Membrane and Soluble Protein Structure Determination by cryo-TEM

Inauguraldissertation

Zur
Erlangung der Würde eines Doktors der Philosophie vorgelegt der
Philosophisch – Naturwissenschaftlichen Fakultät der Universität Basel

von

Ricardo Jorge Diogo Adaixo

aus Portugal

Basel, 2019

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Henning Stahlberg

Prof. Dr. Jan Pieter Abrahams

Basel, 15.10.2019

Prof. Dr. Martin Spiess, The Dean of the Faculty

In Memory of
Abel Adaixo

Table of Contents

Chapter I

1. Introduction

Proteins are biological polymers ubiquitous through all forms of life. Essential processes such as ion conduction, enzymatic catalysis, signal detection and transduction rely on proteins. Structural aspects of the cell such as the cell shape or the compact packing of DNA in a chromosome also require proteins.

While DNA carries the genetic information, which ultimately defines the response of a cell to any given event, virtually all processes in the cell depend on proteins to occur.

1.1. Protein Structure and Biogenesis

Proteins are polymers of small organic molecules called amino acids. Amino acids, as the name suggests, are molecules which possess amine (-NH2) and carboxyl (-COOH) functional groups. The carbon to which this functional groups are bound, also termed carbon alpha, is often a chiral center which allows amino acids to occur in enantiomeric forms L and D, however the L form is predominant and only in rare cases D enantiomers are found in proteins. Carbon alpha can establish two other chemical bonds typically to a hydrogen and the so-called side-chain or R group. Side chains can vary in hydrophobicity, acidity, charge, size and other properties therefore being the basis for the different chemical behavior of each amino acid but also their nomenclature. The 20 standard amino acids are depicted in figure 1.1

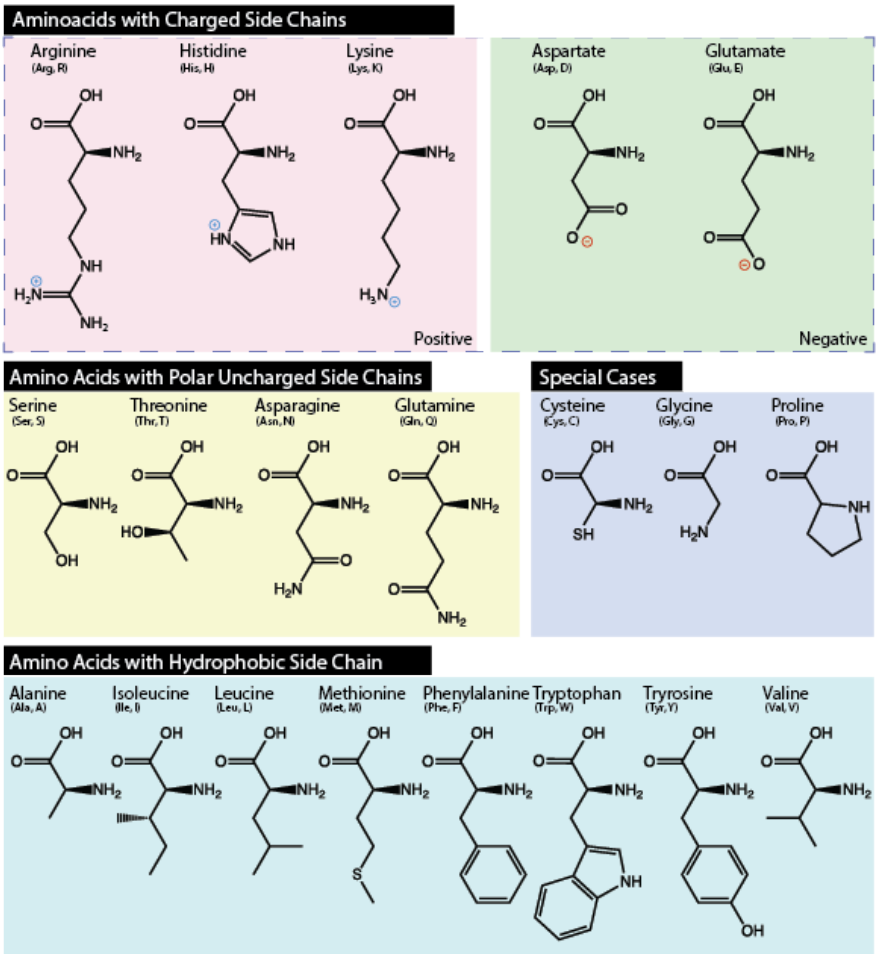

Figure 1.1: The standard L-amino acids categorized according to side chain properties.

During the translation process amino acids are covalently bound by peptide bonds of amino and carboxyl groups of adjacent amino acids forming polypeptides (Alberts *et al.*, 2017). Polypeptides present an amino and a carboxyl group which do not participate in peptide bond formation and define the N an C terminus of the amino acid chain.

The primary structure of a protein is simply defined by the sequence of amino acids. It does not consider the arrangement of amino acids of a protein in space. The secondary structure of a protein considers the formation of hydrogen bond networks between amino acids in the protein sequence at the backbone level. Alpha helices and beta sheets are common secondary structure elements that

arise from repetitive patterns of hydrogen bond networks. Other secondary structure elements with no defined repetitive pattern in hydrogen bonds are called loops.

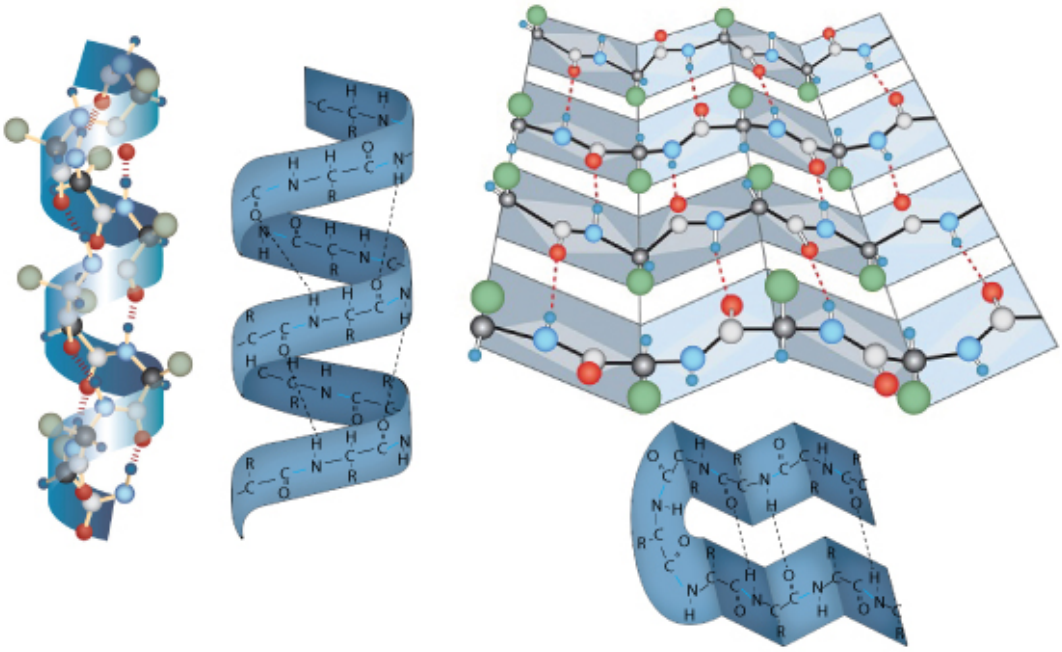

Figure 1.2: Ribbon and ball-and-stick representations of the alpha helix (left) and beta sheet (right) secondary structure motifs; hydrogen bonds are represented as dashed lines. Adapted from Alberts et al., 2017.

The tertiary structure of proteins comprehends the side chain interactions within a polypeptide and how these determine the orientation of the secondary structure elements relative to each other.

The quaternary structure of a protein describes how multiple chains of a protein interact to form the protein complex. In the context of multiple chain proteins, the subunits can be identical or not and the assembled complex may be symmetric or not. The formation of large protein complexes may answer a need of a specific shape (e.g.: a container, a filament, a ring, etc) but other reasons

such as efficiency gain or higher stability are equally valid (Goodsell and Olson, 2000).

The sequence of amino acids forming a protein is coded in DNA. During a process called transcription DNA acts as a template to synthesize messenger RNA. At the ribosome mRNA acts as a template for incoming transfer RNA coupled to amino acids, these will be covalently bound forming the nascent polypeptide chain.

1.1.2. Protein folding

The nascent polypeptide chain experiences multiple conformational states until the final native state of the mature protein is reached. This process, known as folding, is driven by:

(I) the formation of hydrogen bonds.

(II) hydrophobic interaction forces, amino acids with hydrophobic side chains will collapse to the interior of the protein being effectively shielded from the aqueous solvent.

(III) (III) charge compensation, amino acids with charged side chains will be stabilized by formation of salt bridges within the polypeptide chain or by pairing with ions in the aqueous solvent.

The free energy of the protein is reduced from the state of unstructured polypeptide to the final folded state, however the conformational entropy is also reduced during the folding process, this being the major counteracting force during protein folding (Mallamace *et al.*, 2016). The folding process is assisted by a group of proteins called chaperones, these prevent premature aggregation of the nascent chain therefore increasing the probability that the global energy minimum of the folding process is reached.

1.1.3. Post-translational modifications (PTM)

Proteins often exhibit structural features that are not simply the sum of its constituent amino acids. Such features are generally associated with function and

regulation of activity. The most common PTMs involve the covalent addition of a functional group such as a phosphate, an acetyl, a glycosyl or even the bulkier ubiquityl. Other PTMs such as disulfide bond formation, or peptide isomerization do not add functional groups but rather chemically rearrange those present in the protein.

1.1.4. Membrane proteins

The cell can be understood as a compartmented vessel whose boundaries are lipid membranes. Unlike soluble proteins, membrane proteins perform their function embedded in membranes such as channels, receptors or some enzymes. The surface exposed by membrane proteins to the lipidic environment is also hydrophobic.

Currently, the majority of FDA approved drugs target different classes of membrane proteins which motivates further research on them (https://www.proteinatlas.org/humanproteome/tissue/druggable).

The study of membrane proteins is, in most cases, difficult relatively to soluble proteins due to their low expression levels and the necessity of having detergent or other membrane mimicking systems present in the sample (Thonghin *et al.*, 2018).

Despite the obstacles the number of structures of membrane proteins and respective drug complexes is growing at almost exponential rate.

1.1.5. Protein structure determination

Structural biology aims at determining the three-dimensional structure of biological macromolecules and establish relations between the obtained spatial models and the functionality of the studied target.

The information that can be deduced from a protein model is limited by the resolution of such model: at resolutions of ~8 to 15 Å the general domain architecture of a protein can be described, at ~7 Å alpha helices can be identified

as rods, at ~5 Å beta sheets become apparent and at ~3 Å side chains can be assigned.

Currently the most common techniques to obtain high resolution data of a protein target are nuclear magnetic resonance, X-ray diffraction and cryogenic electron microscopy. All these methods require high sample purity and conformational homogeneity which can take several years in some cases.

In X-ray diffraction crystals of the purified protein are prepared typically by raising the concentration of protein to supersaturation. Mild precipitants such as salts and polymers are also used for this purpose. Other factors such as pH, temperature and the presence of additives also influence the formation of protein crystals as well as their final quality. The crystallization process is in essence a trial and error approach since there are no methods to predict which conditions will favor crystal nucleation and growth for a given protein. In practice extensive screening in miniaturized and automatized pipelines is used to obtain protein crystals. Optimization of the initial crystallization conditions is often required to improve the quality of the crystals (McPherson and Gavira, 2014). The very bright and coherent electron beam of a synchrotron accelerator is the most usual X-ray source to collect high resolution data on protein crystals. In order to acquire X-ray diffraction data typically crystals are placed on small loop holders, snap frozen and exposed to the X-ray beam path under a constant cooling gas (usually nitrogen) and snapshots of diffracted X-rays are recorded for different orientations of the crystal. The fact that the crystal is kept at cryogenic temperatures during X-ray exposure minimizes radiation damage and ultimately improves the quality of the recorded data. An electron density map of the crystals' unit cell can be computed from the recorded diffraction patterns. An additional obstacle in X-ray diffraction of protein samples comes from the fact that the collected diffraction patterns lack phase information (these can be understood as Fourier transforms of the unit cell and as such both amplitudes and phases are required in order to recover the real space object, i.e. the density

map within the unit cell). Phases can be obtained by a variety of methods a discussed elsewhere (Taylor, 2010). Very high-resolution density maps can be obtained by X-ray diffraction and to date the vast majority of published protein structures is derived from this method.

Nuclear magnetic resonance is a technique originally used in analytical chemistry and later applied in other fields such as medical imaging or structural biology (Wüthrich, 1991; van Geuns *et al.*, 1999). In NMR, unlike cryo-EM or XRD, the recorded signal is derived from nuclei in the sample which have a non-zero spin number (i.e. 1H or 13C are NMR observable while 2H or 12C are not). In the presence of a strong magnetic field the nuclear spins of atoms have no longer random orientations but instead align with the magnetic field, in this context and considering a ½ spin atom two energetic levels are possible: spin up and spin down, and their energy difference is proportional to the strength of the applied magnetic field. Because the energetic difference between the two mentioned energy levels is small the Boltzmann distribution at room temperature and thermal equilibrium predicts both states are almost equally populated. In fact, the higher energetic state as an excess of less than 0.0001% of the total number of nuclear spins and these are the nuclei which will be observed during the spectroscopy experiment. For this reason, the sensitivity of NMR experiments is inherently low. Radio frequency pulses are applied to the sample in the NMR magnet and the precession frequency of the spin of the observable nuclei is recorded also in the form of a radio frequency. The precession frequency is proportional to the described energetic difference and this is influenced by the chemical environment of any given nucleus. The variety and number of NMR experiments that can be done with proteins is colossal, in comparison with cryo-EM or XRD; for the study of protein's structure NOE experiments can deduce inter atomic distances therefore providing a basis for atomic modelling.
Despite not requiring crystallization protein samples used in NMR experiments are often produced in special cell growth media to enrich the proportion of NMR

active nuclei, a method called isotope labelling. Optimizing growth and protein expression levels in such growth media can be tedious and very expensive depending on the chosen labelling scheme. Protein stability is a limiting factor in NMR spectroscopy since experiments will typically require the sample to remain at room temperature (or close to) for hours up to days. Protein size is also a limiting factor with the current NMR techniques being capable of determining structures of proteins smaller than 50 kDa.

Cryogenic electron microscopy (cryo-EM) is a technique where a thin frozen hydrated sample is imaged by exposure to an electron beam. Cryo-EM is suitable to study a wide variety of samples from small proteins to protein complexes, crystals, fibers, viruses, cells or even tissue sections (Murata and Wolf, 2018). Freezing samples in amorphous ice for cryo-EM was developed in the 1980's by Jacques Dubochet and co-workers providing a solution to image biological samples in a water-based environment in the vacuum of the electron microscope (Adrian *et al.*, 1984). Rapid freezing of the sample and imaging at cryogenic temperatures also minimized the radiation damage to which most biological samples are highly sensitive. Joachim Frank developed algorithms capable of aligning and averaging cryo-EM 2D images and reconstructing the corresponding particle 3D volume (Frank, 2006). Richard Henderson did theoretical predictions on the advantages of using electrons and direct electron detectors to study biological samples (Henderson, 1995; Faruqi and Henderson, 2007). The development of these methods was acknowledged by the Nobel prize in chemistry award in 2017 "for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution"[1]. The availability of complementary metal-oxide-semiconductor (CMOS) direct electron detectors provided means to acquire data in movie mode with higher sensitivity than charge-coupled device (CCD) cameras. Imaging samples in movie mode allows

[1] The Nobel Prize in Chemistry 2017. NobelPrize.org. Nobel Media AB 2019. Sat. 15 Jun 2019

deblurring of the final movie averages, (i.e. compensation of the movement experienced by specimen during irradiation), as well as the implementation of dose weighting schemes which account for accumulation of damage in the sample during irradiation (Grant and Grigorieff, 2015). Coincident with the advent of direct electron detectors was the implementation of maximum likelihood methods to cryo-EM image processing (Sigworth *et al.*, 2010). Altogether, sample vitrification techniques, direct electron detectors and robust and reliable image processing algorithms are at the basis of the current success of cryo-EM in the structural biology field.

1.2. Image Formation and 3D Object Reconstruction in TEM

The first electron microscope was conceived by Ernst Ruska in 1933. Theoretically this device would allow imaging with resolution in the Å range by simply considering Bragg's law and the wavelength of electrons accelerated to hundreds of kilovolts. Experience proved many rounds of iteration over the very first electron microscope were needed in order to obtain atomic resolution (e.g.: higher coherence electron sources, better mechanical stability of the microscope, better performing lenses; Chiu and Glaeser, 1976).

In TEM a highly coherent electron beam is focused to a thin (tens of nanometers) specimen by means of electromagnetic lenses and the resulting transmitted electrons are focused again in order to generate a magnified image of the specimen also called micrograph. The interaction between specimen and the incoming electrons has different discrete outcomes namely:

(I) no interaction, the incoming electron is not scattered by the specimen.

(II) elastic scattering, the incoming electron interacts with the sample and is transmitted without losing energy.

(III) inelastic scattering, the incoming electron interacts and transfers energy to the specimen.

Inelastic scattering events transfer energy to the specimen which is sufficient to ionize atoms and break covalent bonds, therefore low dose setups are essential for imaging biological samples.

Image formation in TEM is the result of the interference between the un-scattered electrons and the elastically scattered electrons. Such interference generates contrast at the image plane which reflects the integrated Coulomb potential of the illuminated specimen area. Inelastically scattered electrons translate into noise at the image plane (Frank, 2006).

For every scattering event about three inelastic scattering events occur, consequently the longer the electron path across the specimen the more likely it is that a single electron experiences inelastic scattering augmenting the noise in the image. The magnitude of the mentioned scattering events is proportional to the atomic number of the scattering center in the specimen. In the context of biological samples, predominantly composed of C, H and O atoms, the scattering magnitude is marginally higher than that of the surrounding aqueous media which results in low signal to noise ratio (SNR).

1.2.1. Contrast Transfer Function

In order to increase contrast images can be collected with a given amount of defocus. This is a rudimentary form of phase contrast imaging. Defocusing, imperfections in electromagnetic lenses and the interference of electrons past the specimen modulate image formation in what we describe as contrast transfer function (CTF). The CTF is a reciprocal space function whose real space counterpart is the point spread function (PSF). It has an amplitude component and a phase component, for thin biological samples the phase component has about 10 times the magnitude of the amplitude component. The CTF depends on several factors such as the wavelength of the electron beam, the applied defocus and the spherical aberration of the electron microscope lenses. In practice a CTF curve displays the phase sign and amplitude of each spatial frequency in the recorded image. The CTF of a typical TEM micrograph with non-zero

defocus is a sinusoidal curve, plotted as function of spatial frequency, with decaying amplitude as higher spatial frequencies are considered. The mentioned decay is not predicted by the CTF itself but results from other factors such as the nature of scattering events at higher angles (discussed elsewhere; Mindell and Grigorieff, 2003) and the modulation transfer function of the camera. As a result of the sinusoidal shape of the CTF some frequencies will be under represented or zeroed, therefore it is common practice to acquire cryo-EM micrographs at different defocus values so that all frequencies are effectively sampled.

1.2.2. Single Particle Reconstruction

Single particle reconstruction (SPR) is a method which allows the construction of a 3D map of an object from its individual projections. At the core of SPR lies the Fourier slice theorem which states: if an image P is obtained by projecting a 3D object V in a given direction d, then the FT of P is a slice of the FT of V which passes through the origin and is perpendicular to the d vector.

The idea is then to determine the d vectors for each P image, do the FT of P and populate the 3D reciprocal space with the obtained image, and finally reverse FT the reciprocal space object to obtain the real 3D object V.

Calculations in reciprocal space are much faster than in real space and another speed up factor comes from being able to consider varying ranges of resolution by ignoring the outermost shells of the reciprocal space volume for example or even something as simple as using polar coordinates instead of cartesian.

The difficulty in SPR is to determine the d vector, or in other words, the set of shifts and angles that relate the projection to the corresponding 3D map. Projection matching (Frank, 2006) was one of the first methods to solve this image alignment problem in cryo-EM with the caveats of being computationally very demanding and being prone to reference bias. More recently maximum-likelihood methods are extensively used to determine orientations of individual projections in SPR. These are less prone to reference bias and instead of relying on a “hard assignment” of each projection to a given d vector the assignment is

done in a probabilistic manner so that the same projection can in fact contribute to multiple slices of the FT of V (Sigworth *et al.*, 2010).

Typically, SPR is performed with a low-resolution map of the object, to avoid reference bias, and increasing the considered number of shells in the reciprocal space volume over iterations (which is equivalent to consider finer sampling of shifts and angles of the projections in real space) until convergence has been reached.

1.3. Cryo-EM Sample Preparation

Sample preparation in cryo-EM does not require crystallization *a priori* nor isotope labelling and can be done with more diluted aliquots than those typically used in XRD or NMR. Given that the right conditions for freezing have been determined a full dataset can be obtained from a few microliters of sample. State-of-the-art methods are pushing the required volumes to the low nanoliter range (Arnold *et al.*, 2017; Schmidli *et al.*, 2019).

Prior to imaging, samples are transferred to a foil supported by a metallic grid. This metal is usually copper although others such as platinum, molybdenum or gold can also be used. The foil is often made of amorphous carbon or gold. Nowadays the diversity of EM grid configurations commercially available is enormous presenting features such as additional graphene or graphene oxide coatings, surface derivatization or self-blotting properties.

Unless freshly produced the amorphous carbon foil of EM grids is hydrophobic. In order to adsorb biological samples in aqueous buffers the grid's surface is rendered hydrophilic prior to sample application. This can be achieved by UV treatment, cleaning with organic solvents or plasma cleaning. The latter is the most usual process and consists of exposing the grid to a plasma in a vacuum chamber where different gases are introduced. The nature and flux of the introduced gases as well as the plasma power and duration tune the surface properties of grids. Gas mixtures such as air, Argon/Hydrogen or Oxygen/Hydrogen are commonly used for plasma cleaning of EM grids.

Cryo-EM grids typically require 2 to 4 µL of biological sample to be deposited into holey EM grids, the excess liquid is blotted resulting in a thin film within the holes of the foil and the grid is rapidly plunge frozen into liquid ethane at liquid nitrogen temperature. After the blotting step the resulting liquid film is prone to evaporation; to minimize this effect, blotting is performed in temperature and humidity-controlled chambers within semi-automated plunge freezing apparatus (such as the FEI Vitrobot or the Leica Plunger). Plunge freezing devices increase reproducibility during sample preparation, by allowing the user to preset parameters such as blotting time and blotting force, and also minimize sample contamination which may happen from exposure to humidity in the air.

In order to obtain high quality frozen specimens (I) the resulting vitreous ice layer must be as thin as possible still accommodating the particles of interest, (II) the concentration of salts, glycerol or sugars in the sample buffer must be minimized since they reduce specimen's contrast, (III) the frozen grid must be devoid of contaminations and (IV) the particles must preserve their native structure during the freezing process ideally adopting random orientations in the vitreous ice. The last point is particularly difficult to fulfill since majority of proteins denature or degrade when exposed to the air-water interfaces in the film (Glaeser, 2017). The disposition of particles in preferred orientations rather than random orientations is also a common problem in cryo-EM sample preparation usually attenuated by carefully tweaking sample buffer composition or by introducing small concentrations of detergents in the sample (Drulyte *et al.*, 2018). Frozen grids with biological samples are stored in liquid nitrogen containers, at temperatures lower than 100 K, until they are imaged, also at cryogenic temperatures.

Robust and reproducible sample preparation is currently considered one of the most significant limitations in cryo-EM.

1.4. Cryo-EM methods

Cryo-EM offers a multitude of experimental setups to retrieve structural information of biological samples. Electron Tomography, 2D crystallography, helical reconstruction and single particle analysis are in essence techniques under the "cryo-EM umbrella".

In 2D crystallography samples, usually membrane proteins in the presence of lipids, are arranged in a 2D periodic lattice. The electron microscope is set up such that the camera records the diffraction plane instead of the image plane. Micrographs are collected at different tilt angles in order to recover as much as possible form the 3D reciprocal space. The diffraction patterns from several 2D crystals can then be merged, indexed and Fourier inverted to generate a real space map. Analogous to XRD no phase information is collected in the mentioned setup, so phases have to be retrieved using complementary methods to perform the Fourier inversion step. 2D crystallography can also be performed in imaging mode where phases are recorded but images will have lower SNR and will be CTF modulated. Another limitation comes from the fact that sample stages of electron microscopes have a limited tilt range, usually -60 to +60 degrees, and as result the reciprocal space can only be partially recorded. This limitation was branded as "the missing wedge problem" and is common to 2D crystallography ("missing cone") and tomography ("missing wedge") techniques.

In tomography micrographs of an area of interest in the specimen are successively recorded with different sample stage tilt angles. The set of acquired images, also known as tilt series, is then compiled and aligned to reconstruct the imaged volume, the obtained 3D reconstruction is referred to as tomogram. It is common practice to spike tomography samples with colloidal gold, as the metallic clusters act as markers to aid tilt series alignment. Tomography is particularly well suited to cell and tissue samples (Baumeister, Grimm and Walz, 1999) but relatively

large protein complexes can also be studied by tomography using subtomogram averaging (Schur *et al.*, 2016).

Helical reconstruction is a method used to reconstruct 3D maps of helical assemblies such as filaments and helical viruses. Such helical assemblies are formed by repeating units geometrically related by a shift in z height, the rise, as well as an angle, the twist (Fromm and Sachse, 2016). One single filament provides many subunits in different orientations which allows structure determination from relatively small datasets. The first map derived from helical reconstruction was the tail of the bacteriophage T4 in 1968 (DeRosier and Klug, 1968), this was possible in part because the computational power requirements to perform helical reconstruction are minimal in comparison to other reconstruction methods like single particle analysis. Recent examples of helical reconstruction methods applied to biological samples with high clinical interest include Fitzpatrick *et al.*, 2017 on tau filaments or Guerrero-Ferreira *et al.*, 2018 on alpha-synuclein fibrils.

Single particle analysis (SPA) is probably the most widespread cryo-EM technique to determine protein structures today. SPA sample preparation was discussed in point 1.3: essentially a homogeneous and highly pure protein sample is vitrified resulting in randomly oriented particles distributed in a thin layer of vitreous ice. A transmission electron microscope operated at cryogenic temperatures is used to collect hundreds to thousands of movies in a low dose setup. It is common practice to use data collection automation software at this stage such as EPU (Thermo Fischer company) or SerialEM (Mastronarde, 2005) as these greatly enhance the throughput of data collection allowing the electron microscope to operate continuously with minimal user intervention. Movies in a dataset are then (I) drift corrected, to account for beam induced motion and stage drift, (II) dose weighted, so that high resolution information is better represented in the final movie aligned average, and (III) CTF is estimated for

the non-dose weighted aligned movie averages (discussed in point 1.2.1.). The output log files of the drift correction and CTF estimation provide evaluation parameters to prune the dataset, it is common practice to exclude movies with high drift profile, poor CTF fit resolution or high astigmatism.

Individual particle locations have to be determined for each aligned average, a process termed particle picking. Different software packages are available to perform automated particle picking such as Gautomatch (developed by Dr Kai Zhang, MRC Laboratory of Molecular Biology, Cambridge, UK, http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/) or crYOLO (Wagner *et al.*, 2019).

The extracted particles have low SNR which prevents identification of high-resolution features or discrimination of "good" from "bad" particles based on single images. However, it is possible to infer over an ensemble of similar particles which have been classified, aligned and averaged so that the SNR of the resulting image is improved. This clustering of particles is known as 2D classification and it is similar to SPR except in the later template projections are all necessarily consistent with a single 3D object whereas in 2D classification templates can (and often will) be independent from each other.

2D class averages allow the user to discriminate between true and false positives obtained at the particle picking stage. Furthermore, after 2D classification one can estimate how homogeneous the dataset is, detect flexibility among domains or identify possible preferential orientation problems. It is common to perform consecutive rounds of 2D classification to prune a dataset.
Classification in 3D uses the same principles described for SPR with the difference that each particle will have an additional parameter assigned, the probability of belonging to each of the 3D templates. Initial 3D seeding models can be obtained from ab-initio methods, for example using the common lines approach

implemented in IMAGIC (Van Heel, 1987) or from similar deposited structures if available.

Once a 3D class with clear secondary structure elements is obtained it can be further refined. During 3D refinement each particle's geometric parameters (x and y shift, phi and psi angles) will be optimized, typically using finer sampling than 3D classification, hopefully resulting in a higher resolution map. It is common practice to limit the resolution content of the raw extracted particles and of the reconstructed map at the end of each iteration according to the estimated resolution of the reconstructed map itself, as this approach prevents noise overfitting.

More recently the refinement of parameters other than geometric is also available in SPA processing software such as RELION-3, these include CTF and beam tilt refinement as well as per-particle drift correction (Zivanov *et al.*, 2018).

Projections of the obtained 3D refined model can be used to generate more accurate 2D templates for automated particle picking if needed.

1.4.1. Resolution Estimation

Estimating resolution on density maps obtained as described in the previous sections is still to date matter of debate. A tempting easy solution would be to derive this estimate directly from the aligned movie averages somehow mimicking the XRD approach where this estimation is derived from the highest resolution reflections considered for structure determination. In SPR however the final resolution of the reconstructed map is highly dependent on the alignment accuracy of the particles. In the case of small proteins (<200 kDa) without prominent features or with high degree of flexibility the micrographs may have high resolution information judging by the outermost ring of the power spectra that is effectively considered by the fitted CTF, but, because of the difficulty of aligning each particle to the 3D reference the reconstructed map may have high angle and shift alignment errors which results in a map with lower resolution than the raw data actually suggests.

Fourier Shell Correlation (FSC) is the most widespread method to estimate resolution of a map obtained by SPR (Harauz and van Heel, 1986). In essence this approach requires that the particles are randomly divided in two groups and independently refined, at least beyond a given resolution cutoff. This resolution cutoff is used simply to avoid diverging orientations of the refined half-maps. The correlation between the refined half-maps can then be plotted as a function of spatial frequency. Theoretical considerations, out of the scope of this section, determine that the nominal resolution of the reconstructed map is the value at which the FSC curve is 0.143 (Rosenthal and Henderson, 2003; van Heel and Schatz, 2005).

It is important to point out that despite being the current standard method for resolution estimation in the cryo-EM field the FSC is in reality a measure of consistency between two maps, typically calculated from random halves of the data. Furthermore, it is prone to inflation due to masking or reference bias (Chen *et al.*, 2013).

Cryo-EM maps can be anisotropic regarding resolution meaning not all regions in a given map are equally well resolved. Three main factors account for this anisotropy: first, the fact that the core of most proteins is better ordered than the solvent exposed surfaces resulting in a radial decay of resolution in the map; second, the fact that proteins often rely on flexible regions to perform their functions and such flexible regions appear poorly resolved using traditional reconstruction methods; third, if the dataset lacks particles in a set of orientations then the corresponding views on the reconstructed model will also be poorly resolved in comparison to the well populated orientations.

1.5. Bibliography

Adrian, M. *et al.* (1984) 'Cryo-electron microscopy of viruses', *Nature*. Nature Publishing Group, 308(5954), pp. 32–36. doi: 10.1038/308032a0.

Alberts, B. *et al.* (2017) *Molecular Biology of the Cell*. doi: 10.1201/9781315735368.

Arnold, S. A. *et al.* (2017) 'Blotting-free and lossless cryo-electron microscopy grid preparation from nanoliter-sized protein samples and single-cell extracts', *Journal of Structural Biology*. Academic Press, 197(3), pp. 220–226. doi: 10.1016/J.JSB.2016.11.002.

Baumeister, W., Grimm, R. and Walz, J. (1999) 'Electron tomography of molecules and cells', *Trends in Cell Biology*. Elsevier Current Trends, 9(2), pp. 81–85. doi: 10.1016/S0962-8924(98)01423-8.

Chen, S. *et al.* (2013) 'High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy.', *Ultramicroscopy*. Elsevier, 135, pp. 24–35. doi: 10.1016/j.ultramic.2013.06.004.

Chiu, W. and Glaeser, R. M. (1976) 'Factors affecting high resolution fixed-beam transmission electron microscopy', *Ultramicroscopy*, 2, pp. 207–217. doi: 10.1016/S0304-3991(76)91334-6.

DeRosier, D. J. and Klug, A. (1968) 'Reconstruction of Three Dimensional Structures from Electron Micrographs', *Nature*. Nature Publishing Group, 217(5124), pp. 130–134. doi: 10.1038/217130a0.

Drulyte, I. *et al.* (2018) 'Approaches to altering particle distributions in cryo-electron microscopy sample preparation.', *Acta crystallographica. Section D, Structural biology*. International Union of Crystallography, 74(Pt 6), pp. 560–571. doi: 10.1107/S2059798318006496.

Faruqi, A. . and Henderson, R. (2007) 'Electronic detectors for electron

microscopy', *Current Opinion in Structural Biology*. Elsevier Current Trends, 17(5), pp. 549–555. doi: 10.1016/J.SBI.2007.08.014.

Fitzpatrick, A. W. P. *et al.* (2017) 'Cryo-EM structures of tau filaments from Alzheimer's disease', *Nature 2017 547:7662*. Nature Publishing Group, 547(7662), p. 185. doi: 10.1038/nature23002.

Frank, J. (2006) *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. doi: 10.1093/acprof:oso/9780195182187.001.0001.

Fromm, S. A. and Sachse, C. (2016) 'Cryo-EM Structure Determination Using Segmented Helical Image Reconstruction', *Methods in Enzymology*. Academic Press, 579, pp. 307–328. doi: 10.1016/BS.MIE.2016.05.034.

van Geuns, R.-J. M. *et al.* (1999) 'Basic principles of magnetic resonance imaging', *Progress in Cardiovascular Diseases*. W.B. Saunders, 42(2), pp. 149–156. doi: 10.1016/S0033-0620(99)70014-9.

Glaeser, R. M. (2017) 'Proteins, interfaces, and cryo-EM grids', *Current Opinion in Colloid and Interface Science*, 34, pp. 1–8. doi: 10.1016/j.cocis.2017.12.009.

Goodsell, D. S. and Olson, A. J. (2000) 'Structural Symmetry and Protein Function', *Annual Review of Biophysics and Biomolecular Structure*. Annual Reviews, 29(1), pp. 105–153. doi: 10.1146/annurev.biophys.29.1.105.

Grant, T. and Grigorieff, N. (2015) 'Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6', *eLife*, 4. doi: 10.7554/eLife.06980.

Guerrero-Ferreira, R. *et al.* (2018) 'Cryo-EM structure of alpha-synuclein fibrils', *eLife*, 7. doi: 10.7554/eLife.36402.

Harauz, G. and van Heel, M. (1986) 'Exact filters for general geometry three dimensional reconstruction.', *Optik.*, 73(4), pp. 146–156.

Van Heel, M. (1987) 'Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction', *Ultramicroscopy*. North-Holland, 21(2), pp. 111–123. doi: 10.1016/0304-3991(87)90078-7.

van Heel, M. and Schatz, M. (2005) 'Fourier shell correlation threshold criteria', *Journal of Structural Biology*. Academic Press, 151(3), pp. 250–262. doi: 10.1016/J.JSB.2005.05.009.

Henderson, R. (1995) 'The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules.', *Quarterly reviews of biophysics*, 28(2), pp. 171–93.

Mallamace, F. *et al.* (2016) 'Energy landscape in protein folding and unfolding.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 113(12), pp. 3159–63. doi: 10.1073/pnas.1524864113.

Mastronarde, D. N. (2005) 'Automated electron microscope tomography using robust prediction of specimen movements', *Journal of Structural Biology*. Academic Press, 152(1), pp. 36–51. doi: 10.1016/J.JSB.2005.07.007.

McPherson, A. and Gavira, J. A. (2014) 'Introduction to protein crystallization', *Acta Crystallographica Section F Structural Biology Communications*. International Union of Crystallography, 70(1), pp. 2–20. doi: 10.1107/S2053230X13033141.

Mindell, J. A. and Grigorieff, N. (2003) 'Accurate determination of local defocus and specimen tilt in electron microscopy', *Journal of Structural Biology*. Academic Press, 142(3), pp. 334–347. doi: 10.1016/S1047-8477(03)00069-8.

Murata, K. and Wolf, M. (2018) 'Cryo-electron microscopy for structural analysis of dynamic biological macromolecules', *Biochimica et Biophysica Acta (BBA) - General Subjects*. Elsevier, 1862(2), pp. 324–334. doi: 10.1016/J.BBAGEN.2017.07.020.

Rosenthal, P. B. and Henderson, R. (2003) 'Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy', *Journal of Molecular Biology*. Academic Press, 333(4), pp. 721–745. doi: 10.1016/J.JMB.2003.07.013.

Schmidli, C. *et al.* (2019) 'Microfluidic protein isolation and sample preparation for high-resolution cryo-EM.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 116(30), pp. 15007–15012. doi: 10.1073/pnas.1907214116.

Schur, F. K. M. *et al.* (2016) 'An atomic model of HIV-1 capsid-SP1 reveals structures regulating assembly and maturation', *Science*. American Association for the Advancement of Science, 353(6298), pp. 506–508. doi: 10.1126/SCIENCE.AAF9620.

Sigworth, F. J. *et al.* (2010) 'An Introduction to Maximum-Likelihood Methods in Cryo-EM', *Methods in Enzymology*. Academic Press, 482, pp. 263–294. doi: 10.1016/S0076-6879(10)82011-7.

Taylor, G. L. (2010) 'Introduction to phasing', *Acta Crystallographica Section D Biological Crystallography*. International Union of Crystallography, 66(4), pp. 325–338. doi: 10.1107/S0907444910006694.

Thonghin, N. *et al.* (2018) 'Cryo-electron microscopy of membrane proteins', *Methods*. Academic Press, 147, pp. 176–186. doi: 10.1016/J.YMETH.2018.04.018.

Wagner, T. *et al.* (2019) 'SPHIRE-crYOLO: A fast and accurate fully automated particle picker for cryo-EM', *bioRxiv*. Cold Spring Harbor Laboratory, p. 356584. doi: 10.1101/356584.

Wüthrich, K. (1991) *NMR of proteins and nucleic acids*. Wiley.

Zivanov, J. *et al.* (2018) 'New tools for automated high-resolution cryo-EM structure determination in RELION-3', *eLife*, 7. doi: 10.7554/eLife.42166.

1.6. Structure and aims of the thesis

The work presented on this thesis is a selection of some of the projects I developed in CINA over the past 4 years; a more complete description of the full project list please is depicted in the "Publication list" section.

The determination of the atomic structure of proteins and protein complexes by cryo-EM is highly dependent on the nature and quality of the specimen embedded in vitreous ice. Whereas high resolution density maps could be determined for high molecular weight soluble proteins, for smaller membrane proteins several obstacles limited the overall quality and resolution of the final maps. Atomic modelling in the latter case is often difficult or not feasible but biologically meaningful information can still be retrieved from such maps without discussing side chain positions.

Chapters I and V provide an introduction and conclusion/outlook, respectively, of the discussed projects while chapters II, III and IV are a set of application cases of cryo-EM SPA to protein samples in very different scenarios.

- Chapter I introduces protein sample preparation as well as methods for EM data acquisition and image processing in the SPA context.

- Chapter II focuses on the sample optimization of the bovine Rhodopsin-Gprotein complex and the impact of using a Fab fragment to improve particle alignment during image processing. (Application case of a small non-symmetric membrane protein complex with a poorly resolved flexible region).

- Chapter III describes the work done on the ExbB-ExbD complex where unexpectedly a different complex stoichiometry was revealed by the

obtained electron density map. Despite having a nominal resolution marginally better than 4Å the map provides sufficient and enough evidence of a 5:2 ExbB-ExbD complex. (Application case of ExbB, a small homo pentameric membrane protein with C5 symmetry which becomes an hetero heptamer with C1 symmetry upon ExbD binding).

- Chapter IV presents the determination of the human Thyroglobulin structure, the refinement approach taken to solve flexible peripheral regions present in this protein and subsequent atomic modelling. (Application case of a large soluble C2 symmetric dimer with an extensive disulfide bond network and a pair of flexible regions whose densities were significantly improved using local refinement methods).

- Chapter V concludes the thesis, elaborates on what are the current bottlenecks in structure determination by cryo-EM and provides an outlook.

1.7. Publication list

1. Tsai C-J, Marino J, Adaixo R, Pamula F, et al. (2019) 'Cryo-EM structure of the rhodopsin-Gαi-βγ complex reveals binding of the rhodopsin C-terminal tail to the gβ subunit' *eLife*, 8, doi: 10.7554/eLife.46041
   (Equal first author; Chapter II on this thesis)

2. Marchesi A, Gao X, Adaixo R, Rheinberger J, Stahlberg H, Nimigean C & Simon Scheuring, (2018) 'An iris diaphragm mechanism to gate a cyclic nucleotide-gated ion channel' *Nature Communications*, 9, doi: 10.1038/s41467-018-06414-8
   (Contributing author)

3. Arnold SA, Albiez S, Bieri A, Syntychaki A, Adaixo R, McLeod RA, Goldie KN, Stahlberg H, Braun T, (2017) 'Blotting-free and lossless cryo-electron microscopy grid preparation from nanoliter-sized protein samples and single-cell extracts' Journal of Structural Biology, 197(3) doi: 10.1016/J.JSB.2016.11.002
   (Contributing author)

4. Biou V, Adaixo R, Chami M, Izadi N, Malosse C, Rooke J, Coureux P, Stahlberg H, Delepelaire P, 'Functional and structural characterization of a new class of ExbB proteins with a long N-terminal periplasmic extension'
   (manuscript in preparation; equal first author; Chapter III on this thesis)

5. Adaixo R, Righetto R, Taylor N, Stahlberg H, 'Cryo-EM Structure of the human Thyroglobulin'
   (manuscript in preparation; Chapter IV on this thesis)

6. Righetto R, Adaixo R, Anton L, Schwede T, Maier T & Stahlberg H 'Revisiting the structure of urease by high-resolution cryo-EM' (manuscript in preparation)

7. Schneider JP, Nazarov S, Adaixo R, Liuzzo M, Ringel PD, Stahlberg H, Basler M, (2019) 'Diverse roles of TssA-like proteins in the assembly of bacterial type VI secretion systems' The EMBO Journal doi: 10.15252/embj.2018100825
(Contributing author)

Chapter II

2. Cryo-EM structure of the rhodopsin-Gαi-βγ complex

This chapter focuses on the prototypical membrane protein rhodopsin in complex with Gαi-β-γ subunits and an engineered Fab fragment. The determined 4.38 Å resolution map revealed a complex with all subunits present (some only partially). The position and contacts of rhodopsin's C terminal tail, an important regulatory element in rhodopsin signaling cascade, could be observed in this complex.

Contribution: cryo-EM dataset acquisition and processing; contributed to atomic modelling of the complex; contributed to manuscript writing, editing and reviewing.

The following section has been published in:

eLife

Volume 8, June 2019, Pages 1-19

DOI: doi.org/10.7554/eLife.46041

Cryo-EM structure of the rhodopsin-Gαi-βγ complex reveals binding of the rhodopsin C-terminal tail to the gβ subunit

Ching-Ju Tsai[1], Jacopo Marino[1], Ricardo Adaixo[2], Filip Pamula[1], Jonas Muehle[1], Shoji Maeda[1], Tilman Flock[1,3], Nicholas MI Taylor[2], Inayatulla Mohammed[2], Hugues Matile[4], Roger JP Dawson[4], Xavier Deupi[1,5], Henning Stahlberg[2], Gebhard Schertler[1,3]

1 Division of Biology and Chemistry / Laboratory of Biomolecular Research, Paul Scherrer Institute, Villigen, Switzerland;

2 Center for Cellular Imaging and NanAnalytics (C-CINA), Biozentrum, University of Basel, Basel, Switzerland;

3 Department of Biology, ETH Zürich, Zürich, Switzerland;

4 Pharma Research and Early Development, Therapeutic modalities, Roche Innovation Center Basel, Hoffmann-La Roche Ltd, Basel, Switzerland;

5 Condensed Matter Theory Group, Paul Scherrer Institute, Villigen, Switzerland

Abstract

One of the largest membrane protein families in eukaryotes are G protein-coupled receptors (GPCRs). GPCRs modulate cell physiology by activating diverse intracellular transducers, prominently heterotrimeric G proteins. The recent surge in structural data has expanded our understanding of GPCR-mediated signal transduction. However, many aspects, including the existence of transient interactions, remain elusive. We present the cryo-EM structure of the light-sensitive GPCR rhodopsin in complex with heterotrimeric Gi. Our density map reveals the receptor C-terminal tail bound to the G β subunit of the G protein, providing a structural foundation for the role of the C-terminal tail in GPCR signaling, and of G β as scaffold for recruiting G α subunits and G protein-receptor kinases. By comparing available complexes, we found a small set of common anchoring points that are G protein-subtype specific. Taken together, our structure and analysis provide new structural basis for the molecular events of the GPCR signaling pathway.

2.1. Introduction

G protein-coupled receptors (GPCRs) are the most diverse class of integral membrane proteins with almost 800 members in humans. GPCRs are activated by a great diversity of extracellular stimuli including photons, neurotransmitters, ions, proteins, and hormones (Glukhova et al., 2018). Upon activation, GPCRs couple to intracellular transducers, including four subtypes of G proteins (G α s, G α i/o, G α q/11, G α 12/13) (Milligan and Kostenis, 2006), seven subtypes of GPCR kinases (GRKs) (Gurevich et al., 2012), and four subtypes of arrestins (Smith and Rajagopal, 2016) (Figure 2.1A), among many other partners (Magalhaes et al., 2012). While most GPCRs are promiscuous and can couple to more than one G protein subtype (Flock et al., 2017), the molecular determinants of G protein recognition are not yet fully understood. Understanding the molecular basis for G protein coupling and selectivity could lead to the design of

drugs that promote specific signaling pathways and avoid unwanted side effects (Hauser et al., 2017).

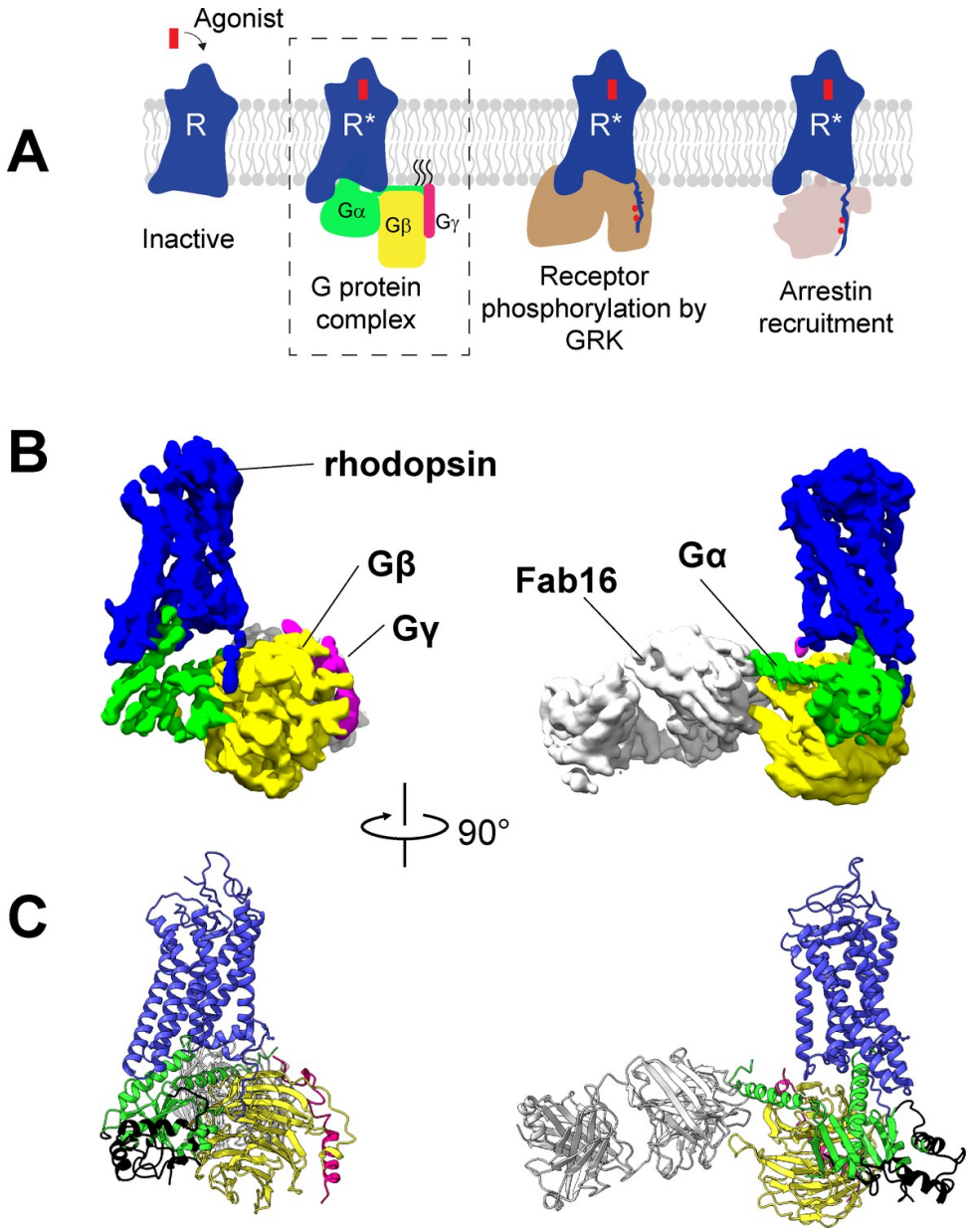

Figure 2.1: Cryo-EM structure of the rhodopsin-Gi-Fab16 complex. (A) GPCR signaling complexes. (B) EM density map of the complex (rhodopsin – blue, G α i - green, G β – yellow, G γ – magenta, Fab16 – white). (C) Atomic model of the complex (same color code as B). The region

of the Ras domain of G α with no corresponding density in the EM map is depicted in back.

The recent surge in the number of structures of GPCR-G protein complexes has greatly expanded our understanding of G protein recognition and GPCR-mediated signal transduction. Out of the 13 structures of GPCR-G protein complexes available, six contain a Gi/o subtype: μ-opioid receptor bound to Gi (Koehl et al., 2018), adenosine A_1 receptor bound to Gi (Draper-Joyce et al., 2018), cannabinoid receptor one bound to Gi (Krishna Kumar et al., 2019), human rhodopsin bound to Gi (Kang et al., 2018), $5HT_{1B}$ receptor bound to Go (García-Nafría et al., 2018b), and bovine rhodopsin bound to an engineered Go (mini-Go) (Tsai et al., 2018). However, the preparation of GPCR-G protein complexes for structural biology still remains challenging (Munk et al., 2019). Nanobodies (Rasmussen et al., 2011), Fab fragments (Kang et al., 2018; Koehl et al., 2018), and mini-G proteins (García-Nafría et al., 2018b; Tsai et al., 2018) have been very important tools to overcome the inherent instability and flexibility of these complexes and obtain near atomic-resolution structures. Importantly, these structures represent snapshots of a particular state of the complex in the signaling cascade, and therefore additional structural data are required to improve our understanding of this process (Capper and Wacker, 2018).

The photoreceptor rhodopsin is one of the best-characterized model systems for studying GPCRs, providing invaluable information on how receptor activation is translated into G protein and arrestin binding and activation (Hofmann et al., 2009; Hofmann and Palczewski, 2015; Scheerer and Sommer, 2017). Rhodopsin has been shown to interact with the G $\beta\ \gamma$ subunit of the G protein heterotrimer to assist binding and activation of the G α subunit (Herrmann et al., 2004; Herrmann et al., 2006). After dissociation of the G $\alpha\ \beta\ \gamma$ heterotrimer, the G β subunit recruits GRKs to the membrane, resulting in the phosphorylation

of the receptor C-terminal tail (C-tail) (Claing et al., 2002; Pao and Benovic, 2002; Pitcher et al., 1998) and binding of arrestin (Goodman et al., 1996). Despite this biochemical evidence, a direct interaction between rhodopsin and G $\beta\ \gamma$ could not be observed in the existing complex (Kang et al., 2018).

Here, we present the cryo-EM structure of bovine rhodopsin in complex with a heterotrimeric Gi. Overall, our structure agrees well with current published structures (Kang et al., 2018; Tsai et al., 2018). Remarkably, the EM density map provides structural evidence for the interaction between the C-tail of the receptor and the G β subunit. The density map also shows that intracellular loops (ICL) 2 and 3 of rhodopsin are at contact distance to G α. This prompted us to perform a comparison of all available structures of GPCR-G protein complexes to generate a comprehensive contact map of this region. We then extended this analysis to the binding interface formed by the C-terminal helix α 5 of G α and found that only a few G protein subtype-specific residues consistently bind to the receptors. These contacts are ubiquitous anchoring points that may be also involved in the selective engagement and activation of G proteins.

2.2. Results

To obtain a rhodopsin-Gi complex suitable for structural studies, we expressed in HEK cells the constitutively active mutant of bovine rhodopsin N2C/M257Y/D282C (Deupi et al., 2012) which binds to the Gi protein heterotrimer (Maeda et al., 2014). The construct of the human G α i1 subunit was expressed in *E. coli* (Sun et al., 2015), while the G β 1 γ 1 subunit was isolated from bovine retinae and thus contains native post-translational modification important for transducin function (Matsuda and Fukada, 2000). We reconstituted the rhodopsin-G α i1 β 1 γ 1 complex in the detergent lauryl maltose neopentyl glycol (LMNG) (Maeda et al., 2014) in presence of Fab16 (Maeda et al., 2018) (Figure 2.1—figure supplement 1), as cryo-EM screening revealed that

the complex without Fab could not be refined to high resolution (Figure 2.1—figure supplements 2 and 3). During image processing (Figure 2.1—figure supplements 3 and 4), 3D classification revealed that the density corresponding to Gα is heterogeneous (Figure 2.1—figure supplement 5), particularly at flexible regions such as the α-helical (AH) domain. The AH domain was then excluded by using a soft mask during refinement, resulting in a map with a nominal global resolution of 4.38 Å (Figure 2.1—figure supplement 3D and E). The EM map was used to build a model of the complex (Figure 2.1, Figure 2.1—figure supplement 6).

2.2.1. Architecture of the rhodopsin-G protein complex

The structure of rhodopsin-Gi-Fab16 reveals the features observed in previously reported GPCR-G protein complexes (Figure 2.1B and C; Figure 2.1—figure supplement 7A). In particular, our cryo-EM structure is in excellent agreement with the crystal structure of the same rhodopsin mutant bound to a mini-Go protein (Tsai et al., 2018), with a nearly identical orientation of the C-terminal α5 helix (Figure 2.1—figure supplement 7B), which contacts transmembrane helices (TM) 2, 3, 5–7 and the TM7/helix 8 (H8) turn of the receptor (Figure 2.1—figure supplement 6B).

2.2.2. Interaction between the C-terminal tail of rhodopsin and Gβ

The EM map reveals a density on the Gβ subunit as continuation of H8 of the receptor (Figure 2.2A), which corresponds to the C-tail of rhodopsin. This feature has only been observed in the recent structure of the M1 muscarinic acetylcholine receptor in complex with G11 (Maeda et al., 2019). We modeled half of the C-tail of rhodopsin (12 out of 25 residues; 324–335) into this density as a continuous stretched peptide with residues G324, P327 and G329 serving as flexible hinges (Figure 2.2B). This allows us to compare the structure of this region in three conformational states of the receptor: inactive (Okada et al., 2004), G protein-bound (this structure), and arrestin-bound (Zhou et al., 2017) (Figure 2.2B).

In the inactive state, the C-tail folds around the cytoplasmic side of rhodopsin, although this is likely due to crystal packing as this region is intrinsically disordered in the absence of a binding partner (Jaakola et al., 2005; Venkatakrishnan et al., 2014).

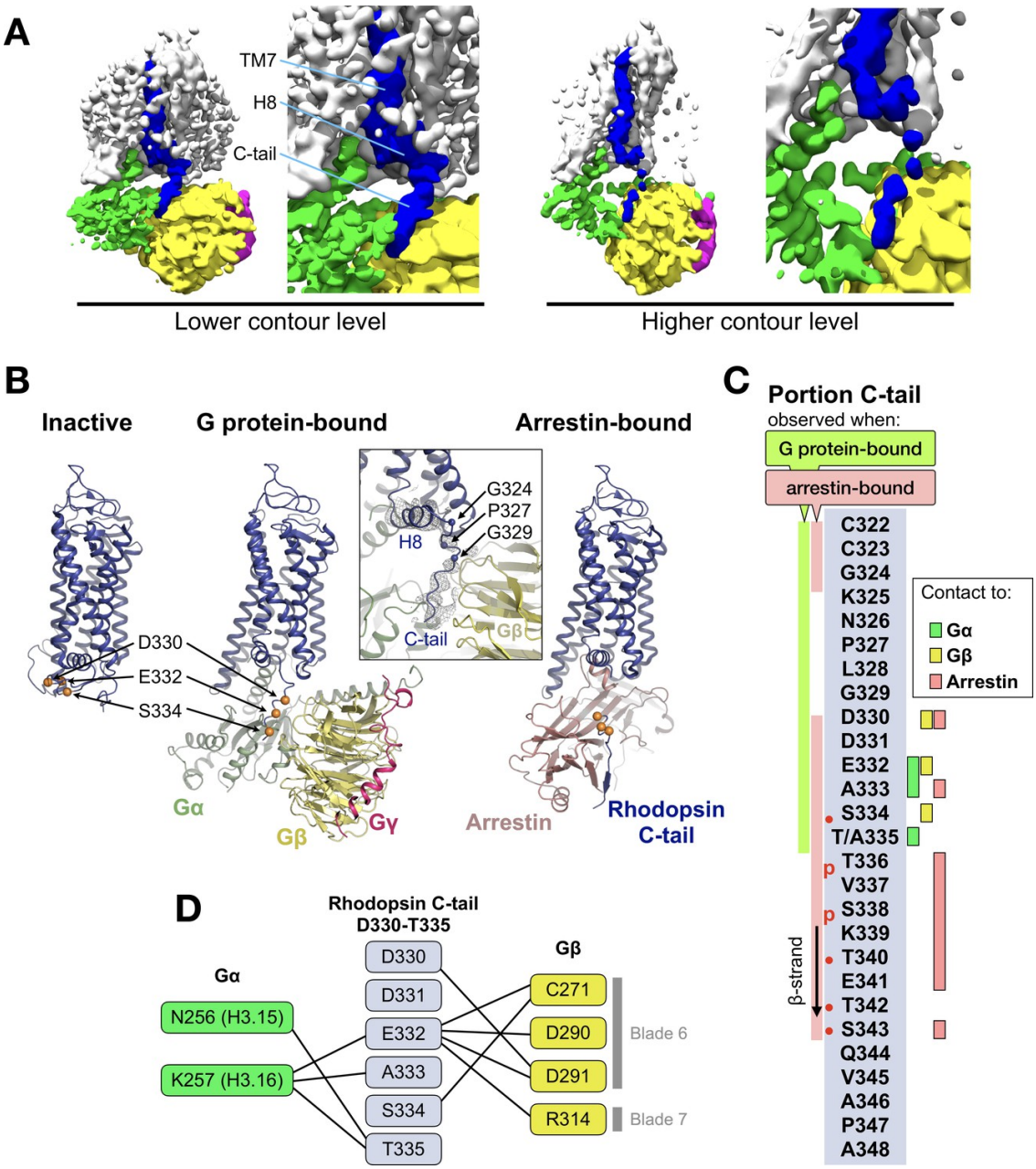

Figure 2.2: The C-terminal tail of rhodopsin. (A) The EM map is contoured at two different levels to show the continuity of the density. The weakening at the end of H8 may arise from impaired interactions of the receptor with the detergent micelle (Glukhova et al., 2018). TM7, H8 and the C-tail of the receptor are colored in blue, G α in green, G β in yellow, and G γ in magenta. (B) Conformational change of the C-tail between three different conformational states of rhodopsin: Inactive state (left, PDB id: 1U19), G protein-bound (center, this work), and arrestin-bound (right, PDB id: 5W0P, chain A). The C α atoms of residues Asp330, Glu332 and Ser334 are shown as orange spheres to help tracking the structural changes in the C-tail. All structures are aligned to rhodopsin. (C) Schematic representation of the rhodopsin C-tail from Cys322 to Ala348. On the left, colored bars indicate the portion of the C-tail visible in this structure (green), and in the arrestin-bound structure (salmon) (PDB id: 5W0P). On the right, the residue-residue contacts between rhodopsin C-tail and G α (marked in green), G β (yellow), and arrestin (salmon) within 4 Å distance are indicated. Thr336 and Ser338 are phosphorylated in the arrestin-bound structure. The predicted phosphorylation sites are marked with red dots. (D) Model of residue-residue interaction between the rhodopsin C-tail and the G protein subunits. $Asn^{H3.15}$ and $Lys^{H3.16}$ of the G α i subunit forms the contact to the C-tail of rhodopsin near Glu332, Ala333 and Thr 335. In this model, the surface region of blades 6 and 7 of G β contact the C-tail via hydrophilic residues Cys271, Asp290, Asp 291, and Arg314.

In our G protein complex, the C-tail stretches over a polar surface on the cleft between G α and G β, interacting with both subunits (Figure 2.2C and D). In G β, the interacting residues are C271, D290, and D291 in blade six and R314

in blade 7 of the β-propeller (Figure 2.2D), which impart a local negative electrostatic potential (Figure 2.2—figure supplement 1). These residues are conserved in Gβ1–4 (Figure 2.2—figure supplement 2B) but not in Gβ5, the least similar to the other Gβ subtypes (Dupré et al., 2009). In Gα, the interacting residues with the C-tail are N256H3.15 and K257H3.16 in helix 3 of the Ras domain (Figure 2.2D), which confer instead a local positive electrostatic potential (Figure 2.2—figure supplement 1). These residues are quite conserved across G protein subtypes except Gq (Figure 2.2—figure supplement 2A). Interestingly, these regions in Gα and Gβ that contact the receptor C-tail are also involved in recognition of GRK2 (Tesmer et al., 2005).

In the rhodopsin-arrestin complex part of the proximal segment of the C-terminus (residues 325–329) is not resolved, but the distal part could be modeled up to S343, eight residues longer than in our G protein complex and including two phosphorylated sites. In the presence of arrestin, the C-tail stretches further, with residues K339-T342 forming a short β-strand antiparallel to the N-terminal β-strand of arrestin (Figure 2.2B).

Thus, it appears that distinct portions of the C-tail are responsible for contacting different intracellular partners. The central part of the C-tail (residues 330–335) can bind to Gα and Gβ, while the distal half of the C-tail (residues 336–343), which contains five (out of six) phosphorylation sites (Azevedo et al., 2015), binds to arrestin (Figure 2.2C).

2.2.3. Comparison of the GPCR-G protein binding interface

As in the other existing GPCR-G protein complex structures, the C-terminal α5 helix of Gα forms the major contact interface to rhodopsin. The α5 helix consists of the last 26 amino acids of Gα (H5.01–26), in which the last five residues fold into a hook-like structure (Tsai et al., 2018). The majority of the

contacts formed by the α5 helix to GPCRs concentrate in the region from H5.11 to H5.26 (Glukhova et al., 2018) (Figure 2.3—figure supplement 1).

We aligned the structures of available complexes using the Cα atoms of H5.11–26. This alignment reduces apparent differences in the binding positions and provides the 'viewpoint of the G protein' (Figure 2.3A). We compiled an exhaustive list of the residue-residue contacts between the receptor and the α5 helix in all the available GPCR/G protein complexes, and observed that the main contacts to the α5 helix are formed by TM5 and TM6, followed by TM3 and TM7/8 (Figure 2.3B, Figure 2.3—figure supplement 2).

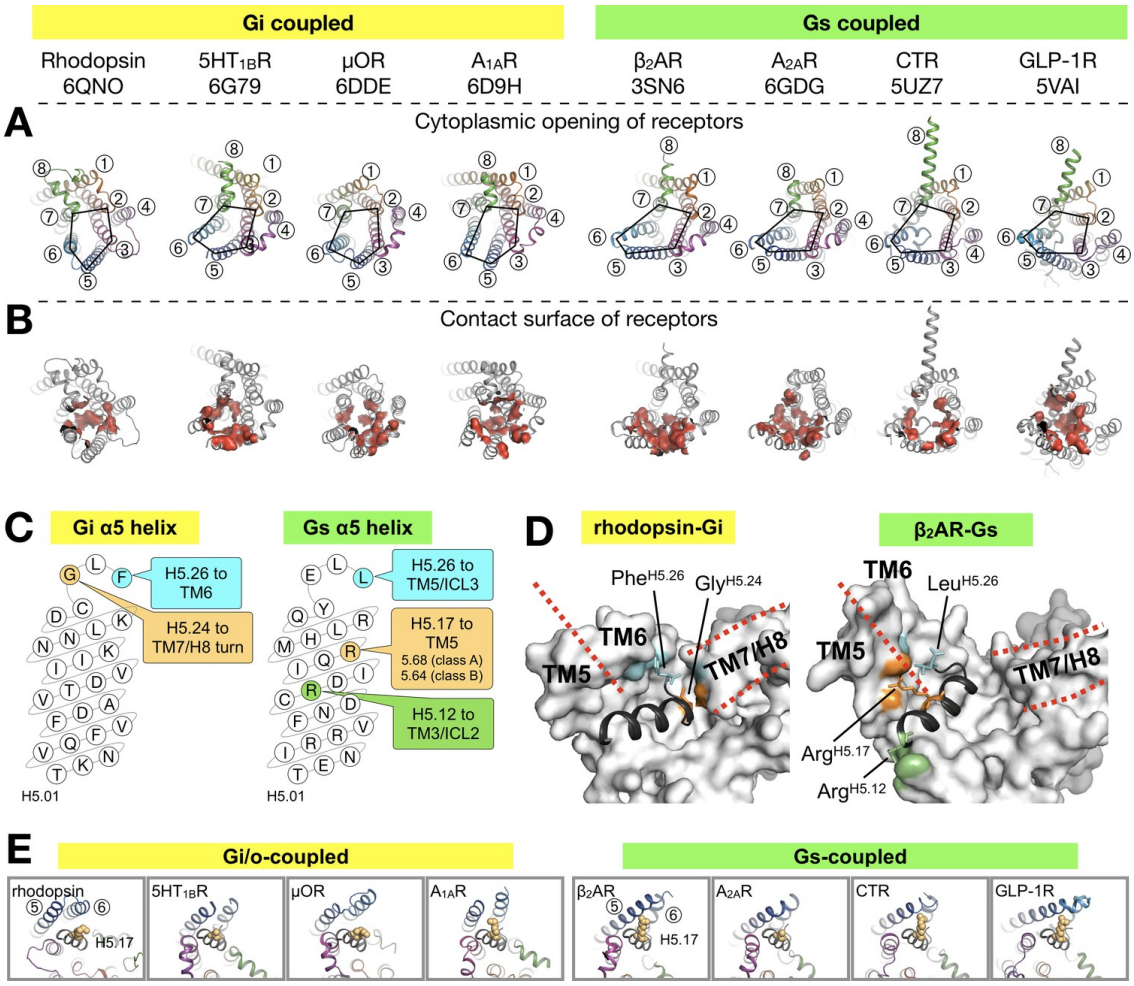

Figure 2.3: Binding of the Gα α5 helix in GPCR-G protein complexes. (A) Overview of a subset of GPCR-G protein complexes used for this analysis. For a complete list of the complexes used, see Figure 2.3—figure supplement 2. The complexes are shown from the cytoplasmic side, and the black pentagons connecting TM2-3-5-6-7 delineate the G protein-binding interface. Receptors are represented in ribbons with their TM helices numbered. (B) The red surfaces depict the contact area (within a distance of 4 Å) between the receptor and the α5 helix of Gα (C) Schematic representation of the α5 helix in Gi- and Gs-subtypes. The residues highlighted, and their respective binding site to the receptor, are conserved among all Gi, and Gs, complexes analyzed. All contacts retrieved from this analysis are reported in Figure 2.3—figure supplement 2. (D) Position of Gi- and Gs-specific contacts shown in the rhodopsin-Gi (this work, PDB id: 6QNO) and the β_2AR-Gs (PDB id: 3SN6) complexes. The cytoplasmic side of the receptors is depicted as gray surface. Red lines mark the border between TM5 and TM6 and delineate H8. (E) Cytoplasmic view of GPCR-G protein complexes showing the interaction between H5.17 (orange spheres) and TM5 and TM6. From left to right, the structures (PDB ids) correspond to: this work, 6G79, 6DDE, 6D9H, 3SN6, 6GDG, 5UZ7, and 5VAI. Throughout the analysis, contacts are defined as atomic distances smaller than 4 Å.

Gi/o-bound complexes have two conserved contacts: one between $Gly^{H5.24}$ and the TM7/H8 turn, and one between $Tyr/Phe^{H5.26}$ and TM6. In Gs-bound complexes, three distinct interactions are found: between $Arg^{H5.12}$ and TM3/ICL2, and between $Leu^{H5.26}$ and $Arg^{H5.17}$ and TM5 (Figure 2.3C,D). In Gi/o complexes, the equivalent $Lys^{H5.17}$ lies instead roughly parallel to TM6. Interestingly, $Arg^{H5.17}$ appears to consistently interact with residue 5.68 in class

A GPCRs, and with Lys$^{5.64}$ in class B GPCRs (Figure 3E, Figure 2.3—figure supplement 2). We suggest that a tight interaction between Arg$^{H5.17}$ and TM5 might be one of the main determinants for the Gs-specific relocation of TM6.

Besides the canonical interaction with the $\alpha 5$ helix, our complex shows that ICL2 and ICL3 of the receptor are at contact distance to Gi (Figure 2.3—figure supplement 3A). In all analyzed structures, we found that ICL2 lies near $\alpha N/\beta 1$ and $\beta 2/\beta 3$ of $G\alpha$, while ICL3 is close to $\alpha 4/\beta 6$ (Figure 2.3—figure supplements 3 and 4). Interestingly, ICL2 folds into an α-helical structure in all the class A receptors except rhodopsin (Figure 2.3—figure supplement 5).

ICL2 does not contribute to binding $G\alpha$ in the structures of $5HT_{1B}$-mini-Go and $A_{1A}R$-Gi. Nevertheless, the contact between $G\alpha$ and ICL2 seems to discriminate between class A GPCRs –which interact via the $\alpha N/\beta 1$ loop– and class B GPCRs -which instead the region around $\beta 1$ and $\beta 2/\beta 3$ (Figure 2.3—figure supplement 4).

ICL3, one of the most diverse regions in GPCRs, is often not completely resolved in the available structures (Figure 2.3—figure supplement 3D), either because it is too flexible or because it has been engineered to facilitate structural determination (Munk et al., 2019). Gi/o-coupled receptors use residues on TM5/ICL3/TM6 to contact the $\alpha 4/\beta 6$ region, while Gs-coupled receptors mainly use TM5/ICL3 (Figure 2.3—figure supplement 4). This difference is due to the larger displacement of TM6 in Gs complexes. In ICL3, residue Tyr$^{G.S6.02}$, highly conserved in all G protein subtypes, engages the receptor in about half of all the complexes. This residue has been shown to be crucial for the stabilization of the rhodopsin-Gi complex (Sun et al., 2015).

Thus, our analysis suggests that ICL2 and ICL3 contribute to the binding interface between receptor and G protein, a feature that may be required to further stabilize the complex during nucleotide exchange.

2.3. Discussion

In this work, we present the cryo-EM structure of the signaling complex between bovine rhodopsin and a Gi protein heterotrimer, stabilized using an antibody Fab fragment (Fab16) (Maeda et al., 2018). Overall, this structure agrees very well with existing complexes. In particular, we found that the binding mode of the G protein Ras domain –including the key C-terminal $\alpha 5$ helix– is virtually identical to our previously reported X-ray structure of the same receptor bound to a mini-Go protein (Tsai et al., 2018) (Figure 2.1—figure supplement 7B).

However, our EM map shows a density on the G β subunit that extends from H8 of the receptor, constituting the proximal segment of the C-terminus (residues G324 to T335) (Figure 2.2). One explanation of why the C-tail is observed in our density map and not in other structures may rely on the nature of the components used to reconstitute the reported GPCR complexes (Figure 2.1—figure supplement 8). Among those, a meaningful comparison may be done with the μ-opioid receptor complex (Koehl et al., 2018), which is bound to a shorter version of our antibody Fab16, but in which the C-terminus of the receptor was partially truncated. In the human rhodopsin complex (Kang et al., 2018), which contains the full length C-tail, a recombinant G $\beta\ \gamma$ was used. Thus, our bovine rhodopsin complex contains the unique combination of a full-length C-tail, G $\beta\ \gamma$ purified from bovine rod outer segments, and Fab16 that may have contributed to trap the transient interaction of the intrinsically disordered C-tail to G β. Remarkably, the recently published structure of the M1 muscarinic acetylcholine receptor in complex with G11 and scFv16 (Maeda et al., 2019)

reveals part of the receptor C-tail bound to the G protein in essentially the same conformation as in our structure (Figure 2.2—figure supplement 4).

The $G\beta\gamma$ subunit engages with a wide range of effectors (Dupré et al., 2009; Khan et al., 2013), including direct interactions with GPCRs. For instance, these can associate prior to trafficking to the plasma membrane (Dupré et al., 2006) where they can remain pre-coupled (Galés et al., 2005). Also, according to a sequential fit model (Herrmann et al., 2004; Herrmann et al., 2006), activated rhodopsin binds to $G\beta\gamma$ assisting to position the Ras domain of $G\alpha$ into proximity of its binding site to the receptor. This would facilitate folding of the C-terminal $\alpha 5$ helix of this subunit leading to nucleotide exchange (Dror et al., 2015; Flock et al., 2015; Kapoor et al., 2009). After dissociation of the G protein heterotrimer, the $G\beta$ subunit recruits GRKs to the membrane, resulting in the phosphorylation of the receptor C-tail (Claing et al., 2002; Pao and Benovic, 2002; Pitcher et al., 1998) and binding of arrestin (Goodman et al., 1996). Interestingly, HEK cells lacking functional $G\alpha$ subunits (but retaining the native $G\beta\gamma$) are still able to recruit arrestin, although they fail to activate ERK and whole-cell responses (Grundmann et al., 2018).

Our findings provide a structural explanation for these roles of $G\beta\gamma$ in GPCR-mediated signaling. We suggest that the observed interaction between the central part of the receptor C-terminus and $G\beta\gamma$ is involved in localizing the G protein heterotrimer to the active receptor first. After dissociation of the $G\alpha$ subunit, a transient $G\beta\gamma$/receptor complex could provide the adequate molecular context to allow receptor phosphorylation by bringing the kinase close to the receptor C-tail (Ribas et al., 2007).

In other class A GPCRs, interactions with $G\beta\gamma$ have been also located in ICL3 (in M2 and M3) muscarinic receptors, involved in the phosphorylation of this

loop (Wu et al., 1998) and potentially in ICL1 (in A_2AR and β_2AR) (García-Nafría et al., 2018a). In our complex, K6612.48 and K6712.49 in ICL1 are indeed close to D312 in $G\beta$ (7 Å between $C\alpha$ atoms); however, we do not observe defined density for the side chains in this region (Figure 2.1—figure supplement 6A). Interestingly, the complexes of the class B GPCRs calcitonin and GLP-1 feature an extended and tilted H8 resulting in extensive contacts with $G\beta\gamma$ that have not been observed in Class A receptors (García-Nafría et al., 2018a). H8 has also been shown to bind $G\beta\gamma$ in the class B PTH1 receptor (Mahon et al., 2006). Thus, it is likely that class A and class B GPCRs use distinct mechanisms to engage $G\beta\gamma$.

The receptor-$G\beta\gamma$ interaction observed in our structure is compatible with proposed models of receptor-GRK recognition (Komolov et al., 2017; Sarnago et al., 2003). Thus, we suggest that the $G\beta\gamma$ subunit can 'tag' activated GPCRs and provide an interface for bringing different effectors ($G\alpha$ subunits first, and GRKs later) near the cytoplasmic domains of GPCRs.

The availability of several GPCR-G protein complexes has greatly advanced our understanding on how receptors activate the $G\alpha$ subunit (Dror et al., 2015; Flock et al., 2015). The binding interface of the receptor is partially formed by ICL2 and ICL3 (Chung et al., 2011; Glukhova et al., 2018; Sun et al., 2015). Accordingly, our EM map shows that these domains are in close proximity to G (Figure 2.3—figure supplement 3). In particular, we found that ICL2 is at contact distance to the αN helix, the $\alpha N/\beta 1$ and the $\beta 2/\beta 3$ $G\alpha$ in most structures (Figure 2.3—figure supplements 3–5). While ICL2 contributes to the binding interface, there are no apparent conserved contacts among complexes (Figure 2.3—figure supplement 4). However, at the interface between receptor and the $\alpha 5$ helix of the G protein, in all Gi complexes $Phe^{H5.26}$ and $Gly^{H5.24}$ contact TM6 and TM7/H8 of the receptor, respectively. Instead, in all Gs complexes we find

that Leu$^{H5.26}$, Arg$^{H5.17}$ and Arg$^{H5.12}$ contact TM5/ICL3, TM5, and TM3/ICL2, respectively. In a recent analysis (Glukhova et al., 2018), it was proposed that residues 5.68 (class A GPCRs) and 5.64 (class B GPCRs) are particularly important for Gs protein binding. We observe that residue 5.68 interacts with H5.16 in both Gi- and Gs-bound complexes, but only with Arg$^{H5.17}$ in all Gs complexes (Figure 2.3—figure supplement 2). While our analysis is limited by the number of available structures, we suggest that the conserved contacts identified between $\alpha 5$ and the receptors are structurally important anchoring points, but we cannot exclude that they might be relevant at the level of recognition between G protein subtypes.

In order to visualize snapshots of the signaling pathway that have not been observed yet, more work on the structural characterization of complexes is still needed, ideally using the same receptor and different transducers. Such complexes are of pivotal importance to decipher the structural basis of the GPCR-mediated signaling cascade.

2.4. Materials and Methods

2.4.1. Protein expression and purification

The N2C/M257Y/D282C mutant of bovine rhodopsin was expressed as described (Deupi et al., 2012) in human embryonic kidney (HEK) 293 GnTI-deficient cells (gift from Philip Reeves), which retain all post-translational modifications of the native protein including palmitoylation of Cys residues at positions 322 and 323 (Standfuss et al., 2011). Cell lines were tested and confirmed to be mycoplasma-free. The human $G\alpha i$ subunit ($G\alpha i1$) with an N-terminal TEV protease-cleavable deca-histidine tag was expressed in the *E. coli* BL21 (DE3) strain (Sigma: CMC0014) and purified as described (Sun et al., 2015). The transducin heterotrimer was isolated from the rod outer segment of bovine retina (W L Lawson Company) and $G\beta 1\gamma 1$ was separated from $G\alpha t$ with Blue Sepharose 6 Fast Flow (GE Healthcare) as described (Maeda et al., 2014). The $G\alpha i1\beta 1\gamma$

1 heterotrimer (Gi) was prepared by mixing equimolar amounts of G α i1 with or without 10xHis-tag and G β 1 γ 1 and incubated at 4°C for 1 hr shortly before use for rhodopsin-Gi complex formation on the 1D4 immunoaffinity column.

2.4.2. Fab16 production

The monoclonal mouse antibody IgG16 was generated as described (Koehl et al., 2018). Large scale production of IgG16 was performed using adherent hybridoma cell culture (Hoffmann-La Roche, collaboration) grown in DMEM medium supplemented with 10% ultra-low IgG fetal bovine serum (FBS) (Gibco, #16250078) and 25 U/ml of mouse interleukin-6 (Invitrogen, #PMC0064) at 37°C and 5% CO_2. Antibody expression was increased by stepwise dilution of FBS concentration down to 2%. After incubation for 10–14 days,~500 ml cell suspension containing the secreted IgG was clarified by centrifugation and subsequent filtration through a 0.45 µm HAWP membrane (Merck Millipore). The filtrate was mixed with an equal volume of binding buffer (20 mM Na_2HPO_4/NaH_2PO_4, pH 7.0) and loaded to a 1 ml HiTrap Protein G Sepharose FF column (GE Healthcare). The column was washed with binding buffer until the UV_{280} absorbance dropped to a stable baseline, and IgG was eluted with 0.1 M glycine-HCl (pH 2.7). Fractions were immediately neutralized with 1 M Tris-HCl (pH 9.6). Fractions containing IgG16 were combined and dialyzed against 20 mM Na_2HPO_4/NaH_2PO_4 (pH 7.0), 1.5 mM NaN_3 using a slide-A-lyzer dialysis cassette (12–14 kDa MWCO, Thermo Scientific) at 4°C for 15 hr. The dialysate was collected and mixed with the immobilized papain resin (0.05 ml resin for 1 mg IgG) (Thermo Scientific, #20341). Papain was activated by the addition of L-cysteine and EDTA to a final concentration of 20 mM and 10 mM respectively. IgG was digested overnight at 37°C with gentle mixing. Afterwards the immobilized papain resin was removed and the digested fraction was mixed with Protein A Sepharose (0.2 ml resin for 1 mg digested IgG, GE Healthcare, #17078001) for 1 hr at RT. Resins were washed with two column volumes (CV) of wash buffer 10 mM Tris (pH 7.5), 2.5 M NaCl. The flow-through and washing

fractions containing Fab16 were collected and dialyzed against PBS supplemented with 1.5 mM NaN_3 using a slide-A-lyzer dialysis cassette (12–14 kDa MWCO) at 4°C. The dialysate was collected and concentrated with a VivaSpin 20 concentrator (10 kDa MWCO, Sartorius) to approximately 1.1 mg/ml. Glycerol was supplemented to the concentrated Fab16 at a final concentration of 10%, and protein was flash frozen in liquid nitrogen and stored at −20°C until use for formation of the rhodopsin-Gi-Fab16 complex.

2.4.3. Fab16 crystallization and structure determination

For its crystallization, Fab16 was further purified by SEC on a Superdex 200 Increase 10/300 GL column (GE Healthcare) equilibrated in buffer 10 mM Tris (pH 7.5), 50 mM NaCl. Fractions containing pure Fab16 were collected and concentrated to approximately 14 mg/ml using a VivaSpin four concentrator (10 kDa MWCO, Sartorius). Fab16 was crystallized by vapor diffusion at 4°C by dispensing 200 nl of protein and 200 nl of crystallization buffer containing 1.5 M malic acid (pH 7.5), 7% (v/v) LDAO in an MRC 2-well crystallization plate (Swissci) using a mosquito crystallization robot. Crystals appeared after one day and grew to full size within 4 days. Crystals were soaked in the reservoir solution supplemented with 20% (v/v) PEG 400 as a cryo-protectant and flash frozen in liquid nitrogen. The X-ray data was collected at the PXI beam line at the Swiss Light Source (SLS). Bragg peaks were integrated using XDS for individual datasets. XSCALE was used to scale and combine six datasets, and the pooled reflection list was further analyzed using the STARANISO server (Global Phasing Ltd.). The STARANISO server first analyzed the anisotropy for each dataset, giving a resolution of 1.90 Å in overall, 1.90 Å in the *0.92 a* - 0.38 c** direction, 2.25 Å in the *k* direction, and 2.13 Å in the *−0.14 a* + 0.99 c** direction under the criterion of CC½=0.3., following by scaling and merging of the reflections. The light chain from PDB id: 1MJU and the heavy chain from PDB id 2AAB without the CDR regions were used for molecular replacement using the program *Phaser-MR* in the Phenix suite. The model was built

automatically using the Phenix *AutoBuild*, and the coordinates were manually adjusted using the visualization program Coot. The structure was refined using the *phenix.refine* to 1.90 Å (Supplementary file 1). The structure factor and the coordinates are deposited to the Protein Data Bank under the accession code 6QNK.

2.4.4. Purification of the rhodopsin-Gi-Fab16 and the rhodopsin-Gi complexes

Buffers for every purification step were cooled to 4°C before use, and the steps after adding retinal were performed under dim red light before light activation of rhodopsin. The stabilized, constitutively active rhodopsin mutant N2C/M257Y/D282C was expressed in HEK293 GnTI$^-$ deficient cells as described (Deupi et al., 2012). The cells were collected from the cell culture by centrifugation and homogenized in PBS buffer with cOmplete EDTA-free protease inhibitor (Roche). The cells were solubilized by supplementing dodecyl maltoside (DDM) (Anatrace, Sol-grade) to final concentrations of cell at 20% (w/v) and of DDM at 1.25% (w/v). After gentle stirring for 1–2 hr, the solubilized fraction was collected after centrifugation at 200,000x g for 1 hr. Solubilized rhodopsin apoprotein was captured in batch using the immunoaffinity Sepharose beads (GE Healthcare, #1043001) coupled to 1D4 antibody for more than 4 hr at the ratio of 5 g cells per ml resin. 1D4 resins were collected and washed with 10 CV of PBS, 0.04% DDM. Afterwards, resins were resuspended in 2 CV of PBS, 0.04% DDM and 75 µM of 9-cis (Sigma) or 11-cis retinal (Roche, or from NIH) for at least 6 hr in the dark. Later resins were washed with 30 CV of buffer A containing 20 mM HEPES pH 7.5, 150 mM NaCl, 1 mM $MgCl_2$, 0.02% (w/v) lauryl-maltose neopentyl glycol (LMNG) to wash off unbound retinal. Washed resins were resuspended in 1 CV of buffer A and mixed with 10xHis-tagged Gi heterotrimer at the ratio of 2 mg Gi per ml resin with the presence of 25 mU/ml apyrase (New England Biolabs). The resuspended resin/Gi mixture was subjected to irradiation for 10–15 min with light that had been filtered through a 495 nm

long-pass filter to induce activation of rhodopsin and G protein binding, followed by a 30 min incubation in the dark to allow full hydrolysis of nucleotide by apyrase. Resins were washed with 8 CV of buffer A to remove unbounded Gi heterotrimer. rhodopsin-Gi complex was eluted three times in batch with 1.5 CV of buffer containing 20 mM HEPES (pH 7.5), 150 mM NaCl, 1 mM $MgCl_2$, 0.02% LMNG, 80 µM 1D4 peptide (TETSQVAPA) for at least 2 hr incubation each time. Elution fractions were combined and incubated for at least 2 hr with Ni-NTA resins (0.5–2 ml), washed with 6 CV of 20 mM HEPES (pH 7.5), 150 mM NaCl, 50 mM imidazole, 0.01% LMNG to remove free rhodopsin. Rhodopsin-Gi was eluted five times with 1 CV of 20 mM HEPES (pH 7.5), 150 mM NaCl, 350 mM imidazole, 0.01% LMNG. Elution fractions were immediately concentrated using an Amicon Ultra concentrator (MWCO 100 kDa) with simultaneous buffer exchange to 20 mM HEPES (pH 7.5), 150 mM NaCl, 0.01% LMNG. Rhodopsin-Gi was mixed with molar excess (1:1.4) of Fab16 and incubated for at least 1 hr. The mixture of rhodopsin-Gi and Fab16 was concentrated using an Amicon Ultra concentrator (MWCO 30 kDa) and loaded to a Superdex 200 Increase 10/300 GL column for SEC with detergent-free buffer containing 20 mM HEPES (pH 7.5), 150 mM NaCl. Protein quality of each fraction was evaluated by UV-VIS measurement and SDS-PAGE. Fractions showing OD_{280}/OD_{380} = ~5.9 (ratio between 280 and 380 nm) was chosen for cryo-EM studies. For preparation of rhodopsin-Gi complex, purified rhodopsin and 10xHis-tag-free Gi heterotrimer were mixed in a test tube in equimolar ratio with the presence of 25 mU/ml of apyrase and incubated for 1 hr at 4°C. The protein mixture was concentrated and purified using a Superdex 200 Increase 10/300 GL column. Fractions showing OD_{280}/OD_{380} = 3 were used to prepare cryo-EM specimens.

2.4.5. Cryo-electron microscopy and image analysis

Purified samples of rhodopsin-Gi with and without Fab16 were plunge-frozen in a Vitrobot MarkIV (FEI Company) operated at 4°C and 90% humidity. A drop of 3.5 µL sample at 0.2 mg/mL was dispensed onto a glow discharged lacey

carbon grid (Ted Pella, Inc) and blotted for 2–3 s before vitrification in liquid ethane. Images were acquired by a Titan Krios microscope operated at 300 kV equipped with a Falcon III, or with a K2 Summit and GIF energy filter (Supplementary file 1). Datasets were pre-processed and pruned in FOCUS (Biyani et al., 2017) using MotionCor2 (Zheng et al., 2017) with dose weighting for movie alignment, and CTFFIND4 (Rohou and Grigorieff, 2015) for micrograph contrast transfer function estimation. Automated particle picking was performed in Gautomatch (http://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/) and all further processing steps were carried in cryoSPARC (Structura Biotechnology Inc) and RELION 2 and 3 (Kimanius et al., 2016; Zivanov et al., 2018). Around 115,000 particles from the three best resolved 3D classes were pooled and subjected to 3D auto-refinement with a soft mask which deliberately excludes the density of the AH domain observed in one of the 3D classes. Map sharpening and modulation transfer function (MTF) correction were performed with RELION post-processing. The resulting map has a nominal resolution of 4.38 Å estimated following the gold standard Fourier Shell Correlation (FSC) at FSC = 0.143. Local resolution estimation was performed using *blocres* (Cardone et al., 2013).

2.4.6. Model building and structure refinement

The initial models of rhodopsin and the Ras-like domain of G α i protein were adapted from the structure of rhodopsin-mini-Go complex (PDB id: 6FUF). The initial model of G β γ was obtained from the crystal structure of guanosine 5'-diphosphate-bound transducin (PDB id: 1GOT). The models were docked into the 3D map as rigid bodies in Chimera (Pettersen et al., 2004). After initial refinement using the phenix.real_space_refine in the Phenix suite (Adams et al., 2010), the C-tail of rhodopsin was modeled by manually building the residues 324–335 in the unassigned density near H8 and G β as a continuation of H8 using Coot (Emsley and Cowtan, 2004). The coordinates of the entire structure were improved by further refinement using phenix.real_space_refine and manual

adjustment in Coot iteratively. This final placement of the C-tail was then validated by using the molecular dynamics flexible fitting (MDFF) method (Trabuco et al., 2008). This technique allows to use electron density data as an external potential added to the molecular mechanics force field, thus taking advantage of the features present in the density while retaining a chemically sound structure. We observed that the modeled C-tail remains stable during the simulation (Figure 2.2—figure supplement 3), supporting a good agreement between the built atomic structure and the density map.

2.4.7. Structure and sequence analysis

Structural models were downloaded from the Protein Data Bank. For comparing structures of the GPCR-G protein complexes, the residue-residue contacts within 4 Å were first identified using PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC). Electrostatic potentials were calculated using the APBS method (Baker et al., 2001) as implemented in PyMOL using a concentration of 0.150 M for the +1 and −1 ion species. The biomolecular surface is colored from red (−5 kT/e) to blue (+5 kT/e) according to the potential on the soluble accessible surface. A sequence alignment of mammalian G α proteins was obtained from the GPCRdb (Pándy-Szekeres et al., 2018). The MDFF simulation (1ns) was performed using Namdinator (Trabuco et al., 2008) with default parameters (temperature: 298 K; G-force scaling factor: 0.3; Phenix RSR cycles: 5).

2.4.8. Figure preparation

Figures were prepared using ChimeraX (Goddard et al., 2018) and PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC).

2.4.9. Data availability

The cryo-EM density map is deposited under accession code EMD-4598 on the EM Data Bank. The related structure coordinates of the rhodopsin-complex

bound to Fab16 (accession code 6QNO) and the crystal structure of Fab16 (accession code 6QNK) are deposited on the Protein Data Bank.

2.5 Supplementary Information

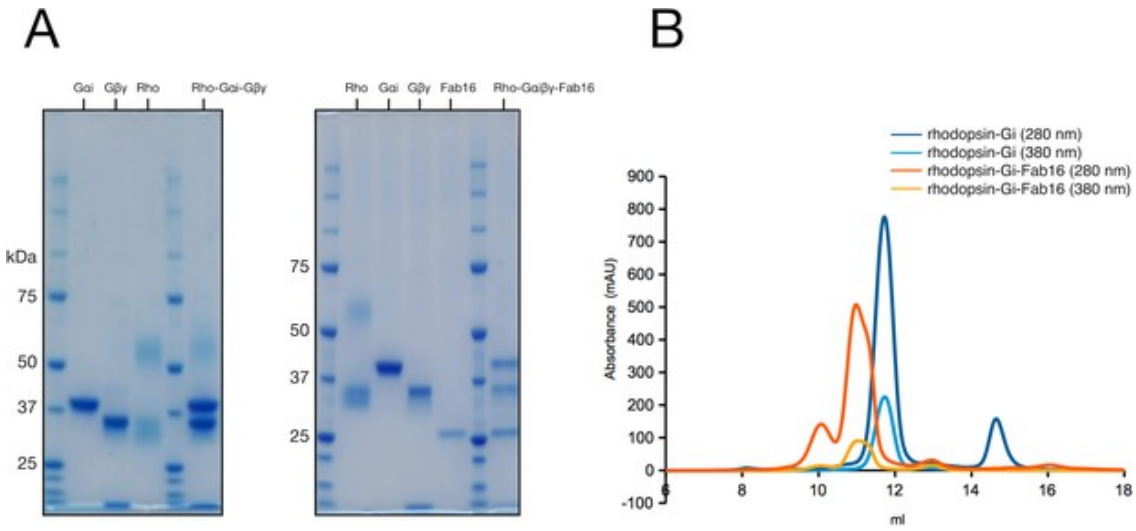

Figure 2.1sup1: Purification of the rhodopsin-Gi and rhodopsin-Gi-Fab16 complexes. (A) SDS-PAGE analysis of individual proteins and main peak fraction of corresponding size-exclusion chromatography (SEC). (B) SEC profiles of rhodopsin-Gαi and rhodopsin-Gi-Fab16 complexes used for the preparation of cryo-grids.

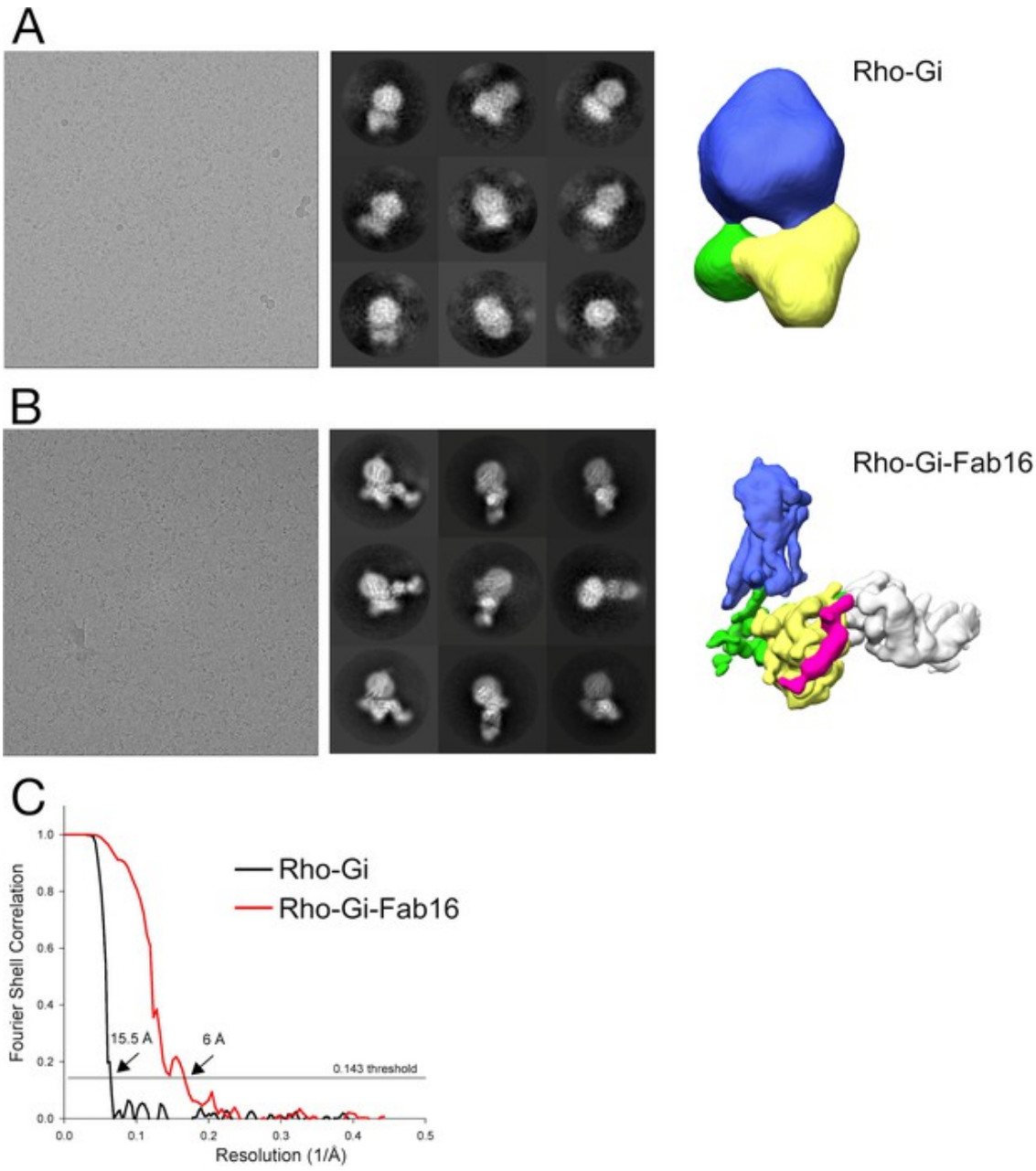

Figure 2.1sup2: Cryo-EM maps of rhodopsin-Gi complexes with and without Fab16.(A) Small dataset of rhodopsin-Gi (without Fab16) obtained using a Falcon III detector. (B) Small dataset of rhodopsin-Gi-Fab16 obtained using a Falcon III detector. (C) Gold-standard FSC curves with resolutions estimated at the 0.143 cut-off for the density maps of the two datasets.

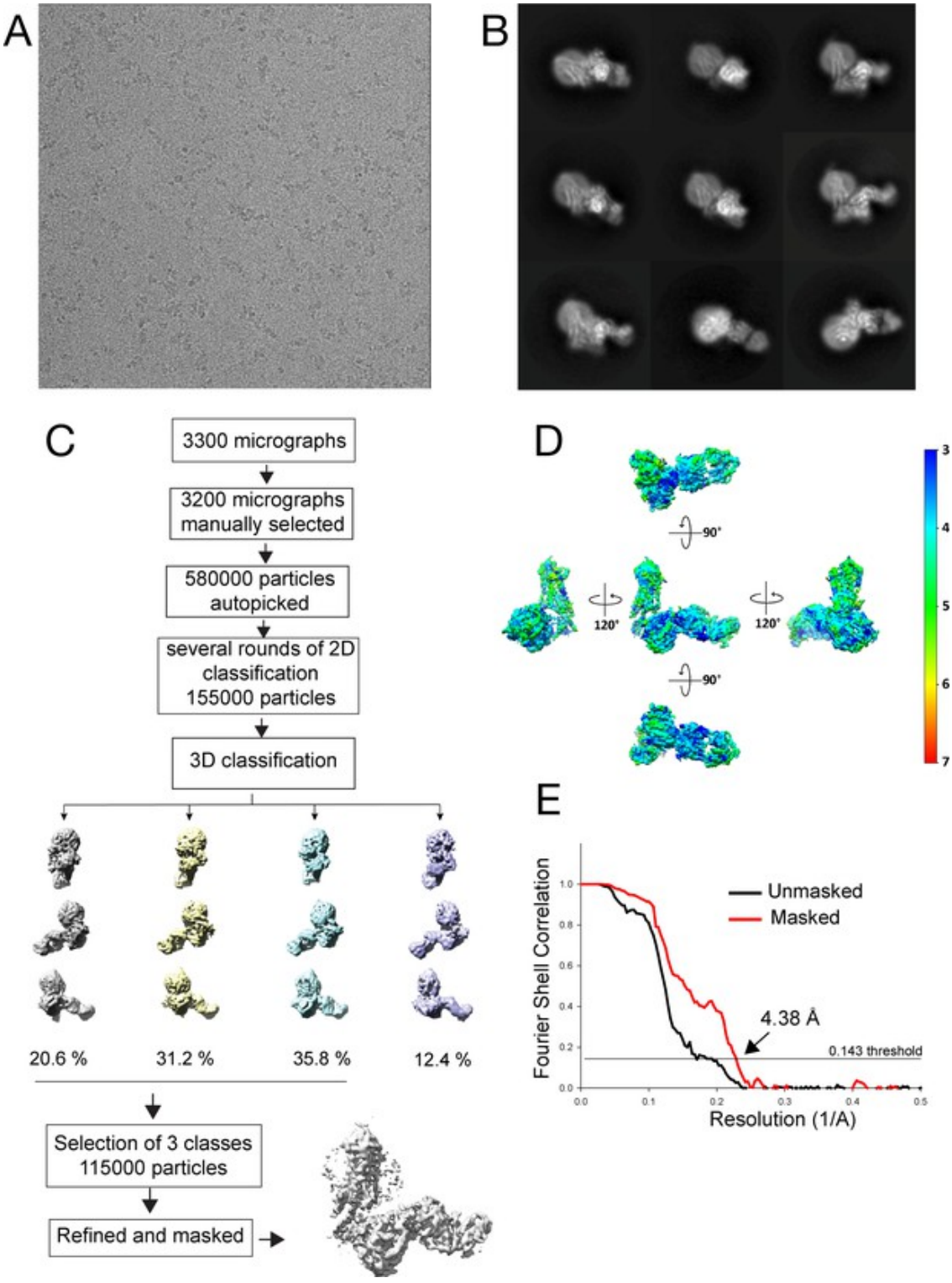

Figure 2.1sup3: Image processing of the rhodopsin-Gi-Fab16 complex acquired with K2. (A) A representative micrograph. (B) Selected 2D class

averages. (C) Image-processing workflow of the data processing. (D) Local resolution of the 3D reconstruction from (C). (E) Fourier Shell Correlation plot of the two half datasets generated in RELION.

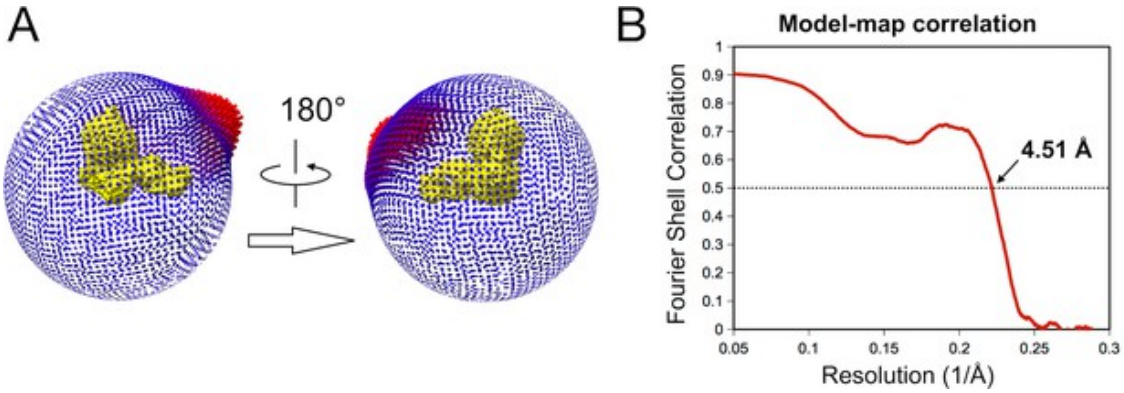

Figure 2.1sup4: Image processing details.(A) Angular distribution of the final set of particles (115'000 particles). Removal of particles that contribute to the over-represented views did not improve neither worsen map quality. (B) Spatial correlation between the 3D map and the atomic model. The correlation calculated by *mtriage* in the Phenix Suite shows the agreement between the 3D map and the atomic model to resolution.

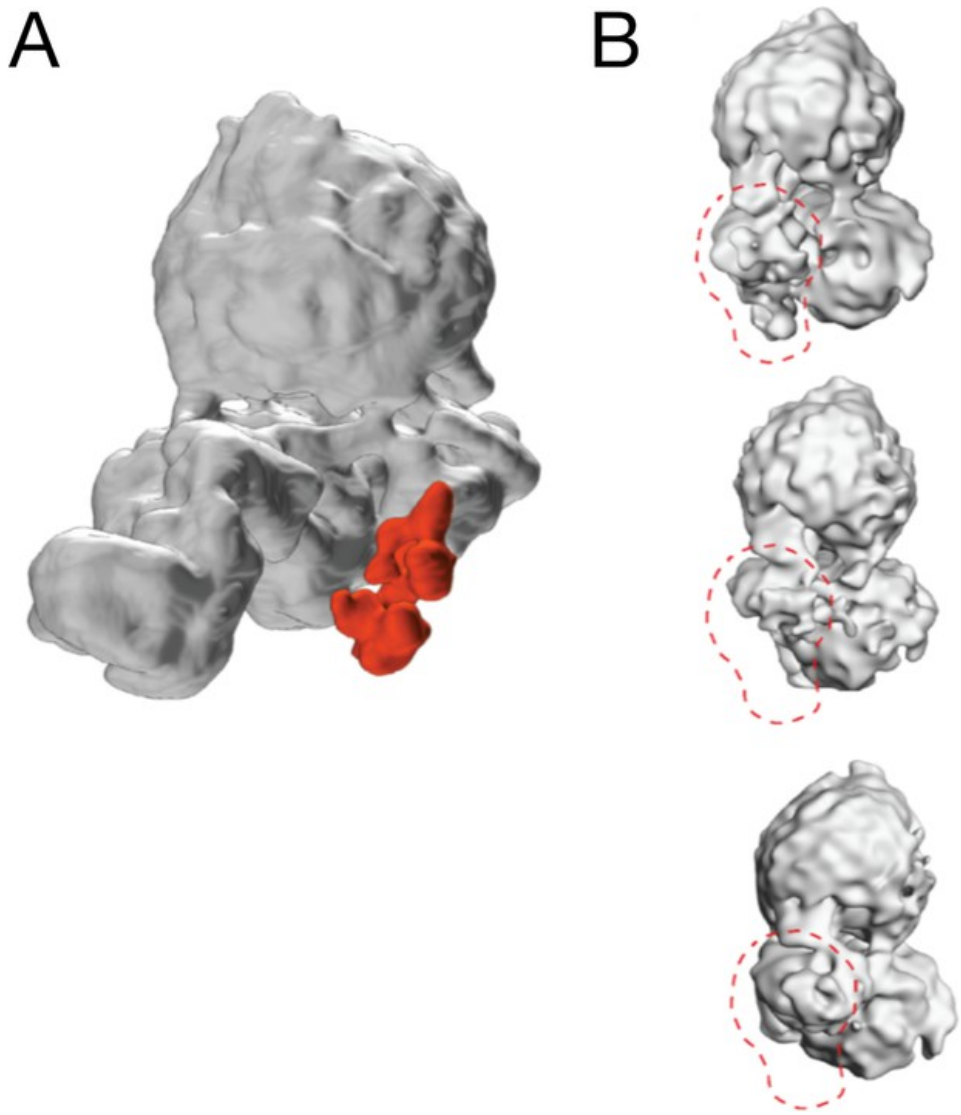

Figure 2.1sup5: 3D classification reveals the flexibility of the AH domain of G α i.(A) Density map of one 3D class obtained during classification in RELION. The AH domain of the G α subunit, highlighted in red, becomes visible only at high threshold when visualizing the density map in Chimera. (B) Overview of the three 3D classes selected after 3D classification (see Figure 1—figure supplement 3C). The region corresponding to the Ras domain of G α i, which displays higher heterogeneity, is indicated with a red circle.

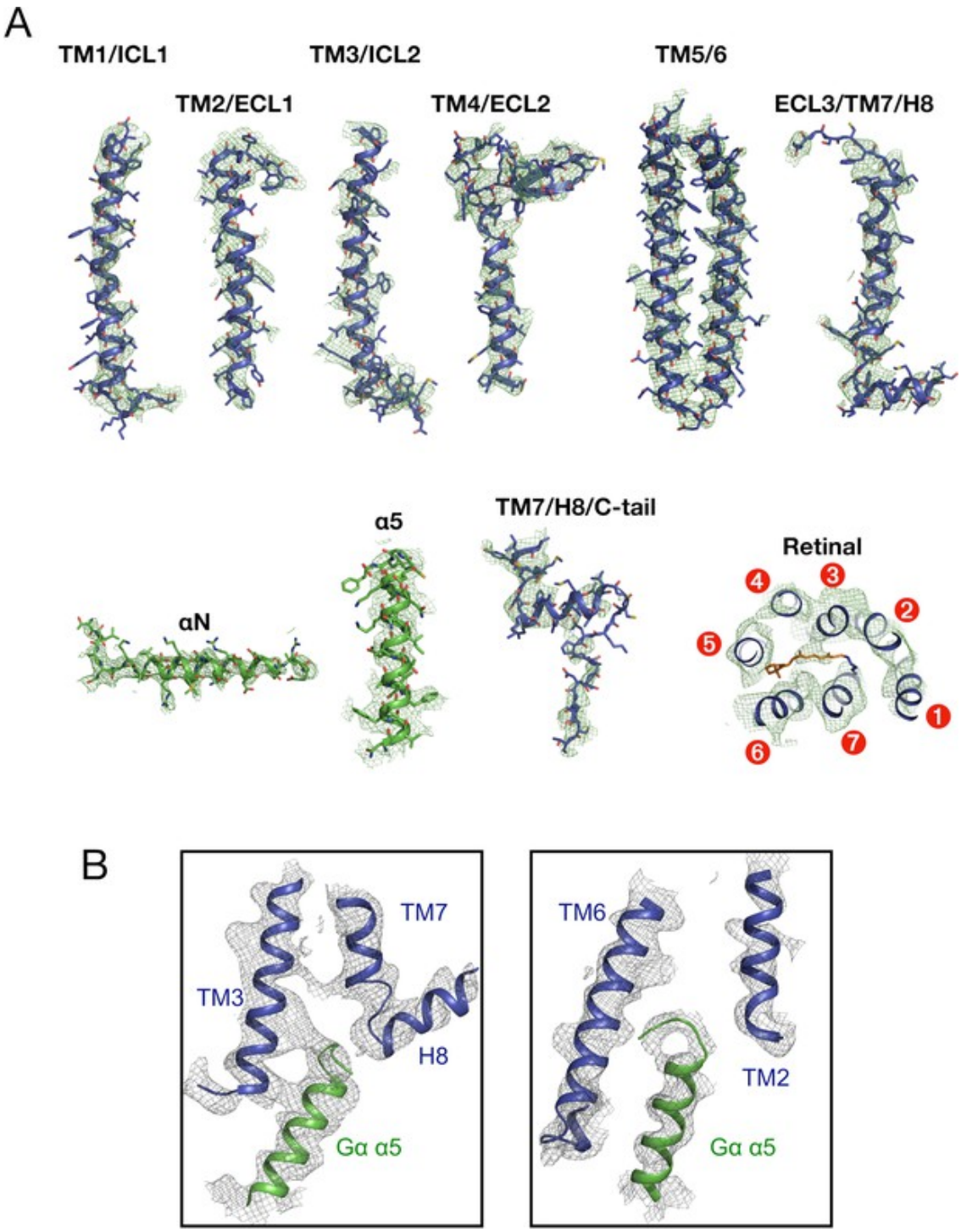

Figure 2.1sup6: Details of the cryo-EM density map of rhodopsin-Gi-Fab16 with a fitted atomic model.(A) The segmented density map shows individual regions of rhodopsin (transmembrane helices (TM),

intracellular and extracellular loops (ICL, ECL), helix 8 (H8) and the C-terminal tail (C-tail), retinal-binding pocket; rhodopsin – blue, retinal – orange) and G α i (α N and α 5 helices, G α i in green). Atoms are colored by type (nitrogen - blue; oxygen - red; sulfur - dark yellow). (B) Binding interface between the α 5 helix of G α i (in green) and the TM helices of rhodopsin (in blue). The maps are plotted at a 10 σ cuf-off using Pymol.

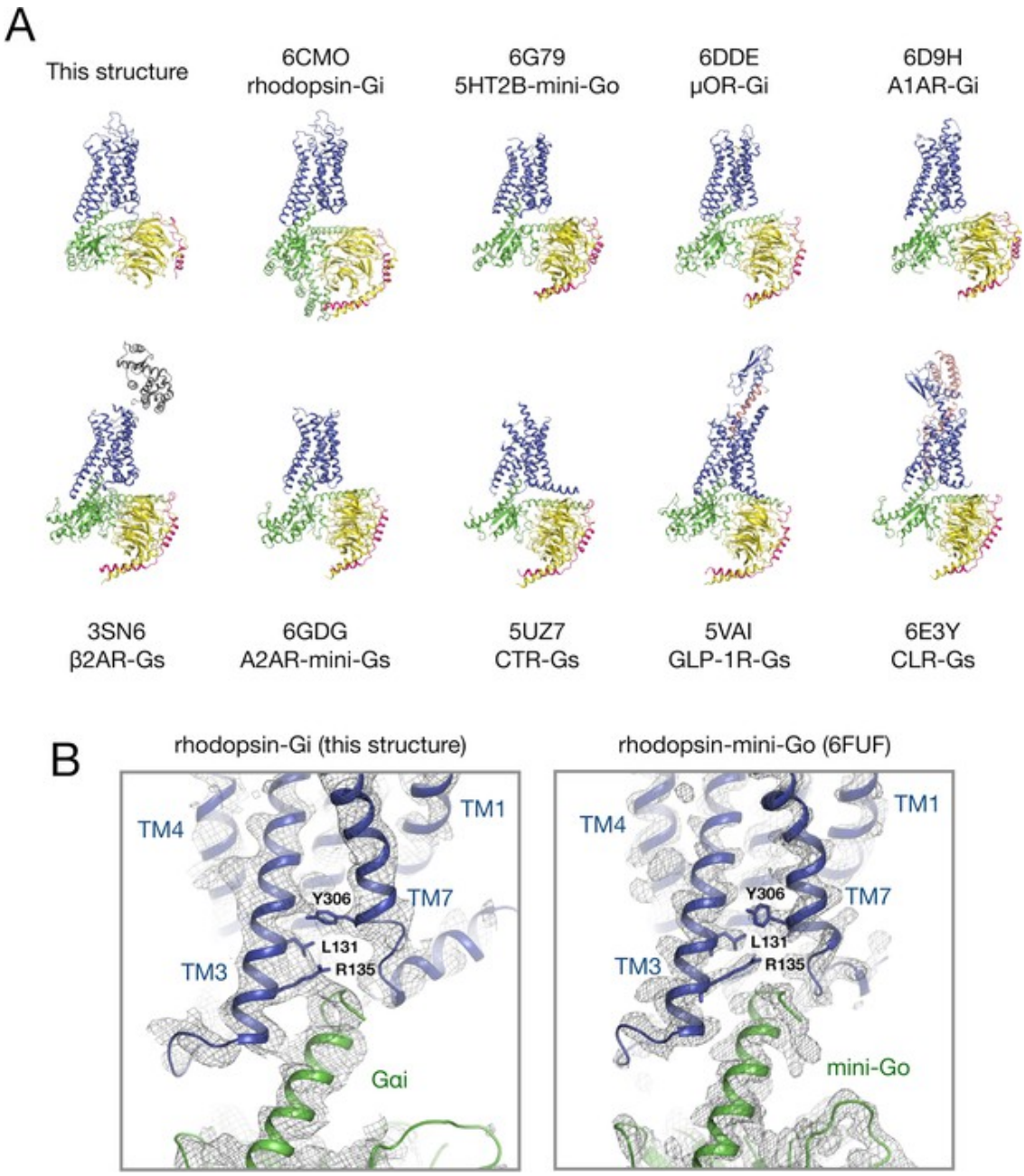

Figure 2.1sup7: Comparison of the bovine rhodopsin-Gi complex to the other GPCR-G protein complexes. (A) The complex structures are aligned to the C α atoms of the rhodopsin-Gi complex (receptor – blue; G α – green; G β – yellow; G γ – magenta; peptide ligand – salmon). (B) Binding interface between the receptor and the G α α 5 helix in the structures of rhodopsin-Gi and rhodopsin-mini-Go (PDB id: 6FUF).

Rhodopsin and G proteins are colored in blue and green, respectively. The side chains of Leu131$^{3.46}$, Arg135$^{3.50}$, Tyr306$^{7.53}$ are plotted in sticks and the density maps are shown as gray meshes.

Complex	PDB	Gα	Uniprot ID	Gβ	Uniprot ID	Gγ	Uniprot ID
bovine Rho-Gi	this work	Human Gαi1	P63096	Bovine Gβ1	P62871	Bovine Gγ1	P02698
human Rho-Gi	6CMO	Human Gαi1	P63096	Rat Gβ1	P54311	Bovine Gγ2	P63212
μOR-Gi	6DDE	Human Gαi1	P63096	Human Gβ1	P62873	Human Gγ2	P59768
5HT1BR-mini-Go	6G79	Human Gαo1 (mini)	P09471	Human Gβ1	P62873	Human Gγ2	P59768
A1AR-Gi	6D9H	Human Gαi2	P04899	Human Gβ1	P62873	Human Gγ2	P59768
CB1R-Gi	6N4B	Human Gαi1	P63096	Human Gβ1	P62873	Human Gγ2	P59768
β2AR-Gs	3SN6	Bovine Gαs1	P04896	Rat Gβ1	P54311	Bovine Gγ2	P63212
A2AR-mini-Gs	6GDG	Human Gαs2 (mini)	P63092	Human Gβ1	P62873	Human Gγ2	P59768
CTR-Gs	5UZ7	Human Gαs2	P63092	Human Gβ1	P62873	Human Gγ2	P59768
GLP-1R-Gs	5VAI	Human Gαs2	P63092	Rat Gβ1	P54311	Bovine Gγ2	P63212
GLP-1R-Gs	6B3J	Human Gαs2	P63092	Human Gβ1	P62873	Human Gγ2	P59768
CLR-Gs	6E3Y	Human Gαs2	P63092	Human Gβ1	P62873	Human Gγ2	P59768

Figure 2.1sup8: Details of the source organism of the G α , G β , and G γ proteins used to obtain GPCR G-protein complexes for structure determination.

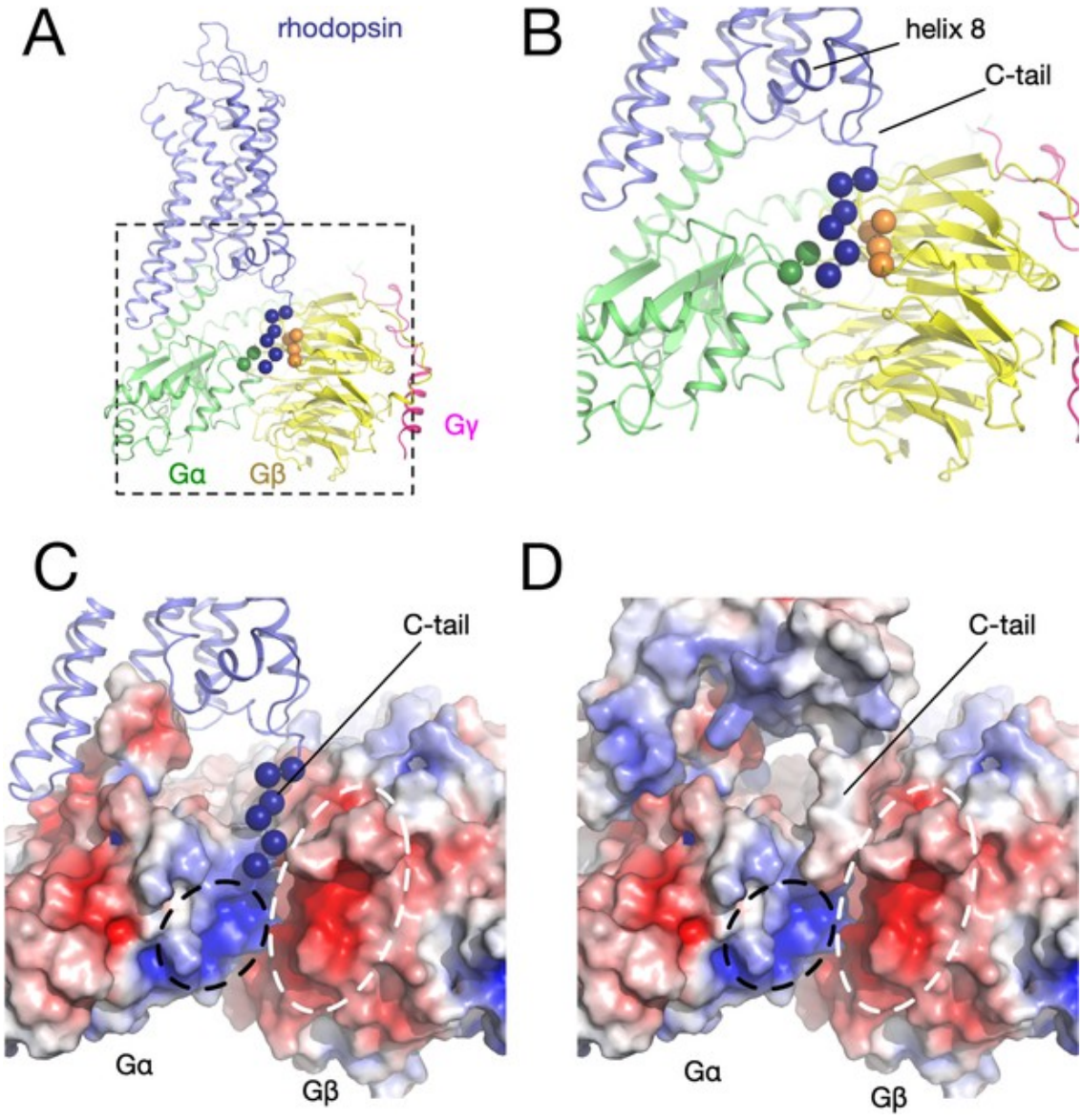

Figure 2.2sup1: Electrostatic potential. (A) Three-dimensional structure of the rhodopsin (blue) - G α (green) – G β γ (yellow and magenta) complex. The alpha carbons of the interacting residues in each component are displayed as spheres. (B) Detail of the interacting region between the receptor C-tail and the G protein. (C) Electrostatic potential mapped on the surfaces of G α (positive near the receptor C-tail, black circle) and G β γ (negative, white circle). (D)

Electrostatic potential mapped on the surfaces of the receptor, G α, and G β γ.

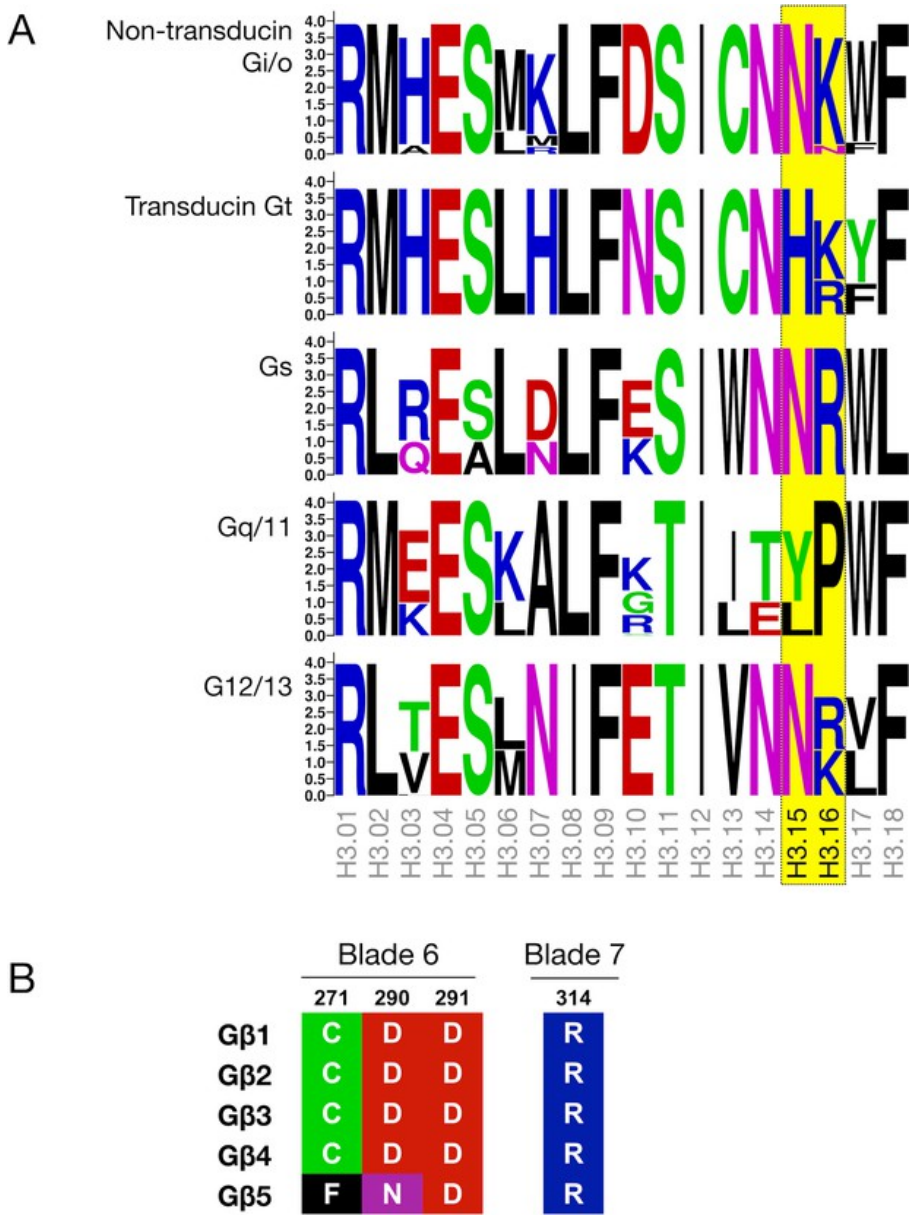

Figure 2.2sup2: Sequence conservation in G α and G β γ. (A) Sequence logo depicting the sequence conservation in the α 3 helix of the Ras domain in different G protein subtypes. The size of the letters represents frequency within a sequence alignment of mammalian G α proteins. Residues at positions H3.15 and H3.16, which can potentially interact

with the C-tail of the receptor, are highlighted in yellow. (B) Alignment of residues C271, D290, and D291 in blade six and R314 in blade 7 of the β-propeller in the five human Gβ subtypes.

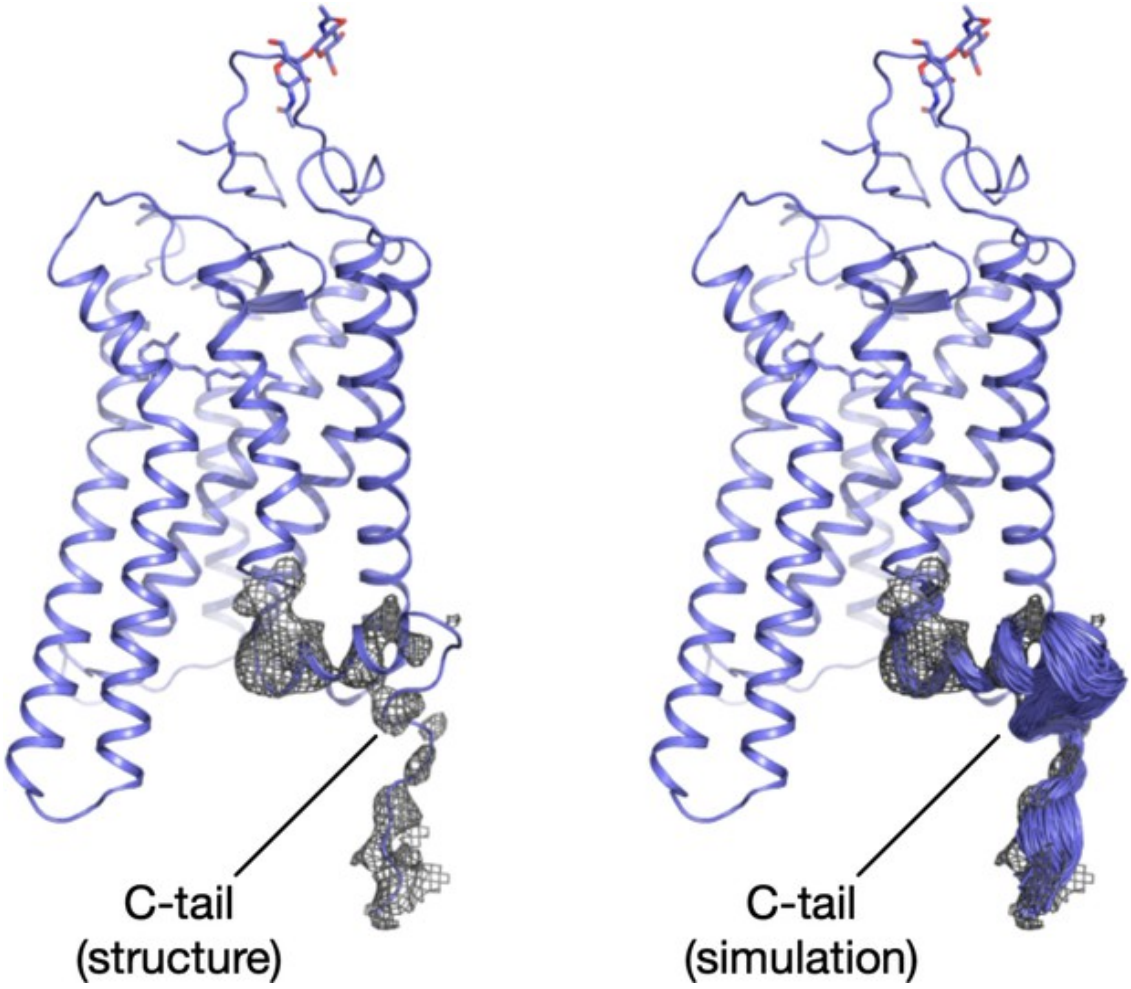

Figure 2.2sup3: Flexible fitting of the C-tail in the electron density. Structural model of rhodopsin displaying the electron density at a $10\,\sigma$ cuf-off around H8 and the C-tail (left), and snapshots of this region during the MDFF simulation (right).

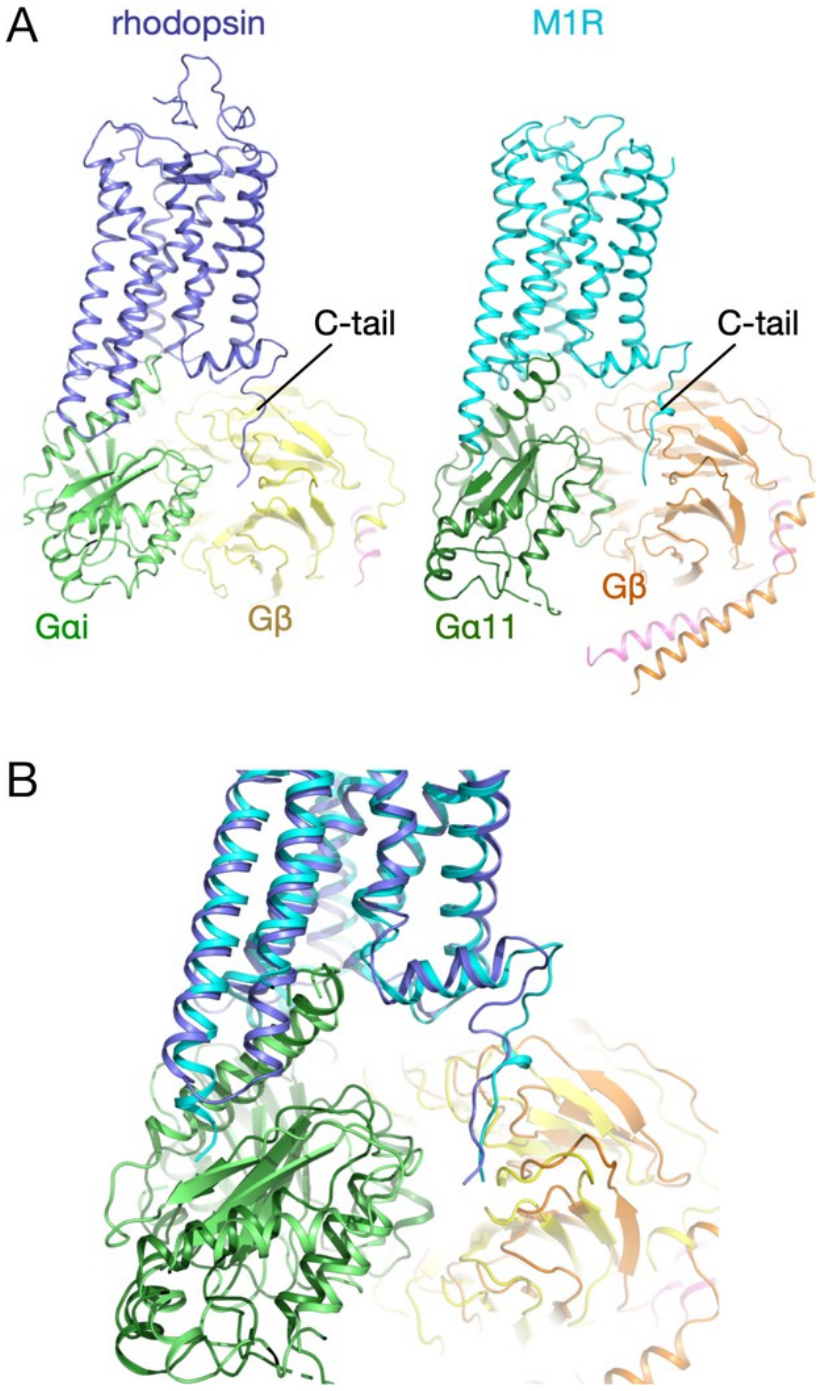
A
rhodopsin
M1R
C-tail
C-tail
Gαi
Gβ
Gα11
Gβ
B

Figure 2.2sup4: Structure of the C-tail in rhodopsin and M1R. (A) Three-dimensional structure of the rhodopsin (blue) - G α i (green) - G β (yellow) complex. (B) Three-dimensional structure of the M1 muscarinic acetylcholine receptor (blue) – G α 11 (green) - G β (orange) complex. (C) Structural superposition of the complexes.

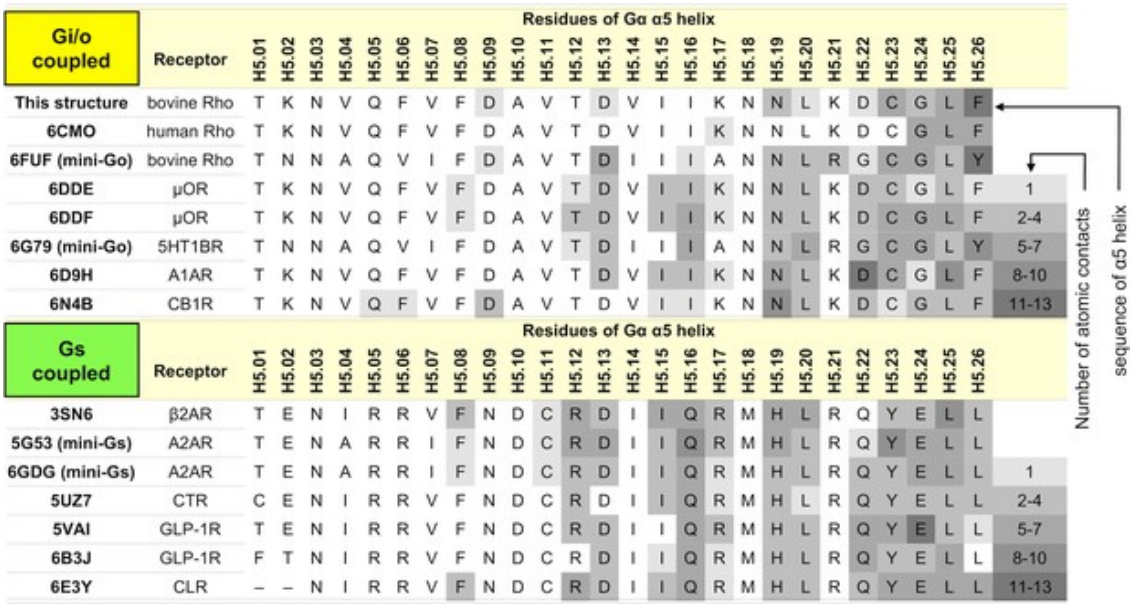

Figure 2.3sup1: Residue-residue contact list between GPCRs and G α H5. 11–26 within 4 Å.

Gα α5:	H5.11	H5.12	H5.13	H5.15	H5.16	H5.17	H5.19	H5.20	H5.22	H5.23	H5.24	H5.25	H5.26
Gi/o coupled													
Rho: This structure	-	-	Q237-5.72	-	-	-	V138-3.53	V139-3.54	L72-2.39	L72-2.39 R135-3.50 V138-3.53	R135-3.50 N310-8.47	R135-3.50 M257-6.40	T242-6.25 K245-6.28 A246-6.29 E249-6.32 K311-8.48
Rho: 6CMO	-	-	-	-	-	S240-ICL3	-	-	-	-	M309-7.56 N310-8.47 K311-8.48	R135-3.50 E249-6.32 M253-6.36	T242-6.25 K245-6.28 A246-6.29 K311-8.48
Rho: 6FUF	-	-	Q237-5.72 S240-ICL3 T243-6.26	-	A233-5.68	-	V138-3.53	V139-3.54 A246-6.29 V250-6.33	L72-2.39	R135-3.50 N310-8.47	N310-8.47 K311-8.48	R135-3.50 M257-6.40	T242-6.25 K245-6.28 A246-6.29 K311-8.48
5HT1BR: 6G79	-	V155-ICL2	R238-5.68	-	I151-3.43 A154-ICL2 R238-5.68	-	A150-3.53 I151-3.54	I151-3.54 A235-5.65 R238-5.68	S372-7.56 N373-8.47	R147-3.50 S372-7.56	K311-6.32 T315-6.36 S372-7.56	I231-5.61 A312-6.33	R238-5.68 I239-5.69 N308-6.29 K311-6.32
μOR: 6DDE	-	P172-ICL2	V262-5.68 R263-ICL3 M264-ICL3	P172-ICL2 L176-ICL2	V169-3.54 P172-ICL2 R258-5.64	M264-ICL3	A168-3.53 R179-ICL2	V169-3.54 L259-5.65	T103-2.39	T103-2.39 R165-3.50	D340-8.47	R277-6.32 M281-6.36	I278-6.33
A1AR: 6D9H	-	-	Q210-5.68 K213-5.71	P112-ICL2 L113-ICL2	P112-ICL2 Q210-5.68	L211-5.69	R108-3.53 P112-ICL2	V109-3.43 I207-5.65	D42-2.37 F45-2.40 R108-3.53 I292-8.47 K294-8.49	R105-3.50 R108-3.53 I292-8.47	R291-7.56	R105-3.50 V203-5.61 I207-5.65 I232-6.33 L236-6.37	K228-6.29
CB1R: 6N4B	-	-	-	L222-ICL2	P221-ICL2	-	S217-3.53 P221-ICL2 Y224-ICL2	S217-3.53 I218-3.54	R150-2.37 S152-2.39	S152-2.39	S401-8.47	R214-3.50 L341-6.33	M337-6.29 R340-6.32
Gs coupled													
β2AR: 3SN6	F139-ICL2	T136-3.55 P138-ICL2 F139-ICL2 K140-ICL2	Q229-5.68 K232-5.71	P138-ICL2 Q142-ICL2	I135-3.54 P138-ICL2 E225-5.64 R228-5.67 Q229-5.68	Q229-5.68 I233-5.72	A134-3.53 Y141-ICL2 Q142-ICL2	I135-3.54 A226-5.65 Q229-5.68	-	R131-3.50	K270-6.32 T274-6.36	I135-3.54 V222-5.61 A226-5.65 A271-6.33 T274-6.36 L275-6.37	L230-5.69
A2AR: 5G53	L110-ICL2	R107-3.55 I108-3.56 P109-ICL2 L110-ICL2	Q207-5.68 Q210-5.71	P109-ICL2	I106-3.54 P109-ICL2 A203-5.64 Q207-5.68	Q207-5.68	A105-3.53 I106-3.54 Y112-ICL2	I106-3.54 A204-5.65	I292-8.47	R102-3.50 A105-3.53 R291-7.56	R291-7.56 R293-8.48 R296-8.51	I200-5.61 A231-6.33 L235-6.37 R291-7.56	L308-5.69 K227-6.29
A2AR: 6GDG	L110-ICL2	R107-3.55 P109-ICL2 L110-ICL2	Q207-5.68	P109-ICL2	I106-3.54 P109-ICL2 Q207-5.68	Q207-5.68	A105-3.53 I106-3.54 Y112-ICL2	I106-3.54 A204-5.65	I292-8.47 R293-8.48	R102-3.50	S234-6.36 R291-7.56 R293-8.48	I200-5.61 A231-6.33 S234-6.36 L235-6.37	L308-5.69 K227-6.29
CTR: 5UZ7	-	V252-ICL2 K326-5.64	-	V252-ICL2	I248-3.58 V249-3.59 V252-ICL2 K326-5.64	K326-5.64	I247-3.57 T254-ICL2	L323-5.61	R180-2.46	R180-2.46 T243-3.53 L244-3.54	L348-6.45 C394-7.60	T345-6.42 L348-6.45	K327-5.65
GLP-1R: 5VAI	-	A256-3.59	K334-5.64	-	L255-3.58	K334-5.64	L254-3.57	V331-5.61	R176-2.46 E408-8.49	R180-2.50 L251-3.54 L356-6.45 L359-6.48	S352-6.41 L401-7.56 Y402-7.57 V405-7.60 N406-8.47 N407-8.48	S352-6.41 T353-6.42	L339-ICL3
GLP-1R: 6B3J	-	-	K334-5.64	S258-ICL2	L255-3.58 K334-5.64	K334-5.64	L254-3.57 L255-3.58	L255-3.58 V331-5.61	R176-2.46	Y250-3.53 L251-3.54	N407-8.48	S352-6.41 L356-6.45	-
CLR: 6E3Y	-	V242-3.59 V243-3.60 V245-ICL2 F246-ICL2	K319-5.64	V245-ICL2 F246-ICL2	I241-3.58 V242-3.59 V245-ICL2 K319-5.64	K319-5.64	L240-3.57	I241-3.58 L316-5.61	R173-2.46	R173-2.46 T236-3.53 L237-3.54	R336-6.40 F387-7.60 N388-8.47 G389-8.48	L316-5.61 R336-6.40 A337-6.41 L341-6.45	L316-5.61 L320-5.65 K333-6.37 R336-6.40

Figure 2.3sup2: Residue-residue contact list between GPCRs and G α H5. 11–26

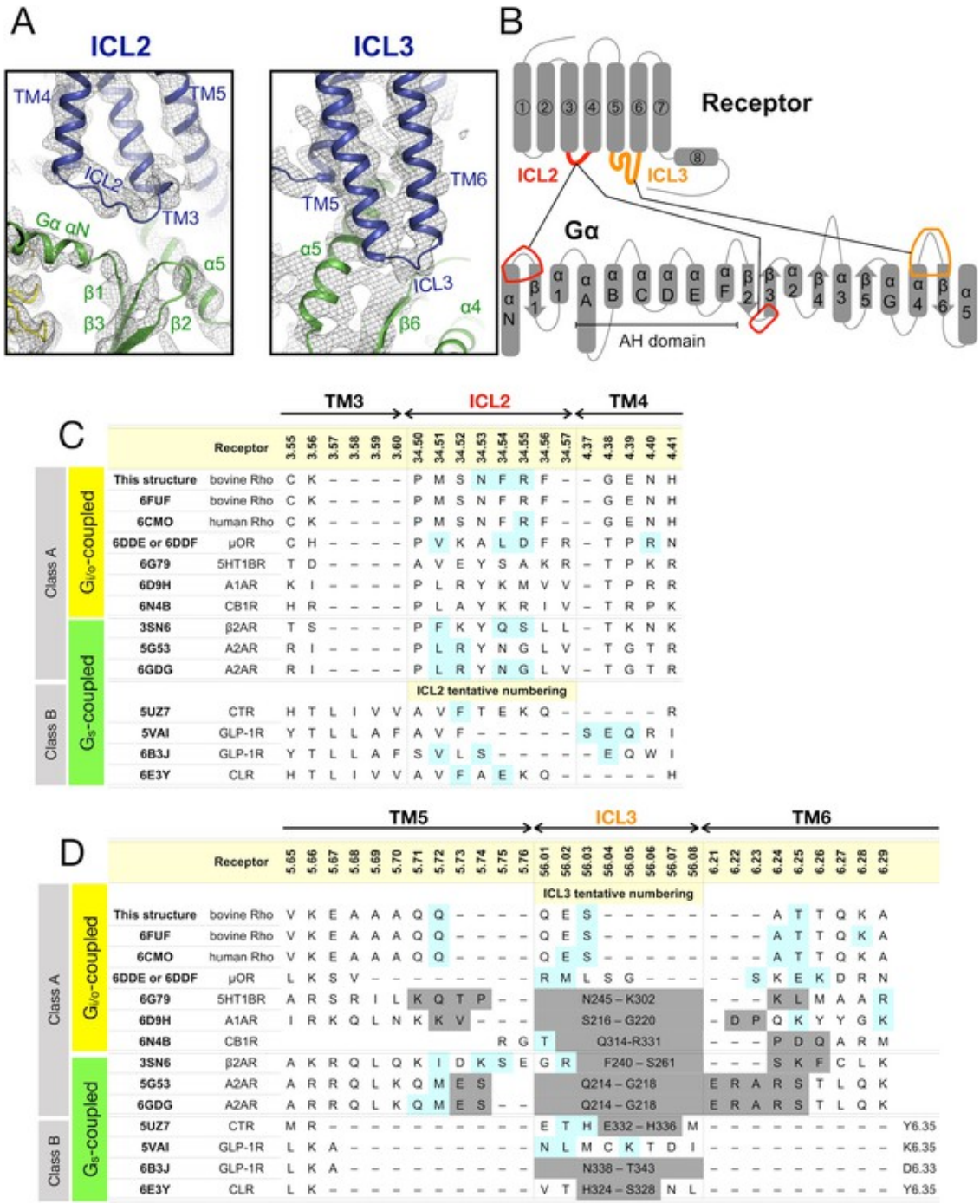

Figure 2.3sup3: Contacts observed between ICL2/ICL3 and the G α subunit. (A) Intracellular loop (ICL) 2 and 3 in the EM density map (B) Schematic overview of the contacts between the intracellular loops of the receptor and G α. (C) Sequence of ICL2 in the available structures of the

GPCR-G protein complexes. Residues that contact to regions of the G protein other than the $\alpha 5$ helix are highlighted in cyan. (D) Same as panel B, but for ICL3. Residues missing from the structure are marked in gray. The residues are numbered using the Ballesteros-Weinstein scheme.

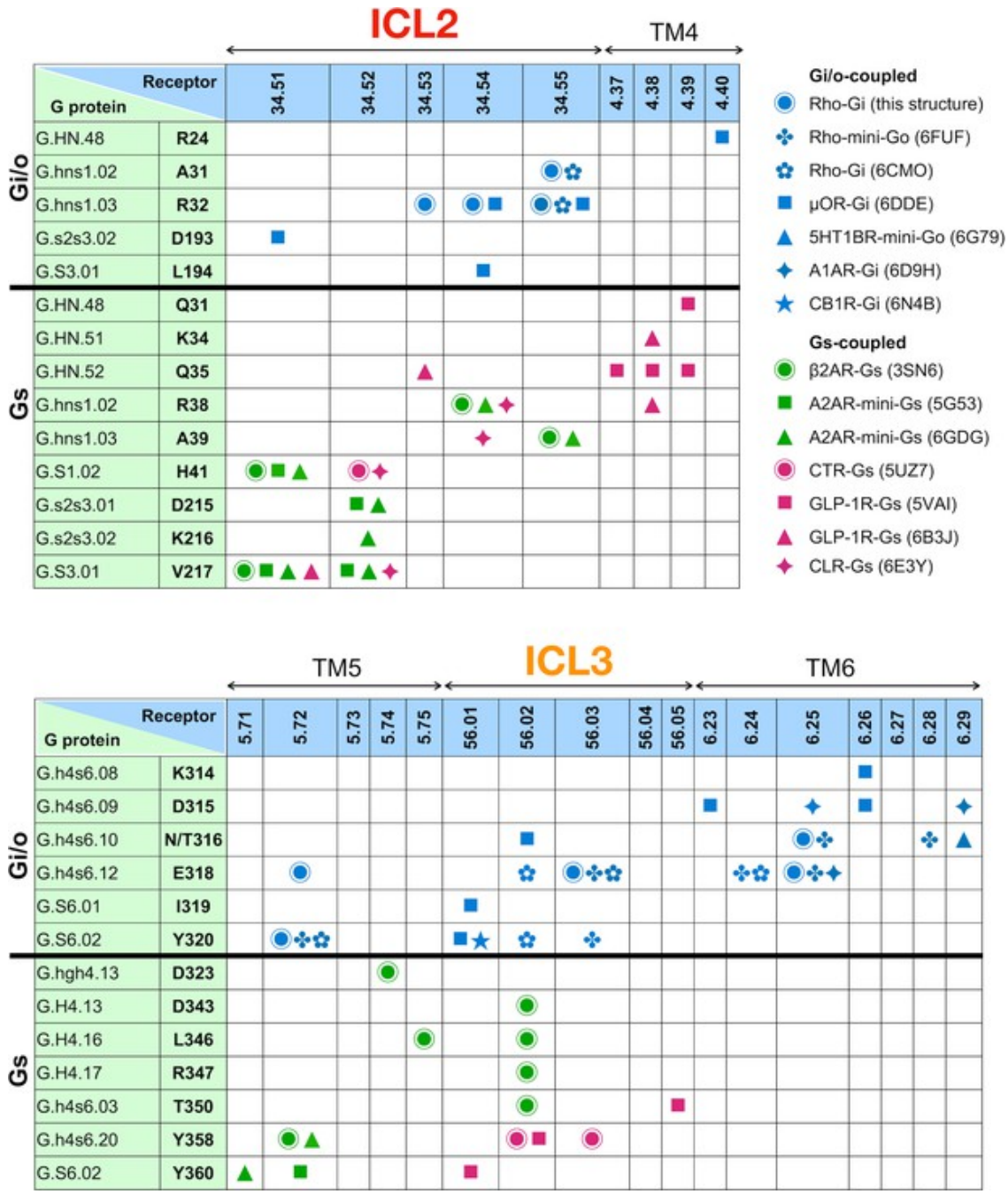

Figure 2.3sup4: Residue-residue contacts between ICL2/3 and G α . Contacts at the regions of ICL2 and 3 are identified using a 4 Å cut-off. Residue-residue contacts are listed for the regions of ICL2 (top panel) and ICL3 (bottom panel). The residues of G α are numbered following the common G α protein numbering (CGN). Contacts observed in the

individual structures are noted as colored bullet symbols (blue – Gi-coupled receptors; green – Gs-coupled, class A receptors; magenta – Gs-coupled, class B receptors).

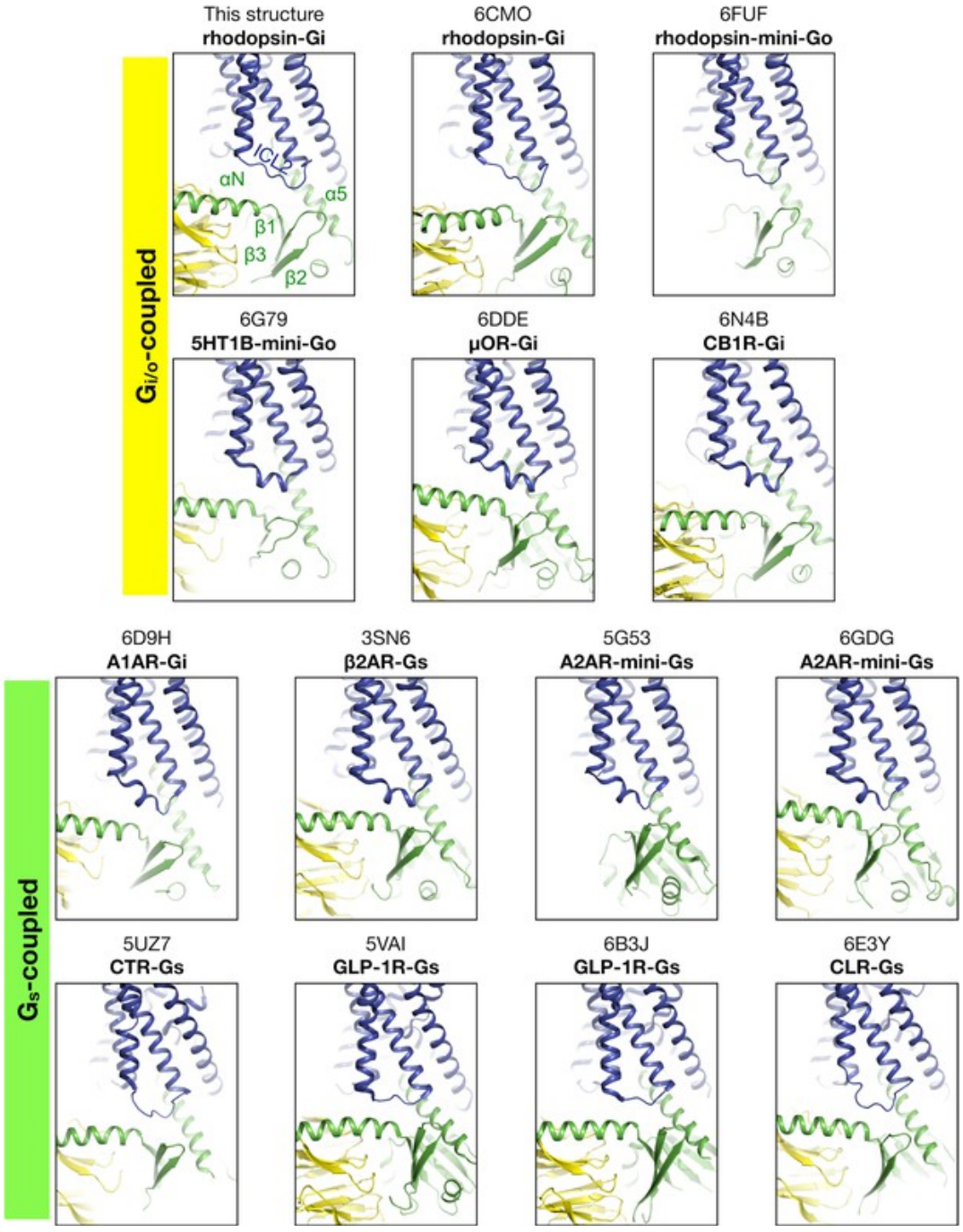

Figure 2.3sup5: Region near ICL2 in available structures. View of the existing GPCR-G protein complex structures centered at the ICL2 region. The structures are aligned to the C α atoms of rhodopsin. The regions of the G protein involved in contacting ICL2 are the αN/β1 and β2/β3 turns, and the α5 helix.

2.6. Bibliography

'PHENIX: a comprehensive Python-based system for macromolecular structure solution'

PD Adams PV Afonine G Bunkóczi VB Chen IW Davis N Echols JJ Headd LW Hung GJ Kapral RW Grosse-Kunstleve AJ McCoy NW Moriarty R Oeffner RJ Read DC Richardson JS Richardson TC Terwilliger PH Zwart (2010)

Acta Crystallographica Section D Biological Crystallography 66:213–221.

doi.org/10.1107/S0907444909052925

C-terminal threonines and serines play distinct roles in the desensitization of rhodopsin, a G protein-coupled receptor

AW Azevedo T Doan H Moaven I Sokal F Baameur SA Vishnivetskiy KT Homan JJG Tesmer VV Gurevich J Chen F Rieke (2015)

eLife 4:e05981.

doi.org/10.7554/eLife.05981

Electrostatics of nanosystems: application to microtubules and the ribosome

NA Baker D Sept S Joseph MJ Holst JA McCammon (2001)

PNAS 98:10037–10041.

doi.org/10.1073/pnas.181342398

Focus: the interface between data collection and data processing in cryo-EM

N Biyani RD Righetto R McLeod D Caujolle-Bert D Castano-Diez KN Goldie H Stahlberg (2017)

Journal of Structural Biology 198:124–133.

doi.org/10.1016/j.jsb.2017.03.007

How the ubiquitous GPCR receptor family selectively activates signalling pathways

MJ Capper D Wacker (2018)

Nature 558:529–530.

doi.org/10.1038/d41586-018-05503-4

One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions

G Cardone JB Heymann AC Steven (2013)

Journal of Structural Biology 184:226–236.

doi.org/10.1016/j.jsb.2013.08.002

Conformational changes in the G protein gs induced by the β 2 adrenergic receptor

KY Chung SG Rasmussen T Liu S Li BT DeVree PS Chae D Calinski BK Kobilka VL Woods RK Sunahara (2011)

Nature 477:611–615.

doi.org/10.1038/nature10488

Endocytosis of G protein-coupled receptors: roles of G protein-coupled receptor kinases and beta-arrestin proteins

A Claing SA Laporte MG Caron RJ Lefkowitz (2002)

Progress in Neurobiology 66:61–79.

doi.org/10.1016/S0301-0082(01)00023-5

Stabilized G protein binding site in the structure of constitutively active metarhodopsin-II

X Deupi P Edwards A Singhal B Nickle D Oprian G Schertler J Standfuss (2012)

PNAS 109:119–124.

doi.org/10.1073/pnas.1114089108

Structure of the adenosine-bound human Adenosine A1 receptor–Gi complex

CJ Draper-Joyce M Khoshouei DM Thal Y-L Liang ATN Nguyen SGB Furness H Venugopal J-A Baltos JM Plitzko R Danev W Baumeister LT May D Wootten PM Sexton A Glukhova A Christopoulos (2018)

Nature 558:559–563.

doi.org/10.1038/s41586-018-0236-6

SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins

RO Dror TJ Mildorf D Hilger A Manglik DW Borhani DH Arlow A Philippsen N Villanueva Z Yang MT Lerch WL Hubbell BK Kobilka RK Sunahara DE Shaw (2015)

Science 348:1361–1365.

doi.org/10.1126/science.aaa5264

Seven transmembrane receptor core signaling complexes are assembled prior to plasma membrane trafficking

DJ Dupré M Robitaille N Ethier LR Villeneuve AM Mamarbachi TE Hébert (2006)

Journal of Biological Chemistry 281:34561–34573.

doi.org/10.1074/jbc.M605012200

The role of gbetagamma subunits in the organization, assembly, and function of GPCR signaling complexes

DJ Dupré M Robitaille RV Rebois TE Hébert (2009)

Annual Review of Pharmacology and Toxicology 49:31–56.

doi.org/10.1146/annurev-pharmtox-061008-103038

Coot: model-building tools for molecular graphics

P Emsley K Cowtan (2004)

Acta Crystallographica. Section D, Biological Crystallography 60:2126–2132.

doi.org/10.1107/S0907444904019158

Universal allosteric mechanism for g α activation by GPCRs

T Flock CNJ Ravarani D Sun AJ Venkatakrishnan M Kayikci CG Tate DB Veprintsev MM Babu (2015)

Nature 524:173–179.

doi.org/10.1038/nature14663

Selectivity determinants of GPCR-G-protein binding

T Flock AS Hauser N Lund DE Gloriam S Balaji MM Babu (2017)

Nature 545:317–322.

doi.org/10.1038/nature22070

Real-time monitoring of receptor and G-protein interactions in living cells

C Galés RV Rebois M Hogue P Trieu A Breit TE Hébert M Bouvier (2005)

Nature Methods 2:177–184.

doi.org/10.1038/nmeth743

Cryo-EM structure of the adenosine A2A receptor coupled to an engineered heterotrimeric G protein

J García-Nafría Y Lee X Bai B Carpenter CG Tate (2018a)

eLife 7:e35946.

doi.org/10.7554/eLife.35946

Cryo-EM structure of the serotonin 5-HT1B receptor coupled to heterotrimeric Go

J García-Nafría R Nehmé PC Edwards CG Tate (2018b)

Nature 558:620–623.

doi.org/10.1038/s41586-018-0241-9

Rules of engagement: gpcrs and G proteins

A Glukhova CJ Draper-Joyce RK Sunahara A Christopoulos D Wootten PM Sexton (2018)

ACS Pharmacology & Translational Science 1:73–83.

doi.org/10.1021/acsptsci.8b00026

UCSF ChimeraX: meeting modern challenges in visualization and analysis

TD Goddard CC Huang EC Meng EF Pettersen GS Couch JH Morris TE Ferrin (2018)

Protein Science 27:14–25.

doi.org/10.1002/pro.3235

Beta-arrestin acts as a clathrin adaptor in endocytosis of the beta2-adrenergic receptor

OB Goodman JG Krupnick F Santini VV Gurevich RB Penn AW Gagnon JH Keen JL Benovic (1996)

Nature 383:447–450.

doi.org/10.1038/383447a0

Lack of beta-arrestin signaling in the absence of active G proteins

M Grundmann N Merten D Malfacini A Inoue P Preis K Simon N Rüttiger N Ziegler T Benkel NK Schmitt S Ishida I Müller R Reher K Kawakami A Inoue U Rick T Kühl D Imhof J Aoki GM König C Hoffmann J Gomeza J Wess E Kostenis (2018)

Nature Communications 9:341.

doi.org/10.1038/s41467-017-02661-3

G protein-coupled receptor kinases: more than just kinases and not only for GPCRs

EV Gurevich JJ Tesmer A Mushegian VV Gurevich (2012)

Pharmacology & Therapeutics 133:40–69.

doi.org/10.1016/j.pharmthera.2011.08.001

Trends in GPCR drug discovery: new agents, targets and indications

AS Hauser MM Attwood M Rask-Andersen HB Schiöth DE Gloriam (2017)

Nature Reviews Drug Discovery 16:829–842.

doi.org/10.1038/nrd.2017.178

Sequence of interactions in receptor-G protein coupling

R Herrmann M Heck P Henklein P Henklein C Kleuss KP Hofmann OP Ernst (2004)

Journal of Biological Chemistry 279:24283–24290.

doi.org/10.1074/jbc.M311166200

Rhodopsin-transducin coupling: role of the galpha C-terminus in nucleotide exchange catalysis

R Herrmann M Heck P Henklein C Kleuss V Wray KP Hofmann OP Ernst (2006)

Vision Research 46:4582–4593.

doi.org/10.1016/j.visres.2006.07.027

A G protein-coupled receptor at work: the rhodopsin model

KP Hofmann P Scheerer PW Hildebrand HW Choe JH Park M Heck OP Ernst (2009)

Trends in Biochemical Sciences 34:540–552.

doi.org/10.1016/j.tibs.2009.07.005

The G protein-coupled receptor rhodopsin: a historical perspective

L Hofmann K Palczewski (2015)

Methods in Molecular Biology 1271:3–18.

doi.org/10.1007/978-1-4939-2330-4_1

G protein-coupled receptors show unusual patterns of intrinsic unfolding

V-P Jaakola J Prilusky JL Sussman A Goldman (2005)

Protein Engineering, Design and Selection 18:103–110.

doi.org/10.1093/protein/gzi004

Cryo-EM structure of human rhodopsin bound to an inhibitory G protein

Y Kang O Kuybeda PW de Waal S Mukherjee N Van Eps P Dutka XE Zhou A Bartesaghi S Erramilli T Morizumi X Gu Y Yin P Liu Y Jiang X Meng G Zhao K Melcher OP Ernst AA Kossiakoff S Subramaniam HE Xu (2018)

Nature 558:553–558.

doi.org/10.1038/s41586-018-0215-y

Structural evidence for a sequential release mechanism for activation of heterotrimeric G proteins

N Kapoor ST Menon R Chauhan P Sachdev TP Sakmar (2009)

Journal of Molecular Biology 393:882–897.

doi.org/10.1016/j.jmb.2009.08.043

The expanding roles of g β γ subunits in G protein-coupled receptor signaling and drug action

SM Khan R Sleno S Gora P Zylbergold JP Laverdure JC Labbé GJ Miller TE Hébert (2013)

Pharmacological Reviews 65:545–577.

doi.org/10.1124/pr.111.005603

Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2

D Kimanius BO Forsberg SH Scheres E Lindahl (2016)

eLife 5:e18722.

doi.org/10.7554/eLife.18722

Structure of the μ-opioid receptor-Gi protein complex

A Koehl H Hu S Maeda Y Zhang Q Qu JM Paggi NR Latorraca D Hilger R Dawson H Matile GFX Schertler S Granier WI Weis RO Dror A Manglik G Skiniotis BK Kobilka (2018)

Nature 558:547–552.

doi.org/10.1038/s41586-018-0219-7

Structural and functional analysis of a β2-Adrenergic Receptor Complex with GRK5

KE Komolov Y Du NM Duc RM Betz J Rodrigues RD Leib D Patra G Skiniotis CM Adams RO Dror KY Chung BK Kobilka JL Benovic (2017)

Cell 169:407–421.

doi.org/10.1016/j.cell.2017.03.047

Structure of a signaling cannabinoid receptor 1-G protein complex

K Krishna Kumar M Shalev-Benami MJ Robertson H Hu SD Banister SA Hollingsworth NR Latorraca HE Kato D Hilger S Maeda WI Weis DL Farrens RO Dror SV Malhotra BK Kobilka G Skiniotis (2019)

Cell 176:448–458.

doi.org/10.1016/j.cell.2018.11.040

Crystallization scale preparation of a stable GPCR signaling complex between constitutively active rhodopsin and G-protein

S Maeda D Sun A Singhal M Foggetta G Schmid J Standfuss M Hennig RJ Dawson DB Veprintsev GF Schertler (2014)

PLOS ONE 9:e98714.

doi.org/10.1371/journal.pone.0098714

Development of an antibody fragment that stabilizes GPCR/G-protein complexes

S Maeda A Koehl H Matile H Hu D Hilger GFX Schertler A Manglik G

Skiniotis RJP Dawson BK Kobilka (2018)

Nature Communications 9:3712.

doi.org/10.1038/s41467-018-06002-w

Structures of the M1 and M2 muscarinic acetylcholine receptor/G-protein complexes

S Maeda Q Qu MJ Robertson G Skiniotis BK Kobilka (2019)

Science 364:552–557.

doi.org/10.1126/science.aaw5188

Regulation of GPCR activity, trafficking and localization by GPCR-interacting proteins

AC Magalhaes H Dunn SS Ferguson (2012)

British Journal of Pharmacology 165:1717–1736.

doi.org/10.1111/j.1476-5381.2011.01552.x

A docking site for G protein β γ subunits on the parathyroid hormone 1 receptor supports signaling through multiple pathways

MJ Mahon TM Bonacci P Divieti AV Smrcka (2006)

Molecular Endocrinology 20:136–146.

doi.org/10.1210/me.2005-0169

Functional analysis of farnesylation and methylation of transducin

T Matsuda Y Fukada (2000)

Methods in Enzymology 316:465–481.

Heterotrimeric G-proteins: a short history

G Milligan E Kostenis (2006)

British Journal of Pharmacology 147:S46–S55.

doi.org/10.1038/sj.bjp.0706405

An online resource for GPCR structure determination and analysis

C Munk E Mutt V Isberg LF Nikolajsen JM Bibbe T Flock MA Hanson RC Stevens X Deupi DE Gloriam (2019)

Nature Methods 16:151–162.

doi.org/10.1038/s41592-018-0302-x

The retinal conformation and its environment in rhodopsin in light of a new 2.2 A crystal structure

T Okada M Sugihara AN Bondar M Elstner P Entel V Buss (2004)

Journal of Molecular Biology 342:571–583.

doi.org/10.1016/j.jmb.2004.07.044

GPCRdb in 2018: adding GPCR structure models and ligands

G Pándy-Szekeres C Munk TM Tsonkov S Mordalski K Harpsøe AS Hauser AJ Bojarski DE Gloriam (2018)

Nucleic Acids Research 46:D440–D446.

doi.org/10.1093/nar/gkx1109

Phosphorylation-Independent desensitization of G Protein-Coupled receptors?

CS Pao JL Benovic (2002)

Science Signaling 2002:pe42.

doi.org/10.1126/stke.2002.153.pe42

UCSF chimera--a visualization system for exploratory research and analysis

EF Pettersen TD Goddard CC Huang GS Couch DM Greenblatt EC Meng TE Ferrin (2004)

Journal of Computational Chemistry 25:1605–1612.

doi.org/10.1002/jcc.20084

G protein-coupled receptor kinases

JA Pitcher NJ Freedman RJ Lefkowitz (1998)

Annual Review of Biochemistry 67:653–692.

doi.org/10.1146/annurev.biochem.67.1.653

Crystal structure of the β 2 adrenergic receptor-Gs protein complex

SG Rasmussen BT DeVree Y Zou AC Kruse KY Chung TS Kobilka FS Thian PS Chae E Pardon D Calinski JM Mathiesen ST Shah JA Lyons M Caffrey SH Gellman J Steyaert G Skiniotis WI Weis RK Sunahara BK Kobilka (2011)

Nature 477:549–555.

doi.org/10.1038/nature10361

The G protein-coupled receptor kinase (GRK) interactome: role of GRKs in GPCR regulation and signaling

C Ribas P Penela C Murga A Salcedo C García-Hoz M Jurado-Pueyo I Aymerich F Mayor (2007)

Biochimica Et Biophysica Acta (BBA) - Biomembranes 1768:913–922.

doi.org/10.1016/j.bbamem.2006.09.019

CTFFIND4: fast and accurate defocus estimation from electron micrographs

A Rohou N Grigorieff (2015)

Journal of Structural Biology 192:216–221.

doi.org/10.1016/j.jsb.2015.08.008

Involvement of intramolecular interactions in the regulation of G protein-coupled receptor kinase 2

S Sarnago R Roca A de Blasi A Valencia F Mayor C Murga (2003)

Molecular Pharmacology 64:629–639.

doi.org/10.1124/mol.64.3.629

Structural mechanism of arrestin activation

P Scheerer ME Sommer (2017)

Current Opinion in Structural Biology 45:160–169.

doi.org/10.1016/j.sbi.2017.05.001

The β-Arrestins: multifunctional regulators of G Protein-coupled receptors

JS Smith S Rajagopal (2016)

Journal of Biological Chemistry 291:8969–8977.

doi.org/10.1074/jbc.R115.713313

The structural basis of agonist-induced activation in constitutively active rhodopsin

J Standfuss PC Edwards A D'Antona M Fransen G Xie DD Oprian GF Schertler (2011)

Nature 471:656–660.

doi.org/10.1038/nature09795

Probing g α i1 protein activation at single-amino acid resolution

D Sun T Flock X Deupi S Maeda M Matkovic S Mendieta D Mayer R Dawson GFX Schertler M Madan Babu DB Veprintsev (2015)

Nature Structural & Molecular Biology 22:686–694.

doi.org/10.1038/nsmb.3070

Snapshot of activated G proteins at the membrane: the Galphaq-GRK2-Gbetagamma complex

VM Tesmer T Kawano A Shankaranarayanan T Kozasa JJ Tesmer (2005)

Science 310:1686–1690.

doi.org/10.1126/science.1118890

Flexible fitting of atomic structures into electron microscopy maps using

molecular dynamics

LG Trabuco E Villa K Mitra J Frank K Schulten (2008)

Structure 16:673–683.

doi.org/10.1016/j.str.2008.03.005

Crystal structure of rhodopsin in complex with a mini-Go sheds light on the principles of G protein selectivity

CJ Tsai F Pamula R Nehmé J Mühle T Weinert T Flock P Nogly PC Edwards B Carpenter T Gruhl P Ma X Deupi J Standfuss CG Tate GFX Schertler (2018)

Science Advances 4:eaat7052.

doi.org/10.1126/sciadv.aat7052

Structured and disordered facets of the GPCR fold

AJ Venkatakrishnan T Flock DE Prado ME Oates J Gough M Madan Babu (2014)

Current Opinion in Structural Biology 27:129–137.

doi.org/10.1016/j.sbi.2014.08.002

Receptor docking sites for G-protein betagamma subunits. Implications for signal regulation

G Wu JL Benovic JD Hildebrandt SM Lanier (1998)

The Journal of Biological Chemistry 273:7197–7200.

doi.org/10.1074/jbc.273.13.7197

MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy

SQ Zheng E Palovcak JP Armache KA Verba Y Cheng DA Agard (2017)

Nature Methods 14:331–332.

doi.org/10.1038/nmeth.4193

Identification of phosphorylation codes for arrestin recruitment by G Protein-Coupled receptors

XE Zhou Y He PW de Waal X Gao Y Kang N Van Eps Y Yin K Pal D Goswami TA White A Barty NR Latorraca HN Chapman WL Hubbell RO Dror RC Stevens V Cherezov VV Gurevich PR Griffin OP Ernst K Melcher HE Xu (2017)

Cell 170:457–469.

doi.org/10.1016/j.cell.2017.07.002

New tools for automated high-resolution cryo-EM structure determination in RELION-3

J Zivanov T Nakane BO Forsberg D Kimanius WJ Hagen E Lindahl SH Scheres (2018)

eLife 7:e42166.

doi.org/10.7554/eLife.42166

Chapter III

3. Structural characterization of the ExbBD complex from *Serratia marcescens*

ExbB and ExbD belong to a highly adaptable protein complex which dissipates proton motive force at the Gram-negative periplasmic membrane and mechanically transmits the gathered energy to a set of proteins in the outer membrane. Recently a handful of structures of *E. coli* ExbB alone and in complex with ExbD were published which elucidate various features of this complex. In our study we found ExbB pentamers only, independently of the presence of ExbD, which contrasts with previous observations on which the oligomer could adopt hexameric or pentameric forms. Additionally, we describe a new arrangement of the ExbBD complex where 5 units of membrane bound ExbB form a cradle to accommodate 2 units of ExbD at the central pore.

Contribution: cryo-EM dataset processing; contributed to manuscript writing and editing.

The following section is adapted from a manuscript in preparation

Running Title:

Functional and structural characterization of a new class of ExbB proteins with a long N-terminal periplasmic extension

Valérie Biou[1]; Ricardo Adaixo[2]; Mohamed Chami[2]; Benoist Laurent[3]; Nadia Izadi[4]; Florence Cordier[4]; Christian Malosse; Julia Chamot-Rooke[5]; Pierre-Damien Coureux[6]; Henning Stahlberg[2]; Philippe Delepelaire[1]

[1] CNRS Université Paris, Institut de Biologie Physico-Chimique, Paris, France; [2] Center for Cellular Imaging and NanAnalytics (C-CINA), Biozentrum, University of Basel, Basel, Switzerland; [3] Bioinformatics platform, Institut de Biologie Physico-Chimique Paris, France; [4] Bioinformatique Structurale/RMN Institut Pasteur, Paris, France; [5] Structural Mass Spectrometry and Proteomics Unit, Institut Pasteur, Paris, France; [6] Laboratoire de Biochimie, Ecole polytechnique, Université, Paris-Saclay, Palaiseau, France

3.1. Introduction

Transport of nutrients across the Gram-negative outer membrane is either a diffusion-facilitated or an active process, powered in that last case by energy of the proton-motive force (PMF) transmitted to specialized outer membrane receptors (the so called TBDT TonB dependent transporters) by a complex of three cytoplasmic membrane proteins, TonB, ExbB and ExbD, that together form the TonB complex (see [1] for a review). In *Escherichia coli* K12, there is only one set of *tonB, exbB* and *exbD* genes, whereas there are eight TBDT, all energized by the same complex. ExbB and ExbD respectively belong to the MotA- TolQ-ExbB and MotB –TolR-ExbD protein families, that are involved in power generation and transmission in various processes (flagellar rotation for MotAB, active transport of molecules across OM receptors for ExbBD and cell division for TolQR), at the expense of dissipation of cytoplasmic membrane PMF[2-5]. In *E. coli*, some cross-complementation has been observed between ExbBD and TolQR[3]. Those proteins form complexes that associate with either TonB (ExbBD) or TolA (TolQR). Besides power generation, ExbB stabilizes ExbD and the ExbBD complex also stabilizes TonB[6]. TonB has one TM segment with the N-terminus inside the cytoplasm, a proline-lysine rich domain that extends into the periplasm, and a C-terminal structured domain that interacts with the so-called TonB box in the N-terminal part of the « plug » of the receptors[7]. ExbB is a cytoplasmic membrane protein with three TM helices[8,9], its N-terminus resides inside the periplasm and its C-terminus in the cytoplasm. ExbB constitutes a « scaffold » onto which ExbD and TonB will associate to fulfil their function. ExbD has one TM segment at the N-terminus with the N-terminus in the cytoplasm, and a C-terminal structured domain in the periplasm[10]. ExbD TM has one essential strictly conserved aspartate residue[11], that is supposed to undergo cycles of protonation/deprotonation that are coupled to PMF dissipation and to conformational changes in the TonB/TonB box-plug domain that ultimately lead to entry of the substrate initially bound to the

extracellular side of the receptor to the periplasmic space. It is also known that the C-terminal domains of TonB and ExbD interact with each other[12,13].

In *Serratia marcescens*, a close relative of *E. coli*, in which there are about twenty potential TBDT's, there are at least two TonB homologs, an ortholog of $TonB_{Ec}$ (having 48% identity with $TonB_{Ec}$[14]) and HasB, a TonB paralog that is strictly dedicated to its cognate outer membrane receptor HasR[15,16]. The Has system comprises the TBDT HasR, which recognizes both free heme and the HasA hemophore. HasA is a small protein secreted by a type I secretion system, and it has a very high affinity for both heme (Kd around 10pM) and the HasR receptor (apparent Kd of a few nM). Upon interaction of heme-HasA with HasR the heme molecule is first transferred from its HasA binding site to its HasR binding site. Heme is subsequently taken up inside the periplasm, and the HasA molecule is released in the extracellular medium for further rounds of heme scavenging. Both events (heme internalization and HasA release) require the action of the TonB/HasB complex and the use of PMF. This system has been reconstituted in a *E. coli* K12 heme auxotroph, in which it allows heme entry powered by the resident TonB complex[17]. HasB, which is encoded in the *has* locus, has been previously been studied in detail, and has only low sequence identity with either *E. coli* or *S. marcescens* TonB. Whereas $TonB_{Sm}$ complements $TonB_{Ec}$ functions, this is not the case for HasB, that is unable to drive heme entry via the HasR receptor in the presence of $ExbBD_{Ec}$. A gain of function mutant of HasB was isolated in *E. coli*, leading HasB to allow heme entry via HasR in the presence of $ExbBD_{Ec}$[15]. HasB has the same topology as TonB, and its C-terminal domain interacts specifically with HasR with a ca. 40-fold higher affinity than the corresponding TonB domain[18].

Those observations prompted us to identify the $ExbBD_{Sm}$ partners from *S. marcescens* that would function with HasB. We identified the resident orthologous $ExbBD_{Sm}$ as able to function with HasB and TonB, and showed that this specificity for HasB arises from a N-terminal periplasmic extension of $ExbB_{Sm}$, absent from $ExbB_{Ec}$. Recently several structures of *E. coli* ExbB and

ExbBD have been solved, either by X-ray crystallography or by cryo-EM[19,20]. In the X-ray crystallography study[19], $ExbB_{Ec}$ appeared as a pentamer, each monomer of ExbB having three TM helices that extend into the cytoplasm, and with a characteristic charge distribution on the cytoplasmic part (figure S3.5). The central pore is apolar, lined by TM helices 2 and 3 of each monomer, creating a large hydrophobic cavity inside the structure. In the ExbBD structure (PDB 5SV1), one TM helix from ExbD was seen inside the pore of ExbB, and, although double electron-electron resonance (DEER) experiments showed that two ExbD molecules were present for each ExbB pentamer, the TM of the other ExbD monomer was not identified in the electron density map. Another study using X-ray crystallography, single-particle cryo-EM and electron diffraction on two-dimensional crystals concluded that ExbB could exist as a pentamer or an hexamer, that the pentameric-hexameric transition was pH dependent, and that the hexameric ExbB could accommodate three ExbD TM segments in its pore[20]. We characterized $ExbB_{Sm}$ and $ExbBD_{Sm}$ by single-particle cryo-EM and determined their structures at 3,2 and 4 Å resolution respectively. We show that in both cases, $ExbB_{Sm}$ behaves as a stable pentamer; in the ExbBDSm complex, we identify two TM helices of $ExbD_{Sm}$ in the pore of ExbB pentamer. We further show that the N-terminal periplasmic extension specific to ExbBSm is able to interact with the C-terminal domain of HasB.

3.2. Results

3.2.1. Identification of $ExbBD_{Sm}$: $ExbBD_{Sm}$ complement $ExbBD_{Ec}$

Sequence analysis of strain Db11, a fully sequenced *S. marcescens* isolate (GenBank: Serratia marcescens subsp. marcescens Db11, complete genome. ACCESSION HG326223, https://www.ncbi.nlm.nih.gov/nuccore/HG326223.1), indicated that there is one putative operon encoding $ExbBD_{Ec}$ orthologs with high sequence identity with *E. coli* (SMDB11_3479(ExbD) and SMDB11_3480(ExbB)). Another locus enclosed five genes, *exbD-tonB-fecA-exbB-cls* like genes and shared low sequence identity with *E. coli* ExbBD. In this

former case the *exbB* like gene is close to and in the opposite orientation to a *metC* homolog, as in *E. coli.* There is no conservation on the 3' side. Comparison of the putative coding sequences indicated that, whereas the putative *exbD* gene encoded a 140 residues long protein (71% identity with the 141 residues long *E. coli*), the putative *exbB* gene encoded a 325 residues long protein, much longer than its *E. coli* counterpart (244 residues, 73% identity for the common part).

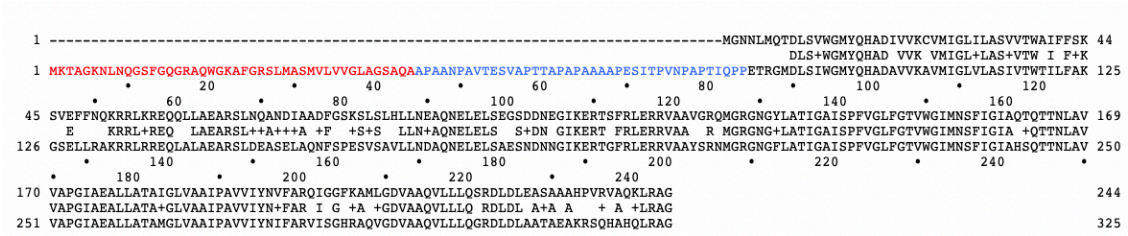

Figure 3.1: alignment of ExbBEc (top) with ExbBSm (bottom), made with DNA Strider v1.4x-2b, CEA. The signal sequence and periplasmic extension sequences of ExbBEc are depicted in red and blue respectively; these are absent in ExbBEc.

Sequence analysis indicated that it contained a putative signal/anchor sequence at the N-terminus followed by a ca. 40 residues long sequence rich in Pro and Ala residues (figure 3.1). The plasmid pexbBD$_{Sm}$/24 was constructed, introducing the two putative genes into pBAD24 under the control of p$_{araBAD}$ promoter and introduced into strain C600*ΔhemAΔtonBΔexbBD*(phasISRADEB/238). This strain could grow in the presence of heme and arabinose, whereas the C600*ΔhemAΔtonBΔexbBD*(phasISRADEB/238 + pexbBD$_{Ec}$/24) could not. The strain C600*ΔhemAΔexbBD*(phasISRADE/238 + pexbBD$_{Ec}$/24) could grow in the presence of heme and arabinose, indicating that pexbBD$_{Ec}$/24 is functional. This indicates that the ExbBD$_{Sm}$ functions functionally associate with HasB to allow HasA mediated heme entry via HasR.

By using growth under iron starvation conditions as a test, we could also show that $ExbBD_{Sm}$ complements $ExbBD_{Ec}$. The *ΔexbBD* strain is more sensitive than its wild-type counterpart to iron starvation (induced for example by dipyridyl) as ExbBD is necessary to transduce PMF to TonB, and TolQR complements only to a very small extent this function. The same holds true for $ExbBD_{Sm}$Delext, indicating that the N-terminal extension is not necessary for complementation.

3.2.2. The N-terminal extension of $ExbB_{Sm}$ is necessary to support HasB mediated heme entry but not TonB mediated heme entry via HasR

Given the presence of this additional domain in $ExbB_{Sm}$, it was quite tempting to speculate that this extension was required to allow function of HasB, and not TonB. We therefore generated a construct where this extension was deleted (see MM for details), pexbBD$_{Sm}$Δext/24. This construct allowed growth of a C600*ΔexbBD* under iron starvation conditions, indicating that this extension is not required for complementation of $ExbBD_{Ec}$. It also allowed growth of C600*ΔhemAΔexbBD*(phasISRADE/238 + pexbBD$_{Sm}$Δext/24) in the presence of heme and arabinose, indicating that it complements $ExbBD_{Ec}$ for HasA mediated heme entry with TonB. However this plasmid was not able to promote growth of the C600*ΔhemAΔtonBΔexbBD*(phasISRADEB/238 + pexbBD$_{Sm}$Δext/24) strain. This likely indicates that the N-terminal extension is strictly required to allow HasB functionality.

Since the putative $ExbB_{Sm}$ gene product included an additional TM segment at its N-terminus as well as a ca. 50-residue long segment potentially residing inside the periplasmic space, we decided to purify the ExbBD complex from *S. marcescens* expressed in *E. coli* to decipher its sequence. The N-terminus of the purified $ExbB_{Sm}$ was APAAN, indicating that the first TM segment was actually a signal sequence and that the cleavage site recognized by the leader peptidase was the canonical one AQA. Furthermore, mass spec data on the denatured

proteins of the complex (supplementary figure S3.1) indicated a mass of 29557 Da for ExbB (theoretical mass 29557 Da) and 16131 for ExbD (theoretical mass 16262 Da), indicating that the initial methionine had been eliminated, consistent with the Ala residue in second position.

3.2.3. « Long » ExbB's are found in several bacterial genera

A BLAST search on bacterial genomes indicated that such long ExbB's are found in different gammaproteobacteria, including Serratia, Yersinia, Pseudomonas, Erwinia/Dickeya, and many genera in the alpha-proteobacteria. Table I in supplementary information lists representative species of alphaproteobacteria where such long ExbB are found.

An alignment (Clustal omega https://www.ebi.ac.uk/Tools/msa/clustalo/, supplementary figure S3.2) also shows that the N-terminal extension, is rich in Ala (24,6%) and Pro (14,5%) residues; it is therefore likely to be unstructured.

3.2.4. Structure of $ExbB_{Sm}$ determined by single particle cryo-EM

Despite the high sequence identity in the C-terminal parts $ExbB_{Sm}$ presented an additional N-terminal sequence relatively to $ExbB_{Ec}$, we therefore aimed at determining its structure. First attempts to get crystals were unsuccessful and we turned to cryo-EM to obtain structural data of $ExbB_{Sm}$. Reconstructions were performed imposing C1 or C5 symmetry and the resulting maps were very similar with a correlation coefficient of 0.98 as calculated in UCSF Chimera. Therefore, we decided to proceed with refinements imposing C5 symmetry. Postprocessing of the obtained map indicated a 3.1Å resolution using the gold standard criterion (figure S3.6).

The resulting map depicts an homo pentamer where side chains are evident and was consequently used to build an atomic model (figure 3.2, left).

The overall structure of ExbBSm monomer is very similar to those of the published structures of ExbB from *E. coli* (Cα RMSD of ca 1Å; figure 3.2, right), as is the pentameric assembly. The periplasmic N-terminal extension did not

yield measurable electron density, precluding its structure determination, and strengthening the hypothesis of an unfolded state. At the C-terminus of each monomer, however, electron density was clearly defined on the cytoplasmic side, allowing model building to the C-terminal residue (Gly 283, see Figure 3.2, right) in the shape of an alpha helical structure [8]. The function of this C-terminal region is not known at the present time. In the 5SV0 X-ray structure, a calcium ion is coordinated by the five Glu 106 of TM2 on the cytoplasmic side. In $ExbB_{Sm}$, this residue is replaced by an Asn, and we do not observe density that could be assigned to a calcium ion. It is however possible that at the C-terminus of $ExbB_{Sm}$, His 276 (Val 239 in $ExbB_{Ec}$) constitutes a potential binding residue for a metal ion.

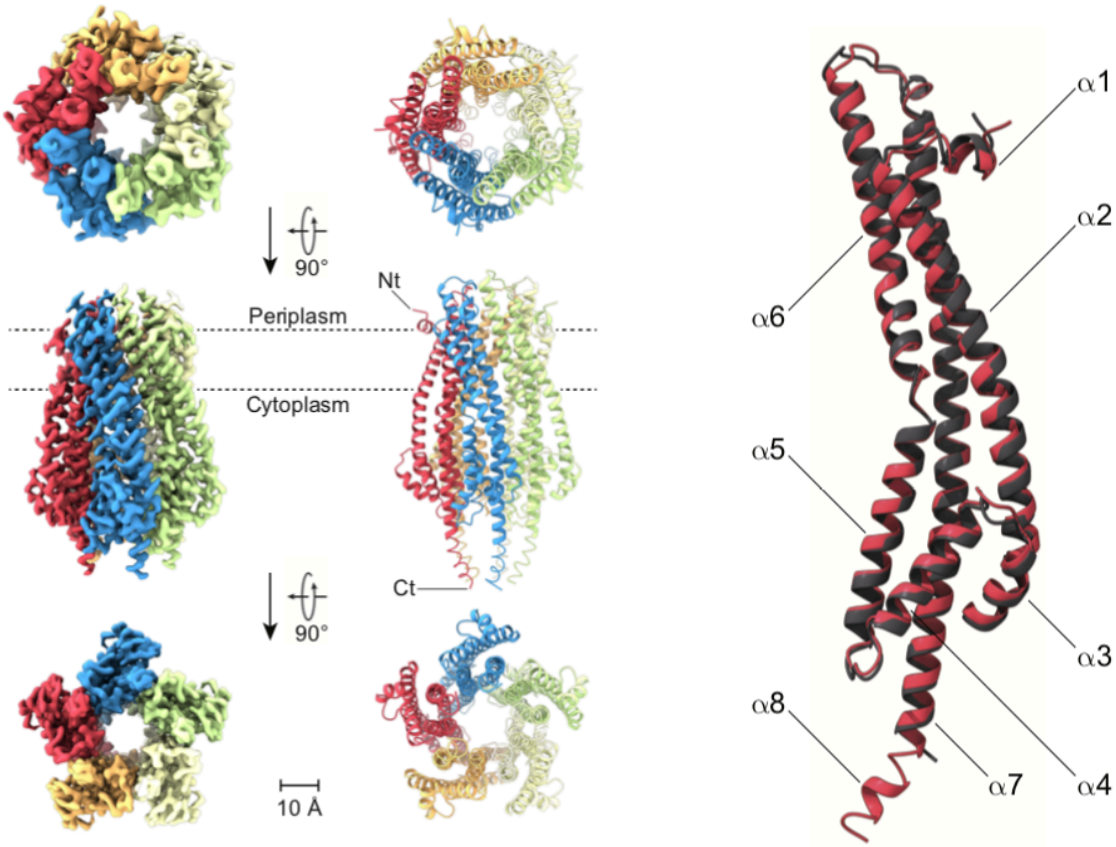

Figure 3.2: Cryo-EM map and atomic model of $ExbB_{Sm}$. (Left) Side by side depiction of the $ExbB_{Sm}$ density map and the corresponding atomic

model in ribbon representation; each $ExbB_{Sm}$ monomer is coloured differently. (Right) Superposition of the $ExbB_{Sm}$ and $ExbB_{Ec}$ monomers in ribbon representation.

Residues lining the inner face of the ExbB pore, within the membrane boundaries, are strongly conserved, indicative of high functional constraints as already mentioned by Celia et al (figure 3.3).

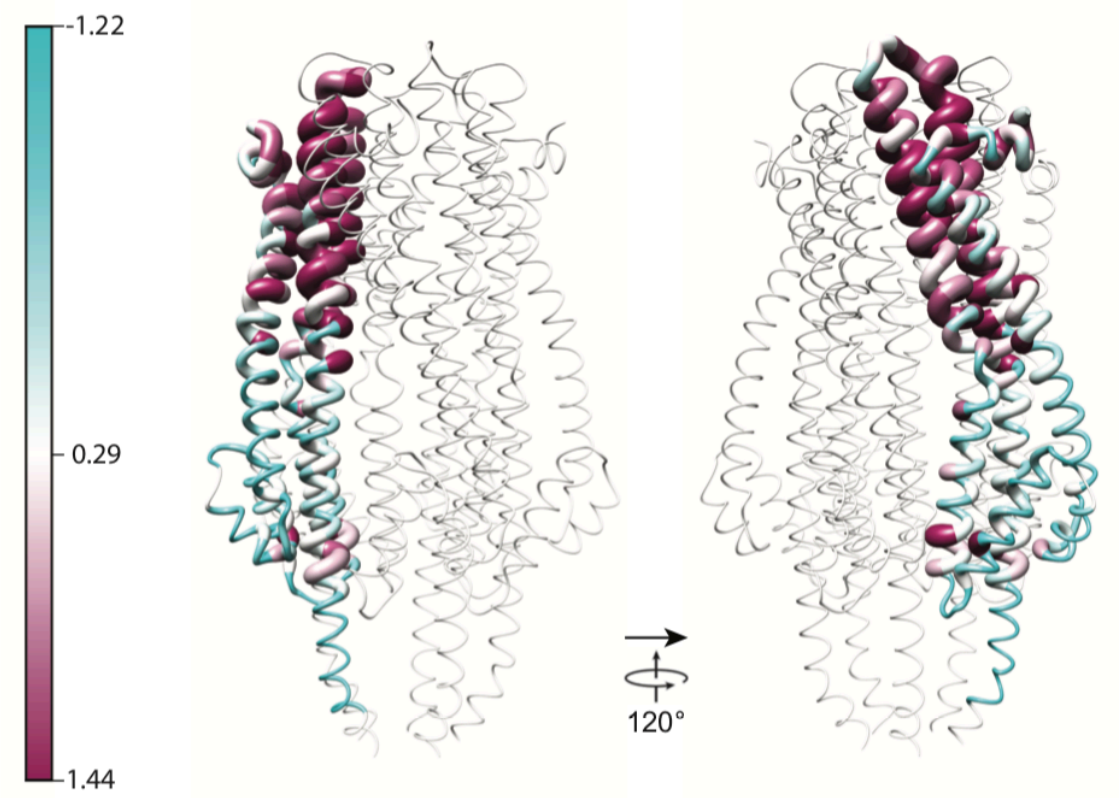

Figure 3.3: Residue conservation in $ExbB_{Sm}$. Conservation plot was obtained from the Consurf server using the 146 sequences mentioned before. Conservation score colour code on the left.

Celia et al. also mentioned the existence of « vents » or « fenestrations » in the structure potentially leading to solvent or ion passage[19]. These vents are located at the junction of adjacent monomers and are also observed in the $ExbB_{Sm}$ structure; however an additional density was clearly observed on the external

side of those vents, with four elongated densities joined by a common head. Considering the lipid composition of bacterial membrane, we built a cardiolipin in this density (figure 3.4). This density is observed for each monomer, whether or not we enforce the C5 symmetry.

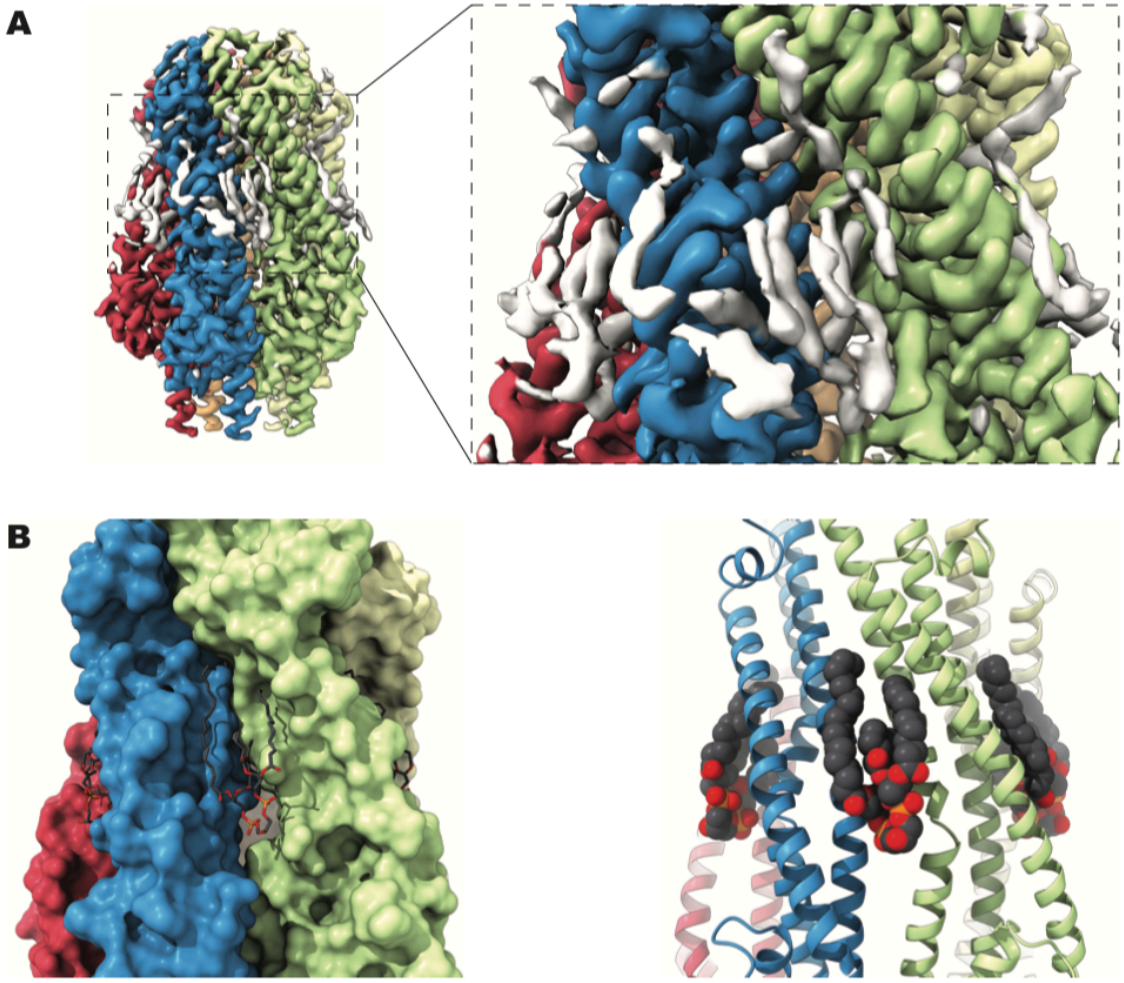

Figure 3.4: Lipid like molecules line the $ExbB_{Sm}$ fenestrations. (A) Cryo-EM map of $ExbB_{Sm}$ in different colours together with cardiolipin in white, the dashed box area was zoomed in to obtain the image on the right; (B) Surface representation of the $ExbB_{Sm}$ in different colours together with cardiolipin depicted in sticks and coloured by atom type. The right image is analogous to the left image but protein is represented as ribbons and cardiolipin as spheres.

Our structure was solved at pH 8, and the obtained 2D and 3D classes were consistent with pentameric structures (we did not observe hexameric forms). This is in contrast with the Maki-Yonekura *et al* study, where ExbB was predominantly observed as an hexameric species in the high pH regime[20]. In the mentioned study protein was solubilised in DDM which was later exchanged for either C8E4 or C10E5. These shorter chain detergents might have led to some destabilization of the pentameric structure and be at the basis of the conversion to the hexameric form. Our preliminary cryo-EM data of $ExbB_{Sm}$ in DDM at two pH values is again consistent with a single pentameric form of $ExbB_{Sm}$ (data not shown).

3.2.5. Structure of $ExbBD_{Sm}$ determined by single particle cryo-EM

We also solved the structure of the $ExbBD_{Sm}$ complex by cryo-EM (figure 3.5). As for ExbB, the complex exhibited a symmetrical peak on SEC. The $ExbBD_{Sm}$ density map has an estimated resolution of 4 Å (figure S3.4) precluding reliable modelling of side chains. At the present moment a tentative preliminary atomic model was built to allow comparison with the remaining ExbB and ExbD structures (figures 3.6 and 3.7).

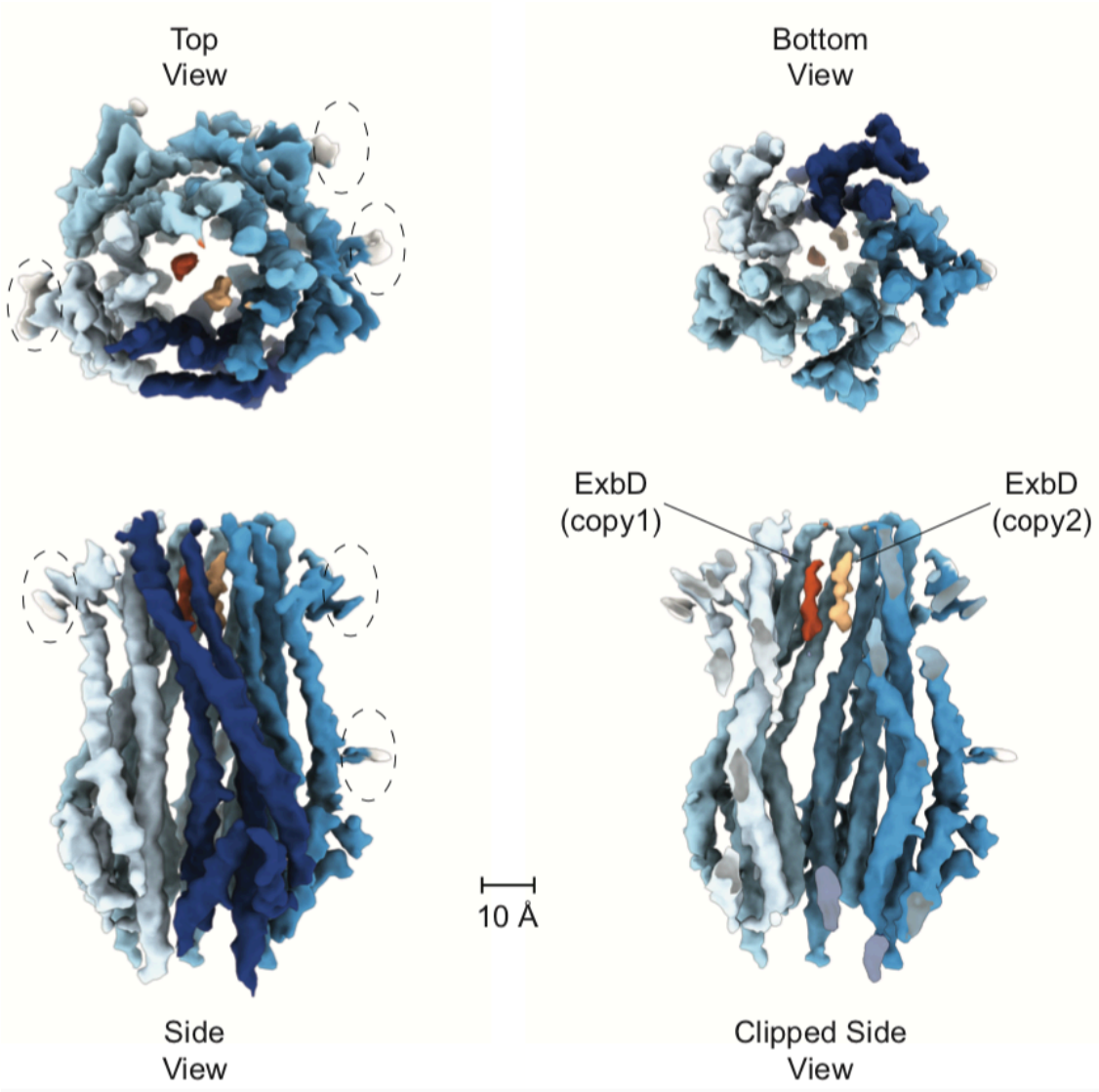

Figure 3.5: Cryo-EM map of the $ExbBD_{Sm}$ complex. Each $ExbB_{Sm}$ monomer is coloured in a different shade of blue while each $ExbD_{Sm}$ helix is coloured in a different shade of orange. Densities inside dashed ovals are artefacts from imperfect signal subtraction of the detergent micelle signal.

Despite the resolution limitations, several distinctive features are worth mentioning: first, similar to the apo form $ExbB_{Sm}$ is as a pentamer in the $ExbBD_{Sm}$ complex; second, we observed a slight « opening » at the periplasmic side of the pore (distance Ala 197 of one subunit-Leu 204 of the facing subunit

varying from 25,5 in $ExbB_{Sm}$ to 29,8Å in $ExbBD_{Sm}$); third, there are two clear densities inside the ExbB pore that we assigned to a pair of ExbD TM segments (figure 3.5).

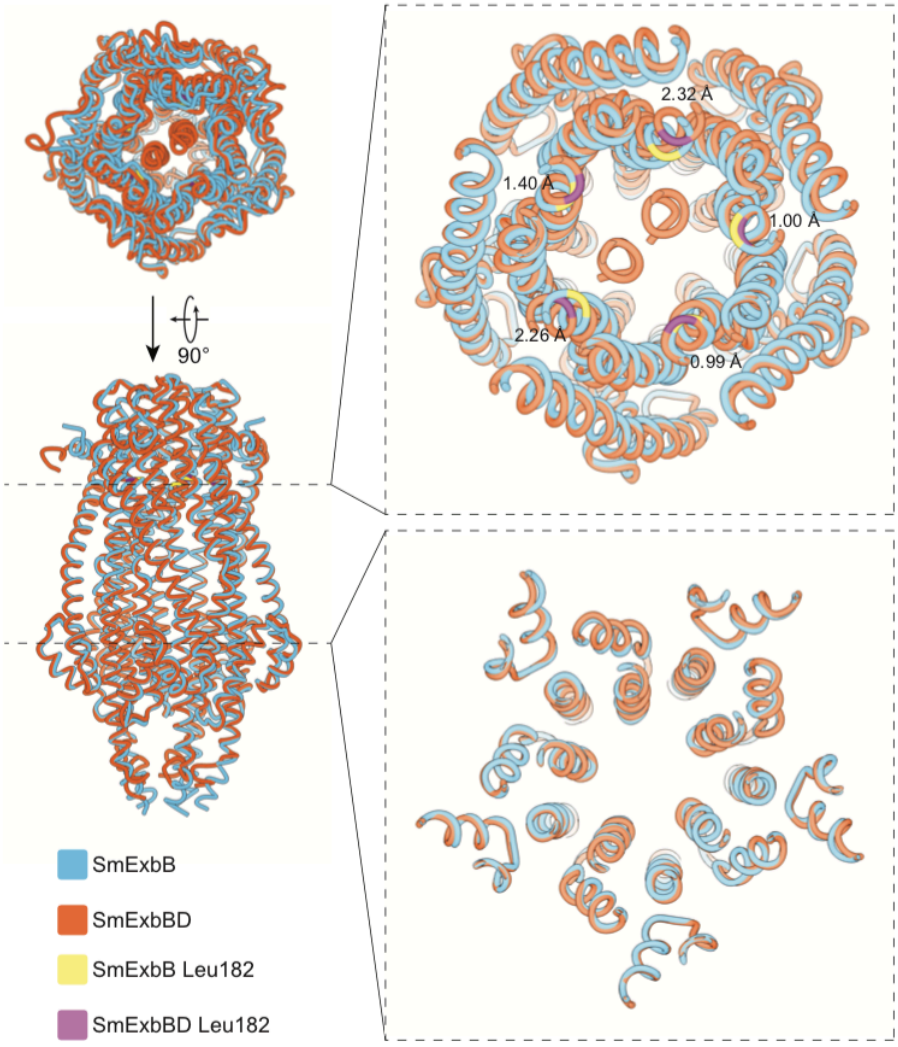

Figure 3.6: Superposition of $ExbB_{Sm}$ and $ExbBD_{Sm}$. Colour code at the bottom left of the figure. Leucine 182 backbone was coloured yellow in the case of $ExbB_{Sm}$ and purple in the case of $ExbBD_{Sm}$ and the carbon alpha distance between the pairs determined and presented (top right).

Even though amino acid side chains are not visible in the density, we can describe the position of $ExbD_{Sm}$ relative to $ExbB_{Sm}$. The first $ExbD_{Sm}$ monomer is close

to the 6 helix of ExbB monomer E, and the second one is nested between helix 6 and 7 of monomer B and helix 6 of monomer C. This indicates two different interactions, that may allow a rotation of the ExbD helix with respect to $ExbB_{Sm}$, involving a set of apolar residues. The position of the ExbD TM helices in the Sm structure also differs from the Ec structure (pdbid 5SV1); while in Ec ExbD occupies the center of the pore (coincident with the pseudo 5-fold axis) in Sm the 2 ExbD TMs are disposed radially to the pseudo 5-fold axis. In the Ec case, binding of a single ExbD TM helix causes no or very little perturbation of the ExbB structure, however, in the Sm complex such perturbations of the ExbB pentamer do occur at the periplasmic end of the pore as mentioned above.

It is important to note that the observed changes on the $ExbB_{Sm}$ structure upon $ExbD_{Sm}$ binding are localised and the carbon alpha rmsd between apo and holo forms for Ec and Sm is bellow 1 Å in both structures; CαRMSD $ExbBD_{Ec}/ExbB_{Ec}$: 0,46Å, CαRMSD $ExbBD_{Sm}/ExbB_{Sm}$: 0,93Å).

Again, as in the ExbB case, we saw no indication of hexameric state of ExbB in those particles.

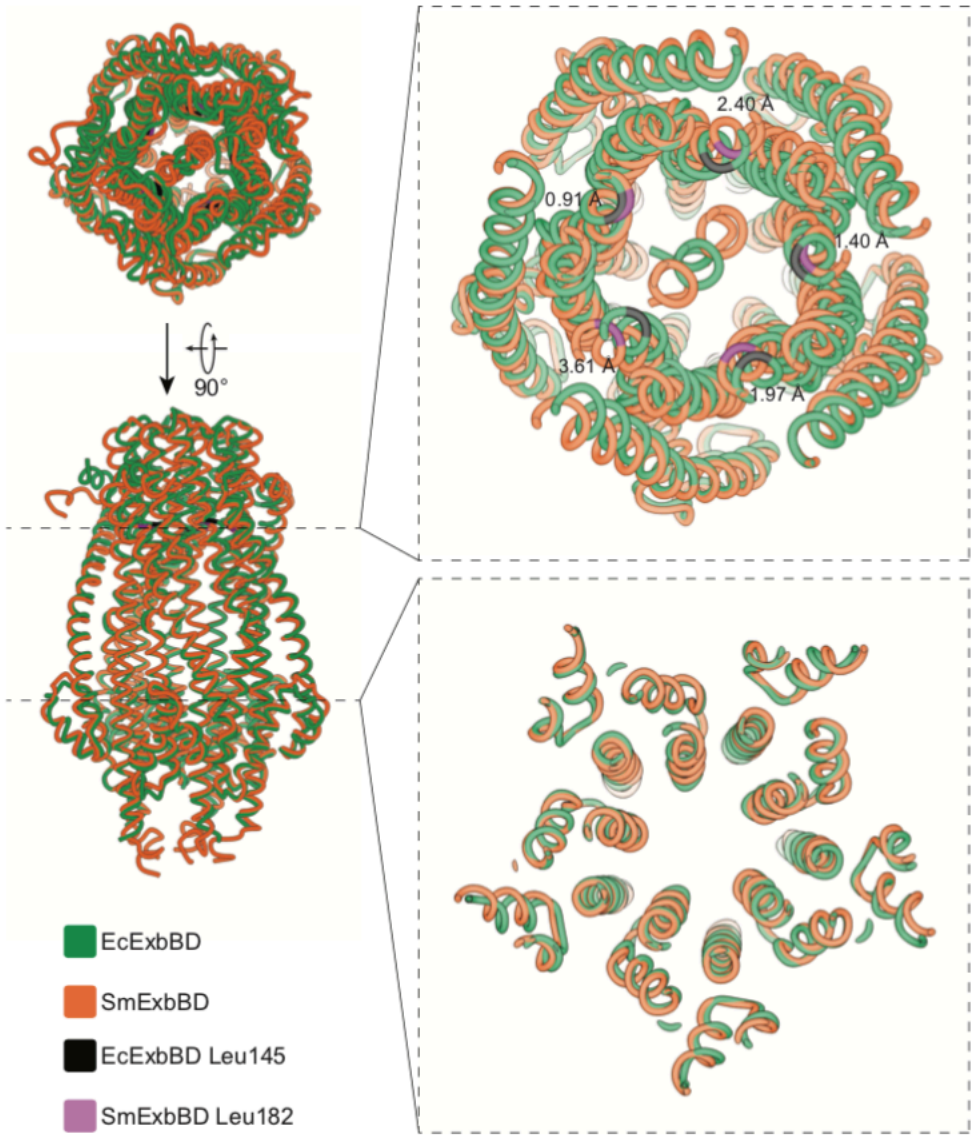

Figure 3.7 : Superposition of $ExbBD_{Sm}$ and $ExbBD_{Ec}$. Colour code at the bottom left of the figure. Leucine 182 backbone was coloured yellow in the case of $ExbB_{Sm}$ and purple in the case of $ExbBD_{Sm}$ and the carbon alpha distance between the pairs determined and presented (top right).

3.3. Conclusions

Our study of two essential components of the heme import system of Serratia marcescens ExbB and ExbD disclosed important information that was not clear in the previous work on E. coli proteins. The complex is formed by an ExbB pentamer that incorporates two ExbD TM monomers in the channel. A slight deformation of the membrane area accompanies ExbD insertion. This arrangement is in agreement with the DEER measurements published in Celia et al that produced two ExbD monomers per complex, despite the crystal structure showing a single TM helix in the ExbB pentamer. The ExbB density exhibits five lipid-like densities that are located in the inner cytoplasmic leaflet of the membrane. Based on their shape we interpret these densities as cardiolipin molecules where the phosphates of the polar head interact with basic residues, while the aliphatic chains interact with hydrophobic residues. However, our structures do not give information about the periplasmic region. The periplasmic domain of ExbD is not present in the density, probably due to a disordered 20 aminoacid linker that allows various orientations of the periplasmic domain with respect to the main protein body. In addition, $ExbB_{Sm}$ has a 40-aminoacid extension that is not visible in the structure. This region is very rich in alanine and proline residues and it is predicted as disordered. SAXS measurements are currently being analyzed and will enable modelling the volume occupied by the periplasmic residues.

In summary, our two structures display new features of the TonB complex, but the precise relative orientation of ExbD with respect to ExbB is not visible due to a limitation in resolution. Improvement of the $ExbBD_{Sm}$ map resolution is required to hypothesize over the molecular mechanism of the complex.

3.4. Materials and Methods

3.4.1. Strains and plasmid construction

Plasmid pBADExbBD$_{Sm}$ was constructed after amplification on the genomic DNA from strain *S. marcescens* Db11 of a ca. 1,42kb fragment using primers ExbBDSm5' GGAGGAATTCACCATGAAAACGGCTGGCAAGAAT and ExbBDSm3' AAGCTTGCATGCCTATTTGGCGGCGCCTTCCA. The PCR product was purified digested with EcoRI and SphI, and ligated with pBAD24 digested with EcoRI and SphI. Correct clones were selected after sequencing of the insert.

Plasmid pBADExbBD$_{Sm}$His6 was constructed by first amplifying on pBADExbBDSm a ca. 0,4kb fragment with the following oligonucleotides SphIHisCtexbDSm
TTGCATGCCTAATGGTGATGGTGATGGTGTTTGGCGCGCCTTC

and BglIIexbDSm GCGCTTAAATGAAGATCTGGACGACAGCGG; after amplification, the fragment was purified, digested with BglII and SphI, and ligated with pBADExbBDSm digested with the same enzymes. The correct clones were selected after sequencing of the insert. In biological tests, this plasmid was undistinguishable from its parent plasmid pBADExbBDSm.

-Plasmid pBADExbBD$_{Sm}$DelExt was constructed after amplification of a ca.1,2kb fragment with oligonucleotide delextexbBSm GAGGAATTCACCATGCCACCGGAAACCCGCGGCAT and ExbBDSm3' (see above); after purification of the PCR fragment, digestion with EcoRI and SphI, the fragment was ligated into pBAD24 digested with the same enzymes. Correct clones were selected after sequencing. -Plasmid pBADExbBD$_{Ec}$ was constructed by amplification on genomic MG1655 DNA of a ca. 1.7kb fragment with the following oligonucleotides ExbBD5c CAGGAGGAATTCACCATGGGTAATAATTTAATGCAGACGGA, and ExbBD3c AAGCTTGCATGCTTACTTCGCTTTGGCGGTTTCTT ; the PCR product was purified, digested with EcoRI and SphI, and ligated with pBAD24

digested with the same enzymes. Correct clones were selected after sequencing.

-Plasmid pBADExbB$_{Sm}$His6 was constructed as follows a PCR fragment was amplified on pBADExbBDSm with the oligos PBADFOR CTGACGCTTTTTATCGCAAC and ExbBHis6 AAGCTTGCATGCCTAATGGTGATGGTGATGGTGCCCCGCCCGCAGTTG, digested with EcoRI and SphI, and ligated with pBAD24 digested with the same enzymes. Correct clones were selected by sequencing.

All constructions were carried out in *E. coli* strain XL1-Blue.

Strains and other plasmids used in this study are shown in Table 1 and are from the lab collection.

3.4.2. Protein expression and purification

The BL21DE3(pExbBD$_{Sm}$His6/pBAD24) or BL21DE3(pExbB$_{Sm}$His6/pBAD24) were grown at 37°C either in TB or MDM medium, and induced with 0,2% arabinose at either 1,5-2OD600nm (TB) or 5-6OD600nm (MDM) and the incubation continued for 3hrs. The cells were harvested by centrifugation (10000g 20minutes 4°C), washed once in 20mMTris-HCl pH8,0 and kept frozen at -80°C. Cells were broken in a Cell disruptor (Constant, UK) at 1kbar (10g of cells in 40ml final of 20mM Tris pH 8.0 containing protease inhibitor (Roche, EDTA free), at 4°C. Dnase and RNase were added (10 and 50µg/ml). After ca. 15 minutes, the solution was centrifuged for 1hr at 100000g, or alternatively 20000g for 45 minutes at 4°C. The pellet (crude membrane preparation) was resuspended in 20mM Tris pH 8.0 plus protease inhibitor cocktail (Roche EDTA free), flash-frozen in liquid N2 and kept at -80°C.

The crude membrane preparation was solubilized in 20mM Tris pH 8.0, 20mM Imidazole, 100mMNaCl, 10% Glycerol, 0,8% of either DDM or LMNG (10g of equivalent whole cells solubilized in 40ml), plus protease inhibitor cocktail (Roche EDTA-free) for 30 minutes at 15°C. After centrifugation (1hr 100000g or

45minutes 20000g), the supernatant was incubated with 2.5ml Ni-Agarose beads preequilibrated with the same buffer except for the detergent concentration (0.015%DDM or 0.0015%LMNG). After three hours of incubation on a rotating wheel at 4°C, the beads were washed three times with 25ml of pre-equilibration buffer, and then eluted with two times 25ml of the pre-equilibration buffer with 200mM Imidazole. The eluate was concentrated and washed on 100kDa cut-off centrifugal device, in pre-equilibration buffer without NaCl and Imidazole. The resulting sample was loaded on a monoQ HR10-100 column equilibrated with 20mM Tris pH 8.0, 10% glycerol, 0.015%DDM or 0.0015%LMNG and eluted with a gradient from 0 to 1M NaCl in the same buffer. The peak fractions were collected and concentrated and the sample was loaded on a Superose 6 increase column equilibrated with 20mM Tris pH8.0, 100mM NaCl, 0.015%DDM or 0.0015%LMNG. The peaks fractions were collected, concentrated and their concentration determined by using the theoretical absorption coefficient of either ExbB5D2 (113790M-1cm-1), or ExbB5 (104850M-1cm-1). They were then kept frozen at -80°C in aliquots until use.

3.4.3. Cryo-EM grid preparation and data acquisition

3 µL of purified ExbBD complex at 1.2 mg/mL was applied to C-Flat 1.2/1.3 holey carbon grids (Protochips Inc., USA) glow discharged in air for 30s. Grids were blotted for 2s at blot force 1 and vitrified in liquid ethane using a Vitrobot mark IV (FEI company) operated at 10 °C and 100% relative humidity.

All data collection was performed with a Titan Krios (ThermoFisher Scientific) operated at 300 kV equipped with a K2 Summit direct electron detector (Gatan Inc., Pleasanton, CA) at the European synchrotron research facility, ESRF (Grenoble, France). Movies were recorded in electron counting mode with EPU software (ThermoFisher Scientific), aligned with Motioncor2 [22] and aligned images were processed with Gctf [23] using Scipion interface[24].

For ExbB-D data collection, 4043 movies were collected in a Titan at a nominal magnification of 139000 with a pixel size of 1.067Å. Movies of 48 frames were acquired using a dose rate of. 4.6 electrons/Å^2/second over 12 seconds yielding a cumulative exposure of 55.2 electrons/Å^2.

For ExbB data collection, 4567 movies were collected at a magnification of 165000x with a nominal pixel size of 0.827Å using a defocus range of -1.2 to -2.5µm. Movies of 56 frames were acquired using a dose rate of 8 electrons/Å^2/second over 7 seconds, yielding a cumulative exposure of 55.95 electrons/Å^2.

3.4.4. Cryo-EM image processing and analysis

3.4.4.1. ExbBD

Movies were drift-corrected and dose-weighted using MotionCorr2 [22]. Aligned dose weighted averages were imported into Cryosparc2[25] and contrast transfer function was estimated using CTFFIND4.1[26]. Micrographs with poor CTF estimation statistics or high drift profiles were discarded. The remaining 3028 micrographs were used for automated particle picking. Particles were extracted, Fourier cropped to 2 Å/px and 2D classified. The best 2D classes were used as templates for automated particle picking resulting in 1.3 Mio particles. After several rounds of classification, the best 600 k particles were submitted to 3D classification by means of multi class Ab-Initio Reconstruction and Heterogeneous Refinement. 158 k particles belonging to best resolved class were corrected for local motion, re-extracted and used in Non-Uniform Refinement. The resulting refined map has nominal resolution of 4.56 Å.

Based on the previous map a soft mask lining the micelle was designed in UCSF Chimera (Pettersen et al., 2004) and used to signal subtract the corresponding micelle density of particles in the refined map.

Localized refinement of the signal subtracted particles produced a map of the

complex with estimated resolution of 3.96 Å judging by FSC at 0.143 criterion.

3.4.4.1. ExbB

For ExbB, aligned movies were processed with Gctf [23] and only images with a resolution higher than 4Å were kept; after visual inspection of images the remaining processing was carried out with Relion-3 [27]. Particles were extracted using a 2-fold binning, issued from a manual picking and a 2-D classification on 50 images. Automatic extraction was performed using the selected 2D classes. After several rounds of 2D and 3D classification, 161k particles were selected for 3D refinement. They were corrected for local motion using Bayesian polishing option in Relion-3 and a post-refined map produced a 3.1Å overall resolution with a 5-fold symmetry.

3.4.5. BLAST search

$ExbB_{Sm}$ was used as a search for BLAST for orthologs in bacterial genomes. Top hits were in Serratia, Yersinia, Dickeya, Erwinia and Pseudomonas. Those genera were later excluded from successive BLAST searches to obtain orthologs in other genera with higher p-value.

3.5. Supplementary Information

Afipia broomeae	Hartmannibacter diazotrophicus	Phreatobacter stygius
Agrobacterium rhizogenes	Hoeflea olei	Phyllobacterium zundukense
Aliihoeflea sp.	Hyphomicrobium nitrativorans	Proteobacteria bacterium
Aminobacter aminovorans	Insolitispirillum peregrinum	Pseudaminobacter manganicus
Amorphus coralli	Kaistia granuli	Pseudochrobactrum asaccharolyticum
Ancylobacter aquaticus	Ketogulonicigenium robustum	Pseudolabrys taiwanensis
Aquamicrobium aerolatum	Labrys okinawensis	Pseudorhizobium pelagicum
Aurantimonas manganoxydans	Magnetospirillum gryphiswaldense	Pseudorhodoplanes sinuspersici
Aureimonas flava	Mangrovicella endophytica	Reyranella massiliensis
Azorhizobium caulinodans	Mesorhizobium loti	Rhizobium tropici
Azospirillum brasilense	Methylobrevis pamukkalensis	Rhodoligotrophos sp.
bacterium A52C2	Methylocapsa acidiphila	Rhodomicrobium udaipurense
Bartonella apis	Methyloceanibacter stevinii	Rhodopseudomonas pseudopalustris
Beijerinckiaceae bacterium	Methylocella silvestris	Rhodospirillum centenum
Blastochloris sp.	Methyloligella halotolerans	Rhodovarius sp.
Bosea vaviloviae	Methylopila sp.	Roseococcus sp.
Bradyrhizobium yuanmingense	Methylovirgula ligni	Roseospirillum parvum
Brucella sp.	Microvirga aerophila	Sinirhodobacter sp.
Chelatococcus asaccharovorans	Neorhizobium alkalisoli	Sinorhizobium sp.
Ciceribacter lividus	Nitratireductor pacificus	Skermanella aerolata
Devosia psychrophila	Nitrobacter hamburgensis	Sphingobium chlorophenolicum
Dongia mobilis	Nitrospirillum amazonense	Sphingomonas sp.
Enterovirga rhinocerotis	Niveispirillum cyanobacteriorum	Sphingopyxis sp.
Erythrobacter	Ochrobactrum thiophenivorans	Sphingorhabdus sp.
Falsochrobactrum ovis	Paenirhodobacter enshiensis	Starkeya novella
Geminicoccus roseus	Paracoccus aminovorans	Tardiphaga robiniae
Granulibacter bethesdensis	Paramesorhizobium deserti	Telmatospirillum siberiense
Haematobacter massiliensis	Parvibaculum lavamentivorans	Variibacter gotjawalensis
Hansschlegelia zhihuaiae	Phaeospirillum fulvum	Xanthobacter autotrophicus

Table I: list of selected species of alphaproteobacteria where long ExbB are found. All those species contain HasB orthologs, except for the ones in red.

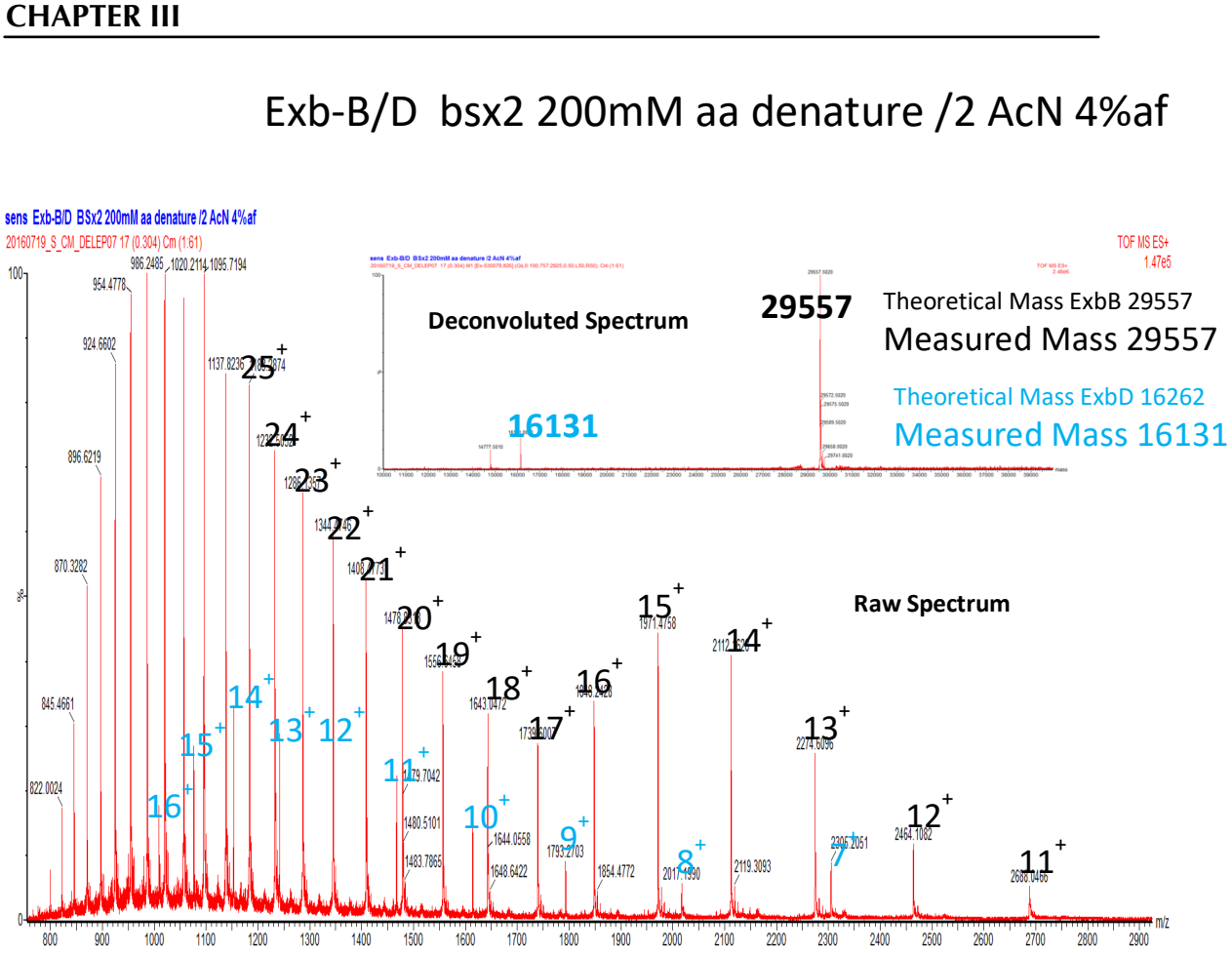

Figure S3.1: Mass spectrometry analysis of the denatured components of the $ExbBD_{Sm}$ complex (m/z) and their deconvolution.

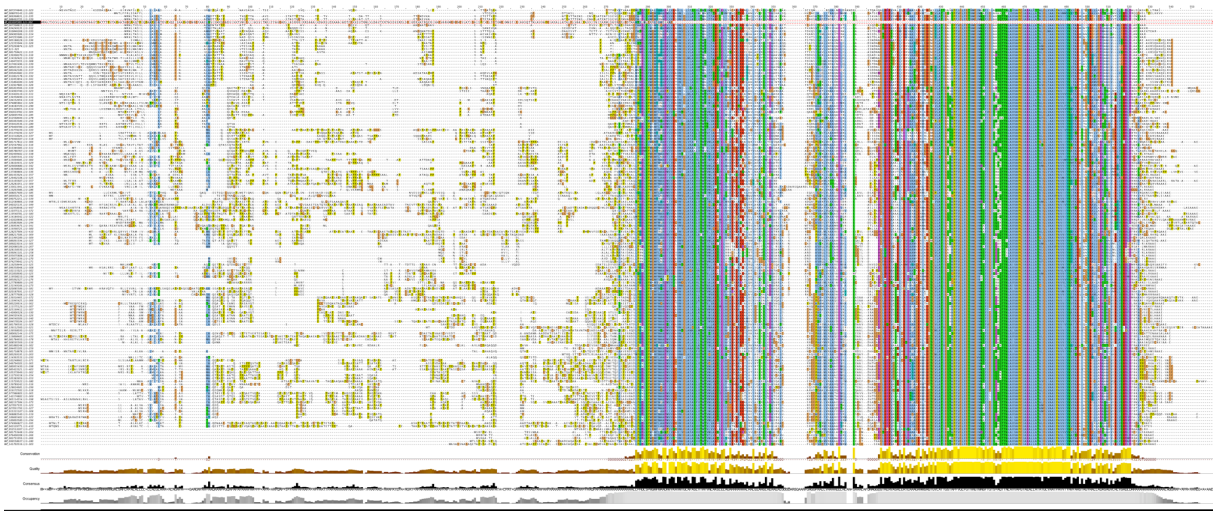

Figure S3.2: Alignment of 146 « long » ExbB's built with JALView[28]. The three very well conserved regions (at the right) correspond to the three TM helices (TM1 being the least conserved), whereas the somewhat less well conserved region corresponds to the cytoplasmic linker between TM1 and TM2.

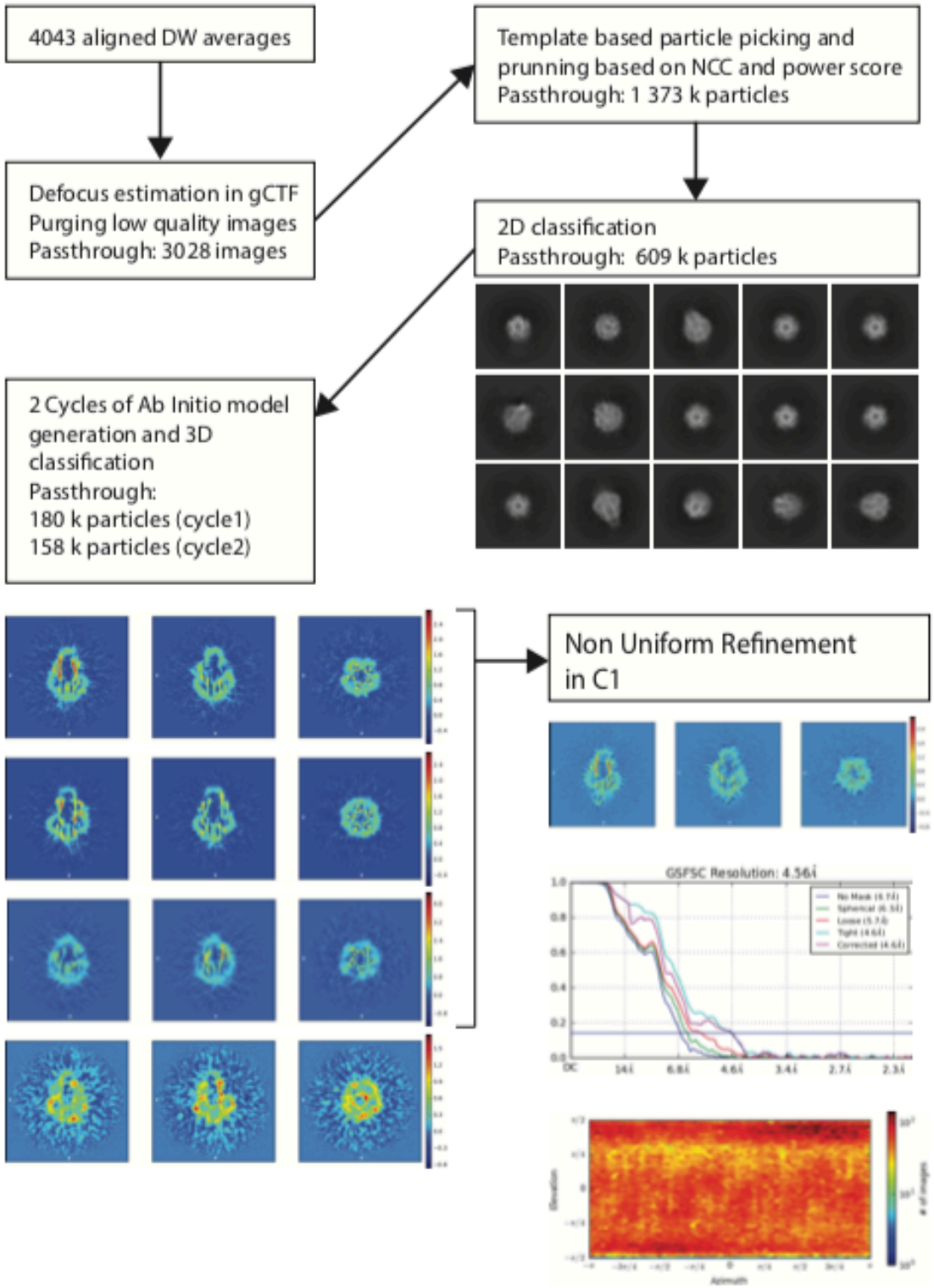

Figure S3.3: processing scheme of the ExbBD$_{Sm}$ dataset before signal subtraction of the detergent micelle.

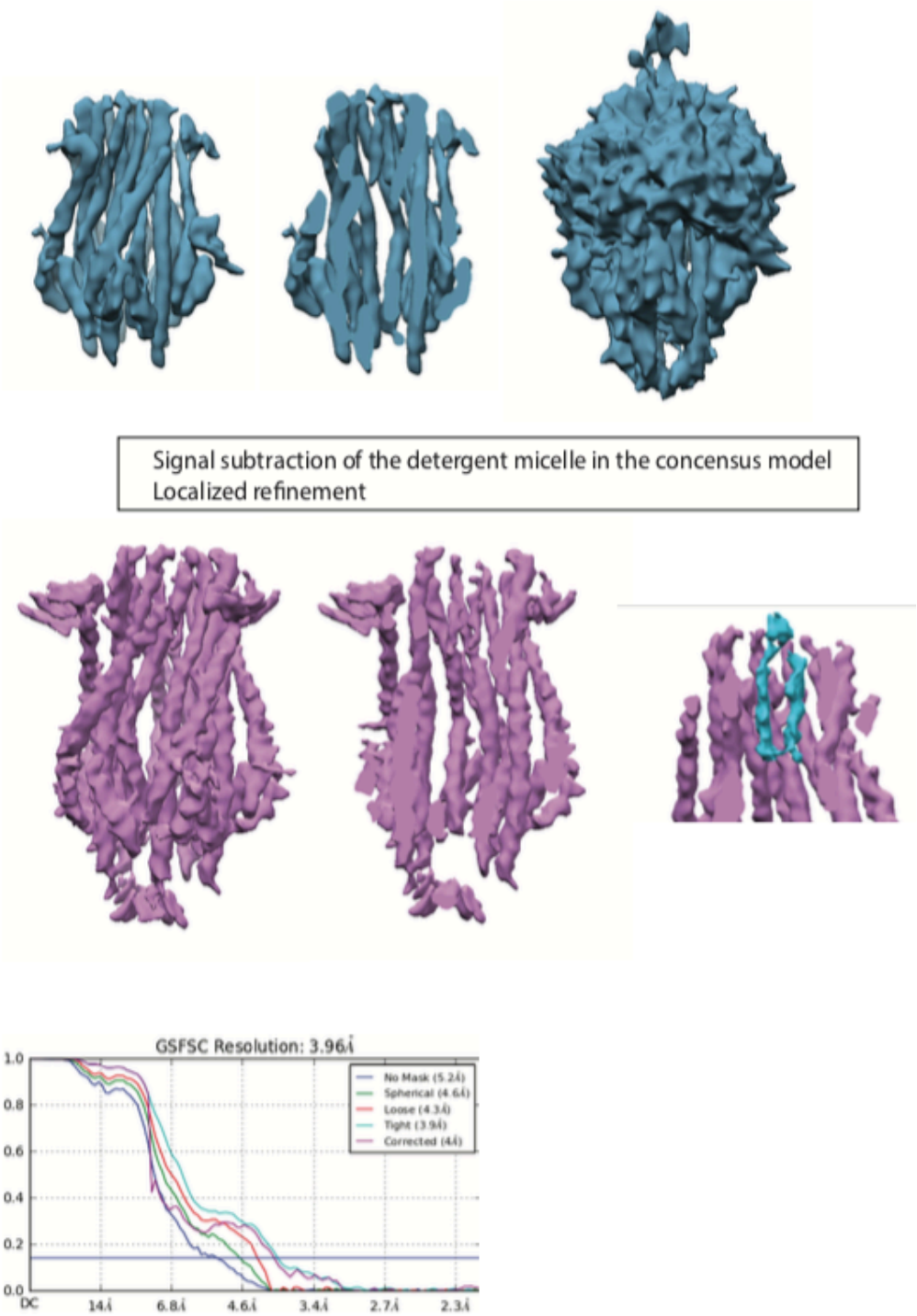

Figure S3.4: (Top) $ExbBD_{Sm}$ density maps before and after signal subtraction of the detergent micelle. (Bottom) FSC plot of the final volume obtained by local refinement.

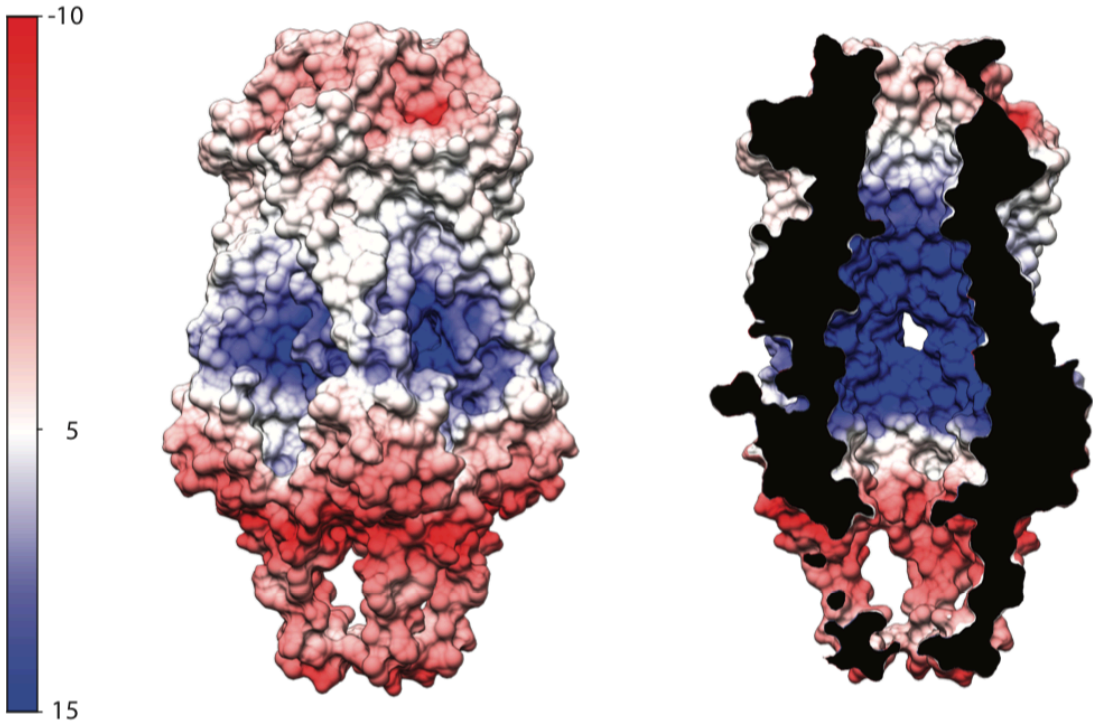

Figure S3.5: surface representation of $ExbB_{Sm}$ atomic model coloured by electrostatic potential (calculated in APBS); a vertical section of the same surface is depicted on the right.

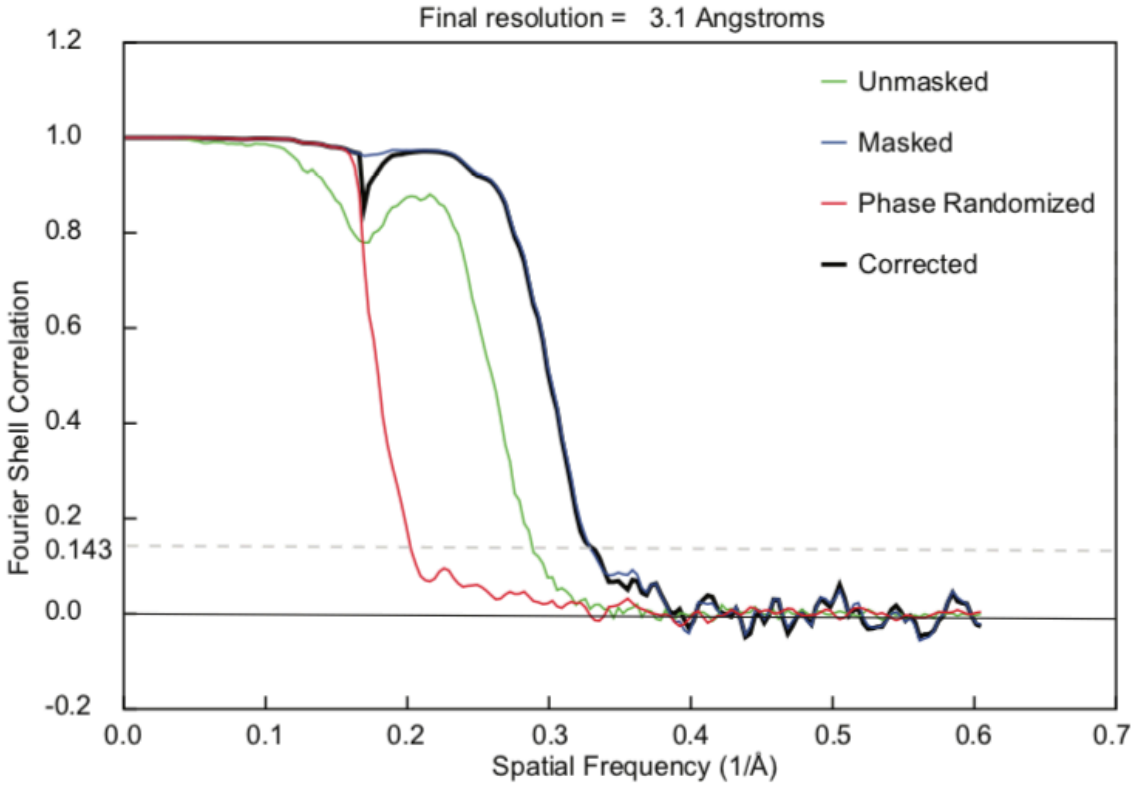

Figure S3.6: FSC plot of the $ExbB_{Sm}$ cryo-EM map.

3.6. Bibliography

1. Noinaj, N., Guillier, M., Barnard, T.J. & Buchanan, S.K. TonB-dependent transporters: regulation, structure, and function. *Annu Rev Microbiol* 64, 43-60 (2010).

2. Cascales, E., Lloubes, R. & Sturgis, J.N. The TolQ-TolR proteins energize TolA and share homologies with the flagellar motor proteins MotA-MotB. *Mol Microbiol* 42, 795-807 (2001).

3. Braun, V. The structurally related exbB and tolQ genes are interchangeable in conferring tonB-dependent colicin, bacteriophage, and albomycin sensitivity. *J Bacteriol* 171, 6387-90 (1989).

4. Eick-Helmerich, K. & Braun, V. Import of biopolymers into Escherichia coli: nucleotide sequences of the exbB and exbD genes are homologous to those of the tolQ and tolR genes, respectively. *J Bacteriol* 171, 5117-26 (1989).

5. Braun, V. & Herrmann, C. Evolutionary relationship of uptake systems for biopolymers in Escherichia coli: cross-complementation between the TonB-ExbB-ExbD and the TolA-TolQ-TolR proteins. *Mol Microbiol* 8, 261-8 (1993).

6. Fischer, E., Gunter, K. & Braun, V. Involvement of ExbB and TonB in transport across the outer membrane of Escherichia coli: phenotypic complementation of exb mutants by overexpressed tonB and physical stabilization of TonB by ExbB. *J Bacteriol* 171, 5127-34 (1989).

7. Roof, S.K., Allard, J.D., Bertrand, K.P. & Postle, K. Analysis of Escherichia coli TonB membrane topology by use of PhoA fusions. *J Bacteriol* 173, 5554-7 (1991).

8. Kampfenkel, K. & Braun, V. Topology of the ExbB protein in the cytoplasmic membrane of Escherichia coli. *J Biol Chem* 268, 6050-7 (1993).

9. Baker, K.R. & Postle, K. Mutations in Escherichia coli ExbB transmembrane domains identify scaffolding and signal transduction functions and exclude participation in a proton pathway. *J Bacteriol* 195, 2898-911 (2013).

10. Garcia-Herrero, A., Peacock, R.S., Howard, S.P. & Vogel, H.J. The solution structure of the periplasmic domain of the TonB system ExbD

protein reveals an unexpected structural homology with siderophore-binding proteins. *Mol Microbiol* 66, 872-89 (2007).

11. Braun, V. et al. Energy-coupled transport across the outer membrane of Escherichia coli: ExbB binds ExbD and TonB in vitro, and leucine 132 in the periplasmic region and aspartate 25 in the transmembrane region are important for ExbD activity. *J Bacteriol* 178, 2836-45 (1996).

12. Ahmer, B.M., Thomas, M.G., Larsen, R.A. & Postle, K. Characterization of the exbBD operon of Escherichia coli and the role of ExbB and ExbD in TonB function and stability. *J Bacteriol* 177, 4742-7 (1995).

13. Higgs, P.I., Myers, P.S. & Postle, K. Interactions in the TonB-dependent energy transduction complex: ExbB and ExbD form homomultimers. *J Bacteriol* 180, 6031-8 (1998).

14. Gaisser, S. & Braun, V. The tonB gene of Serratia marcescens: sequence, activity and partial complementation of Escherichia coli tonB mutants. *Mol Microbiol* 5, 2777-87 (1991).

15. Paquelin, A., Ghigo, J.M., Bertin, S. & Wandersman, C. Characterization of HasB, a Serratia marcescens TonB-like protein specifically involved in the haemophore-dependent haem acquisition system. *Mol Microbiol* 42, 995-1005 (2001).

16. Benevides-Matos, N., Wandersman, C. & Biville, F. HasB, the Serratia marcescens TonB paralog, is specific to HasR. *J Bacteriol* 190, 21-7 (2008).

17. Cescau, S. et al. Heme acquisition by hemophores. *Biometals* 20, 603-13 (2007).

18. Lefevre, J., Delepelaire, P., Delepierre, M. & Izadi-Pruneyre, N. Modulation by substrates of the interaction between the HasR outer membrane receptor and its specific TonB-like protein, HasB. *J Mol Biol* 378, 840-51 (2008).

19. Celia, H. et al. Structural insight into the role of the Ton complex in energy transduction. *Nature* 538, 60-65 (2016).

20. Maki-Yonekura, S. et al. Hexameric and pentameric complexes of the ExbBD energizer in the Ton system. *Elife* 7(2018).

21. Guzman, L.M., Belin, D., Carson, M.J. & Beckwith, J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. *J Bacteriol* 177, 4121-30 (1995).

22. Zheng, S.Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* 14, 331-332 (2017).

23. Zhang, K. Gctf: Real-time CTF determination and correction. *J Struct Biol* 193, 1-12 (2016).

24. Gomez-Blanco, J. et al. Using Scipion for stream image processing at Cryo-EM facilities. *J Struct Biol* 204, 457-463 (2018).

25. Punjani, A., Rubinstein, J.L., Fleet, D.J. & Brubaker, M.A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* 14, 290-296 (2017).

26. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* 192, 216-21 (2015).

27. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* 7(2018).

28. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. & Barton, G.J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189-91 (2009).

29. Baker, N.A., Sept, D., Joseph, S., Holst, M.J. & McCammon, J.A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 98, 10037-41 (2001).

Chapter IV

4. Cryo-EM structure of the human thyroglobulin

This chapter reports on the structure of the human thyroglobulin the central factor in thyroid hormone synthesis obtained by Cryo-EM. This constitutes an application case of SPA applied to a large (660 kDa) C2 symmetric homodimer. The structure provides detailed information about non-homologous insertions in the highly repetitive hTg sequence, and provides a framework to understand hundreds of clinically relevant mutations. The hTg density map also greatly benefited from local refinement of two flexible regions providing a new example of successful application of this relatively new processing method.

Contribution: cryo-EM sample preparation; cryo-EM dataset acquisition and processing; contributed to atomic modelling; contributed to manuscript writing, editing and reviewing.

The following section is adapted from a manuscript in preparation

Running Title:

Cryo-EM structure of the human Thyroglobulin

Ricardo Adaixo[1]; Eva Rebrova[2]; Ricardo Righetto[1];
Henning Stahlberg[1]; Nicholas Taylor[2]

1 Center for Cellular Imaging and Nano Analytics (C-CINA), Biozentrum, University of Basel, Basel, Switzerland;
2 Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

4.1. Introduction

Thyroglobulin (Tg) is a 660 kDa hyperglycosylated protein expressed to high levels in thyrocytes. Dimeric Tg is secreted to the follicular cavity and iodinated to different extents at specific tyrosine residues, a process modulated by the dietary iodine intake. Iodinated Tg is transported to the thyrocyte cytosol by pinocytosis and digested releasing T3 and T4 hormones (Di Jeso and Arvan, 2016). Therefore, Tg is simultaneously a precursor for thyroid hormones (TH) biogenesis and the carrier protein responsible for iodine storage in the follicle colloid. TH are essential to fetal brain development as well as throughout adulthood as metabolism regulators (Pharoah, Buttfield and Hetzel, 1971). Mutations in the Tg sequence are related to increased propensity of thyroid cancer occurrence as well as hypothyroidism and other clinical conditions such as goiter (Alzahrani *et al.*, 2006; Citterio *et al.*, 2013).

Analysis of primary sequences of Tg allowed early identification of internal homology domains classified as type 1, type 2 and type 3 (Mercken *et al.*, 1985) as well as a cholinesterase like domain (ChEL) at the carboxy end of the protein (Mori, Itoh and Salvaterra, 1987). Type 1 repeats occur 11 times in the human Tg (hTg) sequence and are homologous to 1a domains found in other proteins with known structure (Guncar *et al.*, 1999; Pavšič *et al.*, 2014). The ChEL domain also has homologues with structure determined by X-ray diffraction (Sussman *et al.*, 1991). Rather than performing enzymatic functions the ChEL domain assists Tg folding, dimerization and secretion processes (Lee, Jeso and Arvan, 2008).

Despite the extensive biochemical characterization of Tg done in the past decades, the three-dimensional structure of Tg remained unknown (Citterio, Targovnik and Arvan, 2019) limiting our understanding of its function. We present the structure of hTg determined by cryogenic transmission electron microscopy. The obtained density map depicts a dimer with extensive interchain contacts which span more than the ChEL interface. The hTg dimer is decorated

with several disulfide bridges which add structural stability and rigidity to most of the protein, however, the extreme N and C terminal segments as well as two other regions display a higher degree of flexibility which is likely to be related to function.

We provide a comprehensive structural description of the hTg dimer, including an atomic model, in a state we believe to resemble the early cathepsin digestion product where the extreme N and C hormonogenic sequences are proteolytically cleaved.

4.2. Results

4.2.1. Overall hTg architecture

We obtained an homogenous solution of hTg by resuspension and gel filtration of the commercially available lyophilate without performing any in-vitro iodination. The cryo-EM images of plunge frozen vitrified hTg solution displayed randomly oriented particles with size and shape consistent with previous observations in negative stain preparations (Kim *et al.*, 2015). A composite map at 3.2 Å nominal resolution was reconstructed allowing atomic modelling of the hTg dimer to around 90% completeness (2483 modelled residues over 2748 expected residues per chain). The composite map is the result of the consensus map and two local refinement maps elaborated around particularly flexible regions ("wing" and "foot"; described later). Due to their peripheral localization and flexible nature, such regions may have constituted an obstacle for hTg crystallization attempts in the past. The hTg dimer is approximately 250 Å long by 160 Å wide and 110 Å along the C2 symmetry axis. Each chain is formed by regions I, II and III and a C terminal ChEL domain (Figure 4.1). The interface between monomers buries an area of 8,860 $Å^2$ involving all regions except region II.

The hTg structure is annotated similarly to what is reported in the literature, however, region I lacks the so called "linker" between repeats 1.4 and 1.5.

Three types of cysteine rich internal homology repeats are present in hTg: type 1, type 2 and type 3. There are 10 type 1 repeats within region I and an 11th in region II. Three type 2 repeats, each bearing 14 to 17 residues, lie between the hinge region and repeat 1.11. Type 3 repeats are located between repeat 1.11 and the ChEL domain.

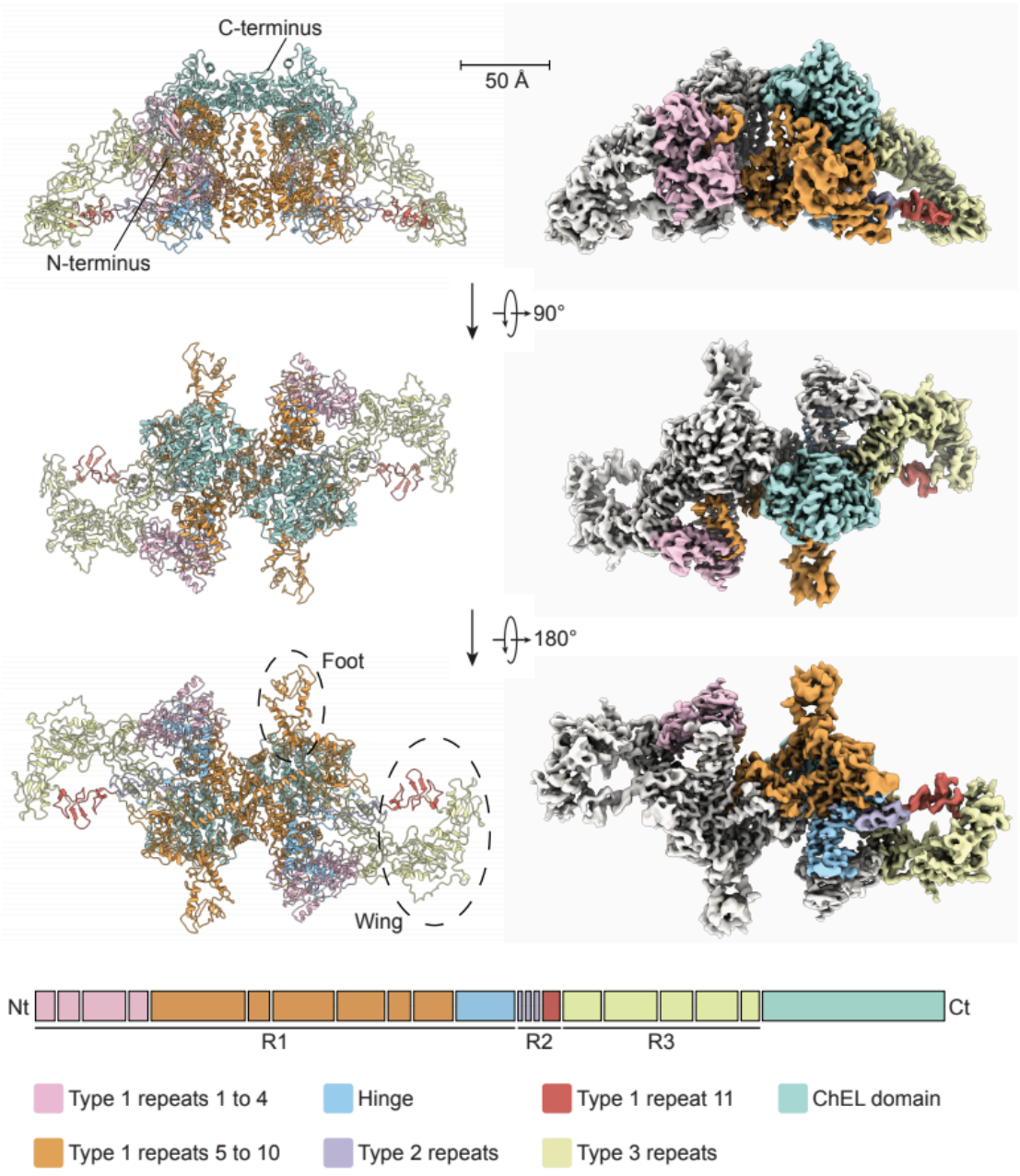

Figure 4.1: Cryo-EM structure of hTg. Ribbon representation of hTG atomic model (top left) and composite map (top right). Linear representation

of hTg domains sequence (middle). The color code of all images is depicted at the bottom.

4.2.2. Type 1 repeats

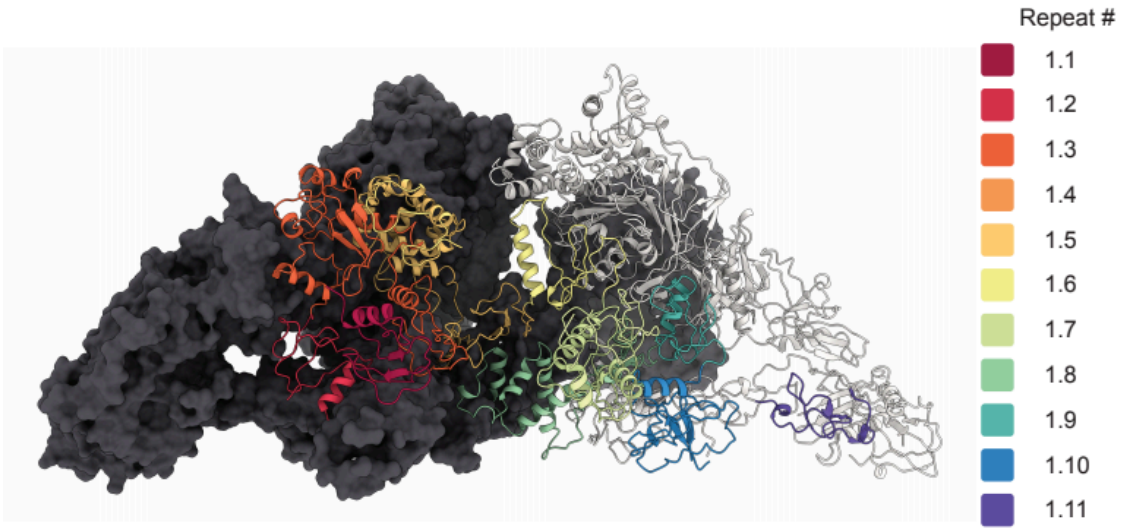

Figure 4.2: Disposition of the different type 1 repeats in the hTg structure. Surface representation of one hTg monomer in the background colored charcoal. Ribbon representation of the second hTg monomer in the foreground with each type 1 repeat colored differently (color code on the right).

The first four type 1 repeats in the proximal region I cluster at the N terminus of each chain and establish significant contacts with region III and the hinge of the opposing chain, as well as intra chain contacts with repeat 1.5 (figures 4.1 and 4.2). Within repeats 1.1 to 1.4 we observe two non-homologous insertions (NHI), namely on loops II and III of repeat 1.3. Repeats 1.5 to 1.10 occupy the central core of hTg and form contacts to all other regions on both chains. Additional NHIs are present in loop I of repeat 1.5, loop II of repeat 1.7 and loop II of repeat 1.8. Insertions of repeats 1.3 and 1.5 are in close proximity and exposed between the proximal region I and the ChEL domain of the opposing

chain. Insertions of repeat 1.8 from both chains lie at the C2 symmetry axis and form a helix bundle providing additional 1710 $Å^2$ interchain contact surface (Figure 4.3).

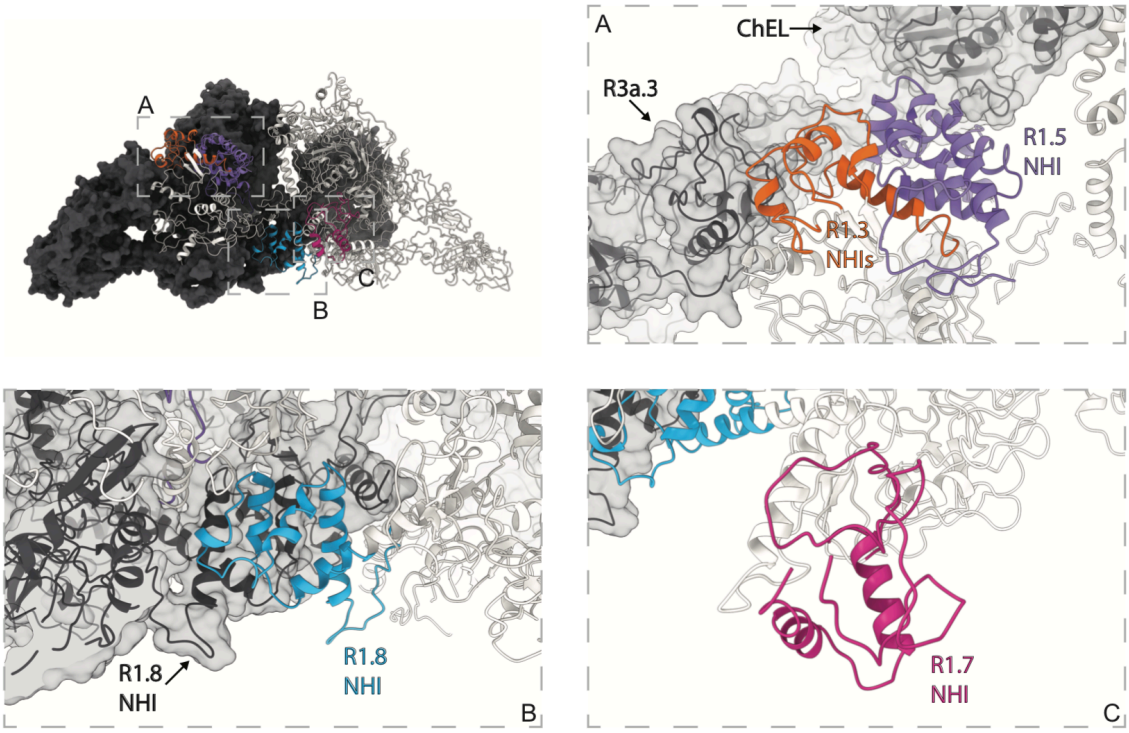

Figure 4.3: Location of type 1 repeat NHIs. (Top left) Surface representation of one hTg monomer in the background colored charcoal. Ribbon representation of the second hTg monomer in the foreground with each NHI colored differently: orange – repeat 1.3 NHIs; purple – repeat 1.5 NHI; blue - repeat 1.8 NHI; wine – repeat 1.7 NHI. Detailed views of each insertion are displayed in the dashed boxes.

Repeat 1.7 exposes a NHI protruding almost radially to the C2 symmetry axis and forms no additional contacts. We named this protrusion as “foot”, it is flexible as suggested by the diffuse density obtained in the consensus map.

Residues 378 to 615 were assigned to the so called “linker region” in previous work while in our structure they form the insertion of repeat 1.5. A consequence of our annotation is that repeat 1.5 encompasses 2 additional cysteines, Cys408 and Cys608, and a total of 4 disulfide bridges, therefore we classify this repeat as

type 1c, as opposed to type 1a and type 1b repeats containing 6 and 4 cysteine residues respectively. The remaining NHIs are devoid of additional cysteine residues.

4.2.3. Type 2 repeats

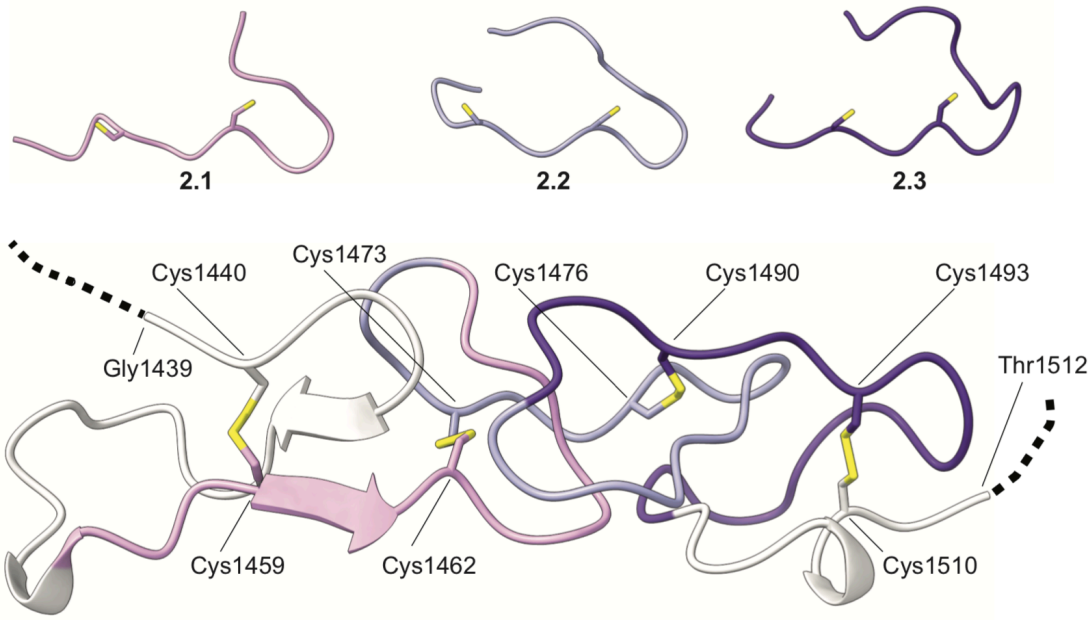

Figure 4.4: Type 2 repeats. (Top) Ribbon representation of the aligned type 2 repeats. (Bottom) Ribbon representation of the type 2 repeats as disposed in the context of the hTg structure. Cysteine residues represented in sticks.

hTg residues 1456 to 1487 comprise three contiguous type 2 repeats flanked by the hinge and repeat 1.11. Each type 2 repeat comprises 2 cysteine residues all of which consumed in disulfide bond formation (figure 4.4). The most N terminal cysteine of each type 2 repeat establishes a disulfide bond (DSB) with the adjacent N terminal domain while the most C terminal cysteine establishes a DSB with the adjacent C terminal domain. Therefore, both repeat 1.11 and hinge region are linked to repeats 2.1 and 2.3 via DSBs contrary to previous hypothesis where all 6 cysteines in type 2 repeats would form DSBs internally.

Repeat 2.1 has a small beta strand between both cysteine residues which is the only secondary structure element found within the set. All type 2 repeats have a shape similar to an arrow head where the pointy edge of one repeat is embraced by the flat base of the following type 2 repeat.

It has been hypothesized the CXXC motif in type 2 repeats constitutes a thioredoxin box which may be required for TG multimerization via intermolecular DSBs (Klein *et al.*, 2000). We did not test this hypothesis, however, all type 2 repeats are extensively exposed to the solvent and therefore could serve as a potential substrate to thioredoxin.

4.2.4. Type 3 repeats

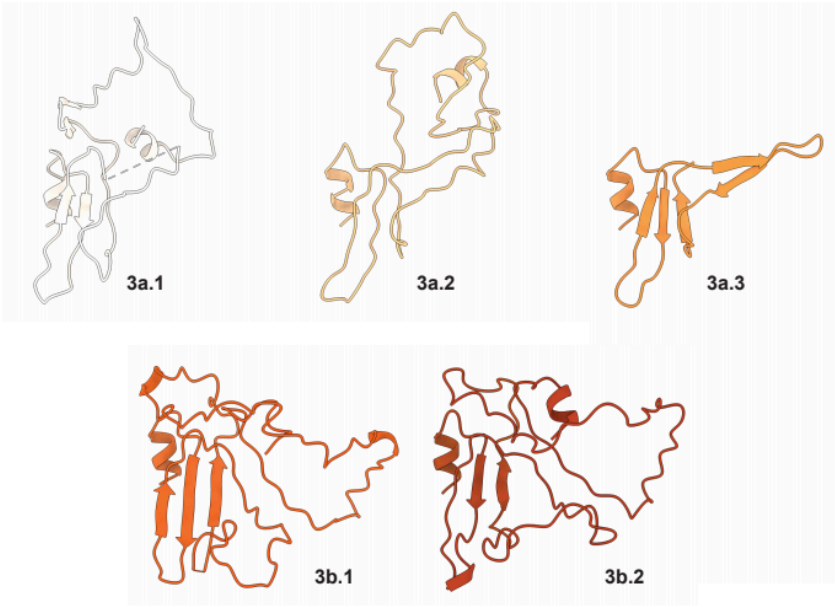

Figure 4.5: Type 3 repeats. Ribbon representation of the aligned type 3 repeats.

Region III comprises a total of 5 type 3 repeats linking the ChEL domain to repeat 1.11. These repeats are formed an alpha helix followed by a 3 stranded beta sheet and can be subdivided into type 3a, bearing 8 cysteines, and type 3b, bearing 6 cysteines (figure 4.5). The loop connecting the third beta strand to the neighbor domain is longer and apparently disordered. We noticed the annotation

of previous authors (Citterio, Targovnik and Arvan, 2019) places the limits of each type 3 repeat within secondary structure elements and therefore does not take into consideration the globular nature of individual domains but, for sake of consistency, we will follow the same annotation.

Repeats 3a3 and 3b2 were reasonably well defined in the consensus map (see materials and methods section) however this was not the case for repeats 3a2, 3b1, 3a1 and down to repeat 1.11. Both 2D classes and consensus refinement maps were fuzzy in the 3a2 to 1.11 region (which we named “wing region”) likely due to increased flexibility. The wing has a “C” shape where the tips seem to act as pivot points, connecting repeats 2.3 to 1.11 at the N terminus and repeats 3a2 to 3b2 at the C terminus. The reason for the increased flexibility of the wing region is unknown to us.

4.2.5. Mapping of proteolysis sites

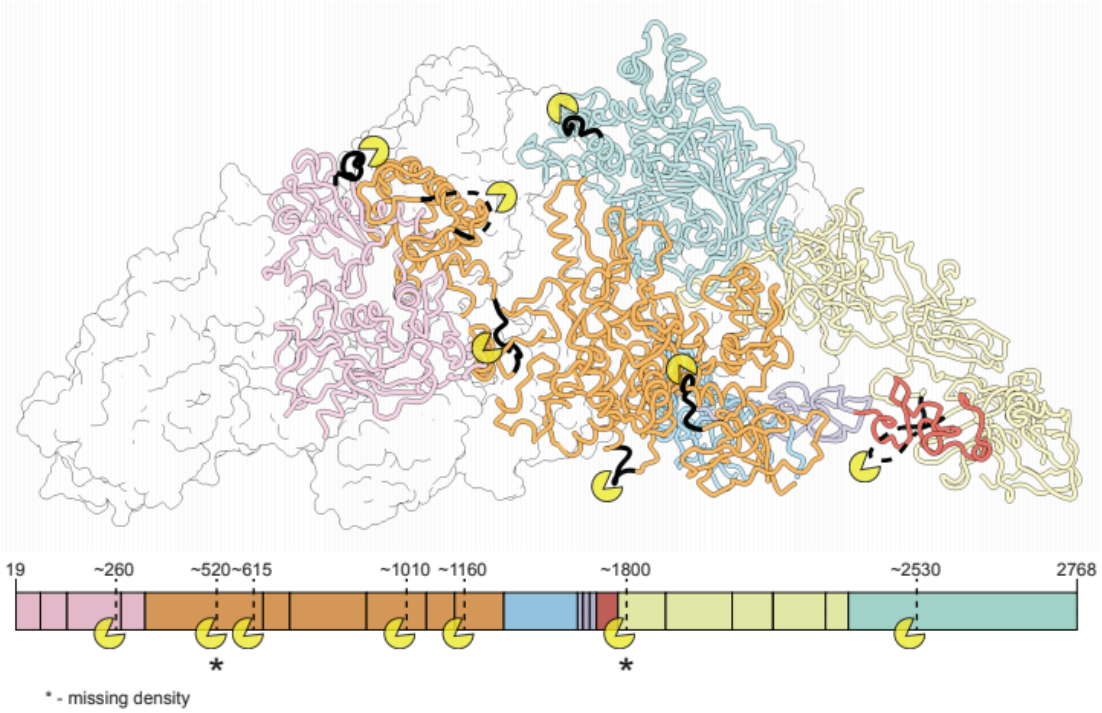

Figure 4.6: Location of proteolysis sites in hTG. (Top) Surface representation of one hTg monomer in the background colored white. Ribbon representation of the second hTg monomer in the foreground with each region colored as in figure 4.1. Approximate location of protease recognition sequences depicted in black with a yellow major sector. (Bottom) Approximate location of protease recognition sequences in the hTg linear diagram depicted in dashed lines with a yellow major sector. Locations marked with an asterisk symbol were not modelled.

The lifecycle of Tg comprises multiple proteolysis events: the cleavage of the N-terminal 19 residues signal peptide; the N and C terminus cleavages which liberate iodopeptides; the limited proteolysis of Tg which releases Tg particles from the colloid agglomerate and finally the digestion of Tg internalized in the thyrocyte (Rousset *et al.*, 1989; Dunn, Crutchfield and Dunn, 1991; Jordans *et al.*, 2009). Cathepsins, a family of cysteine proteases, perform the mentioned

proteolytic attacks on Tg both inside the thyrocyte and in the follicular lumen. The approximate locations of cathepsin proteolysis sites are depicted in Figure 4.6. Digestion of the extreme N- and C-terminus containing thyroid hormones (TH) is considered to be the earliest proteolysis event experienced by mature Tg.

Insertions of repeats 1.3, 1.5 and 1.8 contain motifs targeted by cathepsins. Hence, one function of these insertions is to expose proteolysis prone segments in order to facilitate hTg digestion.

Our hTg atomic model lacks the residues between N456-P547 and T1781-N1814 because the corresponding EM map density was not observed in those regions. Interestingly, two of the proteolysis sites plotted in Figure 4.6 (asterisk marks on sites ~520 and ~1800) lie within these missing segments suggesting that our sample was partially digested at these specific locations.

A sequence of proteolysis events has been suggested (Friedrichs *et al.*, 2003) involving different proteases from the cathepsin family. Crossing the mentioned sequence with our atomic model and mass spectrometry will probably allow us to locate our particular hTg sample within its life cycle.

4.2.6. Mapping of nonsense and missense mutations

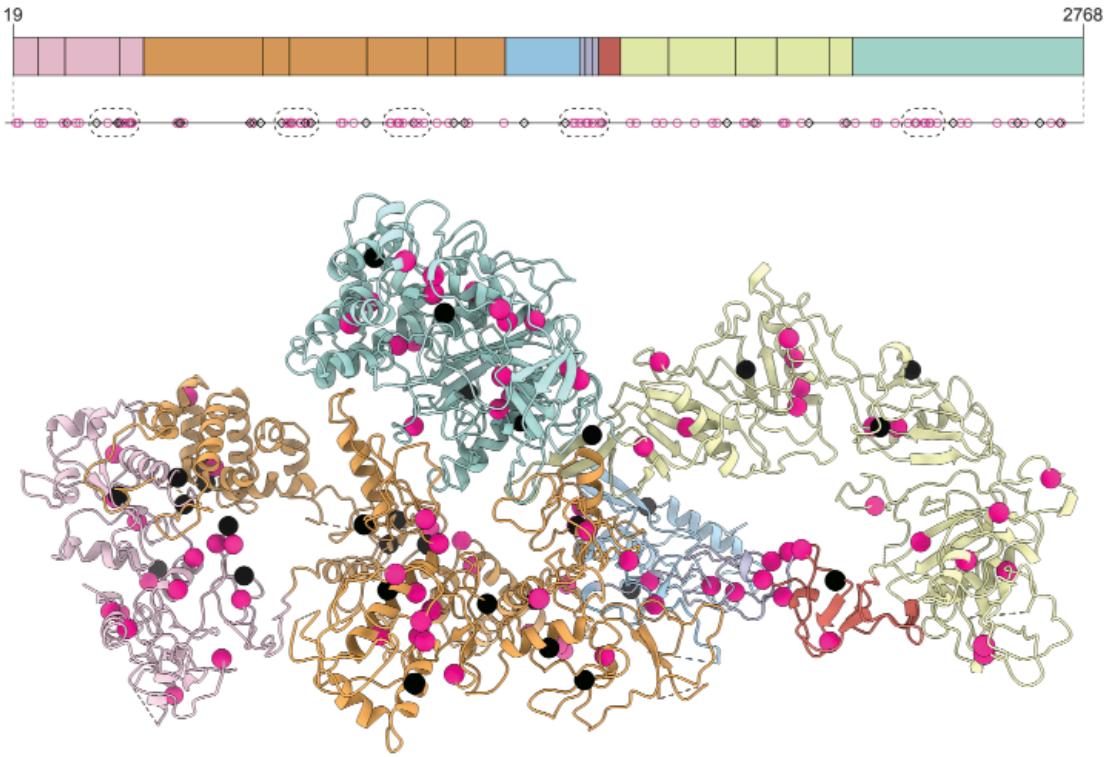

Figure 4.7: Location of nonsense and missense mutations in hTg. (Top) hTg linear diagram depicting the location of nonsense mutations in black squares and missense mutations in pink; domains color code is the same as Figure 4.1. Dashed boxes represent the 5 clusters with the highest density of mutations. (Bottom) Location of the same mutations in the ribbon representation of hTg.

We considered all the nonsense and missense mutations reported in Citterio, Targovnik and Arvan, 2019 and plotted these on our hTg structure (Figure 4.7). The majority of the mutations causing early termination of hTg translation as well as those causing a change in amino acid identity fall within modelled regions; 31 of 36 nonsense mutations and 84 of 90 missense mutations. Interestingly, 4 of the 5 major clusters of the mentioned mutations (figure 4.7 dashed boxes) overlap with the proteolysis sites depicted in figure 4.6, namely those within repeats 1.5, 1.7, 1.8 and ChEL domain.

4.3. Discussion

We determined the atomic structure of hTg based on a composite cryo-EM density map at 3.2 Å nominal resolution. The overall dimeric structure is consistent with previous biochemical experiments. The structural characterization done here also evidences new features which were unknown to the moment: the environment and possible function of the 4 NHIs present in type 1 repeats; the previously annotated linker region which in fact is a NHI of repeat 1.5; the globular nature of type 3 repeats which could be annotated differently and finally the flexible nature of the wing and foot regions.

Except for the case of repeat 1.7, all NHIs are solvent exposed and present peptide motifs recognized by proteases. Interestingly, the NHI of repeat 1.7 is also unique in the sense that the other NHIs establish contacts to adjacent domains other than the type 1 repeat itself. One very speculative hypothesis is that repeat 1.7 NHI could still form inter domain contacts but in the context of hTg multimerization.

The hTg atomic structure is decorated with a variety of post translational modifications which were further studied by MS.

We hope the current structure paves the path to a better description of hTg biology, both in basic sciences as well as diagnostics and therapy areas where our model could be valuable to locate different Fab epitopes.

4.4. Materials and Methods

4.4.1. Sample preparation

Human Thyroglobulin (catalog no. T6830; Sigma-Aldrich) was dissolved in gel filtration buffer (25 mM Tris-HCl pH 7.5, 150 mM NaCl, 1x sodium azide) and injected on a Superdex-200 increase size-exclusion chromatography column connected to an ÄKTA purifier FPLC apparatus (GE Healthcare Bio-Sciences). Peak fractions were pooled and concentrated to 2 mg/mL of protein before plunge freezing.

4.4.3. Cryo-EM sample preparation and data collection

Quantifoil 2/2 400 mesh Cu grids were glow discharged in low pressure air for 30 s. Then 3 µL of concentrated hTg were dispensed to the hydrophilic surface of the grid prior to single side blotting for 2 s and plunge freezing in liquid ethane using a Leica EM GP2 plunger (Leica Microsystems) operating at 20 °C and 80% relative humidity.

Frozen grids were imaged using a Titan Krios (Thermo Fischer Scientific) operating at 300 kV equipped with a Gatan Quantum-LS energy filter (slit width 20 eV; Gatan Inc.) on a K2 Summit direct electron detector (Gatan Inc.). SerialEM (Mastronarde, 2005) was used for automated data collection with 7 acquisitions per hole using beam-image shift (Cheng *et al.*, 2018). Movies were recorded in counting mode with a pixel size of 0.64 Å/px at the sample level. Each movie comprised an exposure of 50 $e^-/Å^2$ fractionated into 50 frames over 10 s.

4.4.4. Image processing and model building

Movies were preprocessed online in FOCUS (Biyani *et al.*, 2017) using MotionCorr2 (Zheng *et al.*, 2017) for drift correction and dose weighting and CTFfind 4.1 (Rohou and Grigorieff, 2015) for contrast transfer function estimation. 7266 movies with estimated CTF resolution better than 4 Å were selected and used for automated particle picking in Gautomatch (Zhang, K.,

https://www.mrc-lmb.cam.ac.uk/kzhang/Gautomatch/) with a CC threshold of 0.4 using a Gaussian blob as template.

Particles were classified in 2D using RELION-3 (Zivanov *et al.*, 2018) and the best classes were selected for ab-initio map generation and auto-refinement. EMAN2 (Tang *et al.*, 2007) was used to create projections of the refined map. Automated particle picking was repeated using 20 Å low pass filtered images of the projections as templates. The new set of particles was pruned by 2D and 3D classification resulting in 37619 particles being allocated to one class with well-defined features and apparent C2 symmetry. This class was further refined imposing C2 symmetry, particles were corrected for beam-induced motion and CTF refined in RELION3. A consensus map with nominal resolution of 3.3 Å was obtained after post processing using an automatically estimated b-factor of -45 $Å^2$.

Particles considered in the consensus map were imported into cryoSPARC v2 (Punjani *et al.*, 2017) for localized refinement with the aim of improving quality of the densities in the "wing" and "foot" regions. Masks around these regions were created in UCSF Chimera (Pettersen *et al.*, 2004) using a local resolution filtered version of the consensus map as template and the volume segmentation tool. Both regions benefited from the local refinement procedure as the final maps have better connectivities and side chain densities than the consensus map in the considered regions. Interestingly the best local refinement maps of the wing region were obtained without performing any prior signal subtraction.

An atomic model of hTg based on the consensus, “foot” and “wing” maps was built in Coot (Emsley et al., 2010). The model covers 90% of the amino acid sequence, lacking mainly loops and the N and C terminus extensions. Maps were merged using phenix.combine_focused_maps and the atomic model real space refined in PHENIX (Adams *et al.*, 2010). The final composite map has a resolution of 3.2 Å based on the FSC curve at 0.143 criterion (Rosenthal and Henderson, 2003).

4.5. Supplementary information

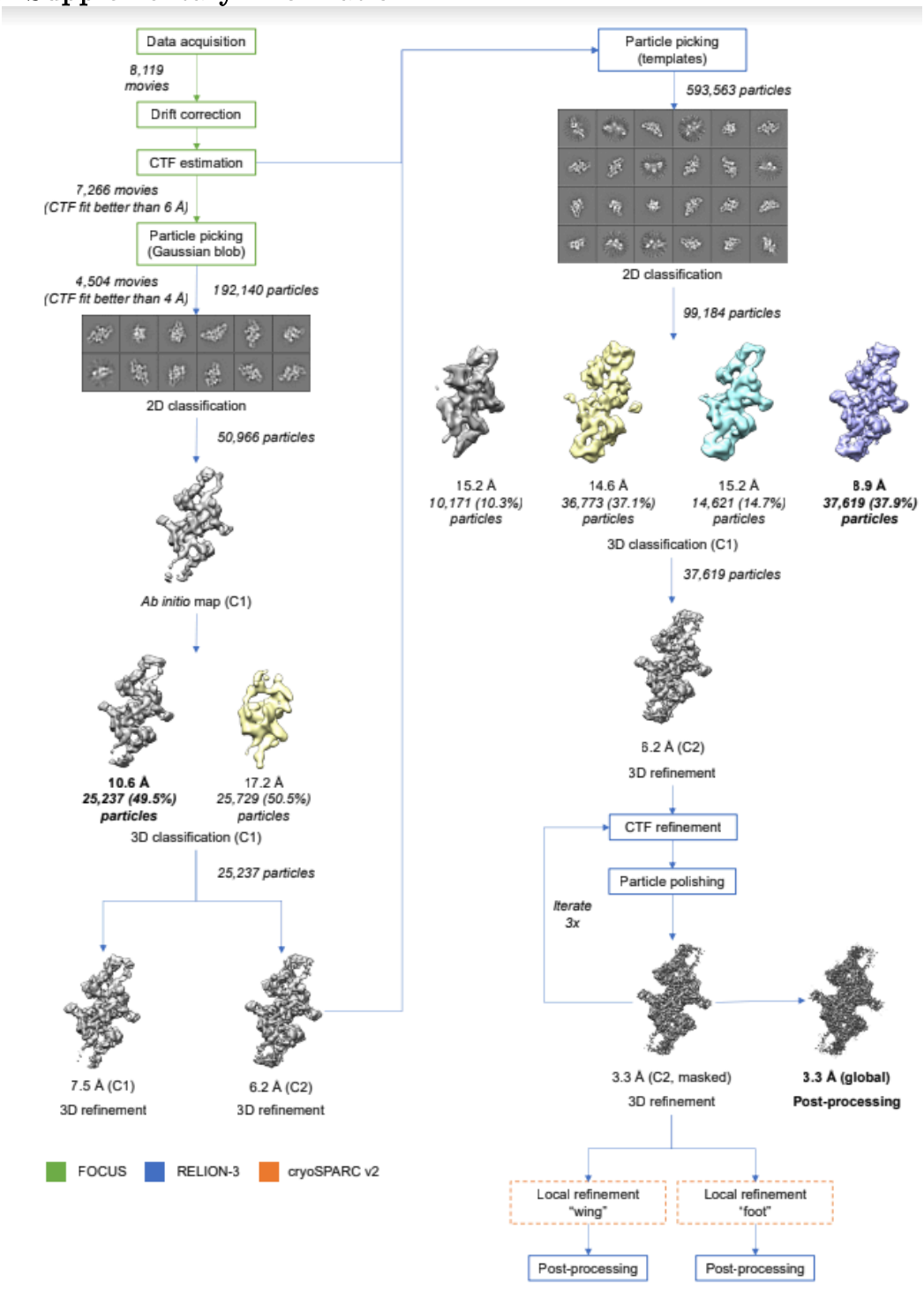

Figure S4.1: processing scheme of the hTg cryo-EM dataset.

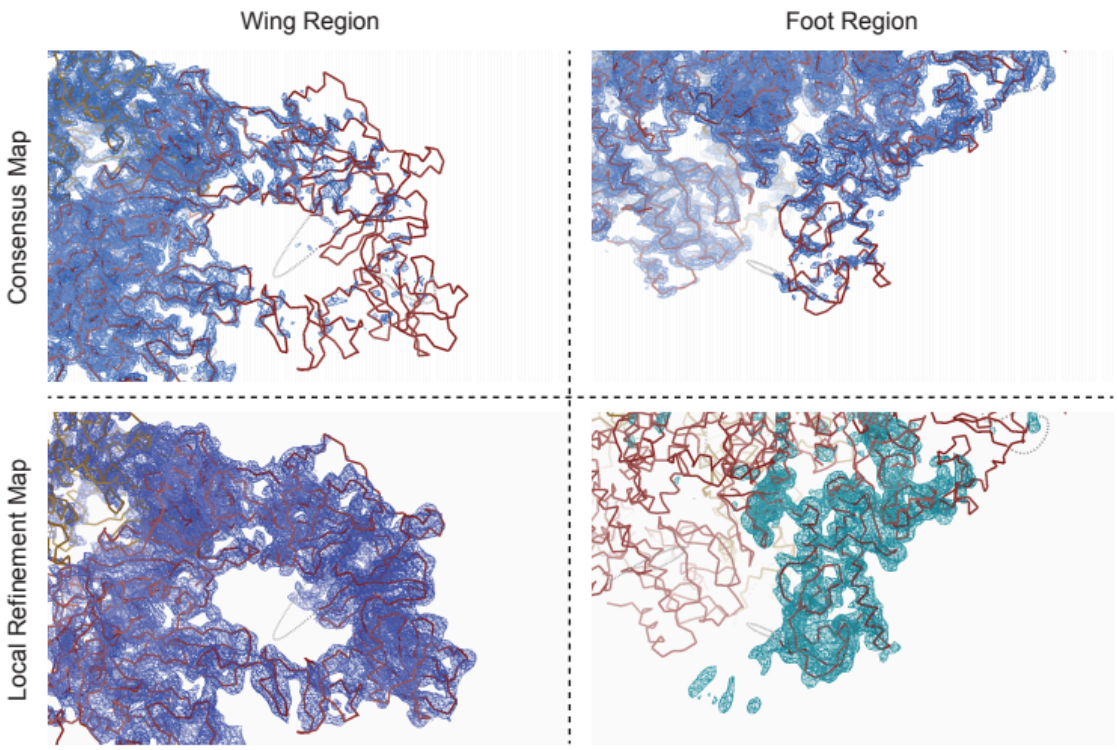

Figure S4.2: comparison of the "foot" and "wing" regions in the consensus map and in the local refinement maps at equivalent threshold levels. Note the foot local refinement was performed with signal subtracted particles while the wing was not.

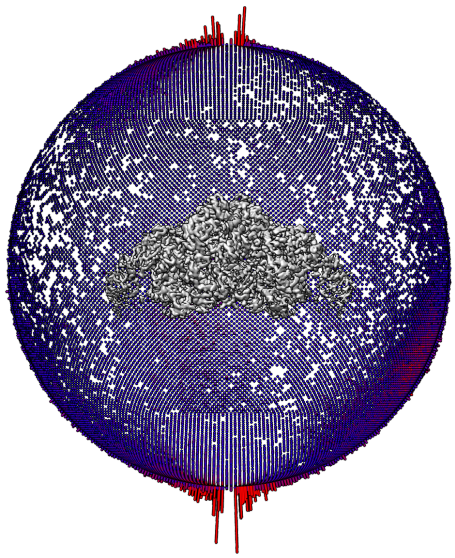

Figure S4.3: Angular plot distribution of the hTg consensus map.

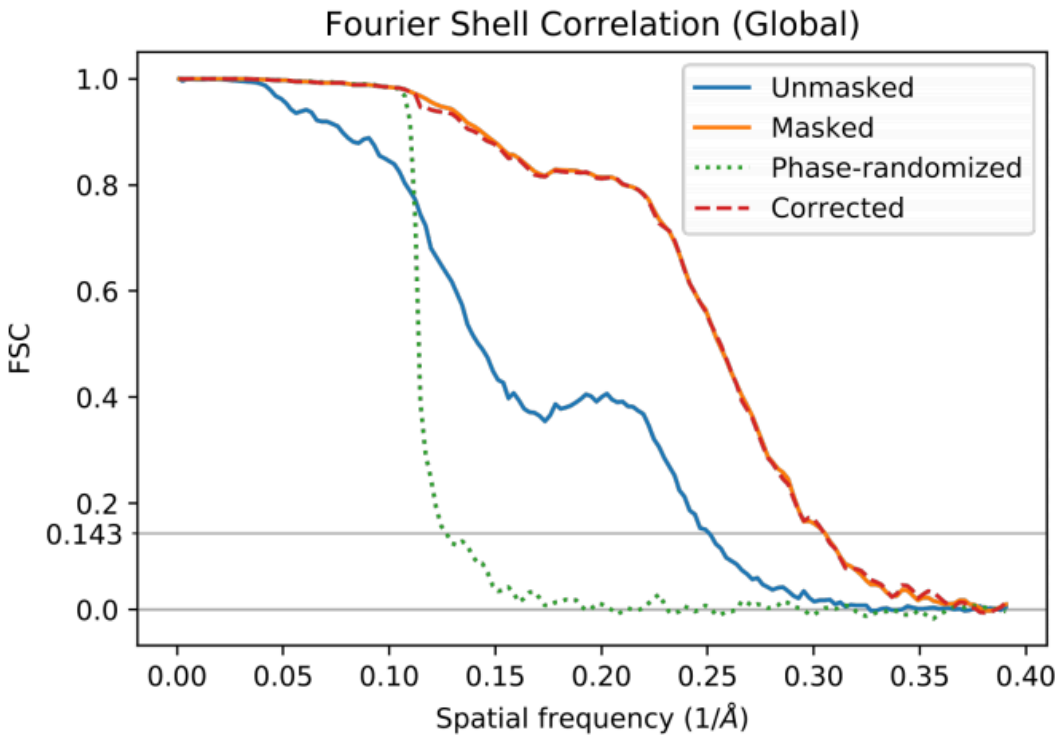

Figure S4.4: FSC plot of the hTg consensus map.

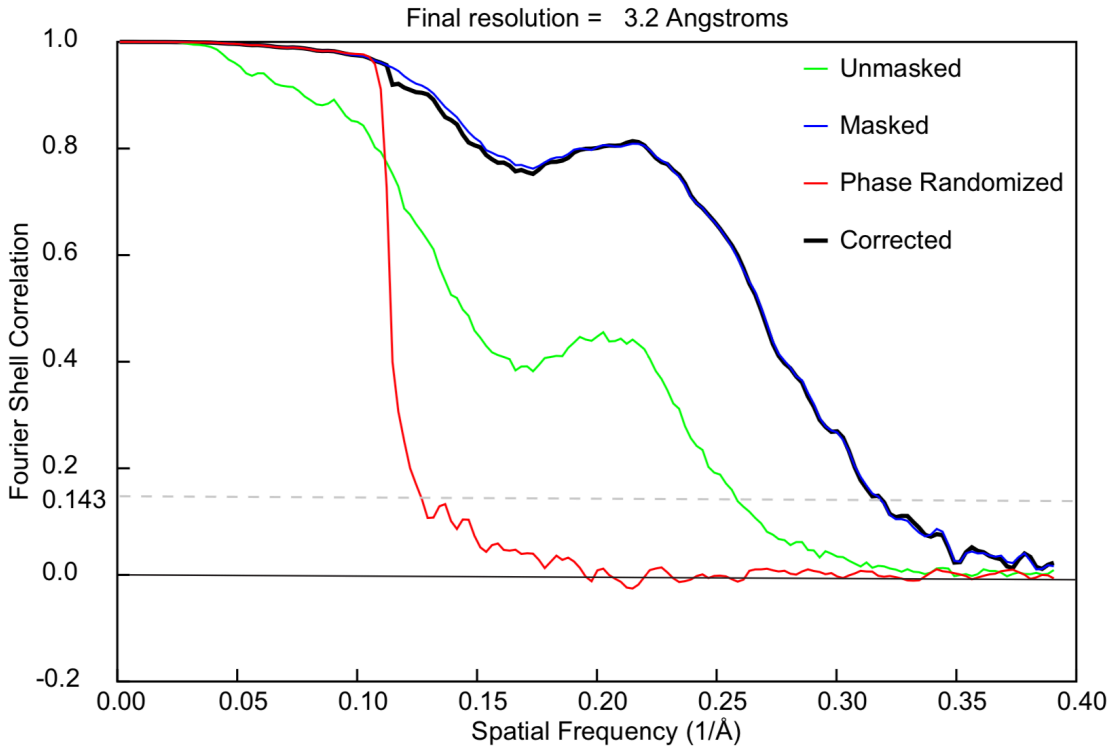

Figure S4.5: FSC plot of the hTg composite map.

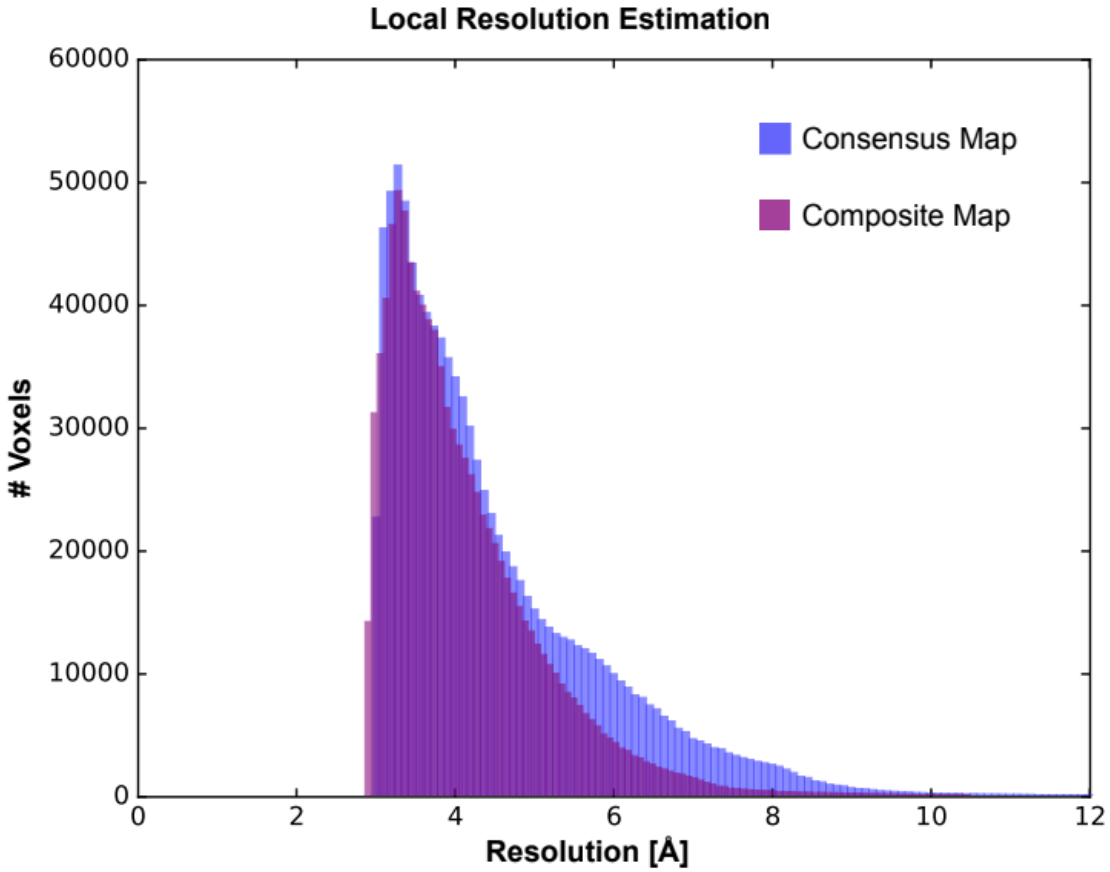

Figure S4.5: local resolution histograms of the consensus and composite hTg maps.

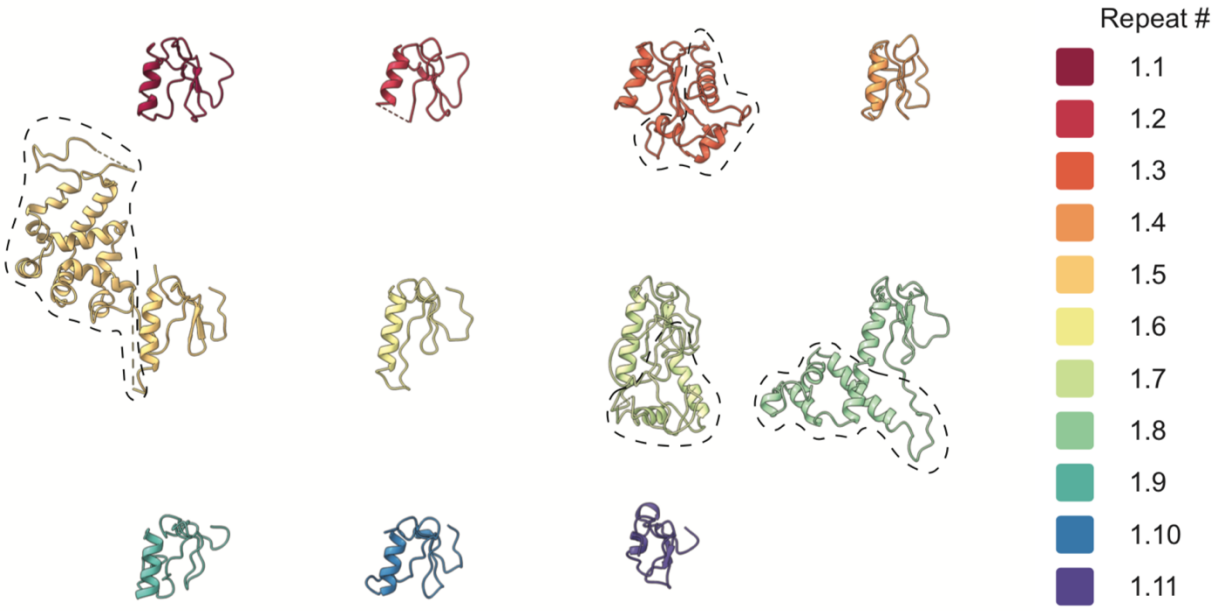

Figure S4.6: Type 1 repeats aligned. Color code for each repeat on the right. NHIs are highlighted with a dashed line.

4.6. Bibliography

Adams, P. D. *et al.* (2010) 'PHENIX: a comprehensive Python-based system for macromolecular structure solution.', *Acta crystallographica. Section D, Biological crystallography*. International Union of Crystallography, 66(Pt 2), pp. 213–21. doi: 10.1107/S0907444909052925.

Alzahrani, A. S. *et al.* (2006) 'Metastatic Follicular Thyroid Carcinoma Arising from Congenital Goiter as a Result of a Novel Splice Donor Site Mutation in the Thyroglobulin Gene', *The Journal of Clinical Endocrinology & Metabolism*. Narnia, 91(3), pp. 740–746. doi: 10.1210/jc.2005-2302.

Biyani, N. *et al.* (2017) 'Focus: The interface between data collection and data processing in cryo-EM', *Journal of Structural Biology*. Academic Press, 198(2), pp. 124–133. doi: 10.1016/J.JSB.2017.03.007.

Cheng, A. *et al.* (2018) 'High resolution single particle cryo-electron microscopy using beam-image shift', *Journal of Structural Biology*. Academic Press, 204(2), pp. 270–275. doi: 10.1016/J.JSB.2018.07.015.

Citterio, C. E. *et al.* (2013) 'New insights into thyroglobulin gene: Molecular analysis of seven novel mutations associated with goiter and hypothyroidism', *Molecular and Cellular Endocrinology*. Elsevier, 365(2), pp. 277–291. doi: 10.1016/J.MCE.2012.11.002.

Citterio, C. E., Targovnik, H. M. and Arvan, P. (2019) 'The role of thyroglobulin in thyroid hormonogenesis', *Nature Reviews Endocrinology*. Nature Publishing Group, 15(6), pp. 323–338. doi: 10.1038/s41574-019-0184-8.

Dunn, A. D., Crutchfield, H. E. and Dunn, J. T. (1991) 'Thyroglobulin processing by thyroidal proteases. Major sites of cleavage by cathepsins B, D, and L.', *The Journal of biological chemistry*, 266(30), pp. 20198–204. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1939080 (Accessed: 8 September 2019).

Friedrichs, B. *et al.* (2003) 'Thyroid functions of mouse cathepsins B, K, and

L.', *The Journal of clinical investigation*. American Society for Clinical Investigation, 111(11), pp. 1733–45. doi: 10.1172/JCI15990.

Guncar, G. *et al.* (1999) 'Crystal structure of MHC class II-associated p41 Ii fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S.', *The EMBO journal*. European Molecular Biology Organization, 18(4), pp. 793–803. doi: 10.1093/emboj/18.4.793.

Di Jeso, B. and Arvan, P. (2016) 'Thyroglobulin From Molecular and Cellular Biology to Clinical Endocrinology', *Endocrine Reviews*. Narnia, 37(1), pp. 2–36. doi: 10.1210/er.2015-1090.

Jordans, S. *et al.* (2009) 'Monitoring compartment-specific substrate cleavage by cathepsins B, K, L, and S at physiological pH and redox conditions', *BMC Biochemistry*. BioMed Central, 10, p. 23. doi: 10.1186/1471-2091-10-23.

Kim, K. *et al.* (2015) 'The Effects of Electron Beam Exposure Time on Transmission Electron Microscopy Imaging of Negatively Stained Biological Samples', *Applied Microscopy*. Korean Society of Microscopy, 45(3), pp. 150–154. doi: 10.9729/AM.2015.45.3.150.

Klein, M. *et al.* (2000) 'The Thioredoxin Boxes of Thyroglobulin: Possible Implications for Intermolecular Disulfide Bond Formation in the Follicle Lumen', *Biological Chemistry*. Walter de Gruyter, 381(7), pp. 593–601. doi: 10.1515/BC.2000.076.

Lee, J., Jeso, B. Di and Arvan, P. (2008) 'The cholinesterase-like domain of thyroglobulin functions as an intramolecular chaperone', *The Journal of Clinical Investigation*. American Society for Clinical Investigation, 118(8), p. 2950. doi: 10.1172/JCI35164.

Mastronarde, D. N. (2005) 'Automated electron microscope tomography using robust prediction of specimen movements', *Journal of Structural Biology*. Academic Press, 152(1), pp. 36–51. doi: 10.1016/J.JSB.2005.07.007.

Mercken, L. *et al.* (1985) 'Primary structure of bovine thyroglobulin deduced from the sequence of its 8,431-base complementary DNA', *Nature*. Nature Publishing Group, 316(6029), pp. 647–651. doi: 10.1038/316647a0.

Mori, N., Itoh, N. and Salvaterra, P. M. (1987) 'Evolutionary origin of cholinergic macromolecules and thyroglobulin.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 84(9), pp. 2813–7. doi: 10.1073/pnas.84.9.2813.

Pavšič, M. *et al.* (2014) 'Crystal structure and its bearing towards an understanding of key biological functions of EpCAM', *Nature Communications*. Nature Publishing Group, 5(1), p. 4764. doi: 10.1038/ncomms5764.

Pettersen, E. F. *et al.* (2004) 'UCSF Chimera?A visualization system for exploratory research and analysis', *Journal of Computational Chemistry*. John Wiley & Sons, Ltd, 25(13), pp. 1605–1612. doi: 10.1002/jcc.20084.

Pharoah, P. O., Buttfield, I. H. and Hetzel, B. S. (1971) 'Neurological damage to the fetus resulting from severe iodine deficiency during pregnancy.', *Lancet (London, England)*, 1(7694), pp. 308–10. doi: 10.1016/s0140-6736(71)91040-3.

Punjani, A. *et al.* (2017) 'cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination', *Nature Methods*. Nature Publishing Group, 14(3), pp. 290–296. doi: 10.1038/nmeth.4169.

Rohou, A. and Grigorieff, N. (2015) 'CTFFIND4: Fast and accurate defocus estimation from electron micrographs', *Journal of Structural Biology*. Academic Press, 192(2), pp. 216–221. doi: 10.1016/J.JSB.2015.08.008.

Rosenthal, P. B. and Henderson, R. (2003) 'Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy', *Journal of Molecular Biology*. Academic Press, 333(4), pp. 721–745. doi: 10.1016/J.JMB.2003.07.013.

Rousset, B. ***et al.*** (1989) 'Thyroid hormone residues are released from thyroglobulin with only limited alteration of the thyroglobulin structure.', *The Journal of biological chemistry*, 264(21), pp. 12620–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2663861 (Accessed: 8 September 2019).

Sussman, J. L. ***et al.*** (1991) 'Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic acetylcholine-binding protein.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 253(5022), pp. 872–9. doi: 10.1126/science.1678899.

Tang, G. ***et al.*** (2007) 'EMAN2: An extensible image processing suite for electron microscopy', *Journal of Structural Biology*. Academic Press, 157(1), pp. 38–46. doi: 10.1016/J.JSB.2006.05.009.

Zheng, S. Q. ***et al.*** (2017) 'MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy', *Nature Methods*. Nature Publishing Group, 14(4), pp. 331–332. doi: 10.1038/nmeth.4193.

Zivanov, J. ***et al.*** (2018) 'New tools for automated high-resolution cryo-EM structure determination in RELION-3', *eLife*, 7. doi: 10.7554/eLife.42166.

Chapter V

5. Conclusions and Outlook

Within the last decade cryo-EM methods applied to biological samples evolved at a stellar pace to what it is today. The possibility of studying macromolecules in a near native environment, requiring minute amount of purified sample and still recover data with enough resolution to build atomic models caused a rapid shift of the structural biology community towards specialization in cryo-EM; this is particularly evident when considering groups migrating from XRD.

Does it deliver? And if yes, to what extent? To answer these questions, I invite the reader to consider chapters II, III and IV of this thesis which constitute application cases of cryo-EM SPA to a set of samples with very different characteristics. Particle size seems to be among the most important factors influencing the quality and resolution of the final density maps. Larger particles also seem to be more forgiving relatively to the freezing conditions. It is estimated that as much as 80% of the particles embedded in the vitreous ice may adopt non-native conformations either by contact with the air water interface or by other unknown phenomena. In my experience, larger soluble proteins like Thyroglobulin (chapter IV), Urease or Beta-Galactosidase (data not shown) result in high quality particles embedded in vitreous ice and require minimal optimization of the freezing setup. The high SNR large particles provide along with symmetry (all proteins in this set are symmetric) account for smooth and fast processing of the raw data, i.e. (1) particles are easy to identify in the aligned averages, (2) 2D and 3D classification methods provide classes which truly represent different states/conformations rather than different views, (3) particle polishing and CTF refinement approaches do improve the resolution of the final maps, (4) refinement of sub regions is possible. In the particular case of Urease the time between “protein in a tube” to “sub 4 Å density map” took us less than

24 h and was performed in the context of master students training (manuscript in preparation).

On the other hand, not so large membrane proteins pose many challenges at different stages which may not be evident in the presented chapters but are the reality for majority of protein samples in the lower molecular weight range. In this regime protein production and purification can often be the first obstacle in the cryo-EM project: the experimenter typically needs ~0.1 mg of protein to start freezing and screening attempts; the sample must be as homogeneous as possible in terms of composition and particle conformation. Techniques for purification and stabilization of proteins are well described and have been used extensively by structural biologists. In fact, it is common to start from a highly homogeneous and pure sample, as evaluated by gel filtration profiles (for example), but for unknown reasons the resulting frozen grids display deteriorated or heterogeneous looking particles, aggregates, low number of particles, preferential orientation problems or any combination of all these. Other problems like preferential adsorption to the grid's carbon film or too thick vitreous ice may also occur but are usually easier to solve. The intuitive response to this discrepancy problem is to blame the freezing process and a typical approach to solve it is to tweak the freezing parameters or the sample itself. This can be achieved by changing grid type, using solid supports (such as graphene or graphene oxide layers), changing the plasma cleaning protocol of the grids, screening parameters on the plunge freezing robot, or, more related to the sample itself, the user may try to change detergent or membrane embedding system, protein concentration, buffer pH and/or salinity, introduce small concentrations detergents or stabilizing agents, introduce ligands or stabilizing large molecules (Fab or Fab fragments) among other. It is important to mention all these approaches are trial and error based, hence likely to be time consuming and with no guarantee of success. Sample preparation, more specifically sample stabilization and freezing, are in my opinion the biggest bottlenecks in the vast majority cryo-EM projects.

For the mentioned reasons grids of small membrane protein samples are often sub optimal. To some extent this lack of quality can be circumvented by acquiring larger datasets over the course of multiple sessions. Chapter II provides an example where addition of a Fab fragment was crucial to stabilize the Rhodopsin-Gp complex which allowed us to obtain model with clear secondary structure assignments but at the same time border-line concerning side chain assignments. On chapter III two cryo-EM maps were obtained for relatively similar samples: the ExbB pentamer and the ExbB pentamer in complex with the ExbD dimer. Sample preparation, freezing and data acquisition methods were very similar in both cases, however for the pentamer alone a map with 3.2 Å resolution was obtained while for in the pentamer plus dimer sample we obtained a 4 Å resolution map. Processing of these 2 datasets was certainly different because in the first case C5 symmetry could be imposed from the very first SPR steps while in the second case we were limited to C1. To my understanding the main cause for the difference in resolution in these 2 maps arises from the symmetry imposition differences themselves. Signal subtraction of the micelle was also crucial to improve the ExbBD density map, still, this was not sufficient to obtain densities for unambiguous side chain assignment. A tentative atomic model of the complex was built and must be interpreted with caution. As in the case of Rhodopsin-Gp-Fab structure, despite not having a clear atomic description of side chains, the final map still allowed us to make some conclusions about the biology of these proteins in particular: the position of the regulatory C terminal tail of Rhodopsin and the distortions caused on the ExbB pentamer upon binding of the ExbD dimer.

In sum, and taking my personal opinion into account, cryo-EM SPA does deliver to the expectations as a powerful technique to solve protein structures, but, as any technique it is not perfect. Better microscopes, better cameras, better processing algorithms made the data acquisition and processing more reliable, on the other hand sample freezing is not as reliable nor reproducible, despite the

development of freezing robots. Furthermore, sample freezing is highly dependent on the sample characteristics, as discussed above, and optimization of this process is a trial and error process. A deeper understanding of the phenomena occurring during protein plunge freezing is definitely needed to eliminate this bottleneck.

Finally, I would like to focus on Thyroglobulin and why we decided to characterize this protein by cryo-EM. Thyroglobulin was extensively studied for more than 50 years, it is clinically relevant and at the same time it is so stable and robust that most size exclusion chromatography calibration kits include a liophilate of this same protein. With such characteristics Thyroglobulin qualifies as a very interesting target for structure determination by cryo-EM SPA, how come this was not done yet?

In my perspective targets as Thyroglobulin are of incredible value to any academic research group for a number of reasons:

(I) Protein purification is easy and high-quality frozen grids can be prepared with minimal optimization allowing the experimenter to focus on subsequent steps; this is particularly useful for the beginner PhD student.

(II) The sample is robust enough to benchmark freezing, data acquisition and processing; analogous to other proteins used as standards in cryo-EM method development such as Beta-galactosidase or Apoferritin.

(III) Before our project the structure of Thyroglobulin was not known, so, contrary to Beta-galactosidase or Apoferritin the new structure can be discussed and linked to its biological function which adds value to the publication arising from this project.

For the mentioned reasons I think targets like Thyroglobulin make the best

projects for the newcomer PhD student.

What will cryo-EM look like in 10 years? Answering the rhetorical question may provide an outlook, speculative sure, but still a good exercise. I think we can expect a slow but constant development of the algorithms used in cryo-EM data processing, even though they are not the limiting factor at the present time. We have seen that it is possible to study flexible regions within proteins and I would not be surprised to see more detailed descriptions of protein motion arising from localized refinement or similar processing methods. Artificial intelligence based algorithms were successfully applied to particle picking and classification of heterogeneous populations of particles, and I think the trend is to extend this approach to other stages of the pipeline (for example curation of dataset aligned averages and/or extracted particles, "guided" automated data collection, etc.).

The development of direct electron detectors revolutionized cryo-EM but we know these could be improved, more specifically, we know not all spatial frequencies are equally well integrated, a characteristic that is easy to appreciate when the modulation transfer function of the detector is considered. At half-Nyquist current detectors are approximately 60% efficient (depending on which study is considered) in detecting the TEM signal and this efficiency drops as Nyquist frequency is approached. In practice this means that much more information could be retrieved from each single image collected by the detector, and the dampening is especially pronounced in the higher resolution range. In sum, better detectors would translate to higher SNR images and would very likely be well received by the cryo-EM community.

Finally, I think sample preparation methods (a topic already explored in this thesis) are currently the major bottleneck in the SPA pipeline. Irrespective of their relative efficiencies all other steps of protein structure determination by cryo-EM can be optimized using a rational approach while plunge freezing remains essentially a trial and error task; somehow resembling protein

crystallization in the XRD field. New methods and devices for freezing cryo-EM grids, which don't use blotting paper and/or exploit microfluidics technology, are slowly becoming available and I hope these add reproducibility and ease to this process.

Acknowledgements

I would like to express my sincere gratitude towards all the people who, one way or another, made this work possible.

My partner Sofia Cunha, for leaving Portugal behind, learning a new language and sharing the good and not so good moments that came with my decision of doing a PhD in Switzerland.

My supervisor Henning Stahlberg, who accepted me in his lab and always provided guidance and a good environment for my training. Thanks to Jan Pieter Abrahams and Tilman Schirmer for accepting the invitation to be in my PhD evaluation board.

To João Morais-Cabral for accepting me in his lab as a master student and providing one-on-one training in x-ray crystallography, along with so much more.

All my colleagues at C-CINA for all the enlightning discussions and help especially: Mohamed Chami, Ricardo Guerrero and Robert McLeod, Thomas Braun, Daniel Castaño-Díez.

Once my bright master student and currently PhD candidate Lena Muckenfuss.

To the researchers outside of C-CINA with whom I had the pleasure to collaborate: Jacopo Marino, Valérie Biou, Philippe Delepelaire, Crina Nimigean, Simon Scheuring, Alexander Schmidt and Sergey Nazarov.

The great friends and colleagues who helped solving all kind of problems and in addition were available for an occasional beer: Ricardo Righetto, Inaytulla Mohammed and Nicholas Taylor.

To the support team in C-CINA which kept all our microscopes in shape and running: Lubomir Kovacik, Ariane Fecteau-Lefebvre, Daniel Coujolle-Bert and Kenneth Goldie as well as Karen Bergmann and Angie Klarer for administrative support.

To my friends and partners in crime Hugo Aragão and Leonildo Delgado. All the amazing friends that I got to meet in Basel and those spread around the globe especially Leonor, Morgane, Maria João, Sérgio, Zé, Nana, Tono, Cristina, Catarina and Filipe.

Finally, to my sister Ana, parents, Maria and João Adaixo, and my grandparents Dulce and Abel Adaixo for their unconditional love, support and for providing a true north in difficult times.