The structure of a natural competency protein from the thermophilic bacterium *Thermus thermophilus*

A thesis submitted to the University of Manchester for the degree of Master of philosophy (MPhil) in Biochemistry.

2015

Matthew Snee

Contents

List of figures

Word count: 197686

Abstract

Natural transformation is an important mode of horizontal gene transmission in bacteria, a process that underlies the ability of these organisms to evolve rapidly and diverge into the vast multitude of forms that are observed in nature. The sharing of bacterial genes has become of increasing significance when studying the causes of bacterial drug resistance. Despite its importance, the state of competency that allows the uptake of DNA from the environment remains poorly understood in gram-negative bacteria. The highly thermophilic eubacterium *Thermus thermophilus* has become a focus of work seeking to elucidate the steps involved in natural transformation, the underlying developmental state of competency, and its link to type IV pilus formation. The protein ComZ has been shown to be required for competency in *T. thermophilus* but contains no detectable homology to other protein components of the apparatus or indeed any known proteins, with the exception of ComZ from very closely related species. During this project, ComZ was expressed, purified to homogeneity and studied via X-ray crystallography and yielded experimental phase data that allowed computation of an electron density map at 3.5 Å resolution. In addition, the thermostability and domain architecture of ComZ was also studied by other methods. An initial model of ComZ is presented here, which shows a two domain structure comprising an alpha/beta domain and a 3 stranded parallel beta helix. This structure suggests that ComZ might have an enzymatic or binding role in *natural competence.*

Matthew Snee

University of Manchester

26/09/2015

Declaration

No part of this thesis has been submitted in support of an application for any degree or qualification of The University of Manchester or any other University or Institute of learning.

Copyright statement

1 Introduction

1.1 Horizontal gene transfer in Bacteria

The first well-known observation of natural transformation was made by Griffith (1928) who noticed the transformation of pneumococcal bacteria from a benign to virulent state. Transformation in the same organism was used by Avery, McLeod and McCarty during the work that proved that DNA was the molecule that encoded heritable information in living creatures (Avery *et al.,* 1944). It is now known that natural transformation using donor DNA harvested directly from the environment is only one of 3 major methods for horizontal transfer of genetic material in bacteria, alongside virus mediated transduction and conjugation, a process that it also known as bacterial sex (Tomoeda *et al.,* 1976). Transduction occurs when genetic material is transferred via a phage and can occur when fragments of the host genome are accidentally packaged into new virions and recombine with the genome of the next infected cell (generalised). Alternatively, it may occur when a phage that relies on integrating its genome with that of the host excises host sequences adjacent to its own genetic material (specialised) (Chan and Botstein, 1976; Sternberg and Maurer, 1991). The evolution of highly virulent strains of pathogens like *Vibrio cholera, Escherichia coli* and *Staphylococcus aureus* has resulted from the expression of virulence factors such as superantigens and shiga toxin that have been transferred by transduction (Brussow *et al.,* 2004). Conjugation involves a specialised pilus to connect two cells, allowing the transfer of plasmids. It is believed that inter-species conjugation is responsible for the transfer of antibiotic resistance genes from their organisms of origin, (species that produce, or coexist with species that produce, antibiotics) to pathogenic organisms (Davies, 1994). The machinery that performs the task of conjugation has two evolutionary sources. The DNA processing components are related to proteins involved in DNA replication, and the delivery system is a specialised form of the type IV secretion system (Waters and Guiney, 1993). Related systems are used to deliver effector molecules into host cells in plant and animal pathogens. Examples include the CagA system of *Helicobacter pylori* and the genetic modification apparatus of *Agrobacterium tumifaciens* (Cascales and Christie, 2003, 2004).

As an extreme *thermophile, T. thermophilus* has become an interesting model for studying the processes of horizontal gene transfer in gram negatives. Kinetic studies by Schwarzenlander and Averhoff (2006) have demonstrated an extraordinary DNA uptake rate of 40,000 bases per second per cell, and genetic studies have demonstrated a large amount of horizontal gene transmission in the evolutionary past of this organism (Omelchenko *et al.,* 2005). Studying the methods that this particular bacterium uses to acquire new genes can therefore yield important information on how these complex systems can function in extreme environments, as well as shed light on generalised principles of horizontal gene transmission in bacteria.

1.2 Natural transformation

Natural transformation, is unique in that there is still widespread debate on its purpose. Attempts to resolve this question have been confounded by the myriad of different ways in which competency for natural transformation is regulated. Some organisms such as *T. thermophilus* are constitutively competent and require no external stimuli (Hidaka *et al.,* 1994), whereas others such as *Haemophilus influenzae* and *Bacillus subtilis* become competent in response to nutrient limitation, suggesting that DNA may be taken up as a source of nutrients (Redfield, 1993). In the aquatic environment, *Vibrio cholerae* down-regulates the extracellular nuclease that normally provides nucleotides as a carbon source from exogenous DNA and becomes competent in response to chitin using a quorum sensing mechanism (Blokesch and Schoolnik, 2008). This observation suggests that DNA is used for adaptation and allows rapid spread of traits that allow these bacteria to make use of scarce resources. The well-known pathogen *Streptococcus pneumonia*e which was studied in the Avery, MacLeod, McCarty experiment has since been shown to become competent in response to stress caused by antibiotics (Prudhomme *et al.,* 2006). This indicates that natural transformation could be used to repair damaged DNA in stressful environments. The DNA for repair theory is bolstered by the observation of DUS uptake sequences in organisms such as *H. influenzae* and various *Neisseria* species, which limit uptake of DNA to closely related organisms. These sequences are disproportionately distributed around genes with functions related to genome maintenance (Davidsen *et al.,* 2004). It is perhaps the case that competency serves different purposes depending on the bacteria in question. It is not easy to assess the evolutionary pressures that maintain competency in *T. thermophilus.* It has been shown that DNA uptake occurs at the rate of 1.5 μg (mg protein)$^{-1}$S^{-1} at optimal concentrations, with no specificity for DNA from its own species or domain (Schwarzenlander and Averhoff, 2006). Whether adaptability is the main reason for maintenance of competency in *T. thermophilus* is not clear. However, it has been shown that horizontal gene transmission has been important in its adaptation to its extreme environment, providing genes encoding thermophilic proteins involved in mineral uptake and even ribosomal components (Omelchenko *et al.,* 2005).

1.3 Type IV pilus Biogenesis

With a few exceptions, DNA uptake in gram negative bacteria has a strong link to biogenesis of type IV pili. Early experiments detected the association between the non-pilated and non-competent phenotypes in multiple competent gram negative species including various species of *Moraxella*, *Neisseria gonorrhoeae,* and *Thermus thermophilus* (Bovre and Froholm, 1972; Biswas *et al.,* 1977; Freidrich *et al.,* 2002). Targeted knock-out studies in *T. thermophilus* have revealed that, although many genes are required for both the morphological phenotype of pilation and the ability to undergo natural transformation, many are only required for a single process. With the exception of ComZ and DprA, these competency-specific proteins are all homologous to pilin proteins and are termed pilin-like (Freidrich *et al.,* 2003). It is not clear whether the DNA uptake apparatus represents a special variant of the type IV pilus, termed a “pseudopilus”, or the activity is present in all the pili

that are observed on the surface of wild type cells. Type IV pili are believed to originate from an inner membrane complex that has been structurally characterised in detail (Karuppiah *et al*., 2013). Assembly of this platform is thought to begin with the assembly of a dimer formed from 2 copies of the membrane protein PilN. This is followed by association with a second protein, PilM, recruited from the cytosol. A second membrane protein, PilO binds to each copy of PilN, disrupting the dimer. This allows the binding of the major pilin, PilA4 and subsequent polymerisation of the pilus fibre (Karuppiah *et al.,* 2013). Additional proteins required include the prepilin peptidase PilD, the outer membrane pore protein PilQ, and an ATPase, PilF, believed to provide the energy for polymerisation or DNA uptake (Schwarzenlander *et al.,* 2009; Burkhardt *et al.,* 2011; Collins *et al.,* 2013).

1.4 DNA uptake

Three different approaches have been employed to study the DNA uptake apparatus of *T. thermophilus*. Firstly information about the role of core components has been inferred from homology to proteins in other organisms such as *N. gonorrhoeae* and the gram positive bacterium *B. subtilis* (Freidrich *et al.,* 2003). Secondly, classical microbiological experiments using knock-out strains have been used to link complex phenotypes with the loss of specific genes (Freidrich *et al.,* 2002, 2003; Schwarzenlander *et al.,* 2009). Thirdly, structural approaches have revealed many of the interactions between the biogenesis platform components, suggesting a mechanism whereby ATP hydrolysis by PilF provides energy for DNA uptake that is transferred through the pilus structure (Karuppiah *et al.,* 2013; Collins *et al.,* 2013). One study by Schwarzenlander *et a*l (2009) investigated knock outs in every gene that has been implicated in competency and attempted to infer the roles of the gene products by measuring the relative amounts of DNAse sensitive and DNAse resistant DNA associated with cells. This study appears highly successful in confirming that many competency genes have equivalent functions to their homologues in other organisms. Examples include directly observed evidence that PilQ is the site of initial DNA binding and ComEC has an inner membrane transport function as has been shown in *N. meningitidis* and *B. subtilis* respectively (Assalkhou *et al.,* 2007; Draskovic and Dubnau, 2005). However, this approach cannot untangle the complex range of possibilities for interactions between DNA and proteins in the periplasmic space. For instance, ComZ was predicted to have a role related to DNA binding, an idea that is problematic when one considers that ComZ lacks the sequence that might allow its release from the inner membrane, and a membrane-bound deep binding protein (ComEA) is already known (Freidrich *et al.,* 2003).

An exhaustive description of the entire DNA uptake machinery is beyond the scope of this thesis, provided here is a brief summary (for a more complete review see Averhoff 2009) outlining the key steps (Fig.1). DNA binds to PilQ and possibly to pilin subunits, PilA1-4 (Schwarzenlander *et al.,* 2009). Internalisation into the periplasm occurs via pilus retraction or conformational changes in the pilus fibre driven by PilF (Collins *et al.,* 2013) allowing the DNA to bind to ComEA, another protein with well characterised homologues in *N. gonorrhoeae* and *B. subtilis* (Chen and Gotschlich, 2001; Provvedi and Dubnau, 2002). It is believed that degradation of one strand of the DNA duplex occurs at some point between binding by ComEA and transfer through the inner membrane channel formed

by ComEC (Assalkhou *et al.,* 2007; Draskovic and Dubnau, 2005). The enzyme, EndA, that carries out this task has been identified and characterised in 2 gram positive species *S. pneumoniae* and *B. subtilis,* but has yet to be found in gram negative bacteria (Lacks *et al.,* 1975; Levine and Strauss, 1965; Provvedi, et al., 2001). Single stranded DNA imported to the cytosol is predicted to bind DprA, the last competence protein in the pathway. This component is predicted to stabilise single-stranded DNA prior to entry into the classical DNA repair/recombination pathway via RecA (Mortier-Barriere *et al.,* 2007;Averhoff, 2009).

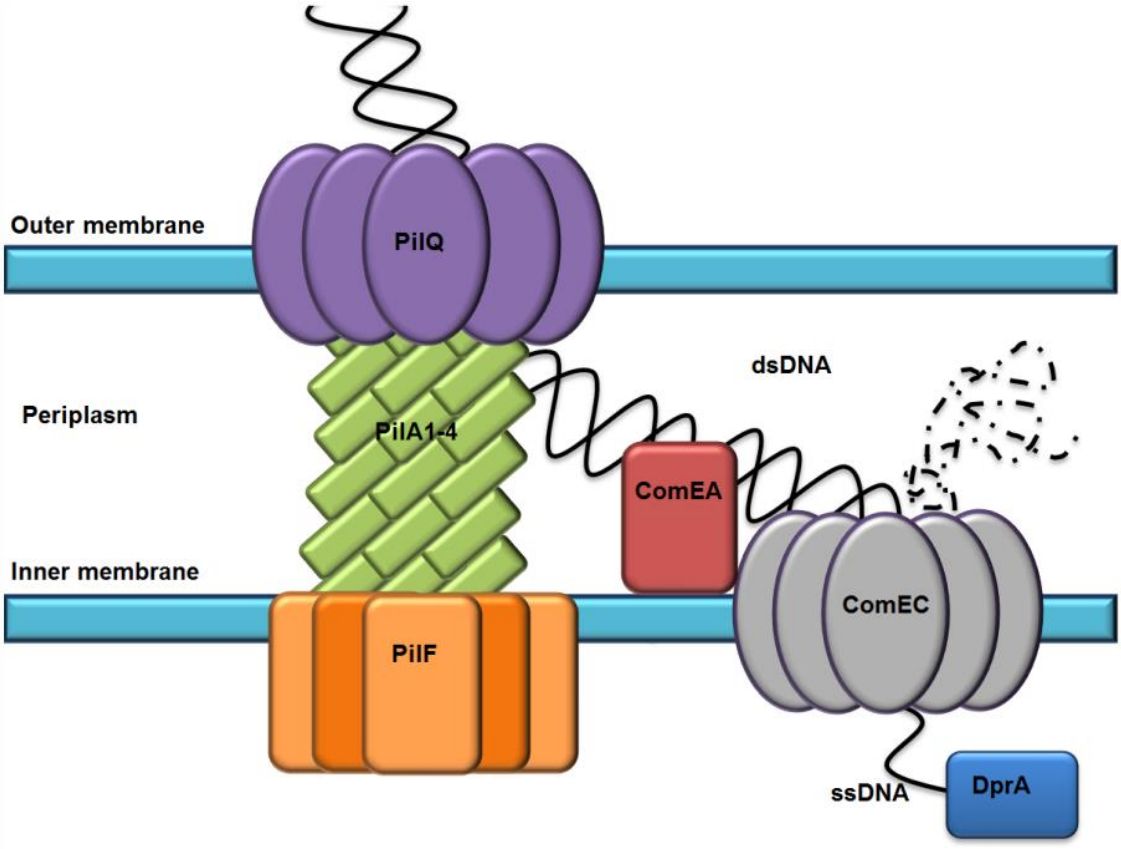


Figure 1*:* A model for DNA uptake in gram negative bacteria *DNA binds to PilQ (purple) and later to the pilin proteins PilA1-4 (green). ATP is hydrolysed by PilF (orange) and DNA is taken into the periplasm. DNA binds ComEA (Red) before travelling through the pore formed by ComEC to enter the cytosol. At some point before entry into the cytosol, one strand is degraded. The final competency protein dprA (blue) binds Single stranded DNA (ssDNA) before it enters the recombination pathway adapted from Freidrich et al (2003).*

1.5 ComZ

During work on natural transformation in *T. thermophilus,* the protein encoded by *orf1698* was identified as being necessary for competency and named ComZ (Freidrich *et al.,* 2003). This gene is co-transcribed along with 4 genes that encode prepilin-like proteins including the major pilin subunit PilA4, but does not appear to be homologous to these, or any other known proteins. Homologues of ComZ do exist in the genomes of related organisms such as *Thermus scotoductus*

(Freidrich *et al.,* 2003; Gounder *et al.,* 2011). The protein is the largest component encoded on its operon with a total of 554 residues (Freidrich *et al.,* 2003). Knockout strains for ComZ in *T. thermophilus exhibit* a pilated phenotype with cells binding slightly more DNA on the cell surface with an equivalent decrease in internalised DNA (Friedrich *et al.,* 2003; Schwarzenlander *et al.,* 2009). From these studies, it was concluded that ComZ must have a role in DNA binding, but it is possible that a loss of activity in any downstream process could cause this intermediate phenotype, which is less pronounced than that caused by loss of PilA4 or PilQ (Schwarzenlander *et al.,* 2009). In this thesis, a structural model of ComZ is presented that suggests some possible functions and may serve as a valuable starting point for future research into its function and interactions.

1.6 Structure as a route to function

X-ray crystallography and electron microscopy have already shed some light on the structural features of the competency machinery of *T. thermophilus* (Karuppiah et al., 2013; Collins et al., 2013) but in ComZ, which lacks obvious sequence homology to any known proteins, structural characterisation has the potential to reveal homology that has been retained despite sequence divergence. Additionally, knowledge of the structure might point to a function via the presence of catalytic motifs or binding pockets. This project aims to determine the function of ComZ using a structural approach. X-ray crystallography projects such as this have 2 major challenges, the need to create good quality diffracting crystals, and the need for phases sufficient to calculate an interpretable electron density map.

Proteins for structural studies are typically produced in recombinant form. This allows effective initial purification by the addition of tags, simplified culture conditions, and the ability to efficiently produce large amounts of protein via overexpression systems (Choi *et al.,* 2006). For soluble proteins of bacterial origin, *E. coli* provides a robust system for converting culture nutrients into recombinant protein, but for proteins that contain disulphide linkages, membrane proteins, and proteins of eukaryotic origin, extra factors require consideration. If the purpose of expressing the protein is to study its structure by X-ray crystallography, the need for very high purity and homogeneity become emphasised (Chayen and Saridakis, 2008). The general rationale for recombinant protein purification is based upon subjecting the protein to a sequence of steps that separate macromolecular components of a solution based on one or more biochemical properties. Theoretically, impurities such as proteins and DNA from the expression organism will only be co-purified with the protein of interest if they are similar with respect to the parameter being used for separation. The generalised probability of one specific impurity being carried through multiple steps can be thought of as the product of multiplying the individual probabilities of a high degree similarity with the target, in relation to each variable being exploited (size, charge, isoelectric point etc.). Therefore, a multi-step purification is much more effective than the sum of its individual steps, although protein-protein binding affinity often allows certain impurities to bind to the target molecule making them harder to remove. Typically, the first stage of preparing soluble recombinant protein from a crude cell extract is purification via its tag. These short sequences are encoded on the

expression vector and usually contain some affinity for a ligand which can be immobilised onto a matrix, over which the cell extract is passed. Increasing concentrations of an elutant, which displaces the bound protein or destabilises the interaction, are used to separate species based on their affinity for the matrix. Examples of this are the polyhistidine tag, which has an affinity for nickel, and is destabilised by imidazole (Hengen, 1995), and the strep tag, a peptide that binds to streptavidin and is removed by application of dissolved biotin (Schmidt and Skerra, 2007). Further steps that may be employed include ion exchange where the protein binds to a charged column and is eluted using increasing concentrations of salt, and size exclusion chromatography where the protein is passed through a dense matrix with pores of sizes that are within similar orders of magnitude to protein molecules. Larger proteins do not readily pass through these pores, and so are eluted faster. Smaller molecules enter the pores easily and so are retained for longer. Size exclusion, with or without dynamic light scattering, can be used to estimate the size of a protein, for instance, to establish if it forms a multimeric complex in solution (Wen *et al.,* 1996). Size exclusion is typically the last step, as it dilutes the protein into a homogenous buffer and removes any elutants. Care must be taken to optimise the conditions for purification and storage of recombinant proteins. At certain pH values, a protein may partially unfold and become prone to aggregating through hydrophobic interactions between exposed core residues. Aggregation can also occur when there is not sufficient salt in the buffer to mask the surface charges. This is a common problem during preparation of a sample for ion exchange which requires removal of salt to allow the protein to bind the column (Cromwell *et al.,* 2006).

1.7 Crystallising a Protein

Although correlations between parameters such and size and shape and the propensity of a protein to crystallise have been observed, and have resulted in attempts to predict the likelihood of crystallisation from sequence (Slabinski *et al.,* 2007), initial screening of a new protein is largely based on trial and error. During crystallisation screening, proteins are tested against large numbers of conditions based on a logical rationale for exploring chemical space, or past successes (Newman *et al.,* 2005; Chayen and Saridakis, 2008; Jancarik and Kim, 1991). Protein crystallisation occurs when the chemical conditions allow for permissive surface charge, protein conformation, protein order, and formation of salt bridges. These, and additional parameters, must coalesce into a combination that makes it thermodynamically favourable for the protein to form an ordered lattice as it is coaxed out of solution. It is possible that for some proteins there are fundamental incompatibilities that make crystallisation impossible using current techniques, whereas others may be crystallised in different crystal forms, representing multiple acceptable combinations of these parameters. Predicting the precise conditions required to crystallise a protein would, most likely, require complete structural modelling of the protein from sequence alone. This, if it were possible, would often remove the need for crystallising the protein at all. Despite the inherent complexity of the problem, several key principles are vital when attempting to create protein crystals. Purity and homogeneity of the protein sample is perhaps the most important. Nucleation of protein crystals is believed to begin with ordered aggregates, which will quickly be poisoned by interactions with proteins co-purified from the

expression cells, or different species of the same protein generated by degradation or other chemical processes (Chayen, 2005). Small globular proteins with a minimal proportion of disordered regions, such as insulin and lysozyme make the best candidates for crystallisation (Fiddis *et al.,* 1978).

During a vapour diffusion of dialysis experiment, the production of crystals is encouraged by gradual equilibration between the level of the precipitant, typically an ionic salt or polyethylene glycol, in the protein/screen solution to that of the condition being tested. In vapour diffusion, which is the basic process underlying the hanging and sitting drop trials, the concentration of protein and precipitant rises as water is lost from the drop to the larger reservoir. When these parameters reach a level that exceeds the solubility limit of the protein, nucleation may occur. If the conditions are not permissible, then protein will begin to precipitate. As protein joins newly nucleated crystals, the concentration will gradually fall to a level where it is said to be metastable. This means that the protein is soluble and, although nucleation is not possible, it is thermodynamically favourable for molecules to join an existing crystal. Further equilibration encourages more protein to join growing crystals until the precipitant levels in the drop are equal to the screen in the reservoir. In ideal circumstances, the nucleation zone is only reached briefly, allowing the formation of a small number of nuclei that quickly soak up additional protein. If most of the equilibration process occurs in the metastable region, then the crystals will be much larger. In practice, once a “hit” condition has been identified, it is advisable to attempt to uncouple nucleation from growth via seeding. This involves the removal of small crystals, optional shattering to multiple nuclei (micro seeding), and introduction of additional protein and screen (Walter *et al.,* 2008). This may allow for better growth because entry into the nucleation zone can be avoided, preventing loss of protein into large numbers of small crystals. Seeding can also be employed to identify conditions that do not allow for nucleation, but may permit better growth than the original condition (Chayen, 2005).

1.8 Phasing

The second challenge, after obtaining diffracting crystals, involves identifying a route to obtaining phase estimates. In the Bragg model of diffraction, spots produced on a detector when a crystal is rotated through an X-ray beam may be treated as reflections from a series of parallel planes (Bragg and Bragg, 1913). Reflections are given miller indices (h,k,l) according to the number of times their associated Bragg planes intersect the faces of the unit cell (a,b,c). The Fourier sum that allows for the calculation of electron density at every point (x,y,z) within the cell, combines information from every reflection. The contribution of each reflection towards the calculated density at a given point is determined by its amplitude (F) and its phase (ϕ).

$$\rho(x,y,z)=1/V \sum_{h}\sum_{k}\sum_{l} \left| F_{hkl} \right| e^{2\pi i \phi_{hkl}} \quad e^{-2\pi i(hx+ky+lz)}$$

Amplitudes are the square roots of the reflection intensities that are measured in the X-ray diffraction experiment and represent the number of electrons in the corresponding plane (Taylor, 2010), but the phases are lost. In real-terms the phase-angle can be imagined as the position of the peak of the wave at the origin of diffraction of each reflection. If the peak is at the origin of diffraction, then the phase angle is 0, if the trough is at this point then the angle is π (180°). If a synchronised (in phase) beam of X-rays could be diffracted onto a phase-sensitive detector, then the phase could be measured, but currently neither of these conditions is fulfilled. Phase angles have no relationship to the amplitudes, except that they are both determined by the contents of the unit cell so some knowledge of these contents must be obtained before phases may be estimated (Taylor, 2010). When the protein target is homologous to a protein with known structure, a technique called molecular replacement can be employed. This approach involves *in silico* transfer of the model of the homologous molecule into the space group of the target molecule, accounting for symmetry operations, and calculating phases which, depending on the level of structural similarity between the 2 molecules, allow for adequate density modelling, refinement, and model building (Chayen and Saridakis, 2008). The advantage of this technique is that the structure of a protein may be solved using only a single native dataset.

1.8.1 Isomorphous Replacement

In projects such as this, where no homology model is available, experimental methods must be employed to gain experimental estimates for the phases. There are a wide range of approaches that have been developed for this purpose, but the most commonly used approaches are single/multiple isomorphous replacement (SIR/MIR) and single/multi wavelength anomalous diffraction (SAD/MAD). Generally both techniques rely on the addition of a heavy element to the crystal to create a derivative. Crystals may be soaked in a solution containing metal ions or halogens or co-crystallised with these species. The aim in isomorphous replacement is to produce a crystal that is identical (isomorphous) to the native, but with the addition of heavy elements in high

occupancy ordered sites (Green *et al.,* 1954; Perutz, 1956). Additional X-ray diffraction experiments are performed using these crystals and data is merged with that of the native. If the merging suggests a suitably high degree of isomorphism, then SIR/MIR phasing can be attempted. In essence the isomorphous phasing methods involve subtracting the native amplitudes from those of the equivalent reflections from the derivative to produce data that consists only of the extra scattering caused by the heavy atoms. Because the substructure typically contains only a small number of atoms, it can be determined using direct methods or Patterson analysis (Kartha and Ramachandran, 1955). Calculation of the Patterson function does not require knowledge of the phases. The peaks of a difference Patterson map calculated in this way represent interatomic vectors between heavy atoms in the substructure. The positions of the atoms in the cell can then be computed using *ab initio* methods aided by Harker analysis which can be used to resolve the positions of atoms related by crystallographic symmetry. Direct methods for solving substructures involve probabilistic analysis of measured intensities to compute phase solutions based on statistical relationships in a way that is not possible for highly complex macromolecules at lower than atomic resolution (Uson and Sheldrick, 1999).

Once the heavy atom substructure has been solved, it possible to resolve the phases of the native reflections. An easy way to visualise this process is to imagine the amplitudes of equivalent reflections from the native and derivative crystals and heavy atom substructure as distance vectors in a 2 dimensional (Argand) diagram with lengths equivalent to their measured amplitudes (F). The derivative and native vectors (Fph and Fp) begin at the origin between the real (x) and imaginary (y) axes and are linked by the vector representing the known amplitude and phase for the heavy atom (Fh α_H). This creates 2 possible triangles of sides Fph, Fp and Fh. The internal angle between Fp and Fh can be solved using the cosine rule, subtracted from 180° and added or subtracted from the known heavy element phase to give an estimate of the protein phase (Fig. 2). Special treatment of the cosine rule to incorporate this, leads to a formula for SIR phasing of:

$$\alpha_p = \alpha_h \pm \cos^{-1}[(F^2_{ph} - F^2_p - F^2_h)/2F_pF_h]$$ (Taylor, 2010; Blundell, 1976)

This method produces 2 estimates for the phase, displaced either side of the heavy atom phase by the same angle. A second derivative can be used to resolve this ambiguity, as only one solution will be common to both. This was the first method used to solve the phase problem in protein crystallography (Green *et al.,* 1954; Harker, 1956).

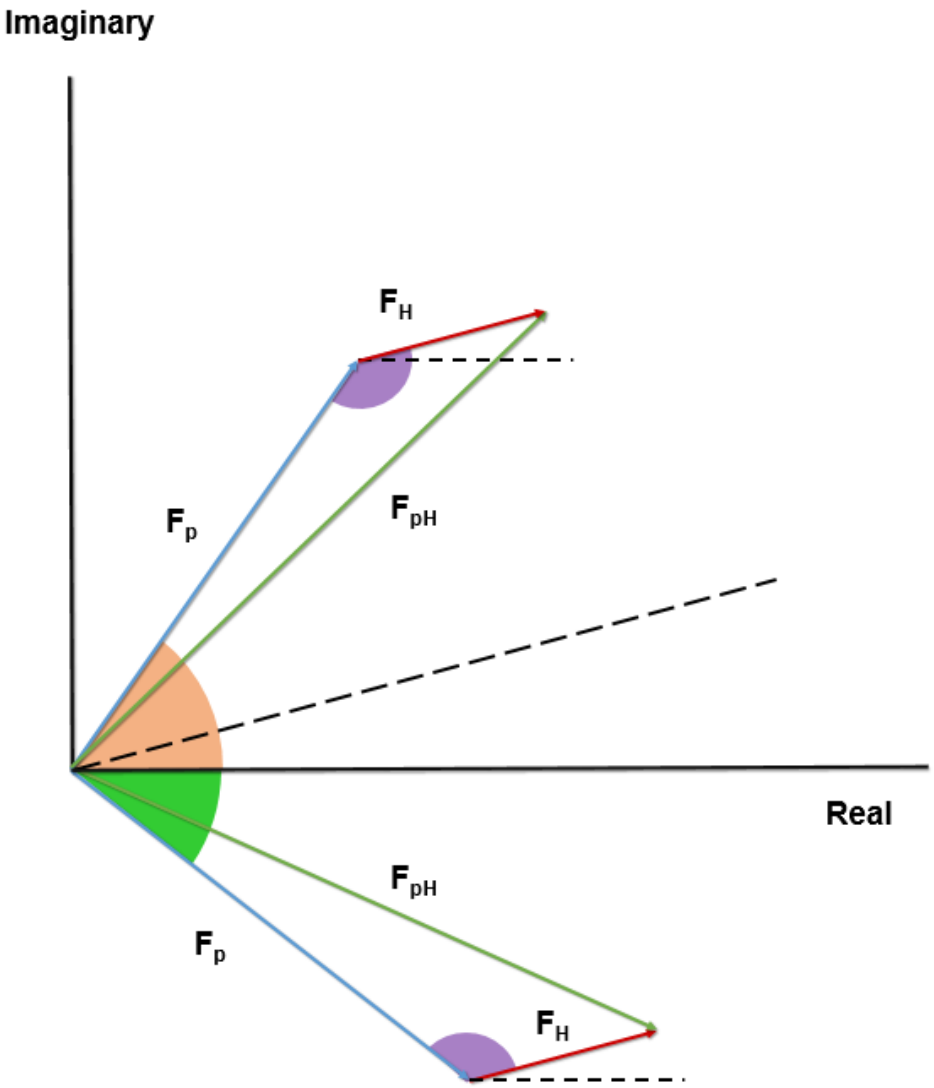


Figure 2*:* Argand diagram for SIR phasing. *Using the amplitudes of the native (F_p), Derivative (F_{pH}) and heavy atom (F_H) and the angle F_H from the real (x) axis, two triangles may be created. The angle in purple can be calculated using the cosine rule. Subtracting this angle from 180 yields the displacement angle (relative to α_H). The two phase estimates produced from SIR (orange and green angles) are displaced symmetrically around F_H shown by the central dotted line (reproduced from Taylor, 2010).*

1.8.2 Anomalous Diffraction

Often, derivatives will not exhibit a suitable level of isomorphism to allow for isomorphous replacement so other methods must be employed. When an X-ray beam is diffracted by a crystal, reflections are produced in Bijvoet pairs which are related through symmetrical inversion through the origin (Friedel pairs), and additional symmetry factors. Under normal conditions, Friedels law is obeyed and these pairs have equal intensities. However, at certain wavelengths which correspond with energies required to promote electrons into higher orbitals, heavy elements within a crystal can absorb the energy of the X-ray beam to produce anomalous differences between pairs. The heavy atom substructure can be solved by direct methods or by analysis of a Patterson map calculated using squared Bijvoet differences which, like the SIR difference Patterson, features peaks corresponding to interatomic vectors. The amplitude of a reflection predicted from the substructure is the sum of normal and anomalous scattering. The anomalous scattering can be described by a real component and an imaginary component with amplitudes that depend only on the specific element and the wavelength used for diffraction. In a vector diagram that describes a reflection predicted by the heavy atom substructure as the sum of its real and imaginary scattering components, the real component (f') is 180° out of phase with the overall substructure-derived vector and therefore is subtracted from its amplitude. The imaginary component (f'') can be assumed to be offset by 90° from that of the overall substructure vector. To visualise the process of SAD phasing, it is useful to imagine another 2 dimensional diagram with one vector, extending from the origin, with amplitude that represents the mean amplitude of 2 Bijvoet pairs. The end of this vector reaches into an area between the circumferences of 2 circles with radii equal to the amplitudes of individual pairs. The substructure (heavy atom) vector (minus the real component of anomalous scattering), which has a known phase and amplitude, must be placed so that the vectors that extend at 90° from the end, representing the imaginary components of anomalous scattering, are able to make contact with the circles (Fig. 3). Anomalous diffraction-based experiments will often use a wavelength that maximises the anomalous scattering values to allow for the best phase estimates.

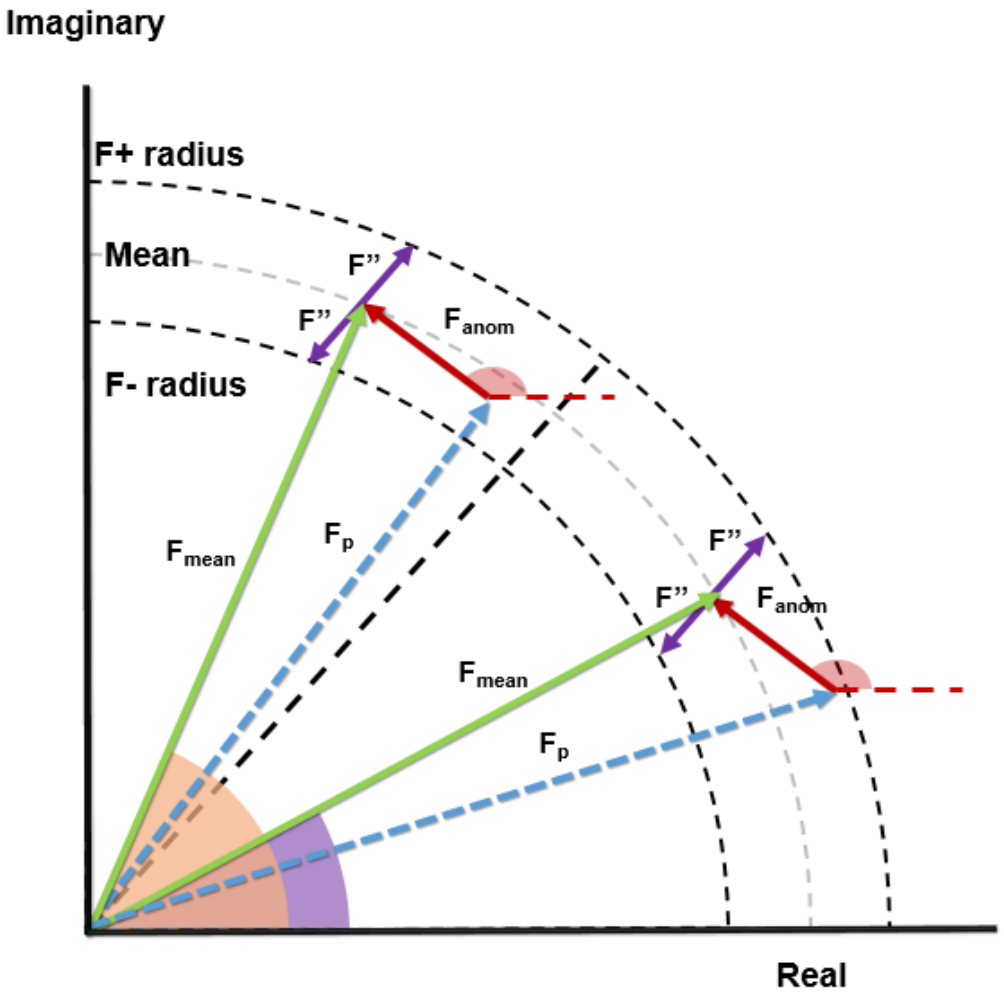


Figure 3*:* Phasing diagram for SAD phasing. *Diagram illustrating the method for phase estimation by SAD. Green arrows represent the amplitude of the derivative reflection (F_{mean}), Red arrows and angles represent the anomalous substructure amplitude and (known) phase minus the real component of anomalous scattering (F_{anom}), and purple arrows represent the imaginary component (f'') of anomalous scattering at the chosen wavelength. Possible solutions for the phase angle of the derivative reflection are shown in orange and purple. The theoretical amplitude and phase of the protein without the anomalous scatterers is shown as a dotted blue arrow. The black dotted line extending from the origin has an angle displaced from the anomalous substructure phase by -90°, phase estimates are distributed symmetrically around this line. Adapted from Dauter (2013).*

1.8.3 Data Quality

Bijvoet differences are typically in the region of a few percent and the margin for error is large. For accurate solutions using this method, data sets with a high completeness and multiplicity (redundancy) are required. Completeness is a measure of the proportion of possible reflections that have been observed and multiplicity is the number of independent measurements of each unique reflection. These statistics can be computed for all the data and for the reflections that are considered anomalous (Evans and Murshudov. 2013; Evans *et al.,* 2011). In order to assess quality it is also critical to study the R_{merge}, R_{symm} and $R_{p.i.m}$ values of the data. The R_{merge} value is the average intensity difference between multiple measurements of the same reflection, which would ideally be 0 in the absence of measurement errors and radiation damage to crystals. R_{symm} gives the average intensity difference across all symmetrically related reflections and would ideally be close to 0, with a small deviation caused by anomalous scattering. The formula for these statistics are the same, except j represents either individual measurements of a single reflection (R_{merge}) or individual symmetry related reflections (R_{symm}) (Weiss 2001).

$$R = \frac{\sum_{hkl} \sum_{j} \left| I_{hkl,j} - \overline{I}_{hkl,j} \right|}{\sum_{hkl} \sum_{j} I_{hkl,j}}$$

The precision indicating merging factor ($R_{p.i.m}$) contains an additional term that adjusts the mean to reflect the increase in accuracy that is achieved with higher redundancy. This merging statistic therefore represents the precision of the mean (Weiss 2001).

$$R_{p.i.m} = \frac{\sum_{hkl} \sqrt{\frac{1}{N-1}} \sum_{j} \left| I_{hkl,j} - \overline{I}_{hkl,j} \right|}{\sum_{hkl} \sum_{j} I_{hkl,j}}$$

In real experiments, R_{merge} and R_{symm} values of 10% or less are desirable. Large R factors at low resolution indicate poor quality, but will increase at higher resolutions even when good quality data is available. With high redundancy, random measurement errors are expected to destructively interfere allowing anomalous differences to be recognised at higher resolution. This is the basis of many SAD experiments including S (sulphur) SAD, where very small anomalous differences caused by native sulphur atoms are used for phasing. Combining data from multiple crystals can be used to negate the problems with radiation damage during collection of large datasets, but this relies on minimal crystal-to crystal variability for best results (Foadi *et al.,* 2013).

The advent of phasing using anomalous differences was important, as it allowed for initial density calculation using just one dataset. Like SIR, SAD produces 2 estimates of the phase, this time, symmetrically distributed around the anomalous phase -90°. Ramachandran and Raman (1956) suggested that choosing the phase closest to that predicted from the substructure would result in the correct estimate a suitably high proportion of the time to allow computation of a meaningful solution for refinement. This is possible because the correct phases correlate with the true electron density, whereas the incorrect solutions are random. It is also possible to use averages of the 2 phases to generate a crude electron density map. Even with the substantial noise generated by the effect of averaging with randomly distributed phases it can be possible to identify the solvent boundary, refine, and use the inverse Fourier to produce more accurate phases (Dauter, 2013). Electron density maps produced using SAD come in pairs, representing the inversion of the anomalous substructure. Identifying the correct hand can be done computationally or manually, as only one should contain recognisable macromolecule-like density regions. In the modern era of crystallography, researchers have access to a wide range of phasing techniques that build on, or combine the 2 aforementioned methods. Single isomorphous replacement with anomalous scattering (SIRAS) techniques use anomalous diffraction to resolve the phase ambiguity that is normally present in SIR experiments, and Multi-wavelength anomalous diffraction (MAD) techniques use multiple wavelengths to increase the accuracy of anomalous-based phase estimates and, by using one wavelength that minimises the anomalous signal, can provide a “native-like” dataset. This means that MAD data can often be treated like a special MIR case (Ramakrishnan and Biou, 1996). More exotic techniques employ the disproportionate accumulation of radiation damage at heavy atom positions to phase datasets (Ramagopal *et al.,* 2005).

1.8.4 Density modification and modelling

A map computed with initial experimental phases may be difficult to interpret even with high quality data. Using computational analysis, phase estimates can be improved by applying generalised knowledge about the cell contents of macromolecular crystals or correlations in reciprocal space (direct methods). Protein crystals have a much higher solvent content than those formed from smaller molecules, with large channels and voids filled with disordered molecules. Sophisticated methods have been developed for identifying the boundary between the solvent and ordered regions. Once a region has been identified as containing solvent it can serve as a reference point because a set of phases that correlates with homogenous density in this region is more likely to be correct than one that does not (Terwilliger, 1999). Another powerful technique involves establishing the translational operations that link multiple molecules in the asymmetric unit and averaging the density of the equivalent regions. This technique becomes more effective when the number of copies in the asymmetric unit is large (Terwilliger, 2004). Because heavy atoms often bind to the same sites in multiple copies of proteins and other macromolecules these symmetry operations can be identified by studying the heavy atom substructure (Lu, 1999). The input of an experienced crystallographer is invaluable in difficult cases where automated methods cannot refine electron density to a level where automatic model building is possible. These cases demonstrate both the

power of computational methods but also their current limitations. A crude map can suggest features such as secondary structural motifs that can allow basic models to be constructed manually in a program such as COOT (Emsley *et al.,* 2010). During this process, the researcher employs statistical tools to guide the building process to ensure the best possible model, and increasing the chances of further refinement. This is done by studying parameters such as the phi and psi angles of the chain and attempting to find an optimal fit to the density map (Murshudov *et al.,* 1997). These models can be imputted into the refinement process allowing the statistical processes to benefit from types of knowledge that have yet to be replicated in mathematical processes.

1.9 Project Aims

In this project the key aim was to gain structural information about ComZ to act as a starting point for investigating its function. In order to accomplish this, ComZ was expressed in recombinant form and purified by metal affinity, ion exchange, and size exclusion chromatography. The structure of ComZ was studied using X-ray crystallography. In addition to this, the domain structure was studied through limited proteolysis, and divalent metal binding was explored using a thermal shift assay.

2 Materials and Methods

2.1 Chemical suppliers

All general laboratory chemicals and reagents including buffers not otherwise mentioned were supplied by Sigma Aldrich chemicals. Selenomethionine was purchased from ACROS organics through Thermo Scientific. Bacterial culture mediums were purchased from ForMedium with the exception of selenomethionine labelling media which is provided by Molecular Dimensions. Plasmid vectors and the methionine-auxotrophic expression strain were provided by Novagen and general expression strains were supplied by New England Biolabs. Crystallisation screens and plates were obtained from Molecular Dimensions. Antibodies for were purchased from Qiagen. Ion exchange and Size exclusion Chromatography (IEC/SEC) was carried out using columns and AKTA purifiers manufactured by GE healthcare systems.

2.2 Bioinformatics

Prediction of transmembrane helices was performed using the THMM server (Krogh *et al., 2001; Sonnhammer et al.,* 1998). Secondary structure prediction was carried out using the PSIPRED method (Jones, 1999; Buchan et al., 2013). Residues encoded by plasmid-derived sequences (presumably unstructured) were excluded from percentage calculations.

2.3 Recombinant expression of ComZ

2.3.1 Strains and plasmids

General expression was carried out in the T7 EX *E. coli* expression strain (New England Biolabs). Expression of selenomethionine labelled protein was performed using B834 (DE3) (Novagen). The construct used was based on a pET22b backbone (Novagen) and encoded 523 residues of Native ComZ and 30 vector derived residues under the control of a T7 promoter. This region was amplified from *T. thermophilus* genomic DNA using primers of sequence 5’-CTTCACCATGGCCATAGAGCTCTGGACCACCCGCAACGAC-‘3 (forward) and 5’-CGGTGTGACTCGAGGCGGCGCTCATAGGAGAGCACCTG-‘3 (reverse). (Karuppiah *et al,* unpublished work). The pET22b plasmid encodes an N-terminal Pelb leader sequence to ensure targeting to the periplasm for disulphide bond formation and a C-terminal polyhistidine (6X) tag for purification purposes.

2.3.2 Transformation

All transformations were carried out by incubating 150ng of plasmid DNA with 50 µl aliquots of competent cells. DNA internalisation was induced by heat shock at 42 °C for 1 minute. Recovery was allowed in 300 µl of Super Optimal Broth (Sigma) for 1 hr at 37 °C and transformants were selected using LB agar containing 100 µg/mL ampicillin.

2.3.3 Culture conditions

For the production of unlabelled protein, the construct was transformed into expression strains using the heat-shock transformation method, and transformants were selected by plating onto LB agar containing 100 µg/ml ampicillin followed by incubation overnight at 37 °C. A number of these colonies were then inoculated into 50 ml LB starter cultures with equal concentrations of antibiotic and incubated at 37 °C with shaking at 200 rpm until optical density at 600 nm (OD600) of 0.6 was reached. At this point, 10 ml of these cultures (approximately 4.8 x 109 cells) was transferred into each of a number larger shake flasks containing 500 ml of terrific broth supplemented with 2 ml of glycerol (Sigma), and ampicillin levels equal to the previous mediums. Growth was carried out in identical conditions to the starter cultures, and was monitored until OD600 reached 1. At this point, the cultures were cooled and protein expression was induced by the addition of Isopropyl β-D-1-thiogalactopyranoside (IPTG) (Sigma) to a concentration of 0.4 mM. Overnight fermentation was carried out at 16 °C with shaking at 200 rpm. Cells were harvested from culture by centrifugation at 11,000Xg for 20 minutes. On occasions when it was not possible to continue to the extraction stage immediately, cell pellets were flash-frozen by immersion in liquid nitrogen and stored at -80 °C.

2.3.4 Selenomethionine labelling culture

The protocol for production of Selenomethionine-labelled protein was largely similar. An identical transformation was carried out using Competent B834 cells and starter culture growth was also carried out. After growth in methionine-containing LB broth, cells were washed by a process by centrifugation at 10,000 xg, and suspension in sterile water. After a second round of centrifugation, the cells were resuspended in Selenomethionine labelled media equivalent to the original volume of LB. The media used for the main fermentation was a commercially available (Molecular Dimensions) product described as being based on M9 salts supplemented with glucose, amino acids (not including methionine), and vitamins. Seleno-L-methionine (ACROS organics) was added to a final concentration of 100 mg/L. To account for the slower growth of bacteria in labelled media, incubation at 37°C was allowed to continue to an OD600 of 1.2 prior to induction, and overnight fermentation was carried out at 27 °C. Cells were harvested using the same method as those grown in unlabelled media. A single litre of labelled culture would typically yield 3-4 grams of cells.

2.4 Extraction and purification of ComZ

2.4.1 Cell lysis

Lysis was carried out using approximately 45 ml of 25 mM Tris HCl at pH 8.0 with 100 mM NaCl. Cell pellets were resuspended in lysis buffer with the addition of a single protease inhibitor tablet, and 50 µl of DNase I (Roche). The cell suspension was sonicated on ice using a high frequency power output of 70 W and 5 second pulses at10 second intervals. Sonication lasted 7 minutes for each 50 ml aliquot of suspension. In order to separate the soluble and insoluble fractions of the lysis extract, centrifugation was carried out at 31,000 xg for 30 mins. The soluble fraction was

passed through a 0.22 µm filter to remove any cell debris that may have become resuspended post-centrifugation.

2.4.2 Metal affinity chromatography

Ni NTA agarose resin (Qiagen) was washed by a process of centrifugation at 1500 xg, removal of supernatant, and resuspension in sterile water in order to remove the 20% ethanol storage solution. This cycle was performed 3 times and after the supernatant had been discarded for the third time, the pellet was resuspended in lysis buffer. Resin was added to filtered cell extract at a ratio of 3 ml of 50% v/v resin per 45 ml of cell extract (corresponding to approximately 15 g of cells). This mixture was allowed to rotate slowly at 4°C for 1 hr, allowing the metal Ions to bind the Polyhistidine tag on ComZ. Elution of ComZ was performed using nickel affinity gravity flow chromatography. Washes were performed using 6 mls of lysis buffer containing 10, 20, 30, 40, 50, and 60 mM imidazole and elution was carried out using the same volume but consisted of 3 fractions containing 200 mM imidazole and 2 containing 500 mM.

2.4.3 Ion exchange chromatography

Fractions carried forward from metal affinity chromatography were dialysed overnight into 25 mM Tris HCl pH 9.0 in dialysis tubing with a 30 kDa cut-off. After dialysis, protein was concentrated to a volume of 10 ml using a centrifuge ultrafiltration column with a 30 kDa molecular weight cut-off. Ion exchange chromatography was carried out using an AKTA prime or HPLC (GE healthcare) system at a flow rate of 1 ml/min. low and high salt buffers used 25 mM Tris HCl pH 9.0 with 0 or 1 M NaCl respectively. Typically, fractions of 0.5-1 ml would be collected over a 2% gradient. Absorbance at 280 nm as well as % of high salt buffer was recorded with respect to elution volume.

2.4.4 Size exclusion chromatography

Fractions produced from Ion exchange were concentrated (or diluted) to 5 ml. size exclusion chromatography was carried out using a column packed with Superdex 75 matrix (GE healthcare) the total column volume was 120 mL and the void volume was 45 mL. The size exclusion (final) buffer was 25 mM Tris HCl pH 8.0 100 mM NaCl. Fractions of 1 ml were collected over a total elution of 130 mL at a flow rate of 1 mL/min. absorbance at 280 nm was recorded as for Ion exchange. When purifying selenomethionine labelled protein, the size exclusion buffer was de-gassed using a vacuum pump to limit oxidation.

2.5 SDS PAGE

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS PAGE) was performed by the addition of SDS loading dye containing beta-mercaptoethanol and denaturation at 90°C for 6 minutes. Separation was carried out using precast gels (Thermo Fisher) and MES SDS running buffer, and were run for 1 hr at 180 V. Gels were either stained using Coomassie blue stain (Thermo Fisher) or used for western blotting. All SDS PAGE gels are presented with labelled molecular weight standards in the far left hand lane.

2.6 Western blotting

Western blotting was carried out using transfer from a SDS PAGE gel to nitrocellulose (Ge healthcare) in 25 mM Tris, 190 mM glycine and 20% methanol. Blocking was performed using 5% bovine serum albumin (Sigma) dissolved in phosphate buffered saline with 0.2% tween (PBST). Membranes were then washed 3 times using PBST with 10 minutes allowed for each wash. The primary antibody was diluted by a factor of 2000 in PBST and added to the membrane overnight at 4 °C. Membranes were then washed, as in the earlier step, and secondary antibody, diluted by a factor of 5000 in PBST was added for 1 hr at room temperature. A final 3 washes were performed to remove unbound antibody, and the blots were developed. The western blots in this thesis were performed with a mouse tetra anti-his primary (Qiagen) and a rabbit anti-mouse IgG secondary antibody labelled with alkaline phosphatase (Sigma). Blots were developed using an exposure time of 3 minutes using alkaline phosphatase substrate (Sigma).

2.7 Determination of protein concentration

Final concentration was confirmed using a NanoDrop spectrophotometer to measure the ratio of absorbance at 260 and 280 nM. These estimates were made more accurate by adjusting using mass and extinction coefficient calculated using protParam (Gasteiger, *et al.,* 2005).

2.8 Thermofluor analysis

To test for changes in melting temperature, ComZ was combined with SYPRO orange dye (Life technologies). This dye increases in fluorescence when it interacts with hydrophobic residues as a protein begins to unfold (Steinberg *et al.,* 1996; Lavinder *et al.,* 2009) and test ligands/buffers to produce a final concentration of 50 µg/ml protein and a dye concentration of 5X the manufacturers stated concentration factor. Samples were then subjected to a melt-curve protocol using a real-time PCR machine using a temperature range of 25-92 °C. Fluorescence was induced by excitation at 300 nm and detection at the emission wavelength of 570nm. Melting temperature was inferred from the lowest value of the first derivative of temperature (°C) vs emission intensity (arbitrary units).

2.9 Enzymatic digestion

Enzymatic digestion of ComZ was carried out in 25 mM Tris HCl pH 8.0 with 25 mM NaCl. Enzymes were supplied by Sigma and were prepared at 1 mg/ml in the same buffer prior to addition to final reactions. Reactions would typically contain 500 µg/ml ComZ and were terminated by addition of SDS PAGE loading dye and heating to 90 °C for 6 minutres prior to SDS PAGE.

2.10 Crystallisation Screening

Crystallisation was carried out using the sitting drop method. Automated pipetting was carried out using a mosquito pipetting robot (TTP labtech). The commercially available screens PACT, JSGQ, Morpheus and SG1 were used for screening for new conditions (Fazio, *et al.,* 2014; Gorrec, 2009; Newman *et al.,* 2005).

2.11 Processing of Crystallographic Data

Processing of data was performed using the CCP4 suite (Winn *et al.,* 2011). Autoindexing and integration of images was carried out using XDS automatic data processing (Kabsch, *et al.,* 2010). Space group (though already known) was predicted using pointless (Evans, 2006/2011). Scaling and merging was performed using aimless (Evans and Murshudov, 2013; Evans *et al.,* 2011). Substructure determination was carried out using the hybrid approach (HYSS) combining Patterson, direct, and phaser completion methods (Grosse-Kunstleve and Adams, 2003). Experimental phases were calculated using Phaser (McCoy *et al.,* 2007) and initial density modification was performed using RESOLVE (Terwilliger, 2000). Model building was carried out using winCOOT (Emsley *et al.,* 2010) Phase transfer from SAD derived models to native data was carried out using MOLREP (Vagin and Teplyakov, 1997) and graphical images of electron density and structural models were presented using CCP4mg (McNicholas *et al*., 2011).

2.12 Outside services

Sequencing data was provided by GATC Biotech using the Sanger method. Construct sequences were determined by combining two Sanger sequences generated using primers complementary to the T7 promoter and terminator regions of the pET22b expression vector (GATC Biotech). Mass spectrometry was performed by Dr Emma Keevil using a 6200 series TOF spectrometer and analysis software from Agilent Technologies. Size exclusion chromatography with dynamic laser light scattering (SEC MALLS) was carried out by Diana Ruiz Nivea using a DAWN HELEOS detector from Wyatt Technologies. Both mass spectrometry and SEC MALLS were performed in the Biomolecular Analysis Core Facility at the University of Manchester. Crystal derivatisation, freezing and data collection were performed by Dr Colin Levy at the Manchester Protein Structure Facility (MPSF). Crystallographic data was collected at the Diamond Light Source synchrotron facility in Oxfordshire.

3 Results

3.1 Bioinformatics and Construct design

The sequence encoding ComZ had been previously amplified from *T. thermophilus* HB27 genomic DNA and ligated into the pET22b expression vector by Karuppiah *et al* (unpublished work). During this process, a region of the coding sequence was omitted, as it was predicted to encode a transmembrane helix (Freidrich *et al*., 2003). Computational analysis of the full ComZ coding sequence was performed to demonstrate the basis for this decision (Krogh *et al., 2001; Sonnhammer et al.,* 1998). This server gives a probability score by assessing the sequence for the ability to form known features of transmembrane helices such as helix caps and loop regions using a hidden Markov model (Sonnhammer *et al.,* 1998). Analysis by this method predicted the presence of a transmembrane helix between residues 5 and 27 with a high degree of certainty (Fig. 4). The remaining sequence is shown as being “outside” (periplasmic) and although THMM is not able to reliably determine protein localisation, this is in keeping with the evidence on the localisation of other components of the competency machinery. Since ComZ is predicted to be anchored to the inner membrane facing the periplasmic space, the amended coding region had been ligated into the vector downstream of its PelB leader sequence using the Msc and Xho1 sites within the multiple cloning region. It was predicted that this would allow any potential disulphide linkages to form, and thus ensure proper folding of the protein. Prior to expression of ComZ the plasmid was digested using Msc1 and Xho1 to confirm the presence of the coding sequence in the expression region of the plasmid. Sequence identity of the construct was confirmed using multiple Sanger sequences (see methods). Details of the polypeptide sequence are shown in Fig. 5. The predicted molecular mass, extinction coefficient at 280 nm, and isoelectric point of the product of this construct were predicted using protParam to be 57155.9 Da, 53915 $M^{-1}cm^{-1}$, and pH 7.18 respectively (Gasteiger, *et al.,* 2005). Secondary structure prediction indicated that ComZ featured a high proportion of beta strands with an N terminal alpha helix (Fig. 4).

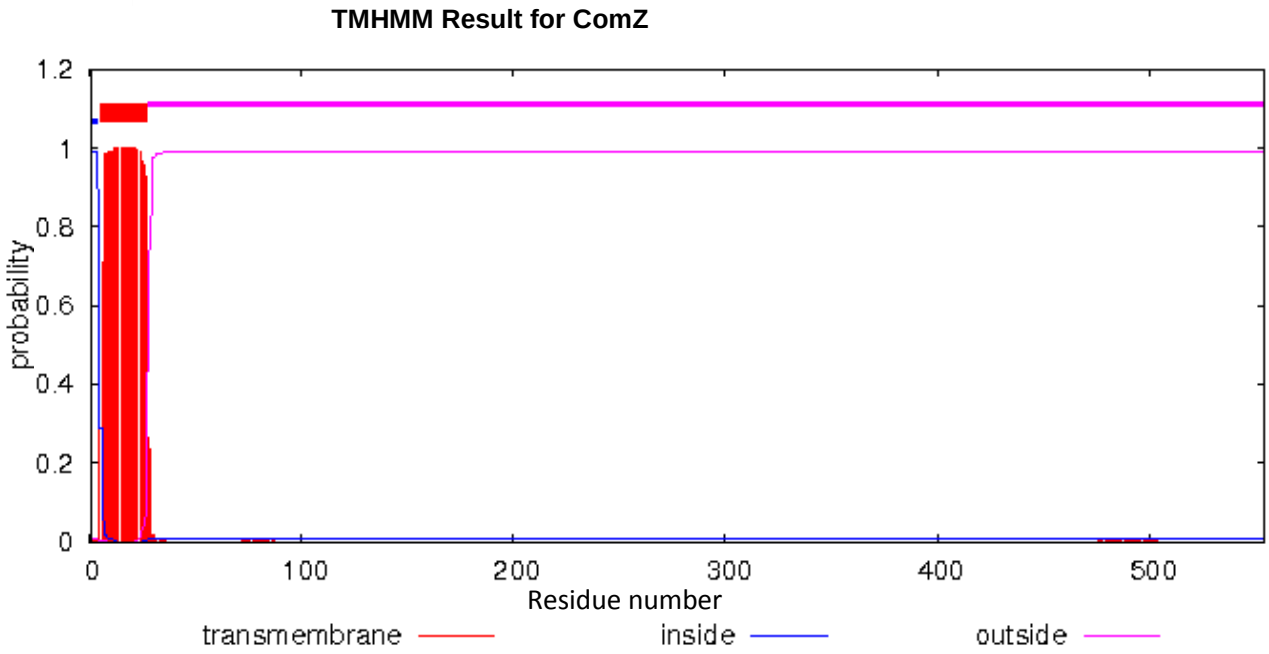


Figure 4: Output from THMM server for prediction of transmembrane helices. *Residue number (X axis) is shown against predicted probability of transmembrane localisation (Y axis) (Krogh et al., 2001; Sonnhammer et al., 1998).*

MNAKGIALVATLALMVVIALLVFGTFFTTQ

0- MKYLLPTAAAGLLLLAAQPAMA/IELWTTRNDTTSVQAFYAAEAGLQKYKA
50- ALFQQYVWREQRGGTGGGGGCFTSLARGLDLDRDGTITPFVNNRLVLAQN
100- EVVTDANGNPVGRYTATLYKDAQDDQLFTLVSEGTSGGAKARVQATFRIS
150- NSDYLEQAIFAGGQANKWLNGGATIRGGVYVVGNPNDPDQYVIEANGNFA
200- LYNRYDLTTYSEVTNRVEPSYRQVQDLCASLRVQYGKISVGGSTQIGEPN
250- NKVKGVFVGRGAQDITGENVGVCRNNKGVCTEAMGGFDLSDPPPFPTLDA
300- KLDSDACSAYPTWRACLQGKAALRIQRIGNILSVASPPNATLSPSCLQAM
350- QSGTLTLDTQSVDCTFTRLDGSRGGFRYTYTGGQELLEVFGDVVLEGIDA
400- VLNRPVDYRAQSGSAKSATLAVLKLGGNGGNLDINGNLLPDATFGLFPNH
450- ALGFVAEGDIYQRGQHVMAPVYAGGTFRVVKGNVLFGSVISNQFCTTSAG
500- NQMSCNASQKAEVVYIRIPKENRPALLPSLRGGKPVFQVLSYERRLEHHHHHH

Predicted transmembrane helix (omitted)

Alpha (10.32% of non-plasmid derived residues)

Beta (34.80% of non-plasmid derived residues)

Pelb Leader sequence (cleavage site denoted by "/")

Polyhistidine tag

Figure 5: Polypeptide sequence of ComZ construct with secondary structure predictions *percentages of predicted alpha (Red) and beta (yellow) secondary structure elements are shown below (Jones, 1999; Buchan et al., 2013).*

3.2 Expression of recombinant ComZ

The expression construct was transformed into expression strains and fermentation was carried out as previously outlined. Growth and expression in native conditions was found to be adequate with each litre of culture yielding around 7g of cells and about 2.5mg of purified ComZ (at the end stage). The selenomethionine labelling protocol yielded approximately 3.5 grams of cells and around 1mg of purified ComZ per litre of culture. Although these expression levels appear low, large proteins are often poorly expressed, especially if they must be targeted to the periplasm.

3.3 Metal Affinity chromatography

Initial purification of ComZ from crude cell lysate was performed using nickel affinity gravity flow chromatography and Fractions were analysed using SDS PAGE (see Fig. 6). From these results it appeared that ComZ was being successfully expressed in a soluble form with the polyhistidine tag. A large number of impurities were visible. These are likely to be *E. coli* proteins, rich in surface-exposed histidine residues. The most abundant impurities are visible at approximately 65 kDa and between 20 and 25 kDa. Protein concentration was measured at this point using a NanoDrop spectrophotometer.

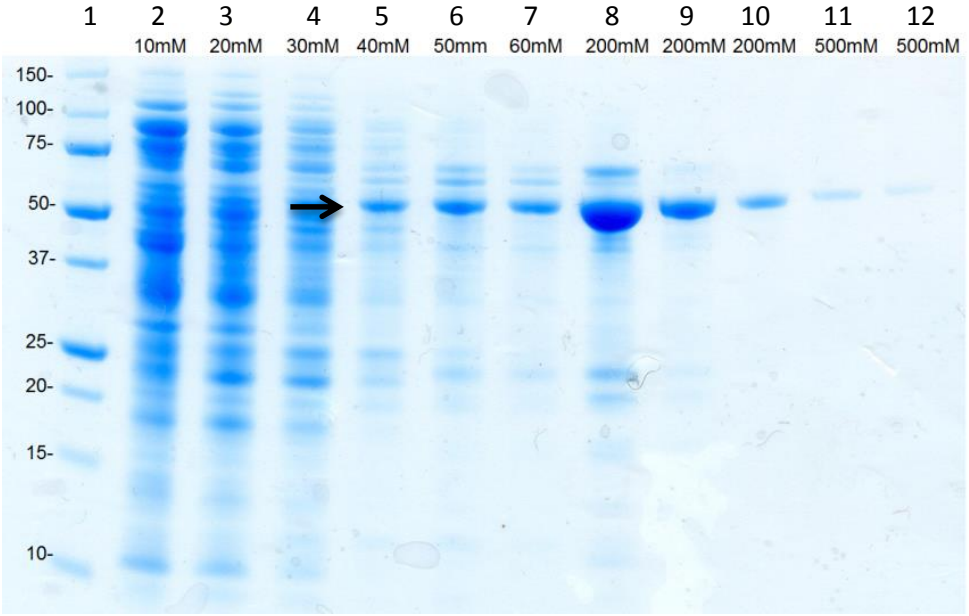
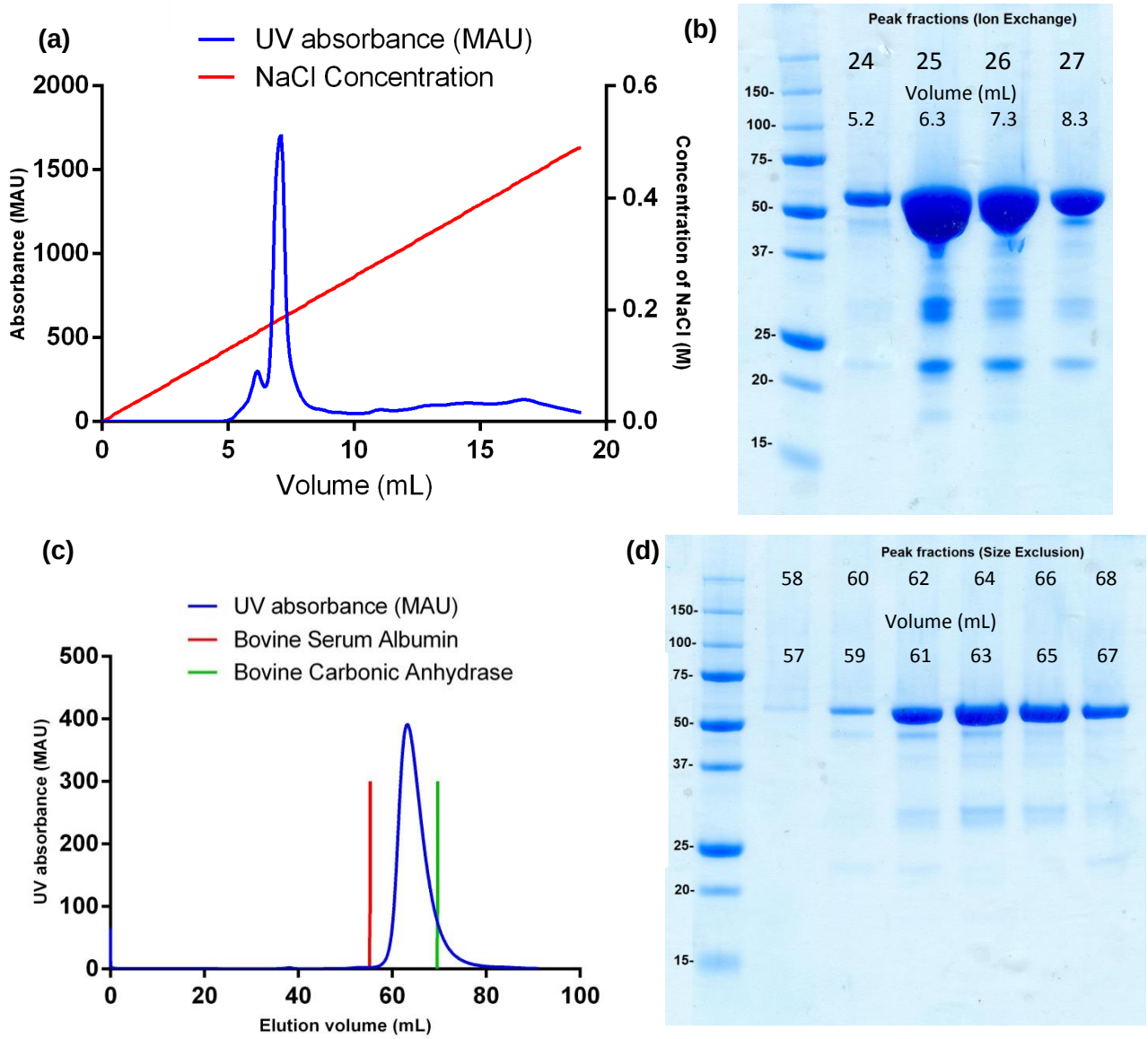


Figure 6: Coomassie stained SDS PAGE gel showing fractions from nickel chelate gravity chromatography. *Marker masses (kDa) are shown at the left of the ladder in lane 1, and imidazole concentrations of the sampled fractions are displayed over lanes 2-12. A band of approximately 57kDa corresponding to ComZ is visible in lanes 5 onwards. ComZ is indicated by an arrow.*

3.4 Ion exchange and Size exclusion Chromatography

Fractions from metal affinity chromatography were selected according to abundance of ComZ relative to other contaminating species visible from SDS page. Only fractions that appeared to contain an equal or greater purity than the first elution fraction (200 mM imidazole) were included. Pooled fractions were dialysed, concentrated and ion exchange chromatography was performed. Peak elution of ComZ at pH 9.0 was observed at 184 mM NaCl. Fractions containing the main-peak (Fig. 7a) were analysed by SDS PAGE (Fig. 7b) before being pooled, concentrated, and subjected to size exclusion chromatography (Fig. 7c). Peak elution occurred at 62.0 mL and was compared to that of 2 molecular weight standards bovine carbonic anhydrase (29kDa) which eluted at 69.70 mL, and bovine serum albumin (66kDa) which eluted at 55.33 mL. These proteins were run under the same buffer conditions as ComZ. Based on these two standards the protein mass was estimated to be around 50 kDa. more accurate measurement was carried out later. Integration of the example size exclusion UV absorbance peak gave a predicted total mass of 13.6 mg. Fractions collected at this stage were, once again, analysed by SDS PAGE (Fig. 7d) and pure samples were combined. The final ‘crystallography grade’ product was again analysed with a final round of SDS PAGE (see Fig. 7e). The final purity was high, but two impurities of approximately 30 kDa were still visible; this could indicate that ComZ was binding *E. coli* proteins, or that it was undergoing proteolytic degradation. Samples were concentrated by centrifuge ultrafiltration as for the earlier steps, up to a concentration of 15 mg/mL. Selenomethionine labelled ComZ behaved identically to native protein during these purification steps.



(e)

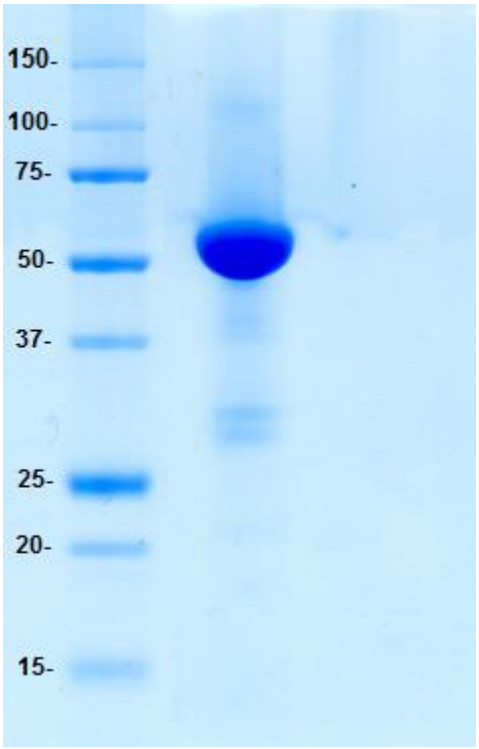


Figure 7: Purification of ComZ

(a) Example Ion exchange chromatograph of ComZ showing absorbance at 280 nm (MAU) and concentration of NaCl (M) against elution volume. *Data shown covers only the elution gradient region. Data points on all graphs are normalised by the subtraction of the lowest absorbance reading from the entire chromatograph to account for mis-calibration of the UV detector.*

(b) Example Coomassie stained SDS PAGE gel showing samples of peak fractions from Ion exchange chromatography

(c) Example *size exclusion chromatograph of ComZ showing absorbance at 280 nm against elution volume*

(d) Examples Coomassie stained SDS PAGE gel showing samples of peak fractions from size exclusion chromatography.

(e) Coomassie stained SDS PAGE gel showing final purity of ComZ. *This sample represents the final crystallography grade sample, but was diluted to 500 µg/ml to produce a clear band.*

3.5 Determination of the mass of recombinant ComZ

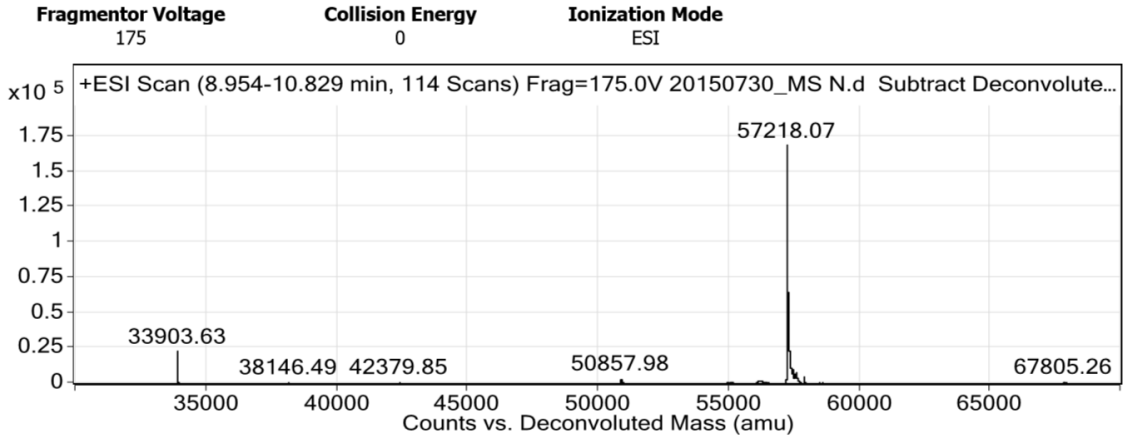
3.5.1 Mass spectrometry

In order to confirm that the construct was being expressed and processed correctly, and to determine if selenomethionine labelling had been successful, the true mass of labelled and unlabelled ComZ was determined using intact mass spectrometry. The mass was found to be 57218.9 Da for native protein and 57409.0 for selenomethionine labelled protein (fully labelled) (see Fig. 8). This difference was equal to that expected for the substitution of sulphur atoms with selenium at the 4 methionine residues (190.1 vs 187.6 da). These 2 values represent the most abundant species, but several additional masses were abundant in both samples. One of these masses represents 3 partially unlabelled species where one methionine is unlabelled (57360.54 Da), but it is unclear what causes the other changes, or if the abundant heavy ComZ species seen in both mass spectrographs are caused by the same events. Testing the combinations of removing various multiples of the S-Se mass difference from the masses of the 2 abundant heavy species in the selenium labelled prep (purple and pink) does not reveal any obvious relationship to the heavy species seen in the unlabelled prep. Subtracting the mass of up to 4 oxygen atoms was also attempted using these combinations, to account for the possibility of increased oxidation at selenium sites. No relationship to native “heavy” ComZ species was noticed but it is possible that the masses shown in purple and pink are related via dual oxidation events, as their mass difference is 32 Da.These heavy species could indicate some form of post translational modification or aberrant processing of the leader sequence could be occurring causing this difference of up to 81 Da (green). Labelling efficiency was calculated using the following formula where N_s is the abundancy of one species, M_s is the mass of a species in the labelled sample, and M_u is the mass of the corresponding species in the unlabelled sample. This relied on recognising equivalent species amongst the list of abundant masses (Fig. 8). This was unambiguous for the masses shown in red, but a theoretical unlabelled mass for the species of unknown identity was unknown.

$$Le = \frac{100 \sum N_s[(m_s - m_u)/48]}{\sum(4N_s)}$$

This produced an estimated labelling efficiency of 75% assuming that all species with a mass higher than 57408.97 Da were fully labelled ($[M_s-M_u/48]=4$). This is not certain, but combining the S-Se mass difference with that of either the pink or purple-labelled species (relative to the fully labelled mass) gives a mass difference that is larger than any seen in the unlabelled protein. There is a substantial “shoulder” visible in the mass spectrograph from the labelled prep above 57408 kDa. If this feature represents a large number of low abundance fully labelled species, then it is possible that the true labelling efficiency is much higher. The calculation takes into account all the listed highly abundant ComZ species, but it is unclear if all species form crystals. If only the unaffected labelled/partially labelled protein forms crystals, then effective labelling was only around 65%.

a Unlabelled

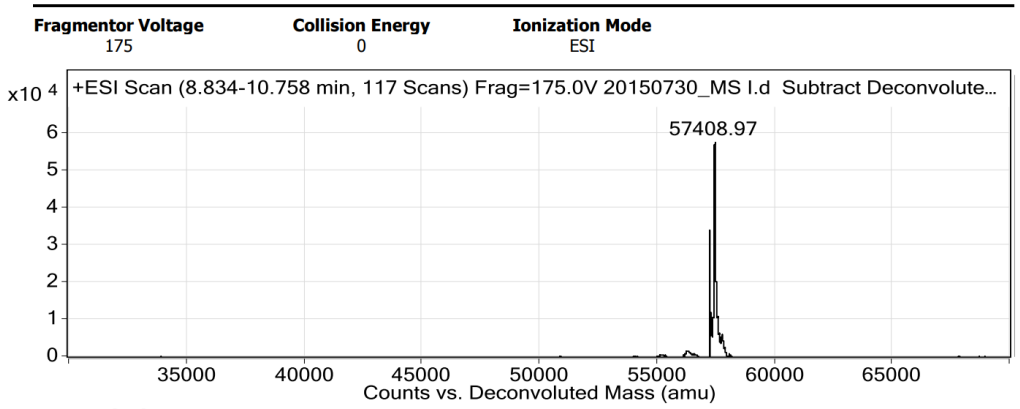


Peak List

m/z	Abund	Score (DB)
33903.63	22320.67	6
57218.07	168230.47	10
57236.24	53586.92	9
57277.39	29898.52	9
57298.64	19074.57	8

ComZ
ComZ + 18da
ComZ + 59da
ComZ + 81da

b Selenomethionine labelled



Peak List

m/z	Abund	Score (DB)	NL
57217.91	33821.77	9	0
57360.54	15861.02	8	3
57408.97	57664.69	9	4
57450.65	26987.89	9	4?
57482.49	15103.8	8	4?

ComZ
ComZ + 42da
ComZ + 74da

Figure 8: Deconvoluted mass spectrographs showing masses of ComZ species.

(a) Unlabelled native protein.

(b) Selenomethionine labelled protein. *Species that are equivalent (aside from labelling) are labelled with the same colour. The number of selenium atoms for each mass is displayed on the far right of the lower table (NL). The mass increase for the heavy ComZ species is relative to the lowest mass of the unlabelled protein (upper) or the fully labelled protein (lower).*

3.5.2 Size exclusion chromatography with multi angle laser light scattering

In order to screen for oligomeric complexes in solution and inform any judgments made about interactions in the crystal form, a sample of ComZ at 600 µg/ml was analysed using size exclusion chromatography with multi angle laser light scattering (SEC MALLS). Analysis was carried out under normal ComZ buffer conditions using a 25 ml column containing Superdex75 matrix (Fig. 9). Light scattering was measured using 18 detectors at angles between 0 and 157.8°. Analysis of the main peak gave a size estimate of 57 kDa (±0.738%) very similar to the result of mass spectrometry. The peak is largely symmetrical suggesting that, under these buffer conditions, ComZ is homogenous and monomeric. One additional minor peak was observed leading to a size estimate of 92 kDa (±10.207%). Because the mass estimate from second peak has a large error value, it is possible that it could represent a dimeric form, but it is perhaps more likely that it was caused by a minor impurity or an aggregate (Fig. 9).

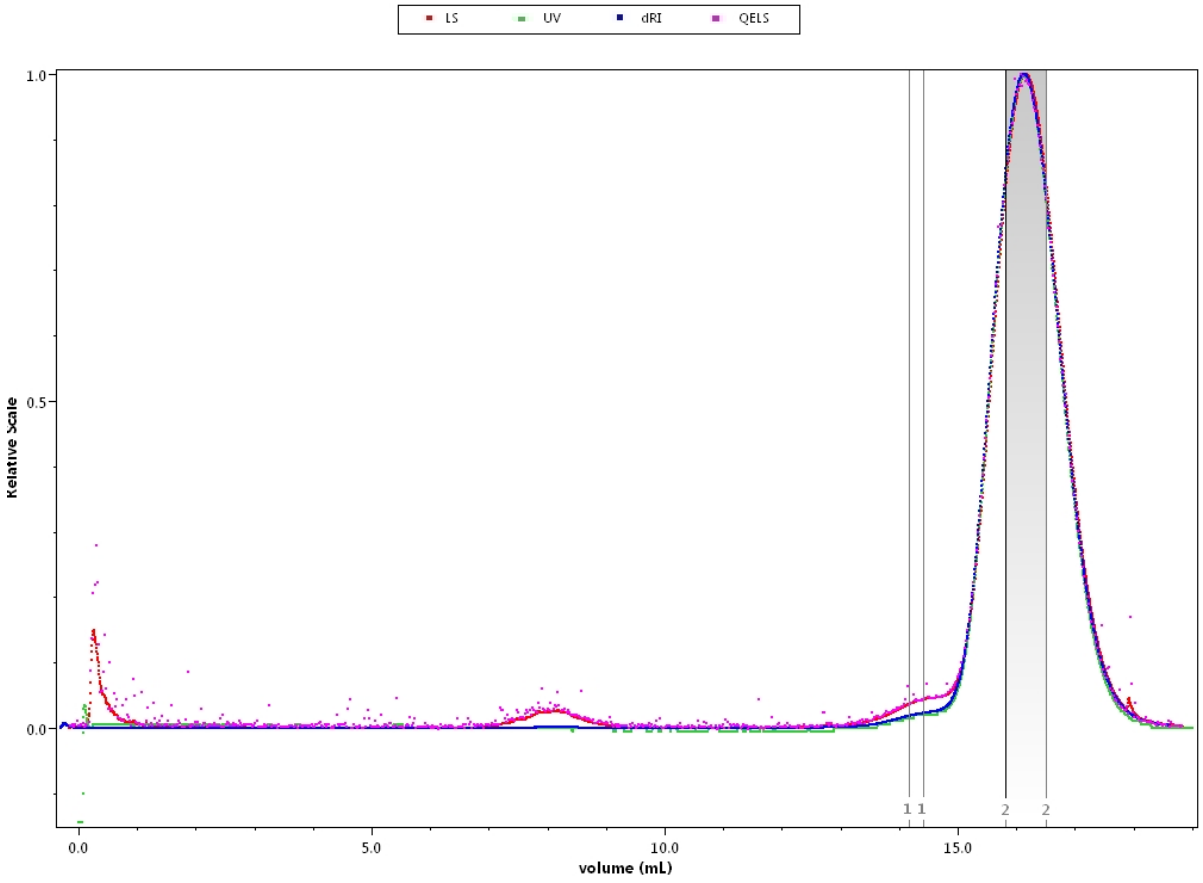


Figure 9: Size exclusion chromatograph with multi angle light scattering (SEC MALLS). *This chromatograph combines the results of light scattering (LS, red), UV absorbance (UV, green), differential refractive index (dRI, blue), and Quasi-elastic light scattering (QELS, pink). Peak 1 produced an estimated average mass of 92 kDa, and peak 2 produced an estimate of 57 kDa. A small light scattering peak is visible at approximately 7 ml, representing aggregates or large particles that elute at the void volume*.

3.6 Thermofluor analysis

To test for correct folding, preservation of thermophillic properties, and to screen for potential interactions with divalent metals, ComZ was subjected to analysis using the SYPRO orange thermal shift assay. This assay detects unfolding through an increase in fluorescence from the dye, and can be used to measure melting temperature (Steinberg *et al* 1996; Lavinder *et al.,* 2009). Firstly, the melting temperature of ComZ was determined by recording fluorescence as the temperature is raised from 25 and 99 °C. The melting temperature was determined by taking the lowest point of the first derivative of the resulting curves (see Fig. 10a), and was shown to be 79.1 °C. This demonstrated that ComZ is relatively thermostable and is likely to be folded correctly. To test the suitability of heating and potential utility of centrifugation as a purification step, 500 µg/mL ComZ was heated to 65 °C for 10 minutes and centrifugation was carried out at 15000 Xg for 10minutes to pellet any large aggregates. ComZ proved resistant to heat denaturation and precipitation at this temperature, and the process appeared to increase purity of the sample (see Fig. 10b).Thermal stability was measured in MES HCl at pH 6.5, Bis-Tris HCl at pH 7.0, and Tris HCl at pH 8.0 and 9.0 at 0, 50, 100, 150 and 200 mM NaCl (Fig. 11a). ComZ was also screened for binding of divalent metal (sulphate) salts at a concentration of 1 mM, and 0 or 100 mM NaCl (see Fig. 11b). A slight increase in melting temperature was observed with higher pH and a slight decrease was observed with rising salt concentrations. The only divalent metal to produce a noticeable effect on the melting temperature of ComZ was zinc, which produced a decrease of 2.2 °C.

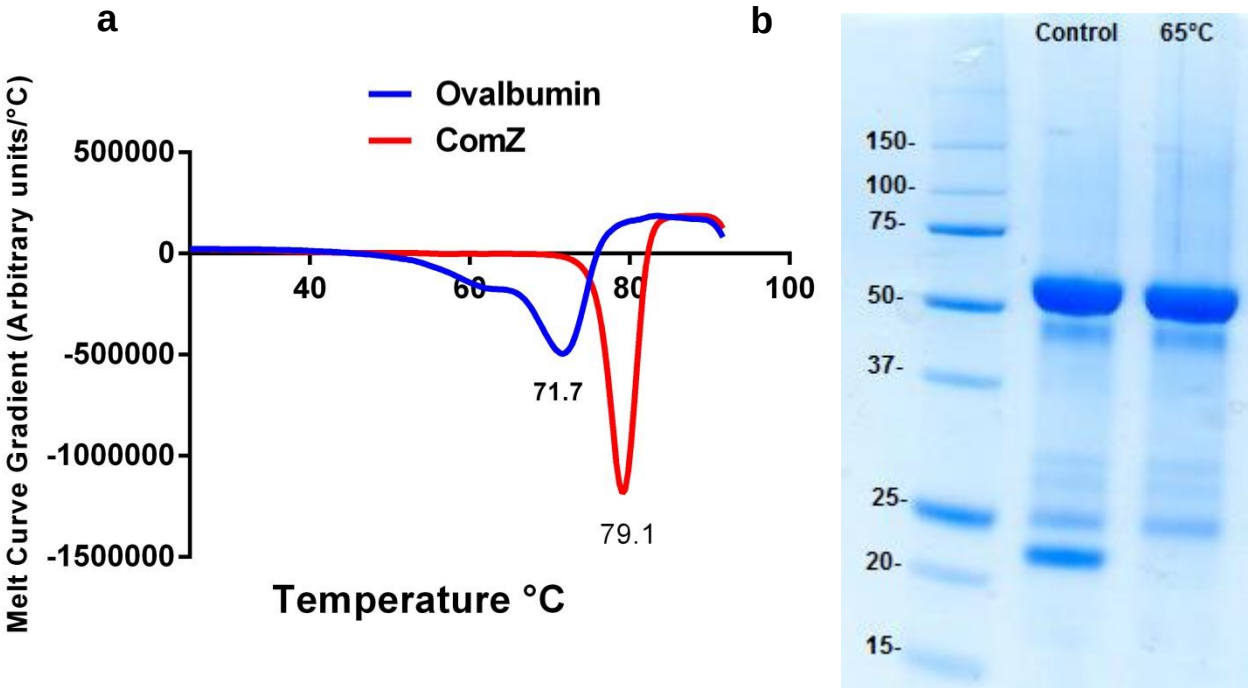


Figure 10: Thermal profile of ComZ and heat denaturation assay

(a) Example of *first derivative of a SYPRO orange fluorescence melt curve showing data from ComZ and ovalbumin control.* *Melting temperatures are shown for each protein. Data points are mean values from 3 separate wells*.

(b) Coomassie stained SDS PAGE gel demonstrating the resistance of ComZ to heat denaturation and precipitation. *An impurity of approximately 22kDa can be removed through heating and centrifugation.*

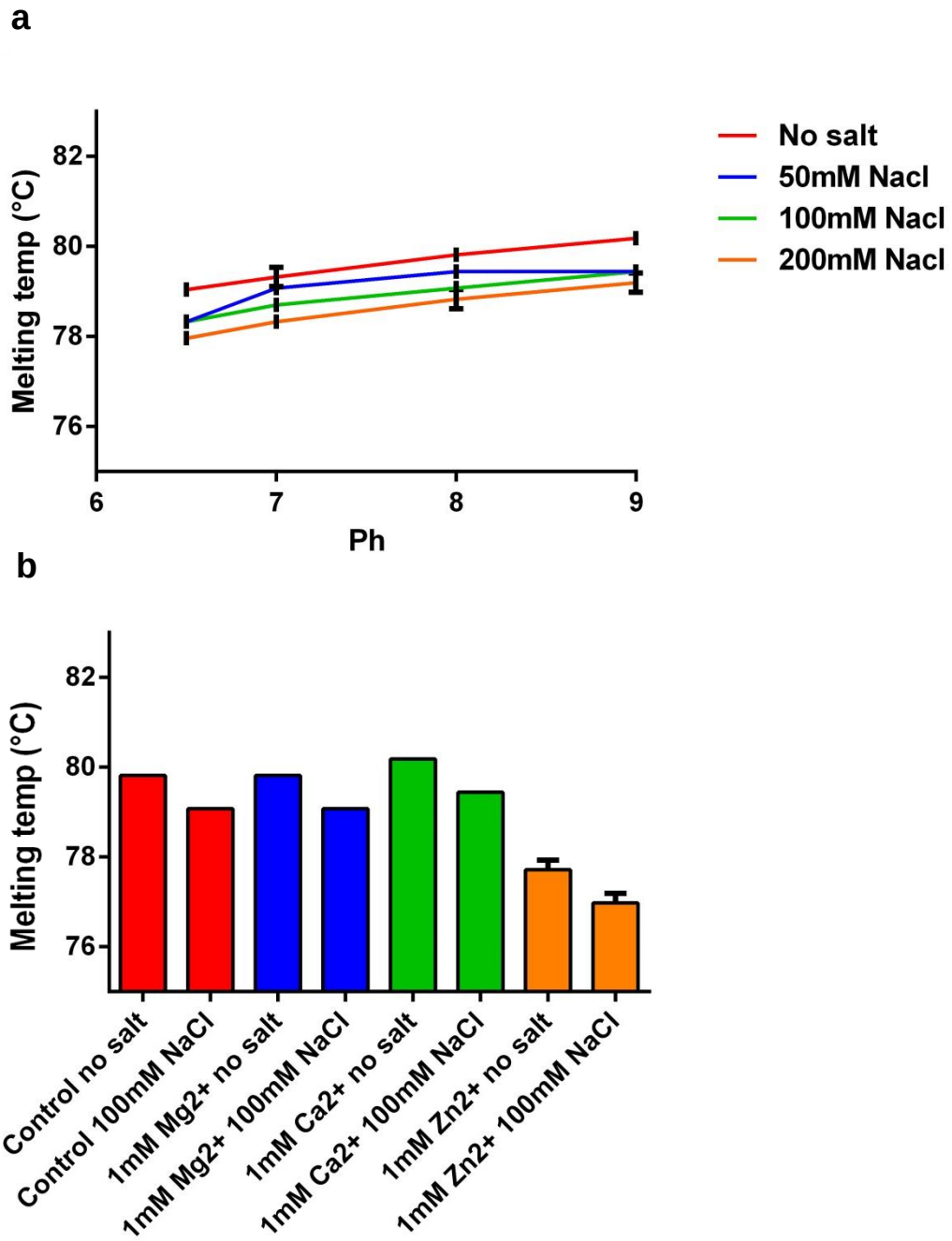


Figure 11: Thermal shift screening of ComZ

(a) Melting temperature of ComZ at varying PH and NaCl concentrations. *Buffers at pH 6.5, 7, 8, and 9 were MES, BisTris HCl, Tris HCl, and Tris HCl respectively*.

(b) Melting temperature of ComZ in the presence of divalent metal ions (sulphate salts) at different concentrations of NaCl. *Error bars are shown only for zinc, as errors in other experiments were lower than the precision limit for detection. A destabilising effect is apparent for zinc, although it is not clear if this is due to a specific interaction with ComZ.*

3.7 Limited Proteolysis

3.7.1 Digestion and Analysis

In order to gain some information on the domain structure of ComZ, samples were subjected to limited proteolysis through enzymatic digestion. Highly folded domains of proteins are often resistant to digestion by enzymes with a low specificity, whereas loop regions that link these domains are more susceptible. This approach can be used to match a domain structure to the sequence of a protein and is also used to remove disordered regions that often inhibit crystallisation (Dale *et al.,* 2003).Initial digestion was carried out using 500 µg/mL ComZ and chymotrypsin at a concentration of between 20 and 500 µg/mL for a period of 1hr (see Fig. 12). Digestion was halted by addition of the SDS PAGE loading buffer and denaturation by heating (as for normal SDS PAGE).

This analysis of digestion fragments showed that even at high concentrations (lanes 4 and 5), ComZ

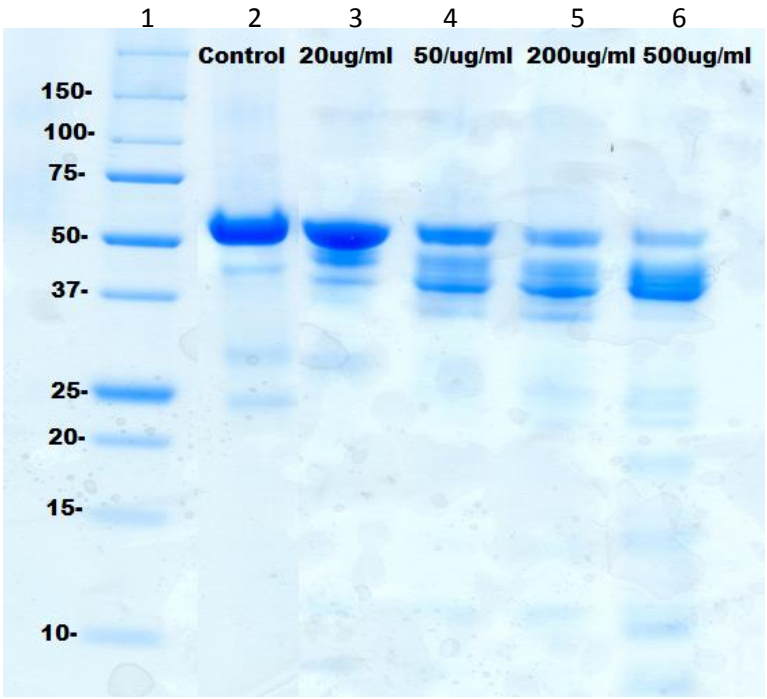


Figure 12: Coomassie stained SDS PAGE gel showing concentration-dependent Proteolytic digestion of ComZ using chymotrypsin.
Concentrations of chymotrypsin used are displayed above lanes 3-6. Lanes have been re-arranged for clarity.

did not form fragments smaller than 40kDa. This suggests the presence of a protease-resistant core region. Because of the structure-orientated nature of the project, it was decided that various proteases would be screened at high concentration. This would allow the identification of highly stable species lacking disordered regions. Removal of flexible regions by proteolysis can improve the crystallisation potential of proteins and has been successfully employed to solve multiple structures (Huber *et al., 1976;* Jurnak *et al.,* 1980). Additionally it was hoped that this approach would also result in the degradation of proteins from *E. coli* that co-purify with ComZ.

Trypsin, chymotrypsin, and papain were used at a concentration of 200 µg/mL for a period of 1, 2, 3, 4 hrs and overnight (o/n). SDS PAGE gels for each set of digestions were run in duplicate to allow for both standard Coomassie blue staining and anti-His western blot analysis (see Fig. 13). All enzymes produced fragments in the region of 43 kDa with the exception of Papain, which appeared to result in the generation of 2 distinct fragments of similar mass. The western blot analysis indicated

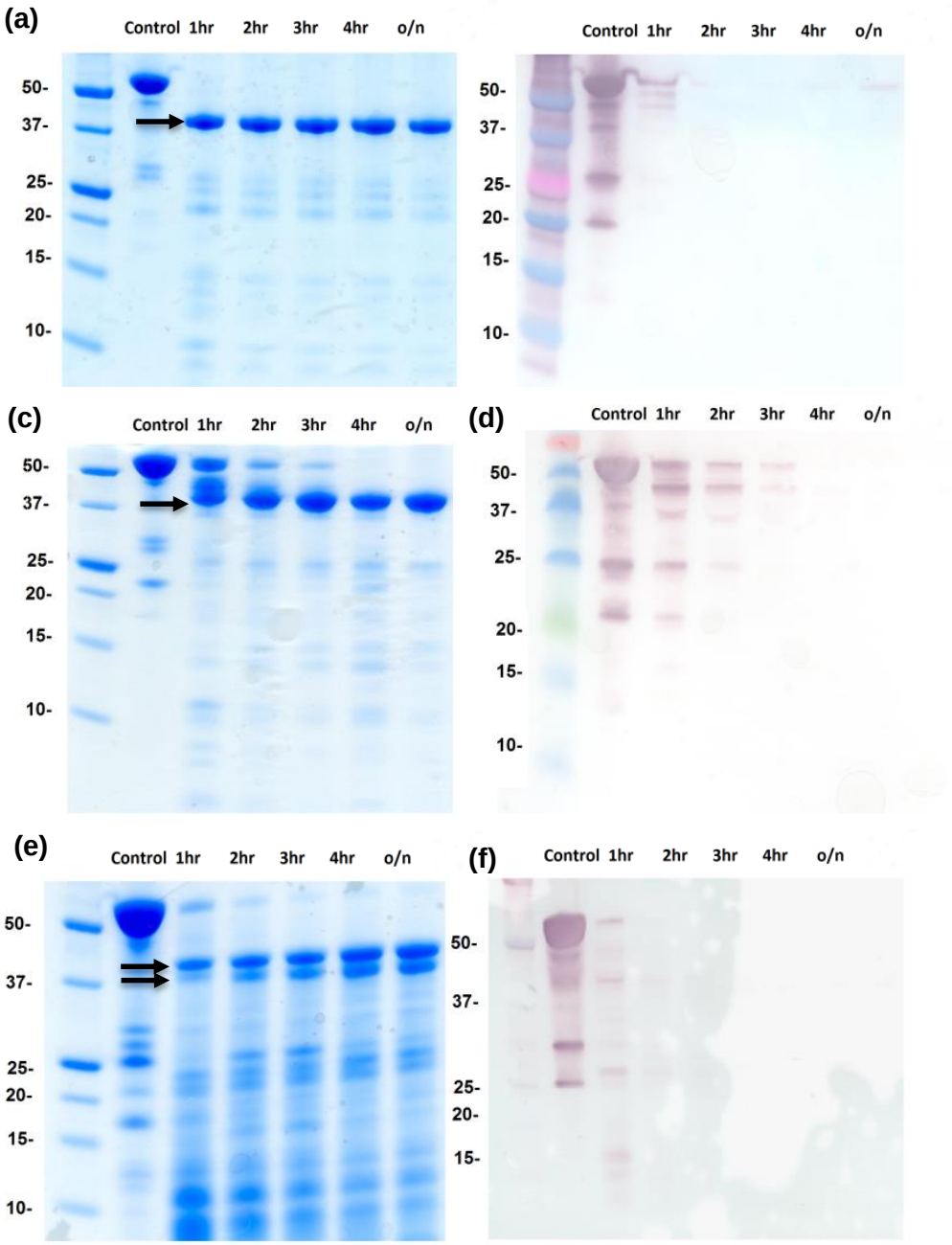
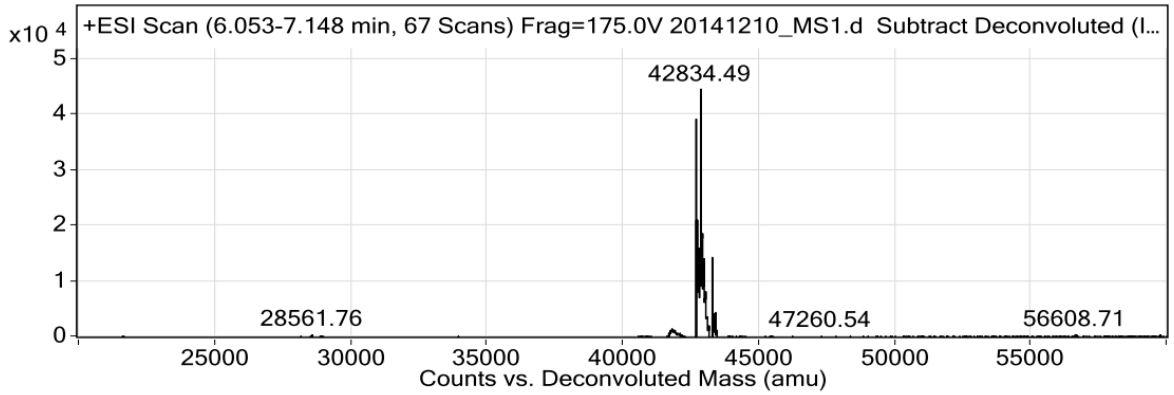


Figure 13: Coomassie stained SDS PAGE gels and Western blot analysis of High concentration Proteolysis of ComZ. ***a*** *and* ***b*** *show the products of digestion at 200 µg/mL trypsin using Coomassie staining and anti-his western blotting respectively,* ***c*** *and* ***d*** *show the equivalent results for chymotrypsin, and* ***e*** *and* ***f*** *show the results from digestion using papain. All digestions yielded fragments of around 40 kDa, with Papain producing two distinct similar sized fragments. All species produced from enzymatic digestion of ComZ appeared to have lost the polyhistidine tag. Fragments are indicated with an arrow.*

that all proteolytic fragments no longer contained the polyhistidine tag, suggesting the loss of either a C terminal region, or C and N terminal regions (Fig. 13b, d, f). Samples of the chymotrypsin and trypsin digestions were subjected to intact mass spectrometry and found to have masses of 42834.49 and 43089.47 Da respectively (Fig. 14). These results indicate the presence of a domain that is resistant to proteolysis. This domain comprises around 75% of the mass of ComZ, but these results cannot be used to confirm if this is the only domain, or if there is a smaller domain that is more susceptible to digestion. The varying population of masses could be result of the different ComZ species shown in Fig. 8 or may be caused by slightly different enzymatic cut sites.

Chymotrypsin

Fragmentor Voltage	Collision Energy	Ionization Mode
175	0	ESI

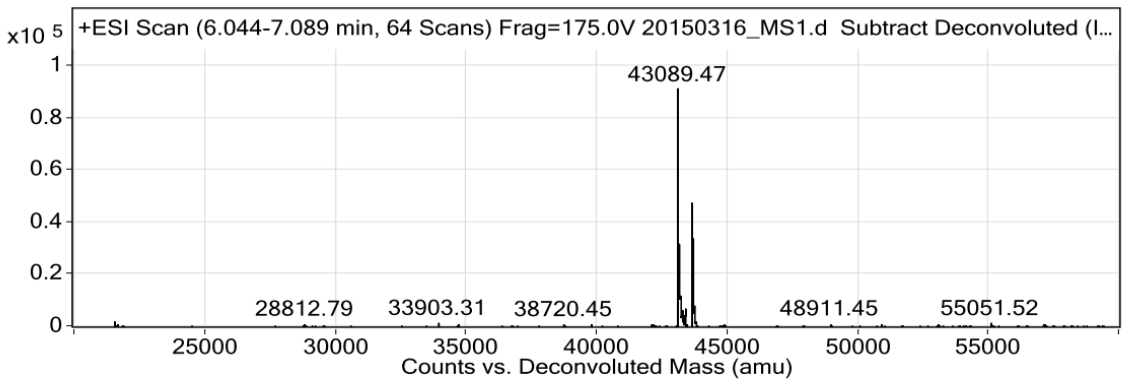


Peak List

m/z	Abund	Score (DB)
42678.49	38617.99	9
42805.37	18048.41	8
42834.49	44599.34	9
42855.08	17644.51	8
42864.77	18692.98	8

Trypsin

Fragmentor Voltage	Collision Energy	Ionization Mode
175	0	ESI



Peak List

m/z	Abund	Score (DB)
43089.47	91193.01	10
43108.89	28970.6	9
43148.43	17954.28	9
43624.98	47285.69	9
43645.37	15799.16	8

Figure 14: Deconvoluted mass spectrograph showing the species present after digestion of ComZ using 200 µg/mL chymotrypsin (upper) and trypsin (lower) for 4 hours. *The 5 most abundant masses are shown in the tables below each spectrograph*

The loss of the C terminal polyhistidine tag during digestion (Fig. 13) allowed for the introduction of a Reverse nickel affinity step to prepare crystallography grade samples of these species for crystallisation trials in place of ion exchange. A simplified version of the metal affinity chromatography protocol was developed featuring the same buffer conditions and 3 washes containing 0, 20 and 40 mM imidazole and 2 elutions containing 500 mM imidazole. These fractions were run through a gravity column as for the earlier protocol. Protein prepared in this way was finally purified using the same size exclusion step as the un-truncated form, peak elution occurred at 64.0 mL representing the slight reduction is mass of the protein. Samples from these steps were analysed using SDS PAGE (see Fig.15).

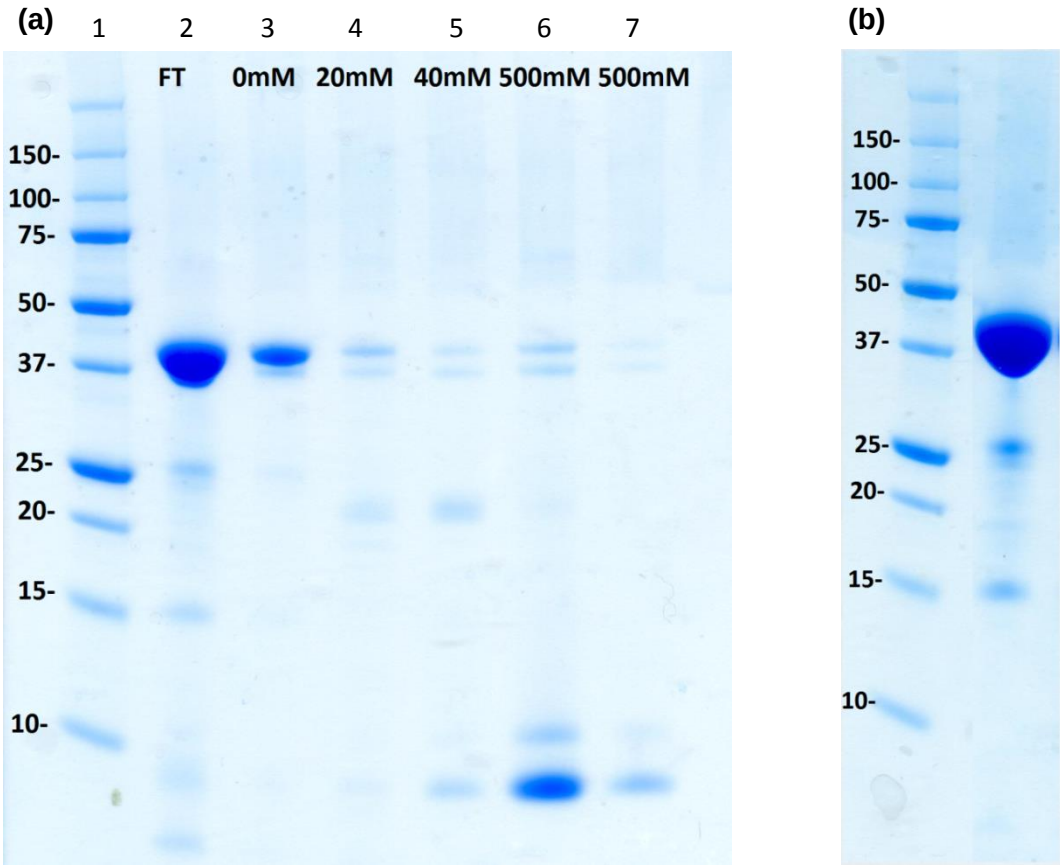


Figure 15: Reverse nickel purification of ComZ fragments

(a) ***Coomassie stained SDS PAGE showing fractions from Reverse nickel purification of the chymotrypsin fragment of ComZ.***

(b) Coomassie stained SDS PAGE showing final product of Chymotrypsin fragment purification. *Size markers are shown to the left of lane 1 in each gel.*

3.8 Crystallography

3.81 Crystallogenesis

Previous unpublished work on ComZ in the Derrick lab had shown that the native form of ComZ crystallised in the space group P 3_1 2 1 in the commercially available (molecular dimensions) condition PACT F4 (Newman *et al.,* 2005). This screen is buffered to pH 6.5 using 100 mM Bis Tris propane and contains 200 mM potassium thiocyanate (KSCN) and 20% w/v PEG 3350 as a precipitant. Previously, this approach had led to the collection of a 2.7 Å native dataset; the aim of the structural element of the project was, therefore, to produce crystals that could be used to obtain experimental phases. Proteolytic fragments from chymotrypsin and trypsin digestion (Fig. 13) were screened for crystallisation using the molecular dimensions screens; PACT, Morpheus, JCSG, and SG1. Screening was carried out using the sitting drop method, initially using a protein concentration of 10 mg/mL in 400 nl drops and later using 2 µl drops at various concentrations. No recognisable crystals were produced from the trypsin or chymotrypsin-treated species in any condition. This suggests that, rather than improving the crystallisation of ComZ, limited proteolysis had prevented the formation of crystals. There are many possible reasons for this. Key factors could be the loss of important residues that interact in the crystal form, or loss of homogeneity through the generation of a population of different fragments via proteolysis at slightly different cut sites. The mass spectrometry results do display a population of different fragments, so this is perhaps the more plausible explanation. By utilising the aforementioned purification techniques, untreated ComZ in native and selenomethionine-labelled form was successfully crystallised. Optimal conditions for the production of the largest possible crystals were determined to be approximately 8 mg/mL with an initial protein to screen ratio of 2:1 (see Fig. 16). Wells containing protein at an initial concentration of 5 or 6 mg/mL typically exhibited large scale nucleation but limited growth, (presumably as precipitant concentration in the drop neared that of the reservoir). Samples with protein introduced at 7,8, or 9 mg/mL displayed limited nucleation but large amounts of growth (per nucleation event), and wells featuring protein at 10, 11, or 12 mg/mL produced high nucleation rates and intermediate growth resulting in large numbers of small and medium sized crystals. Micro-seeding into different conditions from PACT, Morpheus, JCSG and SG2, using crystals grown in F4 was also attempted but was unsuccessful. Due to the fact that only one condition, out of the 480 tested, permitted observable crystallisation, even with seeding, it was decided that it would be useful to explore crystallisation space around this condition. It was hoped that would yield possible methods for co-crystallising ComZ with atoms with accessible wavelengths for single/multi wavelength Anomalous Dispersion (SAD/MAD) or isomorphous replacement (SIR MIR) phasing experiments. Three novel conditions were created by utilising the same buffer and precipitant conditions as in PACT F4 but substituting 200 mM potassium thiocyanate for 200 mM potassium iodide, bromide, or selenocyante (KSeCN). No crystal growth was observed with either bromide or iodide, even when seeding from crystals grown in PACT F4 was employed. Growth was observed with selenocyanate with and without seeding, although the crystals were much smaller. This indicates that the chemical nature of thiocyanate and selenocyanate is important in crystallisation of these proteins. This, in turn suggests

that there might be ordered sites where these ions bind to the protein in the crystalline lattice, making it plausible that crystals grown KSeCN could be used for SAD/MAD experiments in a similar way to selenomethionine labelled protein. Although this is an interesting idea, it is not likely that co-crystallisation of proteins with this highly toxic compound would be seen as an attractive alternative to selenomethionine/selenocysteine labelling, or other well characterised derivatisation compounds except as a last resort. The thiocyanate Ion has been shown to be chaotrophic so it is possible that the binding of the ion allows ComZ to transfer into a conformational or folding state that permits crystallisation. Due to the result of the thermal shift assay that showed a potential interaction with zinc, Crystallisation was also attempted in F4 with the addition of 1 mM zinc sulphate, in the hope that a crystal containing zinc could be used as a SIR/SAD derivative. In this case, no crystals were observed. This added to the evidence that ComZ might interact in some way with zinc but prevented the use of zinc co-crystallisation for creating phasing derivatives.

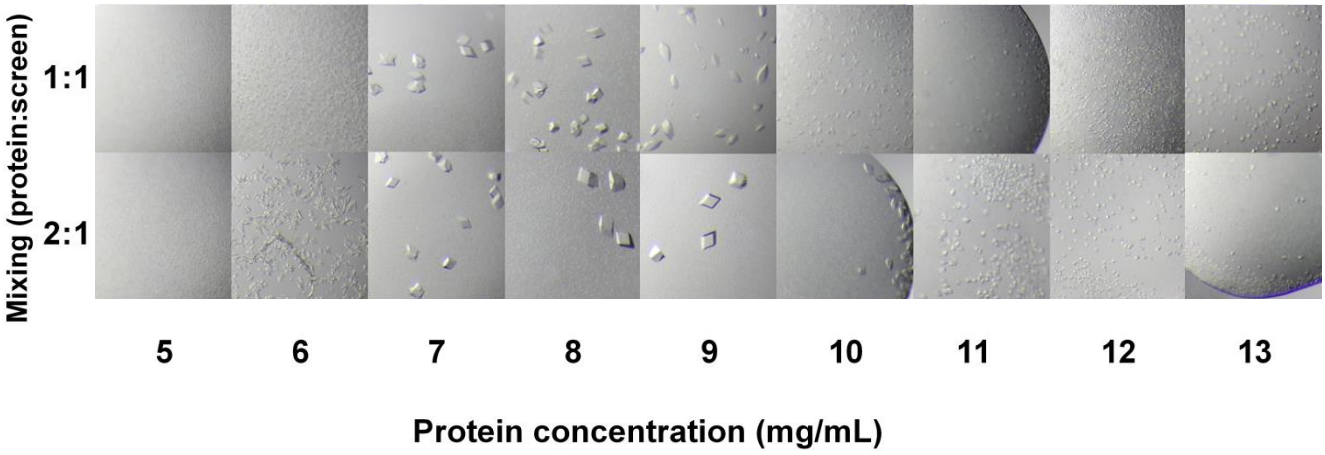


Figure 16: Examples of crystal morphology after growth in PACT F4 at different initial protein concentrations using symmetric and asymmetric mixing. *Images from different conditions have been purposefully selected to illustrate a general trend. All images were taken at the same magnification.*

3.8.2 Derivatisation and data collection

In order to perform experimental phasing experiments, it was necessary to create derivative ComZ crystals. Derivatisation and data collection was carried out by Dr Colin Levy from the Manchester protein structure facility (MPSF). In order to introduce heavy atoms for SAD or SIR/SIRAS experiments Derivatization was performed using soaks of between 2 and 3 hours with 5-10 mM potassium tetrachloroplatinate or platinum thiocyanate and soaks of between 5 and 30minutes with 5-10 mM potassium bromide. Glycerol was added to wells containing derivatives for cryoprotection, and the crystals were then mounted and frozen using liquid nitrogen.

3.8.3 Data analysis

Soaking native crystals of ComZ with potassium tetrachloroplatinate yielded crystals that exhibited significant anomalous signal at a wavelength of 0.97949 Å with a high resolution limit of up to 3.06 Å. Indexing revealed that the unit cell a, b, and c dimensions had increased by 4.19, 4.19 and 4.4 Å respectively (relative to the previously collected native). This observation increased confidence that a "real" derivative had been produced, but also prevented the effective use of single isomorphous replacement (SIR/SIRAS) methods for generation of phase estimates for the native data (Blow and Rossman, 1961). Another problem was posed by the drop-off of anomalous signal after 4.94 Å and rapid increase in R_{merge} values at resolutions approaching the high-resolution limit (see table 1). These issues could be attributed to factors such as suboptimal crystal quality, damage during soaking, and radiation damage during data collection, a phenomenon that disproportionally affects anomalous scattering atoms to such an extent that it can even be used as an experimental phasing tool (Ravelli, *et al.,* 2003; Ramagopal *et al.,* 2005). Fortunately, crystal-to-crystal variability of platinum derivative crystals was low. This observation led to the decision to scale data from multiple crystals into a single high-redundancy dataset. Statistics from this process are shown in table 1. Although R_{merge} values are high, partially due to the high redundancy, the $R_{p.i.m}$ statistics were much lower, especially in the inner shell. This reflects the benefit of high redundancy datasets in overcoming data quality issues. The estimated limit for phasing by SAD was 4.94 Å. The ability to calculate a density map showing secondary structural features was therefore dependent of effective density modification to extend phases to a higher resolution. Crystals of selenomethionine-labelled ComZ diffracted up to 2.6 Å at an X-ray wavelength of 0.92819 Å and yielded datasets with R_{merge} and $R_{p.i.m}$ values of 12.2 and 2.8% respectively with 10 fold multiplicity (table 1). The anomalous correlation was 0.318 indicating that Incorporation of selenomethionine had produced a detectable anomalous signal. The dimensions of the unit cell were not distorted relative to the native dataset. Phasing from this data alone might be problematic, due to the relatively low abundance of methionine in ComZ, but being able to locate the sites could be used to place the methionine residues during later refinement.

3.8.4 Phasing and Density modification

This dataset was then inputted into the Phenix SAD pipeline (HYSS, Phaser and Resolve) in "thorough" mode. Phasing was carried out based on 13 Pt sites and, following density modification, a map at a resolution of 3.5 Å was produced. The figure of merit (FOM) was low (0.322) but not unacceptable for an initial SAD solution. Significantly, hand selection was resolved during phasing with a score (Bayes-CC) of 37.8 for $P3_1\ 2\ 1$ compared to 14.9 for $P3_2\ 2\ 1$ (Terwilliger, *et al.,* 2009). After initial density modification the R factor was 0.387.

Table 1: Crystallographic data quality and merging statistics

(a)Statistics produced from scaling of platinum derivative datasets from 4 crystals (blue). *The program Aimless* was used for scaling *(Evans and Murshudov, 2013; Evans et al., 2011).*

(b) Example data quality statistics from a selenomethionine labelled crystal *the program XDS was used for data processing (Kabsch, et al., 2010).*

a

	Overall	Inner shell	Outer shell
Low resolution limit	109.090	109.09	3.74
High resolution limit	3.50	9.90	3.50
Rmerge (within $I^+$and I^-)	0.388	0.102	2.101
Rmerge (all $I^+$and I^-)	0.395	0.109	2.132
Rmeas (within $I^+$and I^-)	0.394	0.103	2.157
Rmeas (all $I^+$and I^-)	0.399	0.110	2.161
Rpim (within $I^+$and I^-)	0.067	0.017	0.467
Rpim (all $I^+$and I^-)	0.049	0.014	0.336
Rmerge in top intensity Bin	0.093		
Total observation	1625376	78293	156751
Total unique	25782	1267	4568
Mean (I)/sd (I)	13.8	57.0	2.4
Mn (I) half-set correlation CC (1/2)	0.999	0.999	0.856
Completeness (%)	99.9	99.9	99.7
Multiplicity	63.0	61.8	34.3
Anomalous completeness	99.9	99.9	99.7
Anomalous multiplicity	33.1	36.3	17.7
DelAnom correlation between half-sets	0.487	0.871	0.027
Mid-slope of anomalous normal probability	1.194		
Estimate of max resolution for significant anomalous signal	4.94		

b

	Overall	Low	High
Low resolution limit	107.70	107.70	2.65
High resolution limit	2.60	11.63	2.60
Rmerge	0.122	0.030	2.536
Rmeas	0.130	0.034	2.701
Rpim	0.028	0.008	0.629
Completeness %	95.3	92.3	94.4
Multiplicity	20.7	18.4	18.1
Anomalous completeness %	95.3	92.9	94.0
Anomalous multiplicity	10.8	11.2	9.4
Anomalous correlation	0.318	0.648	0.031

3.8.5 Electron density map and Structural model of ComZ

Upon inspection, the electron density map and displayed some discernible features that agreed with predictions from *in silico* secondary structure analysis of the sequence, such as the N terminal alpha helix and predominantly beta-structure. Fig. 17 shows 2 helices on top of a beta sheet region with ten strands. Fig. 18 shows this sheet from below. The alpha helices appear to be N terminal as each features one end with no discernible electron density that could connect it to other regions. This is in agreement with the PSIPRED result shown in Fig. 4 (Jones 1999). Figure 19 shows 2 beta helices. This structural motif appears to be in agreement with the primary structure of the C terminal region which features multiple glycine rich, and twin glycine regions that are likely to represent the loop regions that link parallel beta strands (Bachhawat and Singh, 2007). Model building was performed and the structure was transferred to the native data collected by Karuppiah (unpublished work) set using MOLREP (Vagin and Teplyakov, 1997) to allow for refinement at higher resolution of 2.7 Å. The model presented here is partially refined. Further refinement and model building incorporating the selenomethionine data should eventually yield a fully refined atomic model with a suitably high FOM. The proposed crystal structure of ComZ is that of 2 domains (Fig. 21). Based on gross morphology, the N terminal region displays some similarity to pilin-like proteins like PilA4 (Karuppiah *et al.,* 2013) with one large alpha helix running diagonally across a beta sheet (Fig. 19). In the crystalline form, 2 N terminal regions interact, contributing five beta strands to the large platform-like region (Fig. 17). The N terminal alpha helices sit on top of this structure and give the region an appearance similar to that of the major histocompatibility (MHC) receptor (Bjorkman *et al.,* 1987). The space between the helices contains some currently unmodelled density that may indicate loop regions with high B factors, or a bound ligand (Fig. 17).

The second (C terminal) domain of ComZ is a single beta helix that is positioned, in the crystal, to the side of each beta sheet (see Fig. 21). In the crystal lattice, 2 beta helices extend down onto the platform region (Fig. 17), but are believed to originate from different molecules. Additional regions of density are visible around the main beta helix. Although difficult to model, these areas appear to contain complex chain folding and short alpha-helical motifs that may represent a catalytic centre or binding region (see Fig. 19, 20, 21).

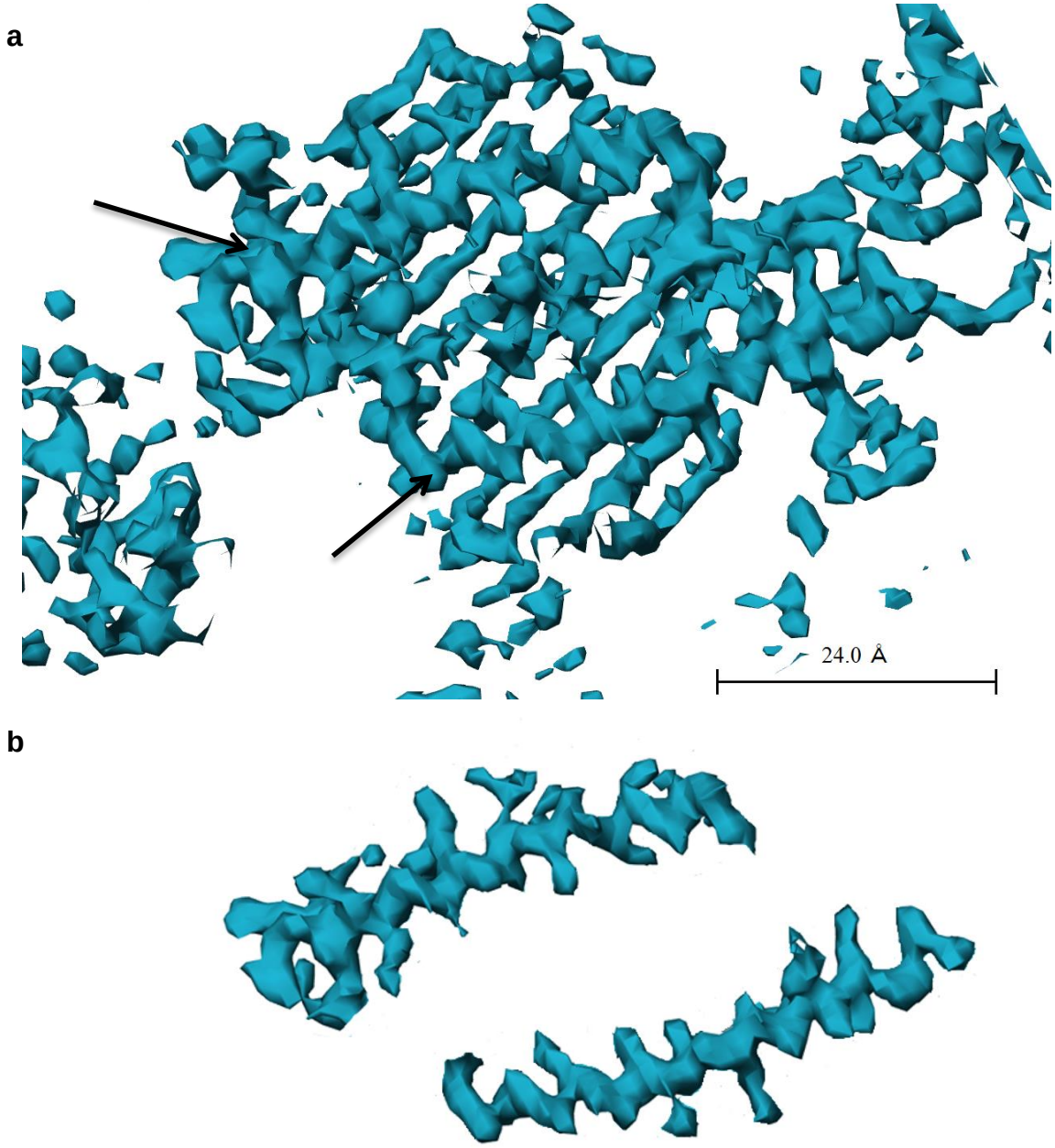


Figure 17: View from above platform-like region showing two N terminal helices.

(a) Full platform region showing groove-like region with unmodelled density. *The beta sheet is visible below. Alpha helices are indicated with arrows. This region is shown "side on" in Fig. 19.*

(b) Electron density of the helices with surrounding features removed. *In this view the main chain of the helix is more clearly visible with protruding side chains*

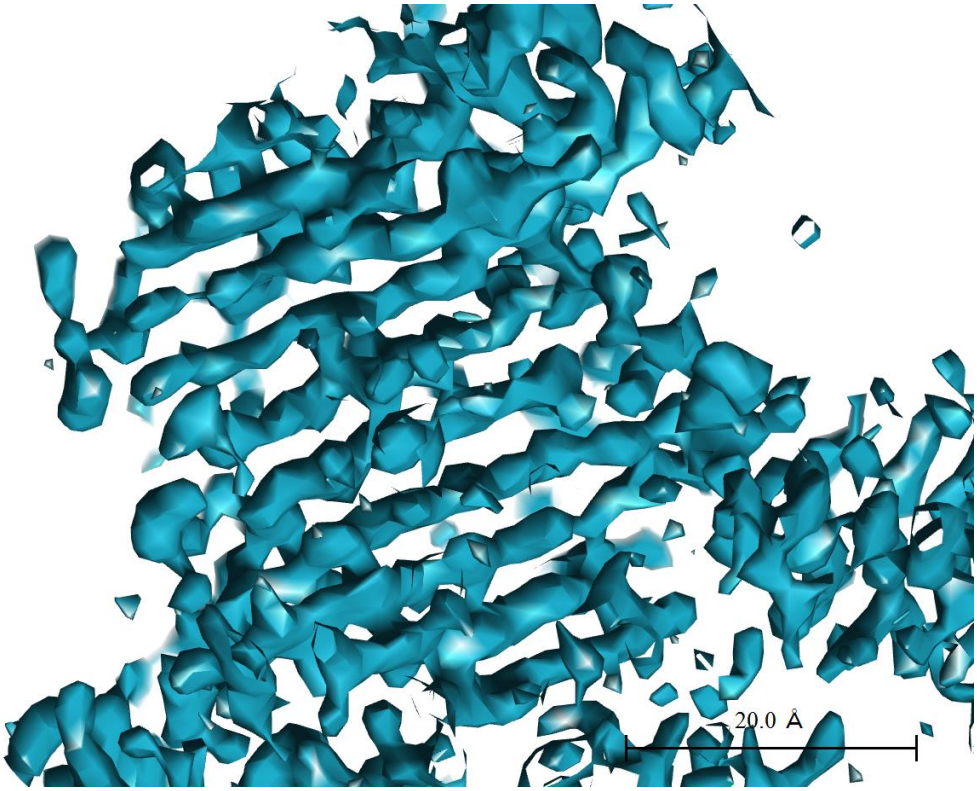


Figure 18*:* View from below ComZ showing 10 stranded beta sheet. *The symmetrical nature of this region indicates that this sheet is formed by the interaction of 5 strands from each molecule.*

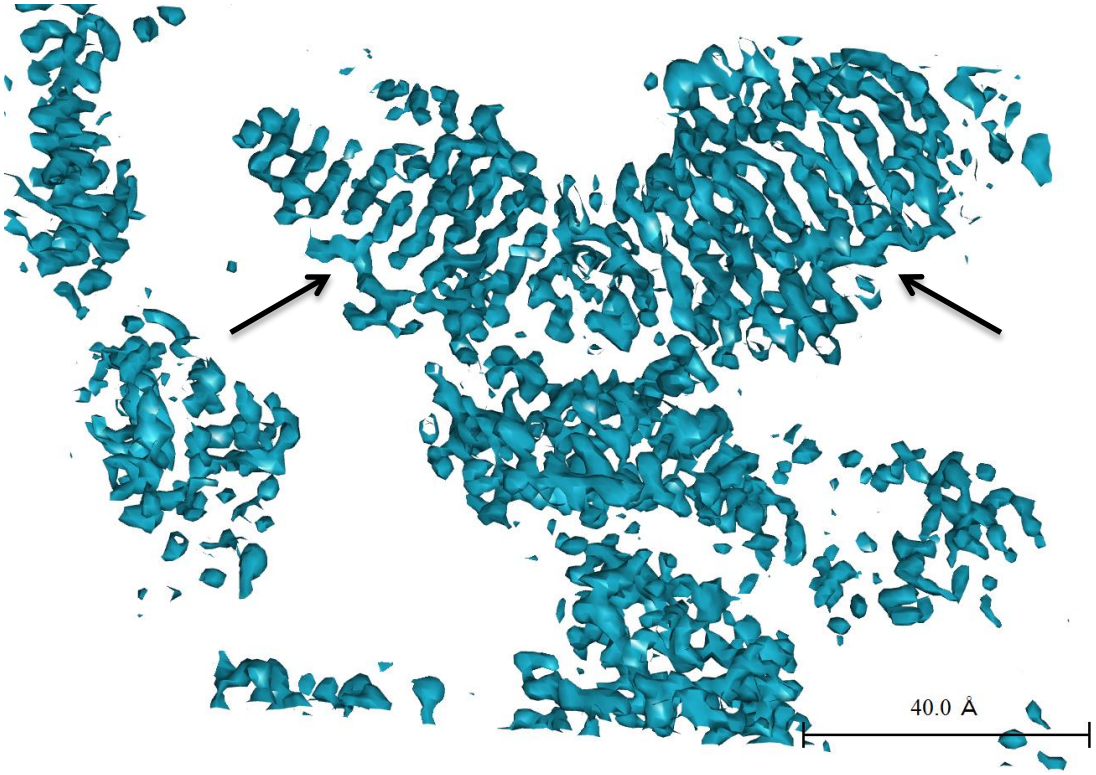


Figure 19*:* "side on" view of the ComZ dimer like platform. *Two beta helices extend down towards the platform but are believed to come from different molecules. Beta helices are indicated using arrows*

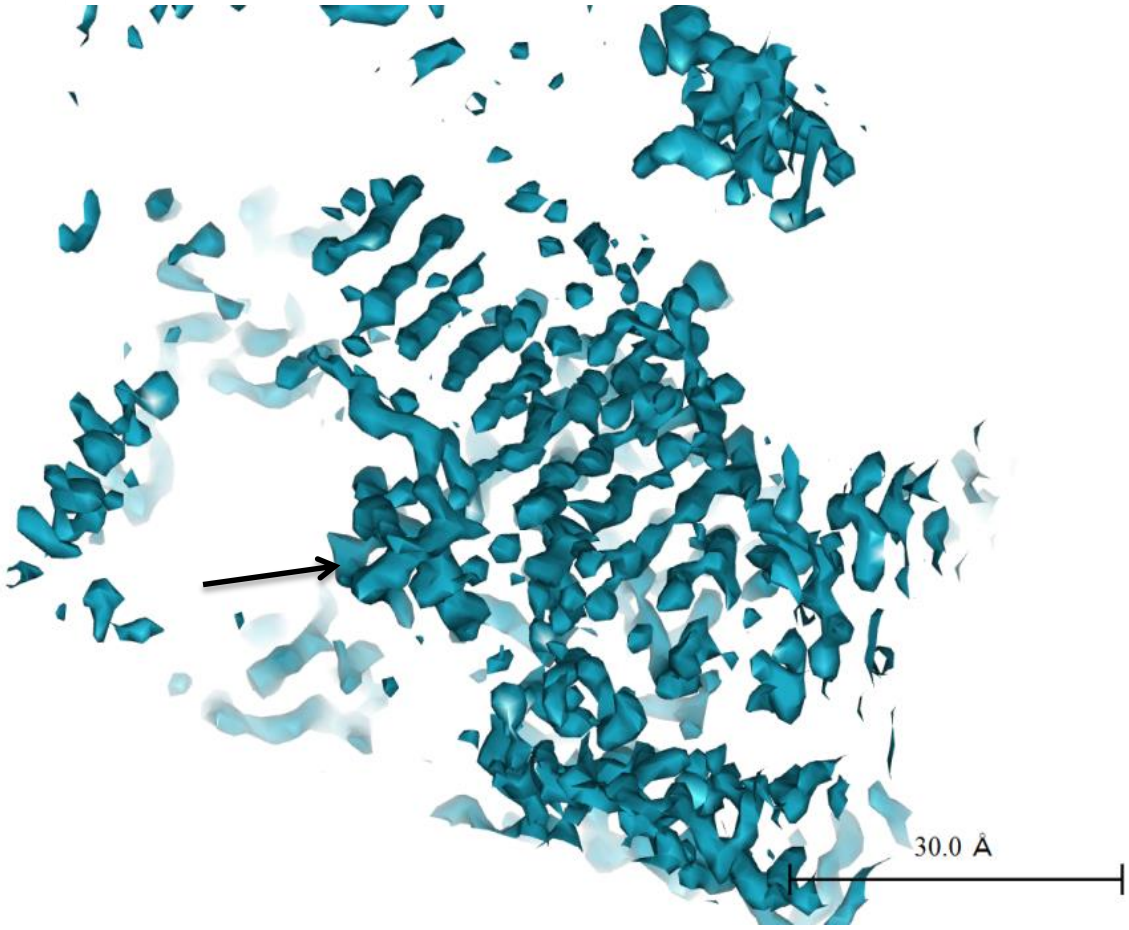


Figure 20*:* beta helix domain showing unknown structural region (left). *This image has been clipped to display the beta strands that make up the back surface of a beta helix. To the left of these elements there is a region of density with unknown structure. This could represent a loop region or the C terminus of the chain. Predicted loop region is indicated by an arrow*

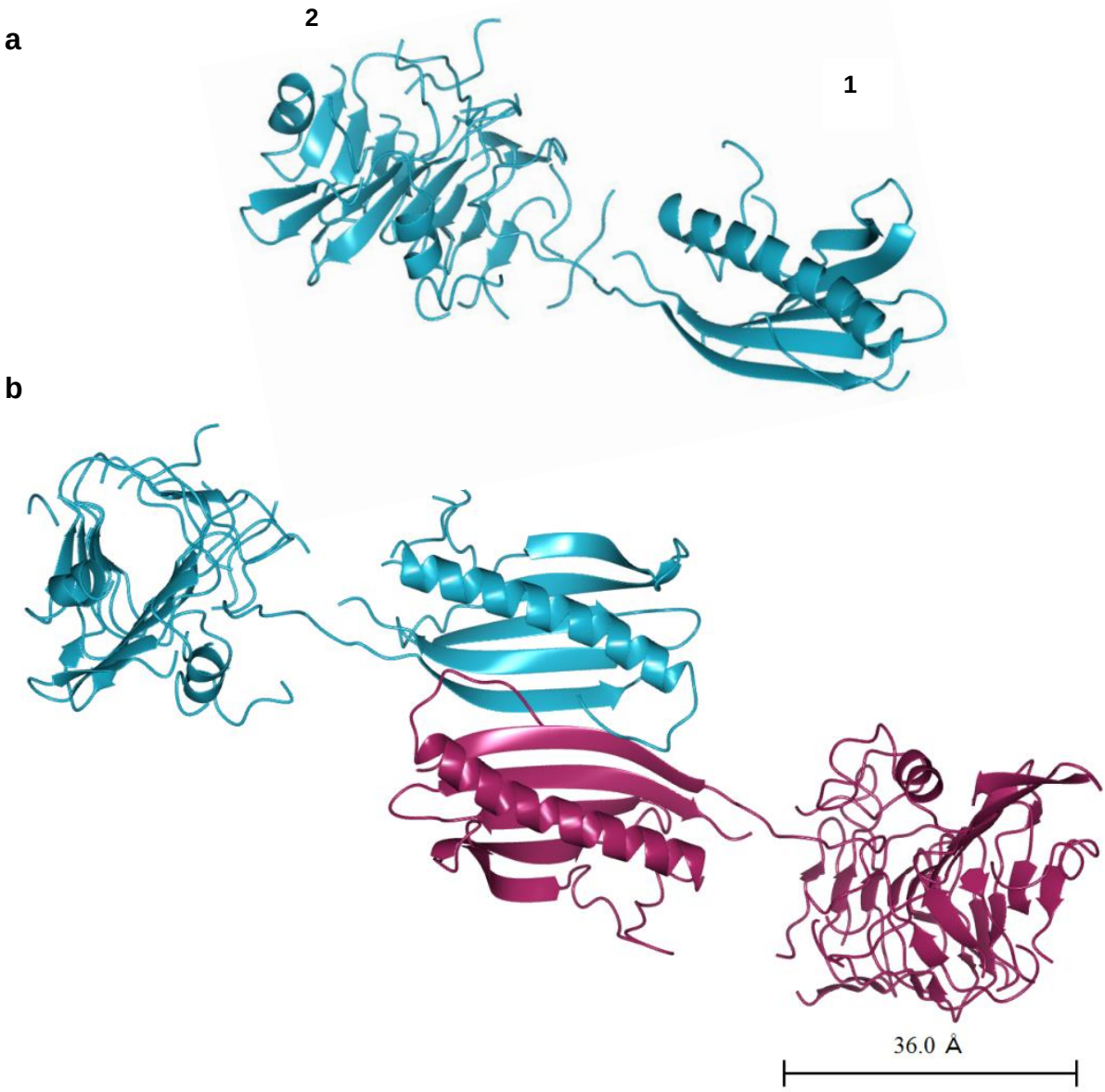


Figure 21*:* Ribbon Diagram showing structural features of ComZ

(a) A single molecule of ComZ. *The alpha helix and beta sheet region at the N terminal (1) is attached via a loop region to the beta helix at the C terminal (2)*

(b) The dimer-like structure present in the crystal lattice. Separate molecules are shown in blue and purple. The 3 parallel beta strands that form the beta helix (2) are also visible. The triangular space is most clear in the left molecule and can be seen in the unbiased density (Fig. 19).

4 Discussion

4.1 Biochemical characterisation of ComZ

Extensive characterisation of ComZ at a biochemical level has yielded information that has proved valuable in producing crystals of a sufficient quality to perform phasing experiments. It has shown that ComZ is able to withstand changes in pH and salinity and this knowledge has led to a fully optimised purification protocol for producing pure homogenous protein. In addition to this, interesting observations that were made during this work may lead to productive areas of experimentation. As expected, ComZ is highly resistant to heat-induced denaturation which not only demonstrates that its folding is maintained by strong intramolecular interactions, but that it folds correctly when produced using an *E. coli* expression system. The melting temperature of ComZ does not appear to be strongly affected by pH but is influenced by zinc. This could suggest that ComZ naturally incorporates a zinc ion, but could also simply be the result of zinc disrupting a polar interaction between residues which might also explain the observed shift in melting temperature. Protease digestion confirms that a large portion of ComZ is tightly folded but crucially, removal of the non-resistant regions does not facilitate easier crystallisation. Exploration of crystal space around the known “hit” condition has revealed an apparent dependence on one family of counter ion as well as revealing the optimum conditions for producing good quality crystals. The absence of crystals in other conditions, coupled to the large scale nucleation and growth at a relatively low concentration, seems to suggest that a unique property of pseudohalogens may be vital for the crystallisation process. This knowledge could prove useful in later-stage work on binding partners or ligands that may require co-crystallisation or soaking experiments.

4.2 Structural model of ComZ

4.2.1 Platform-like domain

This work has given some valuable insight into the structure of this highly unusual protein and will aid future exploration of its function. Due to the high solvent content of the crystal, solvent flattening has been very effective making it is possible to interpret structural features at a resolution and FOM that would typically be very difficult. Perhaps most interesting observation is the apparent similarity of the platform region of ComZ to the MHC receptor which features a similar beta sheet with a groove between twin alpha helices that acts as the binding site for the antigen (Bjorkman *et al.,* 1987). The discovery of convergent evolution giving rise to an MHC-like fold would be an interesting discovery. MHC-like proteins have been identified in viruses, where they are employed to dampen the immune response, but it has been concluded that these are true homologues which have lost sequence homology (Revilleza *et al.,* 2011). It is possible that the similarity between ComZ and these eukaryotic proteins because both bind macromolecules such as DNA or protein. This hypothesis relies heavily on the observed interaction between 2 molecules of ComZ in the crystalline lattice. In studying crystal structures it can be difficult to differentiate between real biological interactions and interactions due to crystal packing. The association between 2 molecules of ComZ

that form the (presumed) dimer structure are far more extensive than any other interactions in the relatively sparse crystal lattice, a fact that has been shown to be a good indicator of a “real” association (Carugo and Argos, 1997). However, a dimeric form of ComZ has not been confirmed during purification by SEC or during DLS. It is possible that this interaction is only observed at concentrations exceeding those reached during these experiments, and/or that a different pH values or chemical conditions are required. This, in turn, could account for the difficulty in identifying additional conditions for crystallisation of ComZ and the failure to crystallise any protease-generated fragments. The 2 counter ions that allowed for crystallisation, thiocyanate and selenocyanate are both Hofmeister chaotropes (Lo Nostro, *et al.,* 2010; Gibb and Gibb, 2011). It is unclear if this is simply a chance observation, or is significant in understanding the packing interactions in the ComZ electron density map. The effect of these agents on macromolecules is complex and it is unlikely that these Ions could cause complete disruption of domain/motif secondary structure in such a thermostable protein. However, because chaotropes are known to promote unfolding and flexibility this observation does add uncertainty to the theory of a dimeric form because the possibility for non-native conformations and interactions is increased.

When ComZ is viewed as a single molecule (Fig. 21a) the N terminal region bears morphological similarity to type IV pilin subunits, which likewise feature an alpha helix at the extreme N terminus running diagonally across a beta sheet (Parge *et al.,* 1995). In *T. thermophilus,* the *comZ* gene is part of an operon containing 4 pilin-like proteins including the major pilin. The structure of this protein was solved by Karuppiah *et al* (2013). It is Plausible that ComZ evolved from a pilin or pilin-like protein and may have a role that involves interaction with other pilins via a mechanism similar to the normal polymerisation of the fibre. This hypothesis is not without its caveats. Firstly, all the genes in this operon display recognisable sequence homology both to each other, and to pilins from other organisms (Freidrich *et al.,* 2002; Freidrich *et al.,* 2003). ComZ has no such homology and furthermore, lacks the cleavage signal for the prepilin peptidase PilD that is believed to allow pilins to leave the inner membrane to join the pilus structure as it polymerises (Freidrich *et al.,* 2003; Karuppiah *et al., 2013*). This enzyme is the only recognisable prepilin peptidase found in *T. thermophilus* (Schwarzenlander *et al.,* 2009). It is possible that the sequence of the N terminal region of ComZ allows for an interaction with pilins at the base of the fibre, and that its sequence has diverged along with its function, whilst other pilins retained conserved sequences due to the need to form a larger repeating structure (sequence divergence would be more deleterious).

4.2.2 Beta Helix domain

The second domain of ComZ is also of great interest. Beta helices were discovered in 1993 in a bacterial pectate lysase (Yoder *et al.,* 1993) and an alkaline protease (Baumann *et al.,* 1993), and represent a significant addition to known folding motifs. These domains are characterised by 2 or more parallel beta strands. The domain present in ComZ appears to belong to the three stranded family, which includes proteins such as the carbonic anhydrase from *Methanosarcina thermophila* (Kisker *et al.,* 1996) and the tailspike protein from *Salmonella typhimurium* phage 22 (Steinbacher *et*

al., 1994). Beta helical proteins with enzymatic function such as the carbonic anhydrase often bind metal ions and feature loop regions that protrude from the main helix (Yoder *et al.,* 1993; Kisker *et al.,*1994; Pickersgill *et al.,* 1994). These features could account for the effect of zinc on ComZ in the thermal shift assay as well as the region of unknown structure shown in Figure 21. However, zinc is known to bind histidine, so it is possible that the thermal analysis result was due to disruption of a pseudo-secondary structural conformation of the polyhistidine tag (Steward *et al.,* 1995). The binding of zinc in the carbonic anhydrase were reliant upon residues from 2 copies of each protein arranged in a homotrimer. No such oligomerisation of the beta helix region of ComZ has yet been observed (Kisker *et al.,* 1996).

The most likely enzymatic roles that can be assigned to ComZ are that of an exonuclease or helicase although there are a wide range of plausible possibilities. It is also possible that ComZ is required for the proper delivery or function of another component but this idea is not supported in the literature. If this was the case, the DNA localisation experiment published by Schwarzenlander *et al* (2009) would have produced similar phenotype for the ComZ mutant and its partner protein. This was not the case. Competency in both gram positive and gram negative bacteria is thought to involve degradation of one strand of duplex DNA to allow for protection and recombination to the chromosome. As mentioned earlier, this exonuclease has been identified in the gram positive pathogen *Streptococcus pneumoniae* (Lacks *et al.,* 1975), and the soil bacterium *Bacillus subtilis* (Levine and Strauss, 1965; Provvedi, *et al.,* 2001). Evidence has also been produced that the *B. subtilus* protein ComFA has a helicase or transport function driven by hydrolysis of ATP (Londono-Vallejo and Dubnau, 1994). Neither of these functions have yet been assigned to a protein in a gram negative organism and it is not understood if they represent fundamental requirements of natural competency, or are organism specific. If ComZ was shown to have either function it would confirm the long-held theory that unwinding and strand degradation are vital steps in the competency pathway. This would be an important discovery, although the impact for wider study of competence might be limited because, as was pointed out by Dubnau (1999), it is highly likely that diverse gram negative bacteria recruit unrelated enzymes that can easily (in evolutionary terms) be co-opted into the competency pathway.

It is not clear from the current map, exactly where the beta helix joins the rest of the molecule. Attempts by Prof Jeremy Derrick (unpublished work) to further refine the model of ComZ to fit the higher resolution native dataset have so-far favoured the conformation shown in Fig. 21. There is an abundance of currently unmodelled density in the groove region, but clearly identifying the position of ever residue in the chain does is not yet possible, perhaps due to high B factors for atoms in this region. Without a complete atomic model with accurate chain-threading, it is impossible to establish the predicted molecular weights of each domain, but it seems plausible that the protease fragments shown in Fig. 13 are produced by cleavage in the disordered region linking the 2 domains, followed by degradation of the alpha helix/beta sheet region. Following on from the hypothesis that the large beta sheet and twin helix domain is formed by the contribution of regions from 2 polypeptides, the beta helix is the only remaining region with a size that can account for the recorded fragment mass

of 42-43kDa. Non covalent protein complexes can remain intact during intact-mass spectrometry, but it does not appear likely that the 42-43kDa fragment is produced from the platform seen in the crystal form. This is due to the monomeric nature of the protein at the low-medium concentrations at which these experiments were performed (as shown by SEC MALLS), and the lack of any multiples of the 57kDa mass in the intact mass spectrogram of the undigested protein. Western blotting demonstrated that the polyhistidine tag at the C terminus was also removed, but this is not unusual as it is not believed to be tightly folded or buried in the protein structure, indeed this would impair initial purification of ComZ during the nickel affinity step.

In conclusion, it appears highly likely that ComZ performs an enzymatic or binding role and is less likely to be purely structural or involved in secretion of other components. It is not currently possible to discard either the dimer/binding groove or pilin-like hypothesis based on merit, however since both ideas point to the beta-helical C terminus as the region containing the core activity, it might be best to concentrate on studying this region in isolation.

4.3 Further work

The electron density map and structure presented here shows only gross morphology of secondary structural elements and is not sufficient to determine the position of all residues. However, transfer to the native data has allowed the determination of chain direction in many regions (Fig. 21). Continued manual model building and multiple cycles of refinement should result in a much more accurate model. Even though the anomalous signal from the selenomethionine data is not strong perhaps due to a low proportion of methionine in the sequence, it should be possible to locate the anomalous scattering selenium atoms. This would allow for positioning of the methionine residues and aid model building. If the strength of the anomalous signal is not high enough, additional steps could be taken to improve the abundance of selenium in the protein such as the dual methionine-cysteine method published by Strub *et al* (2003). Another option is to utilise the selenomethionine data as a derivative for MR SAD, a combination of the molecular replacement and SAD approaches (McCoy *et al.,* 2007). An accurate and complete atomic model would allow for structural homology searches, electrostatic modelling, and docking of specific binding partners in order to guide the search for a function.

This work has led to a refined protocol that exploits experimentally obtained knowledge about ComZ to optimise purity and homogeneity of samples to allow the consistent production of diffracting crystals. However, with such a restricted range of conditions permissible to crystallogenesis, it might prove difficult to perform co-crystallisation studies. Single domains, expressed separately, might exhibit drastically different crystallisation behaviour. Producing the isolated platform region could be very useful for visualising DNA binding that might not be possible in the conformation shown here, which may be a result of a chaotrope-induced distortion of the domain positioning. This would, however, be very challenging, as the platform is thought to be formed by an interaction between 2 chains at a region away from either the C or N terminus. If monomeric forms of the N terminal region could be shown to dimerise, insertional mutagenesis studies could be performed to identify the

crucial interactions. In vitro tests on the full length ComZ protein, or individually expressed domains, could help to demonstrate a function. Electrophoretic mobility shift and exonuclease protection assays can be used to show DNA binding or, if a protein binding partner is theorised, it might be useful to attempt to purify the complex for electron microscopy or further crystallisation studies.

5. References

Assalkhou, R. Balasingham, S. Collins, R. Frye, S. Davidsen, T. Benham, A. Bjoras, M. Derrick, J. Tonjum, T. (2007). The outer membrane secretin PilQ from Neisseria meningitidis binds DNA. *Microbiology*. 153 (1), 1593-1603.

Averhoff, B. (2009). Shuffling genes around in hot environments: the unique DNA transporter of Thermus thermophilus. *FEMS Microbiology Reviews*. 33 (1), 611-629.

Avery, O. MacLeod, C. McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. Induction of transformation by a deoxyribonucleic acid isolated from pneumococcus type II. *Journal of Experimental Medicine*. 79 (2), 137–158.

Bachhawat, N. Singh, B. (2007). Mycobacterial PE_PGRS proteins contain calcium-binding motifs with parallel beta-roll folds. *Genomics Proteomics Bioinformatics*. 5 (3-4), 236-241.

Baumann, U. Wu, S. Flaherty, K. McKay, D. (1993). Three-dimensional structure of the alkaline protease of Pseudomonas aeruginosa: a two-domain protein with a calcium binding parallel beta roll motif. *The EMBO Journal*. 12 (9), 3357-3364.

Biswas, G. Sox, T. Blackman, E. Sparling, P. (1977). Factors affecting genetic transformation of Neisseria gonorrhoeae. *Journal of Bacteriology*. 192 (2), 983-992.

Bjorkman, P. Saper, M. Samraoui B. Bennet, W. Strominger, J. Wiley, D. (1987). Structure of the human class I histocompatibility antigen, HLA-A2. *Nature*. 329 (1), 506 - 512.

Blokesch, M. Schoolnik, G. (2008). The Extracellular Nuclease Dns and Its Role in Natural Transformation of Vibrio cholerae. *Journal of Bacteriology*. 190 (21), 7232-7240.

Blow, D. Rossman, M. (1961). The single isomorphous replacement method. *Acta Crystallographica*. 14 (11), 1195-1202.

Blundell, T. Johnson, L. (1976). *Protein Crystallography*. New York: Academic Press.

Bovre, K. Froholm, L. (1972). Competence in genetic transformation related to colony morphology and fimbriation in three species of Moraxella. *Acta Pathologica Microbiologica Scandinavica Section B Microbiology and Immunology*. 80b (5), 649-659.

Bragg, W. Bragg, W. (1913). The Reflection of X-rays by Crystals. *Proceedings of the Royal Society of London. Series A*. 88 (605), 428-438.

Brussow, H. Canchaya, C. Hardt, W. (2004). Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. *Microbiology and Molecular Microbiology Reviews*. 68 (3), 560-602.

Buchan, D. Minecci, F. Nugent, T. Bryson, K. Jones, D. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research*. 41 (1), 340-348.

Burkhardt, J. Vonck, J. Averhoff, B. (2011). Structure and Function of PilQ, a Secretin of the DNA Transporter from the Thermophilic Bacterium Thermus thermophilus HB27. *Journal of Biological Chemistry*. 286 (1), 9977-9984.

Burley, S. (2000). An overview of structural genomics. *Nature Structural and Molecular Biology*. 7 (1), 932-934.

Carugo, O. Argos, P. (1997). Protein-protein crystal-packing contacts. *Protein Science*. 6 (10), 2261-2263.

Cascales, E. Christie, P. (2003). The versatile bacterial type IV secretion systems. *Nature Reviews Microbiology*. 1 (1), 137-149.

Cascales, E. Christie, P. (2004). Definition of a Bacterial Type IV Secretion Pathway for a DNA Substrate. *Science*. 304 (5674), 1170-1173.

Chan, R. Botstein, D. (1976). Specialized transduction by bacteriophage P22 in Salmonella typhimurium: genetic and physical structure of the transducing genomes and the prophage attachment site. *Genetics.* 83 (3), 433-458.

Chayen, N. (2005). Methods for separating nucleation and growth in protein crystallisation. *Progress in Biophysics and Molecular Biology*. 88 (1), 329-337.

Chayen, N. Saridakis, E. (2008). Protein crystallization: from purified protein to diffraction-quality crystal. *Nature Methods*. 5 (1), 147-153.

Chen, I. Gotschlich, E. (2001). ComE, a Competence Protein from Neisseria gonorrhoeae with DNA-Binding Activity. *Journal of Bacteriology*. 183 (10), 3160-3168.

Choi, J. Keum, K. Lee, S. (2006). Production of recombinant proteins by high cell density culture of Escherichia coli. *Chemical Engineering Science*. 61 (3), 876-885.

Collins, R. Hassan, D. Karuppiah, V. Thistlethwaite, A. Derrick, J. (2013). Structure and mechanism of the PilF DNA transformation ATPase from Thermus thermophilus. *Biochemical Journal*. 450 (2), 417-425.

Cromwell, M. Hilario, E. Jacobson, F. (2006). Protein aggregation and bioprocessing. *The AAPS Journal*. 8 (3), E572-E579.

Dale, G. Oefner, C. D'arcy, A. (2003). The protein as a variable in protein crystallization. *Journal of Structural Biology*. 142 (1), 88–97.

Dauter, Z. (2013). SAD/MAD Phasing. In: Read, R. Urzhumtsev, A. *Advancing Methods for Biomolecular Crystallography*. Dordrecht: Springer. 137-147.

Davidsen, T. Rodland, E. Lagesen, K. Seeberg, E. Rognes, T. Tonjum, T. (2004). Biased distribution of DNA uptake sequences towards genome maintenance genes. *Nucleic Acids Research*. 11 (3), 1050-1058.

Davies, J. (1994). Inactivation of antibiotics and the dissemination of resistance genes. *Science*. 264 (5157), 375-382.

Draskovic, I. Dubnau, D. (2005). Biogenesis of a putative channel protein, ComEC, required for DNA uptake: membrane topology, oligomerization and formation of disulphide bonds. *Molecular Microbiology*. 55 (3), 881-896.

Dubnau. (1999). DNA Uptake in Bacteria. *Annual review of Microbiology*. 53 (1), 217-244.

Emsley, P. Lohkamp, B. Scott, W. Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D Biological Crystallography*. 66 (4), 486-501.

Evans, P. (2006). Scaling and assessment of data quality. *Acta Crystallographica Section D Structural Biology*. 62 (1), 72-82.

Evans, P. (2011). An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D Biological Crystallography*. 67 (1), 282-292.

Evans, P. Murshudov, G. (2013). How good are my data and what is the resolution?. *Acta Crystallographica Section D Biological Crystallography*. 69 (7), 1204-1214.

Fazio, V. Peat, T. Newman J. (2014). A drunken search in crystallization space. *Acta Crystallographica Section F Structural Biology Communications*. 70 (10), 1303-1311.

Fiddis, R. Longman, R. Calvert, P. (1978). Crystal Growth Kinetics of Globular Proteins Lysozyme and Insulin. *Journal of the Chemical Society.* 75 (1), 2753-2761.

Foadi, J. Aller, P. Alguel, Y. Cameron, A. Axford, D. Owen, R. Armour, W. Waterman, D. Iwata, S. Evans, G. (2013). Clustering procedures for the optimal selection of data sets from multiple crystals in macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*. 69 (8), 1617-1632.

Freidrich, A. Prust, C. Hartsch, T. Henne, A. Averhoff, B. (2002). Molecular Analyses of the Natural Transformation Machinery and Identification of Pilus Structures in the Extremely Thermophilic Bacterium Thermus thermophilus Strain HB27. *Applied and Environmental Microbiology*. 68 (2), 745-755.

Friedrich, A. Rumszauer, J. Henne, A. Averhoff, B. (2003). Pilin-Like Proteins in the Extremely Thermophilic Bacterium Thermus thermophilus HB27: Implication in Competence for Natural Transformation and Links to Type IV Pilus Biogenesis. *Applied and Environmental Microbiology.* 69 (7), 3695-36700.

Gasteiger, E. Hoogland, C. Gattiker, A. Duvand, S. Wilkins, M. Appel, R. Bairoch, A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server*. New Jersey: Humana Press. 571-607.

Gibb, C. Gibb, B. (2011). Anion Binding to Hydrophobic Concavity Is Central to the Salting-in Effects of Hofmeister Chaotropes. *Journal of the American Chemical Society*. 133 (19), 7344-7347.

Gorrec, F. (2009). The MORPHEUS protein crystallization screen. *Journal of Applied Crystallography*. 42 (6), 1035-1042.

Gounder, K. Brzuszkiewicz, E. Liesegang, H. Wollherr, A. Daniel, R. Gottschalk, G. Reva, O. Kumwenda, B. Srivastava, M. Bricio, C. Berenguer, J. Van Heerden, E. Litthauer, D. (2011). Sequence of the hyperplastic genome of the naturally competent Thermus scotoductus SA-01. *BMC Genomics*. 12 (577), doi:10.1186/1471-2164-12-577.

Green, D. Ingram, V. Perutz, M. (1954). The Structure of Haemoglobin. IV. Sign Determination by the Isomorphous Replacement Method. *Proceedings of the Royal Society of London A*. 225 (1162), 287-307.

Griffith, F. (1928). The Significance of Pneumococcal Types. *Journal of Hygiene*. 27 (02), 113-159.

Grosse-Kunstleve R. Adams, P. (2003). Substructure search procedures for macromolecular structures. *Acta Crystallographica Section D Structural Biology*. 59 (11), 1966-1973.

Harker, D. (1956). The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. *Acta Crystallographica.* 9 (1), 1-9.

Hengen, P. (1995). Purification of His-Tag fusion proteins from Escherichia coli. *Trends in Biochemical Science*. 20 (7), 285–286.

Hidaka, Y. Hasegawa, M. Nakahara, T. Hoshino, T. (1994). The Entire Population of Thermus thermophilus Cells Is Always Competent at Any Growth Phase. *Bioscience, Biotechnology, and Biochemistry*. 58 (7), 1338-1339.

Huber, R. Deisenhoffer, J. Colman, P. Matsushima, M. (1976). Crystallographic structure studies of an IgG molecule and an Fc fragment. *Nature*. 264 (1), 415-420.

Jancarik, J. Kim, S. (1991). Sparse matrix sampling: a screening method for crystallization of proteins. *Journal of Applied Crystallography*. 24 (1), 409-411.

Jones. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*. 292 (1), 195-202.

Jurnak, F. Mcpherson, A. Wang, A. Rich, A. (1980). Biochemical and structural studies of the tetragonal crystalline modification of the Escherichia coli elongation factor Tu. *Journal of Biological Chemistry*. 255 (1), 6751-6757.

Kabsch W. (2010). XDS. *Acta Crystallographica Section D Biological Crystallography*. 66 (2), 125-132.

Kartha, G. Ramachandran, G. (1955). Applications of the difference-Patterson technique in structure analysis. *Acta Crystallographica*. 8 (4), 195-199.

Karuppiah, V. Collins, R. Thistlethwaite, A. Gao, Y. Derrick, J. (2013). Structure and assembly of an inner membrane platform for initiation of type IV pilus biogenesis. *Proceedings of the National Academy of Sciences of the United States of America.* 110 (48), E4638-E4647.

Kisker, C. Schindelin, H. Alber, B. Ferry, J. Rees, D. (1996). A left-hand beta-helix revealed by the crystal structure of a carbonic anhydrase from the archaeon Methanosarcina thermophila. *The EMBO Journal*. 15 (10), 2323-2330.

Krogh, A. Larsson, B. Von Heijne, G. Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*. 305 (3), 567-580.

Lacks, S. Greenberg, B. Neuberger, M. (1975). Identification of a deoxyribonuclease implicated in genetic transformation of Diplococcus pneumoniae. *Journal of Bacteriology*. 123 (1), 222-232.

Lavinder, J. Hari S. Sullivan, B. Magliery, T. (2009). High-Throughput Thermal Scanning: A General, Rapid Dye-Binding Thermal Shift Screen for Protein Engineering. *Journal of the American Chemical Society*. 131 (11), 3794–3795.

Leahy, D. Hendrickson, W. Aukhil, I. Erickson, H. (1992). Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science*. 258 (5084), 987-91.

Levine, J. Strauss, N. (1965). Lag Period Characterizing the Entry of Transforming Deoxyribonucleic Acid into Bacillus subtilis. *Journal of Bacteriology*. 89 (2), 281-287.

Londono-Vallejo, J. Dubnau, D. (1994). Mutation of the putative nucleotide binding site of the Bacillus subtilis membrane protein ComFA abolishes the uptake of DNA during transformation. *Journal of Bacteriology*. 176 (15), 4642-4645.

Lo Nostro, P. Peruzzi, N. Severi, M. Ninham, B. Baglioni, P. (2010). Asymmetric Partitioning of Anions in Lysozyme Dispersions. *Journal of the American Chemical Society*. 132 (18), 6571-6577.

Lu, G. (1999). FINDNCS: a program to detect non-crystallographic symmetries in protein crystals from heavy-atom sites. *Journal of Applied Crystallography*. 32 (2), 365-368.

McCoy, A. Grosse-Kunstleve, R. Adams, P. Winn, M. Storoni, L. Read, R. (2007). Phaser crystallographic software. *Journal of Applied Crystallography*. 40 (4), 658-674.

McCoy, A. Storoni, L. Read, R. (2004). Simple algorithm for a maximum-likelihood SAD function. *Acta Crystallographica Section D Biological Crystallography*. 60 (7), 1220-1228.

McNicholas, S. Potterton, E. Wilson, K. Noble, M. (2011). Presenting your structures: the CCP4mg molecular-graphics software. *Acta Crystallographica Section D Biological Crystallography*. 67 (4), 386-394.

Mortier-Barriere, I. velten, M. Dupaigne, P. Mirouze, N. Pietrement, O. McGovern, S. Fichant, G. Martin, B. Noirot, P. Polard, P. Claverys, J. (2007). A Key Presynaptic Role in Transformation for a Widespread Bacterial Protein: DprA Conveys Incoming ssDNA to RecA. *Cell*. 130 (5), 824-836.

Murshudov, G. Vagin, A. Dodson, E. (1997). Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallographica Section D Biological Crystallography*. 53 (3), 240-255.

Newman, J. Egan, D. Walter, T. Meged, R. Berry, I. Jelloul, M. Sussman, J. Stuart, D. Perrakis, A. (2005). Towards rationalization of crystallization screening for small- to medium-sized academic laboratories: the PACT/JCSG+ strategy. *Acta Crystallographica Section D Biological Crystallography*. 61 (10), 1426-1431.

Omelchenko, M. Wolf, Y. Gaidamakova, E. Matrosova, V. Valisenko, A. Zhai, M. Daly, M. Koonin, E. Makarova, K. (2005). Comparative genomics of Thermus thermophilus and Deinococcus radiodurans: divergent routes of adaptation to thermophily and radiation resistance. *BMC Evolutionary Biology*. 5 (57), doi:10.1186/1471-2148-5-57.

Parge, H. Forest, K. Hickey, M. Christensen, D. Getzoff, E. Tainer, J. (1995). Structure of the fibre-forming protein pilin at 2.6A resolution. *Nature*. 378 (1), 32-38.

Perutz, M. (1956). Isomorphous replacement and phase determination in non-centrosymmetric space groups. *Acta Crystallographica*. 9 (11), 867-873.

Pickersgill, R. Jenkins, J. Harris, G. Nasser, W. Robert-baudouy, J. (1994). The structure of Bacillus subtilis pectate lyase in complex with calcium. *Nature Structural Biology.* 1 (10), 717-723.

Provvedi, R. Chen, I. Dubnau, D. (2001). NucA is required for DNA cleavage during transformation of Bacillus subtilis. *Molecular Microbiology*. 40 (3), 634-644.

Provvedi, R. Dubnau, D. (2002). ComEA is a DNA receptor for transformation of competent Bacillus subtilis. *Molecular Microbiology*. 31 (1), 271-280.

Prudhomme, M. Attaiech, L. Sanchez, G. Martin, B. Claverys, J. (2006). Antibiotic Stress Induces Genetic Transformability in the Human Pathogen Streptococcus pneumoniae. *Science*. 313 (5783), 89-92.

Ramachandran, G. Raman, S. (1956). A New Method for the Structure Analysis of Non-Centrosymmetric Crystals. *Current Science*. 25 (11), 348-351.

Ramagopal, U. Dauter, Z. Thirumuruhan, R. Fedorov, E. Almo, S. (2005). Radiation-induced site-specific damage of mercury derivatives: phasing and implications. *Acta Crystallographica Section D Biological Crystallography*. 61 (9), 1289-1298.

Ramakrishnan, V. Biou, V. (1996). Treatment of multiwavelength anomalous diffraction data as a special case of multiple isomorphous replacement. *Methods in Enzymology*. 276 (1), 538-557.

Ravelli, R. schroder Leiros, H. Pan, B. Caffrey, M. McSweeney, S. (2003). Specific Radiation Damage Can Be Used to Solve Macromolecular Crystal Structures. *Structure*. 11 (2), 217–224.

Redfield, R. (1993). Genes for Breakfast: The Have-Your-Cake and-Eat-It-Too of Bacterial Transformation. *Journal of Heredity*. 84 (5), 400-404.

Revilleza, M. Wang, R. Mans, J. Hong, M. Natarajan, K. Margulies, D. (2011). How the Virus Outsmarts the Host: Function and Structure of Cytomegalovirus MHC-I-Like Molecules in the Evasion of Natural Killer Cell Surveillance. *Journal of Biomedicine and Biotechnology*. 2011 (1), 1-12.

Schwarzenlander, C. Averhoff, B. (2006). Characterization of DNA transport in the thermophilic bacterium Thermus thermophilus HB27. *FEBS Journal*. 273 (18), 4210–4218.

Schwarzenlander, C. Haase, W. Averhoff, B. (2009). The role of single subunits of the DNA transport machinery of Thermus thermophilus HB27 in DNA binding and transport. *Environmental Microbiology*. 11 (4), 801-808.

Schmidt, T. Skerra, A. (2007). The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nature Protocols*. 2 (1), 1528 - 1535.

Slabinski L. Jaroszewski, L. Rychlewski, L. Wilson I. Lesley, S. Godzik, A. (2007). XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics*. 23 (24), 3403-3405.

SonnHammer, E. Von Heije, G. Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*. 1 (1), 175-182.

Steinbacher, S. Seckler, R. Miller, S. Steipe, B. Huber, R. Reinemer, P. (1994). Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer. *Science*. 265 (5170), 383-386.

Steinberg, T. Jones, L. Haugland, R. Singer, V. (1996). SYPRO Orange and SYPRO Red Protein Gel Stains: One-Step Fluorescent Staining of Denaturing Gels for Detection of Nanogram Levels of Protein. *Analytical Biochemistry*. 239 (2), 223–237.

Sternberg, N. Maurer, R. (1991). Bacteriophage mediated Generalised transduction in Escherichia coli and Salmonella typhimurium. *Methods in Enzymology*. 204 (1), 18-43.

Steward, M. Samson, A. Errington, W. Emmersom, P. (1995). The Newcastle disease virus V protein binds zinc. *Archives of Virology*. 140 (7), 1321-1328.

Strub, M. Hoh, F. Sanchez, J. Strub, J. Bock, A. Aumelas, A. Dumas, C . (2003). Selenomethionine and Selenocysteine Double Labeling Strategy for Crystallographic Phasing. *Structure.* 11 (11), 1359-1367.

Taylor, G. (2010). Introduction to Phasing. *Acta Crystallographica Section D Structural Biology*. 66 (4), 325-338.

Terwilliger, T. (1999). Reciprocal-space solvent flattening. *Acta Crystallographica Section D Biological Crystallography*. 55 (11), 1863-1871.

Terwilliger, T. (2000). Maximum Likelyhood Density modification. *Acta Crystallographica Section D Biological Crystallography*. 56 (8), 965-972.

Terwilliger, T. (2004). SOLVE and RESOLVE: automated structure solution, density modification and model building. *Journal of Synchotron Radiation*. 11 (1), 49-52.

Terwilliger, T. Adams, P. Read, R. McCoy, A. Moriarty, N. Grosse-Kunstleve, R. Afonine, P. Zwart, P. Hung, L. (2009). Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallographica Section D Biological Crystallography*. 65 (6), 582-601.

Tomoeda, M. Inuzuka, M. Date, T. (1976). Bacterial Sex Pili. *Progress in Biophysics and Molecular* Biology. 30 (1), 23-36.

Uson, I. Sheldrick, G. (1999). Advances in direct methods for protein crystallography. *Current Opinion in Sturctural Biology*. 9 (5), 643-648.

Vagin, A. Teplyakov, A. (1997). MOLREP: an Automated Program for Molecular Replacement. *Journal of Applied Crystallograpy*. 30 (1), 1022-1025.

Walter, T. Mancini, E. Kadlec, J. Graham, S. assenberg, R. Sainsbury, S. Owens, R. Stuart, D. Grimers, J. Harlos, K. (2008). Semi-automated microseeding of nanolitre crystallization experiments. *Acta Crystallographica Section F Structural Biology and Crystallization Communications*. 64 (1), 14-18.

Waters, V. Guiney, D. (1993). Processes at the nick region link conjugation, T-DNA transfer and rolling circle replication. *Molecular Microbiology*. 9 (6), 1123-1130.

Weiss, M. (2001). Global indicators of X-ray data quality. *Journal of Applied Crystallography*. 34 (2), 130-135.

Wen, J. Arakawa, T.Philo, J. (1996). Size-Exclusion Chromatography with On-Line Light-Scattering, Absorbance, and Refractive Index Detectors for Studying Proteins and Their Interactions. *Analytical Biochemistry*. 240 (2), 155-166.

Winn, M. Ballard, C. Cotan, K. Dodson, E. Emsley, P. Evans, P. Keegan, R. Krissinel, E. Leslie, A. McCoy, A. McNicholas, S. Murshudov, G. Pannu, N. Potterton, E. Powell, H. Read, R. Vagin, A. Wilson, K. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D Structural Biology*. 67 (4), 235-242.

Yoder, M. Keen, N. Jurnak, F. (1993). New domain motif: the structure of pectate lyase C, a secreted plant virulence factor. *Science*. 260 (1), 1503-1507.