, Wiley
- Chermak *et al*, **PLoS One** 2016;11:e0166460
- Hou *et al*, **PLoS One** 2016;11:e0155251
- Lensink *et al*, **Proteins** 2016;84 Suppl 1; 323
- Lensink and Wodak, **Proteins** 2014;82:3163
- Lensink *et al*, **Proteins** 2014;82:620
- Lensink and Wodak, **Proteins** 2013;81:2082
- Lensink and Wodak, **Proteins** 2010;78:3085
- Lensink and Wodak, **Proteins** 2010;78:3073
- Lensink and Méndez, **Curr Pharm Biotechnol** 2008;9:77
- Lensink *et al*, **Proteins** 2007;69:704
- Méndez *et al*, **Proteins** 2005;60:150

3.2 Protein-lipid interaction studies

Identification of specific lipid-binding sites in integral membrane proteins

Protein-lipid interactions are increasingly recognized as central to the structure and function of membrane proteins. However, with the exception of simplified models, specific protein-lipid interactions are particularly difficult to highlight experimentally.

In this study, we have used molecular modelling and dynamics to characterize a specific protein-lipid interaction between lactose permease (LacY) and phosphatidylethanolamine, or PE. LacY is a paradigm for the major facilitator superfamily (MFS) that represents as much as 25% of all membrane transport proteins, with over 15,000 sequence members identified to date.

By carefully positioning the protein in a selection of lipid matrices and performing molecular dynamics simulations, we identified a specific protein-lipid interaction in LacY between Asp-68 and PE by investigating the existence of lipid-mediated salt bridges of LacY embedded in five different lipid matrices. Our simulations show a consistent and strong hydrogen bond between (non-, mono-, and dimethylated) POPE amine groups and Asp-68 that is significantly weaker in the case of PC (phosphocholine). In every instance, the bond is formed to a free hydrogen of the amine group with the speed of formation being inversely related to the degree of methylation.

The results have high significance: it is the first computational study ever to highlight a specific protein-lipid interaction, showing unambiguously a persistent interaction between a PE lipid and conserved residues of the protein. In addition, the study has led to significant biological insight into the action of membrane transport proteins.

The structure of the CD3$\zeta\zeta$ transmembrane dimer in lipid bilayers

Virtually every aspect of the human adaptive immune response is controlled by T cells. The T cell receptor (TCR) complex is responsible for the recognition of foreign peptide sequences, forming the initial step in the elimination of germ-infected cells. The recognition leads to an extracellular conformational change that is transmitted intracellularly through the Cluster of Differentiation 3 (CD3) subunits of the TCR-CD3 complex. The intracellular domains of the CD3 subunits are of varying length and are connected to the extracellular domain via a helical TM domain spanning the membrane only once.

The structure of CD3$\zeta\zeta$, in a mixture of dodecyl phosphocholine sodium dodecyl sulfate (DPC SDS), has recently been solved using NMR, showing CD3$\zeta\zeta$ to be a covalent dimer with an inter-chain disulfide bond involving conserved cysteines near the membrane surface. Four residues C-terminal to these, a conserved aspartic acid is found at position 6. Several questions as to the structure of the dimer in lipid bilayers remain, such as: *(i)* what is the protonation state of each of the two Aspartates in the dimer, *(ii)* what is their possible role in forming the disulfide-linkage and, *(iii)* do the aspartic acids have any effect on the integrity of the bilayer and on the local environment of the cysteines?

We have used extensive molecular dynamics simulations of the CD3$\zeta\zeta$ transmembrane dimer in lipid bilayers and in three different charge states, to answer these questions.

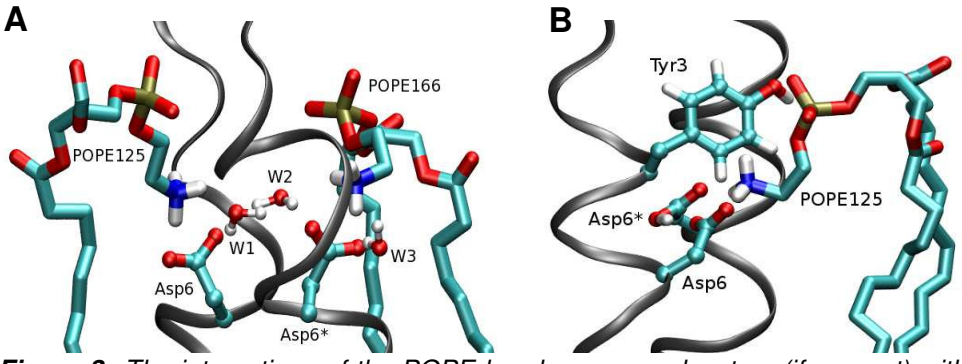


Figure 3: *The interactions of the POPE head groups and waters (if present) with the two aspartic acids for the (A) -2 and (B) -1 charge state. The interaction between Tyr3 and POPE125 is also highlighted.*

The results convincingly demonstrate that the presence of a charged aspartate residue near the membrane surface leads to a deformation of the lipid bilayer, thereby increasing the accessibility of the

cysteines to the machinery catalyzing disulfide bond formation. The Asp6/Asp6$*$ pair is shown to exist in a -1 charge state, meaning both aspartic acids stabilize each other through an inter-Asp hydrogen bond. In addition, we find both residues to be involved in a strong electrostatic interaction with a single lipid head group, also involved the conserved Tyr3. An illustration of this essential interaction is provided in Fig. 3.

The results have led to significant insight in the organization of the transmembrane domain of the TCR complex, including hypotheses on membrane insertion, dimerization, and assembly of a functional TCR-CD3 complex.

Probing the conformation of FhaC with small-angle neutron scattering and molecular modeling

Probing the solution structure of membrane proteins represents a formidable challenge, notably using small-angle scattering. Detergent molecules often present residual scattering contributions even at their match point in small-angle neutron scattering (SANS) measurements.

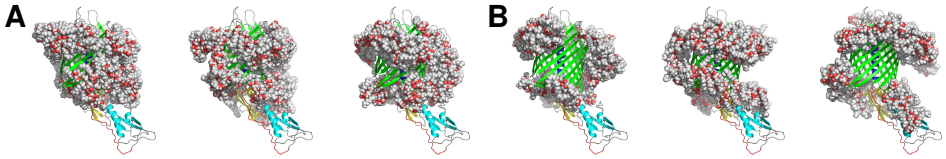


Figure 4: *Several FhaC-detergent arrangements, corresponding to "good" (**A**) and "bad" (**B**) χ^2 fits. FhaC and detergent shown in cartoon and sphere representation, resp.*

Here we studied the conformation of FhaC, the outer-membrane, β-barrel transporter of the *Bordetella pertussis* filamentous haemagglutinin adhesin. SANS measurements were performed on homogeneous solutions of FhaC solubiziled in n-octyl-d17βD-glucoside and on a variant devoid of the first α-helix that critically obstructs the FhaC pore, at two solvent conditions corresponding to the match points of the protein and the detergent, resp. By using molecular modelling and starting from three distinct conformations of FhaC and its variant embedded in lipid bilayers, we generated ensembles of protein-detergent arrangement models. The scattered curves were back-calculated for each model and compared with experimental data. Good fits were obtained for relatively compact, connected detergent belts that however display small detergent-free patches on the outer surface of the β-barrel. The combination of SANS and modelling clearly enabled us to infer the solution structure of FhaC, with H1 inside the pore as in the crystal structure. The computational methodology is relatively CPU-friendly and allows for the generation of a large number of protein-detergent arrangements. The χ^2 fits of back-calculated *vs.* experimental curves are decidedly discriminative and allow for the elimination of both protein conformation and detergent organization.

Our strategy that combines explicit atomic detergent modelling with SANS measurements holds significant potential for structural studies of other detergent-solubilized membrane proteins.

Fusogenic activity of cationic lipids

Cationic lipids have been extensively used as carriers of biologically active molecules (nucleic acids, peptides and proteins) into cells. DiC(14)-amidine (amidine) is a nonphysiological, cationic lipid that forms stable liposomes under physiological pH and temperature.

We characterized the cationic diC(14)-amidine bilayer by mixed DMPC/diC(14)-amidine molecular dynamics simulations, revealing a remarkable fluidity in the hydrophobic bilayer core, with a tendency for strong surface curvature, in agreement with the relatively small size of experimentally formed liposomes. The amidine bilayer shows an interdigitated, nonlamellar bilayer phase, with a bilayer thickness of only 2.7 nm and an average area per lipid of 0.83 nm^2, see also Fig. 5.

By combining FRET and confocal microscopy, we demonstrate that some cationic lipids do not require a co-lipid to fuse efficiently with cells. These cationic lipids are able to self-organize into bilayers that are stable enough to form liposomes, while presenting some destabilizing properties reminiscent of the conically shaped fusogenic co-lipid, DOPE. We therefore analyzed the resident lipid structures into populations of similarly shaped molecules, as opposed to the classical approach of using the static

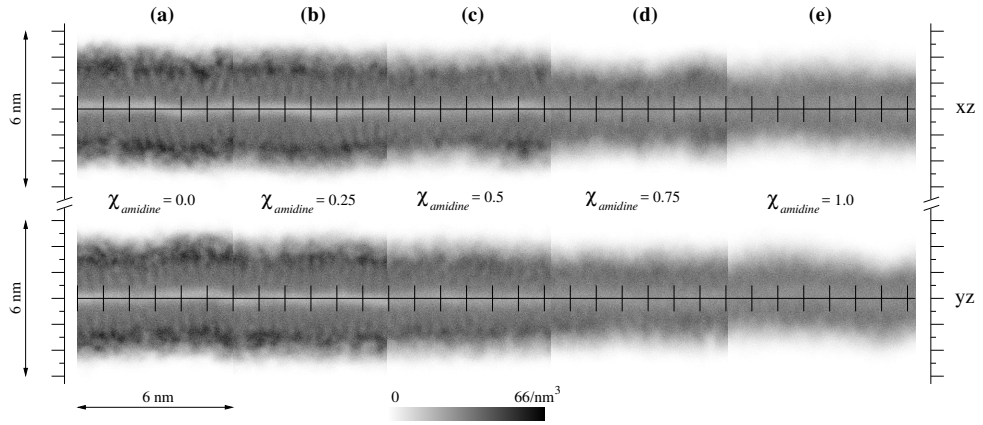


Figure 5: *Bilayer number density map in the xz (top) and (yz) (bottom) plane. Increasing amidine content in panels a-e in steps of 25%. Only bilayer atoms are counted, and all hydrogens are ignored. DMPC head groups contain eight heavy atoms more than amidine.*

packing parameter to define the lipid shapes. Comparison of fusogenic properties with these lipid populations suggests that the ratio of cylindrical *vs.* conical lipid populations correlates with the ability to fuse with cell membranes.

Publications

- Gabel *et al*, **Biophys J** 2014;107:185
- Grimard *et al*, in: **Bacterial membranes: structural and molecular biology**, Caister
- Sharma *et al*, **Biochim Biophys Acta** 2013;1838:739
- Bouchet *et al*, **Chem Phys Lipids** 2012;165:89
- Urbina *et al*, **Biochim Biophys Acta** 2011;1808:2618
- Lensink *et al*, **J Biol Chem** 2010;285:10519
- Lonez *et al*, **Cell Mol Life Sci** 2010;67:483
- Seil *et al*, **Pharmaceuticals** 2010;3:3435
- Lonez *et al*, **Biochim Biophys Acta** 2009;1790:425
- Lensink *et al*, **Langmuir** 2009;25:5230
- Lensink *et al*, **Biophys J** 2005;88:939

3.3 Protein and protein-small molecule modelling

Molecular mechanisms of signal transduction by BvgS, a paradigm for bacterial Venus-flytrap-domain receptors

Two-component systems (TCS) are the predominant signal transduction system in bacteria and enable them to adapt to changes in their environment. Perception of a chemical or physical signal leads to kinase activation and autophosphorylation, inciting the response regulator to mediate a – typical transcriptional – cellular response.

The dimeric BvgS consists of several distinct domains: a periplasmic sensor domain, a transmembrane domain, a PAS domain and several kinase domains, see Fig. 6a. We use computational techniques to study function and dynamics of the BvgS sensor-kinase TCS.

We have performed molecular dynamics simulations of wt BvgS periplasmic domain, totalling 1 μs simulation time. The simulations show highly flexible VFT1 domains, whereas the VFT2 domains change little (Fig. 6c). Individual lobes of each VFT domain are found to move in a concerted manner (Fig. 6b), as also confirmed through the use of both isotropic and anisotropic Gaussian network models. The lobes 1 of both VFT1 modules were found to occupy, independently, the two major eigenmotions of the system.

Using a combination of computational docking studies and experimental evidence, we show that nicotinate – virtually the only known modulator of BvgS activity – affects the behaviour of the entire

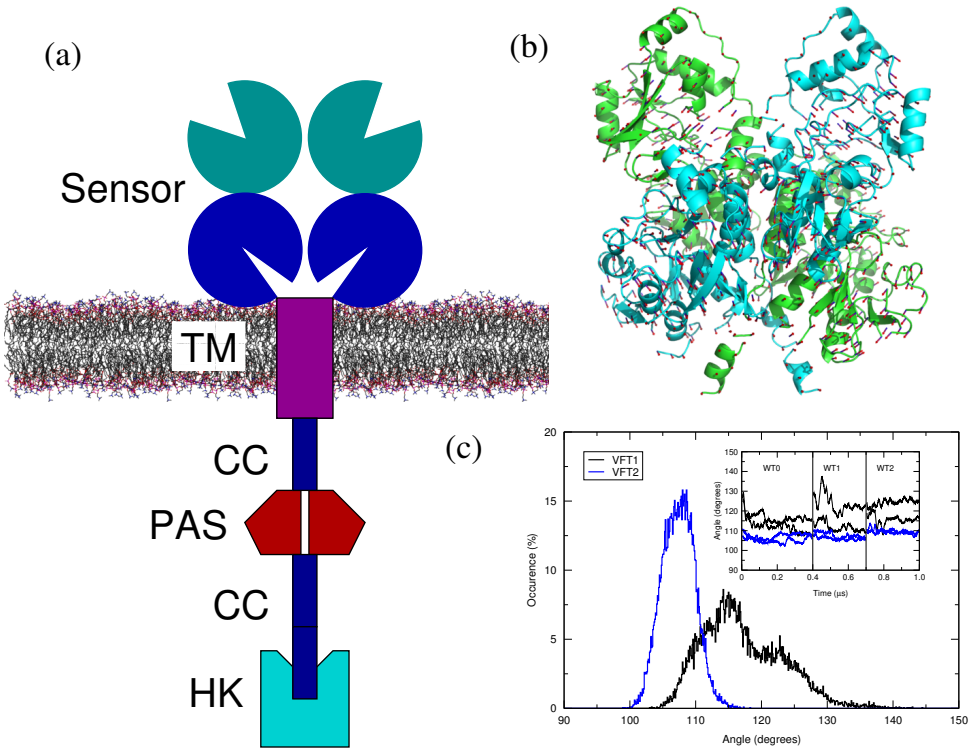


Figure 6: *(a) Modular architecture of BvgS, going from the N-terminal periplasmic sensor domain via a transmembrane (TM) domain, a so-called PAS domain flanked by two coiled-coil (CC) regions to the histidine kinase (HK). (b) Visualization of one of the major normal modes of the sensor domain. (c) Distribution of VFT opening angles over the course of 1 μs molecular dynamics simulation.*

periplasmic domain, by binding in the cleft of VFT2, and not to VFT1. Nicotinamide, a structurally related but uneffective molecule, was found to bind to neither module (VFT1 nor VFT2) convincingly.

Extensive molecular dynamics have been implemented to study the effect of various nicotinate and nicotinamide concentrations on the flexibility and dynamics of the VFT modules. The total simulation time amounts to 10 μs. Presence of either nicotinamide or nicotinate leads to a restriction in flexibility, but the restriction if markedly more evident when an additional nicotinate molecule is occupying the binding pocket. (Un)-binding of nicotinate could be associated to a large motion in the primary eigenvector.

A series of MD simulations have been performed of mutant variants showing modified activity *in vitro*. All simulations show VFT2 opening angles corresponding to the wt ones, however VFT1 angles are markedly different. The simulations explain how point mutations can decouple the motion between VFT1 and VFT2.

The putative structure of the transmembrane domain was investigated using secondary and tertiary structure prediction techniques. These predictions show a high tendency to form an α-helix, with a possibility of kink formation in the lower region of the upper bilayer half. Post-TM-helix and both pre- and post-PAS-domain sequence show a propensity for the formation of coiled-coil structure, starting immediately after the TM helix. Immersion of an in-silico built double helix (random pairing) into a lipid bilayer shows a hydrophobic thickness of 30 Å, corresponding to the optimal hydrophobic thickness of the TM region calculated from a minimization of the free energy of transfer. The tilt angle is close to zero degrees.

Targeting the Ets-1 oncoprotein

The Ets-1 transcription factor is the defining member of the ETS family, which is characterized by a DNA-binding domain, the ETS domain. Ets-1 needs protein partners in order to be activated and two such partners, DNA-repair enzymes DNA-PK and PARP-1, have recently been identified, including their respective domains of interaction. Nothing is known of the molecular details of these interactions, but

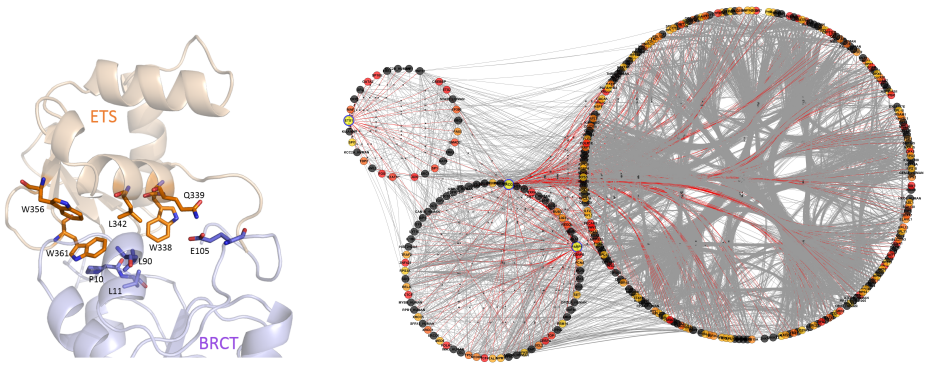


Figure 7: (a, left) *Detailed view of the interacting residues of the ETS/BRCT interface. Trp338 and Trp361 of ETS are buried in a hydrophobic cavity formed by Pro10, Leu11 and Leu90 of BRCT, while Gln339 is hydrogen-bonded to Glu105.* ***(b, right)*** *Illustrative view for the performed network analyses. The nodes organized in circles show Ets-1, PARP-1 and Ku70 as well as their homologues. Lines between the nodes denote interactions between the proteins, with their color – going from blue to red – indicating an up- or down-regulation. A selection of interactions is colored with red lines.*

given the important role that Ets-1 plays in various invasive diseases it is important to characterize the interaction between ETS and its protein partners at the molecular level.

We have identified the interaction surface of the interacting domains between Ets-1 and two established protein partners. The results identify the same interaction surface on ETS for binding to both protein partners, centered on α-helix H1. The interaction surface was rationalized by means of alanine scanning, molecular dynamics simulation, residue centrality analysis and pharmacophore construction. The models highlight a hydrophobic patch, including three tryptophanes (Trp338, Trp356, Trp361) at the center of the interaction interface, see Fig. 7.

We have subsequently identified human proteins containing structure and sequence homologues of the ETS domain and its interacting partner domains (BRCT for PARP-1 and SAP for DNA-PK), and constructed the global protein-protein interaction network, using both known and predicted interactions. Focusing in particular on proteins implicated in DNA repair, we identified several of these that contain domains homologous to BRCT and SAP and that could interact with Ets-1, when these interactions were not known beforehand. Docking simulations between ETS and these new protein partners systematically highlight two of the three aforementioned tryptohpnaes, indicating the possibility of Ets-1 recruiting other DNA-repair proteins.

Bacterial adhesion

Bacteria use adhesins to attach themselves to target cells. The adhesins need to withstand the shear stress excerted by for instance mucosal secretions, but the bacteria also need to be able to detach themselves under low stress in order to spread. The FimH protein, located at the top of *E. coli* type I fimbriae, mediates shear-dependent adhesion to high-mannose glycan structures present on the surface of the host cells.

We have started to look at these aspects from a computational point of view. Initial studies have focused on the threedimensional structure determination of FimH in complex with anti-adhesives. Other studies have investigated the opening and closing of the FimH binding pocket, the dynamical behaviour of selected ligands, and put these in a medical context. This is illustrated in Figure 8.

Three of the currently funded projects in the group involve bacterial adhesion. The team is a partner in the European Training Network “GLYCOVAX”. Our contribution consists of the study of carbohydrate-protein interactions, specifically the application of biophysical techniques (including X-ray diffraction, glycoarrays, SPR, calorimetry, computational modelling and dynamics) to study the interaction between oligosaccharides with antibodies. We are host to the “FimH-Mech” Marie Skłodowska-Curie fellowship, studying the mechanism of FimH bacterial adhesion through the investigation of the shear-force dependence of FimH and FimH variants by modelling the complex with target receptors. It will use state-

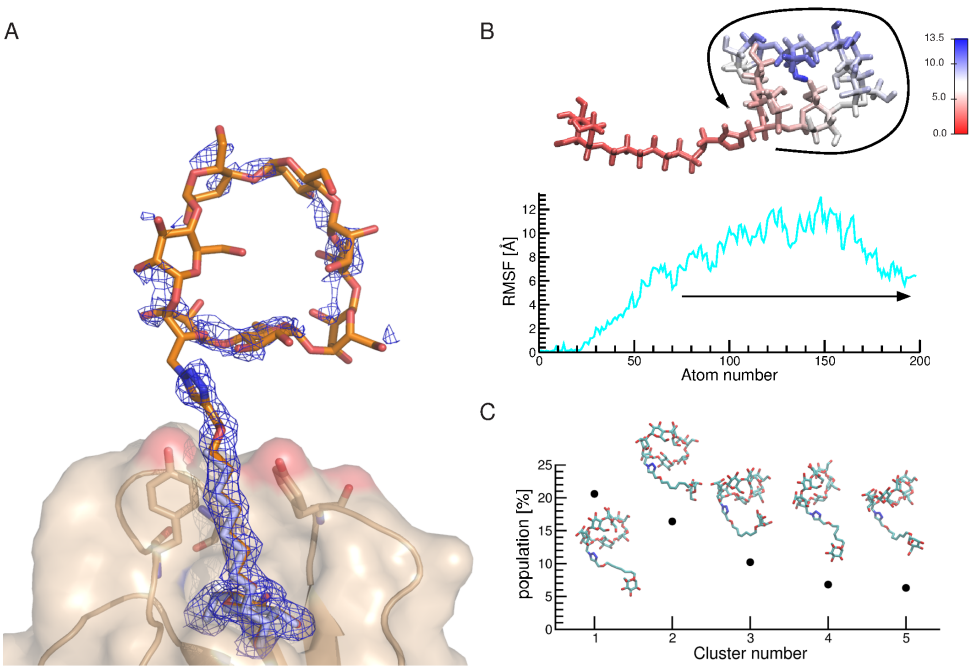


Figure 8: (A) *Crystal structure of FimH co-crystallized with a heptyl-mannose compound. The hydrophobic gate, composed of Tyr38, Tyr137 and Ile52 lines the entry of the binding pocket and is displayed in stick-form. The β-cyclodextrin ring is located outside of the binding pocket. The electron density map is shown on the ligand.* ***(B)*** *Atom-wise root-mean square fluctuations of the compound from a 150 ns molecular dynamics simulation of the complex (FimH and ligand), both in graph form and colored on the ligand (blue equaling more mobile atoms).* ***(C)*** *Major conformations of the heptyl-mannose compound in water, extracted from a cluster analysis of a 150 ns long molecular dynamics trajectory. The center conformation of each cluster is shown.*

of-the-art computational and theoretical techniques, in a back-and-forth interaction with experimental assays. And we are partner in the “HICARE” ANR, which aims to develop an innovative and unprecedented approach where multivalent inhibitor molecules simultaneously interact with several domains of carbohydrate-processing enzymes. In our contribution, we intend to use a powerful blend of analytical techniques with molecular modelling and simulations that is expected to result in quantifiable and predictive biophysical models of the binding modes. The high-resolution models from molecular modelling and quantum-chemical calculations will be combined with X-ray and laser scattering experiments, and ITC thermodynamic and SPR kinetic characterization of the interactions.

Publications

- Touaibia *et al*, **Molecules** 2017;22:E1101
- Rabbani *et al*, **IUCrJ** 2017;4:7
- Lesne *et al*, **MBio** 2016;7:e02089
- De Ruyck *et al*, **Adv Appl Bioinform Chem** 2016;9:1
- De Ruyck *et al*, **IUCrJ** 2016;3:163
- Alvarez Dorta *et al*, **Chembiochem** 2016;17:936
- Dupré *et al*, **PLoS Pathog** 2015;11:e1004700
- Dupré *et al*, **J Biol Chem** 2015;290:23307
- Brysbaert *et al*, **J Phys Condens Matter** 2015;27:064108
- Singh *et al*, **AIMS Biophys** 2015;2:398
- Lensink, **Methods Mol Biol** 2014;1215:109

4. SELECTED PUBLICATIONS

- Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman-Duhovny D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath S, Elber R, Grudinin S, Popov P, Neveu E, Lee H, Baek M, Park S, Heo L, Rie Lee G, Seok C, Qin S, Zhou HX, Ritchie DW, Maigret B, Devignes MD, Ghoorah A, Torchala M, Chaleil RA, Bates PA, Ben-Zeev E, Eisenstein M, Negi SS, Weng Z, Vreven T, Pierce BG, Borrman TM, Yu J, Ochsenbein F, Guerois R, Vangone A, Rodrigues JP, van Zundert G, Nellen M, Xue L, Karaca E, Melquiond AS, Visscher K, Kastritis PL, Bonvin AM, Xu X, Qiu L, Yan C, Li J, Ma Z, Cheng J, Zou X, Shen Y, Peterson LX, Kim HR, Roy A, Han X, Esquivel-Rodriguez J, Kihara D, Yu X, Bruce NJ, Fuller JC, Wade RC, Anishchenko I, Kundrotas PJ, Vakser IA, Imai K, Yamada K, Oda T, Nakamura T, Tomii K, Pallara C, Romero-Durana M, Jimnez-Garca B, Moal IH, Frnandez-Recio J, Joung JY, Kim JY, Joo K, Lee J, Kozakov D, Vajda S, Mottarella S, Hall DR, Beglov D, Mamonov A, Xia B, Bohnuud T, Del Carpio CA, Ichiishi E, Marze N, Kuroda D, Roy Burman SS, Gray JJ, Chermak E, Cavallo L, Oliva R, Tovchigrechko A, Wodak SJ,
Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment,
Proteins 2016;84 Suppl 1:323. **Citations: 22**

  "As of May/June 2017, this highly cited paper received enough citations to place it in the top 1% of the academic field of Biology & Biochemistry based on a highly cited threshold for the field and publication year." – Web Of Science

- Lensink MF, Govaerts C, Ruysschaert J-M,
Identification of specific lipid-binding sites in integral membrane proteins,
J Biol Chem 2010;285:10519. **Citations: 26**

- Copie G, Cleri F, Blossey R, Lensink MF,
On the ability of molecular dynamics simulation and continuum electrostatics to treat interfacial water molecules in protein-protein complexes,
Sci Rep 2016;6:38259. **Citations; 1**

- Gabel F, Lensink MF, Clantin B, Jacob-Dubuisson F, Villeret V, Ebel C,
Probing the conformation of FhaC with small-angle neutron scattering and molecular modeling,
Biophys J 2014;107:185. **Citations: 7**

- Dupré E, Herrou J, Lensink MF, Wintjens R, Vagin A, Lebedev A, Crosson S, Villeret V, Locht C, Antoine R, Jacob-Dubuisson F,
Virulence regulation with Venus flytrap domains: Structure and function of the periplasmic moiety of the sensor-kinase BvgS,
PLoS Pathog 2015;11:e1004700. **Citations: 15**

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A CASP-CAPRI experiment

Marc F. Lensink,[1]* Sameer Velankar,[2] Andriy Kryshtafovych,[3] Shen-You Huang,[4] Dina Schneidman-Duhovny,[5,6] Andrej Sali,[5,6,7] Joan Segura,[8] Narcis Fernandez-Fuentes,[9] Shruthi Viswanath,[10,11] Ron Elber,[11,12] Sergei Grudinin,[13,14] Petr Popov,[13,14,15] Emilie Neveu,[13,14] Hasup Lee,[16] Minkyung Baek,[16] Sangwoo Park,[16] Lim Heo,[16] Gyu Rie Lee,[16] Chaok Seok,[16] Sanbo Qin,[17] Huan-Xiang Zhou,[17] David W. Ritchie,[18] Bernard Maigret,[19] Marie-Dominique Devignes,[19] Anisah Ghoorah,[20] Mieczyslaw Torchala,[21] Raphaël A.G. Chaleil,[21] Paul A. Bates,[21] Efrat Ben-Zeev,[22] Miriam Eisenstein,[23] Surendra S. Negi,[24] Zhiping Weng,[25] Thom Vreven,[25] Brian G. Pierce,[25] Tyler M. Borrman,[25] Jinchao Yu,[26] Françoise Ochsenbein,[26] Raphaël Guerois,[26] Anna Vangone,[27] João P.G.L.M. Rodrigues,[27] Gydo van Zundert,[27] Mehdi Nellen,[27] Li Xue,[27] Ezgi Karaca,[27] Adrien S.J. Melquiond,[27] Koen Visscher,[27] Panagiotis L. Kastritis,[27] Alexandre M.J.J. Bonvin,[27] Xianjin Xu,[28] Liming Qiu,[28] Chengfei Yan,[28,29] Jilong Li,[30] Zhiwei Ma,[28,29] Jianlin Cheng,[30,31] Xiaoqin Zou,[28,29,31,32] Yang Shen,[33] Lenna X. Peterson,[34] Hyung-Rae Kim,[34] Amit Roy,[34,35] Xusi Han,[34] Juan Esquivel-Rodriguez,[36] Daisuke Kihara,[34,36] Xiaofeng Yu,[37] Neil J. Bruce,[37] Jonathan C. Fuller,[37] Rebecca C. Wade,[37,38,39] Ivan Anishchenko,[40] Petras J. Kundrotas,[40] Ilya A. Vakser,[40,41] Kenichiro Imai,[42] Kazunori Yamada,[42] Toshiyuki Oda,[42] Tsukasa Nakamura,[43] Kentaro Tomii,[42,43] Chiara Pallara,[44] Miguel Romero-Durana,[44] Brian Jiménez-García,[44] Iain H. Moal,[44] Juan Férnandez-Recio,[44] Jong Young Joung,[45] Jong Yun Kim,[45] Keehyoung Joo,[45,46] Jooyoung Lee,[45,47] Dima Kozakov,[48] Sandor Vajda,[48,49] Scott Mottarella,[48] David R. Hall,[48] Dmitri Beglov,[48] Artem Mamonov,[48] Bing Xia,[48] Tanggis Bohnuud,[48] Carlos A. Del Carpio,[50,51] Eichiro Ichiishi,[52] Nicholas Marze,[53] Daisuke Kuroda,[53] Shourya S. Roy Burman,[53] Jeffrey J. Gray,[53,54] Edrisse Chermak,[55] Luigi Cavallo,[55] Romina Oliva,[56] Andrey Tovchigrechko,[57] and Shoshana J. Wodak[58,59]*

Additional Supporting Information may be found in the online version of this article.
Grant sponsor: NIH; Grant numbers: R01 GM083960; P41 GM109824; GM058187; R01 GM061867; R01 GM093147; R01 GM078221; R01GM109980; R01GM094123; R01 GM097528; R01GM074255; Grant sponsor: Biotechnology and Biological Sciences Research Council; Grant number: BBS/E/W/10962A01D; Grant sponsor: Research Councils UK Academic Fellowship program; Grant sponsor: Cancer Research UK; Grant sponsor: Klaus Tschira Foundation; Grant sponsor: Platform Project for Supporting in Drug Discovery and Life Science Research; Grant sponsor: Japan Agency for Medical Research and Development; Grant sponsor: Agence Nationale de la Recherche; Grant number: ANR-11-MONU-0006; Grant sponsor: National Research Foundation of Korea (NRF); Grant numbers: NRF-2013R1A2A1A09012229; 2008-0061987; Grant sponsor: BIP; Grant number: ANR-IAB-2011-16-BIP:BIP; Grant sponsor: H2020 Marie Sklodowska-Curie Individual Fellowship; Grant number: 659025-BAP; Grant sponsor: Netherlands Organization for Scientific Research Veni; Grant number: 722.014.005; Grant sponsor: National Science Foundation; Grant numbers: CAREER Award DBI0953839; CCF-1546278; NSF IIS1319551; NSF DBI1262189; NSF IOS1127027; NSF DBI1262621; NSF DBI 1458509; NSF AF 1527292; Grant sponsor: EU; Grant number: FP7 604102 (HBP); Grant sponsor: BMBF; Grant number: 0315749 (VLN); Grant sponsor: Spanish Ministry of Economy and Competitiveness; Grant number: BIO2013-48213-R; Grant sponsor: European Union; Grant number: FP7/2007-2013 REA PIEF-GA-2012-327899; Grant sponsor: National Institute of Supercomputing and Networking; Grant number: KSC-2014-C3-01; Grant sponsor: US-Israel BSF; Grant number: 2009418; Grant sponsor: Regione Campania; Grant number: LR5-AF2008.

*Correspondence to: Marc F. Lensink; University Lille, CNRS UMR8576 UGSF, Lille, F-59000, France. E-mail: marc.lensink@univ-lille1.fr or Shoshana J. Wodak; VIB Structural Biology Research Center, VUB, 1050 Brussels, Belgium. E-mail: shoshana.wodak@gmail.com
Tyler M. Borrman current address is Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD 20850.
Yang Shen current address is Center for Bioinformatics and Genomic Systems Engineering, Department of Electrical and Computer Engineering, Texas a&M University, College Station, TX 77843.
Kenichiro Imai, Toshiyuki Oda, and Kentaro Tomii current address is Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology (AIST), Koto-Ku, Japan.
Kazunori Yamada current address is Group of Electrical Engineering, Communication Engineering, Electronic Engineering, and Information Engineering, Tohoku University, Sendai, Japan.
Iain H. Moal current address is European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom.
Daisuke Kuroda current address is School of Pharmacy, Showa University, Shinagawa-Ku, Tokyo 142-8555, Japan.
Received 29 May 2015; Revised 30 December 2015; Accepted 2 February 2016
Published online 00 Month 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.25007

[1] University Lille, CNRS UMR8576 UGSF, Lille, F-59000, France
[2] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
[3] Genome Center, University of California, Davis, California, 95616
[4] Research Support Computing, University of Missouri Bioinformatics Consortium, and Department of Computer Science, University of Missouri, Columbia, Missouri 65211
[5] Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California 94158
[6] Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California 94158
[7] California Institute for Quantitative Biosciences (QB3), University of California San Francisco, San Francisco, California 94158
[8] GN7 of the National Institute for Bioinformatics (INB) and Biocomputing Unit, National Center of Biotechnology (CSIC), Madrid, 28049, Spain
[9] Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Aberystwyth, SY233FG, United Kingdom
[10] Department of Computer Science, University of Texas at Austin, Austin, Texas 78712
[11] Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, Texas 78712
[12] Department of Chemistry, University of Texas at Austin, Austin, Texas 78712
[13] LJK, University Grenoble Alpes, CNRS, Grenoble, 38000, France
[14] INRIA, Grenoble, 38000, France
[15] Moscow Institute of Physics and Technology, Dolgoprudniy, Russia
[16] Department of Chemistry, Seoul National University, Seoul, 151-747, Republic of Korea
[17] Department of Physics and Institute of Molecular Biophysics, Florida State University, Tallahassee, Florida 32306, USA
[18] INRIA Nancy—Grand Est, Villers-lès-Nancy, 54600, France
[19] CNRS, LORIA, Campus Scientifique, BP 239, Vandœuvre-lès-Nancy, 54506, France
[20] Department of Computer Science and Engineering, University of Mauritius, Reduit, Mauritius
[21] Biomolecular Modelling Laboratory, the Francis Crick Institute, Lincoln's Inn Fields Laboratory, London, WC2A 3LY, United Kingdom
[22] G-INCPM, Weizmann Institute of Science, Rehovot, 7610001, Israel
[23] Department of Chemical Research Support, Weizmann Institute of Science, Rehovot, 7610001, Israel
[24] Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, 301 University Boulevard, Galveston, Texas 77555-0857
[25] Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605
[26] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Saclay, CEA-Saclay, Gif-sur-Yvette, 91191, France
[27] Bijvoet Center for Biomolecular Research, Faculty of Science – Chemistry, Utrecht University, Padualaan 8, Utrecht, 3584 CH, The Netherlands
[28] Dalton Cardiovascular Research Center, University of Missouri, Columbia, Missouri 65211
[29] Department of Physics and Astronomy, University of Missouri, Columbia, Missouri 65211
[30] Department of Computer Science, University of Missouri, Columbia, Missouri 65211
[31] Informatics Institute, University of Missouri, Columbia, Missouri 65211
[32] Department of Biochemistry, University of Missouri, Columbia, Missouri 65211
[33] Toyota Technological Institute at Chicago, 6045 S Kenwood Avenue, Chicago, Illinois 60637
[34] Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907
[35] Bioinformatics and Computational Biosciences Branch, Rocky Mountain Laboratories, National Institutes of Health, Hamilton, Montano 59840
[36] Department of Computer Science, Purdue University, West Lafayette, IN, USA 47907
[37] Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany
[38] Center for Molecular Biology (ZMBH), DKFZ-ZMBH Alliance, Heidelberg University, Heidelberg, Germany
[39] Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Heidelberg, Germany
[40] Center for Computational Biology, The University of Kansas, Lawrence, Kansas 66047
[41] Department of Molecular Biosciences, The University of Kansas, Lawrence, Kansas 66047
[42] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Koto-Ku, Japan
[43] Graduate School of Frontier Sciences, the University of Tokyo, Kashiwa, Japan
[44] Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, C/Jordi Girona 29, Barcelona, 08034, Spain
[45] Center for in-Silico Protein Science, Korea Institute for Advanced Study, Seoul, 130-722, Korea
[46] Center for Advanced Computation, Korea Institute for Advanced Study, Seoul, 130-722, Korea
[47] School of Computational Science, Korea Institute for Advanced Study, Seoul, 130-722, Korea
[48] Department of Biomedical Engineering, Boston University, Boston, Massachusetts
[49] Department of Chemistry, Boston University, Boston, Massachusetts
[50] Institute of Biological Diversity, International Pacific Institute of Indiana, Bloomington, Indiana 47401
[51] Drosophila Genetic Resource Center, Kyoto Institute of Technology, Ukyo-Ku, 616-8354, Japan
[52] International University of Health and Welfare Hospital (IUHW Hospital), Asushiobara-City, Tochigi Prefecture 329-2763, Japan
[53] Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland 21218
[54] Program in Molecular Biophysics, Johns Hopkins University, Baltimore, Maryland 21218
[55] King Abdullah University of Science and Technology, Saudi Arabia
[56] University of Naples "Parthenope", Napoli, Italy
[57] J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, Maryland 20850
[58] Departments of Biochemistry and Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[59] VIB Structural Biology Research Center, VUB Pleinlaan 2, Brussels, 1050, Belgium

ABSTRACT

We present the results for CAPRI Round 30, the first joint CASP-CAPRI experiment, which brought together experts from the protein structure prediction and protein–protein docking communities. The Round comprised 25 targets from amongst those submitted for the CASP11 prediction experiment of 2014. The targets included mostly homodimers, a few homotetramers, and two heterodimers, and comprised protein chains that could readily be modeled using templates from the Protein Data Bank. On average 24 CAPRI groups and 7 CASP groups submitted docking predictions for each target, and 12 CAPRI groups per target participated in the CAPRI scoring experiment. In total more than 9500 models were assessed against the 3D structures of the corresponding target complexes. Results show that the prediction of homodimer assemblies by homology modeling techniques and docking calculations is quite successful for targets featuring large enough subunit interfaces to represent stable associations. Targets with ambiguous or inaccurate oligomeric state assignments, often featuring crystal contact-sized interfaces, represented a confounding factor. For those, a much poorer prediction performance was achieved, while nonetheless often providing helpful clues on the correct oligomeric state of the protein. The prediction performance was very poor for genuine tetrameric targets, where the inaccuracy of the homology-built subunit models and the smaller pair-wise interfaces severely limited the ability to derive the correct assembly mode. Our analysis also shows that docking procedures tend to perform better than standard homology modeling techniques and that highly accurate models of the protein components are not always required to identify their association modes with acceptable accuracy.

Proteins 2016; 00:000–000.

Key words: CAPRI; CASP; oligomer state; blind prediction; protein interaction; protein docking.

INTRODUCTION

Most cellular processes are carried out by physically interacting proteins.[1] Characterizing protein interactions and higher order assemblies is therefore a crucial step in gaining an understanding of how cells function.

Regrettably, protein assemblies are still poorly represented in the Protein Databank (PDB).[2] Determining the structures of such assemblies has so far been hampered by the difficulty in obtaining suitable crystals and diffraction data. But this limitation is being circumvented with the advent of new powerful electron microscopy techniques, which now enable the structure determinations of very large macromolecular assemblies at atomic resolutions.[3]

On the other hand, the repertoire of individual protein 3D structures has been increasingly filled, thanks to large-scale structural genomics projects such as the PSI (http://sbkb.org/) and others (http://www.thesgc.org/). Given a newly sequenced protein, the odds are high that its 3D structure can be readily extrapolated from structures of related proteins deposited in the PDB.[4,5] Moreover, thanks to the recent explosion of the number of available protein sequences, it is now becoming possible to model the structures of individual proteins with increasing accuracy from sequence information alone[6,7] as will be highlighted in the CASP11 results in this issue. Structures from this increasingly rich repertoire may be used as templates or scaffolds in protein design projects that have useful medical applications.[8,9] Larger protein assemblies can be modeled by integrating information on individual structures with various other types of data with the help of hybrid modeling techniques.[10]

Computational approaches play a major role in all these endeavors. Of particular importance are methods for deriving accurate structural models of multiprotein assemblies, starting from the atomic coordinates of the individual components, the so-called "docking" algorithms, and the associated energetic criteria for singling out stable binding modes.[11–13]

Taking its inspiration from CASP, the community-wide initiative on the Critical Assessment of Predicted Interactions (CAPRI), established in 2001, has been designed to test the performance of docking algorithms (http://www.ebi.ac.uk/msd-srv/capri/). Just as CASP has fostered the development of methods for the prediction of protein structures, CAPRI has played an important role in advancing the field of modeling protein assemblies. Initially focusing on protein–protein docking and scoring procedures, CAPRI has expanded its horizon by including targets representing protein-peptide and protein nucleic acids complexes. It has moreover conducted experiments aimed at evaluating the ability of computational methods to estimate binding affinity of protein–protein complexes[14–16] and to predict the positions of water molecules at the interfaces of protein complexes.[17]

Considering the importance of macromolecular assemblies, and the new opportunities offered by the recent progress in both experimental and computational techniques to probe and model these assemblies, a better integration of the different computational approaches for modeling macromolecular assemblies and their building blocks was called for. Establishing closer ties between the CASP and CAPRI communities appeared as an important step in this direction, inaugurated by running a

joint CASP-CAPRI prediction experiment in the summer of 2014. The results of this experiment were summarized at the CASP11 meeting held in Dec 2014 in Cancun Mexico, and are presented in detail in this report.

The CASP11-CAPRI experiment, representing CAPRI Round 30, comprised 25 targets for which predictions of protein complexes were assessed. These targets represented a subset of the 100 regular CASP11 targets. This subset comprised only "easy" CASP targets, those whose 3D structure could be readily modeled using standard homology modeling techniques. Targets that required more sophisticated approaches (*ab-initio* modeling, or homology modeling using very distantly related templates) were not considered, as the CAPRI community had little experience with these approaches. The vast majority of the targets were homo-oligomers. CAPRI groups were given the choice of modeling the subunit structures of these complexes themselves, or using models made available by CASP participant, in time of the docking calculations.

On average, about 25 CAPRI groups, and about 7 CASP groups submitted docking predictions for each target. About 12 CAPRI scorer groups per target participated in the CAPRI scoring experiment, where participants are invited to single out correct models from an ensemble of anonymized predicted complexes generated during the docking experiment.

In total, these groups submitted >9500 models that were assessed against the 3D structures of the corresponding targets. The assessment was performed by the CAPRI assessment team, using the standard CAPRI model quality measures.[18,19] A major issue for the assessment, and for the Round as a whole, was the uncertainties in the oligomeric state assignments for a significant number of the targets. For many of these the assigned state at the time of the experiment was inferred solely from the crystal contacts by computational methods, which can be unreliable.

In presenting the CAPRI Round 30 assessment results here, we highlight this issue and the more general challenge of correctly predicting the association modes of weaker complexes of identical subunits, and those of higher order homo-oligomers. In addition, we examine the influence of the accuracy of the modeled subunits on the performance of the docking and scoring predictions, and evaluate the extent to which docking procedures confer an advantage over standard homology modeling methods in predicting homo-oligomer complexes.

THE TARGETS

The 25 targets of the joint CASP-CAPRI experiment are listed in Table I. Of these 23 are homo-oligomers, with 18 declared to be dimers and five to be tetramers, and two heterocomplexes. Hence for the majority of the targets (23) the goal was to model the interface (or interfaces in the case of tetramers) between identical subunits, whose size varied between 44 and 669 residues but was of ~250 residues on average. The majority of the targets were obtained from structural genomics consortia. They represented mainly microbial proteins, whose function was often annotated as putative.

Since it is not uncommon for docking approaches to use information on the symmetry of the complex to restrain or filter docking poses, predictors needed to be given reliable information on the biologically/functionally relevant oligomeric state of the target complex to be predicted. While self association between proteins is common, with between 50 and 75% of proteins forming dimers in the cell,[20,21] this association depends on the binding affinity between the subunits and on their concentration. Information on the oligomeric state is in principle derived using experimental methods such as gel filtration or small-angle X-ray scattering (SAXS),[22] and is usually communicated by the authors upon submission of the atomic coordinates to the PDB. With a majority of the targets being offered by structural genomics consortia before their coordinates were deposited in the PDB, author-assigned oligomeric states were available to predictors only for a subset (~15) of the targets, and those were often tentative. For the remaining targets, the oligomeric state was inferred from the crystal contacts using the PISA software,[23] which although being a widely used standard in the field, may still yield erroneous assignments in a non-negligible fraction of the cases, as will be shown in this analysis. Such incorrect assignments represented a confounding factor in this CAPRI round, but also allowed to show that docking calculations may help to correct them.

GLOBAL OVERVIEW OF THE PREDICTION EXPERIMENT

As in typical CAPRI Rounds, CAPRI predictor groups were provided with the amino-acid sequence of the target protein (for homo-oligomers), or proteins (for heterocomplexes), and with some relevant details about the protein, communicated by the structural biologists. Using the sequence information, the groups were then invited to model the 3D structure of the protein or proteins, and to derive the atomic structure of the complex. To help with the homology-modeling task, with which CASP participants are usually more experienced than their CAPRI colleagues, 3D models of individual target proteins predicted by CASP participants were made available to CAPRI groups for use in their docking calculations. A good number of CAPRI groups, but not all, took up this offer.

Table I
The CAPRI-CASP11 Targets of CAPRI Round 30

Target ID			Quaternary state			Buried	
CAPRI	CASP	Contributor	Author	PISA	Residues	area (Å^2)	Protein
T68	T0759	NSGC	**1 or 2**	**1**	109	860	Plectin 1 and 2 repeats (HR9083A) of the Human Periplakin
T69	T0764	JCSG	2	**2**	341	2415	Putative esterase (BDI_1566) from Parabacteroides distasonis
T70	T0765	JCSG	**2**	**4**	128	2030	Modulator protein MzrA (KPN_03524) from Klebsiella pneumoniae subsp.
T71	T0768	JCSG	**4**	**4**	170	2380	Leucine rich repeat protein (BACCAP_00569) from Bacteroides capillosus ATCC 29799
T72	T0770	JCSG	2	**2**	488	1120	SusD homolog (BT2259) from Bacteroides thetaiotaomicron
T73	T0772	JCSG	**4**	**4**	265	5900	Putative glycosyl hydrolase (BDI_3914) from Parabacteroides distasonis
T74	T0774	JCSG	**1**	**4**	379	2040	Hypothetical protein (BVU_2522) from Bacteroides vulgatus
T75	T0776	JCSG	2	**2**	256	1040	Putative GDSL-like lipase (PARMER_00689) from Parabacteroides merdae (ATCC 43184)
T77	T0780	JCSG	2	**2**	259	1600	Conserved hypothetical protein (SP_1560) from Streptococcus pneumoniae TIGR4
T78	T0786	Non-SGI	**4**	**4**	264	4160	Hypothetical protein (BCE0241) from Bacillus cereus
T79	T0792	Non-SGI		**2**	80	680	OSKAR-N
T80	T0801	NPPB	**2**	2	376	1960	Sugar aminotransferase WecE from Escherichia coli K-12
T81	T0797	Non-SGI	2	2	44	1070	cGMP-dependent protein kinase II leucine zipper
	T0798		2	2	198		Rab11b protein
T82	T0805	Non-SGI	**2**	2	214	3250	Nitro-reductase rv3368
T84	T0811	NYSGRC		**2**	255	1740	Triose phosphate isomerase
T85	T0813	NYSGRC	**2**	**2**	307	4620	Cyclohexadienyl dehydrogenase from Sinorhizobium meliloti in complex with NADP
T86	T0815	NYSGRC	**2**	**2**	106	470	Putative polyketide cyclase (protein SMa1630) from Sinorhizobium meliloti
T87	T0819	NYSGRC	**2**	**2**	373	3430	Histidinol-phosphate aminotransferase from Sinorhizobium meliloti in complex with pyridoxal-5'-phosphate
T88	T0825	Non-SGI	**2**	**2**	205	1350	WRAP-5
T89	T0840	Non-SGI	1		669	870	RON receptor tyrosine kinase subunit
	T0841		1		253		Macrophage stimulating protein subunit (MSP)
T90	T0843	MCSG	2	**2**	369	2360	Ats13
T91	T0847	SGC	**1**	**2**	176	1320	Human Bj-Tsa-9
T92	T0849	MCSG	**2**	**2**	240	1900	Glutathione S-transferase domain from Haliangium ochraceum DSM 14365
T93	T0851	MCSG	**2**	**2**	456	2680	Cals8 from Micromonospora echinospora (P294S mutant)
T94	T0852	MCSG	**2**	**2**	414	1190	APC103154

Bold numbers under Quaternary State indicate the oligomeric state assignments available at the time of the prediction experiment; 1 (monomer), 2 (dimer), 4 (tetramer); numbers in regular fonts indicate subsequent assignments collected from the PDB entries for the target structures.
NSGC, Northeast Structural Genomics Consortium; JCSG, Joint Center for Structural Genomics; Non-SGI, Non-SGI research Centers and others; NNPB, NatPro PSI:Biology; NYSGRC, New York Structural Genomics Research Center; MCSG, Midwest Center for Structural Genomics; SGC, Structural Genomics Consortium.

In addition to submitting 10 models for each target complex, predictors were invited to upload a set of 100 models. Once all the submissions were completed, the uploaded models were shuffled and made available to all groups as part of the CAPRI scoring experiment. The "scorer" groups were in turn invited to evaluate the ensemble of uploaded models using the scoring function of their choice, and submit their own 10 best ranking ones. The typical timelines per target were about 3 weeks for the homology modeling and docking predictions, and 3 days for the scoring experiment.

Table II
CAPRI Round 30 Experiment Statistics

Target ID				Number of groups				Number of models			
				CAPRI			CASP	CAPRI			CASP
CAPRI	CASP	PDB	[a]	Predictors	Uploaders	Scorers	Predictors	Predictors	Uploaders	Scorers	Predictors
T68	T0759	4q28	2	23	10	12	3	221	1000	120	7
T69	T0764	4q34	2	28	10	14	7	266	1000	132	17
T70	T0765	4pwu	2	23	8	13	5	221	710	130	18
T71	T0768	4oju	3	22	9	14	1	214	810	131	1
T72	T0770	4q69	3	25	11	13	4	244	914	130	11
T73	T0772	4qhz	2	23	11	11	7	221	1195	110	16
T74	T0774	4qb7	2	22	11	10	7	202	911	96	11
T75	T0776	4q9a	1	26	12	12	8	253	840	120	21
T76	T0779			*Cancelled – no structure*							
T77	T0780	4qdy	4	24	12	12	6	229	971	120	12
T78	T0786	4qvu	2	24	10	11	5	229	818	110	15
T79	T0792	5a49	3	25	11	12	9	242	900	120	23
T80	T0801	4piw	1	27	10	12	8	264	911	120	27
T81	T0797	4ojk	1	23	9	11	20	218	641	110	64
	T0798										
T82	T0805	[b]	1	25	10	12	9	242	911	120	27
T83	T0809			*Cancelled – article from different group online*							
T84	T0811	[b]	1	25	10	12	10	241	910	120	28
T85	T0813	4wji	1	25	11	12	8	241	920	120	21
T86	T0815	4u13	2	26	11	12	9	251	1010	119	25
T87	T0819	4wbt	1	24	10	12	9	231	894	120	25
T88	T0825	[b]	1	27	10	13	18	261	910	130	62
T89	T0840	[b]	1	22	9	11	55	211	790	110	243
	T0841										
T90	T0843	4xau	1	23	9	11	9	221	811	110	28
T91	T0847	4urj	1	25	9	11	9	242	798	110	24
T92	T0849	4w66	1	23	9	11	9	225	789	110	33
T93	T0851	4wb1	1	22	9	11	8	213	697	110	27
T94	T0852	4w9r	1	22	9	12	8	215	783	120	21

The number of groups corresponds to registered groups that effectively submitted models for the respective target. The number of models represents submitted models, regardless of quality and includes disqualified models. CAPRI groups are allowed to submit no more than their 10 best models, whereas CASP groups are allowed to submit no more than their 5 best models.
[a]Number of interfaces assessed.
[b]Not yet released.

Table II lists for each target the number of groups submitting predictions and the number of models assessed. On average ~25 CAPRI groups submitted a total of ~230 models per target, and an average of 12 scorer groups submitted a total of ~120 models per target. With the exception of three targets, an average of seven groups registered with CASP submitted a total of anywhere between 1 and 33 models for individual targets. CASP predictors participated in larger numbers in the prediction of T88 (T0825) and of the heterocomplexes (T89 – T0840/T0841 and T81 – T0797/T0798), where the CASP targets were defined as the oligomeric structures.

Table II also lists the uploader groups and the models that they make available for the scoring experiment (100 models per target per uploader group). As detailed above, the uploaded models are complexes output by the docking calculations carried out by individual participants for a given target. Models, uploaded by the different groups, are anonymized, shuffled, and made available to groups solely interested in testing their scoring functions.

SYNOPSIS OF THE PREDICTION METHODS

Round 30 participants used a wide range of modeling methods and software tools to generate the submitted models. In addition, the approaches used by a given group often differed across targets. Here, we provide only a short synopsis of the main methodological approaches. For a more detailed description of the methods and modeling strategies, readers are referred to the extended Methods Abstracts provided by individual participants (see Supporting Information Table S6).

Templates, representing known structures of homologs to a given target, stored in the PDB, were used in a number of ways. Most commonly, they were employed to model the 3D structures of individual subunits. Some

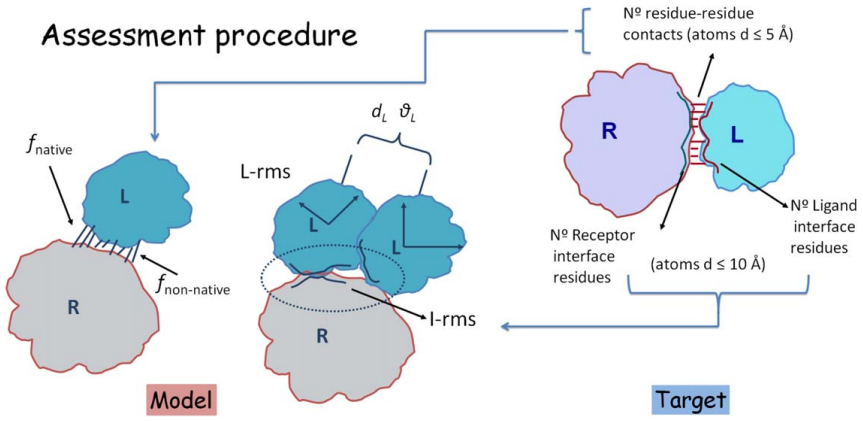


Figure 1

Schematic illustration of the CAPRI assessment criteria. The following quantities were computed for each target: (1) all the residue-residue contacts between the Receptor (R) and the Ligand (L), and (2) the residues contributing to the interface of each of the components of the complex. Interface residues were defined on the basis of their contribution to the interface area, as described in references.[18,19] For each submitted model the following quantities were computed: the fractions *f(nat)* of native and *f(non-nat)* of non-native contacts in the predicted interface; the root mean square displacement (rmsd) of the backbone atoms of the ligand (*L-rms*), the mis-orientation angle θ_L and the residual displacement d_L of the ligand center of mass, after the receptor in the model and experimental structures were optimally superimposed. In addition we computed *I-rms*, the rmsd of the backbone atoms of all interface residues after they have been optimally superimposed. Here the interface residues were defined less stringently on the basis of residue-residue contacts (see Refs. 18,19).

CAPRI participants selected their own templates and used a variety of custom built or well-established algorithms such as Modeller,[24] Swiss-Model,[25] or ROSETTA,[26] to model the subunit structures. Others used the models produced by various servers participating in the CASP11 experiment and made available to CAPRI groups, or servers of other groups (HADDOCK[27]). The quality of the CASP server models was usually first assessed using various criteria and the best quality models were selected for the docking calculations. Some groups selected a single best model for a given target, whereas others used several models (sometimes up to five models). Several groups additionally used loop modeling to adjust the conformation of loops regions, and subjected the subunit models to energy refinement.

The majority of CAPRI participants used protein docking and scoring methods to generate and rank candidate complexes. Many employed their own docking methods, some of which were designed to handle symmetric assemblies, whereas others relied on well-established docking algorithms such as HEX,[28] ZDock,[29] RosettaDock,[30] as well as on docking programs such as MZDock[31] which apply symmetry constraints.

When templates were available for a given target (mostly for homodimers), some participants used the information from these templates (consensus interface residues, contacts, or relative arrangement of subunits) to guide the docking calculations or to select docking solutions. Others used the dimeric templates directly to model the target dimer (template-based "docking"[32–34]). Less than a hand-full of groups employed template-based modeling alone for all or most of the targets.

To model tetrameric targets, most groups proceeded in two steps. They used either known dimeric homologs, or docking methods to build the dimer portion of the tetramer, and then run their docking procedures to generate a dimer-of-dimers, representing the predicted tetramer.

ASSESSMENT PROCEDURES AND CRITERIA

The standard CAPRI assessment protocol

The predicted homo and heterocomplexes were assessed by the CAPRI assessment team, using the standard CAPRI assessment protocol, which evaluates the correspondence between predicted complex and the target structure.[18,19]

This protocol (summarized in Fig. 1) first defines the set of residues common to all the submitted models and the target, so as to enable the comparison of residue-dependent quantities, such as the root mean square deviation (rmsd) of the models versus the target structure. Models where the sequence identity to the target is too

Table III
Summary of CAPRI Criteria for Ranking Predicted Complexes

	Score	*f*(nat)	L-rms		I-rms
***	High	≥ 0.5	≤ 1.0	OR	≤ 1.0
**	Medium	≥ 0.3	< 1.0–5.0]	OR	< 1.0–2.0]
*	Acceptable	≥ 0.1	< 5.0–10.0]	OR	< 2.0–4.0]
	Incorrect	< 0.1	> 10.0	AND	> 4.0

low are not assessed. The threshold is determined on a per-target basis, but is typically set to 70%.

The set of common residues is used to evaluate the two main rmsd-based quantities used in the assessment: the ligand rmsd (*L-rms*) and the interface rmsd (*I-rms*). *L-rms* is the backbone rmsd over the common set of ligand residues after a structural superposition of the receptor. *I-rms* is the backbone rmsd calculated over the common set of interface residues after a structural superposition of these residues. An interface residue is defined as such when any of its atoms (hydrogens excluded) are found within 10 Å of any of the atoms of the binding partner.

An important third quantity whereby models are assessed is *f(nat)*, representing the fraction of native contacts in the target, that is, reproduced in the model. This quantity takes all the protein residues into account. A ligand-receptor contact is defined as any pair of ligand-receptor atoms within 5 Å distance. Atomic contacts below 3 Å are considered as clashes; predictions with too many clashes are disqualified. The clash threshold varies with the target and is defined as the average number of clashes in the set of predictions plus two standard deviations. The quantities *f(nat)*, *L-rms* and *I-rms* together determine the quality of a predicted model, and based on those three parameters models are ranked into four categories: High quality, medium quality, acceptable quality and incorrect, as summarized in Table III.

Applying the CAPRI assessment protocol to homo-oligomers

Evaluating models of homo and heteroprotein complexes against the corresponding target structure is a well-defined problem when the target complex is unambiguously defined, for example, if the target association mode and corresponding interface represents the biologically relevant unit. This is usually, although not always, the case for binary heterocomplexes, but was not the situation encountered in this experiment for the homo-oligomer targets. All except two of the 25 targets for which predictions were evaluated here represent homo-oligomers. For about half of these targets the oligomeric state was deemed unreliable, as it was either only inferred computationally from the crystal structure using the PISA software[23] or because the authors' assignment and inferred oligomeric states, although available, were inconsistent (Table I). Only about 15 targets had an oligomeric state assigned by the authors at the time of the experiment.

To address this problem in the assessment, the PISA software was used to generate all the crystal contacts for each target and to compute the corresponding interface areas. The interfaces were then ranked according to size of the interface. In candidate dimer targets, submitted models were usually evaluated against 1 or 2 of the largest interfaces of the target, and acceptable or better models for any or all of these interfaces were tallied. For candidate tetramer targets, the relevant largest interfaces for each target were identified in the crystal structure, and predicted models were evaluated by comparing in turn each pair of interacting subunits in the model to each of the relevant pairs of interacting subunits in the target (Supporting Information Fig. S1), and again the best predicted interfaces were retained for the tally. One of the two bonafide heterocomplexes was also evaluated against multiple interfaces.

Evaluating the accuracy of the 3D models of individual subunits

Since this experiment was a close collaboration between CAPRI and CASP, the quality of the 3D models of individual subunits in the predicted complexes was assessed by the CASP team using the LGA program,[35] which is the basic tool for model/target comparison in CASP.[36,37] The tool can be run in two evaluation modes. In the sequence-dependent mode, the algorithm assumes that each residue in the model corresponds to a residue with the same number in the target, while in the sequence-independent mode this restriction is not applied. The program searches for optimal superimpositions between two structures at different distance cutoffs and returns two main accuracy scores; GDT_TS and LGA_S. The GDT_TS score is calculated in the sequence-dependent mode and represents the average percentage of residues that are in close proximity in two structures optimally superimposed using four selected distance cutoffs (see Ref. 38 for details). The LGA_S score is calculated in both evaluation modes and represents a weighted sum of the auxiliary LCS and GDT scores from the superimpositions built for the full set of distance cutoffs (see Ref. 35 for details). We have run the evaluation in both modes, but since the CAPRI submission format permits different residue numbering, we used the LGA_S score from the sequence-independent analysis as the main measure of the subunit accuracy assessment. This score is expressed on a scale from 0 to 100, with 100 representing a model that perfectly fits the target. The rmsd values for subunit models cited throughout the text are those computed by LGA software. We verified that for about 80% of the assessed models the GDT-TS and LGA-S scores differed by <15 units, indicating that these

models correspond to near identical structural alignments with the corresponding targets, in line with the fact that the majority of the targets of this Round represent proteins that could be readily modeled by homology. Of the remaining 20% with larger differences between the 2 scores, 18% correspond to disqualified models or incorrect complexes and 2% correspond to acceptable (or higher quality) predicted complexes. Their impact on the analysis is therefore negligible.

Building target models based on the best available templates

In order to better estimate the added value of protein docking procedures and template-based modeling techniques it seemed of interest to build a baseline against which the different approaches could be benchmarked. To this end, the best oligomeric structure template for each target available at the time of the predictions was identified. Based on this template, the target model was built using a standard modeling procedure, and the quality of this model was assessed using the CAPRI evaluation criteria described above.

To identify the templates, the protein structure database "PDB70" containing proteins of mutual sequence identity $\leq$70% was downloaded from HHsuite.[39] The database was updated twice during the experiment (See Supporting Information Table S5 for the release date of the database used for each target). Only homo-complexes were considered for this analysis.

The best available templates were detected in three different ways and target models were generated from the templates as follows: (1) *Detection based on sequence information alone*: For each target sequence, proteins related to the target were searched for in the protein structure database by HHsearch[40] in the local alignment mode with the Viterbi algorithm.[41] Among the top 100 entries, up to 10 proteins that are in the desired oligomer state were selected as templates. When more than two assembly structures with different interfaces were identified, the best ranking one was selected as template. The target and template sequences were aligned using HHalign[40] in the global alignment mode with the maximum accuracy algorithm. Based on the sequence alignments, oligomer models were built using MODELLER.[42] The model with the lowest MODELLER energy out of 10 models was selected for further analysis. (2) *Detection based on the experimental monomer structure*: Proteins with highest structural similarity to the experimental monomer structure were searched for using TM-align.[43] Among the top 100 entries, up to 10 proteins that are in the desired oligomer state were selected as templates as described above. Based on the target-template alignments output by TM-align, models were built using MODELLER, and the lowest energy model was selected as described above. (3) *Detection based on the experimental oligomer structure*: A similar procedure to those described above was applied. Although this time, the best templates were identified by searching for proteins with the highest structural similarity to the target oligomer structure. The search was performed using the multimeric structure alignment tool MM-align.[44] For computational efficiency, MM-align was applied only to the 100 proteins with the highest monomer structure similarity to the target. Models were built using MODELLER based on the alignment output by MM-align.

RESULTS

This section is divided into three parts. The first part presents the prediction results for the 25 individual targets for which the docking and scoring experiments were conducted. In the second part, we present an overview of the results across targets and across predictor and scorer groups, respectively. In the third part, we review the accuracy of the models of individual subunits in the predicted oligomers, and how this accuracy influences the performance of docking procedures.

Prediction results for individual targets

Easy homodimer targets: T69, T75, T80, T82, T84, T85, T87, T90, T91, T92, T93, T94

The 12 targets in this category comprised some of the largest subunits of the entire evaluated target set, with sizes ranging between 176 and 456 residues. Four of the targets were multi-domain proteins (T85, T87, T90, and T93), and one (T82) was an intertwined dimer.

In the following, we present examples of the performance achieved for this category of targets. Detailed results for all the targets of Round 30 can be found in the Supporting Information Table S2, and on the CAPRI website (URL: http://www.ebi.ac.uk/msd-srv/capri/).

An illustrative example of the average performance obtained for this category of targets is that obtained for target **T69 (T0764)**: a 341-residue putative esterase (BDI_1566) from *Parabacteroides distasonis*. The submitted models for this target were evaluated against two interfaces in the crystal structure of this protein, generated by applying the crystallographic symmetry operations listed in the Supporting Information Table S1, and depicted in Figure 2(a): one large interface (2415 Å^2) and a smaller interface (622 Å^2). Good prediction results were obtained only for interface 1. Twenty-eight CAPRI predictor groups submitted a total of 266 models for this homodimer. Of these, 30 were of acceptable quality and 57 were of medium quality. Twelve predictor groups and three docking servers submitted at least one model of acceptable quality or better. Among those, nine groups and one server (CLUSPRO) submitted at least 1 medium quality model. The best performance (10 medium quality

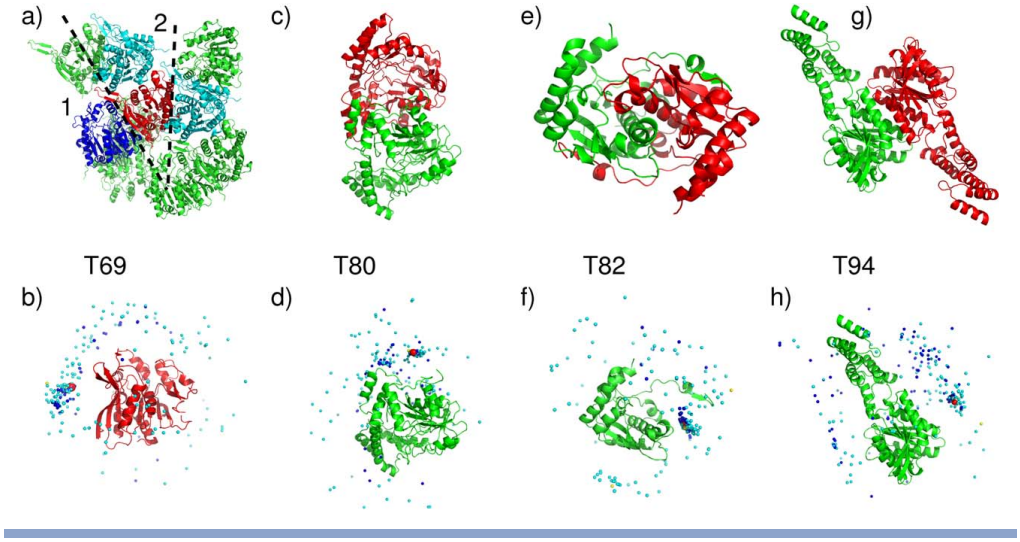


Figure 2

Target structures and prediction results for easy dimer targets. **T69 (T0764),** a Putative esterase (BDI_1566) from *Parabacteroides distasonis*, PDB code 4Q34. (**a**) Target structure, with highlighted interfaces (1,2). (**b**) Global docking prediction results displaying one subunit in cartoon representation, with the center of mass of the second subunit in the target (red sphere), and in docking solutions submitted by CAPRI predictors (light blue spheres), CAPRI scorers (dark blue spheres), and CASP predictors (yellow spheres). **T80 (T0801)**, a sugar aminotransferase WecE from Escherichia coli K-12, PDB code 4PIW. (**c**) Target structure. (**d**) Global docking prediction results by different predictor groups (see legend (b) for detail). **T82 (T0805)** Nitroreductase (structures unreleased). (**e**) Target structure. (**f**) Global docking prediction results by different predictor groups. **T94 (T0852),** uncharacterized protein Coch_1243 from *Capnocytophaga ochracea* DSM 7271, PDB code 4W9R. (**g**) Target structure. (**h**) Global docking prediction results by different predictor groups.

models) was obtained by the groups of Seok, Lee and Guerois, followed closely by the groups of Zou, Shen, and Eisenstein (see Supporting Information Table S2 for the complete ranking)

The best model for this target, obtained by Guerois, had an *f(nat)* value of 49%, and *L-rms* and *I-rms* values of 2.88 and 2.12 Å, respectively (Supporting Information Table S4).

Six groups, registered with CASP, submitted in total 12 models for this target, comprising one acceptable model by the group of Umeyama and one medium quality model by the Baker group. The global landscape of all the predicted models by the different groups is outlined in Figure 2(b).

An even better performance was achieved by the CAPRI scoring experiment (Supporting Information Table S2). Of the 14 groups participating in this experiment, 12 submitted at least two models of medium quality. The best performance was achieved by Kihara (10 medium quality models), closely followed by Zou and Grudinin, with eight and five medium quality models, respectively. As already observed in previous CAPRI evaluations the best performers in the docking calculations were not necessarily performing as well in the scoring experiment, and thus not singling out even their own best models from the uploaded anonymized set of predicted complexes, highlighting yet again the distinct nature of the docking and scoring procedures.

An important factor in the successful predictions was the overall good accuracy of the 3D models used by predictors in the docking calculations (see Fig. 6 and CAPRI website for detailed values). The best models had an LGA_S score of ∼85 (backbone rmsd of ∼3.9 Å), and only a few models had LGA_S scores lower than 40 (backbone rmsd > 10 Å) (values for all models are available on the CAPRI website). The accuracy of the 3D models across targets and its influence on the predictions will be discussed in a dedicated section below.

Very good predictions were obtained for **T82 (T0805)**, the nitro-reductase rv3368, a significantly intertwined dimer with unstructured arms reaching out to the neighboring subunit and a subunit interface area of 3250 $Å^2$ [Fig. 2(e,f)]. The majority of the models of the individual subunits were quite accurate with LGA_S values of 60–85 (backbone rmsd <5 Å) (see CAPRI website). As many as 54 medium quality models and 17 acceptable models were submitted by CAPRI participants, 99 models of acceptable quality or better were submitted by

CAPRI scorer groups, and 11 acceptable models or better were submitted by three CASP groups (Supporting Information Table S2). The high success rate for both complex predictions and subunit modeling stems from the fact that most predictors made good use of known structures of related homodimers in the PDB in which the intertwining mode was well conserved. These known dimer structures were mainly used in templates for modeling the target dimer (template-based docking).

Very similar participation, number of submitted models and performance, was featured in docking predictions for the other targets in this category (see Supporting Information Tables S2 and S3). The models of individual subunits were also of similar accuracy or higher.

Excellent performance was obtained for targets **T80 (T0819)** and **T93 (T0851)** with >100 correct models of which ~70 were of medium quality, followed by targets **T90 (T0843)** and **T91 (T0847)**, for which >100 correct models, comprising ~40 medium quality ones- were submitted. These targets featured subunits sizes of 176–456 residues.

T80 (T0801) was the sugar aminotransferase WecE from E.coli K-12, with 376 residues per subunit. Submitted models were evaluated against one interface (1960 Å^2) between the two subunits of the crystal asymmetric unit [Fig. 2(c)]. A total of 27 CAPRI predictor groups submitted 105 models of acceptable quality or better. The majority of these (71 models) were of medium quality. 12 CAPRI groups participated in the scoring experiment and submitted 120 models, of which about half (51) were of medium quality and 14 were acceptable models. Six CASP participants submitted 11 medium quality models, and two models of acceptable quality. The top ranking CAPRI predictor groups for this target were those of Sali, Guerois, and Eisenstein who submitted 10 medium quality models each. These three groups were closely followed by the groups of Seok, Zou, Shen and Lee, each of whom predicted at least five medium quality models. Each of the three participating servers, HADDOCK, GRAMM-X, and CLUSPRO, submitted at least one acceptable model. The best performers from among the scorer groups were those of Zou and Huang with 10 medium quality models each, followed by Gray, Kihara and Weng with at least 5 medium quality models, and by Fernandez-Recio and Bates with four medium quality models. The global landscape of the predictions for this target is shown in Figure 2(d).

The subunit models for this target were of very high quality, with the best models featuring a LGA_S score of ~95 and a backbone rmsd of 1.3 Å. The quality of the best models for targets T90 and T91 for which a similarly high performance was achieved was only somewhat lower, with LGA_S values of 70–88 and backbone rmsd of 2.0–5.0 Å.

Interestingly, **T91 (T0847),** the human Bj-Tsa-9, was predicted to be a dimer by PISA, but assigned as a monomer by the authors. The good docking performance for this target and the fact that the dimer interface (1320 Å^2) is within the range expected for proteins of this size (176 residues),[45] suggests that this protein forms a dimer.

A somewhat lower performance was achieved for **T92 (T0849)** the glutathione S-transferase domain from *Haliangium ochraceum*), and for **T94 (T0852)**, an uncharacterized 2-domain protein (putative esterase according to Pfam) Coch_1243 from *Capnocytophaga ochracea*. A total of 98 acceptable models were submitted for T92, of which only 12 were of medium quality, but the models were contributed by a large fraction of the participating groups (17 out of 23). On the other hand, the scorer performance was very good with 68 acceptable models of which almost half (33) were of medium quality. These models were contributed across most scorer groups (10 out of 11). CASP participants achieved a particularly good performance. Of the 23 models submitted by CASP groups, 17 were of acceptable quality or better, and those were contributed by six of the seven participating groups. The accuracy of the subunit models was in general lower, with LGA_S ~70 and rmsd ~7 Å for the best models, and LGA_S values of 50 – 60 for most other models.

In T94, predicted complexes were assessed only against the largest interface (1190 Å^2), formed between large domains of the adjacent subunits, as the second largest interface was much smaller (620 Å^2). In total, 97 acceptable homodimer models only, were contributed for this target: 58 models by CAPRI predictors, 37 by CAPRI scorers, and 2 by CASP groups [see Supplementary Table S2, and Fig. 2(g,h) for a pictorial summary]. The lower accuracy of the subunit models for this target (LGA_S score ~58 and rmsd >6 Å, for the best model) may have limited the accuracy of the modeled complexes, without however compromising the task of achieving correct solutions.

Difficult or problematic homodimer targets: T68, T72, T77, T79, T86, T88

This category comprises 6 targets, representing particular challenges to docking calculations for reason inherent to the proteins involved, or targets for which the oligomeric state was probably assigned incorrectly at the time of the experiment.

With the exception of T72, targets in this category are much smaller proteins, than those of the easy dimer targets (Table I). In three of the targets (T68, T79, T86) the largest interface area between subunits in the crystal is small (470–860 Å^2) and their oligomeric state assignments were often ambiguous. In the following, we comment on the insights gained from the results obtained for several of these targets.

No acceptable homodimer models were contributed by CAPRI or CASP groups for targets T68, T77 and T88.

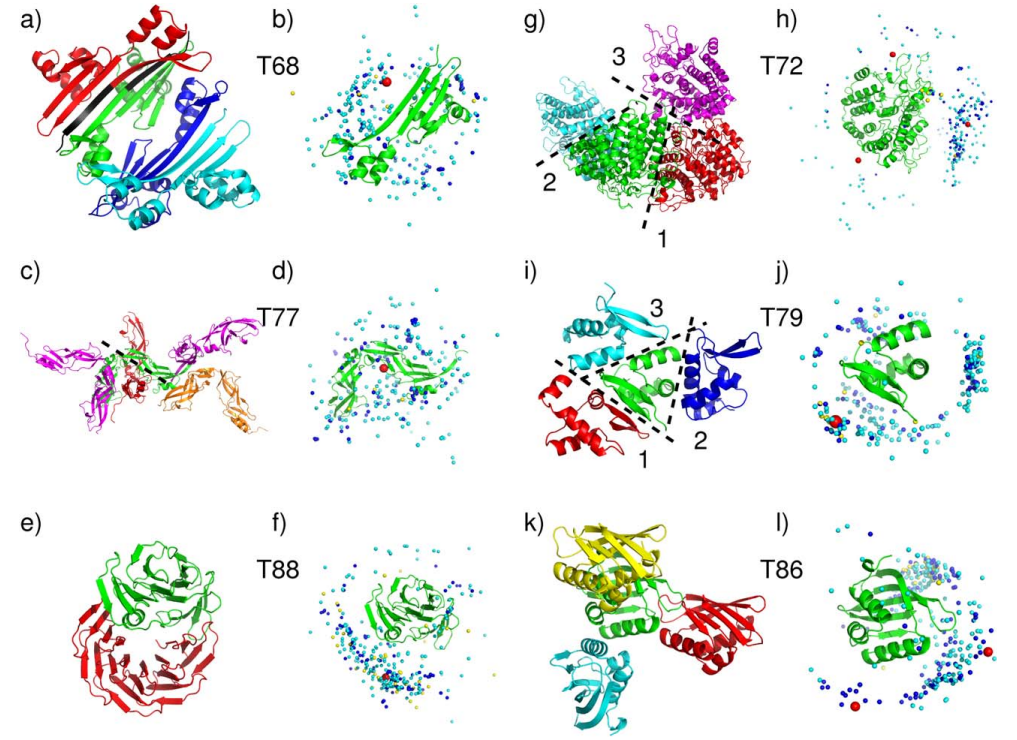


Figure 3

Target structures and prediction results for difficult or problematic dimer targets. **T68 (T0759)**, Plectin 1 and 2 Repeats of the Human Periplakin, PDB code 4Q28. (**a**) Target structure in cartoon representation, displaying 4 subunits in the crystal. The His-Tag sequence, highlighted in black, mediates contacts at the largest interface. (**b**) Global docking prediction results displaying one subunit in cartoon representation, with the center of mass of the second subunit in the target (red sphere), and in docking solutions submitted by CAPRI predictors (light blue spheres), CAPRI scorers (dark blue spheres), and CASP predictors (yellow spheres). **T77 (T0780),** conserved hypothetical protein (SP_1560), *Streptococcus pneumoniae TIGR4* PDB code 4QDY. (**c**) Target structure, highlighting the assessed interface (dashed line). (**d**) Global docking prediction results by different predictor groups (see legend (b) for detail). **T88 (T0825)**, synthetic wrap five protein (structure unreleased). (**e**) Target structure. (**f**) Global docking prediction results by different predictor groups. **T72 (T0772)**, SusD homolog (BT2259) from *Bacteroides thetaiotaomicron* VPI-5482, PDB code 4Q69. (**g**) Target structure, highlighting the three assessed interfaces. (**h**) Global docking prediction results for the three interfaces, by different predictor groups. **T79 (T0792)**, OSKAR-N, PDB code 5a49. (**i**) Target structure, highlighting the three assessed interfaces. (**j**) Global docking prediction results for the three interfaces by different predictor groups. **T86 (T0815)** Putative polyketide cyclase (protein SMa1630) from *Sinorhizobium meliloti*, PDB code 4U13. (**k**) Target structure, showing three interfaces. (**l**) Global docking prediction results for the two interfaces by different predictor groups (the interface with the yellow monomer was not assessed).

The main problem with **T68 (T0759)**, the plectin 1 and 2 repeats of the Human Periplakin, was that the crystal structure contains an artificial N-terminal peptide representing the His-tag (MGHHHHHHS…) that was used for protein purification. The N-terminal segments of neighboring subunits, which contain the artificial peptide, associate to form the largest interface between the subunits in the crystal (1150 Å^2) [Fig. 3(a)]. Submitted model were assessed against this interface and the second largest interface (860Å^2), but not against the 2 much smaller interfaces (240 and 160 Å^2).

Most predictor groups (from both CASP and CAPRI) carried out docking calculations without the His-tag, which they assumed was irrelevant to dimer formation *in-vivo*. They were therefore unable to obtain docking solutions that were sufficiently close to the largest interface of the target [Fig. 3(b)]. As well, no acceptable solutions were obtained for second largest interfaces,

indicating that it too was unlikely to represent a stable homodimer.

The quality of the subunit models was also lower than for many other targets (the best model had an LGA_S score of ~57), as most groups ignored the His-Tag in building the models as well (see Fig. 6 and CAPRI website for details). Considering that the His-Tag containing peptide contributes significantly to the largest subunit interface, the protein is likely a monomer in absence of the artificial peptide. This is in fact the authors' assignment in the corresponding PDB entry (4Q28), and in retrospect this target should not have been considered for the CAPRI docking experiments.

Different factors contributed to the failure of producing acceptable docking solution for **T77 (T0780)**, the conserved hypothetical protein (SP-1560), from *Streptococcus pneumonia* TGR4 [Fig. 3(c,d)]. The protein consists of two YbbR-like structural domains (according to Pfam) arranged in a crescent-like shape. The domains adopt rather twisted β-sheet conformations with extensive stretches of coil, and are connected by a single polypeptide segment, suggesting that the protein displays an appreciable degree of flexibility both within and between the domains. Probably as a consequence of this flexibility, the structures of most templates identified by predictor groups (which approximated only one domain), were not close enough to that of the target (Supporting Information Table S5). As a result, the subunit models were generally quite poor, with the best model featuring an LGS-A score of only ~40 (rmsd ~7 Å). Although the largest interface of the target is of a respectable size (1600 Å^2) and involves intermolecular contacts between one of the domains only, the docking calculations were unable to identify it. The best docking model was incorrect as it displayed an *L-rms* ~19 Å, and an *I-rms* ~10 Å (see Supporting Information Table S4).

A very different issue plagued the docking prediction of **T88 (T0825)**, the wrap5 protein. The information given to predictors was that the protein is a synthetic construct built from 5 sequence repeats, and is similar to 2YMU (a highly repetitive propeller structure). It was furthermore stated that the polypeptide has been mildly proteolyzed, yielding two slightly different subunits, in which the N-terminus of the first repeat was truncated to different extent, and that therefore the dimer forms in a non-trivial way. Predictors were given the amino acid sequence of the two alternatively truncated polypeptides.

It turned out that the longer of the two chains, with the nearly intact first repeat forms the expected 5-blade β-propeller fold, whereas the chain with the severely truncated first repeat forms only four of the blades, with the remainder of the first repeat forming an α-helical segment that contacts the first repeat [Fig. 3(e)].

Both CAPRI and CASP predictor groups were quite successful in building very accurate models for the less truncated subunit (rmsd < 0.5 Å, LGA_S ~90). But subunit models for the more truncated subunit were much poorer (rmsd 6.5–10 Å), and since the helical region of the shorter subunit contributes significantly to the dimer interface, whose total area is not very large (~1300 Å^2), no acceptable docking solutions were obtained [Fig. 3(e,f)].

For the other three targets in this category, T72, T79, and T86, the homodimer prediction performance remained rather poor, with only very few acceptable models submitted. The main issue with **T79 (T0792)**, the OSKAR-N protein, and **T86 (T0815)**, the polyketide Cyclase from *Sinorhizobium meliloti*, was likely their very small subunit interface (Table I). T79 was predicted by PISA to be a dimer, but the area of its largest subunit interface is only 680 Å^2. T86, predicted to be dimeric by both PISA and the authors (as stated in the PDB entry, 4U13), has even smaller size subunit interfaces with the largest one burying no $>$470 Å^2. In both cases these interfaces are much smaller than the average size required in order to stabilize weak homodimers.[46] It is therefore likely that these two proteins are in fact monomeric at physiological concentrations. Furthermore, T79 and T86 are quite small proteins (80 residues for T79, and 100 residues for T86), and it is not uncommon that proteins of this size cannot form large enough interfaces unless they are intertwined.[47]

This notwithstanding, a few acceptable homodimer models were contributed for all three assessed interfaces (interfaces 1,2,3) of T79 (Supporting Information Table S2).

Among predictor groups, 17 acceptable docking solutions (of which five were medium quality models) were obtained for the largest interface (interface 1). Twelve acceptable solutions, of which one medium quality one, were obtained for the second smaller interface (440 Å^2), and no acceptable quality solutions were obtained for the third assessed interface (400 Å^2) [see Fig. 3(i,j) for an overview of the prediction results]. Seven CAPRI predictor groups, 1 CASP group and one server (GRAMM-X) contributed the correct models for interface 1, and seven CAPRI groups submitted acceptable models for interface 2.

Interestingly scorers did less well than predictors for interface 1, but better for interface 2, and two scorer groups submitted two acceptable models for interface 3, whereas none were submitted by predictor groups.

Overall, the models for the T79 subunit were quite accurate, with the best model having and LGA_S score of ~89 and rmsd ~1.9 Å.

Not too surprisingly, the dimer prediction performance for T86 was significantly poorer, with only three acceptable models submitted by CAPRI predictors (Ritchie and Negi) for the largest interface (470 Å^2). Scorers identified five acceptable models for interface 1 (Fernandez-Recio and Gray), and two acceptable (or better) models for interface 2 (Seok and Kihara). None of the 19 models submitted by the seven CASP groups were correct [Fig. 3(k,l) for a pictorial summary].

Different problems likely led to the weak prediction performance for Target **T72 (T0770)**, the SusD homolog (BT2259) from *Bacteroides Thetaiotaomicron*. While the largest subunit interface is of near average size (1120 Å^2), the interface itself is poorly packed and patchy, an indication that it may not represent a specific association. Not too surprisingly, therefore, this led to a poor prediction performance. Overall only three models of acceptable quality were submitted by CAPRI dockers, namely by the HADDOCK and SWARMDOCK servers, and the Guerois group, each contributing 1 such model. The best of these models (contributed by Guerois) had *f(nat)* ~29% and *L-rms* and *I-rms* values of 8.85 and 3.57 Å, respectively. Seven acceptable models were submitted by scorers. Bonvin contributed two models, and the groups of Huang, Grudinin, Gray, Weng and Fernandez-Recio, respectively, submitted one model. The best quality models had *f(nat)* ~18%, and *L-rms* and *I-rms* values of ~7.29 and 4.28 Å, respectively. No acceptable models were submitted by CASP participants. The target structure and the distribution of the all the docking solutions are depicted in Figure 3(g,h).

The accuracy of the subunit models for T72 was reasonable, with the best models having a LGA_S score of ~70 (backbone rmsd ~3.8 Å). The three successful CAPRI predictor groups (HADDOCK, SWARMDOCK and Guerois) all had somewhat lower quality subunit models with LGA_S scores in the range of 55 – 67.

Targets assigned as tetramers: T70, T71, T73, T74, T78

Five targets were assigned as tetramers at the time of the prediction experiment. As described in Assessment Procedure and Criteria, models for tetramer targets were assessed by systematically comparing all the interfaces in each model to all the relevant interfaces in the target, and selecting the best-predicted interfaces. Most predictor groups used a two-step approach to build their models. First they derived the model of the most likely dimer, and then docked the dimers to one another. Some groups imposed symmetry restraints as part of the docking procedures, or combined this approach with the two-step procedure.

In three of the targets (T70, T71, T74) predictors faced the problem that all the pair-wise subunit interfaces were quite small (440–720 Å^2), making it difficult to identify stable dimers to initiate the assembly procedure.

T70 (T0765), the modulator protein MzrA from *Klebsiella Pneumoniae* Sub Species, was assigned as a tetramer at the time of the predictions, but is listed as a dimer (predicted by PISA and assigned by the authors) in the PDB entry (4PWU). Only two of its interfaces in the crystal bury an area exceeding 400 Å^2 [Fig. 4(a)]. The assembly built by propagating these two interfaces appears to form an extensive layered arrangement across unit cells in the crystal, rather than a closed tetramer.

Interestingly, acceptable or better models were submitted only for the smaller interface (475 Å^2) (Supporting Information Table S2). CAPRI predictors submitted 37 acceptable models, of which 27 were of medium quality, and scorers submitted 27 acceptable models (including 21 medium quality ones) [Fig. 4(b)]. Indeed no acceptable models were submitted for the largest interface (560 Å^2), which is assigned as the dimer interface in the PDB entry for this protein.

The failure to model a higher order oligomer for this target was not due to the quality of the subunit models as the latter was quite high (see Fig. 6 and CAPRI website), and is probably rooted in the pattern of contacts made by the protein in the crystal, which suggest that this target is likely a weak dimer. Considering that all the acceptable docking models involve a different interface than that assigned in the corresponding PDB entry, it is furthermore possible that the interface identified in these solutions is in fact the correct one. But given the very small size of either interface, the protein could also be monomeric.

A similar situation was encountered with **T74 (T0774)**, a hypothetical protein from *Bacteroides vulgatus*. Here too the target was assigned as a tetramer by PISA at the time of the predictions, but is listed as a monomer by the authors in the PDB entry (4QB7). Associating the subunits according to the two largest interfaces (520 and 490 Å^2), also produced an open-ended assembly rather than a closed tetramer, and this time no acceptable solutions were produced for either interface, strongly suggesting that the protein is monomeric as specified by the authors. It is noteworthy that the subunit models for this target were particularly poor (LGA_S values ~40, and rmsd ~7 Å), which could also have hampered identifying some of the binding interfaces.

T71 (T0768), the leucine-rich repeat protein from *bacteroides capillosus*, was a difficult case for other reasons. Subunit contacts in the crystal are mediated through three different interfaces, ranging in size from 470 Å^2 to 720 Å^2. A closed tetrameric assembly can be built by combining interfaces 1 and 3, associating the dimer formed by subunits A and B with the equivalent dimer of subunits C and D, as shown in Figure 4(c). Interfaces 1 and 3 were also those for which some acceptable predictions were submitted. One acceptable model was contributed for the largest interface, by the GRAMM-X, an automatic server. Eleven acceptable models were submitted for the third interface (470 Å^2) by 4 CAPRI predictor groups, and six acceptable models were submitted by four CAPRI scorer groups. All the models submitted by a single CASP group were wrong. No group succeeded in building the tetramer that comprises the correct models for interfaces 1 and 3 at the same time. Some models looked promising, but when superimposing equivalent subunits (in the model vs. the target) the neighboring

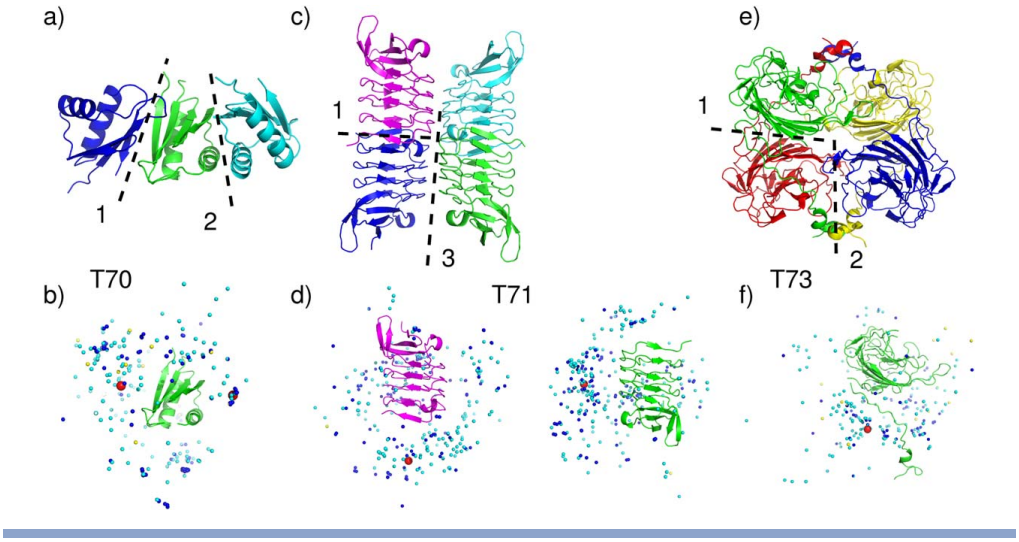


Figure 4

Target structures and prediction results for tetrameric targets. **T70 (T0765)**, Modulator protein MzrA (KPN_03524) from *Klebsiella pneumoniae* subspecies. (**a**) Target structure in cartoon representation, highlighting the two assessed interfaces (dashed lines). (**b**) Global docking prediction results displaying one subunit in cartoon representation, with the center of mass of the second subunit in the target (red spheres), and in docking solutions submitted by CAPRI predictors (light blue spheres), CAPRI scorers (dark blue spheres), and CASP predictors (yellow spheres). **T71 (T0768)** Leucine-rich repeat protein (BACCAP_00569) from *Bacteroides capillosu*, PDB code 4QJU. (**c**) Target structure in cartoon representation, highlighting the two relevant interfaces (interfaces 1 and 3) (dashed lines). (**d**) Global docking prediction results for the assessed interfaces by different predictor groups (monomer color corresponding to (c), that is, the red spheres represent the same, blue, monomer). **T73 (T0772),** Putative glycosyl hydrolase, PDB code 4QHZ. (**e**) Target structure in cartoon representation, highlighting the two assessed interfaces (interface 1 and 2) (dashed lines). (**f**) Global docking prediction results for the assessed interfaces by different predictor groups.

subunit of the model (the one across the incorrectly predicted interface) had its position significantly shifted relative to that in the target, resulting in an incorrect structure of the tetrameric assembly.

The remaining two targets, **T73 (T0772)**, a putative glycosyl hydrolase from *Parabacteroides distaspnos,* and **T78 (T0786),** a hypothetical protein from *Bacillus cereus*, were genuine tetramers assigned as such by both PISA and the authors. Both targets are proteins of similar size (~260 residues) adopting an assembly with classical D_2 symmetry, which comprises two interfaces, a sizable one (>1000 Å^2) and a smaller one. But the main bottleneck for both targets was that their larger interface was intertwined. Available templates did not seem to capture the intertwined associations, as witnessed from the overall poorer models derived for the individual subunits. For both targets, the best models had an LGA_S score ~50 and a backbone rmsd of ~5–10 Å. For T73, a total of only nine acceptable models were submitted by the CAPRI predictor groups of LZERD, Zou and Kihara for the largest interface, and two acceptable models were submitted by the Lee group for the second interface. None of the predicted tetramer models simultaneously captured both interfaces, as illustrated in Figure 4(e,f). For T78, no acceptable solutions were submitted by any of the participating groups, but the subunit models were only marginally more accurate than those of T73.

The conclusions to be reached from the analysis of these five targets are twofold. One is that the oligomeric state assignment for higher order assemblies such as tetramers is more error prone than that of dimer versus monomers. Tetramers often involve smaller interfaces between subunits, especially those formed between individual proteins when two dimers associate, and therefore predictions on the basis of pair-wise crystal contacts such as those by PISA become unreliable. Independent experimental evidence is therefore required to validate the existence of a higher order assembly. The second conclusion to be drawn is that the prediction of higher order assembly by docking procedures remains a challenge. Acceptable models derived for the largest dimer interface are probably not accurate enough to enable the identification of stable association modes between two modeled dimers. This indicates in turn that the propagation of

errors is the problem that currently hampers the modeling of higher order assemblies from the structures of its components in absence of additional experimental information.

Heterocomplex targets: T81, T89

T81 (T0797/T0798) and **T89 (T0840/T0841)** were the only two *bona-fide* heterocomplex targets in Round 30. T81 is the complex between the cGMP-dependent protein Kinase II leucine zipper (44 residues) and the Rab11b protein (198 residues) (PDB code 4OJK). T89 is the complex between the much larger RON receptor tyrosine kinase subunit (669 residues) and the macrophage stimulating protein subunit (MSP) (253 residues).

The crystal structure of T81 features two Rab11b proteins binding on opposite sides of the centrally located leucine zipper, in a quasi-symmetric arrangement, which likely represents the stoichiometry of the biological unit [Fig. 5(a)]. A total of 3 interfaces were evaluated for this targets: Interface 1 (chains C:A, leucine zipper helix 1/one copy of the Rab11b protein), Interface 2 (C:D, leucine zipper helix 1/helix 2), interface 3 (equivalent to interface 1). The two Rab11b/zipper helix interfaces were not exactly identical (780 Å^2 for interface 1 and 630 Å^2 for interface 2). The interface between the helices of the leucine zipper was somewhat larger (780 Å^2). Overall, the interface area of a single copy of the Rab11b protein binding to the leucine zipper dimer measures 1070 Å^2.

Consolidating correct predictions for the equivalent interfaces (Interfaces 1 and 3), the prediction performance for this complex as a whole was disappointing. Only 12 correct models were submitted by the 7 CAPRI predictor groups of Guerois, Seok, Huang, Vajda/Kozakov, SWARMDOCK, CLUSPRO (a server) and Bates. Five of those (submitted by Guerois, Seok and Huang) were of medium quality. The performance of CAPRI scorers was better, with 54 correct models of which 16 of medium quality. All 11 scorer groups contributed these models, and the best scorer performance was achieved by the groups of Bates, followed by those of LZERD, Oliva, Huang, Fernandez-Recio and Seok. The prediction landscape for this target is shown in Figure 5(b).

T89, the RON receptor kinase subunit complex with MSP, was a simpler target, given the clear, binary character of this heterocomplex. But the large size of the receptor subunit, and the relatively small interface it formed with MSP, represented a challenge for the docking calculations. The prediction performance for this complex was quite good overall, with a total of 87 correct models submitted by predictors, representing 41% of all submitted predictor models. Unlike for many other targets of this round, scorers did only marginally better, with 42% of correct models. CASP groups were specifically invited to submit models for this target, and 55 groups did, nearly ten times more than for other targets in this round. But their performance was much poorer than that of CAPRI groups. Only 23 models out of the 223 submitted by CASP groups (10%) were correct, and 6 of these were medium accuracy models.

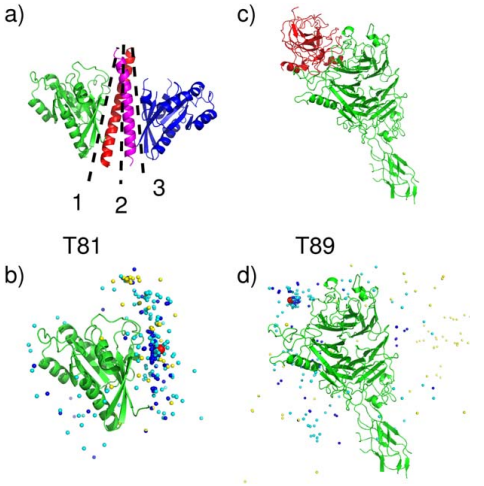


Figure 5

Target structures and prediction results for heterocomplex targets. **T81 (T0797/T0798)**, cGMP-dependent Protein Kinase II Leucin Zipper and Rab11b Protein Complex, PDB code 4OJK. (**a**) Target structure in cartoon representation, highlighting the interface of the leucine zipper dimer (2), and the two equivalent interfaces (1,3), between the zipper dimer and the two Rab11b proteins (dashed lines). (**b**) Global docking prediction results displaying one of the Rab11b subunits in cartoon representation, with the center of mass of the leucine zipper dimer in the target (red sphere), and in docking solutions submitted by CAPRI predictors (light blue spheres), CAPRI scorers (dark blue spheres), and CASP predictors (yellow spheres). **T89 (T0840/T0841)**, complex of the RON receptor tyrosine kinase subunit and the macrophage stimulating protein subunit (MSP) (structure not released). (**c**) Target structure in cartoon representation. (**d**) Global docking prediction results displaying the RON receptor kinas subunit, in cartoon representations, and the center of mass of the MCP proteins in the target and in docking solutions submitted by different predictor groups.

The best performance among CAPRI predictor groups was by the HADDOCK server, followed by the groups of Vakser, Seok, Guerois, Grudinin, Lee, Huang and Tomii (see Supporting Information Table S2). A pictorial summary of the prediction performance for this target is provided in Figure 5(c,d).

Results across targets and groups

Across target performance of CAPRI docking predictions

Results of the docking and scoring predictions for the 25 assessed targets of Round 30, obtained by all groups

that submitted models for at least one target, are summarized in Figure 6 and in the Supporting Information Table S3. For a full account of the results for this Round the reader is referred to the CAPRI web site (http://www.ebi.ac.uk/msd-srv/capri/).

The results summarized in Figure 6 show clearly that the prediction performance varies significantly for targets in the four different categories. As expected, the performance is significantly better for the 12 dimer targets in the "easy" category, than for those in the other categories. For 10 of the 12 "easy" targets, at least 30% of the submitted models per target are of acceptable quality or better, and for most of these (eight out of 10), at least 20% of the models are of medium quality. The accuracy of the subunit models (top panel, Fig. 6) is rather good for most of these targets. With the exception of T93, for which the quality of the subunits models spans a wide range (LGA_S ~40–80), the models of the remaining 11 targets achieve high LGA_S scores with averages of 80 or above.

The two less well-predicted targets in this category are T92 and T94, probably due to the lower quality of the subunit models (average LGA_S < 60) (top panel, Fig. 6).

The docking prediction performance is quite poor for the six "difficult or problematic" dimer targets, where a few acceptable models were submitted for only three of the targets (T72, T79, T86), and no acceptable models were submitted for the remaining three targets. This very poor performance was not rooted in the docking or modeling procedures but rather in the targets themselves. In 4 of the targets in this category (T68, T72, T79, T86) the oligomeric state (dimer in this case), often predicted only by PISA, but sometimes also provided by the authors, was likely incorrectly assigned. In T68, the His-tag used for protein purification and included in the crystallization forms the observed dimer interface, which is therefore most certainly non-native. In T72 the main problem was its very poorly packed and patchy interface, suggesting that the dimer might be a crystal artifact, whereas in T79 and T86, all the pair-wise interfaces in the crystal structure were too small for any of them to represent a stable dimer.

The only genuinely difficult dimer targets were T77 and T88. For T77, the subunits of this flexible 2-domain protein were rather poorly modeled (average LGA_S 30–40), making it difficult to model the "handshake" arrangement of the subunits in the dimer [Fig. 3(c,d)]. In T88, the synthetic wrap5 protein, most predictor groups failed to meet the challenge of correctly modeling the shorter of the two subunits, in turn leading to incorrect solutions for the heterodimer.

As already mentioned, a very poor performance was achieved for the five targets assigned as tetramers at the time of the predictions. This is illustrated at the level of the individual interfaces in these targets (Fig. 6). However, here too the problem was not necessarily rooted in limitations of the docking or modeling procedures. Two of the targets, T70 and T74, seem to have been erroneously assigned as tetramers at the time of the prediction by PISA, as described above. T70 was assigned as a dimer, and T74 as a monomer, by the respective authors in the PDB entry. In agreement with the authors' assignment, no acceptable solutions were identified for any of the interfaces in T74. Somewhat surprisingly, the quality of the subunits models for this target was particularly poor as well (average LGA_S ~30).

In T70, the docking calculations were able to identify only the smaller of the two interfaces as forming the dimer interface (Fig. 6), but this interface differs from the one assigned by the authors. This result leaves open the possibility that this protein may indeed be a weak dimer, in agreement with the author's assignment, albeit a different dimer than the one that they propose. Thus for both of these seemingly erroneously assigned tetrameters, the docking calculations actually gave the correct answer, which supports the author's subsequent assignments, which were not made available at the time of the prediction experiment.

For the other three tetrameric targets, T71, T73 and T78, the poor interface prediction performance reflects the genuine challenges of modeling higher order oligomers. In T71 the small size of the individual interfaces was likely the reason for the paucity of acceptable dimer models, and those were moreover not accurate enough to enable the correct modeling of the higher order assembly (dimer of dimers). In T73 and T78, the very few acceptable models for interfaces in the former, and the complete failure to model any of the interfaces in the latter (Fig. 6), likely stem from the lower accuracy of the corresponding subunit models (average LGA_S ~50–60).

The docking prediction performance was better, but not particularly impressive for the two heterocomplex targets T81 and T89, which represent the type of targets that the CAPRI community commonly deals with. For T81 only ~5% of the submitted models were of acceptable quality or better, whereas for T89 the corresponding model fraction was 40%, similar to that achieved for the easy dimer targets. The poorer performance for T81 can be readily explained by the fact that this target was in fact a hetero tetramer, two copies of the Rab11b protein binding to opposite sides of a leucine zipper, which had to be modeled first.

These results taken together indicate that homology modeling techniques and docking calculations are able to predict rather well the structures of biologically relevant homodimers. In addition we see that the prediction performance for such targets is on average superior than that obtained for heterocomplexes in previous CAPRI rounds, where on average only about 10–15% of the submitted models are correct for any given target (http://onlinelibrary.wiley.com/doi/10.1002/9781118889886.ch4/summary), compared to 25% obtained for the majority of the genuine

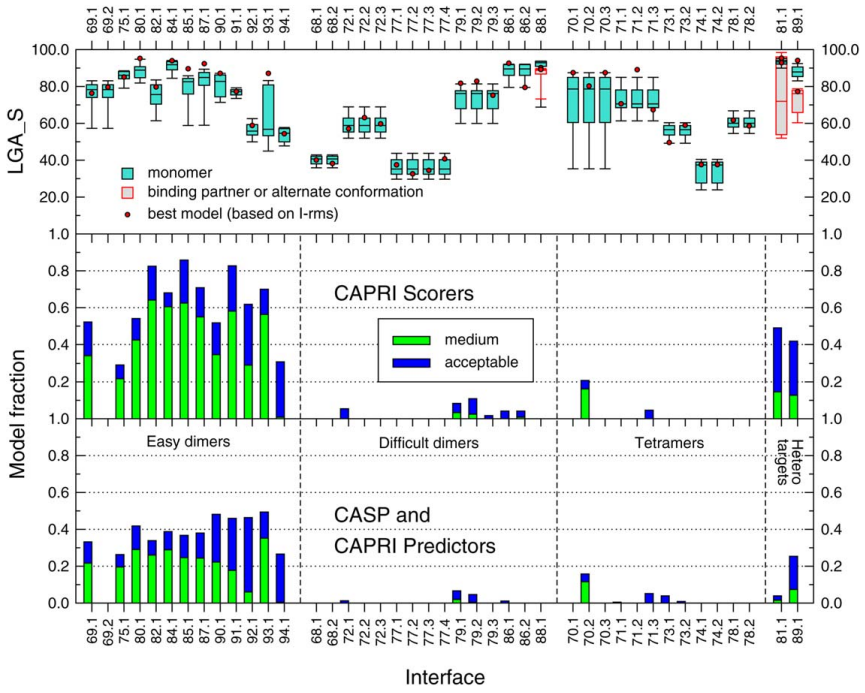


Figure 6

Pictorial summary of the prediction results per assessed interface of the targets in CAPRI Round 30. The lower panel depicts the fraction of models of acceptable and medium quality respectively, submitted by CAPRI and CASP predictor groups, for the 42 assessed interfaces in all 25 targets (listed along the horizontal axis). The digit following the CAPRI target number represents the assessed interface. The symmetry transformation corresponding to the assessed interfaces in each target are listed in the Supporting Information Table S1. The fraction of correct models is shown separately for the four main target categories: Easy dimer targets, difficult (or problematic) dimer targets, tetrameric targets, and heterocomplex targets. The middle panel displays the same data for models submitted for the same interfaces by CAPRI scorer groups. The top panel shows box plots of the LGA_S score values of the subunits in submitted models for the targets listed along the horizontal axis. The LGA_S score is one of the CASP measures of the accuracy of the predicted 3D structure of a protein.[35] The red dots represent the LGA_S score of the subunit structure of the best quality homo or heterocomplex model submitted for each target. The best quality model is defined as the one with the lowest *I-rms* (see **Fig.** 1 for details).

dimer targets in this Round, including both easy and difficult homodimers. This result is not surprising, as interfaces of homodimers are in general larger and more hydrophobic than those of heterocomplexes,[45] properties which should make them easier to predict.

Another noteworthy observation is that docking calculations can often help to more reliably assign the protein oligomeric state, especially in cases where available assignments were ambiguous. Such cases were encountered for several of the difficult or problematic targets, and for targets assigned as tetramers. On the other hand, the main challenge in correctly modeling tetramers is to minimize the propagation of errors caused by even small inaccuracies in modeling individual interfaces, which can in turn be exacerbated by inaccurate 3D models of the protein components.

Across target performance of CAPRI scoring predictions

As shown in the middle panel of Figure 6, CAPRI scorer groups achieved overall a better prediction performance than predictor groups. The scoring experiment involves no docking calculations, and only requires singling out correct solutions from among the ensemble of models uploaded by groups participating in the docking predictions. Clearly, such solutions cannot be identified if the ensemble of uploaded models contains only incorrect solutions. Therefore no correct scoring solutions were submitted by scorers for targets where no acceptable docking solutions were present within the 100 models uploaded by predictor groups for given target.

However, for targets where at least a few correct docking models were obtained by predictors, scorers were

often able to identify a good fraction of these models, as well as other models that were not identified amongst the 10 best models by the groups that submitted them (Fig. 6). This was particularly apparent for the easy dimer targets, where scorers often submitted a significantly higher fraction of acceptable-or-better models (>50%) than in the docking experiment, where this fraction rarely exceeded 40%. A similar result was achieved for the heterocomplexes, and was particularly impressive for T81, where nearly half of the submitted models by scorers were correct, compared to only 5% for the docking predictions.

The seemingly superior performance of scorers over dockers has been observed in previous CAPRI assessments[16,19] where it was attributed in part to the generally poor ranking of models by predictors. Their highest-ranking models are often not the highest-quality models, and acceptable or better models can often be found lower down the list and amongst the 100 uploaded models. Another reason is the fact that the search space that scorers have to deal with is orders of magnitude smaller (a few thousands of models), than the search space dockers commonly sample (tens of millions of models). This significantly increases the odds of singling correct solutions in the scoring experiment.

Clearly however, there is more to the scorers' performance than chance alone, particularly in this CAPRI Round, where the main challenge was to model homo-oligomers. Some groups that have also implemented docking servers had their server perform the docking predictions completely automatically, but carried out the scoring predictions in a manual mode, which still tends to be more robust. In addition, a meta-analysis of the uploaded models, such as clustering similar docking solutions and selecting and refining solutions from the most populate clusters can also lead to improved performance.

This notwithstanding, the actual scoring functions used by scorer groups must play a crucial role. But this role is currently difficult to quantify in the context of this assessment.

Performance across CAPRI and CASP predictors, scorers and servers

The ranking of CAPRI-CASP11 participants by their prediction performance on the 25 targets of Round 30 is summarized in Table IV. The per-target ranking and performance of participants can be found in the Supporting Information Tables S2 and S3.

The ranking in Table IV considers only the best quality model submitted by each group for every target. The ranking in the Supporting Information Table S2 takes into account both the total number of acceptable models, and the number of higher quality models (medium quality ones for this Round, as detailed in the section on assessment criteria). When two groups submitted the same number of acceptable models, the one with more high quality models is ranked higher, and when two groups submitted the same number of high quality models, the group with more acceptable models is ranked higher.

Overall, a total of 11 CAPRI predictor groups submitted correct models for at least 10 targets, and medium quality models for at least seven targets. These groups submitted models for at least 20 of the targets. Among those, the highest-ranking groups in this Round are Seok, Huang, and Guerois, with correct models for 15 or 16 targets, and medium quality models for 12–14 of these targets. These are followed by Zou, Shen and Grudinin (correct models for 11–14 targets, and medium quality models for 10 or 11 of those). The remaining five highest ranking groups, Weng, Vakser, Vajda/Kozakov, Fernandez-Recio and Lee, achieve correct predictions for 10–15 targets and medium quality predictions for 7–9 of those. It is noteworthy that two of the three top ranking predictor groups (Seok and Guerois), and at least one other group (Vakser) made heavy use of template-based modeling, an indication that this approach can be quite effective.

The remaining groups listed in Table IV were ranked lower, as they corrected predicted between 1 and 8 targets only, and produced only a few medium quality models for these targets. However some of these groups submitted predictions for a smaller number of targets. Their performance can therefore not be fairly compared to that of other groups.

Of the 6 CAPRI automatic docking servers ranked in Table IV, HADDOCK and CLUSPRO rank highest, followed by SWARMDOCK, and GRAMM-X.

It is interesting to note that two top ranking CAPRI servers submitted correct predictions for 16 targets, just as many as the top ranking predictor groups. But the latter groups still produce more medium accuracy models (>10) than the servers (no more than 9). Thus as already noted in previous CAPRI assessment, some CAPRI servers perform nearly on par with more manual predictions.

Among the CASP predictor and server groups listed in Table IV, the groups of Umeyama and Dunbrack rank highest, and both would rank among the best CAPRI predictor groups as their success rate (fraction of correct over submitted models) was also high. Of the servers, ROSETTASERVER and SEOK_SERVER rank highest, with a performance level similar to SWARMDOCK. Thirty-nine CASP groups submitting models for 1–5 targets, none of which were correct, are not explicitly listed in the Table.

Lastly, judging also by the best model submitted for each target, CAPRI scorers outperform CAPRI predictors, as already mentioned when analyzing the performance across targets. Highly ranking scorer groups submitted on average correct models for 1–2 more targets than CAPRI predictors, and the number of medium

Table IV
Participant ranking by Target performanceParticipant

	Participated targets	Performance
CAPRI Predictor Ranking		
Seok	25	**15/14****
Huang	25	**16/13****
Guerois	25	**16/12****
Zou	25	**14/11****
Shen	25	**13/11****
Grudinin	24	**11/10****
Weng	25	**13/9****
Vakser	25	**11/9****
Vajda/Kozakov	24	**15/8****
Fernandez-Recio	25	**11/8****
Lee	20	**10/7****
Tomii	20	**8/6****
Sali	12	**6/4****
Negi	25	**7/3****
Eisenstein	6	**3****
Bates	25	**7/2****
Kihara	23	**7/2****
Zhou	25	**4/2****
Tovchigrechko	12	**3/1****
Ritchie	8	**2/1****
Fernandez-Fuentes	14	**1**
Xiao	11	**1**
Gong	8	**0**
Del Carpio	3	**0**
Wade	2	**0**
Haliloglu	1	**0**
CAPRI SERVER Ranking		
HADDOCK	25	**16/9****
CLUSPRO	25	**16/8****
SWARMDOCK	25	**11/4****
GRAMM-X	22	**6/1****
LZERD	25	**3**
DOCK/PIERR	2	**1**
CAPRI Scorer Ranking		
Bonvin	25	**18/14****
Bates	24	**17/13****
Huang, Seok	25	**16/13****
Zou, Kihara	25	**15/12****
Fernandez-Recio	25	**14/12****
Weng	25	**16/11****
Oliva	22	**14/11****
Grudinin	25	**13/10****
Gray	17	**10/7****
LZERD	25	**6****
Lee	5	**3/2****
Sali	1	**0**
CASP Predictor and Server Ranking		
Umeyama	19	**13/8****
ROSETTASERVER	13	**9/8****
Dunbrack	12	**11/6****
SEOK_SERVER	22	**7/5****
Luethy	8	**5/4****
Nakamura	12	**7/3****
Baker	8	**3****
Wallner	2	**1****
Skwark, Lee, RAPTOR-X_Wang, NNS_Lee	1–4	**1**
39 participants not listed	1–5	**0**

For each target only the best quality solution is counted; in total 25 targets were assessed. Column 2 indicates the number of targets for which predictions were submitted. In Column 3, the numbers without stars indicate models of acceptable quality or better, and the numbers with "**" indicate the number of those models that were of medium quality.

quality models that groups submit for these targets is also somewhat higher.

Of the 13 scorer groups that submitted an accurate model for at least one target, 11 have correctly predicted at least 10 targets and submitted medium quality models for seven of those.

The best performing groups are those of Bonvin, Bates, Huang, Seok, Zou and Kihara, followed closely by four other groups that correctly predicted at least 13 targets, and produced medium quality models for at least 10 of these (Table IV).

Factors influencing the prediction performance

Unlike in previous CAPRI rounds, Round 30 comprised solely targets where both the 3D structure of the protein subunits and their association modes had to be modeled. Deriving the atomic coordinates of the predicted homo-oligomers therefore involved a number of steps each requiring the use of specialized software and making strategic choices as to how it should be applied.

As mentioned in Synopsis of the Prediction Methods, the approaches for modeling the subunit structures and generating the oligomer models vary widely amongst predictor groups, and across targets. It is therefore difficult to reliably pinpoint specific factors that contributed or hampered successful predictions. Nonetheless some general trends can be outlined. Even though Round 30 comprised only targets whose subunits could be readily modeled using templates from the PDB, the subunit modeling strategy had an important influence on the final oligomer models. Groups that used several different subunit models for the same target increased their chance of deriving at least an acceptable oligomer model. Such different models were obtained either by using different templates (some groups used as many as five templates for the same target), or by starting from the same template and modifying it by optimizing loop conformations and subjecting it to energy refinements. These optimizations seemed particularly effective when carried out in the context of the oligomers representing the highest-ranking template-based or docking models.

As already mentioned, information on oligomeric templates in the PDB was another important element contributing to improve the prediction performance. This information was the main ingredient for two of the best performing groups that heavily relied on template-based docking. Other groups that performed well used mainly *ab-initio* docking methods of various origins, but either guided the calculations or filtered the results based on structural information from homologous oligomers.

Other important elements, such as selecting representative members of clusters of docking solutions, and the final scoring functions used to rank models and select

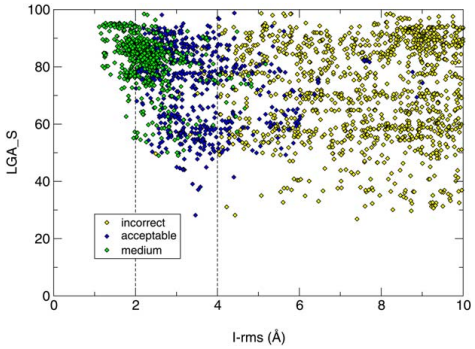


Figure 7

Subunit model accuracy and the quality of predicted complexes in CAPRI Round 30. The CASP LGA_S scores of subunit models in the predicted complexes for the 25 targets in this Round (vertical axis) are plotted as a function of the *I-rms* values (horizontal axis). Each point in this Figure represents one submitted model, and points are colored according to the quality of the predicted complex, respectively, incorrect (yellow), acceptable (blue) and medium (green) quality (see Table I and the text for details).

those to be submitted, also played a role as already mentioned here and in previous CAPRI reports.[19]

In the following we examine in more detail the impact of two important elements of this joint CASP-CAPRI experiment. We evaluate the influence of the accuracy of individual subunits models on the oligomer prediction performance, and estimate the extent to which procedures that rely on docking methodology and those that employ specialized template-based modeling confer an advantage over straightforward homology modeling.

Influence of subunit model accuracy

The subunit models used to derive the models of the oligomers were generated either by CAPRI groups, those with more homology modeling expertise, or borrowed from amongst the models submitted by CASP servers, which were made available to CAPRI groups in time for each docking experiment. The subunit structures in models submitted by CAPRI and CASP groups for all 25 targets of Round 30 were assessed using the standard CASP GDT_TS and LGA_S scores, as well as the backbone rmsd of the submitted model versus the target structures. The values of these measures obtained for models submitted by all participants in Round 30 for each target can be found at the CAPRI website together with the assessment results for this Round.

To gauge the relation between the accuracy of subunit models and the docking prediction performance, the LGA-S scores of subunit models in the predicted complexes for the 25 targets in this Round are plotted in Figure 7 as a function of the *I-rms* value. The LGA_S measure was used because it does not depend on the residue numbering along the chain, which may vary at least in a fraction of the models submitted by CAPRI participants. The *I-rms* measure was used as it represents best the accuracy level of the predicted interface.

Each point in Figure 7 represents one submitted model, and points are colored according to the quality of the predicted complex (incorrect, acceptable and medium quality). The plot clearly shows that medium quality predicted complexes (*I-rms* values between 1 and 3 Å) tend to be associated with high accuracy subunit models (LGA_S values >80). We also see that predicted complexes of acceptable quality (*I-rms* values of 2–4 Å) are associated with subunit models that span a wide range in accuracy levels (LGA_S between 30 and 90). This range is comparable to the subunit accuracy range associated with incorrect models of complexes (*I-rms* >4 Å; see Table III for details on how *I-rms* contributes to rank CAPRI models). Identical trends are observed when plotting the GDT-TS scores as a function of the *I-rms* values for the fraction of the models with correct residues numbering (Supporting Information Fig. S2).

That both accurate and inaccurate subunit models are associated with incorrectly modeled complexes is expected. Inaccurate subunit models may indeed prevent the identification of the correct binding mode, and docking calculations may fail to identify the correct binding mode even when the subunit models are sufficiently accurate. It is however noteworthy that complexes classified as incorrect by the CAPRI criteria do not necessarily represent prediction noise, as a recent analysis has shown that residues that contribute to the interaction interfaces are correctly predicted in a significant fraction of these complexes.[48]

Somewhat less expected is the observation (Fig. 7) that in a significant number of cases, acceptable and to a smaller extent also medium quality docking solutions can be identified even with lower accuracy models of the individual subunits. This is an encouraging observation, as it suggests that docking calculations can lead to useful solutions with protein models built by homology, and that these models need not always be of the highest accuracy. What probably matters more for the success of docking predictions is the accuracy with which the binding regions of the individual components of the complex are modeled, rather than the accuracy of the 3D model considered in its entirety.

Round 30 predictions versus standard homology modeling

To estimate the extent to which docking methods or template-based modeling procedures confer an advantage over straightforward homology modeling, the accuracy of the submitted oligomer models for each target was

Table V
Best available templates detected based on sequence ("Sequence"), experimental monomer structure ("Monomer"), and experimental oligomer structure ("Oligomer")Target

			Best template TM-score (detected template)		
	Target released	Database released	Sequence	Monomer	Oligomer
T68	May 01, 2014	April 24, 2014	0.348 (3njd)	0.370 (3fse)	0.370 (3fse)
T69	May 05, 2014	April 24, 2014	0.852 (1qlw)	0.852 (1qlw)	0.852 (1qlw)
T70	May 06, 2014	April 24, 2014	0.639 (2f06)	0.644 (3c1m)	0.652 (3tvi)
T71	May 07, 2014	April 24, 2014	0.509 (2id5)	0.618 (3jur)	0.618 (3jur)
T72	May 08, 2014	April 24, 2014	0.510 (3otn)	0.510 (3otn)	0.510 (3otn)
T73	May 09, 2014	April 24, 2014	[a]	0.554 (1hql)	0.554 (1hql)
T74	May 12, 2014	April 24, 2014	0.340 (4jrf)	0.340 (4jrf)	0.340 (4jrf)
T75	May 13, 2014	April 24, 2014	0.880 (3rjt)	0.880 (3rjt)	0.880 (3rjt)
T77	May 15, 2014	April 24, 2014	0.393 (2xwx)	0.375 (4iib)	0.375 (4iib)
T78	May 20, 2014	May 17, 2014	0.315 (3c6c)	0.370 (1o0s)	0.403 (2f3o)
T79	May 23, 2014	May 17, 2014	0.440 (2bnl)	0.469 (2xig)	0.471 (2w57)
T80	June 02, 2014	May 17, 2014	0.938 (1mdo)	0.943 (2fnu)	0.943 (2fnu)
T82	June 04, 2014	May 17, 2014	0.846 (4dn2)	0.846 (4dn2)	0.846 (4dn2)
T84	June 09, 2014	May 17, 2014	0.939 (2btm)	0.941 (1b9b)	0.941 (1b9b)
T85	June 10, 2014	May 17, 2014	0.889 (3ggo)	0.889 (3ggo)	0.889 (3ggo)
T86	June 11, 2014	May 17, 2014	0.459 (4h3u)	0.467 (3gzr)	0.470 (3hk4)
T87	June 13, 2014	May 17, 2014	0.922 (3get)	0.922 (3get)	0.922 (3get)
T90	July 03, 2014	June 06, 2014	0.921 (4qgr)	0.927 (2oga)	0.927 (2oga)
T91	July 08, 2014	June 06, 2014	0.750 (4gel)	0.750 (4gel)	0.808 (3hsi)
T92	July 09, 2014	June 06, 2014	0.785 (1tu7)	0.837 (3h1n)	0.837 (3h1n)
T93	July 10, 2014	June 06, 2014	0.896 (4a7p)	0.896 (4a7p)	0.896 (4a7p)
T94	July 11, 2014	June 06, 2014	0.655 (3gff)	0.655 (3gff)	0.655 (3gff)

TM-score of the templates that have the highest TM-score among top 10 selected templates for each target and the PDB IDs of the templates are listed.
[a]No protein with the desired oligomer state was found among the top 100 HHsearch entries.

compared to the accuracy of the models build using the best oligomer templates for that target available in the PDB at the time of the prediction. Only dimer targets (and templates) were considered, given the uncertainty of the oligomeric state assignments for some of the tetrameric targets.

Three categories of the best dimeric templates were considered (see Assessment Procedures and Criteria): templates identified on the basis of sequence alignments alone, templates identified by structurally aligning the target and template monomers, and templates identified by structurally aligning the target and template oligomers. Only the sequence-based template selection reconstitutes the task performed by predictors, to whom only the target sequence was disclosed at the time of the prediction. The resulting templates thus represent the best templates available to predictors during the prediction Round. Obviously, the structurally most similar templates could not be identified by predictors, but are considered here in order to evaluate the advantage, if any, conferred by such templates over those identified on the basis of sequence alignments.

Table V lists the best templates from each category identified for all dimeric targets of Round 30 and the corresponding template-target TM-scores. These templates represent those with the highest TM-score among the best 10 templates from each category detected for a given targets. Not too surprisingly targets with more similar templates, those featuring high TM-scores ($\geq$0.7), are the easy targets, whereas difficult targets are those with poorer templates (lower TM-scores). Many of the best templates from all three categories were also detected and used by predictor groups (see Supporting Information Table S5), even though these groups only had sequence information to identify them during the prediction round.

The accuracy levels of the models built using the three categories of best templates for each target and the best models from each of the participating CAPRI predictor groups submitted for the same target are plotted in Figure 8. The model accuracy is measured by the *I-rms* value, representing the accuracy level of the predicted interface in the complex. Each entry in the Figure represents one model, and for each template category (based on sequence alignments, on structural alignment of the monomers and dimers, respectively), up to 10 best models are shown per target and colored according to the template category.

Inspection of Figure 8 indicates that models submitted by CAPRI predictor groups, a vast majority of which employed docking methods as part of their protocol, tend to be of higher accuracy. For most of the easy targets, the 10 models submitted by CAPRI groups more consistently display lower *I-rms* values then the models built from the best templates. This is the case not only for models derived from the sequence-based templates but also for the most structurally similar templates of

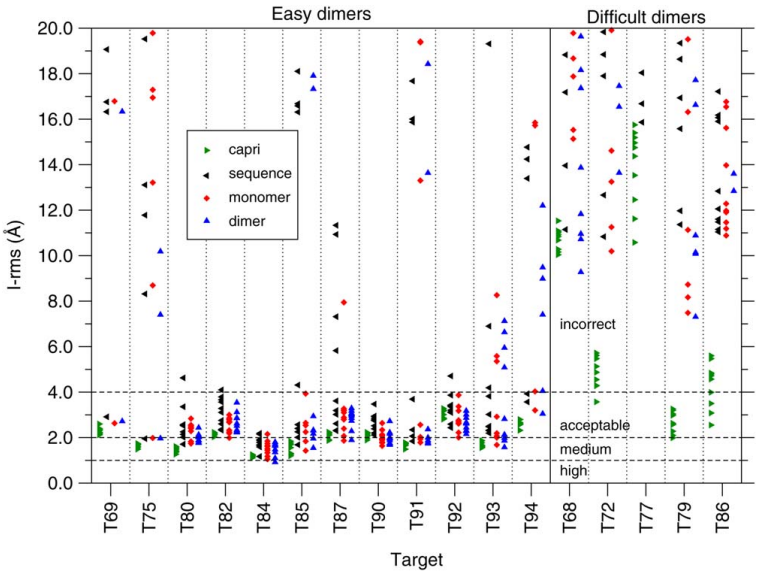


Figure 8

Accuracy of Round 30 homodimer models predicted by protein docking methods and template-based modeling versus models derived by standard homology modeling. The *I-rms* values, representing the accuracy level of the predicted interface, are plotted (vertical axis) for different models for each target (listed on the horizontal axis using the CAPRI target identification). Each point represents one model. The best models submitted by individual CAPRI predictor groups are represented by green triangles. The remaining models are those built in this study by standard homology modeling techniques[42] on the basis of homodimer templates from the PDB. Up to 10 best models are shown per target and template category (see text). Models based on templates identified using sequence information (black triangles), models based structural alignments of individual monomers (red lozenges), and those based on structural alignments of the entire dimers (blue triangles). The targets (only dimers) are subdivided into easy and difficult targets (see text). Dashed horizontal lines represent *I-rms* values delimiting models of high, medium, acceptable and lower (incorrect) quality by CAPRI criteria.

the monomer or dimer categories. Considering only the best models for each targets the performance results are mode balanced. For seven out of the 12 easy targets the best models overall were submitted by CAPRI participants, whereas for the remaining five targets the most accurate models were those derived from the structurally most similar template. Overall however, acceptable or medium quality models were obtained with all the approaches and for nearly all the easy targets.

On the other hand it is remarkable that for three of the difficult targets (T72, T79, and T86), the docking procedures were able to produce acceptable models, with one medium quality model for T79, whereas all the template-based models were incorrect.

Overall these results do confirm that protein docking procedures represent an added value over straightforward template-based modeling. One must recall however, that docking was often combined with template-based restraints and hence, can in general not be qualified as *ab-initio* docking in the context of this experiment. It is also important to note that for two targets, T82 and T85, the highest accuracy models were predicted by the group of Seok, who employed specialized template-based modeling techniques augmented by loop modeling and refinement. But the accuracy of these models was not vastly superior to that of the best docking models.

Lastly, not too surprisingly, oligomer models build using the sequence-based best templates were generally of inferior accuracy than models built from templates of the two other categories. Interestingly, models derived from the most structurally similar dimer templates were not generally more accurate than those derived from the structurally most similar monomers. This may stem from differences in the structural alignments that were used to detect the templates, which in turn could have affected the performance of the homology modeling procedure (MODELLER).

CONCLUDING REMARKS

CAPRI Round 30, for which results were assessed here, was the first CASP-CAPRI experiment, which brought

together the community of groups developing methods for protein structure prediction and model refinement, with groups developing methods for predicting the 3D structure of protein assemblies. The 25 targets of this round represented a subset of the targets submitted for the CASP11 prediction season of the summer of 2014. In line with the main focus of CASP, the majority of these targets were single protein chains, forming mostly homodimers, and a few homotetramers. Only two of the targets were heterodimers, similar to the staple targets in previous CAPRI rounds. Unlike in most previous CAPRI rounds both subunit structures and their association modes had to be modeled for all the targets. Since the docking or assembly modeling performance may crucially depend on the accuracy of the models of individual subunits, the targets chosen for this experiment were proteins deemed to be readily modeled using templates from the PDB. Interestingly, templates were used mainly to model the structures of individual subunits, to limit the sampling space of docking solution or to filter such these solutions. Only a few groups carried out template-based docking for the majority of the targets, and two of those ranked amongst the top performers, indicating that this relatively recent modeling strategy has potential.

As part of our assessment we established that the accuracy of the models of the individual subunits was an important factor contributing to high accuracy predictions of the corresponding complexes. At the same time we observed that highly accurate models of the protein components are not necessarily required for identifying their association modes with acceptable accuracy.

Furthermore, we provide evidence that protein docking procedures and in some cases also specialized template-based methods generally outperform off-the-shelf template-based prediction of complexes. These findings apply to templates identified on the basis of sequence information alone, as well as to templates structurally more similar to the target. The added value of docking methods was particularly significant for the more difficult targets, where the structures of the identified best templates differed more significantly from the target structure

Thus, the assessment results presented here confirm that the prediction of homodimer assemblies by homology modeling techniques and docking calculations is feasible, especially for stable dimers that feature interface areas of 1000–1500 $Å^2$, whose size is comparable or larger than the one associated with transient heterocomplexes. They also confirm that docking procedures can represent a competitive advantage over standard homology modeling techniques, when those are applied without further improvements to model the complex.

On the other hand, difficulties arise when the subunit interface in the target is similar in size to those associated with crystal contacts.[45] Such cases were associated with a number of targets where the oligomeric state assignment was ambiguous or inaccurate. Such ambiguous or inaccurate oligomeric state assignments represented a confounding factor for the docking prediction in this round. The problem arose mainly from the fact that the authors' assignments, usually based on independent experiment evidence, were not available to predictors at the time of the prediction experiment. Instead, predictors were provided with tentative assignments, inferred on the basis of computational analysis of the crystal contacts. Quite encouragingly, for most targets with ambiguous assignment, or for which the tentative assignments were later overruled by the authors upon submission to the PDB, the docking predictions were shown to provide useful information, which often confirmed the final assignment or helped resolve ambiguous ones. This occurred for both homodimer and homotetramer targets.

Lastly, we find that the docking prediction performance for the genuine homodimer targets was superior to that obtained for heterocomplexes in previous CAPRI rounds, in line with the expectation that, owing to their higher binding affinity (and larger and more hydrophobic interfaces), homodimers are easier to predict than heterodimers. Much poorer prediction performance was however achieved for genuine tetrameric targets, where the inaccuracy of the homology-built subunit models and the smaller pair-wise interfaces limited the prediction performance. Accurately modeling of higher order assemblies from sequence information is thus an area where progress is needed.

ACKNOWLEDGMENTS

We are most grateful to the PDBe at the European Bioinformatics Institute in Hinxton, UK, for hosting the CAPRI website. Our deepest thanks go to all the structural biologists and to the following structural genomics initiatives: Northeast Structural Genomics Consortium, Joint Center for Structural Genomics, NatPro PSI:Biology, New York Structural Genomics Research Center, Midwest Center for Structural Genomics, Structural Genomics Consortium, for contributing the targets for this joint CASP-CAPRI experiment. MFL acknowledges support from the FRABio FR3688 Research Federation "Structural & Functional Biochemistry of Biomolecular Assemblies."

REFERENCES

1. Alberts B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 1998;92:291–294.
2. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr Section D Biol Crystallogr 2002;58:899–907.
3. Smith MT, Rubinstein JL. Structural biology. Beyond blob-ology. Science 2014;345:617–619.

4. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci USA 2012;109:9438–9441.
5. Berman HM, Coimbatore Narayanan B, Di Costanzo L, Dutta S, Ghosh S, Hudson BP, Lawson CL, Peisach E, Prlic A, Rose PW, Shao C, Yang H, Young J, Zardecki C. Trendspotting in the Protein Data Bank. FEBS Lett 2013;587:1036–1045.
6. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30:1072–1080.
7. Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife 2014;3:e02030
8. Whitehead TA, Baker D, Fleishman SJ. Computational design of novel protein binders and experimental affinity maturation. Methods Enzymol 2013;523:1–19.
9. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Myers CA, Kamisetty H, Blair P, Wilson IA, Baker D. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol 2012;30:543–548.
10. Ward AB, Sali A, Wilson IA. Biochemistry. Integrative structural biology. Science 2013;339:913–915.
11. Wodak SJ, Janin J. Structural basis of macromolecular recognition. Adv Protein Chem 2002;61:9–73.
12. Ritchie DW. Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci 2008;9:1–15.
13. Vajda S, Kozakov D. Convergence combination of methods in protein-protein docking. Curr Opin Struct Biol 2009;19:164–170.
14. Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC, Demerdash ON, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko JS, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Aze J, Soner S, Ovali SK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Perez-Cano L, Pons C, Fernandez-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AM, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-Rodriguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang SY, Zou X, Wodak SJ, Janin J, Baker D. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 2011;414:289–302.
15. Moretti R, Fleishman SJ, Agius R, Torchala M, Bates PA, Kastritis PL, Rodrigues JP, Trellet M, Bonvin AM, Cui M, Rooman M, Gillis D, Dehouck Y, Moal I, Romero-Durana M, Perez-Cano L, Pallara C, Jimenez B, Fernandez-Recio J, Flores S, Pacella M, Praneeth Kilambi K, Gray JJ, Popov P, Grudinin S, Esquivel-Rodriguez J, Kihara D, Zhao N, Korkin D, Zhu X, Demerdash ON, Mitchell JC, Kanamori E, Tsuchiya Y, Nakamura H, Lee H, Park H, Seok C, Sarmiento J, Liang S, Teraguchi S, Standley DM, Shimoyama H, Terashi G, Takeda-Shitaka M, Iwadate M, Umeyama H, Beglov D, Hall DR, Kozakov D, Vajda S, Pierce BG, Hwang H, Vreven T, Weng Z, Huang Y, Li H, Yang X, Ji X, Liu S, Xiao Y, Zacharias M, Qin S, Zhou HX, Huang SY, Zou X, Velankar S, Janin J, Wodak SJ, Baker D. Community-wide evaluation of methods for predicting the effect of mutations on protein-protein interactions. Proteins 2013;81:1980–1987.
16. Lensink MF, Wodak SJ. Docking, scoring, and affinity prediction in CAPRI. Proteins 2013;81:2082–2095.
17. Lensink MF, Moal IH, Bates PA, Kastritis PL, Melquiond AS, Karaca E, Schmitz C, van Dijk M, Bonvin AM, Eisenstein M, Jimenez-Garcia B, Grosdidier S, Solernou A, Perez-Cano L, Pallara C, Fernandez-Recio J, Xu J, Muthu P, Praneeth Kilambi K, Gray JJ, Grudinin S, Derevyanko G, Mitchell JC, Wieting J, Kanamori E, Tsuchiya Y, Murakami Y, Sarmiento J, Standley DM, Shirota M, Kinoshita K, Nakamura H, Chavent M, Ritchie DW, Park H, Ko J, Lee H, Seok C, Shen Y, Kozakov D, Vajda S, Kundrotas PJ, Vakser IA, Pierce BG, Hwang H, Vreven T, Weng Z, Buch I, Farkash E, Wolfson HJ, Zacharias M, Qin S, Zhou HX, Huang SY, Zou X, Wojdyla JA, Kleanthous C, Wodak SJ. Blind prediction of interfacial water positions in CAPRI. Proteins 2014;82:620–632.
18. Lensink MF, Mendez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins 2007;69:704–718.
19. Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. Proteins 2010;78:3073–3084.
20. Goodsell DS, Olson AJ. Structural symmetry and protein function. Annu Rev Biophys Biomol Struct 2000;29:105–153.
21. Kuhner S, van Noort V, Betts MJ, Leo-Macias A, Batisse C, Rode M, Yamada T, Maier T, Bader S, Beltran-Alvarez P, Castano-Diez D, Chen WH, Devos D, Guell M, Norambuena T, Racke I, Rybin V, Schmidt A, Yus E, Aebersold R, Herrmann R, Bottcher B, Frangakis AS, Russell RB, Serrano L, Bork P, Gavin AC. Proteome organization in a genome-reduced bacterium. Science 2009;326:1235–1240.
22. Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL, 2nd Tsutakawa SE, Jenney FE, Jr., Classen S, Frankel KA, Hopkins RC, Yang SJ, Scott JW, Dillard BD, Adams MW Tainer JA. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods 2009;6:606–612.
23. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372:774–797.
24. Webb B, Sali A. Protein structure modeling with MODELLER. Methods Mol Biol 2014;1137:1–15.
25. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 2006;22:195–201.
26. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D. High-resolution comparative modeling with RosettaCM. Structure 2013;21:1735–1742.
27. de Vries SJ, van Dijk M, Bonvin AM. The HADDOCK web server for data-driven biomolecular docking. Nat Protoc 2010;5:883–897.
28. Macindoe G, Mavridis L, Venkatraman V, Devignes MD, Ritchie DW. HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 2010; 38:W445–449.
29. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 2014;30:1771–1773.
30. Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res 2008;36:W233–238.
31. Pierce B, Tong W, Weng Z. M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. Bioinformatics 2005;21:1472–1478.
32. Szilagyi A, Zhang Y. Template-based structure modeling of protein-protein interactions. Curr Opin Struct Biol 2014;24:10–23.
33. Kallberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J. Template-based protein structure modeling using the RaptorX web server. Nat Protoc 2012;7:1511–1522.
34. Tuncbag N, Keskin O, Nussinov R, Gursoy A. Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement. Proteins 2012;80:1239–1249.
35. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
36. Kryshtafovych A, Monastyrskyy B, Fidelis K. CASP prediction center infrastructure and evaluation measures in CASP10 and CASP ROLL. Proteins 2014;82Suppl 2:7–13.
37. Cozzetto D, Kryshtafovych A, Fidelis K, Moult J, Rost B, Tramontano A. Evaluation of template-based models in CASP8 with standard measures. Proteins 2009;77Suppl 9:18–28.

38. Zemla A, Venclovas, Moult J Fidelis K. Processing and evaluation of predictions in CASP4. Proteins 2001;Suppl 5:13–21.
39. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 2012;9:173–175.
40. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21:951–960.
41. Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure prediction with HHpred. Proteins 2009;77 Suppl 9:128–132.
42. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779–815.
43. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
44. Mukherjee S, Zhang Y. MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res 2009;37:e83
45. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004; 336:943–955.
46. Dey S, Pal A, Chakrabarti P, Janin J. The subunit interfaces of weakly associated homodimeric proteins. J Mol Biol 2010;398:146–160.
47. Mackinnon SS, Malevanets A, Wodak SJ. Intertwined associations in structures of homooligomeric proteins. Structure 2013;21:638–649.
48. Lensink MF, Wodak SJ. Blind predictions of protein interfaces by docking calculations in CAPRI. Proteins 2010;78:3085–3095.

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 285, NO. 14, pp. 10519–10526, April 2, 2010

Identification of Specific Lipid-binding Sites in Integral Membrane Proteins*[S]

Received for publication, September 23, 2009, and in revised form, January 14, 2010 Published, JBC Papers in Press, February 5, 2010, DOI 10.1074/jbc.M109.068890

Marc F. Lensink[‡§1], **Cédric Govaerts**[‡2], **and Jean-Marie Ruysschaert**[‡]

From [‡]*Structure and Function of Biological Membranes and* [§]*Genome and Network Bioinformatics, Université Libre de Bruxelles, Boulevard du Triomphe-CP 263, B-1050 Brussels, Belgium*

Protein-lipid interactions are increasingly recognized as central to the structure and function of membrane proteins. However, with the exception of simplified models, specific protein-lipid interactions are particularly difficult to highlight experimentally. Here, we used molecular dynamics simulations to identify a specific protein-lipid interaction in lactose permease, a prototypical model for transmembrane proteins. The interactions can be correlated with the functional dependence of the protein to specific lipid species. The technique is simple and widely applicable to other membrane proteins, and a variety of lipid matrices can be used.

The molecular mechanisms underlying the influence of the lipid environment on the function of membrane proteins remain unclear. Lipid composition is known to have a modulatory effect on membrane protein activity, and for a number of membrane proteins a clear correlation was found between protein activity and bulk properties of the membrane bilayer such as fluidity (1). Membrane proteins are anchored in their lipid environment through nonspecific protein-lipid interactions (2, 3), and different fluidic properties are therefore expected to play a modulatory role on membrane protein activity. However, there is also increasing evidence for specifically bound lipids that are necessary to achieve biological function (4). For example, the presence of phosphatidylethanolamine (PE)[3] in the bilayer has been found to be essential for the structure and function of a number of transmembrane transporter proteins (5–10). Unfortunately, crystallographic evidence of such a dependence is exceptional. First, membrane protein structure remains rare, with the structures of less than 200 unique proteins deposited in the Protein Data Bank. Second, membrane protein purification and crystallization require the use of detergent, which tends to delipidate the target protein. Thus only a few examples of crystal structures show lipids, known to be important for function, specifically associated to the protein. The yeast cytochrome bc_1 complex contains a number of bound phospholipids, which suggest specific roles both for the structural and functional integrity of the protein (11). The crystal structure of the *Thermochromatium tepidum* photosynthetic reaction center shows one PE and seven detergent molecules on its molecular surface (12). In the crystal structure of the *Rhodobacter sphaeroides* photoreaction center, a specific interaction with cardiolipin has been found (13). This binding site has subsequently been suggested to be a conserved feature of these reaction centers (14). More recently, a tightly bound cholesterol molecule was observed in the crystal structure of a protein G-coupled receptor (15), triggering the question of the general presence of nonannular cholesterol-binding sites for these proteins (16).

Molecular modeling and other computational studies are therefore playing an increasingly relevant role in the study of membrane proteins (17); these include the use of molecular dynamics simulation to study dynamic events and conformational pathways (18). The installment of membrane protein-targeted structural genomics and the rapid development of high throughput structural biology techniques are expected to result in a significant increase in the number of available membrane protein structures in the near future (19). But it is doubtful that these structures will show physiologically associated lipid molecules; the crystallization procedure typically delipidates the protein, and only the strongest bound lipid molecules may withstand the purification procedure and show sufficient electron density to be resolved at the atomic level. Molecular modeling may play an important role here. With the increase in computing power, molecular dynamics simulations are now routinely applied, and physics-based force fields are capable of treating a wide range of molecules and molecular interaction (20, 21).

In this study, we use molecular modeling and dynamics to characterize a specific protein-lipid interaction between lactose permease (LacY) and PE. LacY, which has been crystallized at a resolution of 2.95 Å (22), is a paradigm for the major facilitator superfamily (MFS) (23) that represents as much as 25% of all membrane transport proteins, with over 15,000 sequence members identified to date (24). Members of the MFS show a common architecture (25), and for a number of these transporters the presence of PE in the bilayer is required for proper structure and activity (26). By carefully positioning the protein in a selection of lipid matrices and performing molecular dynamics simulations, we identify a limited number of strong

* The Genome and Network Bioinformatics Laboratory is a partner of the BioSapiens Network of Excellence funded under the Sixth Framework Program of the European Commission (LSHG-CT-2003-503265).

[S] The on-line version of this article (available at http://www.jbc.org) contains supplemental Fig. S1.

[1] Supported by DGTRE Contract 515993 from the Région Wallonne of Belgium. To whom correspondence should be addressed. Tel.: 32-2-650-5411; Fax: 32-2-650-5425; E-mail: marc.lensink@ulb.ac.be.

[2] Chargé de Recherches of the Fonds National de la Recherche Scientifique.

[3] The abbreviations used are: PE, phosphatidylethanolamine; MFS, major facilitator superfamily; POPE, 1-palmitoyl-2-oleoyl-L-phosphatidylethanolamine; POPC, 1-palmitoyl-2-oleoyl-L-phosphatidylcholine; POPG, 1-palmitoyl-2-oleoyl-L-phosphatidylglycerol; PC, phosphatidylcholine; MD, molecular dynamics; GlcDAG, monoglucosyl-diacylglycerol; DGlcDAG, diglucosyl-diacylglycerol.

interactions between specific amino acids and individual lipid molecules. The strength of this technique lies in the fact that it leads to a molecular picture of the lipid-protein interactions involved. The only limitation is the number of membrane protein structures available with a sufficient resolution; a large number of lipid matrices in which proteins can be inserted are available, and the technology required to create new ones mimicking all kinds of lipid membrane compositions is available and operational (20).

MATERIALS AND METHODS

Modeling Setup—We take the LacY crystal structure with the highest available resolution (2.95 Å), representing wild-type LacY in neutral pH, that is open on the cytoplasmic side and therefore susceptible to gradient sensing and sugar transport (22). With lipid chain lengths near C_{18} often being the most favorable for function (27, 28), we employ POPC, POPE, and POPG bilayers that have been extensively used in MD simulations and for which parameter sets are available (29–31). The bilayer thickness of these bilayers (32) shows good agreement with the calculated hydrophobic thickness of LacY (33), minimizing the effect of hydrophobic mismatch. Previous simulations of LacY, investigating the mechanism of sugar transport, have also been performed in a POPE bilayer (34).

Lactose permease (Protein Data Bank code 2cfq) was reoriented with its principal axis aligned with the z axis, rotated 10° over the x axis, and placed in the center of an equilibrated POPE bilayer (30) with all waters removed and the x and y coordinates expanded by a factor of 4. This orientation was found to yield the best final alignment between phosphates and interfacial arginines. The entire system was subjected to 100 steps of steepest descent energy minimization, applying 10^5 kJ/nm^2 position restraints on the protein non-hydrogen atoms. In subsequent iterations, the system was restored to the reference area per lipid by shrinking the lipid x and y coordinates by 2% and deleting all lipids that had their phosphorus atom at a distance closer than 6 Å to any Cα atom of the protein (35). Every iteration was accompanied by 100 steps of steepest descent, and after eight iterations the deflation was increased to 5% per iteration.

The original box size of the resulting system of LacY and 298 POPE lipids was expanded by 1 nm perpendicular to the bilayer surface (now $\sim 10 \times 10 \times 7$ nm^3) and solvated with roughly 9000 water molecules. The atoms in the palmitoyl and oleoyl chains were given a van der Waals radius of 6 Å to avoid solvation of the hydrophobic bilayer core. The system was neutralized by adding chloride ions (36) and was subjected to 1000 steps of steepest descent energy minimization and 100 ps of MD using the weak position restraint (10^3 kJ/nm^2 on the non-hydrogen protein atoms), followed by 10 ns of free MD.

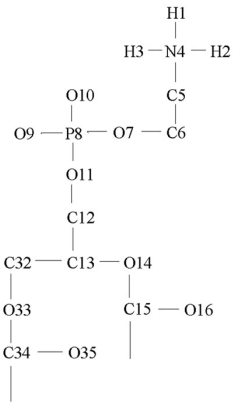
Previous simulations of PE systems have reduced the ethanolamine partial charges to correspond to the lysine parameterization (37). To eliminate artificial association between the ethanolamine moiety and acidic side chains of the protein because of an overpolarization of the lipid partial charges, a new charge set was also developed by fitting atomic point charges to reproduce the electrostatic potential obtained from *ab initio* HF/6–31G* calculations (38), keeping symmetry considerations (39) and summing hydrogen charges onto the nonpolar carbons (40). For all purposes of this work, the results obtained with both charge sets were found to be similar, and all analyses presented henceforth were carried out using the latter charge set. Both charge sets are listed in Table 1.

TABLE 1
Lipid head group partial charges

Modified partial charges for the lipid head group. Although we qualitatively obtained the same results with previously published charges (61), these charges are based on a Mulliken population analysis of electron density (62), whereas a charge fitting to reproduce electrostatic potential is the currently more accepted standard (63). This new charge set avoids preferential binding due to overpolarization.

atom	charge[(a)]	charge[(b)]
$H^{(c)}_{1,2,3}$	0.4	0.252
N_4	−0.5	−0.087
C_5	0.3	0.107
C_6	0.4	0.308
O_7	−0.8	−0.468
P_8	1.7	1.178
$O_{9,10}$	−0.8	−0.787
O_{11}	−0.7	−0.468
C_{12}	0.4	0.16
C_{13}	0.3	0.278
O_{14}	−0.7	−0.48
C_{15}	0.7	0.9
O_{16}	−0.7	−0.62
C_{32}	0.5	0.21
O_{33}	−0.7	−0.48
C_{34}	0.8	0.9
O_{35}	−0.6	−0.62

lipid head group



[a] Original partial charge is shown.
[b] Newly derived partial charge is shown.
[c] Atom is C in the case of methylated PE.

Molecular dynamics simulations were performed with the Gromacs 3.3 package (41), using the Gromos96 43a2 force field (40), with Berger parameters for the lipid tails (42). A time step of 2 fs was employed. All systems were coupled to a temperature bath of $T = 310$ K with a coupling constant $\tau_T = 0.1$ ps. Protein, lipids, and solvent (ions plus water) were coupled independently. Single point charge water was employed (43). Pressure was maintained using semi-isotropic pressure coupling (44) at $p = 1$ bar with a coupling constant of $\tau_p = 1.0$ ps. van der Waals interactions were cut off at a distance of 1.4 nm, and electrostatic interactions were calculated with the particle mesh Ewald method (45), using fourth-order splining and a grid spacing of 0.12 nm. Equations of motion for the water molecules were solved analytically (46), and all covalent bonds in the system were constrained in the MD simulations (47).

The entire procedure of positioning, equilibration, and simulation was repeated using POPE lipids with monomethylated, dimethylated, and trimethylated ethanolamine (the trimethylation resulting in POPC) and in a bilayer consisting of POPG lipids. The POPG bilayer was obtained by mutating every POPE lipid into POPG as before (48). Methylation of the PE lipids was achieved by mutating a lipid amine hydrogen into methyl while at the same time modifying all other related parameters (bond lengths, angles, nonbonded parameters, etc.). Additional *ab initio* calculations had shown it not to be necessary to modify the partial charges for the hydrogen to methyl mutation, which

TABLE 2
Transmembrane helix angles

Residues defining the TM helices and their angle with the bilayer normal (corresponding to the *z* axis, pointing from cytosol to periplasm) are shown. The direction of the helix is taken into account when calculating its angle with the bilayer normal. For reference, the corresponding angles of the x-ray structure, after positioning in the bilayer, but before equilibration, are also listed. TM3 and TM8 show a kink in the middle. Helix 7.c is a short helix located in the cytosol, just before TM7 and under an angle of 97.4 ± 5.7° with it.

Helix	Residues	Angle[a]	Angle[b]
TM1	6–36	32.5	34.3 ± 3.9
TM2	45–68	160.7	159.6 ± 2.8
TM3.1	74–83	6.8	9.0 ± 4.2
TM3.2	86–101	47.3	52.5 ± 7.4
TM4	104–137	144.9	143.3 ± 2.4
TM5	139–165	6.3	9.0 ± 2.4
TM6	166–184	161.4	158.8 ± 5.5
TM7.c	209–217	69.9	76.6 ± 7.0
TM7	219–250	35.0	33.4 ± 2.5
TM8.1	253–275	161.8	160.4 ± 3.8
TM8.2	276–286	117.4	118.6 ± 3.9
TM9	288–307	5.1	5.6 ± 2.5
TM10	311–341	157.5	157.5 ± 2.5
TM11	343–376	4.4	7.0 ± 2.8
TM12	376–397	155.5	156.9 ± 3.1

[a] Initial angle, after positioning in the bilayer, is shown.
[b] Average angle over the combined simulations is shown.

corresponds to the way the original POPE parameters were derived from POPC (30).

Validation of Equilibration Procedure—The orientational evolution of the α-helices in the simulation is used to validate the equilibration procedure and overall structural stability of the system. LacY contains 12 transmembrane helices, two of which with a kink in the middle, and one additional short helix on the cytosolic side. Table 2 shows the residues defining the helices and their angle with the normal bilayer averaged over the four simulations with neutral lipids. Helical axes were calculated using a rotational least squares fitting procedure (49). It is known that TM helices may tilt or flex to match the hydrophobic thickness of a membrane (3, 50). We indeed observed this for the longer helices, *e.g.* tilt for TM1 and TM4 and flex for TM3 and TM8. The shorter TM helices may span the bilayer without a tilt, *e.g.* TM5 and TM9. The observed behavior corresponds to what has been found by NMR experiments and MD simulations of the individual LacY helices TM1 and TM5 (50). A secondary structure analysis shows all helices retain their helical structure throughout the simulations (data not shown). Also, no change in helix orientation (less than 10° range of change in angle) was found (with the exception of TM6 and TM3.2 that varied within a 20° range). In addition, the relatively short helix 7.c showed some degree of motion, with a deviation up to 20° around its average. TM6 is imperfect in the crystal structure, and we find it to fully adopt the α-helical structure in all simulations, most likely due to the stabilizing influence of the lipid environment. The same happens with TM3.2, which is located next to TM6. The average helical angles and standard deviations were not found to differ significantly when calculated over the full trajectory or just the second half of it. Taken together with the correspondence to initial values, we conclude that the positioning and equilibration procedure has had no destabilizing effect on the overall protein structure.

Detection of Lipid-mediated Salt Bridges—The presence of salt bridges and hydrogen bonds between LacY and lipids was determined using simple distance criteria. Residues involved in salt bridges and hydrogen bonds to individual lipids were determined using the criterion that at any point during the combined simulations the donor-acceptor distance fell below the threshold of 4 Å. The upper limit of 4 Å constitutes a weak interaction. By using the obtained residues, and the requirement that two of them must *simultaneously* be bound to the same lipid, lipid-mediated salt bridges were determined using a distance criterion of 4 Å of the residue phosphate (donor-phosphorus), 3 Å of residue amine or residue hydroxy (acceptor-donor), and 4 Å of residue choline (acceptor-nitrogen) distance. Cumulative presence (Δt_{cumul}) of individual lipid-mediated salt bridges was calculated as combined fractional presence over the entire simulation time. The persistence factor F was calculated as shown in Equation 1,

$$F_{atom} = \Delta t_{cumul} \times (MSF_{max} - MSF_{atom} + MSF_{min}) \quad \text{(Eq. 1)}$$

where atom is either the lipid nitrogen, hydroxy oxygen, or phosphorus atom; MSF is the mean square fluctuation, and MSF_{min} and MSF_{max} equal 0.156 and 0.653 nm^2, respectively. Lipid-mediated salt bridges were retained if they showed $\Delta t_{cumul} > 10.0\%$.

Residue Conservation—LacY sequences were obtained by querying the NCBI nonredundant sequence data base of January 22, 2009, with BLAST (51), using the BLOSUM62 matrix and default parameters, and the *Escherichia coli* LacY sequence as query. All sequences with an *E*-value better than 1.0e-10 were retained and aligned using MUSCLE (52), resulting in 92 unique sequences with the following (selected) keyword occurrences: permease, 90; oligosaccharide/H^+, 54; galactoside, 39; hypothetical protein, 38; lactose, 31; transport, 21; symporter,: 20; and lactose-proton, 20.

RESULTS

Interactions between protein and lipids can manifest themselves in a number of ways as follows: lipid acyl chains can settle on a hydrophobic surface patch created by one or two TM helices; the lipid carbonyl or phosphate groups can act as acceptor for hydrogen bonds emanating from the protein; and salt bridges may be formed between opposing charges, *e.g.* phosphate and arginine or lysine. Hydrophobic interactions are the weakest of these, and the strongest interactions are made by salt bridges, which are in fact a particularly strong form of a hydrogen bond.

In the following sections we analyze the occurrence of individual and double hydrogen bonds and salt bridges. We then identify specific protein-lipid interactions from the presence of lipid-mediated salt bridges. The strongest of these is validated by additional simulations in lipid matrices with increasing degrees of methylation and correlated with experimental data and sequence conservation in the family and superfamily.

Hydrophobic Interactions and Single Hydrogen Bonds—Annular lipids were found to show limited diffusion (as calculated from the mean square deviation of lipid tails), in accordance with spectroscopic data (53), but no clear transition in distribution between annular and bulk lipids was observed. We could therefore not identify any single lipid as being associated with the protein on the basis of spatial fluctuation and hydrophobic interaction.

TABLE 3

Lipid-mediated salt bridges

Lipid-mediated salt bridges show a cumulative presence of more than 2%. Bridges run from the acceptor residue over the lipid amine (PE), choline (PC), or primary (PG_1) or secondary (PG_2) hydroxy through phosphate to the donor residue. Lipid matrices used are POPE, POPC, and POPG. Mean square fluctuation was calculated over the lipid nitrogen, hydroxy oxygen, and phosphorus atoms separately. Cumulative presence is Δt_{cumul}, and persistence factor F is calculated as outlined under "Materials and Methods." RMSF, root mean square fluctuation.

Lipid-mediated salt bridge				RMSF (nm^2)		Δt_{cumul}[e]	F_{donor}[f]	$F_{acceptor}$[g]
Acceptor-donor[a]		Lipid[b]		Donor[c]	Acceptor[d]			
						%		
Asp-44	Lys-102	p	PC	0.288	0.214	10.14	5.3	6.0
Asp-44	Asn-102	p	PG_1	0.536	0.375	5.04	1.4	2.2
Asp-44	Asn-102	p	PG_2	0.534	0.375	6.64	1.8	2.9
Asp-68	Arg-344	c	PE	0.505	0.506	9.41	2.9	2.9
Asp-68	Ser-209	c	PE	0.505	0.506	3.42	1.0	1.0
Asp-68	Lys-69	c	PE	0.213	0.213	36.84	22.0	22.0
Asp-68	Lys-69	c	PC	0.318	0.204	28.57	14.0	17.3
Glu-139	Asn-199	c	PE	0.286	0.282	4.97	2.6	2.6
Asn-166	Gln-167	p	PG_2	0.455	0.281	2.89	1.0	1.5
Asp-190	Arg-73	c	PE	0.347	0.331	5.19	2.4	2.5
Asp-190	Lys-74	c	PE	0.347	0.331	10.96	5.1	5.2
Asp-190	Lys-188	c	PC	0.180	0.205	7.39	4.6	4.5
Asp-190	Lys-74	c	PC	0.180	0.205	7.41	4.7	4.5
Asn-199	Arg-73	c	PG_1	0.361	0.271	9.66	4.3	5.2
Asn-199	Arg-73	c	PG_2	0.319	0.271	11.47	5.6	6.2
His-205	Lys-69	c	PG_1	0.340	0.232	3.78	1.8	2.2
Glu-215	Lys-211	c	PE	0.653	0.326	18.86	2.9	9.1
Glu-215	Lys-211	c	PC	0.546	0.277	6.60	1.7	3.5
Glu-215	Arg-218	c	PC	0.303	0.210	6.20	3.1	3.7
Gln-219	Lys-221	c	PC	0.443	0.175	3.71	1.4	2.4
Glu-255	Gln-256	p	PE	0.413	0.198	13.71	5.4	8.4
Glu-255	Arg-259	p	PC	0.267	0.218	2.83	1.5	1.7
Glu-255	Arg-259	p	PG_1	0.355	0.246	2.31	1.0	1.3
Glu-340	Gln-412	c	PG_1	0.286	0.210	2.31	1.2	1.4
Glu-374	Asn-371	p	PE	0.211	0.221	2.80	1.7	1.6
Glu-415	Asn-284	c	PE	0.280	0.228	11.13	5.9	6.5
Glu-415	Lys-335	c	PE	0.280	0.228	12.96	6.9	7.5

[a] Listed are hydrogen bond acceptor residues binding amine, choline, or hydroxy and donor residues binding phosphate.
[b] Membrane side is (c)ytosol or (p)eriplasm, and lipid matrix is PE, PC, or PG.
[c] RMSF of lipid nitrogen or hydroxy oxygen is shown.
[d] RMSF of lipid phosphorus is shown.
[e] Cumulative presence of lipid-mediated salt bridge over simulation is shown.
[f] Lipid hydrogen bond donor persistence factor F_{donor} is shown.
[g] Lipid phosphorus persistence factor $F_{acceptor}$ is shown.

LacY contains 17 negatively and 24 positively charged residues plus a number of otherwise hydrogen bond-capable residues; all these are potential candidates for protein-lipid-specific interactions. We determined for all hydrogen bond-capable residues of the protein whether they were at one time or another favorably bonded to a hydrogen bond-complementary (but not necessarily opposingly charged) lipid head group. For many of the candidate residues, a large population of distances indicative of binding was found, but it is impossible to distinguish between specific and nonspecific protein-lipid interactions. As expected, the strongest interactions observed were those involving negatively (Asp and Glu) or positively charged residues (Arg and Lys), the latter of which are responsible for the anchoring of a membrane protein in its environment (3). Interactions through the phosphate group are not lipid head group-specific, because they do not discriminate between different types of phospholipids. Also, single salt bridges and hydrogen bonds have a limited lifetime, with fluctuations on a 1–5-ns time scale (2). Lipids showing interactions through the amine, hydroxy, or choline group would therefore show too fast an exchange rate with bulk lipids to call them specific.

Lipid-mediated Salt Bridges—Lipids that are simultaneously bound to two different residues, *e.g.* through both their positively and negatively charged moieties, will show significantly longer residence times, and their diffusion away requires a two-step process. We therefore determined for every salt bridge and hydrogen bond, whether the same lipid was simultaneously bound to another salt bridge or hydrogen bond. Although cation-π interactions can be described using the current force field (31), and the PE amine group would be prone to such interactions, we have found no significant lipid-mediated salt bridges involving the π-clouds of aromatic residues. The list of lipid-mediated salt bridges is supplied in Table 3. As additional requirement, we impose the presence of at least 10% of the simulation time (amounting to 1 ns), resulting in a total of nine lipid-mediated salt bridges (listed in Table 4).

Of the nine lipid-mediated salt bridges, five, including the three strongest, showing a $\Delta t_{cumul} > 15\%$, involve both an acidic and basic residue, and four of these occur in the PE bilayer (the fifth one with PC). Two bridges are found involving a PC lipid, and only one with PG. Most lipid-mediated salt bridges show similarly strong interactions, with persistence factor values lying between 5 and 10. An exception is the bridge involving Glu-215 and Lys-211, which shows a weaker interaction on the amine side, with $F_{donor} = 2.9$, related to an increased local fluctuation (Table 3). Unmistakably, however, the interactions with Asp-68 and Lys-69, two neighboring residues, stand out, with significantly larger persistence factors. In addition, this lipid-mediated salt bridge is weaker in the case of PC and may therefore be responsible for PE preference. We validate the results by a closer investigation of this bridge, both in terms of the molec-

ular details of the simulation and sequence conservation in the lactose permease family and major facilitator superfamily.

Binding of PE to Asp-68 and Lys-68 in Lipid Matrices with Increasing Degrees of Methylation—In the independent studies of two MFS transporters, including LacY (54), transporter activity was investigated using PE lipids with various degrees of methylation. In both cases, a triple methylation (*i.e.* PC lipids) abolished activity completely, indicating a resident hydrogen on the ethanolamine moiety of PE to be essential (9, 54). We performed two additional simulations of LacY in lipid bilayer systems composed of mono- and dimethylated POPE lipids and analyzed the four simulations together to provide a molecular basis for PE specificity.

Fig. 1 (*E* and *F*) shows binding of Lys-69 to PE to occur already in the equilibration phase. Weak position restraints keep the protein heavy atoms in place, but the lipid is free to move, and the phosphate group of PE is gradually moving closer to Lys-69 to form a salt bridge at the final stages of energy minimization (POPE) or the first step of position restraint MD (Me-POPE, dimethyl-POPE, and POPC). The bridge is maintained throughout the remainder of the simulations, showing no significant difference in average distance.

The distance evolution between the lipid and Asp-68 is shown in Fig. 1 (*A–D*). Although a few lipids were found to bind to Asp-68 during the course of the simulations, in all four simulations a single lipid is continuously bound to Lys-69, and in every instance this is the same lipid that ultimately binds Asp-68. In the crystal and initial structure, Asp-68 is directed toward the inside of the protein and is hydrogen-bonded to Lys-131. It is remarkable that POPC shows a movement of the choline group away from Asp-68, and all three PE simulations an approach of the amine moiety to Asp-68 (Fig. 1*B*). After the side chain is released, the salt bridge is formed in the initial steps of the free MD simulation for POPE and monomethylated POPE. Data for the full simulation (Fig. 1*D*) shows binding to occur progressively later upon increased methylation. Fig. 1*D* shows initial binding of one of the two amine methyl groups at $t = 1.5$ ns, after which a rotation of the dimethylated amine group at $t = 2$ ns rotates the single hydrogen toward Asp-68. Binding of PC to Asp-68 (Fig. 1*D*) occurs only at $t = 2.4$ ns, and at an equilibrium distance more than 1 Å larger (0.37 *versus* 0.25 nm), making the rather significant difference between strong and weak hydrogen bonding.

Our observation is that a phospholipid (PE or PC) is recruited by Lys-69, followed by a rotation of the Asp-68 side chain to bind its amine or choline group, thus forming a double salt bridge or lipid-mediated salt bridge. Fig. 2 shows an image of this lipid-mediated salt bridge.

Sequence Conservation in LacY and the MFS—The wide variety of substrates of MFS transporters is reflected in the large sequence variation going from one subfamily to the other, but all members of the MFS share a highly conserved cytoplasmic motif (55) that includes the acidic Asp-68 and the basic Lys-69, Arg-73, and Lys-74. Individual mutations have shown Asp-68, but also a resident positive charge, to be required for activity (55, 56). In addition, Asp-68 was found to be related to the proton gradient-sensing mechanism but not to substrate transport (9, 56, 57). Our strategy for detecting lipid-mediated salt bridges directly and unambiguously identifies Asp-68 as the most relevant PE-interacting residue. This is the first study where the PE dependence of LacY (5) and the functional importance of Asp-68 (55) can be directly linked together.

TABLE 4

Lipid-bridged residue-residue contacts

Lipid-bridged residue-residue contacts show $\Delta t_{cumul} > 10.0\%$. The persistence factors F_{donor} and $F_{acceptor}$, calculated as outlined under "Materials and Methods," combine presence with spatial fluctuation. An increase in the cumulative presence and a decrease in the local fluctuation both lead to increased persistence factors.

Acceptor	Lipid	Δt_{cumul}	F_{donor}	$F_{acceptor}$	Donor
		%			
Asp-68	PE	36.8	22.0	22.0	Lys-69
Asp-68	PC	28.6	14.0	17.3	Lys-69
Glu-215	PE	18.9	2.9	9.1	Lys-211
Glu-255	PE	13.7	5.4	8.4	Gln-256
Glu-415	PE	13.0	6.9	7.5	Lys-335
Asn-199	PG	11.5	5.6	6.2	Arg-73
Glu-415	PE	11.1	5.9	6.5	Asn-284
Asp-190	PE	11.0	5.1	5.2	Lys-74
Asp-44	PC	10.1	5.3	6.0	Asn-102

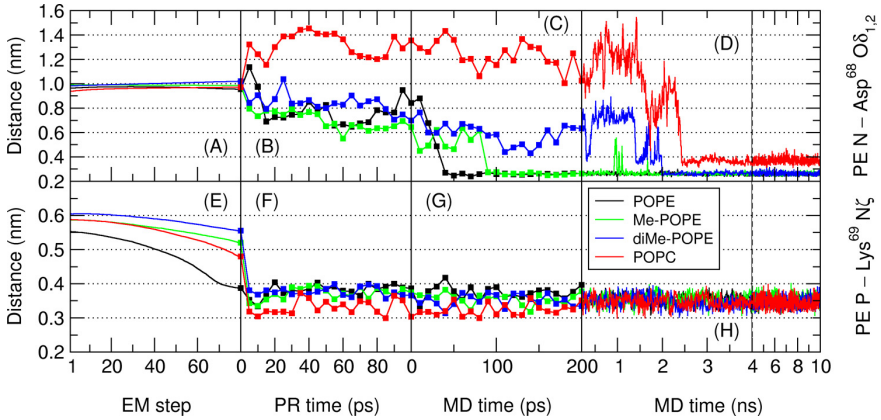


FIGURE 1. **Distances of Asp-68 and Lys-69 to the lipid.** Distance of Asp-68 Oδ atoms to lipid N (*A–D*) and Lys-69 Nζ to lipid P (*E–H*) for the last steps of the equilibration (*A, B, E,* and *F*) and for the MD simulation (*C, D, G,* and *H*) are shown. Weak position restraints on the protein heavy atoms are applied in the equilibration (*A, B, D,* and *E*), and any change in distance there is solely due to lipid movement. In the free simulation (*C, D, G,* and *H*), the entire system is free to move.

For the residues involved in the three strongest lipid-mediated salt bridges of Table 4, residue conservation in lactose permease is listed in boldface in Table 5. As opposed to the MFS, the multiple sequence alignment of (putative) lactose permease sequences is less restrictive and still shows Asp-68 and Lys-69 to remain the only such bridge between two conserved residues. The other lipid-mediated salt bridges are likely to be only phosphate-specific, due to the limited conservation of the acceptor residues involved, *e.g.* Asp-190 and Glu-215. Arg-73 and Lys-74 are located toward the periphery of the protein, at a small distance from Asp-68/Lys-69, and could therefore be

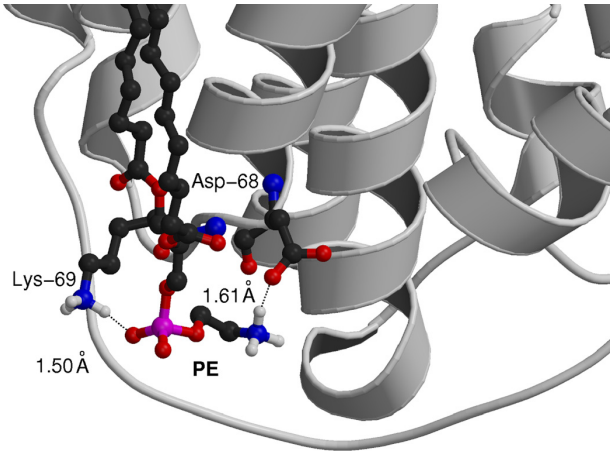


FIGURE 2. **Final structure of the POPE simulation showing the lipid-mediated salt bridge between Asp-68 and Lys-69.** Distances indicated are between the closest hydrogen and oxygen atoms. Only amine hydrogens are shown.

TABLE 5

Residue conservation

Residue conservation is in the lactose permease family of the residues involved in lipid-mediated salt bridges that show a cumulative presence of more than 2%. Those residues that are involved in lipid-mediated salt bridges with a cumulative presence of more than 15% are listed in boldface. Similarity score was calculated by using the following similarity grouping: Ala,Gly,Ile,Leu,Val; Cys,His,Met; Asp,Glu; Arg,Lys; Phe,Trp,Tyr; Asn,Gln; Pro; and Ser,Thr. In the alignment of LacY sequences, 152 out of 417 residues, or 37%, show a sequence similarity score above 75%. The primary residue is the most occurring residue at that position in the alignment, ignoring deletions and insertions.

LacY[a]	Similarity	Primary residue	Occurrence
	%		%
Asp-44	71.7	Glu	40.2
Asp-68	94.6	Asp	**94.6**
Lys-69	80.4	Lys	**67.4**
Arg-73	89.1	Lys	51.1
Lys-74	94.6	Lys	79.4
Asn-102	76.1	Asn	76.1
Glu-139	94.6	Glu	94.6
Asp-142	62.0	Arg	37.0
Asn-166	17.4	Pro	78.3
Gln-167	41.3	Asn	28.3
Lys-188	69.6	Lys	43.5
Asp-190	37.0	Glu	27.2
Asn-199	20.7	Asp	27.2
His-205	2.2	Pro	18.5
Ser-209	58.7	Ser	40.2
Lys-211	62.0	Lys	**44.6**
Glu-215	23.9	Ala	**25.0**
Arg-218	80.4	Lys	47.8
Gln-219	17.4	Met	33.9
Glu-255	46.7	Glu	37.0
Gln-256	23.9	Gln	23.9
Asp-259	34.8	Arg	29.4
Asn-284	67.4	Asn	67.4
Lys-335	89.1	Lys	73.9
Gln-340	29.4	Asn	25.0
Asp-344	82.6	Arg	63.0
Asn-371	10.9	Tyr	15.2
Glu-374	82.6	Asp	75.0
Gln-379	33.7	Gln	32.6
Gln-412	1.1	Arg	5.4
Glu-415	6.5	Glu	6.5

[a] Residue is in the LacY sequence.

involved in the initial recruitment of PE, shuttling a suitable phospholipid toward Asp-68 by the use of its flexible lysine side chain. In most LacY family members, a lysine is in fact found at position 73. Snorkeling actions of the lysine side chain are not an uncommon phenomenon (2, 3) and could in this case be related to lipid recruitment.

DISCUSSION

Specific Interaction between Asp-68 and PE That Contributes to the Gradient-sensing Mechanism—Using molecular modeling and dynamics, we identify a specific protein-lipid interaction in LacY between Asp-68 and PE by investigating the existence of lipid-mediated salt bridges of LacY embedded in five different lipid matrices. Our simulations show a consistent and strong hydrogen bond between (non-, mono-, and dimethylated) POPE amine groups and Asp-68 that is significantly weaker in the case of PC. In every instance, the bond is formed to a free hydrogen of the amine group with the speed of formation being inversely related to the degree of methylation. Asp-68-bound PE is initially bound to Lys-69 through its phosphate entity before being recruited by Asp-68. The nearby residue Lys-74 may have a facilitating function in the recruitment of PE. This lipid-mediated salt bridge is the only such bridge existing between conserved residues in LacY.

Asp-68 and Lys-69 are also part of a highly conserved motif in the MFS. Together with the PE dependence of MFS transporters (26) and the involvement of Asp-68 in the energy-coupling mechanism (9, 56, 57), the interaction between Asp-68 and PE may constitute a specific protein-lipid interaction involved in ΔpH sensing. This suggests that not only secondary and tertiary structure elements (25) but also functional elements such as the proton gradient-sensing mechanism have been conserved in secondary transporters during evolution. The fact that Lys-69 has not been found to be crucial for transport is in accord with our findings; only a single resident positive charge is required in the conserved motif (55), and only a single positively charged residue in principle suffices to recruit a phospholipid that can subsequently be shuttled toward Asp-68.

Role of Alternative Lipids—Although PE is reportedly required for proper LacY topogenesis and functioning (5, 58), it has been proposed that it can be replaced by a nonendogenous lipid, monoglucosyl-diacylglycerol (GlcDAG), a lipid found in *Acholeplasma laidlawii* (59). This may seem incompatible with the hypothesis proposed here considering the structural differences between PE and GlcDAG. We performed docking studies of the glucosyl moiety to identify binding modes with the LacY pocket. Surprisingly, the glucosyl can interact with both Asp-68 and Lys-69 in a PE-like fashion (supplemental Fig. S1). In the absence of the phosphate group, Lys-69 now interacts with the O_4 of the sugar ring, whereas Asp-68 interacts with O_3, leading to a motif conformation very similar to the one observed in the POPE simulation. In addition, O_2 is found to be hydrogen-bonded to the backbone NH of Lys-69. Note that in diglucosyl-diacylglycerol (DGlcDAG), the O_2 atom is involved in the connection between the two sugar rings and would therefore not be available for binding, although binding of the second sugar ring to the motif would be prohibited by the bulkiness of DGlcDAG as a whole. Although full-fledged MD simulations of LacY in GlcDAG and DGlcDAG bilayers should be done (a technically challenging process that goes beyond the scope of this paper), these results provide an explanation on why GlcDAG appears to allow LacY function and not DGlcDAG (64). In contrast, and

in accord with experimental data (54), the simulations of LacY in a POPG bilayer show no affinity of the glycerol side chain to the acidic Asp-68 (Table 1), and it is found to bind the phosphate groups of itself and neighboring lipids.

Implications of Results on the Functioning of Membrane Proteins—Our findings show that specific protein-lipid interactions may be identified through the identification of simultaneous occurrence of multiple (strong) interactions involving the same lipid. The occurrence of such interactions provides further evidence to the fact that specific protein-lipid interactions indeed play a role in the functioning of membrane proteins (4, 10). The significance of membrane proteins that constitute ~25% of genomic sequences (19) is not reflected in the number of membrane structures that are currently available in the Protein Data Bank. Yet membrane proteins represent 70% of current drug targets (60). Specific lipid-binding residues may therefore also present a yet unexplored area for drug targeting.

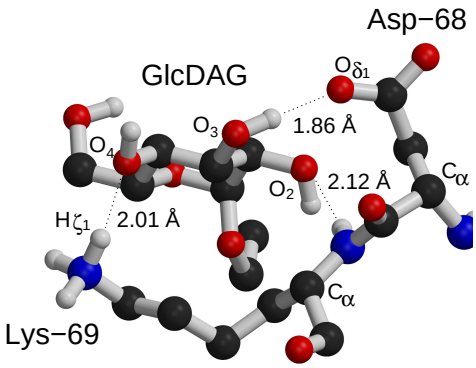
This is the first computational study ever to identify a specific protein-lipid interaction and provide the molecular basis for lipid species specificity at the same time. The technique applied here can be used on any membrane protein structure and embedded in a variety of lipid matrices, including mixtures with lipid species that do not form bilayer structures on themselves like cholesterol and cardiolipin. The approach is a thorough but simple application of molecular modeling and dynamics, a tried and tested technique.

REFERENCES

1. Vigh, L., Escribá, P. V., Sonnleitner, A., Sonnleitner, M., Piotto, S., Maresca, B., Horváth, I., and Harwood, J. L. (2005) *Prog. Lipid Res.* **44,** 303–344
2. Deol, S. S., Bond, P. J., Domene, C., and Sansom, M. S. (2004) *Biophys. J.* **87,** 3737–3749
3. Nyholm, T. K., Ozdirekcan, S., and Killian, J. A. (2007) *Biochemistry* **46,** 1457–1465
4. Lee, A. G. (2005) *Mol. Biosyst.* **1,** 203–212
5. Bogdanov, M., and Dowhan, W. (1995) *J. Biol. Chem.* **270,** 732–739
6. Wang, X., Bogdanov, M., and Dowhan, W. (2002) *EMBO J.* **21,** 5673–5681
7. Zhang, W., Bogdanov, M., Pi, J., Pittard, A. J., and Dowhan, W. (2003) *J. Biol. Chem.* **278,** 50128–50135
8. Zhang, W., Campbell, H. A., King, S. C., and Dowhan, W. (2005) *J. Biol. Chem.* **280,** 26032–26038
9. Hakizimana, P., Masureel, M., Gbaguidi, B., Ruysschaert, J. M., and Govaerts, C. (2008) *J. Biol. Chem.* **283,** 9369–9376
10. Powl, A. M., East, J. M., and Lee, A. G. (2008) *Biochemistry* **47,** 12175–12184
11. Lange, C., Nett, J. H., Trumpower, B. L., and Hunte, C. (2001) *EMBO J.* **20,** 6591–6600
12. Nogi, T., Fathir, I., Kobayashi, M., Nozawa, T., and Miki, K. (2000) *Proc. Natl. Acad. Sci. U.S.A.* **97,** 13561–13566
13. McAuley, K. E., Fyfe, P. K., Ridge, J. P., Isaacs, N. W., Cogdell, R. J., and Jones, M. R. (1999) *Proc. Natl. Acad. Sci. U.S.A.* **96,** 14706–14711
14. Wakeham, M. C., Sessions, R. B., Jones, M. R., and Fyfe, P. K. (2001) *Biophys. J.* **80,** 1395–1405
15. Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K., and Stevens, R. C. (2007) *Science* **318,** 1258–1265
16. Paila, Y. D., Tiwari, S., and Chattopadhyay, A. (2009) *Biochim. Biophys. Acta* **1788,** 295–302
17. Elofsson, A., and von Heijne, G. (2007) *Annu. Rev. Biochem.* **76,** 125–140
18. Aksimentiev, A., Brunner, R., Cohen, J., Comer, J., Cruz-Chu, E., Hardy, D., Rajan, A., Shih, A., Sigalov, G., Yin, Y., and Schulten, K. (2008) *Methods Mol. Biol.* **474,** 181–234
19. Carpenter, E. P., Beis, K., Cameron, A. D., and Iwata, S. (2008) *Curr. Opin. Struct. Biol.* **18,** 581–586
20. Guvench, O., and MacKerell, A. D., Jr. (2008) *Methods Mol. Biol.* **443,** 63–88
21. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O., and Shaw, D. E. (2009) *Curr. Opin. Struct. Biol.* **19,** 120–127
22. Guan, L., Mirza, O., Verner, G., Iwata, S., and Kaback, H. R. (2007) *Proc. Natl. Acad. Sci. U.S.A.* **104,** 15294–15298
23. Kaback, H. R. (2005) *C. R. Biol.* **328,** 557–567
24. Saier, M. H., Jr. (2000) *Microbiology* **146,** 1775–1795
25. Law, C. J., Maloney, P. C., and Wang, D. N. (2008) *Annu. Rev. Microbiol.* **62,** 289–305
26. Dowhan, W., and Bogdanov, M. (2009) *Annu. Rev. Biochem.* **78,** 515–540
27. Starling, A. P., East, J. M., and Lee, A. G. (1995) *Biochem. J.* **310,** 875–879
28. Yuan, C., O'Connell, R. J., Feinberg-Zadek, P. L., Johnston, L. J., and Treistman, S. N. (2004) *Biophys. J.* **86,** 3620–3633
29. Tieleman, D. P., and Berendsen, H. J. (1996) *J. Chem. Phys.* **105,** 4871–4880
30. Tieleman, D. P., and Berendsen, H. J. (1998) *Biophys. J.* **74,** 2786–2801
31. Lensink, M. F., Christiaens, B., Vandekerckhove, J., Prochiantz, A., and Rosseneu, M. (2005) *Biophys. J.* **88,** 939–952
32. Sprong, H., van der Sluijs, P., and van Meer, G. (2001) *Nat. Rev. Mol. Cell Biol.* **2,** 504–513
33. Lomize, M. A., Lomize, A. L., Pogozheva, I. D., and Mosberg, H. I. (2006) *Bioinformatics* **22,** 623–625
34. Yin, Y., Jensen, M. Ø., Tajkhorshid, E., and Schulten, K. (2006) *Biophys. J.* **91,** 3972–3985
35. Kandt, C., Ash, W. L., and Tieleman, D. P. (2007) *Methods* **41,** 475–488
36. Chandrasekhar, J., Spellmeyer, D. C., and Jorgensen, W. L. (1984) *J. Am. Chem. Soc.* **106,** 903–910
37. Marrink, S. J., and Mark, A. E. (2002) *Biochemistry* **41,** 5375–5382
38. Guest, M. F., Bush, I. J., van Dam, H. J., Sherwood, P., Thomas, J. M., van Lenthe, J. H., Havenith, R. W., and Kendrick, J. (2005) *Mol. Phys.* **103,** 719–747
39. Dupradeau, F. Y., Cézard, C., Lelong, R., Stanislawiak, E., Pêcher, J., Delepine, J. C., and Cieplak, P. (2008) *Nucleic Acids Res.* **36,** D360–D367
40. Van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R., and Tironi, I. G. (1996) *Biomolecular Simulation: The GROMOS96 Manual and User Guide,* Hochschulverlag AG an der ETH Zürich, Zürich, Switzerland
41. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005) *J. Comput. Chem.* **26,** 1701–1718
42. Berger, O., Edholm, O., and Jähnig, F. (1997) *Biophys. J.* **72,** 2002–2013
43. Berendsen, H. J., Postma, J. P., Van Gunsteren, W. F., and Hermans, J. (1981) in *Intermolecular Forces* (Pullman, B., ed) pp. 331–342, Reidel, Dordrecht, The Netherlands
44. Berendsen, H. J., Postma, J. P., DiNola, A., and Haak, J. R. (1984) *J. Chem. Phys.* **81,** 3684–3690
45. Essman, U., Perela, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995) *J. Chem. Phys.* **103,** 8577–8592
46. Miyamoto, S., and Kollman, P. A. (1992) *J. Comput. Chem.* **13,** 952–962
47. Hess, B., Bekker, H., Berendsen, H. J., and Fraaije, J. G. (1997) *J. Comput. Chem.* **18,** 1463–1473
48. Lensink, M. F. (2008) *Methods Mol. Biol.* **443,** 161–179
49. Christopher, J. A., Swanson, R., and Baldwin, T. O. (1996) *Comput. Chem.* **20,** 339–345
50. Yeagle, P. L., Bennett, M., Lemaître, V., and Watts, A. (2007) *Biochim. Biophys. Acta* **1768,** 530–537
51. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402
52. Edgar, R. C. (2004) *Nucleic Acids Res.* **32,** 1792–1797
53. Devaux, P. F., and Seigneuret, M. (1985) *Biochim. Biophys. Acta* **822,** 63–125
54. Chen, C. C., and Wilson, T. H. (1984) *J. Biol. Chem.* **259,** 10150–10158
55. Jessen-Marshall, A. E., Paul, N. J., and Brooker, R. J. (1995) *J. Biol. Chem.* **270,** 16251–16257

56. Yamaguchi, A., Ono, N., Akasaka, T., Noumi, T., and Sawai, T. (1990) *J. Biol. Chem.* **265,** 15525–15530
57. Mazurkiewicz, P., Poelarends, G. J., Driessen, A. J., and Konings, W. N. (2004) *J. Biol. Chem.* **279,** 103–108
58. Bogdanov, M., and Dowhan, W. (1998) *EMBO J.* **17,** 5255–5264
59. Xie, J., Bogdanov, M., Heacock, P., and Dowhan, W. (2006) *J. Biol. Chem.* **281,** 19172–19178
60. Lundstrom, K. (2007) *J. Cell. Mol. Med.* **11,** 224–238
61. Chiu, S. W., Clark, M., Balaji, V., Subramaniam, S., Scott, H. L., and Jakobsson, E. (1995) *Biophys. J.* **69,** 1230–1245
62. Mulliken, R. S. (1955) *J. Chem. Phys.* **23,** 1833–1840
63. Bayly, C. I., Cieplak, P., Cornell, W. D., and Kollman, P. A. (1993) *J. Phys. Chem.* **97,** 10269–10280
64. Wikström, M., Kelly, A. A., Georgiev, A., Eriksson, H. M., Klement, M. R., Bogdanov, M., Dowhan, W., and Wieslander, A. (2009) *J. Biol. Chem.* **284,** 954–965

Figure S1: Critical distances in the interaction of GlcDAG with Asp-68 and Lys-69.



www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 01 April 2016
Accepted: 28 October 2016
Published: 01 December 2016

On the ability of molecular dynamics simulation and continuum electrostatics to treat interfacial water molecules in protein-protein complexes

Guillaume Copie[1,2], Fabrizio Cleri[1], Ralf Blossey[2] & Marc F. Lensink[2]

Interfacial waters are increasingly appreciated as playing a key role in protein-protein interactions. We report on a study of the prediction of interfacial water positions by both Molecular Dynamics and explicit solvent-continuum electrostatics based on the Dipolar Poisson-Boltzmann Langevin (DPBL) model, for three test cases: (i) the barnase/barstar complex (ii) the complex between the DNase domain of colicin E2 and its cognate Im2 immunity protein and (iii) the highly unusual anti-freeze protein Maxi which contains a large number of waters in its interior. We characterize the waters at the interface and in the core of the Maxi protein by the statistics of correctly predicted positions with respect to crystallographic water positions in the PDB files as well as the dynamic measures of diffusion constants and position lifetimes. Our approach provides a methodology for the evaluation of predicted interfacial water positions through an investigation of water-mediated inter-chain contacts. While our results show satisfactory behaviour for molecular dynamics simulation, they also highlight the need for improvement of continuum methods.

Water is essential to all life that we know of. Its omnipresence in biological processes is generally assumed and as a consequence its molecular structure is often ignored when devising theories about the molecular details of those processes. At a coarse-grained scale, solvent effects are responsible for phenomena such as electrostatic screening, where they effectively reduce the electrostatic field of the molecules they surround[1], and the hydrophobic effect[2], which leads to self-assembly or aggregation of nonpolar molecules in an attempt to minimize their interaction with water[3]. Insights in the molecular details of the aggregation reveal that the hydrophobic effect is an entropy-driven process, which is fueled by an increased freedom for the water molecules to engage in hydrogen bonding. The difference in hydrogen-bonding behaviour between bulk water and water molecules at the water-detergent interface have been made evident by time-resolved vibrational spectroscopy, showing two distinct but exchanging populations[4].

Water molecules also solvate protein structures. Due to the polar nature of the protein surface, the change in properties between the solvation shell and bulk water is smaller as opposed to the detergent-water interface, but nevertheless distinct: a notable difference in the water-water hydrogen bonded network can be observed. This phenomenon manifests itself in various ways in different systems, as the following short list shows. An investigation of the dynamic properties of water around simple polypeptides shows the formation of a pseudo-rigid structure around the peptide core that exhibits stronger hydrogen bonding and with longer lifetimes[5]. The formation of networks of hydration water molecules around protein structures had been described before from the investigation of cryogenic X-ray structures of bovine beta-trypsin[6]. It was found that the water network allowed exposure of the active site to bulk solvent, thereby avoiding the hampering of its protease activity. Gradients of coupled protein-water motions have also been observed near the MT1 metalloprotease active site[7] and the so-called ice-binding plane of antifreeze proteins[8]. The water shell around the p53 core domain has been described as also

[1]University Lille, CNRS, UMR8520 IEMN, Lille, F-59000, France. [2]University Lille, CNRS, UMR8576 UGSF, Lille, F-59000, France. Correspondence and requests for materials should be addressed to R.B. (email: ralf.blossey@univ-lille1.fr) or M.F.L. (email: marc.lensink@univ-lille1.fr)

consisting of two such regimes: a dynamical one, showing fast exchanges with bulk water that are unambiguously assisted by local protein motions, and a structural one that contributes to the structural integrity of the protein[9].

Crystal structures often show crystal waters associated to the protein. An analysis of water molecules within a 5 Å shell around dimeric crystal structures has shown preferential binding to protein-protein interaction interfaces, whether these are true interfaces or crystal contacts[10]. Water unmistakably influences structure, function and stability of proteins and protein complexes[11–14]. The prototypical example for strong electrostatic binding, the high-affinity barnase/barstar complex, features a large amount of water molecules in its protein-protein interface, a third of which are fully buried[15]. The association between barnase and its inhibitor barstar is an extreme example of hydrophilic association, which is characterized by the anisotropy of interfacial water molecules that contribute to the association funnel[16]. The opposite extreme, hydrophobic association, exhibits a bimodal binding where hydrophobic dewetting takes place after initial long-range electrostatic attraction.

Interfacial water molecules are also crucial to both stability and specificity of colicin DNase-immunity protein complexes[17]. The complex between the DNase domain of colicin E2 and its cognate Im2 immunity protein[18], resolved at 1.72 Å resolution at a temperature of 100K, featured as Target 47 in the CAPRI protein docking experiment[19]. With high-quality template structures of the complex available in the PDB, both cognate and non-cognate, the focus of the experiment lay in the prediction of the water positions at the protein-protein interface. It was clear from the experiment that further work in the prediction of interfacial water molecules was required. Nevertheless, the results were encouraging: several of the conserved water molecules, which define the interface hot spot, were correctly predicted, as was at least one of the water molecules responsible for the specificity for the family of complexes[19]. The more sophisticated methods employed, which often combine the use of classical empirical force fields with additional sampling and energy minimization, were found to be more successful than simpler water placement methodologies.

When studying the microscopic interaction of proteins and protein complexes with its solvating environment, molecular dynamics (MD) simulation seems the method of choice, as it not only allows to sketch a molecular picture of the interactions involved, but also provides a dynamical viewpoint. In terms of protein-water interaction, MD has been used to study the microscopic dynamics of water around unfolded proteins[20] or to look at diffusion around intrinsically disordered proteins[21] but it also allowed to establish the existence of coupled interactions between two distant proteins that were mediated by water[22].

Continuum electrostatics methods, on the other hand, have so far been mainly employed for a quantification of the energetics of protein interactions. Their common assumption relies on a constant permittivity of the solvent, both in the Generalized Born (GB) approach and the Poisson-Boltzmann (PB) theory. The latter relies on the partial differential equation for the electrostatic potential $\phi(\mathrm{r})$.

$$\nabla \cdot \varepsilon(\mathrm{r})\nabla\phi(\mathrm{r}) + G(\phi(\mathrm{r})) = \rho(\mathrm{r}) \tag{1}$$

where $\varepsilon(\mathrm{r})$ is the dielectric function, typically chosen as a constant value inside the protein ($\varepsilon \approx 2$–4) and in water ($\varepsilon \approx 80$); $G(\phi(\mathrm{r}))$ is a generally nonlinear function of the electrostatic potential of the mobile charges; and $\rho(\mathrm{r})$ is the charge distribution of the fixed charges. These theories are computationally less demanding than an all-atom description of solvent effects. However, they have been found to be underperforming for large-scale simulations[23] and unusable for protein design applications[24], due to an underestimation of the hydrophobic forces, leading *e.g.* to burial of salt bridges. Recently, microscopic details of solvent structure have been integrated into the PB-approach leading to formulations of continuum electrostatics with explicit solvent. A simple continuum electrostatics model with explicit water is the Dipolar Poisson-Boltzmann Langevin (DPBL) model[25,26]. While also being a mean-field theory for the electrostatic potential $\phi(\mathrm{r})$ of the system, this model goes beyond the usually employed PB theory—which is also a mean-field theory–by explicitly introducing solvent molecules in the form of point dipoles. In Eq. (1) this amounts to the introduction of a dependence of the dielectric permittivity on a nonlinear function of $\phi(\mathrm{r})$ via

$$\varepsilon(\mathrm{r}) \rightarrow \varepsilon(F(\nabla\phi(\mathrm{r})) \tag{2}$$

where F is a nonlinear function, resulting in terms of higher powers of $\nabla^2\phi$ in the DPBL equations as compared to the PB equation[27].

This approach underlies the Marcus functional employed in studies of electron transfer[28], which is equivalent to the DPBL-model in its linearized form[29,30], and was also employed in the work by Warshel and Levitt[31]. The DPBL-model lies at the basis of a dedicated solver, AquaSol[27], built on an original PB solver[32]. It has previously been applied *e.g.* to the computation of SAXS profiles based on an extension of the AquaSol solver by the AquaSAXS module[33], and it was also used to predict free energies of amyloid fibril aggregates[34].

In this work, we apply both MD simulations and the DPBL model as implemented in the AquaSol server for the prediction of water positions at protein-protein interaction interfaces. We are looking at phenomena related to the motion of water molecules, and use the crystal structure as frame of reference, hence relatively short simulations suffice, as they only need to allow the water molecules to diffuse and no large-scale motions of the proteins are involved. In order to convince ourselves of the correctness of this assumption we have monitored the evolution of the so-called f^w(nat) value (defined below), which for all our chosen complexes rapidly settles around a well-defined mean-value (data not shown).

As our study systems we have chosen three distinct but challenging systems to test the methodologies, which are the two complexes already discussed before: (i) the barnase/barstar complex[15] (hereafter abbreviated by Barnase) and (ii) colicin DNase E2/Im2 protein complex[18] or CAPRI T47[19], and furthermore, (iii) the antifreeze protein Maxi[35,36]. All systems are shown in Fig. 1. Maxi is a four-helix bundle formed by head-to-tail dimerization of two 195-residue polypeptide chains. With a length of close to 15 nm and a diameter roughly one tenth of this value, it exhibits an unusual hydration of its protein core: several hundreds of water molecules form an

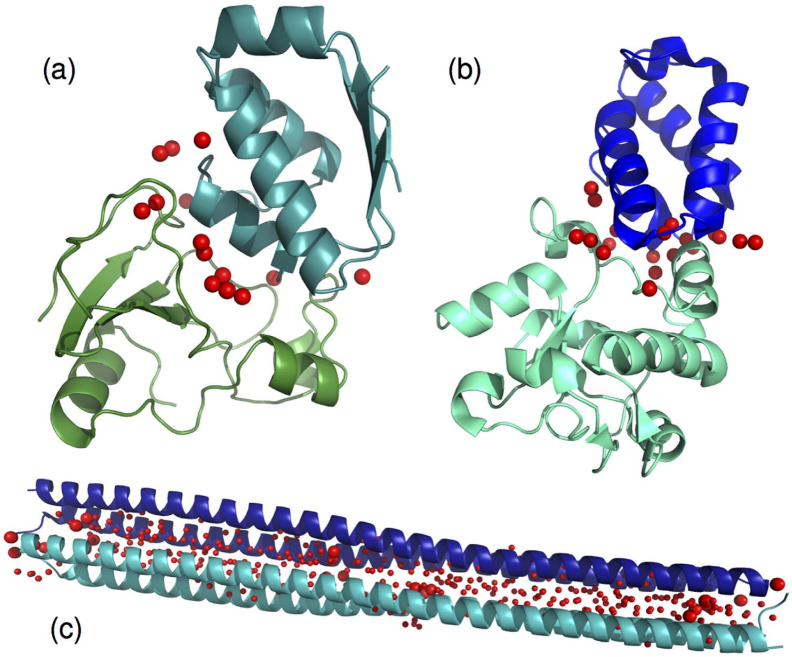


Figure 1. The three protein complexes. (**a**) Barnase/barstar, (**b**) E2/Im2, and (**c**) Maxi. Individual protein chains are colored differently. Red spheres indicate interfacial water molecules for the three systems. The smaller red spheres indicate additional core water molecules for Maxi (see text for definition).

elongated and dynamic water network – a counter example to the common cases in which water is essentially not present *within* a protein core. This case will therefore necessitate a more detailed discussion of the water molecule positions.

As the aim of this study is to investigate the performance of classical mechanical and continuum solvation methodologies in the treatment of interfacial water molecules in protein complexes, our choice reflects three different aspects of relevance for studies of water positions:

- Whereas hydrophobic association results in a protein-protein interface devoid of water molecules, the barnase/barstar complex, which has been extensively studied by computational and experimental means, is considered the prototypical example for hydrophilic association;
- E2/Im2 has been the target of the CAPRI protein docking experiment and therefore offers a direct comparison with a multitude of methodologies used for the prediction of water positions and thus easily allows the research community for additional checks;
- The Maxi protein has been chosen because the behaviour of its core water molecules falls in between the two regimes with the static interfacial waters on the one hand and bulk water on the other. It is as distinctly different from regular protein-protein interfaces as one can get, without becoming bulk water.

Results and Discussion

We first need to clarify the concepts that we used to characterize the different water molecules in the systems. We distinguish between:

- *Water mediated contacts and interfacial waters.* This notion refers to the water molecules that are found in the overlap of two water shells considered around both the receptor and the ligand, hence the waters are shared by both. There are two measures that can be used for the waters within this shell: the first is the recall of native water positions, which is the ratio of predicted waters within a certain distance to those in the PDB template within this overlap region. The second is the recall of water-mediated ligand-receptor contacts f^w(nat), which we use here. A contact occurs when any two (heavy) atoms of a ligand and receptor residue pair are found within a distance of 3.5 Å or less from the same water molecule, see Fig. 2; this definition is the same as in ref. 19, where it has also been shown that these two measures correlate. As Fig. 2 also shows, the number of such contacts can be larger than the actual number of water molecules.
- *Associated waters.* These water positions refer to the water molecules that are found in the water shells surrounding the proteins but exclude the interfacial waters.
- *Core waters.* We introduce this notion for the discussion of the Maxi complex as it contains a large number of waters between the chains. Core waters are those that were shared by the water shells around each of the four chains making up the complex. We define a water molecule to be in the core of the protein if the distance

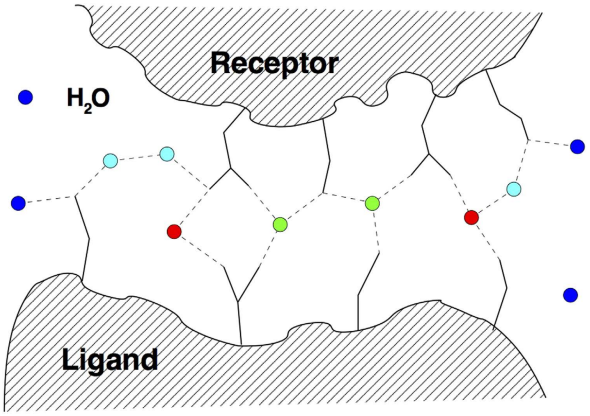


Figure 2. Schematic illustration of water-mediated inter-chain contacts at the interface of a protein-protein complex. Water molecules are indicated as colored circles, with red water molecules engaging in one, and green waters in multiple water-mediated contacts. Blue surface waters are bound to a single one or none of the entities (ligand or receptor) and do not contribute to the water-mediated contact list. Cyan water molecules are engaged in water-mediated ligand-receptor contacts mediated by two water molecules.

System	Barnase	E2/Im2	Maxi
Protein atoms	3159	3612	4760
Water molecules	15659	11298	11662
Ions	4 Na^+	4 Cl^-	2 Na^+, 2 Cl^-
Box size (nm^3)*	$7.9 \times 7.9 \times 7.9$	$6.6 \times 10.1 \times 5.7$	$4.7 \times 4.6 \times 17.8$
Temperature (K)	300	300	273
Interfacial waters[a]	15	22	22
Water-mediated contacts[a]	20	41	29
Diffusion coefficient in bulk[b]	3.8	4.1	2.3
Residence time of associated water	8 ps	8 ps	7 ps

Table 1. Initial and computed characteristics of the MD simulations of the three systems. [a]Values change over the course of the simulation, initial values reported here. [b]Diffusion constant $\times 10^5$ cm^2/s.

between its oxygen atom and at least one of the protein heavy atoms of both α-helices is less than 1.1 nm. If the interfacial water definition were used, the Maxi complex would only contain a 22 of such waters (see also Fig. 1c).

- *Bulk waters.* These are the water molecules that are not influenced by the presence of the complex.

Table 1 summarizes the setup of the MD simulations and the computed interfacial waters and water-mediated contacts for the three complexes. An interfacial water is defined as being in contact (distance less than or equal to 3.5 Å) with both protein chains (ligand and receptor) simultaneously. A water-mediated contact is a ligand-receptor contact running over an interface water molecule. The reference number of contacts is 20 for Barnase, 41 for E2/Im2 and 29 for Maxi. For Barnase, three copies of the complex are found in the crystal structure, we have only retained in the analysis those contacts that occur in all three complexes. For Maxi, chains A and B were chosen as ligand and receptor entities, respectively.

The table also reports on the computed diffusion constants and water residence times. The diffusion constant estimates are in line for Barnase and E2/Im2, and amount to $4 \cdot 10^5$ cm^2/s. Although these values are an overestimation with respect to experimental measurements[37], they correspond well to earlier reported values for SPC water[38,39]. A lower diffusion of $2.3 \cdot 10^5$ cm^2/s is found for Maxi, which can be accounted for by the lower simulation temperature of 273 K. The residence times of associated water molecules lie in the order of 7 to 8 ps and agree with previous calculations[36]. In our comparison with the results of[36], which were obtained with the TIP3 water model, we obtain the same bimodal distribution of intermolecular water angles, and we can also reproduce the particular water network structures inside the protein, so that we are confident that the differences in the water models do not play a role for our results. The lower value found for Maxi can be attributed to an increased fluidity of the environment, which is due to the presence of alanine residues in the protein, which are also accessible from the outside.

Looking more into detail at the interfacial water molecules, we use the recall of water-mediated inter-chain contacts, f^w(nat), to estimate the quality of prediction. The measure is readily calculated from a single MD frame, where the water molecules can either be directly used, or *a posteriori* placed by AquaSol, giving rise to the

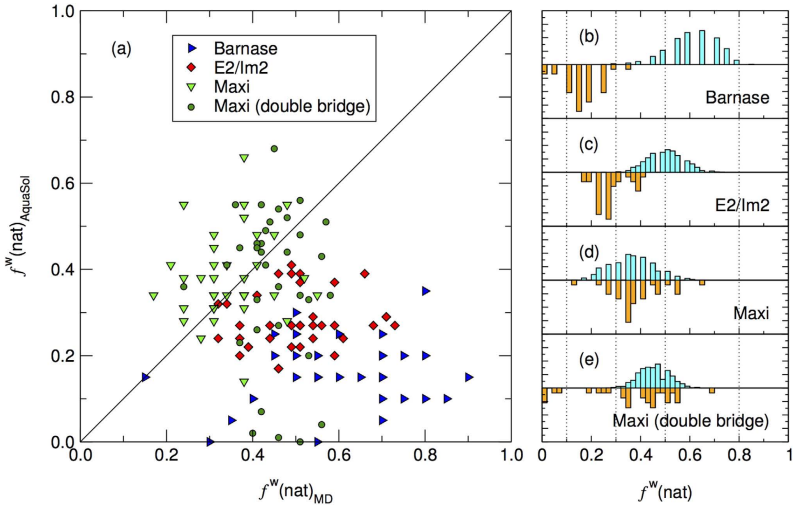


Figure 3. Comparison of MD simulation with AquaSol. (**a**) f^w(nat) values, AquaSol *vs.* MD simulation, for the three complexes. (**a**) Barnase, blue triangles; E2/Im2, red squares; and Maxi, green triangles (single bridge) and green circles (double bridge). For the interpretation, see main text. (**b–e**) Distribution of f^w(nat) values from MD (light blue) and AquaSol (orange). Dotted vertical lines delineate category of prediction quality[19], going from *bad* ($f^w(nat) < 0.1$) to *outstanding* ($f^w(nat) \geq 0.8$).

quantities $f^w(nat)_{MD}$ and $f^w(nat)_{AquaSol}$, resp. $f^w(nat)_{AquaSol}$ values have been calculated for a representative selection of configurations, chosen at random. Those value pairs are shown exhaustively in Fig. 3a. For reasons of clarity, we discuss first the two protein complexes, and then Maxi.

Barnase/barstar and E2/IM2. The MD results for the prediction of interfacial waters in both complexes are very good according to the evaluation scheme of ref. 19. The top panel of Fig. 3b (light blue bars) shows that the f^w(nat) values for Barnase fall in the range $0.5 \leq f^w(nat) < 0.8$ and can therefore be termed "excellent". For E2/Im2 (Fig. 3c), the values are slightly lower, with a distribution centering on $f^w(nat) = 0.5$, balancing between the "good" and "excellent" categories. Such values lie a tenth of a point lower than what was found previously[19], but it should be noted that those simulations held the relative position of the monomers fixed, whereas in our work everything was free to move. One can also assume that our simulation temperature of 300 K will have an adverse effect on f^w(nat) values, as the crystal structure was solved at cryogenic temperatures (100 K).

AquaSol predictions are significantly worse for these two complexes (Fig. 3b and c, orange bars): for the barnase complex they can at best be considered as "fair", while for E2/IM2 the best values are only found at the lower tail of the MD distribution values ("good"). The number of interfacial waters in both complexes differs slightly, between 15 for Barnase and 22 for E2/Im2. However, the number of contacts that these waters mediate is significantly different and goes from 20 for Barnase to 41 for E2/Im2. An investigation of the composition of the protein-protein interface in both complexes reveals a more polar interface for E2/Im2, with a charged/polar/non-polar ratio of 47%/24%/29%, as opposed to 34%/29%/37% for Barnase. In the more polar interface of the E2/Im2 complex, two-thirds of all water-mediated contacts involved charged or polar residues on both sides, whereas this is only half for Barnase (data not shown).

So it would seem that continuum methods show better performance as the interface becomes more polar. This is not surprising, since placement of water molecules in the AquaSol methodology occurs on the basis of electrostatic potential and the larger this potential is, the sooner a water molecule will be placed at that location. More generally stated, since in continuum methodology a more polar environment can be translated into a higher dielectric constant, the more this environment resembles bulk water, the better the performance will be.

The nonetheless disappointing AquaSol results are illustrated in Fig. 4a. Both protein partners are plotted in surface and cartoon representation, and there is only very little overlap between the MD (red) and the AquaSol waters (blue). Figure 4b on the other hand, nicely shows the functioning of AquaSol. The figure shows a top view of E2/Im2, showing the circumferential positioning of clusters of water molecules. For each cluster of two or three MD waters, AquaSol places one or two water molecules. Interestingly, such conservative placement was also observed in ref. 19, where an analysis of different water placement methodologies was performed and typically only one water molecule per cluster of waters was recovered.

Figure 4b also shows three water molecules at the center of the interface, which are not recovered by AquaSol. These waters represent buried water molecules. Table 2 gives a detailed look at the AquaSol predictions for the residues at the interface of the E2/Im2 complex. The Table, which follows Table S1 of ref. 19, shows that most of the water molecules that are in contact with bulk water (*i.e.* they are not buried) can at one point be recovered by AquaSol. However, the table also shows that it is much more difficult to recover contacts involving buried waters. Three of the nine buried waters can be recovered by AquaSol. These waters are essentially buried waters 7, 8, and

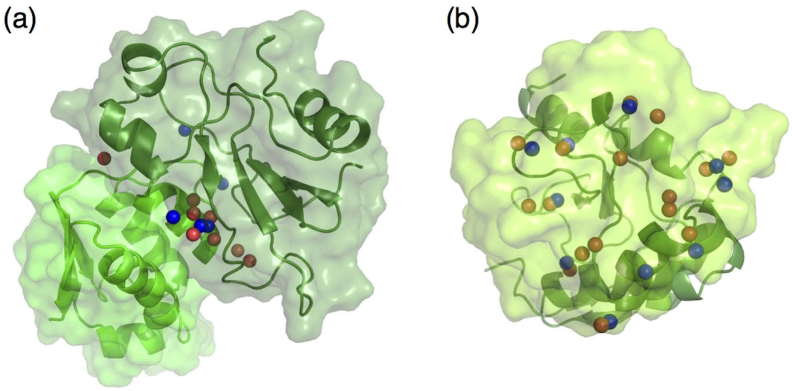


Figure 4. Comparison of interface-water positions generated by AquaSol (blue spheres) to those occurring in MD (red spheres). (**a**) Barnase (dark green) and barstar (light green), both plotted in surface and cartoon representation. (**b**) Top view of DNase E2 (above) in cartoon and Im2 (below) in surface representation. The f^w(nat) values corresponding to these frames are 0.7 (MD) and 0.25 (AquaSol) for Barnase, and 0.73 (MD) and 0.27 (AquaSol) for E2/Im2.

9 of ref. 19, which are the most solvent-exposed ones (the numbering is on increasing exposure). Buried water 2, mediating a contact between Asp-62A and Ser-74B, is only recovered because it forms a cluster with water number 9, and it is not the actual water that is recovered, but its contacts that are captured by the other water molecule. The most buried water molecules (numbers 1–6) are categorically not recovered by AquaSol. This includes buried water 5, which although buried, mediates a highly polar contact, between Asp-33A Oδ_2 and Arg-98B Nη_2.

Maxi. Maxi has been the topic of an MD study which investigated the properties of core water molecules[36]. In our work, we revisit these findings using a different water model, SPC as opposed to TIP3P. Calculating the distribution of angles of water molecules in the protein core, we find a bimodal distribution with peaks at 10 and 46 degrees, in agreement with the values obtained previously[36]. Our simulations also reproduce the pentagonal structures of the water network, albeit not as complete as observed in the crystal structure. Revisiting Table 1, the diffusion constants of bulk water are also in accord, with values of $2.3 \cdot 10^{-5}$ cm^2/s and $2.2 \cdot 10^{-5}$ cm^2/s for 'random' and 'crystal', resp., *vs.* $3.1 \cdot 10^{-5}$ cm^2/s found by Sharp[36]. The difference can be accounted for by the chosen water model. TIP3P, the water model used by Sharp, is known to lead to faster diffusion than SPC[39]. Inside the protein core, a diffusion constant of $0.3 \cdot 10^{-5}$ cm^2/s is found (for 'random', and $0.4 \cdot 10^{-5}$ cm^2/s for 'crystal') *vs.* $0.7 \cdot 10^{-5}$ cm^2/s by Sharp, making the effect of the confinement of the water molecules in the protein core slightly more pronounced in our description. This is also reflected in the residence times of water molecules outside the core region (protein-associated waters), where we obtain a value of 7 ps compared to the value of 8 ps found by Sharp. Residence times of water molecules inside the core were found to be 14 ps on average. In the crystal structure, the number of water molecules as defined to be in the core region is 321. We find an only slightly lower amount of 296 ± 16 water molecules in this region over the course of the simulation.

Of the 321 core water molecules only 22 can be considered interfacial water molecules, and these make a total of 29 water-mediated contacts. In the crystal structure, these molecules are primarily found at the center and edges of the protein structure, as shown in Fig. 1. The recall of these water molecules can on average be called "good", with the majority of recall values lying in the range $0.3 \leq f^w(\text{nat}) < 0.5$, both for MD as well as AquaSol-predicted waters. Nevertheless, this means that 50–70% of contacts are false positives. Looking into detail at the distribution of these waters, we observed that during the course of the simulation, contact-mediating waters are distributed all along the protein, and their number increases substantially, to as many as 193 at the end of the simulation. This cannot be due to the water model, since a quick test with TIP water shows 196 of such waters at the end of a 10 ns simulation. One possible explanation could be that current water models are too tightly packed in the vicinity of proteins, or hydrophobic residues in particular, and, therefore, also at the core of Maxi, which contains a large amount of alanines.

Core waters of Maxi do not show the same behaviour as interfacial water molecules of "regular" protein complexes, but fall into an intermediate regime between interfacial and protein-associated waters, evidenced also by the extended residence time of 14 ps, which is twice as long as the residence time of protein-associated water. In order to capture the more diffuse behaviour of core water molecules, we have devised an extension of the water-mediated inter-chain contacts, called double-bridge contacts, where inter-chain contacts can be mediated by *two* water molecules. This is illustrated in Fig. 2. In the crystal structure, 95 such water molecules can be found, which are involved in 78 contacts. The MD simulation shows an average of 117 ± 17 water molecules involved in double bridge contacts, which is acceptable albeit slightly higher than the crystal structure and reflects the more compact behaviour of water, as mentioned just before.

The distribution of double-bridge f^w(nat) values for Maxi, shown in Fig. 3e, shows them to lie in the range $0.3 \leq f^w(\text{nat}) < 0.6$, which is an improvement over the single-contact f^w(nat) values and only marginally weaker

	Ile 22A	Cys 23A	Arg 24A	Glu 26A	Gly 27A	Glu 30A	Glu 31A	Asp 33A	Asn 34A	Arg 38A	Glu 41A	Ser 50A	Asp 51A	Ile 53A	Tyr 54A	Tyr 55A	Pro 56A	Asp 58A	Asp 62A
Gln 70B																**W**			
Lys 72B															**B1/6**	**W**	*W*	*W*	
Gly 73B														**B2**					*B2*
Ser 74B														**B2**	**B1**				*B2/9*
Asn 75B															**B1/6**				
Thr 77B		*W*																	
Asn 78B	**B4**							**B5**							**B4**				
Lys 81B			*W*	*W*															
Gly 82B						*W*													
Lys 83B					**W**	*W*													
Ala 87B												**B3**	**B3**						
Arg 88B													*W*						
Lys 89B													*W*						
Lys 90B													*W*						
Gln 92B											**B7**	*B7*	**B3**						
Gly 95B							*W*		*W*										
Glu 97B									*W*	*B8*									
Arg 98B						*W*		**B5**	*W*										

Table 2. Table listing the water-mediated ligand-receptor contacts of the E2/Im2 complex, for the frame corresponding to Fig. 4a. Residues involved in such contacts show a **W** at their intersection, and a **B** if the water molecule responsible for the contact is buried. The table and the numbering of the buried waters follow Table S1 and Fig. 1 of ref. 19, resp. Waters that are recovered by AquaSol are listed in Italics.

		f^w(nat			Number of water molecules			
Barnase		**MD**	**PBL**	**Yukawa**	**AquaSol**	**MD**		
	$t = 3.8$ ns	0.85	0.15	0.15	16 (3)	17 (8)		
	$t = 8.1$ ns	0.80	0.30	0.30	16 (6)	22 (10)		
E2/Im2		**MD**	**PBL**	**Yukawa**	**AquaSol**	**MD**		
	$t = 1.8$ ns	0.73	0.27	0.27	31 (10)	27 (17)		
	$t = 2.5$ ns	0.71	0.29	0.29	27 (11)	27 (18)		
Maxi		**MD**	**PBL**	**Yukawa**	**AquaSol**	**MD**	**AquaSol**	**MD**
	$t = 0.5$ ns	0.41	0.34	0.34	42 (9)	28 (10)	323	303
	$t = 5.3$ ns	0.62	0.38	0.38	56 (8)	45 (15)	359	304
	$t = 10.0$ ns	0.66	0.38	0.38	59 (9)	36 (13)	194	224
					Interface		Core	

Table 3. Values of the f^w(nat) coefficient calculated from MD and AquaSol-predicted water positions of selected simulation frames. The selected frames are hand-picked and correspond to some of the best values obtained for MD. The number of interface (and core for Maxi) water molecules is also listed, with the number of these molecules involved in native contacts in parentheses.

than those of E2/Im2. With some exceptions, most of the AquaSol values can be found in the same region, indicating that AquaSol is capable of recovering a great deal of the core water positions of Maxi.

Summarizing. Our findings can be summarized in Table 3, which shows representative frames belonging to the best results obtained for f^w(nat) for the three systems, for both MD simulation and AquaSol.

We conclude from the results that molecular dynamics simulation with explicit solvent is quite capable of treating the dynamics of interfacial water molecules, even though crystal water molecules in the interface had been removed prior to solvation, with later stages of the simulation revisiting the native-like initial organization of the crystal structure, as evidenced by f^w(nat) values around 80% for Barnase and E2/Im2. The AquaSol f^w(nat) values however, are systematically lower than the MD values, and rarely exceed 40%. Inclusion of the Yukawa potential has little to no effect on these values. For both Barnase and E2/Im2, the number of interfacial water molecules obtained by both methods is an overestimation with respect to the crystal structure, but comparable. However, the number of those molecules involved in native contacts is severely inferior for AquaSol in the case of E2/Im2, and even more so for Barnase. The slightly better performance of AquaSol on E2/Im2 is due to the more polar nature of the protein interface.

For Maxi, the story is slightly more complicated. Core waters in Maxi only contribute to a limited number of water-mediated contacts, and this number is severely overestimated in the simulations. Here lies also the reason

for the high recall values, which are simply due to the high number of interfacial waters recovering contacts by chance. But the measure is not meaningless. The 22 interfacial water molecules as defined by our measure mediate true inter-chain contacts, and those contacts are probably important for the structural integrity of the protein. It is reassuring to observe that those waters show "good" recall values, both for MD and AquaSol. Our extended measure of the double-bridge contacts covers better the water molecules in the core of the protein, capturing about a third of the core waters. Also here, f^w(nat) values occupy a satisfying range around 50% recall. Starting the MD simulation with the crystal water positions has little influence on the results, which had been concluded previously for barnase/barstar as well[19].

The simulation of Maxi shows an increase in number of interfacial water molecules, which can only be due to both protein chains getting closer together. We hypothesize that this results from an underestimation of the repulsion between water and hydrophobic residues of the protein, notably alanines. Incidentally, this is likely also the reason for the "excellent" $f^w(\text{nat})_{MD}$ recall values for Barnase. With the number of core waters being in strong agreement with the crystal structure and the protein chains finding themselves closer together, we conclude that the core waters in our simulation show too large a diffusion and are not as *ice*-like as they should be. It has been argued before that simple and local corrections to empirical force fields may significantly improve the solvation of proteins[40], but the Maxi systems shows that particular care should be taken to ensure a proper treatment of the hydrophobic protein-water interactions.

Conclusions. In this work we have investigated the positions of interfacial water molecules from molecular dynamics simulations and an explicit-solvent continuum model. Based on three challenging examples which reflect different types of interfacial waters, we have presented a methodology to classify these waters and quantify the prediction of water positions of the computational approaches. As a general trend, exemplified by the comparison of the results in Fig. 3, we observe that the MD-based recall values are better than those obtained from AquaSol, with the discrepancy largest for the barnase/barstar complex. Given the simulation temperature, the recall values for MD of 40–70%, labelled "good" to "excellent", are satisfactory. For these systems, AquaSol is unable to recover buried water molecules, even when these are involved in highly polar inter-chain contacts. As one would expect from a continuum theory, the agreement with MD is best for Maxi. Here, the overall low level of prediction must be attributed to the more dynamic, 'bulk'-like behaviour of the waters. Future work in improving in particular the continuum approach, which has the advantage of computation speed, must lie both in the development of more sophisticated water models, but also in going beyond the mean-field approximation.

Methods

MD simulations. The three-dimensional coordinates of the systems were retrieved from the Protein Data Bank, entries 1BRS (Barnase/barstar, chains A:D)[15], 3U43 (E2/Im2)[19] and 4KE2 (Maxi)[35]. Missing atoms or residues were added with the Jackal package[41] or interactively with VMD in the case of Barnase by copying from another chain in the unit cell. The systems were prepared in an octahedron periodic box, using a minimum distance of 1 nm between the protein and the boundary. Solvation was achieved using the standard Gromacs solvate tool[42], see also http://www.gromacs.org/. All ions and crystal waters were removed before solvation, including the interfacial water molecules. However, we also prepared simulation runs of Maxi where crystal waters were kept. Both systems are referred to as 'random' and 'crystal', respectively. Systems were made electrostatically neutral with randomly placed Na^+ and Cl^- counterions. Due to its peculiar nature, the neutral system Maxi was charged at a concentration of 7 mmol/L. However, we found none of the counterions to interact significantly with the protein during the course of our simulations. MD simulations were performed with the Gromacs software[42], v5.0.4. using the charmm27 force field[43] and the SPC water model[44].

Prior to data production, the systems were minimized using the steepest descent method until the maximum force on any atom was lower than 10^3 kJ mol^{-1} nm^{-1}. The systems were then equilibrated by a 100 ps *NVT* run to thermalize the system, followed by a 200 ps *NpT* equilibration run to stabilize the volume. During the equilibration, all bonds were constrained using the LINCS algorithm[45]. A time step of 2 fs was employed, Van der Waals interactions were cut off at 1 nm, electrostatic interactions were calculated with PME[46]. Equations of motion for the water molecules were solved with the SETTLE algorithm[47]. The system was coupled to a Berendsen temperature bath[48] of 300 K (Barnase, E2/Im2) or 273K (Maxi), with protein and solvent coupled independently. Pressure was maintained at 1 bar by the Parrinello-Rahman barostat[49]. Data production runs consisted of 10 ns simulations, with bond constraints only on bonds involving hydrogen atoms. The configuration of the system was saved every 10 ps.

Diffusion and residence times. In order to calculate diffusion rates of water molecules, additional short simulations were performed for 300 ps, saving the configuration every 200 fs. The diffusion constant was computed by calculating the mean square displacement of a selected number of water molecules and then using the Einstein relation, using the Gromacs suite of analysis tools. In order to calculate the residence times of protein-associated water molecules, we monitor the total time a first-shell water molecule stays in contact with a protein atom. Any water (oxygen) within a distance of 3.5 Å from a protein heavy atom was considered to be in contact with the protein. This is a slightly simplified approach as employed in ref. 36, since we do not distinguish between different atom types. The residence time is then calculated by averaging over all residence times of associated water molecules.

AquaSol. We obtained the AquaSol software[27] from the authors and installed it locally. We use both the basic DPBL-model as well as the Yukawa functional which takes repulsion between the water molecules into account[50]. Conversion from pdb to pqr format was done using the online pdb2pqr Server[51]. The simulation temperature was the same as for the MD simulation (300 K and 273 K) and the implicit ion concentration was set to match the

volume and number of ions in the MD simulation box. The number of points in the *x*, *y* and *z* dimensions $(2^n + 1)$ was $129 \times 129 \times 129$ for Barnase and E2/Im2, and $65 \times 65 \times 129$ for Maxi. This ensures a resolution of around 1 Å in each dimension. Individual coordinate configurations, chosen at random, were extracted from the MD runs and used as input structure in the AquaSol software[27]. The configurations with the best $f^w(\text{nat})_{MD}$ values are listed in Table 3. A dielectric constant of 3 was employed for the protein interior. We have tested the effect of the chosen value on several residues in order to see how the choice affects the obtained $f^w(\text{nat})$ values. For all residues tested the general trend is observed that for a smaller value ($\varepsilon_p = 2$) the $f^w(\text{nat})$ -values increase while they decrease for a larger value ($\varepsilon_p = 4$), as a consequence of the increased dielectric contrast between the protein and the surrounding solvent for the lower dielectric constant of the protein. We opted for the value of 3 as a compromise on the usual scale of $\varepsilon_p \approx 2-4$ and performed all analyses for this value. The lattice grid size for the solvent was 2.8 Å with a concentration of 55 mol/l and dipole moment of 3.0 debye. Placement of water molecules was done using the method described in the paper of Azuara *et al.*[26] which consists of sorting the density values in descending order. The water molecules are then placed by walking down the list until the density threshold has been reached, which in our case was the reference density of bulk water. At every water molecule placement we eliminate points within 3.0 Å of this position from the list.

References

1. Schutz, C. N. & Warshel, A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins* **44,** 400–417 (2001).
2. Kuntz, I. D. & Kauzmann, W. Hydration of proteins and polypeptides. *Adv. Protein Chem.* **28,** 239–345 (1974).
3. Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nature* **437,** 640–647 (2005).
4. Livingstone, R. A., Nagata, Y., Bonn, M. & Backus, E. H. Two Types of Water at the Water-Surfactant Interface Revealed by Time-Resolved Vibrational Spectroscopy. *J. Am. Chem. Soc.* **137,** 14912–14919 (2015).
5. Czapiewski, D. & Zielkiewicz, J. Structural properties of hydration shell around various conformations of simple polypeptides. *J. Phys. Chem. B* **114,** 4536–4550 (2010).
6. Nakasako, M. Large-scale networks of hydration water molecules around bovine beta-trypsin revealed by cryogenic X-ray crystal structure analysis. *J. Mol. Biol.* **289,** 547–564 (1999).
7. Grossman, M. *et al.* Correlated structural kinetics and retarded solvent dynamics at the metalloprotease active site. *Nat. Struct. Mol. Biol.* **18,** 1102–1108 (2011).
8. Conti Nibali, V. & Havenith, M. New insights into the role of water in biological function: studying solvated biomolecules using terahertz absorption spectroscopy in conjunction with molecular dynamics simulations. *J. Am. Chem. Soc.* **136,** 12800–12807 (2014).
9. Xu, X. *et al.* Water's potential role: Insights from studies of the p53 core domain. *J. Struct. Biol.* **177,** 358–366 (2012).
10. Hong, S. & Kim, D. Interaction between bound water molecules and local protein structures: A statistical analysis of the hydrogen bond structures around bound water molecules. *Proteins* **84,** 43–51 (2016).
11. Zhang, L., Yang, Y., Kao, Y. T., Wang, L. & Zhong, D. Protein hydration dynamics and molecular mechanism of coupled water-protein fluctuations. *J. Am. Chem. Soc.* **131,** 10677–10691 (2009).
12. Ben-Naim, A. Molecular recognition - viewed through the eyes of the solvent. *Biophys. Chem.* **101–102,** 309–319 (2002).
13. Janin, J. Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Structure* **7,** R277–R279 (1999).
14. Heyden, M. Resolving anisotropic distributions of correlated vibrational motion in protein hydration water. *J. Chem. Phys.* **141,** 22D509 (2014).
15. Buckle, A. M., Schreiber, G. & Fersht, A. R. Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. *Biochemistry* **33,** 8878–8889 (1994).
16. Ahmad, M., Gu, W., Geyer, T. & Helms, V. Adhesive water networks facilitate binding of protein interfaces. *Nat. Commun.* **2,** 261 (2011).
17. Meenan, N. A. *et al.* The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proc. Natl. Acad. Sci. USA* **107,** 10080–10085 (2010).
18. Wojdyla, J. A., Fleishman, S. J., Baker, D. & Kleanthous, C. Structure of the ultra-high-affinity colicin E2 DNase-Im2 complex. *J. Mol. Biol.* **417,** 79–94 (2012).
19. Lensink, M. F. *et al.* Blind prediction of interfacial water positions in CAPRI. *Proteins* **82,** 620–632 (2014).
20. Pal, S., Chakraborty, K., Khatua, P. & Bandyopadhyay, S. Microscopic dynamics of water around unfolded structures of barstar at room temperature. *J. Chem. Phys.* **142,** 055102 (2015).
21. Rani, P. & Biswas, P. Diffusion of Hydration Water around Intrinsically Disordered Proteins. *J. Phys. Chem. B* **119,** 13262–13270 (2015).
22. Kuffel, A. & Zielkiewicz, J. Water-mediated long-range interactions between the internal vibrations of remote proteins. *Phys. Chem. Chem. Phys.* **17,** 6728–6733 (2015).
23. Zhou, R. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins* **53,** 148–161 (2003).
24. Jaramillo, A. & Wodak, S. J. Computational protein design is a challenge for implicit solvation models. *Biophys. J.* **88,** 156–171 (2005).
25. Abrashkin, A., Andelman, D. & Orland, H. Dipolar Poisson-Boltzmann equation: ions and dipoles close to charge interfaces. *Phys. Rev. Lett.* **99,** 077801 (2007).
26. Azuara, C., Orland, H., Bon, M., Koehl, P. & Delarue, M. Incorporating dipolar solvents with variable density in Poisson-Boltzmann electrostatics. *Biophys. J.* **95,** 5587–5605 (2008).
27. Koehl, P. & Delarue, M. AQUASOL: An efficient solver for the dipolar Poisson-Boltzmann-Langevin equation. *J. Chem. Phys.* **132,** 064101 (2010).
28. Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. *J. Chem. Phys.* **24,** 966–978 (1956).
29. Felderhof, B. U. Fluctuations of polarization and magnetization in dielectric and magnetic media. *J. Chem. Phys.* **67,** 493–500 (1977).
30. Pujos, J. S. & Maggs, A. C. Legendre Transforms for Electrostatic Energies in *Electrostatics of Soft and Disordered Matter*, (eds. Dean, D., Dobnikar, J., Naji, A. & Podgornik, R.) 69–79 (PanStanford, 2014).
31. Warshel, A. & Levitt, M. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozme. *J. Mol. Biol.* **103,** 227–249 (1976).
32. Holst, M. J. & Saied, F. Numerical solution of the nonlinear Poisson-Boltzmann equation: Developing more robust and efficient methods. *J. Comput. Chem.* **16,** 337–364 (1995).
33. Poitevin, F., Orland, H., Doniach, S., Koehl, P. & Delarue, M. AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformally hydrated atomic models. *Nucl. Acids Res.* **39,** W184–W189 (2011).
34. Smaoui, M. R. *et al.* Computational assembly of polymorphic amyloid fibrils reveals stable aggregates. *Biophys. J.* **104,** 683–693 (2013).

35. Sun, T. *et al.* An antifreeze protein folds with an interior network of more than 400 semi-clathrate waters. *Science* **343,** 795–798 (2014).
36. Sharp, K. A. The remarkable hydration of the antifreeze protein Maxi: a computational study. *J. Chem. Phys.* **141,** 22D510 (2014).
37. Holz, M., Heil, S. R. & Sacco, A. Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate 1H NMR PFG measurements. *Phys. Chem. Chem. Phys.* **2,** 4740–4742 (2000).
38. van der Spoel, D., van Maaren, P. J. & Berendsen, H. J. C. A systematic study of water models for molecular simulation: Derivation of water models optimized for use with a reaction field. *J. Chem. Phys. A* **108,** 10220–10230 (1998).
39. Mark, P. & Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298K. *J. Chem. Phys A* **105,** 9954–9960 (2001).
40. Best, R. B., Zheng, W. & Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **10,** 5113–5124 (2014).
41. Jackal. A. Protein Structure Modeling Package by J. Z. Xiang. http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Jackal (2006).
42. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* **91,** 43–56 (1995).
43. MacKerell, A. D., Banavali, N. & Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **56,** 257–265 (2000).
44. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F. & Hermans, J. Interaction Models for Water in Relation to Protein Hydration. In *Intermolecular* Forces (ed. Pullman, B.) 331–342 (Reidel, 1981).
45. Hess, B., Bekker, H. Berendsen, H. J. C. & Fraajie, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **18,** 1463–1472 (1997).
46. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103,** 8577–8592 (1995).
47. Miyamoto, S. & Kollman, P. A. SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithms for Rigid Water Models. *J. Comp. Chem.* 13, 952–962 (1992).
48. Berendsen, H. J. C., Postma, J., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **81,** 3684–3690 (1984).
49. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52,** 7182–7190 (1981).
50. Koehl, P., Orland, H. & Delarue, M. Beyond the Poisson-Boltzmann model: modeling biomolecule-water and water-water interactions. *Phys. Rev. Lett.* **102,** 087801 (2009).
51. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucl. Acids Res.* **32,** 665–667 (2004).

Acknowledgements

We thank M. Delarue for providing the AquaSol solver and F. Poitevin for help. This work was supported by the ANR project "FSCF", grant number ANR-12-BSV5–0009. Computer resources were provided by the French CINES Montpellier, under contract x2016–077225. Support from the Research Federation FRABio (University Lille, CNRS, FR 3688, "Structural and Functional Biochemistry of Biomolecular Assemblies") is gratefully acknowledged.

Author Contributions

R.B. and M.F.L. designed research, G.C. performed research, all authors discussed the results and wrote the paper.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Copie, G. *et al.* On the ability of molecular dynamics simulation and continuum electrostatics to treat interfacial water molecules in protein-protein complexes. *Sci. Rep.* **6**, 38259; doi: 10.1038/srep38259 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Biophysical Journal Volume 107 July 2014 185–196

Article

Probing the Conformation of FhaC with Small-Angle Neutron Scattering and Molecular Modeling

Frank Gabel,[1,2,3,4,*] Marc F. Lensink,[5,*] Bernard Clantin,[5] Françoise Jacob-Dubuisson,[6,7,8,9] Vincent Villeret,[5] and Christine Ebel[1,2,3,*]

[1]Université Grenoble Alpes, IBS, F-38044 Grenoble, France; [2]CNRS, IBS, F-38044 Grenoble, France; [3]CEA, IBS, F-38044 Grenoble, France; [4]Institut Laue-Langevin, Grenoble, France; [5]CNRS USR 3078, Institut de Recherche Interdisciplinaire, Campus CNRS de la Haute Borne, Université Lille Nord de France, IFR 147, BP 70478, Villeneuve d'Ascq, France; [6]Institut Pasteur de Lille, Centre d'Infection et d'Immunité de Lille, Lille, France; [7]CNRS UMR8204, Lille, France; [8]INSERM U1019, Lille, France; and [9]Université Lille Nord de France, Lille, France

ABSTRACT Probing the solution structure of membrane proteins represents a formidable challenge, particularly when using small-angle scattering. Detergent molecules often present residual scattering contributions even at their match point in small-angle neutron scattering (SANS) measurements. Here, we studied the conformation of FhaC, the outer-membrane, β-barrel transporter of the *Bordetella pertussis* filamentous hemagglutinin adhesin. SANS measurements were performed on homogeneous solutions of FhaC solubilized in n-octyl-d17-βD-glucoside and on a variant devoid of the α helix H1, which critically obstructs the FhaC pore, in two solvent conditions corresponding to the match points of the protein and the detergent, respectively. Protein-bound detergent amounted to 142 $\pm$ 10 mol/mol as determined by analytical ultracentrifugation. By using molecular modeling and starting from three distinct conformations of FhaC and its variant embedded in lipid bilayers, we generated ensembles of protein-detergent arrangement models with 120–160 detergent molecules. The scattered curves were back-calculated for each model and compared with experimental data. Good fits were obtained for relatively compact, connected detergent belts, which occasionally displayed small detergent-free patches on the outer surface of the β barrel. The combination of SANS and modeling clearly enabled us to infer the solution structure of FhaC, with H1 inside the pore as in the crystal structure. We believe that our strategy of combining explicit atomic detergent modeling with SANS measurements has significant potential for structural studies of other detergent-solubilized membrane proteins.

INTRODUCTION

Membrane proteins perform a wide range of functions within cells, are involved in a number of genetic diseases, and have considerable therapeutic importance since 60% of drug targets are membrane receptors or ion channels. Membrane proteins are also essential for the virulence of pathogens. However, their biochemical and structural characterization remains limited compared with that of soluble proteins. For example, membrane proteins account for ~30% of the genomic sequences but represent <1% of protein structures solved at the atomic scale.

This is due to the technical challenges associated with the properties of these macromolecules embedded in lipid membranes, including production in sufficient amounts, solubilization by detergents, purification in a functional form, and crystallization. Detergents are amphiphilic molecules bearing a hydrophilic head and hydrophobic tail. Above their critical micelle concentrations (CMCs), detergents not only occur as monomers but also assemble into micelles and cover the hydrophobic surfaces of membrane proteins, thus allowing their solubilization. Detergent monomers and micelles (which may contain dissolved lipids) coexist in solution with protein-detergent complexes (PDCs; which may also contain lipids and/or cofactors). Thus, membrane protein samples are always complex, multicomponent systems.

Small-angle neutron scattering (SANS) is a unique technique for investigating the structure of complex systems in solution (1,2). It can be used in combination with contrast variation, i.e., by varying the deuterium content of the protein and/or detergent components and/or the solvent using H_2O/D_2O mixtures to mask the signal from one type of component (e.g., the detergent). Structural information about the protein within the PDC can therefore be obtained with minimal contributions from the detergent (for a recent review, see Breyton et al. (3)). Mathematical approaches have been developed over the last two decades to interpret the information from small-angle x-ray scattering (SAXS) or SANS scattering curves (4,5), allowing comparisons between crystal and solution structures as well as ab initio, low-resolution modeling. However, only a few studies have applied these techniques to membrane proteins, despite the considerable interest in investigating

Submitted September 30, 2013, and accepted for publication May 5, 2014.

*Correspondence: frank.gabel@ibs.fr or marc.lensink@iri.univ-lille1.fr or christine.ebel@ibs.fr

Frank Gabel and Marc F. Lensink contributed equally to this work.

Editor: Bert de Groot.

0006-3495/14/07/0185/12 $2.00

http://dx.doi.org/10.1016/j.bpj.2014.05.025

the conformational changes associated with their function. This is due in particular to the intrinsic chemical heterogeneity (aliphatic chains versus hydrophilic heads) of most detergent molecules. As a consequence, their contribution, in general, cannot be completely removed at all scattering angles, even under conditions in which the detergent molecules are globally masked. Therefore, there is a clear need for tools to model detergent organization around membrane proteins for a proper interpretation of SANS data, allowing the discrimination of moderate conformational changes.

TpsB transporters are components of two-partner secretion (TPS) systems in Gram-negative bacteria. They secrete large, mostly β-helical proteins, collectively called TpsA proteins, that generally serve as virulence factors (6). TpsB transporters function as monomers and without accessory factors. The structure of FhaC, the outer-membrane transporter that secretes the *Bordetella pertussis* filamentous hemagglutinin adhesin (FHA) (7), has served as a model for the Omp85 superfamily of protein transporters. The FhaC structure shows a β barrel with an N-terminal extension consisting of an α helix and two periplasmic POTRA domains, each organized around a mixed, three-stranded β sheet and one or two α helices (Fig. 1 *A*). The β barrel consists of 16 antiparallel β strands connected by short turns at the periplasmic side and long loops at the cell surface. The channel within the barrel is occluded by the extracellular loop L6, which folds into the barrel, and by the N-terminal α helix, H1, which spans the channel interior. H1 is joined to the first POTRA domain by a 22-residue-long linker that is not resolved in the x-ray structure, indicating that it is disordered.

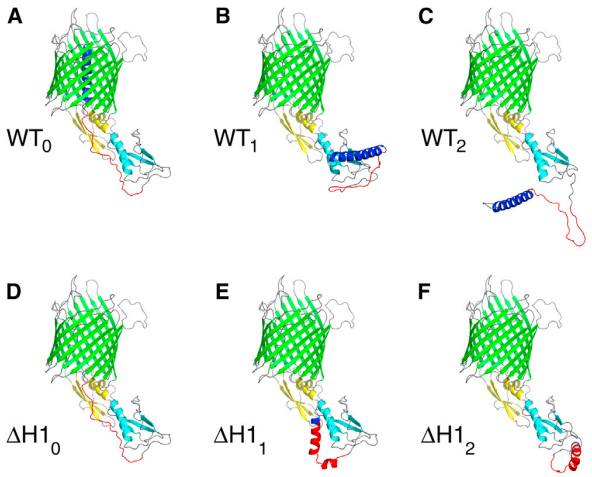


FIGURE 1 The six models of WT FhaC and the FhaC-ΔH1 variant used in the study. Color coding: blue, H1 (residues 1–30); cyan, POTRA1 (residues 53–134); yellow, POTRA2 (residues 135–208); green, β-barrel strands; gray, loops. The linker region, which was not resolved in the crystal structure, was modeled and is colored red (residues 31–52). (*A*) WT FhaC, displayed in cartoon representation. (*B* and *C*) Alternative conformations with H1 outside the β barrel. (*D–F*) Models of FhaC-ΔH1. (*D*) As in (*A*), but with H1 deleted from the structure file. (*E* and *F*) Alternative conformations of FhaC-ΔH1. All figures were created with PyMol (PyMOL Molecular Graphics System, version 1.6.0; Schrödinger, LLC). To see this figure in color, go online.

The crystal structure does not permit us to understand how FhaC mediates the secretion of its partner protein, because the residual opening of the β barrel pore is much too narrow for the translocation of a protein, even in an extended conformation: plain removal of H1 would create a roughly 8-Å-wide pore. Upon reconstitution of FhaC into planar lipid bilayers and application of a transmembrane potential, 8- to 10-Å-wide channels were revealed (7,8). Thus, the FhaC channel appears to be dynamic: in solution, it might open by extrusion of the α helix H1 and/or loop L6, thus creating a protein translocation pathway. L6, which is conserved among TpsB proteins, as well as in the Omp85 superfamily (9), was demonstrated to be a key element for the function of FhaC (7). In addition, earlier work indicated that it might change conformation when its cargo, FHA, is coproduced (10). H1 is also a conserved element in the TpsB family (9). Its deletion had little effect on FhaC secretion activity, although it increased the permeability of the outer membrane to antibiotics (7,8). This indicated that H1 might have a channel-plugging function in the closed conformation and move out of the pore in the course of secretion (the open conformation). Therefore, it is likely that both loop L6 and helix H1 undergo topological rearrangements that play crucial functional roles in the mechanism of secretion.

Our aim in this work was to probe the conformation of FhaC in solution by SANS before FHA binding. At the contrast match point (CMP) of the detergent, we expected to visualize the position of H1 inside or outside the β barrel. We analyzed full-length FhaC and a modified version of the protein harboring a deletion of H1, FhaC-ΔH1, to help decipher the significance of differences between the measured scattering curves. SANS measurements were performed on homogeneous solutions of detergent-solubilized protein in two solvent conditions corresponding to the match points of the protein and the detergent, respectively. The selected detergent, n-octyl-d17-βD-glucoside (d17-OG), has a residual signal even at its match point. We developed a strategy to model individual detergent molecules bound to the surface of the protein explicitly, and thereby compare their contributions to the back-calculated scattering curves from the protein-detergent models in an accurate manner. The combination of SANS and modeling clearly enabled us to infer the location of the H1 helix inside the β barrel of FhaC in solution. The results also provided insight into the structure of the detergent micelle around the protein. We believe that our strategy of combining explicit atomic detergent modeling with SANS measurements will prove to be generally useful for structural studies of detergent-solubilized membrane proteins.

MATERIALS AND METHODS

Detergent samples

d17-OG was purchased from Anatrace (ref. No. O311T). Two samples of d17-OG at 30 mg/mL in H_2O or D_2O were prepared by weighted dissolution using a METTLER AE240 balance (precision 2 10^{-6} g) and mixed or diluted to obtain other concentrations and/or % D_2O.

Production and purification of the proteins

Protocols developed for x-ray studies (7) were adapted to obtain homogeneous solutions of FhaC and FhaC-ΔH1 in perfectly defined solvents containing d17-OG at controlled concentrations. Briefly, the two proteins (8) were overexpressed in *Escherichia coli*, extracted with OG, and purified with cation exchange chromatography in OG (7). Then, each protein was loaded in parallel onto two HisTrap affinity columns and eluted with 1% d17-OG, 0.5 M imidazole pH 6 in either H_2O (H_2O buffer) or D_2O (D_2O buffer). The eluted proteins, at concentrations between 3 and 10 mg/mL, were then dialyzed in their respective elution buffers and eventually diluted in their dialysis buffers.

Analytical ultracentrifugation

Sedimentation velocity experiments were recorded on an analytical ultracentrifuge XLI (acquisition program v4.5; Beckman Coulter, Palo Alto, CA) with a rotor speed of 42,000 rpm, at 20°C, using a rotor Anti-50, and double-sector cells with an optical path length of 3 mm equipped with sapphire windows, with the solvent compartments filled with the sample buffer without detergent. Acquisitions were made using absorbance and interference optics. Analyses in terms of the distribution $c(s)$ of sedimentation coefficients s, and of noninteracting species were done with the program SEDFIT (v 8.9) from P. Schuck (National Institutes of Health; available free of charge at http://www.analyticalultracentrifugation.com/). The parameters of the $c(s)$ analysis were typically 200 particles, with a confidence level of 0.68 for the regularization procedure, and the option "buffer mismatch" for the interference data. d17-OG was investigated using interference optics at 30, 20, and 10 mg/mL in H_2O and D_2O, and the analysis was performed as described previously (11). The samples of FhaC and FhaC-ΔH1 were used following dialysis at concentrations of $\approx$3 mg/mL and $\approx$1.5 mg/mL in D_2O and H_2O buffers, respectively. The solvent density and viscosity of 1.010 g/mL and 1.07 mPa.s for the buffer and of 1.110 g/mL and 1.30 mPa.s for the D_2O buffer were measured at 20°C on a density meter (DMA 5000) and a viscosity meter (AMVn; Anton-Paar, Graz, Austria). Partial specific volumes of 0.728 mL/g, molar masses of 62.10 and 59.83 kDa, and extinction coefficients of 1.36 and 1.423 mL g^{-1} cm^{-1} for FhaC and FhaC-ΔH1, respectively, were calculated with the program SEDNTERP created by D. Hayes et al. (USA; available free at http://bitcwiki.sr.unh.edu/). We used the refractive index increment $\partial n/\partial c$ = 0.187 mL/g, which is typical of membrane proteins (12), and a molar mass in the deuterated solvent increased by a factor of 1.013, as determined from the chemical structure. Details of the analysis, including estimates of the amount of bound detergent from the combination of the absorbance and interference signals, are provided elsewhere (13,14).

SANS experiments and data analysis

All detergent and protein samples were measured in Hellma 100-QS quartz cells (path length = 1 mm) on the small-angle diffractometer D22 at the Institut Laue-Langevin (Grenoble, France). Transmissions T of all samples were recorded systematically. In addition, water (H_2O), boron, and empty-cell references were recorded for detector efficiency, electronic background, and empty-cell subtraction procedures.

Scattering data from the d17-OG detergent were measured for 10 min at concentrations C between 10 and 30 mg/mL in 0%, 25%, 50%, 75%, and 100% D_2O solutions using two instrumental configurations (2 m/2 m and 8 m/8 m collimator/detector distances) with a fixed neutron wavelength λ = 6 Å. One-dimensional (1D) scattering intensities $I(Q)$ were obtained using Institut Laue-Langevin in-house software (15), with the scattering vector $Q = (4\pi/\lambda)\sin(\theta)$, 2θ being the scattering angle. The intensities in the forward scattering direction, $I(0)$, and the radii of gyration, R_G, were extracted by the Guinier approximation (16) using the program PRIMUS (17). The CMC was determined from a linear fit of the concentration series at a given contrast, and the CMP was extracted by plotting $\sqrt{I(0)/[(C-cmc)T]}$, with C being the overall concentration, from the data sets with the highest concentration (i.e., 30 mg/mL), as a function of contrast.

The PDCs (wild-type (WT) FhaC at 3 and 13 mg/mL, and FhaC-ΔH1 at 3 mg/mL) were measured at two different contrasts in buffers with 42% and 90% D_2O (CMP of FhaC and d17-OG, respectively), and d17-OG at 24 mg/mL in buffer with 90% D_2O, in the same experimental setup used for the detergent contrast series. Exposure times were ~60 min per sample. The SANS intensities were back-calculated from the atomic PDCs using the program CRYSON (18,19) in default setup for 42% and 90% D_2O and scored in a least χ^2 fit against the respective experimental SANS data sets.

Creation of detergent topology and structure library

An initial detergent topology was created using the Automated Topology Builder (20) with a PM3 optimized geometry. The topology was translated into a residue building block for use with the Gromacs suite of programs (21) and manually modified to reflect bonded and nonbonded parameters of likewise functional groups from the Gromos 43a2 force field (22). One to five additional exclusions involving polar hydrogen atoms in the sugar ring were introduced. The structure was energy minimized in vacuo using 1000 steps of steepest descent, employing angular removal of mass motion at every step. A molecular-dynamics (MD) simulation of 1 ns was performed using a time step of 2 fs. A Coulomb and van der Waals cutoff distance of 8 nm was set, resulting in an effective complete evaluation of forces. The system was coupled to a temperature bath of 310 K (23). Bonds were constrained to their equilibrium values with LinCS (24). Configurations were saved every 1 ps (500 steps), and thus a library of 1001 detergent structures was obtained.

Modeling of the FhaC linker between H1 and POTRA1

The crystal structure of FhaC (PDB 2qdz) shows disorder for residues 31–52, corresponding to a 22-residue linker region between H1 and the first POTRA domain. With H1 inside the barrel, the linker has to span a distance of ~5 nm, which leaves no possibility to form secondary structure elements. We modeled the atomic coordinates of residues 31–52 using the Jackal package (25), refining only the inserted region. The resulting structure is depicted in Fig. 1 *A*, with the linker colored red. Additional residues that were modeled at the same time were residues 1 and 2, and the extracellular loop L5 formed by residues 384–397. Individual missing atoms were reconstructed as well, without modifying the backbone conformation. The final structure is referred to as WT_0 in the text.

Modeling of FhaC with H1 outside

We investigated the possible orientations of H1 with respect to POTRA1 by submitting the sequence of FhaC, residues 1–133 (H1+linker+POTRA1), to the I-TASSER structure prediction server (26). The resulting five best models were fitted back onto the crystal structure using the POTRA1

domain as reference. Two of the models could be discarded because the fit resulted in steric clashes due to overlapping domains, and a third model corresponded to our previous modeling of the linker region, with a slightly different orientation. The two remaining models were retained. Both models were incorporated into the crystal structure using the Jackal package, replacing residues 1–62 by the model and fitting on POTRA1 (72 residues from 63 onward). In addition, all loop regions were refined. The resulting models (depicted in Fig. 1, *B* and *C*, and referred to as WT_1 and WT_2, respectively, in the text) were prepared for detergent modeling according to the procedure described below.

Modeling of ΔH1 constructs

We investigated the possible orientation of the linker in FhaC-ΔH1 by submitting the sequence of residues 27–133 to I-TASSER and fitted the resulting five best models onto the crystal structure. All five models predicted the linker region to be α helical. One model again led to overlapping domains and three others resembled one another. Thus, two distinctly different models remained for further analysis. No elaborate fitting was required to build complete models, as the start of the POTRA1 domain in both cases overlapped with the end of the modeled structure. In the preparation of the protein in a lipid bilayer, position restraints were removed from the linker atoms (see below). Fig. 1, *D–F*, depict the three FhaC-ΔH1 structures used in the analysis, which are subsequently referred to as $\Delta H1_0$, $\Delta H1_1$, and $\Delta H1_2$. $\Delta H1_0$ corresponds to WT_0 without H1.

Preparation of FhaC in a lipid bilayer

Each of the six FhaC structures was placed in a lipid bilayer of dipalmitoylphosphatidylcholine (DPPC) molecules according to a procedure similar to that described by Lensink et al. (27) and Kandt et al. (28). Briefly, an equilibrated bilayer of DPPC lipids was expanded in the *xy* plane (with the *z* axis being the normal to the bilayer plane) and FhaC/FhaC-ΔH1 was placed in the resulting vacuum at the center. Its position along the *z* axis was such that the water-to-membrane transfer energy was minimized (29,30). In subsequent iterations, the system was restored to the reference area per lipid by shrinking the *x* and *y* coordinates by 2% and deleting all overlapping lipid molecules. The shrinking factor was increased to 5% after eight iterations. Each iteration was accompanied by 100 steps of steepest-descent energy minimization, applying strong position restraining (10^5 kJ/nm^2) on the protein atoms. After ~30 iterations, the system was found to be restored to its original size. The system box was then extended by 1.2 nm in the *z* axis and solvated with ~26,000 water molecules, avoiding solvation of the hydrophobic membrane core by applying a van der Waals radius of 6 Å to the lipid tail atoms. The system was neutralized by adding chloride ions (31) and then subjected to 1000 steps of steepest-descent energy minimization, followed by 100 ps of MD using weak position restraints (10^3 kJ/nm^2) on the nonhydrogen protein atoms to relax the lipids surrounding the protein. The final frame was used as the starting point for the next step.

Modeling of detergent around FhaC

An initial structure of FhaC with detergent was obtained by replacing a fixed number of DPPC lipid molecules by detergent molecules, using the program DOPE (M.F.L.; to be published elsewhere, available upon request). In short, this was done by iteratively replacing a lipid from the input file by a molecule from the detergent library structures by aligning the detergent's molecular structure principal axes to the system axes. Subsequently, the overlap with other molecules in the system was checked and a replacement was accepted if no contacts shorter than 2 Å were detected. If such contacts occurred, the library molecule was rotated in increments of 60° about its *z* axis and contacts were again evaluated. If no solution was found, the next molecule from the library was tried until a solution was found and the total number of molecules to be replaced had been reached. We varied the number of detergent molecules from 120 to 160, in increments of 5. For every such combination (FhaC plus *n* detergent molecules), 20 different detergent configurations were generated. To increase the variance between them, a different random number seed was used for every configuration. From each of the resulting 180 configurations, the solvent was removed (only FhaC and detergent remaining) and the system was subjected to a short energy minimization, followed by an MD simulation in vacuo, and in both cases the protein dynamics were removed from the equations of motion. In essence, this means that the protein coordinates were kept fixed in space while the detergent belt would associate around it. It was found that after ~100 ps of simulation the association was completed, and each system was run for 200 ps to ensure this was done. All end configurations were processed to produce an all-atom system, with deuterium atoms placed on the detergent tails, and nonassociated detergent clusters were removed from the system.

In an alternative approach, we modeled 200 detergent molecules around FhaC using the approach described above, and then reduced the detergent belt by iteratively removing individual detergent molecules until 140 of them remained associated with FhaC. To determine which detergent molecule to remove, we calculated for each detergent molecule the closest distance to any protein atom. The detergent molecule that had the highest value of these, i.e., of which the closest atom was located the farthest away from the protein, was removed. One hundred different detergent configurations were generated using this approach.

RESULTS

Sample design

We chose to investigate the conformation of FhaC solubilized in OG, as this detergent was used for the determination of the high-resolution structure by x-rays. The purification protocol was suitable to provide protein samples at an appropriate concentration for SANS (3–13 mg/mL), with a controlled concentration of detergent micelles and a controlled D_2O content, allowing contrast variation.

To determine the position of the helix inside or outside the barrel in solution, we used full-length FhaC and a variant deleted of the helix H1 (FhaC-ΔH1). Our rationale was that a detectable difference in SANS between FhaC and FhaC-ΔH1 would be indicative of an extended, open conformation in solution, whereas a closed, compact conformation should yield similar results for both proteins. In addition, analysis of FhaC-ΔH1 was expected to provide indications about the position of the linker in the open conformation. A model of WT FhaC with H1 and the linker fully extended was thus generated, and a large change in the calculated radius of gyration (program CRYSON (18,19)) between FhaC in this open conformation and FhaC closed as in the crystal structure was predicted (37 vs. 29 Å) (not shown).

The hydrogenated form of the detergent has a calculated CMP of 19% D_2O, which would provide only a low contrast for the protein (CMP 42% D_2O). The completely deuterated OG, which is commercially available, has a CMP of ~120% D_2O, which cannot be experimentally matched. The tail-deuterated d17-OG has a theoretical CMP of 90% D_2O, in which hydrogenated proteins have a significant contrast

(3). FhaC and FhaC-ΔH1 were thus purified in this detergent, in solvents containing H_2O and 100% D_2O.

Quality control and estimation of bound detergent by analytical ultracentrifugation

The analyses of d17-OG in pure H_2O and D_2O (Fig. S1 and Table S1 in the Supporting Material) yielded a CMC of 9.1 ± 0.4 mg/mL (29.5 mM), an aggregation number of 74 ± 10, and a refractive index increment of 0.135 ± 0.005 mL/g, which are similar to those given in the literature (http://www.affymetrix.com/) (32). The homogeneity of FhaC and FhaC-ΔH1 in d17-OG was excellent, as assessed by analytical ultracentrifugation (AUC). Fig. 2 shows the analysis of the sedimentation velocity profiles for the two proteins in D_2O buffer. Samples in H_2O buffer at 3 and 1.5 mg/mL (not shown) behaved similarly considering the differences in solvent density and viscosity. There are only two detectable contributions. The detergent micelles are detected at 1.6 ± 0.1 and 0.85 ± 0.05 S in H_2O and D_2O buffers (s_{20w} = 2.1 ± 0.1 S) only with interference optics, at concentrations in the 0–1.7 mg/mL range, close to the expected value of 0.9 mg/mL (total concentration C of 10 mg/mL minus CMC). The sedimentation coefficients of the FhaC and FhaC-ΔH1 complexes are indistinguishable: s = 4.76 ± 0.12 S and 2.79 ± 0.08 S in H_2O and D_2O buffers, respectively. The average number of detergent molecules bound to the protein was also determined from the combination of the absorbance and interference signals to be 142 ± 10 mol/mol, which combined with s-values corresponds to a frictional ratio of 1.33 ± 0.03, close to the usual value of 1.25 for a globular compact assembly. Analysis of the sedimentation profiles in terms of noninteracting particles (small monomer or solvent species, detergent micelle, and FhaC complex) give independent estimates of the sedimentation and diffusion coefficients, and thus of the buoyant mass of the complex. The derived average number of bound detergent is 90 ± 20 mol/mol, and, given the s-values, would correspond to a frictional ratio of 1.27 ± 0.04. The difference between these values may be due to uncertainties in the extinction coefficients in the first or second method, or (more likely) to a slight overestimation of the diffusion coefficients.

SANS analyses

d17-OG scattering curves at 30 mg/mL were recorded as a function of contrast (Fig. S2 *A*). The CMC was determined to be 8 ± 1 mg/mL from the d17-OG concentration series (Fig. S2 *B*). Using this value, we determined the aggregation

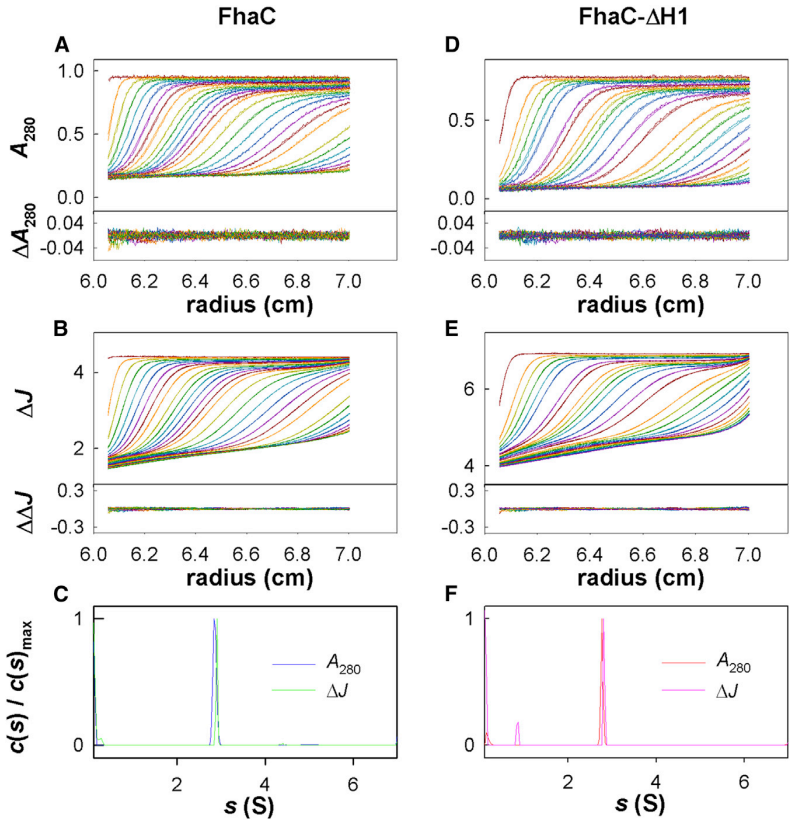


FIGURE 2 Sedimentation velocity of FhaC at 3.3 mg/mL and FhaC-ΔH1 at 3.0 mg/mL in 1% d17-OG, 0.5 M imidazole pH 6, 100% D_2O. (*A*, *B*, *D*, and *E*) Superposition of experimental and fitted sedimentation velocity profiles and their differences (*top and bottom subpanels*), at 280 nm (*A* and *D*), and using interference optics (*B* and *E*). (*C* and *F*) Sedimentation coefficient distributions $c(s)$ normalized to the main protein peak value. To see this figure in color, go online.

number from the 30 mg/mL data in H_2O to be 85, and the detergent CMP to be 90% D_2O (Fig. S2 *C*). The scattering curve of d17-OG at 24 mg/mL at its CMP, in the buffer used for FhaC with 90% D_2O, is presented in Fig. 3. Due to the chemical heterogeneity of d17-OG (tail versus head), there is a significant residual signal at the match point.

The scattering curves of two proteins (FhaC at 3 and 13 mg/mL, and FhaC-ΔH1 at 3 mg/mL) were then measured at 42% and 90% D_2O. At 42% D_2O, the detergent dominates the signal, with the protein being matched; at 90% D_2O (Fig. 3), the protein dominates the signal, with the detergent being matched. All four data sets are of excellent quality in terms of signal/noise ratios and no signs of aggregation were detected, confirming the homogeneity found for all samples in the AUC experiments. To interpret the data with some precision in terms of protein structure, and given that the detergent is not homogeneously masked at its match point, we chose to model atomic detergent molecules explicitly. Note that during sample preparation, we avoided concentration steps using ultrafiltration and used dialysis against solvents with the same detergent and D_2O concentrations to strictly control them. Therefore, there was no need to model detergent micelles, because their contribution was removed by subtraction of the solvent scattering.

Modeling the PDCs

The computational treatment of the template structures (as displayed in Fig. 1), with the number of detergent molecules varying in steps of 5 between 120 and 160, and 20 configurations for each combination of protein and detergent, led to 6 ×180 models of detergent arrangement around FhaC/FhaC-ΔH1. We found the arrangement of detergent to be nonhomogeneous, but nonetheless centered around the β barrel. In addition, detergent could bind to solvent-exposed regions of the protein close to the lipid bilayer, such as POTRA2 or the extracellular loops. Whereas those cases generally resulted in good fits, we found that the structures with detergent bound to POTRA1 (and to H1 in templates WT_1 and WT_2) made for bad χ^2 fits when we back-calculated the theoretical SANS curves (see below).

In an alternative approach (applied only to WT_0), we modeled 140 detergent molecules more compactly onto the FhaC surface by deleting the most distant 60 molecules from an initial amount of 200 detergent molecules (see Figs. 5 *B* and 6 *C*). In general, all of these 100 reduced-detergent-belt models displayed fewer and smaller detergent-free patches (if any) on the β-barrel FhaC surface with respect to the first modeling approach.

SANS analysis of WT FhaC models at 90% D_2O

The scattering curves were back-calculated for all models and compared with the experimental curves using the program CRYSON. The quality of the fit was evaluated by χ^2. Details regarding the individual models are provided in Tables S2 and S4. Fig. 4 A gives an overview of the χ^2 average values for the five best (out of 20) fitting models (complex), calculated for each detergent number (25 best

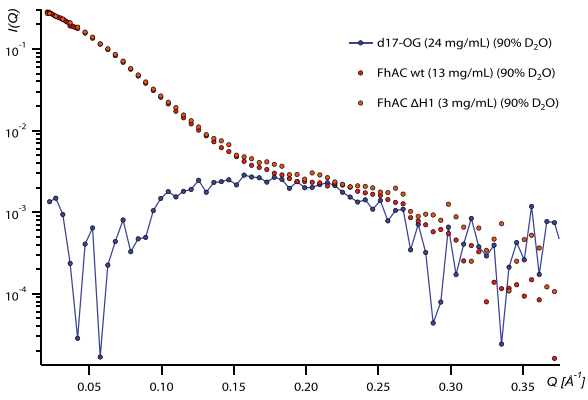


FIGURE 3 SANS curves of d17-OG at the CMP of d17-OG in the FhaC buffer with 90% D_2O. d17-OG in the protein sample is at 10 mg/mL. To see this figure in color, go online.

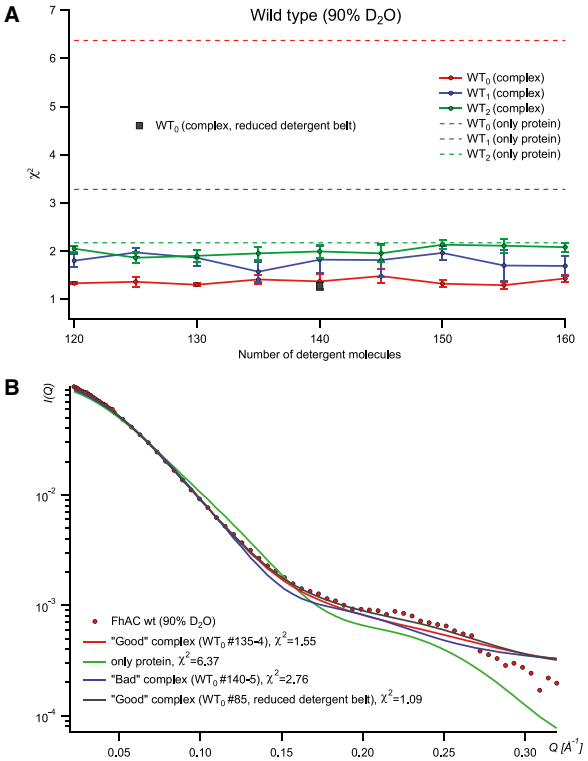


FIGURE 4 Modeling detergent-WT complexes at 90% D_2O. (*A*) χ^2 versus detergent number. (*B*) superposition of experimental and modeled $I(Q)$ curves from good and bad WT_0 models. To see this figure in color, go online.

of 100 for the reduced-detergent-belt models). It also shows the χ^2 values calculated without detergent (only protein). Even though at 90% D_2O the detergent contribution was matched on average, we still observed a marked decrease of χ^2 when detergent was explicitly modeled. Improvement of the fit by detergent modeling appears clearly in Fig. 4 *B*, which shows the superposition of the experimental and fitted scattering curves obtained with and without detergent. Among the three WT FhaC conformations, the one with the H1 helix inside the β barrel, i.e., WT_0, corresponding to the crystal structure, fits the SANS data best. Within each configuration series, the average χ^2-values do not vary a lot as a function of the number of detergent molecules present. This indicates that the 90% D_2O SANS data sets are not very discriminative regarding the shape of the detergent belt, as expected at the detergent match point. A small but significant improvement is observed for the compacted (reduced-detergent-belt) models with respect to their non-compacted counterparts. The remaining minor variations of the χ^2-values can probably be attributed to residual detergent heterogeneities and/or to effects of the modeled hydration shell by CRYSON, which varies as a function of the detergent shape.

SANS analysis of WT FhaC models at 42% D_2O

Good χ^2-values of the models against the SANS data at 42% D_2O (detergent visible, protein matched; Tables S2 and S4; Fig. 5 *A*) are found for individual models in the entire 120–160 detergent molecules range, indicating that the volume of

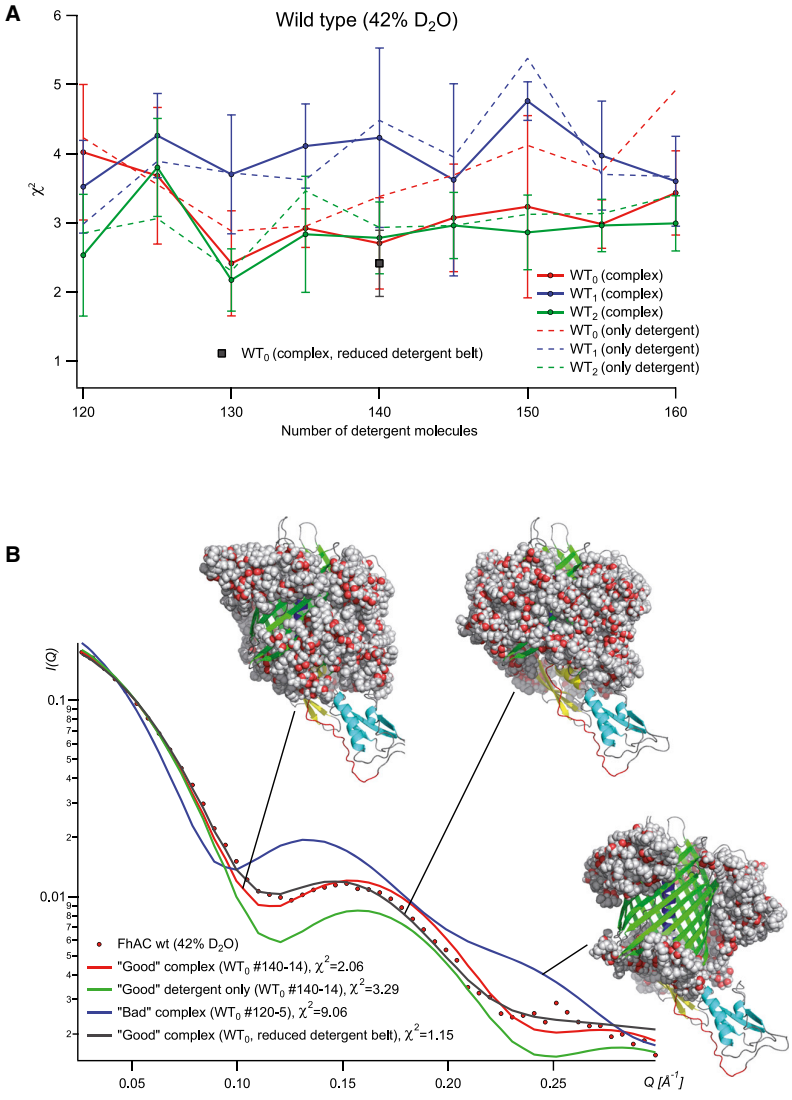


FIGURE 5 Modeling detergent-WT complexes at 42% D_2O. (*A*) χ^2 versus detergent number. (*B*) Superposition of experimental and modeled $I(Q)$ from two WT_0 models. To see this figure in color, go online.

detergent bound to FhaC fits within that range, in good agreement with the AUC results (140 molecules). Fig. 5 A was built considering the five best (out of 20) χ^2 values per FhaC conformation and detergent quantity (25 best out of 100 for the compacted reduced-detergent-belt models). Among the 20 individual models generated for a given, fixed number of detergent molecules, the χ^2-values vary more widely than do their average values for different numbers of detergent molecules. Accordingly, the superposition of fitted and experimental scattering curves allows a clear distinction between appropriate and inappropriate models (Fig. 5 *B*). Compacted reduced-detergent-belt models displayed on average better fits with the SANS data, with the best χ^2 values being close to one. Moreover, the considerable variation of the fit values between different individual detergent models can be used to determine general structural features at low resolution of the detergent belt around FhaC. Fig. 6, *A–C*, show several bad detergent models ($\chi^2 > 8$), several good detergent models ($\chi^2 < 2.5$), and several good compact (reduced-detergent-belt) models, respectively ($\chi^2 < 2.3$; higher-resolution images are provided in the Supporting Material). A comparison of the general features of the detergent models in Fig. 6 reveals that good fits are only obtained for relatively compact, connected detergent belts, whereas models that fit the SANS data very poorly are rather disconnected detergent topologies, with large, central parts of the transmembrane β-barrel wall being detergent free, and detergent molecules accumulating at the top and bottom of the barrel. Several, slightly different, good detergent topologies fit the SANS data equally well (Fig. 6, *B* and *C*).

SANS analysis of the FhaC-ΔH1 models

The SANS curves for the 20 models of FhaC-ΔH1 generated at each of the nine different quantities of detergent and for each of the three protein conformations ($\Delta H1_0$, $\Delta H1_1$, and $\Delta H1_2$; presented in Fig. 1) were back-calculated. A detailed analysis of the χ^2 fits over all models is included in Table S3. For the 90% D_2O data sets (protein visible), the average values of the best five χ^2 fits (for each of the detergent numbers and protein conformations) and a comparison with the χ^2 values from the corresponding protein-only models are shown with selected superposition of experimental and fitted scattering curves in Fig. S3, A and B. Explicit incorporation of detergent in the models (complex) improves the fits significantly with respect to the protein-alone models. However, very good fits can be found for all three protein conformations, and thus our SANS data do not allow us to discriminate among these

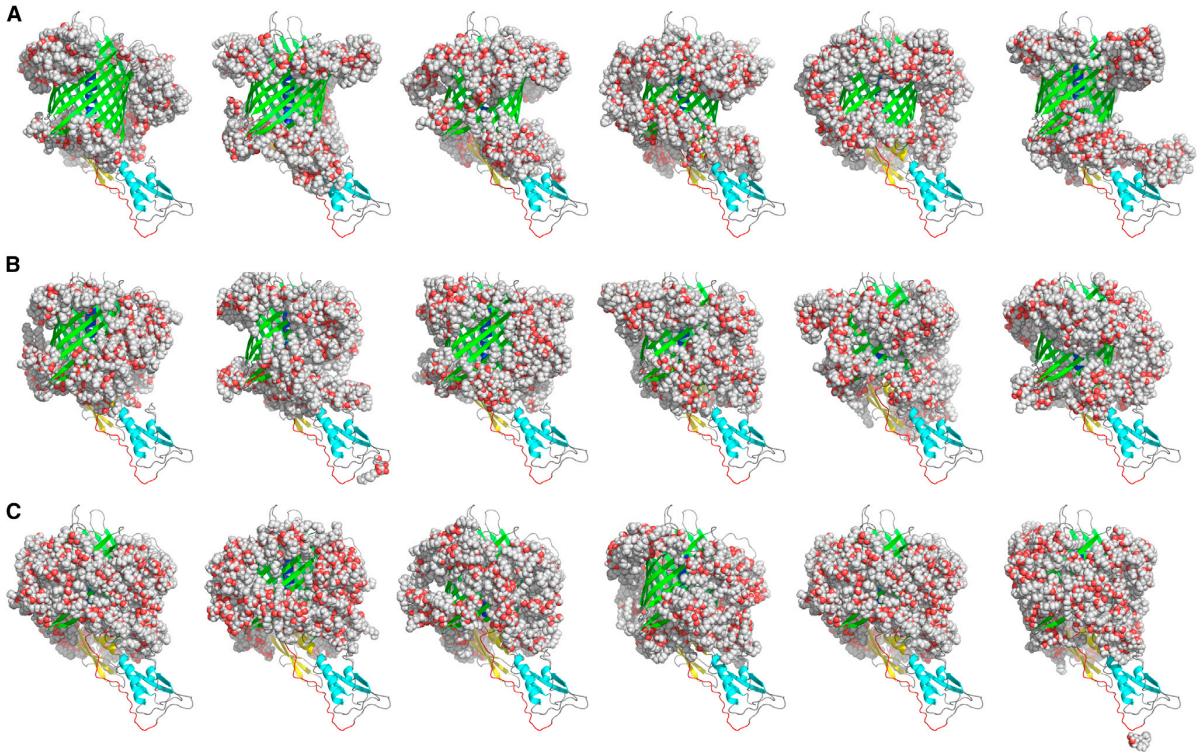


FIGURE 6 (*A–C*) WT FhaC-detergent models (WT_0) corresponding to bad (*A*), good (*B*), and good compact (*C*) fits. Color coding as in Fig. 1; detergent molecules shown in sphere representation. To see this figure in color, go online.

three conformations in solution. The results are more discriminating for the 42% D_2O SANS data sets (detergent visible, protein invisible): the $\Delta H1_1$ models represent the detergent belt better than do the $\Delta H1_0$ and $\Delta H1_2$ series (Fig. S4 A). Most of the 20 models generated with the three distinct conformations fit the experimental data sets poorly, and very few structures yield satisfactory fits against the experiment SANS data. Examples of representative good and bad structures in Figs. S4 B and S5 show that the detergent location around the barrel generally improves the fit, unlike its positioning on the POTRAs. For the truncated FhaC-ΔH1 variant, removing the protein moiety from the models, even if it is globally matched, significantly decreases the quality of the fit.

DISCUSSION

The study of membrane proteins represents a formidable challenge for most biophysical techniques in solution, and in particular small-angle scattering, due to the presence of detergent molecules in the sample. Their contribution is relatively strong in the case of x-rays (SAXS) and needs to be taken into account in explicit protein-detergent models where the protein and detergent parts contribute simultaneously, in a weighted manner, to the scattering curve (33). In contrast to SAXS, neutron scattering (SANS) using contrast variation has the advantage of providing independent structural data about either the protein or the detergent within a complex. For instance, SANS is able to mask the scattering contribution from detergent molecules on average (i.e., the $I(Q = 0)$ intensity is zero) and to focus on the protein topology in situ (34). However, in general, the chemical heterogeneity between detergent head and tail moieties yields residual scattering contributions for $Q > 0$ that contribute to the scattering intensity and affect low-resolution modeling of the protein (35,36). Only some specific surfactants (37) or mixtures of deuterated and hydrogenated detergents (38) can be homogeneously masked (for a review, see Breyton et al. (3)). In this work, we demonstrated that explicit modeling of a PDC combined with SANS and contrast variation can yield excellent results. We were able to probe fine structural details of the protein in the presence of detergent, such as the position of the N-terminal α helix in FhaC, and discard several potential conformations. We observed a significant improvement of the fit when the detergent was explicitly incorporated into the models. Moreover, by obtaining a valid atomic model of the protein (that could be validated against the SANS data), we were able to study the structural properties of an explicit detergent belt and determine its general features. We believe that our approach can be applied to several membrane protein systems in solution, which are otherwise (e.g., by crystallography) very difficult to approach. It allows one to study the structural features of both the protein conformation and the shape of the detergent bound in a complex in situ by simply adjusting the H_2O/D_2O levels to the respective CMPs. However, we would like to stress that three key factors are very important for a successful application of our approach: 1), excellent sample quality (in particular, regarding the monodispersity of the PDCs, which ideally should be checked by AUC before SANS analyses); 2), a sufficient contrast between the protein and detergent parts, which in practice requires the use of either deuterated proteins or deuterated detergents; and 3), a sufficiently wide sampling of the modeled detergent belt.

When applied to membrane proteins, SAXS and SANS methods therefore need to deal with the detergent belt. It has been theorized that the detergent around membrane proteins organizes in a homogeneous, belt-like manner. Theoretical contributions originating from ellipsoidal detergent shapes were calculated in a SAXS study of Photosystem I-detergent complexes (39). It turned out to be impossible to match the SAXS data to the crystal structure embedded in a disk of detergent, which was interpreted as the protein being partially unfolded in solution. A SANS study of Light-Harvesting Complex II in detergent suggested also that the detergent does not homogeneously surround the hydrophobic periphery of the protein (40). A more recent study combined SAXS analyses of Aquaporin-0 with modeling of the detergent organization as an elliptical toroid (33). Although a good correspondence of the theoretical scattering curve with the experimental one was observed, the authors recognized the fact that the detergent corona is unlikely to adopt a static elliptical conformation, as the ellipsoid can be interpreted as resulting from a dynamic detergent behavior, the average of which is measured by SAXS. They recently extended their method by adopting both coarse-grained and atomistic modeling studies, allowing the deformation of an assumedly elliptical detergent toroid (41) and resulting in slightly improved χ^2 fits. In an elegant approach, detergent molecules were guided toward an elliptical detergent organization through nonequilibrium MD simulations with explicit water. Although the procedure gives promising results and will mark an important step in the study of the detergent belt in small-angle scattering experiments, only a single conformation is generated, and simulations in explicit solvent are relatively expensive.

In this study, we did not assume an elliptical organization; rather, we used molecular modeling to generate an ensemble of FhaC-detergent arrangements from their mutual association based on physicochemical interaction parameters. We did ask where the detergent would partition in such a system, and since the detergent was needed to solubilize FhaC in the first place, the obvious answer is, near or around the hydrophobic β barrel. Several tools exist to place a membrane protein in a lipid bilayer for the purpose of MD simulations (28,42), and we decided to adopt an approach in which FhaC is initially placed in a lipid bilayer, a sufficient number of lipid molecules are replaced by detergent, and, after removal of water, a simulation is run in vacuo to let

the detergent molecules associate with FhaC. The procedure is then repeated 20 times to generate an ensemble of FhaC-detergent conformations. The association of detergent with FhaC is illustrated in Movie S1.

This approach does require some CPU-intensive steps. The initial placement in the lipid bilayer uses iterative energy minimization steps and a short MD simulation is run to relax lipids around the protein; the lipid-detergent replacement routines apply a number of mathematical operations to rotate and translate the detergent molecules, fit them to lipids, and evaluate any atomic overlap; and the final conformation is used for a short MD simulation in vacuo. Nevertheless, the total CPU time is limited to ~2 hr per configuration on present-day laptop and desktop machines. The total simulation time required for the six protein conformations, nine different detergent quantities, and 20 different configurations amounts to ~90 days of CPU time, a duration that is quite manageable for even small computing clusters. Previous simulations of OmpA and GpA in detergent (43–46) required an equilibration time of 20–50 ns for the association of detergent with protein. These simulations generally show that the detergent covers at least 80% of the hydrophobic surface, also in a nonhomogeneous manner. If we consider our vacuum simulations, we have a total simulation time of $9 \times 20 \times 200$ ps $= 36$ ns per protein variant. This is comparable in terms of CPU time, but offers 180 different detergent arrangements.

The trade-off between speed and accuracy is obvious: although a large number of protein-detergent arrangements can be produced relatively quickly, a significant fraction of these arrangements are unrealistic. Here, unrealistic arrangements either correspond to detergent binding at the level of the POTRA domains or show large uncovered regions on the hydrophobic β barrel. However, the χ^2-values of the back-calculated curves are decidedly discriminative. This is an important aspect of our approach, as it allows the elimination of bad protein structures as well as bad detergent belts.

In an attempt to reach a more homogeneous distribution, and having confirmed the conformation of WT FhaC in solution as WT_0, we also simulated the association of excess detergent (200 molecules) to FhaC. In subsequent iterative steps, we then removed the detergent molecules located the farthest away from the β barrel until the required number of associated detergent molecules (140) was reached. We produced 100 such configurations. In general, this procedure results in small but significant improvements of the χ^2 fits (see Figs. 4 and 5). The procedure is computationally somewhat more expensive, since the iterative replacement procedure converges more slowly to a solution for the larger detergent amounts. However, we consider this approach to be superior, since for more complicated systems, such as those showing moderate to large conformational changes, it will provide better sampling than simply increasing the number of different configurations. Such cases of limited sampling are readily identified by the χ^2 fits.

We would like to emphasize that what is measured experimentally is not a particular generated conformation, but rather the average signal of a large number of conformations, originating from the many proteins in the sample and from dynamic variations within each sample. The variation in protein-detergent conformations and their theoretical scattering curves led us to include only the five best-fitted models for further consideration. It is not unexpected that the best fits show the detergent to cover the larger surface area, but, interestingly, many of the good models display small detergent-free patches on the outer surface of the β barrel, i.e., the barrel is not completely covered with detergent molecules. Such patches are still observed in our reduced-detergent-belt approach, but they are fewer in number and smaller, and the ensemble of structures shows a good coverage of the β barrel. As to the shape of the detergent belt (40), our SANS data are compatible with a slightly nonhomogeneous distribution of detergent around the hydrophobic β barrel of the protein. Given that SANS provides an average intensity over all particle conformations present at a given moment in solution, it is conceivable that there are indeed several different, interconverting detergent arrangements possible around FhaC in solution, and that these are sampled by our modeling approach. The fact that several slightly different detergent belt structures are in excellent agreement with our SANS data (Fig. 6 *C*) illustrates the accuracy and uniqueness of such a low-resolution approach. Within these limits, a possible interpretation of our data is that a detergent belt is a dynamic entity consisting of an ensemble of slightly different, interchanging conformations.

The results clearly show the feasibility of our approach, as we could confirm that WT FhaC in solution adopts a structure similar to the crystallographic structure and rule out two alternative conformations (WT_1 and WT_2). The approach developed here may apply to membrane proteins of known or unknown structures. It can be used in combination with both SANS and SAXS studies. (In the case of SAXS, the free micelles need to be separated from the complex by using size exclusion chromatography directly on the beamline (33)). The association of detergent is flexible and can use any template structure or detergent molecule (after development of adequate force-field parameters). Subsequent comparison of the theoretical and experimental scattering curves allows the identification of good and bad models, in terms of both protein structure and detergent organization, when measured at the match point of either one. It can thus also be applied to models originating from ab initio modeling tools, such as I-TASSER (26), to help validate or discard them. This flexibility makes it a powerful tool for studying membrane proteins, including their conformational span.

Regarding the protein under study in this work, the crystal structure of full-length FhaC is in excellent agreement with the SANS data in solution. The conformation with the helix inside the pore could be discriminated from two alternative

conformations, with the helix outside the barrel. This result does not solve the question about the conformational changes that occur to open the FhaC channel and to allow the passage of the cargo protein FHA. Open and closed conformations might be in equilibrium, but the open conformations are likely to be poorly populated in a detergent environment. It is possible that an open structure exists or is more populated in a lipid environment when compared with the detergent-solubilized state. The membrane environment was shown to be required for the native conformation of the KvAP voltage-dependant channel (47). Similarly, the kinetics of the photocycle of bacteriorhodopsin depends on its hydrophobic environment (48). The large conformational changes associated with the function of the reticulum Ca^{2+}-ATPase are allowed by small rearrangements of the lipid bilayer and of the protein in its transmembrane part (49). In the case of FhaC, 8- to 10-Å-wide channels were revealed by means of electrophysiology techniques with the protein inserted into a bilayer.

CONCLUSIONS

Despite the ever-growing interest in membrane protein structure and dynamics, it remains notoriously difficult to study them. Solubilization requires the use of amphiphilic molecules (detergent) to act as a buffer between the solvent and the transmembrane domain, but little is known about the arrangement of detergent around the protein. We have successfully applied a combination of SANS and molecular modeling to probe the conformational space of the membrane protein FhaC. Since a residual contribution remained at the CMP of the detergent, we used molecular modeling to generate an ensemble of detergent arrangements around the WT protein and putative alternative conformations. Thus, we were able to confirm that WT FhaC in solution adopts a conformation similar to the x-ray structure, while ruling out alternative conformations at the same time. This study provides valuable insight into the organization of the detergent belt. Modeling studies may employ various detergent molecules and can be combined with both SANS and SAXS studies. The general applicability of this approach makes it an extremely powerful and significant tool that may allow more detailed studies of membrane protein structure and dynamics.

SUPPORTING MATERIAL

Five figures, four tables, one movie, and one zip file are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00559-1.

We thank the Institut Laue-Langevin for beam time on the D22 instrument, and Dr. Phil Callow for local contact during the experiment. We thank Prof. Marc le Maire for helpful discussions.

This work used the AUC platform of the Grenoble Instruct Centre (ISBG; UMS 3518 CNRS-CEA-UJF-EMBL) with support from FRISBI (ANR-10-INSB-05-02) and GRAL (ANR-10-LABX-49-01) within the Grenoble Partnership for Structural Biology (PSB). This work was supported by the CEA, the CNRS, Université Joseph Fourier, and grants from the EU (FP7/2007-2013, grant agreement 226507-NMI3) and the ANR (Programme Blanc DYN FHAC ANR-10-BLAN-1306).

SUPPORTING CITATIONS

References (50,51) appear in the Supporting Material.

REFERENCES

1. Jacrot, B. 1976. The study of biological structures by neutron scattering from solution. *Rep. Prog. Phys.* 39:911–953.
2. Heller, W. T. 2010. Small-angle neutron scattering and contrast variation: a powerful combination for studying biological structures. *Acta Crystallogr. D Biol. Crystallogr.* 66:1213–1217.
3. Breyton, C., F. Gabel, …, C. Ebel. 2013. Small angle neutron scattering for the study of solubilised membrane proteins. *Eur. Phys. J. E Soft Matter.* 36:71.
4. Svergun, D. I. 2010. Small-angle X-ray and neutron scattering as a tool for structural systems biology. *Biol. Chem.* 391:737–743.
5. Petoukhov, M. V., and D. I. Svergun. 2007. Analysis of X-ray and neutron scattering from biomacromolecular solutions. *Curr. Opin. Struct. Biol.* 17:562–571.
6. Jacob-Dubuisson, F., J. Guérin, …, B. Clantin. 2013. Two-partner secretion: as simple as it sounds? *Res. Microbiol.* 164:583–595.
7. Clantin, B., A. S. Delattre, …, V. Villeret. 2007. Structure of the membrane protein FhaC: a member of the Omp85-TpsB transporter superfamily. *Science.* 317:957–961.
8. Méli, A. C., H. Hodak, …, N. Saint. 2006. Channel properties of TpsB transporter FhaC point to two functional domains with a C-terminal protein-conducting pore. *J. Biol. Chem.* 281:158–166.
9. Jacob-Dubuisson, F., V. Villeret, …, N. Saint. 2009. First structural insights into the TpsB/Omp85 superfamily. *Biol. Chem.* 390:675–684.
10. Guédin, S., E. Willery, …, F. Jacob-Dubuisson. 2000. Novel topological features of FhaC, the outer membrane transporter involved in the secretion of the Bordetella pertussis filamentous hemagglutinin. *J. Biol. Chem.* 275:30202–30210.
11. Salvay, A. G., M. Santamaria, …, C. Ebel. 2007. Analytical ultracentrifugation sedimentation velocity for the characterization of detergent-solubilized membrane proteins Ca++-ATPase and ExbB. *J. Biol. Phys.* 33:399–419.
12. Hayashi, Y., H. Matsui, and T. Takagi. 1989. Membrane protein molecular weight determined by low-angle laser light-scattering photometry coupled with high-performance gel chromatography. *Methods Enzymol.* 172:514–528.
13. le Maire, M., B. Arnou, …, J. V. Møller. 2008. Gel chromatography and analytical ultracentrifugation to determine the extent of detergent binding and aggregation, and Stokes radius of membrane proteins using sarcoplasmic reticulum Ca2+-ATPase as an example. *Nat. Protoc.* 3:1782–1795.
14. Ebel, C. 2011. Sedimentation velocity to characterize surfactants and solubilized membrane proteins. *Methods.* 54:56–66.
15. Gosh, R. E., S. U. Egelhaaf, and A. R. Rennie. 2006. A computing guide for small-angle scattering experiments. Institute Laue-Langevin internal report. ILL06GH05T. ftp://ftp.ill.fr/pub/cs/sans/sans_manual.pdf.
16. Guinier, A. 1939. La diffraction des rayons X aux très petits angles; application à l'étude de phénomènes ultramicroscopiques. *Ann. Phys.* 12:166–237.
17. Konarev, P. V., V. V. Volkov, …, D. I. Svergun. 2003. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Cryst.* 36:1277–1282.

18. Svergun, D. I., C. Barberato, and M. H. J. Koch. 1995. CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* 28:768–773.

19. Svergun, D. I., S. Richard, …, G. Zaccai. 1998. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA.* 95:2267–2272.

20. Malde, A. K., L. Zuo, …, A. E. Mark. 2011. An automated force field topology builder (ATB) and repository: version 1.0. *J. Chem. Theory Comput.* 7:4026–4037.

21. Hess, B., C. Kutzner, …, E. Lindahl. 2008. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4:435–447.

22. van Gunsteren, W. F., S. R. Billeter, …, I. G. Tironi. 1996. Biomolecular Simulation, the GROMOS96 Manual and User Guide. vdf Hochschulverlag AG an der ETH, Zürich.

23. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. *J. Chem. Phys.* 126:014101.

24. Hess, B. 2008. P-LINCS: a parallel linear constraint solver for molecular simulation. *J. Chem. Theory Comput.* 4:116–122.

25. Petrey, D., Z. Xiang, …, B. Honig. 2003. Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins.* 53 (*Suppl 6*):430–435.

26. Zhang, Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics.* 9:40.

27. Lensink, M. F., C. Govaerts, and J. M. Ruysschaert. 2010. Identification of specific lipid-binding sites in integral membrane proteins. *J. Biol. Chem.* 285:10519–10526.

28. Kandt, C., W. L. Ash, and D. P. Tieleman. 2007. Setting up and running molecular dynamics simulations of membrane proteins. *Methods.* 41:475–488.

29. Lomize, A. L., I. D. Pogozheva, and H. I. Mosberg. 2011. Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes. *J. Chem. Inf. Model.* 51:930–946.

30. Lomize, M. A., A. L. Lomize, …, H. I. Mosberg. 2006. OPM: orientations of proteins in membranes database. *Bioinformatics.* 22:623–625.

31. Chandrasekhar, J., D. C. Spellmeyer, and W. L. Jorgensen. 1984. Energy component analysis for dilute aqueous solutions of Li+, Na+, F−, and Cl− ions. *J. Am. Chem. Soc.* 106:903–910.

32. le Maire, M., P. Champeil, and J. V. Møller. 2000. Interaction of membrane proteins and lipids with solubilizing detergents. *Biochim. Biophys. Acta.* 1508:86–111.

33. Berthaud, A., J. Manzi, …, S. Mangenot. 2012. Modeling detergent organization around aquaporin-0 using small-angle X-ray scattering. *J. Am. Chem. Soc.* 134:10080–10088.

34. Compton, E. L., E. Karinou, …, A. Javelle. 2011. Low resolution structure of a bacterial SLC26 transporter reveals dimeric stoichiometry and mobile intracellular domains. *J. Biol. Chem.* 286:27058–27067.

35. Johs, A., M. Hammel, …, R. Prassl. 2006. Modular structure of solubilized human apolipoprotein B-100. Low resolution model revealed by small angle neutron scattering. *J. Biol. Chem.* 281:19732–19739.

36. Zimmer, J., D. A. Doyle, and J. G. Grossmann. 2006. Structural characterization and pH-induced conformational transition of full-length KcsA. *Biophys. J.* 90:1752–1766.

37. Breyton, C., A. Flayhan, …, C. Ebel. 2013. Assessing the conformational changes of pb5, the receptor-binding protein of phage T5, upon binding to its Escherichia coli receptor FhuA. *J. Biol. Chem.* 288:30763–30772.

38. Clifton, L. A., C. L. Johnson, …, J. H. Lakey. 2012. Low resolution structure and dynamics of a colicin-receptor complex determined by neutron scattering. *J. Biol. Chem.* 287:337–346.

39. O'Neill, H., W. T. Heller, …, E. Greenbaum. 2007. Small-angle X-ray scattering study of photosystem I-detergent complexes: implications for membrane protein crystallization. *J. Phys. Chem. B.* 111:4211–4219.

40. Cardoso, M. B., D. Smolensky, …, H. O'Neill. 2009. Insight into the structure of light-harvesting complex II and its stabilization in detergent solution. *J. Phys. Chem. B.* 113:16377–16383.

41. Koutsioubas, A., A. Berthaud, …, J. Pérez. 2013. Ab initio and all-atom modeling of detergent organization around Aquaporin-0 based on SAXS data. *J. Phys. Chem. B.* 117:13588–13594.

42. Wolf, M. G., M. Hoefling, …, G. Groenhof. 2010. g_membed: efficient insertion of a membrane protein into an equilibrated lipid bilayer with minimal perturbation. *J. Comput. Chem.* 31:2169–2174.

43. Bond, P. J., J. M. Cuthbertson, …, M. S. Sansom. 2004. MD simulations of spontaneous membrane protein/detergent micelle formation. *J. Am. Chem. Soc.* 126:15948–15949.

44. Friemann, R., D. S. Larsson, …, D. van der Spoel. 2009. Molecular dynamics simulations of a membrane protein-micelle complex in vacuo. *J. Am. Chem. Soc.* 131:16606–16607.

45. Khao, J., J. Arce-Lopera, …, J. P. Duneau. 2011. Structure of a protein-detergent complex: the balance between detergent cohesion and binding. *Eur. Biophys. J.* 40:1143–1155.

46. Neale, C., H. Ghanei, …, R. Pomès. 2013. Detergent-mediated protein aggregation. *Chem. Phys. Lipids.* 169:72–84.

47. Lee, S. Y., A. Lee, …, R. MacKinnon. 2005. Structure of the KvAP voltage-dependent K+ channel and its dependence on the lipid membrane. *Proc. Natl. Acad. Sci. USA.* 102:15441–15446.

48. Gohon, Y., T. Dahmane, …, C. Ebel. 2008. Bacteriorhodopsin/amphipol complexes: structural and functional properties. *Biophys. J.* 94:3523–3537.

49. Sonntag, Y., M. Musgaard, C. Olesen, B. Schiott, J. V. Moller, P. Nissen, and L. Thogersen. 2011. Mutual adaptation of a membrane protein and its lipid bilayer during conformational changes. *Nat. Commun.* 2:304.

50. Solovyova, A., P. Schuck, …, C. Ebel. 2001. Non-ideality by sedimentation velocity of halophilic malate dehydrogenase in complex solvents. *Biophys. J.* 81:1868–1880.

51. Le Roy, A., H. Nury, …, C. Ebel. 2013. Sedimentation velocity analytical ultracentrifugation in hydrogenated and deuterated solvents for the characterization of membrane proteins. *Methods Mol. Biol.* 1033: 219–251.

PLOS | PATHOGENS

RESEARCH ARTICLE

Virulence Regulation with Venus Flytrap Domains: Structure and Function of the Periplasmic Moiety of the Sensor-Kinase BvgS

Elian Dupré[1,2,3,4☯], Julien Herrou[1,2,3,4☯¤], Marc F. Lensink[5], René Wintjens[6], Alexey Vagin[7], Andrey Lebedev[8], Sean Crosson[9], Vincent Villeret[5], Camille Locht[1,2,3,4], Rudy Antoine[1,2,3,4‡*], Françoise Jacob-Dubuisson[1,2,3,4‡*]

1 Center for Infection and Immunity (CIIL), Institut Pasteur de Lille, Lille, France, **2** Center for Infection and Immunity (CIIL), University Lille North of France, Lille, France, **3** UMR 8204, Centre National de la Recherche Scientifique (CNRS), Lille, France, **4** U1019, Institut National de la Santé et de la Recherche Médicale (INSERM), Lille, France, **5** Unité de Glycobiologie Structurale et Fonctionnelle, CNRS UMR8576, University Lille North of France, Villeneuve d'Ascq, France, **6** Laboratory of Biopolymers and Supramolecular Nanomaterials, Université Libre de Bruxelles, Brussels, Belgium, **7** Structural Biology Laboratory, University of York, York, England, United Kingdom, **8** Research Complex at Harwell, Science and Technology Facilities Council Rutherford Appleton Laboratory, Didcot, England, United Kingdom, **9** Department of Biochemistry & Molecular Biology, University of Chicago, Chicago, Illinois, United States of America

☯ These authors contributed equally to this work.
¤ Current Address: Department of Biochemistry & Molecular Biology, University of Chicago, Chicago, Illinois, United States of America
‡ RA and FJD also contributed equally to this work.
* rudy.antoine@pasteur-lille.fr (RA); francoise.jacob@ibl.cnrs.fr (FJD)

Citation: Dupré E, Herrou J, Lensink MF, Wintjens R, Vagin A, Lebedev A, et al. (2015) Virulence Regulation with Venus Flytrap Domains: Structure and Function of the Periplasmic Moiety of the Sensor-Kinase BvgS. PLoS Pathog 11(3): e1004700. doi:10.1371/journal.ppat.1004700

Editor: Craig R. Roy, Yale University School of Medicine, UNITED STATES

Received: November 26, 2014

Accepted: January 14, 2015

Published: March 4, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by Grant N° ANR-13-BSV8-0002-01 from the Agence Nationale de la Recherche (http://www.agence-nationale-recherche.fr/) to FJD. ED acknowledges the receipt of predoctoral fellowships from the French Research Ministry and the Fonds de la Recherche Médicale (FRM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Two-component systems (TCS) represent major signal-transduction pathways for adaptation to environmental conditions, and regulate many aspects of bacterial physiology. In the whooping cough agent *Bordetella pertussis*, the TCS BvgAS controls the virulence regulon, and is therefore critical for pathogenicity. BvgS is a prototypical TCS sensor-kinase with tandem periplasmic Venus flytrap (VFT) domains. VFT are bi-lobed domains that typically close around specific ligands using clamshell motions. We report the X-ray structure of the periplasmic moiety of BvgS, an intricate homodimer with a novel architecture. By combining site-directed mutagenesis, functional analyses and molecular modeling, we show that the conformation of the periplasmic moiety determines the state of BvgS activity. The intertwined structure of the periplasmic portion and the different conformation and dynamics of its mobile, membrane-distal VFT1 domains, and closed, membrane-proximal VFT2 domains, exert a conformational strain onto the transmembrane helices, which sets the cytoplasmic moiety in a kinase-on state by default corresponding to the virulent phase of the bacterium. Signaling the presence of negative signals perceived by the periplasmic domains implies a shift of BvgS to a distinct state of conformation and activity, corresponding to the avirulent phase. The response to negative modulation depends on the integrity of the periplasmic dimer, indicating that the shift to the kinase-off state implies a concerted conformational transition. This work lays the bases to understand virulence regulation in *Bordetella*. As homologous sensor-kinases control virulence features of diverse bacterial

Competing Interests: The authors have declared that no competing interests exist.

pathogens, the BvgS structure and mechanism may pave the way for new modes of targeted therapeutic interventions.

Author Summary

Bacteria make use of two-component transduction systems, composed of a sensor-kinase and a response regulator, to perceive environmental signals and orchestrate an appropriate response. The virulence regulon of the whooping cough agent *Bordetella pertussis* is controlled by the two-component system BvgAS. The sensor-kinase BvgS harbor extra-cytoplasmic Venus flytrap perception domains similar to those found in neuronal receptors, and it is the prototype of a large bacterial protein family. We report the atomic structure of the extra-cytoplasmic moiety of BvgS, which shows a novel dimeric arrangement. We show that the virulent phase of *B. pertussis* that occurs by default corresponds to a specific conformation of BvgS generated by the periplasmic architecture itself and by the differential dynamics of its Venus flytrap domains. The perception of negative signals by the periplasmic domains causes BvgS to shift to a different conformation that corresponds to the avirulent phase of the bacteria. In addition to contributing to our understanding of virulence regulation by *B. pertussis* at a time of whooping cough re-emergence, this study also paves the way to the mechanistic exploration of the homologous sensor-kinases found in various bacterial pathogens.

Introduction

Two-component sensory transduction systems (TCSs) regulate various physiological processes in response to environmental changes [1]. They are abundant throughout the phylogenetic tree except for vertebrates and represent major bacterial signaling pathways [2,3]. TCSs notably regulate the cell cycle, motility, biofilm formation or antibiotic resistance, as well as the virulence of major pathogens [4–8]. TCSs are typically composed of a sensor-kinase activated by environmental stimuli and a response regulator mediating phosphorylation-dependent effects [9,10]. Upon perception of a physical or chemical signal, auto-phosphorylation of a conserved cytoplasmic His residue of the sensor-kinase is followed by transfer of the phosphoryl group to a conserved Asp residue of the response regulator. The phosphorylated response regulator mediates a specific, frequently transcriptional, cellular response [11]. There is considerable diversity among TCSs regarding domain composition and organization [9,10].

Bordetella pertussis, the whooping cough agent, colonizes the upper respiratory tract of humans [12]. Transcription of its virulence regulon is positively regulated by the TCS BvgAS [13]. Over one hundred genes belong to the Bvg regulon, including those coding for the adhesins and toxins and their secretion and assembly machineries [14]. The virulent, Bvg^+ phase, in which phosphorylated BvgA trans-activates the expression of the virulence regulon, is essential for the development of the infection cycle of *B. pertussis* and other pathogenic *Bordetella* species [13,15]. The kinase and phosphotransfer activities of BvgS are maximal (referred to below as the 'kinase-on' state) without specific chemical stimuli and at 37°C, the *B. pertussis* host body temperature, while low temperatures and specific negative modulators turn these activities off in laboratory conditions (referred to below as the 'kinase-off' state). Thus, millimolar concentrations of nicotinate or sulfate ions result in the dephosphorylation of BvgA, switching the bacteria to the avirulent, Bvg^- phase [16,17]. Virulence genes are no longer expressed, while

a smaller set of virulence-repressed genes (*vrg*s) are upregulated [18,19]. At low modulator concentrations, an intermediate Bvg^i phase occurs in which the reduced concentration of phosphorylated BvgA is sufficient to transactivate 'early' virulence genes as well as specific intermediate genes [13,20,21]. Thus, BvgAS operates like a rheostat, determining several states of gene expression that might correspond to distinct temporal or spatial situations in the course of infection. BvgS is composed of periplasmic Venus flytrap (VFT) domains, a transmembrane segment, a PAS domain, and a kinase and additional domains that make up a phosphorelay (Fig. 1A). The cytoplasmic moiety of BvgS dimerizes, similar to the other TCS sensor-kinases [22,23].

BvgS is the prototype of a family of bacterial VFT-domain-containing sensor-kinases [24]. VFT domains have a bi-lobed structure with two mobile jaws delimitating a putative ligand-binding cavity [25,26]. They exist in open and closed conformations that interconvert by clam-shell motions. Typically, binding of a ligand in the cavity stabilizes the closed conformation, which triggers downstream cellular events such as transport or signaling. The periplasmic moiety of BvgS is composed of two VFT domains, membrane-distal VFT1 and membrane-proximal VFT2. We have previously reported the structure of the isolated VFT2 domain and showed that nicotinate and related negative modulators bind to VFT2 [27]. There are currently more than 2000 predicted BvgS homologs, containing from one to five VFT domains. Some of them are found in major pathogens, including *Pseudomonas aeruginosa*, *Vibrio cholerae*, *Yersinia enterocolitica* and *Borrelia burgdorferi*, in which they regulate various responses that contribute to pathogenicity [28–32] (Fig. 1B). Unlike those of classical TCSs, the molecular mechanisms of signal perception and transduction by these VFT-containing sensor-kinases are largely unknown.

In this work, we describe the structure of the periplasmic portion of BvgS, revealing a novel homo-dimeric architecture with two highly intricate polypeptide chains wound around each other. A combination of site-directed mutagenesis, functional analyses *in vivo* and molecular modeling indicated that the integrity of the periplasmic domain is necessary both to maintain BvgS in a kinase-on state by default and to bring about conformational changes that switch the protein to the kinase-off state in response to negative modulation. This study shows that BvgS represents a new paradigm of bacterial two-component sensor-kinases and contributes to our understanding of virulence regulation in *Bordetella*.

Results

Structure of the periplasmic domain of BvgS

The periplasmic domain of BvgS (residues Ala_{29}-Leu_{544}, which includes VFT1 and VFT2) was produced in *Escherichia coli* and crystallized as a recombinant protein with a 60-residue-long GB1 domain at the N terminus and a 6-His tag at the C terminus. The structure was solved to a resolution of 3.1 Å (Fig. 2, S1 Table). BvgS forms intricate butterfly-shaped dimers in which the A and B polypeptide chains ('protomers') wind around each other, with an extensive dimeric interface of ≈ 4000 Å^2. The two protomers overlap with an RMSD of 1.184 Å. The N-terminal GB1 domain and C-terminal His tag are not visible in the electron density maps.

A two-fold symmetry axis runs parallel to the long axis of the BvgS dimer. The N termini of the two protomers are located on the outer surface of the dimer, and their C termini interrupt α helices at the membrane-proximal end of the structure. VFT1 and VFT2 adopt typical Venus flytrap architectures consisting of two α/β subdomains called lobes 1 and 2 (hereafter L1 and L2) separated by a cleft. They have similar topologies with two crossings between the lobes (S1 Fig). The hinge is formed of anti-parallel β strands in VFT2 and flexible loops in VFT1. The VFT2s are followed by the C-terminal (Ct) domains that encompass the Gly_{527}-Pro_{532}

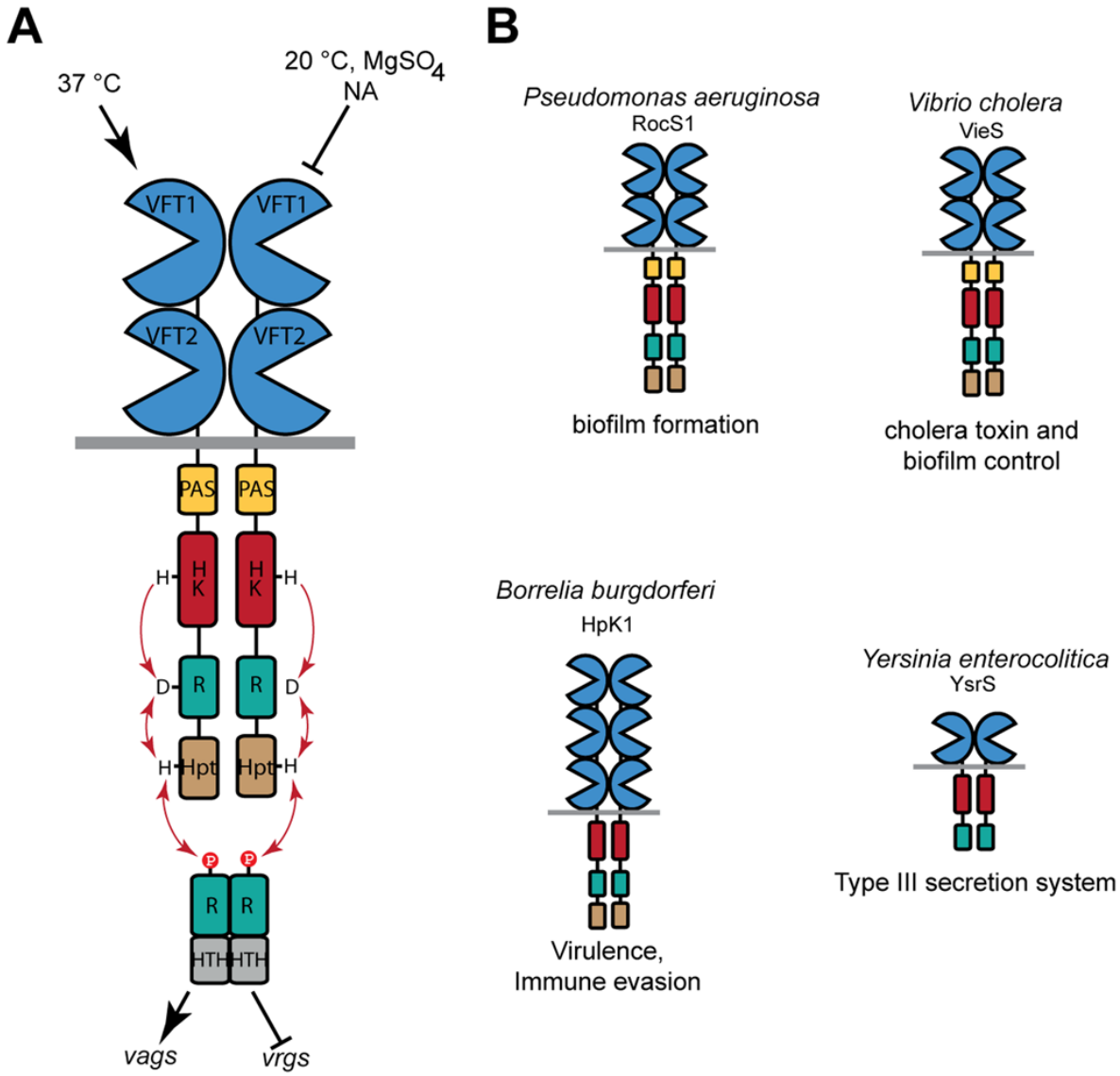


Fig 1. Function of BvgS and selected homologs. A. Schematic representation of virulence regulation by BvgAS in *B. pertussis*. Only the virulent (Bvg^+) and avirulent (Bvg^-) phases of the bacterium are represented for simplicity. Conditions that turn the bacteria to the avirulent phase include low temperatures and negative modulators such as sulfate or nicotinate (NA) ions. The *vags* (virulence-activated genes) are trans-activated by phosphorylated BvgA, while the *vrgs* (virulence-repressed genes) are upregulated in the avirulent phase. An intermediate phase occurs at low modulator concentrations (see text). From N to C terminus, 135 kDa-BvgS is composed of two periplasmic VFT domains, a transmembrane segment, a PAS domain, followed by a histidine-kinase (HK), a receiver (R) and a Histidine phosphotransfer (Hpt) domains that make up a phosphorelay (represented by arrows). BvgA is composed of a receiver domain and a helix-turn-helix DNA-binding domain (HTH). B. Structural organization of selected BvgS homologs, with the same color code as for BvgS. Note that the domain composition varies in the family. The cellular functions regulated by these sensor-kinases are also indicated.

doi:10.1371/journal.ppat.1004700.g001

Ct loops and the H19 Ct helices (Figs. 2 and S1). In the absence of membrane constraints, the H19s adopt divergent orientations in the crystal structure. In full-length BvgS they are predicted to continue across the membrane down to the cytoplasmic PAS domain, with a total length of 60 residues.

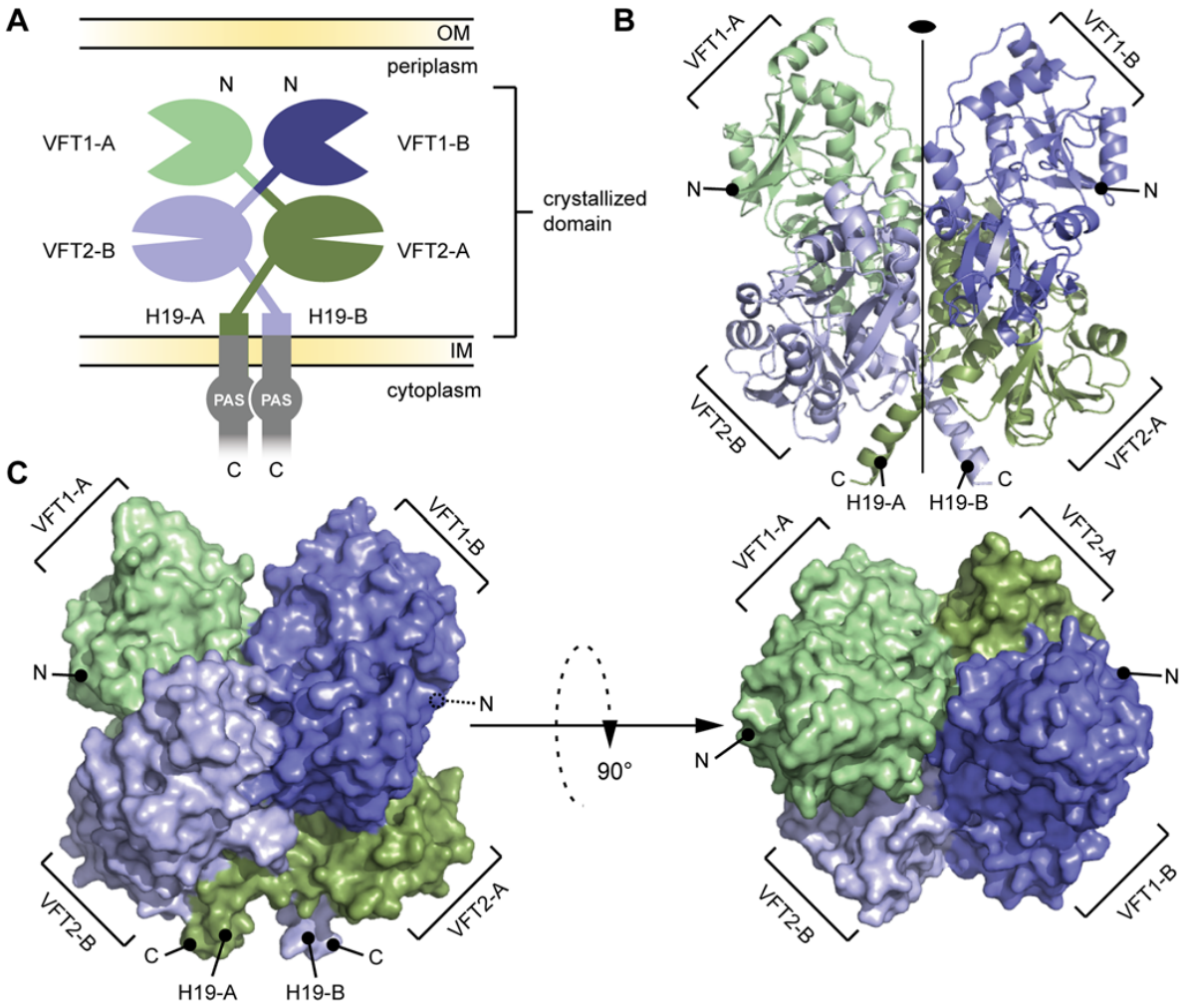


Fig 2. General organization of the BvgS periplasmic domain. A. Schematic representation of the homodimeric BvgS periplasmic portion. The protomers A and B are shown in shades of green and blue, respectively. One protomer consists of two VFT domains and a C-terminal H19 α helix. B. Ribbon representation of the X-ray structure of the BvgS periplasmic domain, the same color code as in (A) is used to show the different VFTs. The two-fold symmetry axis is indicated. C. Surface representation of the periplasmic domain of BvgS. On the left, the view angle is similar to (B), while on the right, a 90° clockwise rotation along the x-axis was applied. N and C denote the N and C termini of the two protomers.

doi:10.1371/journal.ppat.1004700.g002

The two VFT1s are open, while atypically the VFT2s are closed with no ligand in their inter-lobe cavities (Fig. 3), consistent with the structure of VFT2 alone [27]. The VFT1 cavities are each oriented toward the hinge of the VFT2 domain of the other protomer, and the cavities of the VFT2s are each oriented toward the H19 helix of the opposite protomer (Fig. 3).

The $VFT1_{L1s}$ interact with each other through several hydrogen bonds between their H8s, while the VFT2s are not directly interconnected. Both lobes of the VFT1s, $VFT1_{L1}$ and $VFT1_{L2}$, contact the hinge and lobes of VFT2 of the opposite protomer (Fig. 4), forming the largest dimeric interfaces. Other large interfaces occur between $VFT1_{L2}$ and VFT2 of the same protomer, and between $VFT2_{L2}$ and the Ct domains. In particular, both the Ct loop and the N terminus of H19 strongly interact with $VFT2_{L2}$ of the opposite protomer through hydrogen

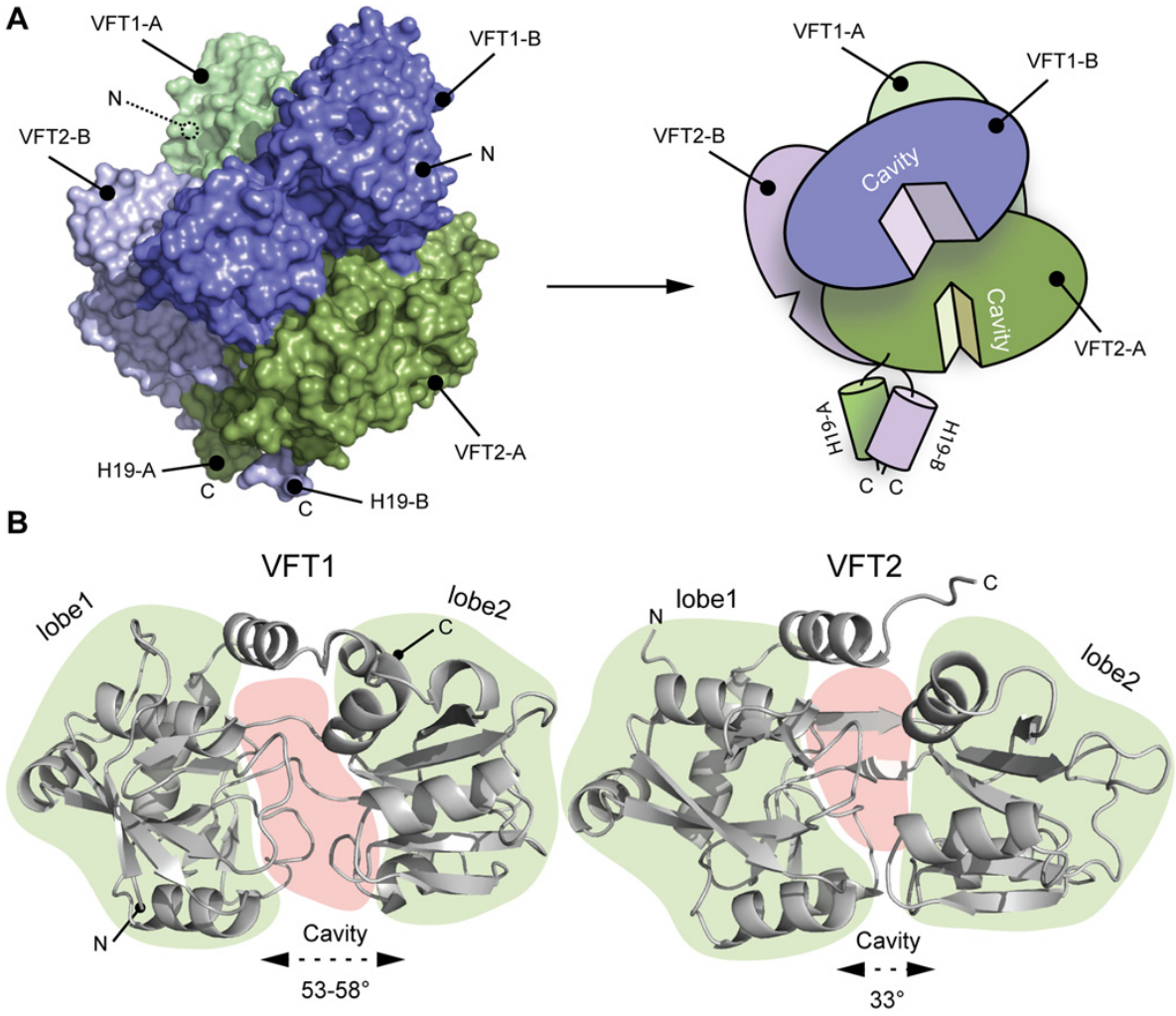


Fig 3. Characterization of the VFT domains. A. Surface and cartoon representation of BvgS showing that VFT1-B is open and VFT2-A is in an apo-closed conformation. B. Ribbon representation of the open VFT1 and closed VFT2 domains. The lobes are delimited in light green and the cavities in light red. The opening angles for the VFTs are given. The linker (H9) joining VFT1 and VFT2 and the Ct loop that follows VFT2 have been included in the representation of the VFT1 and VFT2 domains, respectively. N and C indicate the N and C termini of each protomer (in A) or VFT domain (in B).

doi:10.1371/journal.ppat.1004700.g003

bonds and through π-stacking interactions that involve a conserved residue in the BvgS family, Trp_{535} (Fig. 4).

Conformation and dynamics of the VFT domains

In the crystal structure, the VFT1 domains are open and unliganded, while conversely the VFT2 domains are closed without ligands. We performed normal mode analyses of BvgS motions based on a Gaussian network model to identify the main global motions that are accessible to the protein based on its tridimensional structure. The first, lowest-frequency normal modes are usually most relevant to function. For BvgS, the first two modes of motion consist of large motions of one $VFT1_{L1}$ (S2 Fig). In contrast, the VFT2s move together as a rigid body, as

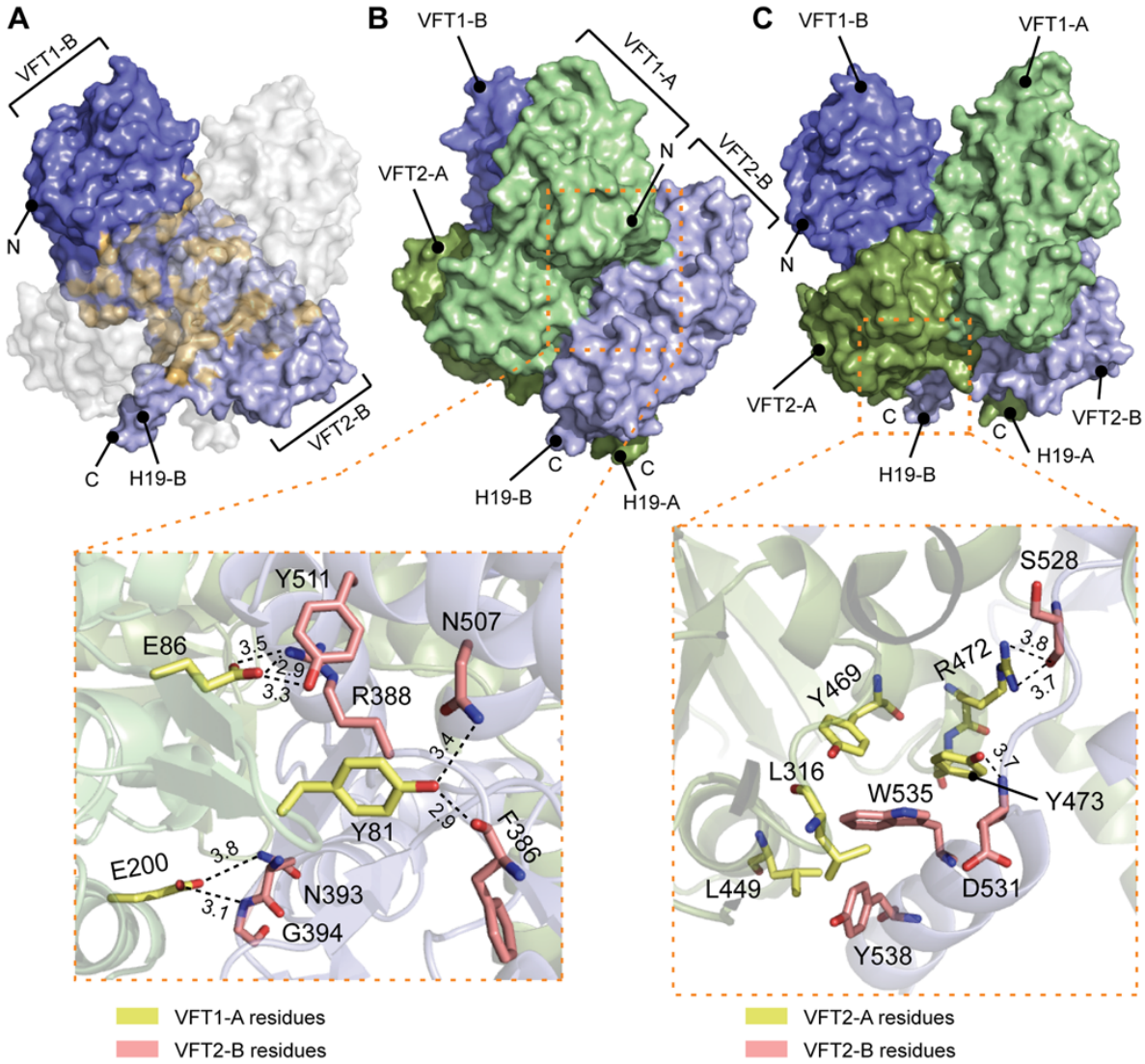


Fig 4. Interfaces between the VFT domains important for the kinase-on state. A. Surface representation of protomer B (in blue); the residues interacting with protomer A are shown in orange. To help visualizing these interactions, a "ghost" protomer A is represented in transparent white on top of protomer B. B. Illustration of the VFT1-VFT2 inter-protomer interface. A side view of BvgS is shown in surface representation, with the VFT1 of one protomer in green and the VFT2 of the other protomer in pale blue. A zoom delimited by a dashed orange box shows specific residues that are critical for BvgS function, as shown by mutagenesis. The side chains of Tyr_{81} and Glu_{86} of the β hairpin in $VFT1_{L1}$ form hydrogen bonds with Phe_{386} and Arg_{388} at one extremity of the VFT2 hinge, and with residues of the α helix H17. Glu_{200} belongs to $VFT1_{L2}$, and its side chain makes hydrogen bonds with Asn_{393} and Gly_{394} at the other extremity of the VFT2 hinge. C. Illustration of the VFT2-Ct domain inter-protomer interface. In the upper panel, BvgS is shown in surface representation, with protomer A in green and protomer B in blue. A zoom shows specific residues involved in critical interactions for BvgS kinase activity. Thus, Trp_{535} from H19 stacks in a hydrophobic and aromatic pocket mainly lined with $VFT2_{L2}$ residues of the other protomer, and Arg_{472} and Tyr_{473} from helix H16 in $VFT2_{L2}$ interact with Ser_{528} and Asp_{531} in the Ct loop of the other protomer. Hydrogen-bond distances are reported in angstroms.

doi:10.1371/journal.ppat.1004700.g004

shown in mode #3. Mode #4 consists of motions of both $VFT1_{L2s}$ together with the VFT2s. Thus, the first lobes of the VFT1s in particular can make large motions, while the VFT2 motions are more restrained and mainly coupled to each other and to those of the VFT1s. This was confirmed by performing molecular dynamics simulations to measure the evolution of the

opening angles of the VFTs over time. In the first parts of the simulations, the VFT1s make clamshell motions, while motions of the VFT2s are limited around their closed conformations (S2 Fig). As the simulations progress the VFT1 mobility is reduced, which suggests that sustained VFT motions may require the feedback from the transmembrane and cytoplasmic portions of BvgS absent from our model. These *in silico* analyses thus indicate that the X-ray structure reflects *bona fide* differences between the VFT1 and VFT2 domains in terms of conformation and dynamics.

We then asked whether VFT1 closing—as might happen upon binding of a ligand—would affect BvgS activity. We locked the VFT1 domains in closed conformations by generating a disulfide (S-S) bonds across their cavity [33,34]. Two residues located on the edges of the lobes were replaced by Cys to obtain $BvgS_{E113C+N177C}$ (S3 Fig). The corresponding point mutations were inserted into the chromosomal *bvg* locus by allelic exchange, and we verified the production of the protein and the formation of the S-S bond by immunoblotting (S4 Fig). The *in vivo* effect of the substitution on BvgS function was then measured by using a reporter system with the *lacZ* gene under the control of the Bvg-regulated *ptx* promoter [35]. *In vivo* formation of the S-S bond in VFT1 abrogates the kinase activity of BvgS (Fig. 5). This phenotype is reverted by the addition of a reducing agent, TCEP, to the growth medium (S5 Fig), which confirms that the S-S bond forms *in vivo* and shows that the loss of function is related to its presence and not to the Cys substitutions.

The VFT2s remain closed even when isolated [27]. Nevertheless, to maintain them closed *in vivo* we also generated an S-S bond between their lobes using a similar method as above, yielding $BvgS_{T355C+D442C}$. We checked that the S-S bond was formed (S4 Fig; see also below). In contrast to VFT1, closing VFT2 was found to have no effect on the BvgS kinase activity as determined with the *ptx-lacZ* reporter (Figs. 5 and S5). Altogether thus, closing of the VFT1 domains and/or restraining their mobility abrogate BvgS kinase activity. In contrast, closed VFT2 domains correspond to the kinase-on state of BvgS. The different conformations and dynamics of the two VFT domains thus contribute to BvgS function.

Importance of periplasmic domain integrity for BvgS kinase activity

B. pertussis is in the virulent, Bvg^+ phase by default at 37°C. To determine the role of the periplasmic domain of BvgS in maintaining this kinase-on state, we loosened the connections between the periplasmic and cytoplasmic moieties of BvgS by replacing Trp_{535} with Ala. This residue is located in the C-terminal helix H19 and it contributes to connecting each H19 to the $VFT2_{L2}$ of the opposite protomer (Fig. 4C). After allelic exchange, the effect of the substitution on BvgS function was measured by using the *ptx-lacZ* reporter system. The $BvgS_{W535A}$ variant has no kinase activity (Fig. 5). The presence of $BvgS_{W535A}$ in *B. pertussis* membranes was verified, showing that the substitution does not affect the structure of the protein in such a way as to prevent its integration in the membrane or to cause its proteolytic degradation *in vivo* (S4 Fig). Thus, the kinase-on state of BvgS depends on tight connections between the periplasmic domains and the transmembrane H19 helices.

To confirm that the periplasmic portion imposes a specific conformation on the cytoplasmic moiety, we introduced other substitutions in the inter-protomer interfaces between the VFT2s and the Ct domains, by targeting residues whose side chains connect the $VFT2_{L2s}$ and the Ct loops that precede the H19s (Fig. 4C). Thus, Arg_{472} and Tyr_{473} located in helix H16 of $VFT2_{L2}$ form hydrogen bonds with residues of the Ct loop of the other protomer. Their simultaneous replacement by Ala abolishes BvgS kinase activity, while the single-substitution variants $BvgS_{R472A}$ and $BvgS_{Y473A}$ are partially active (Figs. 5 and S5). This indicates that the inter-protomer interface between H16 in VFT2 and the Ct loop is critical and that it is maintained

Substitutions	Targeted interaction	Activity
WT		
W535A	VFT2 L2 - C dom (inter)	nd
E113C+N177C	SS bond in VFT1 (intra)	nd a
T355C+D442C	SS bond in VFT2 (intra)	 a
Y81A+E86A	VFT1 L1 - VFT2 hinge (inter)	nd
E200A	VFT1 L2 - VFT2 hinge (inter)	nd
R472A+Y473A	VFT2 L2 - C dom	nd
R472A	VFT2 L2 - C dom	
Y473A	VFT2 L2 - C dom	
R314G+T325G+ D326G+E327G	VFT2 L1 - C dom (inter)	
R160A	VFT1 L2 - VFT2 L1 (intra)	
R234A	VFT1 L2 - VFT2 L1 (intra)	 b
F230A+S287A	VFT1 L2 - VFT2 L1 (intra)	
R526A	VFT2 L2 - C dom (intra)	
Q463A	VFT1 L2 - VFT2 L2 (inter) + intra VFT2	
N231A	VFT1 L2 - C dom (inter)	
S271A+E272A+ R274A+S275A	VFT1 L1 - VFT1 L1 (inter)	 b

Fig 5. *In vivo* effects of the substitutions in BvgS. A *lacZ* reporter gene under the control of the Bvg-regulated *ptx* promoter was used for determination of BvgS kinase activity in standard or modulated culture conditions. Blue and pink bars indicate kinase activity levels of bacteria producing the indicated BvgS variants and grown without or with 8 mM nicotinate, respectively, with the standard errors of the mean calculated from three distinct experiments. The middle column indicates the interfaces in which the targeted interactions are located, with inter- and intra-protomer interfaces designated 'inter' and 'intra', respectively. Nd, no β-gal activity detected; a, wild type activity and/or modulation recovered when cells were grown in the presence of TCEP; b, BvgS variants only responsive to high nicotinate concentrations (20 mM). The full set of data is shown in S5 Fig.

doi:10.1371/journal.ppat.1004700.g005

by partly redundant interactions. In contrast, substitutions at the tip of a β hairpin in $VFT2_{L1}$ whose residues interact with the other face of the Ct loop do not affect BvgS function, as shown with $BvgS_{R324G/T325G/D326G/E327G}$. The effect of disrupting of specific interactions between the $VFT2_{L2s}$ and the Ct loops preceding the H19s is consistent with the effect of the $W_{535}A$ substitution, showing that the kinase-on state depends on VFT2-Ct domain inter-protomer connections.

To identify additional architectural features of the periplasmic dimer critical to maintain BvgS in its kinase-on state, we disrupted specific interactions in other intra-dimer interfaces of BvgS by site-directed mutagenesis. We targeted residues in the large interfaces between the VFT1s and the VFT2s of the opposite protomers (Fig. 4B). The side chains of Tyr_{81} and Glu_{86} in a β hairpin of $VFT1_{L1}$ and that of Glu_{200} in helix H5 of $VFT1_{L2}$ form hydrogen bonds with residues at the N- and C-terminal sides of the first hinge strand of VFT2, respectively. Two BvgS variants, $BvgS_{Y81A+E86A}$ and $BvgS_{E200A}$ were generated and analyzed as above (Figs. 5, S4 and S5). Neither of them is functional, demonstrating that connections between the two lobes of VFT1 and the hinge of VFT2 of the opposite protomer are essential to maintain the kinase-on state of BvgS. In contrast, the replacement of Gln_{463} by Ala in the same large inter-protomer VFT1-VFT2 interface does not affect activity (Figs. 5 and S5). Gln_{463} is part of VFT2 but not located in the hinge, unlike the residues of VFT2 in contact with Tyr_{81}, Glu_{86} and Glu_{200}. The loss of kinase activity of the $BvgS_{Y81A+E86A}$ and $BvgS_{E200A}$ variants might result from the loss of constraints applied by the VFT1 lobes on the VFT2 hinge.

In contrast, disruption of specific interactions in other dimeric interfaces (S3 Fig), including the H8-mediated VFT1-VFT1 inter-protomer interface, the VFT1-VFT2 intra-protomer interfaces, the VFT2-Ct domains intra-protomer interfaces or the VFT1-Ct domains inter-protomer interfaces, does not markedly affect Bvg kinase activity (Figs. 5 and S5).

Altogether, thus, we have identified interactions in the inter-protomer interfaces between VFT1 and the VFT2 hinge and between $VFT2_{L2}$ and the Ct domain that are necessary to maintain BvgS in its kinase-on state. In particular, the substitutions $A_{472}A+Y_{473}A$ and $W_{535}A$ support the idea that the periplasmic domain exerts a strain on the transmembrane domains, causing the cytoplasmic moiety to adopt a specific conformation corresponding to the kinase-on state. The VFT1s contribute to the strain via the close contacts of their two lobes with the hinges of the tight VFT2 domains. Loosening the periplasmic portion or its connections with the transmembrane helices releases the strain, and therefore the cytoplasmic moiety switches to a distinct, kinase-off state.

Modulation by nicotinate requires multiple intra-dimer interactions

Negative modulators turn BvgS to the kinase-off state at millimolar concentrations in laboratory conditions, and they possibly mimic *in vivo* ligands that might decrease or turn off virulence genes expression at specific stages of the infection. The sites of interaction of these negative modulators are mostly unknown. We have shown that nicotinate binds to isolated VFT2, even though additional sites cannot be ruled out in the dimer [29], and therefore we used nicotinate to determine how the periplasmic moiety contributes to the response of BvgS to negative modulation. The ability of the BvgS variants described above to respond to nicotinate was thus assessed.

The $BvgS_{T355C+D442C}$ variant with a S-S bond across the VFT2 cavity variant is unresponsive to nicotinate but reverts to the wild type (wt) modulation phenotype when the growth medium is supplemented with TCEP (Figs. 5 and S5). This confirms the *in vivo* formation of the S-S bond and also shows that it, rather than the Cys substitutions, hampers the response to

nicotinate. The S-S bond might prevent nicotinate from binding or hamper a conformational changes involved in the response to the negative modulator.

A number of other substitutions similarly abrogate the effect of nicotinate (Figs. 5 and S5). Interestingly, both inter-protomer and intra-protomer interactions are required for BvgS response to negative modulation. These interactions map to the $VFT1_{L1}$-$VFT1_{L1}$, $VFT1_{L2}$-$VFT2_{L1}$, and $VFT1_{L2}$-$VFT2_{L2}$ inter-protomer interfaces and to the $VFT1_{L2}$-$VFT2_{L1}$ and $VFT2_{L2}$-Ct domain intra-protomer interfaces (Figs. 5 and S3). Altogether, a large set of both inter- and intra-protomer interactions is required for the response of BvgS to nicotinate. The fact that the response to negative modulation strongly depends on the integrity of the periplasmic moiety indicates that the transition from the kinase-on state to the kinase-off state implies a concerted conformational change.

Function of BvgS heterodimers

The importance of the structural integrity of the periplasmic domain for the kinase-on state and for the transition to the kinase-off state was further probed by generating *in vivo* BvgS heterodimers that harbor one wt periplasmic domain and another one with a substitution. A merodiploid containing two inactive but complementary *bvgS* copies, one with a substitution of the phosphorylable Asp of the receiver domain ($D_{1023}N$) and the other with a substitution of the phosphorylable His of the Hpt domain ($H_{1172}Q$), will form inactive homodimers and active heterodimers (Fig. 6) [36,37]. Indeed, only heterodimers will be able to restore the phsophorylation cascade of BvgS. We set up this merodiploid expression system in *B. pertussis*. As shown in Fig. 6, the homodimers formed by $BvgS_{D1023N}$ or by $BvgS_{H1172Q}$ are inactive using the *ptx* reporter, but the heterodimer $BvgS_{D1023N/H1172Q}$ is functional, displaying kinase activity in the default state and responding to nicotinate like wt BvgS.

We disrupted critical contacts in one side of the dimer by combining the $W_{535}A$ variant with the wt periplasmic protomer. The kinase activity of BvgS was measured using the *ptx-lacZ* system as above. The resulting BvgS homodimer is not functional, similar to the homo-dimeric $BvgS_{W535A}$ variant (Fig. 6). Another variant that harbors the $Y_{81}A+E_{86}A$ substitutions in the inter-protomer VFT1-VFT2 interface was similarly combined with the wt periplasmic moiety. The heterodimer is also not functional, a phenotype similar to that of the $BvgS_{Y81A+E86A}$ homodimer (Fig. 6). Both results support the model that the periplasmic architecture and more specifically the crucial inter-protomer interfaces identified above impose a kinase-on conformation onto the cytoplasmic moiety via the H19 helices. Releasing the strain in one half of the dimer is sufficient to lose the kinase-on conformation.

We also combined the wt periplasmic moiety with that harboring a S-S bond across the VFT1 cavity. The resulting BvgS heterodimer has no kinase activity (Fig. 6). Thus, both protomers must have the proper conformation and dynamics for BvgS function.

We finally used the heterodimer strategy to test the effect of a substitution that makes BvgS unresponsive to nicotinate. We thus combined a protomer harboring a wt periplasmic domain with that harboring the $N_{231}A$ substitution. Asn_{231} from $VFT1_{L2}$ makes interactions with the Ct loop of the other protomer (S3 Fig), and the $BvgS_{N231A}$ homodimer does not respond to nicotinate (Figs. 5 and S5). The recombinant strain expressing the heterodimer has β-galactosidase activity and interestingly, its sensitivity to nicotinate is partially restored. Thus, the heterodimer responds to 20 mM nicotinate, although it is not fully modulated (Fig. 6). This intermediary phenotype indicates that the transition to the kinase-off state requires higher modulator concentrations when the integrity of the periplasmic domain is slightly compromised.

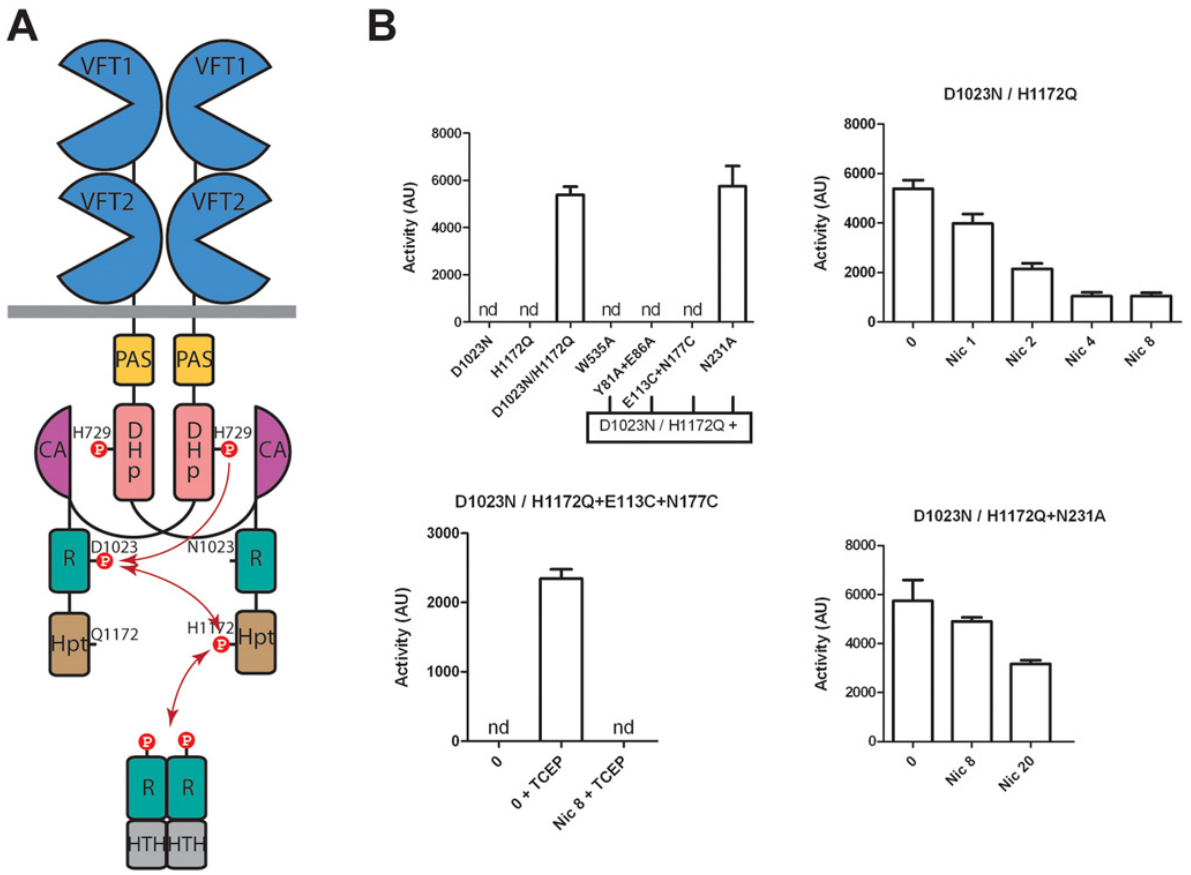


Fig 6. BvgS heterodimers. A. Schematic representation of the BvgS heterodimers. The dimerisation/Histidine phosphotransfer domain (DHp) and the catalytic ATP-binding domain (CA) of the kinase are represented separately to show the phosphorylation cascade (arrows). B. Kinase activity levels as determined using the *ptx-lacZ* reporter for *B. pertussis* harboring the indicated BvgS variants and grown in standard or modulation conditions. The first panel shows the activities of the various strains. The first two express inactive homodimers, and the last four express heterodimers in which one protomer harbors a wt periplasmic portion combined with the $D_{1023}N$ substitution and the other protomer harbors the indicated periplasmic substitution(s) combined with the $H_{1172}Q$ substitution. The last three panels show the β-gal activities of the strains expressing the indicated heterodimers, with the standard errors of the mean calculated from three distinct experiments. Nicotinate (nic) and TCEP were added at the indicated concentrations (in mM). nd, no activity detected.

doi:10.1371/journal.ppat.1004700.g006

Discussion

Although the BvgAS system was identified more than 25 years ago [38], the mode of regulation of *Bordetella* virulence has remained a puzzle. With its kinase-on state by default and its extra-cytoplasmic domain different from those of classical 'PDC' (for PhoB/ DcuS/CitA) TCS sensor-kinases, BvgS was initially considered an oddity. However, the realization that many bacterial sensor-kinases harbor similar sensor domains and the first clues about its structure and mode of action have made BvgS a model for the family [23,24,27]. Importantly, some of the BvgS homologs are found in major pathogens, including other *Bordetella* species as well as *P. aeruginosa*, *E. coli*, *V. cholerae*, *Y. enterocolitica* and *B. burgdorferi*, in which they control programs such as biofilm formation, efflux pump expression, type III secretion, or nutritional adaptation [28–32]. The BvgS structure establishes the foundations to decipher the molecular mode of action of this poorly characterized family of VFT-containing sensor-kinases, and it may pave the way to develop new, highly specific, anti-infective therapeutic strategies [39].

Our functional analyses based on the BvgS structure support the following model. Specific inter-protomer interactions are necessary to maintain the kinase-on state. The tight architecture of the periplasmic moiety together with the differential dynamics of the VFTs imposes a strain onto the transmembrane H19 helices. In response, the cytoplasmic moiety, beginning with the PAS domain, adopts specific conformation and dynamics that support the kinase and phosphotransfer activities of BvgS. The bacteria are thus in the virulent, Bvg^+ phase, and they can establish an infection. Switching BvgS to the kinase-off state involves a conformational change of the periplasmic moiety, which modifies the conformation, and possibly the dynamics, of the downstream cytoplasmic PAS and kinase domains. The roles of the avirulent or intermediate phases of *B. pertussis* are unclear, and *in vivo* stimuli that may trigger the shift to phosphatase or lower kinase states of activity remain to be identified. However, this work shows that the shift to the kinase-off state can easily be hampered by point mutations at various periplasmic sites. That the ability to reversibly perform the shift that regulates BvgS activity has been conserved through evolution supports the importance of the avirulent or intermediate phases in the lifestyle of *B. pertussis*. It also strongly argues that the VFT domains perceive negative *in vivo* signals, which explains the good conservation of their cavities in *Bordetella* [35].

As shown in this work, one can artificially turn BvgS to the kinase-off state by disrupting specific inter-protomer interactions between the VFT2 domains and the H19 helices. The release of constraints on these helices causes the cytoplasmic portion to adopt an alternative conformation in which BvgS functions as a phosphatase. We have shown that other events putatively relevant to BvgS function, i.e. the closing of VFT1 domains, which might mimic the binding of a ligand, or the binding of nicotinate to VFT2 [27], also turn BvgS to the kinase-off state. Both most likely cause conformational—and/or dynamic- changes to the periplasmic domain, with repercussions below the membrane. A number of BvgS variants with looser connections between the VFT domains are blocked in the kinase-on state and cannot respond to nicotinate, which shows that the shift to the kinase-off state implies a concerted conformational change. Modulation therefore facilitates the transition by shifting the equilibrium from the kinase-on to the kinase-off conformations. It is likely that these two stable states will also differ in their dynamics, and we have indeed obtained preliminary indications that VFT1 dynamics is modified in the modulated state. Similarly, VFT1 dynamics probably contributes to the transition, in line with the emerging paradigm that the dynamics of signaling proteins relates to their function [40].

In the default situation—i.e., at 37°C and without modulators-, the equilibrium is strongly shifted towards the kinase-on state of BvgS, which is therefore fully populated, while conversely the equilibrium is strongly shifted towards the kinase-off state in the presence of high modulator concentrations. This two-state model is compatible with intermediate levels of activity of the BvgAS system, such as those obtained at intermediate modulator concentrations [15], in which kinase-on and-off BvgS proteins may co-exist in equilibrium. It is also most likely the case for the $BvgS_{wt/N231A}$ heterodimer, in which the lack of a critical interaction on one side of the dimer hampers the transition, and therefore only a proportion of the BvgS molecules shift to the kinase-off state at high modulator concentrations. The transition between the two conformations will likely imply relative rotation, translation or shearing movements of the helices that join the periplasmic and cytoplasmic domains, similar to what has been proposed in other signaling proteins [41–44].

With its clamshell motions, VFT1 behaves like a typical VFT domain. As stated above, the conservation of the VFT1 cavity residues in *Bordetella* [35] suggests that it binds specific ligand (s) *in vivo*, and if so our results show that ligand binding to VFT1 will likely cause BvgS to shift to the kinase-off state. In contrast, the VFT2s remain closed in the kinase-on state with no *bona fide* ligand in their cavity. Whether nicotinate binding to VFT2 opens the cavity or causes

another type of deformation remains unknown, but the thermal stabilization of VFT2 upon nicotinate binding argues against the former possibility [27]. The crystal structure of the single VFT domain of a BvgS homolog, the HK29 histidine-kinase of *Geobacter sulfurreducens* interestingly shows that this VFT is also closed unliganded [45]. Its hinge is composed of two β strands, like that of VFT2 in BvgS, leading those authors to propose that it might not be able to open. Sequence analyses of BvgS homologs indicate that the regions forming the hinge of the membrane-proximal VFT domain contain fewer Gly and more Pro residues than those of classical VFT domains. Therefore, we speculate that in the BvgS family the membrane-proximal VFT domain should be closed and tight for the regulation of sensor-kinase activity. BvgS also responds to various organic and inorganic ions [46]. The binding of these modulating molecules might not necessarily involve the cavity but possibly also interfaces, as in some other VFT-containing receptors [47,48].

The periplasmic moiety of BvgS adopts a highly compact dimeric structure. The helical and strongly intertwined architecture of BvgS may explain how some of its homologs could be functional with three, four or even five predicted VFT domains in tandem [24]. The multiple VFT domains of these sensor-kinases potentially enable the perception of several chemical signals that must be integrated to determine the appropriate response. A compact structure like that of BvgS appears to be better suited for inter-domain communication than more linear arrangements such as those found in the VFT-based iGlu receptors of higher eukaryotes, which might dissipate information coming from the most distal VFT domains [49,50] (S6 Fig). This study of BvgS will undoubtedly serve as a basis to elucidate the function of the other family members. Not all BvgS homologs are in a kinase-on state by default [51], but our mechanistic model can perfectly accommodate sensor-kinases that are regulated in the opposite manner.

Materials and Methods

Crystallization of BvgS, data collection and processing

The *bvgS* sequence was amplified by PCR and introduced into pGEV2 [52]. The resulting plasmid encodes the periplasmic portion of BvgS (A_{29}-L_{544}) with N-terminal GB1 and C-terminal His tags. The recombinant protein was purified on a Ni^{2+}-Sepharose affinity column (GE Life Sciences) and eluted in 10 mM Tris—HCl (pH 8.8), 500 mM NaCl, 200 mM at 4°C. BvgS was concentrated by ultrafiltration to 20 mg/mL. The initial crystallization screening was carried out using the sitting-drop, vapor-diffusion technique in 96-wells microplates with a Cybi-Workstation (Cybio) and commercial crystallization kits (Nextal-Qiagen and JBSscreen). Extremely fragile crystals were obtained at 19°C by manual refinement in 100 mM sodium acetate (pH 4.6), 1.6 M NaCl, in 5 to 7 days. All manual crystallization attempts were carried out using the hanging-drop, vapor-diffusion technique in 24-well plates. The crystals were soaked in a stepwise fashion to a final concentration of 20% glycerol in the crystallization buffer.

A preliminary diffraction screening was performed on 80 crystals. On the best crystal, diffracting at 3.10 Å, a single diffraction dataset (160 images with an oscillating range of 1°) was collected at an X-ray wavelength of 1.5418 Å and a temperature of 100 K using an in-house Mardtb goniostat and a Mar345 image plate detector. Diffraction images were indexed and scaled using the XDS program package [53]. The crystal belongs to the space group $P2_12_12_1$, with cell parameters a = 72 Å, b = 286 Å and c = 128 Å. According to the calculated Matthews coefficient of 2.52 $Å^3$ Da^{-1}, a solvent content of 51.3% was estimated.

Structure determination and refinement

The crystal structure containing four monomers in the asymmetric unit was determined by molecular replacement using MOLREP [54] and the crystallographic structure of the isolated

VFT2 domain (PDB code: 3MPK) as a search model. Eight copies of the model were located, four occupying the actual positions of VFT2 domains and the other four those of VFT1 domains, whose sequence identity to VFT2 is 24%. The former four copies were positioned using the conventional Patterson search. The latter four copies were found using an iterative procedure alternating refinement of a partial structure with REFMAC [55] and molecular-replacement search in the electron density maps [56]. Subsequent model rebuilding and refinement of the 3.10 Å structure were conducted iteratively using Coot [57] and phenix.refine [58], with the use of local non-crystallographic symmetry restraints. Torsion angles of the structure were optimized by using the Godzilla web server (http://godzilla.uchicago.edu/) [59]. The structure was refined to final R_{work} of 18.1% and R_{free} of 24.4%. The two BvgS homodimers (AB and CD) found in the asymmetric unit can be superimposed with a Cα rmsd of 1.234 Å. A Ramachandran analysis performed with the program Phenix indicated that 94.4% of residues are in preferred conformations and 1.4% in disallowed conformations. The GB1 domains are not seen in the electron density. Analysis of crystal packing revealed an empty space close to the N-terminal segment of each polypeptide chain, indicating that they might be unseen because of crystallographic disorder.

Structure analyses

The 1026-residue AB dimer was used for all analyses. The opening angles for VFT1 and VFT2 were measured using three residues structurally equivalent between the two VFTs, one on the lip of each lobe and one in the hinge. They are Tyr_{70}, Gly_{244} and Ser_{199} for VFT1 and Leu_{314}, Glu_{490} and Pro_{444} for VFT2. The inter-domain interfaces were defined using the PISA server (http://www.ebi.ac.uk/pdbe/prot_int/pistart.html) [60], and http://capture.caltech.edu/ was used to identify cation-π interactions. A model for the closed VFT1 domain was made with Modeller [61] based on its closest homologous structure (PDB code: 1WDN), and residues to be replaced by cysteines were chosen by using http://cptweb.cpt.wayne.edu/DbD/ [62].

Molecular modeling

The methods used for the normal mode analysis and the molecular dynamics simulations with their associated references are described in S1 Protocol. For the analyses of the MD simulations, the opening angles of the VFT domains were calculated based on the geometric centers of the C-α atoms of each lobe and the hinge using a slightly extended definition of the hinge region, encompassing residues 146–151 and 241–246 for that of VFT1, and 390–395 and 486–491 for that of VFT2, to make them less susceptible to noise.

Measurement of BvgS activity

Point mutations were introduced into the chromosome of *B. pertussis* BPSM by allelic exchange [35]. The BvgS sequence corresponds to that of TohamaI except for a Glu residue at position 705, as in most *B. pertussis* strains [35]. A *ptx-lacZ* transcriptional fusion was generated in each recombinant strain [63]. The strains were grown in modified Stainer-Scholte medium [64] non-supplemented or containing 1 to 8 mM of nicotinate. Tris(2-carboxyethyl)phosphine (TCEP, SIGMA) was added to 3–10 mM where indicated. TCEP did not affect the activity or the response to nicotinate of wild type BvgS. The bacteria were grown to mid-exponential phase, harvested by centrifugation, resuspended to an OD_{600} of 5 and broken by using a Hybaid Ribolyser apparatus for 30 s at speed 6 in tubes containing 0.1 mm silica spheres. β-galactosidase activities were measured and calculated as described [63]. Each experiment was performed with 3 different clones at different times. The bars represent the standard errors of the mean.

Detection of inactive BvgS variants

The inactive proteins were detected by immunoblotting of *B. pertussis* membrane extracts using anti-BvgS polyclonal antibodies [23] to verify that the substitution(s) generated no major structural defect that might cause BvgS to misfold and to be degraded intracellularly. $BPSM_{\Delta bvgA}$ and $BPMS_{\Delta bvgS}$ were described previously [23,35].

Construction of heterodimers

To construct a *bvgAS* locus deletion strain from BPSM, a *Tohama* I streptomycin- resistant derivative, sequences on either side of the locus (i.e., the 5' end of the *fhaB* gene and the 3' end of the *bvgR* gene) were amplified by PCR using the pairs of oligonucleotides iEco-up and Xma-lo, and Xho-up and HindIII-lo (S2 Table). All the amplicons were first introduced into pCRII-TOPO (Invitrogen) and sequenced. The amplicons were introduced as EcoRI-XhoI and XhoI-HindIII fragments into pUC19 by performing a triple ligation, yielding $pUC19_{newΔbvgAS}$. The EcoRI-HindIII insert was then introduced as in [35] into pSORTP1, a mobilizable plasmid for allelic replacement, resulting in $BPSM_{newΔ}$.

The *bvgAS* locus was then constructed as a mosaic gene for allelic replacement in $BPSM_{newΔ}$. We replaced the EcoRI-SpeI part of $pUC19_{mos}$ [35] using a triple ligation with the EcoRI-XmaI fragment obtained as above and a XmaI-SpeI fragment generated using the primers XmaI-up and SpeI-lo. In the latter amplicon, a natural EcoRI site was eliminated by site-directed mutagenesis with a synonymous mutation. The XbaI-HindIII part of $pUC19_{mos}$ was replaced by 3 fragments: a XbaI-NcoI PCR fragment generated using the primers XbaI-up and NcoI-lo, a NcoI-XhoI PCR fragment generated using the primers NcoI-up and XhoI-lo, and the XhoI-HindIII fragment described above. The latter fragment contains a natural NcoI site, which was eliminated as above. The final plasmid was called $pUC19_{mint}$. The 5.5-kb EcoRI-HindIII insert of $pUC19_{mint}$ was transferred into pSORTP1 for allelic exchange.

A plasmidic construction of the *bvgAS* locus was also created starting from $pUC19_{mint}$ and replacing the EcoRI-SpeI fragment by that generated using the primers pEcoRI-up and SpeI-lo. The natural EcoRI site of this latter fragment was eliminated as above. Finally, the NcoI-HindIII fragment of $pUC19_{mint}$ was replaced by another fragment generated using the primers NcoI-up and pHindIII-lo, yielding $pUC19_{mpla}$. The 4.7-kb EcoRI-HindIII insert was transferred into pBBR1-MCS4 [65], a low-copy, mobilizable and replicative plasmid.

The residues Asp_{1023} and His_{1172} were replaced by Asn and Gln, respectively, using site-directed mutagenesis (QuikchangeXL, Agilent). The first mutation was inserted in $pUC19_{mint}$ and then in pSORTP1 for allelic replacement in $BPSM_{newΔ}$. The second mutation was inserted in $pUC19_{mpla}$ and then in pBBR1-MCS4, yielding $pBBR_{mpla}$ to be introduced in *Bordetella* as an episome.

Successive conjugations were then performed to generate the merodiploids. The first one introduced pSORTP1 containing the *bvgAS* locus with the $D_{1023}N$ substitution into $BPSM_{newΔ}$, yielding an avirulent strain. Then, pFUS-S1 was integrated to generate the *ptx-lacZ* transcriptional fusion [63], and the resulting strain was finally transformed with $pBBR_{mpla}$ containing *bvgAS* with the $H_{1172}Q$ substitution. The mutations of the periplasmic domain were introduced via restriction fragment exchange in $pUC19_{mpla}$ and then in $pBBR_{mpla}$.

Accession numbers

Atomic coordinates and structure factors for the BvgS periplasmic moiety have been deposited in the Protein Data Bank under the accession number 4Q0C.

Supporting Information

S1 Table. Crystallographic parameters.
(DOCX)

S2 Table. Oligonucleotides used for the construction of the BvgS heterodimers.
(DOCX)

S1 Fig. Sequence of the BvgS periplasmic domain and definition of its secondary structure elements. The α helices (H) and β strands (S) are numbered and colored orange and green, respectively. The lobes and hinges between the two lobes of each VFT domain and the Ct loop are also indicated.
(DOCX)

S2 Fig. Dynamics of BvgS. A. Amplitude profiles for the first four normal modes of motion based on a Gaussian network model. Fluctuations of the A and B protomers are indicated by black and red curves, respectively. Note that a single mode describes fluctuation probabilities for the dimer. Black and blue horizontal lines delineate VFT1 and VFT2, respectively, with thick lines indicating their lobes 2. B. Distributions of the VFT internal angles over three molecular dynamics simulations. The opening angles of the VFT domains were calculated based on the geometric centers of the Cα atoms of the two lobes and the hinge region, and data were collected every 100 ps. Blue and black curves refer to the VFT2 and VFT1 angles, respectively. The vertical red lines indicate the initial values of the opening angles of the four VFT domains (lower than 110°: VFT2s; higher than 120°: VFT1s). The inset shows a running average of the angles (1-ns window) over the simulations called WT0, WT1 and WT2. The horizontal stippled red lines show the initial opening angles of the four VFT domains.
(DOCX)

S3 Fig. Substitutions introduced in BvgS. A. Ribbon representation of the engineered VFT1 and VFT2 Cys variants. The mutated residues are circled in green. The open structure of VFT1 is shown, although the selection of the residues for S-S bond formation was performed using a closed model based on the closest homolog (see Methods). B. Position of the substitutions that make BvgS unresponsive to modulation. One protomer is shown in surface representation, while the other is outlined and colored gray. The pink balls represent the modified residues. A zoom delimited by a dashed orange box shows specific residues whose replacement affects the responsiveness of BvgS to nicotinate but not its kinase activity. Residues Ser_{271} to Ser_{275} are in the α helix H8 that forms the $VFT1_{L1}$-$VFT1_{L1}$ interface. Residues Arg_{160}, Phe_{230}, Arg_{234}, Ser_{287} are in the intra-protomer VFT1-VFT2 interface, and Arg_{526} is in the intra-protomer VFT2-Ct interface. Residues Gln_{463} and Asn_{231} are part of the inter-protomer VFT1-VFT2 and VFT1-Ct interfaces, respectively.
(DOCX)

S4 Fig. Detection of specific BvgS variants in membrane extracts of *B. pertussis* by immunoblotting. ΔS and ΔA represent strains with deletions of *bvgS* and *bvgA*, respectively. In the right panel in A, the $BvgS_{E113C+E177C}$ band was most likely too faint and fuzzy for detection under non-reducing conditions, but the left panel confirms that the protein is produced and membrane-localized as expected. The amounts of BvgS are generally lower in avirulent strains because the *bvgAS* operon is positively auto-regulated. The asterisk in the right panel denotes that the oxidized $BvgS_{T355C+D442C}$ variant migrates slightly faster than the wild type control. Note that *in vivo* S-S bond formation was confirmed by the observation that the recombinant strain producing the $BvgS_{T355C+D442C}$ variant does not respond to nicotinate modulation, unless the S-S bond is reduced (see S5 Fig). The other non-functional BvgS variants are presented

in B, showing that they are all produced and localized in the membrane.
(DOCX)

S5 Fig. β-galactosidase activities of recombinant *B. pertussis* harboring BvgS variants. The histograms show the β-gal activity levels from the Bvg-regulated *ptx-lacZ* fusion in the respective strains grown in different conditions. Nic indicates the addition of nicotinate to the growth medium at the given concentrations (in mM). TCEP was added to 10 mM to the growth medium where indicated. WT corresponds to the TohamaI strain with the $K_{705}E$ substitution in BvgS. The bars represent the standard errors of the mean that were calculated from three different experiments.
(DOCX)

S6 Fig. BvgS represents a distinct paradigm of VFT-containing signal-transduction proteins. Cartoon representations compare the structures of an AMPA receptor in A (pdb code: 3KG2), an NMDA receptor in B (pdb code: 4PE5) and of the periplasmic moiety of BvgS in C. The three proteins are shown at the same scale, with each protomer represented in one color. The AMPA and NMDA receptors are tetrameric, with two VFT domains per protomer. The transmembrane segments forming the ion channels are at the bottom of the structure. The extracytoplasmic face of the membrane is represented as a dashed line. For AMPA, the linkers between the NTD (N-terminal domain) and the ABD (agonist-binding domain) and between the ABD and the trans-membrane domain can be seen in the pink and yellow monomers, respectively.
(DOCX)

S1 Protocol. *In silico* analyses of BvgS-p dynamics and associated references.
(DOCX)

Acknowledgments

We thank A. Wöhlkonig and B. Clantin for advice on crystallogenesis and data collection, E. Haddadian for help and advice on the structural refinement, T. Haliloglu and C. Etchebest for advice on normal mode analyses, and A. Baulard for critical reading of the manuscript.

Author Contributions

Conceived and designed the experiments: ED JH RA FJD. Performed the experiments: ED JH MFL RW AV AL SC. Analyzed the data: ED JH MFL RW AV AL SC RA FJD. Contributed reagents/materials/analysis tools: AV AL SC. Wrote the paper: ED JH MFL RW AV AL VV CL RA FJD.

References

1. Stock AM, Robinson VL, Goudreau PN (2000) Two-component signal transduction. Annu Rev Biochem 69: 183–215. PMID: 10966457
2. Bekker M, Teixeira de Mattos MJ, Hellingwerf KJ (2006) The role of two-component regulation systems in the physiology of the bacterial cell. Sci Prog 89: 213–242. PMID: 17338439
3. Wuichet K, Cantwell BJ, Zhulin IB (2010) Evolution and phyletic distribution of two-component signal transduction systems. Curr Opin Microbiol 13: 219–225. doi: 10.1016/j.mib.2009.12.011 PMID: 20133179
4. Szurmant H (2012) Essential two-component systems of Gram-positive bacteria. In: Two-component systems in bacteria. Gross R & Beier D, editors. Norfolk, UK: Caister Academic Press. pp. 127–147.
5. Smith CS, Vicente JJ, Ryan KR (2012) Cell cycle and developmental regulation by two-component signalling proteins in *Caulobacter crescentus*. In: Two-component systems in bacteria. Gross R & Beier D, editors. Norfolk, UK: Caister Acad Press. pp. 269–291.

6. Keilberg D, Huntley S, Sogaard-Andersen L (2012) Two-component systems involoved in motility and development in *Myxococcus xanthus*. In: Two-component systems in bacteria. Gross R & Beier D, editors. Norfolk, UK: Caister Acad. Press. pp. 293–314.
7. Cotter PA, DiRita VJ (2000) Bacterial virulence gene regulation: an evolutionary perspective. Annu Rev Microbiol 54: 519–565. PMID: 11018137
8. Clarke DJ (2012) The Rcs phospohorelay: biofilm formation and virulence in the Enterobacteriaceae. In: Two-component systems in bacteria. Gross R & Beier D, editors. Norfolk, UK: Caister Acad Press. pp. 333–353.
9. Ulrich LE, Zhulin IB (2007) MiST: a microbial signal transduction database. Nucleic Acids Res 35: D386–390. PMID: 17135192
10. Barakat M, Ortet P, Whitworth DE (2011) P2CS: a database of prokaryotic two-component systems. Nucleic Acids Res 39: D771–776. doi: 10.1093/nar/gkq1023 PMID: 21051349
11. Casino P, Rubio V, Marina A (2010) The mechanism of signal transduction by two-component systems. Curr Opin Struct Biol 20: 763–771. doi: 10.1016/j.sbi.2010.09.010 PMID: 20951027
12. Melvin JA, Scheller EV, Miller JF, Cotter PA (2014) *Bordetella pertussis* pathogenesis: current and future challenges. Nat Rev Microbiol 12: 274–288. doi: 10.1038/nrmicro3235 PMID: 24608338
13. Cotter PA, Jones AM (2003) Phosphorelay control of virulence gene expression in *Bordetella*. Trends Microbiol 11: 367–373. PMID: 12915094
14. Cummings CA, Bootsma HJ, Relman DA, Miller JF (2006) Species- and strain-specific control of a complex, flexible regulon by *Bordetella* BvgAS. J Bacteriol 188: 1775–1785. PMID: 16484188
15. Mattoo S, Foreman-Wykert AK, Cotter PA, Miller JF (2001) Mechanisms of *Bordetella* pathogenesis. Front Biosci 6: E168–186. PMID: 11689354
16. Melton AR, Weiss AA (1993) Characterization of environmental regulators of *Bordetella pertussis*. Infect Immun 61: 807–815. PMID: 8432601
17. Boulanger A, Chen Q, Hinton DM, Stibitz S (2013) In vivo phosphorylation dynamics of the *Bordetella pertussis* virulence-controlling response regulator BvgA. Mol Microbiol 88: 156–172. doi: 10.1111/mmi.12177 PMID: 23489959
18. Knapp S, Mekalanos JJ (1988) Two trans-acting regulatory genes (*vir* and *mod*) control antigenic modulation in *Bordetella pertussis*. J Bacteriol 170: 5059–5066. PMID: 2903140
19. Stenson TH, Peppler MS (1995) Identification of two *bvg*-repressed surface proteins of *Bordetella pertussis*. Infect Immun 63: 3780–3789. PMID: 7558280
20. Cotter PA, Miller JF (1997) A mutation in the *Bordetella bronchiseptica bvgS* gene results in reduced virulence and increased resistance to starvation, and identifies a new class of Bvg-regulated antigens. Mol Microbiol 24: 671–685. PMID: 9194696
21. Stockbauer KE, Fuchslocher B, Miller JF, Cotter PA (2001) Identification and characterization of BipA, a *Bordetella* Bvg-intermediate phase protein. Mol Microbiol 39: 65–78. PMID: 11123689
22. Perraud AL, Rippe K, Bantscheff M, Glocker M, Lucassen M, et al. (2000) Dimerization of signalling modules of the EvgAS and BvgAS phosphorelay systems. Biochim Biophys Acta 1478: 341–354. PMID: 10825546
23. Dupre E, Wohlkonig A, Herrou J, Locht C, Jacob-Dubuisson F, et al. (2013) Characterization of the PAS domain in the sensor-kinase BvgS: mechanical role in signal transmission. BMC Microbiol 13: 172. doi: 10.1186/1471-2180-13-172 PMID: 23883404
24. Jacob-Dubuisson F, Wintjens R, Herrou J, Dupré E, Antoine R (2012) BvgS of pathogenic *Bordetellae*: a paradigm for sensor kinase with Venus Flytrap perception domains. In: Two-component system in bacteria. Gross R & Beier D, editors. Norfolk, UK: Caister Academic Press. pp. 57–83.
25. Quiocho FA, Ledvina PS (1996) Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: variation of common themes. Mol Microbiol 20: 17–25. PMID: 8861200
26. Trakhanov S, Vyas NK, Luecke H, Kristensen DM, Ma J, et al. (2005) Ligand-free and-bound structures of the binding protein (LivJ) of the *Escherichia coli* ABC leucine/isoleucine/valine transport system: trajectory and dynamics of the interdomain rotation and ligand specificity. Biochemistry 44: 6597–6608. PMID: 15850393
27. Herrou J, Bompard C, Wintjens R, Dupre E, Willery E, et al. (2010) Periplasmic domain of the sensor-kinase BvgS reveals a new paradigm for the Venus flytrap mechanism. Proc Natl Acad Sci USA 107: 17351–17355. doi: 10.1073/pnas.1006267107 PMID: 20855615
28. Masuda N, Church GM (2002) *Escherichia coli* gene expression responsive to levels of the response regulator EvgA. J Bacteriol 184: 6225–6234. PMID: 12399493

29. Sivaneson M, Mikkelsen H, Ventre I, Bordi C, Filloux A (2011) Two-component regulatory systems in *Pseudomonas aeruginosa*: an intricate network mediating fimbrial and efflux pump gene expression. Mol Microbiol 79: 1353–1366. doi: 10.1111/j.1365-2958.2010.07527.x PMID: 21205015

30. Martinez-Wilson HF, Tamayo R, Tischler AD, Lazinski DW, Camilli A (2008) The *Vibrio cholerae* hybrid sensor kinase VieS contributes to motility and biofilm regulation by altering the cyclic diguanylate level. J Bacteriol 190: 6439–6447. doi: 10.1128/JB.00541-08 PMID: 18676667

31. Walker KA, Miller VL (2004) Regulation of the Ysa of the type III secretion system of *Yersinia enterocolitica* by YsaE/SycB and YsrS/YsrR. J Bacteriol 186: 4056–4066. PMID: 15205407

32. Caimano MJ, Kenedy MR, Kairu T, Desrosiers DC, Harman M, et al. (2011) The hybrid histidine kinase Hk1 is part of a two-component system that is essential for survival of *Borrelia burgdorferi* in feeding *Ixodes scapularis* ticks. Infect Immun 79: 3117–3130. doi: 10.1128/IAI.05136-11 PMID: 21606185

33. Blanke ML, VanDongen AM (2008) Constitutive activation of the N-methyl-D-aspartate receptor via cleft-spanning disulfide bonds. J Biol Chem 283: 21519–21529. doi: 10.1074/jbc.M709190200 PMID: 18450751

34. Zhu S, Stroebel D, Yao CA, Taly A, Paoletti P (2013) Allosteric signaling and dynamics of the clam-shell-like NMDA receptor GluN1 N-terminal domain. Nat Struct Mol Biol 20: 477–485. doi: 10.1038/nsmb.2522 PMID: 23454977

35. Herrou J, Debrie AS, Willery E, Renaud-Mongenie G, Locht C, et al. (2009) Molecular evolution of the two-component system BvgAS involved in virulence regulation in *Bordetella*. PLoS One 4: e6996. doi: 10.1371/journal.pone.0006996 PMID: 19750014

36. Beier D, Deppisch H, Gross R (1996) Conserved sequence motifs in the unorthodox BvgS two-component sensor protein of *Bordetella pertussis*. Mol Gen Genet 252: 169–176. PMID: 8804390

37. Uhl MA, Miller JF (1996) Integration of multiple domains in a two-component sensor protein: the *Bordetella pertussis* BvgAS phosphorelay. Embo J 15: 1028–1036. PMID: 8605872

38. Arico B, Miller JF, Roy C, Stibitz S, Monack D, et al. (1989) Sequences required for expression of *Bordetella pertussis* virulence factors share homology with prokaryotic signal transduction proteins. Proc Natl Acad Sci USA 86: 6671–6675. PMID: 2549542

39. Gotoh Y, Eguchi Y, Watanabe T, Okamoto S, Doi A, et al. (2010) Two-component signal transduction as potential drug targets in pathogenic bacteria. Curr Opin Microbiol 13: 232–239. doi: 10.1016/j.mib.2010.01.008 PMID: 20138000

40. Smock RG, Gierasch LM (2009) Sending signals dynamically. Science 324: 198–203. doi: 10.1126/science.1169377 PMID: 19359576

41. Gao R, Lynn DG (2007) Integration of rotation and piston motions in coiled-coil signal transduction. J Bacteriol 189: 6048–6056. PMID: 17573470

42. Lowe EC, Basle A, Czjzek M, Firbank SJ, Bolam DN (2012) A scissor blade-like closing mechanism implicated in transmembrane signaling in a *Bacteroides* hybrid two-component system. Proc Natl Acad Sci U S A 109: 7298–7303. doi: 10.1073/pnas.1200479109 PMID: 22532667

43. Airola MV, Sukomon N, Samanta D, Borbat PP, Freed JH, et al. (2013) HAMP Domain Conformers That Propagate Opposite Signals in Bacterial Chemoreceptors. PLoS Biol 11: e1001479. doi: 10.1371/journal.pbio.1001479 PMID: 23424282

44. Brooks AJ, Dai W, O'Mara ML, Abankwa D, Chhabra Y, et al. (2014) Mechanism of activation of protein kinase JAK2 by the growth hormone receptor. Science 344: 1249783. doi: 10.1126/science.1249783 PMID: 24833397

45. Cheung J, Le-Khac M, Hendrickson WA (2009) Crystal structure of a histidine kinase sensor domain with similarity to periplasmic binding proteins. Proteins 77: 235–241. doi: 10.1002/prot.22485 PMID: 19544572

46. Lacey BW (1960) Antigenic modulation of *Bordetella pertussis*. J Hyg 31: 423–434.

47. He XL, Dukkipati A, Wang X, Garcia KC (2005) A new paradigm for hormone recognition and allosteric receptor activation revealed from structural studies of NPR-C. Peptides 26: 1035–1043. PMID: 15911071

48. Mony L, Zhu S, Carvalho S, Paoletti P (2011) Molecular basis of positive allosteric modulation of GluN2B NMDA receptors by polyamines. EMBO J 30: 3134–3146. doi: 10.1038/emboj.2011.203 PMID: 21685875

49. Karakas E, Furukawa H (2014) Crystal structure of a heterotetrameric NMDA receptor ion channel. Science 344: 992–997. doi: 10.1126/science.1251915 PMID: 24876489

50. Sobolevsky AI, Rosconi MP, Gouaux E (2009) X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor. Nature 462: 745–756. doi: 10.1038/nature08624 PMID: 19946266

51. Johnson MD, Bell J, Clarke K, Chandler R, Pathak P, et al. (2014) Characterization of mutations in the PAS domain of the EvgS sensor kinase selected by laboratory evolution for acid resistance in *Escherichia coli*. Mol Microbiol 93: 911–927. doi: 10.1111/mmi.12704 PMID: 24995530

52. Huth JR, Bewley CA, Jackson BM, Hinnebusch AG, Clore GM, et al. (1997) Design of an expression system for detecting folded protein domains and mapping macromolecular interactions by NMR. Protein Sci 6: 2359–2364. PMID: 9385638

53. Kabsch W (2010) XDS. Acta Crystallogr D Biol Crystallogr 66: 125–132. doi: 10.1107/S0907444909047337 PMID: 20124692

54. Vagin A, Teplyakov A (2010) Molecular replacement with MOLREP. Acta Crystallogr D Biol Crystallogr 66: 22–25. doi: 10.1107/S0907444909042589 PMID: 20057045

55. Murshudov GN, Skubak P, Lebedev AA, Pannu NS, Steiner RA, et al. (2011) REFMAC5 for the refinement of macromolecular crystal structures. Acta Crystallogr D Biol Crystallogr 67: 355–367. doi: 10.1107/S0907444911001314 PMID: 21460454

56. Vagin AA, Isupov MN (2001) Spherically averaged phased translation function and its application to the search for molecules and fragments in electron-density maps. Acta Crystallogr D Biol Crystallogr 57: 1451–1456. PMID: 11567159

57. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr D Biol Crystallogr 66: 486–501. doi: 10.1107/S0907444910007493 PMID: 20383002

58. Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, et al. (2012) Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D Biol Crystallogr 68: 352–367. doi: 10.1107/S0907444912001308 PMID: 22505256

59. Haddadian EJ, Gong H, Jha AK, Yang X, Debartolo J, et al. (2011) Automated real-space refinement of protein structures using a realistic backbone move set. Biophys J 101: 899–909. doi: 10.1016/j.bpj.2011.06.063 PMID: 21843481

60. Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372: 774–797. PMID: 17681537

61. Eswar N, Marti-Renom MA, Webb B, Madhusudhan MS, Eramwan D, et al. (2006) Comparative protein structure modelling with Modeler. In: John Wiley and sons I, editor. Current protocols in Bioinformatics. pp. 5.6.1–5.6.30.

62. Dombkowski AA (2003) Disulfide by Design: a computational method for the rational design of disulfide bonds in proteins. Bioinformatics 19: 1852–1853. PMID: 14512360

63. Antoine R, Alonso S, Raze D, Coutte L, Lesjean S, et al. (2000) New virulence-activated and virulence-repressed genes identified by systematic gene inactivation and generation of transcriptional fusions in *Bordetella pertussis*. J Bacteriol 182: 5902–5905. PMID: 11004193

64. Imaizumi A, Suzuki Y, Ono S, Sato Y, Sato H (1983) Heptakis (2,6-O-dimethyl)beta-cyclodextrin: a novel growth stimulant for *Bordetella pertussis* phase I. J Clin Microbiol 17: 781–786. PMID: 6306047

65. Kovach ME, Elzer PH, Hill DS, Robertson GT, Farris MA, et al. (1995) Four new derivatives of the broad-host-range cloning vector pBBR1MCS, carrying different antibiotic-resistance cassettes. Gene 166: 175–176. PMID: 8529885

5. OUTLOOK

As stated before, my research activities are aimed at the development and application of advanced molecular modeling and simulation techniques to study the process of molecular recognition, focusing primarily on three types of fundamental interactions, namely ***(i)*** **protein-protein**, ***(ii)*** **protein-lipid** and ***(iii)*** **protein-carbohydrate interactions**. The *"Computational Molecular Systems Biology"* team that I'm leading works at the interface of biology and physics and employs a pragmatic and collaborative approach, where multi-scale modeling approaches and structural biology techniques are applied to relevant biological questions, typically involving several members of the team. We regularly embark upon new projects, often in collaboration with other research teams, both internal and external to our host institute. Some of these projects later get funded, while some do not.

The community-wide **CAPRI** and CASP experiments define the state-of-the-art in *computational protein docking* and protein structure prediction. As it is a dynamic experiment, acting on the demands of both the docking community as well as the experimental community providing the targets, future developments are difficult to predict. This being said, we are actively trying to increase the weight of targets involving peptides, nucleic acids and carbohydrates, proof of which are the prediction rounds of autumn 2017, featuring no less than 8 protein-carbohydrate complexes.

Carbohydrate-processing enzymes are also the target of the ANR-funded **HICARE** project[(a)], the aim of which is to develop multivalent inhibitor molecules that simultaneously interact with several domains of these enzymes. The project combines organic synthesis, protein expression, molecular and mathematical modeling, and biophysical interaction characterization.

While glycosylation adds significant diversity to the proteome, there is no molecular template that encodes the glycome; suitable glycosylation is rather achieved through temporal and compartmental separation. In collaboration with research teams in Lille[(b)] and Oulu, Finland[(c)], we investigate the role of *macromolecular complex formation*, where enzymatic activity requires the formation of homomeric complexes and heteromeric complexes involving several successively acting **glycosyltransferases**.

Glycosylation is initiated by *O*-GlcNAc transferase (OGT), which transfers a N-acetylglucosamine to a serine or threonine residue. In collaboration with the university of Rouen[(d)], we obtained funding for a two-year MSc bioinformatics apprentice to develop a *motif search engine* for *O*-glycosylation, using both sequence and structure data. OGT also interacts with β**-catenin**, constituting a key interaction in the Wnt signaling pathway, an evolutionary conserved pathway that regulates crucial aspects of cell proliferation and migration. Using a combination of computational techniques and experimental validation[(e)], our goal is to understand this interaction at the molecular level and identify key binding hot spots.

In the ANR-funded **MECA VENUS** project[(f)] we investigate the signal transduction mechanism of the *Bordetella pertussis* BvgS sensor-kinase, which is responsible for the expression of its virulence genes. The *BvgS two-component signaling system*, considered prototypical for bacterial Venus Fly-Trap-containing sensor-kinases, translates an extracellularly perceived signal across the lipid bilayer membrane into a cytoplasmic autophosphorylation and subsequent regulatory response.

Bacteria adhere themselves to host cells using lectin domains, located at the tip of their fimbriae. These domains target host (glyco)-receptors by mimicking inter-cellular binding mechanisms. The EU-funded **FimH-Mech** project[(g)] investigates the *shear-force-dependent binding mechanism* of FimH through modeling of the interaction with membrane-bound target receptors.

Lectin domains, though having lost their carbohydrate-binding function, can also be found in some of the complexes responsible for the **self-incompatibility** response in flowering plants. We investigate the binding mode of these membrane-bound receptor-ligand complexes, investigating in detail their binding specifity and mode of signal transduction[(h)].

There is a clear relation between some of the projects and from their short description it is also obvious that they address one or several of the fundamental *protein-protein*, *protein-lipid* and *protein-carbohydrate* interactions mentioned at the beginning of this section.

It should however be mentioned that we are always open to new and interesting research projects and the projects mentioned above constitute an incomplete list of our current research activities. All projects typically involve several members of the team, and also target several of our on page 6 mentioned research axes: *(i) computational modeling and dynamics of biomolecular systems*, *(ii) structural biology of protein-carbohydrate interactions*, *(iii) protein interaction and regulatory networks* and *(iv) statistical physics of biomolecular interactions*.

[(a)] Coordinated by S. Gouin, Nantes [(b)] A. Harduin [(c)] T. Glumoff & S. Kellokumpu [(d)] M. Bardor [(e)] Collaboration with T. Lefebvre [(f)] Coordinated by F. Jacob-Dubuisson, Lille [(g)] E.-M. Krammer [(h)] Collaboration with V. Castric, Lille

6. Acknowledgements

This work would not have been possible without the unconditional support of family, friends and colleagues. First and foremost thanks to Emmanuelle, Téa and Luca: I don't need to tell you how important you are to me! A "work"-wise special mention goes to my former colleagues Anton, Peter, Danilo, Bert, Pieter, Rudy, David, Bernard, Natasha, Frans, Steve, Janez, Gilles, Benoit, Theo, Gerrit, Serena, Satyan, Weidong, Niko, Olli, Outi, Tuomo, Kalervo, Rik, Bhaumik, Petri, Inari, Antti, Anna-Kaisa, Lloyd, Maija, Jyrki, Stefano, Piero, Marc, Gilles, Gabriela, Ralph, Kurt, Hans, Stein, Bart, Christine, Berlinda, Jacques, Olivier, Nicolas, Benoit, Raul, Fernanda, Didier, Rekins, Raphaël, Morgane, Didier, Karoline, Gipsi, Mathieu, Jean-Valéry, Sylvain, Ariane, Jean, Fabian, Christian, Michel, Patricia, Ana, Cedric, Caroline, Vincent, Rabia, Erik, Fabrice, Emilie, Pierre, Mathieu, Benjamin, Boris, Moussa, Jean-Paul, Martine, and whomever I surely have forgotten to mention here because the list is so long but who all made academic life so much more pleasant. A special thanks goes to past supervisors Herman, David, André, Maryvonne, Shoshana and Jean-Marie, who have shown me how to lead a team, and to everybody in the CAPRI network whom I have come to know so well over the years: it's truly a pleasure working with all of you. A word of thanks is extended also to the various institutions that have funded my research over the years: CECAM, NWO, University of Groningen, Biocenter Oulu, University of Oulu, Région Wallonne, Max Planck Gesellschaft, Région Nord-Pas-de-Calais, Université de Lille, ANR, European Commission and of course CNRS. Also a warm thank you to my promotor Ralf and the jury members Tony, Marc, Savvas, Alessandra and Isabelle, who all selflessly accepted this responsibility. Finally, a special mention to current and past colleagues since my entry in the CNRS, in particular my current group members, Ralf, Guillaume, Julie, Goedele, Stefania, Jérôme and Eva-Maria. You certainly make my job so much more enjoyable. I do hope that the reverse is also true and that I participate in making academic life as interesting for you as it has always been for me.

7. Jury

- Ralf Blossey – University of Lille – promotor
- Tony Lefebvre – University of Lille – referee
- Marc Baaden – University of Paris-Diderot – referee
- Savvas Savvides – University of Ghent – referee
- Alessandra Carbone – University of Paris-Pierre et Marie Curie – jury member
- Isabelle Landrieu – University of Lille – jury member

8. Cover images

I have been lucky enough to occasionally have my publications feature on the cover of the journal in which they were published. That makes for a nice closing of this document.

- Lensink *et al*, **Proteins** 2017;85:359
- Gabel *et al*, **Biophys J** 2014;107:185
- Lonez *et al*, **Biochim Biophys Acta** 2009;1790:425
- Lensink *et al*, **Proteins** 2007;69:704
- Lensink *et al*, **J Mol Biol** 2002;323:99