# *High-Throughput Determination of Mycobacterium smegmatis Protein Complex Structures*

Angela Mary Kirykowicz

**A thesis submitted in fulfilment of the requirements for the degree of Master of Medical Science in Medical Biochemistry in the Department of Integrative Biomedical Sciences**
**9/2/2018**



Supervisor: Jeremy David Woodward

# Table of Contents

**DECLARATION**

I, Angela Mary Kirykowicz, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 09/02/2018

# Acknowledgements

This has been a (very) long journey with many people to thank for their support along the way. I would like to first thank my family for being there for me during the dark moments of this journey; without your emotional support I could not think of how I would have made it through. I would also like to thank my supervisor, Jeremy Woodward, for guidance through this difficult project.

A special thanks to Mohammed Jaffer for his kindness and patience through the (tricky) cryo-EM work. I owe Brandon Weber inspiration through the more difficult moments, with his vibrant philosophy on the scientific enterprise and the utility of learning from failures and mistakes. Both Brandon and Trevor Sewell were kind enough to read a draft of this thesis and offer helpful suggestions on improving the manuscript.

For the mass spectrometry work, I would like to thank the Blackburn Group and of course the Yale MS & Proteomics Resource. Thanks to Madhu Chan for showing me how to use the ultracentrifuge. Lastly, I would like to thank the National Research Foundation and the University of Cape Town for providing the funds which made this work possible.

# List of Abbreviations

| | |
|---|---|
| BN | Blue Native |
| Brf | Bacterioferritin |
| BSA | Bovine serum albumin |
| *B. thetaiotaomicron* | *Bacteroides thetaiotaomicron* |
| CN | Clear Native |
| CTF | Contrast Transfer Function |
| DQE | Detector Quantum Efficiency |
| DyP | Dye-decolourising peroxidase |
| EC | Enzyme Classification |
| *E. coli* | *Escherichia coli* |
| EM | Electron Microscopy |
| EMDB | Electron Microscopy Data Bank |
| Enc | Encapsulin |
| FDR | False discovery rate |
| Flp | Ferritin Family Protein |
| FolB | 7,8-dihydroneopterin aldolase |
| FRET | Fluorescence Resonance Energy Transfer |
| FT | Fourier Transform |
| GSI | Glutamine Synthetase I |
| GST | Glutatione S-transferase |
| KatG | Catalase |
| LB | Luria-Bertani |
| LC | Liquid Chromatography |
| MS | Mass Spectrometry |

| | |
|---|---|
| *Msm* | *Mycobacterium smegmatis* |
| *Mtb* | *Mycobacterium tuberculosis* |
| MW | Molecular Weight |
| *N. europaea* | *Nitrosomonas europaea* |
| ORF | Open-reading frame |
| NMR | Nuclear magnetic resonance |
| NS | Negative Stain |
| PAGE | Polyacrylamide gel electrophoresis |
| PDB | Protein Data Bank |
| *P. furiosus* | *Pyrococcus furiosus* |
| pI | Isoelectric Point |
| PIN | Protein-Protein Interaction Network |
| PPI | Protein-Protein Interaction |
| SDS | Sodium Dodecyl Sulphate |
| SNR | Signal-to-Noise Ratio |
| TAP | Tandem Affinity Purification |
| TB | Tuberculosis |
| TEM | Transmission Electron Microscopy |
| *T. maritima* | *Thermotoga maritima* |
| Y2H | Yeast Two-Hybrid |

# Abstract

Tuberculosis (TB) is an endemic health-crisis, particularly in sub-Saharan Africa. The rise of multi- and extensively-drug resistant *Mycobacterium tuberculosis* (*Mtb*), the causative agent of TB, has led to further developments in understanding the physiology of *Mtb* during infection, as well as searching for novel drug targets, in order to combat the disease. Our understanding of cells, both eukaryotic and prokaryotic, has changed substantially in the last 50 years, incorporating the role of stable and transient protein-protein interactions which govern cell function and behaviour. Although there are many *in vivo* and *in vitro* methods for studying protein-protein interactions, they suffer from the lack of ability to distinguish physiological interactions from interactions that occur which are not physiologically relevant to the cell. Structure-based methods for determining protein interactions have the benefit of screening out false positives whilst simultaneously assessing the possible biological function of the protein complex in question. This study sought to assess different high-throughput methods for capturing stable, water soluble protein complexes from *M. smegmatis* (*Msm*), a close relative of *Mtb*, for structural characterisation by low-resolution transmission electron microscopy (EM). The use of partial biochemical fractionation was assessed, which produced low-resolution structures of glutamine synthetase I, bacterioferritin, and Encapsulin. These structures were unambiguously identified through a combination of fitting of homologous crystal structures into the low-resolution maps, and information obtained by liquid chromatography mass spectrometry (LC-MS/MS) of bands isolated from native- and SDS-PAGE gels. Since Encapsulin is likely to participate in the *Msm* oxidative stress response and functions to enclose the target proteins DyP-type peroxidase (DyP) and ferritin-family protein (BrfB), optimal conditions for cryo-EM were tested for further efforts to obtain a high-resolution structure. Furthermore, hypotheses were generated for the function of *Mtb* and *Msm* Encapsulin based on the *Msm* Encapsulin structure obtained with the aid of a crystal structure homologue; these related to the mode of cargo binding and pore selectivity. A single-step purification method was also assessed through grid blotting on blue native (BN) PAGE using GroEL as a test protein. The hydrophobicity and charge of the EM copper grid was tested to find the optimal grid property for particle transfer. This established that particles of GroEL could be transferred from BN-PAGE onto an EM copper grid and a successful negative

stain reconstruction was obtained. In summary, the pipeline from purifying protein complexes to generating hypotheses based on structure was successfully investigated in *Msm*, which will aid in the production of novel drug targets for *Mtb* as well as in the application to other organisms.

# Chapter I: Literature Review

## 1.1 Understanding Cells Through Protein-Protein Interaction Networks

Bruce Alberts (1998) elegantly summed how our understanding of cells has changed since the 1960's from viewing them as simply "bags of chemicals" (italics mine):

"But, as it turns out, we can walk and we can talk because the chemistry that makes life possible is much more elaborate and sophisticated than anything we students had ever considered. Proteins make up most of the dry mass of a cell. But instead of a cell dominated by randomly colliding individual protein molecules, we now know that nearly every major process in a cell is carried out by assemblies of 10 or more protein molecules. And, as it carries out its biological functions, each of these protein assemblies interacts with several other large complexes of proteins. Indeed, the entire cell can be viewed as a factory that contains an elaborate network of interlocking assembly lines, each of which is composed of a set of *large protein machines*." (Alberts, 1998)

The "interlocking assembly lines" consisting of "large protein machines" is an intriguing idea which is easily represented as a network graph, where "nodes" consist of individual proteins and "edges" their interactions (e.g de Silva & Stumpf, 2005). Of course, such a topological representation is static whereas cells are constantly responding to environmental stimuli (Levy & Pereira-Leal, 2008). Nevertheless, such network graphs, or interaction networks, are still a useful representation of our budding understanding of the web of interconnected proteins which drive the cell.

Network graphs have been used to represent social networks (e.g Watts & Strogatz, 1998), the World Wide Web (e.g Barbasi & Albert, 1999), and biological processes such as metabolic networks (e.g Jeong *et al*, 2000) and protein-protein interaction networks (PINs) (e.g Jeong *et al*, 2001). There are two types of such interaction networks: exponential and scale-free which

are discriminated by their connectivity, or degree, distributions, P(k), the probability that a node has k connections (Albert *et al*, 2000). Exponential networks are characterised by nodes which display a similar number of connections, and hence P(k) is Poisson with an exponential decay for large k (Albert *et al*, 2000) (**Figure 1.1-1a**). In contrast, scale-free networks are characterised by a small number of highly connected nodes, and hence P(k) decays following a power-law, written as $P(k) \sim k^{-\gamma}$ where γ is a constant (Albert *et al*, 2000) (**Figure 1.1-1a**). A scale-free network topology is appealing since the topology is robust to random node removal, but vulnerable to attack of the most highly connected nodes, the so-called "hubs" (Albert *et al*, 2000). By contrast, the topology of exponential networks is equally vulnerable to random node removal and attacks (Albert *et al*, 2000) (**Figure 1.1-1b**).

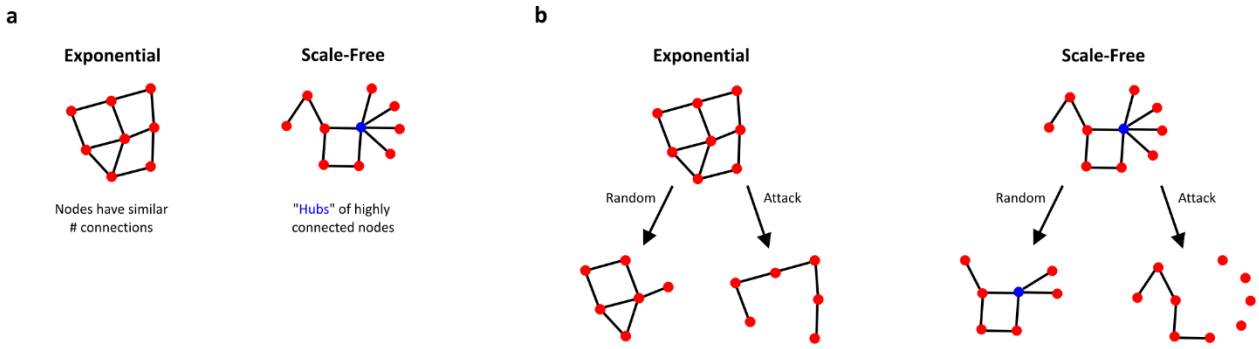



**Figure 1.1-1 Simple exponential and scale-free networks. a**: There are two classes of networks: exponential and scale-free which are discriminated by their node (red e.g protein) connectivity (lines e.g interactions between proteins). **b**: The topology of exponential networks is similar after random node removal (random) and also after the most highly connected nodes are removed first (attack). In contrast, the topology of scale-free networks is robust to random node removal but quickly collapses if the most highly connected nodes are removed first.

In biological terms, this would mean that for a scale-free PIN, most proteins show few connections while the entire network is integrated by a small number of key proteins with a large number of connections. This has important ramifications for targeting disease-causing organisms, where if the key protein "hubs" are known they can be specifically attacked in order to kill the organism (Ideker & Sharan, 2008).

Are biological networks such as PINs scale-free in topology? Jeong *et al* (2000) argues that metabolic networks have scale-free topology based on core metabolic data for 43 different organisms. Jeong *et al* (2001) also argues for a scale-free topology for the PIN of yeast. Both arguments are based on the emergence of an apparent power-law decay for the connectivity distribution. However, as de Silva & Stumpf (2005) note, claims of scale-free topology are based on fitting a power-law distribution to the data without examining competing fits, such as the log-normal distribution. More importantly, understanding the topology of PINs and their applications in drug design is dependent on the quality of the data used to produce the network (de Silva & Stumpf, 2005).

### 1.1.1 Determining Protein-Protein Interactions

The ability of proteins to interact in order to carry out specific functions, such as DNA/RNA or protein synthesis, has many evolutionary advantages over these functions being encoded in a single gene. This includes the promotion of protein stability, reducing transcriptional and translational errors, increasing the likelihood of correct folding, decreasing the probability of an unfavourable interaction, and facilitating the evolution of new functions following gene duplication (Lynch, 2012). In order to infer biological function from a PIN, it is crucial to first determine the protein components of the interaction and assess its likelihood of being correct in the physiological context of the cell.

To arrive at a potential PIN requires determining the protein-protein interactions (PPIs) for the proteome of the organism in question. Either the interactions can be hypothesised based on bioinformatics or determined experimentally. It should be noted that interactions inferred by bioinformatics are not necessarily direct and may have some other relationship, such as the proteins are part of the same enzymatic pathway, which is why the term functionally linked is used to describe such interactions (Eisenberg *et al*, 2000).

To predict PPIs, there are three computational methods: the phylogenetic profile method (Pellegrini *et al*, 1999), the Rosetta stone method (Enright *et al*, 1999), and the gene neighbourhood method (Overbeek *et al*, 1999). The phylogenetic profile method compares the profiles of the presence or absence of specific proteins in different species; if two protein

profiles correlate, then they are said to be functionally linked (Pellegrini *et al*, 1999). The Rosetta stone method compares fused proteins in one organism to possible homologues which are not fused in another organism, in which the unfused proteins are also said to be functionally linked (Enright *et al*, 1999). The gene neighbourhood method compares the position of genes in chromosomes of different organisms; if two genes are always found nearby then their protein products are predicted to be linked (Overbeek *et al*, 1999).

There are many different ways to experimentally determine whether proteins physically interact, although they can be broadly divided into binary or complex (group) interactions. They can further be classed as transient or stable ("permanent") (Levy & Pereira-Leal, 2008). Methods such as yeast two hybrid (Y2H) (Fields & Song, 1989) and fluorescence resonance energy transfer (FRET) (e.g Gordon *et al*, 1998) test whether pairs of proteins interact, while methods such as co-immunoprecipitation (e.g Free *et al*, 2009) and tandem-affinity purification (TAP) (Rigaut *et al*, 1999) test whether multiple proteins are in complex. These methods vary in sensitivity, specificity, and interaction strength (**Table 1.1.1-1**).

**Table 1.1.1-1 Some methods for studying protein-protein interactions**

| Method | Binary or Complex? | Interaction Strength[1] | Transient and/or Stable? | Reference |
|---|---|---|---|---|
| **Yeast Two Hybrid (Y2H)** | Binary | ~ 10 – 100 µM | Transient and Stable | Mackay *et al* (2007a) |
| **Co-immunoprecipitation** | Binary | N/A | Stable | Mackay *et al* (2007a) |
| **Glutathione-*S*-transferase (GST)-pull down** | Complex | ~ 10 nM | Stable | Mackay *et al* (2007a) |
| **Fluorescence Resonance Energy Transfer (FRET)** | Binary | ~ 1 – 10 µM<br>~ 0.01 – 10 mM | Transient and Stable | Martin *et al* (2008)<br>Margineanu *et al* (2016) |
| **Tandem-affinity purification (TAP)** | Complex | Mid-nM range | Stable | Oeffinger (2012) |
| **Chemical Cross Linking** | Complex | Not suitable for low affinity complexes (> 25 µM) | Transient and Stable | Mädler *et al* (2010) |
| **NMR Spectroscopy** | Binary | 0.1 – 1 mM | Transient | Vaynberg & Qin (2006) |

1: Based on minimum dissociation constant ($K_d$)

Any experimental method which seeks to determine PPIs has to account for both false positives and false negatives. A false positive is when an interaction is experimentally determined to occur which does not exist in the cell. In contrast, a false negative is when an interaction is experimentally determined not to occur which does exist in the cell. The unreliability of Y2H results is well-known (e.g Deane *et al*, 2002; Deeds *et al*, 2006); for example, out of approximately 8063 interactions uncovered by Y2H, around 1400 of those are likely to be correct (Deane *et al*, 2002). This has led to the suggestion by Deeds *et al* (2006) that most interactions uncovered by Y2H are nonspecific.

Mackay *et al* (2007a) note that it is dangerously easy to conclude if proteins interact, since these conclusions are based on the requirements of biological plausibility, protein co-expression, and confirmation by an experimental method such as glutathione-*S*-transferase (GST)-pull down. Their own experience suggests that only half of the reported interactions could be validated (Mackay *et al*, 2007a). Others have also reported similar validation results (e.g Deane *et al*, 2002; Bader *et al*, 2004; Tong *et al*, 2004). Most of the controversy (e.g Mackay *et al*, 2007a; Chatr-aryamontri *et al*, 2007; Mackay *et al*, 2007b) seems to stem from the fact that there is no set criteria to single out the possible false positives and negatives in an interaction dataset. Does one use co-expression data in conjunction with comparing paralogous interactions as completed by Deane *et al* (2002)? What if there is no data for the paralogous interactions? Or should one use their own scoring function (e.g Gavin *et al*, 2006)? Most would agree that it is best to use interaction data from multiple experiments from which to base tenuous conclusions (e.g Gavin & Furga, 2003; Titz *et al*, 2004).

### 1.1.2 "Interactomics": Is Bigger Better?

"Interactomics" is another name for the large-scale, or high-throughput, study of PPIs. High-throughput application of Y2H has been used to build PINs for various species (e.g Uetz *et al*, 2000; Ito *et al*, 2001; Rain *et al*, 2001; Li *et al*, 2004; Parrish *et al*, 2007). Since Y2H only gives data for binary protein interactions, an alternative method, known as tandem-affinity purification (TAP), was explored for purifying protein complexes which consist of more than two proteins and are expressed at natural levels (Rigaut *et al*, 1999).

Gavin *et al* (2002) applied TAP to analyse the yeast proteome. A purification tag was attached to 1739 yeast (*Saccharomyces cerevisiae*) genes which were inserted into the yeast chromosome by homologous recombination. Any interacting partners which form stable enough complexes could then be identified by mass-spectrometry (MS) of bands from an SDS-PAGE gel of the purified complexes. This process lead to 589 purified tagged proteins, producing 98 previously identified complexes and 134 novel complexes (Gavin *et al*, 2002). These "non-binary" interactions produce a different picture to those produced by Y2H assays (Gavin *et al*, 2006). Gavin *et al* (2006) applied TAP-MS to all 6466 yeast ORFs of which 1993 ORFs were successfully purified and 88% of these bound to at least one partner. This allowed them to build a picture of the yeast proteome complexes as composed of "core components" (stable PPIs) and "attachments" and "modules" which are comprised of proteins which interact with the core depending on the function of the complex (Gavin *et al*, 2006). We must be cautious when interpreting the results of such large-scale PPI data, since the PIN observed experimentally is heavily influenced by the presence of noise and the fact that they represent a small sample of the entire cell's PIN (e.g Stumpf & Wiuf, 2012).

TAP-MS has been successfully applied to a range of organisms (e.g Butland *et al*, 2005; Krogan *et al*, 2006; Kühner *et al*, 2009). Other methods have been explored which do not require tagging. For example, Havugimana *et al* (2012) identified human soluble protein complexes through purification without a tag in conjunction with identification by MS.

Although one does see the appeal in such large-scale studies, are they more reliable than the known problems with Y2H? Utilising the correct inference is critical in any study considering a null with an alternative hypothesis (Rouder *et al*, 2016). PPI data are implicitly comparing the two hypotheses:

**Null**: There is no interaction

**Alternative**: There is an interaction

As Rouder *et al* (2016) notes, careful consideration of the alternative hypothesis is critical in inferring the correct conclusions. However, the majority of PPI studies are falling down the same logical trap which comes with not specifying a good alternative hypothesis (Rouder *et al*, 2016). In protein interactions, we are not interested in whether or not there is an interaction, but whether or not the interaction is *physiological*. Thus, a null hypothesis must encompass the situation where interactions occur in the lack of a physiological context, such as through aggregation. We can see that although PPI studies can determine whether or not an interaction has occurred, it is much harder to determine if that interaction is physiologically relevant.

### 1.1.3 Seeing Is Believing

Structural data provides compelling evidence for the existence of PPIs which are therefore physiologically relevant to the cell. Edwards *et al* (2002) used the crystal structures of well-known complexes (the protease, RNA polymerase II, and Arp2/3) and first compared them to small-scale PPI experiments completed before these protein complex structures were known. They found that 61% of these small-scale interactions were false positives while 38% were false negatives. In addition, they found that some of the false positives had been 'validated' in other biochemical studies (Edwards *et al*, 2002). Next, they compared the structures to results from small-scale Y2H screens and found that the false negative rate was approximately 43–71%, with it being higher in the two large scale Y2H studies completed by Uetz *et al* (2000) and Ito *et al* (2001). However, the false negative rate for the TAP-MS method (Gavin *et al*, 2002) was relatively low at 15% (Edwards *et al*, 2002).

Under the guise of 'seeing is believing' (Mackay *et al*, 2007a), structure-based methods offers a powerful approach for determining PPIs which are likely to be physiologically relevant to the cell.

### 1.1.4 Tuberculosis: A Health Crisis

Tuberculosis (TB) is an important problem in South Africa as it is the single greatest contributor to mortality and causes of death (Stats SA, 2014). The causative agent of TB, *Mycobacterium tuberculosis* (*Mtb*), is notoriously difficult to kill with current treatments, usually requiring around six months of antibiotic compliance for drug susceptible *Mtb* (Chan, 2002). More worryingly, the spread of multi- (MDR) and extensively- (XDR) drug resistant TB poses a great threat to public health, with most cases occurring in the Eastern Cape, the Western Cape, and KwaZulu Natal. The Eastern Cape saw their cases of particularly dramatic 2.2 fold rise in MDR and XDR cases during the 2006–2009 period (Klopper et al, 2013).

As Lienhardt (2014) urges, "…without continued studies into the molecular nature of TB, no new interventions will become available to health-care professionals." Thus, it is apparent that searching for new effective drugs against TB and understanding its biology is critical.

Computational approaches have been used to identify potential drug targets for *Mtb* in the context of its PIN (Mazandu & Mulder, 2011). However, appealing drug targets were based on their belonging to a scale-free topology, without considering alternative models (e.g Hase *et al*, 2009) for the data.

### 1.1.5 High-Throughput Determination of Complex Structures

Current crystal structures in the Protein Data Bank (PDB) for *Mtb* are biased towards monomers and dimers (**Figure 1.1.5-1**). This contrasts heavily with the picture of a cell comprised of "large protein machines" (Alberts, 1998). In general, crystal structures in the PDB contain more homomers and monomers than heteromers, whereas the opposite is true for structures solved by electron microscopy (EM) (Marsh & Teichmann, 2015). This can be seen as well for the structures available for *Mtb* (**Figure 1.1.5-1**).

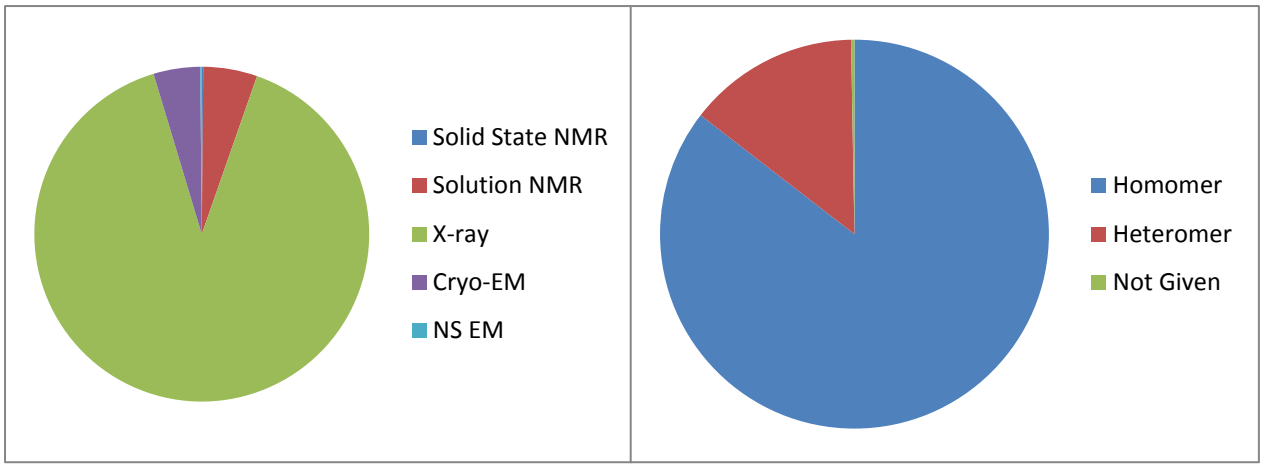



**Figure 1.1.5-1. *Mtb* structures per ORF.** The majority (90%) of *Mtb* structures have been solved by X-ray crystallography, with the remaining 10% solved by nuclear magnetic resonance (NMR) spectroscopy, and negative stain (NS) and cryo-EM (left). Very few (14%) heteromeric structures have been solved, likely as a result of the majority of structures having been solved by X-ray crystallography (right). Data extracted from the PDB as of November 2017. The same biases exist for the available *Msm* structures (see Figure 8-1 in Appendix). A further break-down of the homomer and heteromer composition is available in Figures 8-2 and 8-3 in the Appendix. Note that the data is given for each ORF not per structure.

The crystallisation process in X-ray crystallography presents the major bottleneck for determining atomic resolution structures (Callaway, 2015). Furthermore, large amounts of pure protein is required, typically 1–10 mg (Wlodawer et al, 2013). This is the main reason why high-throughput structure determination methods for protein complexes purified at native concentrations have used single particle EM, since the method only requires as little as 1 µg of protein (Grassucci *et al*, 2007), and is not dependent on producing diffracting crystals (Aloy *et al*, 2004; Han *et al*, 2009).

Aloy *et al* (2004) used TAP in conjunction with low-resolution single particle EM and homology modelling to build models for protein complexes. However, their EM models were not of sufficient resolution to offer validation for the predicted interactions. Han *et al* (2009) purified to near homogeny fifteen protein complexes from the wild type organism *Desulfovibrio vulgaris*, of which eight could be reconstructed by single particle EM. Two of the structures had novel folds which could not be homology modelled (Han *et al*, 2009).

Single particle EM is a much better technique to study large macromolecules, given the difficulties of crystallization, but only eight structures were solved for *Mtb* with this technique (small ribosomal subunit (emd-8646), large ribosomal subunit (emd-8649, emd-8641), 70S ribosome (emd-8648, emd-8645), fatty acid synthase I (emd-2357, emd-2358, emd-2359), 50S ribosome (emd-6177), EspB (emd-6120), the bacterial proteosome activator Bpa (emd-4128), and the heat-shock protein Acr1 (emd-1149)) based on depositions in the Electron Microscopy Databank (EMDB). Although cryo-EM is an excellent technique for studying large macromolecules, only recently has it been able to compete with crystallography for obtaining near-atomic resolutions (Bai *et al*, 2015). More importantly, biochemical purification and grid preparation are still major bottlenecks, requiring time-consuming optimisation for each sample. These factors can explain the current low number of *Mtb* structures solved with this technique. However, despite the challenges involved, cryo-EM remains the only available technique for solving large, and typically heterogeneous, structures (Fernandez-Leiro & Scheres, 2016).

This offers an opportunity to utilise single particle EM, in conjunction with a high-throughput purification strategy, to study Mycobacterial protein complexes. It is useful to think of physiological transient and stable interactions as existing in a continuum of interaction strength, with experimental ambiguity as to where the one ends and the other begins (**Figure 1.1.5-2**). Single particle EM, for this study, is better suited towards studying more stable interactions (**Figure 1.1.5-2**) given that a protein complex must stay intact throughout purification.

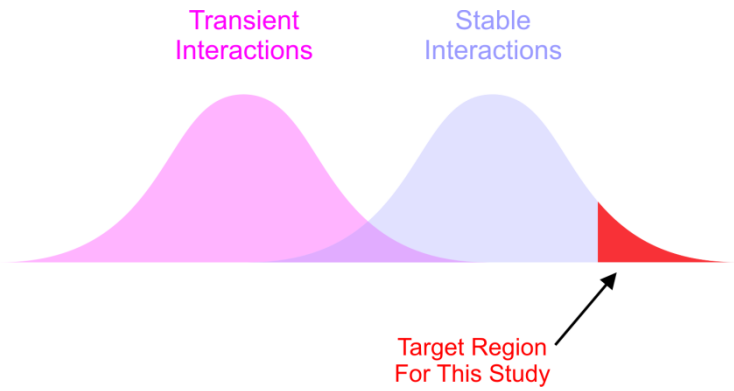



**Figure 1.1.5-2**. **Hypothetical physiological transient and stable interactions obtained experimentally**. Measurements of interaction strength (usually expressed in terms of the dissociation constant, $K_d$), increasing from left to right, impose some ambiguity over the distinction between transient (pink) and stable (blue) interactions. For example, an experimental interaction classed transient may in fact be stable in the cell, while an interaction classed as stable may be transient in the cell. For this study, we will be examining the very stable interactions which are also likely to be very stable in the cell (red).

By using the close relative of *Mtb*, *Mycobacterium smegmatis* (*Msm*), this study will aid in our understanding of the structural underpinnings of Mycobacterial PPIs which can potentially be exploited for drug targets.

### 1.1.6 General Strategy

As mentioned previously, current non-structure based methods of determining PPIs suffer from the lack of a standard solution to distinguish false-positive results from real interactions. Structure-based methods for determing PPIs which utilise X-ray crystallography are also not particulalrly suited for large, complex structures and typically require substantial amounts of purified protein. Relatively recent approaches in using single particle EM for structure-based PPIs have also focused on obtaining near-homogenous samples (e.g Han *et al*, 2009; Kastritis *et al*, 2017), in order to simplify the identification procedure of the protein constitutents for the complex. However, there has been little development in methods which rely on partial fractionation or alternative methods to chromatographic techniques as a means of obtaining protein complexes in a high-throughput manner. Here, we define high-throughput as

techniques which can be accomplished by a single-user. The success of this approach depends heavily on the strategy employed and the information available for the organism under study.

In the current 'post-genomics' era (e.g James, 1997), the success of high-throughput protein purification and identification strategies relies on available sequence information for each ORF in the organism under examination. The complete genome sequence for *Msm* $mc^2155$ was released in 2006 (Fleischmann *et al*, 2006) and updated in 2015 (Mohan *et al*, 2015). Currently known annotated ORFs are available for *Msm* in the database SmegmaList and for *Mtb* in the database Tuberculist (Kapopoulou *et al*, 2011).

The use of the native organism as a source of proteins has worked effectively for the high-throughput crystallisation of proteins from *Escherichia coli* (Totir *et al*, 2012). Totir *et al* (2012) fractionated 120 L of culture in order to purify and reconstruct 23 structures, four of which were novel, although structures >500 kDa failed to crystalise under the conditions tested. The main advantage of using native proteins is that it avoids cloning and expression of thousands of genes. For example, Christendat *et al* (2000) completed a high-througput crystallisation of proteins from a thermophilic archeon; they found that poor expression and solubility accounted for 60% of their recalcitrant proteins. For recombinantly expressed *Mtb* proteins, a significant degree of optimisation is required to achieve sufficient yield and purity for downstream applications, even when *Msm* is used as the expression host (Milewski *et al*, 2016). However, a disadvantage to purifying from the native organism is the reliance on the natural abundance of the proteins, some of which will be present at low copy number (e.g see Vogel & Marcotte, 2012). This is usually compensated by growing a sufficient amount of starting material such as bacteria in cell culture.

## 1.2 Aims and Objectives

The general strategy is summed in **Figure 1.2-1**. The aims of the project was to:

1) Explore a variety of purification methods in order to capture stable, water-soluble protein complexes from *Msm*.

2) Reconstruct these complexes by low-resolution single-particle EM and identify by LC-MS/MS.

3) Develop hypotheses with regards to the biological function of any interesting protein complex(es) captured. Here interesting is defined as a protein complex which possesses drug target potential or is physiologically critical under certain environmental conditions (e.g stress) based on the scientific literature.

This was achieved by:

- Investigating partial biochemical fractionation as the first high-throughput purification technique as well as various methods of identification by LC-MS/MS (Chapter II)
- Examining the use of grid blotting in combination with blue native PAGE as a potentially faster method of purification and reconstruction (Chapter III)
- Completing cryo-EM on an interesting purified protein complex in order to optimize conditions required to obtain a high-resolution structure (Chapter IV)
- Using low-resolution structural information in combination with any available crystal structure homologues to make hypotheses with regards to the function of the identified protein complexes (Chapter V)

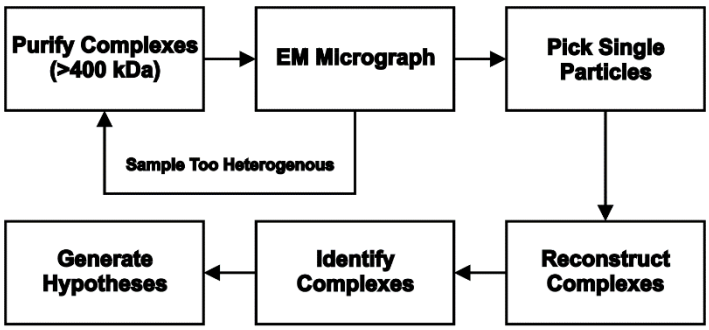



**Figure 1.2-1**. **General strategy for the purification and reconstruction of protein complexes from *Msm*.** The purification of protein complexes were explored through either fractionation (Chapter II), and grid blotting or electro-elution on a blue native PAGE gel (Chapter III). The aim of the purification strategy was to produce a sample which is homogenous enough for reconstruction and identification. Once the identity of the complex is known, hypotheses can be produced as to its function in the cell based on its structure.

# Chapter II: Fractionation

## 2.1 Introduction

Standard biochemical fractionation aims to purify a particular protein target to sufficient homogeneity for a downstream application, such as enzyme analysis or structural determination. In contrast, partial biochemical fractionation aims to reduce the proteome of a target organism to a sufficient degree for a downstream application, such as determining PPIs or solving protein structures in a high-throughput manner. Generally, it is a very successful technique for the purification of multiple proteins (e.g Han *et al*, 2009; Maco *et al*, 2011; Tortir *et al*, 2012; Havugimana *et al*, 2012). For example, Maco *et al* (2011) used sucrose density centrifugation to partially purify protein complexes, based on molecular weight, from mouse macrophages. This yielded 368 unique protein complexes across 29 collected fractions. Although the protein complexes were visualized by single particle EM, the fractions were still too complex to reliably match proteins identified by MS with any putative complexes (Maco *et al*, 2011). Havugimana *et al* (2012) attempted to use multiple biochemical fractionation techniques in combination with MS in order to build a picture of the interaction network for human soluble proteins. They obtained 1,163 fractions and used the co-elution profiles of the identified proteins in order to infer protein interactions; for example, if two proteins were found to co-elute in different purifications they were inferred to interact (Havugimana *et al*, 2012). Of course, this does not necessarily imply that a *direct* interaction is occurring, which is usually validated by cross-linking techniques or, preferably, structural characterization (Edwards *et al*, 2002).

Thus, partial biochemical fractionation seemed an ideal technique to first attempt to purify protein complexes from *Msm*. The strategy followed was one highlighted in **Figure 1.2-1** (see **Chapter I**). The main challenge was to match protein identity with the low-resolution structures obtained. As mentioned previously, Totir *et al* (2012) used partial biochemical fractionation to obtain crystals of varying purity in order to solve low MW (<500 kDa) structures from *E. coli*. They found that only 20% of 23 structures obtained could be identified by MS (Totir *et al*, 2012). The rest were identified through brute-force molecular replacement

using 10,747 structures from the PDB which had >30% sequence identity to an *E. coli* ORF. Furthermore, the 4 novel structures identified underwent further refinement for validation (Totir *et al*, 2012).

Such a strategy is not feasible for low-resolution structures, since the sequence information is not available from the map obtained. However, crystal structure homologues are a powerful tool to solve protein identity since they can reliably be fitted to a low-resolution map. Furthermore, MS/MS data from native- or SDS-PAGE bands can be coupled with low-resolution structural information to reliably match protein identity to the correct complex. The most promising method relied on a correlative approach between the presence or absence of MS/MS peaks and relative abundances of protein complexes derived from the electron microscope through a series of purified fractions.

## 2.2 Materials & Methods

### 2.2.1 Bacterial Growth

A glycerol stock of *Msm* groELΔC (Noens *et al*, 2011) was streaked onto an LB plate and grown over 2 days at 37$^{o}$C. A single colony was used to inoculate a 10 mL starter culture (Middlebrook 7H9 media supplemented with 0.2% glucose, 0.2% glycerol, and 0.05% Tween-80) which was grown for 2 days at 37$^{o}$C with shaking at 120 rpm. The starter culture was then used to inoculate a 1 L culture (Middlebrook 7H9 media supplemented with 0.2% glucose, 0.2% glycerol, and 0.05% Tween-80) which was then grown at 37$^{o}$C with shaking at 120 rpm to the end of stationary phase (~4–5 days). Cells were harvested through centrifugation at 4000g (Beckman, California, USA) for 30 minutes at 4$^{o}$C. The pellet was stored at -80$^{o}$C.

### 2.2.2 Cell Lysis and Ammonium Sulphate Precipitation

The pellet was thawed and resuspended in 25 mL of lysis buffer (50 mM Tris-HCl, 300 mM NaCl, pH 7.2) with protease inhibitor cocktail (Sigma-Aldrich, Missouri, USA). Cells were lysed through 4 x (15 seconds on, 15 seconds off for 4 minutes) on ice using the MiSonix 3000 Sonicator (Cole-Parmer, USA) at 12 W. The mixture was centrifuged at 20,000g (Beckman,

California, USA) for 1 hour at 4°C to pellet cell debris. The supernatant was filtered using a 0.45 µm filter and kept on ice.

Ammonium sulphate cuts were completed on the filtered supernatant (<40%, 40−50%, 50−60%, and >60%). For each cut, the ammonium sulphate was added slowly on ice with continual stirring and incubated for 30 minutes before centrifuging at 9000g (Beckman, California, USA) for 15 minutes. Pellets were clarified by re-suspending in 20 mL of gel filtration buffer (50 mM Tris-HCl, 200 mM NaCl, pH 8.0) and centrifuged at 20,000g (Beckman, California, USA) for 10 minutes at 4°C. The ammonium sulphate cuts were then buffer exchanged to gel filtration buffer using an Amicon® spin-filter with a 100 kDa cut-off (Merck, Darmstadt, Germany).

### 2.2.3 Anion Exchange

Anion exchange was completed using the 20 mL HiPrep Q FF 16/10 column (GE Healthcare Life Sciences, Massachusetts, USA) on a Gilson chromatography system (USA). The column was equilibrated with 5–10 column volumes of start buffer (20 mM Tris-HCl, 20 mM NaCl, pH 8.0) before loading the sample onto the column. Samples were then eluted with 0.5 M NaCl for 3 column volumes, and afterwards a gradient of 0.5 – 1 M NaCl for 19.5 column volumes. The flow rate was 5 mL/min with 60 fractions collected. Fractions were stored at 4°C.

### 2.2.4 Gel Filtration

Both the PWXL5000 and PWXL6000 columns (Tosoh Biosciences, Tokyo, Japan) were calibrated using standards (Tobacco Mosaic Virus (exclusion volume), thyroglobulin (670 kDa), γ globulin (158 kDa), ovalbumin (44 kDa), myoglobin (17 KDa), vitamin B12 (1.35 kDa), and acetone (inclusion volume)). From these results, it was decided that the PWXL5000 column would be more appropriate. The column was equilibrated with gel filtration buffer (50 mM Tris-HCl, 200 mM NaCl, pH 8.0) and run using the Gilson High Performance Liquid Chromatography system (USA) at a flow rate of 0.5 mL/min for 1 column volume. Fractions were stored at 4°C.

### 2.2.5 Sucrose Cushioning

The method was adapted from Peyret (2015). A cell pellet from a 1 L culture was re-suspended in sodium phosphate buffer (0.1 M sodium phosphate, pH 7.2) with protease cocktail inhibitor (Sigma-Aldrich, Missouri, USA). The cells were lysed and spun-down as completed previously (see above) and the supernatant was filtered using a 0.45 µm filter before being added to a 14 mL SW40 ultracentrifuge tube (Beckman, California, USA). A double cushion consisting of 25% (top layer) and 70% (bottom layer) sucrose made in sodium phosphate buffer was produced using a fine needle underneath the supernatant. The tube was spun at 170,462g for 5 hours using a Beckman L7-65 UItracentrifuge (Beckman, California, USA). The layer just above the 70% cushion was extracted and buffer exchanged to gel filtration buffer using an Amicon® spin-filter with a 100 kDa cut-off (Merck, Darmstadt, Germany).

### 2.2.6 Protein Concentration

Protein concentration was determined using the Nanodrop™2000/2000c spectrophotometer (ThermoFisher, Massachusetts, USA) at a wavelength of 280 nm with 1 AU = 1 mg/mL.

### 2.2.7 Negative Stain Electron Microscopy

Selected purified fractions were concentrated to an appropriate volume (concentration ranged from 0.2 to 0.7 mg/mL). Samples were pipetted onto a glow-discharged (in air) copper grid and washed/stained with 5 rounds of 2% uranyl acetate before being air-dried. Images were taken using the Tecnai F20 transmission electron microscope (Phillips/FEI, Eindhoven, The Netherlands) fitted with a CCD camera (4k x 4k) (GATAN US4000 Ultrascan, USA) at 200 kV under normal dose conditions with a defocus of 2.00 µm at the appropriate magnification. The sampling rate was 2.11– or 3.84 Å/pixel.

### 2.2.8 Class Averages

Class averages for the ammonium sulphate cuts were produced in Appion (Lander *et al*, 2009), a reconstruction pipeline accessed through a web-interface which houses a variety of image processing and reconstruction programs such as ACE2, EMAN, and Spider. The Appion pipeline is designed to speed-up the reconstruction process by allowing users to execute programs in a straight-forward manner with easy to access data output.

Briefly, the Contrast Transfer Function (CTF) was estimated using ACE2 (Carragher & Potter, 2009) and poor images excluded based on the presence of astigmatism, bad staining, or noticeable microscope drift. ACE2 is a re-written version of ACE (Mallick *et al*, 2005) but with the added features of astigmatism estimation and CTF correction using either phase-flipping or a Wiener filter. Particles were picked manually and a stack created with CTF correction with a particle binning of 2 (ACE2 Phaseflip of whole image (Carragher & Potter, 2009)). Since high-resolutions are not accessible through negative stain, it is not necessary to perform amplitude correction during CTF correction and hence a Wiener filter was not applied. A Spider reference-free alignment (Frank *et al*, 1996) was completed, averaging all particles in the stack. Afterwards, Spider Coran classification (Frank *et al*, 1996) was completed using appropriate settings and then K-means clustering was completed using selected eigen images.

### 2.2.9 Reconstruction

All reconstructions were completed in the Appion pipeline (Lander *et al*, 2009). The process was the same for producing the ammonium sulphate cut class averages (see above), except hierarchical clustering was used instead. Particles were binned by a factor of 2 for a sampling of 2.11 Å/pixel. The appropriate number of classes was used to complete an initial reconstruction using EMAN Common Lines (Ludtke *et al*, 1999) with the appropriate symmetry imposed. The model was then refined using EMAN model refinement ((Ludtke *et al*, 1999) for 26 iterations with the appropriate symmetry imposed; 20 iterations was used for GSI. Angular sampling was as follows: 5 iterations of $10^{o}$, 5 iterations of $8^{o}$, 10 iterations of $5^{o}$, and 6 iterations of $3^{o}$. For GSI, the angular sampling was 20 iterations of $5^{o}$.

### 2.2.10 Mass Spectrometry

Samples were sent for MS either to the Blackburn Group (in-solution or in-gel LC-MS/MS) (University of Cape Town, South Africa) or to the Yale MS & Proteomics Resource (in gel LC-MS/MS) (Yale School of Medicine, New Haven, USA). Samples were digested with trypsin and analysed on an LTQ Orbitrap (ThermoScientific, Massachusetts, USA). MS/MS spectra were searched using the Mascot algorithm (Hirosawa *et al*, 1993). Peaks with a charge state of +2 or +3 were located first using a signal-to-noise ratio of >1.2. Potential peaks were screened against the NCBInr or SWISS-PROT (Bairoch & Apweiler, 2000) databases.

### 2.2.11 Identification by Mass Spectrometry

The strategy for coupling MS/MS data to protein structure is given in **Figure 2.2.11-1**. The strategy relies on a protein mixture which is reduced enough in complexity in order to correlate relative abundances of the protein complexes present in the electron micrographs with the presence or absence of MS/MS peaks.

### 2.2.12 Bioinformatics

Low-resolution negative stain structures obtained for Encapsulin and glutamine synthetase I were deposited in the EMDB under the accession codes emd-4175 and emd-4186, respectively.

EM models obtained were imported into UCSF-Chimera (Petterson *et al*, 2004) and set to the correct voxel size based on the sampling and binning factors used in model creation. Crystal structural homologues were manually docked into the low-resolution EM maps and the fit refined using the 'Fit in Map' function available in UCSF-Chimera (Petterson *et al*, 2004).

MW estimates for the unknown protein complexes obtained were completed in UCSF-Chimera (Petterson *et al*, 2007). The model was imported and set to the correct voxel size based on the sampling and binning factors used to create the model; the contour level was adjusted until the model had density within a reasonable range (i.e not too little such that

features started to disappear and not too much such that features were smoothed over). Protein mass (in Da) was calculated for the estimated lower and upper contour level limits using the following calculation: 825 * V, where V is the volume (in $nm^3$) of the model density at the specific contour level. See Erickson (2009) for details on the calculation.

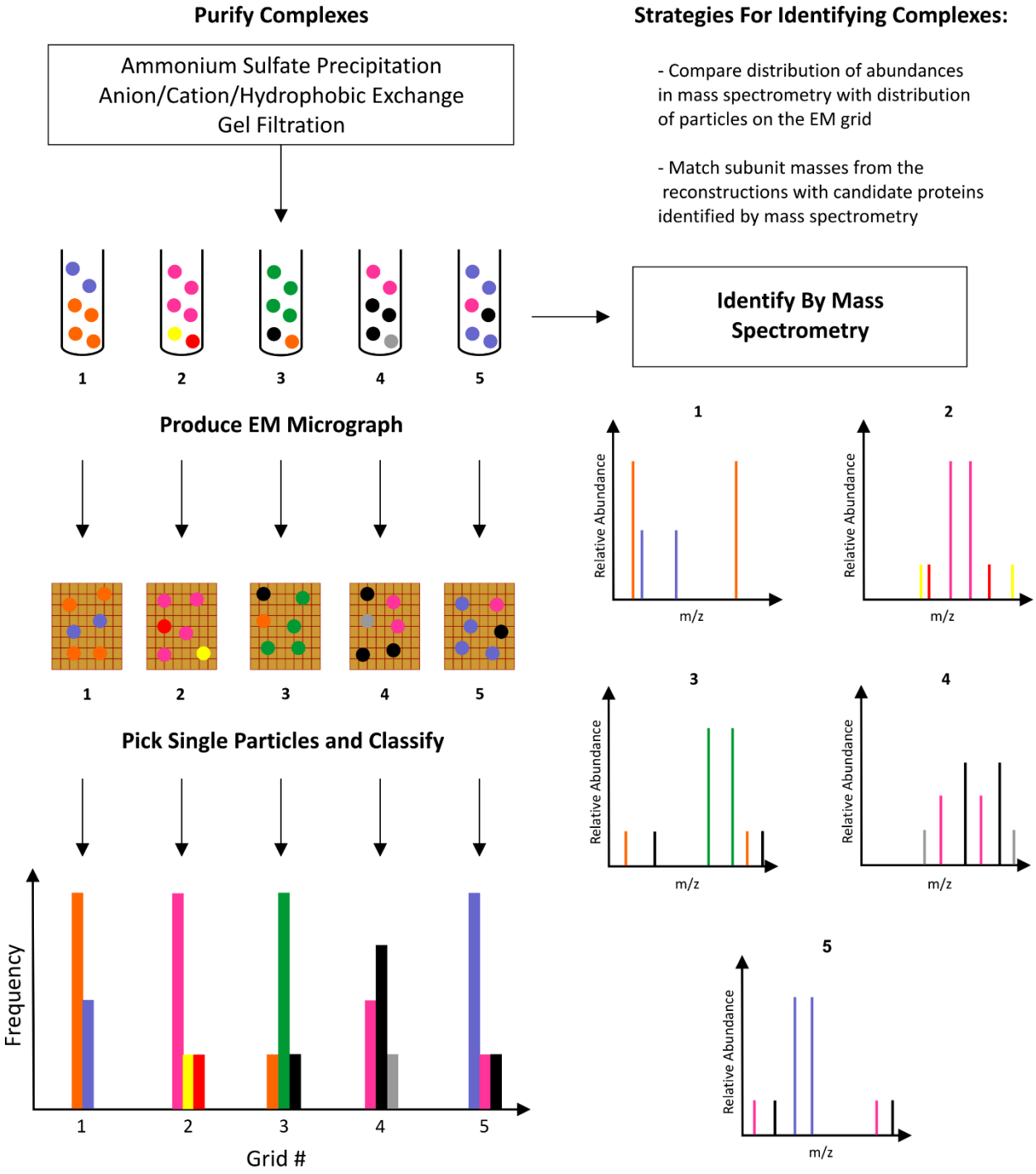



**Figure 2.2.11-1. Strategy for purifying and identifying protein complexes from *Msm.*** Complexes would be purified through different biochemical fractionation steps in order to reduce the complexity of the sample enough to pick and classify single particles as well as identify by LC-MS/MS. Identities of the reconstructed complexes could then be matched by correlating the distribution of the individual complex particle frequencies with that of its presence or absence in MS.

### 2.2.13 Native PAGE

Native PAGE was produced using a continuous Tris-Glycine (pH 8.8) system, where the resolving gel consists of 183−300mM (for 6−15%) Tris-HCl (pH 8.8) and running buffer consists of 25 mM Tris and 192 mM glycine. Non-denaturing sample application buffer was made with 62.5 mM Tris-HCl (pH 6.8), 25% glycerol, and 1% Bromophenol Blue. The running buffer pH was not adjusted. Gels were cast in a Mini Protein 3 Cell (Bio-Rad). Gels were run using pre-cooled running buffer to minimize chance of protein denaturation during the run. Gels were visualized by Acqua stain (Bulldog Bio, New Hampshire, USA).

### 2.2.14 SDS-PAGE

An 8−15% gradient SDS-PAGE (Laemmli, 1970) gel was made by introducing an air-bubble into a pipette containing the 8% (top layer) and 15% (bottom layer) gel mixtures (https://www.youtube.com/watch?v=zu5a-kpMK8k, last accessed February 2018); this was carefully poured into a Mini-Protean 3 cell (Bio-Rad, California, USA). The gel was visualised using a Pierce® Silver Stain for Mass Spectrometry kit (ThermoFisher, Massachusetts, USA). Molecular weight estimates were made using a pre-stained molecular weight marker (New England Biolabs, Massachusetts, USA).

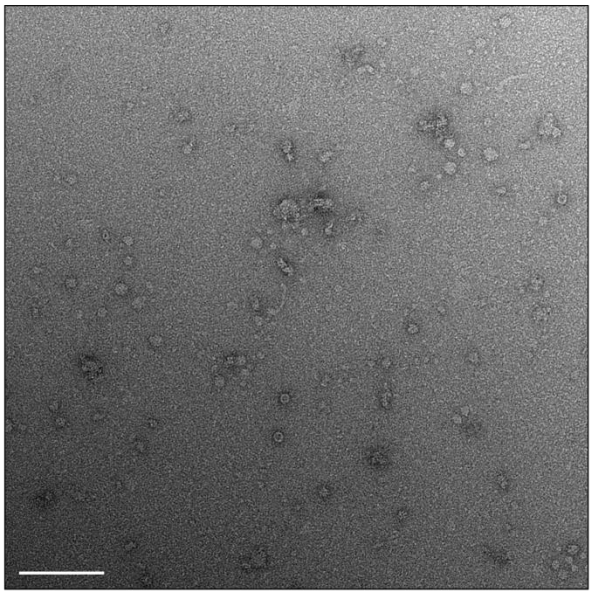
## 2.3 Results & Discussion

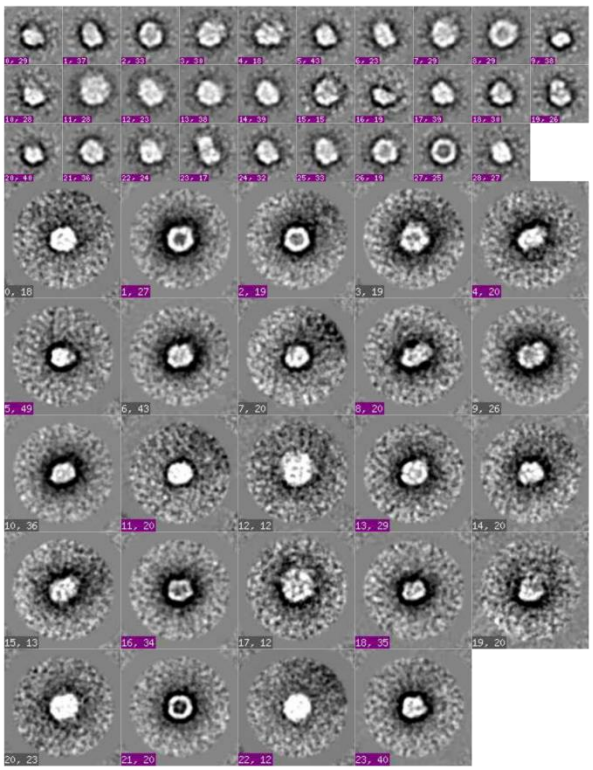
### 2.3.1 Bulk Purification

*Msm* cell culture was exposed to stress by growing to the end of stationary phase. There is evidence that the *Msm* response to stationary phase stress causes the bacteria to become more resistant to other types of stresses, including the oxidative stress response (Smeulders *et al*, 1999), potentially allowing for the purification of protein complexes involved. As a first initial crude purification, *Msm* cell lysate was subjected to four cuts of ammonium sulphate precipitation: <40%, 40−50%, 50−60%, and >60%. The resulting fractions were visualized by negative stain EM and class averages were obtained in order to assess the degree of structural diversity present (**Figure 2.3.1-1**). A class average is composed of a number of aligned particles

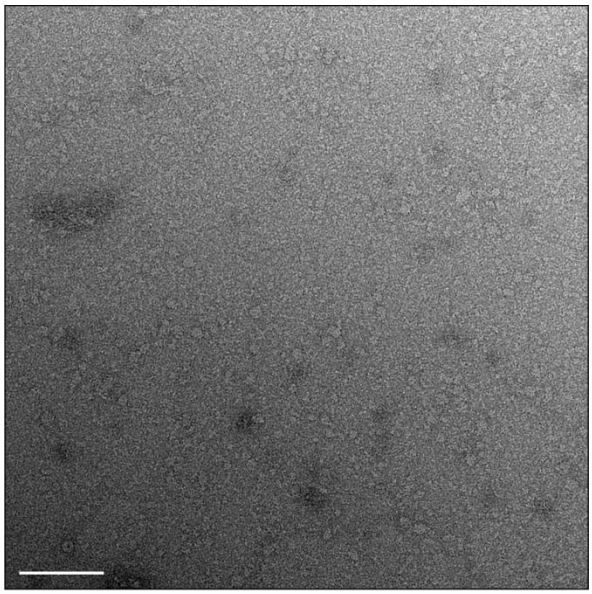
which have similar features for a particular projection/orientation (Frank, 2006). As can be seen in **Figure 2.3.1-1**, a wide-variety of protein complexes appears to be present based on the differing sizes and shapes of the class averages. There does appear to be some bias towards more circular shaped structures, but this may be a result of manual picking and processing of the data. More importantly, as can be seen from these class averages, it becomes difficult to discern particles from different protein complexes and those of different orientations from the same protein complex.

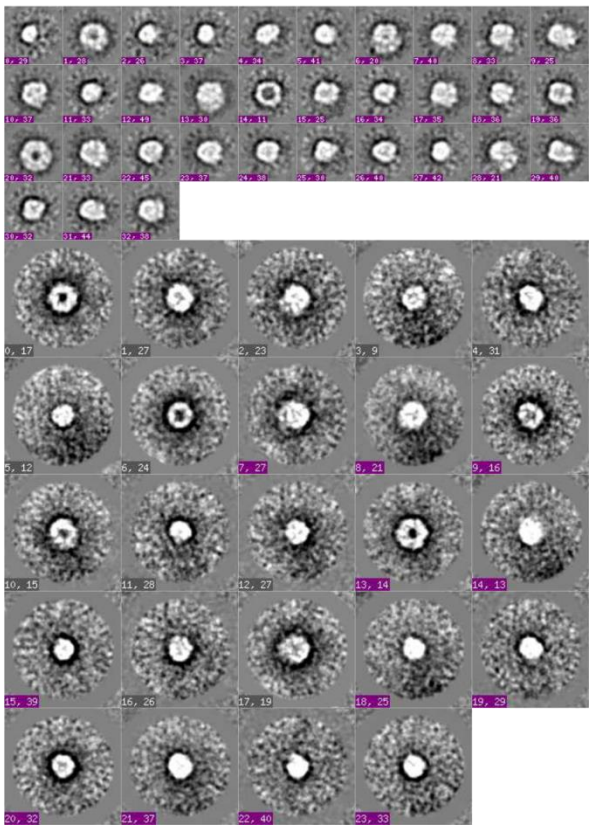
**a** <40%



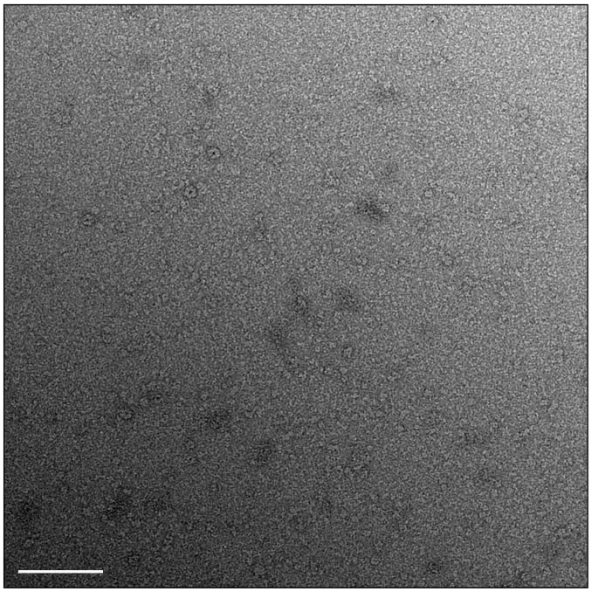


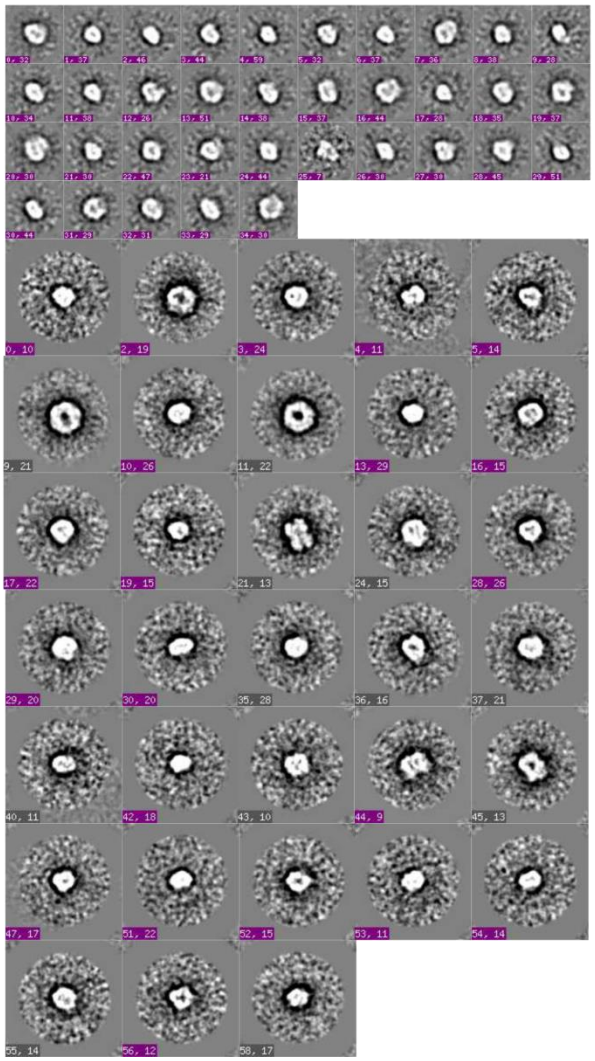
**b** 40–50%



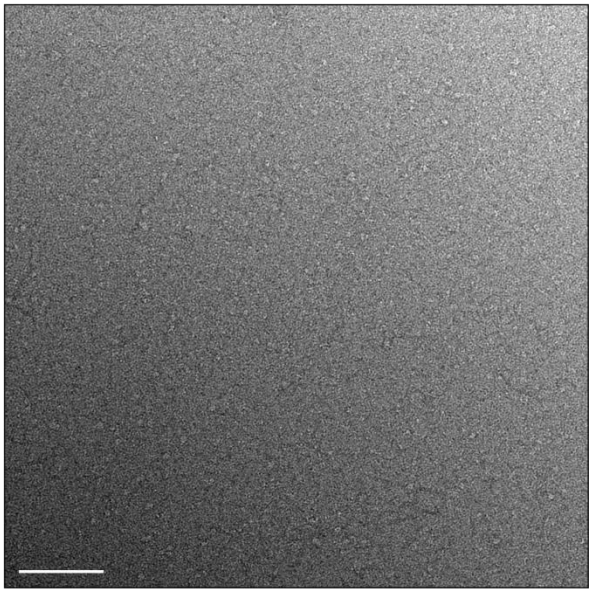


**c** 50–60%





**d** >60%



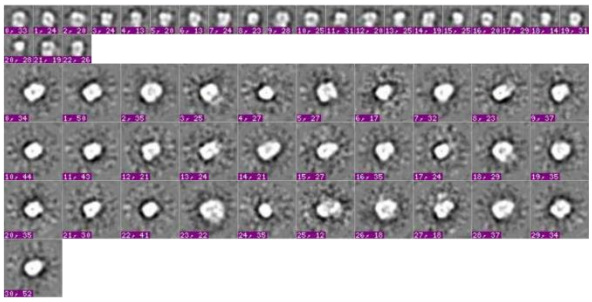


**Figure 2.3.1-1 (previous page). Diversity of protein complexes in *Msm*.** Cell lysate was fractionated by a) <40%, b) 40−50%, c) 50−60%, and d) >60% ammonium sulphate cuts (top row). Particles were picked and assigned to class averages using multivariate statistics through the processing pipeline Appion (bottom row) (Lander *et al*, 2009). Images were taken at x50,000 magnification at a defocus of 2.00 µm using an F20 Tecnai TEM. Scale bars (white) show 100 nm.

A similar method was employed by Maco *et al* (2011) to purify and visualize protein complexes in mouse macrophages using sucrose density centrifugation as a size filter. As mentioned previously, they could not reliably match protein identities found for SDS-PAGE bands of their 29 collected fractions with their class averages obtained from the same fractions. However, they could reasonably identify the presence of the 20S proteosome complex and the small ribosomal subunit. This was based on selecting particles for these complexes to produce a reconstruction of the electron density which reliably matched the input class averages. Evidently, for a reconstruction to be correct, the projections of the model must match closely to the input projections derived from the particle data (Frank, 2006). However, such self-consistency is not sufficient in itself to determine if the resulting model is correct (Frank, 2006). Methods for determining the correctness of an EM model include comparing projections of the model created from untitled particles to those obtained by tilted-particle projections not used in model creation, or comparing the model to one obtained by X-ray crystallography (Frank, 2006).

The feasibility of the approach taken by Maco *et al* (2011) rests on the fact that the structures of these complexes are very well-known and conserved across species (Tanaka, 2009; Melnikov *et al*, 2012), and hence self-consistency of the resulting models is sufficient to make identification. As can be seen in **Figure 2.3.1-1**, a small sample of particles was obtained for each class average. For a successful reconstruction to be attempted, at least ten times the amount of data (as a rough estimate) would need to be required to achieve sufficient orientation sampling (Frank, 2006). There exists computational algorithms for making multiple models when structural heterogeneity is present in the data set, for example the protein complex exists in more than one conformational state or associates with different subunits (e.g Elad *et al*, 2007; Shatsky *et al*, 2010; Elmund & Elmund, 2012). However, it is not clear whether these algorithms would be suitable for reconstructing multiple single-particle

models for different protein complexes, some of which may have similar orientations and hence misclassification of particles poses a significant problem. This problem of making multiple models for different protein complexes from a single dataset was beyond the scope of this work and hence not attempted.

Ammonium sulphate precipitation acts as a crude fractionation step and hence it is expected that this would bias the resulting class averages obtained towards the most abundant complexes present in the cell. However, these protein complexes are most likely to already have been structurally characterized in *Msm*. To obtain rarer protein complexes, more discriminating fractionation methods need to be applied.

Size exclusion chromatography (gel filtration) was implemented by Kastritis *et al* (2017) in order to separate protein complexes in a "single-step" purification. This method has a much higher ability to resolve protein complexes than a crude purification step such as ammonium sulphate precipitation. Gel filtration, although useful as a "cleaning up" step in protein purification, based on its ability to separate by size, the technique has a much lower ability to resolve proteins than other chromatographic methods (Ó'Fágáin *et al*, 2011). For this reason, as a first step, chromatographic techniques which rely on protein binding are preferable when protein complexes are present in low abundance (Ó'Fágáin *et al*, 2011). Hence, anion exchange purification was performed and the resulting peaks analysed by EM (**Figure 2.3.1-2**).

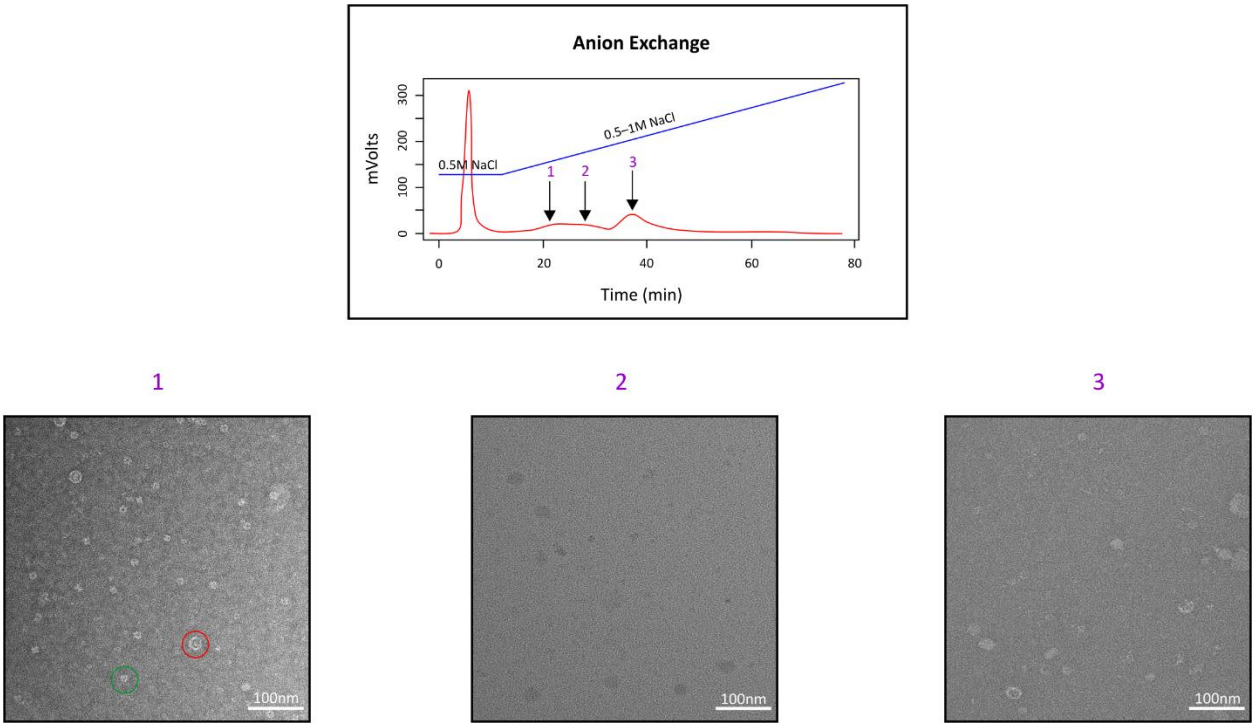
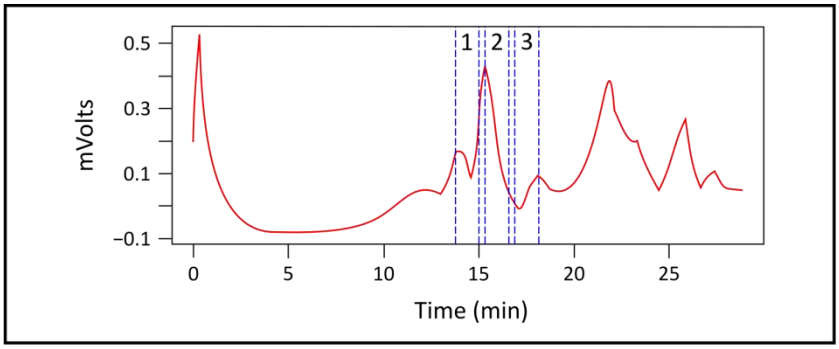
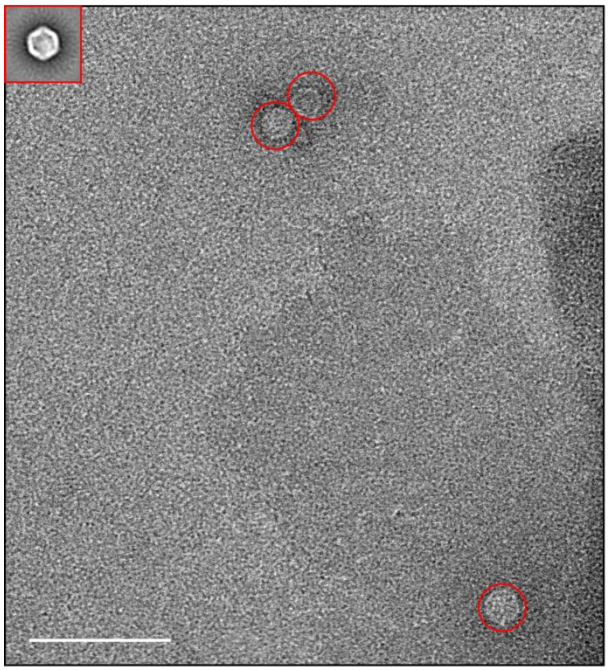



**Figure 2.3.1-2. Fractionation using anion exchange**. Three peaks were recovered from an increasing NaCl gradient (numbered, purple). Inspection by an electron micrograph showed that peak 1 (fractions #15–19) contained two putative complexes (circled red and green respectively). Peak 2 (fractions #20–24) showed no protein complexes while peak 3 (fractions #26–34) contained aggregates. Electron micrographs were taken at a magnification of x80,000 with a 2.00 µm defocus on an F20 Tecnai TEM.

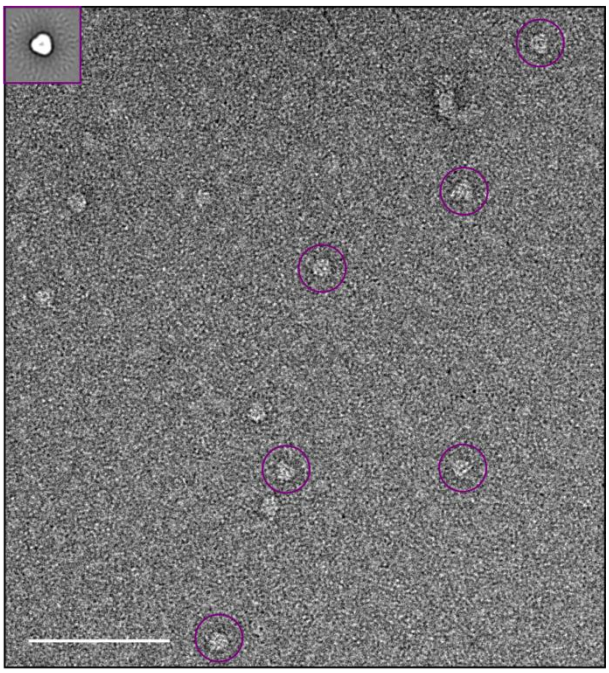
As can be seen in **Figure 2.3.1-2**, three peaks were separated by a gradient in anion exchange. The first peak looked to be the most promising as judged by the electron micrographs; here, two distinct protein complexes appear to be present. Peak 2 showed no protein complexes; it is possible that this peak contained a mixture of small proteins (<100 kDa) which could have been lost while filtering and larger proteins (~100–200 kDa) that are too small to be distinguished from the background carbon on the electron micrograph. Peak 3 appeared to contain mostly aggregates. Hence, Peak 1 was retained for further analysis by gel filtration. This resulted in the appearance of a third distinct protein complex (**Figure 2.3.1-3**).




1



2



3

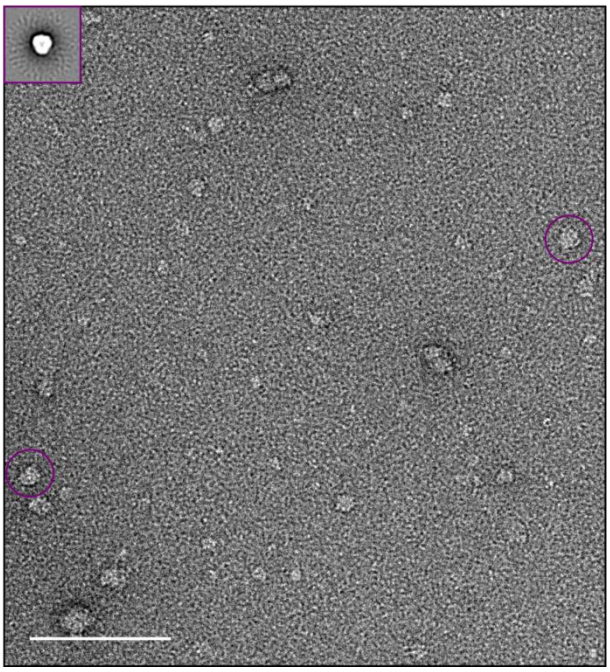


**Figure 2.3.1-3 (previous page). Fractionation using gel filtration after anion exchange.** Three fractions (1–3) were examined from gel filtration of peak 1 from anion exchange (see **Figure 2.3.1-2**). Fraction 1 (#44–48) contained the presence of Enc (red, circled), while fractions 2 (#49–53) and 3 (#54–58) contained a triangle-shaped average protein complex (purple, circled). The average for each particle is given in the top left-hand corner. The white scale bar shows 100 nm. Negative stain electron micrographs were taken at a magnification of x50,000 with a defocus of 2.00 µm on an F20 Tecnai TEM.

Sucrose cushioning is a useful technique in the purification of particularly large protein complexes since it was originally used as a gentle method of purifying viruses and virus-like particles (Peyret, 2015). Hence, a modified method utilizing a double sucrose cushion (Peyret, 2015) was applied to *Msm* cell-lysate. This resulted in the identification of the fourth distinct protein complex (**Figure 2.3.1-4**).

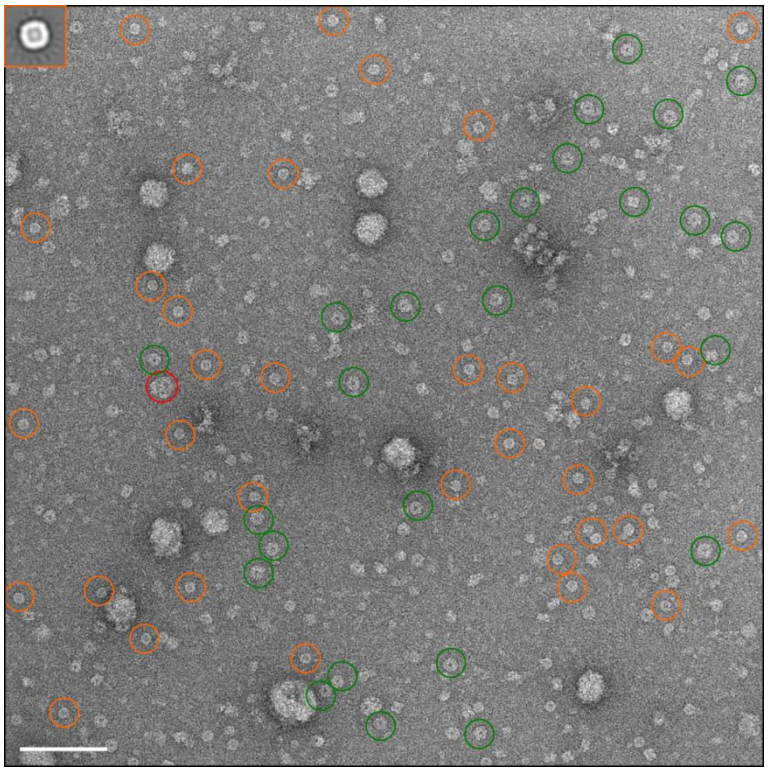


**Figure 2.3.1-4. Fractionation through sucrose cushioning.** BrfB (orange) was purified using a 25% and 70% double sucrose cushion. Also present are likely GSI (green) and Enc (red) particles. Negative stain electron micrograph was taken at x50,000 magnification at a defocus of 2.00 µm on a F20 Tecnai TEM. Scale bar (white) shows 100 nm.

### 2.3.2 Reconstruction Pipeline

Three of the four distinct protein complexes were reconstructed based on the pipeline given in **Figure 2.3.2-1**. These three complexes will later be shown to be: glutamine synthetase I (GSI), Encapsulin (Enc), and bacterioferritin A (BrfA) and/or ferritin-family protein (BrfB). Reconstruction for the one protein complex which showed a triangular-shaped average (**Figure 2.3.1-3**) was abandoned due to biases in orientation (see below). **Figure 2.3.2-1** uses GSI and Enc, both obtained in Peak 1 of anion exchange (**Figure 2.3.1-2**), as examples.

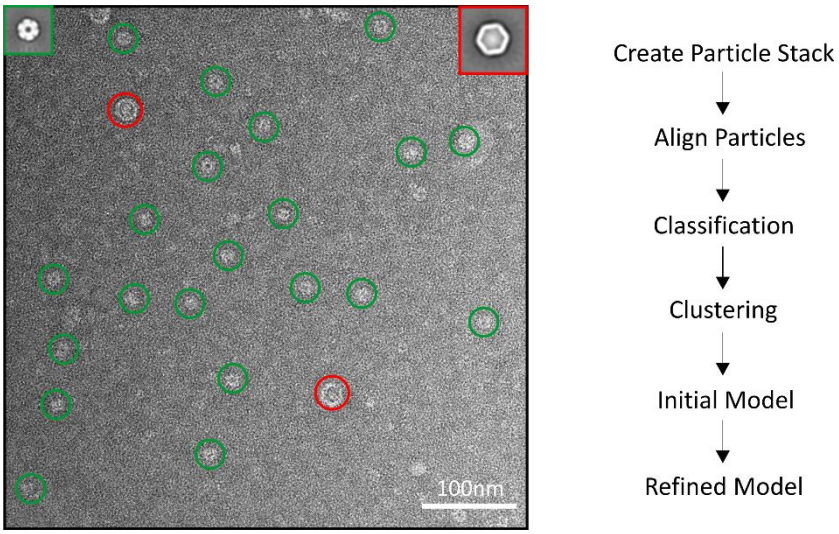



**Figure 2.3.2-1. Reconstruction pipeline.** Particles for the two putative complexes (GSI (green) and Enc (red)) were picked separately and reconstructed in the Appion system (Lander *et al*, 2009) according to the flow shown (right). The average for each putative complex is shown at the top of the electron micrograph: GSI shows a six fold average (green) while Enc shows a hexameric average (red). Further details on the reconstruction are given in the Materials & Methods. Negative stain electron micrograph was taken at a magnification of x80,000 with a 2.00 µm defocus using an F20 Tecnai TEM.

Since each particle is distinct, they can be picked separately which acts as a further *in silico* "purification". GSI (**Figure 2.3.2-1**, green) shows a six-fold symmetry in the averaged stack; this was imposed during initial model creation. Particles for this complex also seemed to show a side view which is present in the initial model. Hence, D6 symmetry was imposed for model refinement. Enc (**Figure 2.3.2-1**, red) showed what is likely to be either icosahedral or

octahedral symmetry in the averaged stack; different initial models were created which imposed either octahedral or icosahedral symmetry (not shown). Both models were refined and the model projections compared to the input class averages. Both models appeared to be self-consistent with the input data (**Figure 2.3.2-2**). As mentioned previously, although self-consistency is required for an EM model to be correct, it is not sufficient in itself. Later structural information showed that Enc is icosahedral (see **2.3.3 Protein Identification Problem**). The triangular-shaped protein complex (**Figure 2.3.1-3**) showed bias towards the end-on view; from the average, this protein complex could either have C3, D3, or tetrahedral symmetry. For C3 symmetry, only end-on views are possible since there are no side views, however this makes reconstruction impossible without tilting the particles to obtain some of the side orientations (Frank, 2006). Due to the uncertainties in particle orientation, this protein complex was not reconstructed or identified. BrfA/B shows a square-like average (**Figure 2.3.1-4**) which could indicate octahedral symmetry and hence this was imposed during initial model creation and refinement. Both *Msm* and *Mtb* have two ferritin-like proteins, BrfA and BrfB, in the genome which have approximately 20% sequence identity. *Msm* BrfA (pdb code 3bkn), and *Mtb* BrfA (pdb code 2wtl) and BrfB (pdb code 3uno) have been reconstructed by X-ray crystallography and all have octahedral symmetry. Since *Msm* BrfB has 72% sequence identity to *Mtb* BrfB, there is a high likelihood that *Msm* BrfB is also octahedral. Hence, at low-resolution BrfA cannot be separated from BrfB if they are both present in the same purified fraction.

The reconstructed protein complexes are given in **Figure 2.3.2-2**. As expected, model projections matched class averages well. For Enc, the similarity between the imposition of icosahedral or octahedral symmetry can be seen (**Figure 2.3.2-2**), emphasizing the need for other methods to determine if an EM model is correct.

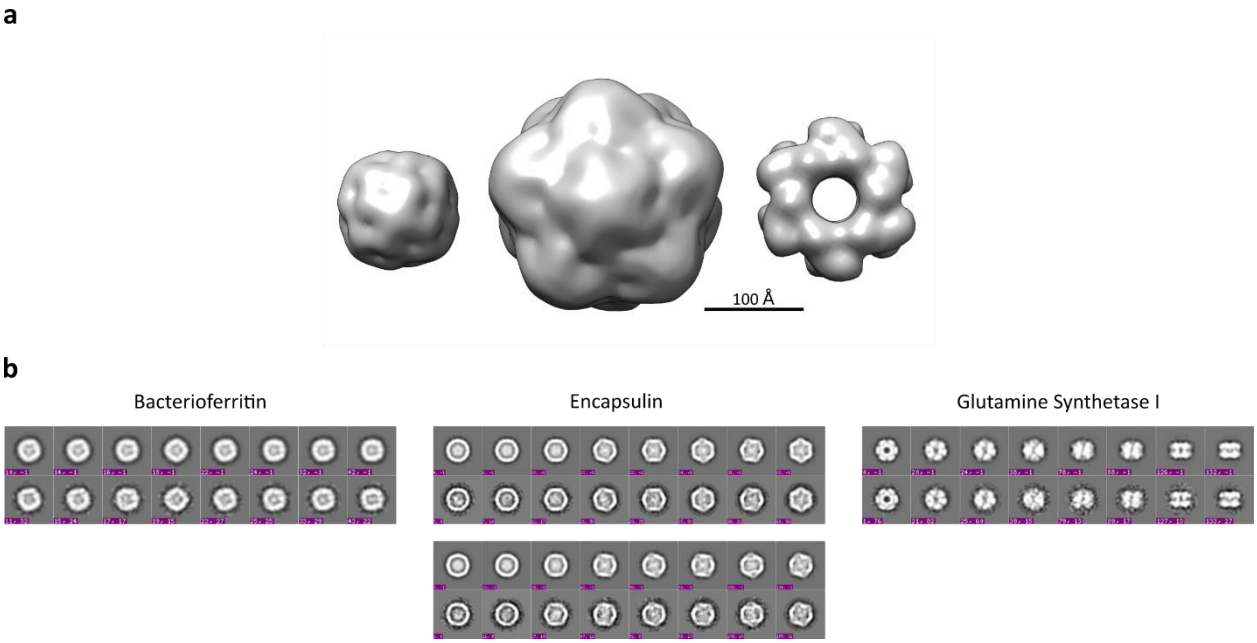



**Figure 2.3.2-2. Model reconstruction and projections.** (a) Refined models of BrfA/B (left), Enc (emd-4175) (middle) and GSI (emd-4186) (right). (b) A sample of pairs of model projections (top panel) and matching class averages (bottom panel). BrfA/B has octahedral symmetry imposed, Enc has icosahedral (top two panels) and octahedral (bottom two panels) symmetry imposed, and GSI has D6 symmetry imposed. All image processing and model creation was completed in Appion (Lander *et al*, 2009), a semi-automated EM reconstruction pipeline. Initial model creation and refinement was completed using EMAN (Ludtke *et al*, 1999).

Once distinct particles can be seen on the electron micrograph, reconstruction is relatively straightforward; the main problem lies in identifying the proteins which constitute the complexes. We shall refer to this as the *identification problem*.

### 2.3.3 Protein Identification Problem

The first method utilised for matching protein identity to structure is highlighted in **Figure 2.2.11-1** (see **Materials & Methods**). Although this technique was promising theoretically, there were several practical problems which rendered the approach unfeasible. First, many small proteins (<200 kDa) are present in the electron micrographs which, due to their size limitations, cannot be reconstructed by negative stain. Secondly, and most importantly, it was

found that a significant proportion of purified proteins were present as aggregates on the electron micrograph. Separating both small proteins and those present in aggregates from intact protein complexes in the "peptide hits" for MS/MS data then becomes a major computational problem. Rather than solving this problem computationally, it is preferable to use other experimental techniques to solve this challenge. This resulted in the coupling of MS/MS data from either native or SDS-PAGE of specific protein bands to the low-resolution structural information of the target complex in order to reliably match protein structure with identity.

Traditional methods for tracking proteins and protein complexes throughout fractionation have relied on SDS and/or native PAGE (e.g Han *et al*, 2009; Maco *et al*, 2011). Other methods have sought to use a correlative approach using MS (e.g Havugimana *et al*, 2012). As we have seen for SDS-PAGE, it is difficult to correlate protein identities to protein structures (i.e Maco *et al*, 2011). In addition, for hetero-complexes the intensity of the constituent proteins will be divided according to their stoichiometry, potentially leading to them being missed. Native PAGE gels have an advantage over SDS-PAGE in that protein complexes are not denatured and hence it can be easier to identify the components of a protein complex by MS/MS. In a native gel, most proteins will run according to their mass if their isoelectric points (pI) are between 3 and 8 since they will be fully deprotonated at the pH of the gel (pH 8.8). However, hydrodynamic size can still have a large influence on how a protein complex migrates through the gel and hence large under- or over-estimates of protein mass can be made; this is usually accounted by utilizing a gradient gel and running it until each protein complex encounters a pore size it cannot enter (Nishizawa *et al*, 1988). The main limitation of native PAGE is that large protein complexes cannot enter the gel matrix. Agarose gels have been used to separate very large protein complexes since larger pore sizes can be made (Righetti, 1989), but this technique suffers from the lack of band resolution (e.g Kim *et al*, 1999).

For these reasons, electron micrographs were the main method of tracking protein complexes throughout fractionation. Since electron micrographs give some information on the relative abundance of a protein complex, it is useful to estimate the expected copy number of a particular sized protein complex. We must make a distinction between protein abundance and protein copy number. Protein copy number is usually defined as the average number of protein particles which exists per cell. Protein abundance, however, is a measure of protein

quantity present in biochemical fractions (Corthals *et al*, 2000). **Figure 2.3.3-1** shows protein copy number as a function of protein concentration. If the protein complex in question is not the most abundant in the cell (and for complexes which are not ribosomes this is not an unreasonable assumption), then the average 1–3 MDa complex will likely have a protein copy number <1000 (**Figure 2.3.3-1**). This imposes limits on the detection by MS/MS for large protein complexes since they will generally be expected to be less abundant than smaller proteins or protein complexes (Fonslow *et al*, 2011).

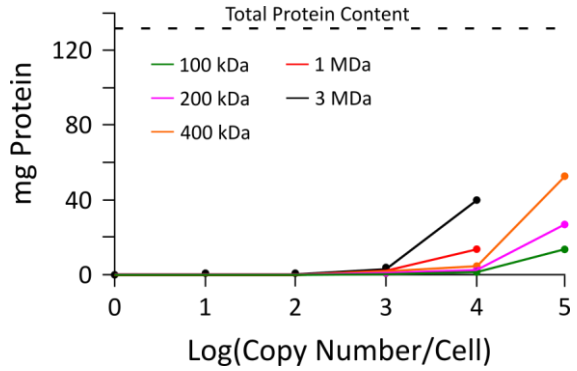



**Figure 2.3.3-1. Theoretical protein copy number as a function of protein concentration**. The amount of protein (mg) is plotted for different sized protein complexes present in different copy numbers for a single cell. Larger protein complexes are present in lower copy numbers compared to smaller protein complexes. Protein amounts for copy numbers that exceed total protein content of a cell are not plotted. Calculations based on ideal dimensions for *E. coli* grown to an $OD_{600}$ of 1.0 (see Appendix for calculation).

Protein abundance becomes critical for reconstructing protein complexes present in low copy number. In standard biochemical fractionation, the protein of interest is enriched relative to other proteins in the cell through a series of well-chosen purification steps, and thus its biochemical abundance is increased even if it is present in low copy number. Likewise, purification steps can be used to purify a mixture of protein complexes which can be visualised by EM even if present at low copy number. However, if the fractionation is too crude, low copy number proteins can be 'crowded out' in terms of abundance, making it difficult to detect and reconstruct. This was observed when using ammonium sulphate precipitation as a crude fractionation step (see **Figure 2.3.1-1**). Enrichment of Enc was

observed after carefully chosen purification steps (see **Figure 2.3.1-2**), even though it appears to be present in low copy number in the cell cytoplasm as it was difficult to detect in the crude ammonium sulphate fractions (data not shown).

It was observed that the electron microscope could detect protein complexes not visible in MS compatible silver stained SDS-PAGE. Since the use of the electron microscope offers an efficient and sensitive tracking technique, it was thought that combining this data with MS/MS data of the same fractions would be a feasible solution to the protein identity problem. Furthermore, the EM reconstructions provided an accurate range for estimation of the molecular weights of the protein complexes, depending on the threshold set (see section **2.2.12** in **Materials & Methods**). A simple equation describes the relationship between protein volume (estimated based on the contour level set) and protein MW (Erickson, 2009), which can be used to calculated a range of possible protein MWs based on reasonable minimum and maximum contour level values that adequately describe the model.

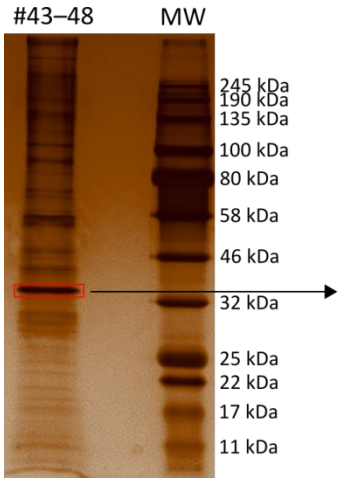
When combined with the knowledge of the symmetry of the protein complex, this can be used as a winnowing tool to eliminate erroneous peptide hits from the MS/MS data. For example, thresholding of the reconstruction of Enc provided MW estimates of 1.7−3.6 MDa, which would correspond to a subunit MW of 71−150 kDa for octahedral symmetry (24 subunits) or 28−60 kDa for icosahedral symmetry (60 subunits). It should be noted that a subunit could consist of more than one protein.

Thus, in-solution LC-MS/MS was conducted on gel filtration fractions where the appearance and disappearance of protein complexes was known by the electron micrographs (see **Figure 2.3.1-3**). However, the MS/MS data for these gel filtration fractions showed no overlapping peptide hits (**Table 8-2**, **Appendix**), in contradiction to the data from the electron micrographs. Furthermore, there were too many peptide hits which could potentially correspond to the subunit MW estimates.

Many large-scale proteomics studies have focused on the sensitivity of identifying proteins by MS/MS, whereby as many peaks from the MS/MS spectra as possible are identified (Cottrell, 2011). This can lead to non-optimal results, especially if the signal-to-noise threshold is set quite low so that “peptide hits” which are incorporated are actually only noise (Wong *et al*, 2010). In addition, optimization of the instrument specificity, whereby the quality of the

MS/MS peaks are considered, is arguably the only requirement one is interested in if the protein complex is present in low abundance. This is usually measured in the false discovery rate (FDR), which estimates the amount of false positives (Cottrell, 2011). However, there is little consideration for measuring the degree of false negatives, which is expected to be more of a problem when proteins are in low abundance (Fonslow *et al*, 2011). Lowering the chance of finding false positives, based on the scoring method used, will always increase the chance of finding false negatives (McHugh & Arthur, 2008), potentially allowing for real peptide hits which corresponds to low abundance proteins to be missed. This could potentially explain why there were no overlapping peptide hits since the electron micrographs showed that there were many other proteins and aggregates present (data not shown).

Due to the many problems faced when using in-solution LC-MS/MS, it was decided to instead use either native- or SDS-PAGE on the purified fractions. LC-MS/MS on native PAGE bands was successfully used to identify Enc in the stacking gel (six unique peptides present) and GSI as a band (12 unique peptides present) in the same native gel (**Table 8-1**, **Appendix**). Since the electron micrographs showed the appearance of a possible cargo for Enc which did not show any reasonable "hits" in the MS/MS data, SDS-PAGE was run on gel filtration fractions which were known to harbour Enc (**Figure 2.3.1-3**). This resulted in the identification of an approximately 40 kDa band, which could either be LppL Protein, ABC Transporter, Saccharopine Dehydrogenase, or DyP-type peroxidase (DyP) (**Figure 2.3.3-2**). Since only DyP agrees with the literature of the known cargoes of Encs (Sutter *et al*, 2008; Contreras *et al*, 2014), this was presumed to be the most likely identity of the band.




| Score | Expectation | Protein ID | Protein Name | MW (kDa) | % Coverage |
|---|---|---|---|---|---|
| 12372 | 0 | A0R0X1 | Major Membrane Protein | 33.706 | 61.2 |
| 1414 | 0 | A0R1H7 | Fatty Acid Synthase | 329.334 | 11.7 |
| 406 | $1.5x10^{-37}$ | A0R4H0 | 29 kDa Antigen Cfp29 | 28.713 | 26.4 |
| 285 | $2.1x10^{-25}$ | A0QYK4 | Sodium:Solute Symporter | 61.554 | 8.2 |
| 147 | $1.5x10^{-11}$ | A0QYY6 | 30S Ribosomal Protein | 53.283 | 8.1 |
| 139 | $8.3x10^{-11}$ | A0QRP7 | TROVE Domain Protein | 61.145 | 8.2 |
| 119 | $8.3x10^{-9}$ | A0QZY7 | LppL Protein | 35.048 | 13.9 |
| 118 | $1.2x10^{-8}$ | A0QRB1 | Amino Acid Carrier Protein | 54.284 | 6.2 |
| 104 | $2.5x10^{-7}$ | A0QT42 | ABC Transporter | 38.744 | 15.3 |
| 102 | $3.9x10^{-7}$ | A0QSL5 | 30S Ribosomal Protein S13 | 14.209 | 21.8 |
| 98 | $1.1x10^{-6}$ | A0R157 | Saccharopine Dehydrogenase | 43.606 | 8.9 |
| 92 | $4.1x10^{-6}$ | A0QZ46 | Proteasome Subunit Alpha | 26.899 | 7.7 |
| 89 | $8.6x10^{-6}$ | A0QT21 | Permease Family Protein | 50.983 | 5.2 |
| 86 | $1.6x10^{-5}$ | A0R4G9 | Dyp-type Peroxidase | 37.193 | 6.4 |

**Figure 2.3.3-2. LC-MS/MS results for the cargo of Enc.** Gel filtration fractions which harbor Enc (#43–48) were run on an 8–15% gradient SDS-PAGE gel. The band harbouring DyP (red box) was cut out and sent for analysis by the Yale MS & Proteomics Resource. MS/MS peaks were analysed using the Mascot search algorithm (table). The score is given as –10log(P) where P is the probability that the match between the experimental mass peak and the database mass is a random event. The expectation value is the number of times we would expect a higher or equal value score by chance. Percentage coverage refers to the number of amino acids which were covered in the protein sequence by the peptide matches. For DyP this means that three peptides matched the protein sequence. Uniprot proteins IDs are given. Molecular weight (MW) marker is shown with corresponding masses. For full MS/MS results see Table 8-3 in Appendix.

Although MS/MS data is useful in confirming suspected protein identities, there are too many superfluous hits which can crowd out the real data (**Figure 2.3.3-2**). For example, many of the protein hits had masses which were excessively too large or too small for the migration of the SDS-PAGE band. In addition, the most abundant protein in the SDS-PAGE band does not necessarily correspond to the greatest number of matching peptides in the MS/MS data since peptide abundance does not always imply protein abundance (Cottrell, 2011). Hence, fitting high-resolution structures into the low resolution maps for these protein complexes aided in matching structure to identity (**Figure 2.3.3-3**). *Msm* GSI has not been solved but *Mtb* GSI is available (pdb code 1hto); fitting of the crystal structure into the low resolution map shows a tight correspondence. This is expected since *Msm* GSI is 84% sequence identical to *Mtb* GSI and hence there is a high chance of conserved quaternary structure (Marsh & Teichmann, 2015). The crystal structure of Enc from *Thermotoga maritima* (pdb code 3dkt), which is 34%

sequence identical to *Msm* Enc, also fits the low-resolution map quite well (**Figure 2.3.3-3**). This is not so surprising since with 30–40% sequence identity, there is a 70% chance of conserved quaternary structure (Marsh & Teichmann, 2015).

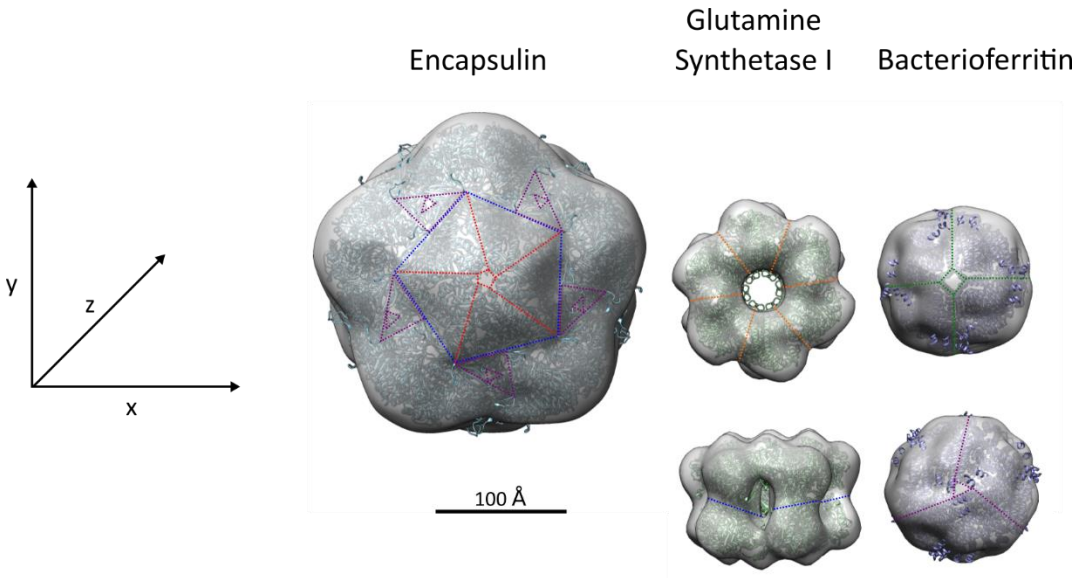



**Figure 2.3.3-3. Fitting of crystal structures into low-resolution maps.** Crystal structures for Enc, GSI, and BrfA/B were obtained from *T. maritima* (pdb code 3dkt) (Sutter *et al*, 2008), *Mtb* (pdb code 1hto) (Gill *et al*, 2002), and BrfB from *Msm* (pdb code 3uno), respectively. The 6-fold (orange), 5-fold (red), 4-fold (green), 3-fold (purple), and 2-fold (blue) symmetry axis are shown as appropriate for each structure. Note that the 2-fold axis for BrfB is not shown.

### 2.3.4 Conclusion

Fractionation is a useful tool for purifying rare protein complexes present in *Msm* cell lysate. The use of anion exchange lead to the purification of GSI and Enc, protein complexes which have not been solved in *Msm*. Furthermore, *Msm* Enc is the first such structure to be solved in Mycobacteria. Sucrose cushioning was also useful as a one-step purification and resulted in the capture of BrfA/B.

Protein identification remains a serious challenge once a structure is obtained through partial biochemical fractionation. Correlating relative abundance of peptide hits obtained by LC-MS/MS with that of the distribution of protein complexes seen on the electron micrographs

in different biochemical fractions was not successful due to a variety of problems. First, no peptide hits were obtained for two of the three biochemical fractions sent for analysis (see **Table 8-2** in **Appendix**) although enough particles were obtained in all three fractions to produce an average of a large and intact protein complex (see **Figure 2.3.1-3**). Secondly, it is not clear how small proteins (<200 kDa) and those present as aggregates could be excluded from the MS/MS data. Thus, protein identification was achieved by MS/MS data of native- or SDS-PAGE gel bands and confirmed by the docking of high-resolution homologues into the obtained low-resolution maps.

Although fractionation is a successful technique in purifying protein complexes it is very time-consuming and laborious to implement. Hence, the use of grid blotting on blue native PAGE was explored in Chapter III as a more high-throughput technique for purifying and identifying protein complexes.

# Chapter III: Blue Native PAGE

## 3.1 Introduction

Native PAGE separates intact proteins based on hydrodynamic size and charge of the protein under the specific running conditions used. Unlike in SDS-PAGE, there is no denaturation of the protein sample (Arndt *et al*, 2012). The technique works well for separating proteins in their native state, usually based on mass for proteins with a pI between 3 and 8 if running at a pH of 8.8. Blue Native (BN) PAGE was originally developed by Schägger & Jagow (1991) as a way of running membrane protein complexes in a native gel. Coomassie G250 is added to the running buffer; the dye binds to the hydrophobic patches of membrane proteins, preventing the protein complex components from dissociating during the run. Since the entire protein is coated with a negatively charged dye, the proteins should run according to their hydrodynamic size rather than their hydrodynamic-size-to-charge ratio, irrespective of their pI. This allows the gel to be run at a neutral pH rather than the alkaline pH required by standard Clear Native (CN) PAGE. The technique is also suitable for water-soluble proteins (Schägger & Jagow, 1991).

Grid blotting was first developed by Knispel *et al* (2012). In its original conception, duplicate proteins are run in a CN-PAGE gel; half of the gel is stained and used as a reference for grid blotting on the unstained portion of the gel, whereby intact particles from the unstained gel then passively diffuse onto an EM grid. Since very small amounts of protein are required for a reasonable representation on the grid (~ 5–10 ng), a reconstruction can easily be obtained. Furthermore, the stained band can be used for identification by MS, thus allowing for an efficient link between the structure determined and the identity of the constituent protein(s) (Knispel *et al*, 2012).

The main disadvantage of the technique is in locating the band, since the reference gel is approximately 5–10% elongated compared to the unstained portion of the gel (Knispel *et al*, 2012). In a complex mixture of proteins, this can seriously hamper accurate location and grid blotting of the correct band without transfer from neighbouring bands. The technique can benefit with the application of BN PAGE in several ways: 1) BN PAGE is run at physiological pH

and this is known to aid in transfer of protein particles (Knispel *et al*, 2012), 2) Complexes will run according to mass rather than hydrodynamic-size-to-charge ratio so the pattern of protein separation is predictable, and 3) Proteins will take up the Coomassie G250 dye without denaturation, allowing for the reliable location of the band for grid blotting without the need for separate reference and blotting lanes.

Grid blotting using BN-PAGE has successfully been tested by Kearns *et al* (2016) on cross-linked p53-DNA complexes. In addition, the complexes were successfully reconstructed to low-resolution using negative stain EM.

In this study, while grid blotting using BN-PAGE was successful with tests on the standard protein GroEL, its application was limited using unknown *Msm* proteins. The main limiting factor was the lack of dye uptake for these unknown proteins, either due to the physio-chemical properties of the proteins which hampered dye uptake, or due to the limited amount of protein present in the gel, or a combination of the two factors.

## 3.2 Materials & Methods

### 3.2.1 Material

Human GroEL was obtained from Sigma-Aldrich (Missouri, USA) at a concentration of 1 mg/mL.

The ammonium sulphate cuts were taken from 20% glycerol stocks thawed from a -20$^{o}$C freezer made as described previously (see section **2.2.2** in **Chapter II Materials & Methods**).

### 3.2.2 Blue Native PAGE

For each grid property tested, an independent BN-PAGE gel was run. Briefly, an 8% resolving gel with a 4% stacking gel was prepared according to standard BN gel procedures (Wittig *et al*, 2006) using a 1% cross-linker. The 6-aminohexanoic acid was omitted from the gel. Samples of GroEL were loaded onto the gel using the same amounts of 1–5 µg found in Knispel *et al* (2012). The gel was run in pre-cooled cathode and anode running buffers according to

the instructions given in Wittig *et al* (2006) until the dye had reached the end of the gel. Samples of GroEL were grid blotted straight after the end of the run.

### 3.2.3 Grid Treatments

To render the grid hydrophilic and negative, the copper grid was glow-discharged in air for 30 seconds. The addition of 5 µl of 5 mM of magnesium acetate to the air glow-discharged grid for 2 minutes was sufficient to make the grid hydrophilic and positive. Grids were glow-discharged in amylamine according to set parameters (Dubochet *et al*, 1982) to make them hydrophobic and positive. For a neutral property, untreated copper grids were used. All grids were treated immediately before grid blotting.

### 3.2.4 Grid Blotting

The grid blotting procedure was completed according to Knispel *et al* (2012). Briefly, the band to be grid blotted was roughened using the tip of a pipette before 5 µl of anode running buffer (25 mM Imidazole, pH 7.0) was added. The treated or untreated grid was then added to band and left for 2 minutes before being stained/washed with 5 rounds of 2% uranyl acetate for negative stain TEM.

### 3.2.5 Electro-elution

Electro-elution was completed using a home-made device produced by Michael and Jeremy Woodward. The device is described in **Results & Discussion**. Electro-elution was conducted using a current of 1.2 V for 10 minutes.

### 3.2.6 Negative Stain Electron Microscopy and Reconstruction

As completed previously (see **Materials & Methods** in **Chapter II**). Reconstruction of GroEL was deposited in the EMDB under the accession code emd-4185.

### 3.2.7 Statistics

Data analysis was conducted in RStudio 1.1.419 (RStudio Team, 2016). Exploratory data analysis for the particle count in each grid property showed that it was non-normal based on the Shapiro-Wilk test for normality (p-value << 0.001). Visualisation of the histograms of the particle count for each grid property showed a positively-skewed distribution. Thus, the particle count data for each grid property was log transformed; the Shapiro-Wilk test confirmed that the log transformed data is likely to be normally distributed (p-value > 0.05). Since the plot for the transfer efficiency of GroEL showed that it is likely to be concentration independent, the Welch two-sample t-test was used to test the hypothesis that there was no difference in transfer efficiency between the grid properties using the log transformed count data from all concentrations; the hypothesis was not rejected for all tests conducted (Hydrophobic Negative vs Hydrophilic Positive (p-value=0.490), Hydrophobic Negative vs Hydrophobic Positive (p-value=0.802), Hydrophilic Positive vs Hydrophobic Positive (p-value=0.899)). Confidence intervals for the mean transfer efficiency for each grid property were produced using the t-distribution on the log transformed data.

## 3.3 Results & Discussion

### 3.3.1 Grid Blotting of GroEL

BN PAGE grid blotting proceeds by a much simpler method (**Figure 3.3.1-1a**) than proposed by Knispel *et al* (2012) due to accurate location of the band. Knispel *et al* (2012) note that parameters such as temperature, extension of the blotting time, and glow-discharge time made no detectable difference in the efficiency of particle transfer. However, the properties of the grid itself on transfer efficiency were not explored. Standard glow-discharge by air renders the grid negatively charged and hydrophilic (Dubochet *et al*, 1982). Coomassie G250

is a hydrophobic and negatively charged molecule at the running pH of 7.0 (Schägger & Jagow, 1991); therefore, it is feasible that standard glow-discharge may not be optimal for the efficient transfer of dye stained particles. Thus, four different properties of the grid were tested for its effect on transfer efficiency of GroEL: hydrophilic and negatively charged, hydrophilic and positively charged, hydrophobic and positively charged, and neutral (no charge).

**Figure 3.3.1-1b** shows the effect of particle transfer based on the properties of the copper grid. No particles were observed for the neutrally charged copper grid. The most striking result is the lack of a correlation between the amount of protein present and the efficiency of transfer. This could potentially be explained by a physical effect, whereby only the GroEL particles at the very surface of the gel can diffuse onto the grid. This effect predicts that any band which is visible, no matter the concentration of protein present, will transfer protein particles at equal efficiency.

There was no statistical difference detected for the mean transfer efficiency between the grid properties. Thus, the main factor underlying the transfer of grid blotted particles appears to be whether or not the grid is charged. The main statistics for each grid property is given in **Table 3.3.1-1**.

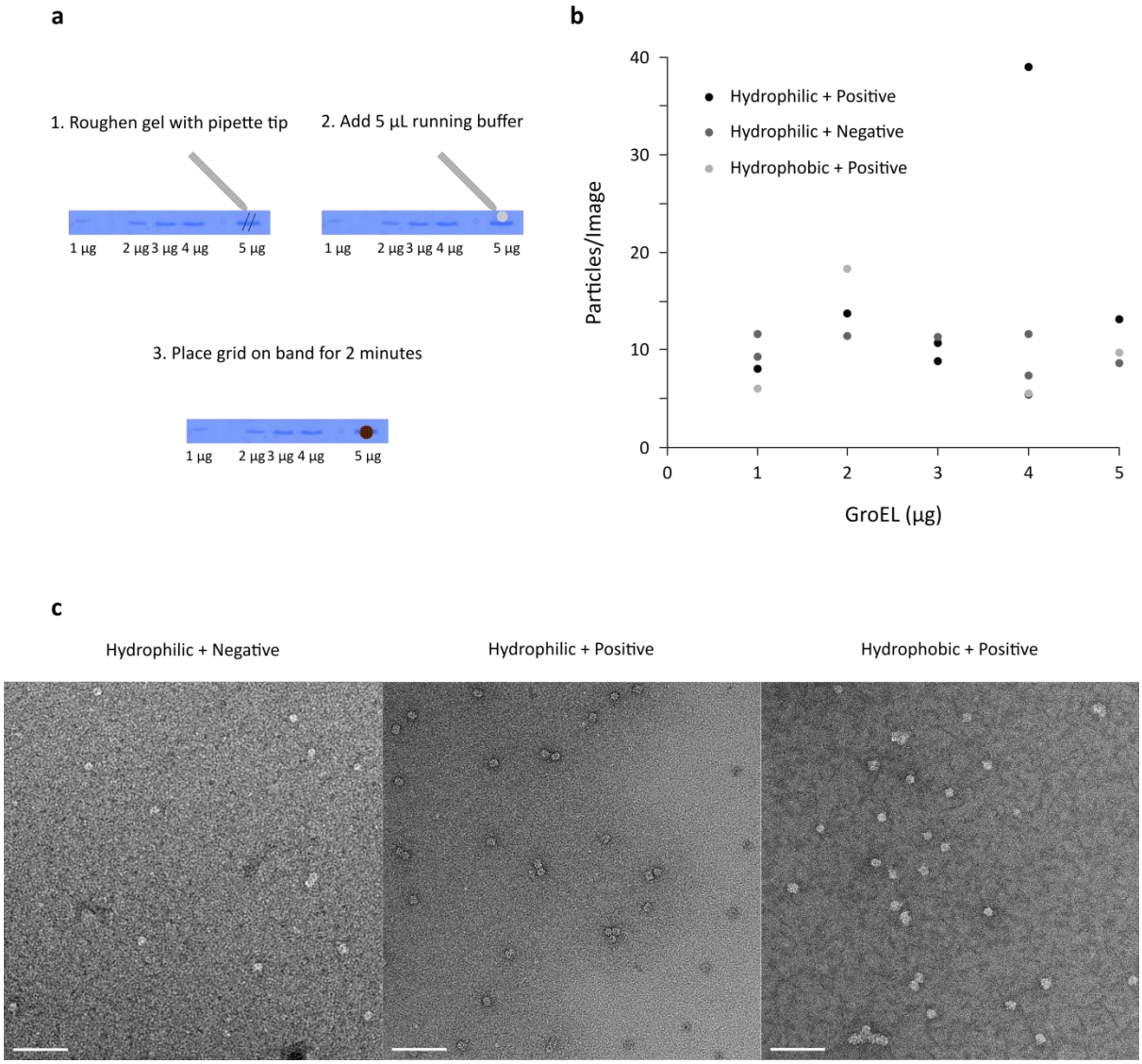



**Figure 3.3.1-1. Effect of properties of copper grid on GroEL particle transfer.** a) BN PAGE grid blotting follows a quick and simple procedure as demonstrated for 5 µg of GroEL. b) There is no relationship between the amount of GroEL present in the BN-PAGE band and subsequent transfer efficiency. Furthermore, the property of the grid has no effect on transfer efficiency. Note the data is given for the mean transfer efficiency for each GroEL concentration under each experiment. Full data provided in Table 8-4 in Appendix. c) An example electron micrograph obtained for each grid property. No particles were seen on an uncharged (neutral) grid. Images were taken at x50,000 magnification at a defocus of 2.00 µm on an F20 Tecnai TEM. White scale bars show 100 nm. Extra example electron micrographs are available in Figures 8-4 to 8-6 in the Appendix.

Since it was not known if dye-uptake could inhibit reconstruction of GroEL, a single-particle negative stain reconstruction was attempted using electron micrographs from the hydrophilic and positively charged grid property (**Figure 3.3.1-2**). This confirmed that dye-uptake during electrophoresis through the gel matrix in BN-PAGE does not hamper the ability to create a reasonable low-resolution reconstruction using these grid-blotted particles. What is unknown is whether or not the dye remains bound to the particle when it diffuses onto the copper grid.

**Table 3.3.1-1. Transfer efficiency statistics for each grid property**

| Grid Property | Mean | Standard Deviation | 99% Confidence Interval |
|---|---|---|---|
| Hydrophilic + Negative | 10.02 | 1.41 | [8.87, 11.32] |
| Hydrophilic + Positive | 9.24 | 2.02 | [6.92, 12.34] |
| Hydrophobic + Positive | 9.51 | 1.80 | [4.74, 19.11] |

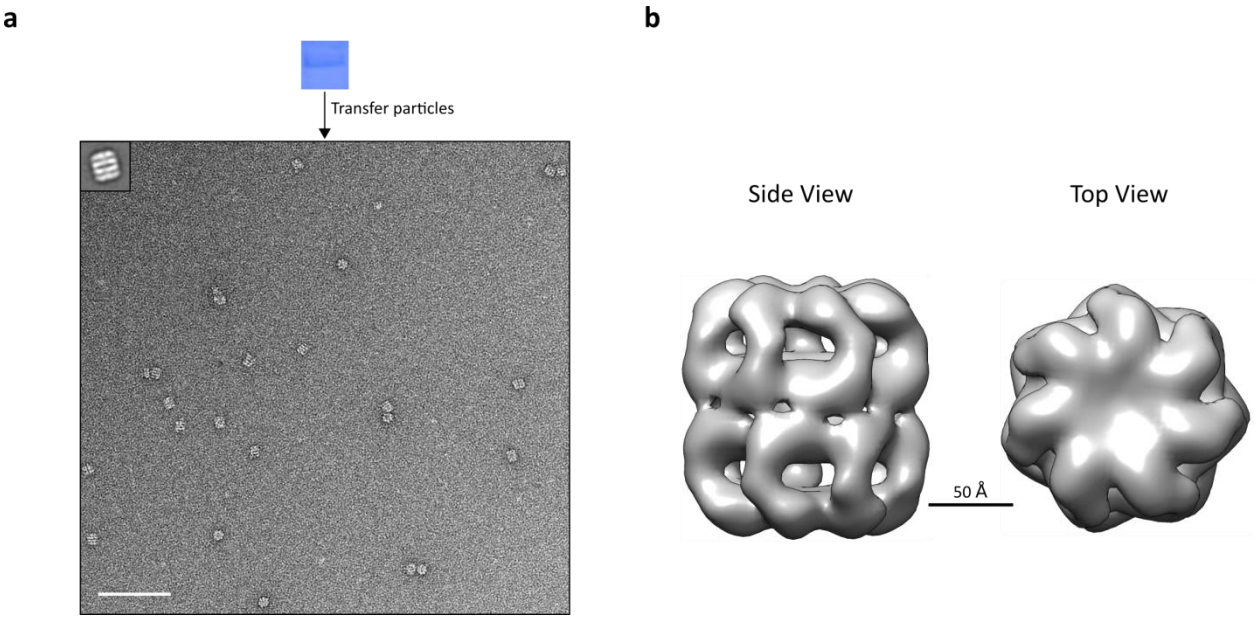



**Figure 3.3.1-2. Grid blotting and reconstruction of GroEL.** a) Protein complex particles from a BN-PAGE gel were transferred onto an EM grid by passive diffusion. The average for GroEL of the picked particles is shown in the top left-hand corner. The white scale bar shows 100 nm. Negative stain electron micrograph taken at a magnification of x50,000 with a defocus of 2.00 µm on an F20 Tecnai TEM. b) A reasonable low-resolution reconstruction can be obtained from the grid-blotted particles.

### 3.3.2 Grid Blotting of Unknown Protein Complexes from *Mycobacterium smegmatis*

Following the success of grid-blotting with the test protein GroEL, ammonium sulphate cuts from *Msm* were run on a BN-PAGE gel and grid-blotted using the hydrophilic and positively charged grid property. No bands in the resolving gel matrix were observed for the 30–40% and 40–50% ammonium sulphate cuts. Only two bands were observed in the 50–60% ammonium sulphate cut. The lack of observed bands may be due to hampered dye-uptake due to a high salt concentration (200 mM NaCl) used in the buffer as salt concentrations which are too high are known to inhibit dye uptake (Wittig *et al*, 2006).

A high salt concentration was used to keep protein complexes intact, as many protein complexes dissociate under low salt conditions. Likewise, other protein complexes only interact under low-salt concentrations (Damodoran & Kinsella, 1980). Evidently, there is a trade-off between dye-uptake and the interaction of protein complexes under specific salt concentrations.

However, the pI of the protein can also have an effect on dye-uptake. The lack of dye-uptake for certain proteins even under low-salt concentrations has been observed by Schägger *et al*

(1994), whereby acidic proteins (pI < 7) were much more likely than basic proteins (pI > 7) to display weak or no dye-uptake under the running conditions tested.

Grid blotting of the bands at the interface of the stacking and resolving gels for the 30–40% and 40-50% ammonium sulphate cuts did not result in the observance of intact particles; the same was found for the two bands in the 50–60% cut. Intact particles were observed for GroEL run on the same gel (**Figure 8-7, Appendix**). Aggregates are expected for grid-blotting from bands located at the interface between the stacking and resolving gels as these cannot enter the gel matrix and hence they could have been transferred. However, aggregation is not expected to be present in bands which can migrate through the resolving gel matrix; thus, it is possible that aggregation could have occurred during the grid-blotting process. Lack of dye-uptake for *Msm* protein complexes during the run could also be due to the low-abundance of proteins present as it was observed for GroEL that the approximate limit of detection is 500 ng per band (data not shown).

In addition, some protein complexes could have dissociated under the running conditions. It was observed that bovine serum albumin (BSA) dissociated when run on a BN-PAGE gel compared to a standard CN-PAGE gel (data not shown); this was not due to lack of dye uptake as the BN-PAGE gel was stained after the run. Schägger *et al* (1994) noted that the dimeric form of BSA bound the dye and remained intact during the BN-PAGE run. It is known that BSA is monomeric in solution and also forms higher-order structures: a dimer, tetramer, and hexamer (Atmeh *et al*, 2007). All forms were seen in the stained CN-PAGE gel and no visible bands were detected in the BN-PAGE gel even after staining. There are many factors underlying the oligomerisation of BSA, with known differences in the proportions of monomeric and higher-order structures in commercially available BSA. Since Schägger *et al* (1994) and this study used commercially available BSA, it is likely that there are differences in the higher-order structures of BSA observed. This study found that there were substantial amounts of higher-order BSA structures, based on the CN-PAGE results, with the monomeric form being the most abundant (data not shown). Similar results were found by Atmeh *et al* (2007). This presents the possibility that the dye interfered with the oligomerisation process of BSA, causing the protein to become unstable and thus propagating dissociation. For reasons unknown, the BSA used by Schägger *et al* (1994) did not dissociate under the running conditions used.

Water-soluble protein complexes have successfully been run on BN-PAGE gels, sometimes requiring some modifications to the procedures depending on the protein complex in question (e.g Eubel *et al*, 2005; Braz & Howard, 2009; Kearns *et al*, 2016). Thus, the success of grid blotting using BN-PAGE will depend on optimising the running conditions for different protein complexes.

### 3.3.3 Electro-elution on Blue Native PAGE

Since the observed distribution of GroEL particles is quite low (**Figure 3.3.1-1**), it was thought that electro-elution on individual bands could improve the transfer efficiency. To electro-elute, an electric current must be passed through the gel which will displace protein complexes from the gel matrix into a small amount of buffer and onto a copper grid. To achieve this, a small device was made consisting of a metal cathode plate housing the BN-PAGE gel, a plastic ring which goes over the target band and holds the copper grid, and a metal anode plate which fits on top of the plastic ring. Buffer would fill the space between the gel and copper grid and between the plastic ring and anode plate which will allow for conductance when the electric current is applied. The current was limited to less than 1.36 V as this is the standard reduction potential of chlorine in which 2 chloride ions bond to form chlorine gas, potentially damaging the grid and electro-eluted proteins when formed. The technique was not successful using the test protein GroEL; transfer of some of the components of the gel matrix may have occurred as the negative control (electro-elution on an empty part of the gel) showed non-protein matter (**Figure 8-8**, **Appendix**).

### 3.3.4 Conclusion

The use of grid blotting on protein bands visualised under non-denaturing BN-PAGE is a successful technique for reconstructing protein complexes efficiently. Partially purified fractions can be run on a BN-PAGE gel and individual bands grid-blotted, allowing for the coupling of the reconstruction of a protein complex with the identity of its protein constituent(s). However, there are some drawbacks to the technique, notably a lack of dye-uptake which hinders the use of the technique for some protein complexes.

In addition, electro-elution was not a successful technique in the transfer of proteins from BN-PAGE to a copper grid. This was due to the transfer of non-protein matter from the gel onto the grid, potentially obscuring any particles present.

# Chapter IV: Cryo-Electron Microscopy

## 4.1 Introduction

Cryo-EM images protein particles in their near-native state by embedding them in vitreous ice. This is achieved by plunge freezing a holey-carbon grid into liquid ethane (Dubochet *et al*, 1988). Recently, there has begun a revolution in the field whereby several protein complexes have now been solved to near atomic ("high") resolution (Bai *et al*, 2015). High-resolution has even recently been obtained on a sub-100 kDa protein complex (Merk *et al*, 2016). For many decades, X-ray crystallography has been considered the 'gold-standard' technique in which to achieve near atomic resolution, but now with advances in hardware and processing software used in cryo-EM this is no longer the case for many protein complexes (Kühlbrandt, 2014). Two major developments came with the use of direct-electron detectors, which greatly improved the detective quantum efficiency (DQE) and is critical for accessing higher resolutions (e.g Veesler *et al*, 2013; for a review see McMullan *et al*, 2016), and the use of imaging particles in movies to correct for the resolution-limiting beam-induced motion (Brilot *et al*, 2012).

Although a high-resolution structure can be crucial to gain biological insight, it is very time-consuming to optimise the conditions necessary in cryo-EM. Factors such as ice-thickness, hole size, electron dose, and defocus range are critical (Cheng *et al*, 2015). As such, native Enc purified from *Msm* (see **Chapter II**) was used in tests for optimal cryo-EM conditions. This was based on its large virus-like size, which makes it much easier to manually pick from other protein contaminants, and its high symmetry, which substantially reduces the number of particles required for a reconstruction. Furthermore, a high-resolution reconstruction of *Msm* Enc would aid in understanding its function (see **Chapter V**).

## 4.2 Materials & Methods

### 4.2.1 Material

Enc particles were purified from *Msm* pellets by anion exchange and gel filtration as described previously (see **Materials & Methods** in **Chapter II**).

### 4.2.2 Vitrification

Quantifoil® holey-carbon grids (Quantifoil Micro Tools, Jena) were glow-discharged in air for 30 seconds. Cryo-EM samples were made in a Vitrobot™ (FEI, USA); the humidity was set to 100% at a temperature of 22$^{o}$C. A blotting time of 3.5 seconds was used. The sample was left for 30 seconds before rapid-plunge freezing into liquid ethane. Grids were stored in liquid nitrogen.

### 4.2.3 Cryo-Electron Microscopy

Concentrations of samples ranged from 0.8 to 1.6 mg/mL. Grids were examined using the Tecnai F20 (Phillips/FEI, Eindhoven) fitted with a CCD camera (4k x 4k) (GATAN US4000 Ultrascan, USA) using a single tilt cryo holder (Gatan, USA). Images were taken in low-dose mode (10−20 $e^{-}/Å^{2}$) at a magnification of x50,000 using a defocus range of 1.5 to 3 µm. A sampling of 2.11 Å/pixel was used and particles were binned by a factor of 3. Processing was completed in Appion (Lander *et al*, 2009) as described previously (see **Materials & Methods** in **Chapter II**). For CTF correction, a Wiener filter was applied (ACE2 Wiener Filter Whole Image (Callagher & Potter, 2009)) and image density was inverted. ACE automatically estimates image astigmatism and defocus using an elliptical averaging function over the power spectrum. The CTF is then calculated using the values which provide the highest confidence (Mallick *et al*, 2005).

For refinement, the initial model underwent a 20 Å low-pass filter before undergoing projection-matching in EMAN (Ludteke *et al*, 1999). Unlike in initial model creation, the EMAN algorithm for model refinement self-generates classes based on the input angular sampling

and the symmetry imposed. High-symmetry particles have fewer classes generated than low-symmetry particles, and more classes are produced as the angular sampling rate is increased.

## 4.3. Theory with Results & Discussion

### 4.3.1 Contrast Transfer Function

Imaging of protein complexes in vitreous ice relies on the ability of biological macromolecules to induce a phase contrast on the passing electron beam. The contrast achieved is much weaker than negative stain, where heavy metal salts coat the protein (Hanszen, 1971).

The contrast transfer function (CTF) describes the phase contrast achieved at different spatial frequencies based on the microscope and defocus used. The function is given as (Wade, 1992):

$$\mathbf{T(k) = -\sin[(\pi/2 * C_s * \lambda^3 * k^4) + (\pi * \Delta f * \lambda * k^2)]}$$

where $C_s$ is the spherical aberration of the microscope (mm), λ is the wavelength of the accelerated electron (pm), Δf is the defocus used (μm), and k is the spatial frequency ($nm^{-1}$). The plot for different defocuses is provided in **Figure 4.3.1-1** based on the F20 Technai TEM microscope. When the amplitude is positive, positive phase contrast occurs and hence atoms appear bright on a dark background, and when the amplitude is negative, negative phase contrast occurs and hence atoms appear dark on a white background. When the curve crosses the x-axis ("the zeros"), no phase contrast occurs and hence this represents a loss of information at that resolution (spatial frequency). To fill in such lost information, different defocuses must be used in which they cross the x-axis at different resolutions (**Figure 4.3.1-1**). Thus, the information lost in one defocus is compensated by another defocus and hence the defocus range used becomes critical at higher-resolutions. In reality, the CTF is significantly dampened at higher spatial frequencies due to factors such as noise and drift (Erickson & Klug, 1971).

The Fourier transform (FT) is another way of visualising the CTF for a particular image; here, "the zeros" are represented as the edges of white Thon rings (**Figure 4.3.1-1**). Thon rings further away from the centre represent higher resolutions (Thon, 1966). The highest

theoretical resolution which can be achieved is based on the sampling of the waveform (in Å/pixel) at a particular magnification; this is known as the Nyquist limit which is twice the sampling (Nyquist, 1928; Shannon, 1948). In reality, Thon rings from a particular image may not extend to the Nyquist limit. In addition, to achieve a particular resolution, a sampling is used which is twice to three times Nyquist limit (Cheng *et al*, 2015). For example, to achieve a resolution of 4 Å, it is safe to use a sampling of 1 Å/pixel, where the highest resolution theoretically possible is 2 Å.

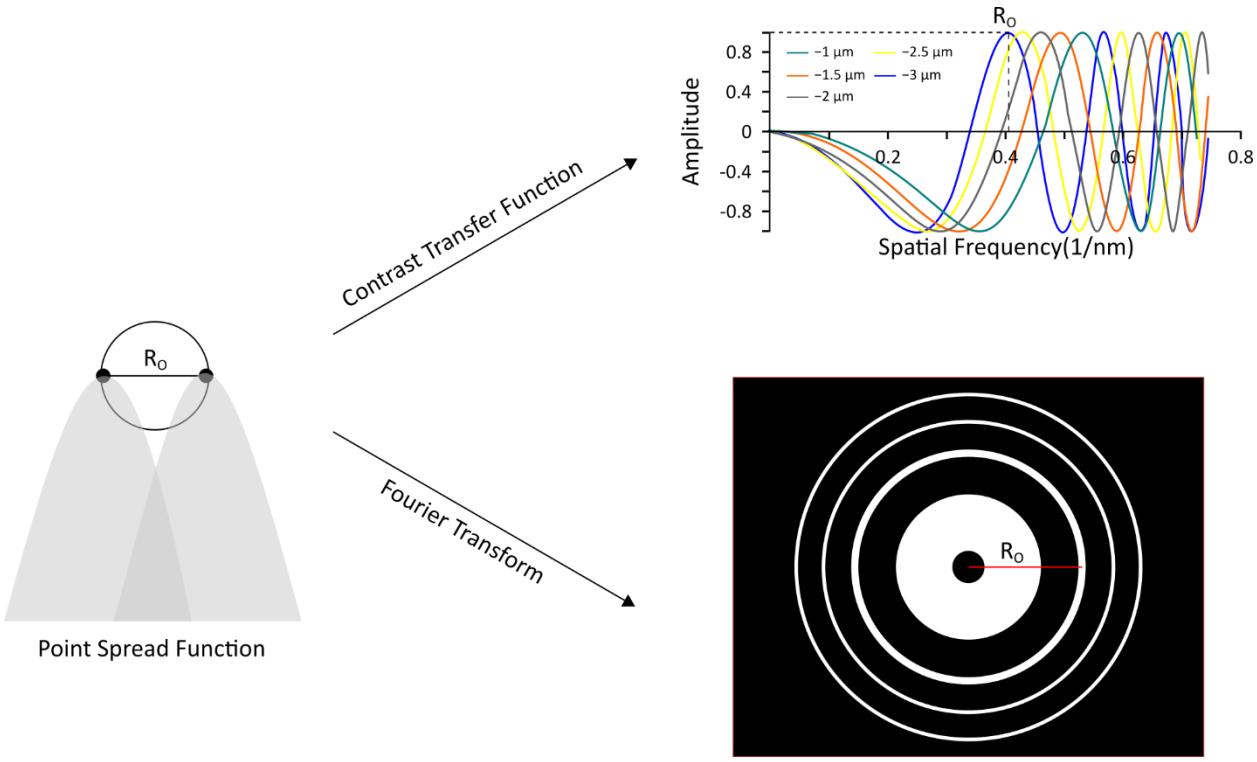



**Figure 4.3.1-1. Minimum resolvable distance.** Resolution is determined by the Rayleigh criterion, where two points in an object are resolvable if their point spread functions only have an overlap which is around 75% of the highest intensity (Liao & Frank, 2010). This is given by the radial cut-off ($R_o$) which can be seen in the CTF (dotted line) and FT (red line). For the CTF, curves for different defocuses are shown but $R_o$ is only given for the 3.00 µm defocus. Note also that the Fourier transform is provided for only the defocus of 3.00 µm.

### 4.3.2 Optimisation of Parameters

Protein concentration and ice-thickness are critical parameters in a cryo-EM experiment. The holey carbon grid typically requires much more protein than negative stain (Cheng *et al*, 2015). Since *Msm* Enc is present in very low abundance in the cell cytoplasm, this posed a significant challenge. In cryo-EM, the vitreous ice should be thin enough to maximise contrast but not too thin that it disappears ("pops") while imaging under the electron beam (Cho *et al*, 2013). Ice-thickness is controlled by the amount of protein solution added onto the grid and the blotting time before plunge-freezing. The humidity is usually kept at 100% to avoid drying out of the solution before it can be blotted off and rapidly frozen (Cho *et al*, 2013). A sample application of 2.5 or 3 µl with a blotting time of 3–6 seconds is standard to achieve a good ice-thickness. Furthermore, the entire procedure from sample application to plunge freezing is semi-automated in a Vitrobot™ to achieve some level of consistency (Cheng *et al*, 2015).

Electron dose is also a critical parameter; the dose should be high-enough to achieve sufficient contrast at a low defocus but not too high that it destroys the protein specimen through radiation damage and thus abolishing access to high-resolution features. A range of 10 to 20 $e^{-}/Å^2$ is typical (Cheng *et al*, 2015). However, at >10 $e^{-}/Å^2$ the resolution achieved will be severely limited by beam-induced radiation damage of the specimen, and it is thus prudent to consult the critical exposure curve developed by Grant & Grigorieff (2015) in order to use the correct electron dose when aiming for a particular resolution.

Finally, the hole-size used can have a substantial impact especially if the protein is present in low concentration; for *Msm* Enc, the smallest hole size (0.61 µm) was used to image the entire area at a high sampling (2.11 Å/pixel) and thus maximising the greatest number of Enc particles per image.

### 4.3.3 Contamination

During sample preparation, a number of contaminants can arise such as ethane or non-vitreous ice (**Figure 4.3.3-1**) (Thompson *et al*, 2016). A number of steps can be taken to minimise the probability of these contaminants occurring, such as ensuring that the sample is not warmed above ~ –160°C after plunge-freezing to avoid the formation of non-vitreous

ice (Thompson *et al*, 2016), or moving the sample from the liquid ethane to the cryo sample holder as quickly as possible to avoid ethane contamination.

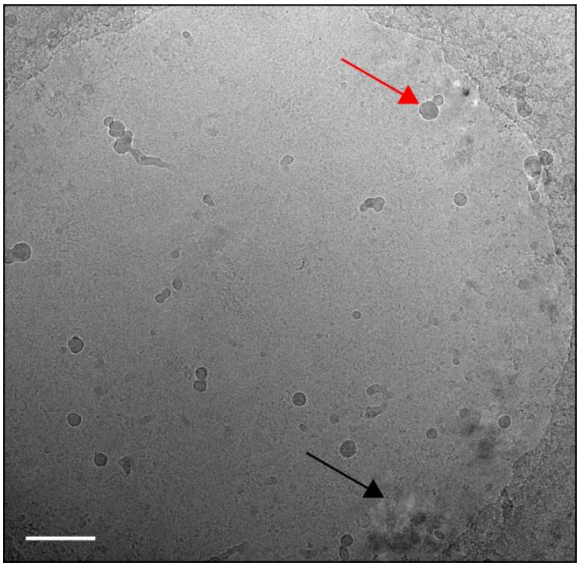


**Figure 4.3.3-1. Contaminants.** An example of possible non-vitreous ice (black arrow) and ethane (red arrow) contamination in a 0.61 µm holey-carbon grid. White scale bar shows 100 nm. Electron micrograph taken at x50,000 magnification on a F20 Tecnai TEM.

### 4.3.4 Reconstruction using Appion

All of the image processing and reconstruction was completed through the Appion pipeline (Lander *et al*, 2009). The main advantage of the Appion system is that standard programmes for image processing and reconstruction are housed together, simplifying the procedures and hence decreasing the time required to obtain a final reconstruction (Lander *et al*, 2009).

Astigmatism and drift are two main factors which worsen a reconstruction and can be easily evaluated in an electron micrograph. A degree of astigmatism will be present due to aberrations in the microscope lens and appears as deviations from a circle of the image Thon rings (Orlova & Saibil, 2011). A measure of astigmatism is provided with an estimation of the electron micrograph CTF (e.g ACE2) (Lander *et al*, 2009), while drift can clearly be seen as "cuts through" the Thon rings in the FT in the direction of the motion (**Figure 4.3.4-1a**).

Electron micrographs with drift or bad astigmatism are discarded. Good electron micrographs will then be used for manual picking of target particles.

A well-known problem in cryo-EM is what is known as “Einstein from noise”; this is when a reconstruction procedure is implemented in which the data consists primarily of noise and yet a “reasonable” model is generated through bias introduced in particle picking using a template (Henderson, 2013). Mao *et al* (2013) claimed that they had solved a 6 Å structure of the HIV-1 envelope glycoprotein trimer, which others subsequently disputed since there was little evidence that their electron micrographs contained any particles (Henderson, 2013; Subramaniam, 2013; van Heel, 2013). Since cryo-EM produces much less contrast than negative stain, to avoid such an outcome, it is prudent to manually pick the target particles, especially if the sample is not homogenous (Henderson, 2013).

Selected particles must undergo CTF correction, which at its minimum involves converting negative phase contrast to positive phase contrast (“phase flipping”) (**Figure 4.3.4-1b**). The dampening effect induced on the CTF by various factors can be accounted to some extent through amplitude correction, but the success of this depends on the signal-to-noise ratio (SNR) since any increase in signal at high spatial frequencies will also be accompanied by an increase in noise (**Figure 4.3.4-1b**) (Orlova & Saibil, 2011).

Resolution is heavily dependent on the SNR. In cryo-EM, the low electron doses required to limit specimen damage also produce a low SNR. This is compensated to a certain extent by using the standard image processing method of averaging particles which boosts the signal. DQE is a measure of the relative ratio of output to input SNR variance and is thus a measure of how efficiently an electron microscope camera detects incoming electrons. Thus, a camera with a high DQE will be able to tolerate lower electron doses without significantly hampering the resultant SNR. Signal can also be boosted by down-sampling the image pixels (pixel integration), which is most useful when the microscope DQE is not high as is the case for a CCD camera (Ruskin *et al*, 2013). Thus, for the sampling of 2.11 Å/pixel used, cryo Enc particles were binned by a factor of 3 in order to boost the SNR and improve the resultant reconstruction.

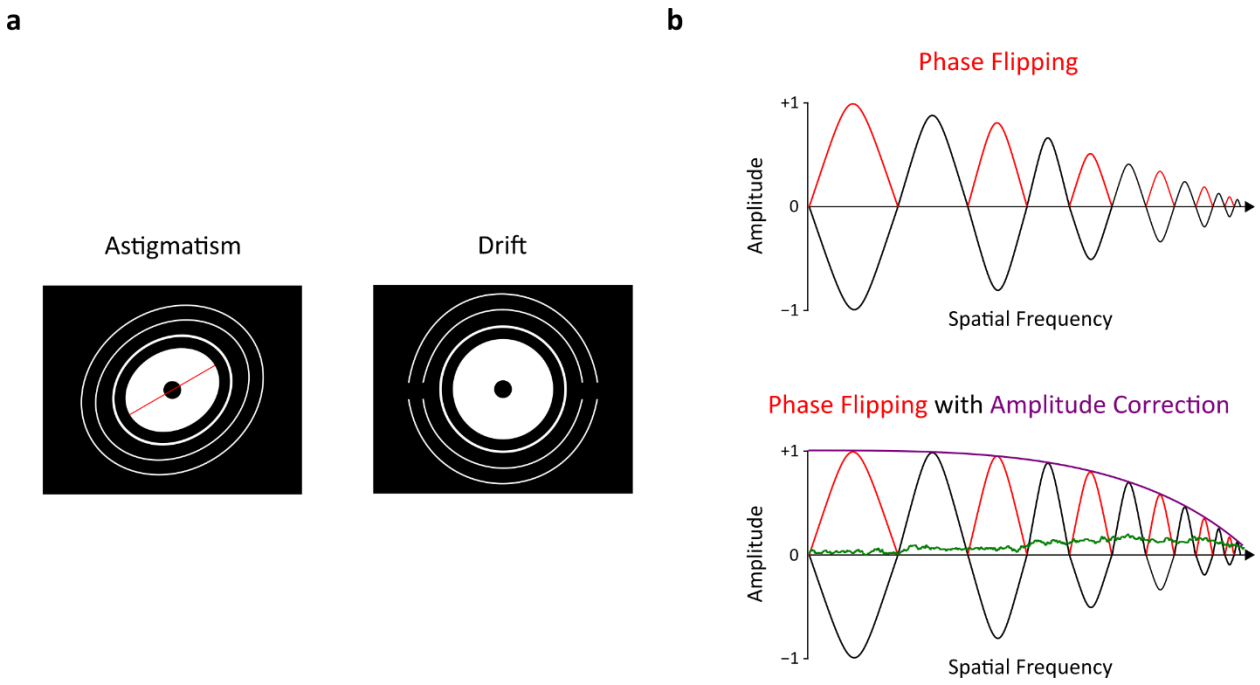



**Figure 4.3.4-1. FT and CTF Correction. a)** Two types of microscope aberrations which can be visualised by the FT are astigmatism and drift. The direction of the astigmatism is shown (red line) while drift shows as "cuts" through the FT. **b)** CTF correction can take the form of phase flipping (top) where negative phase contrast is converted to positive phase contrast, or phase flipping in conjunction with amplitude correction (bottom). Note that in amplitude correction, noise present (green line) is also up-weighted with the amplitudes. Also note that low-resolution information is usually down-weighted in most amplitude correction algorithms.

The particles present on an electron micrograph are 2D projections of the original 3D protein complex. Thus, it is expected that many orientations (projections) are present if there is no bias towards one orientation (De Rosier & Klug, 1968). To obtain the original 3D density of the protein complex, specific particles must be matched to particular orientations. Evidently, this requires that enough orientations are present to reliably reconstruct the original density (De Rosier & Klug, 1968). Many back-projection procedures rely on the projection theorem which states that a 2D FT of a 2D projection gives a central section through the 3D transform of the 3D density (Klug & Crowther, 1972). Thus, how well the FT is sampled will determine the success of a reconstruction (Klug & Crowther, 1972). The number of particles required to reconstruct the density will depend on the angular sampling ($\Delta\phi$) and the number of particles representing a particular view ($N_v$) (Frank, 2006):

$$N_v = (\Delta\phi/2\Pi) \times N_{tot} \times p_v$$

where $N_{tot}$ is the total number of particles obtained and $p_v$ is the probability of obtaining a particular view. It is evident that $p_v$ is unknown in most cases, so the only adjustable parameters are $N_{tot}$ and to a certain degree, $\Delta\phi$. The symmetry of a particle can aid the reconstruction process, since certain orientations will be related, thus reducing the number of particles required to obtain a particular resolution (De Rosier & Klug, 1968). For example, a particle with C3 symmetry means that it must be rotated by $120^o$ in order to arrive at the same position. Hence, any given end-on view of the particle will contain three views of the asymmetric unit. For *Msm* Enc, its icosahedral symmetry means that any given orientation will contain 60 views of the asymmetric unit, drastically reducing the number of particles required to reconstruct its 3D density.

Reconstruction usually proceeds through common lines (angular reconstitution) after particles have been grouped (classed) according to their features. From the projection theorem, two different 2D projections of a 3D object will share a 1D line projection (i.e common line). This information can be used to assign Eulerian angles for particles of unknown orientation. Since two line projections can be related by rotation around a common tilt axis, a third common line is required to unambiguously assign angles (van Heel, 1987).

Algorithms which utilise common lines must produce a sufficient number of class averages (based on object symmetry) which represent different unique views of the 3D object from the particle data, insert these views into the Fourier volume based on common lines, and then apply the inverse FT to produce the 3D density. This will produce an initial model which is refined through projection-matching; the initial model is re-projected and matched to particular particle orientations present in the data. This process is iterated until the model converges to a solution (**Figure 4.3.4-2**) (Ludtke *et al*, 1999). The more symmetry a protein complexes possess, the easier it is to reconstitute since there are many more constraints imposed (Orlova & Saibil, 2011).

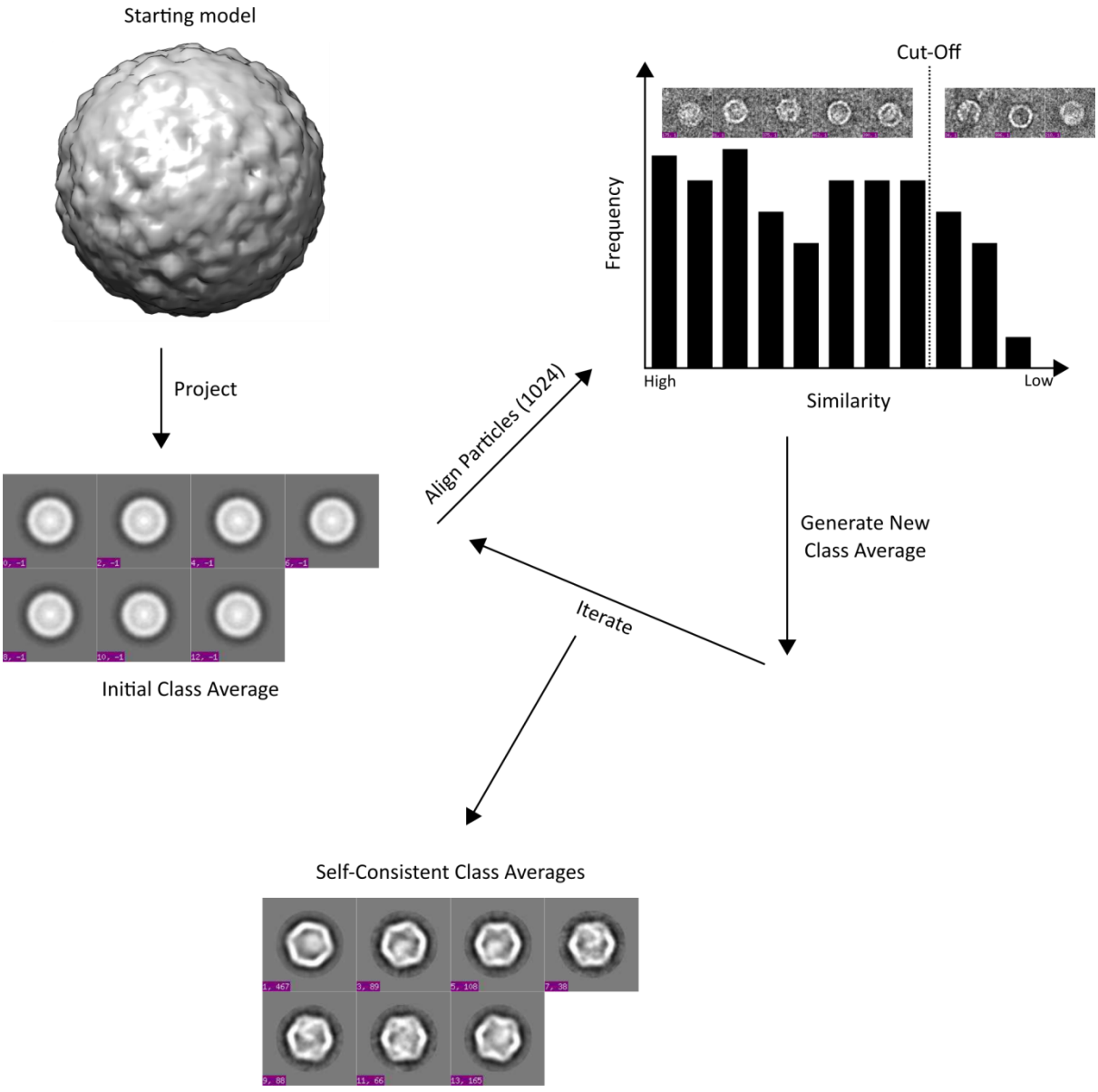



**Figure 4.3.4-2. One round of projection-matching**. An initial model is projected, based on angular sampling rate and symmetry, and aligned to particles in the dataset; bad particles are eliminated using a similarity-score cut-off and not used to generate a new class average, but are allowed to participate in each iteration. Once self-consistent class averages are produced, they are assigned Eulerian angles and a new model is made. This model will then act as the starting model in the next round of projection-matching. Note that this algorithm is from EMAN projection-matching model refinement (Ludtke *et al*, 1999) and other algorithms differ in their approach (Orlova & Saibil, 2011). Note also that data is negative stain single particle for *Msm* Enc.

The cryo-EM reconstruction of *Msm* Enc is given in **Figure 4.3.4-3**.

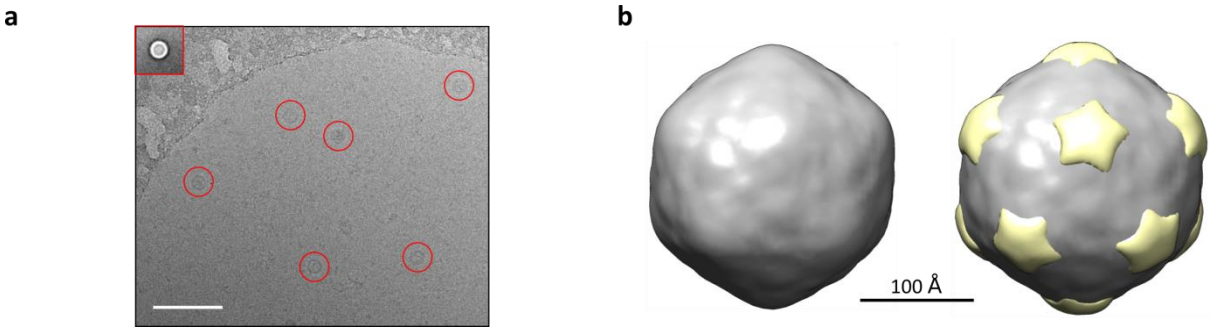



**Figure 4.3.4-3. Cryo-EM of Enc from *Msm*.** a) Enc particles (circled, red) are present in a thin layer of vitreous ice. Conditions were as follows: 2.5 µl of sample was pipetted onto a 0.61 µm Quantifoil holey carbon grid and left for 30 seconds, blotting time of 3.5 seconds, 100% humidity, and a temperature of 22°C. The average for all picked particles is shown in the top left-hand corner. White scale bar shows 100 nm. b) Low resolution reconstruction of Enc with icosahedral symmetry imposed (left). The resulting density was aligned to the negative stain reconstruction (yellow) to show the position of the 5-fold axis (right).

The power spectrums (FTs) obtained for Enc contained few Thon rings (data not shown). Since vitreous ice does not produce Thon rings, the lack of rings could be due to little protein present in the ice (Orlova & Saibil, 2011). It was also found that Enc did not tolerate a drop in the salt concentration (200 mM NaCl) used in the buffer. High salt concentrations are known to diminish contrast in cryo-EM (Bollschweiler *et al*, 2017). Furthermore, only 217 particles were obtained owing to the lack of particles (or none) present in each hole. This was due to the low abundance of Enc relative to other proteins present in the cell cytoplasm. These effects combined to produce a cryo-EM model which is at a lower resolution to the one produced by negative stain (**Figure 4.3.4-3b**). As mentioned previously, picked particles were binned by a factor of 3 in order to boost the SNR and improve the resultant reconstruction. This had a limited effect as the resolution is much lower compared to the reconstruction obtained by negative stain (35 Å vs 25 Å). Based on sampling two times above Nyquist limit, approximately 26 Å was theoretically achievable if sufficient sampling was the sole criteria limiting resolution. The diminished contrast may have had a significant effect, as higher-resolution Thon-rings were not visible, but the low-resolution information would have been

available within the first two Thon rings based on the defocuses used. It is possible that edge-detection used by ACE2 to estimate the CTF was hampered by the weak Thon rings observed, and thus CTF estimation may have failed which would have compromised CTF correction. It was observed that ACE2 estimated the defocus in most of the images to be between 3.00 µm to 4.00 µm, in which the first Thon ring occurs at around 30−40 Å. This would explain the much worse resolution compared to negative stain, since it is likely that the majority of the useable information in the cryo dataset is within the first Thon ring.

Enc does not have a preferred orientation in the vitreous ice, which is known to be a problem for some protein complexes (Cheng *et al*, 2015). Thus, future improvements in resolution will have to focus on boosting the contrast obtained as well as obtaining a larger amount of Enc particles. Since direct-electron detectors have a much higher DQE than CCD camera, it is probable that Enc will benefit from the use of a direct electron detector since contrast will be improved even though a higher salt concentration is required to keep the particles in their optimal state.

### 4.3.5 Using High-Resolution To Solve Protein Identity

Protein identification using low-resolution structures depends on the availability of crystal structure homologues. However, there are many structures without a close homologue in which there exists a solved crystal structure, and thus other methods for protein identification must be used instead. Thus, this section serves as a proposal on the use of cryo-EM as a more accurate and reliable method for solving the protein identification problem (see **section 2.3.3** in **Chapter II Results & Discussion**) for unknown protein complexes which could be complemented with the results from LC-MS/MS.

At low-resolution (20−30 Å), valuable information is available on the number of subunits within the protein complex, how these subunits are related to each other, and the overall shape of the complex which could lead to initial hypotheses relating to its mechanism of function. For example, a 30 Å reconstruction was obtained for a ten-protein kinetechore complex in yeast which allowed the researchers to test hypotheses relating to self-assembly (Wang *et al*, 2007).

Protein topology can be determined with a high-degree of confidence based on sequence alone (e.g Balakrishnan *et al*, 2011; Ovchinnikov *et al*, 2015; Ovchinnikov *et al*, 2017). Furthermore, secondary structure prediction and fold recognition is highly accurate for the majority of proteins, allowing good templates to be found for homology modelling (McGuffin *et al*, 2004). At 4−6 Å, a protein polypeptide chain can be fitted into a cryo-EM map with confidence when the sequence of the protein in question is known (DiMaio *et al*, 2015). Thus, it is theoretically possible to match an unknown protein sequence to a given cryo-EM structure at that resolution. However, it would be too computationally expensive to predict the folds of all protein ORF sequences in a given organism in order to match them to the cryo-EM map. However, several features are available in the cryo-EM map itself which would winnow down the number of possibilities to a manageable number. At 4−6 Å resolution, bulky amino acids are visible along with some other amino acid side chains, depending on their size and the degree of flexibility present in the map. This pattern of side-chains in the cryo-EM map is valuable in screening the proteome of the organism in question, leaving a manageable number of candidates which can be tested against the map through fitting proposed by DiMaio *et al* (2015).

For an intermediate resolution (~7−10 Å) cryo-EM map, the pattern of secondary structure can be determined with a small number of possible topologies (e.g Baker *et al*, 2007). As mentioned previously, *de novo* topology prediction is possible for a protein sequence and thus matching predicted secondary structure topology to that provided by the cryo-EM map can be a feasible method of protein identification. Since cryo-EM maps can provide MW estimates at any resolution, this information can still be used to choose candidates from the target proteome for secondary structure topology prediction. However, the method will evidently be much more computationally expensive if no other features are visible in the cryo-EM map, such as bulky side-chains, which can aid to further cull the number of possibilities.

The strategy given in this section is highlighted in **Figure 4.3.5-1**. It is evident that low-resolution features are available at higher-resolutions and thus higher-resolution maps will contain more information to develop and test hypotheses. Although LC-MS/MS is a good method for determining protein identity, when no other information is available the method suffers from a poor ability to distinguish real hits from noise, most pronounced if the protein in question is present in low abundance. Thus, LC-MS/MS could act as a complementary

method for protein identification, with further confidence in identified hits being given by the information given by cryo-EM maps at different resolutions. This study has demonstrated that such a method is feasible for low-resolution maps where Enc and BrfA/B where unambiguously identified through a combination of these methods (see **Chapter II**). When crystal structures are unavailable, however, the information from a higher resolution (<10 Å) map could significantly benefit the search for protein identities for unknown protein complexes.

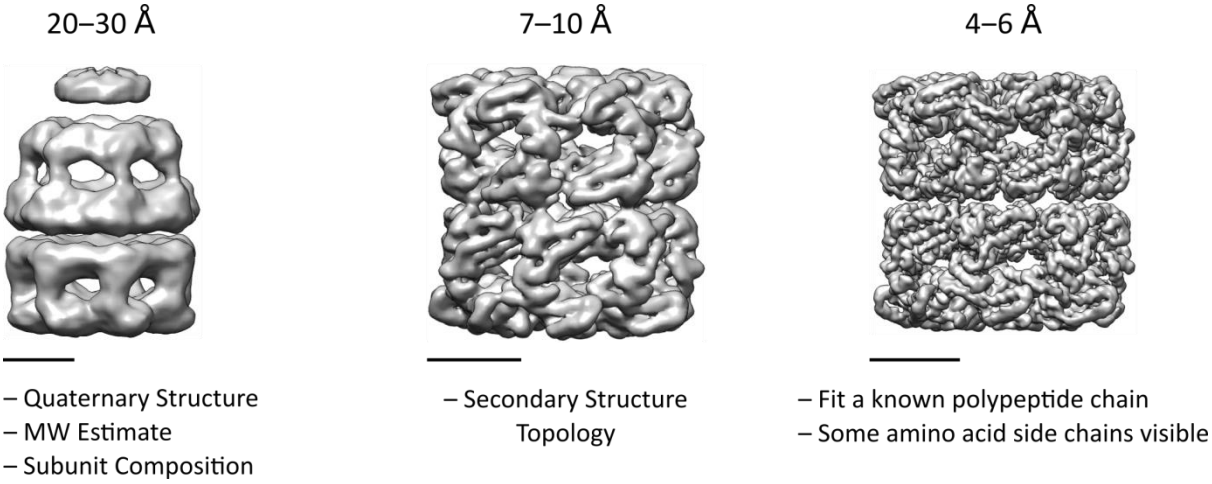



**Figure 4.3.5-1. Structural information at different resolutions**. Low-resolution structures provide information on the overall quaternary structure of a protein, while higher-resolution structures offer more information on the secondary structure of the protein in combination with information gained at low-resolutions. Structures used (from left to right): 23.5 Å GroEL/GroES (emd-1046), 7 Å GroEL (emd-1997), and 3.23 Å GroEL (emd-3407). Scale bar shows 50 Å.

### 4.3.6 Conclusion

*Msm* Enc was successfully reconstructed by cryo-EM. Further improvements in resolution will be aided by the production of more Enc particles, by purifying from cell culture filtrate rather than the cell cytoplasm (see **Chapter V**). The microscope used for the cryo-EM tests had known issues with stage stability (Mohammed Jaffer, personal communication) which may have also contributed to loss of resolution. Boosting the image contrast is required to improve the cryo Enc reconstruction which is likely to benefit from a direct electron detector and a

more stable stage. A method was suggested which could aid in identifying unknown protein complexes in the case where no crystal structure homologues are available to test LC-MS/MS data against an obtained EM map.

# Chapter V: Biological Characteristics

## 5.1 Introduction

From the General Strategy (see **Figure 1.2-1**) introduced in **Chapter I**, once protein complexes have been identified, insights can be gained as to their biological function. GSI, BrfA/B, and Enc were purified through fractionation and identified based on their low resolution structure and LC-MS/MS results of native or SDS-PAGE bands (see **Chapter II**). A literature search can reveal what is already known about their biological function (see below). However, since Enc represents the first such structure solved in Mycobacteria, further insights could be produced with regard to its control over substrate entry (see below).

GSI (EC 6.3.1.2) is an enzyme involved in the ATP-dependent production of the amino acid L-glutamine through the ligation of glutamate and ammonia, and is thus critical in nitrogen recycling of bacterial metabolism. In its active form, GSI consists of two stacked hexameric rings (Krajewski *et al*, 2005). *Mtb* contains four GS genes (GlnA1−4) within its genome, but only GSI (GlnA1) is abundantly expressed and essential for *Mtb* growth (Harth *et al*, 2005). GSI may also be crucial in the synthesis of the *Mtb* cell wall (Hirschfield *et al*, 1990; Harth *et al*, 1994). Because GSI is also essential for *Mtb* virulence (Harth *et al*, 1994; Tullius *et al*, 2003), there has been interest in exploring novel drug targets against the enzyme (Harth & Horwitz, 1999; Harth & Horwitz, 2003; Krajewski *et al*, 2005; Mowbray *et al*, 2014). Furthermore, there is some evidence that GSI is exported into the phagosome during *Mtb* infection (Harth *et al*, 1994) despite the absence of a leader peptide for secretion (Harth & Horwitz, 1997). Further investigation suggests that GSI export may be the result of bacterial leakage or autolysis under over-expression coupled with its extracellular stability (Tullius *et al*, 2001).

Ferritin-like proteins act as storage for insoluble ferric iron ($Fe^{3+}$) and is thus critical for intracellular iron regulation in cells, as the biologically available ferrous iron ($Fe^{2+}$) reacts with hydrogen peroxide ($H_2O_2$) to form the highly toxic hydroxyl radical, $HO^{\cdot}$, via the Fenton reaction (Smith, 2004). In bacteria, they can be divided into three categories: non-heme-binding ferritin (EC 1.16.3.1), heme-binding bacterioferritin (EC 1.16.3.1), and DNA-binding proteins during stationary phase (Dps) (Smith, 2004). In their active forms, ferritin-like

proteins form hollow octahedral or tetrahedral nanocages (Zhang & Orner, 2017). While *Msm* possess Dps which exists in two multimeric forms (Smith, 2004), *Mtb* does not (Khare *et al*, 2017). As mentioned previously, in *Mtb*, BrfA (Rv1876) and BrfB (Rv3841) are octahedral (see **Chapter II**). In *Mtb*, BrfA acts to regulate iron levels in the cell, while BrfB functions to sequester excess iron in order to prevent iron toxicity. In addition, BrfB has a much higher capacity to store iron than BrfA (6000 vs 4500 $Fe^{3+}$/protein), but BrfA is three times faster at releasing stored iron than BrfB (Khare *et al*, 2017). Furthermore, *Mtb* lacking BrfB is susceptible to killing by antibiotics and is unable to persist in infected mice (Pandey & Rodriguez, 2012).

Encs form icosahedral shells with function to encapsulate target proteins via a unique C-terminal extension (Sutter et al, 2008; Nichols et al, 2017). Encs appear to be widely distributed, appearing in 15 bacterial and 2 archael phyla (Giessen & Silver, 2017). Only a handful of Encs have been confirmed structurally in a variety of organisms: *Msm* (this work), *Rhodospirillum rubrum* (He *et al*, 2016), *Rhodococcus erythropolis* N771 (Tamura *et al*, 2014), *Mtb* (Contreras *et al*, 2014), *Myxococcus xanthus* (McHugh *et al*, 2014), *Thermotoga maritima* (Sutter *et al*, 2008), *Pyrococcus furiosus* (Akita *et al*, 2007), *Streptomyces griseus* (Saito *et al*, 2003), and *Brevibacterium linens* (Valdés-Stauber & Scherer, 1994).

The first crystal structure of a large “virus-like” particle was solved in *P. furiosus* by Akita *et al* (2007), which was subsequently found by Sutter *et al* (2008) to be related to their “virus-like” particle, which they named Enc, in *B. linens*. The structure of Enc is arranged as either 60 subunits composed of 12 pentamers (T=1 icosahedron) or 180 subunits composed of 12 pentamers and 20 hexamers (T=3 icosahderon) (Nichols *et al*, 2017) (**Figure 5.1-1**). The Enc fold is homologous to the HK97 major phage capsid protein; a “fossil” provirus found in *Sulfolobus solfataricus* was homologous to the Enc in *P. furiosus*, suggesting an evolutionary relationship (Heinemann *et al*, 2011). Encs are not the only proteinaceous organelles; carboxysomes are much larger (~100–200 nm) microcompartments involved in containing carbon fixation reactions. However, there is no evidence that carboxysomes are related to viruses (Kerfeld *et al*, 2005).

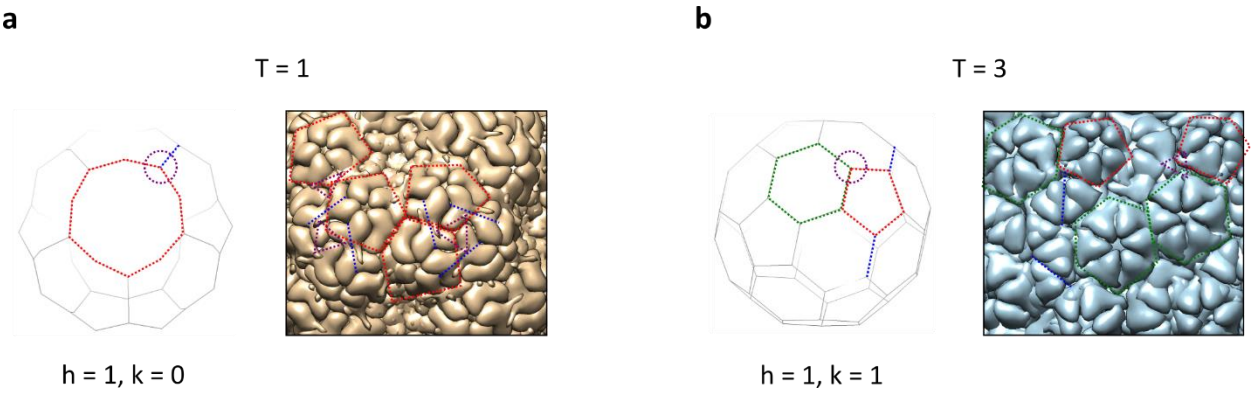



**Figure 5.1-1. Packing arrangements of icosahedrons.** Icosahedron subunits associate in specific ways, depending on the triangulation (T) number determined by the formulae $T = h^2 + 2hk + k^2$, where h and k are non-negative integers (Prasad & Schmid, 2012). For T = 1 icosahedrons (a), subunits associate as pentamers with 5-fold (red), 3-fold (purple), and 2-fold (blue) symmetries. For T = 3 icosahedrons (b), subunits associate in pentamers and hexamers and display an additional 6-fold (green) symmetry axis. Packing arrangements are shown for Encs *T. maritima* (brown) (pdb 3dkt) and *P. furiosus* (light blue) (pdb 2e0z). Note that models are not shown to scale. Ideal icosahedrons were produced in UCSF-Chimera (Petterson *et al*, 2004).

Encapsulated proteins are usually dye-decolourising peroxidise (DyP) (EC 1.11.1.7) or ferritin-like protein (Flp) found immediately upstream from Enc in the genome (Sutter *et al*, 2008). In *Mtb*, Enc (Cfp29, Rv0798c) was first discovered in the filtrate and membrane fraction of cell cultures and is able to elicit long-lived memory immunity in mice (Rosenkrands *et al*, 1998). Recently, it was discovered that *Mtb* Enc encapsulates two other target proteins in addition to DyP (Rv0799c): BrfB and 7,8-dihydroneopterin aldolase (EC 4.1.2.25) (FolB, Rv3607c) (Contreras *et al*, 2014). In *Msm*, only DyP and BrfB contain a C-terminal extension and hence are likely to be encapsulated.

Sutter *et al* (2008) originally noted the correspondence between the location of the *T. maritima* Enc pores and the binding of the cargo. The 5-fold and 3-fold axis of Enc, along with a 2-fold interaction between two subunits, harbour pores around 5 Å in diameter in an otherwise solid compartment. In addition, the partial density of the C-terminal extension was found in pockets corresponding to all of the symmetry-related axis of *T. maritima* Enc. Further observations found that *N. europaea* Flp, which is related to the *T. maritima* Flp, forms a decameric structure with 5-fold symmetry (Chang *et al*, 2005). In addition, the low-resolution

structure of *B. linens* DyP was found to be hexameric with 3-fold symmetry (Sutter *et al*, 2008). These observations lead to Sutter *at al* (2008) to propose the hypothesis that the cargoes of Enc bind at a specific pore, based on the matching symmetry of the pore with that of the C-terminal extension.

There is considerable interest in the biotechnological applications of Enc (e.g Tamura *et al*, 2014; Moon *et al*, 2014; Choi *et al*, 2016; Cassidy-Amstutz *et al*, 2016; Sonotaki *et al*, 2017), although there are many unresolved questions regarding the logistics of binding of the cargo proteins and the mediation of substrate access into the Enc lumen. Since mediation of substrate access is critical to developing more sophisticated biotechnological tools which utilise the encapsulating function of Enc, experimental and bioinformatics methods were used in this study to propose a mode of binding of *Msm* and *Mtb* Enc cargo proteins along the same lines as that developed by Sutter *et al* (2008) i.e that the cargo proteins of *Msm* and *Mtb* Enc are hypothesised to bind at a specific Enc pore, corresponding to their respective symmetries.

## 5.2 Materials & Methods

### 5.2.1 Reconstruction of Encapsulated Dye-Decolourising Peroxidase

All processing was completed in Appion (Lander *et al*, 2009). A sub-stack was created in which particles that were broken, deformed, or may contain BrfB were deleted. This left 207 particles. Enc was masked out using a rectangular box with a Gaussian drop-off in intensity. Class averages were produced as described previously (see **Materials & Methods** in **Chapter II**) using hierarchical clustering. An initial model was created using EMAN Common Lines (Ludtke *et al*, 1999) with C3 symmetry imposed. This model was refined using EMAN projection-matching (Ludtke *et al*, 1999) with D3 symmetry imposed for 26 iterations. For refinement, a 15 Å low pass filter was used and a mask radius of 70 Å applied. Angular sampling rate was: 5 iterations of $10^{o}$, 5 iterations of $8^{o}$, 10 iterations of $5^{o}$, and 6 iterations of $3^{o}$.

### 5.2.2 Export of Encapsulin

Since previous work suggests that Cfp29 (Enc) is exported in *Mtb* (Rosenkrands *et al*, 1998), the possible export of *Msm* Enc was investigated. Starter cultures were made as described previously (see **Materials & Methods** in **Chapter II**). 100 mL of Middlebrook 7H9 media supplemented with 0.2% glucose, 0.2% glycerol, and 0.05% Tween-80 was inoculated with 10 mL of starter culture. The culture was grown at 37$^{o}$C with shaking at 120 rpm until an $OD_{600}$ of 1.2 was reached. The culture was spun-down at 4000g (Beckman, California, USA) and supernatant filtered on ice using a 0.45 μm filter. 50 mL of filtered supernatant was used for purification by anion exchange.

### 5.2.3 Anion Exchange

Completed as described previously (see **Materials & Methods** in **Chapter II**).

### 5.2.4 Negative Stain Electron Microscopy

Completed as described previously (see **Materials & Methods** in **Chapter II**).

### 5.2.5 SDS-PAGE

Completed as described previously (see **Materials & Methods** in **Chapter II**).

### 5.2.6 Phylogenetic Analysis

Alignments of protein sequences were produced in UCSF-Chimera (Petterson *et al*, 2004) and exported to MEGA6 (Tamura *et al*, 2013) for phylogenetic analysis. Alignment of DNA sequences was completed in MEGA6 using MUSCLE (Edgar, 2004) with default parameters. For protein sequences, a neighbour joining-tree was produced using p-distance to model amino acid substitution; the rate of substitution was assumed to be uniform and the pattern among lineages homogenous; gaps or missing data were deleted in the analysis. For DNA

sequences, a minimal evolution tree was constructed using p-distance to model nucleotide substitutions; only transitions were included while the rate of substitution was assumed to be uniform and homogenous across lineages; gaps or missing data were deleted from the analysis. Trees were bootstrapped using 1000 replicates.

### 5.2.7 Homology Modelling

The crystal structure of Enc from *T. maritima* (pdb code 3dkt) (Sutter *et al*, 2008) was used as a template; *Msm* and *Mtb* Enc are 30.6% and 31.3% sequence identical, respectively. The models were produced in UCSF-Chimera (Petterson *et al*, 2004) using MODELLER (Sali & Blundell, 1993). No attempt was made to model loops.

### 5.2.8 Electrostatic Potentials

Electrostatic potentials were visualised in *T. maritima* Enc (Sutter *et al*, 2008) using the Coulombic surface representation option in UCSF-Chimera (Peterson *et al*, 2004). Since the crystallisation conditions were conducted at pH 5.1 (Sutter *et al*, 2008), all histidines in the structure were protonated for accurate analysis. *Msm* and *Mtb* Enc homology models underwent the same procedure. Minor differences between electrostatic potential maps were noted between this work and that produced by Sutter *et al* (2008), likely due to the fact that the Poisson-Boltzmann equations were utilised in their program which is known to be more accurate (Gorham *et al*, 2011).

## 5.3 Results & Discussion

### 5.3.1 The Primary Cargo of *Mycobacterium smegmatis* Encapsulin Is Dye-Decolourising Peroxidase

Although *Msm* Enc has C-terminal extensions for both DyP and BrfB, presence of either cargo had not yet been confirmed. Examination of particles from negative stain EM showed the presence of a small protein with preference towards one-side of Enc, which may be the result of binding on the 3-fold axis (**Figure 5.3.1-1a**). Native mass spectrometry of *T. maritima* and *B. linens* Encs showed that only a single-hexameric DyP is accommodated (Snijder *et al*, 2016), which fits with geometric constraints observed. The cargo of *Msm* Enc was reconstructed through common lines by masking out the Enc shell through a rectangular mask with a Gaussian drop-off (**Figure 5.3.1-1a**). Particles which looked empty, broken, or may contain BrfB cargo where not included; this left 207 particles. Most of the particles present were deformed or broken, while much fewer contained possible BrfB cargo than DyP cargo. Since DyP is located immediately up-stream from Enc in the genome, while BrfB is much further away, its greater rate of encapsulation would be consistent with the spatial organisation in the genome. In addition, it is not clear that the stress conditions imposed would necessarily induce the production of encapsulated BrfB, given that it likely functions to alleviate iron toxicity in *Mtb* (Khare *et al,* 2017).

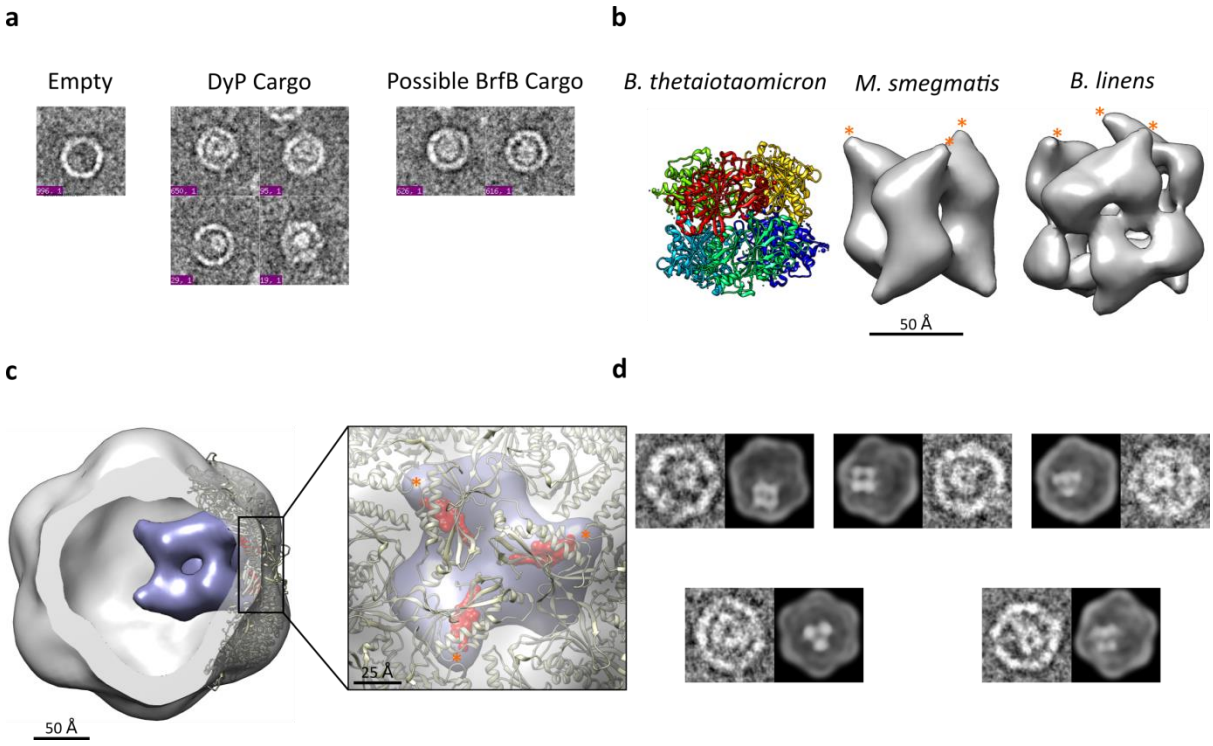



**Figure 5.3.1-1. Primary cargo of *Msm* Enc.** a) *Msm* Enc particles can either be empty or contain DyP or BrfB cargo. The main differentiating feature between the cargoes is that BrfB almost fills the entire lumen at all orientations, while DyP does not. BrfB also appears square-like, owing to its octahedral symmetry, while the 3-fold and 2-fold symmetries of DyP can be seen depending on the particle orientation. b) Prelimary reconstruction of encapsulated DyP (middle) compared to other solved DyP crystal (left, pdb code 2gvk) or low-resolution EM (right, emd-1530) (Sutter *et al*, 2008) structures. The position of the C-terminal extension is starred (orange). Note that the C-terminal extension for *B. thetaiotaomicron* DyP was not built into the crystal structure. c) Prelimary model of DyP (purple) docked into *Msm* Enc (grey), which fills a substantial part of the Enc lumen. It is clear from the docking that only a single DyP hexamer is likely to be accommodated in the Enc lumen. The C-terminal extension from the *T. maritima* crystal structure (brown) (pdb code 3dkt) (red) aligns well with the location of the DyP C-terminal extension (starred, orange). d) Phantom view of the preliminary model of DyP docked into *Msm* Enc shows a good correspondence to Enc negative stain particles at different orientations.

Comparison of this preliminary structure with other solved DyP structures (**Figure 5.3.1-1b**) show that its C-terminal extension lies at an angle to its 3-fold axis, in a similar arrangement to *B. linens* DyP. The arrangement of subunits is most similar to *B. thetaiotaomicron* DyP, although they lie at a much steeper angle; this could be due to bias in the reconstruction or to conformational changes upon binding to Enc. Since so few particles were used in the

reconstruction, there is a much higher-chance of bias being present in the model and hence the reconstruction is most likely not entirely correct.

Docking of DyP to the inside of *Msm* Enc shows a tight correspondence between the probable site of the C-terminal extension and the symmetry of the 3-fold pore (**Figure 5.3.1-1c**). Since the C-terminal extension of DyP most probably binds to the pocket in *Msm* Enc in a similar manner to the Flp cargo in *T. maritima* Enc, this was likely masked out during the reconstruction process. However, the docked model shows that it is feasible for the DyP C-terminal extension to bind with a similar mechanism to the Flp cargo, with the symmetry of DyP matching that of the 3-fold pore. Phantom views of the docked DyP model show a good correspondence to experimental Enc particles (**Figure 5.3.1-1d**) at different orientations. Thus, there is some evidence that DyP binds to the 3-fold pore of Enc.

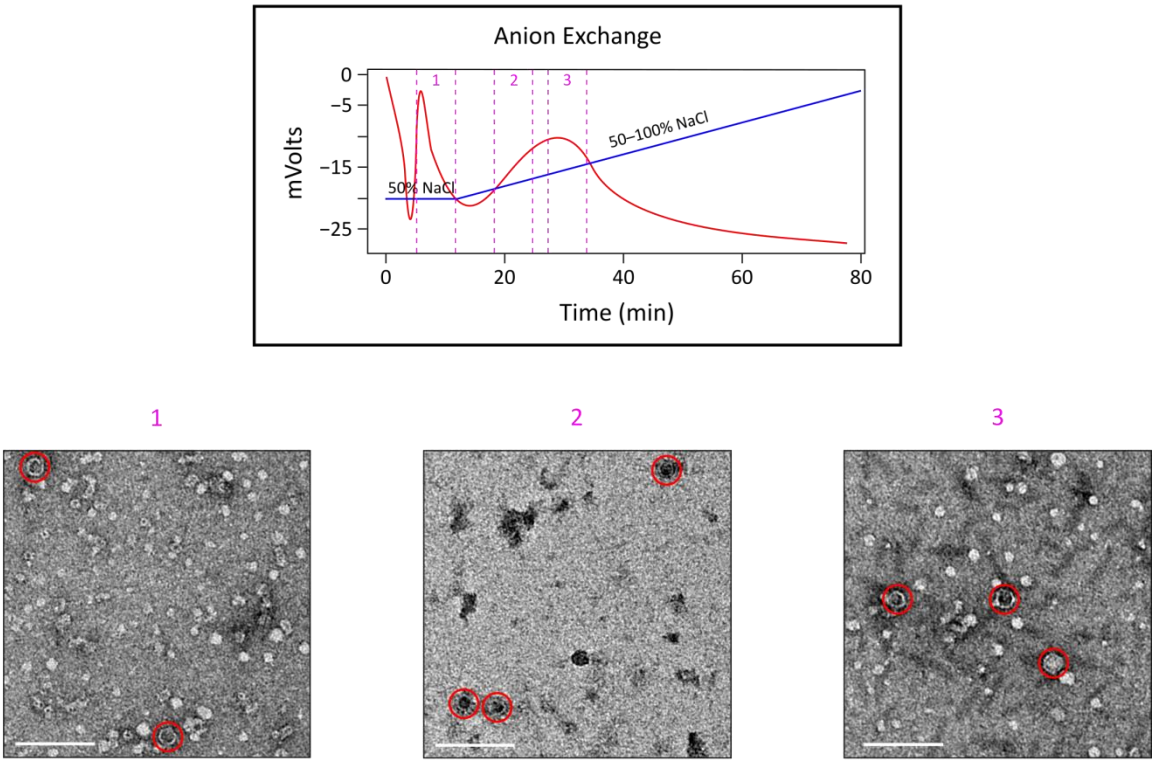
### 5.3.2 *Mycobacterium smegmatis* Encapsulin is Exported

Since there is previous evidence suggested that *Mtb* Enc is exported to the cell culture medium during growth (Rosenkrands *et al*, 1998), this was investigated for *Msm* Enc. A 50 mL volume of cell culture was grown and the cell supernatant fractionated by anion exchange; fractions were examined under the electron microscope and SDS-PAGE (**Figure 5.3.2-1**). A rough calculation of the number of Enc particles found in the cell culture supernatant compared to the cell lysate using the electron micrographs (accounting for differences in magnification, concentrations used, minor differences in purification methodology, etc) found that there are approximately 12 times more Enc particles in the cell culture supernatant than the cell lysate. Although DyP was detected with Enc in two out of the three anion exchange fractions run on the SDS-PAGE gel, BrfB was not (**Figure 5.3.2-1b**). Since BrfB has 4 times as much protein per complex compared to DyP based on their respective symmetries (24 vs 6 subunits per complex), if BrfB was encapsulated at a similar rate to DyP it would be expected to appear on the SDS-PAGE gel along with Enc and DyP. These results suggest that BrfB is encapulsated at a much lower rate than DyP, at least under the growth conditions tested.

Surprisingly, there were many other proteins present in the cell culture supernatant which appeared on the electron micrographs (**Figure 5.3.2-1a**). The appearance of Enc and other proteins in the cell culture supernatant suggest that Enc may be exported during *Msm* growth; the culture was only grown to mid-log phase and so autolysis of cells is not expected to contribute heavily to the presence of Enc and other proteins. However, since samples of *Msm* were not plated out to track the survival of the bacteria as they grew in the culture media, it is possible that significant auto-lysis occurred which would not show on the $OD_{600}$ reading. Thus, it is possible that the proteins observed in the cell culture supernatant are the result of bacterial autolysis.

All bacteria use a variety of pathways to export proteins to the extracellular environment or to the cell wall (see Kostakioti *et al* (2005) for a review). The *Mtb* export mechanisms are under considerable interest since there is evidence that a specific export pathway, the SecA2 system, has a role in promoting *Mtb* virulence by allowing the export of specific proteins which prevent phagosome maturation (Sullivan *et al*, 2012). In *Mtb*, the majority of proteins are exported via the SecA1 and Tat pathways, while a small number are exported via the SecA2 and ESX pathways. The SecA1 pathway functions as a housekeeping mechanism as it is essential for Mycobacterial survival, while the other export pathways are required for *Mtb* virulence or drug resistance (Lignon *et al*, 2012). All export pathways utilise an N-terminal signal peptide or motif to recognise proteins, although for the SecA2 system not all exported proteins have a signal peptide. Folded proteins are usually exported via the Tat pathway in contrast to the SecA1 pathway which only recognises unfolded proteins (Lignon *et al*, 2012). In addition, the SecA2 pathway may also be capable of exporting folded proteins (Feltcher *et al*, 2012). *Msm* and *Mtb* Cfp29 (Enc) lack a signal peptide (Rosenkrands *et al*, 1998), although this is not necessarily required for export. Considering that other proteins involved in the oxidative stress response, such as catalase (KatG) and superoxide dismutase (SodA), are exported by the SecA2 system (Lignon *et al*, 2012), and that this pathway can export folded proteins even without a signal peptide (Feltcher *et al*, 2012), Enc and encapsulated cargo may be exported specifically via this pathway. This has interesting ramifications for exploring whether or not Enc and encapsulated cargo are among the critical exported proteins required for *Mtb* survival in the host macrophage.

**a**




**b**

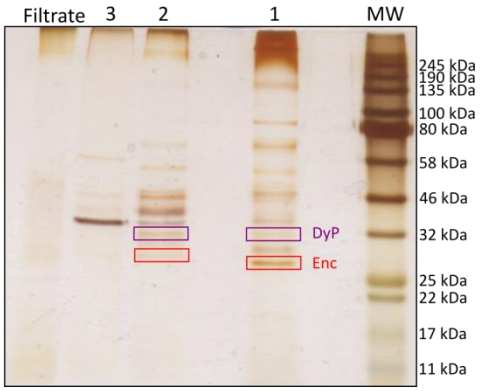



**Figure 5.3.2-1. Export of Enc**. a) Cell culture filtrate was separated by anion exchange which yielded three fractions (1–3, pink) which contained Enc (red circle). Images were taken at a magnification of x53,000 on the T20 Technai TEM. Scale bars show 100 nm. b) A silver-stained 8–15% gradient SDS-PAGE gel showed that fraction 1 contained the predominant amount of Enc. A mass corresponding to that of BrfB (20 kDa) was not visible. Since BrfB is present in 24 copies while DyP has 6 copies per biological unit, based on their respective symmetries, and DyP is clearly present in the gel, this suggests that DyP is encapsulated at a greater rate than BrfB.

### 5.3.3 Phylogeny

A BLAST search using *Msm* or *Mtb* Enc amino acid sequence was conducted to find other possible Encs in *Mycobacteria*. Many of these are labelled as bacteriocins and as such may have possible anti-microbial activity. *B. linens* Linocin was originally classified as a bacteriocin based on its ability to inhibit growth of *Listeria* (Valdés-Stauber *al*, 1991). Interestingly, while Linocin (Enc) from *B. linens* could inhibit the growth of *Listeria monocytogenes* strains to varying degrees, no inhibition was observed by Cfp29 (Enc) from *Mtb* (Rosenkrands *et al*, 1998). In addition, purified *B. linens* Enc or Enc with bound DyP shows no bacteriocin activity against *Listeria ivanovii* (Sutter, PhD Thesis, 2008). The bacteriocin from *Msm* has been shown to have anti-tumour properties in cell culture (Saito & Watanabe, 1979; Saito & Watanabe, 1981). This purified bacteriocin is probably not Enc since its predicted MW based on migration in gel filtration is considerably smaller (75 kDa) (Saito *et al*, 1979) than is expected for its known large size (1.7 MDa). Bacteriocins shown to inhibit *Mtb* growth are small peptides (~3−6 kDa) (Sosunov *et al*, 2007). The mode of action for some small bacteriocins is known, involving interference in DNA, RNA, and protein metabolism (Cotter *et al*, 2013). It is not known how Encs such as the one from *B. linens* exerts its possible anti-microbial activity. Thus, it is possible that Encs do not have anti-microbial activity. A phylogenetic tree was produced for the putative Mycobacterial Encs (**Figure 5.3.3-1**).

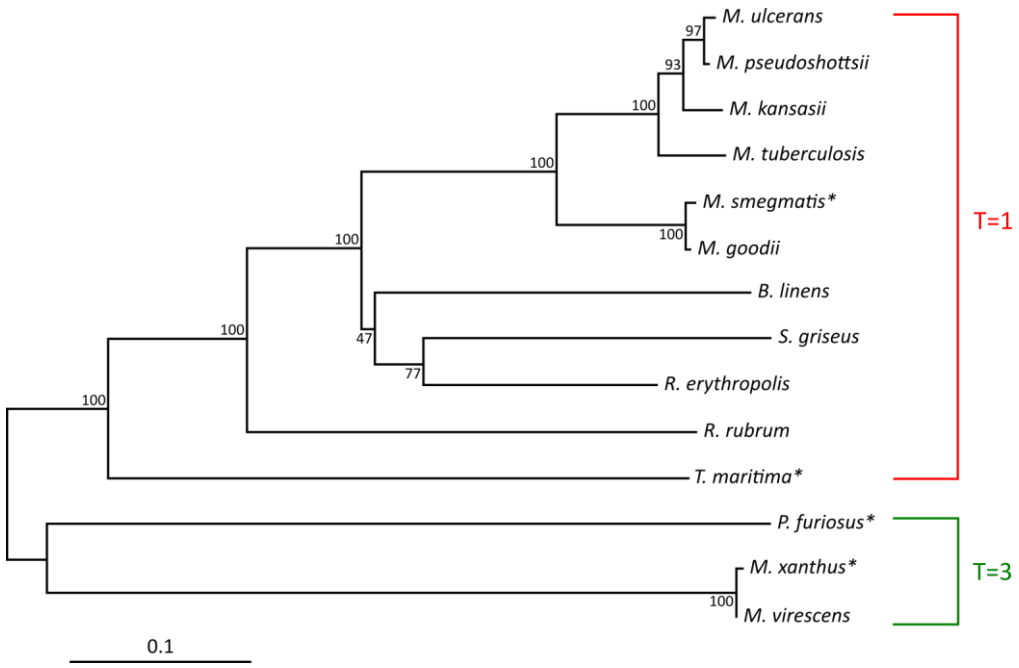



**Figure 5.3.3-1. Unrooted neighbour-joining tree for putative Mycobacterial Encs.** Mycobacterial Encs shown, including *Msm* and *Mtb*, are part of the T=1 icosahedrons. Only four Enc structures have been solved to date (*). Bootstrap values are given for 1000 replicates. Scale shows 0.1 amino acid substitutions.

As can be seen in **Figure 5.3.3-1**, T=1 and T=3 Encs inhabit distinct clusters. T=1 icosahedrons have a 5-fold, 3-fold, and 2-fold axis, while T=3 icosahedrons have an additional quasi-6-fold axis. T refers to the triangulation number in icosahedrons with respect to packing arrangements of subunits required to generate icosahedrons of more than 60 subunits (Prasad & Schmid, 2012) (see **Figure 5.1-1**). The T=3 Enc of *P. furiosus* consist of three icosahedrally independent subunits, with root mean squared (r.m.s) deviations in their structures of 1−1.33 Å (Akita *et al*, 2007). The changes required to accommodate more subunits in T=3 icosahedrons occurs across all three regions of the monomer (**Figure 5.3.3-2**). *P. furiosus* Enc accommodates the quasi-6-fold interaction in the A-domain, where *T. maritima* Enc has its 5-fold interaction, by shortening α6 by one turn and lengthening Loop13 by 4 amino acids compared to α8 and Loop12 in *T. maritima* Enc. The 5-fold interaction for *P. furiosus* Enc also occurs in the A-domain; in the crystal structure, the interaction at the 5-fold in the A-domain has a break in Loop13 which connects α6 and α7, most likely as a result of conformational flexibility to accommodate a 5-fold rather than 6-fold interaction. The 3-fold interaction for both *P. furiosus* and *T. maritima* Enc occurs in the P-domain, whereas the 2-fold interaction occurs across the E-loop and P-domain (**Figure 5.3.3-2c**). The positions of the

E-loop is the most striking difference between *P. furiosus* and *T. maritima* Enc and forms a much tighter 2-fold interaction the latter structure (Sutter *et al*, 2008) (**Figure 5.3.3-2c**).

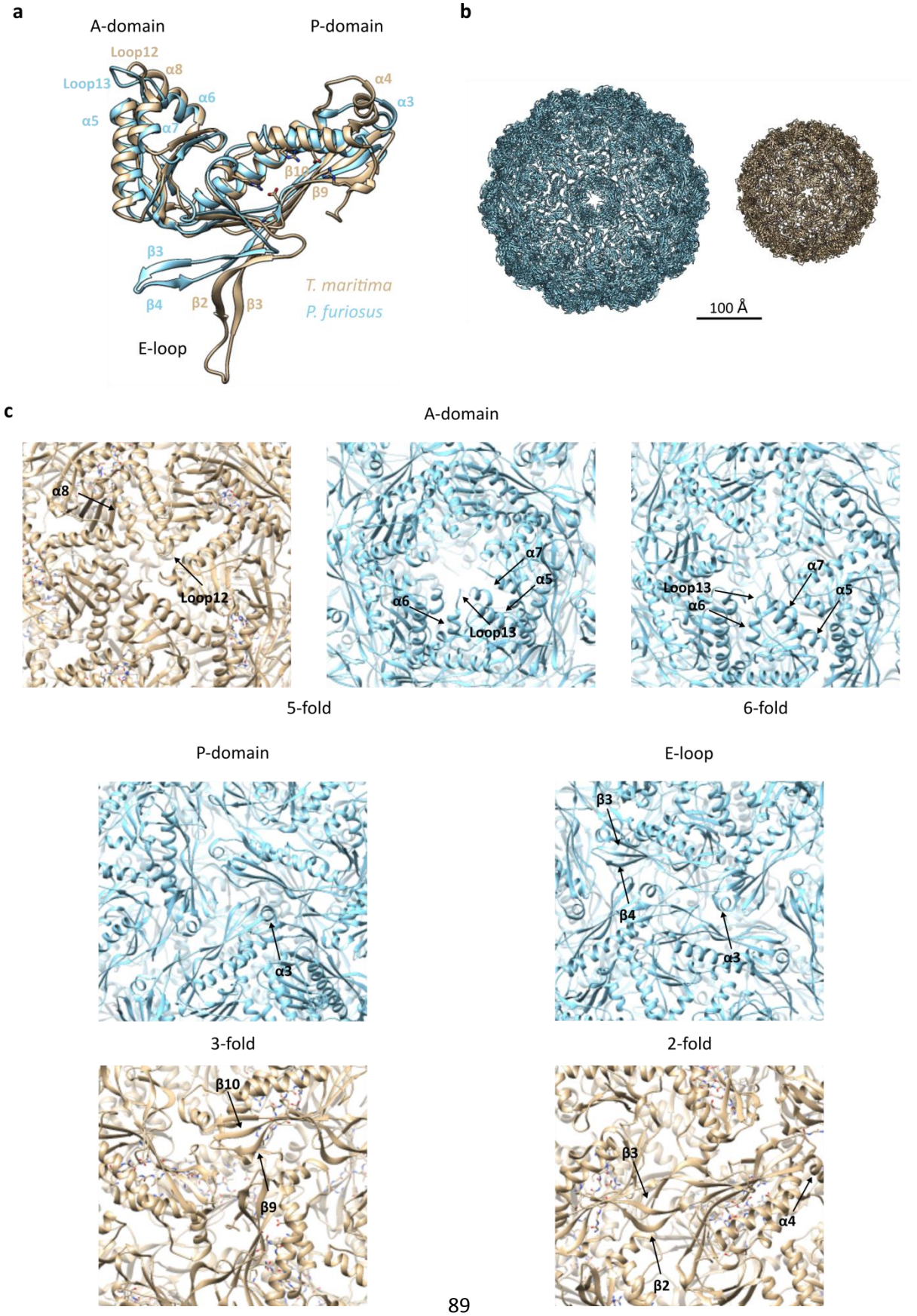


**Figure 5.3.3-2 (previous page). T=1 and T=3 Icosahedral Encs.** a) The monomer of each Enc is divided into three domains: the axial (A) domain, the peripheral (P) domain, and the extension (E) loop (Akita *et al*, 2007). b) Comparative sizes of T=1 (*T. maritima* Enc) (pdb code 3dkt) (Sutter *et al*, 2008) and T=3 (*P. furiosus* Enc) (pdb code 2e0z) (Akita *et al*, 2007) icosahedrons. c) Interactions occurring across the A-domain form the 5-fold axis and additional quasi 6-fold axis, while the P-domain and E-loop form the 3-fold and 2-fold axis.

### 5.3.4 Cargo Binding

For this section, we will go through a detailed explanation of the binding mechanism of the C-terminal extension proposed by Sutter *et al* (2008). Since there is some evidence that *Msm* Dyp binds at the 3-fold axis of Enc, we will also examine how this mechanism could apply to the cargo proteins of *Msm* and *Mtb* Enc. This section draws heavily on **Figure 5.3.4-1** and **Figure 5.3.4-2**, and so careful evaluation of these figures is recommended.

The C-terminal extension is quite conserved for DyP and Flp cargoes across a range of species (Sutter *et al*, 2008). However, there are substantial differences in the C-terminal extension sequences for the BrfB or FolB cargoes (Giessen *et al*, 2007) (**Figure 5.3.4-1a**). This relative conservation of the DyP C-terminal extension is likely to be a result of more evolutionary constraints imposed upon the DyP protein compared to the BrfB and FolB cargoes, rather than a specific selective pressure for DyP C-terminal extensions to remain conserved. From the phylogenetic trees constructed for all three cargo proteins (**Figure 5.3.4-1a**), BrfB and FolB undergo approximately twice the amino acid substitution rate as DyP, which indicates fewer constraints over the entirety of their protein sequences.

As can be seen in the phylogenetic tree for the BrfB cargo, there is a distinct lack of C-terminal sequences for BrfA or (heme-binding) bacterioferritin (**Figure 5.3.4-1a**). In addition, it appears that *T. maritima* contains two Flps; one (here named Flp1) which clusters with C-terminal extension containing BrfBs (**Figure 5.3.4-1a**), and one (here named Flp2) found by Sutter *et al* (2008) that is not related to either BrfA or BrfB shown in the Ferritin phylogenetic tree (**Figure 5.3.4-1a**). Furthermore, Sutter *et al* (2008) noted that Flp2 has 56% homology to *N. europaea* Flp, whose structure had been solved by X-ray crystallography (Chang *et al*, 2005) and which they subsequently used to propose that Flp2 also binds on the 5-fold axis of Enc under the assumption that their quaternary structures are conserved. In support of this assumption,

Flp2 was detected by MS analysis of dissolved crystals of native *T. maritima* Enc and thus the 10 peptide C-terminal extension found in the Enc crystal structure is likely to be from encapsulated Flp2 (Sutter *et al*, 2008).

Recently, it was found that *R. rubrum* Enc could encapsulate an Flp, distinct from its heme-binding bacterioferritin which is not encapsulated (**Figure 5.3.4-1a**), which is 53% sequence identical to *T. maritima* Flp2 (He *et al*, 2016). It was also found to form a decameric structure with 5-fold symmetry (He *et al*, 2016). Thus, it appears to be that the evolution of BrfB and non-related ferritins to be encapsulated by Enc, as shown by the lack of C-terminal extensions for BrfA and other related ferritins, could have functional significance.

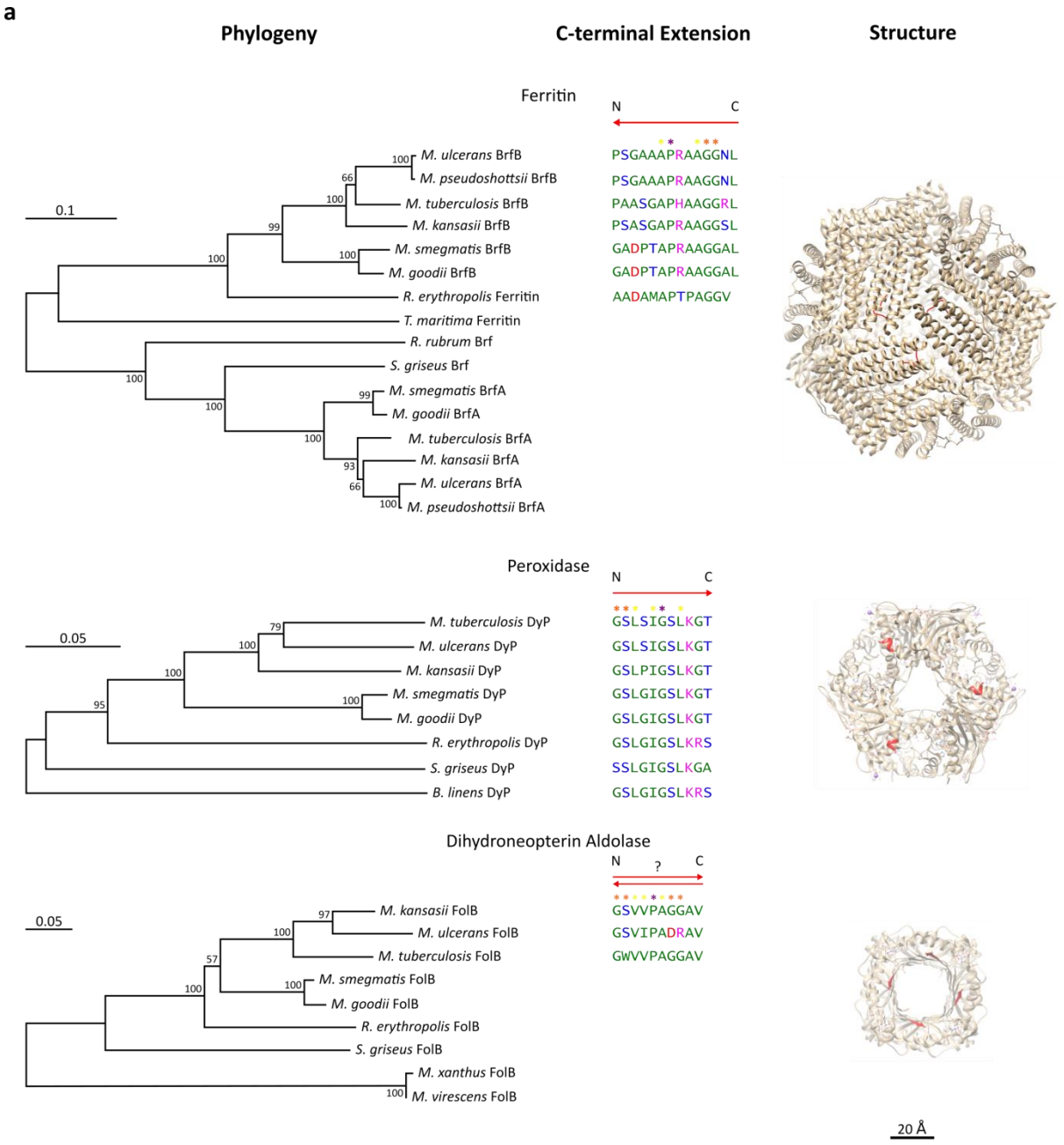

a
Phylogeny
C-terminal Extension
Structure
Ferritin
N
C
M. ulcerans BrfB
M. pseudoshottsii BrfB
M. tuberculosis BrfB
M. kansasii BrfB
M. smegmatis BrfB
M. goodii BrfB
R. erythropolis Ferritin
T. maritima Ferritin
R. rubrum Brf
S. griseus Brf
M. smegmatis BrfA
M. goodii BrfA
M. tuberculosis BrfA
M. kansasii BrfA
M. ulcerans BrfA
M. pseudoshottsii BrfA
0.1
PSGAAAPRAAGGNL
PSGAAAPRAAGGNL
PAASGAPHAAGGRL
PSASGAPRAAGGSL
GADPTAPRAAGGAL
GADPTAPRAAGGAL
AADAMAPTPAGGV
Peroxidase
M. tuberculosis DyP
M. ulcerans DyP
M. kansasii DyP
M. smegmatis DyP
M. goodii DyP
R. erythropolis DyP
S. griseus DyP
B. linens DyP
0.05
GSLSIGSLKGT
GSLSIGSLKGT
GSLPIGSLKGT
GSLGIGSLKGT
GSLGIGSLKGT
GSLGIGSLKRS
SSLGIGSLKGA
GSLGIGSLKRS
Dihydroneopterin Aldolase
?
M. kansasii FolB
M. ulcerans FolB
M. tuberculosis FolB
M. smegmatis FolB
M. goodii FolB
R. erythropolis FolB
S. griseus FolB
M. xanthus FolB
M. virescens FolB
0.05
GSVVPAGGAV
GSVIPADRAV
GWVVPAGGAV
20 Å
b
M. tuberculosis
M. kansasii
M. ulcerans
M. pseudoshottsii
M. goodii
M. smegmatis
R. erythropolis
B. linens
S. griseus
R. rubrum
M. xanthus
M. virescens
T. maritima
P. furiosus
0.05
Corynebacteriales
Micrococcales
Streptomycetales
Alphaproteobacteria
Deltaproteobacteria
Thermotogales
Thermococcales
Actinobacteria
Proteobacteria
Thermotogae
Euryarchaeota
Bacteria
Archaea

**Figure 5.3.4-1 (previous page). Cargo proteins of Enc.** a) The phylogeny, C-terminal extension, and structure are given for the three known cargo proteins of *Mtb*. Binding of the cargo protein to the inside of Enc is determined by the C-terminal extension, which is dominated by non-polar amino acids (green) with interspersed with mostly conserved polar (blue), positively charged (pink), or negatively charged (red) amino acids. The direction of binding is determined by two N- or C-terminal residues (orange star) while a central residue (purple star) separating two hydrophobic residues (yellow star) aids in positioning. The direction of binding for the FolB cargo is ambiguous. Binding of the C-terminal extension (red) is hypothesised to occur along either the 3-fold or 4-fold axis of the cargo protein. Ferritin cargo protein may also bind along its 2-fold axis (not shown). Note that the C-terminal extension is only visible for *Mtb* ferritin (pdb code 3uno) and was not built into the crystal structure of peroxidise (pdb code 2gvk) (Zubieta *et al*, 2007) and was cleaved from FolB (pdb code 1nbu) (Goulding *et al*, 2005). Also note that the peroxidase shown is from the closest structural homolog, *B. thetaiotaomicron*. b) Phylogenetic relationship between organisms that harbour known and putative Enc and cargo proteins, based on 16S rRNA gene sequence. While the peroxidase cargo is found in the Actinobacteria phylum, the ferritin cargo is restricted to the Corneybacteriales order, and the FolB cargo is specific to slow-growing Mycobacteria. For the phylogenetic trees, scale bars show amino acid or nucleotide substitutions.

The C-terminal extension is composed of mainly non-polar amino acid residues interspersed with mostly conserved charged or polar residues (**Figure 5.3.4-1a**). Binding of the C-terminal extension in *T. maritima* Enc occurs in a hydrophobic pocket comprised of Phe30, Leu34, Leu233, and Ile249 stabilised by a salt bridge between Arg37 and Asp232. The extension binds in a specific direction determined by the two N-terminal Gly residues, and is registered through two central hydrophobic residues (Leu and Ile) separated by a Gly (Sutter *et al*, 2008). The binding of the C-terminal extension into the *T. maritima* Enc pocket is shown in **Figure 5.3.4-2**. From this figure, the *Mtb* and *Msm* Enc homology models superimposed on the *T. maritima* Enc structure show that salt bridge between Arg37 and Asp232 is conserved and the hydrophobic pocket is mostly conserved. Minor changes to amino acid residues lining the hydrophobic pocket could reflect a shift in C-terminal binding for the cargoes, or they could be expected due to sequence drift and thus carry no functional significance.

The DyP C-terminal extension is orientated by a Gly-Ser or Ser-Ser on the N-terminal end, and there is an additional hydrophobic residue (Leu) on the N-terminal end separated from the second hydrophobic residue (Ile), which sits next to the central Gly, by either a Ser, Gly, or Pro (**Figure 5.3.4-1a**). In addition, the third hydrophobic residue (Leu) on the C-terminal end is wedged between a conserved Ser and Lys (**Figure 5.3.4-1a**).

Since Ferritin C-terminal extension has a completely conserved double Gly residues on their C-terminal end, they likely bind in the opposite direction to DyP cargo (**Figure 5.3.4-1a**). Furthermore, the "kink" induced by the central Gly (Sutter *et al*, 2008) (**Figure 5.3.4-2**) is performed instead by a conserved proline; the central hydrophobic residue on the C-terminal end (Ala) is separated from the proline by a positively charged or polar amino acid, while the other conserved central hydrophobic residue (Ala) sits next to the central Pro on the N-terminal end (**Figure 5.3.4-1a**).

FolB C-terminal extension differs substantially from that found in the *T. maritima* Enc crystal structure and the DyP or Ferritin cargoes; there are none or few charged or polar residues and two sets of N- and C-terminal residues which could serve to orientate the sequence (**Figure 5.3.4-1a**). Furthermore, there is a Pro which separates the hydrophobic residues Ala and Val/Ile and so could act as a register in either direction (**Figure 5.3.4-1a**). Thus, we hypothesise that this ambiguity in the direction of binding may aid in placement of FolB at a pseudo 4-fold axis (see below). It is also possible that the cargo is not held at a specific pore or that binding may occur in a different pocket.

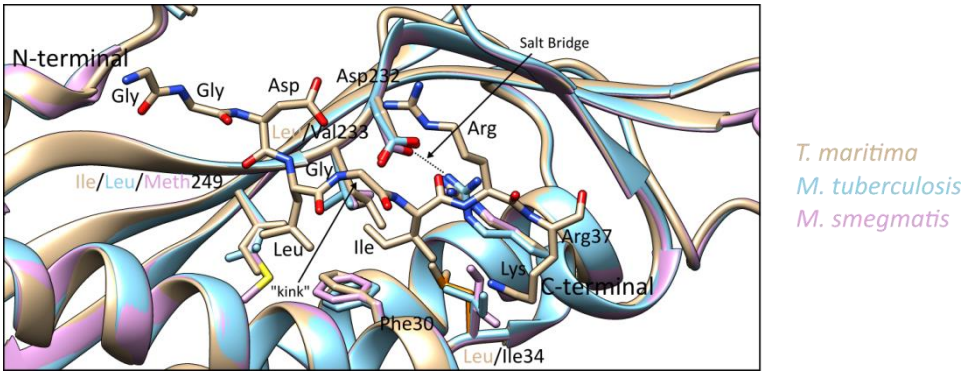



**Figure 5.3.4-2. Binding of the C-terminal Extension**. The C-terminal extension in the *T. maritima* Enc crystal structure binds to a hydrophobic pocket consisting of Phe30, Leu34, Leu233, and Ile249; the pocket is stabilised by a salt bridge between Arg37 and Asp232. Direction of binding is determined by two N-terminal glycine residues, while two hydrophobic residues (Ile and Leu) separated by a central glycine act as the register (Sutter *et al*, 2008). The "kink" introduced by the central glycine seems to act in positioning the extension, by bringing the two hydrophobic residues into close contact with hydrophobic residues lining the pocket. *Mtb* and *Msm*

homology models for Enc retain the salt bridge and hydrophobic nature of the pocket, with changes in the amino acid type reflecting those required to accommodate a different C-terminal extension.

### 5.3.5 Pore Selectivity

This section examines possible mechanisms for control of substrate entry, based on pore electrostatics. From the previous section, we hypothesised that the cargo proteins of *Msm* and *Mtb* Enc bind at a specific pore symmetry based on prelimary evidence that DyP could bind to the 3-fold pore of *Msm* Enc, and inspected how such binding could occur using inference from bioinformatics. Now, we examine the relationship between substrate entry and cargo binding. We also examine the mechanisms of substrate control for three other icosahedral proteinaceous compartments, which suggests common mechanisms for control of substrate entry.

The C-terminal extension binds in different orientations in the *T. maritima* Enc structure, based on the pore symmetry. **Figure 5.3.5-1a** shows that the N- and C-terminal ends of the extension lie in opposite orientations for the 5-fold and 3-fold axis in Enc. Likewise, it can be seen from the DyP and BrfB C-terminal extension sequences that they could bind in opposite directions based on analysing the sequences and comparing it to the known mechanism of binding, as shown in the previous section. However, BrfB also has 3-fold symmetry which poses the question of the orientation of its C-terminal extension on the 3-fold pore. Since the pocket is the same, and can clearly accommodate either C-terminal extension orientation as shown by the differences in binding between the 5-fold and 3-fold pores in *T. maritima* Enc (**Figure 5.3.5-1a**), it is not impossible for the BrfB extension to bind in the opposite direction of the DyP extension. As mentioned previously, the FolB C-terminal extension is only available at its 4-fold axis (**Figure 5.3.4-1a**); this could bind to a pseudo 4-fold axis in Enc as the distances correspond (**Figure 5.3.5-1a**). This hypothesis would explain the ambiguous directionality present in the FolB C-terminal extension; since the pseudo 4-fold axis is composed of two sets of two-fold axes which lie in opposite directions, the binding of the FolB C-terminal extensions must be flexible enough to be able to accommodate either orientation (**Figure 5.3.5-1a**).

Although Sutter *et al* (2008) noted differences in electrostatic potentials between the pores on the outside of Enc, there are also differences in electrostatic potentials between the inside

and outside of the Enc lumen (**Figure 5.3.5-1b**). The 5-fold pore is mostly neutral on the outside but positively charged on the inside, while the 3-fold pore is slightly negatively charged with positively charged lysines lying perpendicular to the pore on the outside which changes to being neutral on the inside. The 2-fold pore does not seem to discriminate in the electrostatic charges between the inside and outside of the Enc lumen (**Figure 5.3.5-1b**). The changes in electrostatic potential suggest a hypothesis for controlling substrate entry and exit; that the substrate can enter through a pore but that the changes in electrostatic potential prevent the substrate from leaving the Enc lumen.

Examination of the electrostatic potentials in *Msm* and *Mtb* Enc homology models show that they are not substantially different to that of *T. maritima* Enc (see **Figure 8-9** in **Appendix**); they both lack the positively charged lysines on the 3-fold pore, but retain a negative charge on the outside and a positive charge on the inside of the 5-fold pore. In addition, since the 2-fold pores lie close to the flexible E-loop region, there may be substantial differences in pore size which is difficult to assess through a homology model.

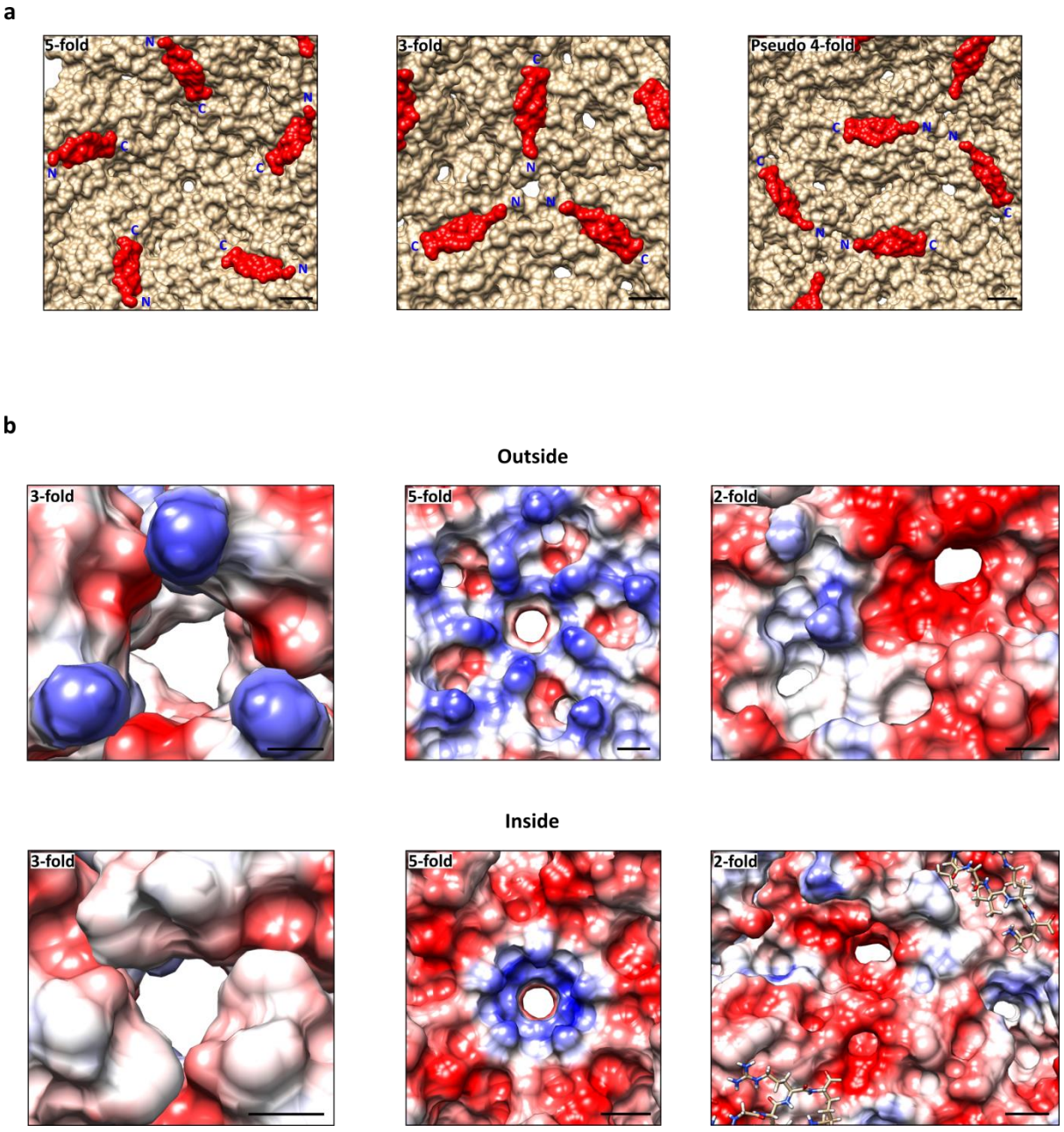



**Figure 5.3.5-1. Relationship between pore selectivity and C-terminal extension binding.** a) Positioning of C-terminal extension (red) around the symmetry-related pores. The orientation of the 5-fold and 3-fold pores differs, while a pseudo 4-fold binding could feasibly occur. Scale bars show 10 Å. b) Electrostatic potential of amino acids around the outside and inside of the pores of *T. maritima* Enc crystal structure (Sutter *et al*, 2008) (pdb code 3dkt), coloured from a red (-10 kcal/mol·e$^-$) to blue (+10 kcal/mol·e$^-$) gradient. White indicates no charge. For the 5-fold and 3-fold pores, there are clear changes in the charges of amino acids from the outside to the inside, suggesting a mechanism for allowing substrates to enter and preventing exit from the lumen. Scale bar shows 5 Å.

The differences in pore electrostatic potentials and cargo binding lead to the hypothesis that each pore has evolved to admit a specific substrate into the Enc lumen. However, since the substrate of DyP, hydrogen peroxide, could react with $Fe^{2+}$, which binds to BrfB, via the Fenton reaction to produce harmful hydroxyl radicals (Smith, 2004), it does not seem likely that these substrates would be admitted to the Enc lumen at the same time. But both substrates are small enough (<2 Å) to access a neutral or negatively charged (in the case of $Fe^{2+}$) pore, thus opening the question of how Enc controls substrate access beyond the chemistry of the pore. In addition, the 7,8-dihydroneopterin substrate of FolB and ABTS dye substrate of DyP are too large (>5 Å) to fit through any of the *T. maritima* Enc pores. However, FolB and DyP retain their activity while encapsulated (Contreras *et al*, 2014), suggesting further mechanisms of substrate access.

Some have hypothesised that Enc undergoes disassembly to allow for substrate access in the case of lignin (Rahmanpour & Bugg, 2013), but this does not seem energetically likely and would not solve the problem of the production of harmful hydroxyl radicals. Instead, it is possible that a better hypothesis involves binding of the cargo at a specific pore which induces conformational changes in *Msm* and *Mtb* Enc, such that the pore allows substrate access for the bound protein while closing other pores where no binding has occurred.

Controlling substrate access is a problem for RNA containing viruses, which must shield their dsRNA genome from the host cytoplasm in an icosahedral protein shell to avoid host defence mechanisms, but still admit substrates to gain access and products to leave in order to complete its life-cycle (Bamford, 2002). Coordination of transcription of the dsRNA genome of these viruses, such as the Bluetongue virus, is tightly regulated, involving the enzymes of transcription and the proteins composing the viral capsid (Diprose *et al*, 2001). The mRNA produced during transcription is exported from the protecting icosahedral shell through the 5-fold pore; this pore is composed of positively charged, basic amino acids and undergoes expansion in the presence of $Mg^{2+}$, which activates all pores in the viral capsid. Interactions between the amino acids lining the pore and mRNA ensure that exit is very specific, preventing the 5'-end of the mRNA from forming tertiary structure before it has left as this would block the pore. Other pores in the icosahedral shell allow for entry of substrates and exit of reaction by-product; a pore between trimers of the capsid could allow nucleotides to enter, while the 2-fold pore appears to allow for the exit of phosphate (Diprose et al, 2001).

As mentioned previously, carboxysomes are large proteinaceous compartments which sequester enzymes and substrates involved in carbon fixation reactions, such as ribulose bisphosphate carboxylase oxygenase (RuBisCO) and carbonic anhydrase (Kerfeld *et al*, 2005). The hexamers composing carboxysomes have 6-fold, 3-fold, and 2-fold pores with diameters of ~7 Å, ~6 Å, and ~4 Å, respectively. These pores allow for the transport of negatively charged metabolites, such as bicarbonate and other substrates, but not uncharged molecules such as carbon dioxide and oxygen (Kerfeld *et al*, 2005). The enzyme carbonic anhydrase converts intracellular bicarbonate to $CO_2$, and thus the shell acts to keep $CO_2$ in a high-enough concentration in the compartment for RuBisCO to fix while preventing oxygen from competing for binding sites (Dou *et al*, 2008). Further investigation established that carboxysomes could be composed of pentamers in addition to the hexamers solved previously, forming an icosahedral polyhedron. The pentamers interact via their C-terminal regions and form pores at their 5-fold axis of symmetry ~ 3.5–5 Å in diameter. Many of the packing arrangements for hexameric and pentameric subunits are unresolved, but basic geometry suggests that pentamers act as vertices and the hexamers constitute the edges of the icosahedron; one such packaging arrangement features additional pores at the interface between pentamer and hexamer (Tanaka *et al*, 2008). The carboxysome from *Prochlorococcus* is capable of operating a gated pore in its pseudo-hexamer subunits; the 3-fold pore can be open (~14 Å in diameter) or closed depending on the orientations of the side chains of two conserved residues, Glu120 and Arg121. The presence of the larger pore in the *Prochlorococcus* carboxysome is thought to be a result of evolutionary pressure in this species to ease the passage of larger substrates, such as ribulose biphosphate, which can only enter the smaller pores of other carboxysomes with difficulty (Klein *et al*, 2009).

Recently, putative components of a bacterial microcompartment (BMC) for *Msm* were solved by X-ray crystallography. The biological role of this BMC remains unknown, but the operons associated with the shell components include a short chain alcohol dehydrogenase, a class III aminotransferase, an aldehyde-alcohol dehydrogenase, and a protein distantly homologous to aminoglycoside phosphotransferases (Mallette & Kimber, 2017). Like other BMCs, such as carboxysomes, this BMC likely forms a large (~ 100 nm) icosahedron, with hexamers (composed of the shell protein MSM0272) forming the edges and pentamers (composed of the shell protein MSM0273) as the vertices. The pentamers contain a hydrophilic and

positively charged central 5-fold pore ~ 5 Å in diameter, while the hexamers contain a central 6-fold pore of a similar diameter characterised by a ΦGZGX motif, where Φ is a small hydrophobic residue, X is any residue, and Z is the critical pore-lining residue. For other BMCs, the pore lining residue tends to be small and polar, commonly composed of serine or glycine, but for *Msm* BMC the pore lining residue is aspartic acid. Two nearby glutamic acids ensure that the outside (convex side) of the pore is highly acidic and negatively charged. There are two other shell proteins in *Msm* BMC (MSM0271 and MSM0275) which form a dimer of trimers, necessitating a complicated packing arrangement as the proposed model. These two shell proteins have no accessible pores, implying that substrate access and any possible product exit is solely determined by the 5-fold and 6-fold pores (Mallette & Kimber, 2017). An alternative hypothesis was proposed by Mallette & Kimber (2017) whereby the flexibility of the shell proteins MSM0271 and MSM0275 suggests that asymmetric pores could form which would provide a greater degree of specificity than currently provided by the two known symmetric pores. Furthermore, these shell proteins could also act as gated pores for small molecules (Mallette & Kimber, 2017).

The above three examples of the regulation of substrate entry and product exit in icosahedral shells of dsRNA viral capsids and BMCs demonstrates that there is a tight-coupling between the chemical nature of the pore, and even pore size, with the chemistry of the substrates and products involved in the sequestered metabolic reactions. Thus, substrate control can be highly regulated through pore chemistry and size without recourse to energetically unfavourable disassembly of the icosahedron as proposed by Rahmanpour & Bugg (2013).

### 5.3.6 Gene Essentiality

We have examined how the mechanism of DyP, BrfB, and FolB encapsulation could occur and propose that the differentiation in electrostatic potentials between the inside and outside of the Enc lumen could control substrate entry and prevent exit. It was previously mentioned that *Msm* and *Mtb* Enc could be exported from the cell and the *Mtb* exports proteins to the host macrophage to aid in survival. Now, we examine the evidence for the essentiality of Enc and cargo proteins to the cell.

Recent evaluation of essential genes for *Mtb* growth *in vitro* using saturation transposon mutagenesis found that Cfp29 (Enc), DyP, and BrfB are non-essential while knock-out of FolB confers a growth defect (DeJesus *et al*, 2017). This contradicts with a previous study which found that Enc and FolB are essential for growth, while DyP and BrfB are non-essential (Sassetti *et al*, 2003). Gene essentiality was assessed by comparing growth of *Mtb* with transposon insertions with those harbouring the same genes only fluorescently tagged; if a gene is essential for growth then it should be significantly under-represented in the transposon insertion library than in the tagged library (Sassetti *et al*, 2003). However, the Enc gene was close to the cut-off applied to determine if a gene is essential. Furthermore, there is previous unrecognised bias in the *Himar1* transposon insertion process (DeJesus *et al*, 2017) which may have affected the previous results. However, the essential genes required for normal "housekeeping" functions may differ substantially to those required for a particular part of the *Mtb* life-cycle, such as a role in infection and propagation in host macrophages.

Survival of *Mtb* within the host macrophage requires a variety of strategies, from inhibiting fusion of the phagosome to the lysome, to hijacking host signalling pathways. *Mtb* can inhibit lysosomal fusion only in non-activated macrophages and hence prevent the production of bactericidal reactive oxygen species. However, *Mtb* can still survive within this hostile environment, through proteins involved in the oxidative stress response, such as KatG, which are exported to the macrophage (Pieters, 2008). Enc, DyP, BrfB, and FolB were not predicted to be essential for *Mtb* survival *in vivo* based on transcription differences, using saturation transposon mutagenesis, between *Mtb* infected mice and *Mtb* grown *in vitro* (Sassetti & Rubin, 2003). However, interpreting knock-out studies is difficult as compensatory effects can occur. In the data for Sassetti & Rubin (2003), BrfB appeared to compensate for the loss of BrfA as it had a greater number of transcripts *in vivo* than *in vitro* (the opposite was the case for BrfA). Since loss of BrfB means that *Mtb* cannot persist in infected mice (Pandey & Rodriguez, 2012), these results are consistent with a mechanism in which BrfB can compensate for loss of BrfA but BrfA cannot compensate for the loss of BrfB.

Analysis of mRNA transcription reads when *Mtb* was exposed to different *in vitro* stresses found that Enc is down-regulated under nitric oxide stress (Namouchi *et al*, 2016). On the other hand, analysis of transcripts following mice infected with *Mtb* over a 14-day period found that BrfB is one of the genes which are important in maintaining infection in the host

macrophage (Rohde *et al*, 2012). In addition, DyP is up-regulated in clinical strains of *Mtb* which are resistant to the antibiotic rifampin compared to wild type *Mtb* which lacks the resistance, although the difference is not statistically significant (Bisson *et al*, 2012).

Given the diversity of environments that *Mtb* must survive under, including the poorly understood physiological changes required to inhabit the granuloma (Pieters, 2008), it is feasible to presume that the export of Enc and encapsulated cargo is required at some stage of the life cycle, most likely under some kind of stress condition given that these genes are not essential for housekeeping functions, with the exception of FolB (see above).

The evidence we have examined previously suggest that DyP and BrfB are utilised under specific stress conditions, such as oxidative stress in the case of DyP and iron-toxicity in the case of BrfB, which would be consistent with results suggesting that they are not essential for normal cell functioning when no stress is present. However, given the importance of FolB in possible "house-keeping" functions of the cell, it is prudent to ask the question of why this enzyme is encapsulated, if the encapsulation only occurs under specific stress conditions, and whether it remains encapsulated or is released when conditions become more favourable. Considering that FolB undergoes oligomerisation upon substrate binding, forming an active octamer from an inactive tetramer (Goulding *et al*, 2005), it is possible that FolB is encapsulated after activation. On the other hand, since its substrate can act as a potent antioxidant, and the inactive tetramer forms a partial active site and hence cannot bind the substrate (Goulding *et al*, 2005), it is possible that Enc also encapsulates inactive FolB, thus allowing the substrate to alleviate oxidative stress without participating in the folate biosynthesis pathway. Interestingly, Goulding *et al* (2005) found that after cleavage of their C-terminal His-tag to improve crystal diffraction, the last 13 amino acids in the C-terminal end had also been removed by thrombin due to a natural cleavage site found between Arg120 and Gly121. This would allow one to hypothesis that Enc could function to protect active and inactive FolB during oxidative stress, which is subsequently released by cleavage upon turnover of Enc.

### 5.3.7 Conclusion

Enc, GSI, and BrfB have important functions in both *Msm* and *Mtb*. The crystal of Enc from *T. maritima* and the low-resolution structure from *Msm* where used to gather insights with regards to forming hypotheses on Enc pore selectivity and cargo binding of the encapsulated proteins DyP, BrfB, and FolB in *Mtb*. We hypothesis that *Msm* and *Mtb* Enc pores function to admit specific substrates, based on the differences in electrostatic potentials observed between the outside and inside of the Enc shell. In addition, there are species specific changes to these electrostatic potentials, based on differences observed between the Encs of *T. maritima* and *Msm* and *Mtb*. This likely reflects selective pressures for the uptake of different substrates, since *Mtb* Enc encapsulates a third protein, FolB, not enclosed in *T. maritima* Enc. As proposed for *T. maritima* Enc by Sutter *et al* (2008), we further hypothesise that the encapsulated targets of *Msm* and *Mtb* Enc bind at a specific pore corresponding to their respective symmetries; DyP and BrfB could bind on the 3-fold axis of Enc while FolB could bind on a pseudo-4-fold axis.

A preliminary reconstruction of encapsulated DyP was completed and docked into *Msm* Enc, which showed that it could feasibly bind at the 3-fold pore. Further work will be needed to test the hypothesis that the cargo proteins of *Msm* and *Mtb* Enc bind at specific Enc pores. This will likely involve using cryo-EM to gain a higher-resolution structure which may elucidate if the C-terminal extension binds in the same hydrophobic pocket in *Msm* and *Mtb* Enc as it does in *T. maritima* Enc. Examining high-resolution structures of empty Enc with cargo loaded Enc will also be critical in testing the hypothesis that the Enc pores only open where a specific cargo has bound.

Under the conditions tested, DyP appears to be the primary encapsulated protein in *Msm* Enc. Furthermore, like *Mtb* Enc, *Msm* Enc is exported into the cell culture medium during growth. Export of Enc in *Mtb* may aid in certain stress conditions faced while inhabiting host macrophages.

# Chapter VI: General Discussion and Future Directions

## 6.1 General Discussion

Recently, Jonas & Kording (2017) sought to find if modern neuroscience techniques could be used to understand the fundamental workings of a man-made microprocessor. They found that the current techniques used by neuroscientists, such as studying defective circuits, reconstructing the connectome of the microprocessor transistors, or studying the "on-to-off" transition for individual transistors, does not lead to meaningful knowledge of how microprocessors work as a system. Although the methods of neuroscientists described, such as examining defective brains, reconstructing the connectome of neurons, or the "spikes" of individual neurons can give valuable information, it is taken for granted in the implication that these techniques will eventually lead to fundamental insights into the workings of the brain.

Likewise, it is often assumed that the molecular biology techniques currently employed, from knock-out studies of genes to mapping the PPIs of cells, will eventually lead to a fundamental understanding of how a cell works. As Jonas & Kording (2017) have shown, this is not necessarily the case. The main goal for the generation of PIN maps has been to describe all possible PPIs under a variety of conditions – such as diseased and non-diseased states (Fessenden, 2017). Under the current paradigm, it is thought that the main limiting factor to understanding how PINs drive the cell is the lack of complete "interactomic" data (Fessenden, 2017). But this is the same logical fallacy identified by Jonas & Kording (2017) in noting that mapping all neuron connections under different states will not necessarily divulge the complexities of the brain.

Structural information is very valuable in understanding the mechanism of individual protein complexes. Since the 1960's, we have begun to appreciate that the individual cell components work in a highly-complex and coordinated manner (Alberts, 1998). Efforts at examining PPIs in a high-throughput manner have sought to capture as many as possible through various purification techniques in combination with MS data (e.g Gavin *et al*, 2002; Butland *et al*, 2005; Krogan *et al*, 2006; Kühner *et al*, 2009; Han *et al*, 2009; Maco *et al*, 2011; Kristensen *et al*, 2012; Havugimana *et al*, 2012; Wan *et al*, 2015). Recently, Kastritis *et al* (2017) attempted

to couple structural data from EM with complexes separated by size-exclusion chromatography and identified by correlative MS in the thermophilic organism *Chaetomium thermophilum*. They purified 1,176 proteins across 30 fractions, which represents 27.4% of the expressed proteome, and managed to find 108 interconnected protein complexes which could be clustered into 27 communities (protein complexes which interact). However, like Maco *et al* (2011), they were only able to identify well-known protein complexes through 2D classification; fatty acid synthase, 20S proteosome, 60S ribosome, and 40S ribosome (Kastritis *et al*, 2017). Thus, it is very challenging in matching protein structure to protein identity with confidence. We have seen in **Chapter II**, that the protein identification problem was overcome to some degree through coupling MS data derived from both native and SDS-PAGE with structural data from EM. In addition, in **Chapter IV** a method was presented in which protein complexes without high-resolution X-ray crystallography structural homologues could feasibly be identified through polypeptide chain-fitting or topology prediction of a high-resolution structure obtained by cryo-EM.

Methods which have sought to produce proteins or protein complexes for structural study have tended to rely on high-throughput cloning (e.g Christendat *et al*, 2000; Totir *et al*, 2012; Milewksi *et al*, 2016). A recent method used antibody-display to capture proteins for reconstruction by negative stain EM (Hubert *et al*, 2014). Other methods have attempted to use chemical cross-linking to produce spatial constraints in order to map protein complex topologies (e.g Shi *et al*, 2015; Kastritis *et al*, 2017), but is not suitable for studying low-affinity complexes (Mädler *et al*, 2010). In **Chapter III**, we examined the use of grid blotting in combination with blue native PAGE as an alternative high-throughput method for purifying proteins for structural study. Recently, a modified grid blotting technique on BN-PAGE was used to visualize by negative stain EM the various states of p53 with bound DNA (Kearns *et al*, 2016). Considering that the cost of determining the crystal structure of a novel drug target for soluble bacterial proteins is, on average, $140,000 with a success rate of 35% (Stevens, 2003), grid blotting in combination with structural determination by cryo-EM has the potential to be a cheaper alternative.

Higher-order structures are often used to sequester metabolic reactions, such as carbon-fixation reactions in carboxysome shells (Kerfeld *et al*, 2005) or proteins involved in *de novo* purine biosynthesis, dubbed “purinosomes” (Chitrakar *et al*, 2017). In **Chapter V**, the biology

of Mycobacterial Encs was examined and predictions made with regards to the mode of binding of their cargo proteins based on the *T. maritima* Enc crystal structure (Sutter *et al*, 2008). Furthermore, the question of substrate access was also addressed. However, since there are substantial differences between *Msm* and *T. maritima* Enc, both in terms of bound cargo and pore chemistry, there is a need to produce a high-resolution structure of *Msm* Enc through cryo-EM. This was examined in **Chapter IV**, which established the conditions required to produce a good sample for data collection.

## 6.2 Future Directions

Further work would utilise the fact that *Msm* Enc appears to be exported into the culture medium during cell growth (see **Chapter V**) in order to obtain more material for cryo-EM trials. Although optimal conditions were found related to hole-size, humidity, temperature, and ice-thickness, the final resolution of the cryo-EM structure obtained was hampered by low Enc concentration and contrast (see **Chapter IV**). A sample with a higher-Enc concentration could be obtained by purifying from the cell culture filtrate rather than the cell cytoplasm which contains considerably fewer Enc particles. Furthermore, contrast can be improved by fixing the stability of the microscope as well as using a direct-electron detector to overcome the diminished contrast which may be caused by the high salt concentration used in the buffer to maintain stable Enc particles.

Hypotheses were generated related to the mode of binding of the cargo proteins (see **Chapter V**) which would benefit from a higher-resolution structure; this would resolve the question of the binding of cargo proteins at a particular point of symmetry in the Enc lumen. In addition, by solving to high-resolution empty Enc particles as well as those which harbour DyP would address the question of pore flexibility upon cargo binding.

The strategies suggested can easily be applied to other organisms. The use of microscopy, rather than SDS- or native-PAGE gels, offers a far more sensitive technique to track protein complexes, particularly if they are present in low copy number. In addition, the use of high-resolution cryo-EM promises to be an improvement upon the current methods of protein identification.

# References


Akita, F., Chong, K.T., Tanaka, H., Yamashita, E., Miyazaki, N., Nakaishi, Y., Suzuki, M., Namba, K., *et al*. 2007. The Crystal Structure of a Virus-like Particle from the Hyperthermophilic Archaeon *Pyrococcus furiosus* Provides Insight into the Evolution of Viruses. *Journal of Molecular Biology*. 368: 1469-83.

Alberts, B. 1998. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*. 92(3): 291-4.

Albert, R., Jeong, H. & Barabási, A.L. 2000. Error and attack tolerance of complex networks. *Nature*. 406: 378-82.

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.C., Bork, P., *et al*. 2004. Structure-Based Assembly of Protein Complexes in Yeast. *Science*. 303: 2026-9.

Arndt, C., Koristka, S., Bartsch, H., Bachmann, M. 2012. Native polyacrylamide gels. *Methods in Molecular Biology.* 869: 49-53.

Atmeh, R.F., Arafa, I.M., Al-Khateeb, M. 2011. Albumin Aggregates: Hydrodynamic Shape and Physico-Chemical Properties. *Jordan Journal of Chemistry*. 2(2): 169-182.

Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J. 2004. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*. 22(1): 78-85.

Bai, X., McMullan, G. & Scheres, S.H.W. 2015. How cryo-EM is revolutionizing structural biology. *Trends in Biochemical Sciences.* 40(1): 49-57.

Bairoch, A. & Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*. 28(1): 45-8.

Baker, M., Ju, T. & Chiu, W. 2007. Identification of Secondary Structure Elements in Intermediate Resolution Density Maps. *Structure*. 15(1): 7-19.

Balakrishnan, S., Kamisetty, H., Carbonell, J.G., Lee, S.I., Langmead, C.J. 2011. Learning generative models for protein fold families. *Proteins*. 79: 1061-78.

Bamford, D.H. 2002. Those magnificent molecular machines: logistics in dsRNA virus transcription. *EMBO Reports*. 3(4): 317-18.

Barabási, A.L. & Albert, R. 1999. Emergence of Scaling in Random Networks. *Science*. 286: 509-12.

Bisson, G.P., Mehaffy, C., Broeckling, C., Prenni, J., Rifat, D., Lun, D.S., Burgos, M., Weissman, D., *et al*. 2012. Upregulation of the Phthiocerol Dimycocerosate Biosynthetic Pathway by Rifampin-Resistant, *rpoB* Mutant *Mycobacterium tuberculosis*. *Journal of Bacteriology*. 194(23): 6441-52.

Bollschweiler, D., Schaffer, M., Lawrence, C.M., Engelhardt, H. 2017. Cryo-electron microscopy of an extremely halophilic microbe: technical aspects. *Extremophiles*. 21: 393-8.

Braz, V.A. & Howard, K.J. 2009. Separation of Protein Oligomers by Blue Native Gel Electrophoresis. *Analytical Biochemistry*. 388(1): 170-2.

Brilot, A.F., Chen, J.Z., Cheng, A., Pan, J., Harrison, S.C., Potter, C.S., Carragher, B., Henderson, R., et al. 2012. Beam-induced motion of vitrified specimen on holey carbon film. *Journal of Structural Biology*. 177: 630-7.

Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., *et al*. 2005. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature*. 433: 531-7.

Callaway, E. 2015. The revolution will not be crystallised. *Nature*. 525: 172-4.

Carragher, B. & Potter, C. 2009. ACE2. URL: http://emg.nysbc.org/redmine/projects/software/wiki/ACE2 [Last accessed February 2018]

Cassidy-Amstutz, C., Oltrogge, L., Going, C.C., Lee, A., Teng, P., Quintanilla, D., East-Seletsky, A., Williams, E.R, *et al*. 2016. Identification of a Minimal Peptide Tag for *in Vivo* and *in Vitro* Loading of Encapsulin. *Biochemistry*. 55: 3461-8.

Chan, ED. 2002. Current medical treatment for tuberculosis. *BMJ*. 325(7375): 1282-6.

Chang, C., Evdokimova, E., Savchenko, A., Joachimiak, A, & Midwest Centre for Structural Genomics. 2005. Crystal structure of protein NE0167 from *Nitrosomonas europaea*, PDB 1ZPY. doi: 102210/pdb1zpy/pdb.

Chatr-aryamontri, A., Ceol, A., Licata, L., Cesareni, G. 2007. Protein interactions: integration leads to belief. *Trends in Biochemical Science*. 33(6): 241-2.

Cheng, Y., Grigorieff, N., Penczek, P.A., Walz, T. 2015. A primer to single-particle cryo-electron microscopy. *Cell*. 161: 438-49.

Chitrakar, I., Kim-Holzapfel, D.M., Zhou, W., French, J.B. 2017. Higher order structures in purine and pyrimidine metabolism. *Journal of Structural Biology*. 197: 354-364.

Cho, H.J., Hyun, J.K., Kim, J. G., Jeong, H.S., Park, H.N., You, D.J., Jung, H.S. 2013. Measurement of ice thickness on vitreous ice embedded cryo-EM grids: investigation of optimizing condition for visualizing macromolecules. *Journal of Analytical Science and Technology*. 4:7.

Choi, B., Moon, H., Hong, S.J., Shin, C., Do, Y., Ryu, S., Kang, S. 2016. Effective Delivery of Antigen−Encapsulin Nanoparticle Fusions to Dendritic Cells Leads to Antigen-Specific Cytotoxic T Cell Activation and Tumor Rejection. *ACS Nano*. 10: 7339-50.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., *et al*. 2000. Structural proteomics of an archaeon. *Nature Structural Biology*. 7(10): 903-9.

Contreras, H., Joens, M.S., McMath, L.M., Le, V.P., Tullius, M.V., Kimmey, J.M., Bionghi, N., Horwitz, M.A., et al. 2014. Characterization of a *Mycobacterium tuberculosis* Nanocompartment and Its Potential Cargo Proteins. *Journal of Biological Chemistry*. 289(26): 18279-89.

Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F., Sanchez, J.C. 2000. The dynamic range of protein expression: A challenge for proteomic research. *Electrophoresis*. 21: 1104-15.

Cotter, P.D., Ross, R.P. & Hill, C. 2013. Bacteriocins — a viable alternative to antibiotics? *Nature Reviews Microbiology*. 11: 95-105.

Cottrell, J.S. 2011. Protein identification using MS/MS data. *Journal of Proteomics*. 74: 1842-51.

Damodoran, S. & Kinsella, J.E. 1980. The Effects of Neutral Salts on the Stability of Macromolecules. *The Journal of Biological Chemistry*. 256(7): 3394-8.

Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D. 2002. Protein Interactions: Two methods for assessment of the reliability of high-throughput observations. *Molecular & Cellular Proteomics*. 1(5): 349-56.

Deeds, E.J., Ashenburg, O. & Shakhnovich, E.I. 2006. A simple physical model for scaling in protein–protein interaction networks. *Proceedings in the National Academy of Sciences USA*. 103(2): 311-6.

DeJesus, M.A., Gerrick, E.R., Xu, W., Park, S.W., Long, J.E., Boutte, C.C., Rubin, E.J., Schnappinger, D., *et al*. 2017. Comprehensive Essentiality Analysis of the *Mycobacterium tuberculosis* Genome via Saturating Transposon Mutagenesis. *American Society for Microbiology*. 8: e02133-16.

De Rosier, D.J. & Klug, A. Reconstruction of Three Dimensional Structures from Electron Micrographs. *Nature*. 217: 130-4.

Diprose, J.M., Burroughs, J.N., Sutton, G.C., Goldsmith, A., Gouet, P., Malby, R., Overton, I., Ziéntara, S., *et al*. 2001. Translocation portals for the substrates and products of a viral transcription complex: the bluetongue virus core. *The EMBO Journal*. 20(24): 7229-39.

DiMaio, F, Song, Y, Li, X, Brunner, MJ, Xu, C., Conticello, V., Egelman, E., Marlovits, T.C., *et al*. 2015. Atomic-accuracy models from 4.5-Å cryo-electron microscopy data with density-guided iterative local refinement. *Nature Methods*. 12(4): 361-5.

Dou, Z., Heinhorst, S., Williams, E.B., Murin, C.D., Shively, J.M., Cannon, G.C. 2008. $CO_2$ Fixation Kinetics of *Halothiobacillus neapolitanus* Mutant Carboxysomes Lacking Carbonic Anhydrase Suggest the Shell Acts as a Diffusional Barrier for $CO_2$. *The Journal of Biological Chemistry*. 283(16): 10377-84.

Dubochet, J., Groom, M. & Mueller-Neuteboom, S. 1982. The Mounting of Macromolecules for Electron Microscopy with Particular Reference to Surface Phenomena and the Treatment of Support Films by Glow Discharge. In *Advances in Optical and Electron Microscopy*. R. Barer & V.E. Cosslett, Eds. London: Academic Press. 8: 107-35.

Dubochet, J., Adrian, M., Chang, J.J., Homo, J.C., Lepault, J., McDowall, A.W., Schultz, P. 1988. Cryo-electron microscopy of vitrified specimens. *Quarterly Review of Biophysics*. 21(2): 129-228.

Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 32(5): 1792-7.

Edwards, A.M., Kus, B., Jansen, R., Greenbaum, D., et al. 2002. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends in Genetics*. 18(10): 529-36.

Eisenberg, D., Marcotte, E.M., Xenarios, I., Yeates, T.O. 2000. Protein function in the post-genomic era. *Nature*. 405: 823-6.

Elad, N., Clare, D.K., Saibil, H.R., Orlova, E.V. 2008. Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections. *Journal of Structural Biology*. 162: 108-20.

Elmund, D. & Elmund, H. 2012. SIMPLE: Software for *ab initio* reconstruction of heterogeneous single-particles. *Journal of Structural Biology*. 180: 420-7.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature*. 402: 86-90.

Erickson, H.P. & Klug, A. 1971. Measurement and Compensation of Defocusing and Aberrations by Fourier Processing of Electron Micrographs. *Philosophical Transactions of the Royal Society of London*. 261(837): 105-18.

Erickson, H.P. 2009. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. *Biological Procedures Online*. 11: 32-51.

Eubel, H., Braun, H.P. & Millar, A.H. 2005. Blue-native PAGE in plants: a tool in analysis of protein-protein interactions. *Plant Methods*. 1:11.

Fernandez-Leiro, R. & Scheres, S.H.W. 2016. Unravelling biological macromolecules with cryo-electron microscopy. *Nature*. 537: 339-46.

Feltcher, M.E., Gibbons, H.S., Lignon, L.S., Braunstein, M. 2012. Protein Export by the Mycobacterial SecA2 System Is Determined by the Preprotein Mature Domain. *Journal of Bacteriology*. 195(4): 672-81.

Fessenden, M. 2017. Protein maps chart the causes of disease. *Nature*. 549: 293-5.

Fields, S. & Song, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature*. 340: 245-6.

Fleischmann, R.D., Dodson, R.J., Haft, D.H., Merkel, J.S., Nelson, W.C., Fraser, C.M. 2006. Mycobacterium smegmatis str. MC2 155 chromosome, complete genome. Available from: https://www.ncbi.nlm.nih.gov/nuccore/NC_008596 [last accessed February 2018].

Frank J., Radermacher M., Penczek P., Zhu J., Li Y., Ladjadj M., Leith A. 1996. SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *Journal of Structural Biology*. 116(1): 190-9.

Frank, J. 2006. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. United States of America: Oxford University Press.

Fonslow, B.R., Carvalho, P.C., Academia, K., Freeby, S., Xu, T., Nakorchevsky, A., Paulus, A., Yates III, J.R. 2011. Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. *Journal of Proetome Research*. 10(8): 3690-700.

Free, R.B., Hazelwood, L.A. & Sibley, DR. 2009. Identifying Novel Protein-Protein Interactions Using Co-Immunoprecipitation and Mass Spectroscopy. *Current Protocols in Neuroscience*. Chapter 5: Unit 5.28.

Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., *et al*. 2002. Functional organisation of the yeast proteome by systematic analysis of protein complexes. *Nature*. 415(6868): 141-7.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., *et al*. 2006. Proteome survey reveals modularity of yeast cell machinery. *Nature*. 440(7084): 631-6.

Gavin, A.C. & Superti-Furga, G. 2003. Protein complexes and proteome organization from yeast to man. *Current Opinion in Chemical Biology*. 7(1): 21-7.

Giessen, T.W. & Silver, P.A. 2017. Widespread distribution of encapsulin nanocompartments reveals functional diversity. *Nature Microbiology*. 2: 17029.

Gill, H.S., Pfuegl, G.M. & Eisenberg, D. 2002. Multicopy crystallographic refinement of a relaxed glutamine synthetase from *Mycobacterium tuberculosis* highlights flexible loops in the enzymatic mechanism and its regulation. *Biochemistry*. 41: 9863-72.

Gordon, G.W., Berry, G., Liang, X.H., Levine, B., Herman, B. 1998. Quantitative Fluorescence Resonance Energy Transfer Measurements Using Fluorescence Microscopy. *Biophysics Journal*. 74(5): 2702-13.

Gorham, R.D., Kieslich, C.A., Nichols, A., Sausman, N.U., Foronda, M., Morikis, D. 2011. An evaluation of Poisson–Boltzmann electrostatic free energy calculations through comparison with experimental mutagenesis data. *Biopolymers*. 95(11): 746-54.

Goulding, C.W., Apostol, M.I., Sawaya, M.R., Phillips, M., Parseghian, A., Eisenberg, D. 2005. Regulation by Oligomerization in a Mycobacterial Folate Biosynthetic Enzyme. *Journal of Molecular Biology*. 349: 61-72.

Grant, T. & Grigorieff, N. 2015. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife*. 4: e06980.

Grassucci, R.A., Taylor, D.J. & Frank, J. 2007. Preparation of macromolecular complexes for cryo-electron microscopy. *Nature Protocols*. 2(12): 3239-46.

Han, B.G., Dong, M., Liu, H., Camp, L., et al. 2009. Survey of large protein complexes in *D. vulgaris* reveals great structural diversity. *Proceedings in the National Academy of Sciences USA*. 106(39): 16580-5.

Harth, G., Clemens, D.L. & Horwitz, M.A. 1994. Glutamine synthetase of *Mycobacterium tuberculosis*: Extracellular release and characterisation of its enzymatic activity. *Proceedings in the National Academy of Sciences USA*. 91: 9342-6.

Harth, G. & Horwitz, M. 1997. Expression and Efficient Export of Enzymatically Active *Mycobacterium tuberculosis* Glutamine Synthetase in *Mycoe01bacterium smegmatis* and Evidence That the Information for Export is Contained within the Protein. *The Journal of Biological Chemistry*. 272(36): 22728-35.

Harth, G. & Horwitz, M.A. 1999. An inhibitor of exported *Mycobacterium tuberculosis* glutamine synthetase selectively blocks the growth of pathogenic mycobacteria in axenic culture and in human monocytes: extracellular proteins as potential novel drug targets. *Journal of Experimental Medicine*. 189(9): 1425-36.

Harth, G. & Horwitz, M.A. 2003. Inhibition of Mycobacterium tuberculosis Glutamine Synthetase as a Novel Antibiotic Strategy against Tuberculosis: Demonstration of Efficacy In Vivo. *Infection and Immunity*. 71(1): 456-64.

Harth, G., Masleš-Galić, S., Tullius, M.V., Horwitz, M.A. 2005. All four Mycobacterium tuberculosis glnA genes encode glutamine synthetase activities but only GlnA1 is abundantly expressed and essential for bacterial homeostasis. *Molecular Microbiology*. 58(4): 1157-72.

Heinemann, J., Maaty, W.S., Gauss, G.H., Akkaladevi, N., Brumfield, S.K., Rayaprolu, V., Young, M.J., Lawrence, C.M., *et al*. 2011. Fossil record of an archaeal HK97-like provirus. *Virology*. 417: 362-8.

Hanszen, K.J. 1971. The optical transfer theory of the electron microscope: fundamental principles and applications. *Advances in Optical and Electron Microscopy*. 4: 1-84.

Hase, T., Tanaka, H., Suzuki, Y., Nakagawa, S., Kitano, H. 2009. Structure of Protein Interaction Networks and Their Implications on Drug Design. *PLoS Computational Biology*. 5(10): e1000550.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., et al. 2012. A Census of Human Soluble Protein Complexes. *Cell*. 150: 1068-1081.

He, D., Hughes, S., Vanden-Hehir, S., Georgiev, A., Altenbach, K., Tarrant, E., Mackay, C.L., Waldron, K.J., *et al*. 2016. Structural characterization of encapsulated ferritin provides insight into iron storage in bacterial nanocompartments. *eLife*. 5: e18972.

Henderson, R. 2013. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proceedings of the National Academy of Sciences of the United States of America*. 110(45): 18037-41.

Hirosawa, M., Hoshida, M., Ishikawa, M., Toya, T. 1993. MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming. *Computer Applications in Biosciences*. 9(2): 161-7.

Hirschfield, G.R., McNeil, M. & Brennan, P.J. 1990. Peptidoglycan-Associated Polypeptides of *Mycobacterium tuberculosis*. *Journal of Bacteriology*. 172(2): 1005-13.

Hubert, A., Mitani, Y., Tamura, T., Boicu, M., Nagy, I. 2014. Protein complex purification from *Thermoplasma acidophilum* using a phage display library. *Journal of Microbiological Methods*. 98: 15-22.

Ideker, T. & Sharan, R. 2008. Protein networks in disease. *Genome Research*. 18(4): 644-52.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings in the National Academy of Sciences USA*. 98(8): 4569-74.

James, P. 1997. Protein identification in the post-genome era: the rapid rise of proteomics. *Quarterly Review of Biophysics*. 30(4): 279-331.

Jeong, H., Tombor, B., Albert, R., Oltavi, Z.N., Barabási, A.L. 2000. The large-scale organization of metabolic networks. *Nature*. 407: 651-4.

Jeong, H., Mason, S.P., Barabási, A.L., Oltavi, Z.N. 2001. Lethality and centrality in protein networks. *Nature*. 411: 41-2.

Jonas, E. & Kording, K.P. 2017. Could a Neuroscientist Understand a Microprocessor? *PLoS Computational Biology*. 13(1): e1005268.

Kapopoulou, A., Lew, J.M. & Cole, S.T. 2011. The MycoBrowser portal: A comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis (Edinburgh, Scotland)*. 91(1): 8-13.

Kastritis, P.L., O'Reilly, F.J., Bock, T., Li, Y., Rogon, M.Z., Buczak, K., Romanov, N., Betts, M.J., *et al*. 2017. Capturing protein communities by structural proteomics in a thermophilic eukaryote. *Molecular systems biology*. 13: 936.

Kearns, S., Lurz, R., Orlova, E.V., Okorokov, A.L. 2016. Two p53 tetramers bind one consensus DNA response element. *Nucleic Acids Research.* 44(13): 6185-99.

Kerfeld, C.A., Sawaya, M.R., Tanaka, S., Nguyen, C.V., Phillips, M., Beeby, M., Yeates, T.O. 2005. Protein Structures Forming the Shell of Primitive Bacterial Organelles. *Science*. 309(5736): 936-8.

Kim, R., Yokota, H. & Kim, S.H. 1999. Electrophoresis of Proteins and Protein−Protein Complexes in a Native Agarose Gel. *Analytical Biochemistry*. 282: 147-9.

Khare, G., Nangpal, P. & Tyagi, A.K. 2017. Differential Roles of Iron Storage Proteins in Maintaining the Iron Homeostasis in *Mycobacterium tuberculosis*. PLoS One. 12(1): e0169545.

Klein, M.G., Zwart, P., Bagby, S.C., Cai, F., Chisholm, S.W., Heinhorst, S., Cannon, G.C., Kerfeld, C.A. 2009. Identification and Structural Analysis of a Novel Carboxysome Shell Protein with Implications for Metabolite Transport. *Journal of Molecular Biology*. 392: 319-33.

Klopper, M., Warren, R.M., Hayes, C., van Pittius, N.C.G., Streicher, E.M., Müller, B., Sirgel, F.A., Chabula-Nxiweni, M., *et al*. 2013. Emergence and spread of extensively and totally drug-resistant tuberculosis, South Africa. *Emerging Infectious Diseases*. 19(3): 449-55.

Klug, A. & Crowther, R.A. 1972. Three-dimensional image reconstruction from the viewpoint of information theory. *Nature*. 238: 435-40.

Knispel, R.W., Kofler, C., Boicu, M., Baumeister, W., Nickell, S. 2012. Blotting protein complexes from native gels to electron microscopy grids. *Nature Methods*. 9(2): 182-4.

Kostakioti, M., Newman, C.L., Thanassi, D.G., Stathopoulos, C. 2005. Mechanisms of Protein Export across the Bacterial Outer Membrane. *Journal of Bacteriology*. 187(13): 4306-14.

Krajewksi, W.W., Jones, T.A. & Mowbray, S.L. 2005. Structure of *Mycobacterium tuberculosis* glutamine synthetase in complex with a transition-state mimic provides functional insights. *Proceedings in the National Academy of Sciences USA*. 102(30): 10499-504.

Kristensen, A.R., Gsponer, J. & Foster, L.J. 2012. A high-throughput approach for measuring temporal changes in the interactome. *Nature Methods*. 9(9): 907-9.

Krogan, NJ, Cagney, G, Yu, H, Zhong, G, *et al*. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature. 440: 637-43.

Kühlbrandt, W. 2014. The resolution revolution. *Science*. 343: 1443-4.

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., *et al*. 2009. Proteome Organization in a Genome-Reduced Bacterium. *Science*. 326: 1235-40.

Laemmli, U.K. 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*. 227(5259): 680-685.

Lander, G.C., Stagg, S.M., Voss, N.R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., *et al*. 2009. Appion: an integrated, database-driven pipeline to facilitate EM image processing. *Journal of Structural Biology*. 166(1): 95-102.

Levy, E.D. & Pereira-Leal, J.B. 2008. Evolution and dynamics of protein interactions and networks. *Current Opinion in Structural Biology*. 18: 349-57.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., *et al*. 2004. A Map of the Interactome Network of the Metazoan *C. elegans*. *Science*. 303(5657): 540-3.

Liao, H.Y. & Frank, J. 2010. Definition and estimation of resolution in single-particle reconstructions. *Structure*. 18(7): 768-75.

Lienhardt, C. 2014. Fundamental research is the key to eliminating TB. *Nature*. 507: 401.

Lignon, L.S., Hayden, J.D. & Braunstein, M. 2012. The Ins and Outs of *Mycobacterium tuberculosis* Protein Export. *Tuberculosis (Edinb.)*. 92(2): 121-32.

Ludtke S.J., Baldwin P.R. & Chiu W. 1999. EMAN: semiautomated software for high-resolution single-particle reconstructions. *Journal of Structural Biology.* 128(1): 82-97.

Lynch, M. 2012. The Evolution of Multimeric Protein Assemblages. *Molecular Biology & Evolution*. 29(5): 1353-66.

Mackay, J.P., Sunde, M., Lowry, J.A., Crossley, M., Matthews, J.M. 2007a. Proteins interactions: is seeing believing? *Trends in Biochemical Science*. 32(12): 530-1.

Mackay, J.P., Sunde, M., Lowry, J.A., Crossley, M., Matthews, J.M. 2007b. Response to Chatr-aryamontri *et al*.: Protein interactions: to believe or not to believe? *Trends in Biochemical Science*. 33(6): 242-3.

Maco, B., Ross, I.L., Landsberg, M.J., Mouradov, D., et al. 2011. Proteomic and electron microscopy survey of large assemblies in macrophage cytoplasm. *Molecular & Cell Proteomics*. 10(6): M111.008763.

Mädler, S., Seitz, M., Robinson, J., Zenobi, R. 2010. Does Chemical Cross-Linking with NHS Esters Reflect the Chemical Equilibrium of Protein–Protein Noncovalent Interactions in Solution? *Journal of the American Society of Mass Spectrometry*. 21(10): 1775-83.

Mallette, E. & Kimber, M.S. 2017. A Complete Structural Inventory of the Mycobacterial Microcompartment Shell Proteins Constrains Models of Global Architecture and Transport. *The Journal of Biological Chemistry*. 292(4): 1197-1210.

Mallick, S.P., Carragher, B., Potter, C.S., Kriegman, D.J. 2005. ACE: Automated CTF Estimation. *Ultramicroscopy*. 104 (1): 8-29.

Mao, Y., Wang, L., Gu, C., Herschhorn, A., Désormeaux, A., Finzi, A., Xiang, A.H., Sodroski, J.G. 2013. Molecular architecture of the uncleaved HIV-1 envelope glycoprotein trimer. *Proceedings in the National Academy of Sciences USA*. 110(30): 12438-43.

Margineaunu, A., Chan, J.J., Kelly, D.J., Warren, S.C., Flatters, D., Kumar, S., Katan, M., Dunsby, C.W., *et al*. 2016. Screening for protein-protein interactions using Förster resonance energy transfer (FRET) and fluorescence lifetime imaging microscopy (FLIM). *Scientific Reports*. 6: 28186.

Marsh, J.A. & Teichmann, S.A. 2015. Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annual Review of Biochemistry*. 84: 551-75.

Martin, S.F., Tatham, M.H., Hay, R.T., Samuel, I.D.W. 2008. Quantitative analysis of multi-protein interactions using FRET: Application to the SUMO pathway. *Protein Science*. 17: 777-84.

Mazandu, G.K. & Mulder, N.J. 2011. Generation and Analysis of Large-Scale Data-Driven *Mycobacterium tuberculosis* Functional Networks for Drug Target Identification. *Advances in Bioinformatics.* 2011: 1-14.

McGuffin, L.J., Street, S.A., Bryson, K., Sorensen, S.A., Jones, D.T. 2004. The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. *Nucleic Acids Research*. 1(32): D196-9.

McHugh, L. & Arthur, J.W. 2008. Computational Methods for Protein Identification from Mass Spectrometry Data. *PLoS Computational Biology*. 4(2): e12.

McHugh, C.A., Fontana, J., Nemecek, D., Cheng, N., Aksyuk, A.A., Heymann, J.B., Winkler, D.C., Lam, A.S., *et al*. 2014. A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *The EMBO Journal*. 33(17): 1896-1911.

McMullan, G., Faruqi, A.R. & Henderson, R. 2016. Direct Electron Detectors. *Methods in Enzymology*. 579: 1-17.

Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L, Yusupova, G., Yusupov, M. 2012. One core, two shells: bacterial and eukaryotic ribosomes. *Nature Structural & Molecular Biology*. 19: 560-7.

Merk, A., Bartesaghi, A., Banerjee, S., Falconieri, V., Rao, P., Davis, M.I., Pragani, R., Boxer, M.B., *et al*. 2016. Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. *Cell*. 165: 1-10.

Milewski, M.C., Broger, T., Kirkpatrick, J., Filomena, A., Komadina, D., Schneiderhan-Marra, N., Wilmanns, M., Parret, A.H.A. 2016. A standardized production pipeline for high profile targets from *Mycobacterium tuberculosis*. *Proteomics Clinical Applications*. 10(9-10): 1049-57.

Mohan, A., Padiadpu, J., Baloni, P., Chandra, N. 2015. Complete Genome Sequences of a *Mycobacterium smegmatis* Laboratory Strain (MC2 155) and Isoniazid-Resistant (4XR1/R2) Mutant Strains. *Genome Announcement*. 3(1): e10520-14.

Moon, H., Lee, J., Kim, H., Heo, S., Min, J., Kang, S. 2014. Genetically engineering encapsulin protein cage nanoparticle as a SCC-7 cell targeting optical nanoprobe. *Biomaterials Research*. 18:21.

Mortality and causes of death in South Africa, 2013: Findings from death notification/ Statistics South Africa. Pretoria: Statistics South Africa, 2014.

Mowbray, S.L., Kathiravan, M.K., Pandey, A.A., Odell, L.R. 2014. Inhibition of Glutamine Synthetase: A Potential Drug Target in *Mycobacterium tuberculosis*. *Molecules*. 19: 13161-76.

Namouchi, A., Gómez-Muñoz, M., Frye, S.A., Moen, L.V., Rognes, T., Tønjum, T., Balasingham, S.V. 2016. The *Mycobacterium tuberculosis* transcriptional landscape under genotoxic stress. *BMC Genomics*. 17: 791.

Nichols, R.J., Cassidy-Amstutz, C., Chaijarasphong, T., Savage, D.F. 2017. Encapsulins: molecular biology of the shell. *Critical Reviews in Biochemistry and Molecular Biology*. 52(5): 583-94.

Nishizawa, H., Kita, N., Okimura, S., Takao, E., Abe, Y. 1988. Determination of molecular weight of native proteins by polyacrylamide gradient gel electrophoresis. *Electrophoresis*. 9: 803-6.

Noens, E.E., Williams, C., Anandhakrishnan, M., Poulsen, C., Ehebauer, M.T., Wilmanns, M. 2011. Improved mycobacterial protein production using a *Mycobacterium smegmatis* groEL1ΔC expression strain. *BMC Biotechnology*. 11:27.

Nyquist, H. 1928. Certain Topics in Telegraph Transmission Theory. *Transactions of the American Institute of Electrical Engineers E.* 47(2): 617-44.

Oeffinger, M. 2012. Two steps forward—one step back: Advances in affinity purification mass spectrometry of macromolecular complexes. *Proteomics*. 12: 1591-1608.

Ó'Fágáin, C., Cummins, P.M. & O'Connor, B. 2011. Gel-Filtration Chromatography. *Methods in Molecular Biology*. 681: 25-33.

Orlova, E.V. & Saibil, H.R. 2011. Structural Analysis of Macromolecular Assemblies by Electron Microscopy. *Chemical Reviews*. 111(12): 7710-48.

Ovchinnikov, S., Kinch, L., Park, H., Liao, Y., Pei, J., Kim, D.E., Kamisetty, H., Grishin, N.V., et al. 2015. Large-scale determination of previously unsolved protein structures using evolutionary information. *eLife*. 4: e09248.

Ovchinnikov, S., Park, H., Varghese, N., Huang, P.S., Pavlopoulos, G.A., Kim, D.E., Kamisetty, H., Kyrpides, N.C., *et al*. 2017. Protein structure determination using metagenome sequence data. *Science*. 355: 294-8.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N. 1999. The use of gene clusters to infer functional coupling. *Proceedings in the National Academy of Sciences USA*. 96: 2896-2901.

Pandey, R. & Rodriguez, G.M. 2012. A Ferritin Mutant of *Mycobacterium tuberculosis* Is Highly Susceptible to Killing by Antibiotics and Is Unable To Establish a Chronic Infection in Mice. *Infection and Immunity*. 80(10): 3650-9.

Parrish, J.R., Yu, J., Liu, G., Hines, J.A., Chan, J.E., Mangiola, B.A., Zhang, H., Pacifico, S., *et al*. 2007. A proteome-wide protein interaction map for *Campylobacter jejuni*. *Genome Biology*. 8(7): R130.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings in the National Academy of Sciences USA*. 96: 4285-8.

Petterson, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*. 25(13): 1605-12.

Peyret, H. 2015. A protocol for the gentle purification of virus-like particles produced in plants. *Journal of virological methods*. 225: 59-63.

Pieters, J. 2008. *Mycobacterium tuberculosis* and the Macrophage: Maintaining a Balance. *Cell Host & Microbe*. 3: 399-407.

Prasad, B.V.V. & Schmid, M.F. 2012. Principles of Virus Structural Organization. *Advances in Experimental Medicine and Biology*. 726: 17-47.

Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., *et al*. 2001. The protein-protein interaction map of *Helicobacter pylori*. *Nature*. 409: 211-5.

Rahmanpour, R. & Bugg, T.D.H. 2013. Assembly in vitro of *Rhodococcus jostii* RHA1 encapsulin and peroxidase DypB to form a nanocompartment. *FEBS Journal*. 280: 2097-104.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., Séraphin, B. 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnology*. 17: 1030-2.

Righetti, P.G. 1989. Of matrices and men. *Journal of Biochemical and Biophysical Methods*. 19: 1-20.

Rohde, K.H., Veiga, D.F.T., Caldwell, S., Balázsi, G., Russell, D.G. 2012. Linking the Transcriptional Profiles and the Physiological States of *Mycobacterium tuberculosis* during an Extended Intracellular Infection. *PLoS Pathogens*. 8(6): e1002769.

Rosenkrands, I., Rasmussen, P.B., Carnio, M., Jacobsen, S., Theisen, M., Andersen, P. 1998. Identification and Characterization of a 29-Kilodalton Protein from *Mycobacterium tuberculosis* Culture Filtrate Recognized by Mouse Memory Effector Cells. 66(6): 2728-35.

Rouder, J.N., Morey, R.D., Verhagen, J., Province, J.M., Wagenmakers, E.J. 2016. Is There a Free Lunch in Inference? *Topics in Cognitive Science*. 8(3): 520-47.

RStudio Team. 2016. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/ [Last accessed February 2018]

Ruskin, R.S., Yu, Z. & Grigorieff, N. 2013. Quantitative characterization of electron detectors for transmission electron microscopy. *Journal of Structural Biology*. 184(3): 10.1016/j.jsb.2013.10.016.

Saito, H. & Watanabe, T. 1979. Effect of a Bacteriocin Produced by *Mycobacterium smegmatis* on Growth of Cultured Tumor and Normal Cells. *Cancer Research*. 39: 5114-7.

Saito, H., Watanabe, T. & Tomioka, H. 1979. Purification, Properties, and Cytotoxic Effect of a Bacteriocin from *Mycobacterium smegmatis*. *Antimicrobial Agents and Chemotherapy*. 15(4): 504-9.

Saito, H. & Watanabe, T. 1981. Effects of a Bacteriocin from *Mycobacterium smegmatis* on BALB/3T3 and Simian Virus 40-Transformed BALB/c Mouse Cells. *Microbiology and Immunology*. 25(1): 13-22.

Saito, N., Matsubara, K., Watanabe, M., Kato, F., Ochi, K. 2003. Genetic and Biochemical Characterization of EshA, a Protein That Forms Large Multimers and Affects Developmental Processes in *Streptomyces griseus*. *Journal of Biological Chemistry*. 278(8): 5902-11.

Sali, A. & Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*. 234(3): 779-815.

Sassetti, C.M., Boyd, D.H., Rubin, E.J. 2003. Genes required for mycobacterial growth defined by high density mutagenesis. *Molecular Microbiology*. 48(1): 77-84.

Sassetti, C.M. & Rubin, E.J. 2003. Genetic requirements for mycobacterial survival during infection. *Proceedings in the National Academy of Sciences USA*. 100(22): 12989-94.

Schägger, H. & von Jagow, G. 1991. Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Analytical Biochemistry*. 199(2): 223-31.

Schägger, H., Cramer, W.A. & von Jagow, G. 1994. Analysis of molecular masses and oligomeric states of protein complexes by blue native electrophoresis and isolation of membrane protein complexes by two-dimensional native electrophoresis. *Analytical Biochemistry*. 217: 220-30.

de Silva, E. & Stumpf, M.P.H. 2005. Complex networks and simple models in biology. *Journal of the Royal Society Interface*. 2: 419-30.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell System Technical Journal*. 27: 379-423.

Shatsky, M., Hall, R.J., Nogales, E., Malik, J., Brenner, S.E. 2010. Automated Multi-model Reconstruction from Single-Particle Electron Microscopy Data. *Journal of Structural Biology*. 170(1): 98-108.

Shi, Y., Pellarin, R., Fridy, P.C., Fernandez-Martinez, J., Thompson, M.K., Li, Y., Wang, Q.J., Sali, A., *et al*. 2015. A strategy for dissecting the architectures of native macromolecular assemblies. *Nature Methods*. 12(12): 1135-8.

Smeulders, MJ, Keer, J, Speight, RA, Williams, HD. 1999. Adaptation of *Mycobacterium smegmatis* to Stationary Phase. *Journal of Bacteriology*. 181(1): 270-83.

Smith, J.L. 2004. The Physiological Role of Ferritin-Like Compounds in Bacteria. *Critical Reviews in Microbiology*. 30: 173-85.

Snijder, J., Kononova, O., Barbu, I.M., Uetrecht, C., Rurup, W.F., Burnley, R.J., Koay, M.S.T., Cornelissen, J.J.L.M., *et al*. 2016. Assembly and Mechanical Properties of the Cargo-Free and Cargo-Loaded Bacterial Nanocompartment Encapsulin. *Biomacromolecules*. 17: 2522-9.

Sonotaki, S., Takami, T., Noguchi, K., Odaka, M., Yohda, M., Murakami, Y. 2017. Successful PEGylation of hollow encapsulin nanoparticles from Rhodococcus erythropolis N771 without affecting their disassembly and reassembly properties. *Biomaterials Science*. 5(6): 1082-9.

Sosunov, V., Mischenko, V., Eruslanov, B., Svetoch, E., Shakina, Y., Stern, N., Majorov, K., Sorokoumova, G., *et al*. 2007. Antimycobacterial activity of bacteriocins and their complexes with liposomes. *Journal of Antimicrobial Chemotherapy*. 59: 919-25.

Stevens, R.C. 2003. The cost and value of three-dimensional protein structure. *Drug Discovery World*. 35-48.

Stumpf, M.P.H. & Wiuf, C. 2010. Incomplete and noisy network data as a percolation process. *Journal of the Royal Society Interface*. 7(51): 1411-19.

Subramaniam, S. 2013. Structure of trimeric HIV-1 envelope glycoproteins [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America*. 110(45): E4172–E4174.

Sullivan, J.T., Young, E.F., McCann, J.R., Braunstein, M. 2012. The *Mycobacterium tuberculosis* SecA2 System Subverts Phagosome Maturation To Promote Growth in Macrophages. *Infection and Immunity*. 80(3): 996-1006.

Sutter, M. 2008. Structural Basis of Enzyme Encapsulation into a Bacterial Nanocompartment. Ph.D Thesis. ETH Zürich.

Sutter, M., Boehringer, D., Gutmann, S., Günther, S., Prangishvili, D., Loessner, M.J., Stetter, K.O., Weber-Ban, E., *et al*. 2008. Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology*. 15(9): 939-47.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*. 30(12): 2725-9.

Tamura, A., Fukutani, Y., Takami, T., Fuijii, M., Nakaguchi, Y., Murakami, Y., Noguchi, K., Yohda, M., *et al*. 2014. Packaging Guest Proteins into the Encapsulin Nanocompartment from Rhodococcus erythropolis N771. *Biotechnology and Bioengineering*. 112(1): 13-20.

Tanaka, S., Kerfeld, C.A., Sawaya, M.R., Cai, F., Heinhorst, S., Cannon, G.C., Yeates, T.O. 2008. Atomic-Level Models of the Bacterial Carboxysome Shell. *Science*. 319: 1083-6.

Tanaka, K. 2009. The proteasome: Overview of structure and functions. *Proceedings of the Japan Academy, Series B Physiological and Biological Sciences*. 85(1): 12-36.

Titz, B., Schlesner, M., Uetz, P. 2004. What do we learn from high-throughput protein interaction data? *Expert Review of Proteomics*. 1(1): 111-21.

Thompson, R.F., Walker, M., Siebert, C.A., Muench, S.P., Ranson, N.A. 2016. An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods*. 100: 3-15.

Thon, F. 1966. Imaging properties of the electron microscope near the theoretical limit of resolution. *Proceedings of the Sixth International Conference for Electron Microscopy*. Kyoto, Japan. p. 23.

Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Young, J., Berriz, G.F., Brost, R.L., *et al*. 2004. Global Mapping of the Yeast Genetic Interaction Network. *Science*. 303: 808-13.

Totir, M., Echols, N., Nanao, M., Gee, C.L., Moskaleva, A., Gradia, S., Iavarone, A.T., Berger, J.M., *et al*. 2012. Macro-to-micro structural proteomics: native source proteins for high-throughput crystallization. *PLoS One*. 7(2): e32498.

Tullius, M.V., Harth, G. & Horwitz, M.A. 2001. High Extracellular Levels of *Mycobacterium tuberculosis* Glutamine Synthetase and Superoxide Dismutase in Actively Growing Cultures Are Due to High Expression and Extracellular Stability Rather than to a Protein-Specific Export Mechanism. *Infection and Immunity*. 69(10): 6348-63.

Tullius, M.V., Harth, G. & Horwitz, M.A. 2003. Glutamine Synthetase GlnA1 Is Essential for Growth of *Mycobacterium tuberculosis* in Human THP-1 Macrophages and Guinea Pigs. *Infection and Immunity*. 71(7): 3927-36.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., *et al*. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*. 403(6770): 623-7.

Valdés-Stauber, Götz, H. & Busse, M. 1991. Antagonistic effect of corneyform bacteria from red smear cheese against *Liseria* species. *International Journal of Food Microbiology*. 13(2): 119-30.

Valdés-Stauber, N. & Scherer, S. 1994. Isolation and Characterization of Linocin M18, a Bacteriocin Produced by *Brevibacterium linens*. *Applied and Environmental Microbiology*. 60(10): 3809-14.

Van Heel, M. 1987. Angular reconstitution: A Posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy*. 21: 111-24.

Van Heel, M. 2013. Finding trimeric HIV-1 envelope glycoproteins in random noise. [Letter to the editor]. *Proceedings of the National Academy of Sciences of the United States of America*. 110(45): E4175–E4177.

Vaynberg, J. & Qin, J. 2006. Weak protein–protein interactions as probed by NMR spectroscopy. *Trends in Biotechnology*. 24(1): 22-7.

Veesler, D., Campbell, M.G., Cheng, A., Fu, C., Murez, Z., Johnson, J.E., Potter, C.S., Carragher, B. 2013. Maximizing the potential of electron cryomicroscopy data collected using direct detectors. *Journal of Structural Biology*. 184: 193-202.

Vogel, C. & Marcotte, E.M. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*. 13(4): 227-32.

Wade, R.H. 1992. A brief look at imaging and contrast transfer. *Ultramicroscopy*. 46: 145-56.

Watts, D.J. & Strogatz, S.H. 1998. Collective dynamics of ‘small-world’ networks. *Nature*. 393: 440-2.

Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., *et al*. 2015. Panorama of ancient metazoan macromolecular complexes. *Nature*. 525: 339-44.

Wang, H.W., Ramey, V.H., Westermann, S., Leschziner, A.E., Welburn, J.P.I., Nakajima, Y., Drubin, D.G., Barnes, G., *et al*. 2007. Architecture of the Dam1 kinetochore ring complex and implications for microtubule-driven assembly and force-coupling mechanisms. *Nature Structural & Molecular Biology*. 14: 721-6.

Wittig, I., Braun, H.P, Schägger, H. 2006. Blue native PAGE. *Nature Protocols*. 1: 418-28.

Wlodawer, A., Minor, W., Dauter, Z., Jaskolski, M. 2013. Protein crystaloography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination. *FEBS Journal*. 280: 5705-36.

Wong, C.C.L., Cociorva, D., Venable, J.D., Xu, T., Yates III, J.R. 2009. Comparison of Different Signal Thresholds on Data Dependent Sampling in Orbitrap and LTQ Mass Spectrometry for the Identification of Peptides and Proteins in Complex Mixtures. *Journal of the American Society for Mass Spectrometry*. 20(8): 1405-14.

Zhang, Y. & Orner, B.P. 2011. Self-Assembly in the Ferritin Nano-Cage Protein Superfamily. *International Journal of Molecular Sciences*. 12: 5406-21.

Zubieta, C., Krishna, S.S., Kapoor, M., Kozbial, P., McMullan, D., Axelrod, H.L., Miller, M.D., Abdubek, P., *et al*. 2007. Crystal structures of two novel dye-decolorizing peroxidases reveal a β-barrel fold with a conserved heme-binding motif. *Proteins*. 69: 223-33.

# Appendix

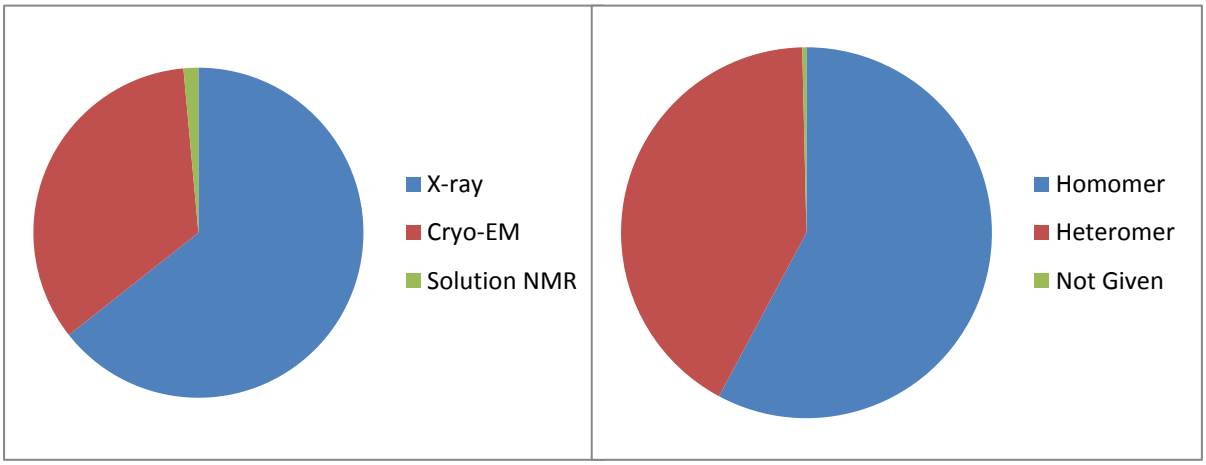



**Figure 8-1. *Msm* structures per ORF.** 64% of the Msm structures have been solved by X-ray crystallography, while 34% and 1% were solved by cryo-electron microscopy (EM) and nuclear magnetic resonance (NMR), respectively (left). 58% and 42% of these structures are in the form of homomers and heteromers, respectively. The heteromeric figure is inflated due to the presence of several large heteromeric structures (Figure i in Appendix) since the data is counted based on the ORF, not per structure. Data obtained from the Protein Data Bank as of November 2017.

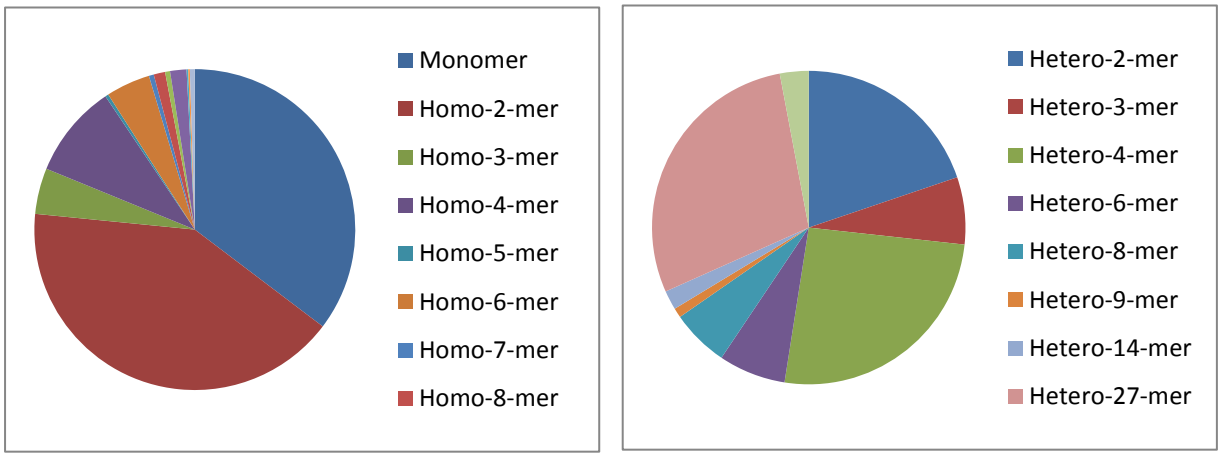



**Figure 8-2. Composition of *Mtb* homomeric (left) and heteromeric (right) structures per ORF.** Data based on depositions in the Protein Data Bank as of November 2017.

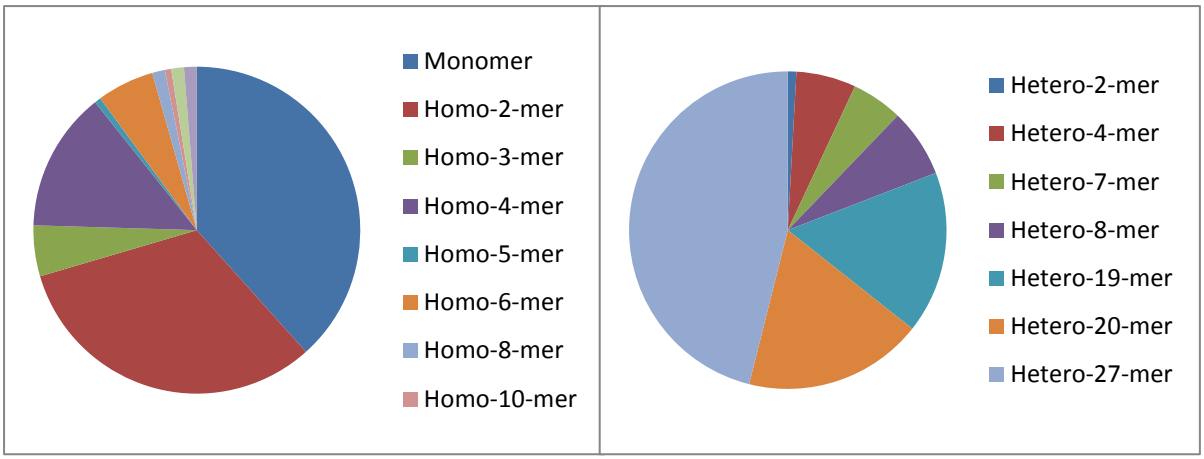



**Figure 8-3. Composition of *Msm* homomeric (left) and heteromeric (right) structures per ORF.** Data based on depositions in the Protein Data Bank as of November 2017.

**Table 8-1.** LC-MS/MS results from native PAGE

| Protein ID[1] | Protein Name | # Unique Peptides | Peptides Gel Band[2] | Peptides Stacking Band[3] | MW (kDa) | % Coverage | Score |
|---|---|---|---|---|---|---|---|
| A0QQU5 | 60 kDa chaperonin 1 | 1 | 1 | 1 | 56.487 | 3.9 | 27.641 |
| A0QS98 | Elongation factor Tu | 1 | 1 | 1 | 43.735 | 6.8 | 126.56 |
| A0QWB3 | Aldehyde dehydrogenase | 1 | 1 | 0 | 50.132 | 2.7 | 10.694 |
| A0R073 | Uncharacterized protein | 2 | 1 | 1 | 31.81 | 15.6 | 11.531 |
| A0R079 | Glutamine synthetase | 12 | 12 | 0 | 53.591 | 27.8 | 99.681 |
| A0R0B3 | Meromycolate extension acyl carrier protein | 1 | 0 | 1 | 10.737 | 30.3 | 8.385 |
| A0R0X1 | Major membrane protein I | 4 | 0 | 4 | 33.727 | 20.2 | 30.681 |
| A0R0X2 | Cysteine desulfurase | 4 | 1 | 3 | 59.621 | 7.3 | 25.658 |
| A0R4H0 | 29 kDa antigen Cfp29 | 6 | 0 | 6 | 28.73 | 29.4 | 176.48 |

1: Uniprot; 2: in resolving gel matrix; 3: at the top of stacking gel matrix

**Table 8-2**. In-solution LC-MS/MS results for gel filtration fractions

| Peptide Sequence[1] | Protein ID[2] | Score | MW (kDa)[3] | #43–48 | #49–53[4] | #54–58[5] |
|---|---|---|---|---|---|---|
| AAAAWKDGVFADEVVPVSIPQRK | A0R1Y7 | 15.296 | 40.067 | 1 | 0 | 0 |
| AAADVGAPAAVSR | A0QT96 | 129.89 | 58.872 | 1 | 0 | 0 |
| AADLLSGPWREK | A0R2H8 | 29.926 | 58.434 | 1 | 0 | 0 |
| AAEAAAGIPASR | A0R510 | 54.066 | 47.478 | 1 | 0 | 0 |
| AAGTDADLAAVAKK | A0QX35 | 57.434 | 64.866 | 1 | 0 | 0 |
| AATLTAIASANYVAR | A0QYF7 | 77.062 | 100.837 | 1 | 0 | 0 |
| ADAEQIIAR | A0QU32 | 61.999 | 25.728 | 1 | 0 | 0 |
| ADDIAALMTLEMGK | A0R3N8 | 49.5 | 50.785 | 1 | 0 | 0 |
| ADFDVDSSGAFTR | A0R1B3 | 54.486 | 94.098 | 1 | 0 | 0 |
| ADLLGTDR | REV__A0R1D8 | 49.358 | 70.071 | 1 | 0 | 0 |
| ADTYESVTVEVK | A0R5Y0 | 97.456 | 28.550 | 1 | 0 | 0 |
| AEPVAHTPDPTRPLAGVR | A0QV70 | 63.145 | 43.306 | 1 | 0 | 0 |
| AFDTQAGPAITSAAR | A0R758 | 95.273 | 59.357 | 1 | 0 | 0 |
| AGGGIVEDVDGNR | A0QWJ0 | 61.353 | 46.926 | 1 | 0 | 0 |
| AGLPGSTEDKGAQAAAAALSTAVTLR | A0QWV1 | 40.428 | 16.187 | 1 | 0 | 0 |
| AGPALACGNAFILKPSER | A0QSJ2 | 65.105 | 54.304 | 1 | 0 | 0 |
| AGPALACGNAFILKPSERDPSVPVR | A0QSJ2 | 105.79 | 54.304 | 2 | 0 | 0 |
| AHDYEEALSLPTK | A0QSJ2 | 80.469 | 54.304 | 2 | 0 | 0 |
| AIDLTPAAVGPGVVPPANLR | A0R4S7 | 95.273 | 32.324 | 1 | 0 | 0 |
| AIGEVFDLRPAAIVR | A0QWT3 | 13.746 | 42.590 | 1 | 0 | 0 |
| AINDNDLAVTAVLSGNR | A0QX20 | 105 | 102.171 | 1 | 0 | 0 |
| ALAEVYAEDDSKEK | A0R609 | 68.44 | 81.115 | 1 | 0 | 0 |
| ALDAAHAAAPAWGK | A0QQW5;A0QSN7 | 125.74 | 55.126; 55.979 | 2 | 0 | 0 |
| ALKEDPGFAEISGYTEDSGEGR | A0QND7 | 35.246 | 31.537 | 1 | 0 | 0 |
| ALLGVGAHDIIGVEAK | A0R5M3 | 115.37 | 46.291 | 2 | 0 | 0 |
| ALLPVLTGDKSPAAQSDTSTDALVR | A0R3N9 | 78.837 | 59.630 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| ANTTAESLAGLKPAFRK | A0R1Y7 | 12.715 | 40.067 | 1 | 0 | 0 |
| APAEYAHDEKAFGR | A0QT01 | 61.167 | 50.068 | 1 | 0 | 0 |
| APFEPLTPGGFRAPNTNFYR | A0QT01 | 87.49 | 50.068 | 1 | 0 | 0 |
| APGLAELPPAATEEEALAELR | A0QYF7 | 138.73 | 100.837 | 1 | 0 | 0 |
| APTGLAALR | A0QP38 | 62.464 | 35.043 | 1 | 0 | 0 |
| ASGDTPVLFDSGIR | A0QZB3 | 85.813 | 42.747 | 1 | 0 | 0 |
| ATARPTFDDDLVTDQVR | A0R716 | 176.32 | 39.498 | 2 | 0 | 0 |
| ATARPTFDDDLVTDQVREQLR | A0R716 | 85.518 | 39.498 | 2 | 0 | 0 |
| ATGVTVDDVAK | A0QYF7 | 91.265 | 100.837 | 1 | 0 | 0 |
| ATGVTVDDVAKR | A0QYF7 | 95.477 | 100.837 | 2 | 0 | 0 |
| ATSGDNHLGGDDWDDR | A0QQC8 | 36.112 | 66.647 | 1 | 0 | 0 |
| AVAEVYAQSDNGER | A0QXX7 | 62.303 | 81.998 | 1 | 0 | 0 |
| AVENFPISFR | A0R2U8 | 58.981 | 49.763 | 1 | 0 | 0 |
| AVLRPGDHVVIPDDAYGGTFR | A0R2X3 | 33.37 | 40.458 | 1 | 0 | 0 |
| AVNAGGVATSALEMQQNASR | A0R3E3 | 21.865 | 48.595 | 1 | 0 | 0 |
| AYRVIDFRRHDK | A0QSD4 | 13.378 | 30.358 | 1 | 0 | 0 |
| CDSIISVGGGSSHDAAK | A0R5M3 | 167.3 | 46.291 | 2 | 0 | 0 |
| CIDGLVANEER | A0R2U8 | 85.29 | 49.763 | 1 | 0 | 0 |
| CLQGVVDGELTAVEAAK | A0R2Q5 | 96.059 | 42.321 | 1 | 0 | 0 |
| DAAATPTVTATR | A0R6D7 | 77.662 | 38.961 | 1 | 0 | 0 |
| DAAESVALYRGEK | A0R6D7 | 45.161 | 38.961 | 1 | 0 | 0 |
| DADGKPDGTTAAAVQQEAAR | A0QV88 | 93.029 | 43.922 | 3 | 0 | 0 |
| DADGKPDGTTAAAVQQEAARR | A0QV88 | 27.924 | 43.922 | 1 | 0 | 0 |
| DAGISVSDIDHVVLVGGSTR | A0QQC8 | 140.74 | 66.647 | 1 | 0 | 0 |
| DAMFAGEHINTSEDR | A0R3N9 | 72.29 | 59.630 | 2 | 0 | 0 |
| DAVQPVHTVYIGAADADEHTPR | A0R471 | 67.877 | 43.440 | 2 | 0 | 0 |
| DEIFGPVLTVR | A0QT04 | 36.693 | 51.916 | 1 | 0 | 0 |
| DFKPYRNELIISTK | A0R716 | 61.344 | 39.498 | 3 | 0 | 0 |
| DFLVPAR | A0QWN3 | 50.12 | 65.170 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| DGAEFVIPTMK | A0QSJ2 | 71.349 | 54.304 | 1 | 0 | 0 |
| DGDAARDFVSR | A0QSJ2 | 143.96 | 54.304 | 2 | 0 | 0 |
| DGPTYWATGETVR | A0QXA3 | 108.47 | 50.988 | 1 | 0 | 0 |
| DGVFADEVVPVSIPQRK | A0R1Y7 | 86.772 | 40.067 | 1 | 0 | 0 |
| DKLDTYVRPLLAISAER | A0QQH0 | 58.83 | 36.595 | 1 | 0 | 0 |
| DKVASYIDAGEAAGAK | A0QV51 | 157.33 | 52.259 | 2 | 0 | 0 |
| DLEPLKSQTLSDAEEK | A0R5L3 | 59.116 | 62.870 | 2 | 0 | 0 |
| DLTSAPCLALSHR | REV__A0QWC3 | 26.969 | 33.118 | 1 | 0 | 0 |
| DLYDTAGIR | A0QRE7 | 63.565 | 49.194 | 1 | 0 | 0 |
| DPAELGPEPER | A0QX35 | 112.75 | 64.866 | 1 | 0 | 0 |
| DPLVPNQVK | A0R3C8 | 92.187 | 35.559 | 1 | 0 | 0 |
| DQHPAPLDPNFTGVGR | A0QVK0 | 33.685 | 28.338 | 1 | 0 | 0 |
| DRILDIGYDSSTK | A0QWT3 | 76.073 | 42.590 | 2 | 0 | 0 |
| DSDMGPLVTK | A0QV51 | 72.643 | 52.259 | 1 | 0 | 0 |
| DSGIDLWR | A0R716 | 51.787 | 39.498 | 1 | 0 | 0 |
| DTLTVGDQSYEIYR | A0QX20 | 184.24 | 102.171 | 1 | 0 | 0 |
| DVDVVTFTGSTAVGRK | A0QT04 | 112.5 | 51.916 | 1 | 0 | 0 |
| DVTGLTMTHCVPNER | A0R5L3 | 65.704 | 62.870 | 2 | 0 | 0 |
| DVVVCAAGSMPGDLHK | A0R189 | 73.435 | 69.226 | 1 | 0 | 0 |
| EAGLPDGVFNVLQGDKTAVDELLTNPK | A0QV51 | 33.835 | 52.259 | 2 | 0 | 0 |
| EDIKEAQNNGGSICR | A0R4S6 | 38.995 | 37.265 | 1 | 0 | 0 |
| EEGGALTLSR | A0QWX7 | 55.064 | 32.518 | 1 | 0 | 0 |
| EELDVEGPR | A0R1B3 | 98.629 | 94.098 | 1 | 0 | 0 |
| EETPFFTGPR | A0QW25 | 163.99 | 26.305 | 1 | 0 | 0 |
| EEVAAIITAEHGK | A0QV51 | 71.176 | 52.259 | 1 | 0 | 0 |
| EFGFTPEAVAAAAER | A0QWY0 | 83.314 | 75.155 | 1 | 0 | 0 |
| EGLAILDSALDVADEHTV | A0QV52 | 83.397 | 49.414 | 1 | 0 | 0 |
| EGTEGPYTGNGGALR | A0QUY3 | 110.39 | 35.859 | 1 | 0 | 0 |
| EGWYTEKPTK | A0R742 | 39.266 | 38.403 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| EITPVTLPDGTVVSKDDGPR | A0R2Y1 | 73.783 | 42.591 | 1 | 0 | 0 |
| EKVAAATMLGQSK | A0R2H8 | 67.456 | 58.434 | 1 | 0 | 0 |
| ELAESSPSIVTPLNSAIGYEEAAK | A0R2U8 | 179.81 | 49.763 | 1 | 0 | 0 |
| ELALTGKDIDAARAEK | A0R5Y0 | 41.482 | 28.550 | 1 | 0 | 0 |
| ELDVAVTAQTAR | A0R3C8 | 112.13 | 35.559 | 1 | 0 | 0 |
| ELGVDPAKVNVNGGAIAIGHPIGMSGAR | A0R1Y7 | 71.974 | 40.067 | 1 | 0 | 0 |
| ELIQEVADEAIGTR | A0R577 | 62.408 | 56.871 | 1 | 0 | 0 |
| ENGDVDLDDLR | A0QYF7 | 51.927 | 100.837 | 1 | 0 | 0 |
| EPAATPQGSAPASAAPETGGDSDEVTELK | A0QQC9 | 105.5 | 23.005 | 2 | 0 | 0 |
| EPTDRRPGR | A0R367 | 37.021 | 24.948 | 1 | 0 | 0 |
| ETFVPPTNHTYPHMEA | A0R2H8 | 40.715 | 58.434 | 1 | 0 | 0 |
| EVEETMHFAVVAGVR | A0QXY0 | 35.037 | 34.534 | 1 | 0 | 0 |
| EVGANIDRYHTYPR | A0R2H8 | 29.937 | 58.434 | 1 | 0 | 0 |
| EVIVTNTLPITEDK | A0R3C8 | 122.7 | 35.559 | 1 | 0 | 0 |
| EVIVTNTLPITEDKR | A0R3C8 | 117.09 | 35.559 | 2 | 0 | 0 |
| EWAAYNPQR | A0QSJ2 | 72.138 | 54.304 | 1 | 0 | 0 |
| FGGDVSHLNLHK | A0QYF7 | 21.759 | 100.837 | 1 | 0 | 0 |
| FIDLVNQNADELAR | A0QSJ2 | 160.18 | 54.304 | 1 | 0 | 0 |
| FRTPSLRQPGGR | A0QVI1 | 28.096 | 19.744 | 1 | 0 | 0 |
| GAASDGGGSKVPEETLAK | A0QQC8 | 29.687 | 66.647 | 1 | 0 | 0 |
| GAFDEATQLVAEARELLDSSPRHSWLQYAR | A0QYV7 | 15.046 | 63.358 | 1 | 0 | 0 |
| GAPDAIVVVR | A0QWG2 | 39.022 | 77.011 | 1 | 0 | 0 |
| GEVAYGAEFFR | A0R3N8 | 119.62 | 50.785 | 1 | 0 | 0 |
| GFHGDQVAALK | A0R742 | 56.043 | 38.403 | 1 | 0 | 0 |
| GGAIVDEPATAEALATVVLR | A0R471 | 55.841 | 43.440 | 1 | 0 | 0 |
| GGGGGEDDDLPGASAAGQER | A0QZ48 | 127.71 | 6.954 | 2 | 0 | 0 |
| GKLSETDKSGLLAR | A0QZQ9 | 49.149 | 30.707 | 1 | 0 | 0 |
| GLDLVASGK | A0QXD8 | 48.423 | 39.356 | 1 | 0 | 0 |
| GLEVGQTFLHENR | A0QU43 | 63.415 | 47.292 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| GLIGDKLSLEELDR | A0R2U8 | 48.527 | 49.763 | 1 | 0 | 0 |
| GLTSGYSPLGAMVASDR | A0QT01 | 63.966 | 50.068 | 1 | 0 | 0 |
| GLVPALR | A0QW47 | 52.231 | 38.863 | 1 | 0 | 0 |
| GMPNAISVLAVAER | A0R2U7 | 46.592 | 35.867 | 2 | 0 | 0 |
| GQYAAGWQGGEK | A0QWX8 | 81.865 | 57.941 | 1 | 0 | 0 |
| GSGIIEELTGK | A0R5M3 | 48.091 | 46.291 | 1 | 0 | 0 |
| GSSSGPVGMILTR | A0QWY0 | 79.82 | 75.155 | 1 | 0 | 0 |
| GTFNVANPVGSLAPTDGSDVPADK | A0R7H5 | 127.32 | 106.556 | 1 | 0 | 0 |
| GTYVPAAEVIER | A0R7H5 | 70.977 | 106.556 | 1 | 0 | 0 |
| GVAEVPLANR | A0QX20 | 109.86 | 102.171 | 1 | 0 | 0 |
| GVISDPAAPFGGIKESGFGR | A0R3N8 | 132.24 | 50.785 | 2 | 0 | 0 |
| GVRVEVDSSDDR | A0QWG2 | 49.423 | 77.011 | 1 | 0 | 0 |
| GVTGALIDDGRLR | A0R3N8 | 179.42 | 50.785 | 2 | 0 | 0 |
| HEYGNGVAIFTR | A0QSJ2 | 136.5 | 54.304 | 2 | 0 | 0 |
| HFGPRYNPWDER | A0R665 | 22.474 | 55.099 | 1 | 0 | 0 |
| HGDGITPPIITR | A0QT01 | 153.12 | 50.068 | 1 | 0 | 0 |
| HGDGITPPIITRGEGVK | A0QT01 | 56.768 | 50.068 | 1 | 0 | 0 |
| HVGTDWNIEIDDK | A0QQC8 | 107.85 | 66.647 | 2 | 0 | 0 |
| HVMSDITGVAEK | A0QT04 | 50.194 | 51.916 | 2 | 0 | 0 |
| IAADLPDRTAGVDYPAGTTAR | A0QYT3 | 56.055 | 88.651 | 1 | 0 | 0 |
| IADIADPLPR | A0R1B3 | 69.812 | 94.098 | 1 | 0 | 0 |
| IAELTESGTVATGSAQK | A0QV70 | 117.03 | 43.306 | 1 | 0 | 0 |
| IAFTGETTTGR | A0QQW5;A0QSN7 | 74.165 | 55.126; 55.979 | 1 | 0 | 0 |
| IAGAAIEAGADLDEAER | A0R758 | 102.97 | 59.357 | 1 | 0 | 0 |
| IDALLTQVDADLAAR | A0R471 | 109.6 | 43.440 | 1 | 0 | 0 |
| IDGGYQTAPAGGSR | A0QV89 | 33.088 | 51.246 | 1 | 0 | 0 |
| IDPETGEVR | A0QQC9 | 101.43 | 23.005 | 1 | 0 | 0 |
| IEHDTMGEVRVPK | A0R2U8 | 32.61 | 49.763 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| IGDPALAETQLGPVISER | A0R7D8 | 113.76 | 53.553 | 1 | 0 | 0 |
| IGKDVQAAIK | A0QX46 | 56.729 | 117.324 | 2 | 0 | 0 |
| IGSMFACNDFGYVPDIITSAK | A0QT01 | 104.41 | 50.068 | 2 | 0 | 0 |
| IIDVVDTGAK | A0R2H8 | 88.177 | 58.434 | 1 | 0 | 0 |
| IKGVSFVGSTPIAK | A0QV51 | 62.088 | 52.259 | 1 | 0 | 0 |
| ILSYIEIGKSEGAK | A0QSN7 | 93.839 | 55.979 | 1 | 0 | 0 |
| IQEGSGLSKEEIDR | A0QQC8 | 70.98 | 66.647 | 1 | 0 | 0 |
| ITDATNGTDPLACIK | A0QWX7 | 132.76 | 32.518 | 1 | 0 | 0 |
| IVCTLGPATSTDETVR | A0QXA3 | 113.24 | 50.988 | 1 | 0 | 0 |
| IVDLPDTSTNDVNK | A0QWX7 | 110.66 | 32.518 | 1 | 0 | 0 |
| IVDLPDTSTNDVNKK | A0QWX7 | 103.67 | 32.518 | 2 | 0 | 0 |
| IVGEVTAER | A0QSV0 | 63.419 | 55.815 | 1 | 0 | 0 |
| IVKEQADKILGK | A0R417 | 103.26 | 48.522 | 2 | 0 | 0 |
| KEEVAAIITAEHGK | A0QV51 | 158.11 | 52.259 | 4 | 0 | 0 |
| LARRASQASRSNPEAR | A0R025 | 29.807 | 16.100 | 1 | 0 | 0 |
| LEEVAVPQR | A0QXD8 | 48.561 | 39.356 | 1 | 0 | 0 |
| LGGVAVR | A0QRR2 | 44.803 | 37.318 | 1 | 0 | 0 |
| LPELWGGSADLAGSNNTTIK | A0QWY0 | 187.37 | 75.155 | 1 | 0 | 0 |
| LPLILSDGHLR | A0R5Y0 | 76.221 | 28.550 | 2 | 0 | 0 |
| LPTIAAYAYK | A0R417 | 71.877 | 48.522 | 1 | 0 | 0 |
| LRDLSWTPDTDVEVTPVAADTEEGR | A0QWG2 | 70.465 | 77.011 | 1 | 0 | 0 |
| LSPSTGAEALAVNR | REV__A0QT49 | 36.334 | 54.414 | 1 | 0 | 0 |
| LVDTEDTVR | A0R189 | 105.57 | 69.226 | 1 | 0 | 0 |
| LVGETGGKDFVLAHSSAHPDVLR | A0R2H8 | 121.47 | 58.434 | 1 | 0 | 0 |
| MDAITDVPTPANEPIHDYAPGSQER | A0R2H8 | 149.74 | 58.434 | 2 | 0 | 0 |
| MQGAITAVADCR | A0R5Y0 | 48.004 | 28.550 | 1 | 0 | 0 |
| NDVDKFTRAEQDEYAAQSHQK | A0R1Y7 | 159.26 | 40.067 | 2 | 0 | 0 |
| NGDGSAGANGAVVLR | A0QX46 | 162.38 | 117.324 | 1 | 0 | 0 |
| NGEVLVGQPAK | A0QQC8 | 106.6 | 66.647 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| NGEVLVGQPAKNQAVTNVDR | A0QQC8 | 143.93 | 66.647 | 2 | 0 | 0 |
| NINEFEGFAK | A0R5M3 | 76.064 | 46.291 | 1 | 0 | 0 |
| NLLVSYSNK | A0QSN7 | 67.993 | 55.979 | 1 | 0 | 0 |
| NTDAVIQPTTGGGR | A0R5Z8 | 45.257 | 33.931 | 1 | 0 | 0 |
| NVVAAGLAER | A0QWT3 | 46.158 | 42.590 | 1 | 0 | 0 |
| PEAVIVATAR | A0R2Y1 | 66.568 | 42.591 | 1 | 0 | 0 |
| PIATPEVYAEMLDR | A0QQH0 | 89.548 | 36.595 | 2 | 0 | 0 |
| QAAAASAARPQDVVK | A0R1B3 | 77.062 | 94.098 | 2 | 0 | 0 |
| QAAVAAGIPWDVAALSINK | A0R1Y7 | 119.87 | 40.067 | 1 | 0 | 0 |
| QGDPLDTETMIGAQASNDQLEK | A0QSN7 | 241.29 | 55.979 | 1 | 0 | 0 |
| QIEAGIERVK | A0R2U8 | 50.108 | 49.763 | 1 | 0 | 0 |
| QLGTPDVIPPADVRRLFDR | A0QT52 | 27.967 | 25.202 | 1 | 0 | 0 |
| QPFQQVIK | A0QQC8 | 51.572 | 66.647 | 1 | 0 | 0 |
| QPIPVLEGTDPEGVAR | A0QWY0 | 111.94 | 75.155 | 1 | 0 | 0 |
| QSLEAALAAVEEAR | A0R2X3 | 107.09 | 40.458 | 1 | 0 | 0 |
| REELDVEGPR | A0R1B3 | 66.56 | 94.098 | 1 | 0 | 0 |
| RGGGGGEDDDLPGASAAGQER | A0QZ48 | 70.353 | 6.954 | 2 | 0 | 0 |
| RIDGAYGDR | A0QYF7 | 0 | 100.837 | 1 | 0 | 0 |
| RPQDRIELTDAK | A0QX20 | 29.299 | 102.171 | 1 | 0 | 0 |
| SADITETPAWQALSDHHAEIGDR | A0R3N9 | 55.208 | 59.630 | 1 | 0 | 0 |
| SAEKLVDTEDTVR | A0R189 | 99.752 | 69.226 | 1 | 0 | 0 |
| SDDVEDADALR | A0R471 | 158.76 | 43.440 | 1 | 0 | 0 |
| SEQQPVEPPVAK | A0R4H9 | 90.108 | 80.453 | 1 | 0 | 0 |
| SFADVPLEGETPAAAATPEAR | A0QX36 | 136.33 | 80.584 | 1 | 0 | 0 |
| SFTDDDDALR | A0QT04 | 54.982 | 51.916 | 1 | 0 | 0 |
| SGAVVTPTIIR | A0QUC9 | 55.461 | 50.326 | 1 | 0 | 0 |
| SGDLVYNSLLCIDR | A0R703 | 49.5 | 27.814 | 1 | 0 | 0 |
| SGGGVDPLTDAPAPITPQQR | A0QWN3 | 163.99 | 65.170 | 2 | 0 | 0 |
| SLQGSSAIEEDRNK | A0QX35 | 87.001 | 64.866 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| SPAAQSDTSTDALVR | A0R3N9 | 92.112 | 59.630 | 1 | 0 | 0 |
| SPNIFFNNVLAQADDYQDK | A0QSN7 | 144.38 | 55.979 | 1 | 0 | 0 |
| SRPAVCSGHSAITDLR | A0R187 | 60.307 | 30.880 | 1 | 0 | 0 |
| SSEADIDKALDAAHAAAPAWGK | A0QSN7 | 118.44 | 55.979 | 2 | 0 | 0 |
| SSFYAETEEQESQR | A0QV52 | 69.451 | 49.414 | 1 | 0 | 0 |
| STADVVSSAPELAR | A0QP38 | 105.46 | 35.043 | 1 | 0 | 0 |
| STGGTLELTDVETPPPDR | A0QXY0 | 95.428 | 34.534 | 1 | 0 | 0 |
| STGGTLELTDVETPPPDRGQVR | A0QXY0 | 134.07 | 34.534 | 2 | 0 | 0 |
| TADAIASEGTPADVVPHK | A0R3N9 | 73.466 | 59.630 | 2 | 0 | 0 |
| TAGVDYPAGTTAR | A0QYT3 | 126.71 | 88.651 | 1 | 0 | 0 |
| TAVASSAPAAPIR | A0QWG2 | 104.05 | 77.011 | 2 | 0 | 0 |
| TAVASSAPAAPIRVPAGTTAGAAVR | A0QWG2 | 115.38 | 77.011 | 4 | 0 | 0 |
| TAVDSFEAQAAR | A0R2U8 | 104.39 | 49.763 | 1 | 0 | 0 |
| TAVNDRPDTTWHNPLR | A0QWX8 | 121.27 | 57.941 | 1 | 0 | 0 |
| TAYPKPAAPNFPER | A0QX46 | 23.596 | 117.324 | 1 | 0 | 0 |
| TDVSAQPPDPDDNR | A0R5L3 | 41.164 | 62.870 | 1 | 0 | 0 |
| TDVSAQPPDPDDNRDVLTDR | A0R5L3 | 151.74 | 62.870 | 3 | 0 | 0 |
| TEANIVNFR | A0R7H5 | 54.611 | 106.556 | 1 | 0 | 0 |
| TENATSNAQLVR | A0R5Z8 | 42.001 | 33.931 | 1 | 0 | 0 |
| TEPATTPTTPDEQIPR | A0R7H5 | 151.55 | 106.556 | 1 | 0 | 0 |
| TEVPELVGVSR | A0R1B3 | 85.807 | 94.098 | 1 | 0 | 0 |
| TGDGTKDSDMGPLVTK | A0QV51 | 101.62 | 52.259 | 2 | 0 | 0 |
| TGKPAALVPLAR | A0QWX7 | 73.233 | 32.518 | 2 | 0 | 0 |
| TGSIYIVKPK | A0QYF5 | 29.551 | 78.277 | 1 | 0 | 0 |
| TGYQYASGETAETVDPAR | A0QRD3 | 74.678 | 34.303 | 1 | 0 | 0 |
| TGYTTYDGGFVNTASTK | A0R417 | 131.32 | 48.522 | 1 | 0 | 0 |
| TITESVCTPEHQR | A0R4P5 | 72.006 | 38.186 | 2 | 0 | 0 |
| TITPPSGAPHPGQPAWNTQR | A0R5Q2 | 45.707 | 66.058 | 1 | 0 | 0 |
| TLFIVASK | A0R3N9 | 45.68 | 59.630 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| TLGPFTWLK | A0R401 | 36.557 | 47.780 | 1 | 0 | 0 |
| TLGSESVPLDATAAGAGK | A0QUC9 | 48.112 | 50.326 | 1 | 0 | 0 |
| TMAPAFR | A0QT01 | 44.611 | 50.068 | 1 | 0 | 0 |
| TPDGEGELTLPGR | A0QYF5 | 152.7 | 78.277 | 1 | 0 | 0 |
| TPIATASECDAAIAR | A0QXH6 | 42.095 | 46.415 | 1 | 0 | 0 |
| TQDDSHEPVTITDK | A0QQC9 | 228.62 | 23.005 | 4 | 0 | 0 |
| TQDDSHEPVTITDKR | A0QQC9 | 181.71 | 23.005 | 6 | 0 | 0 |
| TRDPLVPNQVK | A0R3C8 | 84.054 | 35.559 | 2 | 0 | 0 |
| TTADITQTAPTDLGAK | A0QT01 | 196.24 | 50.068 | 2 | 0 | 0 |
| TTADITQTAPTDLGAKANR | A0QT01 | 155.08 | 50.068 | 1 | 0 | 0 |
| TTDQITVEQLLVNGR | A0R7D8 | 64.121 | 53.553 | 1 | 0 | 0 |
| TTNIDDPTDPR | A0R5U7 | 72.643 | 23.301 | 1 | 0 | 0 |
| TTPSVVAFAR | A0QQC8 | 63.419 | 66.647 | 1 | 0 | 0 |
| TTQIQHFINGR | A0QSJ2 | 158.37 | 54.304 | 2 | 0 | 0 |
| TTSAGVQNIGGAQR | A0R4S7 | 68.847 | 32.324 | 1 | 0 | 0 |
| TTTWDAAETTIPEASEGSR | A0R758 | 98.592 | 59.357 | 1 | 0 | 0 |
| TVFSRPGAADAR | A0QQW5 | 50.354 | 55.126 | 1 | 0 | 0 |
| VAAATMLGQSK | A0R2H8 | 134.25 | 58.434 | 1 | 0 | 0 |
| VAADVLPGAMIRR | A0R1D0 | 37.191 | 22.373 | 1 | 0 | 0 |
| VAETIQSGMVGINR | A0R3N8 | 164.64 | 50.785 | 3 | 0 | 0 |
| VALKVEEVDGDDVVCTVTEGGPVSNNK | A0QXA3 | 49.101 | 50.988 | 1 | 0 | 0 |
| VALVAADDSGR | A0QUC6 | 42.599 | 14.661 | 1 | 0 | 0 |
| VASYIDAGEAAGAK | A0QV51 | 107.11 | 52.259 | 1 | 0 | 0 |
| VDDDHVSVSCDEATTDAHIDAVIK | A0QYF7 | 92.01 | 100.837 | 2 | 0 | 0 |
| VDVGDQNVDGAPR | A0QYT3 | 101.93 | 88.651 | 1 | 0 | 0 |
| VELNRPSETVTLSRPQDGITATLSR | A0QWX7 | 17.158 | 32.518 | 1 | 0 | 0 |
| VFFTTGGGEAVESAWK | A0QT01 | 26.541 | 50.068 | 1 | 0 | 0 |
| VGDWATAASEQER | A0QQU1 | 69.786 | 31.959 | 2 | 0 | 0 |
| VGEGLGITVDDVR | A0R5L3 | 65.179 | 62.870 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| VGEQVIR | A0QQC9 | 55.441 | 23.005 | 1 | 0 | 0 |
| VGGLELTEAGR | A0QS37 | 40.002 | 16.561 | 1 | 0 | 0 |
| VGHDAGAGEVVVLR | A0QWX7 | 138.63 | 32.518 | 2 | 0 | 0 |
| VGPNYLQLPVNRPK | A0R5L3 | 67.726 | 62.870 | 2 | 0 | 0 |
| VGSGEWPVDDNPLR | A0QYF7 | 152.64 | 100.837 | 1 | 0 | 0 |
| VIPGIYGHFAGGDANPEDNK | A0R742 | 100.11 | 38.403 | 2 | 0 | 0 |
| VLSDALGEVTR | A0QV51 | 41.117 | 52.259 | 1 | 0 | 0 |
| VPADAVLGEVDR | A0QU43 | 63.415 | 47.292 | 1 | 0 | 0 |
| VPGRADEVVAAAK | A0QYF7 | 56.122 | 100.837 | 1 | 0 | 0 |
| VPGSAMEVR | A0QWX8 | 91.853 | 57.941 | 1 | 0 | 0 |
| VPKDALWR | A0R2U8 | 37.727 | 49.763 | 1 | 0 | 0 |
| VQRPLINPSDS | A0R189 | 88.37 | 69.226 | 1 | 0 | 0 |
| VRREELDVEGPRTEVPELVGVSR | A0R1B3 | 14.584 | 94.098 | 1 | 0 | 0 |
| VSLDEATPVANGVLTTNTEEQAR | A0QWV1 | 26.181 | 16.187 | 1 | 0 | 0 |
| VTDNPLFTPLDQPR | A0QV70 | 94.584 | 43.306 | 1 | 0 | 0 |
| VVAAIAEQAAK | A0QV52 | 83.647 | 49.414 | 1 | 0 | 0 |
| VVDVPYAEIVASVSSASAGPGTR | A0R4S7 | 121.82 | 32.324 | 2 | 0 | 0 |
| VVEACNDLHSAGR | A0QQH0 | 147.74 | 36.595 | 3 | 0 | 0 |
| VVEGTLAADLK | A0QYZ6 | 52.579 | 16.098 | 1 | 0 | 0 |
| VVEHEALSDETLR | A0R7H5 | 57.885 | 106.556 | 2 | 0 | 0 |
| VVNTVLADLGHETLDTSDYR | A0QWT3 | 31.917 | 42.590 | 1 | 0 | 0 |
| VWEYNLPARYER | A0R5M3 | 69.03 | 46.291 | 2 | 0 | 0 |
| WGDEPIER | A0R5L3 | 48.036 | 62.870 | 1 | 0 | 0 |
| WLDPSHGGINLGFPQNK | A0QV51 | 57.175 | 52.259 | 1 | 0 | 0 |
| WPSGIKDGAEFVIPTMK | A0QSJ2 | 62.162 | 54.304 | 1 | 0 | 0 |
| YAGKGEVIKGDDKTIR | A0R5M3 | 46.89 | 46.291 | 1 | 0 | 0 |
| YDNYIGGEWVAPVEGR | A0QSN7 | 113.24 | 55.979 | 1 | 0 | 0 |
| YFENPTPVTGQVFCEVAR | A0QQW5;A0QSN7 | 31.881 | 55.126; 55.979 | 1 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| YTEDGPHAELLGEK | A0QND7 | 80.632 | 31.537 | 1 | 0 | 0 |
| YVGVSSYSAAK | A0R716 | 79.986 | 39.498 | 1 | 0 | 0 |

1: Contaminant sequences or those found in blank runs excluded

2: Uniprot

3: Molecular weight (MW) of protein ID

4: Only contaminant peptides present

5: No peptides present

**Table 8-3**. SDS-PAGE LC-MS/MS results

| Score | Expectation | Protein ID | Protein Name | MW (kDa) | % Coverage | Comment |
|---|---|---|---|---|---|---|
| 12372 | 0 | A0R0X1 | Major membrane protein I | 33.706 | 61.2 | |
| 1414 | 0 | A0R1H7 | Fatty acid synthase | 329.334 | 11.7 | |
| 406 | $1.50x10^{-37}$ | A0R4H0 | 29 kDa antigen Cfp29 | 28.713 | 26.4 | |
| 285 | $2.10x10^{-25}$ | A0QYK4 | Sodium:solute symporter | 61.554 | 8.2 | |
| 147 | $1.50x10^{-11}$ | A0QYY6 | 30S ribosomal protein S1 | 53.283 | 8.1 | |
| 139 | $8.30x10^{-11}$ | A0QRP7 | TROVE domain protein | 61.145 | 8.2 | |
| 119 | $8.30x10^{-9}$ | A0QZY7 | LppL protein | 35.048 | 13.9 | |
| 118 | $1.20x10^{-8}$ | A0QRB1 | Amino acid carrier protein | 54.284 | 6.2 | |
| 104 | $2.50x10^{-7}$ | A0QT42 | ABC transporter | 38.744 | 15.3 | |
| 102 | $3.90x10^{-7}$ | A0QSL5 | 30S ribosomal protein S13 | 14.209 | 21.8 | |
| 98 | 0.0000011 | A0R157 | Saccharopine dehydrogenase | 43.606 | 8.9 | |
| 92 | 0.0000041 | A0QZ46 | Proteasome subunit alpha | 26.899 | 7.7 | |
| 89 | 0.0000086 | A0QT21 | Cytosine/purine/uracil/thiamine/allantoin permease family protein | 50.983 | 5.2 | |
| 86 | 0.000016 | A0R4G9 | Dyp-type peroxidase | 37.193 | 6.4 | |
| 85 | 0.00002 | A0QVU2 | 35 kDa protein | 30.319 | 19.4 | |
| 84 | 0.000024 | A0R152 | Ribonuclease E | 112.695 | 4.7 | |
| 82 | 0.000044 | A0R5H5 | Anion-transporting ATPase | 36.813 | 9.9 | |
| 82 | 0.000046 | A0QTI0 | L-seryl-tRNA(Sec) selenium transferase | 43.941 | 4.9 | |
| 80 | 0.000073 | A0QS66 | DNA-directed RNA polymerase subunit beta' | 146.422 | 3 | |
| 65 | 0.002 | A0QUM6 | Hydrogenase-2, small subunit | 35.147 | 3.4 | Tentative 1 significant peptide |
| 61 | 0.0052 | A0R150 | 50S ribosomal protein L27 | 9.225 | 8 | Tentative 1 significant peptide |

| | | | | | | |
|---|---|---|---|---|---|---|
| 61 | 0.0054 | A0R052 | Ubiquinol-cytochrome c reductase cytochrome b | 60.171 | 3.8 | |
| 60 | 0.0065 | A0QTH0 | Cationic amino acid transporter | 52.088 | 3 | Tentative 1 significant peptide |
| 58 | 0.011 | A0QSL8 | DNA-directed RNA polymerase subunit | 37.896 | 3.4 | Tentative 1 significant peptide |
| 54 | 0.025 | A0R3S7 | Non-homologous end joining protein Ku | 35.69 | 8.5 | |
| 54 | 0.028 | A0QRS0 | ABC-type transport system periplasmic substrate-binding protein | 44.507 | 5.1 | |
| 54 | 0.025 | P60281 | DNA-directed RNA polymerase subunit beta | 128.451 | 3.9 | |
| 53 | 0.033 | A0R311 | Sodium/proline symporter | 53.6 | 1.6 | Tentative 1 significant peptide |
| 49 | 0.076 | A0QQF8 | Glycosyl hydrolase family protein 76 | 43.431 | 2.8 | Tentative 1 significant peptide |
| 48 | 0.11 | A0QX32 | Band 7 protein | 43.956 | 5.1 | |
| 47 | 0.14 | A0R1B7 | HNH endonuclease family protein | 21.487 | 8.9 | Tentative 1 significant peptide |
| 43 | 0.34 | A0QWJ2 | Protein translocase subunit SecD | 63.607 | 2.4 | |
| 42 | 0.39 | A0QU53 | Putative acyl-CoA dehydrogenase | 43.481 | 3.7 | |
| 39 | 0.92 | A0QR89 | Geranylgeranyl reductase | 43.068 | 1.8 | Tentative 1 significant peptide |

| | | | | | | |
|---|---|---|---|---|---|---|
| 36 | 1.7 | A0QTV6 | Monooxygenase | 42.711 | 2.6 | Tentative 1 significant peptide |
| 36 | 1.8 | A0QTD8 | RNA polymerase sigma-F factor | 27.867 | 5.6 | Tentative 1 significant peptide |
| 34 | 2.4 | A0R2I2 | 2,4-dienoyl-coA reductase | 72.839 | 1.2 | Tentative 1 significant peptide |
| 34 | 2.4 | A0R310 | Aerobic C4-dicarboxylate transport protein | 47.488 | 1.8 | Tentative 1 significant peptide |
| 22 | 44 | A0QS34 | Uncharacterized protein | 32.942 | 3.2 | Tentative 1 significant peptide |
| 0 | 0 | unassigned | unassigned | | | |

**Table 8-4. Number of GroEL particles counted for each grid property**

| Grid Property | Experiment | GroEL (µg) | Image | Count |
|---|---|---|---|---|
| Hydrophilic + Positive | 1 | 1 | 1 | 6 |
| Hydrophilic + Positive | 1 | 1 | 2 | 5 |
| Hydrophilic + Positive | 1 | 1 | 3 | 9 |
| Hydrophilic + Positive | 1 | 1 | 4 | 8 |
| Hydrophilic + Positive | 1 | 1 | 5 | 12 |
| Hydrophilic + Positive | 1 | 2 | 1 | 23 |
| Hydrophilic + Positive | 1 | 2 | 2 | 6 |
| Hydrophilic + Positive | 1 | 2 | 3 | 16 |
| Hydrophilic + Positive | 1 | 2 | 4 | 10 |
| Hydrophilic + Positive | 1 | 3 | 1 | 15 |
| Hydrophilic + Positive | 1 | 3 | 2 | 10 |
| Hydrophilic + Positive | 1 | 3 | 3 | 7 |
| Hydrophilic + Positive | 1 | 4 | 1 | 53 |
| Hydrophilic + Positive | 1 | 4 | 2 | 41 |
| Hydrophilic + Positive | 1 | 4 | 3 | 35 |
| Hydrophilic + Positive | 1 | 4 | 4 | 27 |
| Hydrophilic + Positive | 1 | 5 | 1 | 9 |
| Hydrophilic + Positive | 1 | 5 | 2 | 7 |
| Hydrophilic + Positive | 1 | 5 | 3 | 10 |
| Hydrophilic + Positive | 1 | 5 | 4 | 22 |
| Hydrophilic + Positive | 1 | 5 | 5 | 12 |
| Hydrophilic + Positive | 1 | 5 | 6 | 19 |
| Hydrophilic + Positive | 1 | 5 | 7 | 13 |

| | | | | |
|---|---|---|---|---|
| Hydrophilic + Positive | 2 | 3 | 1 | 9 |
| Hydrophilic + Positive | 2 | 3 | 2 | 6 |
| Hydrophilic + Positive | 2 | 3 | 3 | 7 |
| Hydrophilic + Positive | 2 | 3 | 4 | 5 |
| Hydrophilic + Positive | 2 | 3 | 5 | 8 |
| Hydrophilic + Positive | 2 | 3 | 6 | 7 |
| Hydrophilic + Positive | 2 | 3 | 7 | 7 |
| Hydrophilic + Positive | 2 | 3 | 8 | 23 |
| Hydrophilic + Positive | 2 | 3 | 9 | 9 |
| Hydrophilic + Positive | 2 | 3 | 10 | 7 |
| Hydrophilic + Positive | 2 | 4 | 1 | 11 |
| Hydrophilic + Positive | 2 | 4 | 2 | 6 |
| Hydrophilic + Positive | 2 | 4 | 3 | 6 |
| Hydrophilic + Positive | 2 | 4 | 4 | 5 |
| Hydrophilic + Positive | 2 | 4 | 5 | 7 |
| Hydrophilic + Positive | 2 | 4 | 6 | 2 |
| Hydrophilic + Positive | 2 | 4 | 7 | 4 |
| Hydrophilic + Positive | 2 | 4 | 8 | 2 |
| Hydrophilic + Positive | 2 | 4 | 9 | 4 |
| Hydrophilic + Positive | 2 | 4 | 10 | 7 |
| Hydrophilic + Negative | 1 | 1 | 1 | 10 |
| Hydrophilic + Negative | 1 | 1 | 2 | 12 |
| Hydrophilic + Negative | 1 | 1 | 3 | 9 |
| Hydrophilic + Negative | 1 | 1 | 4 | 6 |
| Hydrophilic + Negative | 1 | 4 | 1 | 10 |

| | | | | |
|---|---|---|---|---|
| Hydrophilic + Negative | 1 | 4 | 2 | 6 |
| Hydrophilic + Negative | 1 | 4 | 3 | 6 |
| Hydrophilic + Negative | 2 | 1 | 1 | 15 |
| Hydrophilic + Negative | 2 | 1 | 2 | 8 |
| Hydrophilic + Negative | 2 | 1 | 3 | 6 |
| Hydrophilic + Negative | 2 | 1 | 4 | 6 |
| Hydrophilic + Negative | 2 | 1 | 5 | 18 |
| Hydrophilic + Negative | 2 | 1 | 6 | 9 |
| Hydrophilic + Negative | 2 | 1 | 7 | 15 |
| Hydrophilic + Negative | 2 | 1 | 8 | 14 |
| Hydrophilic + Negative | 2 | 1 | 9 | 10 |
| Hydrophilic + Negative | 2 | 1 | 10 | 15 |
| Hydrophilic + Negative | 2 | 2 | 1 | 16 |
| Hydrophilic + Negative | 2 | 2 | 2 | 18 |
| Hydrophilic + Negative | 2 | 2 | 3 | 11 |
| Hydrophilic + Negative | 2 | 2 | 4 | 12 |
| Hydrophilic + Negative | 2 | 2 | 5 | 13 |
| Hydrophilic + Negative | 2 | 2 | 6 | 9 |
| Hydrophilic + Negative | 2 | 2 | 7 | 7 |
| Hydrophilic + Negative | 2 | 2 | 8 | 8 |
| Hydrophilic + Negative | 2 | 2 | 9 | 9 |
| Hydrophilic + Negative | 2 | 2 | 10 | 11 |
| Hydrophilic + Negative | 2 | 3 | 1 | 15 |
| Hydrophilic + Negative | 2 | 3 | 2 | 16 |
| Hydrophilic + Negative | 2 | 3 | 3 | 13 |

| | | | | |
|---|---|---|---|---|
| Hydrophilic + Negative | 2 | 3 | 4 | 14 |
| Hydrophilic + Negative | 2 | 3 | 5 | 8 |
| Hydrophilic + Negative | 2 | 3 | 6 | 15 |
| Hydrophilic + Negative | 2 | 3 | 7 | 7 |
| Hydrophilic + Negative | 2 | 3 | 8 | 9 |
| Hydrophilic + Negative | 2 | 3 | 9 | 10 |
| Hydrophilic + Negative | 2 | 3 | 10 | 6 |
| Hydrophilic + Negative | 2 | 4 | 1 | 16 |
| Hydrophilic + Negative | 2 | 4 | 2 | 14 |
| Hydrophilic + Negative | 2 | 4 | 3 | 11 |
| Hydrophilic + Negative | 2 | 4 | 4 | 9 |
| Hydrophilic + Negative | 2 | 4 | 5 | 13 |
| Hydrophilic + Negative | 2 | 4 | 6 | 8 |
| Hydrophilic + Negative | 2 | 4 | 7 | 10 |
| Hydrophilic + Negative | 2 | 4 | 8 | 15 |
| Hydrophilic + Negative | 2 | 4 | 9 | 12 |
| Hydrophilic + Negative | 2 | 4 | 10 | 8 |
| Hydrophilic + Negative | 2 | 5 | 1 | 6 |
| Hydrophilic + Negative | 2 | 5 | 2 | 12 |
| Hydrophilic + Negative | 2 | 5 | 3 | 10 |
| Hydrophilic + Negative | 2 | 5 | 4 | 8 |
| Hydrophilic + Negative | 2 | 5 | 5 | 4 |
| Hydrophilic + Negative | 2 | 5 | 6 | 9 |
| Hydrophilic + Negative | 2 | 5 | 7 | 11 |
| Hydrophilic + Negative | 2 | 5 | 8 | 11 |

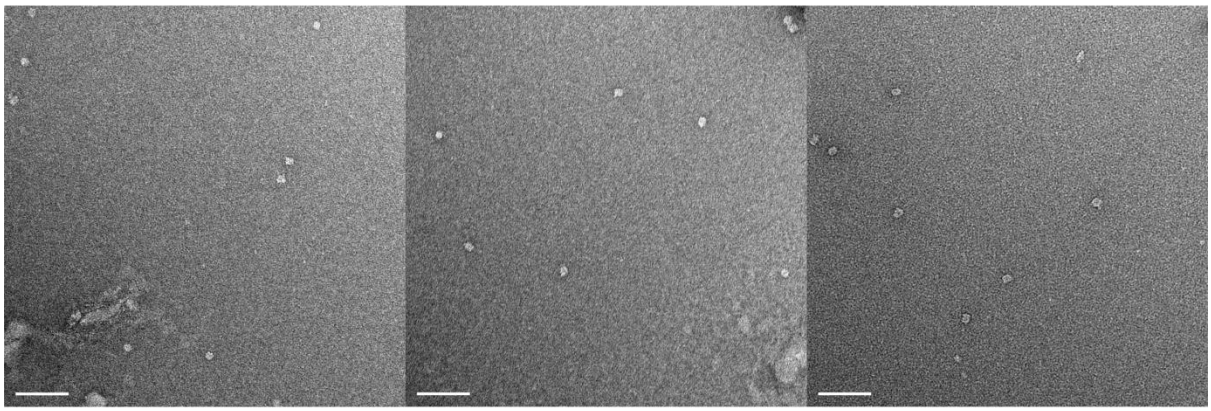
| | | | | |
|---|---|---|---|---|
| Hydrophilic + Negative | 2 | 5 | 9 | 9 |
| Hydrophilic + Negative | 2 | 5 | 10 | 6 |
| Hydrophobic + Positive | 1 | 1 | 1 | 6 |
| Hydrophobic + Positive | 1 | 2 | 1 | 29 |
| Hydrophobic + Positive | 1 | 2 | 2 | 16 |
| Hydrophobic + Positive | 1 | 2 | 3 | 10 |
| Hydrophobic + Positive | 1 | 4 | 1 | 7 |
| Hydrophobic + Positive | 1 | 4 | 2 | 4 |
| Hydrophobic + Positive | 1 | 5 | 1 | 13 |
| Hydrophobic + Positive | 1 | 5 | 2 | 9 |
| Hydrophobic + Positive | 1 | 5 | 3 | 7 |

**Calculation of mg Protein Per Protein Copy Number**

mg Protein =

[(#Cells at $OD_{600}$) x mL grown x (Protein Copy #) x (Mass Protein (Da)) x 1000]/$6.023x10^{23}$

1 µg



4 µg

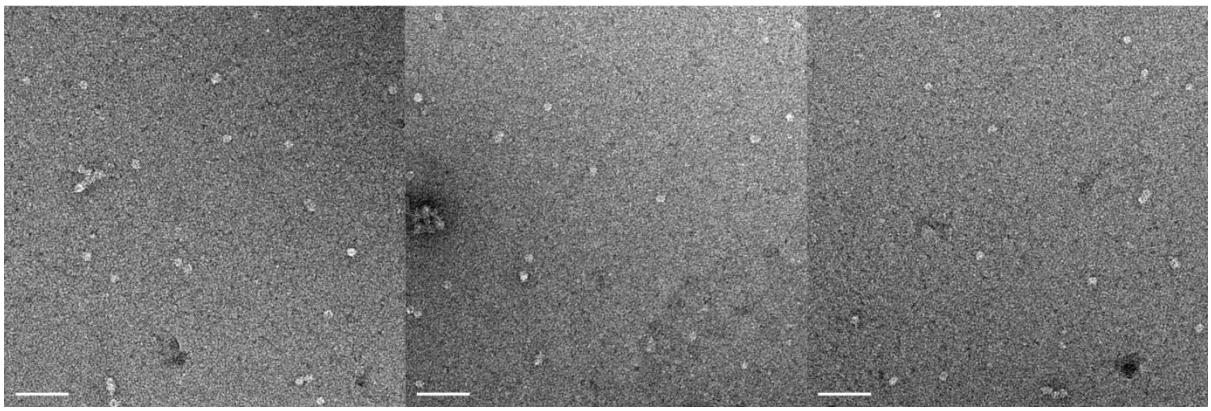
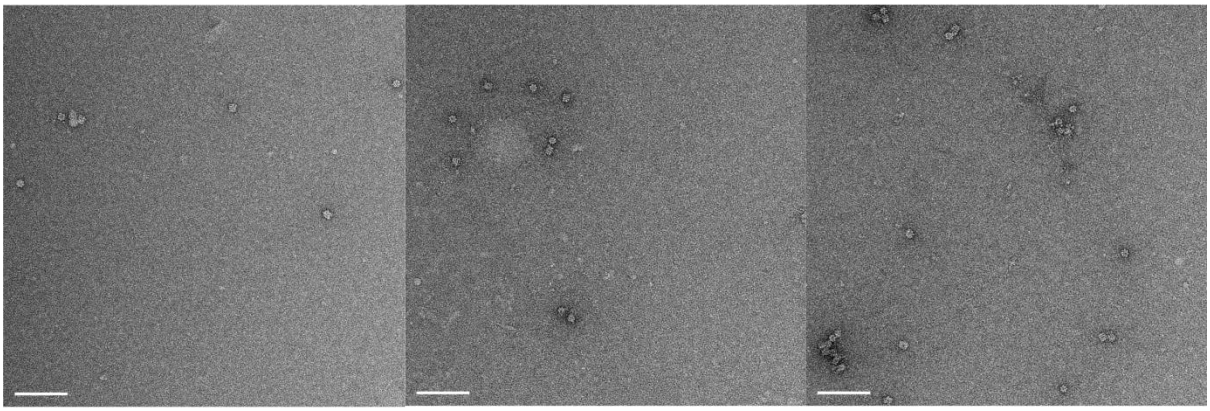
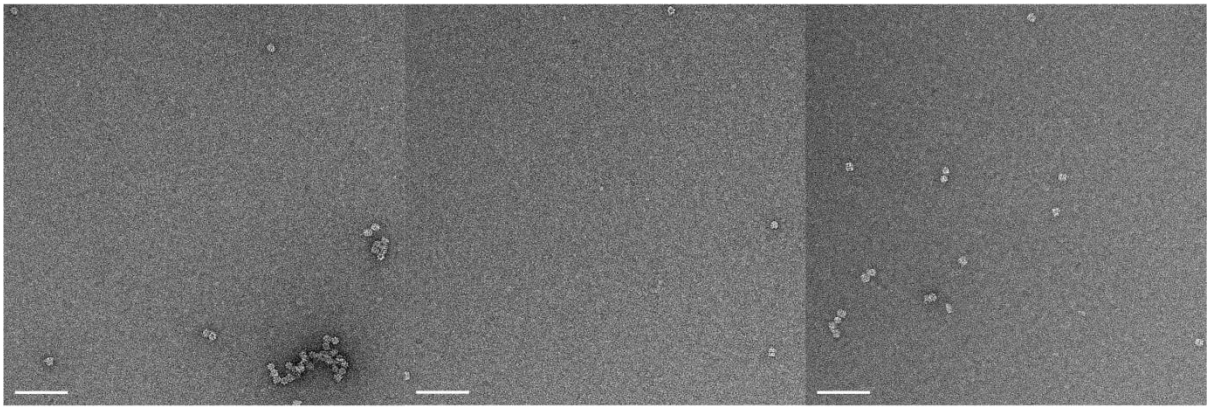


**Figure 8-4**. Distribution of GroEL on hydrophilic and negatively charged copper grids after grid blotting. Scale bar (white) shows 100 nm. Electron micrographs taken at x50,000 magnification on an F20 Technai microscope at a defocus of 2.00 µm.

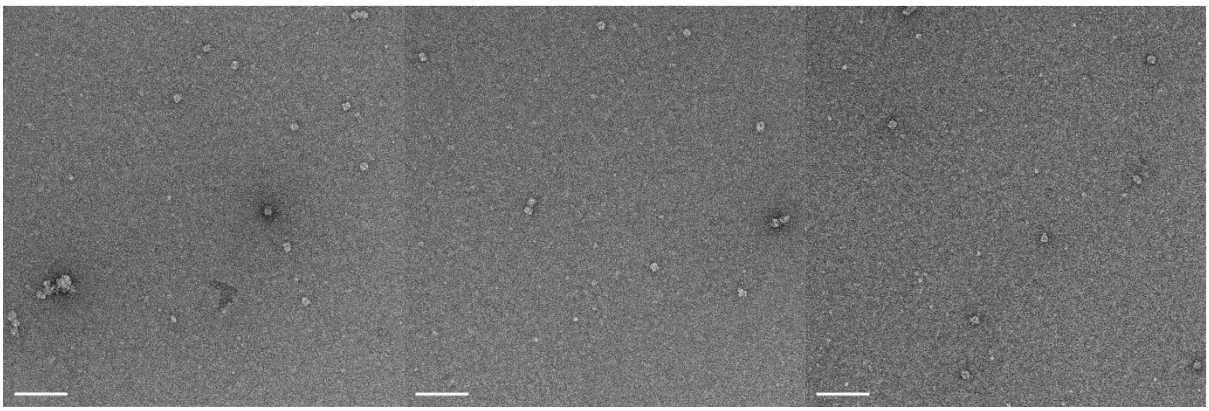
1 µg



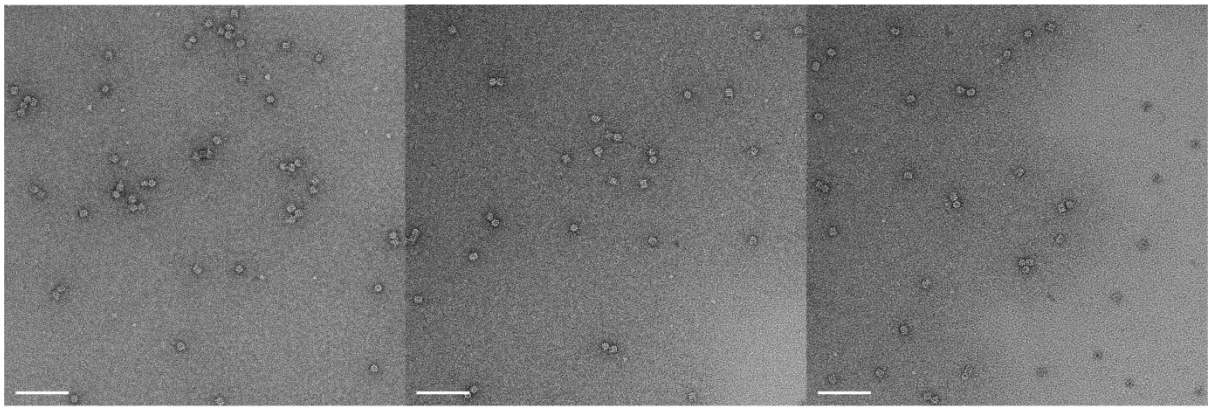
2 µg



3 µg



4 µg



5 µg

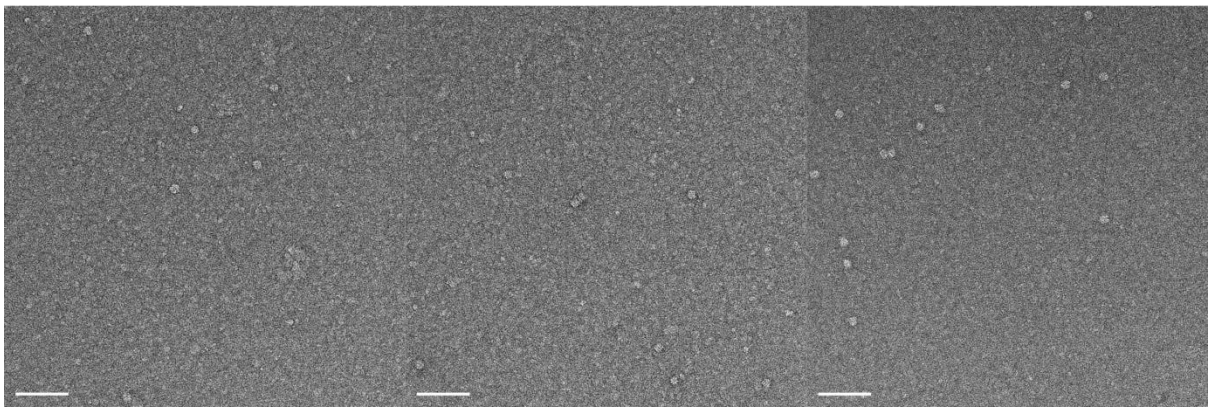


**Figure 8-5**. Distribution of GroEL on hydrophilic and positively charged copper grids after grid blotting. Scale bar (white) shows 100 nm. Electron micrographs taken at x50,000 magnification on an F20 Technai microscope at a defocus of 2.00 µm.

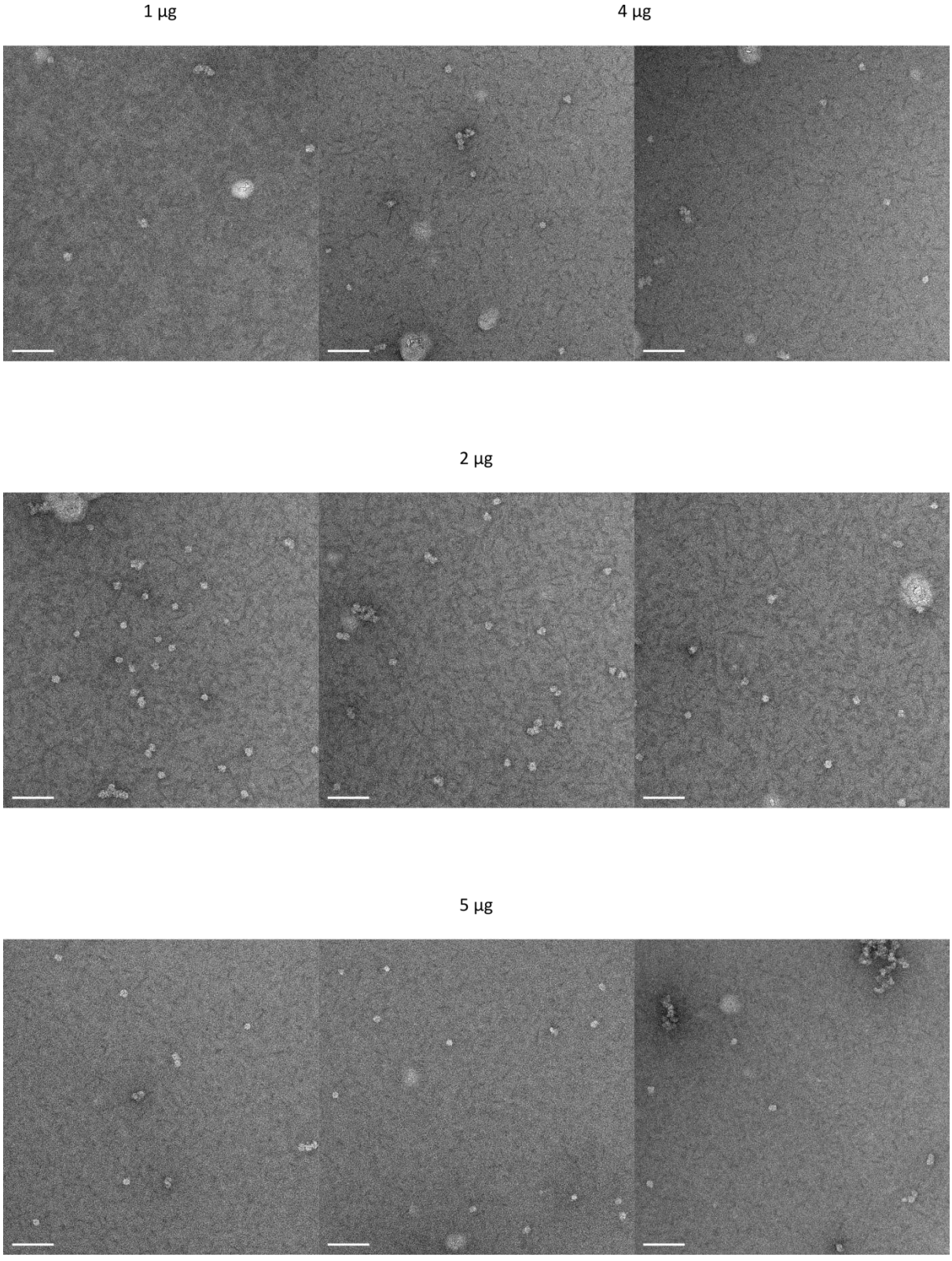



**Figure 8-6**. Distribution of GroEL on hydrophobic and positively charged copper grids after grid blotting. Scale bar (white) shows 100 nm. Electron micrographs taken at x50,000 magnification on an F20 Technai microscope at a defocus of 2.00 µm.

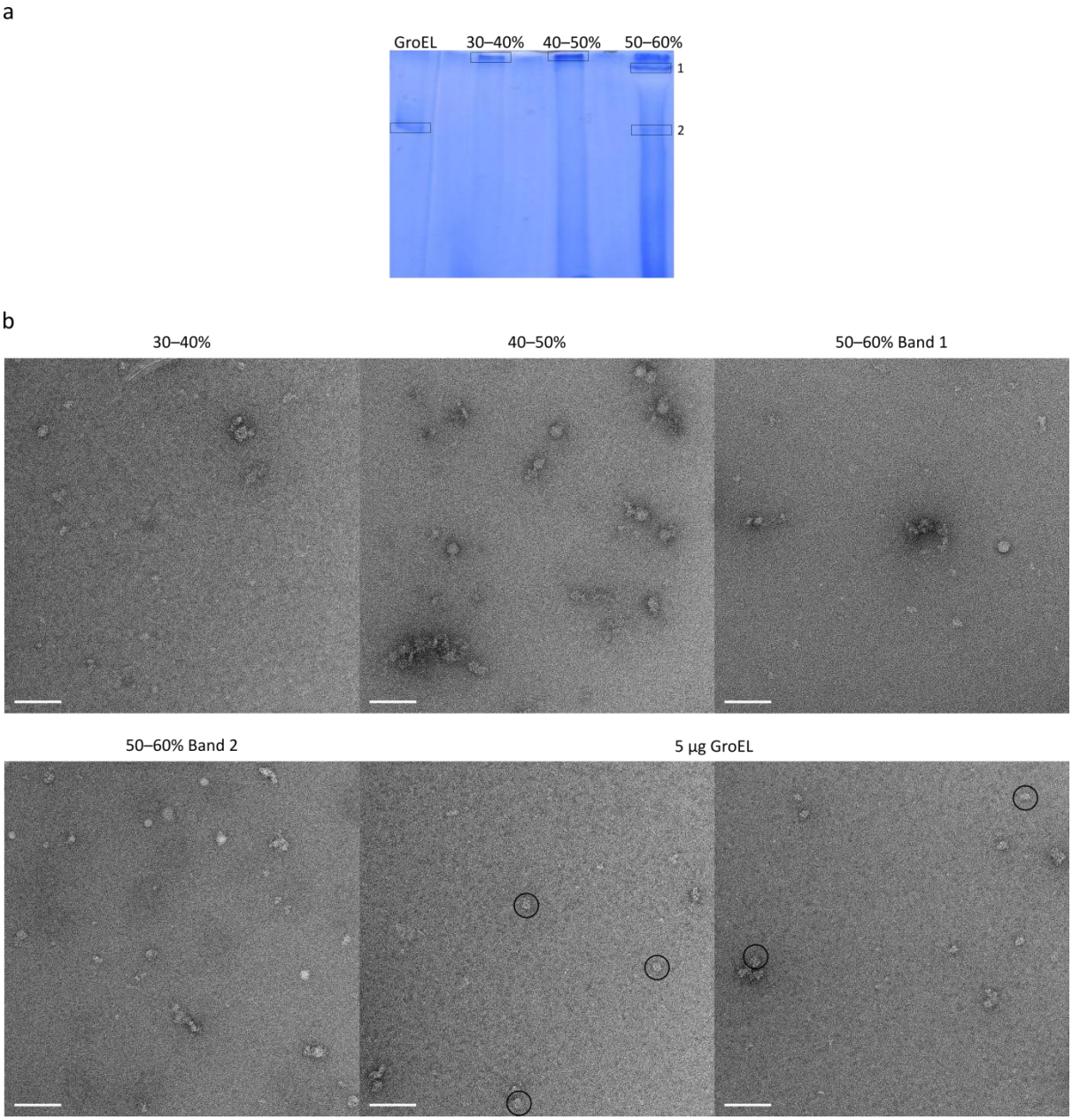



**Figure 8-7. Grid blotting *Msm* ammonium sulphate cuts**. a) 30−40%, 40−50%, and 50−60% ammonium sulphate cuts were run on a blue native PAGE gel and visible bands grid blotted (black box). 5 µg of GroEL was used as a control. b) No intact protein particles were observed on an electron micrograph for the ammonium sulphate cuts. Intact GroEL particles were observed (circled). Electron micrographs taken at x50,000 magnification on an F20 Technai microscope at a defocus of 2.00 µm. White scale bar shows 100 nm.

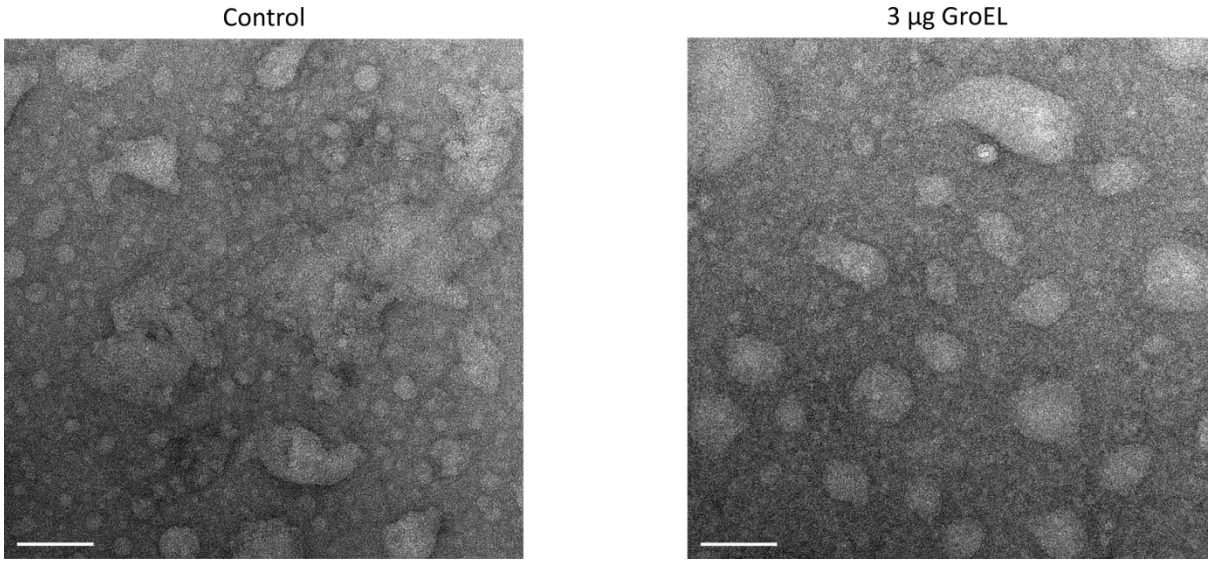
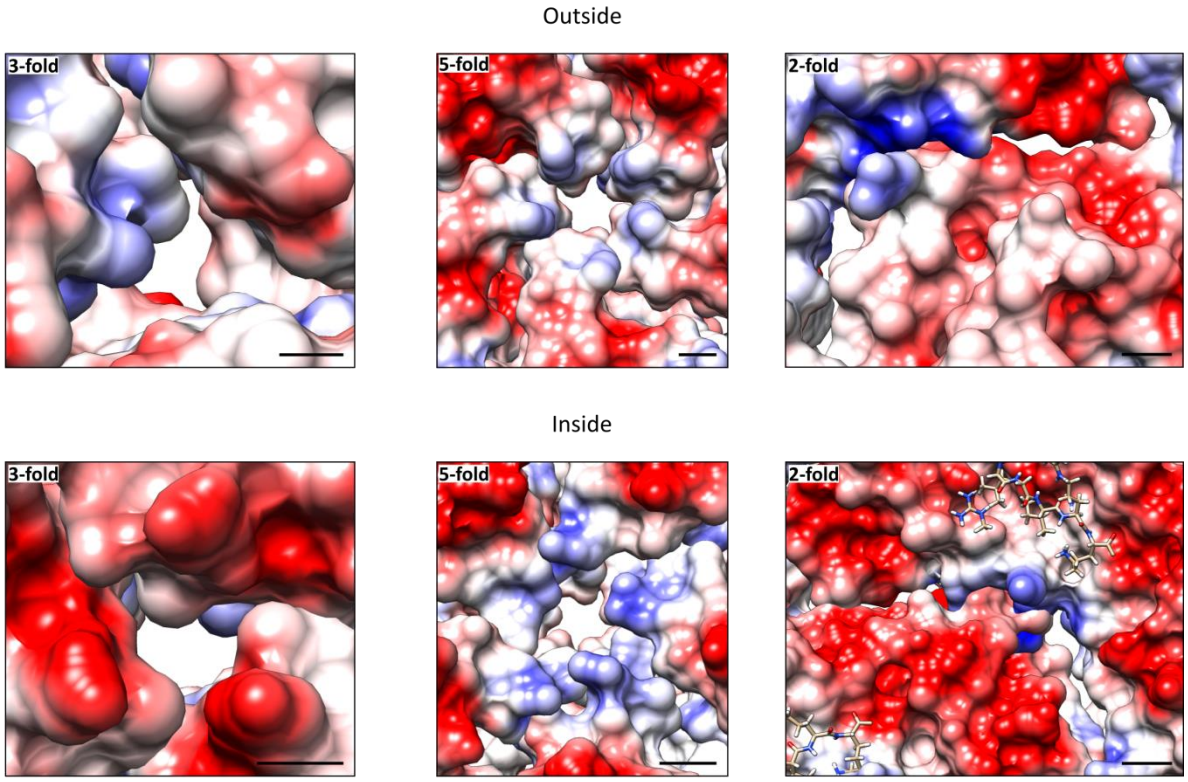



**Figure 8-8. Electro-elution on blue native PAGE.** The technique was not successful as electro-elution for 10 minutes on the control (empty part of the gel) yielded the same non-protein material as that of a band containing GroEL particles. White scale bar shows 100 nm. Electron micrographs taken at x50,000 magnification on an F20 Technai at a defocus of 2.00 µm.

**a**


Outside
3-fold
5-fold
2-fold
Inside
3-fold
5-fold
2-fold

**b**

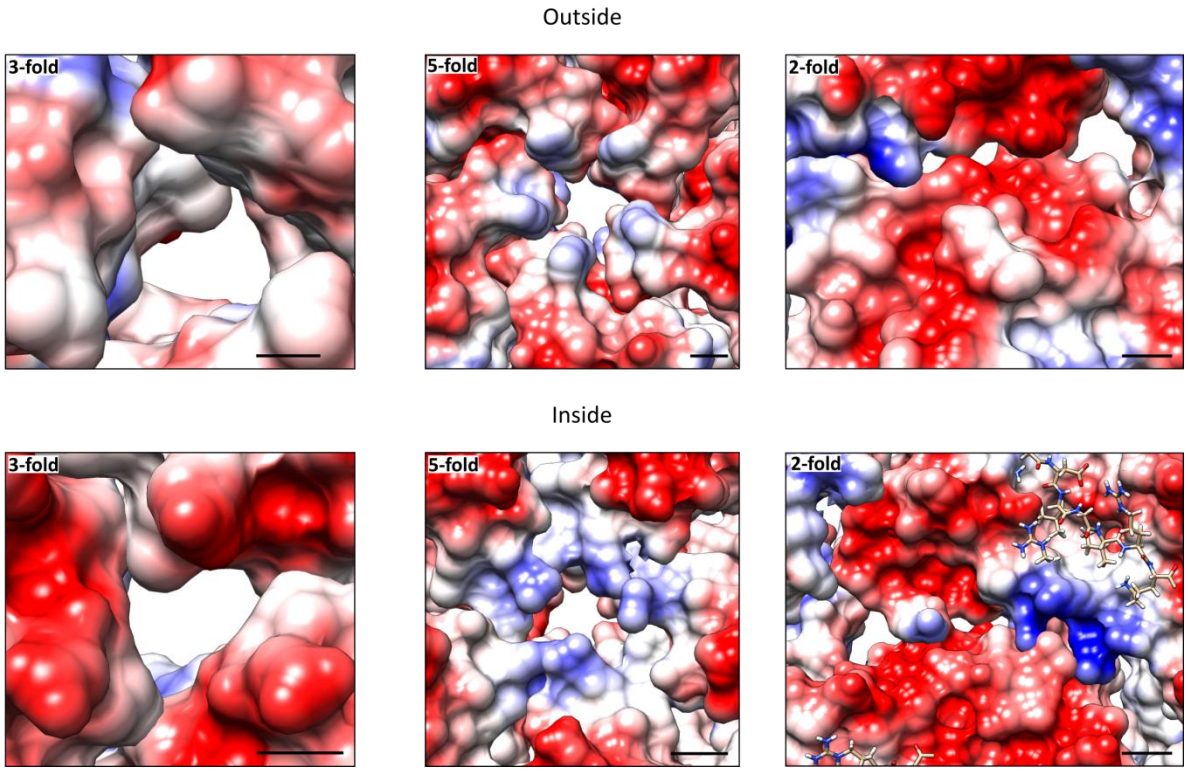

Outside
3-fold
5-fold
2-fold
Inside
3-fold
5-fold
2-fold

**Figure 8-9. Electrostatic potential (previous page).** A red (-10 kcal/mol·$e^-$) to blue (+10 kcal/mol·$e^-$) gradient is given for *Msm* (a) and *Mtb* (b) Enc pores. White indicates no charge. Black scale bars show 5 Å. The symmetry is not exact as they occur across loops in the *T. maritima* Enc crystal structure which is difficult to model.