



UNIVERSITAT DE  
BARCELONA

## Structural characterization of the T7 bacteriophage portal protein

Montserrat Fàbrega Ferrer

**ADVERTIMENT.** La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX ([www.tdx.cat](http://www.tdx.cat)) i a través del Dipòsit Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

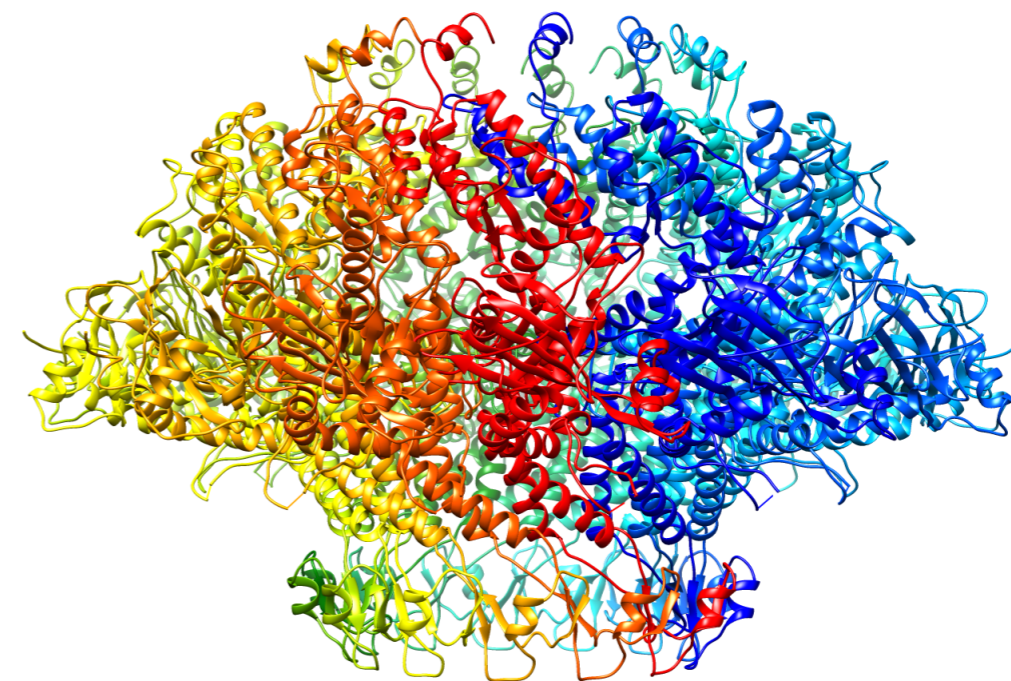
**ADVERTENCIA.** La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR ([www.tdx.cat](http://www.tdx.cat)) y a través del Repositorio Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

**WARNING.** On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX ([www.tdx.cat](http://www.tdx.cat)) service and by the UB Digital Repository ([diposit.ub.edu](http://diposit.ub.edu)) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.



UNIVERSITAT DE  
BARCELONA

## Structural characterization of the T7 bacteriophage portal protein



Montserrat Fàbrega Ferrer  
PhD thesis

Montserrat Fàbrega Ferrer

Structural characterization of the T7 bacteriophage portal protein







UNIVERSITAT DE  
BARCELONA

Facultat de Farmàcia i Ciències de l'Alimentació

Departament de Bioquímica i Fisiologia

---

# Structural characterization of the T7 bacteriophage portal protein

---

PhD thesis

**Montserrat Fàbrega Ferrer**

Co-directed by Prof. Miquel Coll Capella  
and Dr. Cristina Machón Sobrado



INSTITUT  
DE RECERCA  
BIOMÈDICA



Institut de Biologia Molecular de Barcelona  
Molecular Biology Institute of Barcelona  CSIC

Barcelona, 2017





UNIVERSITAT DE  
BARCELONA

---

Thesis submitted by Montserrat Fàbrega Ferrer, enrolled in the *Biotechnology* program at the University of Barcelona, for the degree of Doctor of Philosophy.

This work was carried out in the *Structural Biology of Protein & Nucleic Acid Complexes and Molecular Machines Group* at the Institute for Research in Biomedicine (IRB-Barcelona) and the Molecular Biology Institute of Barcelona (IBMB-CSIC), under the supervision of Prof. Miquel Coll Capella and Dr. Cristina Machón Sobrado. and with Josefa Badia Palacín as tutor.

---

**Montserrat Fàbrega Ferrer**  
IRB Barcelona  
IBMB-CSIC

**Josefa Badia Palacín**  
University of Barcelona

**Miquel Coll Capella**  
IRB Barcelona  
IBMB-CSIC

**Cristina Machón Sobrado**  
IRB Barcelona  
IBMB-CSIC

Barcelona, 2017



*Als meus pares*





# Acknowledgements

First of all, I would like to thank all of the present and past members of the “Structural biology of protein & nucleic acid complexes and molecular machines” group. I am extremely grateful to Prof. Miquel Coll and Dr. Cristina Machón, the co-directors of the thesis. I would like to thank Prof. Miquel Coll for accepting me in his research group and giving me the opportunity to work in this challenging project. Thanks for all the great advice, support and trust. I also would like to thank Dr. Cristina Machón, for being there day after day, and for all the scientific discussions and the personal support. I would like to acknowledge Albert, Esther and Rosa, who have been in the group during all my PhD student period, and who made all this road much more easy and comfortable. The group would not be the same without you. Thanks Albert for being there always ready to help in everything. Thanks Esther for all your help and support, you are the best lab technician that anyone could imagine. Thanks Rosa for all your work with the T7 portal protein and for your kind advice and help during crystallization, I learned a lot from you. I would also like to thank the past lab members: Marta, Simone, Salvatore, Roeland, Radek, Zuzanna, Juliana, Robert, Andrés, Clara, Sara, Rhys... And, of course, the present ones: Dani, Mireia, Jorge, Sara, Olga and Lionel. Thank you all for all the ideas, support, dinners, barbecues and many more activities that we have shared together. I also would like to thank all the people that worked with the T7 bacteriophage portal protein before I arrived to the lab, specially Prof. Cristina Vega and Dr. Francisco J. Fernández.

Secondly, I am deeply indebted to our collaborators from CNB-CSIC in Madrid. I would like to thank Prof. Jose L. Carrascosa for accepting me at his lab during many weeks. Thanks also to all the people from the S0 lab. Among them, I am especially indebted to Dr. Ana Cuervo, who introduced me in the exciting cryo-EM technique,

and who made me feel like at home. Thank you both for all the scientific discussions and the personal support.

I would also like to express my gratitude to many people from IRB-Barcelona and IBMB-CSIC:

- To Isabel, Laia and Roman from the purification service.
- To Joan and Xandra from the PAC. Especial thanks to Joan, who has helped a lot during the structure solution process.
- To Esther, Leonor, Clara, Patricia and Leyre, for their help.
- To Prof. Ignacio Fita, for his advices on structure solution.
- To the members of the other crystallography groups, for all the seminars, discussions, synchrotron trips and activities we have shared together.
- To the members of my TAC committee, for their help: Dr. Núria Verdaguer, Dr. Maria Macias and Dr. Jordi Bernués.

I would also like to acknowledge Prof. Josep Vendrell and Dr. Irantzu Pallarès from the Autonomous University of Barcelona, who gave me the chance to start in the research field years ago.

This study was supported from the *Ministerio de Economía y Competitividad* of Spain, from which I also was the recipient of a FPI fellowship.

Finally, I would like to dedicate this work to my family and friends, for their support and love. Thanks to my parents and to my grandmother for their faith in me. Thanks to Sergi for sharing all these years with me. Thanks Alba, Maria, Laia and Cristina, the weeks were much easier knowing that I would meet you each Friday night. Thanks also to Jaume, Muntsa and Pol for your unconditional support. Thanks also to Marina, Núria, Mireia and Montserrat for all the moments shared together. Thanks to all the people from “La Llanterna” for the magic moments we share together (especially to Joan for the Friday trips!). I learned a lot from all of you.

# Abstract

The *Escherichia coli* infecting T7 bacteriophage shares a common dsDNA packaging mechanism with other bacteriophages of the *Caudovirales* order, Herpesviruses and Adenoviruses. The packaging machinery comprises the portal protein and the terminase complex. The portal protein is a channel located at a unique portal vertex that provides a conduit for DNA translocation, while the terminase complex recognizes a long concatemer of DNA, performs the nuclease catalytic activity and hydrolyses ATP. Available structural information about portal proteins describes them as oligomeric rings with an axial channel. High quality samples suitable for structural characterization of the portal protein of T7 bacteriophage were obtained and characterized. Both X-ray crystallography and cryo-electron microscopy (cryo-EM) data were collected, and an initial model built on the 5.8Å cryo-EM map was used to phase the crystallographic data, which allowed the building of a model of a tridecameric particle at 2.8Å resolution. The T7 portal particle is 170Å tall and 110Å wide toroidal protein with a central channel that ranges from 23Å to 95Å in diameter. Four domains have been identified in the structure: the wing, the stem, the clip and the crown. The  $\alpha 10$ -tunnel loop valve is proposed to play an important functional role. During packaging, it may adapt while DNA is translocated and rotated, and once the genome has been packed the side chain of tunnel loop residue Arg368 may be able to seal the channel and stabilize the DNA inside the capsid before tail assembly. Interestingly, these mechanisms would not only imply the flexibility of a loop region, but also the kink of the longer helix of the portal structure,  $\alpha 10$ .

## Keywords

dsDNA viral packaging; portal protein; T7 bacteriophage; structural biology; X-ray crystallography; cryo-electron microscopy;  $\alpha 10$ -tunnel loop valve



# Table of contents

Acknowledgements.....	vii
Abstract.....	ix
Keywords.....	ix
Table of contents .....	xi
List of figures.....	xv
List of tables .....	xvii
List of abbreviations and symbols.....	xix
Amino acids abbreviations.....	xxiii
Preface .....	xxv
<b>Chapter I: Introduction.....</b>	<b>1</b>
<b>I.1 The T7 bacteriophage .....</b>	<b>3</b>
1.1.1 Viruses, bacteriophages and the T7 bacteriophage.....	3
1.1.2 Taxonomy and phylogeny .....	4
1.1.3 Genome organization.....	6
1.1.4 Morphology and structure.....	7
1.1.5 Viral infection cycle .....	10
1.1.6 Biotechnological applications.....	11
<b>I.2 Viral genome packaging.....</b>	<b>12</b>
1.2.1 Viral assembly strategies.....	12
1.2.2 The viral assembly pathway of large dsDNA viruses.....	12
1.2.2.1 The packaging proteins.....	14
1.2.2.2 Strategies of DNA processing .....	14
1.2.3 The terminase proteins .....	15
1.2.3.1 The small terminase subunit .....	16
1.2.3.2 The large terminase subunit.....	18
1.2.4 The portal protein .....	19
1.2.4.1 Structure.....	19
1.2.4.2 Structure-function relationship .....	26
1.2.5 Properties of the DNA packaging motor .....	29
1.2.6 Models for portal protein dsDNA translocation.....	30
1.2.7 Biotechnological and biomedical interest.....	33
<b>I.3 The T7 packaging machinery.....</b>	<b>34</b>
1.3.1 DNA processing.....	34
1.3.2 The terminase proteins .....	35
1.3.2.1 The small terminase subunit .....	35

1.3.2.2	The large terminase subunit.....	35
1.3.3	The portal protein .....	36
<b>Chapter 2:</b>	<b>Objectives .....</b>	<b>3</b>
<b>Chapter 3:</b>	<b>Materials and methods.....</b>	<b>41</b>
3.1	Sample preparation and analysis.....	43
3.1.1	Sample preparation.....	43
3.1.1.1	Cloning.....	43
3.1.1.2	Plasmid purification and sequencing.....	43
3.1.1.3	Bacterial strains.....	44
3.1.1.4	Competent cells preparation.....	44
3.1.1.5	Bacterial transformation .....	45
3.1.1.6	Culture media.....	45
3.1.1.7	Protein expression .....	46
3.1.1.8	Protein electrophoresis .....	47
3.1.1.9	Protein purification.....	48
3.1.1.10	Protein concentration and quantification.....	49
3.1.2	Sample analysis.....	49
3.1.2.1	Mass spectrometry analysis.....	49
3.1.2.2	Dynamic light scattering.....	50
3.2	Crystallization and X-ray diffraction analysis.....	51
3.2.1	Crystallization and X-ray data collection .....	54
3.2.1.1	Protein crystallization screening .....	54
3.2.1.2	Protein crystallization optimization.....	55
3.2.1.3	Crystal mounting and freezing.....	55
3.2.1.4	Derivative-crystals for experimental phasing.....	56
3.2.1.5	X-ray data collection .....	56
3.2.2	X-ray diffraction analysis .....	56
3.2.2.1	Data processing and analysis.....	56
3.2.2.2	Matthews coefficient calculation .....	57
3.2.2.3	Calculation of the self-rotation function .....	57
3.3	Cryo-EM studies .....	58
3.3.1	Preparation of the grids and data collection.....	61
3.3.1.1	Preparation of holey grids with thin carbon backing.....	61
3.3.1.2	Negative staining.....	61
3.3.1.3	Vitrification optimization .....	62
3.3.1.4	Data collection .....	63
3.3.2	Cryo-EM data processing.....	63
3.3.2.1	Movie alignment.....	63
3.3.2.2	Contrast transfer function correction.....	64
3.3.2.3	Particle picking .....	64
3.3.2.4	Initial volume.....	65
3.3.2.5	Classification of the particles .....	65
3.3.2.6	Volume reconstruction .....	65
3.3.2.7	Calculation of local resolution.....	65
3.4	Structure determination .....	66
3.4.1	Structure solution and refinement.....	66

3.4.1.1	Preliminar model building and refinement.....	66
3.4.1.2	Previous crystallographic data .....	66
3.4.1.3	Crystallographic analysis .....	66
3.4.1.4	Molecular replacement.....	66
3.4.1.5	Density modification and phase extension .....	66
3.4.1.6	Crystallographic model building and structure refinement	67
3.4.1.7	Model refinement into the cryo-EM volume.....	67
3.4.2	Model validation and analysis .....	68
3.4.2.1	Model validation.....	68
3.4.2.2	Structure visualization and analysis.....	68
<b>Chapter 4:</b>	<b>Results and discussion .....</b>	<b>73</b>
4.1	Sample preparation and analysis.....	71
4.1.1	Protein expression.....	71
4.1.2	Protein purification .....	72
4.1.3	Sample characterization.....	75
4.2	Crystallization and X-ray diffraction analysis.....	77
4.2.1	Crystallization and X-ray diffraction.....	77
4.2.2	X-ray data processing analysis .....	79
4.2.3	Structure solution trials.....	81
4.2.3.1	Experimental phasing .....	81
4.2.3.2	MR.....	83
4.3	Cryo-EM studies .....	84
4.3.1	Negative staining .....	84
4.3.2	Vitrification .....	84
4.3.3	Cryo-EM data collection .....	86
4.3.4	CTF estimation and particle picking .....	86
4.3.5	Calculation of an initial volume.....	90
4.3.6	Extensive particle classification .....	92
4.3.7	Structure refinement.....	93
4.4	Structure determination .....	98
4.4.1	Model building into the cryo-EM map.....	98
4.4.2	Crystallographic data analysis.....	99
4.4.3	Structure solution .....	105
4.4.4	Refinement into the EM volume.....	110
4.5	Structural analysis .....	114
4.6	Comparison with other portals.....	122
4.7	Functional model .....	126
<b>Chapter 5:</b>	<b>Conclusions.....</b>	<b>133</b>
<b>Bibliography</b>	.....	<b>xxix</b>





# List of figures

Figure 1.1 T7 bacteriophage viral particles.....	3
Figure 1.2 Prohead, internal core and portal of the T7 bacteriophage.....	8
Figure 1.3 T7 bacteriophage tail.....	9
Figure 1.4 Assembly pathway of large dsDNA viruses.....	13
Figure 1.5 Cartoon representation of P22 TerS at 1.75Å resolution.....	17
Figure 1.6 T4 TerL structures.....	18
Figure 1.7 HCMV terminase structures.....	19
Figure 1.8 Cartoon representation of $\phi$ 29 portal at 2.1Å resolution.....	21
Figure 1.9 Cartoon representation of SPP1 portal at 3.4Å resolution. ....	22
Figure 1.10 Cartoon representation of P22 portal protein at 3.25Å/7.5Å resolution.....	24
Figure 1.11 Cartoon representation of T4 portal protein at 3.6Å resolution.....	25
Figure 1.12 HSV-1 portal at 16Å.....	26
Figure 1.13 Cartoon representation of P22 portal protein during packaging at 3.3Å resolution. ....	27
Figure 1.14 The tunnel loops mechanistic model of DNA translocation.	31
Figure 1.15 Geometric basis for DNA rotation at low capsid filling. ....	32
Figure 1.16 T7 concatemers processing and packaging model.....	34
Figure 1.17 T7 bacteriophage TerL EM structures.....	35
Figure 1.18 T7 portal protein cryo-EM structure at 8Å resolution.....	36
Figure 3.1 Protein crystallization phase diagram. ....	51
Figure 3.2 Single-particle cryo-EM workflow. ....	59
Figure 3.3 FEI Vitrobot. ....	62
Figure 3.4 FEI Talos Arctica.....	63
Figure 4.1 SDS-PAGE analysis of portal protein expression.....	71
Figure 4.2 HisTrap chromatogram and SDS-PAGE analysis. ....	72
Figure 4.3 Size-exclusion chromatograms and SDS-PAGE analysis.....	74
Figure 4.4 DLS size distribution.....	76
Figure 4.5 Optimized hexagonal crystal.....	77
Figure 4.6 Diffraction of hexagonal crystals at ALBA synchrotron.....	78
Figure 4.7 P6 <sub>3</sub> 22 self-rotation function sections.....	81
Figure 4.8 Scanning of a tantalum bromide derivative crystal. ....	82
Figure 4.9 Negative staining.....	84
Figure 4.10 Vitrification optimization. ....	85
Figure 4.11 Grid, square and holes.....	86
Figure 4.12 Calculated defocus by ctffind4. ....	87
Figure 4.13 Calculated resolution by ctffind4.....	88
Figure 4.14 Example of movie with drift. ....	89
Figure 4.15 Aligned movie with particles picked. ....	90
Figure 4.16 2D classification of manually picked particles.....	91

Figure 4.17 Ransac initial volumes.....	91
Figure 4.18 Polished initial volume.....	92
Figure 4.19 2D classification.....	93
Figure 4.20 Gold standard FSC refinement curve.....	94
Figure 4.21 Local resolution slices.....	95
Figure 4.22 Cryo-EM 3D model at 5.8Å.....	96
Figure 4.23 Detail of the channel cavities.....	97
Figure 4.24 Initial model building.....	98
Figure 4.25 P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> SRF sections.....	102
Figure 4.26 P4 <sub>2</sub> 2 <sub>1</sub> 2 SRF sections.....	104
Figure 4.27 P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub> MR solution.....	105
Figure 4.28 Model building.....	107
Figure 4.29 X-ray model Ramachandran plot of chain A.....	108
Figure 4.30 Structure solution and dimensions.....	109
Figure 4.31 X-ray model fitted into the cryo-EM map.....	110
Figure 4.32 Cryo-EM model Ramachandran plot of chain A.....	112
Figure 4.33 Superposition of the X-ray and cryo-EM models.....	113
Figure 4.34 Domains of the T7 portal protein monomer.....	114
Figure 4.35 Model sequence and structural summary.....	115
Figure 4.36 Secondary structure elements.....	116
Figure 4.37 Wing β-sandwich.....	117
Figure 4.38 Interaction between monomers.....	119
Figure 4.39 Surface charge distribution.....	120
Figure 4.40 Comparison of bacteriophage portal particles.....	123
Figure 4.41 Comparison monomeric portal proteins.....	125
Figure 4.42 Model of the portal protein with DNA during packaging.....	126
Figure 4.43 Model of interaction with DNA.....	127
Figure 4.44 α10-tunnel loop valve.....	128
Figure 4.45 α10-tunnel loop valve movement.....	129

# List of tables

Table 1.1	Caudovirales families. ....	4
Table 1.2	Taxonomic classification of T7. ....	5
Table 1.3	T7 bacteriophage genes. ....	6
Table 1.4	Bacteriophage terminase proteins. ....	15
Table 1.5	Herpesvirus terminase proteins. ....	16
Table 1.6	Portal proteins. ....	19
Table 3.1	Expression vector. ....	43
Table 3.2	Cell strains. ....	44
Table 3.3	Crystal screenings used. ....	54
Table 4.1	MW calibration standards run on the Superose 6 column. ....	75
Table 4.2	Hexagonal crystals XDS data processing. ....	79
Table 4.3	P <sub>6</sub> <sub>3</sub> <sub>2</sub> <sub>2</sub> V <sub>M</sub> analysis results. ....	80
Table 4.4	Previous crystallization and freezing results. ....	100
Table 4.5	Bar crystals data reprocessing with XDS. ....	101
Table 4.6	P <sub>2</sub> <sub>1</sub> <sub>2</sub> <sub>1</sub> <sub>2</sub> <sub>1</sub> V <sub>M</sub> analysis results. ....	101
Table 4.7	Prismatic crystals data reprocessing with XDS. ....	103
Table 4.8	P <sub>4</sub> <sub>2</sub> <sub>2</sub> <sub>1</sub> <sub>2</sub> V <sub>M</sub> analysis results. ....	103
Table 4.9	X-ray refinement statistics for the P <sub>4</sub> <sub>2</sub> <sub>2</sub> <sub>1</sub> <sub>2</sub> dataset. ....	106
Table 4.10	X-ray model validation. ....	107
Table 4.11	Cryo-EM refinement. ....	111
Table 4.12	Cryo-EM model validation. ....	111
Table 4.13	Portal protein dimensions. ....	122



# List of abbreviations and symbols

Abbreviation or symbol	Meaning
[v/v]	Volume per volume
[w/v]	Weight per volume
%	Percentage
2D	Two-dimensional
3D	Three-dimensional
Å	Armstrong
APS	Ammonium persulfate
ATP	Adenosine triphosphate
bp	Base pair
cal	Calorie
CTF	Contrast transfer function
Da	Dalton
DLS	Dynamic light scattering
DM	Density modifications
DNA	Deoxyribonucleic acid
dsDNA	Double-stranded DNA
dsRNA	Double-stranded RNA
DTT	Dithiothreitol
e <sup>-</sup>	Electron
Ed.	Editor
<i>E. coli</i>	<i>Escherichia coli</i>
EDTA	Ethylenediaminetetraacetic acid
EM	Electron microscopy
ESRF	European Synchrotron Radiation Facility
eV	Electron volt
$f_0$	Normal scattering term
$f'$	Anomalous scattering dispersion term
$f''$	Anomalous scattering absorption term
$F$	Structure factor
$F_{obs}$	Observed structure factor
$F_{calc}$	Calculated structure factor
FSC	Fourier shell correlation
g	Gram
gp	Gene product

<b>h</b>	Hour
<b>HCMV</b>	Human cytomegalovirus
<b>HisTag</b>	Histidine tag
<b>HSV-1</b>	Herpes simplex virus-1
<b>ICTV</b>	International Committee on Taxonomy of Viruses
<b>IMAC</b>	Immobilized metal ion affinity chromatography
<b>IPTG</b>	Isopropyl $\beta$ -D-1-thiogalactopyranoside
<b>Iter</b>	Iteration
<b>L</b>	Litre
<b>LB</b>	Luria-Bertrani
<b>LC-MS/MS</b>	Liquid chromatography tandem-mass spectrometry
<b>M</b>	Molar
<b>m</b>	Meter
<b>MAD</b>	Multi-wavelength anomalous dispersion
<b>MES</b>	2-( <i>N</i> -mopholino)ethanesulfonic acid
<b>Min</b>	Minute
<b>MIR</b>	Multiple isomorphous replacement
<b>MIRAS</b>	Multiple isomorphous replacement with anomalous scattering
<b>MOPS</b>	3-( <i>N</i> -morpholino) propanesulfonic acid
<b>MR</b>	Molecular replacement
<b>MS</b>	Mass spectrometry
<b>MW</b>	Molecular weight
<b>N</b>	Newton
<b>NCS</b>	Non-crystallographic symmetry
<b>°</b>	Degree
<b>O.D.</b>	Optical density
<b>PEG</b>	Polyethilene glycol
<b>°C</b>	Degree Celsius
<b>pRNA</b>	Packaging RNA
<b>Psi</b>	Pounds per square inch
<b>Px</b>	Pixel
<b>r.m.s.d.</b>	Root mean square deviation
<b>RNA</b>	Ribonucleic acid
<b>Rpm</b>	Revolutions per minute
<b>s</b>	Second
<b>SAD</b>	Single-wavelength anomalous dispersion
<b>SeMet</b>	Selenomethionine

<b>SDS</b>	Sodium dodecyl sulphate
<b>SIR</b>	Single isomorphous replacement
<b>SIRAS</b>	Single isomorphous replacement with anomalous scattering
<b>SRF</b>	Self-rotation function
<b>ssDNA</b>	Single-stranded DNA
<b>ssRNA</b>	Single-stranded RNA
<b>TEMED</b>	N,N,N,N-tetramethylethylendiamine
<b>TerL</b>	Large terminase subunit
<b>TerS</b>	Small terminase subunit
<b>Tfb I</b>	Transformation buffer I
<b>Tfb II</b>	Transformation buffer II
<b>UV</b>	Ultraviolet
<b>V</b>	Volt
<b>V<sub>M</sub></b>	Matthews coefficient
<b>x g</b>	Times of standard gravity
<b>λ</b>	Wavelength





# Amino acids abbreviations

Amino acid	One letter code	Three letter code
Alanine	A	Ala
Arginine	R	Arg
Asparagine	N	Asn
Aspartic acid	D	Asp
Cysteine	C	Cys
Glutamic acid	E	Glu
Glutamine	Q	Gln
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Leucine	L	Leu
Lysine	K	Lys
Methionine	M	Met
Phenylalanine	F	Phe
Proline	P	Pro
Serine	S	Ser
Threonine	T	Thr
Tryptophan	W	Try
Tyrosine	Y	Tyr
Valine	V	Val



# Preface

How large double-stranded DNA viruses are able to fill the capsids with their genomes to a near-crystalline density and deliver it into host cells at high efficiency is a topic that has fascinated scientists for a long time. It is not surprising that because of the effectiveness of the packaging process some mechanisms and components are conserved in phylogenetically distant viruses such as bacteriophages from the *Caudovirales* order and eukaryotic-infecting viruses of biomedical interest like Herpesviruses and Adenoviruses. Despite the importance and relevance of this topic, there are still many important biochemical and mechanistic questions about this process waiting to be answered.

Regarding T7 bacteriophage, although this virus and its encapsidation mechanism have been in the focus of many structural and functional studies, the atomic structure of their packaging proteins remains unknown. This PhD thesis covers the work done for the structural characterization and analysis of the T7 bacteriophage portal protein, one of the packaging machinery components, from sample preparation to structure discussion.

Obtaining good quality diffracting crystals and phasing the data is a bottleneck for many structural projects, especially when dealing with big protein complexes as in our case. In the last years the so-called resolution revolution in the cryo-electron microscopy technique has changed the structural biology field, and this method showed up as a feasible alternative for our project. Finally, an interesting combination of both techniques allowed us to solve the 3D structure of the portal protein at 2.80 Å resolution.

This study was done at the Institute for Research in Biomedicine (IRB-Barcelona) and the Institut de Biologia Molecular de Barcelona (IBMB-CSIC) under the supervision of Prof. Miquel Coll and Dr. Cristina Machón. For the electron microscopy studies, we have collaborated with the research group of Prof. José L. Carrascosa at the Centro Nacional de Biotecnología (CNB-CSIC) in Madrid.



This PhD thesis is organized in the following sections:

- The **Introduction** summarizes previous knowledge about the viral organism under study and the large double-stranded DNA viruses packaging mechanism.
- The **Objectives** section lists the goals of the project.
- The **Materials and methods** chapter lists and describes the materials, instruments and common techniques used throughout the study. The chapter is separated in four parts. The first one describes the sample preparation and analysis steps required before structural studies. The second one is dedicated to crystallization and X-ray data collection and processing. The next section reports the cryo-electron microscopy studies. Finally, the last section shows the process of obtaining the final atomic model, combining data from cryo-electron microscopy and X-ray crystallography experiments.
- The **Results and discussion** section shows the experimental data produced during the study and their interpretation, both from a biological and a methodological point of view. The chapter is organized in the same parts as the Materials and methods section, plus some additional ones with further discussion.
- Finally, the **Conclusions** section summarizes the main findings of this PhD thesis.



# Chapter I:

## Introduction





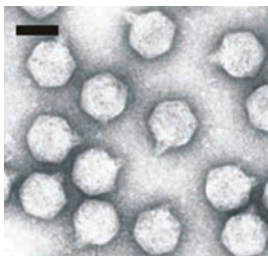
# 1.1 The T7 bacteriophage

## 1.1.1 Viruses, bacteriophages and the T7 bacteriophage

Viruses are infectious agents that are only able to replicate inside living cells of other organisms. All types of life forms can be infected by different types of viruses; from microorganisms, both bacteria and archaea, to animals and plants (Koonin *et al.*, 2006). Viruses are the most abundant type of biological entity on Earth, and can be found in almost every ecosystem (Lawrence *et al.*, 2009; Edwards *et al.*, 2005).

Bacteriophages, also called phages, are viruses that infect bacteria and use them as host cells to multiply. They were first discovered approximately a hundred years ago by Frederick Twort and Félix d’Hérelle, being described by the second as “virus parasitic on bacteria” (Twort, 1915; d’Hérelles, 1917). In fact, the word *bacteriophage* literally means “bacteria eater” in Greek. Early research efforts were mainly focused in phage therapy, exploring their potential medical use to kill pathogenic bacteria. However, the discovery of antibiotics in 1928 changed the scope of investigations in the field. From then on, bacteriophages were studied as model organisms to investigate basic viral biology (Keen, 2015).

One of the model phages that has been extensively studied is the T7 bacteriophage (Figure 1.1). Although probably close relatives had been used in previous studies, this *Escherichia coli* infecting virus was first identified and mentioned in the literature in 1945 (Demerec and Fano, 1945). It is able to infect strains of *E. coli* (B, C and K-12), *Shigella* spp. and *Salmonella enterica* (Lindberg, 1973). Because of its manageable genome size when compared with other bacteriophages, it has been a perfect choice for genetic and biochemical analysis with the goal of obtaining information about basic genetic processes (Studier, 1972).



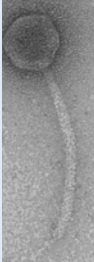
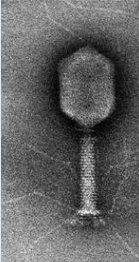
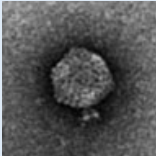
**Figure 1.1** T7 bacteriophage viral particles. Electron microscopy (EM) images of T7 bacteriophage viral particles negatively stained. Scale bar corresponds to 50 nm. (Cuervo *et al.*, 2014)

## 1.1.2 Taxonomy and phylogeny

All the viruses are classified by the International Committee on Taxonomy of Viruses (ICTV) according to their genome characteristics, morphology, viral particle size and infecting host (Web 1). Among viruses that infect Bacteria and Archaea, different types of genomes can be found: linear double-stranded DNA (dsDNA), circular dsDNA, circular single-stranded DNA (ssDNA), segmented double-stranded RNA (dsRNA) and linear single-stranded RNA (ssRNA). Morphologically, prokaryote-infecting viruses can be either enveloped or non-enveloped, and they can have many different shapes: isometric, spherical, ovoid, rod-shaped, bottle-shaped, lemon-shaped, filamentous, pleomorphic or tailed.

There are nineteen families recognized by the ICTV able to infect Bacteria and Archaea. The *Caudovirales* order comprises the so-called dsDNA tailed bacteriophages. It is a very large order, that accounts for 96% of prokaryote-infecting viruses, and it is divided in three families according to the morphology of viral tails (Table 1.1):

**Table 1.1** *Caudovirales* families. Morphology, abundance and examples (Web 1).

Family	<i>Siphoviridae</i>	<i>Myoviridae</i>	<i>Podoviridae</i>
Tail morphology	Long non-contractile	Long contractile	Short non-contractile
Percentage	62%	24%	14%
Examples of species and their host bacteria	SPP1 ( <i>Bacillus subtilis</i> )	T4 ( <i>E. coli</i> )	P22 ( <i>Salmonella typhimurium</i> ) φ29 ( <i>Bacillus spp.</i> )
Negative staining EM images	SPP1 (Alonso <i>et al.</i> , 2006) 	T4 (Kutter <i>et al.</i> , 2013) 	P22 (King <i>et al.</i> , 1976) 

The T7 bacteriophage has a short non-contractile tail, and therefore is classified into the *Podoviridae* family (Table 1.2).

**Table 1.2** Taxonomic classification of T7. According to the ICTV 10th report (Web 1).

<b>Group</b>	I (dsDNA)	<b>Subfamily</b>	<i>Autographivirinae</i>
<b>Order</b>	<i>Caudovirales</i>	<b>Genus</b>	<i>T7virus</i>
<b>Family</b>	<i>Podoviridae</i>	<b>Species</b>	<i>Escherichia virus T7</i>

T7 and T3 bacteriophages are closely related, and both present high sequence identity on many genes (Pajunen *et al.*, 2002). They share many characteristics and in some available bibliography they are reviewed jointly.

Regarding other dsDNA tailed bacteriophages, it is important to highlight that although viruses from the *Caudovirales* order have many things in common there is a huge genetic diversity among them. The comparison of their genomes has revealed their highly mosaic nature. Homologous recombination can occur between phage genes that diverged recently and new ones and can be obtained both from hosts or from other viruses. This situation makes taxonomy difficult, because the ongoing divergence is superimposed with novel junctions as a consequence of the frequent horizontal genetic transfer events (Casjens, 2005).

One relevant phylogenetic observation for our study is the linkage between tailed-bacteriophages and eukaryotic infecting Herpesviruses. As protein structures are more conserved than genetic sequences, structural comparisons can often be useful to analyse distant phylogenetic relationships. Based on that, a common origin for *Caudovirales* bacteriophages and Herpesviruses could be suggested by the structural analysis of their capsid protein topologies and virion architectures, which shows unexpected similarities (Bamford *et al.*, 2005). Both types of viruses are part of the so-called HK97 lineage, as they share the HK97 basic folding unit in their major capsid proteins (Baker *et al.*, 2005; Veesler and Cambillau, 2011). Moreover, similarities between *Caudovirales* and Herpesviruses are also found when their capsid assembly and DNA packaging mechanisms are compared. It is considered that probably these structures related with primordial capsid and packaging functions were derived from an ancient common ancestor (Rixon and Schmidt, 2014).

### 1.1.3 Genome organization

The T7 bacteriophage mature genome has 39,937 base pairs (bp), but it is replicated as an end-to-end polymer or concatemer. After maturation cleavage, the genome has a nucleotide sequence of 160 bp repeated in both ends (Studier, 1972).

T7 genes are named according to their order from one end of the genome to the other, considering that the left end is the first one being injected into bacterial host cells and that non-integral numbers correspond to genes discovered after initial numbering (Studier, 1969; Molineux, 2001).

The functions of many genes have been determined with the help of conditional-lethal mutants (Serwer, 2005). T7 bacteriophage genes have been classified in three different classes (Table 1.3).

**Table 1.3 T7 bacteriophage genes.** Summary of some relevant genes and their functions. They appear classified according to their class. (Adapted from Studier and Dunn, 1983.)

Class	Gene number	Function
I	1	RNA polymerase
II	2	Inactivates host RNA polymerase
	2.5	ssDNA-binding protein
	3	Endonuclease
	4	Primase-helicase
	5	DNA polymerase
	6	5' to 3' exonuclease
III	7.3	Tail protein
	8	Portal protein
	9	Scaffolding protein
	10	Capsid proteins
	11 and 12	Tail proteins
	14, 15 and 16	Inner core proteins
	17	Tail fiber protein
	18	Small terminase subunit
	19	Large terminase subunit

T7 bacteriophage genes are clustered in three classes that are expressed in different stages of infection (Studier, 1972):

- Class I: Early expression genes, including the RNA polymerase and some genes codifying for proteins able to inactivate host ones and produce a favourable environment for viral multiplication.
- Class II: Injected right after the class I genes, encoding proteins related with viral genome replication.
- Class III: Codify for capsid assembly and DNA packaging proteins.

#### **1.1.4 Morphology and structure**

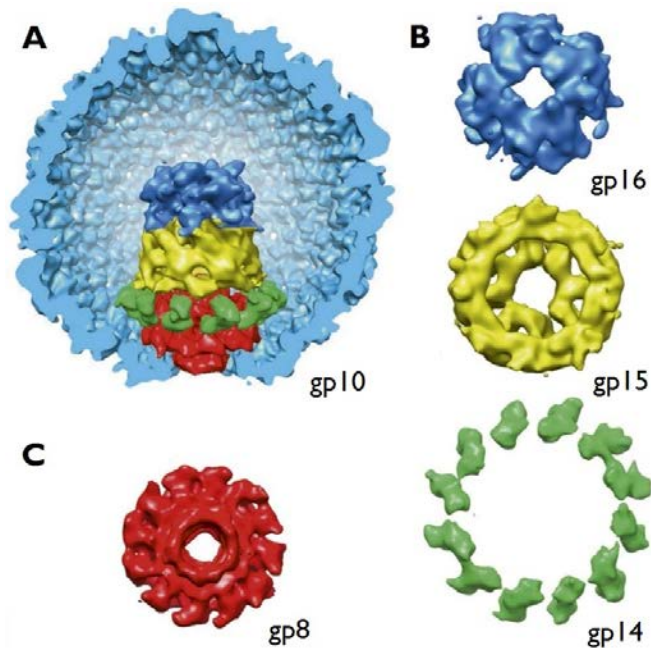
Outside bacterial cells, bacteriophage genomes are surrounded by a protein capsid that protects them. The T7 bacteriophage shares with the other *Caudovirales* viruses two main structural elements (Cuervo and Carrascosa, 2012a):

1. **Head**: It is an icosahedral structure with a diameter of about 60 nm (Stroud *et al.*, 1981). It is formed by capsomers, which are repetitive structural units of pentamers and hexamers of the major capsid protein gp10A with HK97 fold (Agirrezabala *et al.*, 2007). In some bacteriophages, other proteins associate with major capsid proteins to stabilize capsomer connections. In T7 there is gp10B minor capsid protein, which is produced by a read-through that occurs at a frequency of 10% (Condron *et al.*, 1991). Hexamers build capsid faces and edges, while pentamers are located at all the five-fold vertices except one, where the dodecameric portal protein is found (Figure 1.2). In T7 this ring-shaped assembly is codified by the gp8 gene (Cerritelli and Studier, 1996b). It participates in viral morphogenesis and builds a channel through which DNA is translocated during genome encapsidation and ejection. Sometimes it is also called connector, because in mature virions it links the head with the tail. Inner scaffolding protein gp9, is only present in immature capsids before DNA encapsidation (also called procapsids or proheads) and helps during morphogenesis. It interacts both with the portal and with shell protein capsomers. The later ones undergo conformational changes that are thought to produce a stability increase of mature capsids (Cerritelli *et al.*, 2003). Other changes during prohead

maturation occur in the inner core proteins, which form a complex associated with the portal protein. In T7 it is a ring-shaped complex of three proteins (Agirrezabala *et al.*, 2005a; Guo *et al.*, 2013). Each one of them has a different symmetry arrangement (Figure 1.2):

- Tip (gp16): With 4-fold symmetry, it is located in the more internal part of the capsid, distal to the portal complex.
- Bowl (gp15): Joins gp16 and gp14 and has 8-fold symmetry.
- Adaptor (gp14): Complex with 12-fold symmetry that links gp15 with the portal ring.

In mature capsids the 40 kbp genome is wrapped around the long axis of the inner core in six co-axial shells in a quasi-crystalline packing (Cerritelli *et al.*, 1997).

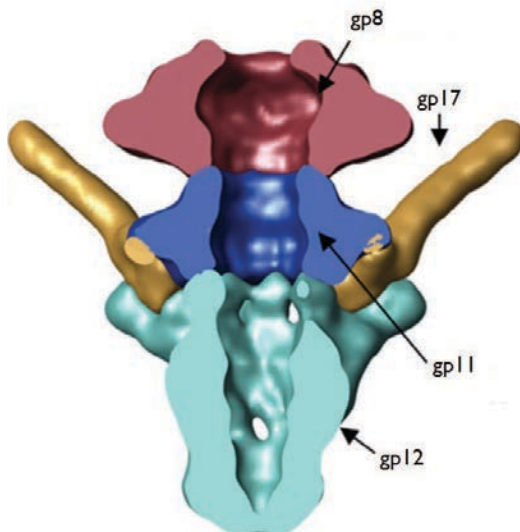


**Figure 1.2 Prohead, internal core and portal of the T7 bacteriophage.**  
 (A) Side view of a cryo-EM reconstruction of a T7 procapsid. Resolution depends on the protein. gp10 (13Å) appears in light blue, gp16 (20Å) in dark blue, gp15 (17Å) in yellow, gp14 (21Å) in green and gp8 (17Å) in red.  
 (B) Axial views of the inner core proteins gp16, gp15 and gp14.  
 (C) Axial view of the portal assembly (gp8).  
 (Adapted from Guo *et al.*, 2013.)

2. **Tail:** Assembled after genome packaging, it is used during infection to recognise the bacterial host cell and to deliver the genome efficiently into its cytoplasm. In T7, the tail is formed by three proteins attached to the outer part of the portal (Cuervo *et al.*, 2013b). Their locations and specific symmetries have been characterized (Figure 1.3):

- **Adaptor (gp11):** This protein builds the part of the tail in direct contact with the portal. Also called gatekeeper, it has dodecameric symmetry.
- **Nozzle (gp12):** It is the biggest protein forming the tail channel and forms a conical domain with 6-fold symmetry.
- **Fibre (gp17):** Each of the copies of gp12 is associated with a thin fibre, which is a trimer of the gp17 protein (Steven *et al.*, 1988). The C-terminal part of the protein is specialized in host recognition, and interacts with bacterial lipopolysaccharide by its tip domain (Garcia-Doval and van Raaij, 2012).

The protein gp7.3 could also form part of the T7 tail, but its precise location is still unknown (Kemp *et al.*, 2005). The core complex, the portal and the tail build a continuous channel where the left end of the viral DNA is located (Agirrezabala *et al.*, 2005a).



**Figure 1.3 T7 bacteriophage tail.** Side view of a cryo-EM reconstruction at 16Å resolution. gp8 appears in red, gp11 in dark blue, gp12 in green and gp17 in orange. (Adapted from Cuervo *et al.*, 2013b.)



### 1.1.5 Viral infection cycle

Another possible way to classify bacteriophages is depending on their infection cycle. All of them use bacterial host cell machineries to amplify their genomes and produce new viral particles, but two different strategies for that have been described. Lytic or virulent viruses multiply right after infection, killing and lysing host cells as soon as new viral particles have been assembled. In a different manner, temperate viruses are able to remain in latency as prophages, either with their genome integrated in the bacterial one or maintained like a plasmid. During the lysogenic state, they replicate their genomes within the bacterial cell without killing it. Only when the bacterial cell is under stress, the activation of some viral genes triggers the beginning of the lytic cycle (Madigan *et al.*, 2010).

The T7 bacteriophage presents a lytic cycle with two different steps (Studier, 1972):

1. **Host cell recognition and genome internalization:** T7 fibers bind the bacterial receptor and change their orientation during adsorption. This process triggers conformational changes that lead to the opening of the tail channel and the translocation of proteins that build an extended structure for DNA injection between both cell membranes (Hu *et al.*, 2013). It is thought that the proteins building the extended tail are the inner core proteins, and this hypothesis would explain the catalytic activity of gp16, which is able to break the cell wall peptidoglycan (Moak and Molineux, 2000). DNA translocation during ejection is an enzyme driven not continuous process, as translocation and transcription are coupled: first the bacterial, and afterwards the viral RNA polymerase, act as motors for DNA injection (Molineux, 2001). Once the DNA has been completely ejected, the extended phage tail disassembles, and the cell membrane reseals (Hu *et al.*, 2013).
2. **Morphogenesis of new viral particles:** After genome internalization into the cytoplasm viral genes are expressed and the bacteriophage genome is replicated. T7 follows the general assembly process for large dsDNA viruses, reviewed in detail on section 1.2.2. DNA packaging and capsid maturation produce new infective particles, cell lysis occurs and the cycle is restarted.

### 1.1.6 Biotechnological applications

Bacteriophage T7 has been used for many biotechnological applications. Some of them are listed below:

- Expression vectors: Molecular biology and biochemical studies of the T7 bacteriophage lead to a deep knowledge about its promoters, ribosome binding sites and details about the specific function of many genes. This information has been used to develop a whole cloning system, where the T7 promoter and T7 RNA polymerase are used. Improved protein expression vectors have also been developed, for instance using T7 lysozyme to reduce basal activities of T7 RNA polymerase when working with toxic proteins (Studier, 1991).
- Bacteriophage-display system: This method is useful for identifying peptides or proteins with certain binding properties starting from a DNA library. Molecules are displayed on the surface of the phage fused to viral capsid proteins, and they can be selected according to their binding affinities to particular targets (Rosenberg *et al.*, 1996).
- Phage therapy: Defined as the usage of bacteriophages to treat bacterial infections, it was developed at the Pasteur Institute in Paris. However, since the discovery of chemical antibiotics in the 1940s it has been ignored in Western countries. Nowadays emerging bacterial resistances to antibiotics lead to a growing interest for the potential use of phages to complement antibiotics in the fighting against infections (Kutter *et al.*, 2013). In fact, in 2006 the United States Food and Drug Administration approved a phage preparation to be added to meat and poultry products to fight against human infections by *Listeria monocytogenes* (Lang *et al.*, 2006). Lately, tackling *E. coli* infections by phage therapy strategies has been reconsidered, and T7 bacteriophage could be an option for that (Brüssow, 2005).

## 1.2 Viral genome packaging

### 1.2.1 Viral assembly strategies

Viruses protect their genetic information inside multifunctional protein containers, which are usually built from a limited set of proteins in order to consume the minimum genetic information. The viral particles have to be easily-assembled in a short time and must be able to selectively incorporate the viral genetic material. Many viruses follow a co-assembly strategy by which the viral particles are formed when the genetic material and the capsid proteins interact with each other. This strategy is followed by many RNA and ssDNA viruses. However, restrictions imposed by the properties of dsDNA limit the possible ways to enclose the genomes of the viruses with this type of nucleotide molecules (Cuervo *et al.*, 2013a).

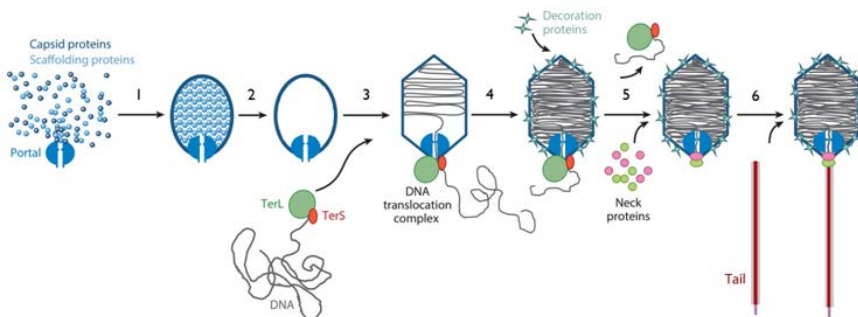
Large dsDNA viruses pack their genomes into the capsids to a near-crystalline density (Earnshaw and Casjens, 1980). Bacteriophages of the *Caudovirales* order, Herpesviruses and Adenoviruses share basic aspects of a complex dsDNA packaging mechanism (Rixon and Schmidt 2014; Ahi *et al.*, 2016).

### 1.2.2 The viral assembly pathway of large dsDNA viruses

The viral assembly pathway of large dsDNA viruses has been extensively characterized and reviewed in many articles (Rao *et al.*, 2015, and references cited therein). The general pathway can be divided into the following six steps (Figure 1.4):

1. **Procapsid assembly:** During formation of the immature procapsid dodecameric portal protein nucleates the coassembly of the capsid protein and the scaffolding protein. A 12:5 symmetry mismatch is created at a unique 5-fold vertex of the capsid.
2. **Procapsid maturation and expansion:** The scaffolding protein is cleaved and the resulting peptides diffuse out of the procapsid. An empty, rounded, thick-walled mature procapsid structure is formed.

3. **Packaging initiation:** The terminase recognizes the viral genome, usually synthesized by the host cell replication machinery as a head-to-tail polymer (concatemer). The terminase makes an endonucleolytic cut and remains bound to the new DNA free end. The DNA-terminase complex docks on the portal protein assembling an oligomeric ring motor.
4. **DNA translocation:** Using the energy from ATP hydrolysis, the terminase catalyses DNA translocation. When about 10-25% of the viral genome has been packed, the procapsid expands leading to a bigger, thinner and more angular capsid shell. The inner capsid volume also increases matching the viral genome size. Decorating proteins may bind to the capsid surface to reinforce the structure.
5. **Packaging termination:** After encapsidating the whole viral genome, the terminase makes a second cut on the DNA and dissociates from the capsid. However, it remains bound to the newly formed concatemer end, ready to bind to another procapsid and encapsidate the next genome. The portal protein prevents DNA loss from the pressurized capsid, and in *Caudovirales* the assembly of neck proteins also participate in the sealing.
6. **Attachment of the tail:** In the case of bacteriophages, infective virions are produced once the tail is assembled. A tail-like assembly has also been described in the case of Herpesviruses, whose function is still unknown (Schmid *et al.*, 2012).



**Figure I.4** Assembly pathway of large dsDNA viruses.

Steps follow the same numbering code as in the text.

TerL and TerS refer to large and small terminase subunits, respectively.

(Adapted from Rao and Feiss, 2015.)

### 1.2.2.1 *The packaging proteins*

The packaging machinery is mainly composed by two components: the portal protein and the terminase proteins. Although homolog proteins from different dsDNA viruses are highly divergent in terms of sequence similarity, structural comparisons among them suggest that they probably use a similar underlying mechanism that has evolved in different manners, depending on the specific characteristics of each viral system and packaging mechanism (Casjens, 2005; Hendrix, 2002; Casjens, 2011).

Available information about DNA processing and packaging proteins will be summarized on the following sections, focusing on:

- Bacteriophages for which there is available atomic structural information about their portal protein:  $\phi$ 29, SPPI, P22 and T4.
- The most studied Herpesviruses: herpes simplex virus-1 (HSV-1) and human cytomegalovirus (HCMV).

T7 bacteriophage will be reviewed in detail in section 1.3.

### 1.2.2.2 *Strategies of DNA processing*

Bacteriophage genomes can have different types of ends, which correlate with different terminase cleaving and packaging mechanisms (Casjens and Gilcrease, 2009 and references cited therein).

Some of the types of DNA ends are summarized below:

1. **Terminally redundant and circularly permuted ends:** Viral DNA is replicated as a concatemer. For the packaging initiation cleavage, a specific site (*pac*) is recognized. However, the location of subsequent cleavages is not sequence specific, and depends on the available volume inside of the head. The packaged DNA is typically between 102% and 110% of the genome length, and the ends of the genomes packaged in serial packing events moved along the sequence. Viruses belonging to this class are called headful packaging phages and include SPPI, P22 and T4.

2. **Short exact direct repeated ends:** In this case, viral genomes are replicated also as concatemeric molecules. Direct double-stranded repeats of a few hundred bp are present at both ends of the genome and are generated in concert with DNA packaging. T7 bacteriophage has this type of genome ends.
  
3. **With terminal proteins:** Viral DNA is replicated in form of monomeric genomes, and terminal proteins are covalently bound to their ends. The only bacteriophages known to have terminal proteins are  $\phi 29$  and its relatives.

Regarding Herpesvirus genomes, they have two packaging sequences, *pac1* and *pac2*, that are found near the genomic ends. At one end, *pac2* mediates initiation and indicates packaging directionality. At the other end, *pac1* terminates packaging after a unit-length genome has been encapsidated (Brown *et al.*, 2002).

### 1.2.3 The terminase proteins

As mentioned before, the main known functions of the terminase proteins during encapsidation are recognizing the DNA, catalysing the nuclease activity and providing the energy for DNA translocation hydrolysing ATP.

In bacteriophages there are two different terminase proteins, which are called the small and large terminase subunits (Table 1.4). This nomenclature derives from the phage  $\lambda$  system, the first described (Mousset and Thomas, 1969).

**Table 1.4 Bacteriophage terminase proteins.** List of terminases and their monomeric molecular weight (MW).

Virus	Small terminase (TerS)	MW	Large terminase (TerL)	MW
T7	gp18	10.15 kDa	gp19	66.26 kDa
$\phi 29$	-	-	gp16	38.96 kDa
SPP1	gp1	16.33 kDa	gp2	48.84 kDa
P22	gp3	18.65 kDa	gp2	57.59 kDa
T4	gp16	18.39 kDa	gp17	69.76 kDa

The interaction between both proteins varies depending on the viral system (Casjens, 2011). In some cases, like in P22 infected cells, both proteins can be found as a complex (Poteete and Botstein, 1979). However, in other cases, for instance T4, both proteins do not seem to interact in a tightly manner (Al-Zahrani *et al.*, 2009).

As observed on Table 1.4,  $\phi$ 29 and relative phages only have one packaging protein, but there is another molecule that participates in the process, the packaging RNA (pRNA). It is a 174 nucleotides long phage encoded RNA, required for the assembly of the ATPase terminase gp16 on the molecular motor (Ding *et al.*, 2011).

In the case of Herpesviruses, three different subunits form the terminase complex (Table 1.5). The interaction between the terminase proteins (TRM1, TRM2, TRM3) has been detected by immunoprecipitation in cytoplasmatic and nuclear lysates of infected cells (Yang *et al.*, 2007). Although it is not the largest component of the terminase complex, there is structural evidence supporting that TRM3 is the equivalent to the TerL from bacteriophages (Nadal *et al.*, 2010; Selvarajan Sigamani *et al.*, 2013). TRM1 would be equivalent to the small terminase subunit in bacteriophages, while TRM2 would not have any homolog in prokaryotic viruses (Sankhala *et al.*, 2016).

**Table 1.5** Herpesvirus terminase proteins. *List of terminases and their monomeric MW.*

Virus	TRM1	MW	TRM2	MW	TRM3	MW
HSV-1	pUL28	85.48 kDa	pUL33	14.44 kDa	pUL15	80.92 kDa
HCMV	pUL56	95.56 kDa	pUL51	16.98 kDa	pUL89	77.08 kDa

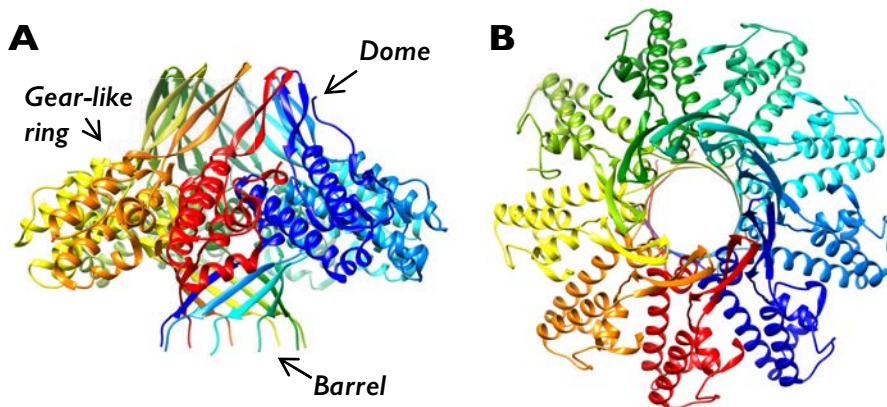
### 1.2.3.1 *The small terminase subunit*

TerS is the most variable packaging protein in terms of amino acid sequence (Casjens and Thuman-Commike, 2011). It is able to recognize the DNA and is required for packaging initiation, although it is not clear whether it has any role in DNA translocation (Casjens, 2011). In phages where TerS and TerL form stable complexes, like in SPPI, it may be present on the motor during DNA translocation (Oliveira *et al.*, 2005).

In other cases, like in the T4 phage system, although TerS does not have enzymatic activity, and it is not required for packaging during *in vitro* experiments, it has a key role coordinating the TerL ATPase, translocase and nuclease functions (Zhang *et al.*, 2011; Al-Zahrani *et al.*, 2009).

P22 gp3 structure shows that it can assemble as a nonameric ring, but mutants with the ability to assemble in decamers have also been described (Roy *et al.*, 2012; Nemecek *et al.*, 2008). SPP1 gp1 and T4 gp16 can probably form octameric assemblies, but the central domain of a gp16 close homolog has also been crystallized forming undecameric and dodecameric assemblies (Chai *et al.*, 1995; Lin *et al.*, 1997; Sun *et al.*, 2012). Therefore, experimental data suggests a less constrained spatial organization of this protein if compared with TerL and the portal (Casjens, 2011).

The nonameric crystallographic structure of P22 gp3 shows three different domains: a  $\beta$ -stranded dome, a gear-like ring and a  $\beta$ -barrel (Figure 1.5). The ring has a central channel of 23Å diameter, about the same one as a double helix B-DNA. The last 23 residues are essential for binding the DNA and for the assembly with TerL (Roy *et al.*, 2012).



**Figure 1.5** Cartoon representation of P22 TerS at 1.75Å resolution.  
(A) Lateral view of the nonamer with the three domains indicated (rainbow colouring per monomer).  
(B) Axial view of the nonamer (rainbow colouring per monomer).  
(Data from Roy *et al.*, 2012.)

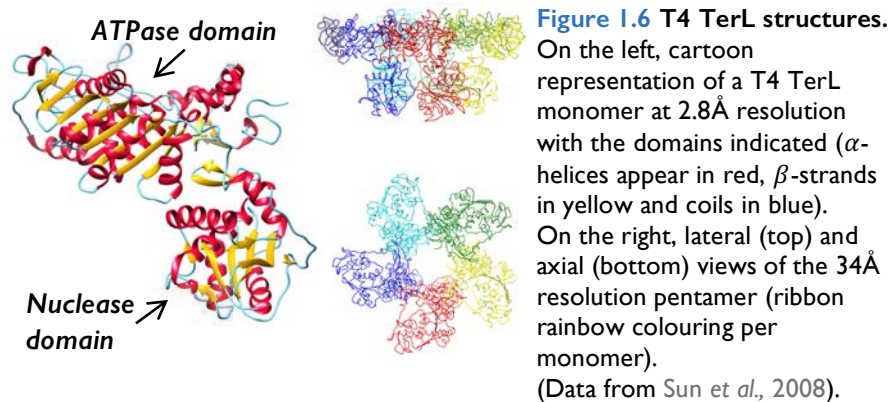
T4 gp16 is thought to have three domains: the N-terminal domain for interaction with the DNA, the helical central oligomerization domain and the C-terminal domain for interaction with TerL (Sun *et al.*, 2012).



### 1.2.3.2 The large terminase subunit

TerL of bacteriophages such as SPPI and P22 have been described as monomers, when expressed alone (Nemecek *et al.*, 2007; Gual *et al.*, 2000; Sun *et al.*, 2008). The crystallographic structure of T4 gp17 shows two different domains closely located on a “tense state” (Figure 1.6):

- **ATPase domain:** Located at the N-terminal, has a  $\beta$ -sheet core and provides the energy for DNA packaging. It contains some typical ATPase features: a Walker A motif, a Walker B motif, an adenine binding motif and a catalytic carboxylate (Walker *et al.*, 1982).
- **Nuclease domain:** C-terminal domain that contains the nuclease active site which participates in genome translocation.



However, a 34Å resolution cryo-EM reconstruction of gp17 with the T4 procapsid shows a pentameric arrangement with monomers spatially separated, in a “relaxed state” (Figures 1.6). A packaging mechanism driven by electrostatic forces and alternation between states has been suggested (Sun *et al.*, 2008).  $\phi$ 29 cryo-EM data of an active packaging complex also showed a pentameric arrangement of TerL, with both domains able to interact with the DNA (Mao *et al.*, 2016). The nuclease activity of TerL depends on the presence of divalent metal ions. The HCMV terminase nuclease has an RNaseH/integrase-like fold (Figure 1.7). Two manganese ions are present in its active site (Nadal *et al.*, 2010). SPPI and P22 structures show manganese and magnesium atoms in their active sites respectively, while in HSV-I the activity requires magnesium (Smits *et al.*, 2009; Roy and Cingolani, 2012; Selvarajan Sigamani *et al.*, 2013).



**Figure 1.7 HCMV terminase structures.** HCMV pUL89 nuclease domain at 3.2Å resolution ( $\alpha$ -helices appear in red,  $\beta$ -strands in yellow, coils in blue and manganese atoms in purple). (Data from Nadal *et al.*, 2010).

## 1.2.4 The portal protein

Some of the known functions of the portals are nucleating the capsid assembly and providing a channel for DNA encapsidation and ejection.

### 1.2.4.1 Structure

The portal proteins of different viruses do not show any significant sequence homology, but they all display a conserved hollow cylindrical ring-shaped architecture with a central channel for DNA passage (Cuervo and Carrascosa, 2012b). Although they are incorporated to the procapsids as dodecamers, it has been described that after overexpression portals can also contain eleven or thirteen monomers (Tsuprun *et al.*, 1994; Dube *et al.*, 1993; van Heel *et al.*, 1996; Sun *et al.*, 2015; Trus *et al.*, 2004). Dodecameric assemblies are probably formed at the beginning of procapsid assembly, requiring the interaction of the portal with the scaffolding and the major capsid proteins (Lurz *et al.*, 2001). The monomeric molecular mass of the most studied portal proteins goes from the 35.88 kDa of the  $\phi$ 29 gp10 to the 82.74 kDa of the P22 gp1 (Table 1.6). Dodecameric assemblies range from 430.56 kDa to 992.88 kDa.

**Table 1.6 Portal proteins.** Summary of their monomeric MW.

Virus	Portal	MW
T7	gp8	59.12 kDa
$\phi$ 29	gp10	35.88 kDa
SPP1	gp6	57.22 kDa
P22	gp1	82.74 kDa
T4	gp20	61.03 kDa
HSV-1	pUL6	74.09 kDa
HCMV	pUL104	78.52 kDa

The first atomic structure available of a dodecameric portal protein was the smallest one, from  $\phi 29$  (Simpson *et al.*, 2000; Guasch *et al.*, 2002). The external diameter is 146Å at the proximal end, which faces the interior of the capsid. On the distal end, facing the exterior, the diameter is 77Å. The height of the portal protein is 75Å (Figure 1.8A).

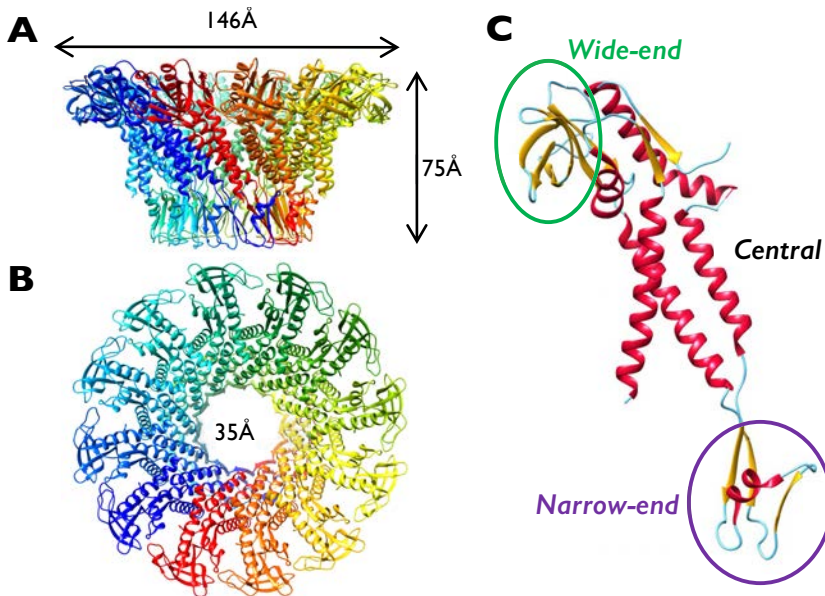
The smaller diameter of the central channel is about 35Å at the distal end, enough to accommodate a dsDNA, which has an average diameter of 23Å (Figure 1.8B). The electrostatic potential of its internal surface is highly electronegative, but there are two electropositive lysine rings separated 20Å from each other.

Three domains were described for each gp10 monomer according to the crystallographic structure (Figure 1.8C):

- Wide-end domain: It is a SH3-like domain composed of six  $\beta$ -strands which are located at the exterior part of the proximal end.
- Central domain: It builds the channel walls and connects the internal face of the wide-end domain with the narrow-end domain, and contains five  $\alpha$ -helices and two  $\beta$ -strands. Three helices form a bundle which is laterally inclined about 45° from the 12-fold axis of the particle. There is a flexible loop facing the channel between two  $\alpha$ -helices that could not be traced.
- Narrow-end domain: This domain is located at the distal end, and contains three  $\beta$ -strands and one  $\alpha$ -helix.

The first 16 and the last 25 residues are not present in the structure, and probably correspond to flexible regions.

Contacts between monomers are of different types. At the narrow-end there is a mixed  $\beta$ -sheet formed with two strands from one monomer and another strand from the adjacent one. Moreover, there are many hydrophobic contacts and hydrogen bonds between subunits.



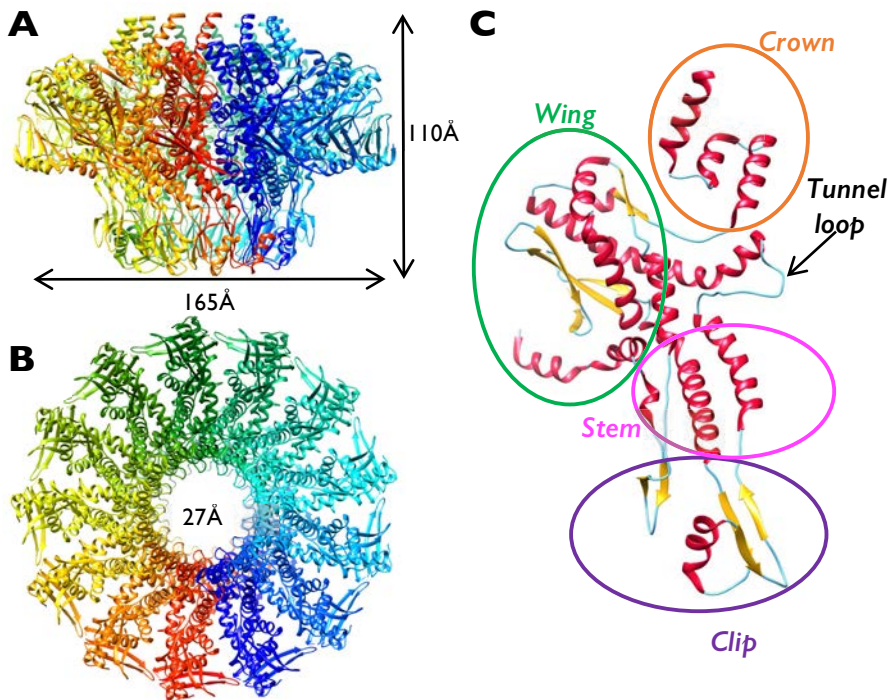
**Figure 1.8** Cartoon representation of  $\phi 29$  portal at 2.1 Å resolution. (A) Lateral view of the dodecamer with portal dimensions indicated (rainbow colouring per monomer). (B) Axial view of the dodecamer with the diameter of the channel indicated (rainbow colouring per monomer). (C) Monomeric gp10 with the three domains indicated.  $\alpha$ -helices appear in red,  $\beta$ -strands in yellow and coils in blue. (Data from Guasch *et al.*, 2002.)

The SPPI portal protein gp6 was also solved by X-ray crystallography, showing a tridecameric assembly (Lebedev *et al.*, 2007). In this case the maximum external diameter is 165 Å and the length is 110 Å (Figure 1.9A). The most constricted part of central channel is about 27 Å of diameter and it is delimited by the tunnel loop, which protrudes into the channel (Figure 1.9B). A pseudoatomic dodecameric model was also built, where the diameter of the channel was predicted to be around 18 Å.

Although some flexible regions could not be traced, four domains were described (Figure 1.9C):

- Wing domain: It is the outer part of the molecule on the proximal end, and is equivalent to the wide-end domain from  $\phi 29$  portal protein, but larger. It is mainly composed by  $\alpha$ -helices and a distal  $\beta$ -sheet.

- Stem domain: It connects the wing with the clip, and matches well with the central domain of gp10, because it also contains tilted  $\alpha$ -helices that build the wall of the channel.
- Clip domain: With an  $\alpha/\beta$  fold, it forms the base of the portal. It corresponds to the narrow-end domain from  $\phi$ 29 portal protein.
- Crown domain: This helical domain is not present on the gp10 structure. Located at the inner proximal end, it is composed by three  $\alpha$ -helices. It corresponds to the C-terminal end of the protein, and 40 residues are not visible on the structure probably because they are disordered.



**Figure 1.9** Cartoon representation of SPPI portal at 3.4 Å resolution.

(A) Lateral view of the tridecamer with dimensions indicated (rainbow colouring per monomer).

(B) Axial view of the tridecamer with the diameter of the channel indicated (rainbow colouring per monomer).

(C) Monomeric gp6 with the four domains and the tunnel loop indicated.

$\alpha$ -helices appear in red,  $\beta$ -strands in yellow and coils in blue.

(Data from Lebedev *et al.*, 2007.)

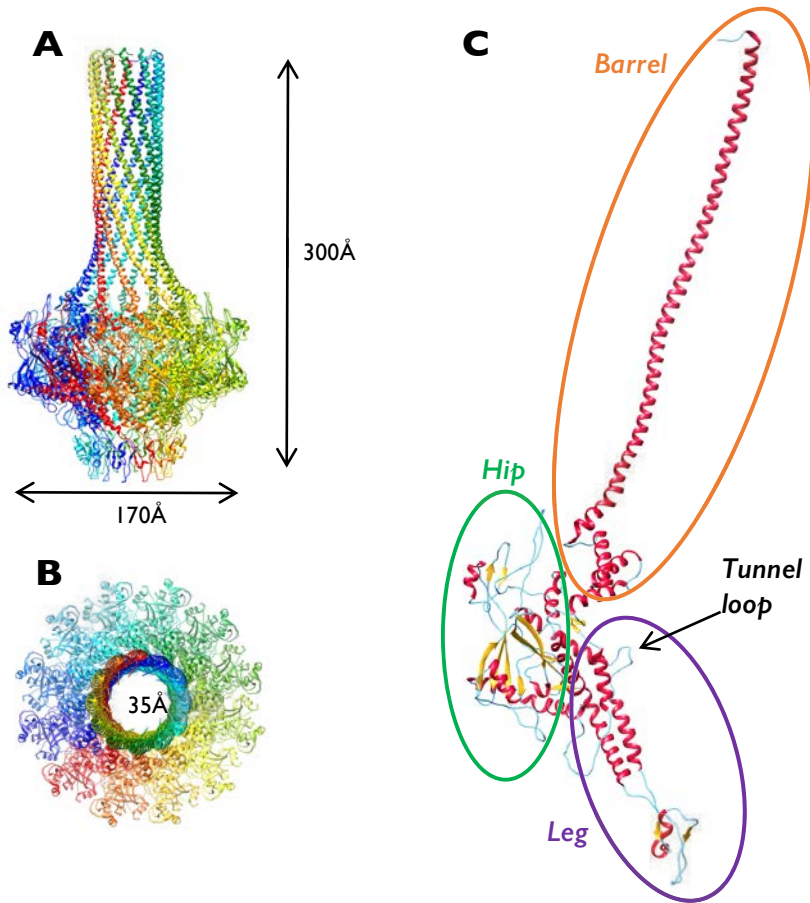
The structure of gpI from P22 was also solved by X-ray crystallography (Olia *et al.*, 2011). Although the full length protein only could be solved at 7.5Å resolution, the protein core in complex with the gatekeeper protein gp4 diffracted up to 3.25Å. So far, this is the largest portal protein described, with an external maximum diameter of the dodecameric particle of about 170Å and a total height of 300Å (Figure 1.10A). The narrowest diameter of the central channel is 35Å (Figure 1.10B). The central channel has five negatively charged rings of glutamate residues.

Three domains were described for each monomer (Figure 1.10C):

- Hip domain: It corresponds to the SPP1 gp6 wing domain, and has an  $\alpha/\beta$  fold with a  $\beta$ -barrel like structure formed by two sheets of eight  $\beta$ -strands that cross each other.
- Leg domain: This domain is mostly helical with an extended  $\alpha/\beta$  domain. The  $\alpha$ -helices that build the channel are tilted by around 30°, and the equivalent central areas to the tunnel loop defined in SPP1 gp6 give a constriction of the channel of 45Å diameter. There are two vestibules of 75Å of diameter immediately above and below these loops. The equivalent domains in SPP1 gp6 are the stem and the clip.
- Barrel domain: It is clearly the most unusual feature of P22 portal protein. It consists on a 200Å long  $\alpha$ -helical stretch of about 120 residues that contains a glutamine-enriched sequence. The narrowest central channel with a diameter of 35Å corresponds to this domain. The barrel domain also includes the equivalent crown region of SPP1 gp6.

The first two domains together form the core of the protein, from which there is atomic resolution structural information.

There is a large interaction interface between monomers where six lysines or arginines of one monomer interact with ten glutamates or aspartates of the neighbouring one.



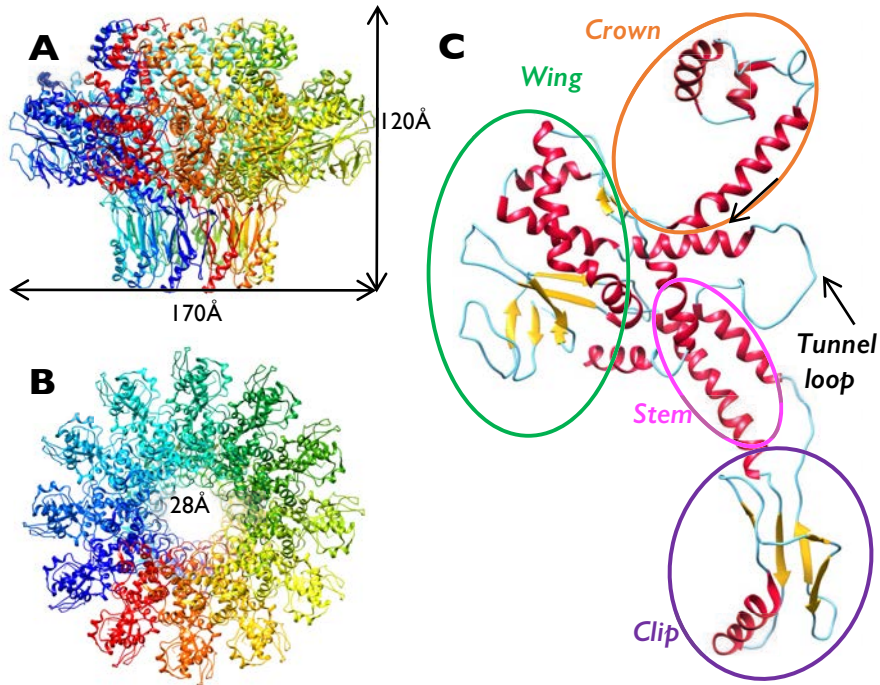
**Figure 1.10** Cartoon representation of P22 portal protein at 3.25Å/7.5Å resolution.

- (A) Lateral view of the dodecamer with dimensions indicated (rainbow colouring per monomer).
- (B) Axial view of the dodecamer with the diameter of the channel indicated (rainbow colouring per monomer).
- (C) Monomeric gpI with the three domains and the tunnel loop indicated.  $\alpha$ -helices appear in red,  $\beta$ -strands in yellow and coils in blue. (Data from *Olia et al.*, 2011.)

Finally, the atomic dodecameric structure of the portal protein of T4 bacteriophage is also available, but in this case it was obtained by high-resolution cryo-EM (*Sun et al.*, 2015). The T4 portal particle has a diameter between 80Å and 170Å, with a length of 120Å (Figure 1.11A). The narrowest diameter of the central channel is 28Å (Figure 1.11B).

Protein domains are equivalent to those from SPPI gp6 (Figure 1.11C).

Contacts between DNA and the protein could occur in three different points of the structure: in the tunnel loop, in a channel loop that connects the wing and the stem, and in an inner clip loop.



**Figure 1.11** Cartoon representation of T4 portal protein at 3.6Å resolution.

(A) Lateral view of the dodecamer with dimensions indicated (rainbow colouring per monomer).

(B) Axial view of the dodecamer with the diameter of the channel indicated (rainbow colouring per monomer).

(C) Monomeric gp20 with the four domains and the tunnel loop indicated.

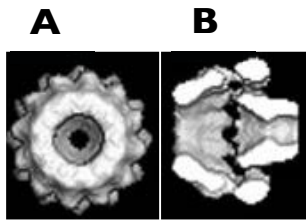
$\alpha$ -helices appear in red,  $\beta$ -strands in yellow and coils in blue.

(Data from Sun *et al.*, 2015.)

There is a conserved central domain found in all the viral portal proteins with two helices and an extended  $\alpha/\beta$  domain. However, when compared, the structures show many different features, for instance a proximal end that varies significantly (Cuervo and Carrascosa, 2012b).



Although there are not atomic resolution structures available, it is known from some low resolution cryo-EM structures that Herpesviruses portal proteins are also ring-shaped dodecamers (Trus *et al.*, 2004; Dittmer and Bogner, 2005). Like the bacteriophage portal proteins, these assemblies have different domains that build axial channel with peripheral flanges (Figure 1.12). Approximate dimensions are similar to those of bacteriophage portal proteins: HCMV portal assembly has a diameter of 170Å, while HSV-I portal assembly is 155Å wide with a height of 130Å.



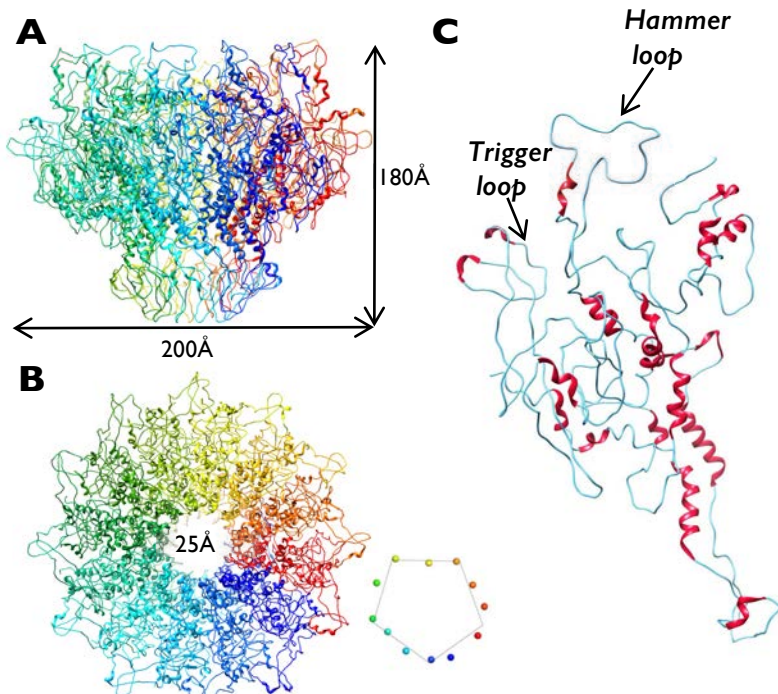
**Figure 1.12** HSV-I portal at 16Å.  
(A) Axial view of the UL6 dodecamer.  
(B) Lateral view of half UL6 dodecamer.  
(Data from Trus *et al.*, 2004.)

#### 1.2.4.2 Structure-function relationship

It has been hypothesized that  $\phi 29$  gp10 would represent the minimum structure able to carry the main portal protein functions: interaction with the terminase proteins, DNA packaging and retention, and interactions with tail proteins. Its central and narrow-end domains, which correspond to the P22 leg and to the stem and clip in SPPI and T4, are the best conserved among portal proteins (Cuervo and Carrascosa, 2012b).

Mutational studies of this region of the SPPI portal protein showed that the change of some residues can inhibit the terminase ATPase activation by the portal protein, without preventing the interaction between both proteins. Thus, these mutations probably avoid necessary portal conformational changes (Oliveira *et al.*, 2006). On the other hand, it has been described that the immobilization of SPPI portal helix  $\alpha 5$  inhibits DNA packaging (Cuervo *et al.*, 2007). The loops that protrude into the channel are also thought to have an important role, as deletion and mutation of charged residues located in the loops of the  $\phi 29$  portal affect DNA retention inside the capsids (Grimes *et al.*, 2011). In the case of T4 portal protein, the tunnel loop is thought to close the channel and stop the DNA from coming out once the capsid has been filled (Sun *et al.*, 2015; Padilla-Sanchez *et al.*, 2014).

Based on structural data of P22 portal protein core in what is thought to be the DNA packaging conformation, it has been suggested that the distal part of the DNA-channel may present a *quasi* 5-fold surface (Figure 1.13). This transient asymmetrical structure might be key for the interaction with the TerL pentamer during DNA translocation, and a conformational change to the 12-fold symmetric oligomer might result in a loss of affinity for TerL and allow the nuclease cleavage of DNA (Lokareddy *et al.*, 2017).



**Figure 1.13** Cartoon representation of P22 portal protein during packaging at 3.3 Å resolution.

(A) Lateral view of the dodecamer with dimensions indicated (rainbow colouring per monomer).

(B) Axial view of the dodecamer with the diameter of the channel indicated (rainbow colouring per monomer). Side chain oxygen atoms of Asn380 are depicted as balls to show the *quasi* 5-fold arrangement of the portal.

(C) Monomeric gpI with the trigger and the hammer loop indicated.  $\alpha$ -helices appear in red and coils in blue.

(Adapted from Lokareddy *et al.*, 2017.)

Regarding the surface charge of the channel, comparison of all available portal structures shows that it is mainly electronegative with some ring areas of positive charges. In some cases, the entrance and the exit of the channel are especially electronegative (Cuervo and Carrascosa, 2012b).

Additional domains present on other portals would be related with other features or functions (Cuervo and Carrascosa, 2012b). Some of the structure-function relationships that have been described so far are the following ones:

- φ29 wide-end: This region would be involved in the connexion to head components, interacting with the scaffolding protein (Fu *et al.*, 2010).
- SPPI crown: The crown domain is related with the incorporation of the protein into the procapsids. Thus, this may be interacting with the major capsid protein and/or the scaffolding protein (Isidro *et al.*, 2004).
- P22 hip: This domain of the portal interacts with the surrounding scaffolding proteins (Chen *et al.*, 2011). It has been proposed that the newly packaged DNA starts the conformational change of the portal by changing the position of the trigger loop, whose conformation during packaging is not compatible with DNA spooling around the portal vertex. A 90° swing of the loop would destabilize the hammer loop, which unfolds and transmits the conformational change to the barrel (Lokareddy *et al.*, 2017).
- P22 barrel: Mutational studies suggest that this domain helps ordering the genome into the capsid during packaging. It acts as a headful sensing valve and also has a role during DNA ejection regulating the delivery pressure (Tang *et al.*, 2011; Moore and Prevelige, 2002). When the capsid is filled with DNA, structural changes on the hip domain are thought to allow the rotation of the barrel domain, which becomes folded. The conformational change would afterwards be transmitted to the leg domain, leading to the symmetrization of the whole particle (Lokareddy *et al.*, 2017).
- T4 wing: Biochemical experiments showed that the N-terminal end of gp20, which is located in the wing domain, is probably involved in the interaction between the portal and the terminase (Dixit *et al.*, 2011).

### 1.2.5 Properties of the DNA packaging motor

In terms of biophysical properties of the DNA translocation process,  $\phi 29$  has been the most extensively studied bacteriophage system. Single molecule experiments demonstrated that packaging complexes are one of the stronger force-generating biological motors known, as they can work against loads of up to 57 pN. The internal force when capsids are full has been calculated around 50 pN (Smith *et al.*, 2011).

The step size has been defined as the number of bp translocated per molecule of ATP hydrolysed. In the case of  $\phi 29$  it is about 2 bp per ATP molecule (Chemla *et al.*, 2005).

The packaging rate decreases as the procapsid is being filled (Smith *et al.*, 2011). It has been suggested that there is some sort of biological sensor that detects the density and conformation of the DNA that has already been packaged, and slows the motor by allosteric regulation of their interactions with ATP. This signal would be transmitted from the inside of the shell to the motor, with the aim of continuously regulating its speed, in response to changes in packaged DNA density or conformation (Berndsen *et al.*, 2015).

The terminase rotates the DNA during packaging, and it has been observed that the rotation per bp increases with capsid filling, while the motor step size decreases. This compensation would preserve motor coordination, allowing one subunit to contact the DNA in a periodic, and specific way. Moreover, when capsids are highly filled, the ATP-binding rate is downregulated and long packaging stops appear (Liu, 2014).

The internal capsid force when DNA is packaged may be available for the initial steps of the ejection process (Smith *et al.*, 2011). However, biochemical experiments done in SPPI showed that pressure itself only ejects 17% of the genome, and an additional force would be needed for the whole genome externalization (São-José *et al.*, 2007).

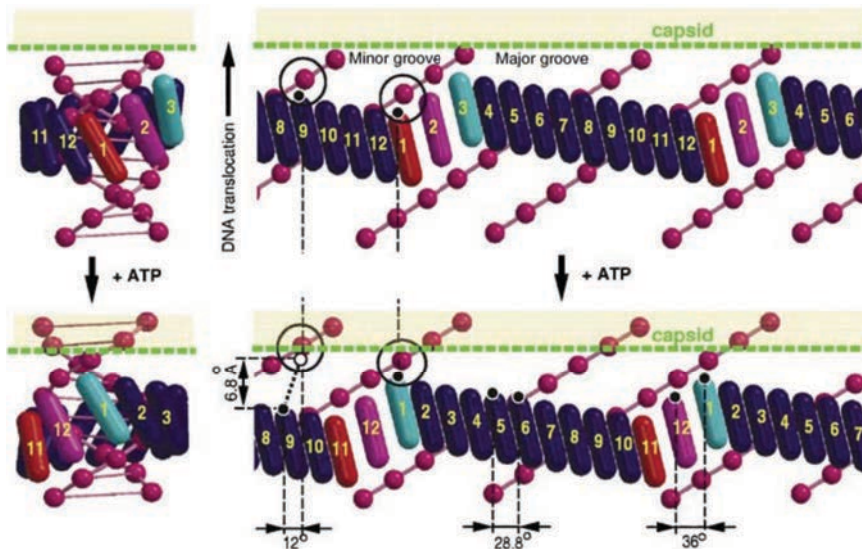
Similar single molecule studies to those described for  $\phi 29$  have been done in the T4 system. In this case, similar forces have been detected, up to 60 pN, what suggests that high force generation is a common property of viral DNA packaging motors. However, the translocation and ATP turnover

rates in this case are much higher, probably in order to make the process more time-efficient, as T4 genome is much bigger. Large dynamic changes in velocity were detected, suggesting multiple active conformational states that would lead to different translocation rates (Fuller *et al.*, 2007).

### **1.2.6 Models for portal protein dsDNA translocation**

There has been a lot of debate regarding how the viral dsDNA molecular motor works. Here there is a summary of the most relevant models that have been proposed, indicating for which specific virus they were described. It is thought that relative rotation of the DNA and the portal protein is necessary for DNA packaging (Lebedev *et al.*, 2007). First, it was suggested that during DNA translocation the portal protein could rotate inside the capsid vertex. In fact, low-energy barriers allow rotation between symmetry mismatching protein rings, which would be the case of the dodecameric portal and the pentameric TerL assemblies that nowadays have been structurally characterized (Hendrix, 1978). Different models were proposed:

- Peristaltic pump ( $\phi$ 29): This model suggests synchronous movements within all the subunits, controlled by the inclination of the channel helices (Simpson *et al.*, 2000).
- Electrostatic interactions ( $\phi$ 29): In this model the portal remains rigid while rotating, and lysine side-chain nitrogen atoms that form rings in the inner part of the central channel interact with the DNA (Guasch *et al.*, 2002).
- Tunnel loops (SPPI): The helical DNA is embraced by an undulating belt of loops that tightly embraces it. During translocation, the rotational symmetry is broken, and the tunnel loops can have three different structural states that ensure engagement of the portal with the DNA at all process stages and that propagate sequentially along the belt (Figure 1.14). On each translocating cycle four events take place: cyclic permutation of loop positions, translocation of 2 bp, 12° rotation of the portal and hydrolysis of one ATP molecule (Lebedev *et al.*, 2007).

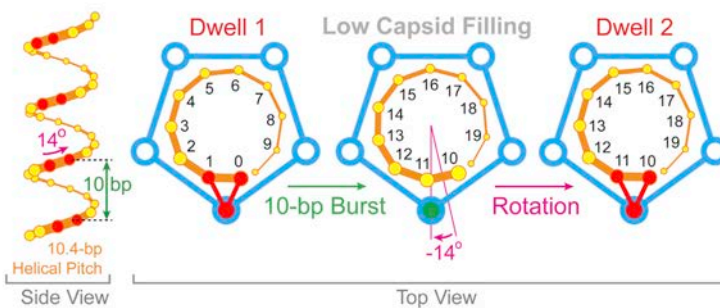


**Figure 1.14** The tunnel loops mechanistic model of DNA translocation. The figure shows the portal-DNA complex before, on the top, and after ATP hydrolysis, on the bottom. The left part shows a three-dimensional model whereas the right figure depicts the slice structure. Cylinders represent tunnel loops and pink spheres DNA phosphates. Two DNA phosphates are circled as reference points. Numbers mark specific tunnel loops, while colours represent different structural states. The three mentioned states required for packaging (magenta, red and blue) propagate along the loop circle. The larger separation between the loops occupying this three states is also drawn, and it is thought to be key to allow them to deep into the major groove. Transition between the two states requires a  $12^\circ$  rotation of the portal relative to the DNA. In the specific cycle depicted, loops 2 to 12 move with respect to the DNA (some example angles are shown in the figure), while loop 1 remains on the same place with respect to the DNA, and moves by  $6.8\text{\AA}$  relative to the capsid together with it. (Taken from Lebedev *et al.*, 2007.)

The tunnel loop links helices  $\alpha 5$  and  $\alpha 6$ , whose relative orientation approximately perpendicular is crucial for the position of the loop in the channel. In the SPPI portal protein structure the lateral chain of V347 tunnel loop residue induces a kink on  $\alpha 6$ . Tunnel loops and TerL are thought to communicate each other during packaging. Signal transduction would imply alterations on the position of  $\alpha 5$  after movements of residues from the tunnel loop and alterations on the kink angle of helix  $\alpha 6$ . Models with an extended  $\alpha 6$  have been computed, but it has not been observed experimentally (Oliveira *et al.*, 2006; Lebedev *et al.*, 2007).

Biochemical experiments on  $\phi 29$  and T4 bacteriophage channel loops suggest that they are not completely essential for DNA translocation, only for DNA retention after packaging (Grimes *et al.*, 2011; Padilla-Sanchez *et al.*, 2014). In spite of that, the available structure of T4 portal protein shows a channel diameter and the presence of a loop that do not discard the tunnel loop mechanism to explain packaging (Sun *et al.*, 2015). Regarding P22, the larger central diameter of its portal channel seemed incompatible with the tunnel loops packaging mechanism (Olia *et al.*, 2011). However, the recent structure published showing the conformation of the protein during packaging conformation has a quite smaller diameter that could agree with the mechanism (Lokareddy *et al.*, 2017).

Rotation of the portal protein with respect to the procapsid was questioned after some biochemical studies (Baumann *et al.*, 2006; Hugel *et al.*, 2007). Nonetheless, rotation of the DNA during packaging was observed in  $\phi 29$ . The DNA rotation would be induced by TerL, in order to produce periodic and specific DNA-protein contacts during packaging (Figure I.15). Dwell phases where ATP binds to the terminase, and burst phases where it is hydrolyzed and converted into mechanical force to translocate DNA, alternate in this packaging model (Liu *et al.*, 2014).



**Figure I.15 Geometric basis for DNA rotation at low capsid filling.**

The side view on the left represents a 5'-3' strand of a B-DNA backbone. The pentameric TerL forms specific contacts with pairs of phosphates (example shown in red) every 10 bp. On the right, top view TerL is represented in blue and the DNA in orange, viewed from inside the capsids. It can be observed that in dwell 1, the same TerL subunit contacts two consecutive phosphates. After a 10 bp burst, a  $14^\circ$  clockwise rotation brings the DNA and the motor again into the necessary relative orientations for dwell 2, contacting the same TerL subunit. At high capsid fillings the model would be similar, but it requires higher DNA rotations, as in the burst step only 9 bp are translocated due to the higher internal capsid pressure.

(Adapted from Liu *et al.*, 2014.)

Therefore, models that propose a relative rotation of portal and DNA would still work, with some adaptations, assuming that the moving part of the process is the DNA and not the protein.

### **1.2.7 Biotechnological and biomedical interest**

Potential biotechnological applications have been described for viral portal proteins, being  $\phi 29$  the best characterized system because of its simple and small structure. Portals could be ideal candidates as biological pores able to work as a nanomachine valve for many applications, from drug controlled loading and release, to DNA delivery (Cuervo and Carrascosa, 2012b). One feature that makes connectors especially interesting for this purpose is that their gating is reversible and might be induced by ligand binding and/or voltage (Geng *et al.*, 2011). Moreover, it is important to mention that the portal has been successfully integrated into artificial lipid bilayers retaining its capability of translocating DNA (Wendell *et al.*, 2009).

Furthermore, some modifications of the protein lead to the formation of arrangements of seven portal rings. These particles could be used in therapeutic treatments as a vehicle for delivering radiopharmaceuticals, therapeutic DNA or enzymatic inhibitors. Other potential application of the particles could also be using them for delivering specific molecules as reporters for *in vivo* diagnostics and imaging (Green *et al.*, 2010).

On the other hand, antiviral agents against human infecting herpesviruses that target assembly steps happening on the nucleus can be designed. Although many molecular aspects of the DNA encapsidation process are still poorly understood, the terminase complex is already a pharmacological target and many efforts have been put into the design of drugs against its activity (Baines, 2011). Letermovir is a compound that targets pUL56 HCMV terminase subunit and which is already into phase III clinical studies (Goldner *et al.*, 2011; Web 2). The portal protein might also be a pharmacological target, either inhibiting a putative movement of the portal protein during DNA translocation within the capsid or against the formation of the complex between the portal protein and the terminase. Although there is no high resolution available of Herpesvirus portal proteins, detailed information about the bacteriophages portal structure and their roles during packaging, including possible conformational changes, could be used as a starting point for this purpose.

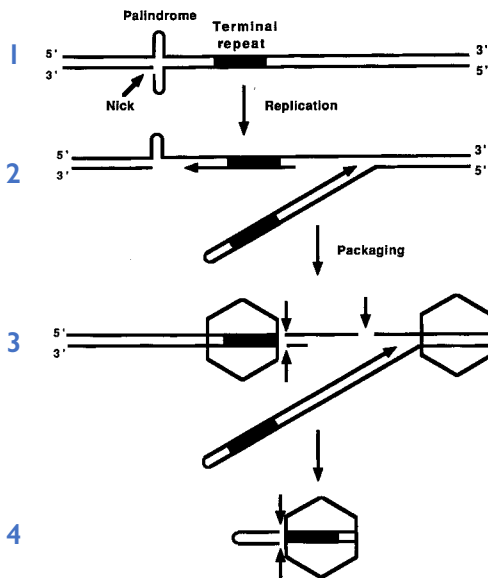


## I.3 The T7 packaging machinery

### I.3.1 DNA processing

As mentioned before, T7 bacteriophage genome replicates forming a concatemeric molecule and, once packaged, presents short exact direct repeated ends that are generated during DNA encapsidation. For duplicating the terminal repeat, a short palindromic hairpin structure formed 190 bp upstream from the left end mature direct repeated end is required (Chung *et al.*, 1990). The process of generation of a mature right end terminal direct repeat can be summarized on the following steps (Figure I.16):

1. **Nicking:** Produced in the palindromic hairpin sequence.
2. **Replication:** A branched concatemer is formed, and the mature right end is created.
3. **Packaging and trimming:** While the capsid is being filled with DNA, the gp3 endonuclease may be involved in trimming the replication forks from the DNA.
4. **Hairpin removal:** Finally, concatemer processing is completed when the mature left end is produced by the removal of the hairpin end.



**Figure I.16**

**T7 concatemers processing and packaging model.** Steps follow the same numbering code as in the text. dsDNA is depicted as two lines, with the representations of the palindromic hairpin and the terminal repeats indicated. Procapsids are drawn as hexagons. Long arrows indicate replicating DNA strands. Small arrows indicate endonuclease cuts. (Adapted from Chung *et al.*, 1990.)

### 1.3.2 The terminase proteins

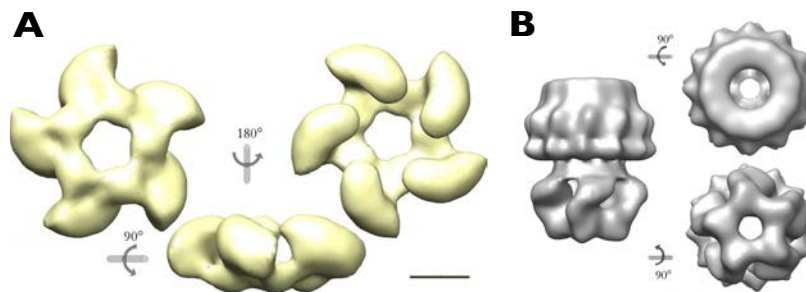
T7 TerS and TerL proteins are, respectively, gp18 and gp19.

#### 1.3.2.1 *The small terminase subunit*

It has been observed that gp18 is essential for DNA packaging in T7 bacteriophage. If the protein is not active, the viral infection gets stuck on the concatemer stage. Gel filtration chromatography after overexpression suggest that the protein is octameric, although this may not be its active state (White and Richardson, 1987).

#### 1.3.2.2 *The large terminase subunit*

The structure of the TerL protein from T7 was determined by negative staining EM, both alone and in complex with the portal. It is a pentameric structure with a central channel, which is wide enough for DNA passing. Coupling between the portal and the terminase leads to the formation of a continuous channel, and the interface between both showed some structural differences that could be related with their interaction. Conformations of TerL are not the same when comparing the isolated protein with the complex (Figure 1.17). The transition between both protein conformations can be achieved by subunit rotation, and could be related with the T7 packaging mechanism (Daudén *et al.*, 2013).



**Figure 1.17**

#### **T7 bacteriophage TerL EM structures.**

(A) Three different orientations of the pentameric gp19 negative staining EM reconstruction at a resolution of 16Å.

(B) Three different orientations of the TerL-portal negative staining EM reconstruction at a resolution of 30Å.

Both scale bars represent 50Å.

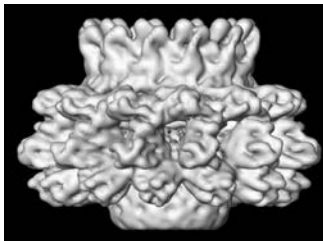
(Adapted from Daudén *et al.*, 2013.)

### 1.3.3 The portal protein

The name of T7 bacteriophage portal protein is gp8. Portal and core proteins are key during procapsid morphogenesis. Their absence produces the reaction between incomplete procapsids to produce non-functional polycapsids or the creation of closed capsid shells (Cerritelli and Studier, 1996a).

It was observed that when overexpressed the protein assemblies were polymorphic, with two distinct populations of twelve and thirteen subunits (Kocsis *et al.*, 1995). Moreover, the distribution of masses determined by scanning transmission electron microscopy is consistent with that (Cerritelli and Studier, 1996b). Gel shift-assays with nucleotides showed that the portal is able to bind linear, circular and supercoiled DNA, while monomers are not. As expected, neither the full-length or the monomeric portals have ATPase activity (Cerritelli and Studier, 1996a).

A cryo-EM reconstruction at 8Å resolution of the dodecameric assembly is available (Agirrezabala *et al.*, 2005a). The overall morphology of the T7 portal is the same as the one of the other portal proteins that have been characterized: a ring assembly with a central channel along its longitudinal axis. However, when this structure is compared with other models of the T7 portal protein in complex with tail or core proteins its morphology is different (Figure 1.18).



**Figure 1.18**  
T7 portal protein cryo-EM structure at 8Å resolution.  
View of the gp8 dodecameric assembly.  
(Adapted from Agirrezabala *et al.*, 2005b.)

Therefore, a high-resolution model would be key to deepen the knowledge about dsDNA packaging in this specific viral system.

## Chapter 2:

## Objectives



The objectives of this thesis were:

- 1.** To solve the 3D high-resolution structure of the T7 bacteriophage portal protein using the available techniques for this type of samples: X-ray crystallography and cryo-EM.
- 2.** To analyze the T7 bacteriophage portal protein structure and compare its features with the other available portal protein structures, in order to distinguish and explain common traits conserved among portals and particularities of the specific T7 bacteriophage viral system.
- 3.** To propose a model for the role of the T7 bacteriophage portal protein during dsDNA packaging.



## Chapter 3:

# Materials and methods





## 3.1 Sample preparation and analysis

This section lists the materials and explains the biochemical methods used for protein sample preparation and analysis before structural studies. All reagents for media and buffer preparation were purchased from BioRad, Fermentas, Fluka, Invitrogen, Merck, New England BioLabs, Panreac, Roche and Sigma.

### 3.1.1 Sample preparation

#### 3.1.1.1 *Cloning*

The *gp8* gene was inserted into the pET28a vector, between the *Nco*I and *Not*I restriction sites with a C-terminal histidine tag (Table 3.1). The protein was expressed fused to the following sequence: AAALHHHHHH. Cloning was done by Francisco J. Fernández.

**Table 3.1** Expression vector. *For protein overexpression.*

<b>Vector name</b>	pET28a
<b>Promoter</b>	T7 promoter
<b>Terminator</b>	T7 terminator
<b>Protein tags</b>	N-terminal histidine tag (HisTag)(x6)/Thrombin/T7 tag C-terminal HisTag (x6)
<b>Antibiotic resistance</b>	Kanamycin
<b>Reference</b>	Novagen

Expected protein parameters were computed with ProtParam (Gasteiger *et al.*, 2005).

#### 3.1.1.2 *Plasmid purification and sequencing*

Plasmid purification was performed using the Qiagen Miniprep kit (Qiagen). A table-top centrifuge was used according to manufacturer's instructions. DNA concentration and purity were checked using a NanoDrop 1000 spectrometer following manufacturer's instructions (ThermoFisher Scientific). Clones were checked by DNA sequencing (Macrogen).

### 3.1.1.3 Bacterial strains

Different *E. coli* strains were used for cloning and expression (Table 3.2).

**Table 3.2** Cell strains. Used during cloning and protein expression.

Cell strain	Genotype	Remarks
DH5 $\alpha$	F <sup>-</sup> $\phi$ 80dlacZ $\Delta$ M15 $\Delta$ (lacZYA-argF)U169 <i>endA1 recA1</i> <i>hsdR17(r<sub>K</sub>-m<sub>K</sub><sup>+</sup>) deoR thi-1 supE44 <math>\lambda</math><sup>-</sup></i> <i>gyrA96 relA1</i>	Strain for general cloning. Invitrogen
BL21(DE3)	B F <sup>-</sup> <i>dcm ompT hsdS</i> (r <sub>B</sub> -m <sub>B</sub> ) <i>gal</i> $\lambda$ (DE3)	Strain for protein expression. BL21-derived with a chromosomal copy of the gene for T7 RNA polymerase. Invitrogen

### 3.1.1.4 Competent cells preparation

Two buffers were prepared and sterilized by filtration:

- **Transformation buffer I (TfbI):** 30 mM KOAc, 100 mM RbCl, 10 mM CaCl<sub>2</sub>, 50 mM MnCl<sub>2</sub>, 3 mM [Co(NH<sub>3</sub>)<sub>6</sub>]Cl<sub>3</sub>, 15% glycerol
- **Transformation buffer II (TfbII):** 10 mM 3-(*N*-morpholino) propanesulfonic acid (MOPS), 75 mM CaCl<sub>2</sub>, 10 mM RbCl, 15% glycerol

*E. coli* cells were grown overnight in LB, at 37°C and with an agitation of 220 rpm. The overnight culture was diluted 1:100, and grown again in the same conditions until the optical density (O.D.) reached a value of 0.25-0.3 (around 2h). After 5 min on ice, cells were centrifuged for 5 min at 4,000 x g and 4°C. Supernatant was discarded and cells were resuspended in five times less volume than the initial of Tfb I. After 5 min on ice, the cells were centrifuged again for 5 min at 4,000 x g and 4°C. The supernatant was discarded and the cells pellet resuspended gently in ten times less volume of Tfb II than the volume used previously of Tfb I. The cells were incubated on ice for 15 min, aliquoted, flash-frozen in liquid nitrogen and stored at -80°C.

### 3.1.1.5 Bacterial transformation

Approximately 50 ng of the pET28a-gp8 plasmid were mixed with 50  $\mu$ l of competent *E. coli* cells. After 30 min of incubation on ice, a 45 s heat shock at 42°C was performed in a water bath. Then, samples were placed back on ice for 2 min. 900  $\mu$ l of LB media were added, and samples were incubated at 37°C with 300 rpm of agitation for 1h. Finally, they were plated on LB-agar supplemented with kanamycin, for resistance selection, and kept at 37°C overnight.

### 3.1.1.6 Culture media

The following medias were prepared for growing *E. coli* bacterial cultures:

- **Plates:** Luria-Bertrani (LB) agar (1% [w/v] tryptone, 0.5% [w/v] yeast extract, 1% [w/v] NaCl, 1.5% [w/v] agar, 0.001M NaOH)
- **Liquid cultures:** LB media (1% [w/v] tryptone, 0.5% [w/v] yeast extract, 1% [w/v] NaCl)
- **Liquid cultures for selenomethionine (SeMet) derivative protein production:** The stock solutions listed below were prepared.
  - Salt solution: 0.16 M  $K_2HPO_4$ , 0.06 M  $KH_2PO_4$ , 0.03 M  $(NH_4)_2SO_4$ , 5 mM  $Na_3C_6H_5O_7$ , 2 mM  $MgSO_4 \cdot 7H_2O$
  - Glucose solution: 2 M glucose (sterilized by filtration)
  - Amino acids solution: Dissolve at 60-80°C with stirring and adjust pH at 7.5 (sterilized by filtration)
    - Ala: 0.2 mg/ml
    - Arg: 0.2 mg/ml
    - Asn: 0.2 mg/ml
    - Asp: 0.2 mg/ml
    - Cys: 0.2 mg/ml
    - Gln: 0.2 mg/ml
    - Glu: 0.2 mg/ml
    - His: 0.2 mg/ml
    - Ile: 0.5 mg/ml
    - Leu: 0.5 mg/ml
    - Lys: 0.5 mg/ml
    - Phe: 0.5 mg/ml

- Pro: 0.2 mg/ml
- Ser: 0.2 mg/ml
- Thr: 0.5 mg/ml
- Trp: 0.2 mg/ml
- Tyr: 0.2 mg/ml
- Val: 0.5 mg/ml
- SeMet solution: 10 mg/ml
- Thiamine solution: 4 mg/ml (sterilized by filtration)
- Thymine solution: 4 mg/ml (sterilized by filtration)

Prepare the following mixture for a liter of culture:

- 200 ml of salt solution
- 200 ml of amino acids solution
- 16 ml of glucose solution
- 8 ml of thiamine solution
- 8 ml of thymine solution
- 5 ml of SeMet solution

Media were autoclaved before using, except where sterilizing filtration is indicated. All the media were supplemented with kanamycin antibiotic:

- Stock solution: 50 mg/ml (in MilliQ water)
- Working concentration: 50 µg/ml

### 3.1.1.7 Protein expression

Native protein expression was performed in BL21(DE3) *E. coli* cells. Precultures were grown in 500 ml flasks with 100 ml of LB. One colony of transformed bacteria and kanamycin were added to the flask, which was incubated overnight at 37°C at 220 rpm of agitation speed. The following day, the expression cultures were grown in two liter flasks containing 500 ml of LB supplemented with kanamycin. Preculture was used as inoculum (4 ml per flask). When an O.D. of 0.6 was reached, expression was induced adding 0.4 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG). After the addition of IPTG, cultures were kept for 3h at 37°C and 220 rpm of agitation speed. Then, cells were centrifuged at 5,000 × g for 20 min at 4°C. The cell pellet was frozen at -20°C.

For expression of the SeMet derivative protein some changes were introduced to the protocol. It is possible to suppress methionine biosynthesis in BL21(DE3) when grown in a minimal media with certain

amino acids. Precultures were grown in regular LB media, and dilution was performed as before, but in the methionine biosynthesis suppressive media. After IPTG induction cultures were kept overnight at 37°C and 220 rpm of agitation speed.

### 3.1.1.8 Protein electrophoresis

To analyze protein samples, denaturing protein acrylamide gel electrophoresis in presence of sodium dodecyl sulphate (SDS-PAGE) was performed. Gels were prepared using the following recipe:

- **Separative gel:** 10% [w/v] acrylamide, 0.27% [w/v] bis-acrylamide, 0.1% [w/v] SDS in 0.38 M Tris-HCl (pH 8.8). Polymerization in presence of 0.5% [w/v] initiator ammonium persulfate (APS) and 0.05% [v/v] crosslinking reagent N, N, N', N'-tetramethylethylenediamine (TEMED).
- **Stacking gel:** 5% [w/v] acrylamide, 0.13% [w/v] bis-acrylamide, 0.1% [w/v] SDS in 0.13 M Tris-HCl (pH 6.8). Polymerization in presence of 0.75% [w/v] APS and 0.125% [v/v] TEMED.

SDS-PAGE was carried out using BioRad electrophoresis tanks with the following running buffer: 25 mM Tris-HCl, 0.2 M glycine and 0.1% [w/v] SDS. BioRad power sources were used to run the gel applying a 200V current. Protein samples were diluted to be in the following loading buffer: bromophenol blue in 25 mM Tris-HCl (pH 6.8), 5% [w/v] SDS, 10% [v/v] glycerol and 5% [v/v]  $\beta$ -mercaptoethanol. After that, they were boiled for 10 min. MW marker SeeBlue Pre-Stained Standard (Invitrogen) was used.

Finally, the following solutions were employed to stain and destain the gels:

- **Coomasie staining solution:** 0.25% [w/v] Coomasie Blue R-250 (Sigma) in 10% [v/v] isopropanol, 10% [v/v] acetic acid
- **Coomasie destain solution:** 10% [v/v] isopropanol, 10% [v/v] acetic acid

### 3.1.1.9 Protein purification

Protein purification was done by a three-step chromatographic protocol on ÄKTA Purifier systems using the following columns, according to manufacturer instructions (GE Healthcare):

- HisTrap HP (5 ml)
- Sephacryl S-400 (16/60)
- Superose 6 (10/300)

The gp8 protein purification protocol is detailed below:

- 1. Immobilized metal ion affinity chromatography (IMAC):** Bacterial pellets coming from the protein expression cultures were resuspended in lysis buffer: 500 mM NaCl, 20 mM imidazole, 3 mM  $\beta$ -mercaptoethanol and 20 mM Tris-HCl (pH 8.0) supplemented with 40  $\mu$ g/ml of DNase I and Complete Protease Inhibitor Cocktail Tablets (Roche). The cells were lysed using a cell disruptor (PECF Constant Systems Ltd.) operated at 20 kpsi and centrifuged at 30,000  $\times$  g for 30 min. After that, the supernatant was filtered and loaded on an equilibrated HisTrap column. The column was washed with 10 column volumes (CV) of binding buffer: 500 mM NaCl, 20 mM imidazole, 3 mM  $\beta$ -mercaptoethanol, 20 mM Tris-HCl (pH 8.0). The protein was eluted in a linear gradient of 20 CV from 20 mM to 350 mM imidazole. The elution buffer was the following: 500 mM NaCl, 350 mM imidazole, 3 mM  $\beta$ -mercaptoethanol, 20 mM Tris-HCl (pH 8.0). A 5 CV wash was done with elution buffer and a 5 CV reequilibration with binding buffer.
- 2. Gel filtration (Sephacryl S-400):** Eluted protein fractions from the affinity column were concentrated and loaded into the first gel filtration column. Protein was eluted with 1.2 CV of elution buffer: 500 mM NaCl, 5 mM dithiothreitol (DTT), 2 mM ethylenediaminetetraacetic acid (EDTA), 20 mM Tris-HCl (pH 8.0).

3. **Gel filtration (Superose 6):** The central fractions of the peak eluted from the Sephacryl S-400 column were concentrated and loaded into the second gel filtration column. Protein was eluted with 1.2 CV of the same elution buffer as in the previous column. MW standards were used on this step according to manufacturer instructions (GE Healthcare).

Before loading samples into any gel filtration column protein samples were either filtrated or centrifuged at 16,000 × g for 15 min at 4°C. Protein purity was analyzed after each step observing the gel filtration chromatogram and the SDS-PAGE gel.

#### *3.1.1.10 Protein concentration and quantification*

After purification, protein samples were concentrated, quantified and freshly used for further techniques.

For protein concentration Vivaspin devices (GE Healthcare) of 6 ml and 20 ml with 30,000 Da MW cut-off were employed, according to manufacturer's instructions. Protein concentration was measured using the Bradford protein quantification assay (Bradford, 1976). A reagent dye is required for the assay (BioRad).

### **3.1.2 Sample analysis**

#### *3.1.2.1 Mass spectrometry analysis*

Mass spectrometry (MS) analysis was performed in the Proteomics Platform of Barcelona Science Park. A nanoAcquity liquid chromatographer (Waters) and a LTA-Orbitrap Velos (Thermo Scientific) mass spectrometer were used.

SeMet proteins were analyzed by MS in order to confirm the proper incorporation of the modified amino acid into the sample. Proteins were in-gel digested with trypsin and then analyzed by liquid chromatography tandem-MS (LC-MS/MS).



### 3.1.2.2 *Dynamic light scattering*

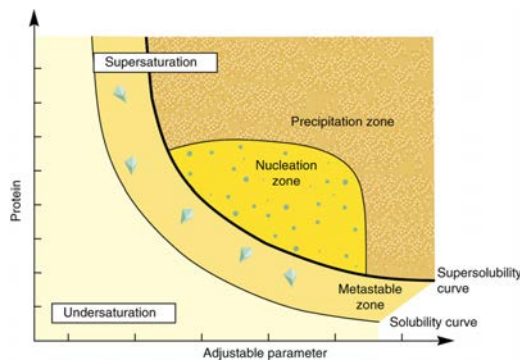
A Zetasizer Nano ZS from the Automated Crystallography Platform of Barcelona Science Park was used, according to manufacturer's instructions (Malvern).

Dynamic light scattering (DLS) is a useful technique in structural biology to calculate the size and size-distribution of solution samples in the submicron region. Therefore, it can be used to check the aggregation state of one sample and/or its homogeneity (Stetefeld *et al.*, 2016).

## 3.2 Crystallization and X-ray diffraction analysis

Up to now, X-ray crystallography is the technique that has given more high-resolution information by far in the structural biology field. The process of obtaining an atomic structure is summarized in the steps below (Egli, 2016 and references cited therein):

1. **Sample preparation:** Large amounts of pure material are required. Molecules should be well structured and not floppy.
2. **Crystallization:** Although there are many crystallization protocols, vapor diffusion technique (both in hanging-drop and sitting-drop), is one of the most common techniques, and the one used in this project. The principle of crystallization relies on keeping on a same closed environment the drop with a certain precipitant and a reservoir with the same chemical at a higher concentration. As water diffuses from the drop to the reservoir the concentration of the precipitant on the drop increases and slowly lowers the protein solubility. Protein ideally changes from an unsaturated phase diagram region to a labile, supersaturated area where spontaneous nucleation occurs. If the system remains in the metastable zone crystals grow (Figure 3.1).



**Figure 3.1 Protein crystallization phase diagram.**

Protein concentration is represented on the y axis, while the adjustable parameter on the x axis can be for instance the precipitant concentration.

(Taken from Khurshid *et al.*, 2014.)

Crystallization is still a trial and error approach, and therefore requires extensive initial screenings of conditions.

3. **Data collection:** Crystals are shot with X-rays. Diffraction patterns corresponding to the X-rays that have been scattered by the sample electrons are collected. Nowadays, this process is usually done at synchrotrons, which emit intense X-rays in a tangential manner respect to a closed circle where accelerated electron beams are travelling. Crystals are cryo-cooled with nitrogen to minimize radiation damage and crystal desiccation, during storage (liquid nitrogen) and data collection (nitrogen gas). Cryo-protectants are required for avoiding the crystallization of solvent molecules, which would interfere with protein diffraction.
4. **Data processing:** Reflection spots in each frame are indexed, crystals and detector parameters are refined, and diffraction peaks are integrated. A relative scale between measurements is established, parameters are refined using the total dataset, and the frames are merged. Statistical analysis of the reflections is done to evaluate the dataset.
5. **Phasing:** During data processing, the amplitude of structure factors ( $F$ ) is calculated, but alone it is insufficient for building a model, because phases of the reflections are also required. There are five basic phasing strategies: single and multiple isomorphous replacement (SIR and MIR), single- and multi-wavelength anomalous dispersion (SAD and MAD), a combination of both, named SIR and MIR with anomalous scattering (SIRAS and MIRAS), molecular replacement (MR) and direct methods. The first three are experimental techniques, while the last ones are done *in silico*. In this project, MR, SAD and MAD have been tried (Taylor, 2010 and references cited therein):
  - **MR:** Requires a similar model structure to calculate the initial phases, usually with a sequence identity above 25%. A Patterson map of interatomic vectors is calculated from the experimental data and from the model. Both maps are first rotated and then translated to correctly locate the search model with respect to the origin of the new unit cell. Initial phases are calculated from the resulting location.
  - **SAD and MAD:** Require derivative crystals with heavy atoms introduced, in order to measure the effect on diffraction of the anomalous scattering of an atom, at certain wavelengths. The atomic scattering factor contains three components: the

normal scattering term  $f_0$ , which depends on the scattering angle, and two wavelength dependent terms,  $f'$  and  $f''$  (respectively the dispersive term and the absorption term).  $f'$  and  $f''$  represent the anomalous scattering at the absorption edge, when an electron is promoted from an inner shell by X-ray energy.  $f'$  is the derivative of  $f''$ . At the synchrotron, absorption curves can be determined experimentally by a fluorescence scan. For MAD phasing, at least two different data are collected at the following points:

- $\lambda_1$ : Absorption peak at ( $f''$  maximum)
- $\lambda_2$ : Inflection point of  $f''$  ( $f'$  minimum)
- $\lambda_3$ : Remote wavelength to maximize dispersive difference to  $\lambda_2$

Data from only one wavelength, normally at  $\lambda_1$ , may be enough for SAD phasing.

6. **Refinement:** During refinement, changes are applied to the coordinates ( $x$ ,  $y$  and  $z$ ) and temperature  $B$ -factors from atoms of the model, in order to reduce the difference between calculated and observed amplitudes. It is an iterative process of manual building and fitting, automatic optimization according to X-ray data, and geometric constrains, and electron density map calculation from the improved model. The  $R$ -work value is used as a guide: between 20% and 30% is usually acceptable for a final model (the higher the resolution, the lower the  $R$ -work is expected). The  $R$ -free value is used as an independent measure of the quality of the fit, as it comes from a test dataset of reflections not included in the refinement. It will be higher than the  $R$ -work, but differences above 5% may indicate over-refinement or errors. Sum electron density maps ( $2F_{\text{obs}}-F_{\text{calc}}$ ) and difference density maps ( $F_{\text{obs}}-F_{\text{calc}}$ ) are observed to refine: the first one should look like the model and the second indicates missing and/or misplaced atoms.
7. **Validation and analysis:** Ramachandran plots are useful to detect problematic areas where backbone torsion areas deviate from the expected ones. During model analysis, it is important to take into consideration those flexible parts, such us the N-terminal and the C-terminal ends, are often not visible. Zones with high  $B$ -factors are also

related with flexible areas. On the other hand, it is important to bear in mind that crystal packing forces may have an effect on the structure of the macromolecules.

This section lists the materials and the methods from protein crystallization to data processing and preliminary analysis. Methods and materials for subsequent steps are explained on section 3.4.

### 3.2.1 Crystallization and X-ray data collection

#### 3.2.1.1 *Protein crystallization screening*

Screening experiments were performed at the Automated Crystallography Platform of the IBMB/IRB at the Barcelona Science Park (Table 3.3).

**Table 3.3** Crystal screenings used.

*Names of the screens and commercial screens in which they are based.*

Name	Screen	Conditions	Reference
PAC1	Crystal Screen I	48	Hampton Research
	Crystal Screen II	48	
PAC2	Wizard I	48	Emerald Bio
	Wizard II	48	
PAC3	Index	96	Hampton Research
PAC5	A/S Ion Screen	48	Hampton Research
	Ammonium sulphate	24	
	Quick phosphate	24	
PAC6	PEG6000	24	Hampton Research
	PEG6000/LiCl	24	
	PEG 400	24	
	PEG 4000/LiCl	24	
PAC9	Natrix	48	Hampton Research
	Complex screen	40	
PAC21	PACT premier HT-96	96	Molecular Dimensions
PAC22	Pi-PEG Screen	96	Jena Bioscience
PAC23	Pi – minimal screen	96	Jena Bioscience
PACPlus	JCSG-Plus	96	Jena Bioscience
PACTOP 96	TOP 96	96	Anatrace Microlytic

96-well sitting drop MRC plates were used for the screenings (Molecular Dimensions). Reservoirs contained 100  $\mu$ l of reservoir solution. Drops were prepared mixing 100 nl of protein and 100 nl of reservoir solution.

Reservoirs of the crystallization plates were prepared with a Freedom EVO robot (TECAN). Crystallization drops were afterwards set up using a Cartesian dispensing robot for microscale liquid handling for high-throughput crystal screening (Cartesian Technologies). Screenings plates were incubated afterwards both at 20°C and 4°C.

This step is necessary to screen many putative crystallization conditions in order to identify promising ones.

### *3.2.1.2 Protein crystallization optimization*

Once promising conditions showing small crystals were identified, the following step consists on optimizing them. By screening around the condition, crystal size and shape can be improved. For instance, salt and precipitant concentrations can be varied around the starting condition, as well as pH.

The best conditions identified on the screenings were optimized on 24-well hanging drop plates (Jena Biosciences). The reservoir contained 1 ml of the crystallization condition. Crystallization optimizations were performed mixing manually 1  $\mu$ l of protein and 1  $\mu$ l of reservoir condition.

### *3.2.1.3 Crystal mounting and freezing*

Protein crystals were fished using nylon cryo-loops (Molecular Dimensions). Fished crystals were soaked in reservoir solutions with increasing amounts of cryo-protectant before being flashed frozen into liquid nitrogen.

Cryo-protectants must be optimized for each crystallization condition, and were checked at the Automated Crystallography Platform of the Barcelona Science Park, before going to the synchrotron.

#### 3.2.1.4 *Derivative-crystals for experimental phasing*

Two different strategies have been used to introduce heavy atoms in the crystals for experimental phasing:

- **SeMet:** This strategy is based on introducing modified amino acids with heavy atoms in the culture media. They are incorporated into the overexpressed protein. SeMet are methionines with the sulphur substituted by selenium.
- **Heavy atoms:** In this case heavy atoms are introduced during protein crystallization, and not during its expression. This can be done by adding them directly to the crystallization solution or soaking the crystals once they appear. In our case soakings with different concentrations and incubation times were tried: from 0.4 mM to 1 mM, and from 3 h to 20 h. The following heavy atom clusters were purchased and tested (Jena Biosciences):
  - Metatungstate:  $\text{Na}_6[\text{H}_2\text{W}_{12}\text{O}_{40}] \times \text{H}_2\text{O}$
  - Paratungstate:  $(\text{NH}_4)_{10}[\text{H}_2\text{W}_{12}\text{O}_{42}] \times \text{H}_2\text{O}$
  - Phosphotungstate:  $\text{Na}_3[\text{PW}_{12}\text{O}_{40}] \times \text{H}_2\text{O}$
  - Tantalum bromide:  $[\text{Ta}_6\text{Br}_{12}]^{2+} \times 2 \text{ Br}^-$

#### 3.2.1.5 *X-ray data collection*

X-ray data collection was done in the ALBA synchrotron (Cerdanyola del Vallès, Spain) at the tunable beamline XALOC and in the European Synchrotron Radiation Facility (ESRF, Grenoble, France) at the ID23-2 and ID30A-3 microfocus beamlines.

### 3.2.2 X-ray diffraction analysis

#### 3.2.2.1 *Data processing and analysis*

Diffraction data were indexed and integrated with XDS, and afterwards scaled, reduced and merged using XSCALE (Kabsch *et al.*, 2010). Solution trials and X-ray data analysis detailed on the following page were carried out using the CCP4 suite of crystallographic programs (Potterton *et al.*, 2003).

### 3.2.2.2 *Matthews coefficient calculation*

The Matthews coefficient ( $V_M$ ) corresponds to the crystal volume per unit of MW (Matthews, 1968). Based on a survey of different parameters of the available crystallographic structures, it allows the estimation of the number of molecules per asymmetric unit during the first steps of macromolecule structure solution (Kantardjieff and Rupp, 2003). To estimate the number of molecules per asymmetric unit the `matthews_coef` program was used (Matthews, 1968).

### 3.2.2.3 *Calculation of the self-rotation function*

Intramolecular vectors for each molecular orientation found in the crystallographic unit cell are contained in Patterson functions. The auto-correlation function of a Patterson function with a rotated version of itself is called self-rotation function (SRF).

Multi-subunit proteins often present local symmetry or non-crystallographic symmetry (NCS). In these cases, SRF are useful to determine which is the order and orientation of local symmetry axes (Rossmann and Blow, 1962).

To calculate the SFR the MOLREP program was employed (Vagin and Teplyakov, 2010).

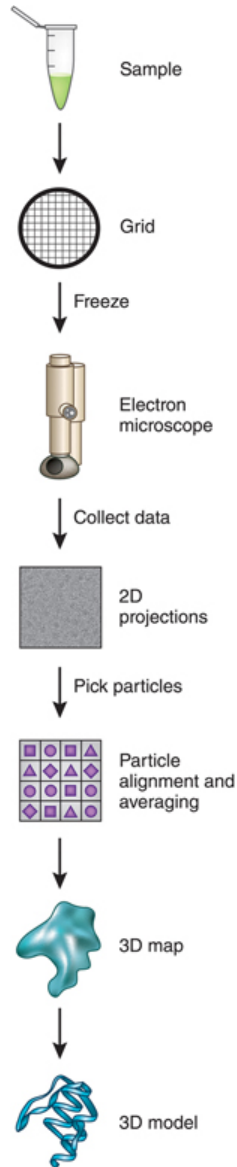


## 3.3 Cryo-EM studies

Single-particle cryo-EM technique has lived a resolution revolution in the recent years (Nogales, 2016 and references cited therein). This technique is based on the fact that by collecting many 2D projections of a macromolecule in different directions, a 3D volume of it can be reconstructed computationally. Nowadays, it is possible to obtain structures at resolution that allow to build atomic models into the maps. However, the process of obtaining a cryo-EM structure is still long (Doerr, 2016). It can be summarized on the following steps (Figure 3.2):

1. **Sample preparation:** A purified sample of the interest protein or complex is required.
2. **Vitrification:** The sample is applied to a grid that usually contains holes in a carbon film, which is supported on a metal frame. The grid is plunged frozen into a cryogen (for instance liquid ethane) and flash-frozen trapping the particles into a thin layer of vitreous ice. Ideally, particles in all the orientations should be present. Vitreous ice prevents evaporation in the microscope high-vacuum conditions and protects the sample from radiation damage.
3. **Data collection:** 2D images are collected on 200 or 300 kV electron microscopes using low electron doses to avoid damages.
4. **Particle picking:** Because of the low electron doses images are too noisy to obtain high-resolution information. It is important to have enough number of particles to improve the signal. The particles have to be individually picked from the micrographs.
5. **Alignment and averaging:** Individual particles are subsequently aligned, classified and averaged in order to obtain the best set of data, to proceed with the volume reconstruction.
6. **Calculation of a 3D map:** Image-processing programs are able to align the images of the different views and merge the data and calculate a 3D map. This map can be iteratively refined and validated using software tools.

7. **Calculation of a 3D model:** On the last step, a protein model is built into the 3D map.



**Figure 3.2 Single-particle cryo-EM workflow.**  
Overall process, from initial sample to atomic model  
(Taken from *Doerr, 2016.*)

The most relevant recent technical advances that lead to the resolution improvement are related to instrumentation and software (Nogales, 2016 and references cited therein):

- **Development of direct electron detectors:** They have limited noise, high contrast and preserve high-resolution information. Moreover, they present a faster read-out. Instead of collecting micrographs nowadays movies are recorded. To overcome the problem of the beam-induced motion, which introduces blurring in the images, the total dose is divided in small doses, and many frames are collected. Computational programs can be used afterwards to correctly align the frames and reduce the blurring effect.
- **Advances on the processing programs:** Specific software has been developed and adapted to the new type of data in order to improve all the steps of data processing. The most relevant feature is the appearance of user-friendly programs able to deal with heterogeneity in the samples, such as RELION. This type of software classifies heterogeneous datasets in structurally homogeneous subsets, and reconstructs independently high-resolution structures of each of them.

Moreover, the combination of the need of fewer particles to build high resolution structures due to new detectors and the automation of data collection and processing, leads to a reduction in the time required for solving a structure.

In summary, cryo-EM present some advantages when it is compared with other structural techniques, such as X-ray crystallography, from a methodological point of view:

- **Amount of sample:** Lower amount of sample is required for cryo-EM than for crystallization.
- **Crystallization bottleneck:** Crystallization, which is not always a straight-forward process, is not necessary.

- **Heterogeneity:** Cryo-EM can be used with samples in which multiple conformations or compositions coexist. This is particularly interesting, because it allows the study of conformational transitions, which gives a deeper biological understanding of protein function and mechanism.

For these reasons, although difficult samples such as large complexes, integral membrane proteins, polymers or macromolecular assemblies with different compositions and/or conformations can be studied with X-ray crystallography, cryo-EM appears as an interesting alternative.

This section lists the materials and explains the methods for the single particle cryo-EM studies that have been done in this project. Model building and refinement are detailed in section 3.4.

### **3.3.1 Preparation of the grids and data collection**

#### *3.3.1.1 Preparation of holey grids with thin carbon backing*

Copper R2/2 holey carbon grids (Quantifoil) were used both in negative staining and cryo-EM experiments. In some cases, they were covered with an extra thin carbon layer (Grassucci *et al.*, 2007).

#### *3.3.1.2 Negative staining*

Negative staining was performed to characterize the sample and determine the concentration of protein required for starting vitrification optimization. In this case, grids with an extra thin carbon layer were used. Hydrophilic grids surfaces were obtained using glow discharger Emitech K100X (Quorum Technologies). After grids were glow discharged for 15 s, they were incubated on top of a 5  $\mu$ l protein drop for 1 min. Excess of liquid was dried by the grid laterally with an absorption paper. The grid was washed 3 times by leaving it on top of a 100  $\mu$ l drop of MilliQ water and immediately drying it with an absorption paper. Finally, the grid was left for 1 minute in a 1% [w/v] uranyl acetate solution and dried again. After drying them at room temperature for at least 10 min, grids were observed on a 100 kV JEM-1011 microscope (JEOL). The microscope was equipped with a Erlangshen ES100W CCD camera (Gatan Inc.).

### 3.3.1.3 Vitrification optimization

Vitrification was performed with a Vitrobot Mark IV (FEI) according to manufacturer's instructions (Figure 3.3). The Vitrobot instrument automates the vitrification process, in order to provide fast and reproducible sample preparation conditions. It works at constant physical and mechanical conditions, in terms of temperature, relative humidity, blotting conditions and freezing velocity. In our case, the Vitrobot was operated always at 95% of relative humidity.



**Figure 3.3** FEI Vitrobot.  
Used for sample vitrification.  
(Taken from *Web 3*.)

The overall process consists on the following steps:

1. Preparation of the ethane, which is cooled using liquid nitrogen
2. Glow discharge of the grids
3. Insertion of the grid into the chamber
4. Incubation of the grid with the sample
5. Blotting to remove the excess of liquid
6. Plunge freezing in liquid ethane
7. Transfer of the grid to liquid nitrogen, where it is stored

The following variables were tested during vitrification optimization:

- Extra carbon layer: Present (could be clear or dark) or not present
- Glow discharge time: 15 s - 1 min
- Temperature: 10°C - 22°C
- Protein concentration: 0.07 mg/ml - 0.2 mg/ml (with carbon);  
0.6 mg/ml - 1.1 mg/ml (without carbon)
- Time of sample incubation: 1 min - 3 min
- Blotting force: (-10) – (+5)
- Blotting time: 2.5 s - 4.5 s

Grids were properly stored in liquid nitrogen until they were checked with a 200 kV Tecnai F20 microscope (FEI) equipped with an Eagle CCD camera (Gatan Inc.).

#### 3.3.1.4 Data collection

Cryo-EM data were collected from an optimized vitrified grid on a 200 kV Talos Arctica equipped with a Falcon II direct electron detector from FEI (Figure 3.4). According to the map of the entire grid, which is called atlas, suitable squares and holes were manually selected for automated data collection. Before starting data collection, the beam was aligned and astigmatism and drift were checked. The dose during data collection was of 15.2 e-/Å<sup>2</sup>s. The pixel size was of 1.37Å/px and the spherical aberration of 2.7 mm. Images were collected at a magnification rate of 73,000X.



**Figure 3.4** FEI Talos Arctica.  
Used for data collection.  
(Taken from *Web 3*.)

### 3.3.2 Cryo-EM data processing

All cryo-EM data processing was performed using Scipion, a software framework that integrates many programs needed during cryo-EM structure solution (de la Rosa-Trevín *et al.*, 2016).

#### 3.3.2.1 Movie alignment

Frames from the movies were aligned using MotionCor2 (Zheng *et al.*, 2017). This algorithm is able to correct anisotropic image motion, at a single pixel level, across the whole frame. The software combines iterative patch-based motion detection with temporal and spatial constraints and

dose weighting. It is able to work in a wide range of datasets, in terms of defocus or short integration times.

### 3.3.2.2 *Contrast transfer function correction*

The contrast transfer function (CTF) describes how aberrations modify the image of a sample in an electron microscope. Biological samples have very low amplitude contrast, because electrons interact weakly with the light atoms. To solve this problem, phase contrast can be generated by defocusing the microscope. If we consider the recorded image as a CTF-degraded true object, CTF correction allows the true object to be reverse-engineered. Therefore, CTF correction is vital to obtain high resolution structures.

Some available programs directly discard the images that do not have good enough CTF and show, for instance, drift or astigmatism problems. `xmipp3` - `ctf` estimation and `ctffind4` were used for that purpose (de la Rosa-Trevín *et al.*, 2013; Rohou and Grigorieff, 2015). Apart from the data automatically eliminated by software, some data quality minima were fixed, and the following images were eliminated:

- `_xmipp3_ctfCritPsdCorr90` below 0.80, to avoid drifted images
- `_defocusRatio` above 1.05, to avoid astigmatic images

### 3.3.2.3 *Particle picking*

The following software was used to pick the particles (de la Rosa-Trevín *et al.*, 2013):

- **xmipp3 - manual-picking (step 1):** Particle picking can be done either manually or with supervised picking support. Initially, all the picking has to be manual. Afterwards, during supervised picking, the program is able to pick the particles itself, but manual corrections are expected, in order to polish the picking algorithm and continue with the second step.
- **xmipp3 - auto-picking (step 2):** Particle picking is done automatically using the previous training done in step 1.

#### 3.3.2.4 *Initial volume*

An initial volume was calculated using the `xmipp3 - ransac` program, which is able to compute an initial 3D model starting from a set of projections or classes (de la Rosa-Trevín *et al.*, 2013). Filtered versions of the initial volume are used during 3D classifications and reconstructions.

#### 3.3.2.5 *Classification of the particles*

Particles were extensively classified in many rounds of 2D classifications, in order to select the best subset of particles to reconstruct the protein volume. Two different programs were used: `xmipp3 - cl2d` and `relion - 2D` classification (de la Rosa-Trevín *et al.*, 2013; Scheres, 2012).

`xmipp3 - cl2d` was used during the initial classification cycles to remove the worse particles, because it is faster. RELION is now the reference software in particle classification, which uses a Bayesian approach to infer parameters of a statistical model from the data.

3D classifications were performed with `relion - 3D` (Scheres, 2012).

#### 3.3.2.6 *Volume reconstruction*

The 3D map was refined using RELION (Scheres, 2012). `relion - auto-refine` includes a gold-standard Fourier shell correlation procedure, which is intended to prevent overfitting and stop refinement when necessary. Therefore, the program is able to yield high-quality reconstructions and reliable resolution estimates. The resulting volume was treated with `relion - post-processing`, which is a protocol that performs automated masking, estimates overfitting, modulation transfer function (related with contrast and resolution) and *B*-factor sharpening.

#### 3.3.2.7 *Calculation of local resolution*

Local resolution of the cryo-EM map was evaluated with `resmap - local resolution` (Kucukelbir *et al.*, 2014).



## 3.4 Structure determination

On this section, there is a list of the materials and methods employed for the resolution of the atomic structure of the T7 bacteriophage portal protein, combining data from cryo-EM and X-ray crystallography.

### 3.4.1 Structure solution and refinement

#### 3.4.1.1 *Preliminar model building and refinement*

Coot was used for interpretation of the cryo-EM map and for building of a preliminar monomeric model for the portal protein (Emsley and Kowtan, 2004). The oligomeric model was created with the help of Joan Pous, obtaining a rotation matrix from the cryo-EM volume. Finally, the model was refined using Phenix real-space refinement (Afonine *et al.*, 2013).

#### 3.4.1.2 *Previous crystallographic data*

Some crystallographic data of the portal protein was previously available in the lab. For these data, sample preparation and crystallization were done by Rosa Pérez-Luque, X-ray data collection was performed by Francisco J. Fernández and Miquel Coll and data processing was done by Cristina Vega.

#### 3.4.1.3 *Crystallographic analysis*

$V_M$  and SRF were analyzed with the programs described in section 3.2.

#### 3.4.1.4 *Molecular replacement*

MR against crystallographic data was performed to obtain the initial phases using PHASER (McCoy *et al.*, 2007). The model built into the cryo-EM map was used as initial search ensemble.

#### 3.4.1.5 *Density modification and phase extension*

Density modifications (DM) were performed using the DM program (Cowtan, 1994). DM apply real space constraints to a density map to improve the phases.

Constraints can be based on various known features of the protein electron density map. Three different protocols were applied:

- **Solvent flattening:** Density in solvent regions is almost constant.
- **Histogram mapping:** Protein maps have similar density histograms.
- **NCS-averaging:** Very similar densities should be found in NCS related regions.

DM is normally applied using first low-resolution data and subsequently applied to higher resolution information in what is called a phase extension protocol.

With the help of Joan Pous, rotation matrices were calculated and masks for solvent flattening and NCS-averaging were generated from the experimental EM volume. Many trials with several variations in the setting parameters were done until finding the best conditions.

#### *3.4.1.6 Crystallographic model building and structure refinement*

REFMAC5 was employed for structure refinement (Murshudov *et al.*, 1997). Rigid body refinement was used first, and then restrained refinement was performed. NCS restraints were progressively relaxed. TLS refinement has also been applied, which defines groups of atoms as rigid bodies and allows to model their anisotropic displacements at medium to low resolution. Coot was used for interpretation of the electron density and model building. NCS maps were calculated with Coot, and were very helpful during the process.

#### *3.4.1.7 Model refinement into the cryo-EM volume*

REFMAC from the CCP-EM package was used for structure refinement (Brown *et al.*, 2015; Burnley *et al.*, 2017). This REFMAC version was modified for optimal refinement of atomic models into cryo-EM maps. Coot was used for visualization of maps, model building and rigid body refinement in real space.

## 3.4.2 Model validation and analysis

### 3.4.2.1 *Model validation*

MolProbity is a structure-validation web service that evaluates model quality at global and local levels (Chen *et al.*, 2010). It is based on the analysis of covalent-geometry, torsion-angle, hydrogen placement and all-atom contact analysis. It was first used for crystallographic models, but nowadays it is also used for validation of cryo-EM structures.

RAMPAGE has been used to obtain Ramachandran plots (Lovell *et al.*, 2003).

### 3.4.2.2 *Structure visualization and analysis*

Model structure visualization and elaboration of figures were performed with Coot and UCSF-Chimera (Pettersen *et al.*, 2004). ENDscript was used to analyse the structure, giving information about secondary structure, accessibility, hydropathy and both crystallographic and non-crystallographic protein-protein contacts between subunits (Robert and Gouet, 2014). Dali and MATRAS were used for structure comparison with known structures (Holm and Rosenström, 2010; Kawabata, 2003).

## Chapter 4:

# Results and discussion

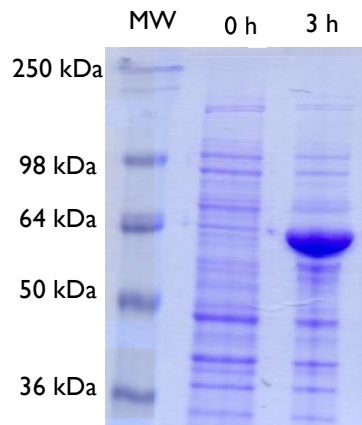


## 4.1 Sample preparation and analysis

### 4.1.1 Protein expression

gp8 was overexpressed in *E. coli* with a C-terminal HisTag, with a total monomeric MW of 60.4 kDa and a predicted theoretical isoelectric point of 4.85.

SDS-PAGE analysis of *E. coli* culture samples before and after induction showed the overexpression of a protein of the expected MW (Figure 4.1).



**Figure 4.1** SDS-PAGE analysis of portal protein expression.

MW: MW marker (MW of each band is indicated on the left column of the figure).

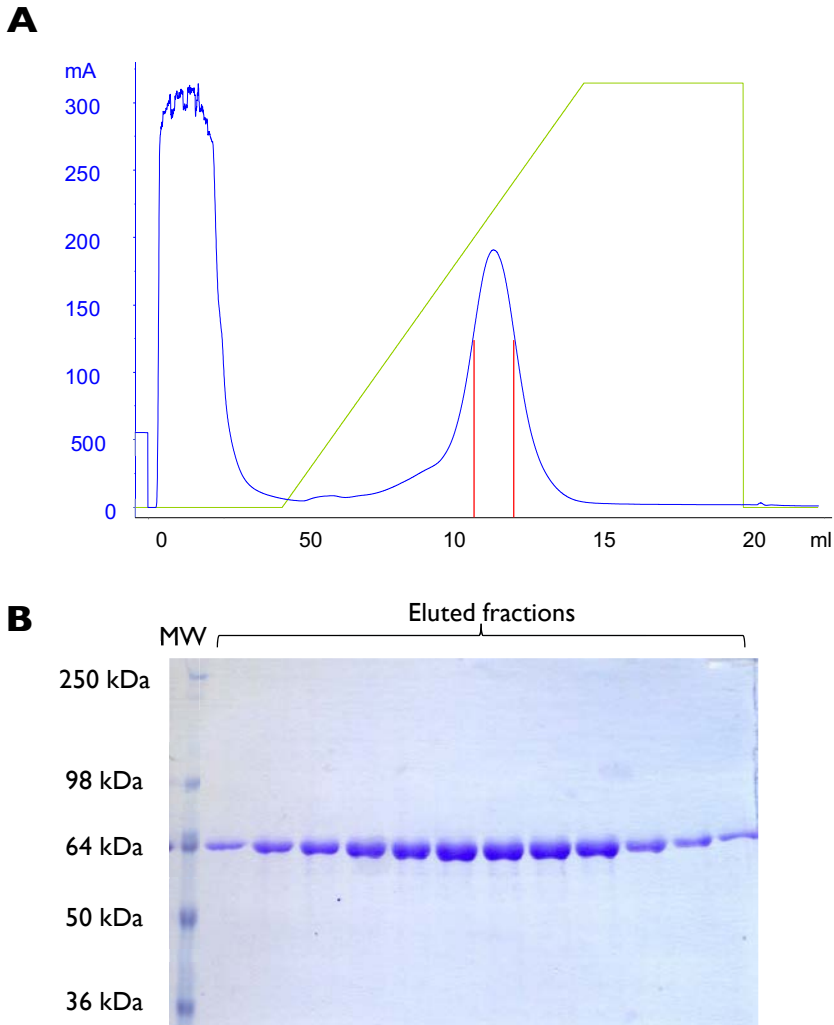
0 h: Sample of cells at the induction moment.

3 h: Sample of cells after induction time.

The overexpression of the T7 bacteriophage portal protein presented a high yield. This high yield might be favored because of expressing the protein in conditions that are close to the *in vivo* ones, with *E. coli* as expression system and at 37°C of temperature. Protein expression results were identical both for the native and the SeMet derivative samples, with yields equally high, but *E. coli* cultures required longer time for growing in the case of the SeMet. This is because the medium used for the cultures was not as rich as LB. However, with longer expression times results were the same.

## 4.1.2 Protein purification

The first step during protein purification was a HisTrap IMAC, which gave a unique peak (Figure 4.2A).



**Figure 4.2 HisTrap chromatogram and SDS-PAGE analysis.**

(A) Chromatogram: x axis corresponds to elution volume in ml and y axis to absorbance at 280 nm in mAU. Blue line represents UV absorbance at 280 nm. Green line represents percentage of elution buffer, from 0% to 100%. The area of the peak loaded on the SDS-PAGE is delimited by red vertical lines.

(B) SDS-PAGE: MW marker on the first lane and eluted fractions from the part of the peak indicated in the chromatogram.

Protein binds to the nickel resin of the chromatography column and elutes at an approximate imidazole concentration of 240 mM. Chromatograms and SDS-PAGE analysis showed that the sample that elutes from the IMAC column is already pure and quite concentrated.

Initially, a two-step chromatography protocol was performed, with first an IMAC and then a Superose 6 column for gel filtration. However, it was published that in the case of the purification of the T4 bacteriophage portal protein, the addition of a Sephacryl S-400 column between the other two was crucial for improving the homogeneity of the sample (Sun *et al.*, 2015). According to the cited data, by loading on the Superose 6 only the central fractions of the Sephacryl S-400 elution peak, samples were enriched in dodecameric assemblies, reducing the proportion of undecameric and tridecameric portals. Even though this step reduces the total yield of the purification, it was added to the purification protocol for the T7 portal protein to improve the quality of the final sample.

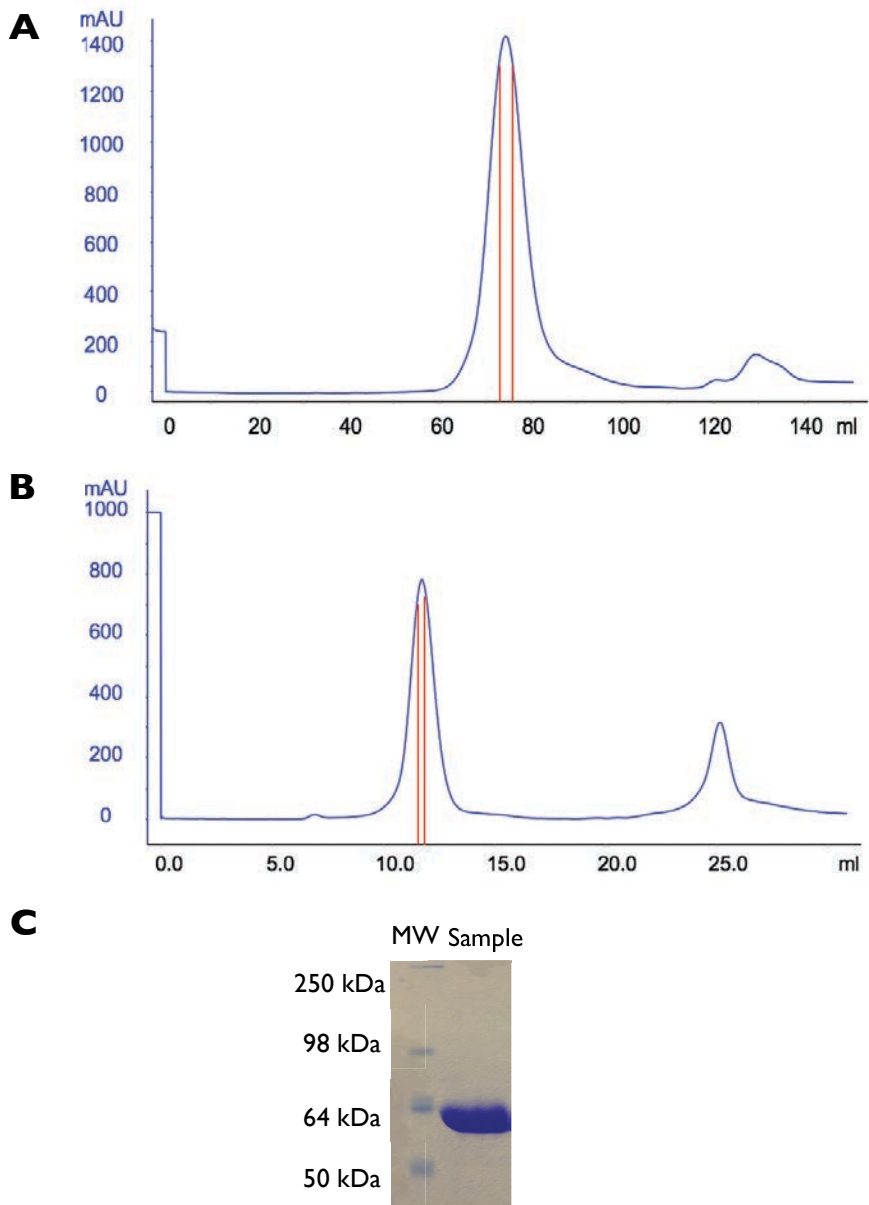
Samples like the ones shown on Figure 4.2B were joined and concentrated to 5 ml to be loaded on a Sephacryl S-400 column. This first size-exclusion chromatography gave a unique peak at 76 ml, and the central 2 ml were concentrated and loaded on the Superose 6 column (Figure 4.3A).

The second size-exclusion chromatography showed also a unique peak at 11.86 ml. One ml corresponding to its central area was concentrated, quantified and used for further techniques and structural studies (Figure 4.3B and 4.3C).

Both gel filtration steps gave as a result unique peaks, which is a good sign of the quality of the sample (Figure 4.3). Samples could be concentrated without aggregation problems up to 16 mg/ml. Different concentrations below this value were used depending on the experimental technique.

There were no differences between native and SeMet proteins in terms of purification.





**Figure 4.3** Size-exclusion chromatograms and SDS-PAGE analysis.

(A) Sephacryl S-400 column: x axis corresponds to elution volume in ml and y axis to UV absorbance at 280 nm in mAU. Red vertical lines indicate the area of the peak loaded on the Superose 6 column.

(B) Superose 6 column: x axis corresponds to elution volume in ml and y axis to absorbance at 280 nm in mAU. Red vertical lines indicate the area of the peak concentrated and used for further studies.

(C) SDS-PAGE: MW markers and sample after size-exclusion chromatographies.

### 4.1.3 Sample characterization

MW standards for calibration of size exclusion columns were run on the Superose 6 column (Table 4.1).

**Table 4.1** MW calibration standards run on the Superose 6 column. Name of the standard, MW and elution volume.

Protein standard	MW (kDa)	Elution volume (ml)
Blue dextran	2,000	7.21
Thyroglobulin	669	11.55
Ferritin	440	13.79
Aldolase	158	15.24
Conalbumin	75	15.68
Ovoalbumin	44	16.63

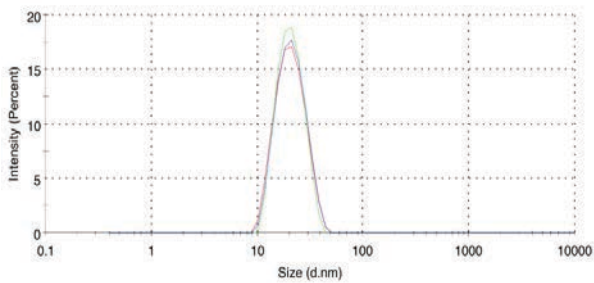
According to the results, an approximate MW of 771 kDa can be calculated for the T7 portal protein sample, in the expected range. A dodecamer would have a total MW of 724.8 kDa, while a tridecamer would be of 786.2 kDa. Therefore, the experimental MW would correspond to 12.8 monomers per particle. It is important to take into account that this type of measurements for such big complexes are not exact, and do not allow to extract clear conclusions. Therefore, portal proteins in the sample could be dodecameric, tridecameric, a mixture of both, or even mixtures of other oligomers of similar order.

SeMet pure samples were analyzed by LC-MS/MS in order to check the presence of the derivative amino acid. Results confirmed that SeMet had incorporated properly during protein expression.

13 mg/ml fresh pure protein samples gave a DLS monodisperse profile. An estimated diameter of the particles of 212Å was obtained (Figure 4.4). Regarding the DLS analysis, two main ideas can be extracted:

- The first one, is that the approximate diameter of the T7 portal particles, of 212Å, seems to be of the same order as the ones of the other portal proteins whose structure is known. The diameter is slightly bigger, but this could be due to the experimental error of the technique.

- On the other hand, the monodisperse behavior of the sample is a good sign for indicating that it is suitable for structural characterization.



**Figure 4.4 DLS size distribution.**

Three independent recordings appear in blue, green and red. Size is represented in a logarithmic scale of the diameter in nm, while intensity is a percentage.

In summary, it seems that both native and SeMet sample are pure, homogeneous, and present the expected characteristics when analyzed by DLS and size-exclusion chromatography.

## 4.2 Crystallization and X-ray diffraction analysis

### 4.2.1 Crystallization and X-ray diffraction

Protein crystallization screenings yielded many conditions with crystals, of different shapes and sizes, and in a wide range of protein concentrations, from 1 mg/ml to 16 mg/ml. However, many of them did not diffract at all and only one hexagonal crystal form diffracted at a reasonable resolution.

The initial screening condition in which hexagonal crystals appeared was the following one, which corresponds to the 8<sup>th</sup> tube from Wizard II screening:

- 0.2 M NaCl
- 0.1 M Na/K phosphate pH 6.2
- 10% [w/v] polyethyleneglycol (PEG) 8000

The screening plate was placed at 20°C. Optimization was performed varying the following parameters:

- NaCl concentration: 0.1 M to 0.3 M
- pH: 6 to 7
- PEG concentration: 5% to 17.5%
- Protein concentration: 1 mg/ml to 16 mg/ml
- Temperature 4°C and 20°C
- Volume of the drops (in  $\mu\text{l}$  of protein + reservoir): 1+1, 1+2, 2+1, 2+2



**Figure 4.5 Optimized hexagonal crystal.**

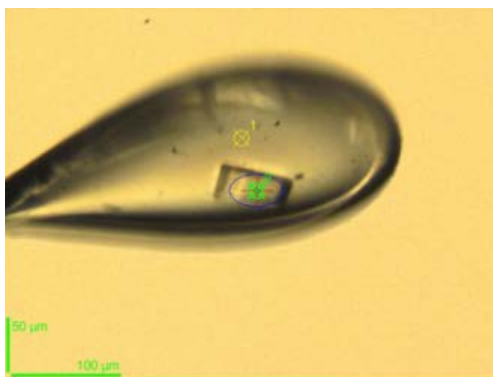
Crystals appeared around five days after setting up the drops. This crystal appeared at 0.2 M NaCl, 0.1 M Na/K phosphate pH 6.2 and 10% [w/v] PEG 8000.

Optimized crystals were around  $50 \times 50 \times 30 \mu\text{m}$  big and had well-defined sharp edges.

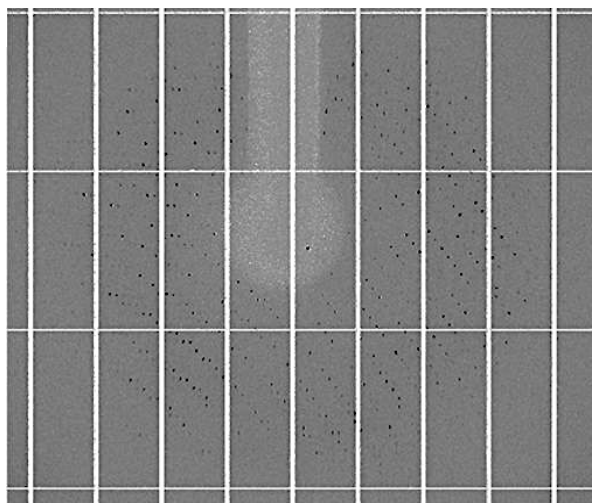
Several crystals were cryo-protected to test their quality diffraction at the synchrotron. Different cryo-buffers were optimized at the Automated Crystallography Platform, and the following were used during cryo-protection as they did not show any ice rings:

- 30% glycerol
- 30% ethylene glycol
- 35% [w/v] PEG 8000

**A**



**B**



**Figure 4.6** Diffraction of hexagonal crystals at ALBA synchrotron.

- (A) Centering of an hexagonal crystal, which appears on side-view. Dimensions are indicated in the left bottom corner in  $\mu\text{m}$ .  
(B) Example of diffraction pattern collected.

Although many crystal forms were tested both at ESRF and ALBA only hexagonal crystals diffracted at resolutions above 4Å (Figure 4.6). Only one out of several crystals presented a diffraction pattern and only one complete dataset could be collected at XALOC beamline in ALBA synchrotron, corresponding to the following optimization condition and cryo-protected with 30% glycerol:

- 0.2 M NaCl
- 0.1 M Na/K phosphate pH 6.2
- 11% [w/v] PEG 8000
- 4 mg/ml of protein, 1  $\mu$ l + 1  $\mu$ l

#### 4.2.2 X-ray data processing analysis

Data was processed with XDS (Table 4.2).

**Table 4.2 Hexagonal crystals XDS data processing.** *Crystallographic processing parameters. Outer shell parameters are indicated between parenthesis.*

Parameters	Values
Wavelength (Å)	0.97949
Resolution range (Å)	45.27 – 3.80 (4.03 – 3.80)
Space group	P6 <sub>3</sub> 22
Unit cell	a=b=245.85Å c=241.00Å $\alpha=\beta=90^\circ$ $\gamma=120^\circ$
Total reflections	628,888 (53,001)
Unique reflections	42,290 (6,321)
Multiplicity	14.8 (6.2)
Completeness (%)	98.6 (93.4)
Mean I/ $\sigma$ (I)	7.63 (0.3)
R-meas	0.31 (7.16)
CC <sub>1/2</sub>	99.7 (10)

From now on, this dataset will be named P6<sub>3</sub>22. Diffraction data was cut at 3.8Å, and although some parameters in the outer shell of P6<sub>3</sub>22 dataset are far from standard, it is not strange to be less strict in terms of statistics in the case of crystals of large complexes, which tend to have big unit cells and to diffract poorly. High local symmetry averaging can also compensate for the poor diffraction at high resolution.

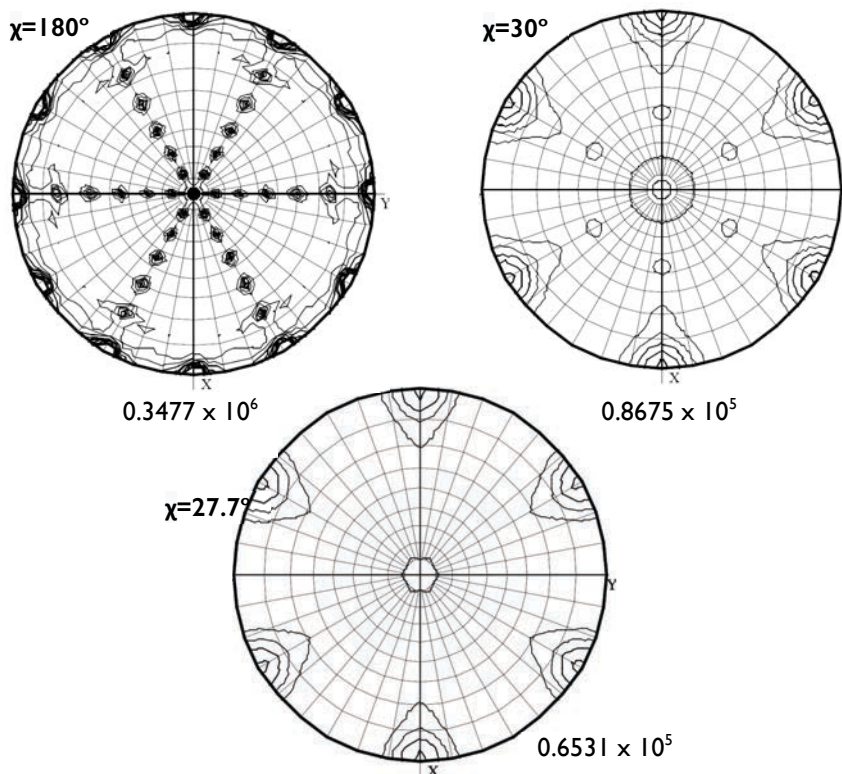
Asymmetric unit content analysis by calculation of the  $V_M$  is shown in Table 4.3, which gives the most probable number of monomers in the asymmetric unit. 6, 7 or 8 monomers per asymmetric unit would be possible with solvent percentage ranging from 43% to 57%. It is reasonable to think that 6 may be a more feasible option, as it would correspond to half connector assembly, and one of the binary crystallographic axis could reconstitute the whole ring. Moreover, it has been described that crystals of big assemblies that have an internal channel may have high solvent contents. That would also be in agreement with the hypothesis of 6 monomers per asymmetric unit, which corresponds to an hypothetical solvent content of 57.56%.

**Table 4.3**  $P6_322$   $V_M$  analysis results. Probabilities,  $V_M$  and percentages of solvent associated to each number of hypothetic monomers per asymmetric unit.

Number of monomers per asymmetric unit	$V_M$	Percentage of solvent	Probability
4	4.35	71.75	0.02
5	3.48	64.69	0.07
6	2.90	57.56	0.20
7	2.49	50.50	0.35
8	2.18	43.50	0.28
9	1.93	36.43	0.07

The SRF gives information about the local symmetry present in the crystals. Figure 4.7 shows the SRF of this dataset at sections  $\chi=180^\circ$ ,  $\chi=30^\circ$  and  $\chi=27.7^\circ$  and its peaks and maxima. It seems that the portal rings present in the  $P6_322$  crystal are dodecameric with the 12-fold axis in the  $ab$  plane and coincident with a crystallographic two-fold axis:

- Tridecamers are less probable as the intensity of the peaks at  $\chi=30^\circ$  is higher than at  $\chi=27.7^\circ$ . Peaks at  $30^\circ$  account for dodecameric symmetry, while at  $27.7^\circ$  would correspond to tridecameric symmetry.
- Moreover, as expected, there are 12 peaks corresponding to 2-fold symmetries at  $\chi=180^\circ$ , in a perpendicular manner with respect to the dodecameric axes.



**Figure 4.7** P<sub>6</sub><sub>3</sub>22 self-rotation function sections.  
 Maximum values are indicated on each section.  
 $\chi=180^\circ$ ,  $\chi=30^\circ$  and  $\chi=27.7^\circ$  sections are shown.

Therefore, the hypothesis of the asymmetric unit containing six monomers is confirmed by the SRF results.

## 4.2.3 Structure solution trials

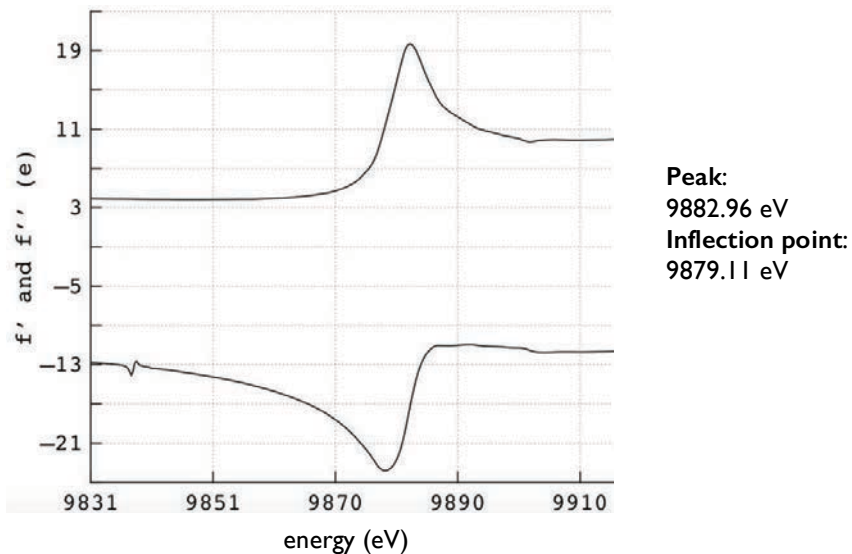
### 4.2.3.1 *Experimental phasing*

Two different strategies were tried:

- **SeMet:** Many crystallization conditions were identified, but none of them gave diffraction above 15Å. Hexagonal crystals of SeMet protein never appeared, although extensive optimizations were performed around the condition of native protein crystallization.



- **Heavy atoms:** Hexagonal P<sub>6</sub><sub>3</sub>22 crystals were soaked with heavy atom clusters to try to phase the data (metatungstate, paratungstate, phosphotungstate and tantalum bromide). Fluorescence scanning at the synchrotron suggested that they had been properly incorporated into the crystal lattice (Figure 4.8). However, none of the tungstate derivative crystals diffracted. Only tantalum bromide soaked crystals diffracted, but below 8Å.



**Figure 4.8** Scanning of a tantalum bromide derivative crystal.  $f'$  and  $f''$  are shown at the region of the tantalum L-III edge. Energies corresponding to the peak and the inflection point are indicated.

In summary, all the experimental phasing trials failed because of the difficulty of obtaining well-diffracting crystals. SeMet derivative protein does not produce well-diffracting hexagonal crystals either. On the other hand, soaking with heavy atom clusters seems to allow their incorporation into the crystal, but derivatization seems to hamper the already poor diffraction.

#### 4.2.3.2 MR

Different models were used for MR trials, both the available structures of other portals and threading models based on them. The available T7 portal protein cryo-EM map was also tried as a model. However, none of the trials was successful, only solutions with low Z-scores and non-interpretable electron density maps were obtained.

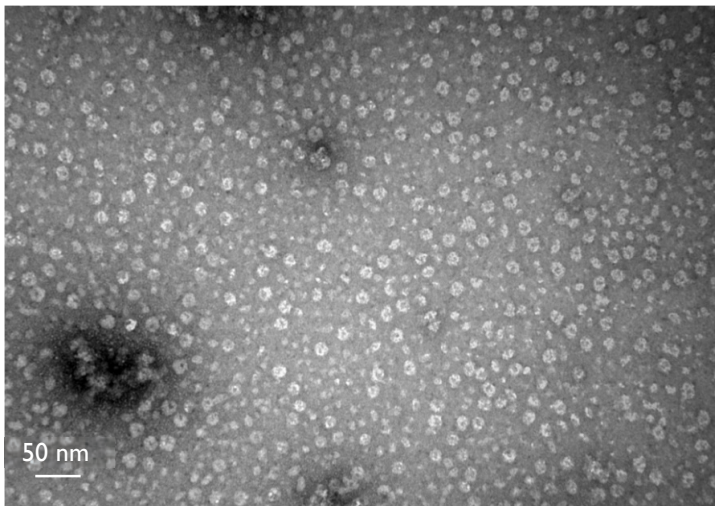
Failure of MR trials is not surprising due to the low sequence identity between the different portal proteins, below 15%. As it has been detailed in the introduction, although portal proteins have some common structural features, there are considerable differences between them, not only in details but also in terms of subdomain and secondary structure architecture.

Regarding the low-resolution cryo-EM map previously available for the T7 portal protein, it was not accurate enough to allow the phasing of the data.

## 4.3 Cryo-EM studies

### 4.3.1 Negative staining

Negative staining images showed the expected ring-shaped assemblies with a central channel. Samples with different concentrations were visualized. The protein does not seem to be aggregated, therefore single particle cryo-EM studies seemed feasible with this sample.

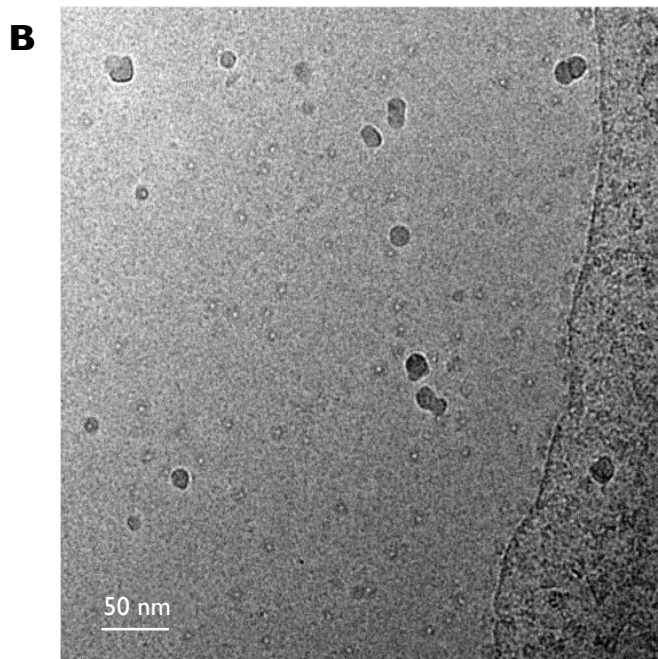
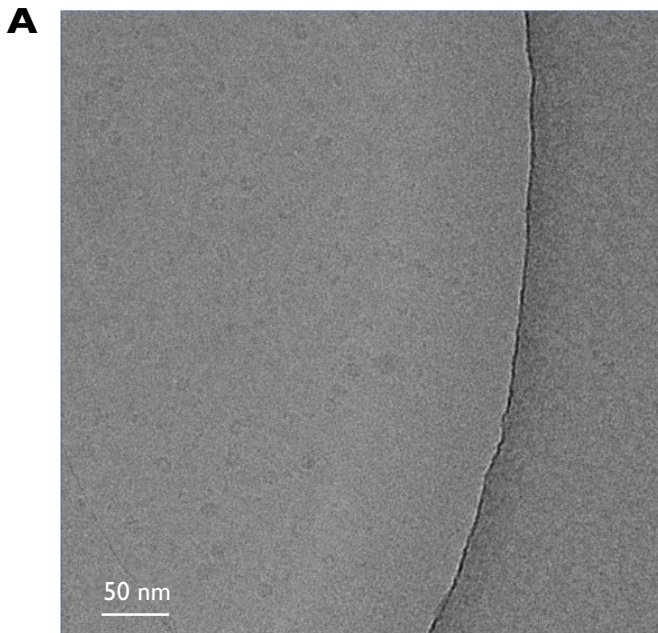


**Figure 4.9 Negative staining.** Ring-shaped complexes observed by negative staining in a 100 kV JEOL JEM-1011 microscope operated at a magnification of 60K. Sample concentration was 0.2 mg/ml. Scale barr indicates 50 nm.

### 4.3.2 Vitrification

Negative staining images suggested that a concentration of 0.2 mg/ml may be a good starting point for vitrification optimization, when working with grids with an extra carbon layer.

During vitrification optimization the key point was the type of grid used. Avoiding extra layers of carbon, images improved considerably in terms of background and contrast (Figure 4.10). However, the protein concentration had to be increased from about 0.2 mg/ml to 2.5 mg/ml or 3 mg/ml. As the sample can be expressed and purified with good yields having to work with higher protein concentrations was not a problem.



**Figure 4.10** Vitrification optimization.

(A) Grid with an extra carbon layer.

(B) Grid without an extra carbon layer.

Images taken with a 200 kV Tecnai F20 at 110K magnification.

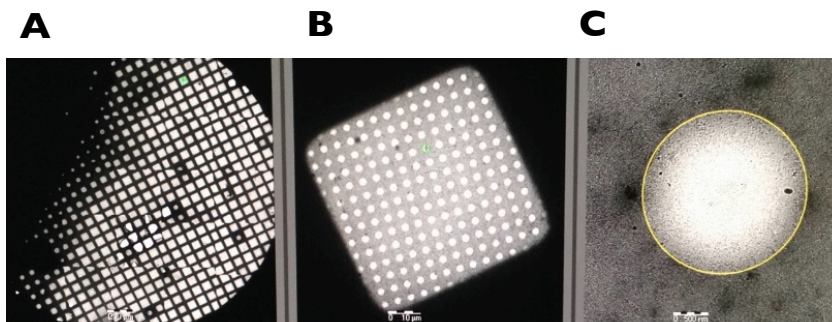
Scale bars indicate 50 nm.

After many rounds of optimization of the vitrification procedure the following conditions were determined to provide the best grids for data collection using a Vitrobot:

- Grids: without carbon after 1 min of glow discharge
- Incubation: with a 3 mg/ml sample for 3 min at 10°C and 95% humidity
- Blotting force: -3
- Blotting time: 3.5 s

### 4.3.3 Cryo-EM data collection

During data collection 1,065 movies could be recorded from one grid, which presented a good ice distribution, with only a small proportion of the surface covered by thick ice (Figure 4.11). Inspection of the grid allowed the selection of enough squares and holes to collect data that once processed, it would give a high-resolution structure. Each movie contained 26 frames.



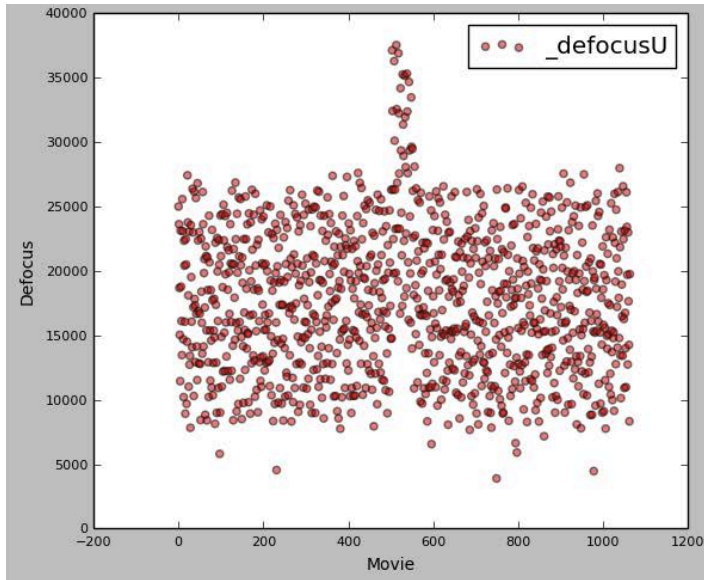
**Figure 4.11** Grid, square and holes

- (A) General view of the grid, with a square marked in green.  
(B) Zoom of the selected square on (A) with a marked hole in green.  
(C) Zoom of the selected hole on (B).

### 4.3.4 CTF estimation and particle picking

After movie alignment, CTF correction and particle picking were done in parallel:

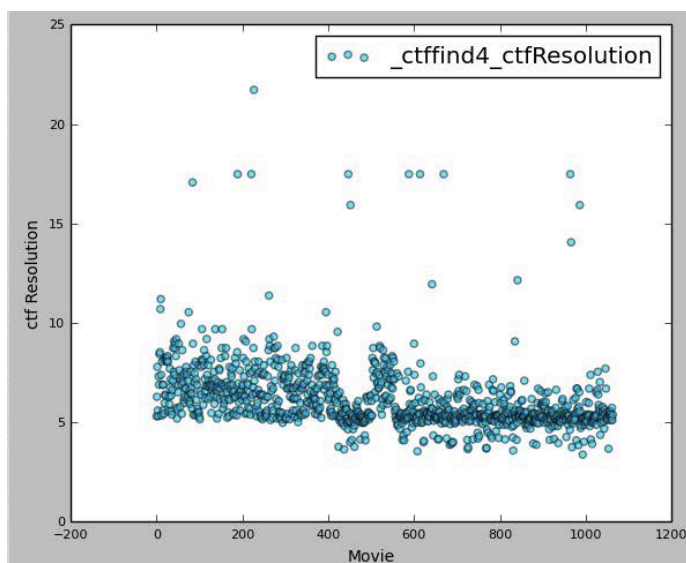
- **CTF estimation:** Defocus and resolution from each movie were estimated. As required for 3D reconstruction, movies were recorded at different defocus values. Only a small proportion of micrographs (around movie number 500) presented higher defocus than expected, probably due to a change on the thickness of the ice in a specific part of a square (Figure 4.12).



**Figure 4.12** Calculated defocus by `ctffind4`.  
x axis corresponds to the movie number.  
y axis corresponds to defocus in Å.

Regarding the calculated resolution of the micrographs, most of them were around  $5\text{Å}$ , although the last ones seemed to have slightly better calculated resolutions, around  $3\text{Å}$  (Figure 4.13).

The main problem encountered during CTF estimation was the relatively high number of movies presenting drift problems. The frames of these movies could not be aligned and the output micrograph remained blurred (Figure 4.14). When the power spectra of these specific micrographs were checked the presence of vanishing Thon rings in one direction was evident.



**Figure 4.13** Calculated resolution by ctffind4.

x axis corresponds to the movie number.

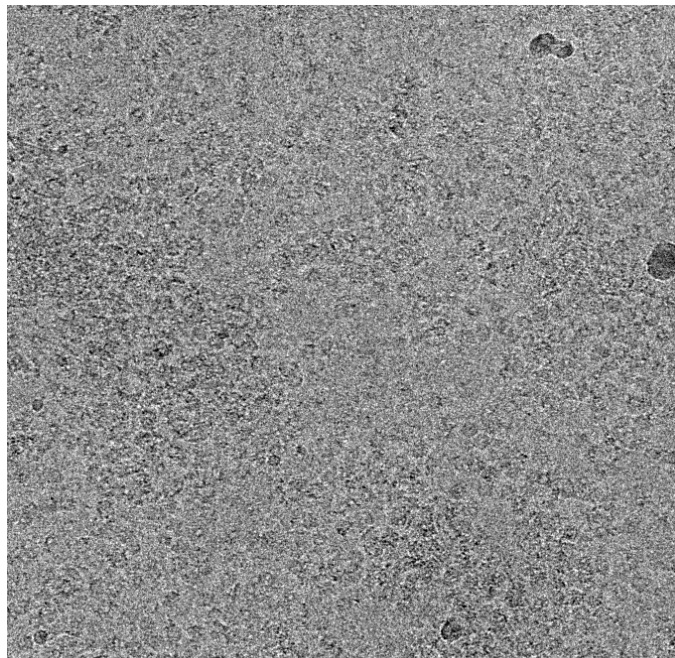
y axis corresponds to resolution in Å.

In order to discard the data with drift in a more objective manner than visual inspection, the `_xmipp3_ctifCritPsdCorr90` value was used and it was set up to discard micrographs with a value lower than 0.80. This value quantifies the correlation of Thon rings at 90° rotation, and therefore it detects drift problems. Almost 200 movies were discarded by applying this criterion.

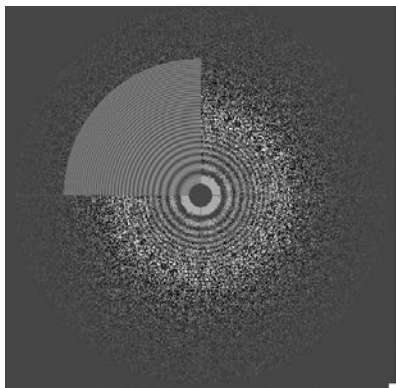
Evaluation of astigmatism was done in a similar way, by checking the `_defocusRatio` parameter and considering a threshold of 1.05 for using the data. In this case, only 16 movies were discarded. However, the software itself discarded another 16 movies more.

In summary, 195 micrographs with drift were manually discarded, 16 were discarded because of astigmatism issues and the program discarded 16 more. In total, 837 movies were left for particle picking.

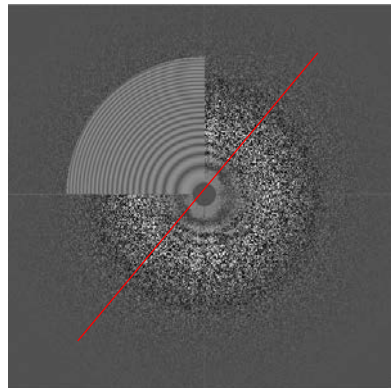
**A**



**B**



**C**

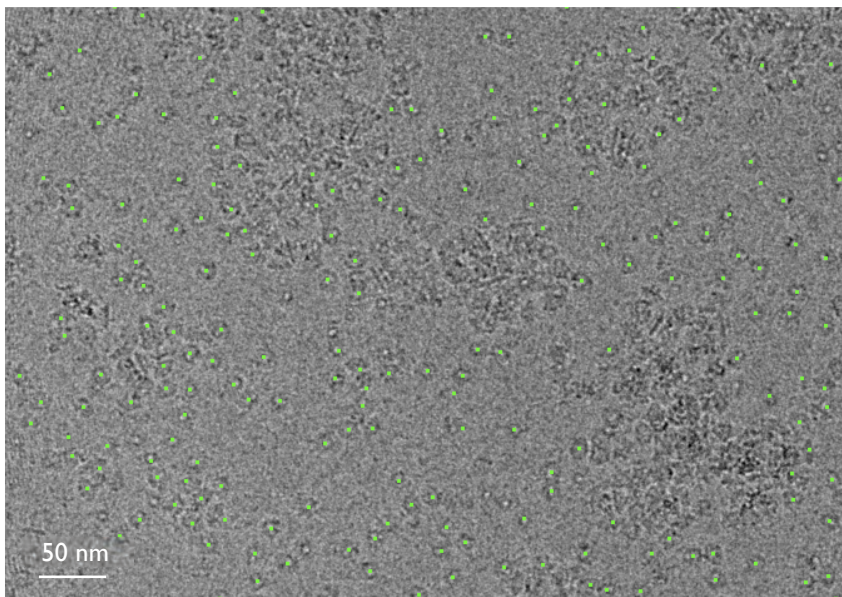


**Figure 4.14** Example of movie with drift.

- (A) The image is blurred, and it is not possible to easily pick the particles.
- (B) Power spectrum image of a movie without drift. Concentric Thon rings have the same intensity in all the directions.
- (C) Power spectrum image of a movie with drift. Thon rings disappear in one direction, marked with a red line.



- **Particle picking:** It was done on the aligned movies, where individual portal assemblies could be observed clearly. Even though in some areas background was a bit noisy and sample was too concentrated, automatic picking could be done. First, around 2,000 particles were picked manually in order to teach the program how to pick the protein (Figure 4.15). Once the training was completed, the program automatically picked 500,000 particles. Boxes of 160 x 160 px were used for picking. Although the protein has a preferred axial orientation, enough lateral and partially lateral views seemed to be present in the micrographs to obtain a 3D reconstruction.

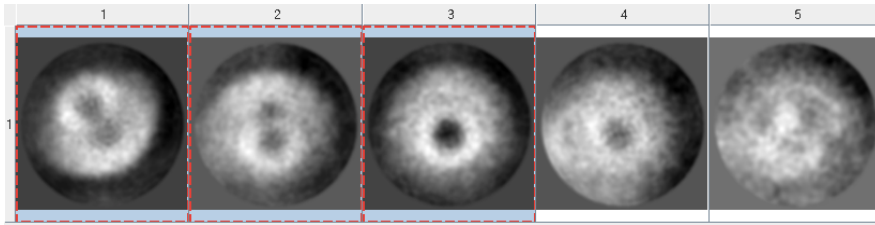


**Figure 4.15** Aligned movie with particles picked.

Example of alignment result and manual particle picking. Each green dot corresponds to one particle. Scale barr indicates 50 nm.

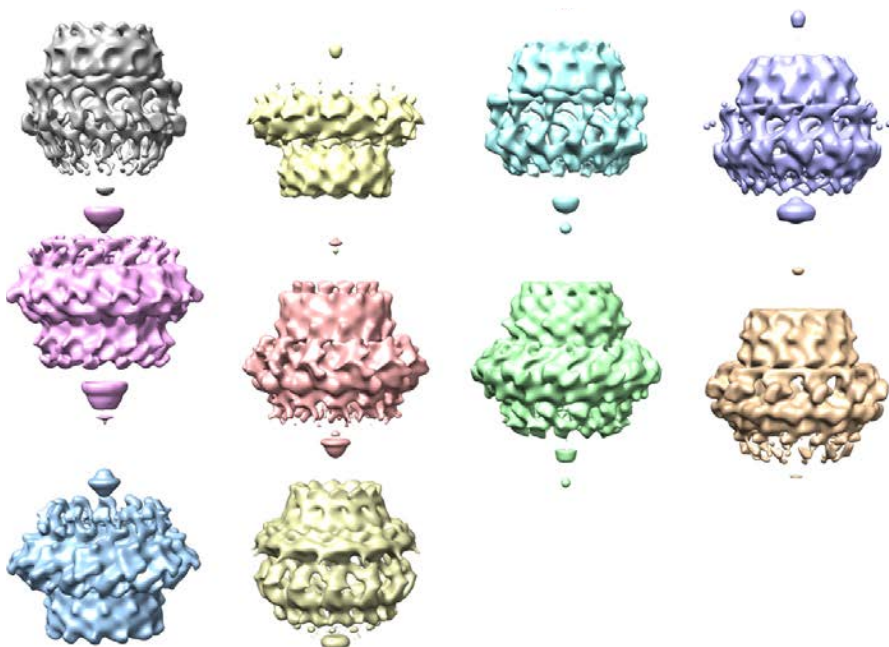
#### **4.3.5 Calculation of an initial volume**

The 2000 particles manually picked were 2D classified with RELION (Figure 4.16). Selected classes were used to obtain an initial volume of the protein with Ransac (Figure 4.17). C12 symmetry was initially imposed, according to what our crystallographic data suggested.



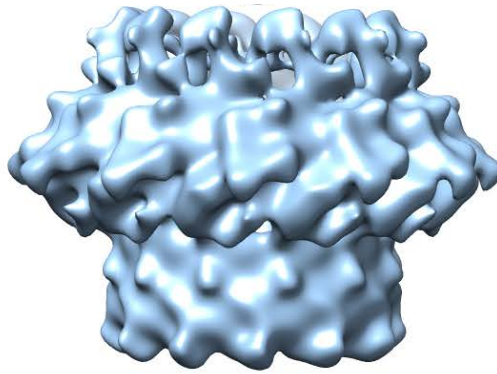
**Figure 4.16** 2D classification of manually picked particles.  
First three classes were selected for initial volume calculation.

The program proposed 10 different initial volumes, which shared a common architecture, indicating coherence on the data (Figure 4.17).



**Figure 4.17** Ransac initial volumes.  
Lateral views of the output volumes.

A specific one was chosen based on the similarity to other portal proteins, the continuity of the map at both ends, and because it showed fewer artificial densities at the channel area than the others. It was prepared for further steps by removing the artifact volumes at both sides of the channel (Figure 4.18).



**Figure 4.18 Polished initial volume.**  
Without central artifact volumes.

### **4.3.6 Extensive particle classification**

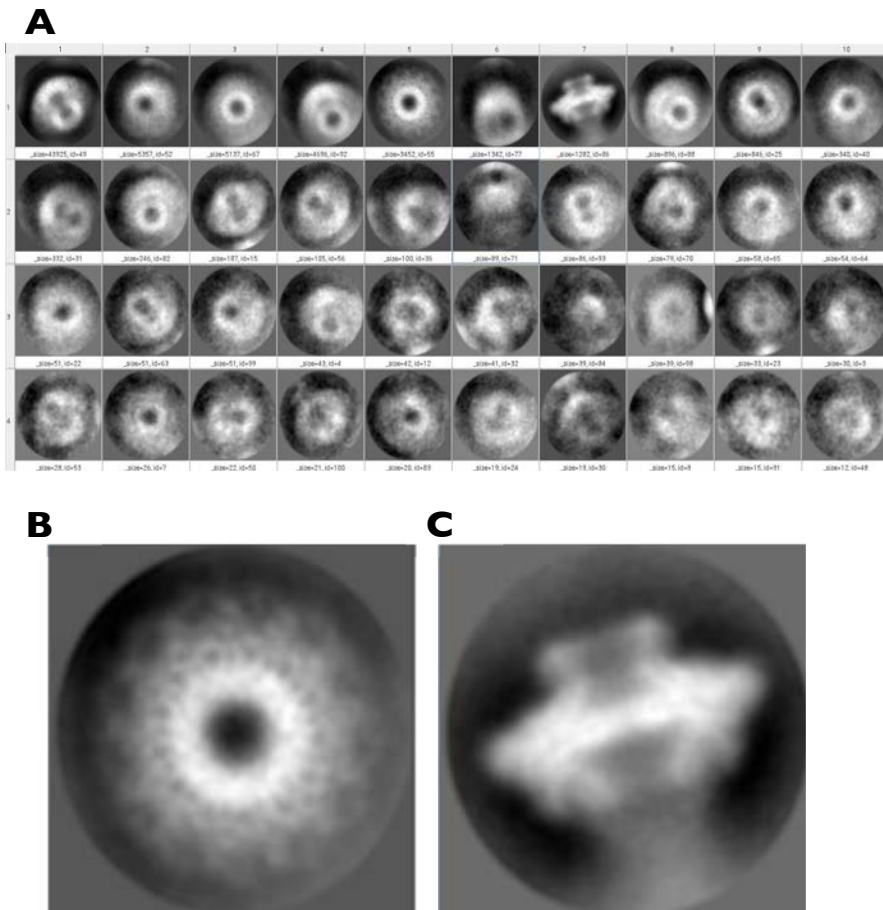
The objective of particle classification is to take only the best particles to reconstruct the EM volume up to the highest resolution.

As the initial number of particles was quite large several rounds of 2D classification were required:

- First, to remove some dusts and contaminations that the picking program selected by mistake.
- Afterwards to distinguish different views of the protein (in the case of 2D classifications) and choose the particles that reconstructed better volumes (in the case of 3D classifications).

Initial classifications were done with xmipp3 - c12d software, which is less hardware-demanding than RELION, the one used on later steps when fewer particles were examined.

Final 2D class averages indicated that the collected data is of high-resolution (Figures 4.19B and 4.19C). Frontal views allowed to distinguish individual monomers, and particles formed by 13 monomers instead of 12 were observed (Figure 4.19B).

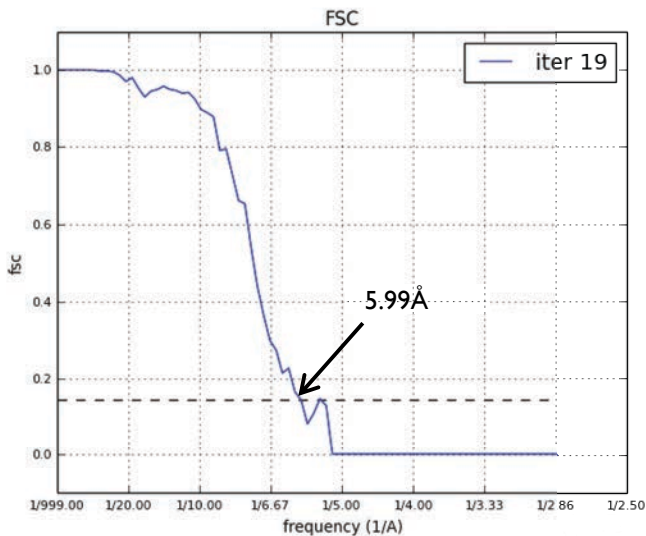


**Figure 4.19 2D classification.**

- (A) Overview of one of the last 2D classification rounds.
- (B) Zoom of a 2D class average of frontal views with C13 symmetry.
- (C) Zoom of a 2D class average of lateral views.

### 4.3.7 Structure refinement

Initial refinement trials with 1,000 lateral views of the particle and imposing C13 symmetry gave a preliminary model at 7.8Å resolution. Starting from a C13 initial volume filtered at 8Å and increasing the number of particles to 12,000, the model resolution improved to 5.99Å after 19 iterations of refinement (Figure 4.20).



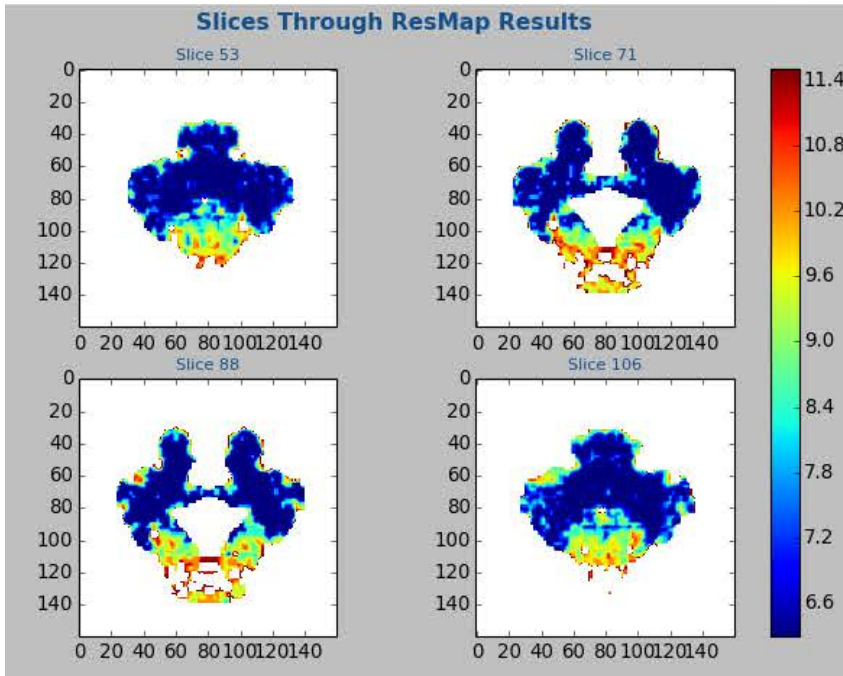
**Figure 4.20** Gold standard FSC refinement curve. Calculated from RELION auto-refine. Overall resolution is calculated to be 5.99Å based on an FSC 0.15 threshold. Auto-refinement stopped at iteration (iter) 19.

However, during alignment of the particles, RELION reported some errors, that prevented obtaining an atomic resolution map. This could be explained by:

- **Heterogeneity of the sample:** A small population of complexes with a different oligomerization state (probably dodecamers) may be interfering in the process.
- **Micrographs background:** The noisy background, most probably denatured protein, may be hampering the alignment.
- **Sample too concentrated:** As particles are too close to each other, fragments of neighboring subunits may interfere with the reconstruction.

Finally, a map at 5.8Å resolution was obtained after post-processing.

Local resolution of the map was calculated (Figure 4.21). Slices showed that the core of the particle has better resolution than the edges, being especially low in the area that would correspond to the crown domain in other portals (see below).



**Figure 4.21** Local resolution slices.

Data at higher resolution appears in blue, while areas in red are the ones with lower structural detail.

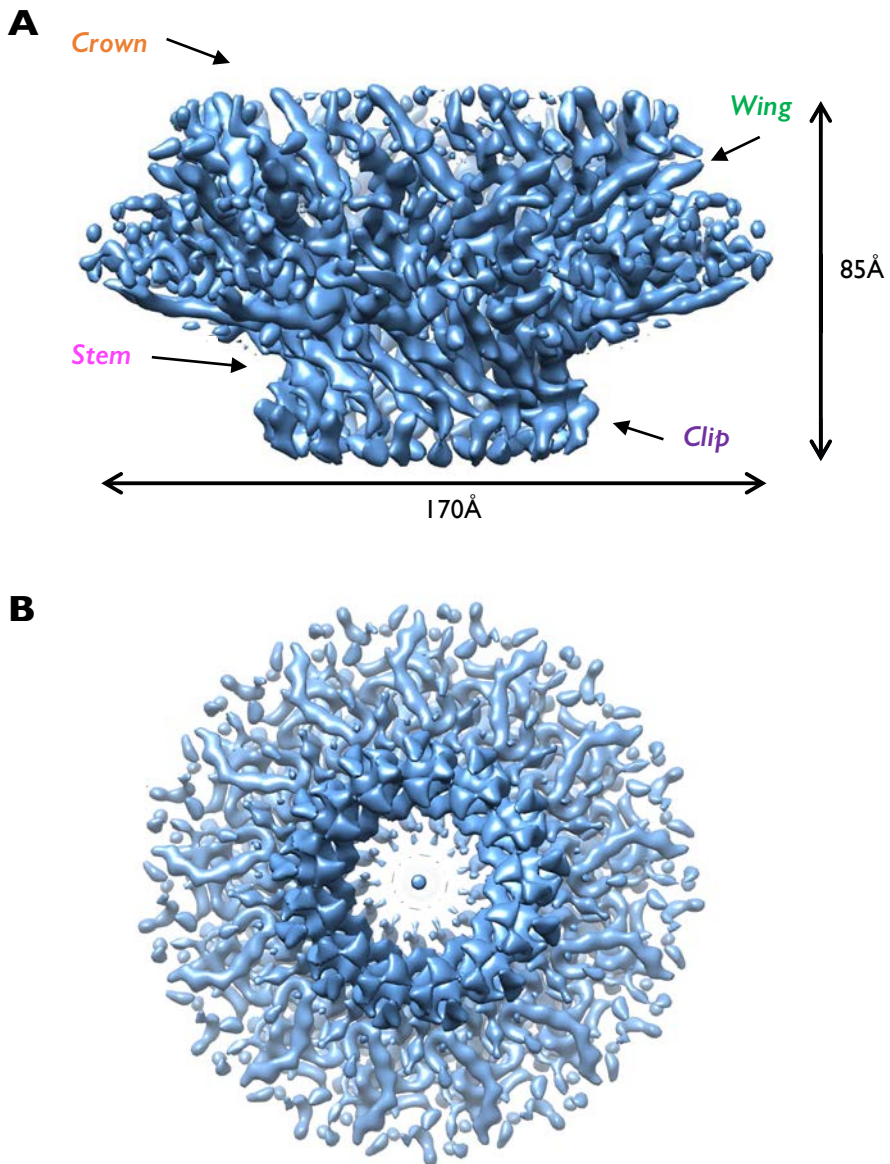
These results suggest that the crown area is highly flexible and may have different conformations or even be partially disordered.

The cryo-EM map of the T7 bacteriophage portal protein shows a particle with a diameter of 170Å, a height of 85Å and a central channel of 25Å (Figure 4.22).

Four domains equivalent to those found in SPP1 and T4 portal proteins can be distinguished on the structure:

- **Crown:** Top part of the structure, which faces the inner of the capsid shell. This area may be disordered, and cannot be seen clearly in the map.
- **Wing:** External part of the assembly, which contains what seems to be a  $\beta$ -strand.

- **Stem:** Ring of tilted  $\alpha$ -helices, two per monomer, which gives a total of 26.
- **Clip:** Bottom-end domain, which seems to have some  $\beta$ -strands.

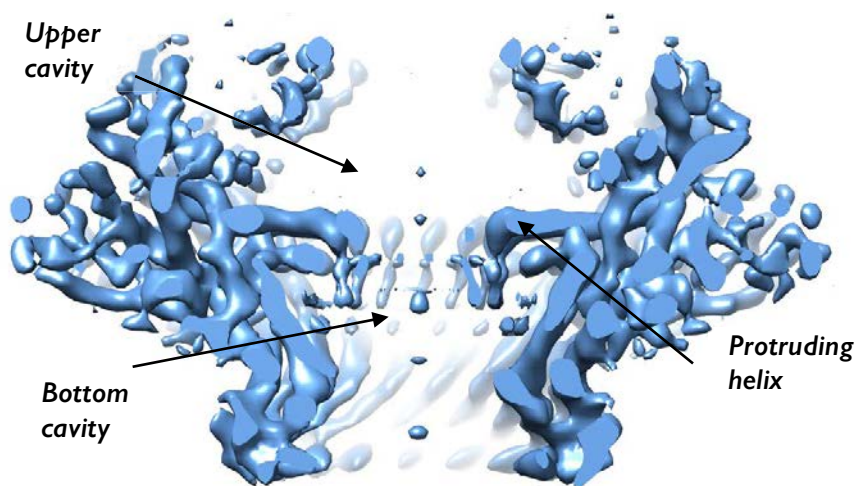


**Figure 4.22** Cryo-EM 3D model at 5.8Å.

(A) Side view with dimensions and domains indicated.

(B) Frontal view with the narrowest channel diameter indicated.

Both the domains and the shape of the internal channel are different in this structure with respect to the previously reported cryo-EM map of gp8 protein. Regarding the channel, each monomer has a long horizontal helix that protrudes into it, forming a ring that delimitates two cavities, one above and one below (Figure 4.23).



**Figure 4.23** Detail of the channel cavities. Central slice of a lateral view of the particle. Protruding helix and cavities are indicated.

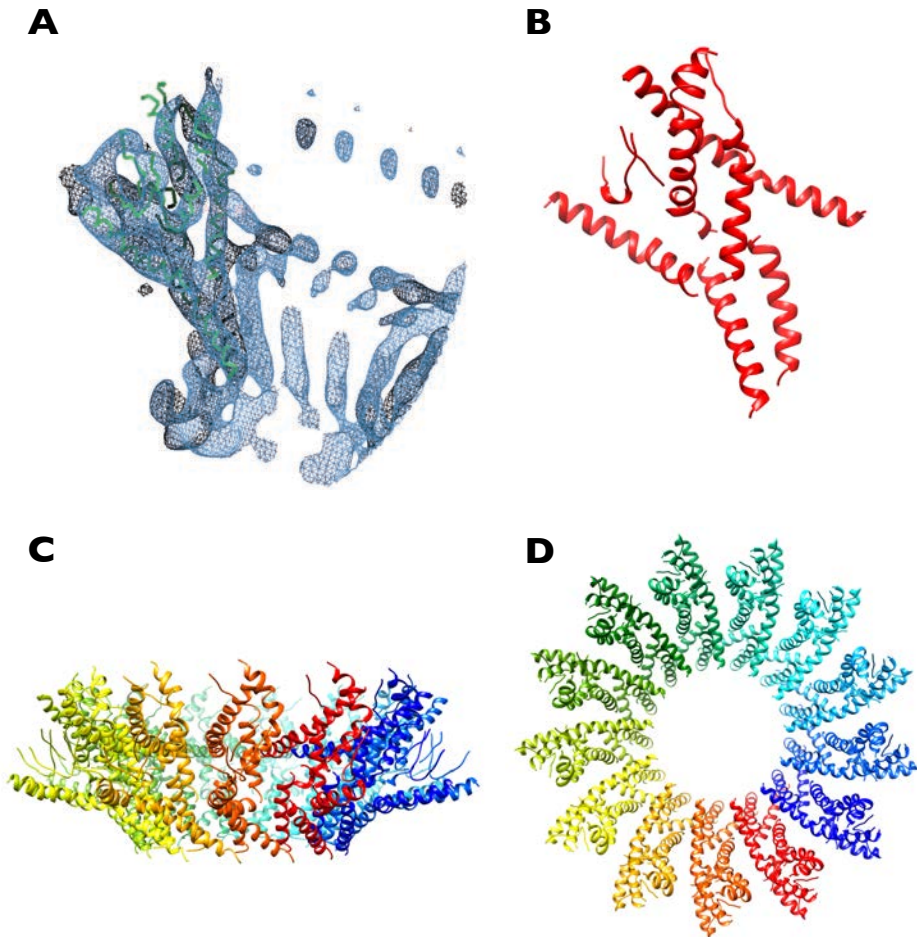
The volume has specific features which are different from the other portals proteins, and keeps the interest on having atomic details of this structure.



## 4.4 Structure determination

### 4.4.1 Model building into the cryo-EM map

Almost half of the protein, mainly  $\alpha$ -helices, could be built as non-connected polyAla chains and was refined against the cryo-EM map (Figure 4.24).



**Figure 4.24 Initial model building.**

- (A) Fitting of the model into the cryo-EM map.
- (B) Cartoon representation of one initial model monomer.
- (C) Cartoon representation of the lateral view of the tridecameric portal initial model (rainbow colouring per monomer).
- (D) Cartoon representation of the axial view of the tridecameric portal initial model (rainbow colouring per monomer).

The initial model was built in Coot, using a tool that allows the *ab initio* building of helices on suitable densities. After that, the individual helices were refined separately in real-space as rigid bodies. Then, an initial model of the tridecameric T7 bacteriophage portal protein was generated applying 13-fold symmetry and refined against the cryo-EM map.

The model consists mainly on 9  $\alpha$ -helices per monomer, located on the wing and stem domains (Figure 4.24B).

As lateral chain densities were not clear in the map, residues could not be correctly assigned and the protein was built with polyAla chains. Moreover, the sequence of the residues could not be established because the connectivity between the built regions was not clear. Therefore, although the cryo-EM volume gives some relevant structural information, it is not sufficient for obtaining an atomic model of the T7 portal protein.


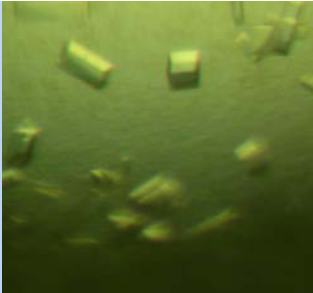
The strategy to follow in order to obtain atomic resolution was to go back now to crystallographic data, and solve the structure using the cryo-EM experimental initial model as MR search ensemble.

P6<sub>3</sub>22 data analysis showed that the hexagonal crystals contained dodecameric portals, and the cryo-EM model corresponds to a tridecameric assembly. Therefore, previously collected crystallographic data were reprocessed and analyzed in order to try to find suitable datasets for MR trials.

#### 4.4.2 Crystallographic data analysis

Diffraction data above 4Å coming from two crystal forms previously obtained in the lab was reanalyzed during the project. Crystallization results regarding these bar and prismatic crystals are summarized below (Table 4.4).

**Table 4.4 Previous crystallization and freezing results.** *Optimized crystallization condition, protein concentration, type of optimization plate, optimization temperature, pictures and cryo-protectant.*

	Bar crystals	Prismatic crystals
Optimized crystallization condition	0.2M CaCl 0.1M HEPES pH 7.5 18% [w/v] PEG 400	15% tacsimate 0.1M HEPES pH 7 12% [w/v] PEG 3350
Protein concentration	4.4 mg/ml	8.5 mg/ml
Type of optimization plate	Hanging drop	Hanging drop
Optimization temperature	20°C	20°C
Picture		
Cryo-protectant	30% [w/v] PEG 400	20% glycerol

Regarding the bar crystals, 270 images were collected, data was reprocessed with XDS, and statistics improved when considering only the first 125 images. The space group is  $P2_12_12_1$  and the resolution  $3.74\text{\AA}$  (Table 4.5).

**Table 4.5** Bar crystals data reprocessing with XDS. Crystallographic processing parameters. Outer shell parameters are indicated in parenthesis.

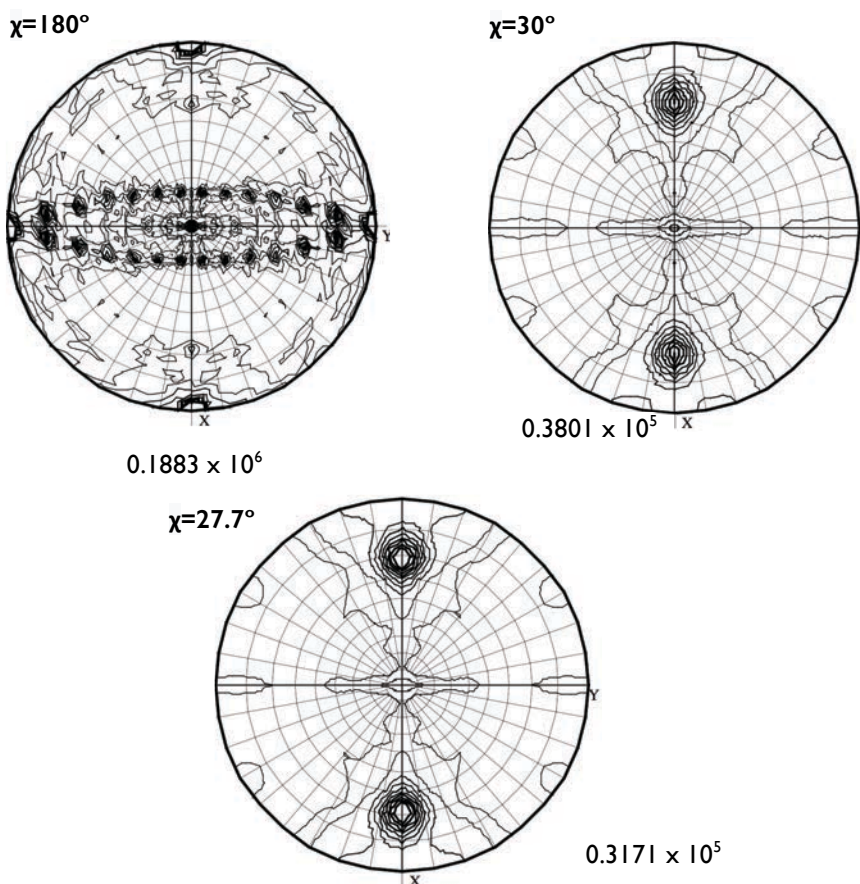
Parameters	Values
Wavelength (Å)	0.9791
Resolution range (Å)	25.94 – 3.74 (3.87 – 3.74)
Space group	P2 <sub>1</sub> 2 <sub>1</sub> 2 <sub>1</sub>
Unit cell	a=119.85Å b=238.57Å c=265.61Å $\alpha=\beta=\gamma=90^\circ$
Total reflections	303,103 (27,368)
Unique reflections	78,146 (7,193)
Multiplicity	3.9 (3.8)
Completeness (%)	98.17 (91.75)
Mean I/ $\sigma$ (I)	9.05 (1.21)
Wilson B-factor	130.85
R-merge	0.162 (1.29)
R-meas	0.189 (1.50)
R-pim	0.097 (0.757)
CC <sub>1/2</sub>	99.3 (46)
CC*	99.8 (79.4)

From now on, this dataset will be named P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub>. Asymmetric unit content analysis by calculating the V<sub>M</sub> suggested the presence of 13 or 14 monomers per asymmetric unit (Table 4.6).

**Table 4.6** P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub> V<sub>M</sub> analysis results. Probabilities, V<sub>M</sub> and percentages of solvent associated to each number of hypothetical monomers per asymmetric unit.

Number of monomers per asymmetric unit	V <sub>M</sub>	Percentage of solvent	Probability
8	3.93	68.72	0.01
9	3.49	64.81	0.03
10	3.14	60.89	0.05
11	2.86	56.98	0.10
12	2.62	53.07	0.15
13	2.42	49.16	0.20
14	2.25	45.25	0.20
15	2.10	41.34	0.15
16	1.96	37.43	0.07
17	1.85	33.52	0.02

The SRF of this dataset at  $\chi=180^\circ$ ,  $\chi=30^\circ$  and  $\chi=27.7^\circ$  sections showed the following peaks and maxima (Figure 4.25):



**Figure 4.25**  $P2_12_12_1$  SRF sections. Maximum values are indicated on each section.  $\chi=180^\circ$ ,  $\chi=30^\circ$  and  $\chi=27.7^\circ$  sections are shown.

The SRF shows more intense peaks at  $\chi=27.7^\circ$  than at  $30^\circ$ , therefore tridecamers seem to be the most probable complex present in the crystal. These results suggest that there is one portal particle per asymmetric unit, and the solvent content of the crystal would be 49.16%.

Regarding the prismatic crystals, six datasets were collected from different spots situated along the longest axis of a unique crystal. 10-image chunks per dataset were integrated, and afterwards scaled and merged together (Table 4.7).

**Table 4.7** Prismatic crystals data reprocessing with XDS. Crystallographic processing parameters. Outer shell parameters are indicated in parenthesis.

Parameters	Values
Wavelength (Å)	0.8726
Resolution range (Å)	28.407 – 2.80 (2.90 – 2.80)
Space group	P4 <sub>2</sub> 2 <sub>1</sub> 2
Unit cell	a=b=261.46Å c=255.94Å $\alpha=\beta=\gamma=90^\circ$
Total reflections	881,091 (11,104)
Unique reflections	193,734 (8,818)
Multiplicity	4.5 (1.3)
Completeness (%)	89.8 (41.2)
Mean I/ $\sigma$ (I)	3.1 (0.1)
R-merge	0.626 (4.883)
R-meas	0.698 (6.735)
R-pim	0.303 (4.618)
CC <sub>1/2</sub>	93.6 (-0.5%)

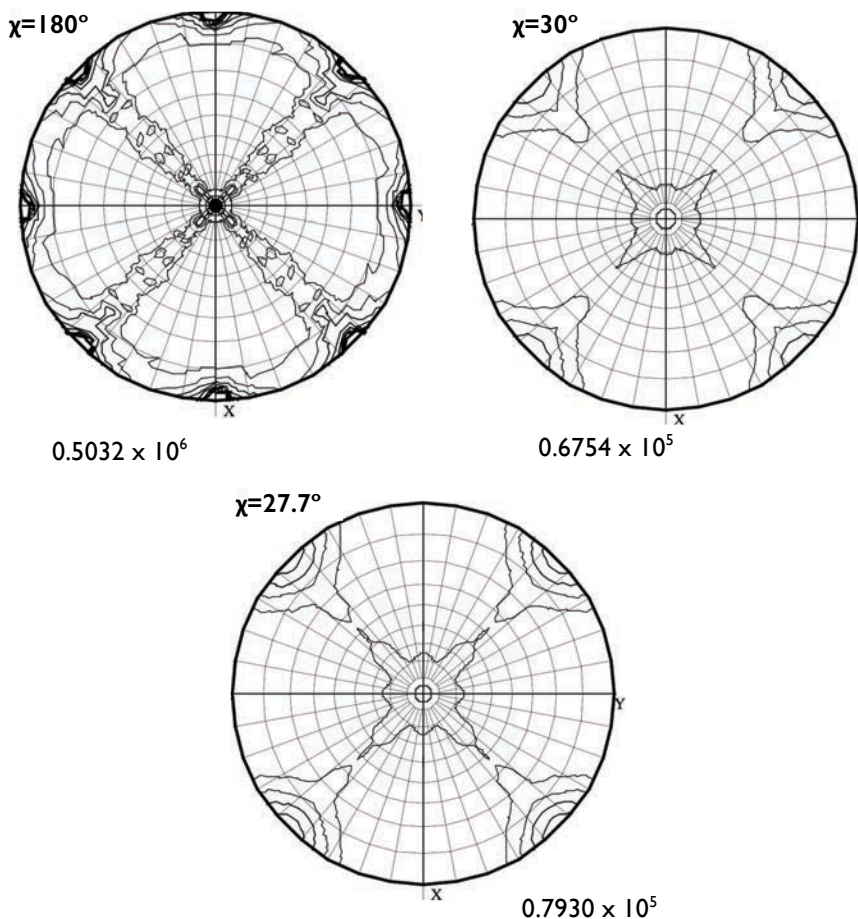
This dataset will be named P4<sub>2</sub>2<sub>1</sub>2.

**Table 4.8** P4<sub>2</sub>2<sub>1</sub>2 V<sub>M</sub> analysis results. Probabilities, V<sub>M</sub> and percentages of solvent associated to each number of hypothetic monomers per asymmetric unit.

Number of monomers per asymmetric unit	V <sub>M</sub>	Percentage of solvent	Probability
9	4.02	69.45	0.01
10	3.62	66.05	0.02
11	3.29	62.66	0.03
12	3.02	59.26	0.06
13	2.79	55.87	0.10
14	2.59	52.47	0.14
15	2.41	49.08	0.17
16	2.26	45.68	0.18
17	2.13	42.29	0.14
18	2.01	38.89	0.09
19	1.91	35.50	0.04
20	1.81	32.10	0.01

Regarding the  $P4_22_12$  dataset, the  $V_M$  for different asymmetric unit contents are shown in Table 4.8. Results suggests the presence of 13-17 monomers per asymmetric unit (Table 4.8). However, according to the SRF peaks, the most probable situation would be crystals with a 55.87% of solvent with 13 subunits (Figure 4.26).

The SRF of this dataset at  $\chi=180^\circ$ ,  $\chi=30^\circ$  and  $\chi=27.7^\circ$  sections shows the following peaks and maxima (Figure 4.26):



**Figure 4.26**  $P4_22_12$  SRF sections. Maximum values are indicated on each section.  $\chi=180^\circ$ ,  $\chi=30^\circ$  and  $\chi=27.7^\circ$  sections are shown.

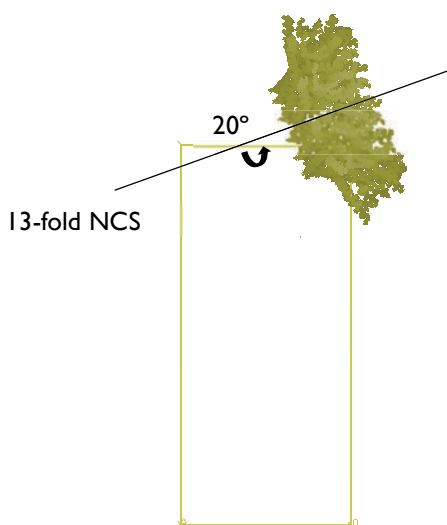
The SRF indicates a 13-fold symmetry of the particle with the local axis slightly offset with respect to the crystallographic two-fold axis.

Data processing of both datasets  $P2_12_12_1$  and  $P4_22_12$ , especially the later one, showed parameters in the outer shell that are not standard (Tables 4.5 and 4.7). As mentioned before, this can be explained because the project deals with a large complex, and this type of samples tend to crystallize forming big unit cells and give poor diffraction.

#### 4.4.3 Structure solution

The model built into the cryo-EM map was used as a search model for MR trials using the crystallographic data described in the previous section. A solution with a TFZ-score 10.7 was found for the  $P2_12_12_1$  dataset.

The location of the symmetry axis of order 13 in agreement with what the SRF suggested confirmed the solution (Figure.27). The peak at the  $27.7^\circ$  section appears at a  $\varphi$  of  $0^\circ$ , which indicates that the symmetry axes of order 13 is located on the  $xz$  or  $ac$  plane. It is inclined  $20^\circ$  respect to the  $x$  ( $a$ ) axis, as shown in Figure 4.27.



**Figure 4.27**  $P2_12_12_1$  MR solution.

Location of the portal particle in the unit cell. The  $ac$  plane is represented, with  $a$  horizontal and  $c$  vertical. The 13-fold axis is located on the plane and inclined  $20^\circ$  with respect to  $a$ .



Although the map was not completely clear, it could be improved after some rigid body refinement cycles with the whole particle and the monomers separately. Moreover, a big improvement on the map was obtained using DM procedures. A phase extension DM protocol by resolution steps was performed, starting at a resolution of 7.9Å. During this process, it was also crucial the information provided by the cryo-EM studies, as the masks used for solvent flattening and NCS-averaging were obtained from those data. During the DM protocol, both masks were updated every 50 and 20 cycles, respectively. The NCS-average correlation between related areas during DM cycles increased as expected. The total number of cycles was 104, which gave a final average correlation between the 13 monomers of 0.889.

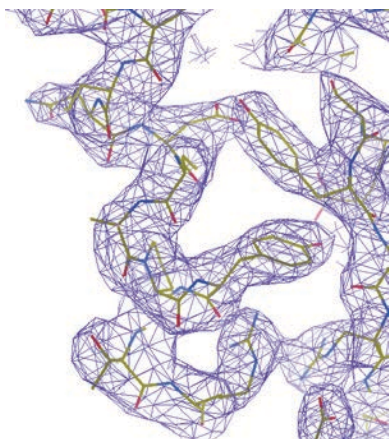
The map that was obtained allowed the building and initial refinement of almost the whole protein, including the clip and part of the crown areas that were not present on the initial cryo-EM model. However, sequence assignment was still difficult because the electron density did not show clearly some lateral chains. After many refinement cycles, a model of almost the whole protein was built.

The new tridecameric model was used as a search ensemble for MR with the P4<sub>2</sub>2<sub>1</sub>2 2.8Å resolution data and a solution was found with a Z-score of 77.5. Model building and refinement could be completed using this data (Table 4.9).

**Table 4.9** X-ray refinement statistics for the P4<sub>2</sub>2<sub>1</sub>2 dataset. *Crystallographic refinement parameters.*

Parameters	Values
Resolution (Å)	2.80
Total reflections	881,091
R-work/ R-free	0.2655 / 0.2881
Number of atoms	49,452
Average B factor	67.61
r.m.s.d. Bond lengths (Å)	0.010
r.m.s.d. Bond angles (°)	1.570

In this case, the electron density map allowed the assignment of the sequence, as lateral chains could be observed clearly (Figure 4.28).



**Figure 4.28 Model building.** Detail of the electron density map and the model built into it, showing some lateral chains.

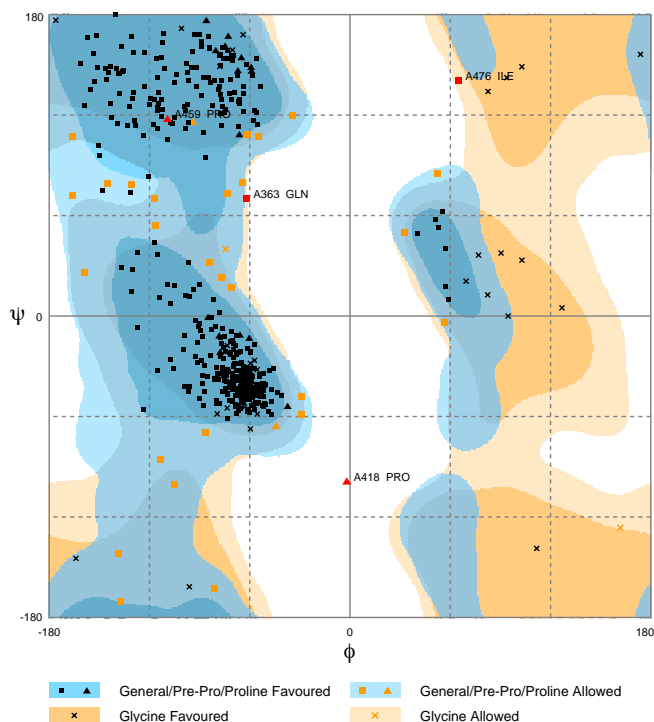
Refinement parameters, with an R-work of 0.2655 and an R-free of 0.2881, are acceptable taking into account the resolution and the size of the complex.

The crystallographic model was validated using MolProbity, a software that checks some parameters that indicate the quality of the model (Table 4.10). One of them is the overall score, which represents the expected experimental resolution for a model of that quality. Ideally, it should be lower than the real resolution. In our case, both have the same value, 2.8Å. In general, values are acceptable for the resolution and considering that the portal is a very big particle.

**Table 4.10 X-ray model validation. MolProbity statistics.**

Parameters	Values
Ramachandran outliers	1.31%
Ramachandran favored	91.69%
C-beta outliers	64
Clashscore	6.29
Overall score	2.80

Ramachandran plot analysis was also performed with RAMPAGE (Figure 4.29). The Ramachandran plot of a monomer showed a small percentage of outliers (0.8%).

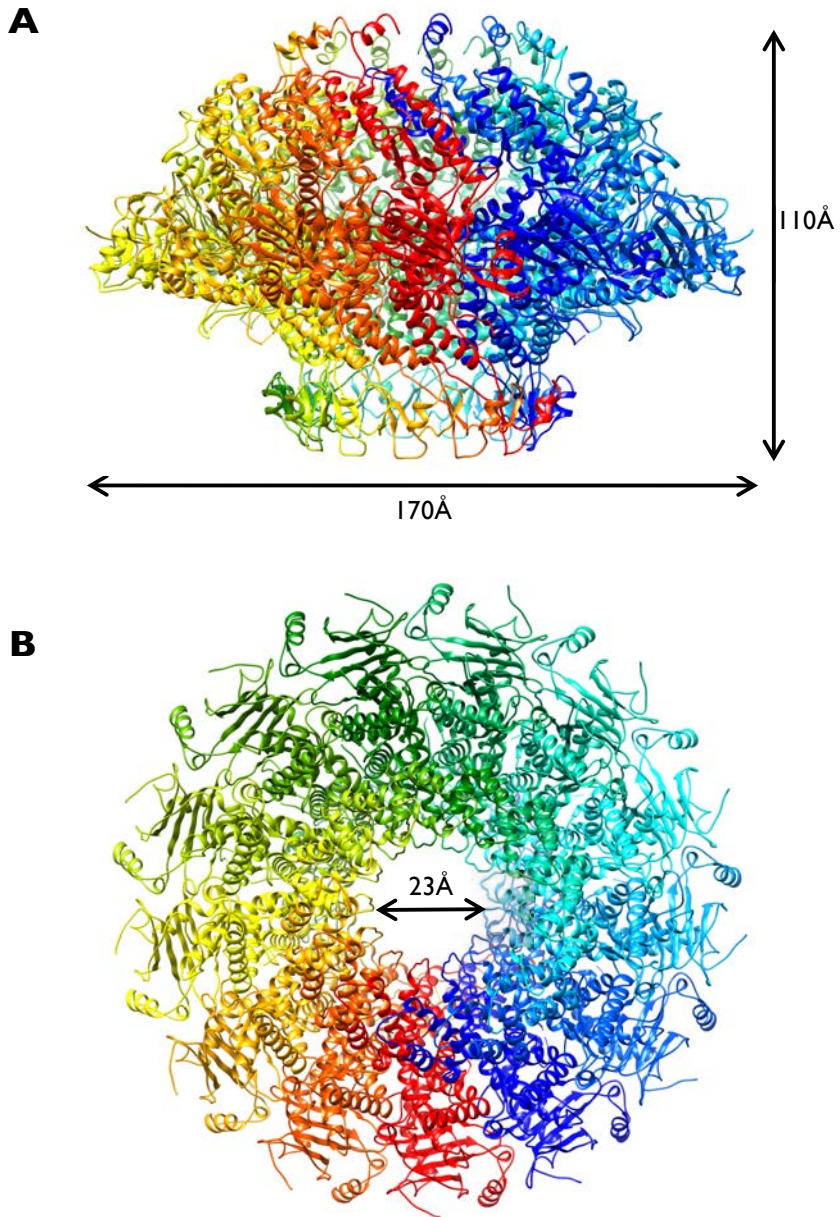


**Figure 4.29** X-ray model Ramachandran plot of chain A.

Favoured residues represented in black (93%), allowed residues in orange (6.2%) and outlier residues in red (0.8%).

The outlier residues correspond to amino acids located near the C-terminal part of the protein, which forms the crown domain (Asn363, Pro418 and Ile476). The map of this part of the protein is significantly worse than the rest, which is in agreement with what was observed in the cryo-EM structure. Therefore, this domain seems to be more flexible than the others.

The overall architecture shown by the X-ray structure of the T7 portal assembly is similar to the one of the cryo-EM volume (Figure 4.30). The height of the particle is 110Å, with a total diameter of 170Å and a channel with a diameter of 23Å at its narrowest point.



**Figure 4.30** Structure solution and dimensions.

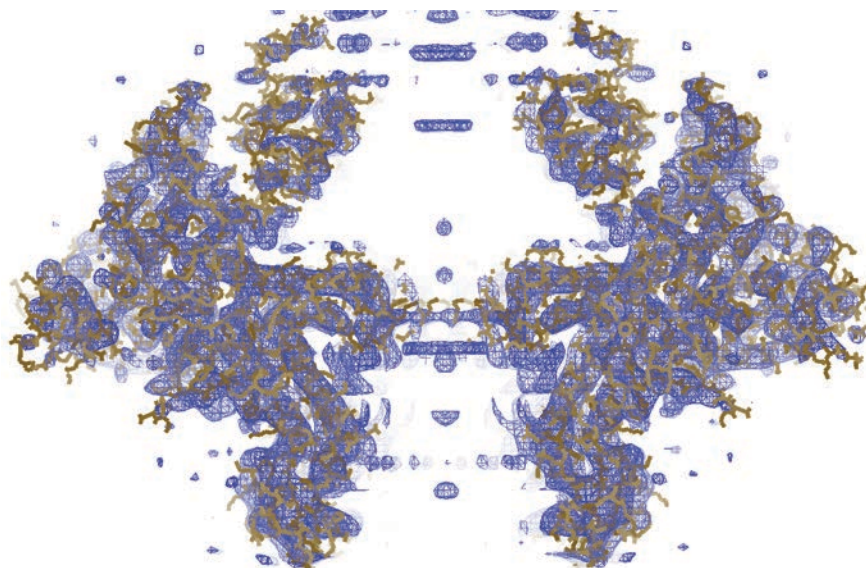
- (A) Cartoon representation of the lateral view of the tridecameric portal (rainbow colouring per monomer).
- (B) Cartoon representation of the axial view of the tridecameric portal (rainbow colouring per monomer).

All the previous portal protein structures were obtained by X-ray crystallography, except the T4 one, that was obtained by cryo-EM. The T7 bacteriophage portal structure is the first one that has been solved using a combination of crystallographic and cryo-EM data. In our case, combining both techniques has been crucial for obtaining an atomic resolution structure. On one hand, it was not possible to phase the crystallographic data, and on the other hand the cryo-EM model obtained did not have enough resolution.

Therefore, this project is an example of a new strategy for phasing crystallographic data, based on the building of a preliminary model in a cryo-EM maps that afterwards can be used for MR.

#### 4.4.4 Refinement into the EM volume

The crystallographic X-ray structure obtained from the P4<sub>2</sub>2<sub>1</sub>2 dataset model could be fitted into the cryo-EM volume (Figure 4.31).



**Figure 4.31** X-ray model fitted into the cryo-EM map.  
Lateral view of the tridecameric portal.

It was first manually docked into the volume, and afterwards rigid body refined per monomer and domain.

The model was then refined against the cryo-EM map. Although the refinement parameters are far from optimal, they are reasonable considering the low resolution of the cryo-EM map (Table 4.11).

**Table 4.11 Cryo-EM refinement. CCP-EM statistics.**

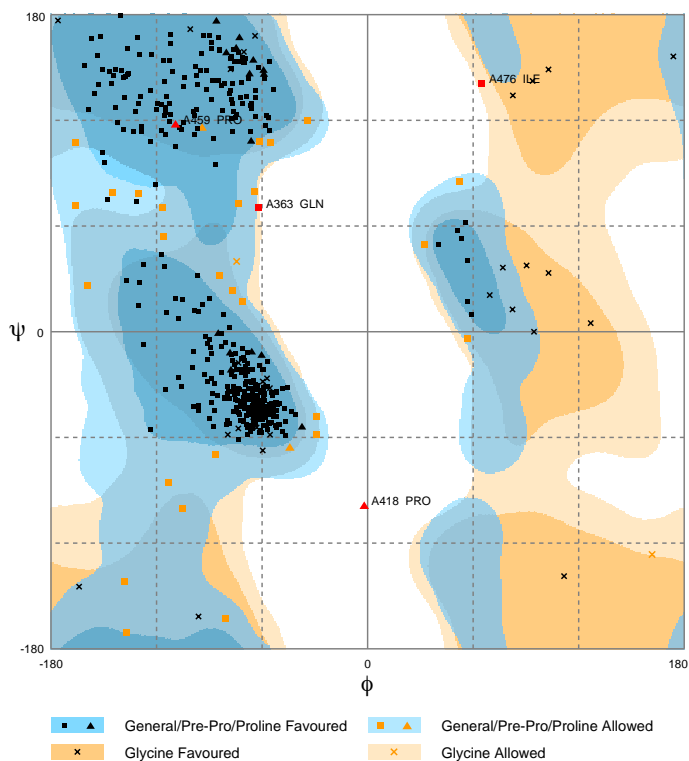
Parameters	Values
Fourier shell correlation (FSC) average	0.6452
R-factor	0.6733
r.m.s.d. Bond lengths (Å)	1.6931
r.m.s.d. Bond angles (°)	0.0115
r.m.s.d. Chiral	0.1126

The model refined into the cryo-EM volume was also checked with MolProbity (Table 4.12). As the only change in the model was rigid body refinement, statistics are quite similar to the previous ones.

**Table 4.12 Cryo-EM model validation. MolProbity statistics.**

Parameters	Values
Ramachandran outliers	1.44%
Ramachandran favored	91.36%
C-beta outliers	104
Clashscore	10.46
Overall score	3.00

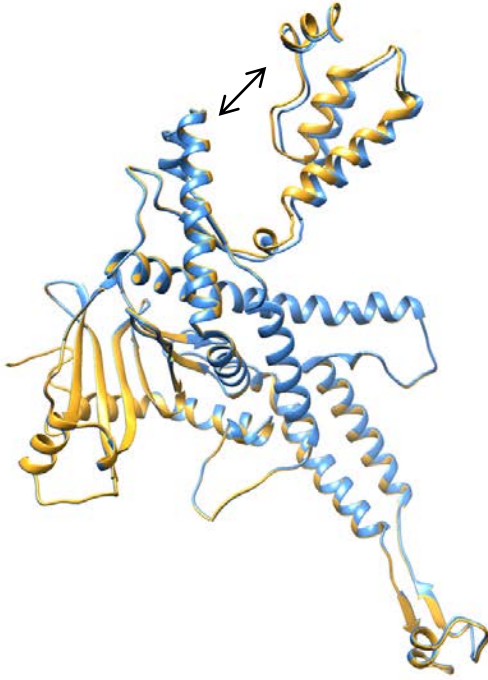
A Ramachandran plot was also obtained with RAMPAGE which showed the same Ramachandran outliers as the X-ray model in which it is based (Figure 4.32).



**Figure 4.32** Cryo-EM model Ramachandran plot of chain A. Favoured residues represented in black (%), allowed residues in orange (%) and outlier residues in red (%).

The resultant cryo-EM model does not offer additional high-resolution information about the particle, but it allows the comparison in terms of domain disposition between two models obtained by different structural techniques, from the same biological sample.

Superposition of one monomer from each model shows that both structures are very similar, and present an r.m.s.d. of 0.530Å (Figure 4.33). However, slight differences relative to the disposition of the crown domain with respect to the rest of the structure can be observed.



**Figure 4.33** Superposition of the X-ray and cryo-EM models.

Models are shown in cartoon representation. The orange molecule is the X-ray model, while the blue one is the portal structure fitted into the cryo-EM volume. The arrow indicates the putative movement of the crown domain.

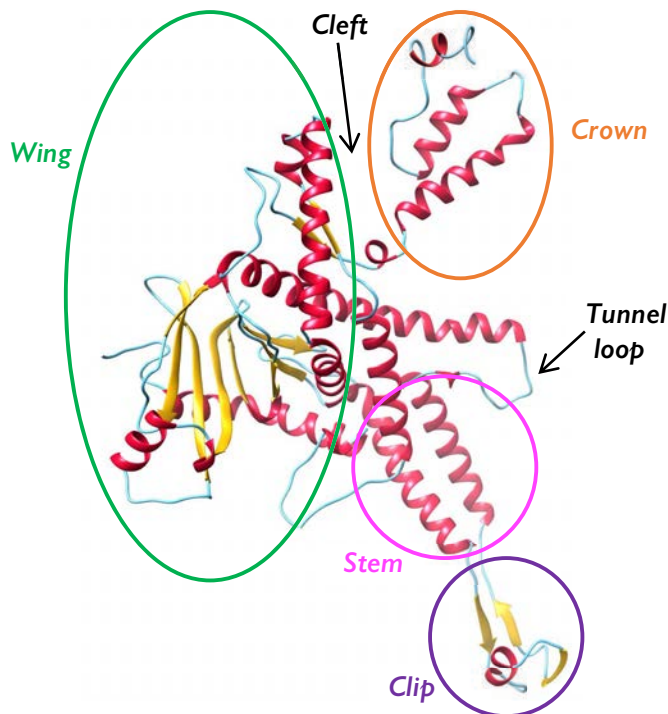
r.m.s.d. = 0.530Å



## 4.5 Structural analysis

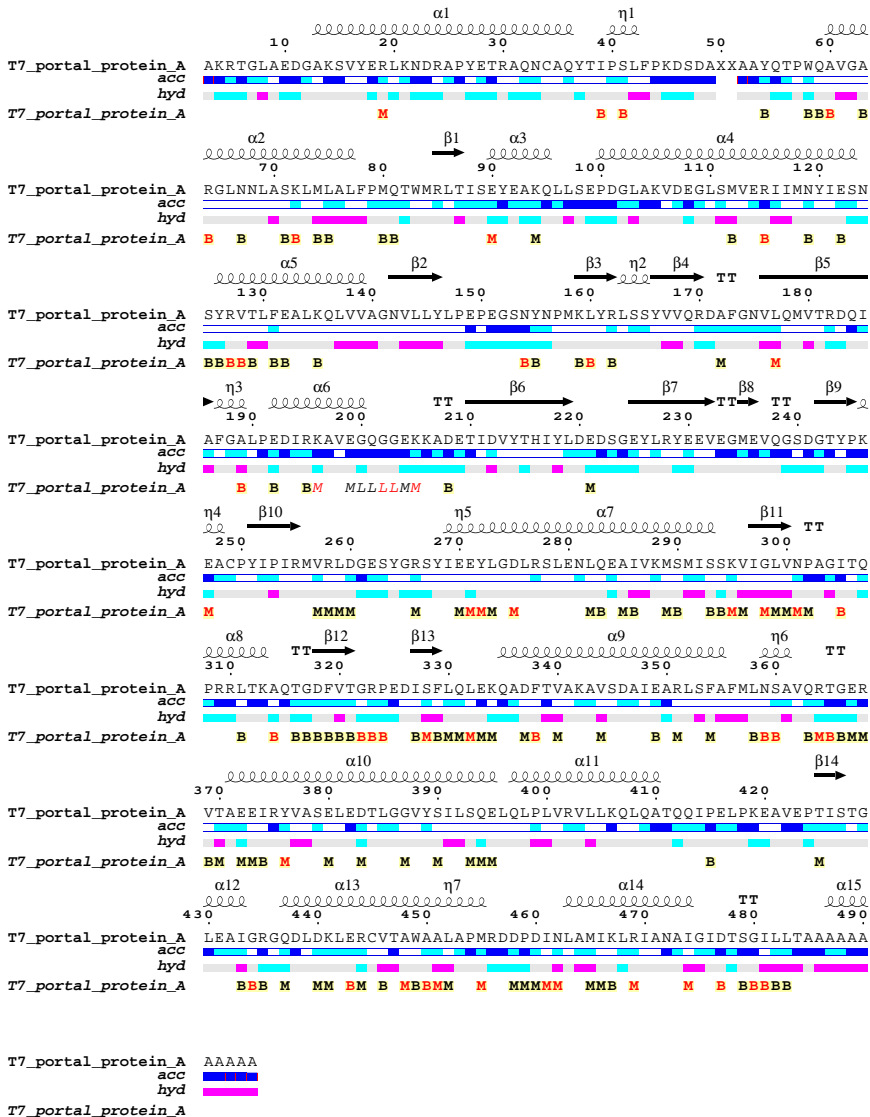
The overall shape of the T7 portal protein by the atomic X-ray structure shows a ring-like assembly of 13 subunits with an axial central channel.

The diameter of the particle is 170Å and its weight 110Å. Regarding the channel diameter, it varies from 95Å in the wider part to 23Å in the narrowest part. These dimensions are similar to the ones described in the cryo-EM section. Differences are a consequence of poor densities in the crown domain and in the tunnel loop, which gave a smaller height and a bigger channel diameter in the cryo-EM results analysis. The four domains already described on the cryo-EM volume can also be identified (Figure 4.34).



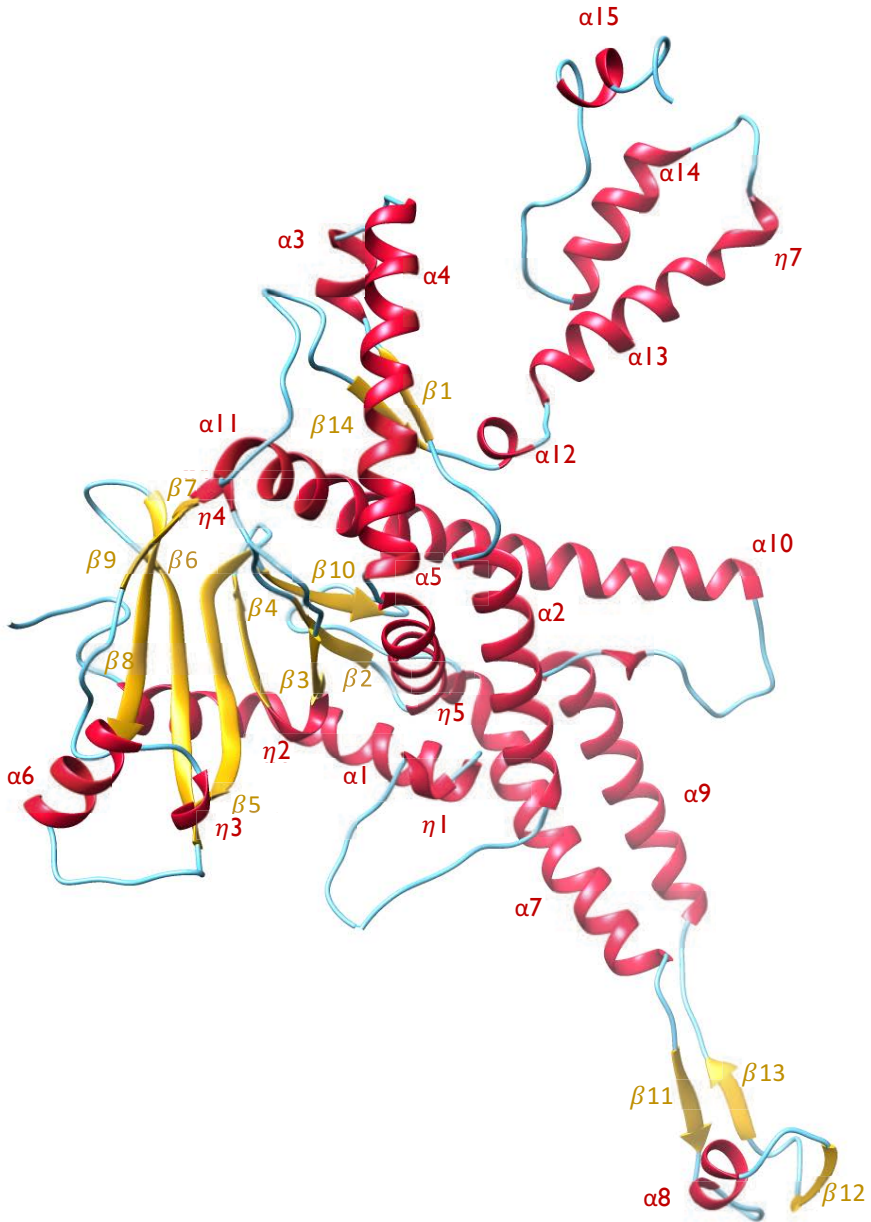
**Figure 4.34** Domains of the T7 portal protein monomer.

Cartoon representation of a T7 portal protein monomer with the four domains, the tunnel loop and the cleft indicated. Helices appear in red,  $\beta$ -strands in yellow.



**Figure 4.35 Model sequence and structural summary.**

Above the sequence,  $\alpha$ -helices appear as medium squiggles,  $3_{10}$  helices as small squiggles,  $\beta$ -strands as arrows and  $\beta$ -turns as TT letters. Below the sequence, relative accessibility (from white buried to blue accessible) and hydrophathy (pink for hydrophobic, grey for intermediate and cyan for hydrophilic) are represented. At the bottom, contacts with other monomers are summarized. Non-crystallographic contacts appear as bold letters with a yellow background. Crystallographic contacts appear as italic letters. In both cases red letters correspond to contacts below  $3.2\text{\AA}$  and black letters to contacts between  $3.2\text{\AA}$  and  $5\text{\AA}$ .



**Figure 4.36** Secondary structure elements.  
 Monomeric T7 portal protein with the secondary structure elements indicated.  
 Helices appear in red,  $\beta$ -strands in yellow and coils in blue.

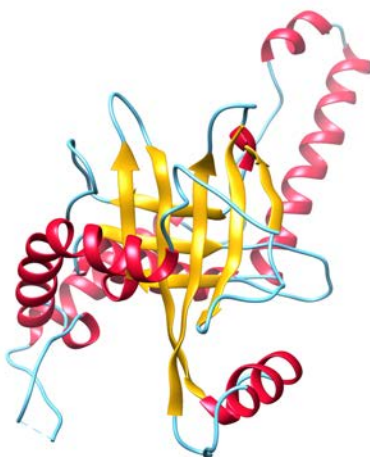
The ENDSCRIPT tool was used to analyse monomer A (Figure 4.35). Some residues are not present in the model, and in some cases, they were built as alanines because the density was not enough to fit the side chains:

- **N-terminal**: The first two residues are not visible at all, and the third, which should be Glu, appears as Ala.
- **Flexible loop**: Residue 49 should be a Gln and it is an Ala, residues 50 and 51 are not visible at all (Ala-Ser), and 52 and 53 should be Thr-Asp and are Ala-Ala.
- **C-terminal**: The residues between 485 and 495 have been built as a polyAla chain due to poor density. From the 496 to the last residue, 536, amino acids are not visible at all.

Regarding the secondary structure, each monomer has 15  $\alpha$ -helices, 14  $\beta$ -strands and 7  $\eta$ -helices  $3_{10}$  (Figure 4.36).

Detailed information about the secondary structure elements forming each of the domains is now available:

- **Wing**: It is the biggest domain, which protrudes outwards in the middle part of the assembly, and it contains the N-terminal end of the protein, which is located at the most outer part.  $\beta$ -strands from  $\beta 2$  to  $\beta 10$  build a significant part of the domain, forming two tilted antiparallel  $\beta$ -sheets. These two sheets, rotated by an angle of  $90^\circ$  one respect to the other, form a sandwich (Figure 4.37).



**Figure 4.37** Wing  $\beta$ -sandwich.  
Detail of the wing domain.  
Helices appear in red,  $\beta$ -strands in yellow and coils in blue.

Connections between the strands consist of  $3_{10}$  helices, coils,  $\beta$ -turns, and one  $\alpha$ -helix, the external  $\alpha 6$ , between  $\beta 5$  and  $\beta 6$ .  $\beta 9$  is in the most external part of the  $\beta$ -sheet, and  $\beta 10$  interacts with  $\beta 2$ , connecting with the bottom stem and clip domains. The domain contains also some long helices, for instance  $\alpha 1$  builds the “floor” of the domain going from the exterior to the interior of the particle,  $\alpha 2$  goes parallel to the channel upwards, and  $\alpha 4$ - $\alpha 5$  go down forming an angle of about  $120^\circ$  one respect to the other. After the tunnel loop, a long kinked  $\alpha$ -helical structure (named in global as  $\alpha 10$ ) built by  $\alpha 10$  (25 residues) and  $\alpha 11$  (14 residues) is found perpendicular to the channel axis and goes from the inner of the channel to the outer part of the wing. A flexible loop goes out at the bottom part of the domain. The  $\beta 14$  strand interacts in an antiparallel manner with the  $\beta 1$ , and connects  $\alpha 11$  with  $\alpha 12$ , and the crown domain.

- **Stem:** The stem domain is built by two  $\alpha$ -helices, that connect the wing and clip domains.  $\alpha 7$  goes from the wing to the clip, while  $\alpha 9$  goes from the clip to the wing.
- **Clip:** Bottom domain, found between  $\alpha 7$  and  $\alpha 9$ . It is built by three  $\beta$ -strands, with  $\beta 11$  and  $\beta 13$  from the same monomer forming an antiparallel  $\beta$ -sheet, and  $\beta 12$  from the contiguous subunit interacting in a parallel manner with  $\beta 11$ . The short helix  $\alpha 8$  links  $\beta 11$  and  $\beta 12$ .
- **Crown:** It is the top C-terminal domain, which is linked to the wing by  $\alpha 12$ . It is formed by the helices  $\alpha 13$ ,  $\eta 7$ ,  $\alpha 14$  and  $\alpha 15$ . Comparison of X-ray and cryo-EM structures suggests that the protein may be flexible from  $\alpha 12$  to its C-terminal end, and the crown domain may have some freedom of movement making the cleft smaller or bigger.

The shape of the internal channel suggested by the cryo-EM volume is confirmed by the X-ray model. The long horizontal helix that protrudes into the channel is  $\alpha 10$ , which separates two cavities:

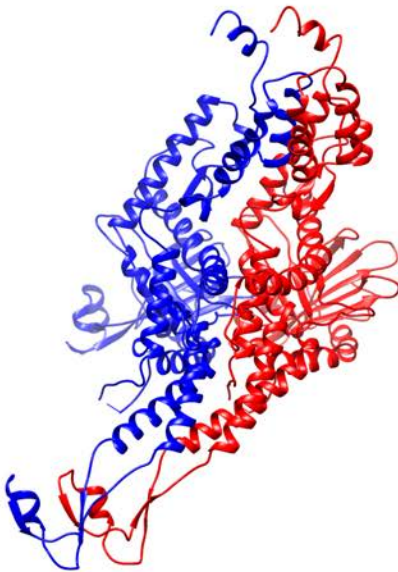
- **Upper cavity:** Delimited mainly by  $\alpha 12$ ,  $\alpha 13$  and  $\eta 7$  from the crown domain and it has a conical shape.
- **Bottom cavity:** With an inverted conical shape, it is delimited mainly by  $\alpha 9$  from the stem domain.

The model also shows another significant structural feature, a deep cleft between the wing and crown domains, which are quite separated and only joined by a coil and the short  $\alpha 12$ .

Non-crystallographic contacts between contiguous monomers are established by many parts of the protein and all the domains are involved in them at some extent. They are especially significant on the following parts (Figure 4.35):

- Wing:  $\alpha 5$
- Stem:  $\alpha 7$
- Clip: Intermolecular  $\beta$ -sheet
- Tunnel loop: All residues involved except one
- Crown:  $\alpha 13$  and  $\alpha 14$

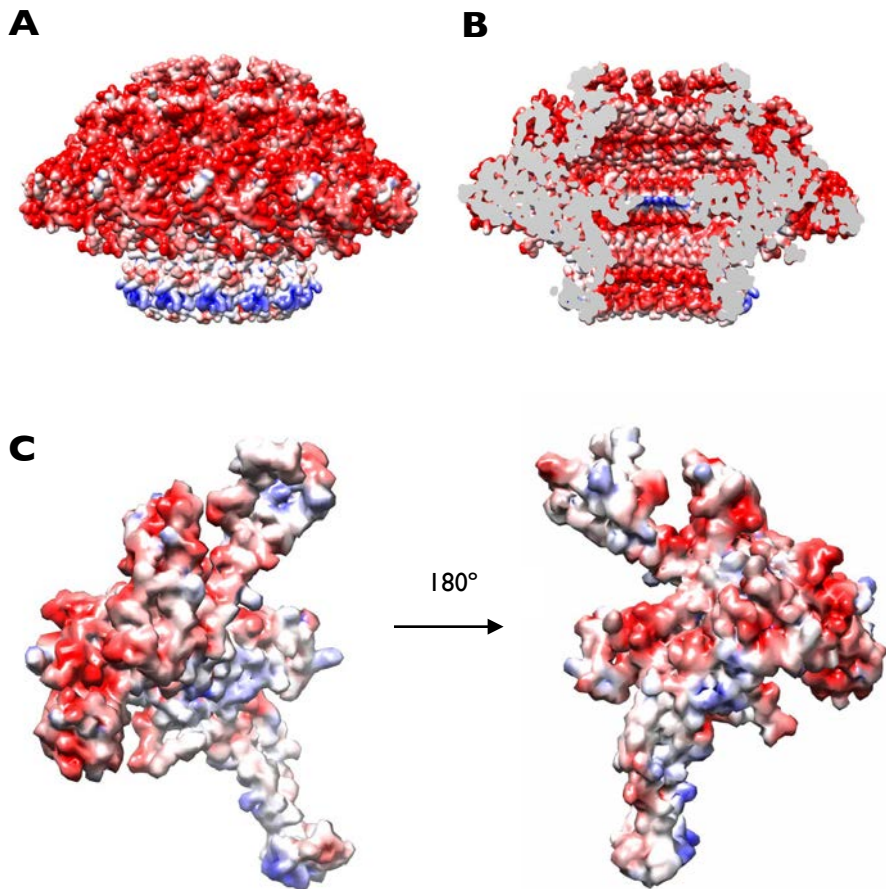
The monomers that form the tridecameric ring are tilted with respect to the symmetry axis (Figure 4.38). The stem helices are tilted around  $45^\circ$ .



**Figure 4.38 Interaction between monomers.**

Two contiguous monomers are represented in blue and red cartoon, seen from the interior of the channel.

Regarding contacts induced by the crystallization in the  $P4_212$  crystals, they are concentrated on the external part of the wing, on the  $\alpha 6$  and the following coil. The presence of two contiguous Gly in this highly accessible coil (Gly202 and Gly203) allows a tight packing of the crystal by means of contacts that involve three different monomers (Figure 4.35).



**Figure 4.39** Surface charge distribution.

(A) Electrostatic potential at the external surface of the tridecamer.

(B) Section of the tridecameric complex showing the electrostatic potential of the portal axial channel.

(C) Lateral views of the electrostatic potential on one monomer.

In the three pictures, blue colour represents a  $10 \text{ kcal}/(\text{mol} \cdot e)$  positive potential, while red represents a  $-10 \text{ kcal}/(\text{mol} \cdot e)$  negative potential.

Electrostatic potentials of the protein surface are mainly negative (Figure 4.38). The tridecameric particle has almost all its external surface negatively charged, with the exception of the clip domain, which is mainly positive (Figure 4.39A).

The analysis of the inner channel shows also a predominantly negative surface, even in the clip domain, with the only exception of the protruding area in the middle of the complex (Figure 4.39B). This positively charged ring corresponds to the Arg368 residue, whose side chain electronic density is partially visible on the map and which is found in the tunnel loop only three residues before the beginning of  $\alpha 10$ .

The surface charge distribution of a single monomer shows some complimentary positively and negatively charged areas that are buried in the interacting surface between monomers. At the same time, some hydrophobic areas without charge are also present in the interactions (Figure 4.39C). Thus, both hydrophilic and hydrophobic contacts stabilize the protomer-protomer interactions.



## 4.6 Comparison with other portals

The dimensions of the T7 portal protein complex are of the same order as the ones from other portal proteins, except for the  $\phi 29$  one which is smaller and the height in the P22 protein, due to the presence of an elongated  $\alpha$ -barrel domain (Table 4.13).

**Table 4.13 Portal protein dimensions.** Summary of the particle dimensions of the different portal structures detailed in the introduction. T7 values are the experimental ones obtained from the crystallographic map. The channel diameter corresponds to the narrowest part.

Virus	Diameter (Å)	Height (Å)	Channel diameter (Å)
T7	170	110	23
$\phi 29$	146	75	35
SPP1	165	110	27
P22	170	300	25/35*
T4	170	120	28

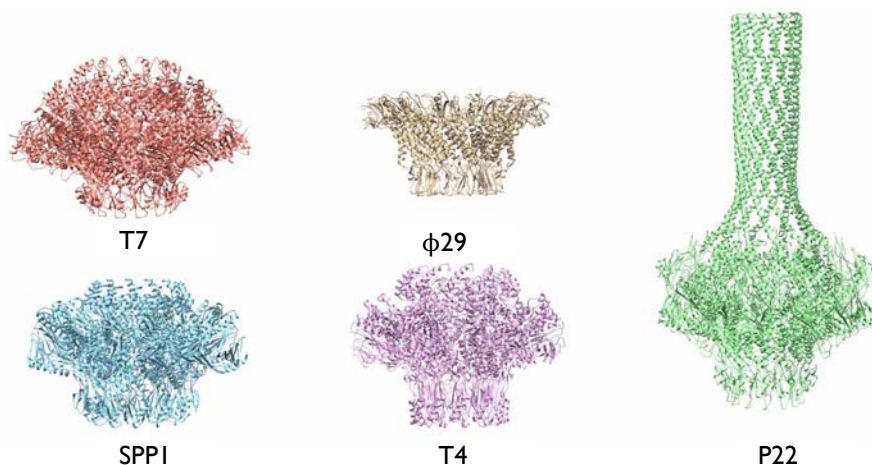
\*During packaging/in mature capsids, respectively.

Although the overall shape of all the portal proteins whose atomic structure is known consists of a ring with a central channel, they present many structural differences when they are compared with each other. The T7 bacteriophage portal protein is not an exception, and possesses some unique characteristics (Figure 4.40).

The  $\phi 29$  portal is the smallest one, and shows some similarities in terms of global shape with the clip, stem and bottom of the wing domains of the T7 portal. However, the T7 wing and crown domains have no counterpart on the  $\phi 29$  portal. On the other hand, the P22 portal is the biggest one, and its characteristic  $\alpha$ -barrel domain differentiates it from the rest of portals. It is not clear whether this domain is not present in the T7 portal protein or it is not built in the available model.

Therefore, at first sight, the SPP1 and T4 portals are the most similar ones to the T7 particle regarding the domain shape, and they also have similar particle sizes. However, they also present some differences: the T4 portal is taller and narrower than the other two and the T7 portal protein exhibits a unique conical shape on its wing domain, which does not share

with any other portal particle. The T7 and SPPI structures are the only ones which correspond to tridecameric assemblies, while the others are dodecameric particles.



**Figure 4.40** Comparison of bacteriophage portal particles.

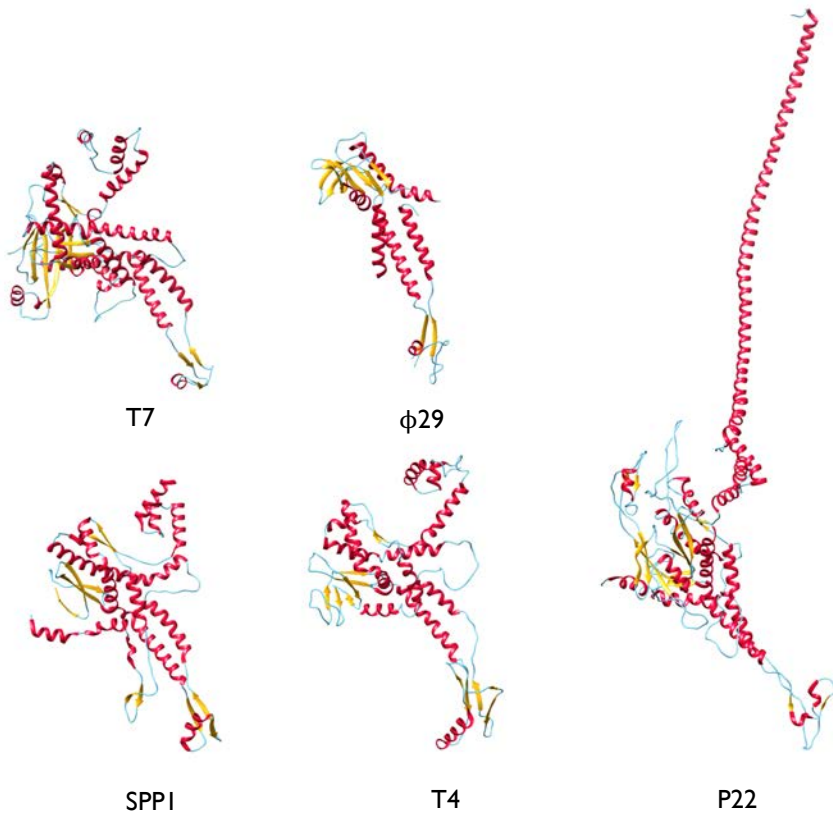
Lateral view of the available atomic structures in cartoon representation: T7 (red),  $\phi$ 29 (brown), P22 (green), SPPI (blue) and T4 (pink).

Structural comparison of a monomer with the available biological structures using the Dali and MATRAS servers retrieves, in first place, with significant scores, the known portal protein structures. Observation of the monomeric structures allows the comparison of each of the T7 portal protein domains with the equivalent domains on the other portal particles, in order to highlight their similitudes and differences (Figure 4.41):

- **Wing:** The T7 portal has a large distal  $\beta$ -sandwich on the bottom part of this domain, which would correspond to the one present on the  $\phi$ 29 wide-end, P22 hip, and SPPI and T4 wing domains. Therefore, it seems to be a conserved feature among portals. Structural similarity using the Dali and MATRAS servers with only this domain shows that the  $\beta$ -sandwich is similar to those found in some fatty acid binding proteins, ion exchangers or in myelin proteins. The presence of a protruding loop on the bottom part of the domain would also have a counterpart on P22 and SPPI, although in the latter case it is bigger and with an ordered secondary structure. All the structures display  $\alpha$ -helices

perpendicular to the channel, after the tunnel loop, equivalent to  $\alpha 10$ . However, in the other structures it is either shorter or kinked. All the portals have the N-terminal end in this domain.

- **Stem:** The helices that build the wall of the channel are a conserved domain. Similar domains are also present on the SPPI and T4 structures, and corresponding helices are found on the central domain of the  $\phi 29$  protein and on the leg domain of the P22 portal. In all the cases, the stem  $\alpha$ -helices are tilted. The tunnel loop connecting the stem with the wing  $\alpha 10$  is observed in all the structures except for the  $\phi 29$  protein, but a flexible loop facing the channel that could not be traced may be equivalent to this structure. Therefore, it seems to be a conserved feature.
- **Clip:** It is also quite conserved, consisting on both  $\alpha$ -helices and  $\beta$ -strands. The clip is also present in the SPPI and T4 structures and it is equivalent to the narrow-end domain of  $\phi 29$  portal and the bottom part of the leg domain of the P22 portal.
- **Crown:** The C-terminal domain displays the most differences, as it is not present at all in the  $\phi 29$  portal, and has the barrel shape in the P22 one. However, the crown domain of the T7 portal protein is similar to that of the T4 and SPPI proteins, consisting on three  $\alpha$ -helices that go towards the top of the complex, which are also found in the bottom part of the P22 barrel domain. In all the three cases the crown domain is separated from the wing domain by a cleft, which could allow certain mobility of the domain with respect to the rest of the protein. The presence of flexible areas that could not be traced after these helices in the SPPI, T4 and T7 structures also suggests the flexibility of this area. Secondary structure predictions show that this is mainly  $\alpha$ -helical, and it might become ordered forming an  $\alpha$ -barrel under certain physiological conditions. In a similar way to what has been described in the case of the P22 portal, when primary sequences of this area are checked they show a tendency to present repetitive amino acids (Gln in P22, Gln and Ala in T7, Asp and Glu in SPPI, Gln and Glu in T4). In all the cases, the C-terminal of the protein is located in this domain.



**Figure 4.41** Comparison of monomeric portal proteins.

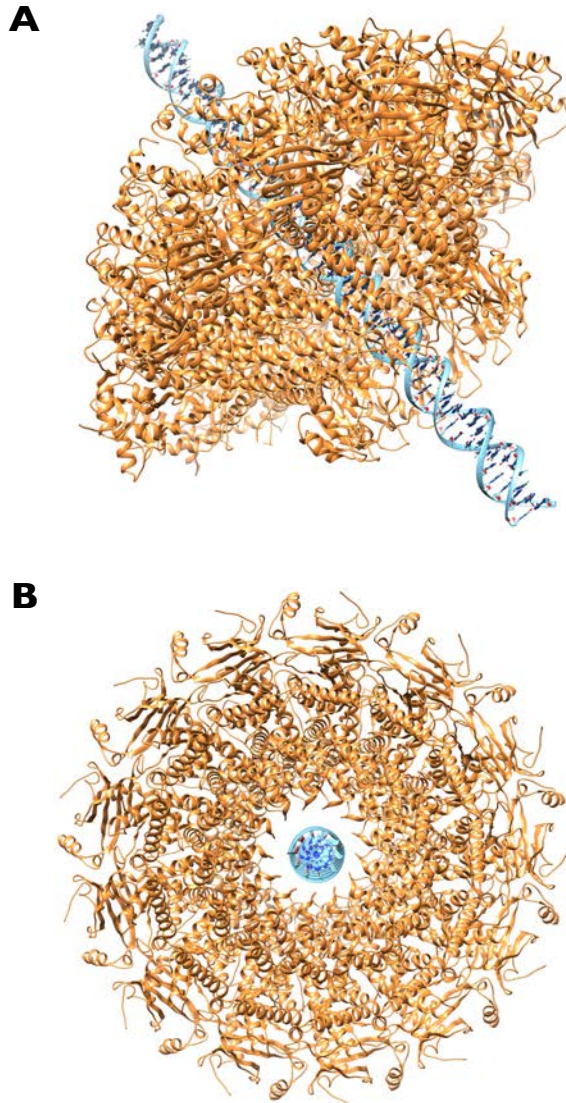
Available atomic structures in cartoon representation.

Helices appear in red,  $\beta$ -strands in yellow and coils in blue.

Analyzing the cavities created by the presence of the  $\alpha 10$  in the middle of the structure, the bottom one can be observed in all the cases. However, the upper one is not present in  $\phi 29$  and it is significantly smaller in the SPPI and T4 cases, due to a different angle of the crown domain with respect to the rest of the protein. Regarding the electrostatic potential of the channel surface, the T7 portal has the same features as the other particles: mainly an electronegative surface with positive rings (Figure 4.39B).

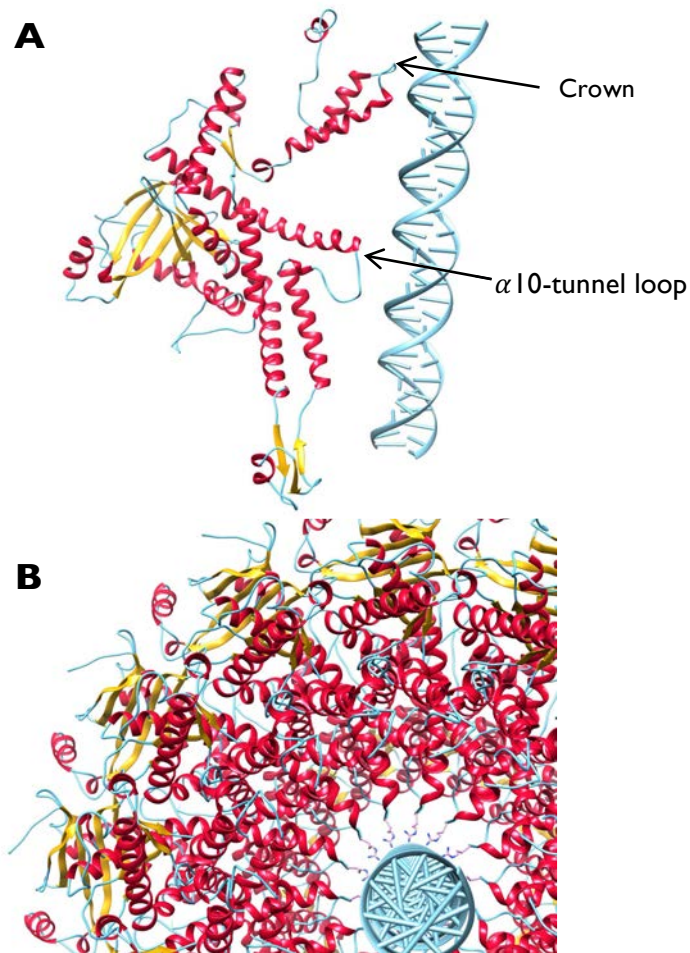
## 4.7 Functional model

A model of the protein-DNA complex during translocation was obtained by placing an idealized B-DNA molecule into the axial channel of the protein (Figure 4.42).



**Figure 4.42** Model of the portal protein with DNA during packaging. The protein structure is represented in cartoon (orange), while B-DNA is depicted as ribbons for the phosphate backbone and as rings for sugars and bases.

The model shows that DNA can pass through the channel, with possible protein-DNA contacts occurring in two points of the structure: in the crown domain and in the  $\alpha$ 10-tunnel loop region (Figure 4.43A). The closest contact between the DNA and the portal takes place at Arg368 from the  $\alpha$ 10-tunnel loop, which is located very close to the DNA phosphate backbone (Figure 4.43B).



**Figure 4.43 Model of interaction with DNA.**

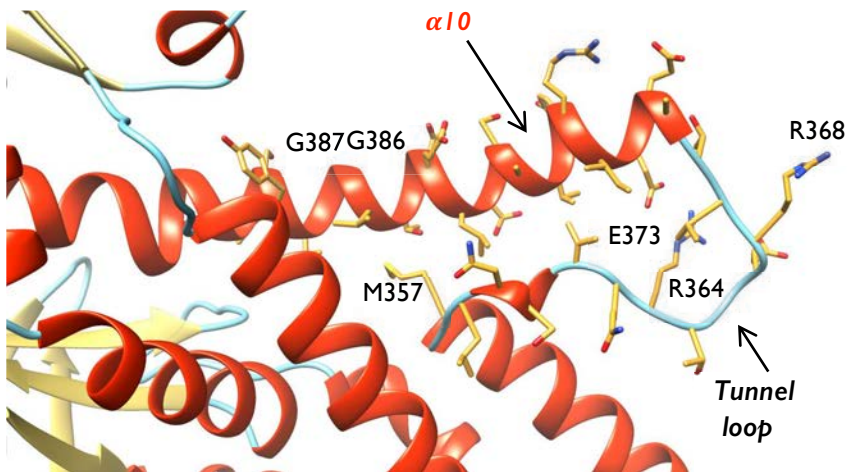
The protein structure is represented in cartoon (helices in red,  $\beta$ -strands in yellow and coils in blue). B-DNA is depicted as ribbons for the phosphate backbone and as ladders for sugars and bases.

(A) Interaction between DNA and one monomer.

(B) Axial view of the DNA-protein complex with tunnel loop Arg368 side-chain atoms shown pointing towards the channel.

Residue Arg368 faces the solvent channel, and it has not a defined density for part of its lateral chain. The diameter of 23Å has been measured from  $C\beta$  to  $C\beta$ , but depending on the disposition of the Arg368 side chains towards the channel the channel diameter may be reduced to 18Å. Therefore, portal conformations allowing and not allowing the passing of the DNA through the channel could both be possible. We propose that the disposition of this lateral chain may be directly involved in two of the functions of the T7 portal protein:

- During DNA translocation: It could interact directly with the negatively charge phosphate backbone of the DNA.
- After translocation: Depending on the disposition of the lateral chain, it may seal the channel. However, it has been described that after packaging the 5' end of the T7 genome, this remains into the tail, ready to be injected. In that case, Arg368 may have an important role stabilizing the 5' end of the genome into the tail, before its injection is triggered once a new bacterial cell is infected.



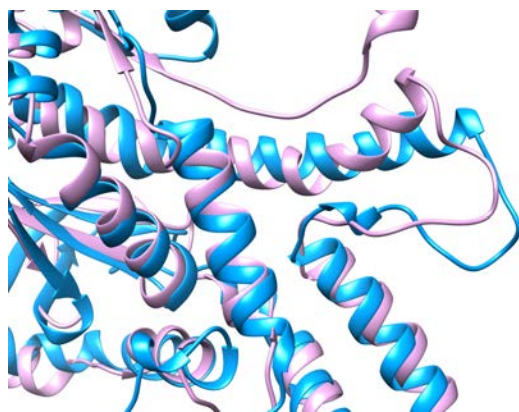
**Figure 4.44**  $\alpha 10$ -tunnel loop valve.

The protein structure is represented in cartoon (helices in red,  $\beta$ -strands in yellow and coils in blue). Side-chain atoms of the tunnel loop and the initial part of  $\alpha 10$  are also shown. Relevant residues for discussion are indicated.

It has also been hypothesized that the whole tunnel loop may have an important role both during DNA translocation and retention. Therefore, the conformational changes suggested may not only affect the Arg368 lateral chain, but the whole tunnel loop and the beginning of the  $\alpha 10$  area. In our structure, the tunnel loop is stabilized by a salt bridge between Arg364 and Glu373. However, some side chain densities are not well defined in the map, which indicates that it is a flexible loop. Therefore, we suggest that changes on the salt bridge may imply large movements in this region, that we call the  $\alpha 10$ -tunnel loop valve (Figure 4.44).

For the SPPI portal protein, it has also been observed that a residue from the tunnel loop induces a kink on  $\alpha 6$ , which is equivalent to the T7 portal  $\alpha$ -valve ( $\alpha 10$ ). In the case of our structure, the kink of the helix may occur on the Gly386 and Gly387 area, and may be induced by Met357. The two adjacent glycine residues in the middle of an helix may allow the kink of the helix, due to their conformational flexibility.

Structural data from our collaborators of the portal protein in complex with tail proteins seem to confirm the hypothesis about the role of the salt bridge and the two glycine residues (Cuervo and Carrascosa, unpublished data). When the salt bridge is disrupted, the position of the loop changes dramatically, and  $\alpha 10$  is kinked at the predicted area, towards the upper cavity. Models with an extended helix after the tunnel loop had been previously computed, but had not been observed experimentally. When the T7 and the SPPI portal proteins are superimposed, it can be observed that the helix after the tunnel loop is more extended in the T7 structure than in the SPPI model (Figure 4.45).



**Figure 4.45  $\alpha 10$ -tunnel loop valve movement.** Detail of the superposition of T7 and SPPI portal protein monomers represented in cartoon (T7 protein in blue, SPPI protein in pink). r.m.s.d. = 1.222Å



The T7 portal protein structure seems to support the hypothesis that the tunnel loop may have an active role both DNA translocation and retention.

Considering all the available information, we propose a model for the role of the T7 portal protein during DNA packaging similar to the one suggested for the SPPI system. However, according to what has been observed in the case of  $\phi 29$ , in this model the DNA and not the portal would rotate. DNA would interact with the flexible  $\alpha 10$ -tunnel loop valve region, which may have different conformations in order to interact with the genome and allow its translocation.

Therefore, the three steps occurring on each translocating cycle would be:

1. Conformational change of the  $\alpha 10$ -tunnel loop valve
2. Translocation and rotation of the DNA
3. ATP hydrolysis (by the terminase complex)

In this model, the terminase would provide the energy for the translocation and push the DNA through the portal particle pore, while the portal ring would accommodate its passage by adapting the  $\alpha 10$ -tunnel loop valve region to interact with the rotating DNA phosphate backbone.

As well as the indicated functions of the  $\alpha 10$ -tunnel loop valve and its role during packaging and translocation, other structure-function relationships can be hypothesized from our structural data, previous knowledge from other viral portal proteins, and additional data from our collaborators.

The crown domain would be involved in many interactions of the portal protein. Although in our structure it is a highly flexible region, it may get ordered when contacts between molecules are established. During procapsid formation, it may interact with the major capsid protein and/or the scaffolding protein. After that, it is supposed to be involved on the interaction with the core proteins, that assemble over the portal.

Regarding the wing, it is the most external domain, and it could be involved in interactions with the terminase complex. The flexible loop present on the bottom of this domain might be performing this contact.

Considering the location of the clip at the most external part of the portal respect to the procapsid, this domain probably interacts with the terminase during genome packaging. Once DNA has been packaged, this domain interacts with the gatekeeper protein during tail assembly. According to data from our collaborators, in the mature virions tail the gatekeeper monomers have a  $\beta$ -strand that interacts with the ones from the clip. The interaction portal-gatekeeper also involves the stem domain, as gpII monomers have a C-terminal  $\alpha$ -helix that embrace the stem domain of the portal (Cuervo and Carrascosa, unpublished data).



## Chapter 5:

# Conclusions



The conclusions of this thesis are:

1. The T7 bacteriophage portal protein crystallizes both in the dodecameric and tridecameric oligomeric forms.
2. A strategy that combined both X-ray crystallography and cryo-EM data was used in order to solve the tridecameric structure of the T7 bacteriophage portal protein at 2.8Å resolution.
3. Initial models obtained from medium-resolution cryo-EM maps can be used as MR ensembles in order to phase crystallographic data.
4. The T7 bacteriophage tridecameric portal protein is a ring-shaped particle 170Å tall and 110Å wide. The diameter of the inner channel varies from 23Å to 95Å, and helix  $\alpha 10$  protrudes into it defining two cavities above and below it.
5. The T7 portal protein has four domains: the wing, the stem, the clip and the crown. The tunnel loop connects the stem domain with  $\alpha 10$  from the wing, which protrudes into the channel, delimitating two cavities above and below it.
6. The side chain of Arg368 from the tunnel loop faces the channel, and it could interact with the DNA phosphates during translocation and may also be related with genome retention into the capsid after packaging.
7. Our structure shows the tunnel loop stabilized by a salt bridge and an extended conformation of the  $\alpha 10$ . Alternative conformations of the tunnel loop without the salt bridge are possible, and they may induce the kink of  $\alpha 10$ .
8. The  $\alpha 10$ -tunnel loop region may present different conformations related with the function of the portal protein during DNA translocation and retention after packaging

9. A mechanism for DNA packaging has been proposed, in which DNA translocation and rotation occur while the  $\alpha$ 10-tunnel loop valve interacts with the DNA, and ATP is hydrolysed by the terminase complex that pushes the genome.

# Bibliography

Afonine P.V., Headd J.J., Terwilliger T.C. and Adams P.D. (2013) *New tool: phenix.real\_space\_refine*. Computational Crystallography Newsletter **4**: 43-4.

Agirrezabala X., Martín-Benito J., Castón J. R., Miranda R., Valpuesta J.M. and Carrascosa J.L. (2005a) *Maturation of phage T7 involves structural modification of both shell and inner core components*. EMBO J. **24**: 3820-9.

Agirrezabala X., Martín-Benito J., Valle M., González J.M., Valencia A., Valpuesta J.M. and Carrascosa J.L. (2005b) *Structure of the connector of bacteriophage T7 at 8Å resolution: structural homologies of a basic component of a DNA translocating machinery*. J Mol Biol **347**: 895-902.

Agirrezabala X., Velázquez-Muriel J.A., Gómez-Puertas P., Scheres S.H., Carazo J.M. and Carrascosa J.L. (2007) *Quasi-atomic model of bacteriophage T7 procapsid shell: insights into the structure and evolution of a basic fold*. Structure. **15**: 461-72.

Ahi Y.S. and Mittal S.K. (2016) *Components of adenovirus genome packaging*. Front Microbiol. **7**: 1503.

Alonso J.C., Tavares P., Lurz R. and Trautner T.A. (2006) Bacteriophage SPPI. In Calendar R. (Ed.), *The bacteriophages*. 2<sup>nd</sup> Edition. UK: Oxford University Press. p. 331-49.

Al-Zahrani A.S., Kondabagil L., Gao S., Kelly N., Ghosh-Kumar M. and Rao V.B. (2009) *The small terminase, gp16, of bacteriophage T4 is a regulator of the DNA packaging motor*. J Biol Chem. **284**: 24490-500.

Baker M.L., Jiang W., Rixon F.J. and Chiu W. (2005) *Common ancestry of herpesviruses and tailed DNA bacteriophages*. J Virol. **79**: 14967-70.

Baines J.D. (2011) *Herpes simplex virus capsid assembly and DNA packaging: a present and future antiviral drug target*. Trends Microbiol. **10**: 606-13.

Bamford D.H., Grimes J.M. and Stuart D.I. (2005) *What does structure tell us about virus evolution?* Curr Opin Struct Biol. **15**: 655-63.

Baumann R.G., Mullaney J. and Black L.W. (2006) *Portal fusion protein constraints on function in DNA packaging of bacteriophage T4*. Mol Microbiol. **61**: 16-32.

Berndsen Z.T., Keller N. and Smith D.E. (2015) *Continuous allosteric regulation of a viral packaging motor by a sensor that detects the density and conformation of packaged DNA*. Biophys J. **108**: 315-24.



Bradford M.M. (1976) *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding*. *Analyt Biochem.* **72**: 248-254.

Brown J., McVoy M.A. and Homa F.L. (2002) Packaging DNA into herpesvirus capsids. In Bogner A.H.E. (Ed.) *Structure-function relationships of human pathogenic viruses*. USA: Kluwer Academic/Plenum Publishers. p. 111-55.

Brown A., Long F., Nicholls R.A., Toots J., Emsley P., Murshudov G. (2015) *Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions*. *Acta Crystallogr D Biol Crystallogr.* **71**: 136-53.

Brüssow H. (2005) *Phage therapy: the Escherichia coli experience*. *Microbiology.* **151**: 2133-40.

Burnley T., Palmer C.M. and Winn M. (2017) *Recent developments in the CCP-EM software suite*. *Acta Crystallogr D Struct Biol.* **73**: 469-77.

Casjens S.R. (2005) *Comparative genomics and evolution of the tailed-bacteriophages*. *Curr Opin Microbiol.* **8**: 451-8.

Casjens S.R. and Gilcrease E.B. (2009) Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. In Clokie M.R.J. and Kropinski A.M. (Eds.), *Bacteriophages: Methods and protocols, Volume 2: Molecular and applied aspects*. USA: Humana Press. **502**: 91-111.

Casjens S.R. (2011) *The DNA-packaging nanomotor of tailed bacteriophages*. *Nat Rev Microbiol.* **9**: 647-57.

Casjens S.R. and Thuman-Commike P.A. (2011) *Evolution of mosaicly related tailed bacteriophage genomes seen through the lens of phage P22 virion assembly*. **411**: 393-415.

Cerritelli M.E. and Studier F.W. (1996a) *Assembly of T7 capsids from independently expressed and purified head protein and scaffolding protein*. *J Mol Biol.* **258**: 286-98.

Cerritelli M.E. and Studier F.W. (1996b) *Purification and characterization of T7 head-tail connectors expressed from the cloned gene*. *J Mol Biol.* **258**: 299-307.

Cerritelli M.E., Cheng N., Rosenberg A.H., McPherson C.E., Booy F.P. and Steven A.C. (1997) *Encapsidated conformation of bacteriophage T7 DNA*. *Cell.* **91**: 271-80.

Cerritelli M.E., Conway J.F., Cheng N., Trus B.L. and Steven A.C. (2003) *Molecular mechanisms in bacteriophage T7 procapsid assembly, maturation, and DNA containment*. In Chiu W. and Johnson J.E. (Ed.), *Advances in protein chemistry: Virus structure*. USA: Academic Press. **64**: 301-23.

Chai S., Lurz R., Alonso J.C. (1995) *The small subunit of the terminase enzyme of bacillus subtilis bacteriophage SPP1 forms a specialized nucleoprotein complex with the packaging initiation region.* J Mol Biol. **252**: 386-98.

Chemla Y.R., Aathavan K., Michaelis J., Grimes S., Jardine P.J., Anderson D.L. and Bustamante C. (2005) *Mechanism of force generation of a viral DNA packaging motor.* Cell. **122**: 683-92.

Chen D.H., Baker M.L., Hryc C.F., DiMaio F., Jakana J., Wu W., Dougherty M., Haase-Pettingell C., Schmid M.F., Jiang W., Baker D., King J.A. and Chiu W. (2011) *Structural basis for scaffolding-mediated assembly and maturation of a dsDNA virus.* Proc Natl Acad Sci U S A. **108**: 1355-60.

Chen V.B., Arendall W.B.3rd., Headd J.J., Keedy D.A., Immormino R.M., Kapral G.J., Murray L.W., Richardson J.S. and Richardson D.C. (2010) *MolProbity: all-atom structure validation for macromolecular crystallography.* Acta Crystallogr D Biol Crystallogr. **66**: 12-21.

Chung Y.B., Nardone C. and Hinkle D.C. (1990) *Bacteriophage T7 DNA packaging. III. A "hairpin" end formed on T7 concatemers may be an intermediate in the processing reaction.* J Mol Biol. **216**: 939-48.

Condrón B.G., Atkins J.F. and Gesteland R.F. (1991) *Frameshifting in gene 10 of bacteriophage T7.* J Bacteriol. **173**: 6998-7003.

Cowtan K. (1994). "*dm*": *An automated procedure for phase improvement by density modification.* Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography. **31**: 34-8.

Cuervo A., Vaney M.C., Antson A.A., Tavares P. and Oliveira L. (2007) *Structural rearrangements between portal protein subunits are essential for viral DNA translocation.* J Biol Chem. **282**: 18907-13.

Cuervo A. and Carrascosa J.L. (2012a). *Bacteriophages: structure.* In eLS. UK: John Wiley & Sons, Ltd. p. 1-11.

Cuervo A. and Carrascosa J.L. (2012b) *Viral connectors for DNA encapsulation.* Curr Opin Biotechnol. **23**: 529-36.

Cuervo A., Daudén M.I. and Carrascosa J.L. (2013a) *Nucleic acid packaging in viruses.* In Mateu M.G. (Ed.) *Structure and physics of viruses.* Subcell Biochem Netherlands: Springer Netherlands. **68**: 361-94.

Cuervo A., Pulido-Cid M., Chagoyen M., Arranz R., González-García V.A., Garcia-Doval C., Castón J.R., Valpuesta J.M., van Raaij M.J., Martín-Benito J. and

Carrascosa J.L. (2013b) *Structural characterization of the bacteriophage T7 tail machinery*. J Biol Chem. **288**: 26290-9.

Cuervo A., Dans P.D., Carrascosa J.L., Orozco M., Gomila G. and Fumagalli L. (2014) *Direct measurement of the dielectric polarization properties of DNA*. Proc Natl Acad Sci U S A. **111**: E3624-30.

Daudén M.I., Martín-Benito J., Sánchez-Ferrero J.C., Pulido-Cid M., Valpuesta J.M. and Carrascosa J.L. (2013) *Latge terminase conformational change induced by connector binding in bacteriophage T7*. J Biol Chem. **288**: 16998-7007.

d'Hérelle F. (1917) *Sur un microbe invisible antagoniste des bacilles dysentériques*. Critical Reviews Academic Science Paris. **165**: 373.

de la Rosa-Trevín J.M., Quintana A., Del Cano L., Zaldívar A., Foche I., Gutiérrez J., Gómez-Blanco J., Burguet-Castell J., Cuenca-Alba J., Abrishami V., Vargas J., Otón J., Sharov G., Vilas J.L., Navas J., Conesa P., Kazemi M., Marabini R., Sorzano C.O. and Carazo J.M. (2016) *Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy*. J Struct Biol. **195**: 93-9.

de la Rosa-Trevín J.M., Otón J., Marabini R., Zaldívar A., Vargas J., Carazo J.M. and Sorzano C.O. (2013) *Xmipp 3.0: an improved software suite for image processing in electron microscopy*. J Struct Biol. **184**: 321-8.

Demerec M. and Fano U. (1945) *Bacteriophage-resistant mutants in Escherichia coli*. Genetics. **30**: 119-36.

Ding F., Lu C., Zhao W., Rajashankar K.R., Anderson D.L., Jardina P.J., Grimes S. and Ke A. (2011) *Structure and assembly of the essential RNA ring component of a viral DNA packaging motor*. Proc Natl Acad Sci U S A. **108**: 7357-62.

Dittmer A. and Bogner E. (2005) *Analysis of the quaternary structure of the putative HCMV portal protein pUL104*. Biochemistry **44**: 759-65.

Dixit A., Ray K, Lakowicz J.R. and Black L.W. (2011) *Dynamics of the T4 bacteriophage DNA packasome motor: endonuclease VII resolvase release of arrested Y-DNA substrates*. J Biol Chem. **286**: 18878-89.

Doerr A. (2016) *Single-particle cryo-electron microscopy*. Nature Methods. **13**: 23.

Dube P., Tavares P., Lurz R. and van Heel M. (1993) *The portal protein of bacteriophage SPPI: a DNA pump with 13-fold symmetry*. EMBO J. **12**: 1303-9.

Earnshaw W.C. and Casjens S.R. (1980) *DNA packaging by the double-stranded DNA bacteriophages*. Cell **21**: 319-31.

Edwards R.A. and Rohwer F. (2005) *Viral metagenomics*. Nat Rev Microbiol. **3**: 504-10.

Egli M. (2016) *Diffraction techniques in structural biology*. *Curr Protoc Nucleic Acid Chem.* **65**: 7.13.1-41.

Emsley P. and Cowtan K. (2004) *Coot: model-building tools for molecular graphics*. **60**: 2126-32.

Fu C.Y., Uetrecht C., Kang S., Morais M.C., Heck A.J., Walter M.R. and Prevelige P.E.Jr. (2010) *A docking model based on mass spectrometric and biochemical data describes phage packaging motor incorporation*. *Mol Cell Proteomics.* **9**: 1764-73

Fuller D.N., Raymer D.M., Kottadiel V.I., Rao V.B. and Smith D.E. (2007) *Single phage T4 DNA packaging motors exhibit large force generation, high velocity, and dynamic variability*. *Proc Natl Acad Sci U S A.* **104**: 16868-73.

Garcia-Doval C. and van Raaij M.J. (2012) *Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers*. *Proc Natl Acad Sci U S A.* **109**: 9390-5.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D. and Bairoch A. (2005) *Protein identification and analysis tools on the ExPASy server*. In Walker J. (Ed.), *The proteomics protocols handbook*. USA: Humana Press. p. 571-607.

Geng J., Fang H., Haque F., Zhang L. and Guo P. (2011) *Three reversible and controllable discrete steps of channel gating of a viral DNA packaging motor*. *Biomaterials.* **32**: 8234-42.

Goldner T., Hewlett G., Ettischer N., Ruebsamen-Schaeff H., Zimmermann H. and Lischka P. (2011) *The novel anticytomegalovirus compound AIC246 (Letermovir) inhibits human cytomegalovirus replication through a specific antiviral mechanism that involves the viral terminase*. *J Virol.* **85**: 10884-93.

Grassucci R.A., Taylor D.J. and Frank J. (2007) *Preparation of macromolecular complexes for cryo-electron microscopy*. *Nat Protoc.* **2**: 3239-46.

Green D.J., Wang J.C., Xiao F., Cai Y., Balhorn R., Guo P. and Cheng R.H. (2010) *Self-assembly of heptameric nanoparticles derived from tag-functionalized phi29 connectors*. *ACS Nano.* **4**: 7651-9.

Grimes S., Ma S., Gao J., Atz R. and Jardine P.J. (2011) *Role of the phi29 connector channel loops in late-stage DNA packaging*. *J Mol Biol.* **410**: 50-9.

Gual A., Camacho A.G. and Alonso J.C: (2000) *Functional analysis of the terminase large subunit, G2P, of Bacillus subtilis bacteriophage SPP1*. J Biol Chem. **275**: 35311-9.

Guasch A., Pous J., Ibarra B., Gomis-Rüth F.X., Valpuesta J.M., Sousa N., Carrascosa J.L. and Coll M. (2002) *Detailed architecture of a DNA translocating machine: the high resolution structure of the bacteriophage phi29 connector particle*. J Mol Biol. **315**: 663-76.

Guo F., Liu Z., Vago F., Ren Y., Wu W., Wright E.T., Serwer P. and Jiang W. (2013) *Visualization of uncorrelated, tandem symmetry mismatches in the internal genome packaging apparatus of bacteriophage T7*. Proc Natl Acad Sci U S A. **110**: 6811-6.

Hendrix R.W. (2002) *Bacteriophages: evolution of the majority*. Theor Popul Biol **61**: 471-80.

Hendrix R.W. (1978) *Symmetry mismatch and DNA packaging in large bacteriophages*. Proc Natl Acad Sci U S A. **75**: 4779-83.

Holm L. and Rosenström P. (2010) *Dali server: conservation mapping in 3D*. Nucleic Acids Res. **38**: W545-9.

Hu B., Margolin W., Molineux I.J. and Liu J. (2013) *The bacteriophage T7 virion undergoes extensive structural remodelling during infection*. Science. **339**: 576-9.

Hugel T., Michaelis J., Hetherington C.L., Jardine P.J., Grimes S., Walter J.M., Falk W., Anderson D.L. and Bustamante C. (2007) *Experimental test of connector rotation during DNA packaging into bacteriophage phi29 capsids*. PLoS Biol. **5**: e59.

Isidro A., Santos M.A., Henriques A.O. and Tavares P. (2004) *The high-resolution functional map of bacteriophage SPP1 portal protein*. Mol Microbiol. **51**: 949-62.

Kabsch W. (2010) *XDS*. Acta Crystallogr D Biol Crystallogr. **66**: 125-32.

Kantardjieff K.A. and Rupp B. (2003) *Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals*. Protein Sci. **12**: 1865-71.

Kawabata T. (2003) *MATRAS: a program for protein 3D structure comparison*. Nucleic Acids Res. **31**: 3367-9.

Keen E.C. (2015) *A century of phage research: bacteriophages and the shaping of modern biology*. Bioessays. **37**: 6-9.

Kemp P., Garcia L.R. and Molineux I.J. (2005) *Changes in bacteriophage T7 virion structure at the initiation of infection*. Virology. **340**: 307-17.

- Khurshid S., Saridakis E., Govada L. and Chayen N.E. (2014) *Porous nucleating agents for protein crystallization*. *Nature Protocols*. **9**: 1621-33.
- King J., Botstein D., Casjens S., Earnshaw W., Harrison S. and Lenk E. (1976) *Structure and assembly of the capsid of bacteriophage P22*. *Philos Trans R Soc Lond B Biol Sci*. **276**: 37-49.
- Kocsis E., Cerritelli M.E., Trus B.L., Cheng N. and Steven A.C. (1995) *Improved methods for determination of rotational symmetries in macromolecules*. *Ultramicroscopy*. **60**: 219-28.
- Koonin E.V., Senkevich T.G. and Dolja V.V. (2006) *The ancient Virus World and evolution of cells*. *Biol Direct*. **19**: 1-29.
- Kucukelbir F.J., Sigworth F.J. and Tagar H.D. (2014) *Quantifying the local resolution of cryo-EM density maps*. *Nat Methods*. **11**: 63-5.
- Kutter E., Gvasalia G., Alavidze Z. and Brewster E. (2013) Phage therapy. In Grassberger M., Sherman R.A., Gileva O.S., Kim C.M.H. and Mumcuoglu K.Y. (Eds.), *Biotherapy – History, principles and practice*. Netherlands: Springer Netherlands. p.191-213.
- Lang L.H. (2006) *FDA approves use of bacteriophages to be added to meat and poultry products*. *Gastroenterology*. **131**: 1370.
- Lawrence C.M., Menon S., Eilers B.J., Bothner B., Khayat R., Douglas T. and Young M.J. (2009) *Structural and functional studies of archaeal viruses*. *J Biol Chem*. **284**: 12599-603.
- Lebedev A.A., Krause M.H., Isidro A.L., Vagin A.A., Orlova E.V., Turner J., Dodson E.J., Tavares P., Antson A.A. (2007) *Structural framework for DNA translocation via the viral portal protein*. *EMBO J*. **26**: 1984-94.
- Lin H., Simon M.N. and Black L.W. (1997) *Purification and characterization of the small subunit of phage T4 terminase, gp16, required for DNA packaging*. *J Biol Chem*. **272**: 3495-501.
- Lindberg A.A. (1973) *Bacteriophage receptors*. *Annu Rev Microbiol*. **27**: 205-41
- Liu S., Chistol G., Hetherington C.L., Tafoya S., Aathavan K., Schnitzbauer J. Grimes S., Jardine P.J. and Bustamante C. (2014) *A viral packaging motor varies its DNA rotation and step size to preserve subunit coordination as the capsid fills*. *Cell*. **157**: 702-13.
- Lokareddy R.K., Sankhala R.S., Roy A., Afonine P.V., Motwani T., Teschke C.M., Parent K.N., Cingolani G. (2017) *Portal protein functions akin to a DNA-sensor that couples genome-packaging to icosahedral capsid maturation*. *Nat Commun*. **8**: 14310.

Lovell S.C., I.W. Davis, Arendall III W.B., de Bakker P.I.W., Word J.M., Prisant M.G., Richardson J.S. and Richardson D.C. (2003) *Structure validation by  $C\alpha$  geometry:  $\phi/\psi$  and  $C\beta$  deviation*. Proteins. **50**: 437-50.

Lurz R. Orlova E.V., Günther D., Dube P., Dröge A., Weise F., ven Heel M. and Tavares P. (2001) *Structural organisation of the head-to-tail interface of a bacterial virus*. J Mol Biol. **310**: 1027-37.

Madigan M.T., Martinko J.M., Stahl D.A. and Clark D.P. (2010). *Brock Biology of microorganisms* (13th edition) USA: Pearson **9**: 249-54.

Mao H., Saha M., Reyes-Aldrete E., Sherman M.B., Woodson M., Atz R., Grimes S., Jardine P.J. and Morais M.C. (2016) *Structural and molecular basis for coordination in a viral DNA packaging motor*. Cell Rep. **14**: 2017-29.

Matthews B.W. (1968) *Solvent content of protein crystals*. J Mol Biol. **33**: 491-7.

McCoy A.J., Grosse-Kunstleve R.W., Adams P.D., Winn M.D., Storoni L.C. and Read R.J. (2007) *Phaser crystallographic software*. J Appl Crystallogr. **40**: 658-74.

Moak M. and Molineux I.J. (2000) *Role of the Gp16 lytic transglycosylase motif in bacteriophage T7 virions at the initiation of infection*. Mol Microbiol. **37**: 345-55.

Molineux I.J. (2001) *No syringes please, ejection of phage T7 DNA from the virion is enzyme driven*. Mol Microbiol. **40**: 1-8.

Moore S.D. and Prevelige P.E.Jr. (2002) *Bacteriophage P22 portal vertex formation in vivo*. J Mol Biol. **315**: 975-94.

Mousset S. and Thomas R. (1969) *Ter, a function which generates the ends of the mature  $\lambda$  chromosome*. Nature. **221**: 242-4.

Murshudov G.N., Vagin A.A. and Dodson E.J. (1997) *Refinement of macromolecular structures by the maximum-likelihood method*. Acta Crystallogr D Biol Crystallogr. **53**: 240-55.

Nadal M., Mas P.J., Blanco A.G., Arnan C., Solà M., Hart D.J. and Coll M. (2010) *Structure and inhibition of herpesvirus DNA packaging terminase nuclease domain*. Proc Natl Acad Sci U S A. **107**: 16078-83.

Nemecek D., Gilcrease E.B., Kang S., Prevelige P.E.Jr., Casjens S. and Thomas G.J.Jr. (2007) *Subunit conformations and assembly states of a DNA-translocating motor: the terminase of bacteriophage P22*. J Mol Biol. **374**: 817-36.

- Nemecek D., Lander G.C., Johnson J.E., Casjens S.R. and Thomas G.J.Jr. (2008) *Assembly architecture and DNA binding of the bacteriophage P22 terminase small subunit*. *J Mol Biol.* **383**: 494-501.
- Nogales E. (2016) *The development of cryo-EM into a mainstream structural biology technique*. *Nat Methods.* **13**: 24-7.
- Olia A.S., Prevelige P.E.Jr., Johnson J.E. and Cingolani G. (2011) *Three-dimensional structure of a viral genome-delivery portal vertex*. *Nat Struct Mol Biol.* **18**: 597-603.
- Oliveira L., Alonso J.C. and Tavares P. (2005) *A defined in vitro system for DNA packaging by the bacteriophage SPP1: insights into the headful packaging mechanism*. *J Mol Biol.* **353**: 529-39.
- Oliveira L., Henriques A.O., Tavares P. (2006) *Modulation of the viral ATPase activity by the portal protein correlates with DNA packaging efficiency*. *J Biol Chem.* **281**: 21914-23.
- Padilla-Sanchez V., Gao S., Kim H.R., Kihara D., Sun L., Rossmann M.G. and Rao V.B. (2014) *Structure-function analysis of the DNA translocating portal of the bacteriophage T4 packaging machine*. **426**: 1019-38.
- Pajunen M.I., Elizondo M.R., Skurnik M., Kieleczawa J. and Molineux I.J. (2002) *Complete nucleotide sequence and likely recombinatorial origin of bacteriophage T3*. *J Mol Biol.* **319**: 1115-32.
- Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C. and Ferrin T.E. (2004) *UCSF Chimera– a visualization system for exploratory research and analysis*. *J Comput Chem.* **25**: 1605-12.
- Poteete A.R. and Botstein D. (1979) *Purification and properties of proteins essential to DNA encapsulation by phage P22*. *Virology* **95**: 565-73.
- Potterton E., Briggs P., Turkenburg M. and Dodson E. (2003) *A graphical user interface to the CPP4 program suite*. *Acta Crystallogr D Biol Crystallogr.* **59**: 1131-7.
- Rao V.B. and Feiss M. (2015) *Mechanisms of DNA packaging by large double-stranded DNA viruses*. *Anny Rev Virol.* **2**: 351-78.
- Rixon F.J. and Schmidt M.F. (2014) *Structural similarities in DNA packaging and delivery apparatuses in Herpesvirus and dsDNA bacteriophages*. *Curr Opin Virol.* **5**: 105-10.
- Robert X. and Gouet P. (2014) *Deciphering key features in protein structures with the new ENDscript server*. *Nucleic Acids Res.* **42**: W320-4.



- Rohou A. and Grigorieff N. (2015) *CTFFIND4: Fast and accurate defocus estimation from electron micrographs*. *J Struct Biol.* **192**: 216-21.
- Rosenberg A., Griffin K., Studier F.W., McCormick M., Berg J., Novy R. and Mierendorf R. (1996) *T7Select phage display system: A powerful new protein display system based on bacteriophage T7*. *Novation, Newsletter of Novagen, Inc.* **6**: 1-6.
- Rossmann M.G. and Blow D.M. (1962) *The detection of sub-units within the crystallographic asymmetric unit*. *Acta Cryst.* **15**: 24-31.
- Roy A., Bhardwaj A., Datta P., Lander G.C. and Cingolani G. (2012) *Small terminase couples viral DNA binding to genome-packaging ATPase activity*. *Structure.* **20**: 1403-13.
- Roy A. and Cingolani G. (2012) *Structure of P22 headful packaging nuclease*. *J Biol Chem.* **287**: 28196-205.
- Sankhala R.S., Lokareddy R.K. and Cingolani G. (2016) *Divergent evolution of nuclear localization signal sequences in herpesvirus terminase subunits*. *J Biol Chem.* **291**: 11420-33
- São-José C., de Frutos M., Raspaud E., Santos M.A. and Tavares P. (2007) *Pressure built by DNA packing inside virions: enough to drive DNA ejection in vitro, largely insufficient for delivery into the bacterial cytoplasm*. *J Mol Biol.* **374**: 346-55.
- Schmid M.F., Hecksel C.W., Rochat R.H., Bhella D., Chiu W. and Rixon F.J. (2012) *A tail-like assembly at the portal vertex in intact herpes simplex type-1 virions*. *PLoS Pathog.* **8**: e1002961.
- Scheres S.H.W. (2012) *RELION: Implementation of a Bayesian approach to cryo-EM structure determination*. *J Struct Biol.* **180**: 519-53.
- Selvarajan Sigamani S., Zhao H., Kamau Y.N., Baines J.D. and Tang L. (2013) *The structure of the herpes simplex virus DNA-packaging terminase pUL15 nuclease domain suggests an evolutionary lineage among eukaryotic and prokaryotic viruses*. *J Virol.* **87**: 7140-48.
- Serwer P. (2005) *T3/T7 DNA packaging*. In Catalano C.E. (Ed.), *Viral genome packaging machines: genetics, structure, and mechanism*. Molecular Biology Intelligence Unit. USA: Landes Bioscience. p. 59-79.
- Simpson A.A., Tao Y., Leiman P.G., Badasso M.O., He Y., Jardine P.J., Olson N.H., Morais M.C., Grimes S., Anderson D.L., Baker T.S. and Rossmann M.G. (2000) *Structure of the bacteriophage phi29 DNA packaging motor*. *Nature.* **408**: 745-50.

- Smith D.E., Tans S.J., Smith S.B., Grimes S., Anderson D.L. and Bustamante C. (2011) *The bacteriophage phi29 portal motor can package DNA against a large internal force*. *Nature*. **413**: 748-52.
- Smits C., Chechik M., Kovalvskiy O.V., Shevtsov M.B., Foster A.W., Alonso J.C. and Antson A.A. (2009) *Structural basis for the nuclease activity of a bacteriophage large terminase*. *EMBO Rep*. **10**: 592-8.
- Stetefeld J., McKenna S.A. and Patel T.R. (2016) *Dynamic light scattering: a practical guide and applications in biomedical sciences*. *Biophys Rev*. **8**: 409-27.
- Steven A.C., Trus B.L., Maizel J.V., Unser M., Parry D.A., Wall J.S., Hainfeld J.F. and Studier F.W. (1988) *Molecular substructure of a viral receptor-recognition protein. The gp17 tail-fiber of bacteriophage T7*. *J Mol Biol*. **200**: 351-65.
- Stroud R.M., Serwer P. and Ross M.J. (1981) *Assembly of bacteriophage T7. Dimensions of the bacteriophage and its capsids*. *Biophys J*. **36**: 743-57.
- Studier F.W. (1969) *The genetics and physiology of bacteriophage T7*. *Virology*. **39**: 562-74.
- Studier F.W. (1972) *Bacteriophage T7*. *Science*. **176**: 367-76.
- Studier F.W. and Dunn J.J. (1983) *Organization and expression of bacteriophage T7 DNA*. *Cold Spring Harbor Symp Quant Biol*. **47**: 999-1007.
- Studier F.W. (1991) *Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system*. *J Mol Biol*. **219**: 37-44.
- Sun L., Zhang X., Gao S., Rao P.A., Padilla-Sanchez V., Chen Z., Sun S., Xiang Y., Subramaniam S., Rao V.B. and Rossmann M.G. (2015) *Cryo-EM structure of the bacteriophage T4 portal protein assembly at near-atomic resolution*. *Nat Commun*. **6**: 7548.
- Sun S., Kondabagil K., Draper B., Alam T.I., Bowman V.D., Zhang Z., Hegde S., Fokine A., Rossmann M.G. and Rao V.B. (2008) *The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces*. *Cell*. **135**: 1251-62.
- Sun S., Gao S., Kondabagil K., Xiang Y., Rossmann M.G. and Rao V.B. (2012) *Structure and function of the small terminase component of the DNA packaging machine in T4-like bacteriophages*. *Proc Natl Acad Sci U S A*. **109**: 817-22.

Tang J., Lander G.C., Olia A.S., Li R., Casjens S., Prevelige P.Jr., Cingolani G., Baker T.S. and Johnson J.E. (2011) *Peering down the barrel of a bacteriophage portal: the genome packaging and release valve in P22*. *Structure*. **19**: 496-502.

Taylor G.L. (2010) *Introduction to phasing*. *Acta Crystallogr D Biol Crystallogr*. **66**: 325-338.

Trus B.L., Cheng N., Newcomb W.W., Homa F.L., Brown J.C. and Steven A.C. (2004) *Structure and polymorphism of the UL6 portal protein of herpes simplex virus type 1*. *J Virol*. **78**: 12668-71.

Tsuprun V., Anderson D. and Egelman E.H. (1994) *The bacteriophage phi 29 head-tail connector shows 13-fold symmetry in both hexagonally packed arrays and as single particles*. *Biophys J*. **66**: 2139-50.

Twort F.W. (1915) *An investigation on the nature of ultramicroscopic viruses*. *Lancet*. **2**: 1241.

Vagin A. and Teplyakov A. (2010) *Molecular replacement with MOLREP*. *Acta Crystallogr D Biol Crystallogr*. **66**: 22-5.

van Heel M., Orlova E.V., Dube P. and Tavares P. (1996) *Intrinsic versus imposed curvature in cyclical oligomers: the portal protein of bacteriophage SPPI*. *EMBO J*. **15**: 4785-8.

Veesler D. and Cambillau C. (2011) *A common evolutionary origin for tailed-bacteriophage functional modules and bacterial machineries*. *Microbiol Mol Biol Rev*. **75**: 423-33.

Walker J.E., Saraste M., Runswick M.J. and Gay N.J. (1982) *Distantly related sequence in the  $\alpha$ - and  $\beta$ - subunits of ATP synthase, myosin, kinases and other ATP requiring enzymes and a common nucleotide binding site*. *EMBO J*. **1**: 945-51.

Wendell D., Jing P., Geng J., Subramaniam V., Lee T.J., Montemagno C. and Guo P. (2009) *Translocation of double-stranded DNA through membrane-adapted phi29 motor protein nanopores*. *Nat Nanotechnol*. **4**: 765-72.

White J.H. and Richardson C.C. (1987) *Gene 18 protein of bacteriophage T7. Overproduction, purification, and characterization*. *J Biol Chem*. **262**: 8845-50.

Yang K., Homa F. and Baines J.D. (2007) *Putative terminase subunits of herpes simplex virus 1 form a complex in the cytoplasm and interact with portal protein in the nucleus*. *J Virol*. **81**: 6419-33.

Zhang Z., Kottadiel V.I., Vafabakhsh R., Dai L., Chemla Y.R., Ha T. and Rao V.B. (2011) *A promiscuous DNA packaging machine from bacteriophage T4*. *PLoS Biol*. **9**: e1000592.

Zheng S.Q., Palovcak E., Armache J.P., Verba K.A., Cheng Y. And Agard D.A. (2017) *MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy*. *Nat Methods*. **14**: 331-2.

## Web references

**Web 1.** Virus taxonomy: the classification and nomenclature of viruses. The online (10<sup>th</sup>) report of the International Committee on Taxonomy of Viruses, *International Committee on Taxonomy of Viruses*

<[https://talk.ictvonline.org/ictv-reports/ictv\\_online\\_report/](https://talk.ictvonline.org/ictv-reports/ictv_online_report/)>

**Web 2.** Merck company webpage. News: *Merck Announces Pivotal Phase III Study of Letermovir, an Investigation Antiviral Medicine for Prevention of Cytomegalovirus (CMV) Infections in High-Risk Bone Marrow Transplant Patients, Met Primary Endpoint*

<<http://www.mrknewsroom.com/news-release/corporate-news/merck-announces-pivotal-phase-3-study-letermovir-investigational-antivir>>

**Web 3.** FEI company webpage

<<https://www.fei.com>>